

Run Run Shaw Library

香港城市大學
City University of Hong Kong

Copyright Warning

Use of this thesis/dissertation/project is for the purpose of private study or scholarly research only. ***Users must comply with the Copyright Ordinance.***

Anyone who consults this thesis/dissertation/project is understood to recognise that its copyright rests with its author and that no part of it may be reproduced without the author's prior written consent.

CITY UNIVERSITY OF HONG KONG
香港城市大學

Estimation of Linkage Disequilibrium from High-Throughput Sequencing Data: Applications to Evolutionary Inference of Viruses

高通量定序資料之連鎖不平衡估計：病毒演化推論之應用

Submitted to
Department of Electrical Engineering
電機工程系
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
哲學博士學位

by

Abdullah Syed Muhammad Umer

September 2026
二零二六年九月

Copyright

Abdullah Syed Muhammad Umer, 2026

All rights reserved.

Declaration

I declare that I have worked on my Ph.D. thesis titled "Estimation of Linkage Disequilibrium from High-Throughput Sequencing Data: Applications to Evolutionary Inference of Viruses" by myself and I have used only the sources mentioned at the end of the thesis. As the sole author of the thesis, I declare that the thesis does not infringe any copyrights, any help received in preparing this thesis have been acknowledged in this thesis.

I declare that the substance of this thesis has not already been submitted for any degree and is not currently being submitted for another degree or qualification.

I authorize the City University of Hong Kong to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize the City University of Hong Kong to reproduce the thesis by photocopy or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

In Hong Kong on September 10, 2026

_____

Abstract

Viral pathogens have been responsible for millions of deaths and continue to burden healthcare systems globally. The development of effective antiviral therapies requires accurate characterization of these pathogens. Phenotypic changes in pathogens need to be studied in the context of genomic evolution, a process driven by fundamental evolutionary parameters such as selection, mutation, recombination, structural variation, and others. Accurate inference of evolutionary parameters is of paramount importance for the characterization of diseases, development of targeted therapies, and understanding pathogen dynamics. Many viral pathogens lack proofreading mechanisms that can correct mutations during replication, which leads to the accumulation of multiple mutations during evolution. While most mutations are deleterious, some confer evolutionary advantages such as enhanced replication, increased transmissibility, or improved immune evasion. Identification of these beneficial mutations is a crucial step in developing antiviral therapies; however, the presence of multiple simultaneous mutations complicates the inference problem. Viruses can acquire multiple neutral mutations before a beneficial mutation appears, and all mutations occurring on the same genomic background with a beneficial mutation increase in prevalence with time, making the inference of the actual beneficial mutation difficult. Mutations occurring on the same genomic background are in genetic linkage, and the non-random association of such mutations is quantified by linkage disequilibrium (LD). Failure to account for LD can confound fitness inference, as neutral mutations may appear beneficial simply because of co-occurrence in a selective sweep.

Viral populations are typically studied via high-throughput sequencing (HTS), characterized into next generation sequencing (NGS) and third generation sequencing (TGS). While HTS can sample individual mutations millions of times, it is limited in the extent of LD it can observe. This is because the read-length, i.e. the number of concurrent nucleotides that can be observed together, is limited in HTS. The read-length varies from a few hundred base pairs (bp) in NGS to a few thousand bp in TGS. Because read lengths are often significantly shorter than the complete genomes, HTS cannot directly observe long-range inter-read LD. In this thesis, we address the problem of LD estimation from HTS data and its subsequent downstream applications in evolutionary biology.

In the first project, we developed MPL-R (marginal path likelihood with population reconstruction), a pipeline for LD estimation and fitness inference from NGS data. In this work, we introduced a bootstrap aggregating (bagging) approach to haplotype

reconstruction which significantly reduced the computational complexity traditionally associated with recovering full-length sequences from NGS reads. We performed analytical validation using simulated ground truth data and demonstrated that MPL-R is robust to changes in population and sequencing parameters, scaling to systems with genome lengths up to 3000 bp, a threshold previously unattainable by existing LD-aware methods due to scalability or accuracy limitations. Technical validation on a clinical human immunodeficiency virus 1 (HIV-1) dataset and biological concordance analysis of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) datasets yielded insights consistent with known biological results, while further investigation revealed how inference from only observable intra-read LD values yields results that are even worse compared to LD-independent inference.

The second project focuses on TGS data, which in addition to observing individual mutations in detail, offers improved resolution of structural variants (SVs) and long-range LD due to relatively longer read length compared to NGS. Currently, no software exists that utilizes TGS reads to infer these parameters simultaneously. To bridge this gap, we developed SLAFIT (viral fitness inference and structural variant detection from TGS data), the first integrated pipeline for simultaneous estimation of LD, fitness, and structural variation from TGS data. SLAFIT utilizes haplotype reconstruction to obtain full length sequences for LD estimation and implements a parallelized implementation of the marginal path likelihood method to improve the scalability of fitness inference. The computational efficiency of SLAFIT enables it to handle systems with genomes lengths up to 18,000bp in a few minutes. Analytical validation of the pipeline on simulated ground truth data demonstrated robustness to read degradation and outperformed NGS-based inference due to the larger read length of TGS. Technical validation of the pipeline on real data of enterovirus C104 and herpes simplex virus, the pipeline demonstrated biologically consistent results.

In the final project, we developed LDEN (linkage disequilibrium estimation from NGS data), a method of LD estimation from NGS data that bypasses population reconstruction. While reconstruction-based methods are effective, these are computationally intensive and can introduce errors into the reconstructed sequences. Recognizing that the LD matrix is fundamentally a covariance matrix, we utilized the property that the LD matrix needs to be positive semidefinite. While the observable banded LD matrix obtained from the NGS reads is not itself positive semidefinite, it can be converted into one via alternating eigenvalue modifications and projections. This is a known technique in finance which is used to complete correlation matrices, which we modified for completing banded

covariance matrices obtained from NGS data. Analytical as well as technical validation was performed and LDEN was found to be orders of magnitude faster compared to the haplotype reconstruction-based scheme. The accuracy of LDEN was also higher since it does not introduce errors associated with the haplotype reconstruction step.

In summary, we developed three methods for the estimation of LD and subsequent analysis from HTS data. The pipelines MPL-R and SLAFIT are based on haplotype reconstruction and focus on LD estimation from NGS and TGS data respectively, whereas LDEN estimates LD directly from NGS reads data. All methods were comprehensively evaluated on both simulated and real data. The results demonstrated that the methods are robust, efficient, and more accurate compared to existing tools. By advancing both haplotype reconstruction-based analysis as well as mathematical matrix completion techniques, this work provides a comprehensive toolkit for accurate and efficient LD estimation from HTS data and subsequent downstream analysis.

Panels

The qualifying panel of the student is composed of:

Supervisor(s)

Prof. SUN Yanni	Department of Electrical Engineering City University of Hong Kong

Qualifying Panel Member(s)

Prof. LI Fuyong	Department of Infectious Diseases and Public Health City University of Hong Kong
Dr. TO Kelvin Kai Wang	Department of Microbiology The University of Hong Kong

This thesis has been examined and approved by the following examiners:

Prof. SUN Yanni	Department of Electrical Engineering City University of Hong Kong
Prof. LI Runsheng	Department of Infectious Diseases and Public Health City University of Hong Kong
Prof. WONG Ka Chun	Department of Computer Science City University of Hong Kong
Prof. DU Xiangjun	School of Public Health (Shenzhen) Sun Yat-Sen University

Acknowledgements

I am thankful to the Almighty for granting me the resources and ability to undertake this work and to complete it to the best of my abilities.

I would like to express my sincerest gratitude to my supervisor, Professor Yanni Sun, for her constant guidance, patience, and support throughout my time here. Prof. Yanni's invaluable guidance has shaped every project I have carried out and has massively improved each work. Her wisdom and technical acumen have spotted pitfalls before I even started working in a direction and saved countless hours of work, for which I am extremely grateful. She has always empathized whenever expected outcomes were not met and has assisted me with every resource at her disposal. When I once mentioned having issues with my computer, she immediately handed me a brand-new laptop so that I would not have to spend time debugging on my own. On several other occasions, despite her own busy schedule, she has gone through raw data herself to help me identify sources of error in my work. I will always be grateful to her for her approach of managing problems with a focus on finding the solution rather than why a solution has not been found. This has consistently inspired confidence in me and helped me tackle challenges, even when projects did not work as expected. Throughout my research, she has always been very patient and understanding about the difficulties encountered and has often gone out of her way to provide guidance, even on weekends, for which I am extremely grateful. I am thankful not only for her technical guidance but also for her instilling critical thinking in me, which has proved immensely valuable. I am also grateful to her for the multitude of thoughtful gifts she has gifted us on various occasions.

I also express my gratitude to the examiners, Prof. Runsheng Li, Prof. Xiangjun Du, and Prof. Ka Chun Wong, for their insightful comments that significantly improved this thesis. I am grateful to my previous supervisors and collaborators, Prof. Kelvin Kai-Wang

To, Prof. Matthew R. McKay, Prof. Muhammad Saqib Sohail, Prof. Raymond H. Y. Louie, and Prof. John P. Barton, not only for their guidance and assistance in my current projects, but also for imparting invaluable research skills that have helped me immensely in my work. I am thankful to members of my qualifying panel Prof. Kelvin Kai-Wang To, and Prof. Fuyong Li for their support during my enrolment. I am grateful to my teachers and mentors Prof. Muhammad Salman, Prof. Asim Aziz, Prof. Tahir Zaidi, Prof. Muhammad Murtaza Khan, Prof. Shoaib Ehsan, and Prof. Muhammad Zeeshan for their support in every step of my academic journey since my undergraduate education.

I would express my deepest gratitude to Dr. Dehan Cai and Dr. Jiayu Shang who have always rendered valuable guidance, explained ideas in detail, and assisted with the very minutest of details of my work. I am grateful to them for always being there to assist whenever I faced bottlenecks in my work and offering tremendous assistance in resolving technical issues. I would like to express my gratitude to my seniors Dr. Xubo Tang, Dr. Herui Liao, Dr. Runzhou Yu, and Dr. Yongxin Ji, for their guidance and support in my projects and for sharing very useful resources which greatly helped in my learning. I express immense gratitude to Dr. Muhammad Ashraf Abdelhamid Abdelgawad and Dr. Abdurrashid Ibrahim Sanka for their valuable guidance during my oral defense; from taking out time to go over the presentation multiple times and providing very insightful comments to providing insightful ideas about improving the work before and after the defense. I am thankful to the members of our group, Dr. Guowei Chen, Mr. Cheng Peng, Ms. Jiaojiao Guan, Ms. Yilin Zhu, Dr. Caifang Wen, Mr. Wei Zou, and Mr. Fuchuan Qu for their detailed guidance regarding project modules and software they were working on. I am especially grateful to Mr. Fuchuan Qu for going out of his way to translate an entire presentation in English when he realized my translation software was not working properly. I would like to thank Mr. Ziyi Huang, Mr. Wei Jiang, Mr. Shen Yang, Mr. Zewei Li, Mr. Jiajin Wei, and

Mr. Yi Lu for their friendship and assistance with both research and non-research matters. I am grateful to my colleagues Mr. Mingchen Qi, Dr. Zhili Su, Dr. Guangyao Liu, Mr. Changlin Wu, Dr. Xinyu Luo, Dr. Kaicheng Wang, Dr. Shufan Chen, Dr. Hanke Feng, Dr. Zhicheng Lin, Mr. Yubin Xiong, Mr. Zhihui Wang, Ms. Jingyi Shi, Dr. Xu Han, Mr. Lei He, Mr. Hankuan Gao, and Mr. Zhengnan Fu for sharing the workload and providing guidance and support during teaching assistance duties. I am especially grateful to Dr. Zhili Su and Ms. Jingyi Shi for kindly assisting me in detail with my queries regarding the Chinese language. I am thankful to my colleagues in room 2386, Dr. Yue Liu, Dr. Tianlu Ma, Dr. Xiaosheng Wang, Dr. Ben Zhang, Dr. Chen Chen, Dr. Yibo Wang, Mr. Jingchun Xiang, Mr. Sheng Ren, Mr. Junhui Yang, Mr. Zhaozheng Zhu, Mr. Hao Guo, Mr. Hanqing Cao, Mr. Linshan Jia, Dr. Yang Zhao, Mr. Qinhong Yan, Dr. Wenwen Liu, Dr. Xueli Chen, and Dr. Kun Zhang for maintaining a very cordial environment in the lab and involving me in every celebration. I am also thankful to my colleagues in room 2388 for especially visiting and sharing snacks whenever they were celebrating any occasion. I am grateful to my past colleagues at The University of Hong Kong, Mr. Kingman Chan, Dr. Jonathan Daniel Ip, Dr. Thrimendra Kaushika Dissanayake, Dr. Shusheng Wu, Dr. Wan Mui Chan, Mr. Allen Wing Ho Chu, Mr. Xiankun Wang, Ms. Xiaojuan Zhang, Ms. Xiaochun Xie, Dr. Linlei Chen, Dr. Carol Ho Yan Fong, Dr. Hanjun Zhao, Mr. Brian Chan, Mr. Dr. Hazel Wu, Dr. Lu Lu, and Ms. Charlotte Choi whose guidance regarding real-data experimentation helped me greatly in designing my bioinformatics software.

I would like to thank my friends Dr. Ahmed Abdul Quadeer, Dr. Muhammad Saqib Sohail, Dr. Ahmar Siddiqui, Dr. Salman Tariq, Dr. Asad Iqbal, Dr. Farhan Khawar, Dr. Fedi Zouari, Dr. Abdul Waheed, Dr. Mohammed Effat, Dr. Mohsin Abbas, Dr. Zubair Ahmad, Dr. Parvej Alam, Mr. Syed Areeb Ali, Mr. Muhammad Saif Ashfaq, Dr. Muhammad Ashraf Abdelhamid Abdelgawad, and Mr. Ahmed Anis for their assistance during my initial days

here. I am grateful to Mr. and Mrs. Gao and everyone in the Sanyi family, as well as Dr. Ahmed Abdul Quadeer, Dr. Muhammad Saqib Sohail, Dr. Syed Faraz Ahmed, Dr. Usman bin Shahid, Dr. Farooq Riaz, Dr. Muhammad Shakeel, Mr. Asif Rahman, Dr. Syed Awais Wahab Shah, Dr. Muhammad Waqar, Dr. Usman Majeed, Dr. Bilal Asif, Mr. Arafat Asghar, Dr. Ikramullah Shah, Dr. Abdullah Tahir, Dr. Naveed Ahmad, Dr. Husnain Ali, Mr. Muhammad Usman, and Dr. Haroon Younas and their respective families. I express my gratitude to my MPhil colleagues Dr. Qingbei Cheng, Dr. Liusha Yang, Dr. Nicolas Auguin, Dr. David Morales-Jimenez, Dr. Laureano Moreno Pozas, and Dr. Hang Zhang. I am grateful to my friends Mr. Istikhar, Mr. Waqas Ahmed, Dr. Abdelrahman Bakr Mohammed Abdelnaby Eldaly, Dr. Muhammad Irfan, Mr. Omar Mohamed Salama Sayed Moustafa, Dr. Ihsan ul Haq, Dr. Nauman Mubarak, Dr. Faisal Rehman, Dr. Usman Bhutta, Dr. Asim Atta, Dr. Shubham Sharma, Mr. Humza Javed, Dr. Kishwar Khan, Dr. Ehsan ul Haq, Dr. Biswadeep Khan, Dr. Rana Rab Nawaz, Mr. Maqsood, Mr. Kamran Khaliq, Mr. Waseem, Prof. Zheyang Zheng, Dr. Zheshen Gao, Dr. Chengjie Luo, Dr. Awais Farid, Dr. Neelkanth Kundu, Dr. Soumitra Pal, Dr. Radik Srazhidinov, and Mr. Furqan Chaudhry and Mr. Thomas Narayana Swamy for their friendship.

I am thankful to City University of Hong Kong for granting me the opportunity and the resources to carry out this work and for the generous funding of my study program. I would like to thank the esteemed faculty members Prof. Paul Corrigan, Prof. Kelvin Shiu Yin Yuen, Prof. Hing Cheung So, Prof. Jie Chen, Prof. Wai Kit Ming, Prof. Anne Conan, Prof. Tu Tu Zaw Win, Prof. Arindam Basu, Prof. Leanne Lai-Hang Chan, and Prof. Haoliang Li for designing highly insightful course materials and for providing guidance that added immensely to my knowledge. I am grateful to Prof. Lip Ket Chin, Prof. Hei Wong, Prof. Yue Zhu, Prof. Wing Shing Chan, Mr. W. Y. Szeto, Mr. Vincent W H Mok, Mr. Anthony K M Hung, and Ms. Claire for their guidance and support during my teaching

assistant duties. I would like to express immense gratitude to Dr. Hung Chim and Ms. Shirely S. K. Ng, as well as Mr. Kenny K. H. Pang and Jacob K. Y. Lai from the Hong Kong University of Science and Technology for their assistance with computational platforms, including rendering assistance even during weekends. I am also grateful to the staff at the Department of Electrical Engineering for their continuous assistance with administrative matters. I would like to thank Ms. Lucia Mak, Miss Connie Lee, Mr. W. M. Shum, Mr. P. L. Ng, and Mr. Kelvin Chan for their assistance throughout my time here. I express heartfelt gratitude to Miss Connie Lee for her detailed guidance during the submission of the thesis and for ensuring everything went smoothly. I am also thankful to the staff at the Chow Yei Ching School of Graduate Studies, especially Ms. Mimi Lee, for continuous assistance and guidance regarding documentation. I pay my gratitude to the university for providing excellent medical and dental facilities on campus, and to all the staff at these facilities who were always there to assist whenever I fell ill. I am grateful to the staff of the Facilities Management Office, Campus Development Office, and Security Office for the continuous monitoring of university facilities and the timely arrangements of repairs. I am particularly grateful to the staff of the Security Office, who have always been first responders whenever assistance was required during holidays or after hours. I am also thankful to Ms. Maggie from the help desk of the Facilities Management Office for arranging urgent maintenance work whenever it was required. I pay my heartfelt gratitude to the landscaping, catering, and sanitation staff for their tireless work in ensuring the smooth functioning of the campus.

I would like to express gratitude to my family for their support, guidance, encouragement, and prayers. I am grateful to my parents, Dr. and Mrs. Pervez, for always keeping me in their prayers and offering words of love, support, and encouragement throughout my life. I thank my brothers, Ali and Saad, for their guidance regarding academic and non-academic matters, and for providing tremendous assistance whenever I visit home.

I am grateful to my wife for her support, patience, and understanding throughout my studies. I am also thankful to my extended family, my uncles, aunts, and cousins, who have always supported me with their love and prayers.

Table of contents

Abstract iv

Panels vii

Acknowledgements viii

List of figures xvii

List of tables xix

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Genomic sequencing . . . 5
  - 1.2.1 Next generation sequencing . . . 6
  - 1.2.2 Third generation sequencing . . . 7
- 1.3 Challenge and motivation for presented work . . . 9
- 1.4 Thesis objective . . . 9
- 1.5 Organization of the thesis . . . 10

2 Linkage-aware inference of fitness from short-read time-series genomic data 12
- 2.1 Introduction . . . 12
  - 2.1.1 Overview . . . 12
  - 2.1.2 Background . . . 12
  - 2.1.3 Related work . . . 13
- 2.2 Methodology . . . 15
  - 2.2.1 Overview . . . 15
  - 2.2.2 Quality control . . . 16
  - 2.2.3 Read aligning and sorting . . . 18
  - 2.2.4 Bagging and population reconstruction . . . 18
  - 2.2.5 Estimation . . . 20
  - 2.2.6 Estimation with a 4-allelic model . . . 22
  - 2.2.7 Simulated data . . . 23
  - 2.2.8 HIV-1 data . . . 26
  - 2.2.9 SARS-CoV-2 data . . . 28
- 2.3 Results . . . 30
  - 2.3.1 Performance on simulated data . . . 30
  - 2.3.2 Comparison between MPL-R and MPL (identity covariance) . . . 39
  - 2.3.3 Naïve estimation of ICM from reads . . . 41
  - 2.3.4 Parallelized population reconstruction . . . 43
  - 2.3.5 Comparison with other methods . . . 44

2.3.6 Investigation of difference in performance . . . 46
2.3.7 Results on HIV-1 patient data . . . 50
2.3.8 Results on SARS-CoV-2 patient data . . . 51
2.4 Discussion . . . 53

3 Viral fitness inference and structural variant detection from TGS data 59
3.1 Introduction . . . 59
3.1.1 Overview . . . 59
3.1.2 Background . . . 59
3.1.3 Related work . . . 62
3.2 Methodology . . . 62
3.2.1 Overview . . . 62
3.2.2 Read alignment and SV detection . . . 63
3.2.3 LD estimation and fitness inference . . . 64
3.2.4 Parallel processing . . . 64
3.2.5 Overview of evaluation . . . 66
3.2.6 Simulated data . . . 67
3.2.7 EV-C104 data . . . 69
3.2.8 HSV-1 data . . . 71
3.3 Results . . . 71
3.3.1 LD estimation and fitness inference via SLAFIT is robust to varying sequencing parameters . . . 71
3.3.2 Longer read length of TGS reads yields improved fitness estimates compared to NGS reads . . . 75
3.3.3 SLAFIT outperforms existing LD-aware fitness inference . . . 77
3.3.4 SLAFIT reveals insights from prolonged infection data consistent with biological knowledge . . . 78
3.3.5 SV detection via cuteSV is robust and consistent with known results . . . 80
3.4 Discussion . . . 81

4 Linkage disequilibrium estimation from NGS data 85
4.1 Introduction . . . 85
4.1.1 Overview . . . 85
4.1.2 Background . . . 85
4.1.3 Related work . . . 86
4.2 Methodology . . . 87
4.2.1 Overview . . . 87
4.2.2 Read alignment and partial LD calculation . . . 88
4.2.3 Alternating projections method . . . 90
4.2.4 LD estimation from paired-end reads . . . 91
4.2.5 Simulated data . . . 92
4.2.5.1 Variation in population diversity . . . 93

4.2.5.2 Variation in genome length 93
4.2.5.3 Wright-Fisher time-series simulation 93
4.2.5.4 NGS reads simulation 94
4.2.5.5 Sequencing parameter sensitivity analysis 94
4.2.5.6 Statistical confidence 95
4.2.6 SARS-CoV-2 data 95
4.3 Results 96
4.3.1 Overview 96
4.3.2 The error correction scheme improves the accuracy of the intra-read LD values 96
4.3.3 LDEN demonstrates fast convergence 97
4.3.4 LDEN recovers valid LD matrices from NGS data 98
4.3.5 The performance of LDEN is correlated with the accuracy of the banded LD information 100
4.3.6 LDEN outperforms population-reconstruction based estimation . . 102
4.3.7 LDEN is scalable with increasing genome length. 103
4.3.8 Paired-end reads improve the accuracy of LDEN. 104
4.3.9 LDEN reveals known biological insights from patient data 105
4.4 Discussion 105

5 Conclusion and future work **108**
5.1 Conclusion. 108
5.2 Future work 109
5.2.1 Incorporation of epistasis 110
5.2.2 Direct LD estimation from TGS data 110

References **111**

Publications **123**

List of figures

Figure 1.1 Phylogenetic tree showing the diversity in HIV. 2
Figure 1.2 Illustration of viral evolution 3
Figure 1.3 Relative prevalence of variants of SARS-CoV-2 with time . . 4
Figure 1.4 Distribution of read lengths obtained via HiFi sequencing. . . 8
Figure 2.1 Illustration of the MPL-R pipeline 19
Figure 2.2 Structures of fitness landscape used in the simulations 24
Figure 2.3 Sweep of the regularization parameter 25
Figure 2.4. Entropy values of the single mutant allele frequencies for the HIV-1 data 26
Figure 2.5 SARS-CoV-2 data indicates presence of linkage. 28
Figure 2.6. Entropy values of the single mutant allele frequencies for the SARS-CoV-2 29
Figure 2.7 MPL-R has good classification performance 30
Figure 2.8. Estimated selection coefficients show good classification performance and correlation with each other 31
Figure 2.9 MPL (identity covariance) gets outcompeted by MPL-R for loci evolving under genetic linkage 32
Figure 2.10 Performance improvement of MPL-R over MPL (identity covariance) is most significant for stronger selection coefficients. 33
Figure 2.11. MPL (banded) performs poorly regardless of read length . . 34
Figure 2.12. MPL-R recovers linkage effects across the entire sequence . 35
Figure 2.13. Robustness of MPL-R to changes in data parameters 36
Figure 2.14. MPL-R consistently performs better than a random predictor when applied to datasets with changing data parameters 37
Figure 2.15. Increasing the genome length decreases the accuracy of the inferred covariance values 38
Figure 2.16. Bagging approach to reconstruction improves the wall clock time without affecting the classification performance 42
Figure 2.17. Decrease in the size of the bootstrap samples decreases the wall clock time with a gradual decrease in the classification accuracy 44
Figure 2.18. MPL-R outperforms available inference methods 45
Figure 2.19. MPL-R consistently performs better than available inference methods. 46
Figure 2.20. The performance of MPL (identity covariance) degrades when mutational flux is ignored 47
Figure 2.21. MPL-R correctly infers biological insights from intrahost HIV-1 patient data 48
Figure 2.22. MPL-R completes the analysis on the HIV-1 data in a reasonable amount of time 50
Figure 2.23. MPL-R identifies a higher number of beneficial mutations in the RBD 51
Figure 2.24. Reconstruction of double mutant allele frequencies is accurate for one bootstrap sample 53
Figure 2.25. The reconstructed double mutant allele frequencies are accurate for all bootstrap samples 54
Figure 2.26. MPL-R faithfully reconstructs the triple mutant allele frequencies 55

Figure 2.27. The reconstruction performed via Quasirecomb is accurate . 56
Figure 2.28. The covariance matrices of allele frequencies are accurate. . 57
Figure 3.1 SLAFIT pipeline schematic. 63
Figure 3.2 Parallelized implementation of MPL supplied with SLAFIT requires lower wall clock time compared to the sequential implementation 65
Figure 3.3 Clinical timeline of EV-C104 infection data. 68
Figure 3.4 Read-length used in the EV-C104 infection data is much shorter than the genome length. 69
Figure 3.5 EV-C104 infection data indicates the presence of LD 70
Figure 3.6 Bi-allelic model for the EV-C104 infection data shows good association with the 4-allelic model 71
Figure 3.7 SLAFIT can perform over a range of sequencing parameters . 72
Figure 3.8 The strain-level fitness inference of SLAFIT is strongly associated with the ground-truth with varying data parameters 73
Figure 3.9 SLAFIT accurately recovers long-range LD values 74
Figure 3.10 LD inference via SLAFIT is accurate throughout the evolutionary history 75
Figure 3.11 SLAFIT is able to reconstruct fitness values from TGS reads more accurately as compared to an NGS-based scheme 76
Figure 3.12 The association between the estimates of SLAFIT and the ground truth values is stronger as compared to a short-read based scheme 77
Figure 3.13 Fitness values inferred via SLAFIT have lower relative error with respect to the ground truth as compared to LB 78
Figure 3.14 SLAFIT infers a number of nonsynonymous SNVs as non-neutral (blue) which are identified as neutral by SLAFIT(LI) 80
Figure 3.15 LD patterns in the EV-C104 infection data show evidence of clonal interference. 81
Figure 3.16 Prevalence values of haplotypes reconstructed from the EV-C104 infection data indicate decreasing diversity with time. 82
Figure 3.17 Distribution of read lengths in the HSV-1 dataset. 83
Figure 4.1 Proposed error correction scheme yields accurate intra-read LD 97
Figure 4.2 LDEN demonstrates fast convergence 98
Figure 4.3 LD matrices estimated by LDEN are positive semidefinite . . 99
Figure 4.4 The performance of LDEN is correlated with the accuracy of the information captured by the reads 100
Figure 4.5 The information observable via reads depends on the read length and the genome length 101
Figure 4.6 LDEN outperforms MPL-R in both estimation accuracy and computational speed 102
Figure 4.7 LDEN is scalable with increasing genome length 103
Figure 4.8 LDEN yields more accurate results using paired-end reads . . 104
Figure 4.9 LDEN identifies beneficial mutations in SARS-CoV-2 105

List of tables

Table 2.1. Mutation rates used in the analysis of HIV-1 patient data . . . 27
Table 2.2. MPL-R reports similar insights as MPL 49
Table 2.3. Mutations inferred as beneficial by MPL-R map to known beneficial mutations 52
Table 3.1 Nonsynonymous SNVs identified as highly fit via linkage-aware inference occur in computationally predicted B-cell epitopes 79
Table 3.2 SVs detected from HSV-1 data under varying maximum read length thresholds 82
Table 4.1 Prevalence values of the four variants used in datasets 1 to 4 . 93

Chapter 1

Introduction

1.1 Background

Viral pathogens are the causative agents of numerous diseases including hepatitis, acquired immune deficiency syndrome (AIDS), influenza, and coronavirus disease 2019 (COVID-19). These and other viral diseases overburden healthcare systems, substantially degrade the quality of life of patients, and have collectively resulted in millions of deaths. For example, AIDS has caused a cumulative loss of approximately 40 million lives since its discovery in 1981 [1]. More recently, the COVID-19 pandemic directly resulted in nearly 7 million confirmed casualties, inflicted economic losses on several industries, and left a significant portion of patients suffering from persistent long-term effects, referred to as post-acute sequelae of SARS-CoV-2 (PASC) [2]. The development of effective antiviral therapies not only relies on understanding viral behavior in cell culture but also on a comprehensive knowledge of how the viral genome evolves, as phylogenetic changes are fundamentally driven by genomic changes, called mutations. Many ribonucleic acid (RNA) and single-strand deoxyribonucleic acid DNA (ssDNA) viruses lack proofreading mechanisms within their replication pipelines, which results in high mutation rates and accumulation of multiple mutations. Fig. 1.1 illustrates the extensive diversity in 4,686 HIV genomes, a direct consequence of mutations acquired during viral evolution.

While the majority of mutations are deleterious to the virus and strains that acquire these quickly die out via purifying selection, some mutations result in an evolutionary advantage to the virus. For example the A163G mutation in HIV reduces recognition by cytotoxic T lymphocytes (CTLs), thereby facilitating immune evasion [3], while the K417N mutation in

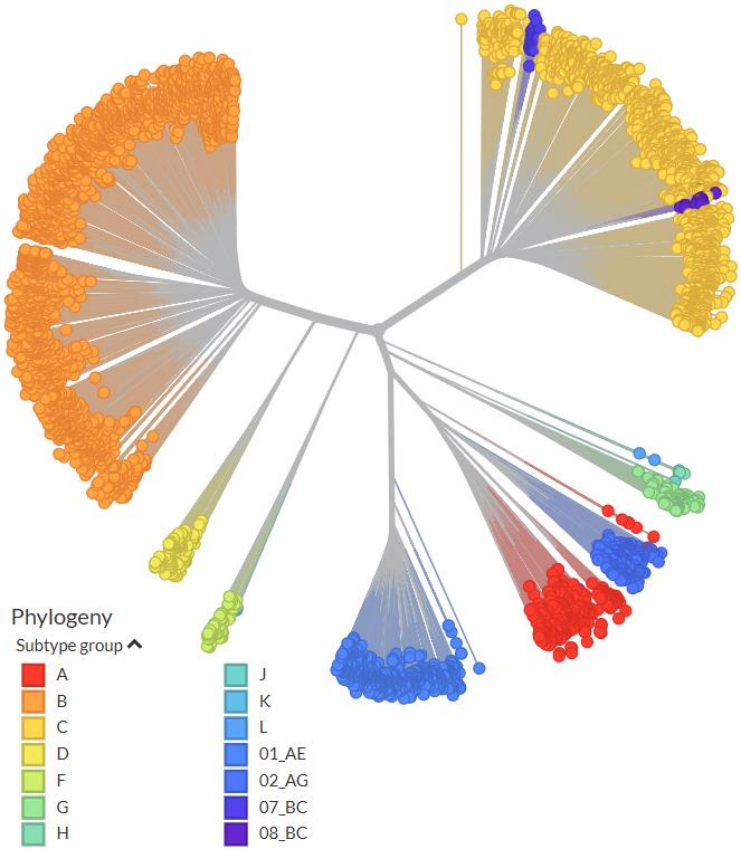


Figure 1.1 Phylogenetic tree showing the diversity in 4,686 HIV genomes sampled between January 1981 and November 2019. Figure generated by NextStrain [5] using data from Los Alamos National Laboratory [6].

SARS-CoV-2 assists in escape from neutralizing antibodies [4]. In addition to point mutations acquired during replication, viruses can also undergo evolution via processes such as reassortment and recombination (Fig. 1.2). Reassortment involves the exchange of genomic segments between different variants of the virus co-infecting the same host, which can lead to the emergence of novel combinations of mutations. These combinations may be fitter, i.e. better capable of survival, compared to the parent variants. Notably, reassortment in the H1N1 variant of influenza was responsible for the 2009 pandemic that claimed more than 200,000 lives [7]. Recombination, by contrast, involves template switching during viral replication, producing progenies that contain chimeras of genetic material from distinct parental sequences. Prominent examples of recombination are the XAY variant of SARS-CoV-2 [8], the G19080 variant of

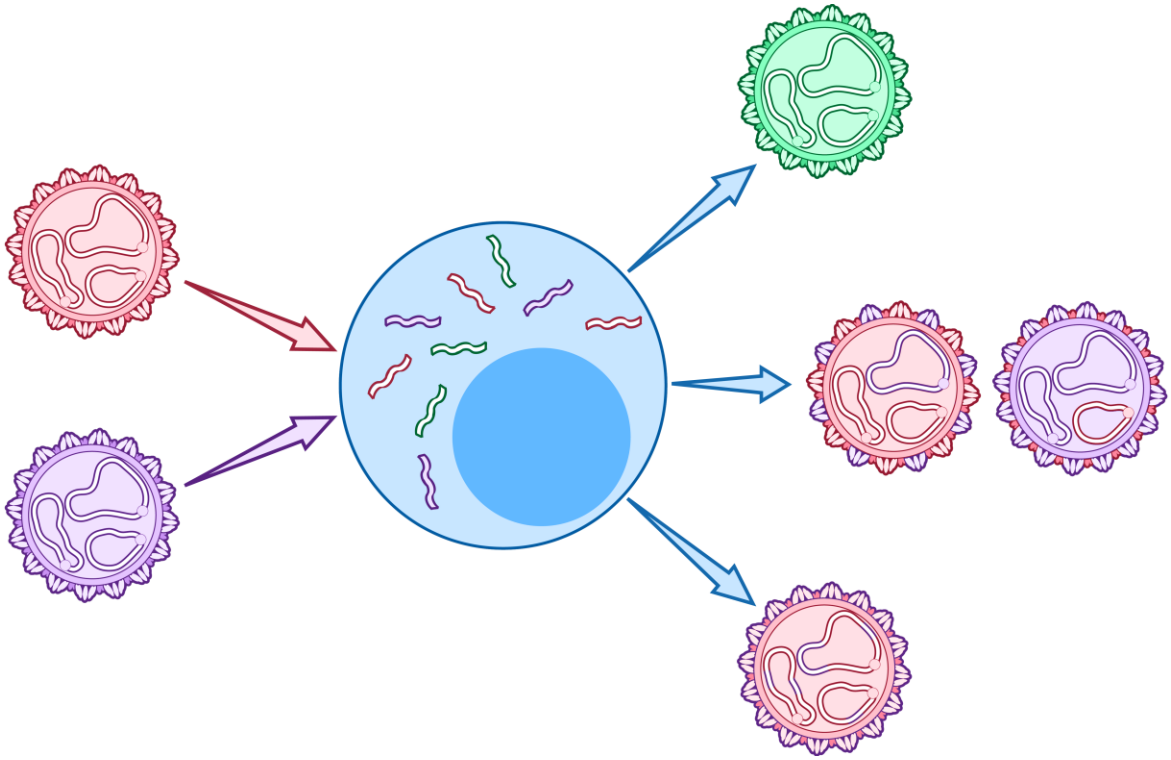


Figure 1.2 Illustration of viral evolution via mutations (*top*), reassortment (*middle*), recombination (*bottom*). Created using template figures sourced from the National Institute of Allergy and Infectious Diseases [11], [12].

the herpes simplex virus (HSV) [9], and variants arising from the vaccine strain of the infectious laryngotracheitis virus [10].

Regardless of their origin, whether through point mutations, reassortment, or recombination, the identification and subsequent therapeutic targeting of beneficial mutations is a critical step in designing effective antiviral interventions. However, this task is substantially complicated by the presence of multiple simultaneous mutations within viral genomes, which are often prevalent in nature. When multiple mutations co-occur on a genome simultaneously, all mutations that appear on the same genomic background as a beneficial mutation can increase in prevalence as evolution progresses. Conversely, if two distinct beneficial mutations arise on two competing variants of the virus, one variant with the less advantageous mutation will get outcompeted and die out despite its mutation conferring a genuine evolutionary advantage to

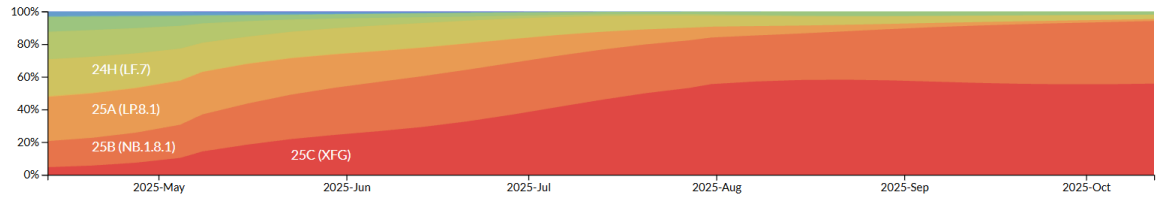


Figure 1.3 Relative prevalence of variants of SARS-CoV-2 varies with time as variants with less advantageous mutations get outcompeted by fitter variants. Figure generated by NextStrain [13] using data from the Global Initiative on Sharing All Influenza Data [14].

the virus in isolation (Fig. 1.3). This non-random association of mutations, termed LD, is shown by multiple studies to be prevalent in nature. LD has been demonstrated to not only influence the evolution of viruses [13], [14], [15], but also drug resistance of the parasite *Plasmodium falciparum* [16], protection against malaria in humans [17], and advancement of cancer [18]. When evolution is influenced by LD, observation of only the single-point statistics of the mutations, also called single mutant allele frequencies, is insufficient to correctly characterize the nature of each mutation. For instance, any mutation that increases in prevalence solely because it is linked with a beneficial mutation would be incorrectly classified as beneficial if analysis is performed solely on the basis of the single mutant allele frequencies. Likewise, any mutation that decreases in prevalence due to competition with fitter mutations would be incorrectly classified as deleterious by single mutant allele frequency-based analyses. Therefore, single mutant allele frequencies cannot reliably differentiate beneficial mutations from neutral or deleterious ones.

Complete characterization of mutations requires knowledge of not only how individual mutations evolve, but also the interactions between mutations. These interactions are captured by the pairwise prevalence of mutations, or the double mutant allele frequencies, which quantify how pairs of mutations evolve together. The single and double mutant allele frequencies are integrated into the coefficient of linkage disequilibrium (D), mathematically given as

$$D = f_{ij} - f_i f_j,$$

where f_i and f_j respectively represent the single mutant allele frequencies at loci i and j, and f_{ij} represents the double mutant allele frequency at the locus pair (i, j). LD is also quantified via normalized metrics such as and Lewontin's D' [19] and correlation coefficient r^2, but throughout this thesis, D will be used. LD has been employed in a number of studies to identify beneficial mutations and has consistently yielded more accurate insights compared to LD-independent methods which rely only on changes in the single mutant allele frequencies [13], [20], [21]. In addition to the classification of mutations, LD has proven invaluable for the estimation of population divergence, inference of historical population sizes, carrying out genome-wide association studies (GWAS), and performing genomic prediction [22], [23], [24], [25]. Thus, accounting for LD is not merely an optional refinement but a fundamental requirement for robust evolutionary inference.

1.2 Genomic sequencing

Accurate estimation of LD is fundamentally dependent on the accurate resolution of the viral genome, a process known as genomic sequencing. The pioneering work on genomic sequencing was published by Frederick Sanger in 1977 [26], a breakthrough that revolutionized molecular biology and led to the resolution of a number of genomes such as that of the bacteriophage φX174 [27], *Saccharomyces cerevisiae* (brewer's yeast) [28], *Drosophila melanogaster* (fruit fly) [29], and ultimately the human genome [30].

Sanger sequencing, also referred to as the chain termination method, involves using a single-stranded DNA template to generate DNA fragments of different lengths through the action of deoxynucleotides and DNA polymerase. The sequencing is performed in four separate batches, each containing one of four deoxynucleotide triphosphates (dNTPs), corresponding to four nucleotides, i.e. deoxyadenosine triphosphate (dATP), deoxycytidine triphosphate (dCTP), deoxyguanosine triphosphate (dGTP), and deoxy thymidine triphosphate (dTTP). The splitting

of the reaction in batches is to ensure that the DNA fragment terminates in a different nucleotide. The resulting DNA fragments are analyzed via gel electrophoresis or capillary electrophoresis to identify their composition. Sanger sequencing can sequence genomic fragments up to 900 base pairs (bp) in length [31] and has a very low error rate of 0.001% [32], which is why it has remained the gold standard for validating sequencing results from other technologies. Sanger sequencing is reported to have. However, Sanger sequencing is slow and expensive, rendering it impractical for large-scale sequencing projects, motivating the development of newer sequencing technologies. These advanced technologies, collectively referred to as HTS, are further subdivided into NGS and TGS. A brief description of each technology is presented below.

1.2.1 Next generation sequencing

Next generation sequencing (NGS) addressed the slow speed of Sanger sequencing by the adoption of massively parallelized sequencing [33]. This approach dramatically decreases both the running time and cost at the expense of shorter read-length and platform dependent error rate. NGS includes legacy technologies such as Roche/454 pyrosequencing [34], Ion torrent sequencing [35], as well as the modern Illumina sequencing technology.

Illumina sequencing has remained the dominant platform for NGS sequencing due to its high throughput and low error rates. This platform adopts a strategy of sequencing by synthesis in which multiple copies of the template genome are generated and amplified via a process known as bridge amplification [36]. The sequencing process begins with sequences undergoing cleavage and adapter attachment, which are then affixed onto a substrate called a flow cell that contains sequence fragments complimentary to the template fragments called oligonucleotides. Bridge amplification is performed in which each fragment forms a bridge with the complimentary fragment and gets copied. Repeated amplification cycles generate multiple copies of the affixed fragments to amplify the signal from the template sequence fragments.

Illumina uses an optical signal to detect added nucleotides, but it does so for the entire flow cell in parallel, which leads to fast turnaround times.

1.2.2 Third generation sequencing

NGS offers fast turnaround times and low error rates, but it is constrained by short read-lengths, typically up to 800bp [37]. The limitation of short read-lengths arises in part from the finite number of PCR cycles possible before degradation of the polymerase compromises amplification efficiency. Third generation sequencing (TGS) was developed specifically to address the read-length limitation associated with NGS. Unlike NGS, TGS involves direct sequencing of the template, bypassing the issues associated with PCR amplification, and yielding reads that can be longer than 1Mbp [38] in length.

However, TGS reads do suffer from higher error rates compared to NGS reads, with error profiles dependent on the speed at which the template sequence is read. Higher sequencing speeds blur the detection signal, thereby missing individual nucleotides or miscalling multiple nucleotides as one. Conversely, slower speeds report individual nucleotides multiple times [39]. Furthermore, because TGS doesn't employ PCR amplification, it lacks the ability to average out errors over multiple copies as done in NGS. As a consequence of the presence of errors in TGS reads, error correction is a standard component in bioinformatic pipelines that handle TGS data. TGS encompasses two main technologies, Pacific Biosciences single-molecule, real-time (SMRT) sequencing [40] and Oxford Nanopore Technology (ONT) sequencing [41].

In PacBio sequencing, adapter sequences are ligated to both ends of the template sequence, and the resulting construct is inserted into small wells called zero-mode waveguides (ZMW). Each ZMW contains an individual DNA polymerase molecule immobilized at its base. The polymerase binds to the adapter sequences and initiates synthesis of a complimentary strand via the incorporation of individual nucleotides. The incorporation of each nucleotide emits an

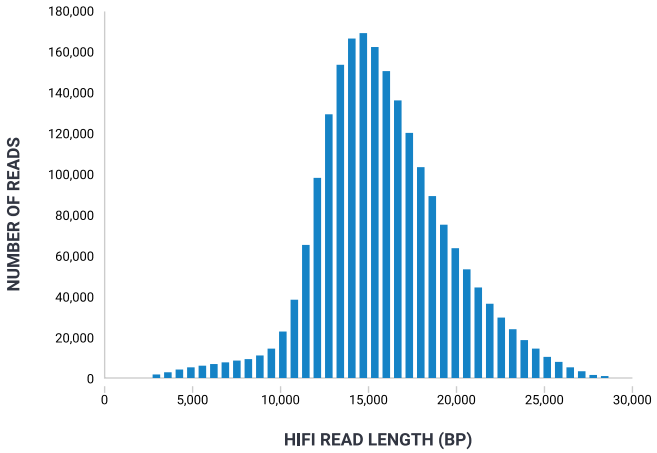


Figure 1.4 Distribution of read lengths obtained via HiFi sequencing. Figure generated by PacBio [46].

optical signal which is observed in real time and interpreted to resolve the sequence. This technology achieves read lengths of up to a few kbp with error rates less than 1% [42] . Error rates can be further reduced by reading the template sequence multiple times and generating a consensus sequence. A complete sequencing of a template is referred to as a "pass" and multiple passes can be made to create a high-fidelity (HiFi) sequence, in a process known as circular consensus sequencing (CCS) or HiFi sequencing. The redundancy provided by multiple passes reduces the error rate to as low as 0.1% [43]. However, HiFi sequencing sacrifices read length to achieve better error performance. The DNA polymerase used for copying the template degrades with time, and multiple passes can only be made on relatively shorter template sequences as compared to a single pass. Nevertheless, HiFi sequencing is still able to generate reads of lengths exceeding 25kbp (Fig. 1.4) [44].

ONT sequencing operates on an entirely different principle from PacBio sequencing. Adapter-ligated templates are passed through a staphylococcal α-haemolysin (αHL) protein channel, or nanopore, embedded in a synthetic membrane. As individual nucleotides travel through the pore, disruptions in an externally applied electric current are created, which are specific to individual nucleotides. These variations in the electric current are measured and interpreted to identify the individual nucleotides passing through the pore in real-time [45]. Adapter ligation

is performed as these not only assist in the passage of the template sequence through the pore but also ensure that the sequence passes in the correct orientation. Because ONT does not rely on polymerase which can undergo degradation over time, it achieves read lengths up to 440kbp [46]. ONT sequencing is particularly renowned for its portability and capability to perform real-time sequencing in field settings [47], [48]. It has been successfully deployed in the field during outbreak investigations, including the 2014 Ebola virus outbreak in West Africa [49], [50] and the 2018 Lassa fever outbreak in Nigeria [51].

1.3 Challenge and motivation for presented work

Although the read lengths of HTS, especially that of TGS, have steadily increased through the use of improved reagents as well as novel software tools, these technologies remain unable to observe LD information across entire viral genomes. The maximum possible read lengths of NGS still fall short of the lengths of most viral genomes and consequently cannot observe LD beyond a few hundred bp. While TGS reads, especially ultralong reads from ONT sequencing, do provide long-range LD information, the low throughput and high error rates still continue to pose challenges for meaningful LD estimation. Therefore, there is a pressing need for software pipelines capable of extracting meaningful long-rage LD information from HTS data while accounting for platform-specific error profiles and coverage limitations.

1.4 Thesis objective

In this thesis, we address the problem of estimating long-range LD from viral HTS data. This problem is of interest in several fields or research such as the characterization of viral pathogens, rational design of antiviral therapeutics, surveillance and control of epidemics, and the fundamental understanding of viral evolutionary dynamics. An abundance of viral data obtained via HTS already exists for major pathogens including SARS-CoV-2 [52], HIV [53], influenza [54], hepatitis C virus (HCV) [55], hepatitis B virus (HBV) [56], and enterovirus C104 [57]. Despite the availability of these and similar datasets, no existing tool can estimate

LD from these in a reasonable amount of time, particularly when distant loci are involved. Existing LD estimation methods are severely limited either by poor scalability to genome sizes or by restrictive model assumptions that render these unsuitable for viral data characterized by high mutation rates and the presence of multiple competing variants [20], [58], [59], [60], [61], [62]. Consequently, there is a need for scalable, robust, and efficient software tools that can utilize HTS data to estimate long-range LD. This thesis seeks to fill this gap by developing novel computational pipelines that address the scalability limitations and model constraints of existing approaches, thereby enabling the estimation of meaningful long-range LD information from viral populations across entire genomes.

1.5 Organization of the thesis

Chapter 1 introduces viral pathogens and their impact on human population. Mechanisms of mutation acquisition are described followed by a discussion on the need for LD estimation to enable accurate understanding of mutations. A brief summary of past and existing sequencing technologies is presented after which challenges associated with LD estimation from HTS data are explained.

Chapter 2 presents MPL-R, a pipeline for the estimation of long-range LD from NGS data and its subsequent utilization in inferring mutational fitness. MPL-R takes in NGS reads from a time-series experiment as input, performs haplotype reconstruction to obtain LD information across the entire genome, then uses the estimated LD to compute the fitness of individual mutations. A novel bootstrap aggregating approach to population reconstruction is proposed to reduce computation time.

Chapter 3 builds upon the concepts introduces in Chapter 2 and exploits the longer read-lengths of TGS data for long-range LD estimation, LD-aware fitness inference, and inference of SVs. The proposed software pipeline, SLAFIT, uses haplotype reconstruction for LD estimation and integrates this with an off-the-shelf software package for SV detection. SLAFIT

can operate on data from a single time-point or from multiple time-points in an evolutionary experiment. A novel, parallelization-based approach to fitness inference is introduced, which makes SLAFIT scalable with respect to genome length and population size.

Chapter 4 addresses the computational limitations faced by population reconstruction-based LD estimation methods. We introduce LDEN, a software tool that estimates long-range LD directly from NGS data without performing population reconstruction. We utilize ideas from finance and convex optimization to complete the unobservable part of the LD matrix directly from the LD information observable from the NGS reads.

Chapter 5 summarizes the content of the thesis and presents some possible directions of future research.

Chapter 2

Linkage-aware inference of fitness from short-read time-series genomic data

2.1 Introduction

2.1.1 Overview

Inferring the fitness effect of mutations is a basic problem in understanding the evolution of populations over time. When multiple mutations are present in a population simultaneously, genetic linkage comes into play, and the fate of an individual mutation depends on both its fitness as well as the background on which it occurs. Accurate inference of fitness effects for evolutionary systems with multiple competing mutations is therefore contingent on resolving the confounding effects of genetic linkage, captured by the covariance between allele-pairs. Increasingly, evolutionary studies are using short-read sequencing technologies to produce detailed snapshots of evolving populations. This presents a problem as the frequencies of allele-pairs are not known beyond the read-length, hampering any attempt to resolve the effects of genetic linkage between pairs of loci residing on different reads. Here we present a computationally efficient pipeline for inferring selection from short-read time-series data with partial allele-pair frequency information, whilst accounting for linkage. Simulation results show that the method has good performance and is scalable to systems with several thousand variants. Additionally, we demonstrate the pipeline's utility on real datasets of within-host HIV and SARS-CoV-2 evolution, showcasing its applicability in resolving linkage effects from complex evolutionary histories.

2.1.2 Background

The evolution of a population (organisms of the same species living and reproducing in a particular environment) is shaped by the complex interplay of various evolutionary

phenomenon such as selection, mutation, drift, recombination, genetic linkage, and others. Genetic linkage—the tendency of mutations to be inherited together—becomes an important factor in shaping evolution when multiple loci are simultaneously polymorphic in the population. Experimental and observational evidence shows that genetic linkage is an important evolutionary factor in several real evolving populations [14], [63], [64]. Genetic linkage has been shown to affect drug resistance in the protozoan parasite *Plasmodium falciparum* [16], malaria protection in humans [17], cancer progression [18], and evolution of viruses such as HIV [13] and influenza [14], [15]. Genetic linkage may result in hitchhiking [65], positive or negative (clonal) interference [66] or background selection [67], [68].

Researchers have long studied population-level time-series genomic data to identify and understand the evolutionary phenomenon responsible for the observed patterns of genetic variation. These efforts have been boosted by the developments in short-read sequencing technologies [33], [69], [70], [71], [72] which have enabled the collection of high-resolution time-series genomic data. Examples of such data include evolve-and-resequence studies under lab settings of viruses [73], [74], bacteria [75], [76], yeast [64], [77], [78], [79], and fruit flies [80], [81], and studies of in vivo evolution of pathogens including HIV [82], [83], influenza [84], [85], [86], HCV [87], [88], SARS-CoV-2 [89], [90] and cancer [91], [92]. These and other similar evolutionary data provide an opportunity to observe evolving populations in remarkable detail.

2.1.3 Related work

Several methods exist in the literature that can use the mutant allele frequency trajectories observable in data obtained from short-read sequencing platforms to infer evolutionary parameters in scenarios where genetic linkage between loci is negligible [93], [94], [95], [96], [97], [98], [99], [100], [101], [102], [103], [104]. However, in populations where genetic linkage influences evolution, relying solely on mutant allele frequencies without considering

interactions between alleles can lead to inaccurate inference of evolutionary parameters. Linkage-aware methods [13], [58], [59], [60], [61], [62] can analyze the evolution of multiple loci collectively to disentangle the effects of genetic linkage. However, despite their utility for analyzing linked loci [58], [59], [60], [61], [62] these methods are limited to small system size [58], [60], lack the ability to quantify fitness effect of individual mutations [59], or incur a high computational cost [58], [59], [60], [61], [62]. In contrast, Sohail *et al.*, 2021 [13] proposed marginal path likelihood (MPL), a fast and scalable framework that leads to closed-form analytical estimates of selection coefficients from time-series data while accounting for the effects of genetic linkage in addition to those of mutation, drift and recombination. MPL has found applications in different scenarios, such as fitness inference in the presence of sampling noise [105], fitness inference in the presence of epistasis [106], understanding HIV-1 escape [107], and analyzing SARS-CoV-2 data [108], [109].

In the MPL framework, the linkage between multiple alleles is taken into account through the covariance matrix of the mutant allele frequencies at each time-point. The entries of this covariance matrix are a function of single and double mutant allele frequencies and are equivalent to a classical LD metric [110]. Inferring this covariance matrix is a significant challenge for short-read sequencing technologies, which are commonly used to observe evolving populations. While short-read sequencing provides accurate estimates of the single mutant allele frequencies, only the double mutant allele frequencies between allele pairs residing on the same read are directly observable. A recent approach [20] presented a method to estimate mutant allele covariance matrices from the product of the change in single mutant allele frequencies. This approach is sensitive to the frequency of temporal sampling since changes in the single mutant allele frequencies become difficult to observe as the temporal sampling step increases. The methodological challenge of robustly inferring double mutant

allele frequencies from short-read data across general evolutionary scenarios remains to be addressed.

2.2 Methodology

2.2.1 Overview

This work builds upon the MPL framework and presents a novel pipeline for the inference of selection from short-read time-series data. The proposed computational pipeline, termed marginal path likelihood with population reconstruction (MPL-R), takes in short-read time-series data, infers the double mutant allele frequencies for all allele-pairs beyond the read length (inter-read allele-pairs), and uses these to infer selection. The "R" in MPL-R represents population reconstruction, whereas "MPL" represents marginal path likelihood as used in the study that proposed this method [13]. The main idea is to reconstruct full-length sequences of an evolving population from short reads, from which the double mutant allele frequencies can be computed. However, population reconstruction is a challenging problem to solve [111]. The reconstruction process may result in the reconstruction of erroneous haplotypes, especially those inferred to have a low frequency in the population and is particularly error-prone when the actual population consists of multiple haplotypes [112]. We show in this work that for the purpose of inferring selection, it is not necessary to reconstruct the actual population exactly, and one only needs to obtain a probable set of sequences approximating the actual population. We use an off the shelf population reconstruction algorithm, Quasirecomb [113], and employ a novel reconstruction strategy based on bootstrap aggregating (bagging) [114], which allows for leveraging parallel processing on subsampled data leading to reduced memory requirements and wall clock time without significantly affecting the accuracy of the final estimates. We demonstrate that Quasirecomb, while making some errors in reconstructing full populations, does an excellent job at producing accurate double mutant allele frequencies. We test the performance of MPL-R on simulated data and show that the pipeline accurately infers selection

over a range of parameters. Interestingly, we found that inference using partial covariance information (e.g., using only covariance values that could be measured directly from short-read data) yields worse results than ignoring covariances entirely. Comparison with a simpler MPL implementation that naively employs observed double mutant allele frequencies while disregarding all unobservable double mutant allele frequencies (corresponding to allele-pairs residing on different reads), has vastly suboptimal performance. This emphasizes the importance of constructing double mutant allele frequencies both within and outside sequencing reads, as achieved with MPL-R. We further demonstrate the performance of MPL-R by applying it to genetic data from a cohort of HIV-1 infected patients, demonstrating that it can capture similar biological insights from short-read data as obtainable with full length sequences.

The main challenge addressed in this work is the inference of inter-read covariance information (which is not directly observable from short-read data) required by the MPL framework. To infer the unobserved inter-read covariance information, MPL-R performs population reconstruction from short-read data at each sampled time-point using Quasirecomb. MPL-R employs a novel reconstruction strategy based on bagging. The single and double mutant allele frequencies are then computed from the reconstructed sequences, which are then used to obtain the estimates of selection coefficients adjusted for linkage effects. The workflow of the proposed MPL-R pipeline is shown in Fig. 2.1.

2.2.2 Quality control

Several processes involved in the generation of NGS reads can affect the mutant allele counts, which in turn modify both the single and double mutant allele frequencies. Before running the MPL-R pipeline, the effect of each process needs to be investigated and corrected.

One such process is amplification, which is used to enhance the signal from the template molecule. Amplification can selectively inflate the count of some alleles, a phenomenon known

as allelic bias, and can lead to the formation of chimeric reads, where two distinct template molecules are stitched together. Both artifacts can produce false SNVs or distort the counts of true SNVs. To address amplification-induced errors, quality control pipelines need to check for evidence of such artifacts and remove the affected reads. Filtering out reads with unusual lengths can help remove chimeric reads, whereas discarding low-quality reads (e.g., those with a Phred score below 20) can remove errors arising due to allelic bias. Subsequent steps in the MPL-R pipeline, such as read alignment, error modeling of Quasirecomb, and thresholding based on single mutant allele frequencies, can also help mitigate amplification-related errors.

Another source of error is target capture, which is used when the concentration of the target RNA in the sample is low. Briefly, target capture involves fragmenting the genome, incorporating probe molecules to facilitate attachment to magnetic beads, binding to the beads, washing away unbound material, and finally amplifying the captured fragments. Target capture can lead to over- or under-counting of SNVs which can affect subsequent analyses. Such biases can be mitigated by checking the coverage profile of the reads. If unexpected coverage fluctuations are observed, the affected region can be corrected or removed from subsequent analysis.

Reverse transcription also introduces errors due to the reverse transcriptase misincorporating erroneous nucleotides. Reverse transcription also exhibits errors due to template switching, wherein the enzymes disengage from one template molecule to engage with another. Template switching leads to chimeras that artificially create or break the covariance between mutations, affecting linkage-aware analysis. The artifacts created by reverse transcription errors can be mitigated via quality filtering prior to running the pipeline in which reads with ambiguous mapping and unusual fragment lengths are discarded.

Basecalling errors constitute another source of inaccuracy in the counts. These occur when the sequencing platform records an ambiguous signal and the basecalling software must make a

probabilistic decision in favor of a particular base, which may be incorrect. Basecalling errors can be mitigated during quality control by trimming low-quality regions of reads with a Phred score below some threshold. These errors can also be addressed during read alignment by enabling the quality control parameter of the aligner. Additionally, the filtering based on the minimum single mutant allele frequency threshold can also help exclude artifacts caused by basecalling errors.

In addition to basecalling errors, mapping bias can also introduce errors in allele counts. This occurs because reads containing mutations may map less well to the reference sequence compared to reads identical to the reference, leading to under-representation of mutant alleles. To mitigate mapping bias, the k-mer length during alignment can be reduced, which forces the alignment software to search for shorter matches. This ensures that reads with mutations are not discarded solely because of differences from the reference.

Finally, read duplication, where multiple reads originate from the same template molecule, can inflate the allele counts and distort LD estimates. Duplicate reads can be removed via standard duplicate removal software such as Picard [115] or Sambamba [116] prior to running the MPL-R pipeline.

2.2.3 Read aligning and sorting

The input to the pipeline consists of short-read data sampled at $K + 1$ time-points t_k, for $k \in \{0, \cdots, K\}$ from an evolving population. For each time-point, the reads are aligned against a user-supplied reference sequence using Burrows-Wheeler Aligner-Maximal Exact Matches (BWA-MEM) [117] and the resulting K Binary Alignment Map (BAM) files are sorted using samtools [118].

2.2.4 Bagging and population reconstruction

The next step performs population reconstruction from short-read data. This is a computation

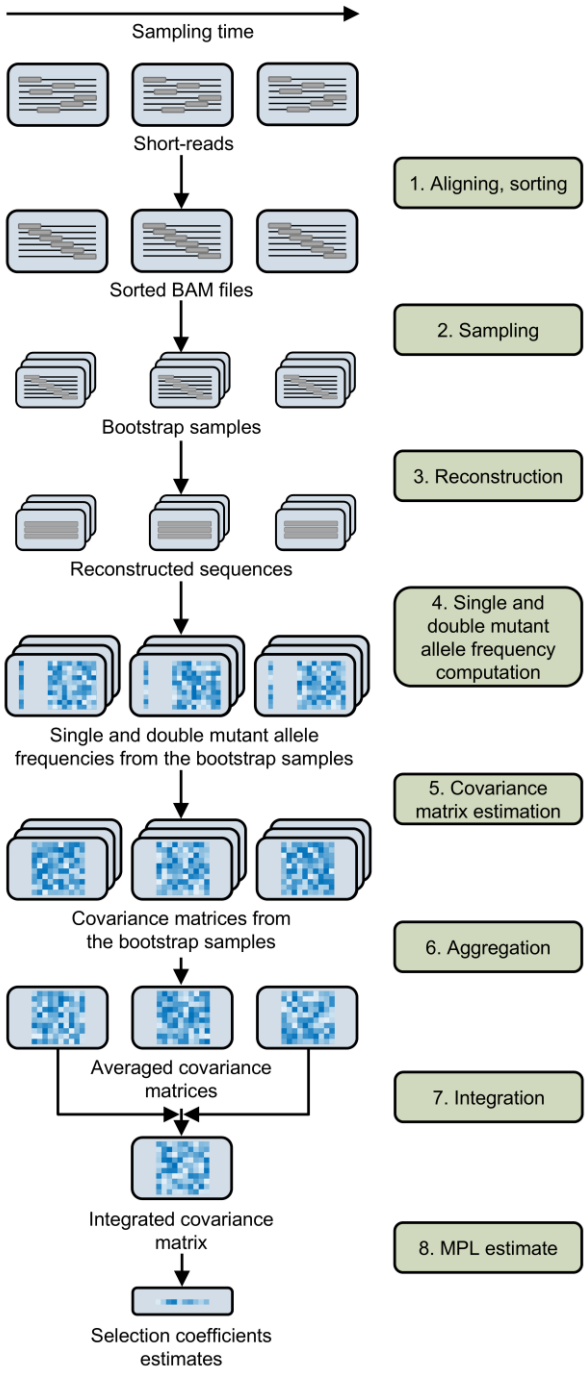


Figure 2.1 Illustration of the MPL-R pipeline. 1. Short reads from a longitudinal experiment are aligned to a reference sequence and sorted to obtain sorted Binary Alignment Map (BAM) files. 2. Bootstrap samples are generated from the sorted BAM files by random sampling with replacement. 3. Population reconstruction is performed on the bootstrap samples to obtain a set of full-length sequences and their corresponding frequencies in the population. 4. For each sampled time-point, the single and double mutant allele frequencies are obtained from the reconstructed sequences. 5. The covariance matrix of the mutant allele frequencies is computed for every bootstrap sample using the single and double mutant allele frequencies. 6. For each sampled time-point, the estimate of the covariance matrix of mutant allele frequencies of the actual population is obtained by averaging the covariance matrices of all M bootstrap samples. 7. The integrated covariance matrix (ICM), which quantifies the linkage effects, is obtained by scaling and summing these covariance matrices over all sampled time-points. 8. Accounting for linkage allows us to obtain accurate estimates of selection coefficients. The mutant allele frequencies obtained in Step 4 are also aggregated according to Equation 2.1 and used in Step 8, though not depicted here for simplicity.

and memory intensive procedure that presents a challenge, particularly as the number of reads becomes large. For efficient reconstruction, we pursue a bagging approach. We generate M bootstrap samples from each sorted BAM file such that each sample comprises of only a fraction of the total reads in the data (by random sampling with replacement). We then perform reconstruction using Quasirecomb on each bootstrap sample in parallel to obtain a probable set of full-length sequences constituting the population from which the short-read data is sampled.

This approach is memory efficient [119], while facilitating parallel processing to reduce the wall clock time required for the reconstruction process.

2.2.5 Estimation

The estimation step involves the calculation of a covariance matrix of mutant allele frequencies from each of the M reconstructed populations, each composed of N sequences of length L. The $L \times L$ covariance matrix $C^m(t_k)$ of the mutant allele frequencies of the m-th bootstrap sample has (i, j)-th entry

$$C_{ij}^m(t_k) = \begin{cases} x_i^m(t_k)\big(1 - x_i^m(t_k)\big), & i = j \\ x_{ij}^m(t_k) - x_i^m(t_k)x_j^m(t_k), & i \neq j, \end{cases}$$

where $x_i^m(t_k), x_{ij}^m(t_k)$ are the single mutant allele frequency and the double mutant allele frequency of the m-th bootstrap sample at locus i and locus-pair (i, j) at time t_k respectively. The covariance matrix of mutant allele frequencies, at time t_k, $C(t_k)$, is then obtained by averaging the covariance matrices of all M bootstrap samples (the aggregating step of bagging), i.e.,

$$C(t_k) = \frac{1}{M}\sum_{m=1}^{M} C^m(t_k).$$

Scaling and summing the covariance matrices of mutant allele frequencies over the first K sampled time-points yields the integrated covariance matrix (ICM), C_{int}, of size $L \times L$, given as

$$C_{\text{int}} = \sum_{k=0}^{K-1} \Delta t_k C(t_k),$$

where $\Delta t_k = t_{k+1} - t_k$. The ICM captures linkage effects in the data over the course of evolution and is used to obtain the estimates of the selection coefficients, $\hat{s}_i$, given by (Sohail *et al.*, 2021 [13])

$$\hat{s}_i = \sum_{j=1}^{L} (C_{\text{int}} + \gamma I)^{-1}_{ij} \times \left[x_j(t_K) - x_j(t_0) - \mu_{\text{fl},j} \right]. \qquad (2.1)$$

Here, I is an $L \times L$ identity matrix, $x_j(t_0)$, $x_j(t_K)$ are the single mutant allele frequencies at the first and the last sampled time-points of the j-th locus, respectively, and γ is a regularization parameter on the selection coefficients. Mathematically, $\gamma = 1/(N\sigma^2)$ where N is the population size and σ^2 is the variance of the Gaussian prior distribution for the selection coefficients (see Sohail *et al.*, 2021 [13] for more details). In practice, one does not need to set the value of σ^2 as it is already incorporated into the inference when the value of γ is chosen (Simulated data). The regularization parameter ensures that the ICM, C_{int}, is invertible. Regularization also suppresses the inference of large selection coefficients that are not well-supported by data. Moreover, $\mu_{\text{fl},j}$ is the integrated mutational flux, representing the net change in allele frequencies due to random mutations. It is given by

$$\mu_{\text{fl},j} = \mu_{\text{WT,mut}} \sum_{k=0}^{K-1} \Delta t_k \left(1 - x_j(t_k)\right) - \mu_{\text{mut,WT}} \sum_{k=0}^{K-1} \Delta t_k \, x_j(t_k),$$

where $\mu_{\text{WT,mut}}$ and $\mu_{\text{mut,WT}}$ are the mutation probabilities of the wild type (WT) allele to mutant allele and mutant allele to WT allele respectively. The single mutant allele frequencies of the j-th locus at time t_k, $x_j(t_k)$, are obtained by averaging the mutant allele frequencies of all M bootstrap samples,

$$x_j(t_k) = \frac{1}{M} \sum_{m=1}^{M} x_j^m(t_k). \qquad (2)$$

The number of polymorphic loci is smaller than the sequence length in real datasets. As only the polymorphic loci contribute towards parameter estimation, a practical methodology for parameter inference is to extract a submatrix and subvector corresponding to the polymorphic loci from the ICM and the numerator term in Equation 2.1 respectively. This approach is faster

with less storage requirements. In the pipeline, the ICM and the numerator term in Equation 2.1 are prefiltered based on loci which have mutant alleles appearing at frequencies of $\geq 1\%$ and inference is performed on this subset of loci. The mutant allele frequency at a locus is defined as the ratio between the number of reads carrying the mutation at the locus and the total number of reads covering the locus.

2.2.6 Estimation with a 4-allelic model

Estimation can be performed with a 4-allelic model by using Equation 12 from Sohail *et al.*, 2021 [13], which gives the estimate of the selection coefficient $\hat{s}_{i,\alpha}$ of allele α at locus i as

$$\hat{s}_{i,\alpha} = \sum_{j=1}^{L}\sum_{\beta=1}^{4}\left[\sum_{k=0}^{K-1}\Delta t_k C(t_k) + \gamma I\right]^{-1}_{ij,\alpha\beta} \times \left[x_{j,\beta}(t_K) - x_{j,\beta}(t_0) - \sum_{k=0}^{K-1}\Delta t_k \sum_{\delta=1}^{4}\{\mu_{\delta\alpha}x_{j,\delta}(t_k) - \mu_{\alpha\delta}x_{j,\alpha}(t_k)\}\right].$$

Here $x_{j,\beta}(t_K)$ and $x_{j,\beta}(t_0)$ represent the single mutant allele frequencies of allele β at locus j at the first and last time-points. $x_{j,\delta}(t_k)$ and $x_{j,\alpha}(t_k)$ respectively represent the single mutant allele frequency of the alleles δ and α at locus j at time-point t_k. $\mu_{\delta\alpha}$ is the mutation probability from allele δ to α and $\mu_{\alpha\delta}$ is the mutation probability from allele α to δ. The (i,j)-th entry of the covariance matrix for allele pair (α,β) is calculated for the m-th bootstrap sample as

$$C^m_{ij,\alpha\beta}(t_k) = \begin{cases} x^m_{i,\alpha}(t_k)\left(1 - x^m_{i,\beta}(t_k)\right), & i = j \\ x^m_{ij,\alpha\beta}(t_k) - x^m_{i,\alpha}(t_k)x^m_{j,\beta}(t_k), & i \neq j, \end{cases}$$

where $x^m_{i,\alpha}(t_k)$ and $x^m_{i,\beta}(t_k)$ are respectively the single mutant allele frequencies of alleles α and β of the m-th bootstrap sample at locus i at time t_k. $x^m_{ij,\alpha\beta}(t_k)$ represents the double mutant allele frequency of the allele pair (α,β) and locus-pair (i,j) at time t_k. $C(t_k)$ is obtained by averaging $C^m(t_k)$ over all M bootstrap samples as described in Section 2.2.4.

Before choosing between using a bi-allelic or 4-allelic model for a particular dataset, the respective entropy values $H_j^{bin}(t_k)$ and $H_j^{ter}(t_k)$ can be computed at each locus as

$$H_j^{bin}(t_k) = -\sum_{\beta \in \{WT, mut\}} x_{j,\beta}(t_k) \log x_{j,\beta}(t_k)$$

and

$$H_j^{ter}(t_k) = -\sum_{\beta=1}^{4} x_{j,\beta}(t_k) \log x_{j,\beta}(t_k) .$$

A comparison of $H_j^{bin}(t_k)$ and $H_j^{ter}(t_k)$ over all values of t_k can then be used to determine the information loss when using a bi-allelic model.

2.2.7 Simulated data

We simulated a Wright-Fisher model with selection and mutation for performing analytical validation. We considered a population of $N = 1000$ individuals, each represented by a $L = 500$ length binary string g_i^a, with zero representing the WT allele and one representing the mutant allele at locus i in sequence a. We note that the bi-allelic model used here is for ease of exposition and the framework readily extends to a 4-allelic model as shown in Section 2.2.5. We assumed an additive fitness model with the total fitness of sequence a given by

$$f_a = 1 + \sum_{i=1}^{L} g_i^a s_i,$$

where s_i is the selection coefficient of the mutant allele at locus i. The forward and backward mutation probabilities were assumed to be equal for simplicity and were set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral respectively. The beneficial and deleterious mutant alleles had the same magnitude of selection coefficients, but they had opposite signs, while the selection coefficients of neutral alleles had

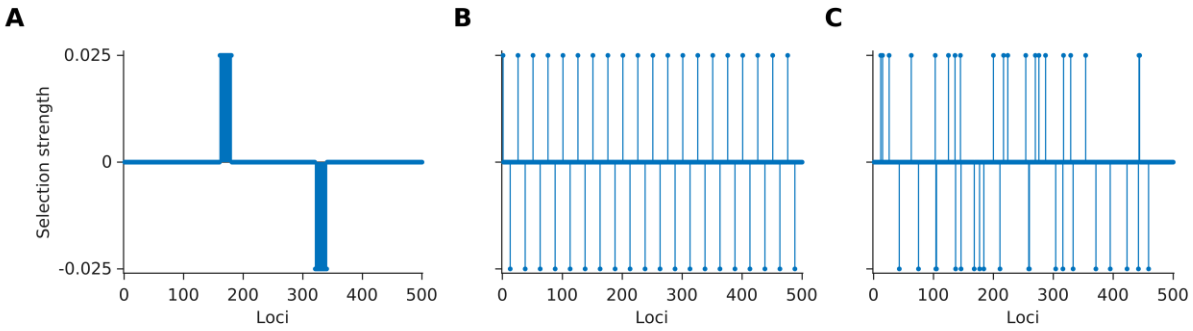


Figure 2.2 Structures of fitness landscape used in the simulations. **(A)** The data set "block" contained a single block of beneficial alleles grouped together at loci 161 to 180, a block of deleterious alleles at the loci 321 to 340, and neutral alleles at all the remaining loci. **(B)** The data set "comb" contained a repeating structure of 1 beneficial-11 neutral-1 deleterious-12 neutral alleles. **(C)** The data set "random" had the beneficial, deleterious, and neutral alleles distributed randomly across the length of the sequence.

a magnitude of zero. We simulated three structures of the underlying fitness landscape: "block" (Fig. 2.2A), "comb" (Fig. 2.2B), and "random" (Fig. 2.2C). Unless otherwise stated, the underlying fitness landscape used the "comb-like" structure. We generated five sets of data, each with a different value of the strength of selection in $\{0.005, 0.010, 0.025, 0.050, 0.075\}$, two sets with the probability of recombination 10^{-6} and 10^{-5}, two sets with $L = 1500, L = 3000$, and four sets with the number of founder sequences (sequences present in the population at the initial time-point) in the range 1 to 4. The founder sequences were randomly generated with the constraint that all founder sequences were within a Hamming distance of 10 to each other. We used 400 generations for inference with the population sampled at every $\Delta t = 50$ generations.

We simulated short-read sequencing of the population using the ART simulator [120]. At each time-point of the sets with $L = 500$, we generated $R = 15{,}000$ single-end reads of 150 bp using the error profile of the Illumina HiSeq 2500 platform, giving a coverage of 4500X. Throughout the text, coverage refers to the population-level coverage, defined as $C = lR/L$, where $C =$ coverage, $l =$ read length, $R =$ number of reads, and $L =$ genome length. We generated $R = 45{,}000$ and $R = 90{,}000$ reads at each time-point for the sets with $L = 1{,}500$

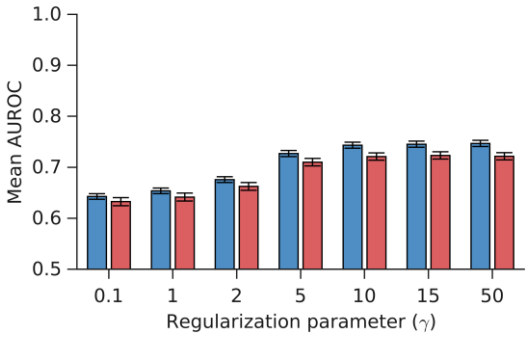


Figure 2.3 Exact value of the regularization parameter does not affect classification performance. The classification performance quantified as the mean AUROC is not sensitive to the value of the regularization parameter for values between 5 and 50. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations. Results are for MPL-R with $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point. The error bars indicate the standard error of the mean.

and $L = 3{,}000$ to ensure consistency of coverage. A total of 100 Monte Carlo runs were generated for each data set.

For inference, we generated $M = 3$ bootstrap samples by sampling (with replacement) from each sorted BAM file, with each bootstrap sample containing a third as many reads, i.e., $R_b = R/3 = 5000, 15000, 30000$, for $L = 500, 1500, 3000$, as contained in the sorted BAM file. This amounted to a coverage of 1500X. The regularization parameter was set to $\gamma = 10$; though our results were insensitive to this choice, with MPL-R showing similar performance for a wide range of regularization values (Fig. 2.3). We measured the performance of MPL-R to detect beneficial mutant alleles from the rest (neutral/deleterious) and deleterious mutant alleles from the rest (neutral/beneficial) in terms of the area under the receiver operating characteristic curve (AUROC). The AUROC was calculated on selection coefficients corresponding to mutant alleles that appeared at frequencies of $\geq 1\%$. On average, the mutant

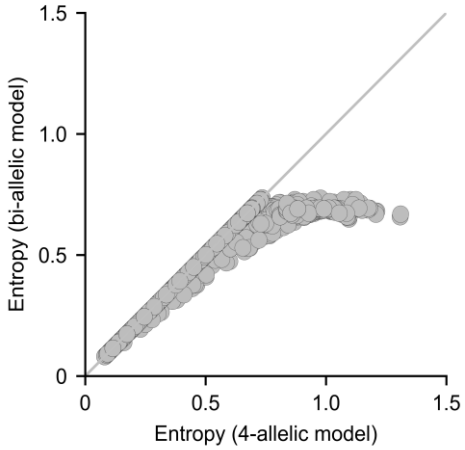


Figure 2.4. Bi-allelic model for the HIV-1 data results in loss of information. Entropy values of the single mutant allele frequencies computed via a bi-allelic model show marked deviation from the entropy values computed via a 4-allelic model.

allele frequency trajectories of 14 of the 20 deleterious mutations crossed this threshold for the single-founder dataset with $L = 500$, $s = 0.050$, and no recombination.

2.2.8 HIV-1 data

We performed technical validation on a real dataset of HIV-1 *in vivo* evolution to show the ability of our proposed pipeline to accurately infer selection from short-read data. As a benchmark, we used the analysis of the half-genome length time-series HIV-1 sequence data performed in Sohail *et al.*, 2021 [13]. The HIV-1 dataset and its pre-processing were explained in detail in Sohail *et al.*, 2021 [13]. Briefly, the dataset comprised Sanger-sequenced data from 13 HIV-1 infected patients sampled at different time-points as part of the Center for HIV/AIDS Vaccine Immunology (CHAVI 001) and Centre for the AIDS Programme of Research in South Africa (CAPRISA 002) studies in the United States, Malawi and South Africa. Sequence data was downloaded from Los Alamos National Laboratory HIV Sequence Database (www.hiv.lanl.gov). The sampling times and number of sequences at each time-point are described in Supplementary Table 1 in Sohail *et al.*, 2021 [13]. The population size N, was not explicitly assumed as it was already incorporated into the value of γ. Model selection was performed by comparing the entropy values of the single mutant allele frequencies for both bi-

Ref nt \ Mut nt	A	C	G	T	- (gap)
A	0	9.0×10^{-7}	6.0×10^{-6}	7.0×10^{-7}	1.0×10^{-9}
C	5.0×10^{-6}	0	5.0×10^{-7}	1.2×10^{-5}	1.0×10^{-9}
G	1.6×10^{-5}	1.0×10^{-7}	0	2.0×10^{-6}	1.0×10^{-9}
T	3.0×10^{-6}	1.0×10^{-5}	3.0×10^{-6}	0	1.0×10^{-9}
- (gap)	1.0×10^{-9}	1.0×10^{-9}	1.0×10^{-9}	1.0×10^{-9}	0

Table 2.1. Mutation rates used in the analysis of HIV-1 patient data. Each entry represents the mutation rate from the nucleotide in the row to the nucleotide in the column.

allelic and 4-allelic models and a 4-allelic model was chosen (Fig. 2.4). Mutation probabilities for all nucleotide transitions were obtained from Figure 1C in Zanini *et al.*, 2017 [82] and mutation probabilities to and from gap states were assumed to be 1×10^{-9} per locus per day following Sohail *et al.*, 2021 [13]. These values have been summarized in Table 2.1. We analyzed the group specific antigen (gag) protein to demonstrate the functionality of MPL-R.

We simulated sequencing of the gag protein (length 1,500bp) using the ART simulator with the error profile of the Illumina HiSeq 2500 to generate single-end reads of length 150bp. For ease of exposition, we set the sequencing parameters to 15 reads per sequence and 10,000 sequences resulting in $R = 150{,}000$ reads and an average coverage of 13,500 X. After aligning and sorting the reads, we obtained $M = 3$ bootstrap samples for each time-point by sampling with replacement keeping the bootstrap sample size $R_b = R/3 = 50{,}000$.

Population reconstruction was performed as explained previously. The reconstructed sequences from each bootstrap sample were supplied to the MPL-R pipeline, preprocessed, and analyzed using the same parameter values used in Sohail *et al.*, 2021 [13] to obtain the ICM and the numerator term from Equation 2.1. The estimates of the ICM and the numerator term were averaged over all bootstrap samples, and the estimates of the selection coefficients obtained by the utilization of Equation 2.1. We also obtained the estimates of the selection coefficients from the full-length gag sequences via MPL for benchmarking.

The performance of MPL-R was compared against MPL via the proportion of biologically important mutations (those residing in CD8+ T cell epitopes, or nonsynonymous reversions

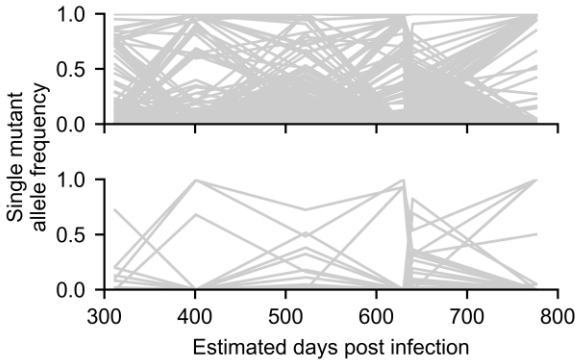


Figure 2.5 SARS-CoV-2 data indicates presence of linkage. Single mutant allele frequency trajectories obtained from the complete SARS-CoV-2 genome (*top* panel) and RBD (*bottom* panel) show multiple cooccurring mutations, suggesting the presence of genetic linkage.

outside CD8+ T cell epitopes, or nonsynonymous reversions within CD8+ T cell epitopes) in the top 5% strongest selection coefficients. Statistical significance of the proportions of biologically important mutations was computed via fold enrichment, which quantifies the likelihood of obtaining a particular proportion by pure chance. Mathematically, if n_{obs} mutations of a particular property are observed in a category with N_{cat} mutations, and there are n_{tot} mutations of that particular property in the total N_{tot} mutations, then the ratio $(n_{\text{obs}}/N_{\text{cat}})/(n_{\text{tot}}/N_{\text{tot}})$ specifies the fold enrichment. The greater the value of fold enrichment relative to 1, the higher the confidence that n_{obs} is not observed by chance. Two-sided Fisher's exact test was used to calculate the p-values of all values of fold-enrichment.

2.2.9 SARS-CoV-2 data

To demonstrate the working of the pipeline on another dataset, we performed biological concordance analysis on a real dataset of SARS-CoV-2 that was evolving in an immunocompromised patient infected with HIV [52]. The dataset consisted of 6 samples of 70bp paired-end reads sequenced via the Illumina NextSeq 500 platform from 312 to 776 days since the estimated date of SARS-CoV-2 infection. The average number of reads *R*, was 2,786,519 with a standard deviation of 1,008,062, which translates to a coverage of 4,658X with a standard deviation of 1,242X. The receptor-binding domain (RBD) of the Spike

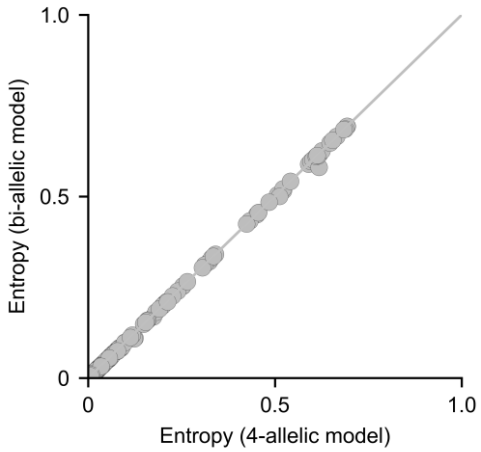


Figure 2.6. Bi-allelic model for the SARS-CoV-2 data shows minimal loss of information. Entropy values of the single mutant allele frequencies computed via a bi-allelic model show near-complete mapping to the entropy values computed via a 4-allelic model.

glycoprotein was chosen for analysis as it harbors biologically important mutations associated with antibody escape [121], [122], [123]. The coordinates of the RBD were obtained from Barnes *et al.*, 2020 [124]. A plot of the single mutant allele frequencies (Fig. 2.5) reported multiple simultaneous mutations across the genome and the RBD, which indicated linkage, and suggested that linkage-aware analysis could be informative. The reads were aligned against the SARS-CoV-2 reference sequence (GenBank accession no. MN908947.3) and sorted. $M = 3$ bootstrap samples were obtained at each time-point consisting of an average of $R_b = R/3$ reads as described before. Population reconstruction was performed between loci $22{,}544$ to $23{,}161$ to obtain full-length sequences of length 618bp of the RBD. A bi-allelic model of WT/mutant was chosen based on the entropy values of the single mutant allele frequencies (Fig. 2.6) and selection coefficients were estimated as described before. The mutation probability $\mu = 1.85 \times 10^{-6}$ per locus per day was obtained from Velasquez-Reyes *et al.*, 2025 [52] and used in estimation. The mutation probability from the WT allele to mutant allele was assumed equal to μ and the mutation probability from the mutant allele to WT allele was assumed equal to 3μ. The value of the regularization parameter was taken as $\gamma = 10$. Estimation was performed on only those loci where the single mutant allele frequency was $\geq 1\%$. Synonymous and non-

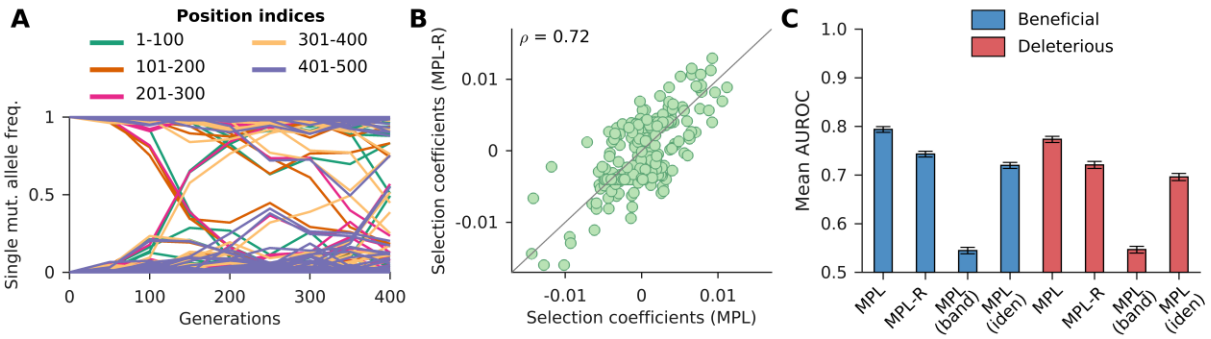


Figure 2.7 MPL-R has good classification performance. **(A)** Plot of single mutant allele frequency trajectories indicates the presence of polymorphic alleles across the entire length of the sequence. **(B)** Selection coefficients estimated from full-length sequences (i.e., all double mutant allele frequencies known) and those estimated from reconstructed double mutant allele frequencies were highly correlated (Pearson correlation coefficient $\rho = 0.72$ with a p-value $< 10^{-100}$). Results shown here are for a typical Monte Carlo run. **(C)** The mean AUROC performance of distinguishing beneficial mutant alleles from the rest (deleterious and neutral), and deleterious mutant alleles from the rest (neutral and beneficial) of the proposed method (MPL-R) is compared against three methods: MPL, MPL (banded), and MPL (identity covariance). Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations. Results for MPL-R are with $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point. The error bars indicate the standard error of the mean.

synonymous mutations were classified via sc2calc [125]. After estimation, the performance of MPL-R was demonstrated by its ability to correctly identify known beneficial mutations in the RBD.

2.3 Results

2.3.1 Performance on simulated data

The proposed computational pipeline MPL-R infers selection from short-read time series data. Initially, we tested the pipeline on simulated time-series data (Methods). For comparison, we also evaluated two alternative schemes which did not require population reconstruction. The first scheme, MPL (banded), used the double mutant allele frequencies directly observable from the short-read data (i.e., the intra-read double mutant allele frequencies) to compute the

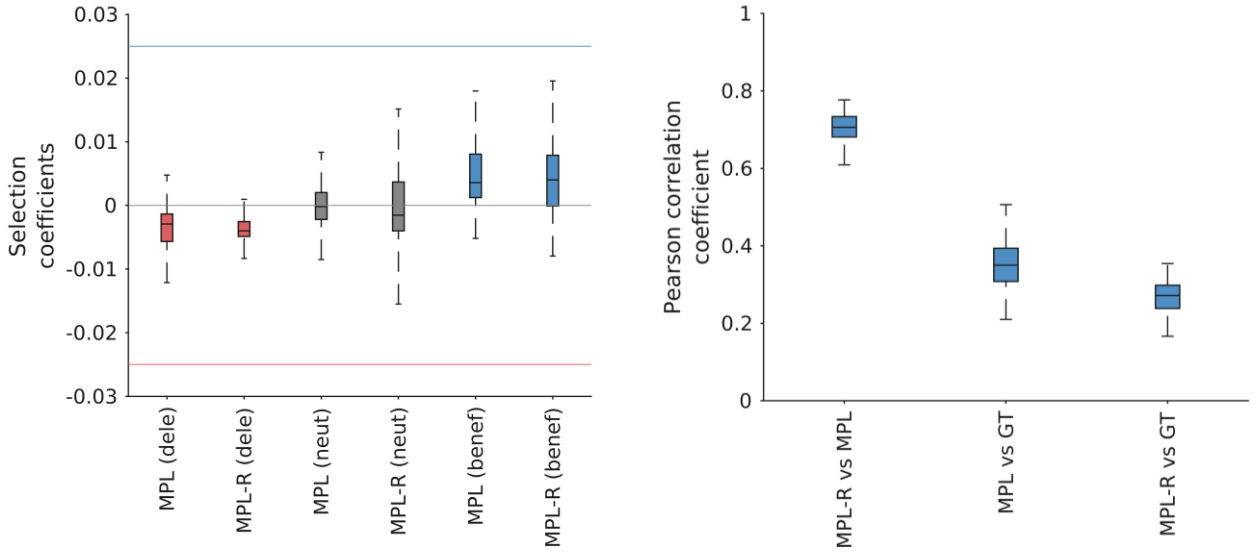


Figure 2.8. Estimated selection coefficients show good classification performance and correlation with each other, but poor correlation with the ground truth ones. Boxplots of the estimated selection coefficients are shown for 100 Monte Carlo runs (*left* panel). The horizontal lines represent the ground truth selection values with the colors corresponding to the individual categories. The Pearson correlation coefficient between the ground truth (GT) and estimated selection coefficients and between the estimated selection coefficients is shown for 100 Monte Carlo runs as well (*right* panel). Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations.

covariance matrices and set the covariance matrix entries corresponding to the inter-read allele pairs to zero. The second scheme, termed MPL (identity covariance), used only the single mutant allele frequency information, and considered the loci to be under independent evolution. For benchmarking, we used MPL (Sohail *et al.*, 2021) [13] on the ground truth full-length sequences, with access to complete knowledge of the single and double mutant allele frequencies.

We considered a system with 500 loci that had simultaneously polymorphic loci spanning the entire length of the sequence. Analysis of the simulated data showed that only around 10% of

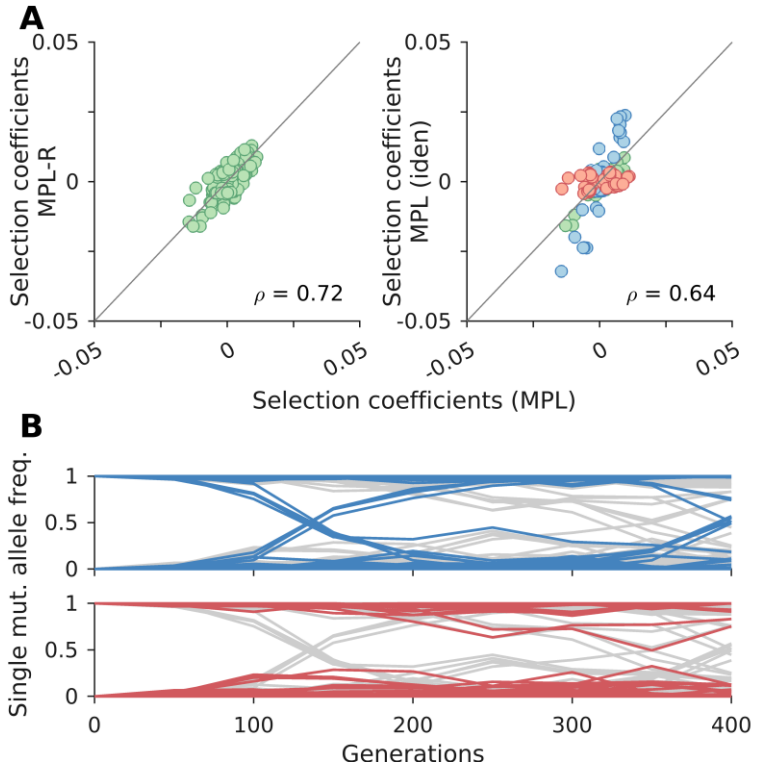


Figure 2.9 MPL (identity covariance) gets outcompeted by MPL-R for loci evolving under genetic linkage: **(A)** Selection coefficients estimated from MPL-R and MPL (*left* panel) showed higher correlation (Pearson correlation coefficient $\rho = 0.72$ with p-value $= 1.49 \times 10^{-58}$) compared to the selection coefficients obtained from MPL (identity covariance) and MPL (*right* panel) (Pearson correlation coefficient $\rho = 0.64$ with p-value $= 1.28 \times 10^{-43}$). MPL (identity covariance) overestimates some selection coefficients (blue) and underestimates some (red). **(B)** Single mutant allele frequency trajectories for overestimated selection coefficients show a rising/falling trend (*top* panel). Single mutant allele frequency trajectories of underestimated selection coefficients start and end near a boundary. Results shown here are for a typical Monte Carlo run which consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations. Results for MPL-R are with $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point.

pairs of polymorphic loci lay within one read length (150 bp) of each other; that is, most loci contributing to linkage effects are out-of-read polymorphic loci not directly observable with short-read data (Fig. 2.7A). We performed analysis using the schemes described and obtained estimates of the selection coefficients.

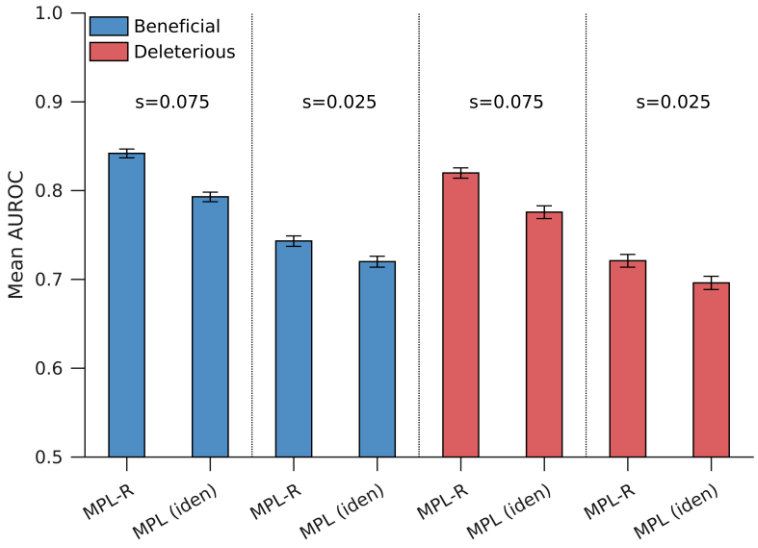


Figure 2.10 Performance improvement of MPL-R over MPL (identity covariance) is most significant for stronger selection coefficients: Plot of the mean AUROC for beneficial loci (*left* sections) and deleterious loci (*right* sections) shows that the difference between the classification accuracy of MPL-R and MPL (identity covariance) increase with increase in the level of linkage. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ or $+0.075/-0.075/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations. Results for MPL-R are with $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point. The error bars indicate the standard error of the mean.

Our results showed a high positive correlation ($\rho = 0.72$, p-value $< 10^{-100}$) between the selection coefficients estimated by the proposed MPL-R method and those estimated by the benchmark MPL method (Fig. 2.7B). Boxplots of the selection coefficients estimated by MPL-R and MPL showed clustering in the correct respective categories [Fig. 2.8 (*left* panel)], and summary statistics indicated that the estimates of MPL-R showed strong correlation with the estimates of MPL [Fig. 2.8 (*right* panel)]. The estimates of MPL-R and MPL clustered close to zero because of strong regularization [Fig. 2.8 (*left* panel)] and exhibited regression error with respect to the ground truth selection coefficients [Fig. 2.8 (*right* panel)], however, in terms of classification performance, both methods performed well. We observed that the

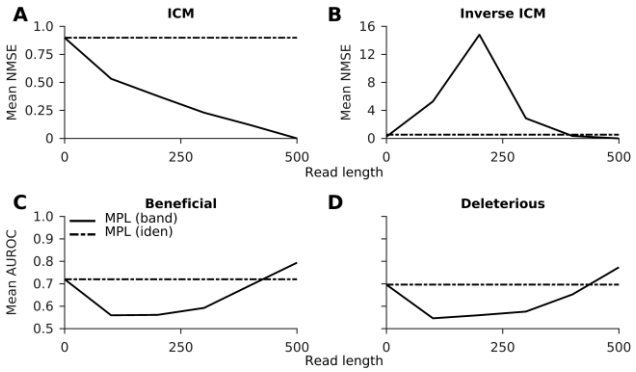


Figure 2.11. MPL (banded) performs poorly regardless of read length: **(A)** The mean NRMSE of the ICM decreases with an increase in the read length. **(B)** The mean NRMSE of the inverse ICM first increases, then decreases with an increase in the read length. Plots of the mean AUROC for beneficial mutant alleles **(C)** and deleterious mutant alleles **(D)** show that MPL (banded) has poor classification performance compared to MPL (identity covariance). Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations.

classification performance of the MPL-R method was close to that of the MPL method and was better than MPL (identity covariance) and MPL (banded) (Fig. 2.7C). The reduced performance of MPL-R as compared to MPL is because inference from short-read data has several sources of errors (sequencing errors, non-uniform sequencing coverage, and population reconstruction errors) not encountered in case of MPL (which assumes complete and accurate knowledge of full-length sequences). MPL (identity covariance) underperforms because it does not explicitly account for genetic linkage effects; thus, it over/underestimates selection coefficients of mutant alleles evolving under genetic linkage (Fig. 2.9). Specifically, MPL (identity covariance) overestimates the selection coefficients of neutral alleles hitchhiking on beneficial ones and underestimates the magnitudes of the selection coefficients of mutant alleles evolving under clonal interference or negative selection. Incorporating the inferred inter-read covariances allows MPL-R to account for genetic linkage information and obtain more accurate selection coefficient estimates. This performance difference becomes more significant for selection

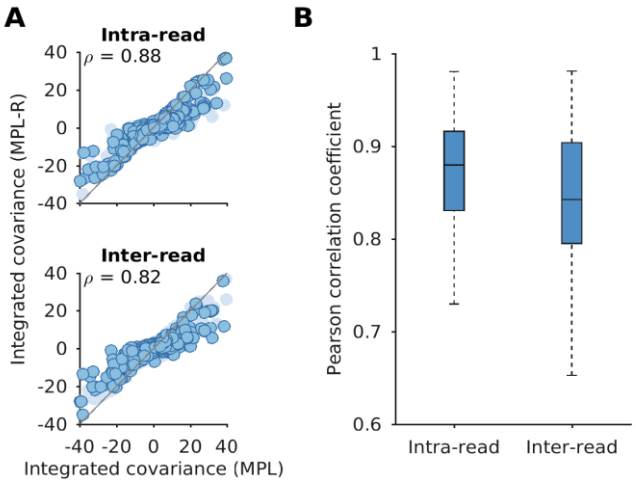


Figure 2.12. MPL-R recovers linkage effects across the entire sequence. The entries of the ICM calculated from the reconstructed double mutant allele frequencies exhibit high correlation with the entries of the ICM calculated from the double mutant allele frequencies obtained from the full-length sequences. **(A)** Intra-read covariance values (*top* panel) show good correlation (Pearson correlation coefficient $\rho = 0.88$ with p-value $< 10^{-100}$). The inter-read covariance values (*bottom* panel) are highly correlated as well which demonstrate the ability of the method to infer linkage patterns not observable in individual short-reads (Pearson correlation coefficient $\rho = 0.82$ with p-value $< 10^{-100}$). In each panel, the dark markers indicate the entries of the ICM in that category, and the light markers denote the rest of the entries. Results are shown here for a typical Monte Carlo run. **(B)** The summary statistics of the Pearson correlation coefficient for inter and intra-read covariance values indicate a consistent trend across 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations. Results are for MPL-R with $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point.

coefficients with larger magnitudes (Fig. 2.10), due to the stronger linkage effects in these cases. A more detailed explanation of the performance difference between MPL-R and MPL (identity covariance) is given in Section 2.3.2.

Surprisingly, we observe that the MPL (banded) scheme, which attempts to account for within-read linkage effects while ignoring out-of-read effects, is outperformed by the MPL (identity covariance) scheme which ignores linkage effects completely. This suggests that the naïve approach of applying the MPL framework by substituting observable (within-read) double mutant allele frequencies and setting all unobservable double mutant allele frequencies to the

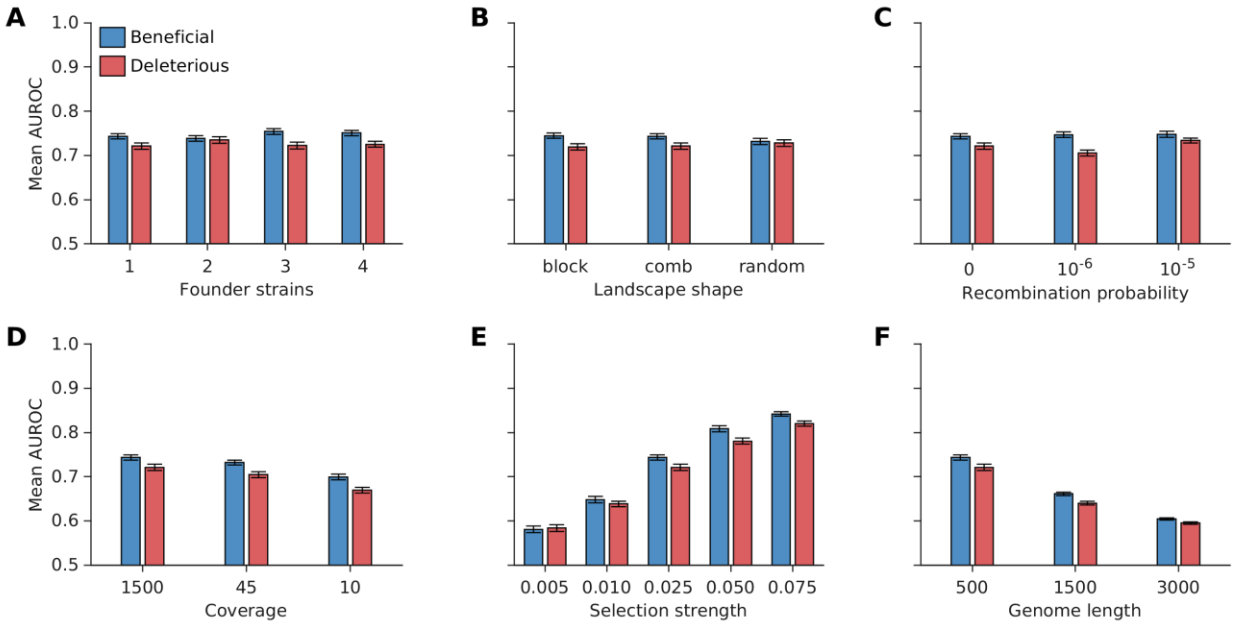


Figure 2.13. Robustness of MPL-R to changes in diversity in data, distribution of fitness effect across the length of the sequence, recombination probability, strength of selection, length of the genome, and coverage. The classification performance of MPL-R quantified as the mean AUROC remained largely unaffected for **(A)** increase in population diversity, controlled here by varying the number of founder strains (frequency of each founder $\geq 10\%$), **(B)** various structures of the underlying fitness landscape, **(C)** increase in recombination probability, and **(D)** sequencing coverage. The classification performance was affected by **(E)** decrease in selection strength, and **(F)** increase in the length of the genome. The structure of the data sets "comb", "block", and "random" is presented in Fig. 2.2. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Unless mentioned otherwise, alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. Unless mentioned otherwise, the fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. Unless mentioned otherwise, one founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations. Results are for MPL-R with $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point. The error bars indicate the standard error of the mean.

product of the single mutant allele frequencies, is not a desirable approach. This has been discussed in detail in Section 2.3.3. Briefly, while the ICM in MPL (banded) captures more information from the ICM obtained using the full-length sequences compared to the ICM in MPL (identity covariance), the same is not true for the respective inverse ICMs (Fig. 2.11). Because of the inverse ICM term in Equation 1, MPL (identity covariance) outperforms MPL (banded). This result highlights the need for reconstructing full-length sequences from short-

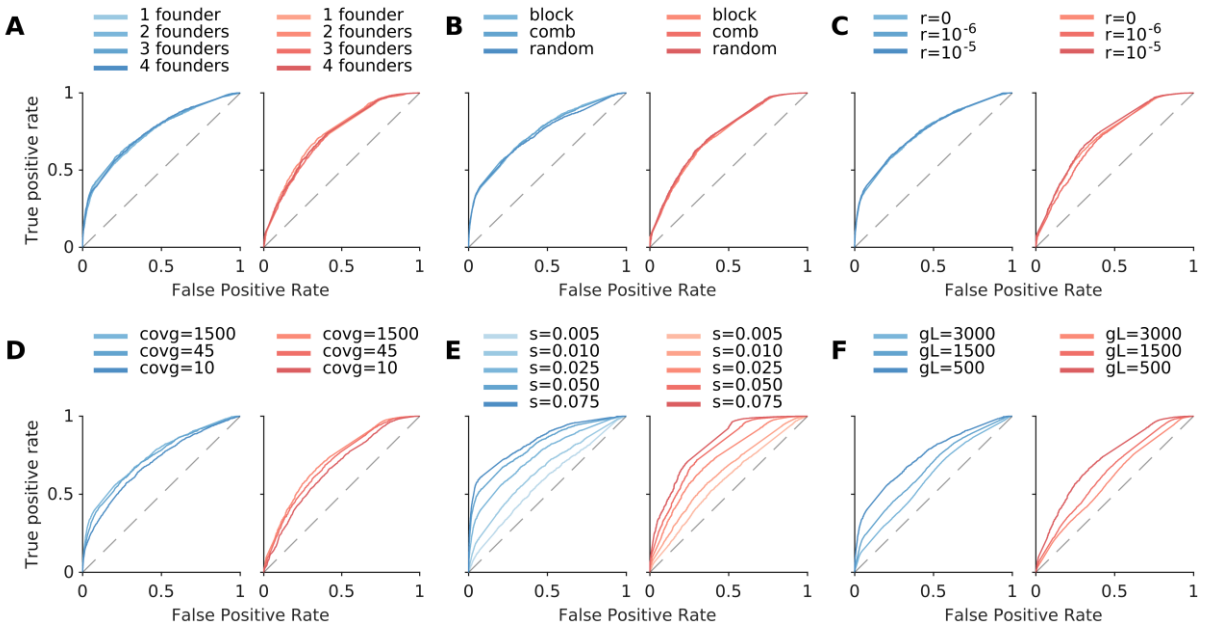


Figure 2.14. MPL-R consistently performs better than a random predictor when applied to datasets with changing diversity in data, distribution of fitness effect across the length of the sequence, recombination probability, strength of selection, length of the genome, and coverage. The classification performance of MPL-R quantified as the mean receiver operating characteristic curve (ROC) consistently remained better than a random predictor for **(A)** increase in population diversity, controlled here by varying the number of founder strains (frequency of each founder $\geq 10\%$), **(B)** various structures of the underlying fitness landscape, **(C)** increase in recombination probability, **(D)** sequencing coverage, **(E)** decrease in selection strength, and **(F)** increase in the length of the genome. The structure of the data sets "comb", "block", and "random" is presented in Fig. 2.2. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Unless mentioned otherwise, alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. Unless mentioned otherwise, the fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. Unless mentioned otherwise, one founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations. Results are for MPL-R with $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point.

read data to infer unobserved inter-read covariance values, as incomplete linkage information obtained from short-read data (only using intra-read covariance information) is not sufficient for accurate inference of selection coefficients.

The reconstruction step of MPL-R (Step 3 in Fig. 2.1) inferred the inter-read double mutant allele frequencies. These, in turn, were used to compute the ICM, which accounts for the genetic linkage present in the data. Our results showed that the proposed MPL-R method was

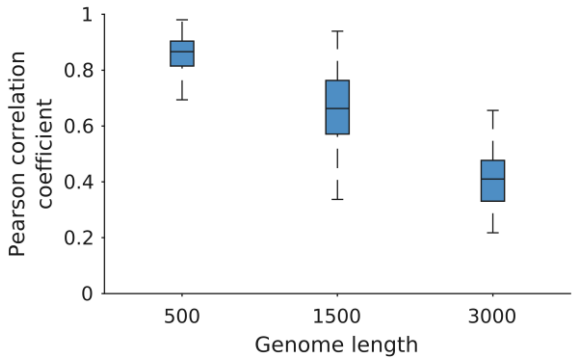


Figure 2.15. Increasing the genome length decreases the accuracy of the inferred covariance values. The Pearson correlation coefficient between the reconstructed and ground truth covariance values indicate a consistent decreasing trend with the increase in genome length. The results are plotted for 100 Monte Carlo runs, each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of bi-allelic (WT and mutant) loci of the indicated genome length, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$, $60/60/1380$, and $120/120/1260$, loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations. Results are for MPL-R with $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point.

able to accurately account for linkage effects across the length of the sequence, faithfully inferring ICM entries for both the intra and inter-read pairs (Fig. 2.12).

Further simulations showed that the performance of MPL-R was robust to the level of diversity in the data (Fig. 2.13A and Fig. 2.14A), the distribution of fitness effects across the length of the sequence (Fig. 2.13B and Fig. 2.14B), the level of recombination (Fig. 2.13C and Fig. 2.14C), and sequencing coverage (Fig. 2.13D and Fig. 2.14D). The performance changed with changes in selection strength (Fig. 2.13E and Fig. 2.14E), and the length of the genome (Fig. 2.13F and Fig. 2.14F). The robustness to sequencing coverage (Fig. 2.13D and Fig. 2.14D) needs to be interpreted while considering that reads are generated from the entire population of size $N = 1000$. If the reads are obtained from a subset of the population, as done in real sequencing runs, the performance may show greater degradation as the coverage is reduced. The degradation of performance at low values of the strength of selection in Fig. 2.13E and Fig. 2.14E is because very weak selection becomes hard to differentiate from neutral evolution,

as the fluctuations in the mutant allele frequencies due to selection and random genetic drift become challenging to distinguish. The degradation associated with the increase in the length of the genome (Fig. 2.13F and Fig. 2.14F) is explained by the fact that more linkage information becomes unobservable as the genome length increases relative to the read length. Population reconstruction and consequently inference of the unobservable values in the ICM become more difficult for larger sequences observed with a fixed read length. Fig. 2.15 demonstrates the degradation in the estimates of the ICM with the increase in the length of the genome while keeping the read length constant.

2.3.2 Comparison between MPL-R and MPL (identity covariance)

The selection coefficients estimated by MPL-R are relatively strongly correlated with those estimated by MPL (Pearson correlation coefficient $\rho = 0.72$ with p-value $= 1.49 \times 10^{-58}$) [Fig. 2.9A (*left* panel)] as compared to the selection coefficients estimated by MPL (identity covariance) (Fig. 2.9A (*right* panel)) (Pearson correlation coefficient $\rho = 0.64$ with p-value $= 1.28 \times 10^{-43}$). We can explain the performance difference between MPL-R and MPL (identity covariance) by analyzing the loci for which the latter performs comparatively worse. The relative error between the estimates of MPL (identity covariance) and MPL, and between MPL-R and MPL is given by

$$e_{i,\mathrm{MPL\ (iden)}} = \left|\left(s_{i,\mathrm{MPL}} - s_{i,\mathrm{MPL\ (iden)}}\right)/s_{i,\mathrm{MPL}}\right|,$$

and

$$e_{i,\mathrm{MPL-R}} = \left|\left(s_{i,\mathrm{MPL}} - s_{i,\mathrm{MPL-R}}\right)/s_{i,\mathrm{MPL}}\right|$$

respectively. $s_{i,\mathrm{MPL\ (iden)}}$, $s_{i,\mathrm{MPL-R}}$, and $s_{i,\mathrm{MPL}}$ denote the selection coefficient for the mutant allele at locus i estimated by MPL (identity covariance), MPL-R, and MPL respectively, $e_{i,\mathrm{MPL\ (iden)}}$ and $e_{i,\mathrm{MPL-R}}$ represent the relative error in the estimate of MPL (identity

covariance) and MPL-R with respect to the estimate of MPL at locus i. $|\cdot|$ represents the absolute value. We first isolate all loci for which $e_{i,\mathrm{MPL\ (iden)}} > e_{i,\mathrm{MPL-R}}$, and then further classify those loci based on whether $s_{i,\mathrm{MPL\ (iden)}}$ is over or underestimated relative to $s_{i,\mathrm{MPL}}$. The scatterplot in Fig. 2.9A (*right* panel) shows examples of overestimation $\left(\left|s_{i,\mathrm{MPL\ (iden)}}\right| > \left|s_{i,\mathrm{MPL}}\right|\right)$ highlighted as blue and examples of underestimation $\left(\left|s_{i,\mathrm{MPL\ (iden)}}\right| < \left|s_{i,\mathrm{MPL}}\right|\right)$ highlighted as red. The corresponding trajectories for the over and underestimated selection coefficients are presented in the same color in Fig. 2.9B against all other trajectories colored gray. It can be observed that MPL (identity covariance) mostly overestimates the selection coefficients of loci for which the single mutant allele frequency trajectories are increasing or decreasing with time [Fig. 2.9B (*top* panel)]. This corresponds to the situation in which only looking at the term $x_j(t_K) - x_j(t_0)$ in Equation 2.1 suggests a large value of the selection coefficient because of the large difference between the initial and final values of the trajectory, but the actual value of the selection coefficient, after correction for linkage, is small. These trajectories represent single mutant alleles that are inherently neutral but are hitchhiking on neighboring alleles because of genetic linkage. As MPL (identity covariance) does not correct for linkage, it overestimates the selection coefficients at these loci. It can also be observed that MPL (identity covariance) mostly underestimates the selection coefficients of loci for which the single mutant allele frequency trajectories start near the fixation/elimination boundary, increase or decrease for some time, and then end up at or close to the initial fixation/elimination boundary again [Fig. 2.9B (*bottom* panel)]. This corresponds to the situation in which only looking at the term $x_j(t_K) - x_j(t_0)$ in Equation 2.1 suggests the selection coefficient to be zero because the initial and final values of the trajectories are similar, but the actual value of the selection coefficient, after correction for linkage, is non-zero. These trajectories represent single mutant alleles that are under positive selection but get outcompeted by alleles with stronger values of the selection coefficient (clonal interference) or are under negative selection

but rise because of genetic linkage with positively selected neighboring alleles. Again, as MPL (identity covariance) does not correct for linkage, it underestimates the selection coefficients at these loci.

Trajectories driven by stronger selection coefficients can be hypothesized to increase the difference between MPL-R and MPL (identity covariance). To demonstrate the effect of the selection strength on the estimates of MPL-R and MPL (identity covariance), we compare the results of the datasets with $s = 0.075$ and $s = 0.025$ respectively. We can observe that the difference between MPL-R and MPL (identity covariance) is greater when $s = 0.075$ as compared to when $s = 0.025$ (Fig. 2.10). The dataset with $s = 0.075$ has a higher magnitude of the selection coefficients, and when MPL (identity covariance) underestimates a large selection coefficient, the error is of a greater magnitude compared to the error that occurs when underestimates the selection coefficients for the dataset with $s = 0.025$.

2.3.3 Naïve estimation of ICM from reads

Here we describe a counterintuitive observation we made about the behavior of the ICM used in MPL (banded). MPL (banded) incorporates more information about the covariance between mutant allele frequencies than MPL (identity covariance), leading to a better estimate of the ICM. This can be seen from Fig. 2.11A, which plots the normalized root mean square error (NRMSE) of the banded ICM as a function of increasing read-length (i.e., band length). Here, we calculate the NRMSE as

$$NRMSE = \left|\left|ICM_{\text{full}} - ICM_{\text{band}}\right|\right| / \left|\left|ICM_{\text{full}}\right|\right|,$$

where ICM_{full} denotes the ICM computed from full-length sequences, ICM_{band} denotes the ICM of the banded case computed from the reads, and $||\cdot||$ denotes the Frobenius norm. When the read-length is zero in Fig. 2.11A, the ICM of the banded case (solid line) reduces to the diagonal matrix of the identity covariance case (dashed line). While when the read-length is equal to the sequence length ($L = 500$) in Fig. 2.11A, $ICM_{\text{band}} = ICM_{\text{full}}$. That is,

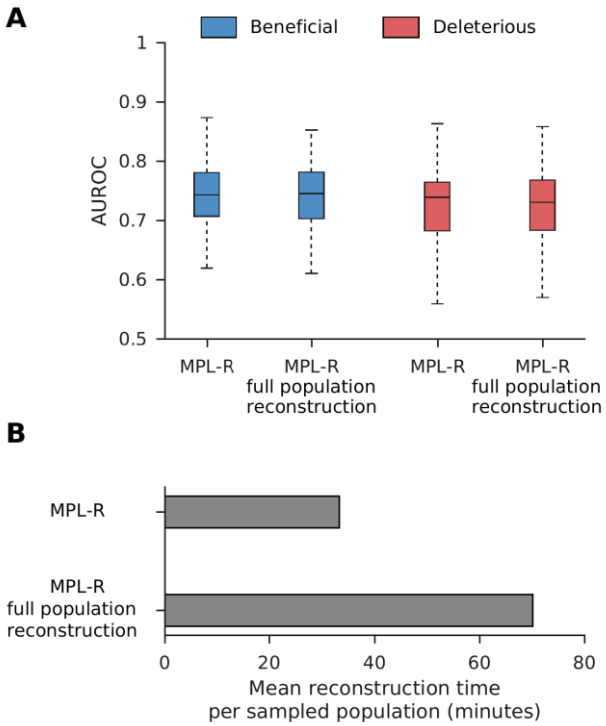


Figure 2.16. Bagging approach to reconstruction improves the wall clock time without affecting the classification performance. **(A)** Classification performance of the MPL-R method with bagging compared with a version that does not use bagging. Results of MPL-R are for $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point used in MPL-R (full population reconstruction). **(B)** Average wall clock time of the reconstruction step of the proposed MPL-R method with bagging compared with MPL-R (full population reconstruction). Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations.

incorporating more data (due to increased read length) successively improves the estimate of the ICM.

However, note that the estimate of the selection coefficients in Equation 2.1 of the thesis involves an inverse of the regularized ICM. Plotting the NRMSE of the inverse of the regularized ICM as a function of read length shows that the minimum NRMSE occurs for the $ICM_{\text{band}} = ICM_{\text{full}}$ case as expected (Fig. 2.11B). Interestingly, we see that the NRMSE of the identity covariance case is lower than the banded covariance case. This is likely due to the non-

linear nature of the inverse function. This inaccuracy in the inverse of the regularized ICM in the banded case at intermediate read-lengths leads to poor performance of MPL (banded) in classifying beneficial/deleterious mutants from the rest (Fig. 2.11C and D). While it would be interesting to study this phenomenon in more detail, a more rigorous analysis is beyond the scope of the current manuscript.

2.3.4 Parallelized population reconstruction

The reconstruction step (Step 3 in Fig. 2.1) of MPL-R is by far the most time-consuming. However, the results above show that this step is necessary to resolve the effects of genetic linkage. The bagging approach used by MPL-R provides an efficient way to complete this step by forming M bootstrap samples at each sampled time-point, where each bootstrap sample consists of a fraction of the total number of reads available at that time-point. These bootstrap samples can be reconstructed on separate processors in parallel. Our simulations showed that the bagging approach employed by MPL-R resulted in a similar classification performance (Fig. 2.16A), and improved wall clock time as compared to MPL-R (full population reconstruction)—a scheme that performs population reconstruction on all the reads available at a sampled time-point (Fig. 2.16B). The wall clock time of the reconstruction step decreased by a factor of approximately R/R_b, where R is the total number of reads available at a sampled time-point and R_b is the number of reads in a bootstrap sample. Simulations showed that decreasing R_b, for the same number of bootstrap samples M, resulted in a proportional decrease in the wall clock time of the reconstruction step without affecting the classification accuracy (Fig. 2.17). The minimum size of the bootstrap samples was empirically observed to be dependent on the sequencing coverage. When the bootstrap sample size was reduced further, some MC runs started to have loci with zero coverage. This caused the reconstruction to terminate prematurely. Thus, the bootstrap sample size needs to be chosen in a way that ensures that no locus in the region being reconstructed has zero coverage.

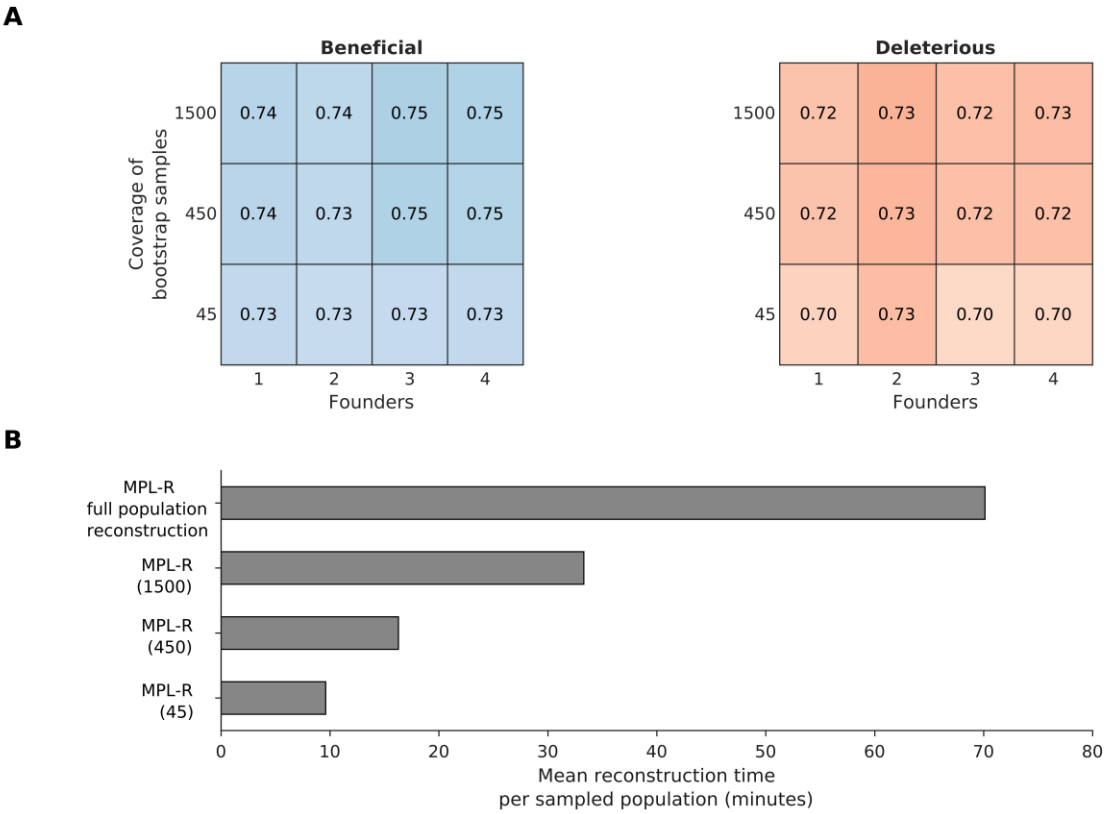


Figure 2.17. Decrease in the size of the bootstrap samples decreases the wall clock time with a gradual decrease in the classification accuracy. **(A)** Comparison of the mean AUROC of MPL-R for different sizes of the bootstrap samples. The bootstrap sample sizes are indicated in terms of the average coverage. The classification accuracy is robust to the bootstrap sample size. **(B)** Average reconstruction time of the population at each time point for MPL-R without bagging, and MPL-R with different sizes of the bootstrap samples indicated in the parenthesis. Results of MPL-R are for $M = 3$ bootstrap samples. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations.

2.3.5 Comparison with other methods

We sought to compare the proposed MPL-R method with other linkage-aware methods in literature; namely IM [61], E&R [60], Evoracle [62], and Li and Barton, 2023 [20]. The method of Li and Barton, 2023 was the only linkage-aware method able to perform inference on the dataset, but it had poor classification performance, likely due to the large temporal sampling step used in the data. The other methods were hampered by impractically large wall clock

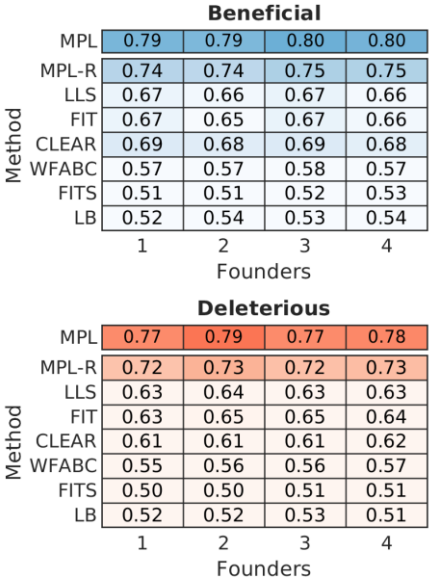


Figure 2.18. MPL-R outperforms available inference methods. The heatmap shows mean AUROC values of detecting beneficial alleles from the rest (deleterious and neutral) and detecting deleterious alleles from the rest (beneficial and neutral). A method with complete knowledge of all double mutant allele frequencies (MPL) is presented as a benchmark. Results of MPL-R are for $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point. Results for the rest of the methods are based on the single mutant allele frequencies computed from the short reads obtained from the ART simulator. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations.

times. For example, the IM method, the fastest of these methods, took ~4.5 hours for a single Monte Carlo run on an Intel Xeon 2.80 GHz CPU. Hence, a comparison of MPL-R's performance with these methods was not feasible. MPL has been shown to outperform IM and E&R on smaller systems [13], and since MPL-R has comparable performance as MPL, we expect MPL-R to outperform both IM and E&R. Other methods in the literature, e.g., LLS [103], FIT [96], WFABC [98], CLEAR [102], and FITS [104] use only the single mutant allele frequency information and do not seek to resolve linkage effects. Simulation results show that MPL-R uniformly outperformed these methods in terms of classification performance (Fig. 2.18 and Fig. 2.19). These methods did not require population reconstruction, and hence their

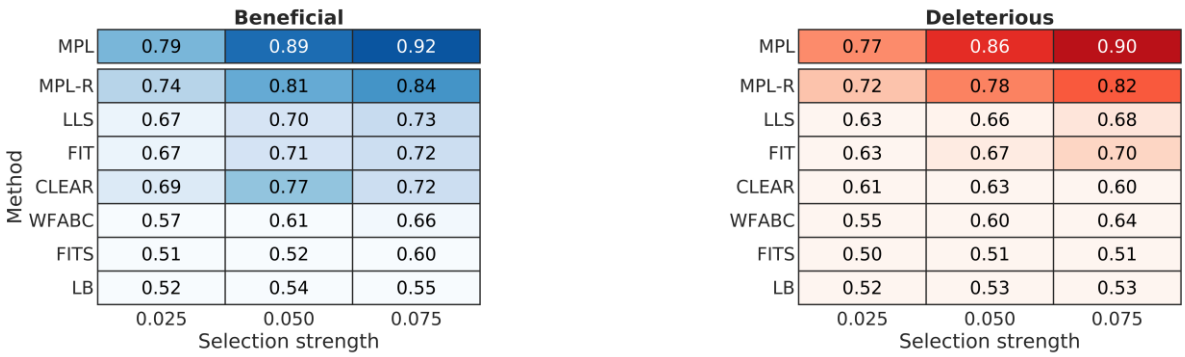


Figure 2.19. MPL-R consistently performs better than available inference methods. Comparison of classification performance (measured in AUROC) of MPL-R and state-of-the-art methods for varying selection strength. A method with complete knowledge of all double mutant allele frequencies (MPL) is presented as a benchmark. Results for MPL-R are with $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point Results on the rest of the methods are on the single mutant allele frequencies computed from the short reads obtained from the ART simulator. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations.

wall clock times were in most cases significantly less than that of MPL-R. Nevertheless, the superior classification performance of MPL-R compensates for the additional computational requirements. Section 2.3.6 presents further details of these alternative methods. It demonstrates empirically that MPL-R, even without compensating for genetic linkage, achieves a performance gain by explicitly accounting for mutations via the mutational flux term in Equation 2.1.

2.3.6 Investigation of difference in performance

The existing time-series methods infer selection based on the rate of change (derivative) of single mutant allele frequencies. Such a model assumes selection as the driving force behind changes in the single mutant allele frequency, and selection inference boils down to the inverse problem of observing changes in the single mutant allele frequencies to quantify the magnitude and sign of selection. This approach will infer the selection coefficients as positive when the

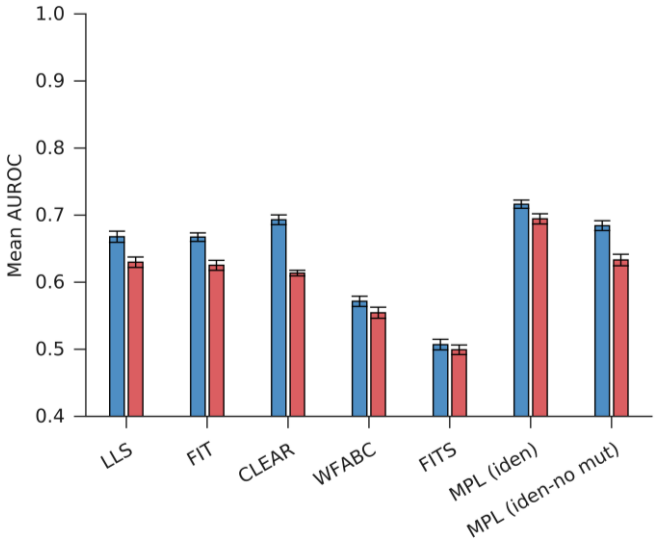


Figure 2.20. The performance of MPL (identity covariance) degrades when mutational flux is ignored: Plot of the mean AUROC for beneficial loci (*blue*) and deleterious loci (*red*) shows that the performance of MPL (identity covariance-no mutational flux) degrades because of ignoring mutational flux. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations. The error bars indicate the standard error of the mean.

single mutant allele frequencies of the mutant alleles undergo an increase as evolution progresses, and negative if the single mutant allele frequencies decrease with time. This methodology does not explicitly model changes in the single mutant allele frequencies due to mutations, and may not be well-suited to analyze data that incorporates mutation. In all our tests, we simulated random mutations in the population, which may explain the degradation in performance of methods based on the rate of change of single mutant allele frequencies. In the following, we shall briefly describe the working principle of each of the competing methods and present an empirical test showing how the incorporation of mutation in the estimation yields better performance.

The method LLS [103] logit-transforms the single mutant allele frequency trajectories and fits a linear model with least squares regression (LLS) to the transformed trajectories. The slope of

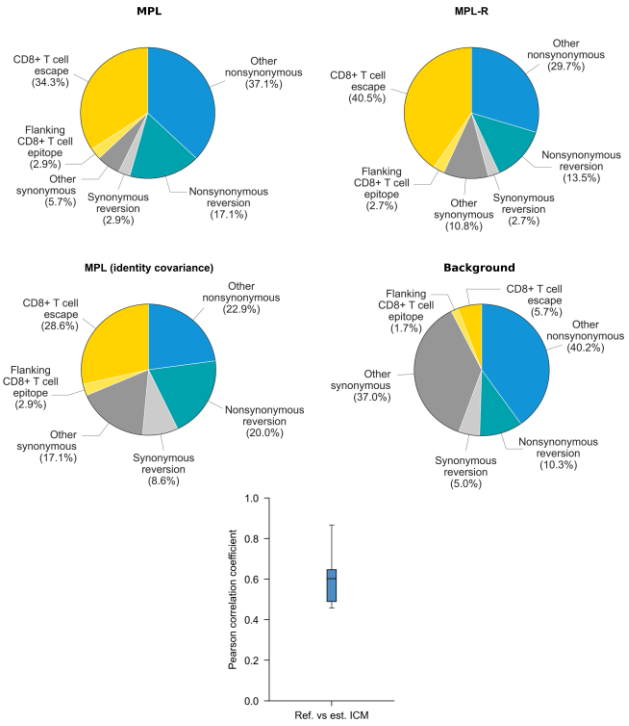


Figure 2.21. MPL-R correctly infers biological insights from intrahost HIV-1 patient data. The proportion of escape mutations and nonsynonymous reversions in the top 5% strongest selection coefficients identified by MPL-R (*top-right* panel) are similar to MPL (*top-left* panel). MPL (identity covariance) is not able to accurately reproduce the insights obtained via MPL (*middle-left* panel). The total composition of the mutations is shown as the background (*middle-right* panel). The ICMs estimated by MPL-R demonstrate moderate correlation with the ground truth ICMs (*bottom* panel).

the fit gives the selection coefficient. CLEAR [102] and WFABC [98] both model the change in the single mutant allele frequencies as a function of the selection coefficients (s) and the dominance parameter (h). For CLEAR, the calculation of the selection coefficients is performed via maximum likelihood estimation given the observed single mutant allele frequencies. WFABC infers the selection coefficients via approximate Bayesian computation (ABC) [126] in which repeated simulations are performed to approximate the likelihood function. FITS [104] is also based on Approximate Bayesian Computation (ABC). The method FIT [96] infers the selection coefficients based on the increments in the single mutant allele frequency trajectories at different time intervals. The assumption is that if the single mutant allele frequency increments over time belong to a zero-mean normal distribution, the

	CD8+ T cell escape mutations			Nonsynonymous reversions outside CD8+ T cell epitopes			Nonsynonymous reversions within CD8+ T cell epitopes		
	%	Fold enrichment	p-value	%	Fold enrichment	p-value	%	Fold enrichment	p-value
MPL-R	40.5	23	$< 10^{-13}$	13.5	21	$< 10^{-5}$	10.8	358	$< 10^{-8}$
MPL	34.3	18	$< 10^{-9}$	17.1	27	$< 10^{-7}$	11.4	382	$< 10^{-8}$
MPL (iden)	28.6	14	$< 10^{-7}$	22.9	33	$< 10^{-9}$	11.4	382	$< 10^{-8}$

Table 2.2. MPL-R reports similar insights as MPL. Percentage of mutations, fold enrichment, and p-values reported by MPL-R, MPL, and MPL (identity covariance) are given for mutations in the listed categories in the top 5% mutations with the strongest selection coefficients.

underlying selection is neutral, as there is no net increase/decrease in the single mutant allele frequency trajectory.

In summary, all these methods have different working principles, implementations, and parameters optimized for the implementations. It is difficult to make an exact comparison about what factors contribute to the difference in performance with MPL (identity covariance), but there is one fundamental difference between the working principle of MPL (identity covariance) and the rest of the methods. MPL (identity covariance) not only models the change in the single mutant allele frequency trajectories as a function of selection, but it also considers the change in single mutant allele frequency trajectories because of mutation via the mutational flux term (Equation 2.1). Intuitively, the mutational flux of a mutation quantifies the tendency of other alleles to acquire that mutation. To empirically demonstrate the effect of mutational flux on the classification performance, we compare the results of MPL (identity covariance) with another implementation in which the mutational flux is not considered, hereby referred to as MPL (identity covariance-no mutational flux). Because only the change in single mutant allele frequencies is used to infer selection coefficients for MPL (identity covariance-no mutational flux), the method makes more mistakes compared to MPL (identity covariance), which takes into account the mutational flux. This resulted in degraded performance of MPL (identity covariance-no mutational flux), which was more comparable to that of CLEAR, which

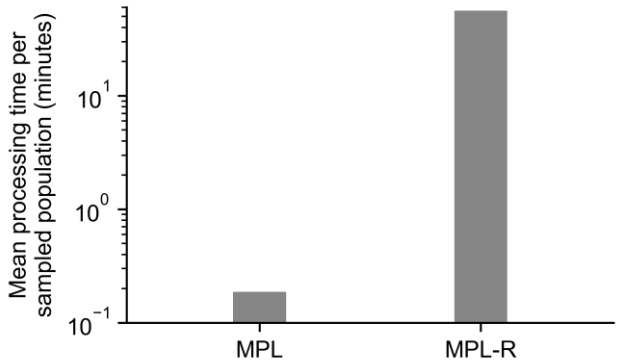


Figure 2.22. MPL-R completes the analysis on the HIV-1 data in a reasonable amount of time. The time required by MPL-R to analyze the HIV-1 data is less than an hour. Though it takes much longer than MPL, it needs to be taken into consideration that full-sequence data required by MPL may not always accessible to the end user, and MPL-R might the only available tool that can perform inference.

also doesn't take into account mutational flux (Fig. 2.20). It is also important to mention here that incorporation of the effects of mutation into the estimator is one reason, but it may not be the only reason for the difference in performance in the methods. It is hard to make a direct comparison between these methods and pinpoint the exact reasons behind the differences in performance, but we have empirically shown that incorporation of mutation does seem to add statistical power to the estimation.

2.3.7 Results on HIV-1 patient data

Next, we analyzed longitudinal data of HIV-1 evolution in 13 patients using MPL-R. We identified the selection coefficients of mutant alleles associated with biological phenomenon such as CD8+ T cell escape, synonymous/nonsynonymous mutations, and reversions using methodology outlined in Sohail *et al.*, 2021 [13]. Fig. 2.21. shows the composition of the top 5% mutations with the strongest selection coefficients estimated by MPL (*top-left* panel), MPL-R (*top-right* panel), and MPL (identity covariance) (*middle-left* panel), and the total composition of the mutations (*middle-right* panel). The percentages, enrichment, and p-values of CD8+ T cell escape mutations, nonsynonymous reversions outside CD8+ T cell epitopes, and nonsynonymous reversions within CD8+ T cell epitopes in the top 5% strongest mutations

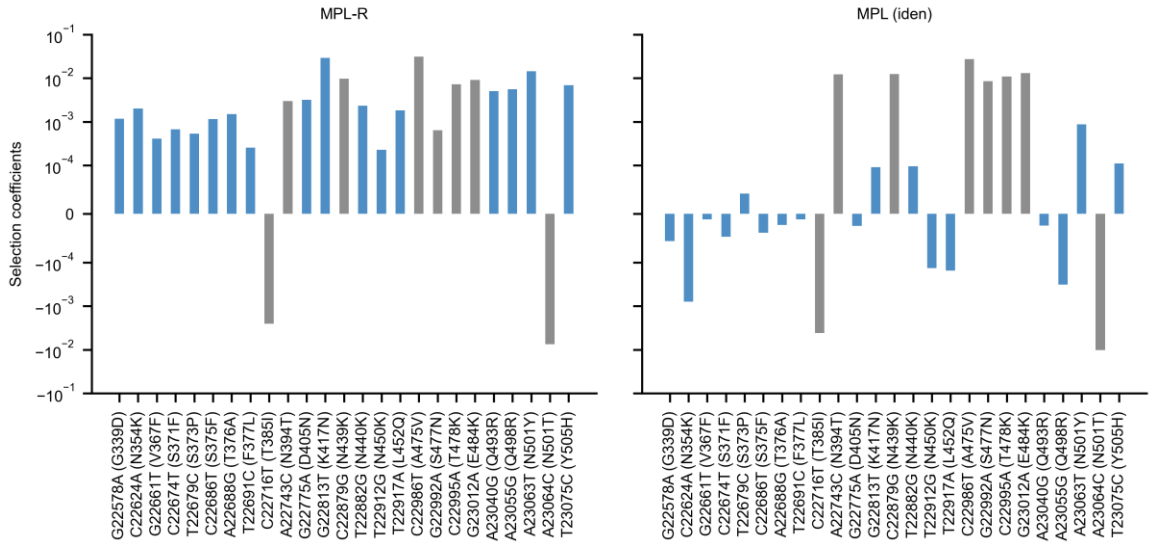


Figure 2.23. MPL-R identifies a higher number of beneficial mutations in the RBD. MPL-R identifies many known beneficial mutations (*left* panel), while MPL (identity covariance) infers several beneficial mutations as neutral or deleterious (*right* panel). Mutations where the estimates of MPL-R and MPL (identity covariance) differ by an order of magnitude or have opposite signs are highlighted in blue. Experimentally-reported phenotypic effects of the mutations are summarized in Table 2.3.

are reported in Table 2.2. MPL-R and MPL both report higher fractions of CD8+ T cell escape mutations and lower fractions of nonsynonymous reversions outside CD8+ T cell epitopes as compared to MPL (identity covariance). Resolution of unobservable linkage information (*bottom* panel) enabled MPL-R to yield insights that were more in-line with those obtained via full-length sequences, which were in turn, had been shown to be consistent with known biological results (Sohail *et al.*, 2021) [13]. In terms of timing performance, MPL-R was able to perform the analysis in a reasonable amount of time (Fig. 2.22).

2.3.8 Results on SARS-CoV-2 patient data

Next, to demonstrate MPL-R's ability to account for linkage while inferring fitness effects of mutations from empirical short read data we analyzed intrapatient data of SARS-CoV-2 evolution. We used LoFreq [127] to count the number of nonsynonymous mutations that exhibited at least 1% variation in the single mutant allele frequency. LoFreq reported 510 such nonsynonymous mutations across the entire genome, of which 26 were in the RBD. The number of mutations in the RBD was 2.6 times (two-sided Fisher's exact test p-value

Mutation	Experimentally reported phenotypic effects
S371F	Antibody escape [128], [129]
F377L	Antibody escape [130]
D405N	Antibody escape [129]
K417N	Antibody escape [4], [131]
N440K	Antibody escape [132]
N450K	Antibody escape [133]
Q493R	Antibody escape [134]
Q498R	Antibody escape [135], host-binding enhancement [136]
Y505H	Antibody escape [137]
G339D	T-cell response weakening [138], antibody escape [128]
N354K	Immunogenicity reduction and antibody escape [139]
L452Q	Stability enhancement [140], antibody escape [129], [141], infectivity enhancement [142]
V367F	Infectivity enhancement [143]
S375F	Infectivity enhancement [144]
S373P	Host-binding enhancement [145]
N501Y	Host-binding enhancement [146]
T376A	Spike cleavage reduction [147]

Table 2.3. Mutations inferred as beneficial by MPL-R map to known beneficial mutations. Nonsynonymous mutations in the RBD which are inferred as beneficial by MPL-R and neutral or deleterious by MPL (identity covariance) map to mutations reported to be beneficial to the virus.

2.71×10^{-5}) higher than expected by random chance. An elevated number of mutations in the RBD suggested immune pressure on the virus. This was further supported by analysis of the $\mathrm{dN/dS}$ (ratio of non-synonymous and synonymous mutations), which evaluated to a mean value of 7.789 ± 2.706 for the RBD. The selection coefficients corresponding to nonsynonymous mutations in the RBD, when inferred by MPL-R (which accounts for linkage effects) had higher magnitudes as compared to those inferred by MPL (identity covariance); a method that does not account for linkage effects (Fig. 2.23). Specifically, MPL-R inferred stronger positive selection coefficient estimates, classifying most mutations in the RBD as beneficial, while MPL (identity covariance) inferred relatively weaker estimates of selection coefficients, classifying most of these mutations as weakly beneficial or deleterious. We searched existing literature for all mutations in Fig. 2.23 for which MPL-R and MPL (identity covariance) reported selection coefficients with opposite signs or differing by more than one

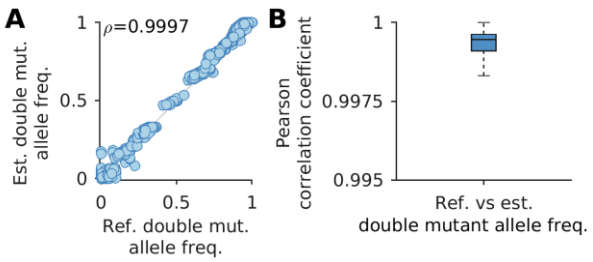


Figure 2.24. Reconstruction of double mutant allele frequencies is accurate. **(A)** Double mutant allele frequencies estimated after reconstruction show high correlation with the actual double mutant allele frequencies (Pearson correlation coefficient $\rho = 0.9997$ with p-value $< 10^{-100}$). **(B)** The summary statistics of the Pearson correlation coefficient between the reference and estimated double mutant allele frequencies indicate a consistent trend across all Monte Carlo runs. Results are for MPL-R with $M = 1$ bootstrap sample of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations.

order of magnitude. Table 2.3 shows that all RBD mutations classified as beneficial by MPL-R and underestimated by MPL (identity covariance) have been reported to be beneficial to the virus. The majority of these mutations benefit the virus by their contribution in antibody escape, whereas some mutations enhance infectivity or binding to the host (see Table 2.3 for details). The observation that MPL-R can identify more biologically-validated beneficial mutations as compared to MPL (identity covariance) demonstrates the advantage of the incorporation of linkage information.

2.4 Discussion

Short-read sequencing is a popular sequencing technology that is used extensively [33], [148]. In this study, we have developed MPL-R, a computational pipeline for the inference of selection coefficients from short-read time-series data and demonstrated its accuracy over a range of parameters. Our analysis revealed novel insight into the importance of the off-diagonal ICM entries. We have shown that if the ICM entries corresponding to the unobservable inter-read allele-pairs are set to zero, the inference performance deteriorates substantially compared

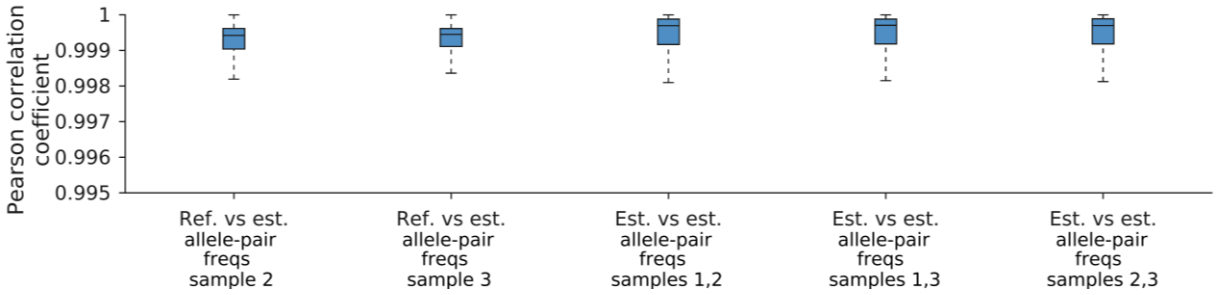


Figure 2.25. The reconstructed double mutant allele frequencies are accurate. The double mutant allele frequencies calculated from the reconstructed and actual populations are highly correlated. The double mutant allele frequencies calculated from different bootstrap samples also exhibit strong correlation. Results are for MPL-R with $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations.

to models that assume independent evolution. We also presented an empirical result which demonstrates that incorporation of the mutational flux term assists in more accurate inference. This is an added advantage of MPL-based methods over methods that do not explicitly model mutation.

Any inference method based on population reconstruction is bound to face two key issues, which MPL-R addresses. First, the reconstruction time scales linearly with the number of reads, which implies that population reconstruction becomes increasingly difficult as sequencing platforms improve and read throughput increases. Here the engineering novelty of MPL-R comes into play. To our knowledge, MPL-R demonstrated for the first time the utility of a divide-and-conquer bagging approach of population reconstruction for inference of selection coefficients. MPL-R is designed to efficiently perform inference by making use of bagging, which ensures that the volume of reads does not significantly affect reconstruction time. MPL-R also distributes memory consumption over multiple reconstruction operations which can be useful when dealing with large datasets. The second issue addressed is that all reconstruction

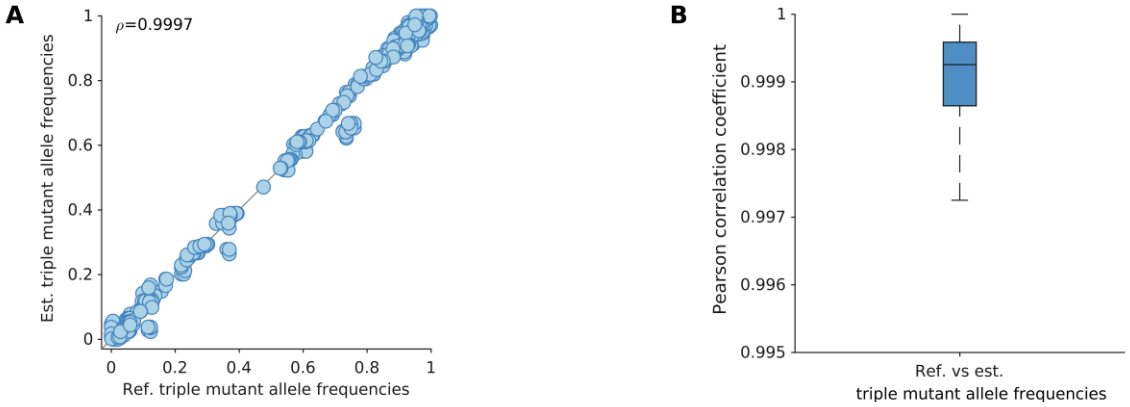


Figure 2.26. MPL-R faithfully reconstructs the triple mutant allele frequencies. **(A)** Triple mutant allele frequencies calculated from the reconstructed population are highly correlated with the actual triple mutant allele frequencies (Pearson correlation coefficient $\rho = 0.9997$ with p-value $< 10^{-100}$). **(B)** The summary statistics of the Pearson correlation coefficient between the estimated and actual triple mutant allele frequencies indicate a consistent trend across all Monte Carlo runs. Results are for $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations.

algorithms struggle with accurate population reconstruction, particularly when the underlying population contains multiple haplotypes or haplotypes with low frequencies [111]. Haplotype reconstruction is also challenging for regions with low coverage, where distinguishing actual nucleotides from noise becomes difficult. All these factors lead to the inclusion of erroneous nucleotides which can lead to erroneous estimates of the calculated statistics. These errors can be mitigated by the minimum single mutant allele frequency threshold, which filters out small perturbations in the single mutant allele frequencies. Also, MPL-R works at the level of single and double mutant allele frequencies so is expected to stay immune to errors in higher-order statistics. Our simulations showed that the double mutant allele frequencies reconstructed by the pipeline were accurate (Fig. 2.24 and Fig. 2.25). The ICM entries computed from the reconstructed double mutant allele frequencies were accurate as well, as shown in Fig. 2.12A. This is sufficient for the purpose of inferring selection coefficients as the MPL framework only

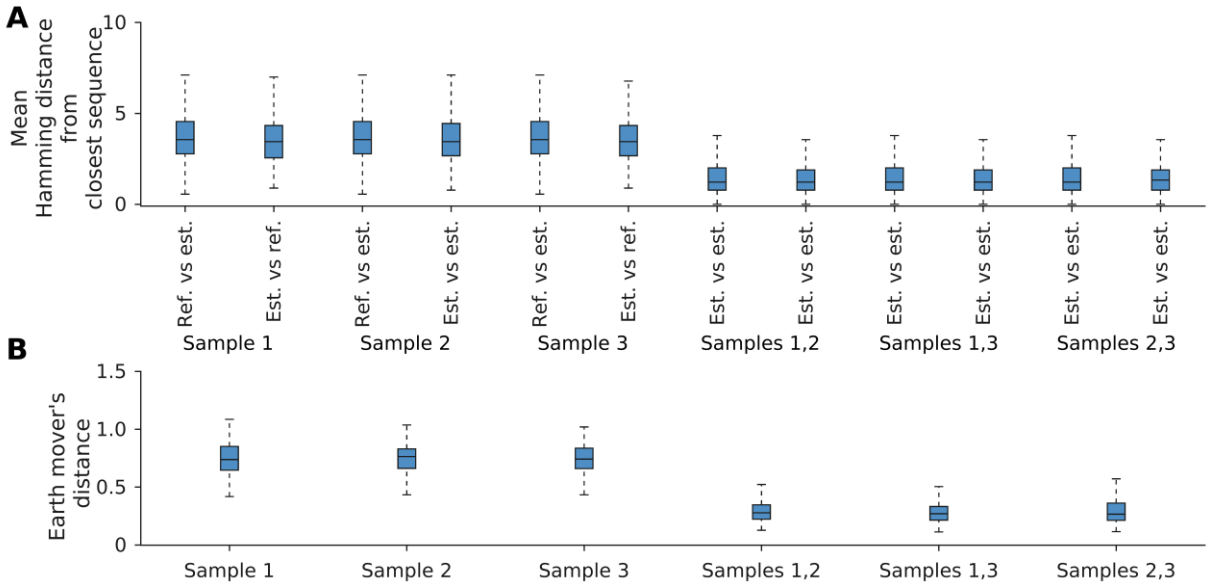


Figure 2.27. The reconstruction performed via Quasirecomb is accurate. **(A)** The mean Hamming distance between the actual Wright Fisher population and the closest sequences in the reconstructed population and vice versa is small. Sequences in the reconstructed populations differ from each other by a small Hamming distance as well. **(B)** The earth mover's distance between the reference Wright Fisher population and reconstructed populations and between the reconstructed populations is small. Results are for $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations.

requires knowledge of single and double mutant allele frequencies, while errors in frequencies of higher-order tuples do not affect the accuracy of the estimated selection coefficients [13]. The higher-order frequencies may be important for more complex models, such as those incorporating epistasis [106]. Further work can explore the extent to which higher order frequencies are preserved by population reconstruction using Quasirecomb or other approaches. An initial analysis of triple mutant allele frequencies suggests that those obtained after reconstruction were strongly correlated with the triple mutant allele frequencies of the original population (Fig. 2.26), though more extensive future investigation is warranted. In our tests, although the reconstruction was not perfect, Quasirecomb did a fairly good job of

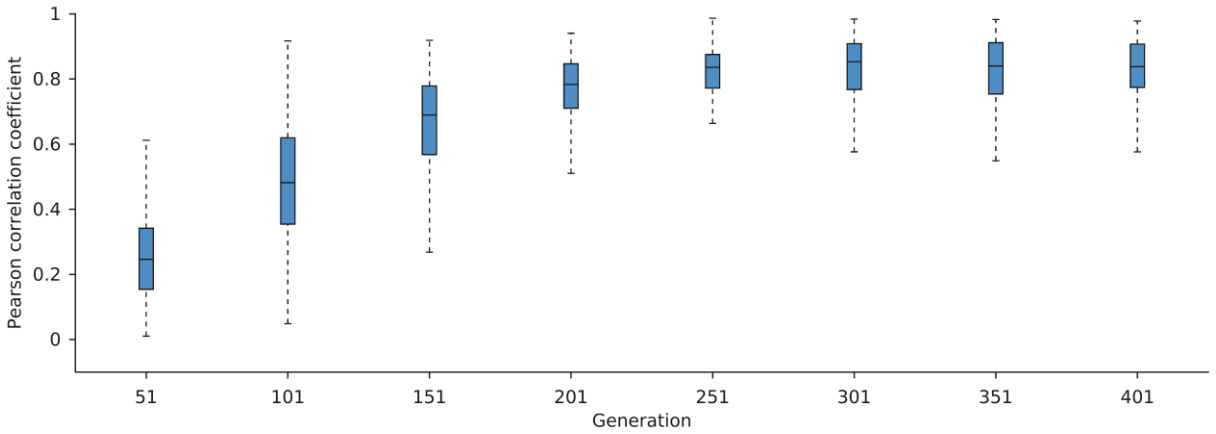


Figure 2.28. The covariance matrices of allele frequencies calculated using reconstructed double mutant allele frequencies at each time point are accurate. The covariance matrices of allele frequencies calculated from the reconstructed and actual populations exhibit good correlation across the sampled time-points. The correlation at the initial time points is poor because the single mutant allele frequencies at these time points are small (Fig. 2.7A), and hence the covariance matrices have negligible magnitudes. Results are for $M = 3$ bootstrap samples each of size $R_b = R/3$ reads (coverage 1500X), where R is the total number of reads available at each sampled time-point. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of $L = 500$ bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $20/20/460$ loci were beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$ respectively. The fitness landscape had a repeating comb-like structure shown in Fig. 2.2B. One founder sequence was used to generate each population, which was allowed to evolve for 400 generations and sampled at $\Delta t = 50$ generations.

reconstructing the full-length sequences. A distribution of the mean Hamming distance between the reconstructed sequences and the closest sequence in the reference population had a median of around 3.5, which showed an average of around one erroneous bp per read-length (Fig. 2.27A). The distribution of the mean Hamming distance between the reference sequences and the closest sequence in the reconstructed populations also had a median of around 3.5, indicating that all the reference sequences were reconstructed with few errors. The low value of the median Earth Mover's distance [149] between the reference and reconstructed populations and between the reconstructed populations also suggested good reconstruction (Fig. 2.27B).

The covariance matrices calculated from the reconstructed double mutant allele frequencies showed strong correlation with the covariance matrices calculated from the double mutant allele frequencies obtained from the full-length sequences (Fig. 2.28). The initial time-points demonstrated weak correlation because the single mutant allele frequencies at these time-points were small (Fig. 2.7A), which resulted in covariance matrix entries with negligible magnitudes. At later time-points, the single mutant allele frequencies had larger magnitudes, which led to relatively larger covariance matrix entries, and stronger correlation. All simulations were carried out using single-end reads, however, MPL-R can handle paired-end reads as well. The advantage of paired-end reads is that read pairs can be combined to generate read fragments that are longer than the read length. These fragments provide double mutant allele frequency information between allele pairs at much greater distances compared to single-end reads. We have utilized this property in the SARS-CoV-2 analysis in Section 2.3.8.

MPL-R provides users with an easy-to-use bioinformatics pipeline to infer selection coefficients from short-read time-series data. The pipeline not only combines the features of its constituent tools but introduces the functionality of handling data transfer, intermediate processing, and function calls internally to minimize the workload of the end-user. The pipeline is flexible in terms of the components used and is scalable to the data sizes tested. MPL-R has been tested on data from a single sequencing platform, yet the reconstruction scheme based on Quasirecomb is general and can reconstruct data from several different sequencing platforms. The proposed MPL-R method offers a novel method to perform fitness inference of genetic time-series data and can be applied to study the complex evolutionary dynamics of populations of evolving microbes. Future work can explore fitness inference from long-read sequencing data and address the incorporation, testing, and optimization of software for long-read processing into the pipeline.

Chapter 3

Viral fitness inference and structural variant detection from TGS data

3.1 Introduction

3.1.1 Overview

Evolution of organisms is driven by genetic variation across multiple scales, from individual nucleotides (SNVs) to large groups of nucleotides (SVs). Interactions between SNVs, commonly quantified by LD, can shape evolutionary dynamics and must be accounted for when inferring fitness from sequencing data. TGS technologies enable detailed characterization of genomic variation but limited read lengths often hinder the resolution of LD information across entire genomes. This work presents a computational pipeline that overcomes this limitation by reconstructing genome-wide haplotypes, inferring LD values across the entire genome, and estimating fitness. The pipeline also detects SVs using an established software tool, providing a comprehensive characterization of genomic variation. Results from both simulated and real datasets demonstrate that the pipeline is fast, robust, and capable of yielding biologically consistent insights. These findings support its utility across diverse experimental settings.

3.1.2 Background

Single nucleotide variants (SNVs) and structural variants (SVs) are both manifestations of genomic evolution. SNVs involve changes at individual loci whereas SVs are defined as genomic changes of at least 50bp in length [150]. SNVs can alter phenotypes such as antimicrobial resistance in bacteria [151], and immune escape in viruses [152], [153]. SNVs can also alter essential biological functions, affect the structure of the organism, and modify other factors essential for survival, thereby directly affecting the replicative potential of an organism. SVs such as insertions, deletions, translocations, inversions, and duplications,

similarly influence pathogenicity and antimicrobial resistance in pathogens [154], [155]. SVs are known to cause diseases, alter heritability, and affect environmental adaptation [156], [157], [158].

In addition to SNVs and SVs, genomic evolution is governed by the non-random association between SNVs, termed LD. LD is observed abundantly in nature and has been found to affect evolution in a number of human pathogens. LD in the influenza virus enables it to accumulate beneficial SNVs, giving rise to novel viral strains resistant to existing vaccines [14]. LD in HIV-1 allows the virus to evade recognition by CD8+ T cells, leading to chronic infection [13]. Quantification of LD is crucial for the complete characterization of SNVs acquired by pathogens. LD between neutral and beneficial SNVs can cause the neutral SNVs to increase in prevalence while hitchhiking on the beneficial SNVs. Conversely, beneficial SNVs can decrease each other's prevalence because of LD. The true fitness, or replicative potential, of SNVs in LD cannot be ascertained without quantifying the effect of LD on SNV prevalence [13]. LD, in addition to characterizing fitness, is required in the estimation of population divergence, calculation of population size, carrying out genome-wide association studies, and performing genomic prediction [22], [23], [24], [25].

Genomic evolution is studied via genomic sequencing which provides detailed information about the composition of genomes and their temporal evolution. Genomic sequencing was first introduced as Sanger sequencing [26], which was followed by NGS [33], and TGS. TGS technologies comprising PacBio sequencing and ONT, enable detailed observations of genomic evolution [40], [41], [159]. PacBio sequencing is known for its high accuracy which enables precise SV detection, targeted sequencing, accurate sequence assembly, and full-length transcriptome recovery [160], [161], [162], [163], [164]. ONT sequencing is popular for its large read lengths that can exceed 1M base pairs (bp) [38], low turnaround time, and low coverage requirement, which makes it ideal for use in onsite sequencing with limited time and

resource availability [49], [165], [166], [167]. ONT sequencing has also been used to study a number of viral pathogens of clinical interest such as HIV-1 [168], HCV [169], HBV [170], and SARS-CoV-2 [171].

We present SLAFIT, a pipeline for the detection of SVs, LD estimation, and fitness inference from TGS data. The pipeline takes TGS reads sampled at single or multiple evolutionary time-points as input and outputs SNVs, SVs, and LD matrices at each time-point, together with fitness estimates when time-series data is available. In addition to unifying these analyses within a single framework, SLAFIT extends existing methodologies by enabling full-genome LD estimation from TGS reads and improving the efficiency of fitness inference. Because individual TGS reads provide partial information across a limited part of the genome, SLAFIT leverages a haplotype reconstruction based-approach. This method infers full-length haplotypes and their prevalence values that best explain the TGS reads, thereby enabling the calculation of LD across the entire genome. Furthermore, to accommodate high-throughput sequencing volumes, we developed a multiprocessing- and GPU-based implementation of a fitness inference framework, MPL [13], that reduces computational time compared to the original framework, particularly as genome size and number of sequences increase. We demonstrate the performance of SLAFIT for LD estimation on simulated data and show that fitness inference is robust across a range of read-sampling conditions. We also evaluate the false discovery rate of SV detection on simulated data and demonstrate that the method performs well.

To demonstrate the real-world utility of SLAFIT, we applied the pipeline to data from two viruses. We used SLAFIT for fitness inference on an EV-C104 dataset. EV-C104 is a single-stranded ribonucleic acid (ssRNA) virus that rapidly accumulates SNVs via error-prone replication and recombination. This virus lacks an effective vaccine and can cause chronic infection in immunocompromised patients. We utilized longitudinal infection data from this

virus and showed that SLAFIT infers SNV fitness values that strongly align with computational predictions of immune escape. We also performed SV detection on data from HSV-1. This virus is one of the most common human pathogens and does not have a vaccine. Identification of variation in the viral genome may lead to better understanding of its structure and vaccine development. We demonstrate that the proposed SV detection scheme is robust against reduction in the read-length, capturing structural variation profiles that match known baselines.

3.1.3 Related work

While TGS technologies can provide real-time, high-resolution information about SNVs and SVs, these struggle with observing LD across the complete genome [172], [173]. This limitation arises because the reads are typically smaller than the genome, which limits the observable LD information to only those SNV pairs that are covered by the same read. The LD information remains unobservable for SNV pairs separated by distances exceeding the read length. Existing methods for LD inference from NGS data [58], [62] can theoretically be adapted to work on TGS data, but these have slow computational speeds, model constraints, and scalability limitations, typically handling systems with no more than a few tens of loci. To our knowledge, no method exists that takes TGS reads as input, detects both SNVs and SVs, and calculates the unobservable LD between SNVs. Consequently, LD information from the growing volume of TGS data remains only partially observed. Thus, there is a pressing need for a method that can estimate LD across the entire genome from TGS data in addition to detecting SNVs and SVs.

3.2 Methodology

3.2.1 Overview

SLAFIT addresses two main problems, first being the reconstruction of unobservable LD values between SNVs that exist at distances greater than the read-length (referred to as inter-read LD values). To calculate the inter-read LD values, SLAFIT first uses RVHaplo [174] to

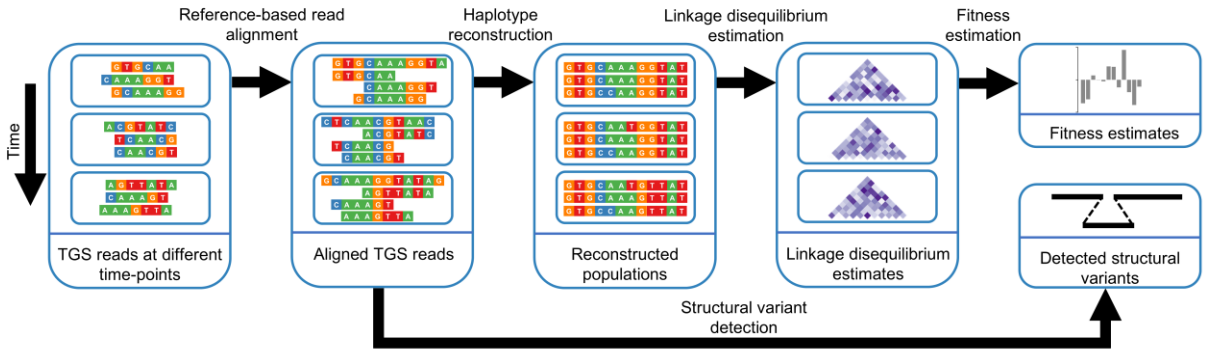


Figure 3.1 SLAFIT pipeline schematic. 1. TGS reads are obtained at each time-point from an evolutionary study. 2. TGS reads are aligned to a reference sequence to obtain Sequence Alignment Map (SAM) files at each time-point. 3. Haplotype reconstruction of the SAM files yields a set of full-length sequences and their corresponding prevalence values at each time-point. 4. LD matrices are obtained from the reconstructed sequences. 5. Fitness estimates are obtained via the MPL equation and SV detection is performed via cuteSV.

perform haplotype reconstruction. RVHaplo was selected as it has been shown to be sequencing platform-independent and has demonstrated good performance on real datasets of HIV-1 and HCV. RVHaplo takes in TGS reads and returns reconstructed haplotypes as well as the corresponding haplotype prevalence values, which are used to calculate the LD matrix. When time-series data is available, the LD matrices are aggregated over all time-points via the proposed implementation of the MPL framework to obtain fitness values. The second problem addressed is the detection of SVs from TGS reads. cuteSV [175] is used for SV detection as it has been shown to consistently perform well over a range of test cases [173]. The SLAFIT pipeline is presented in Figure 3.1.

3.2.2 Read alignment and SV detection

The input to the pipeline are TGS reads sampled either at a single time-point, or at $K+1$ time-points $\{t_0, t_1, \ldots, t_{K-1}, t_K\}$ in a time-series experiment. These reads are first aligned to a reference sequence, which is either user-supplied or obtained by taking the consensus of the reads sampled at the first time-point. Reference-based alignment of the input reads is performed via minimap2 [176]. Subsequently, SV detection is performed on the aligned reads via cuteSV to obtain SVs captured by the reads.

3.2.3 LD estimation and fitness inference

Haplotype reconstruction via RVHaplo returns full-length reconstructed sequences and the prevalence values of the individual sequences. If the data contains a single dominant haplotype and other haplotypes occur at prevalence values that cannot be detected by RVHaplo, the software prompts the user that reconstruction cannot proceed. In this case the consensus sequence is used as a full-length reconstructed sequence as this is a good approximation for the actual population. The reconstructed sequences then are aligned against the reference sequence via Multiple Alignment using Fast Fourier Transform (MAFFT) [177]. The aligned, reconstructed sequences are supplied to the inference framework in which the entry of the LD matrix at loci i and j, D_{ij} is computed as

$$D_{ij}(t_k) = f_{ij}(t_k) - f_i(t_k)f_j(t_k),$$

where $f_i(t_k)$ and $f_j(t_k)$ represent the single mutant allele frequencies at loci i and j respectively, $f_{ij}(t_k)$ represents the double mutant allele frequency at loci (i,j), and t_k represents time-point k. Fitness inference is subsequently performed via our proposed parallelized implementation of MPL. The estimated fitness values, single mutant allele frequencies, and the SNV counts are output by the pipeline.

3.2.4 Parallel processing

Scalability can become a challenge during the analysis of large volumes of TGS data, for example, technologies such as PacBio High-Fidelity (HiFi) sequencing can generate over one million reads of lengths up to 10kbp [178]. Although these reads typically sequence the same genomic region covered by individual reads and are not affected by the limitation of unobservable inter-read LD information, the substantial redundancy in sampling introduces the challenge of scalability. Addressing this challenge is therefore essential for efficient estimation of observable LD from TGS data, independent of the unobservable inter-read LD estimation problem.

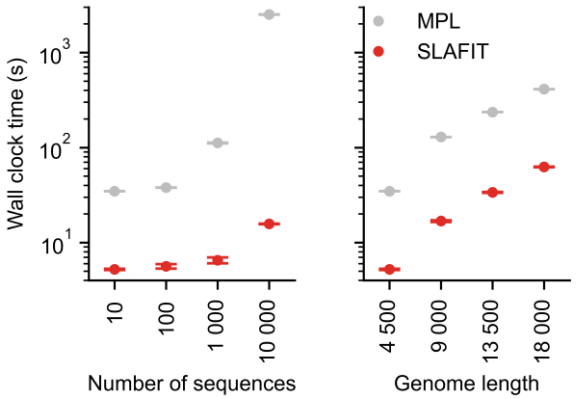


Figure 3.2 Parallelized implementation of MPL supplied with SLAFIT requires lower wall clock time compared to the sequential implementation. The wall clock time required by both implementations of MPL increases with the length (*left* panel) and number of sequences (*right* panel), with the implementation used in SLAFIT consistently taking less time. Results are shown for 100 Monte Carlo runs. For the tests involving sequence length, each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals. For the tests involving number of sequences, the sequence length was fixed at $L = 4500$ bp. The genomes consisted of bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. For genome lengths $\{4500, 9000, 13\,500, 18\,000\}$ alleles at $180/180/4140$, $360,360,8280$, $540,540,12420$, $720,720,16560$ loci were respectively beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$. The beneficial and deleterious loci were uniformly distributed across the entire genome. One founder sequence was used to generate each population, which was allowed to evolve for 550 generations and sampled at $\Delta t = 50$ generations. TGS reads of mean length 1500 bp were simulated with 10% error rate and coverage 50X.

To improve scalability, we made two modifications to the MPL framework. First, the single- and double-mutant allele frequencies at each time-point are calculated on independent processor cores. Second, matrix inversion and multiplication in the MPL equation (Equation 1 in Sohail *et al.* 2021 [178]) are carried out on a GPU, which is better suited to handle large-scale matrix operations. Together, these enhancements offer improved scalability compared to the original implementation, particularly as the number of sequences increases (Fig. 3.2). We also tested SLAFIT on a real patient dataset of SARS-CoV-2 in which reads were obtained at multiple time-points during infection [179]. The SLAFIT pipeline did not run into any computational bottlenecks during the analysis. The results of the analysis have not been presented here as the dataset had very few SNVs from which we were unable to detect LD

information. Based on Fig. 3.2, we expect SLAFIT to handle systems of larger sizes as well, however, care should be taken to ensure all matrices are within the limits of the system memory.

3.2.5 Overview of evaluation

We performed analytical validation of SLAFIT on simulated data and also performed biological concordance analysis on real data. The simulated data tests were designed to answer three main questions: (i) how robust SLAFIT is to degraded sequencing information, (ii) how well the method performs compared to existing approaches, and (iii) how well it scales to larger systems. The real data tests were designed to investigate whether SLAFIT was able to infer biologically-meaningful insights from the data. To assess robustness, we first used simulated data with known ground truth population composition and corresponding strain-level fitness values. In practice, these parameters are not observed directly and must be inferred from sequencing reads, which contain only partial and noisy information about the underlying genomes. We refer to this as information loss during read generation. This information loss arises because reads span only part of a genome, sequencing coverage may vary, and sequencing errors can corrupt the observed information. To systematically study the effect of information loss, we generated TGS reads under varying read lengths, coverage levels, and error rates, and then applied SLAFIT to infer strain-level fitness from the resulting reads. This allowed us to evaluate performance as the information available in the reads became increasingly degraded. Because the accuracy of fitness inference hinges on the recovery of LD, we also evaluated the accuracy of LD inference under these settings. We then compared SLAFIT against an NGS based scheme to quantify the gain in performance when using TGS reads. Furthermore, we performed comparative benchmarking to compare SLAFIT with a state-of-the-art linkage-aware fitness inference method. To assess scalability, we further examined performance as the genome length and number of sequences were increased. Finally, we applied SLAFIT to a real dataset to demonstrate its practical utility in generating

biologically meaningful insights. We evaluated the SV detection component separately. First, using simulated data with no SVs, we estimated the false positive rate. We then tested SV detection on real data while varying TGS read length to simulate degraded input quality. These experiments showed that the method could detect a known SV even when the length of the SV was substantially longer than the maximum read length.

3.2.6 Simulated data

We simulated data with parameters guided by evolving viral populations. Data generation was performed via the Wright-Fisher model, which assumes non-overlapping populations of a finite number of individuals [180], [181]. Unless otherwise specified, a population size of $N = 1000$ individuals of length $L = 4500$ bp was chosen. At the first time-point of evolution, N copies of a random binary sequence of length L, consisting of $0/1$ were generated. The base 'A' was assigned to 0, and the bases {C,G,T} were randomly assigned to 1. This population was cloned and mutated based on the Wright-Fisher evolutionary dynamics. Every cloning step corresponded to one generation. At each generation, selective pressure was simulated on predefined loci to choose for fitter mutations. This evolution was simulated for 550 generations, from which $n = 10$ individuals were sampled every $\Delta t = 50$ generations, starting at the 51 st generation. The subsampling of sequences and time emulates real sequencing, in which a sample cannot represent the entire population, and the samples cannot be taken at every generation. For every sequence represented by a binary string g_i^a, where mutant/WT alleles were represented by $0/1$, the sequence fitness f_a was assumed to follow an additive model, given by

$$f_a = 1 + \sum_{i=1}^{L} g_i^a s_i,$$

where s_i represents the selection coefficient of the mutant allele at locus i. The additive fitness model is a simplification that assumes no epistasis between loci. The MPL equation is based

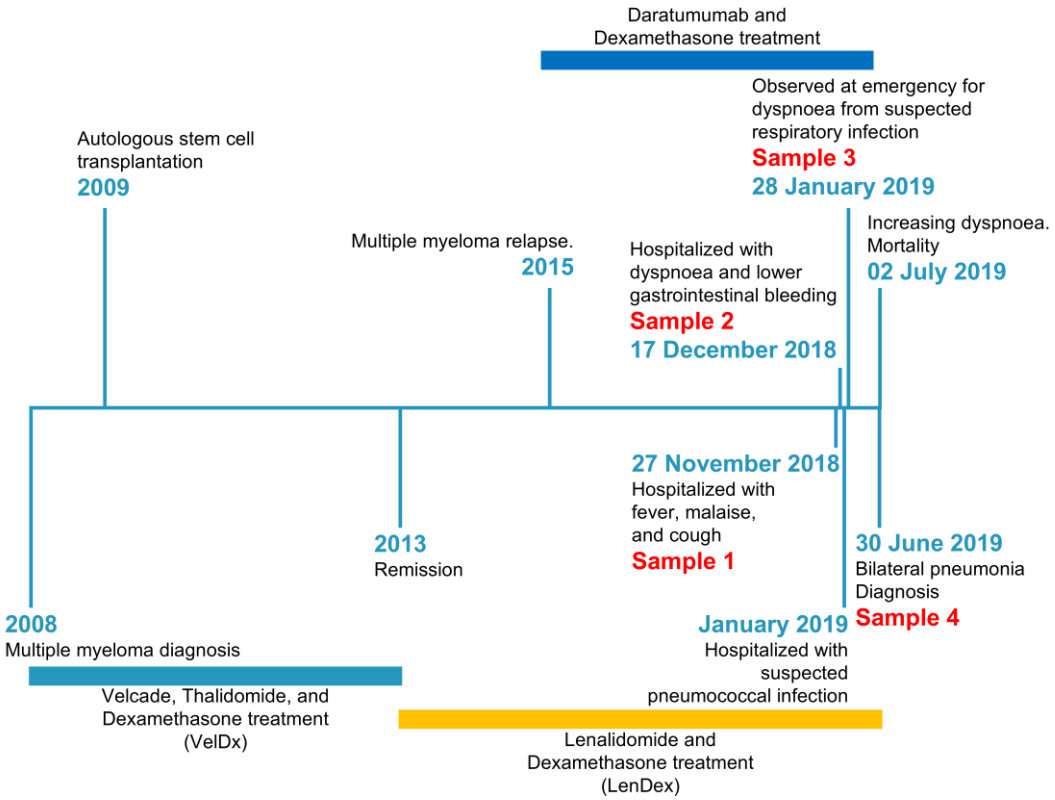


Figure 3.3 Clinical timeline of EV-C104 infection data.

on the additive fitness model, and we use the model for consistency. The mutation probabilities $\mu_{\mathrm{WT,mut}}$ and $\mu_{\mathrm{mut,WT}}$ were assumed to be equal and assigned a value $\mu = 10^{-4}$. The number of beneficial, deleterious, and neutral mutations were $\{180,180,4140\}$. The beneficial and deleterious mutations were assigned the same magnitude of selection coefficients, i.e. 0.025 with opposite signs, while the neutral mutations were assigned zero-valued selection coefficients. The beneficial and deleterious mutations were uniformly distributed across the length of the genome. We generated three datasets with genome-length $\{9000,13500,18000\}$, and three datasets with number of sequences $\{100,1000,10000\}$. For datasets with genome length $\{9000,13500,18000\}$, the number of beneficial, deleterious, and neutral mutations were $\{360,360,8280\}$, $\{540,540,12420\}$, and $\{720,720,16560\}$, respectively. All datasets were generated without SVs. 100 Monte Carlo runs were generated for each dataset. TGS was simulated for the error profile of the Oxford Nanopore R10.4.1 system via badread [182]. For each time-point, reads of mean length 1500 bp (0 standard deviation) were generated with a

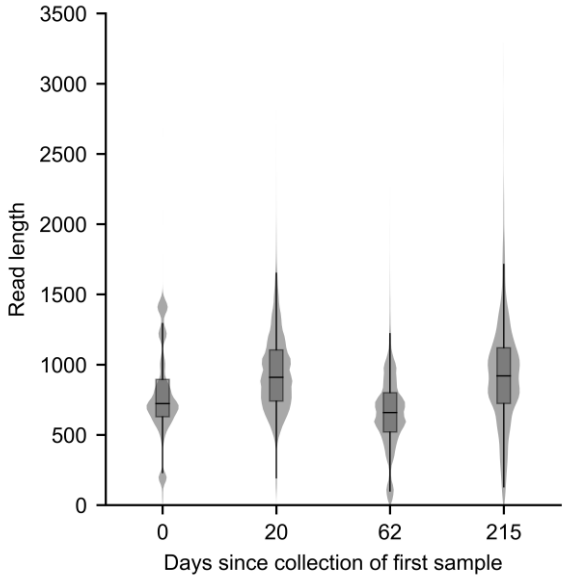


Figure 3.4 Read-length used in the EV-C104 infection data is much shorter than the genome length. TGS reads used to sample the EV-C104 infection data left much of the long-range LD information unobservable due to much shorter lengths compared to the genome-length.

coverage of 50X, and mean/maximum identity $90/100$, with a standard deviation of 0. The parameters "junk reads", "random reads", "chimeras", and "glitches" were set to 0, and no adapter sequences were used on either end. These parameters correspond to a fairly clean sequencing run. For testing robustness to read parameters, we generated additional datasets with different parameters corresponding to sequencing runs with relatively degraded quality. Two datasets were generated with read length $\{500,1000\}$, two datasets with coverage $\{25\text{X}, 35\text{X}\}$, and two datasets were generated with mean identity $\{85,80\}$. For inference, we set the value of the regularization parameter γ in the MPL framework to 20. Estimation was performed on only those loci on which the single mutant allele frequency exhibited at least 10% variation, which is the least count when sampling $n = 10$ individuals. The single mutant allele frequencies were computed from the simulated population data.

3.2.7 EV-C104 data

We performed SNV-level fitness inference on a real dataset of prolonged EV-C104 infection [57] to demonstrate the ability of SLAFIT to infer biologically meaningful insights. The

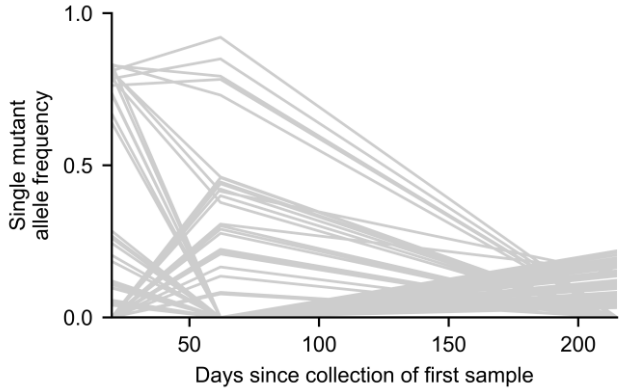


Figure 3.5 EV-C104 infection data indicates the presence of LD. Multiple co-occurring SNVs are observed in the EV-C104 infection data which suggests the existence of LD.

clinical timeline of the infection is presented in Figure 3.3. Briefly, the dataset was obtained over a course of 216 days from a patient with multiple myeloma relapse. The data consists of $57{,}899 \pm 38{,}364$ ONT MinION reads of length 889 ± 107 bp with a coverage of $7027\text{X} \pm 5137\text{X}$. The read-length used is much shorter than the genome-length (7,390 bp) which means that much of the long-range LD information is unobservable (Figure 3.4). Following Cassidy et al. 2022 [57], the consensus sequence of the first time-point (GenBank accession no. MZ092702) was used as reference and reads for all time-points were aligned against it as explained in the Methods section. Single mutant allele frequency trajectories, obtained via LoFreq [127] indicated multiple co-occurring SNVs, which suggested the existence of LD in the data (Figure 3.5). Haplotype reconstruction was performed to obtain full-length sequences of the coding region ($703-7323$ bp) of the genome and selection coefficients obtained via a bi-allelic (WT/mutant) model. The choice of a bi-allelic model was made because the entropy values of the single mutant allele frequencies indicated that this model accurately captured most of the variation in the data (Figure 3.6). Selection coefficient estimation was performed via MPL with the regularization parameter $\gamma = 20$. SNV probability of 2.97×10^{-3} per year estimated by Cassidy et al. 2022 [57] was used for WT-mutant transitions and three times this probability was used for the mutant-WT transitions to account for transitions from three possible mutants to the WT. Inference was performed on

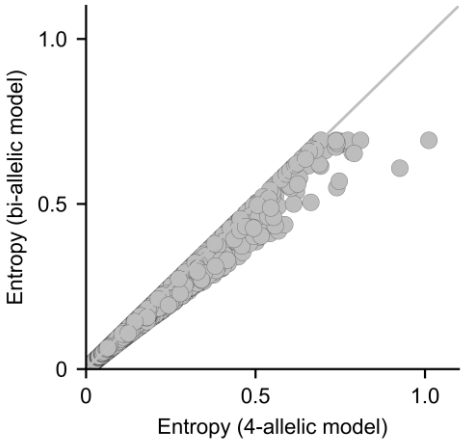


Figure 3.6 Bi-allelic model for the EV-C104 infection data shows good association with the 4-allelic model. Entropy values of the single mutant allele frequencies computed via bi- and 4-allelic models are strongly correlated.

those loci where the single mutant allele frequency exhibited at least 1% variation. This threshold was chosen based on the sensitivity of RVHaplo, which has been demonstrated to accurately recover variants with as little as 1% prevalence.

3.2.8 HSV-1 data

We used a real dataset of HSV-1 [183] to demonstrate the SV detection scheme of SLAFIT. The dataset comprised 104,524 ONT reads with mean length 1,294 bp. 15 reads were ultralong, spanning more than 140 kbp. The reads were aligned against the HSV-1 reference sequence (NCBI accession no. KU310657.1) and SV detection performed as described before.

3.3 Results

3.3.1 LD estimation and fitness inference via SLAFIT is robust to varying sequencing parameters

In real-life sequencing scenarios, control over sequencing parameters is limited by time and resource constraints. For archival samples, analysis must rely on already sequenced datasets without any control over the sequencing parameters. We tested the performance of the fitness inference scheme of SLAFIT for various values of sequencing parameters to check the robustness to varying levels of noise in the data. We performed inference on the sequences of

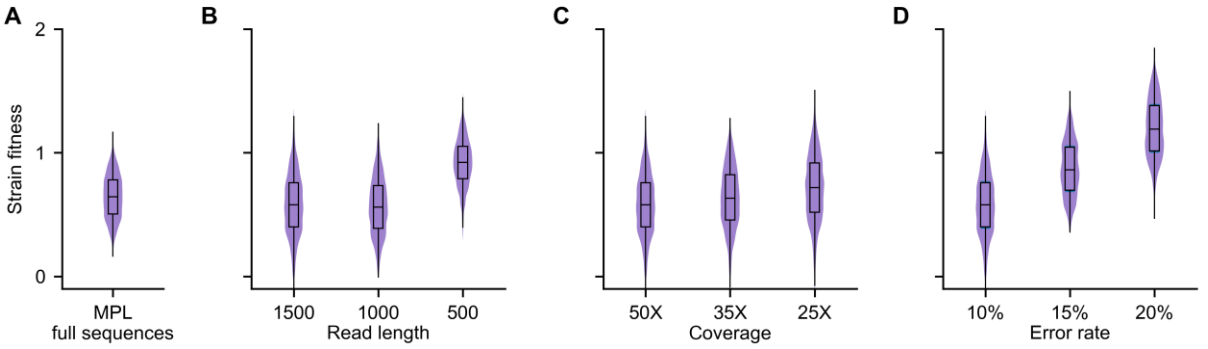


Figure 3.7 SLAFIT can perform over a range of sequencing parameters. **(A)** A distribution of the strain-level fitness obtained from the ground truth population is shown. The strain-level fitness, estimated via SLAFIT is shown while varying the **(B)** read length, **(C)** sequencing coverage, and **(D)** error rate. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of length $L = 4500$ bp. The genomes consisted of bi-allelic (WT and mutant) loci with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $180/180/4140$ loci were respectively beneficial/deleterious/neutral with selection coefficients $+0.025/\text{-}0.025/0$. The beneficial and deleterious alleles were uniformly distributed across the entire genome. One founder sequence was used to generate each population, which was allowed to evolve for 550 generations and sampled at $\Delta t = 50$ generations. Unless otherwise stated, TGS reads of mean length 1500 bp were simulated with 10% error rate and coverage 50X.

the ground truth population to use as a baseline (Fig. 3.7A), and varied the read length (Fig. 3.7B and Fig. 3.8A), coverage (Fig. 3.7C and Fig. 3.8B), and error rate (Fig. 3.7D and Fig. 3.8C). SLAFIT performed well for some degradation in the read quality. The performance demonstrated significant degradation when the error rate was set to 20% or a read length of 500 bp was used. The errors in the reads or the information lost due to the smaller read length propagate into the reconstructed haplotypes and their prevalence values, which affect the estimated LD and the estimates of the selection coefficients. Though these values of error rate and read-length are used for benchmarking purposes and may not be encountered when using modern sequencing machines. These results show that accuracy of fitness inference is directly correlated with the data quality and can serve as a guideline for designing sequencing experiments.

In all the tested datasets, SNVs were spread across the genome, and long-range LD values were not directly observable via TGS reads. Nevertheless, the haplotype reconstruction scheme of

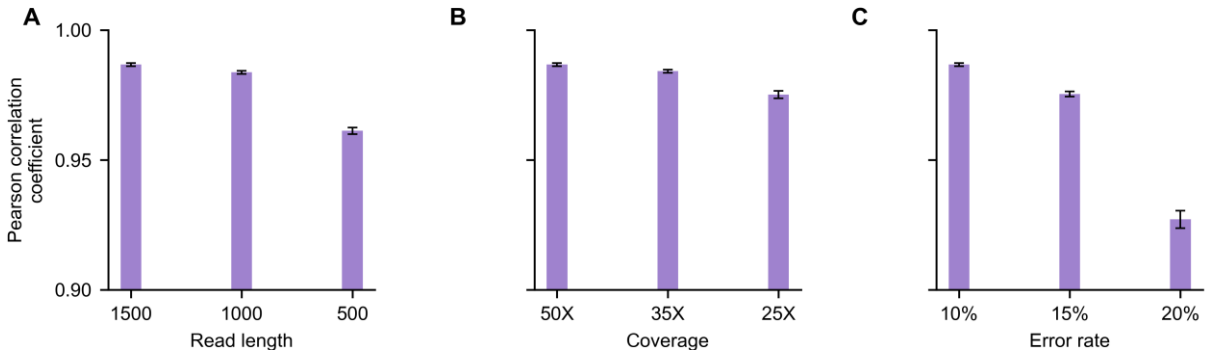


Figure 3.8 The strain-level fitness inference of SLAFIT is strongly associated with the ground-truth with varying **(A)** read length, **(B)** sequencing coverage, and **(C)** error rate. The bars show the mean Pearson correlation coefficient over 100 Monte Carlo runs. The error bars show the standard error of the mean. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of length $L = 4500$ bp. The genomes consisted of bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $180/180/4140$ loci were respectively beneficial/deleterious/neutral with selection coefficients $+0.025/\text{-}0.025/0$. The beneficial and deleterious loci were uniformly distributed across the entire genome. One founder sequence was used to generate each population, which was allowed to evolve for 550 generations and sampled at $\Delta t = 50$ generations. Unless otherwise stated, TGS reads of mean length 1500 bp were simulated with 10% error rate and coverage 50X.

the pipeline accurately reconstructed the full-length sequences, from which the unobservable long-range LD values were faithfully recovered, enabling accurate fitness inference. Figure 3.9 shows Circos plots [184], [185] for LD values aggregated over the entire evolutionary history. Figure 3.9 A and B demonstrate that the ground truth and inferred aggregated LD are similar whereas the inferred selection coefficients have the sam e sign. Figure 3.9 C presents the summary statistics of the comparison of the inferred and ground truth aggregated LD over 100 Monte Carlo runs. We also quantified the accuracy of LD recovery via the Pearson correlation coefficient between the LD values obtained from the reconstructed sequences and the LD values obtained from the sequences of the ground truth population (Figure 3.10). The recovered LD values exhibited strong correlation with the ground truth LD values. The relatively weaker correlation observed at the initial generations is because of lower variation in the data which affects haplotype reconstruction. The correlation values of the first two time-points are not

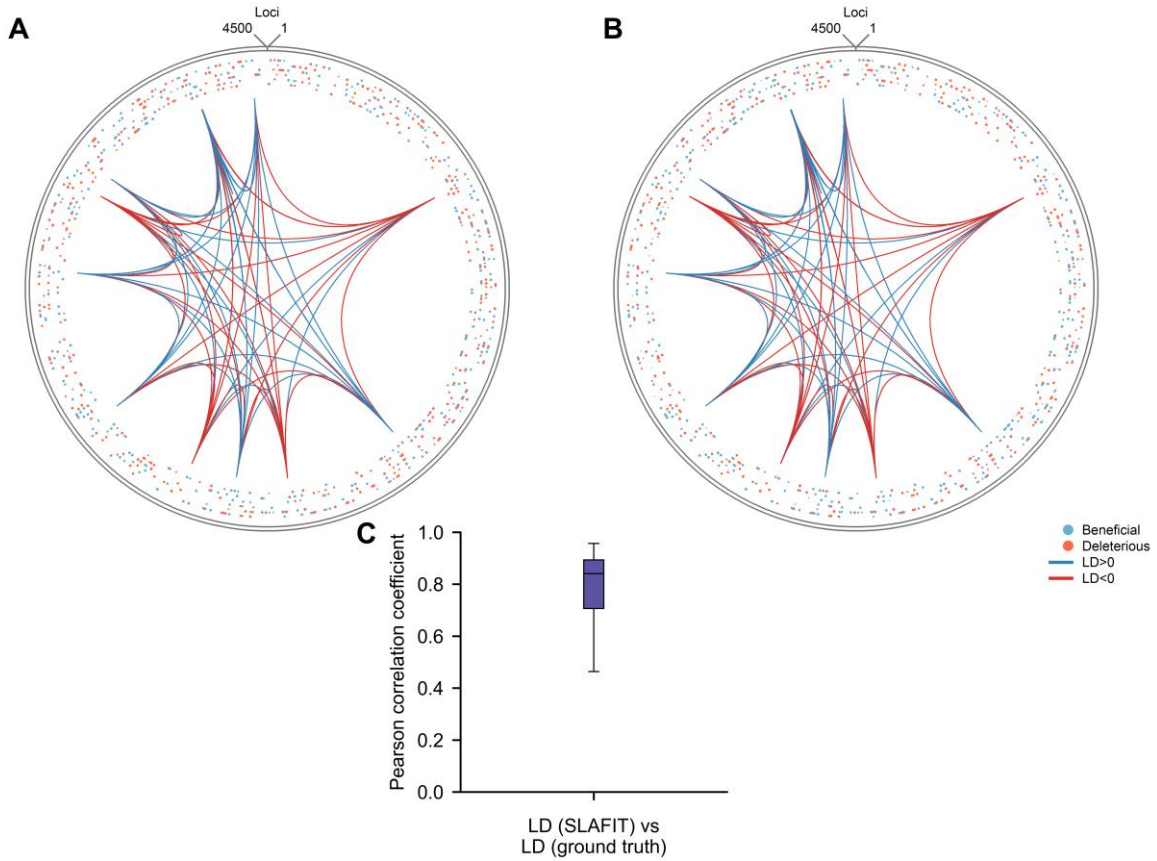


Figure 3.9 SLAFIT accurately recovers long-range LD values. **(A)** Aggregated LD values obtained from the ground truth population and selection coefficients inferred from those. **(B)** Aggregated LD values obtained via SLAFIT and selection coefficients inferred from those. **(C)** Aggregated LD values inferred via SLAFIT show good correlation with the ground truth aggregated LD values. Results in (A) and (B) are shown for a typical Monte Carlo run. log magnitudes of the inferred selection coefficients are indicated as the radii of the circles. Blue circles represent positive values, whereas red circles represent negative ones. Links between loci indicate LD values aggregated over all time-points. The widths of the link indicate the magnitude of the aggregated LD, and blue/red colors indicate positive/negative sign of the aggregated LD. Only the strongest 65 aggregated LD values are shown to avoid clutter. Results in (C) are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of genome length $L = 4500$ bp. The genomes consisted of bi-allelic (WT and mutant) loci with equal forward and backward SNV probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $180/180/4140$ loci were respectively beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$. The beneficial and deleterious mutants were uniformly distributed across the entire genome. One founder sequence was used to generate each population, which was allowed to evolve for 550 generations and sampled at $\Delta t = 50$ generations. Unless otherwise stated, TGS reads of mean length 1500 bp were simulated with 10% error rate and 50X coverage.

shown as haplotype reconstruction failed to return haplotypes for most Monte Carlo runs due to low variation, and consequently, the consensus sequences of the reads were used in subsequent analysis of those time-points. The lack of variation at initial time-points is

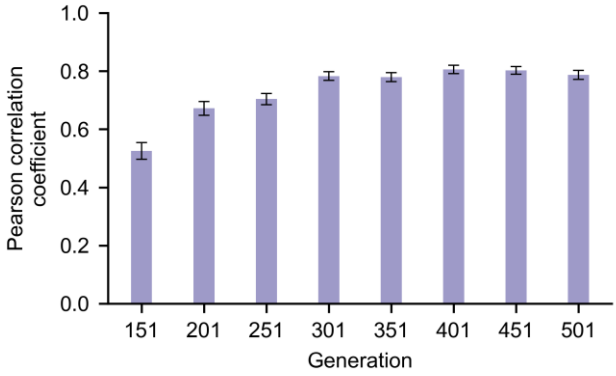


Figure 3.10 LD inference via SLAFIT is accurate throughout the evolutionary history. The later time-points exhibit stronger values of the Pearson correlation coefficient because these contain more variation which improves the accuracy of the haplotype reconstruction process. The bars show the mean Pearson correlation coefficient over 100 Monte Carlo runs. The error bars show the standard error of the mean. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of length $L = 4500$ bp. The genomes consisted of bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $180/180/4140$ loci were respectively beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$. The beneficial and deleterious loci were uniformly distributed across the entire genome. One founder sequence was used to generate each population, which was allowed to evolve for 550 generations and sampled at $\Delta t = 50$ generations. TGS reads of mean length 1500 bp were simulated with 10% error rate and coverage 50X.

characteristic of natural infections in which the infecting strain initially dominates the population. For these time-points, other strains are either not present or present at very low prevalence values and consequently, the consensus strains can approximate the population well. For subsequent time-points, enough variation exists for population reconstruction to identify unique haplotypes.

3.3.2 Longer read length of TGS reads yields improved fitness estimates compared to NGS reads

We compared strain-level fitness estimates reported by TGS and NGS to demonstrate the advantage of the longer reads of TGS. We obtained TGS reads and performed haplotype reconstruction as described in the Methods section. We then used ART [120] to obtain single-end NGS reads with a coverage similar to that of the TGS data. We simulated 14500 reads of 150 bp using the error profile of the Illumina HiSeq 2500 platform and applied MPL-R [186]

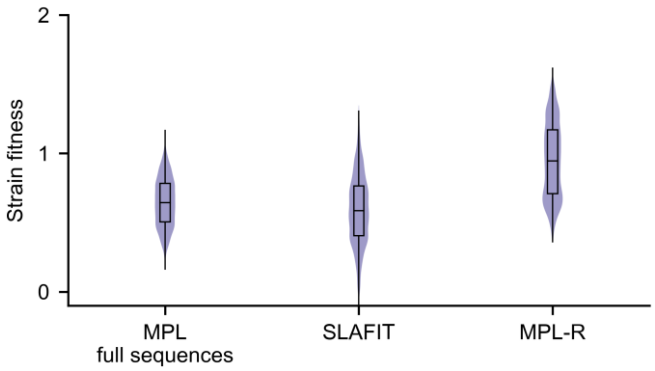


Figure 3.11 SLAFIT is able to reconstruct fitness values from TGS reads more accurately as compared to an NGS-based scheme. The distributions are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of genome length $L = 4500$ bp. The genomes consisted of bi-allelic (WT and mutant) loci with equal forward and backward SNV probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $180/180/4140$ loci were respectively beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$. The beneficial and deleterious mutants were uniformly distributed across the entire genome. One founder sequence was used to generate each population, which was allowed to evolve for 550 generations and sampled at $\Delta t = 50$ generations. TGS reads of mean length 1500 bp were simulated with 10% error rate and 50X coverage. $14{,}500$ single-end NGS reads of 150 bp were simulated to obtain similar coverage as the TGS reads.

to obtain the fitness estimates. Fig. 3.11 demonstrates the distribution of the strain-level fitness estimates obtained via SLAFIT and MPL-R. SLAFIT yielded estimates that more closely matched those obtained from the ground truth population as compared to MPL-R. The Pearson correlation coefficients of the respective estimates also indicated that SLAFIT yielded fitness values that were more strongly associated with the ground truth as compared to MPL-R (Fig. 3.12). To ensure a fair comparison, we attempted to use the same haplotype reconstruction software for both TGS and NGS data, but RVHaplo performed poorly on NGS data, and Quasirecomb [113], the software used in MPL-R, was too computationally expensive to run on TGS data. This is expected as software designed for either technology is optimized to handle the data of only its intended technology. Therefore, we used each reconstruction software for the specific technology for which it was developed. While the usage of different software for haplotype reconstruction introduces a confounder, the difference in performance between the two technologies is qualitatively similar to what is observed when TGS reads of different length

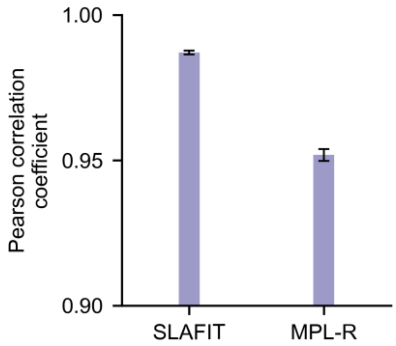


Figure 3.12 The association between the estimates of SLAFIT and the ground truth values is stronger as compared to an NGS based scheme. The bars indicate the mean Person correlation coefficient values. The error bars indicate the standard error of the mean. The statistics are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of length $L = 4500$ bp. The genomes consisted of bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $180/180/4140$ loci were respectively beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$. The beneficial and deleterious loci were uniformly distributed across the entire genome. One founder sequence was used to generate each population, which was allowed to evolve for 550 generations and sampled at $\Delta t = 50$ generations. TGS reads of mean length 1500 bp were simulated with 10% error rate and coverage 50X. 14500 single-end NGS reads of read length 150 bp were simulated to obtain similar coverage as the TGS reads.

are used, which indicates that the read length is likely the major factor for the difference in performance between the data from the two technologies. This can be explained in terms of the error rates of both technologies as well. NGS typically has a lower error rate compared to TGS, so for the same read length, NGS is expected to yield better performance compared to TGS. Hence the observed difference in performance is explained by TGS capturing more information from the reads due to its greater read length.

3.3.3 SLAFIT outperforms existing LD-aware fitness inference

Inference methods that take into account LD are designed only for NGS data with a few SNVs and scale poorly when applied to larger systems [61], [62], [187]. These methods have been shown to require several hours to run on a single Monte Carlo run [186], [188] and were not used for comparison. The only LD-aware method that can process sequences of length 4500 bp without requiring a large wall clock time is the method of Li and Barton 2023 [20]

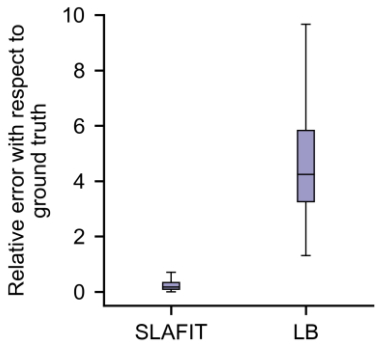


Figure 3.13 Fitness values inferred via SLAFIT have lower relative error with respect to the ground truth as compared to LB, the only LD-aware inference method that can scale up to systems of kpb scale. The relative error is defined as the absolute difference between the strain fitness inferred by the method under test and MPL on the ground truth sequences divided by the absolute value of the fitness inferred by MPL on the ground truth sequences. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of evolving populations of $N = 1000$ individuals of length $L = 4500$ bp. The genomes consisted of bi-allelic (WT and mutant) loci, with equal forward and backward mutation probabilities set to $\mu = 10^{-4}$ per locus per generation. Alleles at $180/180/4140$ loci were respectively beneficial/deleterious/neutral with selection coefficients $+0.025/-0.025/0$. The beneficial and deleterious loci were uniformly distributed across the entire genome. One founder sequence was used to generate each population, which was allowed to evolve for 550 generations and sampled at $\Delta t = 50$ generations. TGS reads of mean length 1500 bp were simulated with 10% error rate and 50X coverage.

(subsequently referred to as LB). We used single mutant allele frequencies obtained from the TGS reads as input to LB, but it performed poorly compared to SLAFIT (Fig. 3. 13). Thus, to our knowledge, SLAFIT remains the only method that can efficiently perform accurate LD-aware inference.

3.3.4 SLAFIT reveals insights from prolonged infection data consistent with biological knowledge

We applied SLAFIT on prolonged intrahost EV-C104 infection data. Haplotype reconstruction of the data returned {15,13,8,11} unique haplotypes at each of the time-points respectively. We analyzed the data via SLAFIT and as a benchmark, also performed LD-independent inference via SLAFIT by setting the off-diagonal entries of the LD matrix to zero. A plot of the ratio of the selection coefficients for nonsynonymous SNVs inferred via SLAFIT and SLAFIT (LD-independent) indicates that SLAFIT reports much higher selective pressure on

NT mutation	AA mutation	Gene	Predicted B-cell epitope	Gene function
C3086T	T214I	VP1	SVVPLNTVDA GASNKY	Receptor-binding domain
A3313G	T290A	VP1	RAADYVGRSV DYNQNSPVMI ATAD	Receptor-binding domain
A5158G	R34G	3A	VDSEEVREY	Viral replication
T6272C	L112P	3D	TDGLEALDLT	Viral replication

Table 3.1 Nonsynonymous SNVs identified as highly fit via linkage-aware inference occur in computationally predicted B-cell epitopes.

five SNVs as compared to SLAFIT (LD-independent) (Fig. 3.14). Experimental evidence of the significance of these SNVs is unavailable and the virus has no known lab-validated epitopes. Therefore, to validate our findings, we predicted the B-cell epitopes of the virus via the Immune Epitope Database (IEDB) [189] using the Bepipred Linear Epitope Prediction 2.0 tool [190] with default threshold of 0.5. Four of the five SNVs classified as beneficial by SLAFIT and neutral by SLAFIT (LD-independent) were observed to occur in the predicted B-cell epitopes (Table 3.1). This is consistent with biological knowledge that epitopes are under immune pressure and nonsynonymous SNVs in epitopes have high fitness as these likely assist in immune escape. This demonstrates that SLAFIT, via the recovery of unobservable LD information, can infer insights from real data that are missed if LD is not considered.

Here we must clarify that the epitopes were predicted computationally due to the unavailability of experimentally-determined epitopes of the virus. These predictions were generated to identify whether the nonsynonymous SNVs lie in regions predicted to be surface-exposed and immunogenic. However, these predictions represent computational concordance, not experimental validation. We therefore use the results of these predictions as a biologically informed benchmark to assess the validity of our results, rather than as concrete evidence of immune escape. Future work employing n eutralization assays can experimentally validate the significance of these SNVs.

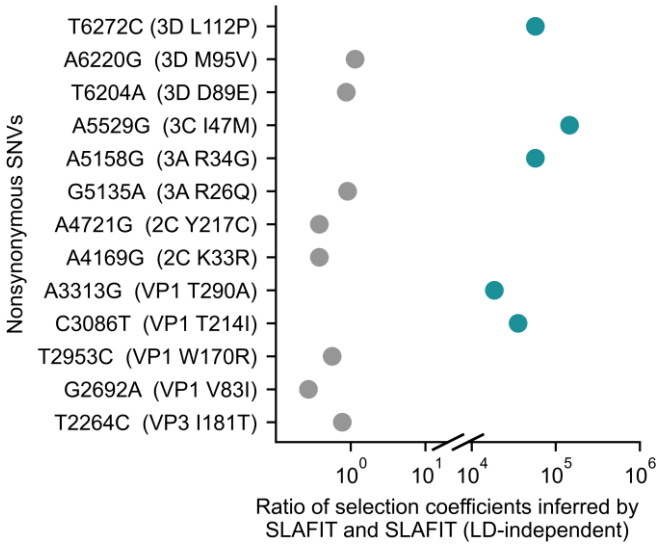


Figure 3.14 SLAFIT infers a number of nonsynonymous SNVs as non-neutral (blue) which are identified as neutral by SLAFIT(LI). Four of these SNVs map to B-cell epitopes predicted via IEDB and are described in Table 3.1.

The LD values obtained via SLAFIT, aggregated over the complete evolutionary history, were mostly negative, which suggested clonal interference between nonsynonymous SNVs (Figure 3.15). This observation suggested competition between strains within the population, which was also evident by the decreasing number of reconstructed haplotypes with time. The decreasing single mutant allele frequencies (Figure 3.5) and the decreasing haplotype diversity (Figure 3.16) also indicated strains being outcompeted and removed from the population.

3.3.5 SV detection via cuteSV is robust and consistent with known results

We performed SV detection on the dataset of HSV-1 by considering all available reads. The SV in the U_L region reported in Saranathan *et. al* 2022 [183] was not detected, which is expected, since the number of reads that can cover the entire length of the U_L region (110,026 bp) was 19 (Figure 3.17). We were able to detect SVs in the U_S region, out of which two SVs matched the inversion in the U_S region reported in the original study. This dataset, however, had ultralong reads (exceeding the length of the U_S region), which are not always available in TGS datasets. To validate the robustness of SV detection to more realistic reads, we sequentially removed reads greater than a fraction $(0.75, 0.50, 0.25,$ and $0.125)$ of the length

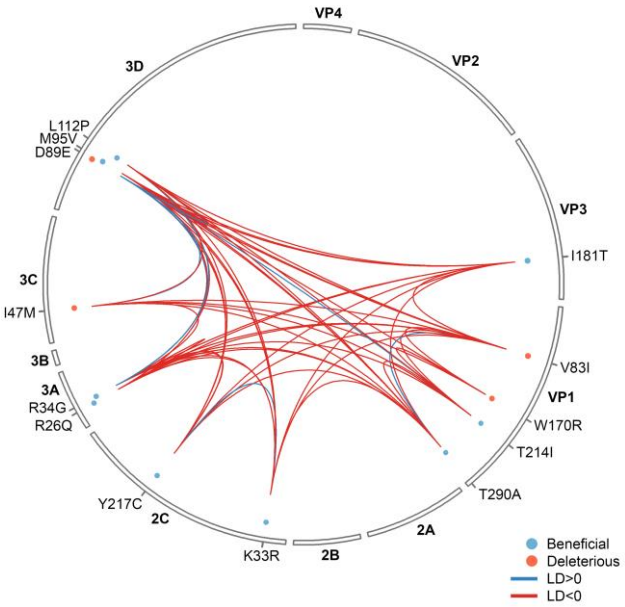


Figure 3.15 LD patterns in the EV-C104 infection data show evidence of clonal interference. log magnitudes of the inferred selection coefficients are indicated as the radii of the circles. Blue circles represent positive values, whereas red circles represent negative ones. Links between loci indicate LD values aggregated over all time-points. The widths of the link indicate the magnitude of the aggregated LD, and blue/red colors indicate positive/negative sign of the aggregated LD.

of the U_S region and repeatedly performed SV detection. The inversion in the U_S region was detected even when the maximum length of the reads was as low as 0.25 that of the region. This shows that the detected SVs are consistent with the SV reported in the original study. The SVs have been summarized in Table 3.2.

3.4 Discussion

In this work, we have presented SLAFIT for the detection of SVs and computation of fitness from TGS data. We have demonstrated the performance of SLAFIT on simulated data under varying read lengths, coverage values, and error rates. We have shown that fitness inference via SLAFIT is accurate, completes in a smaller wall clock time compared to MPL, and outperforms MPL-R, and LB, the only other LD-aware methods capable of scaling up to sequence lengths in the kbp range. SLAFIT yielded biologically consistent insights from EV-C104 infection data, detecting four strong nonsynonymous SNVs in computationally predicted

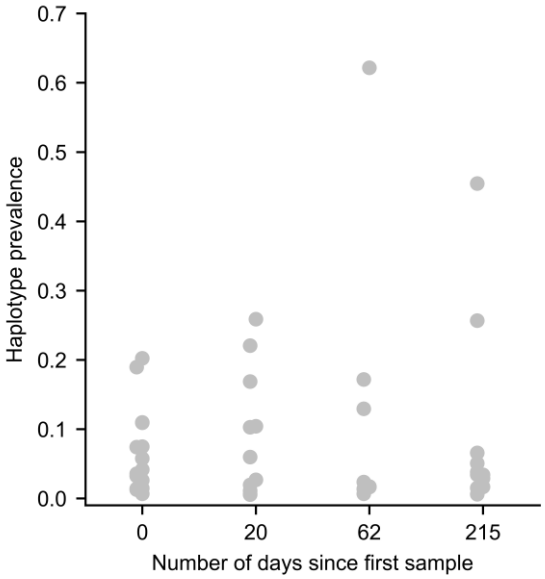


Figure 3.16 Prevalence values of haplotypes reconstructed from the EV-C104 infection data indicate decreasing diversity with time. The first two time-points show multiple haplotypes with similar prevalence values, whereas the last two time-points show a segregation between dominant haplotypes and the rest of the population.

B-cell epitopes that were missed by an LD-independent approach. The SV detection scheme used in SLAFIT, cuteSV, detected SVs in an HSV-1 dataset even when the maximum read length was as short as one-quarter of the length of the target region. These results demonstrate that SLAFIT enables evolutionary analysis from TGS data that is not possible with other methods.

In terms of limitations, SLAFIT is limited by the usage of an additive fitness model. Fitness may not always be additive in nature and epistatic interactions between SNVs warrant the

Max read length (bp)	SV detected	Start	End	Read support
NA (reference)	U_S gene	137,276	150,359	NA
184,240	SV 1	135,457	150,377	14
184,240	SV 2	137,200	150,373	11
9,813	SV 1	135,457	150,377	14
6,542	SV 1	137,200	150,373	11
3,271	SV 1	137,200	150,373	11
1,635	SV 1	137,197	150,376	10

Table 3.2 SVs detected from HSV-1 data under varying maximum read length thresholds. The U_S gene (reference) is shown for comparison.

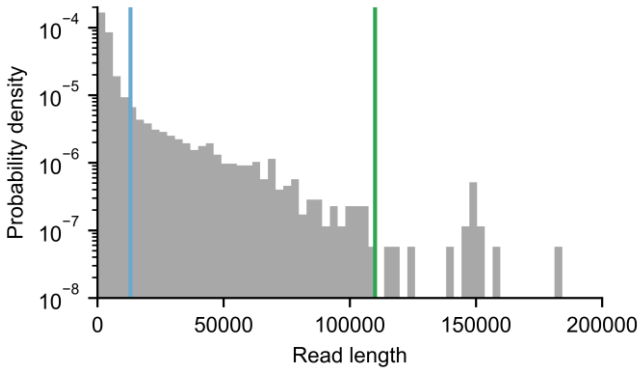


Figure 3.17 Distribution of read lengths in the HSV-1 dataset. Multiple reads are longer than the length of the U_S region (blue line) and can capture SV in it. Some ultralong reads are longer than the length of the U_L region (green line).

incorporation of triple- and quadruple-mutant allele frequencies. While mathematical frameworks exist for the incorporation of epistasis [106], [191], efficiently computing and storing quadruple-mutant allele frequencies for high-throughput data of systems of the scale of kbp remains a challenge. Further research can attempt to address this limitation.

Another limitation is the use of the 1% threshold for differentiating between true variants and sequencing errors. While this threshold is consistent with published benchmarks on RVHaplo, sequencing conditions and strain abundance may vary between datasets. To obtain an accurate threshold for a particular dataset under a particular set of sequencing conditions, simulations may need to be carried out to ascertain the exact prevalence threshold below which it becomes challenging to distinguish actual variants from errors. Future work can investigate this question.

One of the most valuable tools available to clinicians is whole-genome third-generation sequencing of time-series patient samples [192], [193]. Commonly used pipelines are limited to consensus-sequence based analysis and cannot resolve the interaction between mutations at the strain level. SLAFIT provides an integrated pipeline that can obtain accurate insights from TGS data at a finer scale. The core of fitness inference of SLAFIT is RVHaplo, which enables the pipeline to be able to run on a number of sequencing machines. Fitness inference via SLAFIT is also designed to be modular, so users can swap individual software components

with more familiar ones. The SV detection scheme of SLAFIT is based on a robust software, cuteSV, whose performance has been experimentally validated on multiple real datasets. By enabling strain-level evolutionary inference, SLAFIT enables more precise monitoring of pathogen evolution in real-world settings.

Chapter 4

Linkage disequilibrium estimation from NGS data

4.1 Introduction

4.1.1 Overview

Viral populations exist as quasispecies of closely-related variants. Only those variants that possess combinations of mutations conferring an overall fitness advantage survive and replicate. Disentangling the effects of individual mutations from their genetic background is therefore essential for understanding the true selective nature of mutations. This inference problem becomes particularly challenging due to LD, the non-random association between mutations, which can cause neutral mutations to appear fit and vice versa. LD inference can be performed via population reconstruction, which reconstructs complete haplotypes before obtaining LD. However, this approach is computationally slow, memory-intensive, and prone to errors, particularly in regions of low coverage. In this work, we present a novel, population reconstruction-independent method of LD estimation. Our approach formulates LD estimation as a matrix completion problem and solves it via alternating projections on constraints of positive semidefinite matrices and known entries within a band. We demonstrate that our method yields higher accuracy while being two orders of magnitude faster compared to population reconstruction-based estimation. We demonstrate the performance of the method on both simulated as well as real data showing that it can estimate LD from a variety of populations.

4.1.2 Background

Viral evolution is characterized by unusually high mutation rates, generally resulting from the lack of high-fidelity proofreading mechanisms capable of correcting mutations during replication. As a consequence, viral populations rapidly acquire mutations and exist as

quasispecies of diverse groups of genetically distinct variants competing for the same resources. Within such viral populations, only those variants that possess combinations of mutations conferring an overall fitness advantage over other combinations survive in the long-term. The identification of variants harboring advantageous mutations, followed by the characterization of those mutations, is a crucial step in understanding viral pathogens. However, the presence of interactions between mutations renders this problem particularly challenging. Mutations can both compete with and enhance each other, meaning that the observed prevalence of an individual mutation reflects not only its inherent fitness but also the genomic background on which it occurs. Neutral or even deleterious mutations may hitchhike alongside beneficial mutations and consequently appear at high prevalence in the population. Conversely, beneficial mutations may be purged from the population due to competition with fitter mutations. When mutations interact with each other in non-random ways, these are said to be in LD. LD is prevalent in nature and has been documented in a number of viral pathogens of clinical interest, such as HIV [13], HCV [194], influenza [14], [15], and SARS-CoV-2 [186]. Accurately estimating LD is therefore essential for understanding the inherent selective nature of individual mutations and for distinguishing truly virulent variants from those that merely ride to prominence due to the presence of linked loci.

4.1.3 Related work

Viral populations are typically sequenced via NGS and numerous viral datasets sequenced using this technology are available in public repositories. NGS however suffers from the technological limitation of read length, which is dictated by the sequencing platform and the specific reagent chemistry employed. The read length of NGS is typically much shorter than viral genomes, for example the 300 bp read length of the Illumina MiSeq platform is insufficient to read the entire HIV genome, which is 9719 bp in length, in a single read.

Consequently, the short read-length of NGS renders the majority of the LD information across the viral genome unobservable.

Existing methods of LD estimation [60], [61], [62] suffer from poor scalability and cannot be applied even to small viral genomes. To address this limitation of LD estimation methods, we had previously presented MPL-R [186], which introduced parallelized reconstruction of subsampled populations, yielding a fast and scalable method for estimating unobservable inter-read LD values. MPL-R, however, is affected by two notable limitations.

The first limitation is that despite the speedup attained via parallelized population reconstruction, the reconstruction process remains computationally intensive and consumes large amounts of memory. More fundamentally, while population reconstruction solves the problem of LD estimation, it does so via the estimation of higher order statistics which are not actually needed for this task. LD is a two-point metric that only requires computation of single- and double-mutant allele frequencies. Population reconstruction, by contrast, reconstructs all higher-order frequencies across the entire genome, which not only increases the wall clock time but also generates extraneous data not necessary for the problem at hand. The second limitation of MPL-R is that population reconstruction introduces errors of its own, particularly in genomic regions where the coverage is low. Thus, MPL-R, despite being the only method that can efficiently estimate LD from NGS data, leaves considerable room for improvement.

4.2 Methodology

4.2.1 Overview

This work addresses the LD estimation problem without relying on population reconstruction. The proposed scheme, which we refer to as LDEN (LD estimation from NGS data), takes NGS reads as input, estimates intra-read LD values from the reads, and then uses these known LD values to estimate the unobservable inter-read LD across the entire genome.

The estimation of unknown inter-read LD values is based on the alternating projections method (APM) [195], which was modified in this work to the LD matrix completion problem. APM was designed to use partial complete correlation matrices to approximate full correlation matrices that best describe the observed correlations. Because LD matrices are essentially covariance matrices, APM can be modified for estimation of unobservable LD as well. We also exploited the banded structure of the observed LD matrix into APM to speed up computations. Specifically, we formulate the problem as finding an LD matrix that contains the observed LD values for locus pairs within one read-length of each other, and approximations of the unobservable LD values such that the overall matrix is positive semidefinite.

We demonstrate that LDEN not only provides accurate estimates of unobservable LD values but is also orders of magnitude faster than population reconstruction-based estimation. This computational advantage stems from the fact that LDEN operates directly on observable two-point statistics, bypassing the requirement of reconstructing full sequences or higher-order statistics. We further demonstrate the performance of LDEN on populations with different values of evolutionary parameters as well as under different levels of sequencing noise in the NGS reads. Finally, we demonstrate the practical utility of LDEN in inferring biologically meaningful insights from real data of SARS-CoV-2 infection.

4.2.2 Read alignment and partial LD calculation

NGS reads are first aligned to a user-supplied reference sequence to obtain BAM files which are then sorted and indexed. The LDEN pipeline uses these aligned reads to construct a read-alignment matrix in which the presence or absence of a read covering each locus is encoded. Read coverage i.e. the number of reads covering a locus, is used to filter out loci that are covered by very few reads. This coverage threshold is user-defined and was chosen as 5 reads in our tests. For loci with sufficient coverage, alleles are encoded to numerical values based on a binary (WT/mutant) or quaternary (A,C,G,T) encoding scheme depending on the variation in

the data. Single mutant allele frequencies are calculated at each locus by dividing the observed counts of the mutant allele by the read coverage. Loci exhibiting low variation loci are removed from subsequent analysis, as these don't contribute much to LD estimation and unnecessarily increase the number of computations. A threshold of 1% on the single mutant allele frequencies was chosen to differentiate between variable and conserved loci.

In NGS data, coverage varies across loci, which complicates the computation of LD. In the equation for the coefficient of LD, $D = f_{ij} - f_i f_j$, if f_i is calculated over n_i reads covering locus i, f_j over n_j reads covering locus j, and f_{ij} over n_{ij} reads covering both loci, and if $n_i \neq n_j \neq n_{ij}$, then the resulting value of D may violate the axioms of probability due to inconsistent denominators. To circumvent this issue, we compute f_i, f_j, and f_{ij} using only n_{ij} reads that cover both loci simultaneously, ensuring the denominators are consistent. Although this means that the calculated f_i and f_j are not universal but rather pairwise specific, it preserves the probabilistic interpretation of D.

When the number of reads covering a pair of loci becomes small, the resulting LD estimates are affected by statistical noise of the order of $1/n_{ij}$. This can lead to severe overestimation or underestimation of true LD values. As the coverage of NGS typically follows a trapezoidal profile, with loci close to the ends of the genome having lower coverage values compared to loci close to the center, this problem requires explicit treatment. To minimize the effects of low coverage, we employ a shrinkage estimator of the form

$$\hat{f}_{ij} = w.f_{ij}^{observed} + (1 - w).f_i f_j,$$

where $f_{ij}^{observed}$ represents the observed double mutant allele frequency, w is a coverage-dependent shrinkage term, and $\hat{f}_{ij}$ is the shrunken estimate of the double mutant allele frequency used in subsequent analysis. Intuitively, w quantifies how informative $f_{ij}^{observed}$ is. For high values of coverage, w approaches 1, and $\hat{f}_{ij}$ is essentially taken as $f_{ij}^{observed}$, whereas

when the coverage is low, $f_{ij}^{observed}$ is assigned a lower weight and $\hat{f}_{ij}$ is shrunk towards the product $f_i f_j$ (i.e. towards linkage equilibrium) to avoid sampling errors. The shrinkage weight w is defined as

$$w = \frac{n_{ij}}{n_{ij}+\lambda},$$

where λ is a user-defined shrinkage factor. In our tests, we set λ empirically based on the noise levels in the sequencing data.

4.2.3 Alternating projections method

The partial LD matrix obtained directly from the reads does not satisfy the positive semidefinite property required for covariance matrices, and hence, is not a valid covariance matrix. However, it can be used to obtain a covariance matrix that best explains the observed LD values while satisfying the positive semidefinite constraint. Formally, for the observed LD matrix D, and a candidate covariance matrix C, the problem is to minimize the distance $\gamma(D)$ between D and C such that

$$\gamma(D) = \min\{\|D - C\|\},$$

where the norm represents the squared Frobenius norm, i.e. $\|D\| = \sum_{i,j} d_{ij}^2$. The matrix that achieves this minimum distance is the required covariance matrix. Such a matrix is unique and its existence is a known result in convex optimization [196]. This matrix can be obtained by exploiting two properties, i.e. the LD matrix must positive semidefinite and its entries within the read length must equal the observed LD values, mathematically

$$S = \{Y = Y^T \in \mathbb{R}^{n\times n} : Y \succeq 0\},$$

$$B = \left\{Y = Y^T \in \mathbb{R}^{n\times n} : y_{ij} = d_{ij}, j - i < r\right\}.$$

Here r represents the read length, n is the total number of loci, and $Y \succeq 0$ indicates that Y is positive semidefinite. Both S and B are closed convex sets, therefore their intersection is also

a closed convex set. The matrix that satisfies both constraints can be obtained iteratively by a alternately projecting D on to S and B.

In the alternating projection method, an eigendecomposition of D is first obtained and all negative eigenvalues are set to a small positive value to enforce the positive semidefinite constraint. This is followed by replacing the observed LD entries, corresponding to intra-read loci pairs, back into the iterate, thereby enforcing the banded constraint. These two projections are alternated until the difference between the Frobenius norm of successive iterates falls below a specified convergence threshold, ε, or the maximum number of iterations, M, is reached. Empirical tests on convergence discuss these in detail in the Results section. The complete algorithm is presented below.

Initialize $Y_0 = D$

for $k = 1, 2, \ldots, M$

$Y_{k,j-i<r} = D_{j-i<r}$ for all $j - i < r$

$U\Lambda U^T = Y_k$

$\Lambda_{ii} = \max(\Lambda_{ii}, 10^{-8})$ for all i

$Y_k = U\Lambda U^T$

if $\|Y_k - Y_{k-1}\| < \varepsilon$

break

Upon convergence, the final iterate Y_k is the required completed LD matrix. It is positive semidefinite, has all observed intra-read LD values, and has the remaining entries consistent with the defined constraints.

4.2.4 LD estimation from paired-end reads

In addition to single-end sequencing, many NGS sequencing platforms support paired-end sequencing, in which both ends of a genomic fragment are sequenced. Paired-end sequencing provides two key advantages of single-end sequencing: (i) it allows the observation of the same

genomic fragment from both directions, which provides more information about the composition of the fragment; and (ii) these reads capture LD across longer ranges compared to single-end reads, as the two reads in a pair originate from the same fragment and can span regions that cannot be covered by single-end reads. Consequently, incorporating paired-end reads into the LDEN framework is expected to improve the accuracy of the recovered LD matrix by increasing the proportion of the directly observable LD entries.

We developed a two-step approach to extend the LDEN framework to handle paired-end data. In the first step, the intra-read LD is calculated exactly as described for the single-end case. In the second step, we exploit the fragment-level LD observed by read pairs. For any locus pair, whether the individual loci are separately covered by single reads from the pair or by both reads, we utilize this information to obtain the double mutant allele frequency for the pair. This approach extends the range of the observable LD beyond the length of a single read, increasing the genomic distance across which LD can be directly observed.

Similar to single-end reads, paired-end reads are affected by sampling noise, particularly when the number of read pairs covering a locus pair is small. We therefore employ a pairwise coverage threshold, considering only those locus pairs that were covered by at least five read pairs. Locus pairs covered by fewer than five pairs are excluded from the observed set of loci and are left for imputation by the alternating projections algorithm. Additionally, we apply the same linear shrinkage estimator described in Section 4.2.2 to the double mutant allele frequencies and use the same value of the shrinkage factor as the singe-end reads case. This ensures consistent regularization in the two-step approach and uniformly mitigates the effects of statistical noise in regions with low coverage.

4.2.5 Simulated data

We simulated data based on realistic viral populations to perform analytical validation of LDEN. A total of 6 primary datasets were generated, each with varying population parameters.

Dataset	Founder 1	Founder 2	Founder 3	Founder 4
1	$U(0.80,0.90)$	$U(0.01,0.05)$	Rest/2	Rest/2
2	$U(0.40,0.45)$	$U(0.40,0.45)$	Rest/2	Rest/2
3	$U(0.25,0.30)$	$U(0.25,0.30)$	$U(0.25,0.30)$	Rest
4	$U(0.20,0.25)$	$U(0.20,0.25)$	$U(0.20,0.25)$	Rest

Table 4.1 Prevalence values of the four variants used in datasets 1 to 4. $U(a,b)$ represents a uniform distribution over the interval $[a,b]$. "Rest" indicates the remaining prevalence after assigning prevalence values to all other variants.

The total number of sequences in every dataset was set to 1000.

4.2.5.1 Variation in population diversity

For datasets 1 through 4, a random founder sequence of length 500 bp was generated by appending nucleotides uniformly at random, with each nucleotide having equal probability of occurrence, i.e. 0.25 for A, C, G, and T. From this founder sequence, 3 variant sequences were generated by randomly mutating 1%, 2%, and 5% of the loci, respectively. Co-infection dynamics were simulated by assigning prevalence values to each variant according to the distributions shown in Table 4.1.

In dataset 1, a single dominant variant (founder 1) comprises the majority of the population, with the remaining prevalence distributed among the other variants. Dataset 2 represents a population with two co-dominant founders. Dataset 3 contains three founders with equal prevalence, whereas dataset 4 comprises four founders with similar prevalence.

4.2.5.2 Variation in genome length

In datasets 5 and 6, the prevalence values were kept the same as those of dataset 1, while the genome length was increased to 1000 bp and 1500 bp respectively. These datasets were generated to evaluate the performance of LDEN as genome size was increased.

4.2.5.3 Wright-Fisher time-series simulation

We also simulated a time-serial dataset via a Wright-Fisher evolutionary model. A population of 1000 sequences, each of genome length 500 bp was considered with 460 neutral loci under

zero selection, and $20/20$ randomly distributed selected loci with beneficial/deleterious mutations of selection strength ± 0.025 respectively. Mutations were simulated with a probability of 10^{-4} mutations per sequence per generation. The population was allowed to evolve for 400 generations and sampling was performed at every 50 generations, yielding a total of 9 observed generations.

4.2.5.4 NGS reads simulation

NGS reads were simulated via the ART simulator. Single-end reads of length 150 bp were generated using the error profile of the Illumina HiSeq 2500 platform. For datasets 1 to 4, the reads-per-sequence parameter was set to 5 to generate 5000 reads with a coverage of 1500X. For datasets 5 and 6, the reads-per-sequence parameter was set to 8 and 10 respectively, and 500 reads were randomly removed from dataset 5 via custom code to achieve the same coverage of 1500X as datasets 1 to 4. For the time-series dataset, the reads-per-sequence parameter was set to 15 to obtain 15,000 reads with a coverage of 4500X. We also simulated paired-end NGS reads of 150 bp for dataset 1 by setting the mean fragment length to 200 ± 50. The reads-per-sequence parameter to 5 to obtain 5000 reads and custom code was used to select 2500 read pairs from the generated reads to get the same coverage of 1500X as the single-end reads dataset.

4.2.5.5 Sequencing parameter sensitivity analysis

To investigate the sensitivity of LDEN to sequencing parameters, we generated additional datasets based on dataset 1 with varying read length and coverage values. Specifically, two datasets were generated with read lengths 50 bp and 100 bp, while keeping the coverage fixed at 1500X. Three datasets were generated with coverage values 50X, 500X, and 4500 X, keeping the read length fixed at 150 bp.

4.2.5.6 Statistical confidence

For each dataset and experimental condition, 100 independent Monte Carlo runs were simulated. This replication was done to establish statistical confidence in the results.

4.2.6 SARS-CoV-2 data

To perform biological concordance analysis of LDEN on real patient data, we performed LD estimation on data of a patient with long-term SARS-CoV-2 infection [52]. The dataset was obtained from a patient who was immunocompromised due to HIV infection. Due to the patient's degraded immune response, the SARS-CoV-2 virus was able survive and evolve *in-vivo* for over 2 years. This dataset was previously analyzed using MPL-R, which revealed considerable LD within the viral population, therefore we used this dataset to compare the performance of LDEN against MPL-R as a benchmark.

The dataset comprises 6 longitudinal samples obtained from the patient between 312 and 776 days following the estimated date of initial SARS-CoV-2 infection. Sequencing was performed via the Illumina NextSeq 500 platform to obtain 70 bp paired-end reads. The mean number of reads per sample was $2{,}786{,}519 \pm 1{,}008{,}062$, corresponding to an average coverage of $4{,}658\text{X} \pm 1{,}242\text{X}$. For consistency with the previous analysis done via MPL-R, the RBD of the Spike glycoprotein was analyzed, the genomics coordinates of which were obtained from Barnes *et al.* [124]. The reads were aligned against the SARS-CoV-2 reference sequence (GenBank accession no. $\text{MN}\,908947.3$) via BWA-MEM and sorted via samtools. LD estimation was performed across the genomic region spanning loci 22,544 to 23,161, which corresponds to the 618 bp length of the RBD. A bi-allelic model of WT/mutant was adopted and selection coefficients were estimated via MPL using the parameters described in Abdullah *et al.* [186].

4.3 Results

4.3.1 Overview

We evaluated the ability of LDEN in estimating LD from both simulated and real viral sequencing data. Simulated population data was used to evaluate whether sequencing errors and non-uniform coverage affected the estimation of intra-read LD from the reads. We tested the convergence of the algorithm and used the obtained parameters to run subsequent tests. Change in performance of the LDEN was measured as population parameters were varied. Genetic diversity, controlled by varying the number of founder sequences, and genome length were varied and LDEN was used to estimate LD. Tests were carried out on simulated reads to check for variation in performance as the sequencing quality degraded, specifically, by varying read length and sequence coverage. We then compared the performance of LDEN against population reconstruction-based inference on time-series evolutionary data. Finally, we demonstrated the utility of LDEN on a real dataset of SARS-CoV-2 evolution from a long-term immunocompromised patient.

4.3.2 The error correction scheme improves the accuracy of the intra-read LD values

We tested whether our proposed scheme correctly suppresses sequencing errors and mitigates the effects of non-uniform coverage when calculating intra-read LD from NGS reads. Figure 4.1 presents a boxplot of the Pearson correlation coefficient between the intra-read LD calculated from the reads using our proposed scheme, and the actual LD calculated for the same locus pairs from the full-length reference sequences. We first calculated the LD from the reads without either of the 1% frequency threshold and the shrinkage factor. These results are labeled "Unfiltered". We then applied the 1% frequency threshold and observed a significant improvement in accuracy. Then the shrinkage factor λ was introduced and varied from 1 to 15. It can be observed that the proposed scheme with the frequency threshold and shrinkage helps

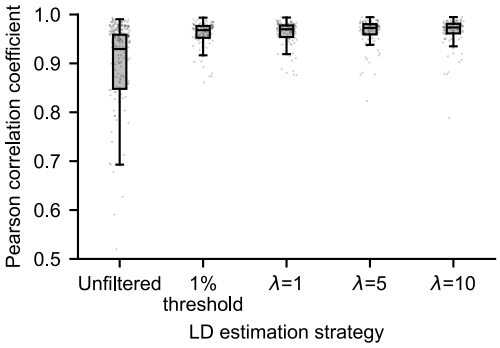


Figure 4.1 Proposed error correction scheme yields accurate intra-read LD. Boxplots of the Pearson correlation coefficient between the intra-read LD values obtained directly from the NGS reads and the LD values obtained from the full-length sequences demonstrate that LD values estimated using the proposed correction scheme are more accurate compared to those obtained without the correction. "Unfiltered" refers to no correction. "1% threshold" refers to filtering out loci with less than 1% variation in the single mutant allele frequencies. The rest of the plots show the results with the variation filtering as well as linear shrinkage. The results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of a population of 1000 sequences derived from 4 variants. One founder of length 500 bp sequence was generated randomly while 3 variants were generated from the founder sequence by randomly mutating 1%, 2%, and 5% of the loci, respectively. Prevalence values of each variant follow dataset 1 in Table 4.1. Analysis was performed on single-end NGS reads of 150 bp length with a coverage of 1500X.

improve the accuracy of the LD calculated from the reads, with consistently high correlation with the reference across all Monte Carlo runs. It can also be observed that the exact value of λ is not important and any value about 5 gives similar results. Based on these results, we used $\lambda = 5$ in all our tests.

4.3.3 LDEN demonstrates fast convergence

We set the convergence threshold $\varepsilon = 10^{-8}$ and chose the maximum number of iterations M from $\{100,1000,10000,100000\}$. For $M = \{100,1000\}$, the convergence threshold was never reached and iteration always stopped when the number of iterations reached M, whereas for $M = \{10000,100000\}$, some Monte Carlo runs reached the convergence threshold (Fig. 4.2A). The Frobenius distance between the last and the previous iterate decreased rapidly with an increase in M (Fig. 4.2B), however in terms of correlation with the ground truth LD matrix, the performance saturated for $M = 1000$. The wall clock time increased as M was increased, however it never exceeded a few tens of seconds even for $M = 100000$. Overall, it can be

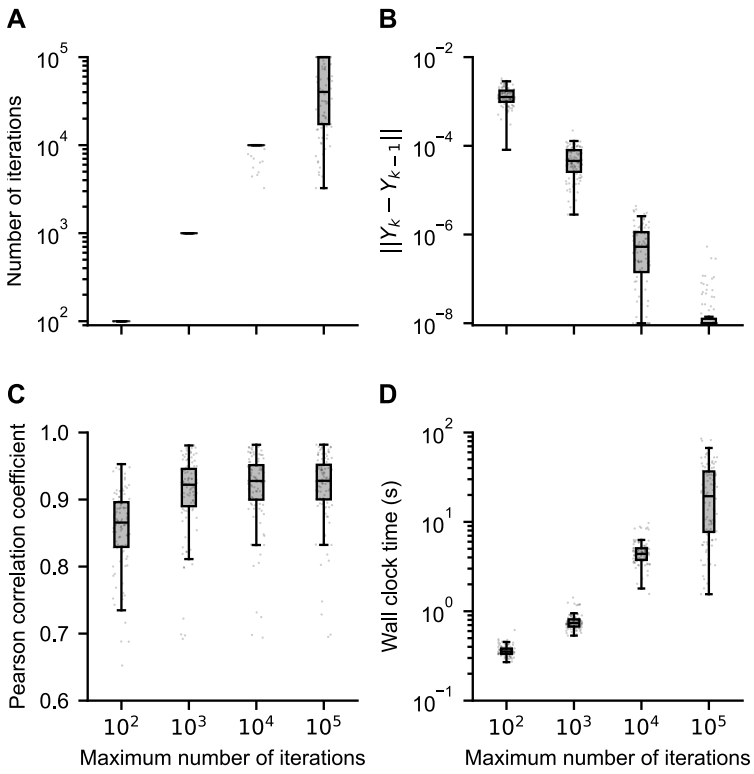


Figure 4.2 LDEN demonstrates fast convergence. **(A)** The maximum number of iterations is reached before the Frobenius distance between the last two iterates reaches the threshold of 10^{-8} for low values of the maximum number of iterations. **(B)** The Frobenius distance between the last two iterates decreases as the maximum number of iterations is increased. **(C)** The Pearson correlation coefficient between the estimated LD matrices and the reference LD matrices do not improve much with the increase in the number of iterations after reaching a peak value. **(D)** The wall clock time increases as the maximum number of iterations is increased but does not exceed a few minutes. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of a population of 1000 sequences derived from 4 variants. One founder of length 500 bp sequence was generated randomly while 3 variants were generated from the founder sequence by randomly mutating 1%, 2%, and 5% of the loci, respectively. Prevalence values of each variant follow dataset 1 in Table 4.1. Analysis was performed on single-end NGS reads of 150 bp length with a coverage of 1500X.

observed that LDEN rapidly converges to LD matrices that are strongly correlated with the reference LD matrix, and increasing the number of iterations does not improve the results much. Based on these results, we set $M = 1000$ in all subsequent tests.

4.3.4 LDEN recovers valid LD matrices from NGS data

The banded LD matrices directly obtained from NGS reads do not true covariance matrices as these do not satisfy the positive semidefinite constraint required for such matrices. Furthermore, these matrices may also be mathematically inconsistent because replacing

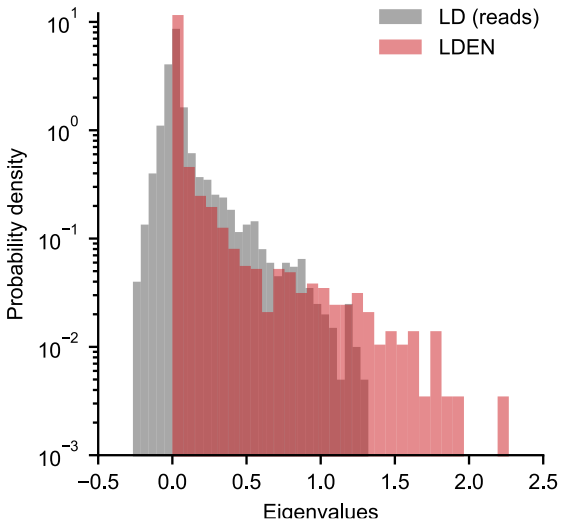


Figure 4.3 LD matrices estimated by LDEN are positive semidefinite. Histograms showing the eigenvalues of the banded LD matrix obtained directly from the NGS reads (gray) and the LD matrix estimated by LDEN (red). The banded LD matrix contains negative eigenvalues and is not positive semidefinite, whereas the LD matrix estimated by LDEN has all nonnegative eigenvalues. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of a population of 1000 sequences derived from 4 variants. One founder of length 500 bp sequence was generated randomly while 3 variants were generated from the founder sequence by randomly mutating 1%, 2%, and 5% of the loci, respectively. Prevalence values of each variant follow dataset 1 in Table 4.1. Analysis was performed on single-end NGS reads of 150 bp length with a coverage of 1500X.

unobserved LD values with zeros can violate the theoretical bounds dictated by probability theory. A comparison between the eigenvalues of the banded LD matrices obtained directly from the reads and those of the LD matrices reconstructed via LDEN is shown in Fig. 4.3. It can be observed that the banded LD matrices exhibit negative eigenvalues, which are impossible for a covariance matrix. On average, around $35.83 \pm 4.01\%$ of eigenvalues were negative across all Monte Carlo runs. In contrast, all eigenvalues of the LD matrices recovered via LDEN are nonnegative, satisfying the positive semidefinite property. Thus, LDEN successfully transforms the banded, potentially inconsistent LD matrices observed from NGS reads into valid covariance matrices suitable for downstream evolutionary analysis.

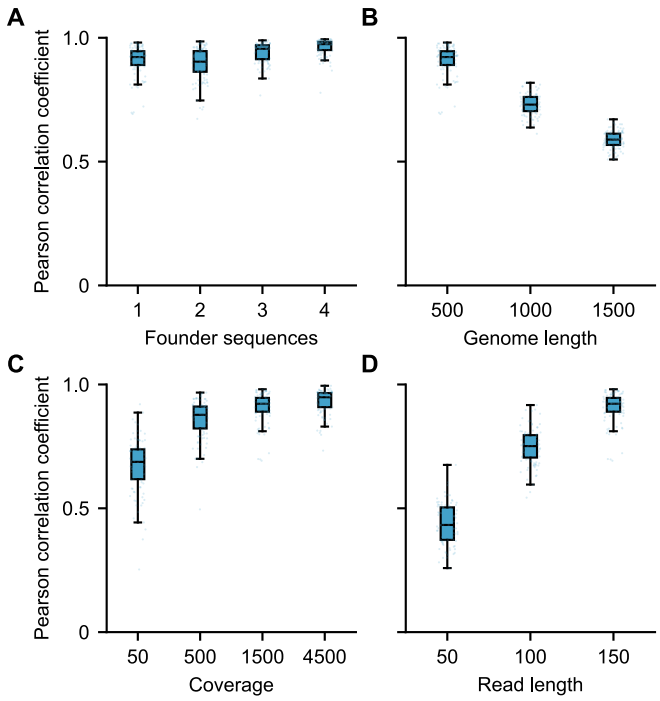


Figure 4.4 The performance of LDEN is correlated with the accuracy of the information captured by the reads. **(A)** Performance remains similar for different levels of variation in the data. **(B)** Performance degrades as the genome length is increased for the same read length. **(C)** Increase in coverage improves performance. **(D)** Increasing read length improves performance. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of a population of 1000 sequences derived from 4 variants. One founder sequence was generated randomly while 3 variants were generated from the founder sequence by randomly mutating 1%, 2%, and 5% of the loci, respectively. Prevalence values of each variant are listed in Table 4.1. Analyses on coverage and read length were performed on dataset 1. Analyses on founder sequences and genome length were performed on simulated single-end NGS reads of 150 bp length with a coverage of 1500X.

4.3.5 The performance of LDEN is correlated with the accuracy of the banded LD information

LD depends on the level of genetic variation in the population, which in turn is influenced by the number of infecting strains, the rates of mutation, recombination, and reassortment of the virus, as well as the time elapsed since infection. LD can also exist over large distances due to three-dimensional interactions between different nucleotides in the viral genome. In addition to the parameters of the population, LD estimation depends on how well the sequencing reads sample the population. Reads may have different levels of coverage which can introduce various degrees of statistical noise into the data. Read lengths also vary across sequencing

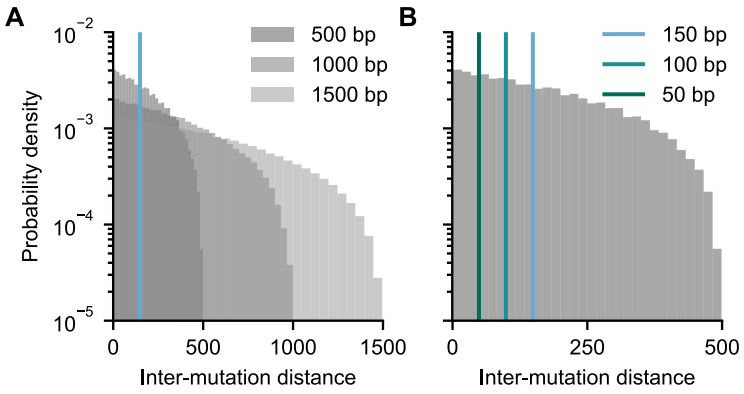


Figure 4.5 The information observable via reads depends on the read length and the genome length. **(A)** For a fixed read length of 150 bp (vertical blue line), increasing the genome length increases the number of unobservable LD values, as inter-mutation distances increases grow larger relative to read length. LD values corresponding to loci pairs to the left of the vertical line are observable via individual reads, whereas LD values to the right are unobservable. **(B)** When read length is increased while the genome length is held constant, more LD values become observable. The vertical lines represent the read length. LD values to the left of each line are observable via individual reads, whereas LD values to the right are unobservable. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of a population of 1000 sequences derived from 4 variants. One founder sequence was generated randomly while 3 variants were generated from the founder sequence by randomly mutating 1%, 2%, and 5% of the loci, respectively. Prevalence values of each variant are listed in Table 4.1. Genome lengths of the sequences were $\{500, 1000, 1500\}$. Analyses were performed on simulated single-end NGS reads of 150 bp length with a coverage of 1500X.

platforms and determine the extent of LD information individual reads can capture. We evaluated the robustness of LDEN to varying population and read parameters and established benchmarks for the performance of the method. LDEN was observed to be robust to the level of variation in the data (Fig. 4.4A), however its performance degraded as the genome length was increased while keeping the read length constant (Fig. 4.4B). The performance of LDEN improved as coverage (Fig. 4.4C) or read length (Fig. 4.4D) was increased. This trend in performance can be understood by the fact that higher coverage reduces statistical noise in the observable LD whereas the longer reads enable the observation of more LD information. To explain how longer genome length and shorter read length render more LD information unobservable, we refer to Fig. 4.5. When the read length is fixed and the genome length is increased, the number of unobservable LD values increases as indicated by the bulk of the histograms appearing towards the right of the vertical line (Fig. 4.5A). Conversely, when the

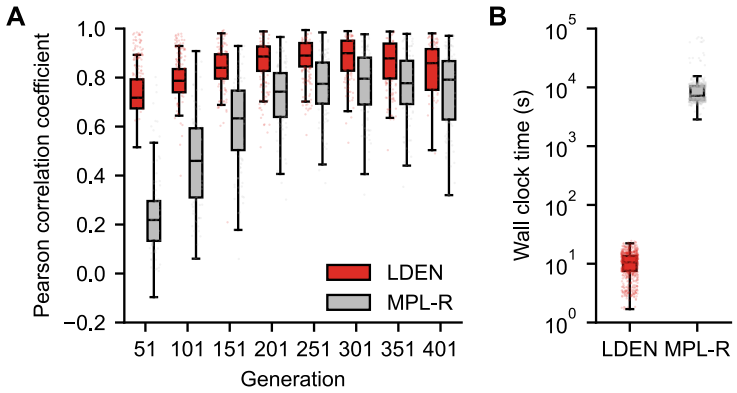


Figure 4.6 LDEN outperforms MPL-R in both estimation accuracy and computational speed. **(A)** LDEN achieves higher accuracy in LD estimation compared to MPL-R across time-series data. **(B)** LDEN completes the entire processing pipeline two orders of magnitude faster than MPL-R. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of a population of 1000 sequences evolved from a single randomly generated founder genome of length 500 bp. Mutation was simulated at a probability of 10^{-4} per base per generation. 460 loci were neutral and $20/20$ were under beneficial/deleterious selection with selection coefficients ± 0.025. Single-end NGS reads of 150 bp length with a coverage of 4500X were simulated for each time-point. MPL-R was run on 3 bootstrap samples each of coverage 1500X and the results were aggregated.

genome length is fixed and the read length is decreased, more of the distribution lies towards the right side of the respective vertical line, meaning a greater proportion of LD information becomes unobservable (Fig. 4.5B). Overall, LDEN performs well as long as the NGS reads capture sufficient information from the data. When the genome length is small relative to the read length, or when the read length or coverage is increased, the quality of the information observed by the reads improves, which in turn enhances the accuracy of the LD estimation. These results may be used as guidelines for designing future NGS experiments.

4.3.6 LDEN outperforms population-reconstruction based estimation

In our previous work on LD estimation, we demonstrated that population reconstruction struggles to accurately estimate LD when genetic diversity is low (Fig. 2.28 in Chapter 2). To compare the performance of MPL-R with LDEN, we applied both methods on time-series NGS data. LDEN consistently outperformed MPL-R in terms of the accuracy of the estimated LD (Fig. 4.6A). In addition to superior accuracy, LDEN completed the entire processing pipeline, from NGS reads to LD estimation, in a fraction of a second, which was 2 orders of magnitude

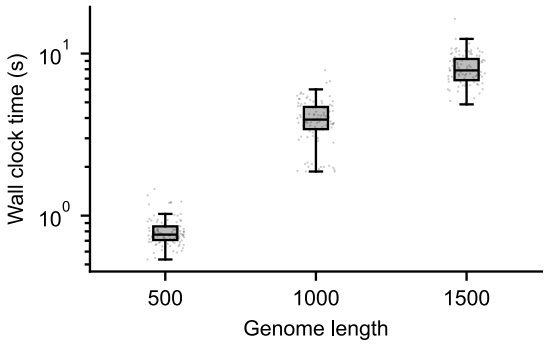


Figure 4.7 LDEN is scalable with increasing genome length. The wall clock time of LDEN increases with the genome length, but it does not increase significantly. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of a population of 1000 sequences derived from 4 variants. One founder sequence was generated randomly while 3 variants were generated from the founder sequence by randomly mutating 1%, 2%, and 5% of the loci, respectively. Prevalence values of each variant are listed in Table 4.1. Genome lengths of the sequences were {500, 1000, 1500}. Analyses were performed on simulated single-end NGS reads of 150 bp length with a coverage of 1500X.

faster compared to MPL-R (Fig. 4.6B). The improved accuracy and computational efficiency of LDEN can be explained by its fundamentally different approach to LD estimation as compared to MPL-R. LDEN bypasses population reconstruction entirely and is not affected by population reconstruction errors. LDEN is also faster because it estimates LD directly from the NGS data, whereas MPL-R first performs population reconstruction, which is a computationally intensive process, and subsequently estimates LD from the reconstructed population, which requires much longer wall clock time compared to LDEN.

4.3.7 LDEN is scalable with increasing genome length

We evaluated the scalability of LDEN by comparing the wall clock time for different genome lengths. LDEN required a median wall clock time of 1.15 seconds to process genomes of length 1500 bp, the largest genome length tested (Fig. 4.7). This was significantly faster than the wall clock time of MPL-R, which required more than 30 minutes to process genomes of only 500 bp. We did not test LDEN on longer genomes because the read length of 150 bp used in the tests is too short relative to larger genome lengths, which renders most of the LD information unobservable. Even for a genome length of 1500 bp, the median Pearson correlation

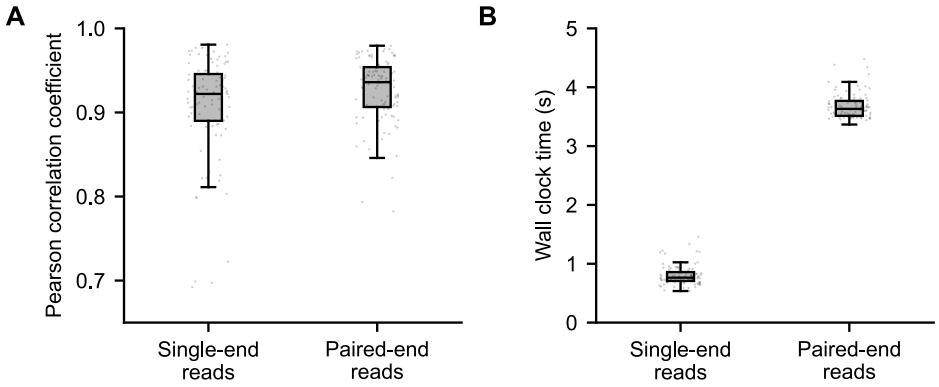


Figure 4.8 LDEN yields more accurate results using paired-end reads. **(A)** The accuracy of LD estimation is higher when using paired-end reads as compared to single-end reads. **(B)** The wall clock time is higher when using paired-end reads but the estimation still completes in less than 5 seconds. Results are shown for 100 Monte Carlo runs. Each Monte Carlo run consisted of a population of 1000 sequences derived from 4 variants. One founder sequence was generated randomly while 3 variants were generated from the founder sequence by randomly mutating 1%, 2%, and 5% of the loci, respectively. Prevalence values of each variant are listed in Table 4.1. Genome lengths of the sequences were $\{500, 1000, 1500\}$. Both single- and paired-end reads had a length of 150 bp length with a coverage of 1500X. The mean fragment length for paired-end reads was 200 ± 50.

coefficient of the recovered LD was only slightly above 0.5, indicating that the limiting factor in scalability is not the wall clock time of the method but rather the amount LD information that can be observed by individual reads.

4.3.8 Paired-end reads improve the accuracy of LDEN

We compared the performance of LDEN on single-end and paired-end reads to assess whether the incorporation of fragment-level linkage information improves the accuracy of LD estimation (Fig. 4.8). The median Pearson correlation coefficient between the LD matrix recovered from paired-end reads and the ground truth was higher than that between the LD matrix obtained via single-end reads and the ground truth (Fig. 4.8A). This improvement indicates the greater amount of directly observable LD information captured by paired-end reads, which extends the observable genomic distance beyond the length of individual reads.

In terms of computational cost, LD estimation using paired-end reads was relatively slower than the single-end case, which is expected given the two-step procedure required to process fragment-level information (Fig. 4.8B). Nevertheless, the median wall clock time remained

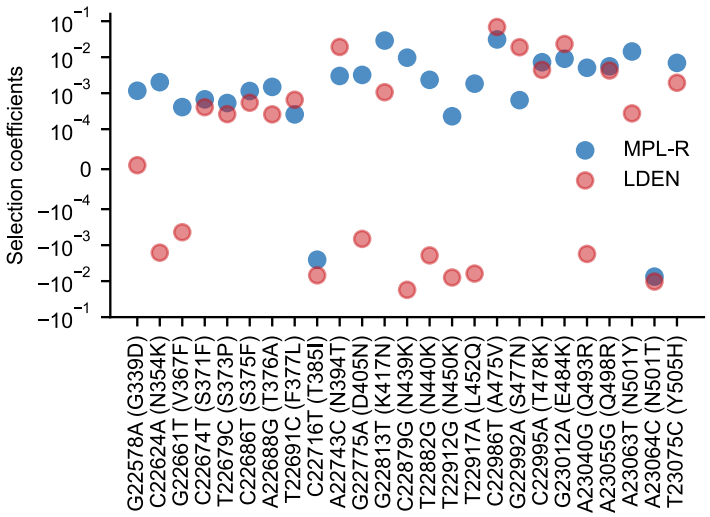


Figure 4.9 LDEN reports high magnitudes of selection coefficients for known beneficial mutations in the RBD of SARS-CoV-2, most of which are in agreement with the magnitudes reported by MPL-R.

modest at approximately 3.6 seconds per Monte Carlo run, which is negligible compared to performing haplotype reconstruction. These results demonstrate that paired-end reads offer improved accuracy of LD recovery at the cost of a minimal increase in the wall clock time.

4.3.9 LDEN reveals known biological insights from patient data

We analyzed time-series data of long-term SARS-C oV-2 infection via LDEN. The details of the preprocessing steps have been described in Abdullah *et al.* [186], and a brief summary follows. We used LoFreq [127] to obtain the single mutant allele frequencies and filtered mutations that exhibited at least 1% variation. A total of 26 nonsynonymous mutations were observed in the RBD, for which we estimated the selection coefficients via both MPL-R and LDEN. LDEN reported similar selection pressure as MPL-R on 14 of the 26 mutations (Fig. 4.9). These 14 mutations are known in literature to have phenotypic effects on the virus. A summary of the phenotypic effects of these mutations is presented in Table 2.3 in Chapter 2.

4.4 Discussion

In this work, we have presented LDEN, a novel method for direct estimation of LD from NGS data without relying on population reconstruction. We have demonstrated that LDEN performs

well under conditions of high read coverage and read length, which are the primary determinants of observable LD information. We have also demonstrated that LDEN substantially outperforms population reconstruction-based LD estimation both in terms of accuracy of the estimated LD values as well as the wall clock time, achieving a speedup of two orders of magnitude. Finally, we reanalyzed longitudinal SARS-CoV-2 data from an immunocompromised patient with LDEN and demonstrated that the inferred insights are qualitatively similar to those obtained via MPL-R, with 14 of the 26 nonsynonymous mutation and demonstrating concordant estimates that correspond to known beneficial mutations.

The performance of LDEN under high coverage and long read length can be understood by the underlying parameters that change with these. High coverage reduces the statistical noise in the observed LD values. This statistical noise scales as $\frac{1}{n_{ij}}$, thereby reducing the error in the observed LD values. Longer read lengths, conversely, increase the range of observable LD entries that satisfy $j - i < r$, which renders more of the LD matrix directly observable rather than in need of estimation.

The superior performance of LDEN compared to MPL-R is attributed to fundamental differences in problem formulation. MPL-R solves a harder problem than is necessary, and reconstructs complete haplotypes, and thereby higher-order statistics, and obtains LD from these. This approach is computationally expensive and requires subsampling and parallelization. Moreover, population reconstruction is sensitive to regions of low coverage and introduces errors that propagate to the LD estimates. LDEN, by contrast, directly computes the required two-point statistics and formulates the task as a convex matrix completion problem. This change of formulation is advantageous because it is computationally tractable, it imposes only those constraints that are mathematically required for the task, and it is guaranteed to converge to the unique global minimum because of the intersection of two

convex sets. This fundamentally different and more efficient approach leads to the observed two orders of magnitude speedup.

The overlap in the results of LDEN and MPL-R in identifying 14 known beneficial mutations, including well-characterized immune escape variants, in the SARS-CoV-2 RBD provides biological validation of the method. The fact that LDEN estimates similar selection coefficients as MPL-R for these mutations indicates that the speedup offered by LDEN comes at no cost of accuracy in identifying strongly selection mutations.

While LDEN addresses some limitations of existing approaches, some challenges remain. APM, which forms the base for LDEN, is known to become slow to converge as the scale increases. This is because optimizing for one constraint may violate the other and vice versa, leading to a "zig-zagging" behavior which takes longer to stabilize for larger systems. The convergence can be accelerated via the incorporation of Dykstra's correction [197], which maintains a memory term that stores the difference between the input matrix and its projection onto the positive semidefinite constraint. At each iteration, this memory term is subtracted from the current matrix before the next projection, effectively correcting for overshoot and helping reach convergence faster. While the current implementation of LDEN works well for systems having up to 1500 bp genome length, future work can focus on implementing Dykstra's correction to attain even greater speedup, particularly for large genomes for which the current implementation may become slow.

Chapter 5

Conclusion and future work

5.1 Conclusion

Viral pathogens pose a persistent challenge to human health and healthcare systems. These pathogens typically exhibit high mutation rates and undergo continuous evolution, leading to the emergence of more virulent and infectious variants that can reduce the effectiveness of existing therapies. Accurate characterization of viral pathogens therefore requires a deep understanding of genomic evolution, which is fundamentally confounded by the persistent presence of genetic linkage. Genetic linkage can mask immune escape mutations and boost the prevalence of neutral mutations, making the identification of effective antiviral targets particularly challenging. Existing sequencing technologies provide only partial information about genetic linkage due to their inherently short read-lengths, leaving much of the LD information across the genome unobservable. Consequently, there is a pressing need for tools capable of resolving unobservable linkage patterns from viral sequencing data to enable accurate characterization of viral mutations.

This work presents a suite of tools for estimation of LD from HTS data and subsequent evolutionary inference. We have utilized ideas from probability theory, finance, convex optimization, and computer science to present efficient, user-friendly software tools for the analysis of HTS data. The presented works can estimate LD from HTS data accurately and efficiently, a capability currently not offered by existing methods. All tools have been rigorously tested on both simulated and real datasets. Importantly, our tools do not require the design of new experiments and can run on already existing sequencing data.

In **Chapter 2**, we presented MPL-R, a tool that can recover unobservable linkage information from NGS data and perform linkage-aware fitness inference. MPL-R accurately recovers LD

across the entire genome and does so in a reasonable amount of time. We tested MPL-R on several simulated datasets and observed it to be robust to a range of population parameters. MPL-R outperformed multiple linkage-independent methods and one linkage-aware method in our comparative tests. We also demonstrated the performance of MPL-R on real patient HIV and SARS-CoV-2 data. The software was successfully able to identify known strongly selected mutations in CD8+ T-cell epitopes in HIV and RBD in SARS-CoV-2.

In **Chapter 3**, we introduced SLAFIT, an integrated pipeline for the detection of structural variants and estimation of LD and fitness from TGS data. We introduced a parallelized implementation of a linkage-aware fitness estimation algorithm that significantly improved the wall clock time. We tested the tool for fitness inference over various levels of sequencing noise and observed it to be robust. The pipeline was also able to infer biologically consistent insights from real patient data, demonstrating its practical utility.

In **Chapter 4**, we presented LDEN, a novel method of estimating LD from NGS data based on techniques from finance and convex optimization. LDEN is robust to genomic variation in the population and provides accurate estimates when the sampling of the population via reads was sufficient. Compared to MPL-R, LDEN yields higher accuracy in LD estimation from time-series data and performs estimation two orders of magnitude faster. The insights inferred by LDEN on real patient data were qualitatively similar to those obtained via MPL-R, validating its biological relevance.

5.2 Future work

In this work, we present three tools for LD estimation from HTS data. Although the presented works have demonstrated significant improvement in the fine-scale analysis of HTS data, several challenges need to be addressed. We present some directions for future research that can build upon the frameworks developed in these tools to address more complicated evolutionary and computational challenges.

5.2.1 Incorporation of epistasis

The methods developed in this work assume an additive model and focus on pairwise LD as a measure of non-random association between mutations. In reality, fitness may not always be additive, and epistatic interactions determine the overall strain fitness. Epistasis plays an important role in viral evolution and can render mutations, that are not inherently fit, to become highly fit when occurring together.

Future work could extend the LD estimation framework to explicitly model epistasis and estimate third and fourth order LD. This would require posing the LD matrix completion problem as a higher-order matrix completion problem. This problem will present its own challenges, such as the exponentially increased memory requirement and the $O(N^4)$ computations.

5.2.2 Direct LD estimation from TGS data

LDEN has been developed for NGS data, but the availability of TGS platforms presents an opportunity for developing a similar method for this data. TGS generates much longer reads that can capture LD across large portions of the genome, but the higher error rates and platform-specific error profiles require specialized treatment. Moreover, TGS data typically doesn't have as deep coverage as NGS, so statistical noise needs to be considered as well.

Future research could build upon LDEN and handle TGS data. An obvious solution could be to preprocess TGS reads prior to LD estimation, but another interesting direction of research is to incorporate error correction into the LD estimation pipeline itself, for example, by using basecalling quality scores in the shrinkage estimator.

REFERENCES

[1] UNAIDS, "Global HIV & AIDS statistics — Fact sheet." [Online]. Available: https://www.unaids.org/en/resources/fact-sheet

[2] N. Moghimi *et al.*, "The Neurological Manifestations of Post-Acute Sequelae of SARS-CoV-2 Infection," *Curr. Neurol. Neurosci. Rep.*, vol. 21, no. 9, p. 44, Sep. 2021, doi: 10.1007/s11910-021-01130-1.

[3] R. P. Payne *et al.*, "Differential Escape Patterns within the Dominant HLA-B*57:03-Restricted HIV Gag Epitope Reflect Distinct Clade-Specific Functional Constraints," *J. Virol.*, vol. 88, no. 9, pp. 4668–4678, May 2014, doi: 10.1128/JVI.03303-13.

[4] M. Yuan *et al.*, "Structural and functional ramifications of antigenic drift in recent SARS-CoV-2 variants," *Science*, vol. 373, no. 6556, pp. 818–823, Aug. 2021, doi: 10.1126/science.abh1139.

[5] Nextstrain, *Genomic epidemiology of HIV Env*. [Online]. Available: https://nextstrain.org/groups/LANL-HIV-DB/HIV/env?l=unrooted

[6] Los Alamos National Laboratory, "HIV Sequence Database." [Online]. Available: https://hiv.lanl.gov

[7] R. J. Garten *et al.*, "Antigenic and Genetic Characteristics of Swine-Origin 2009 A(H1N1) Influenza Viruses Circulating in Humans," *Science*, vol. 325, no. 5937, pp. 197–201, Jul. 2009, doi: 10.1126/science.1176225.

[8] "Network for Surveillance-South Africa (NGS-SA). SARS-CoV-2 Sequencing Update 15 July 2022 (NGS-SA, 2022)." [Online]. Available: https://www.nicd.ac.za/wp-content/uploads/2022/07/Update-of-SA-sequencing-data-from-GISAID-15-July-2022.pdf

[9] D. M. Koelle *et al.*, "Worldwide circulation of HSV-2 × HSV-1 recombinant strains," *Sci. Rep.*, vol. 7, no. 1, p. 44084, Mar. 2017, doi: 10.1038/srep44084.

[10] O. Fakhri *et al.*, "Superinfection and recombination of infectious laryngotracheitis virus vaccines in the natural host," *Vaccine*, vol. 38, no. 47, pp. 7508–7516, Nov. 2020, doi: 10.1016/j.vaccine.2020.09.064.

[11] Nextstrain, *Genomic epidemiology of SARS-CoV-2 with subsampling focused globally over the past 6 months*. [Online]. Available: https://nextstrain.org/ncov/gisaid/global/6m

[12] Global Initiative on Sharing All Influenza Data, "SARS-CoV-2 database." [Online]. Available: https://www.gisaid.org/

[13] M. S. Sohail, R. H. Y. Louie, M. R. McKay, and J. P. Barton, "MPL resolves genetic linkage in fitness inference from complex evolutionary histories," *Nat. Biotechnol.*, vol. 39, no. 4, pp. 472–479, Apr. 2021, doi: 10.1038/s41587-020-0737-3.

[14] N. Strelkowa and M. Lässig, "Clonal Interference in the Evolution of Influenza," *Genetics*, vol. 192, no. 2, pp. 671–682, Oct. 2012, doi: 10.1534/genetics.112.143396.

[15] M. Łuksza and M. Lässig, "A predictive fitness model for influenza," *Nature*, vol. 507, no. 7490, pp. 57–61, Mar. 2014, doi: 10.1038/nature13087.

[16] M. T. Alam *et al.*, "Selective Sweeps and Genetic Lineages of Plasmodium falciparum Drug -Resistant Alleles in Ghana," *J. Infect. Dis.*, vol. 203, no. 2, pp. 220–227, Jan. 2011, doi: 10.1093/infdis/jiq038.

[17] A. Sanchez-Mazas *et al.*, "The HLA -B landscape of Africa: Signatures of pathogen-driven selection and molecular identification of candidate alleles to malaria protection," *Mol. Ecol.*, vol. 26, no. 22, pp. 6238–6252, Nov. 2017, doi: 10.1111/mec.14366.

[18] K. Lee *et al.*, “Benign polymorphisms in the *BRCA* genes with linkage disequilibrium is associated with cancer characteristics,” *Cancer Sci.*, vol. 115, no. 12, pp. 3973–3985, Dec. 2024, doi: 10.1111/cas.16364.
[19] R. C. Lewontin, “THE INTERACTION OF SELECTION AND LINKAGE. I. GENERAL CONSIDERATIONS; HETEROTIC MODELS,” *Genetics*, vol. 49, no. 1, pp. 49–67, Jan. 1964, doi: 10.1093/genetics/49.1.49.
[20] Y. Li and J. P. Barton, “Estimating linkage disequilibrium and selection from allele frequency trajectories,” *Genetics*, vol. 223, no. 3, p. iyac189, Mar. 2023, doi: 10.1093/genetics/iyac189.
[21] B. Lee *et al.*, “Inferring effects of mutations on SARS-CoV-2 transmission from genomic surveillance data,” *Nat. Commun.*, vol. 16, no. 1, p. 441, Jan. 2025, doi: 10.1038/s41467-024-55593-0.
[22] E. Santiago, I. Novo, A. F. Pardiñas, M. Saura, J. Wang, and A. Caballero, “Recent Demographic History Inferred by High-Resolution Analysis of Linkage Disequilibrium,” *Mol. Biol. Evol.*, vol. 37, no. 12, pp. 3642–3653, Dec. 2020, doi: 10.1093/molbev/msaa169.
[23] I. Novo, E. Santiago, and A. Caballero, “The estimates of effective population size based on linkage disequilibrium are virtually unaffected by natural selection,” *PLOS Genet.*, vol. 18, no. 1, p. e1009764, Jan. 2022, doi: 10.1371/journal.pgen.1009764.
[24] C. Veller and G. M. Coop, “Interpreting population- and family-based genome-wide association studies in the presence of confounding,” *PLOS Biol.*, vol. 22, no. 4, p. e3002511, Apr. 2024, doi: 10.1371/journal.pbio.3002511.
[25] D. Zhu *et al.*, “Genomic prediction based on selective linkage disequilibrium pruning of low-coverage whole-genome sequence variants in a pure Duroc population,” *Genet. Sel. Evol.*, vol. 55, no. 1, p. 72, Oct. 2023, doi: 10.1186/s12711-023-00843-w.
[26] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors,” *Proc. Natl. Acad. Sci.*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977, doi: 10.1073/pnas.74.12.5463.
[27] F. Sanger *et al.*, “Nucleotide sequence of bacteriophage φX174 DNA,” *Nature*, vol. 265, no. 5596, pp. 687–695, Feb. 1977, doi: 10.1038/265687a0.
[28] A. Goffeau *et al.*, “Life with 6000 Genes,” *Science*, vol. 274, no. 5287, pp. 546–567, Oct. 1996, doi: 10.1126/science.274.5287.546.
[29] M. D. Adams *et al.*, “The Genome Sequence of *Drosophila melanogaster*,” *Science*, vol. 287, no. 5461, pp. 2185–2195, Mar. 2000, doi: 10.1126/science.287.5461.2185.
[30] International Human Genome Sequencing Consortium *et al.*, “Initial sequencing and analysis of the human genome,” *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001, doi: 10.1038/35057062.
[31] C. Cheng, Z. Fei, and P. Xiao, “Methods to improve the accuracy of next-generation sequencing,” *Front. Bioeng. Biotechnol.*, vol. 11, p. 982111, Jan. 2023, doi: 10.3389/fbioe.2023.982111.
[32] L. K. Vestergaard, D. N. P. Oliveira, C. K. Høgdall, and E. V. Høgdall, “Next Generation Sequencing Technology in the Clinic and Its Challenges,” *Cancers*, vol. 13, no. 8, p. 1751, Apr. 2021, doi: 10.3390/cancers13081751.
[33] S. Goodwin, J. D. McPherson, and W. R. McCombie, “Coming of age: ten years of next-generation sequencing technologies,” *Nat. Rev. Genet.*, vol. 17, no. 6, pp. 333–351, Jun. 2016, doi: 10.1038/nrg.2016.49.
[34] M. Margulies *et al.*, “Genome sequencing in microfabricated high-density picolitre reactors,” *Nature*, vol. 437, no. 7057, pp. 376–380, Sep. 2005, doi: 10.1038/nature03959.

[35] J. M. Rothberg *et al.*, “An integrated semiconductor device enabling non-optical genome sequencing,” *Nature*, vol. 475, no. 7356, pp. 348–352, Jul. 2011, doi: 10.1038/nature10242.
[36] D. R. Bentley *et al.*, “Accurate whole human genome sequencing using reversible terminator chemistry,” *Nature*, vol. 456, no. 7218, pp. 53–59, Nov. 2008, doi: 10.1038/nature07517.
[37] T. Hu, N. Chitnis, D. Monos, and A. Dinh, “Next-generation sequencing technologies: An overview,” *Hum. Immunol.*, vol. 82, no. 11, pp. 801–811, Nov. 2021, doi: 10.1016/j.humimm.2021.02.012.
[38] G. A. Logsdon, M. R. Vollger, and E. E. Eichler, “Long-read human genome sequencing and its applications,” *Nat. Rev. Genet.*, vol. 21, no. 10, pp. 597–614, Oct. 2020, doi: 10.1038/s41576-020-0236-x.
[39] F. J. Rang, W. P. Kloosterman, and J. De Ridder, “From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy,” *Genome Biol.*, vol. 19, no. 1, p. 90, Dec. 2018, doi: 10.1186/s13059-018-1462-9.
[40] J. Eid *et al.*, “Real-Time DNA Sequencing from Single Polymerase Molecules,” *Science*, vol. 323, no. 5910, pp. 133–138, Jan. 2009, doi: 10.1126/science.1162986.
[41] D. Deamer, M. Akeson, and D. Branton, “Three decades of nanopore sequencing,” *Nat. Biotechnol.*, vol. 34, no. 5, pp. 518–524, May 2016, doi: 10.1038/nbt.3423.
[42] J.-Y. Lee, “The Principles and Applications of High-Throughput Sequencing Technologies,” *Dev. Reprod.*, vol. 27, no. 1, pp. 9–24, Apr. 2023, doi: 10.12717/DR.2023.27.1.9.
[43] B. Wang *et al.*, “Long and Accurate: How HiFi Sequencing is Transforming Genomics,” *Genomics Proteomics Bioinformatics*, vol. 23, no. 1, p. qzaf003, May 2025, doi: 10.1093/gpbjnl/qzaf003.
[44] “What are the advantages of HiFi sequencing?,” PacBio. [Online]. Available: https://www.pacb.com/technology/hifi-sequencing/how-it-works/
[45] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley, “Continuous base identification for single-molecule nanopore DNA sequencing,” *Nat. Nanotechnol.*, vol. 4, no. 4, pp. 265–270, Apr. 2009, doi: 10.1038/nnano.2009.12.
[46] D. Lu *et al.*, “Nanopore ultra-long sequencing and adaptive sampling spur plant complete telomere-to-telomere genome assembly,” *Mol. Plant*, vol. 17, no. 11, pp. 1773–1786, Nov. 2024, doi: 10.1016/j.molp.2024.10.008.
[47] A. L. Greninger *et al.*, “Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis,” *Genome Med.*, vol. 7, no. 1, p. 99, Dec. 2015, doi: 10.1186/s13073-015-0220-9.
[48] M. Loose, S. Malla, and M. Stout, “Real-time selective sequencing using nanopore technology,” *Nat. Methods*, vol. 13, no. 9, pp. 751–754, Sep. 2016, doi: 10.1038/nmeth.3930.
[49] T. Hoenen *et al.*, “Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool,” *Emerg. Infect. Dis.*, vol. 22, no. 2, Feb. 2016, doi: 10.3201/eid2202.151796.
[50] J. Quick *et al.*, “Real-time, portable genome sequencing for Ebola surveillance,” *Nature*, vol. 530, no. 7589, pp. 228–232, Feb. 2016, doi: 10.1038/nature16996.
[51] L. E. Kafetzopoulou *et al.*, “Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak,” *Science*, vol. 363, no. 6422, pp. 74–77, Jan. 2019, doi: 10.1126/science.aau9343.
[52] J. M. Velasquez-Reyes *et al.*, “Characterisation of a persistent SARS-CoV-2 infection lasting more than 750 days in a person living with HIV: a genomic analysis,” *Lancet Microbe*, vol. 6, no. 9, p. 101122, Sep. 2025, doi: 10.1016/j.lanmic.2025.101122.

[53] C. Vellas *et al.*, “Comparison of short-read and long-read next-generation sequencing technologies for determining HIV-1 drug resistance,” *J. Med. Virol.*, vol. 96, no. 10, p. e29951, Oct. 2024, doi: 10.1002/jmv.29951.
[54] J. Cane *et al.*, “Nanopore sequencing of influenza A and B in Oxfordshire and the United Kingdom, 2022–23,” *J. Infect.*, vol. 88, no. 6, p. 106164, Jun. 2024, doi: 10.1016/j.jinf.2024.106164.
[55] H. R. Lapointe *et al.*, “Validation of a Genotype-Independent Hepatitis C Virus Near-Whole Genome Sequencing Assay,” *Viruses*, vol. 13, no. 9, p. 1721, Aug. 2021, doi: 10.3390/v13091721.
[56] S. F. Lumley *et al.*, “Whole genome sequencing of hepatitis B virus using tiled amplicon (HEPTILE) and probe based enrichment on Illumina and Nanopore platforms,” *Sci. Rep.*, vol. 15, no. 1, p. 5795, Feb. 2025, doi: 10.1038/s41598-025-87721-1.
[57] H. Cassidy *et al.*, “Exploring a prolonged enterovirus C104 infection in a severely ill patient using nanopore sequencing,” *Virus Evol.*, vol. 8, no. 1, p. veab109, Mar. 2022, doi: 10.1093/ve/veab109.
[58] Z. He, X. Dai, M. Beaumont, and F. Yu, “Detecting and Quantifying Natural Selection at Two Linked Loci from Time Series Data of Allele Frequencies with Forward-in-Time Simulations,” *Genetics*, vol. 216, no. 2, pp. 521–541, Oct. 2020, doi: 10.1534/genetics.120.303463.
[59] V. Buffalo and G. Coop, “The Linked Selection Signature of Rapid Adaptation in Temporal Genomic Data,” *Genetics*, vol. 213, no. 3, pp. 1007–1045, Nov. 2019, doi: 10.1534/genetics.119.302581.
[60] J. Terhorst, C. Schlötterer, and Y. S. Song, “Multi-locus Analysis of Genomic Time Series Data from Experimental Evolution,” *PLOS Genet.*, vol. 11, no. 4, p. e1005069, Apr. 2015, doi: 10.1371/journal.pgen.1005069.
[61] C. J. R. Illingworth, “Fitness Inference from Short-Read Data: Within-Host Evolution of a Reassortant H5N1 Influenza Virus,” *Mol. Biol. Evol.*, vol. 32, no. 11, pp. 3012–3026, Nov. 2015, doi: 10.1093/molbev/msv171.
[62] M. W. Shen, K. T. Zhao, and D. R. Liu, “Reconstruction of evolving gene variants and fitness from short sequencing reads,” *Nat. Chem. Biol.*, vol. 17, no. 11, pp. 1188–1198, Nov. 2021, doi: 10.1038/s41589-021-00876-6.
[63] M. Hegreness, N. Shoresh, D. Hartl, and R. Kishony, “An Equivalence Principle for the Incorporation of Favorable Mutations in Asexual Populations,” *Science*, vol. 311, no. 5767, pp. 1615–1617, Mar. 2006, doi: 10.1126/science.1122469.
[64] G. I. Lang *et al.*, “Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations,” *Nature*, vol. 500, no. 7464, pp. 571–574, Aug. 2013, doi: 10.1038/nature12344.
[65] J. M. Smith and J. Haigh, “The hitch-hiking effect of a favourable gene,” *Genet. Res.*, vol. 23, no. 1, pp. 23–35, Feb. 1974, doi: 10.1017/S0016672300014634.
[66] P. J. Gerrish and R. E. Lenski, “The fate of competing beneficial mutations in an asexual population,” *Genetica*, vol. 102–103, no. 0, pp. 127–144, Mar. 1998, doi: 10.1023/A:1017067816551.
[67] B. Charlesworth, M. T. Morgan, and D. Charlesworth, “The effect of deleterious mutations on neutral molecular variation.,” *Genetics*, vol. 134, no. 4, pp. 1289–1303, Aug. 1993, doi: 10.1093/genetics/134.4.1289.
[68] B. Charlesworth, “The effect of background selection against deleterious mutations on weakly selected, linked variants,” *Genet. Res.*, vol. 63, no. 3, pp. 213–227, Jun. 1994, doi: 10.1017/S0016672300032365.

[69] M. L. Metzker, “Sequencing technologies — the next generation,” *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 31–46, Jan. 2010, doi: 10.1038/nrg2626.
[70] R. J. Elshire *et al.*, “A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species,” *PLoS ONE*, vol. 6, no. 5, p. e19379, May 2011, doi: 10.1371/journal.pone.0019379.
[71] F. Ozsolak and P. M. Milos, “RNA sequencing: advances, challenges and opportunities,” *Nat. Rev. Genet.*, vol. 12, no. 2, pp. 87–98, Feb. 2011, doi: 10.1038/nrg2934.
[72] E. L. Van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, “Ten years of next-generation sequencing technology,” *Trends Genet.*, vol. 30, no. 9, pp. 418–426, Sep. 2014, doi: 10.1016/j.tig.2014.07.001.
[73] S. Cordey *et al.*, “Rhinovirus Genome Evolution during Experimental Human Infection,” *PLoS ONE*, vol. 5, no. 5, p. e10588, May 2010, doi: 10.1371/journal.pone.0010588.
[74] F. Bertels, C. Leemann, K. J. Metzner, and R. R. Regoes, “Parallel Evolution of HIV-1 in a Long-Term Experiment,” *Mol. Biol. Evol.*, vol. 36, no. 11, pp. 2400–2414, Nov. 2019, doi: 10.1093/molbev/msz155.
[75] B. Van Den Bergh *et al.*, “Frequency of antibiotic application drives rapid evolutionary adaptation of Escherichia coli persistence,” *Nat. Microbiol.*, vol. 1, no. 5, p. 16020, Mar. 2016, doi: 10.1038/nmicrobiol.2016.20.
[76] J. Barroso-Batista *et al.*, “Specific Eco-evolutionary Contexts in the Mouse Gut Reveal Escherichia coli Metabolic Versatility,” *Curr. Biol.*, vol. 30, no. 6, pp. 1049-1062.e7, Mar. 2020, doi: 10.1016/j.cub.2020.01.050.
[77] E. M. Frenkel, B. H. Good, and M. M. Desai, “The Fates of Mutant Lineages and the Distribution of Fitness Effects of Beneficial Mutations in Laboratory Budding Yeast Populations,” *Genetics*, vol. 196, no. 4, pp. 1217–1226, Apr. 2014, doi: 10.1534/genetics.113.160069.
[78] K. J. Fisher, S. W. Buskirk, R. C. Vignogna, D. A. Marad, and G. I. Lang, “Adaptive genome duplication affects patterns of molecular evolution in Saccharomyces cerevisiae,” *PLOS Genet.*, vol. 14, no. 5, p. e1007396, May 2018, doi: 10.1371/journal.pgen.1007396.
[79] Y. Raynes, C. S. Wylie, P. D. Sniegowski, and D. M. Weinreich, “Sign of selection on mutation rate modifiers depends on population size,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 13, pp. 3422–3427, Mar. 2018, doi: 10.1073/pnas.1715996115.
[80] P. OROZCO-terWENGEL, M. Kapun, V. Nolte, R. Kofler, T. Flatt, and C. Schlötterer, “Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles,” *Mol. Ecol.*, vol. 21, no. 20, pp. 4931–4941, Oct. 2012, doi: 10.1111/j.1365-294X.2012.05673.x.
[81] P. C. Griffin, S. B. Hangartner, A. Fournier-Level, and A. A. Hoffmann, “Genomic Trajectories to Desiccation Resistance: Convergence and Divergence Among Replicate Selected *Drosophila* Lines,” *Genetics*, vol. 205, no. 2, pp. 871–890, Feb. 2017, doi: 10.1534/genetics.116.187104.
[82] F. Zanini *et al.*, “Population genomics of intrapatient HIV-1 evolution,” *eLife*, vol. 4, p. e11282, Dec. 2015, doi: 10.7554/eLife.11282.
[83] I. Setliff *et al.*, “Multi-Donor Longitudinal Antibody Repertoire Sequencing Reveals the Existence of Public Antibody Clonotypes in HIV-1 Infection,” *Cell Host Microbe*, vol. 23, no. 6, pp. 845-854.e6, Jun. 2018, doi: 10.1016/j.chom.2018.05.001.
[84] M. Imai *et al.*, “Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets,” *Nature*, vol. 486, no. 7403, pp. 420–428, Jun. 2012, doi: 10.1038/nature10831.

[85] J. T. McCrone, R. J. Woods, E. T. Martin, R. E. Malosh, A. S. Monto, and A. S. Lauring, “Stochastic processes constrain the within and between host evolution of influenza virus,” *eLife*, vol. 7, p. e35962, May 2018, doi: 10.7554/eLife.35962.
[86] K. S. Xue, L. H. Moncla, T. Bedford, and J. D. Bloom, “Within-Host Evolution of Human Influenza Virus,” *Trends Microbiol.*, vol. 26, no. 9, pp. 781–793, Sep. 2018, doi: 10.1016/j.tim.2018.02.007.
[87] R. A. Bull *et al.*, “Sequential Bottlenecks Drive Viral Evolution in Early Acute Hepatitis C Virus Infection,” *PLoS Pathog.*, vol. 7, no. 9, p. e1002243, Sep. 2011, doi: 10.1371/journal.ppat.1002243.
[88] H. Takeda *et al.*, “Evolution of multi-drug resistant HCV clones from pre-existing resistant-associated variants during direct-acting antiviral therapy determined by third-generation sequencing,” *Sci. Rep.*, vol. 7, no. 1, p. 45605, Mar. 2017, doi: 10.1038/srep45605.
[89] S. A. Kemp *et al.*, “SARS-CoV-2 evolution during treatment of chronic infection,” *Nature*, vol. 592, no. 7853, pp. 277–282, Apr. 2021, doi: 10.1038/s41586-021-03291-y.
[90] F. A. T. Boshier *et al.*, “Evolution of viral variants in remdesivir-treated and untreated SARS-CoV-2-infected pediatrics patients,” *J. Med. Virol.*, vol. 94, no. 1, pp. 161–172, Jan. 2022, doi: 10.1002/jmv.27285.
[91] A. Fischer, I. Vázquez-García, C. J. R. Illingworth, and V. Mustonen, “High-Definition Reconstruction of Clonal Composition in Cancer,” *Cell Rep.*, vol. 7, no. 5, pp. 1740–1752, Jun. 2014, doi: 10.1016/j.celrep.2014.04.055.
[92] N. Ishaque *et al.*, “Whole genome sequencing puts forward hypotheses on metastasis evolution and therapy in colorectal cancer,” *Nat. Commun.*, vol. 9, no. 1, p. 4782, Nov. 2018, doi: 10.1038/s41467-018-07041-z.
[93] J. P. Bollback, T. L. York, and R. Nielsen, “Estimation of 2 *N* e *s* From Temporal Allele Frequency Data,” *Genetics*, vol. 179, no. 1, pp. 497–502, May 2008, doi: 10.1534/genetics.107.085019.
[94] A.-S. Malaspinas, O. Malaspinas, S. N. Evans, and M. Slatkin, “Estimating Allele Age and Selection Coefficient from Time-Serial Data,” *Genetics*, vol. 192, no. 2, pp. 599–607, Oct. 2012, doi: 10.1534/genetics.112.140939.
[95] I. Mathieson and G. McVean, “Estimating Selection Coefficients in Spatially Structured Populations from Time Series Data of Allele Frequencies,” *Genetics*, vol. 193, no. 3, pp. 973–984, Mar. 2013, doi: 10.1534/genetics.112.147611.
[96] A. F. Feder, S. Kryazhimskiy, and J. B. Plotkin, “Identifying Signatures of Selection in Genetic Time Series,” *Genetics*, vol. 196, no. 2, pp. 509–522, Feb. 2014, doi: 10.1534/genetics.113.158220.
[97] M. Lacerda and C. Seoighe, “Population Genetics Inference for Longitudinally-Sampled Mutants Under Strong Selection,” *Genetics*, vol. 198, no. 3, pp. 1237–1250, Nov. 2014, doi: 10.1534/genetics.114.167957.
[98] M. Foll, H. Shim, and J. D. Jensen, “WFABC : a W right– F isher ABC -based approach for inferring effective population sizes and selection coefficients from time-sampled data,” *Mol. Ecol. Resour.*, vol. 15, no. 1, pp. 87–98, Jan. 2015, doi: 10.1111/1755-0998.12280.
[99] H. Topa, Á. Jónás, R. Kofler, C. Kosiol, and A. Honkela, “Gaussian process test for high-throughput sequencing time series: application to experimental evolution,” *Bioinformatics*, vol. 31, no. 11, pp. 1762–1770, Jun. 2015, doi: 10.1093/bioinformatics/btv014.
[100] A. Ferrer-Admetlla, C. Leuenberger, J. D. Jensen, and D. Wegmann, “An Approximate Markov Model for the Wright–Fisher Diffusion and Its Application to

Time Series Data," *Genetics*, vol. 203, no. 2, pp. 831–846, Jun. 2016, doi: 10.1534/genetics.115.184598.
[101] J. G. Schraiber, S. N. Evans, and M. Slatkin, "Bayesian Inference of Natural Selection from Allele Frequency Time Series," *Genetics*, vol. 203, no. 1, pp. 493–511, May 2016, doi: 10.1534/genetics.116.187278.
[102] A. Iranmehr, A. Akbari, C. Schlötterer, and V. Bafna, "CLEAR : Composition of Likelihoods for Evolve and Resequence Experiments," *Genetics*, vol. 206, no. 2, pp. 1011–1023, Jun. 2017, doi: 10.1534/genetics.116.197566.
[103] T. Taus, A. Futschik, and C. Schlötterer, "Quantifying Selection with Pool-Seq Time Series Data," *Mol. Biol. Evol.*, vol. 34, no. 11, pp. 3023–3034, Nov. 2017, doi: 10.1093/molbev/msx225.
[104] T. Zinger, M. Gelbart, D. Miller, P. S. Pennings, and A. Stern, "Inferring population genetics parameters of evolving viruses using time-series data," *Virus Evol.*, vol. 5, no. 1, p. vez011, Jan. 2019, doi: 10.1093/ve/vez011.
[105] Q. Cheng, M. S. Sohail, and M. R. McKay, "Selection Estimation from Genetic Time-Series Data: Effects of Limited Sampling and Genetic Drift," *Mol. Biol. Evol.*, vol. 42, no. 12, p. msaf301, Nov. 2025, doi: 10.1093/molbev/msaf301.
[106] M. S. Sohail, R. H. Y. Louie, Z. Hong, J. P. Barton, and M. R. McKay, "Inferring Epistasis from Genetic Time-series Data," *Mol. Biol. Evol.*, vol. 39, no. 10, p. msac199, Oct. 2022, doi: 10.1093/molbev/msac199.
[107] Y. Gao and J. P. Barton, "A binary trait model reveals the fitness effects of HIV-1 escape from T cell responses," *Proc. Natl. Acad. Sci.*, vol. 122, no. 8, p. e2405379122, Feb. 2025, doi: 10.1073/pnas.2405379122.
[108] H.-L. Zeng, C.-L. Yang, B. Jing, J. Barton, and E. Aurell, "Two fitness inference schemes compared using allele frequencies from 1068 391 sequences sampled in the UK during the COVID-19 pandemic," *Phys. Biol.*, vol. 22, no. 1, p. 016003, Jan. 2025, doi: 10.1088/1478-3975/ad9213.
[109] B. Lee *et al.*, "Inferring effects of mutations on SARS-CoV-2 transmission from genomic surveillance data," *Nat. Commun.*, vol. 16, no. 1, p. 441, Jan. 2025, doi: 10.1038/s41467-024-55593-0.
[110] P. W. Hedrick, "Gametic Disequilibrium Measures: Proceed With Caution," *Genetics*, vol. 117, no. 2, pp. 331–341, Oct. 1987, doi: 10.1093/genetics/117.2.331.
[111] S. Posada-Cespedes, D. Seifert, and N. Beerenwinkel, "Recent advances in inferring viral diversity from high-throughput sequencing data," *Virus Res.*, vol. 239, pp. 17–32, Jul. 2017, doi: 10.1016/j.virusres.2016.09.016.
[112] M. Schirmer, W. T. Sloan, and C. Quince, "Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes," *Brief. Bioinform.*, vol. 15, no. 3, pp. 431–442, May 2014, doi: 10.1093/bib/bbs081.
[113] A. Töpfer, O. Zagordi, S. Prabhakaran, V. Roth, E. Halperin, and N. Beerenwinkel, "Probabilistic Inference of Viral Quasispecies Subject to Recombination," *J. Comput. Biol.*, vol. 20, no. 2, pp. 113–123, Feb. 2013, doi: 10.1089/cmb.2012.0232.
[114] L. Breiman, "Bagging Predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1023/A:1018054314350.
[115] *Picard*. [Online]. Available: http://broadinstitute.github.io/picard
[116] A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins, "Sambamba: fast processing of NGS alignment formats," *Bioinformatics*, vol. 31, no. 12, pp. 2032–2034, Jun. 2015, doi: 10.1093/bioinformatics/btv098.
[117] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," 2013, *arXiv*. doi: 10.48550/ARXIV.1303.3997.

[118] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.
[119] L. Breiman, “Pasting Small Votes for Classification in Large Databases and On-Line,” *Mach. Learn.*, vol. 36, no. 1–2, pp. 85–103, Jul. 1999, doi: 10.1023/A:1007563306331.
[120] W. Huang, L. Li, J. R. Myers, and G. T. Marth, “ART: a next-generation sequencing read simulator,” *Bioinformatics*, vol. 28, no. 4, pp. 593–594, Feb. 2012, doi: 10.1093/bioinformatics/btr708.
[121] D. Focosi, F. Maggi, M. Franchini, S. McConnell, and A. Casadevall, “Analysis of Immune Escape Variants from Antibody-Based Therapeutics against COVID-19: A Systematic Review,” *Int. J. Mol. Sci.*, vol. 23, no. 1, p. 29, Dec. 2021, doi: 10.3390/ijms23010029.
[122] X. Tang, Z. Qian, X. Lu, and J. Lu, “Adaptive Evolution of the Spike Protein in Coronaviruses,” *Mol. Biol. Evol.*, vol. 40, no. 4, p. msad089, Apr. 2023, doi: 10.1093/molbev/msad089.
[123] Y. Cao *et al.*, “Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution,” *Nature*, vol. 614, no. 7948, pp. 521–529, Feb. 2023, doi: 10.1038/s41586-022-05644-7.
[124] C. O. Barnes *et al.*, “Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies,” *Cell*, vol. 182, no. 4, pp. 828-842.e16, Aug. 2020, doi: 10.1016/j.cell.2020.06.025.
[125] van Zwetselaar M, *sc2calc - SARS-CoV-2 Genome Coordinate Converter*. (Oct. 20, 2025). [Online]. Available: https://github.com/zwets/sc2calc
[126] M. Sunnåker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz, “Approximate Bayesian Computation,” *PLoS Comput. Biol.*, vol. 9, no. 1, p. e1002803, Jan. 2013, doi: 10.1371/journal.pcbi.1002803.
[127] A. Wilm *et al.*, “LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets,” *Nucleic Acids Res.*, vol. 40, no. 22, pp. 11189–11201, Dec. 2012, doi: 10.1093/nar/gks918.
[128] Y. Cao *et al.*, “Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies,” *Nature*, vol. 602, no. 7898, pp. 657–663, Feb. 2022, doi: 10.1038/s41586-021-04385-3.
[129] Y. Cao *et al.*, “BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection,” *Nature*, vol. 608, no. 7923, pp. 593–602, Aug. 2022, doi: 10.1038/s41586-022-04980-y.
[130] J. Huo *et al.*, “A potent SARS-CoV-2 neutralising nanobody shows therapeutic efficacy in the Syrian golden hamster model of COVID-19,” *Nat. Commun.*, vol. 12, no. 1, p. 5469, Sep. 2021, doi: 10.1038/s41467-021-25480-z.
[131] Z. Wang *et al.*, “mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants,” *Nature*, vol. 592, no. 7855, pp. 616–622, Apr. 2021, doi: 10.1038/s41586-021-03324-6.
[132] P. R. Rani *et al.*, “Symptomatic reinfection of SARS-CoV-2 with spike protein variant N440K associated with immune escape,” *J. Med. Virol.*, vol. 93, no. 7, pp. 4163–4165, Jul. 2021, doi: 10.1002/jmv.26997.
[133] Z. Liu *et al.*, “Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization,” *Cell Host Microbe*, vol. 29, no. 3, pp. 477-488.e4, Mar. 2021, doi: 10.1016/j.chom.2021.01.014.

[134] D. Focosi *et al.*, “Emergence of SARS-COV-2 Spike Protein Escape Mutation Q493R after Treatment for COVID-19,” *Emerg. Infect. Dis.*, vol. 27, no. 10, pp. 2728–2731, Oct. 2021, doi: 10.3201/eid2710.211538.
[135] Z. Cui *et al.*, “Structural and functional characterizations of infectivity and immune evasion of SARS-CoV-2 Omicron,” *Cell*, vol. 185, no. 5, pp. 860-871.e13, Mar. 2022, doi: 10.1016/j.cell.2022.01.019.
[136] Q. Hong *et al.*, “Molecular basis of receptor binding and antibody neutralization of Omicron,” *Nature*, vol. 604, no. 7906, pp. 546–552, Apr. 2022, doi: 10.1038/s41586-022-04581-9.
[137] W. B. Wang *et al.*, “Identification of key mutations responsible for the enhancement of receptor-binding affinity and immune escape of SARS-CoV-2 Omicron variant,” *J. Mol. Graph. Model.*, vol. 124, p. 108540, Nov. 2023, doi: 10.1016/j.jmgm.2023.108540.
[138] Y. Li *et al.*, “T-cell responses to SARS-CoV-2 Omicron spike epitopes with mutations after the third booster dose of an inactivated vaccine,” *J. Med. Virol.*, vol. 94, no. 8, pp. 3998–4004, Aug. 2022, doi: 10.1002/jmv.27814.
[139] P. Liu *et al.*, “Spike N354 glycosylation augments SARS-CoV-2 fitness for human adaptation through structural plasticity,” *Natl. Sci. Rev.*, vol. 11, no. 7, p. nwae206, Jun. 2024, doi: 10.1093/nsr/nwae206.
[140] T. N. Starr *et al.*, “Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution,” *Science*, vol. 377, no. 6604, pp. 420–424, Jul. 2022, doi: 10.1126/science.abo7896.
[141] Q. Li *et al.*, “The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity,” *Cell*, vol. 182, no. 5, pp. 1284-1294.e9, Sep. 2020, doi: 10.1016/j.cell.2020.07.012.
[142] X. Deng *et al.*, “Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant,” *Cell*, vol. 184, no. 13, pp. 3426-3437.e8, Jun. 2021, doi: 10.1016/j.cell.2021.04.025.
[143] J. Ou *et al.*, “V367F Mutation in SARS-CoV-2 Spike RBD Emerging during the Early Transmission Phase Enhances Viral Infectivity through Increased Human ACE2 Receptor Binding Affinity,” *J. Virol.*, vol. 95, no. 16, pp. e00617-21, Jul. 2021, doi: 10.1128/JVI.00617-21.
[144] I. Kimura *et al.*, “The SARS-CoV-2 spike S375F mutation characterizes the Omicron BA.1 variant,” *iScience*, vol. 25, no. 12, p. 105720, Dec. 2022, doi: 10.1016/j.isci.2022.105720.
[145] B. Zheng *et al.*, “S373P Mutation Stabilizes the Receptor-Binding Domain of the Spike Protein in Omicron and Promotes Binding,” *JACS Au*, vol. 3, no. 7, pp. 1902–1910, Jul. 2023, doi: 10.1021/jacsau.3c00142.
[146] F. Tian *et al.*, “N501Y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2,” *eLife*, vol. 10, p. e69091, Aug. 2021, doi: 10.7554/eLife.69091.
[147] B. Hu *et al.*, “Spike mutations contributing to the altered entry preference of SARS-CoV-2 omicron BA.1 and BA.2,” *Emerg. Microbes Infect.*, vol. 11, no. 1, pp. 2275–2287, Dec. 2022, doi: 10.1080/22221751.2022.2117098.
[148] A. M. Nafea *et al.*, “Application of next-generation sequencing to identify different pathogens,” *Front. Microbiol.*, vol. 14, p. 1329330, Jan. 2024, doi: 10.3389/fmicb.2023.1329330.
[149] J. Huddleston *et al.*, “Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution,” *eLife*, vol. 9, p. e60067, Sep. 2020, doi: 10.7554/eLife.60067.

[150] P. H. Sudmant *et al.*, “An integrated map of structural variation in 2,504 human genomes,” *Nature*, vol. 526, no. 7571, pp. 75–81, Oct. 2015, doi: 10.1038/nature15394.
[151] E. Christaki, M. Marcou, and A. Tofarides, “Antimicrobial Resistance in Bacteria: Mechanisms, Evolution, and Persistence,” *J. Mol. Evol.*, vol. 88, no. 1, pp. 26–40, Jan. 2020, doi: 10.1007/s00239-019-09914-3.
[152] A. M. Carabelli *et al.*, “SARS-CoV-2 variant biology: immune escape, transmission and fitness,” *Nat. Rev. Microbiol.*, Jan. 2023, doi: 10.1038/s41579-022-00841-7.
[153] E. Adachi *et al.*, “Hepatitis B Virus Reactivation After Switch to Cabotegravir/Rilpivirine in Patient with Low Hepatitis B Surface Antibody,” *Emerg. Infect. Dis.*, vol. 30, no. 8, Aug. 2024, doi: 10.3201/eid3008.240019.
[154] S. A. Bergin *et al.*, “Systematic Analysis of Copy Number Variations in the Pathogenic Yeast Candida parapsilosis Identifies a Gene Amplification in *RTA3* That is Associated with Drug Resistance,” *mBio*, vol. 13, no. 5, pp. e01777-22, Oct. 2022, doi: 10.1128/mbio.01777-22.
[155] A. Canalda-Baltrons *et al.*, “Genome graphs reveal the importance of structural variation in Mycobacterium tuberculosis evolution and drug resistance,” *Nat. Commun.*, vol. 16, no. 1, p. 10746, Nov. 2025, doi: 10.1038/s41467-025-65779-9.
[156] R. L. Collins and M. E. Talkowski, “Diversity and consequences of structural variation in the human genome,” *Nat. Rev. Genet.*, vol. 26, no. 7, pp. 443–462, Jul. 2025, doi: 10.1038/s41576-024-00808-9.
[157] Y. Zhou *et al.*, “Graph pangenome captures missing heritability and empowers tomato breeding,” *Nature*, vol. 606, no. 7914, pp. 527–534, Jun. 2022, doi: 10.1038/s41586-022-04808-9.
[158] X. An *et al.*, “Genomic structural variation is associated with hypoxia adaptation in high-altitude zokors,” *Nat. Ecol. Evol.*, vol. 8, no. 2, pp. 339–351, Jan. 2024, doi: 10.1038/s41559-023-02275-7.
[159] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, “The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community,” *Genome Biol.*, vol. 17, no. 1, p. 239, Dec. 2016, doi: 10.1186/s13059-016-1103-0.
[160] M. O. Pollard, D. Gurdasani, A. J. Mentzer, T. Porter, and M. S. Sandhu, “Long reads: their purpose and place,” *Hum. Mol. Genet.*, vol. 27, no. R2, pp. R234–R241, Aug. 2018, doi: 10.1093/hmg/ddy177.
[161] Y. Chen *et al.*, “Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak,” *Nat. Commun.*, vol. 14, no. 1, p. 283, Jan. 2023, doi: 10.1038/s41467-023-35996-1.
[162] S. V. Iyer, S. Goodwin, and W. R. McCombie, “Leveraging the power of long reads for targeted sequencing,” *Genome Res.*, vol. 34, no. 11, pp. 1701–1718, Nov. 2024, doi: 10.1101/gr.279168.124.
[163] K. Shafin *et al.*, “Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes,” *Nat. Biotechnol.*, vol. 38, no. 9, pp. 1044–1053, Sep. 2020, doi: 10.1038/s41587-020-0503-6.
[164] L. Zong *et al.*, “An optimized workflow of full-length transcriptome sequencing for accurate fusion transcript identification,” *RNA Biol.*, vol. 21, no. 1, pp. 1199–1208, Dec. 2024, doi: 10.1080/15476286.2024.2425527.
[165] A. Rhoads and K. F. Au, “PacBio Sequencing and its Applications,” *Genomics Proteomics Bioinformatics*, vol. 13, no. 5, pp. 278–289, Oct. 2015, doi: 10.1016/j.gpb.2015.08.002.
[166] M. Bloomfield, S. Hutton, C. Velasco, M. Burton, M. Benton, and M. Storey, “Oxford nanopore next generation sequencing in a front-line clinical microbiology

laboratory without on-site bioinformaticians," *Pathology (Phila.)*, vol. 56, no. 3, pp. 444–447, Apr. 2024, doi: 10.1016/j.pathol.2023.07.014.
[167] E. C. Patterson, S. Morrison-Lanjouw, M. A. Jobling, and J. H. Wetton, "Rapid on-site universal vertebrate species identification via multi-barcode nanopore sequencing," Jun. 08, 2025, *Genetics*. doi: 10.1101/2025.06.04.657926.
[168] M. Mori *et al.*, "Nanopore Sequencing for Characterization of HIV-1 Recombinant Forms," *Microbiol. Spectr.*, vol. 10, no. 4, pp. e01507-22, Aug. 2022, doi: 10.1128/spectrum.01507-22.
[169] N. Riaz *et al.*, "Adaptation of Oxford Nanopore technology for hepatitis C whole genome sequencing and identification of within-host viral variants," *BMC Genomics*, vol. 22, no. 1, p. 148, Dec. 2021, doi: 10.1186/s12864-021-07460-1.
[170] J. B. Stenbäck *et al.*, "Accurate and cost-efficient whole genome sequencing of hepatitis B virus using Nanopore," Aug. 12, 2024, *Infectious Diseases (except HIV/AIDS)*. doi: 10.1101/2024.08.12.24311345.
[171] C. M. Voloch *et al.*, "Intra-host evolution during SARS-CoV-2 prolonged infection," *Virus Evol.*, vol. 7, no. 2, p. veab078, Dec. 2021, doi: 10.1093/ve/veab078.
[172] S. Pei, T. Liu, X. Ren, W. Li, C. Chen, and Z. Xie, "Benchmarking variant callers in next-generation and third-generation sequencing analysis," *Brief. Bioinform.*, vol. 22, no. 3, p. bbaa148, May 2021, doi: 10.1093/bib/bbaa148.
[173] Z. Liu, Z. Xie, and M. Li, "Comprehensive and deep evaluation of structural variation detection pipelines with third-generation sequencing data," *Genome Biol.*, vol. 25, no. 1, p. 188, Jul. 2024, doi: 10.1186/s13059-024-03324-5.
[174] D. Cai and Y. Sun, "Reconstructing viral haplotypes using long reads," *Bioinformatics*, vol. 38, no. 8, pp. 2127–2134, Apr. 2022, doi: 10.1093/bioinformatics/btac089.
[175] T. Jiang, S. Liu, S. Cao, and Y. Wang, "Structural Variant Detection from Long-Read Sequencing Data with cuteSV," in *Variant Calling*, vol. 2493, C. Ng and S. Piscuoglio, Eds., in Methods in Molecular Biology, vol. 2493. , New York, NY: Springer US, 2022, pp. 137–151. doi: 10.1007/978-1-0716-2293-3_9.
[176] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, Sep. 2018, doi: 10.1093/bioinformatics/bty191.
[177] K. Katoh, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002, doi: 10.1093/nar/gkf436.
[178] T. Hon *et al.*, "Highly accurate long-read HiFi sequencing data for five complex genomes," *Sci. Data*, vol. 7, no. 1, p. 399, Nov. 2020, doi: 10.1038/s41597-020-00743-4.
[179] P. J. Halfmann *et al.*, "Evolution of a globally unique SARS-CoV-2 Spike E484T monoclonal antibody escape mutation in a persistently infected, immunocompromised individual," *Virus Evol.*, vol. 9, no. 2, p. veac104, Jul. 2023, doi: 10.1093/ve/veac104.
[180] R. A. Fisher, "XXI.—On the Dominance Ratio," *Proc. R. Soc. Edinb.*, vol. 42, pp. 321–341, 1923, doi: 10.1017/S0370164600023993.
[181] S. Wright, "EVOLUTION IN MENDELIAN POPULATIONS," *Genetics*, vol. 16, no. 2, pp. 97–159, Mar. 1931, doi: 10.1093/genetics/16.2.97.
[182] R. Wick, "Badread: simulation of error-prone long reads," *J. Open Source Softw.*, vol. 4, no. 36, p. 1316, Apr. 2019, doi: 10.21105/joss.01316.
[183] R. Saranathan *et al.*, "Capturing Structural Variants of Herpes Simplex Virus Genome in Full Length by Oxford Nanopore Sequencing," *Microbiol. Spectr.*, vol. 10, no. 5, pp. e02285-22, Oct. 2022, doi: 10.1128/spectrum.02285-22.

[184] M. Krzywinski *et al.*, “Circos: An information aesthetic for comparative genomics,” *Genome Res.*, vol. 19, no. 9, pp. 1639–1645, Sep. 2009, doi: 10.1101/gr.092759.109.
[185] M. Harrower and C. A. Brewer, “ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps,” *Cartogr. J.*, vol. 40, no. 1, pp. 27–37, Jun. 2003, doi: 10.1179/000870403235002042.
[186] S. M. U. Abdullah, M. S. Sohail, R. H. Y. Louie, Y. Sun, J. P. Barton, and M. R. McKay, “Linkage-aware inference of fitness from short-read time-series genomic data,” *Virus Evol.*, vol. 12, no. 1, p. veag027, Jan. 2026, doi: 10.1093/ve/veag027.
[187] J. Terhorst, C. Schlötterer, and Y. S. Song, “Multi-locus Analysis of Genomic Time Series Data from Experimental Evolution,” *PLOS Genet.*, vol. 11, no. 4, p. e1005069, Apr. 2015, doi: 10.1371/journal.pgen.1005069.
[188] Y. Li and J. P. Barton, “Correlated Allele Frequency Changes Reveal Clonal Structure and Selection in Temporal Genetic Data,” *Mol. Biol. Evol.*, vol. 41, no. 4, p. msae060, Apr. 2024, doi: 10.1093/molbev/msae060.
[189] R. Vita *et al.*, “The Immune Epitope Database (IEDB): 2024 update,” *Nucleic Acids Res.*, vol. 53, no. D1, pp. D436–D443, Jan. 2025, doi: 10.1093/nar/gkae1092.
[190] M. C. Jespersen, B. Peters, M. Nielsen, and P. Marcatili, “BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes,” *Nucleic Acids Res.*, vol. 45, no. W1, pp. W24–W29, Jul. 2017, doi: 10.1093/nar/gkx346.
[191] K. S. Shimagaki and J. P. Barton, “Efficient epistasis inference via higher-order covariance matrix factorization,” *GENETICS*, vol. 230, no. 4, p. iyaf118, Aug. 2025, doi: 10.1093/genetics/iyaf118.
[192] Y. J. Baek *et al.*, “Within-Host Evolution of SARS-CoV-2 in a B-Cell Depleted Patient With Successful Treatment,” *J. Korean Med. Sci.*, vol. 38, no. 22, p. e175, 2023, doi: 10.3346/jkms.2023.38.e175.
[193] E. G. Quaranta *et al.*, “SARS-CoV-2 intra-host evolution during prolonged infection in an immunocompromised patient,” *Int. J. Infect. Dis.*, vol. 122, pp. 444–448, Sep. 2022, doi: 10.1016/j.ijid.2022.06.023.
[194] A. Houldsworth *et al.*, “Haplotype analysis finds linkage disequilibrium in the IL-12 gene in patients with HCV: Haplotype Analysis Finds Linkage Disequilibrium,” *J. Med. Virol.*, vol. 87, no. 7, pp. 1207–1217, Jul. 2015, doi: 10.1002/jmv.24179.
[195] N. J. Higham, “Computing the nearest correlation matrix--a problem from finance,” *IMA J. Numer. Anal.*, vol. 22, no. 3, pp. 329–343, Jul. 2002, doi: 10.1093/imanum/22.3.329.
[196] J. E. Falk, “Optimization by Vector Space Methods (David G. Luenberger),” *SIAM Rev.*, vol. 12, no. 2, pp. 315–316, Apr. 1970, doi: 10.1137/1012072.
[197] R. L. Dykstra, “An Algorithm for Restricted Least Squares Regression,” *J. Am. Stat. Assoc.*, vol. 78, no. 384, pp. 837–842, Dec. 1983, doi: 10.1080/01621459.1983.10477029.

PUBLICATIONS

I have worked on the following works during my study

1. **S. M. U. Abdullah**, M. S. Sohail, R. H. Y. Louie, Y. Sun, J. P. Barton, and M. R. McKay, "Linkage-aware inference of fitness from short-read time-series genomic data," *Virus Evolution*, vol. 12, no. 1, p. veag027, 2026.

2. **S. M. U. Abdullah**, D. Cai, J. D. Ip, K. K. W. To, and Y. Sun, "SLAFIT: Viral fitness inference and structural variant detection from TGS data," (under review at *Bioinformatics.* 2026)

3. **S. M. U. Abdullah**, Y. Sun, "LDEN: linkage disequilibrium estimation from next-generation sequencing data," (in preparation 2026)

The following two works are not part of this thesis

4. R. Yu, **S. M. U. Abdullah**, Y. Sun, "HMMPolish: a coding region polishing tool for TGS-sequenced RNA viruses," *Briefings in bioinformatics*, vol. 24, no. 5, p. bbad264, 2023.

5. J. D. Ip, A. W. H. Chu, W. M. Chan, R. C. Y. Leung, **S. M. U. Abdullah**, Y. Sun, and K. K. W To, "Global prevalence of SARS-CoV-2 3CL protease mutations associated with nirmatrelvir or ensitrelvir resistance," *EBioMedicine*, vol 91, 2023.