UCLA

UCLA Electronic Theses and Dissertations

Title

Quantifying Viral Replication with Genetic Barcoding and Fitness Profiling

Permalink

https://escholarship.org/uc/item/2rv0g3t1

Author

Zhang, Tianhao

Publication Date

2023

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Quantifying Viral Replication with Genetic Barcoding and Fitness Profiling

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Molecular Biology

by

Tianhao Zhang

2023

© Copyright by

Tianhao Zhang

2023

ABSTRACT OF THE DISSERTATION

Quantifying Viral Replication with

Genetic Barcoding and Fitness Profiling

by

Tianhao Zhang

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2023

Professor Jerome A. Zack, Chair

Virus replication is a stochastic process at a delicate equilibrium between many intrinsic and extrinsic factors. On the one hand, a single copy of viral genomic material can produce thousands of copies of progenies. The descendants consist of a simplistic molecular machinery that can rapidly spread to new environments and infect new hosts. On the other hand, viral replication and assembly is an error prone process that produces a large fraction of defective particles. The high error rate of virus replication and the huge fluctuation of the population size make it difficult to track and study the life history of viruses.

In this thesis, I will introduce two powerful platforms that can accurately trace the viral life history and their applications. The genetic barcode labels each individual viral clone with a unique piece of genetic information. Coupling with the next generation sequencing, the evolutionary history of every viral lineage can be reconstructed. Secondly, the fitness profiling platform combines high throughput mutagenesis with high throughput

sequencing to quantify the frequency change of mutants after various screening conditions. It efficiently measures the mutational effects on various viral systems.

I will first apply the fitness profiling platform to measure the mutational effect of drug resistant mutations on HIV-1 replication. We found the pervasive positive genetic interactions between the drug resistant mutations, indicating a potential mechanism that HIV-1 maintains drug resistance. It uses multiple mutations to compensate for the fitness cost of drug resistance.

Next, I modified the fitness profiling platform to quantify the mutational effect on the SARS-CoV-2 nucleocapsid protein. We measured the stability of ~8000 mutant proteins and found a continuous increasing trend of protein stability after the virus first transmitted to the human species.

Thirdly, I introduced the construction of the HIV-1 barcode library and its application in single cell multi-omics. We developed a new sequencing method that can simultaneously sequence the integration site of HIV-1 and the alternative splicing forms of the viral mRNA. The barcoded HIV-1 library helps illustrate how viral integration affects virus transcription and splicing.

Then I applied the barcoded HIV-1 library in humanized mice and used it to trace the clonal expansion of latently infected T cells. We designed a sequencing pipeline that can quantify the re-seeding events of each viral barcode and the clonal expansion events of each infected T cell clone. Combined with viral RNA barcode amplicon sequencing, we found the genomic features leading to less viral transcription and more T cell clonal expansion.

Lastly, I presented the technical improvement that enabled accurate sequencing of the genetic barcodes and viral mutants. We benchmarked error-free sequencing methods for the next generation sequencing. I also introduced a mathematical framework for estimating the mutational effect from high throughput fitness profiling data.

The dissertation of Tianhao Zhang is approved.

Irvin SY. Chen

Otto Orlean Yang

Alexander Hoffmann

Ren Sun

Jerome A. Zack, Committee Chair

University of California, Los Angeles

2023

TABLE OF CONTENTS

List of Figures

List of Tables

Acknowledgements

When I began my graduate studies, I was a science enthusiast with a keen interest in the natural history of the biology world. My goal was to acquire all human understanding of species, akin to that of a 19th-century naturalist. However, under the mentorship of Dr. Ren Sun, I transformed into a grounded and dedicated biology researcher. Dr. Sun is an exceptional leader who supports, dreams, invests, and befriends with warmth. His brilliant ideas and unique problem-solving approaches have made a significant impact in the fields of virology and immunology. He encouraged me to test my ideas, even using his personal salary savings to support my research. Above all, Dr. Sun taught me to focus not just on my own capabilities, but also on the needs of society. He instilled in me a sense of care and responsibility as a scientist.

During my graduate studies, I received invaluable support from my committee members. After Dr. Sun retired, Dr. Jerome Zack took me under his wing and his lab collaborated with me on multiple projects. He provided thoughtful intellectual guidance, financed costly experiments, and expedited my research with his technicians' assistance. Without his unwavering support, I could not have completed my degree. Dr. Otto Yang was an invaluable member of my committee, offering insightful critiques and critical suggestions regarding my projects. He is a true expert in HIV biology and served as an encyclopedia of knowledge for me. Dr. Irvin Chen played a key role in my projects, enlightening me with the most pertinent questions in the field. I gained exposure to essential HIV techniques and acquired necessary materials from his lab when I began my projects. Dr. Alexander Hoffmann, a renowned immunologist and computational biologist in my committee, provided unique perspectives on my research and offered thoughtful career development

guidance. Finally, Dr. Sri Kosuri expanded my knowledge on sequencing techniques and commercial translation, broadening my skill set and experience in the field.

The UCLA research community is a powerful and compassionate group of individuals. I am grateful for the support and encouragement that I have received from my colleagues, including Dr. Jocelyn Kim, who has been a tremendous collaborator and friend. She is a tenacious and resilient researcher who has helped me to overcome obstacles in challenging projects. When I fell ill, she was the first person I called. Dr. Matthew Marsden has been a kind and patient collaborator who has demonstrated tremendous fortitude and determination in proving the effectiveness of our methods. Dr. Masakazu Kamata has been an excellent teacher, sharing his knowledge of HIV experiment details generously with me. Dr. Xinmin Li and his core lab have been incredibly supportive, providing discounts and free sequencing runs while troubleshooting my experiments with care and patience. Dr. Ting-Ting Wu has been a knowledgeable advisor, offering insightful suggestions on my projects. Dr. Bin Liu and Dr. Ramin Salehi-Rad have been wonderful neighbors and have provided valuable support in the field of cancer biology. Dr. Rui Li has educated me in cancer epidemiology research. Dr. Jing Wen organized my graduation party. Dr. Alex Gorin has been an intelligent and responsive collaborator on CTL projects. Dr. Peter Bradley recruited me to the molecular biology graduate program, and the administrative team, Ashley Straight and Helen Houldsworth, have been tremendously helpful and considerate.

My lab mates are not just colleagues, they are my companions, my friends, and my family. Dr. Nicholas Wu is the individual who showed me the ropes in science. He taught me the fundamentals, from culturing bacteria to reading sequencing files, and shared all his

codes generously with me. Moreover, he provided me with suggestions on how to think, write, and communicate with others as a graduate student. I gained the initial momentum for my research from him. Dr. Yushen Du is my most frequent co-author and a great friend. She has rescued me from despair and depression on numerous occasions. Dr. Yuan Shi is an excellent supervisor and my closest friend in the lab. We share many common interests. Dr. Lei Dai is my mentor in evolutionary biology, and he always satisfies my curiosity and enthusiasm for evolution. Dr. Dangyang Gong is the molecular biology expert who has provided me with countless helpful suggestions. Dr. Haigen Huang took care of me during the pandemic and even gave me haircuts. Dr. Shih-hsin Chiu has invested a tremendous amount of time in the barcode virus project. Our lab manager, Hong Jiang, not only provides administrative support but also helps me with my benchwork.

I am grateful to my family for their unwavering support throughout these years. My wife, Donghui Wei, calls me every day to cheer me up and offer her encouragement. My parents have always been there for me, providing both financial and spiritual support. Even my grandparents worry about my safety and mental health, and I appreciate their concern and care.

In addition, there are a few more people that I would like to acknowledge. Dr. Sunnie Yoh from Dr. Sumit Chanda's lab has been an invaluable external collaborator. She generously provided me with reagents and taught me about studying HIV-1 innate immune response. Dr. Huachun Liu was a wonderful mentor for my TA job, providing me with her lesson plans and guiding me through the process. Matthew Kostelny spent a lot

of time working on my bench for the barcode project, and I am grateful for his dedication. Dr. Zhe Jing always takes the time to stop by and share helpful life hacks.

There are so many collaborators, lab mates, visiting scholars, and friends who have been a part of my journey, and I am grateful for the positivity and kindness they have shown me. Although I cannot list them all, each connection and experience has touched my heart and convinced me of the importance of devoting my life to meaningful and impactful scientific research.

Notes about the chapters in the thesis:

Chapter 2 is an adaptation of the manuscript: Zhang, T. H., Dai, L., Barton, J. P., Du, Y., Tan, Y., Pang, W., ... & Sun, R. (2020). Predominance of positive epistasis among drug resistance-associated mutations in HIV-1 protease. *PLoS genetics*, *16*(10), e1009009.

Chapter 6 is an adaptation of the manuscript: Zhang, T. H., Wu, N. C., & Sun, R. (2016). A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing. *BMC genomics*, *17*(1), 1-9.

Vita

Education

Fudan University | Bachelor of Biological Science | 2011 - 2015

University of California, Los Angeles | PhD candidate in Molecular Biology | 2015 – 2023

Personal Statement

I am a graduate student in Molecular Biology. My research interest is mainly the molecular evolution of viruses. I am motivated to know how stochastic events like viral mutations and random integrations affect the clonal and population dynamics during viral evolution. I have developed many skills in the graduate school. I developed a few computational pipelines dealing with the deep mutational scanning data and the HIV-1 integration site data. I am proficient in analyzing genomics, epigenomics, transcriptome and single cell multi-omics datasets. I also adapted deep learning models in our recent publications. My wet lab skills include designing and constructing large mutational libraries, establishing next generation sequencing workflows, BSL3 virology and other commonly used genetics and cell biology experiments. I am a very collaborative person in the lab. I support my lab members in their research projects and also acquired various skills and experience from them. I successfully tutored some undergraduate students and research assistants. When Dr. Ren Sun is not in the US, I organize the lab meetings and other lab routines. I am also an independent researcher. I am the only person working on HIV-1 in the lab. I independently designed my projects, interpretated data and wrote manuscripts.

Research Experience

1. Map the mutational effect on viral proteins in massive parallel. I used random mutagenesis or combinatorial mutagenesis to generate large viral mutant libraries. Coupled with the next generation sequencing, I measured the fitness cost of the mutations in various

conditions. The pipeline can be adapted to multiple viral platforms, including HIV-1, Influenza A, Zika Virus and SARS-CoV-2, even where a reverse genetics system was not available.

a. **Zhang, T. H.**, Dai, L., Barton, J. P., Du, Y., Tan, Y., Pang, W., ... & Sun, R. (2020). Predominance of positive epistasis among drug resistance-associated mutations in HIV-1 protease. *PLoS genetics*, *16*(10), e1009009.

b. **Zhang, T. H.**, Du, Y. Hong, M., Huang, H., Hong, J., … & Sun, R. (2022) Increasing abundance of SARS-CoV-2 N protein during transmission in human revealed by a high-throughput mutagenesis screening. In submission.

2. Trace viral clonal dynamics *in vivo* and *in vitro* using genetic barcoded HIV-1. I made barcoded HIV-1 libraries based on NL4-3, NFNSX and SHIV-AD8. In collaboration with labs proficient in retrovirus animal models, I measured the *in vivo* viral diversity in various conditions. I also developed a multi-omics method to study how viral integration affects its transcription and splicing.

a. Kim, J. T., **Zhang, T. H.,** Carmona, C., Lee, B., Seet, C. S., Kostelny, M., ... & Zack, J. A. (2022). Latency reversal plus natural killer cells diminish HIV reservoir in vivo. *Nature communications*, *13*(1), 1-14.

b. **Zhang, T. H.**, Du, Y., Shi, Y., Qiu, S., … & Sun, R. (2022) Genetic barcoded HIV-1 revealed the positional effect of integration site to transcription and splicing. (Thesis, in preparation)

3. Use high accuracy deep sequencing methods to study virus mutation *in vivo*. I benchmarked the error rate of different high accuracy deep sequencing methods. Then I analyzed the low-frequency mutations in various clinical samples.

a. **Zhang, T. H.,** Wu, N. C., & Sun, R. (2016). A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing. *BMC genomics*, *17*(1), 1-9.

A full list of publication can be found in:

https://scholar.google.com/citations?user=PdCr1SUAAAAJ&hl=en

Chapter 1

Introduction

1. Development of the viral genetic barcodes

In the perspective of reductionism, the advancement of biological research was accompanied by the improvement of tracing single elements in a complex system. Discovery of GFP enabled real time tracking of single cells in complex tissues[1]. DNA sequencing methods identified mutations that emerged and fixed in evolutionary history, reconstructed the tree of life[2]. Molecular cloning and genome editing explained the importance of genetic elements by specifically disturbing their functions and observing phenotypes of the mutants[3].

Genetic barcodes refer to a series of technologies that can use a unique piece of genetic material to track the life history of a certain individual. For example, Schepers et al. used lentivirus to deliver a set of semi-random nucleotide barcodes into the T cell genomes to trace their lineage in mice[4]. Gerrits et al. inserted a random nucleotide sequence into hematopoietic stem cells' genomes to study their development[5]. These pioneer works inspired many applications of random nucleotide genetic barcodes in various systems [6-10]. Other genetic information can also be used as a barcode to track cell lineages. The VDJ region mutates during lymphocyte development and can be used to study its lineage[11]. Spontaneous mutations can be used to infer phylogenetic trees of bacteria, virus and cancer cell evolution [12-14]. The integration sites of retroviruses and transposons are consistent during development and can trace the cell lineages [15-18]. The inducible CRISPR mutagenesis can generate insertion and deletion barcodes in situ[22,23]. Recent years witnessed the development of using transcriptomic signatures to infer developmental history [19-21]. A plethora of methods tracked the life history of cells in an organism or an individual in a population.

Viral genetic barcode is a specific application of the genetic barcode method, but it also has many unique properties due to the special properties of the virus life history. Firstly, viruses have a large population size compared to the relatively small population of stem cells and lymphocytes. An acutely infected patient of Influenza A virus or SARS-CoV-2 can have 10 million to 100 billion virions *in vivo* [24,25]. Even a latently infected HIV-1 patient can have millions of dormant proviruses in different organs [26,27]. A genetic barcode with low information density will not be able to track the dynamics of such large populations. Secondly, the mutation rate of viruses is several magnitudes higher than that of prokaryotic or eukaryotic organisms. RNA viruses and retroviruses, in particular, create ~1 substitution mutation on its genome every time they replicate [28-30]. Viruses with multiple copies of genetic material easily recombine within the host cells [35-37]. On the one hand, these activities create an abundance of genetic information that can be used to reconstruct the phylogenetic history. On the other hand, they disturb the original genetic barcodes, posing challenges to accurate long-time tracking of a viral lineage. Moreover, unlike haploid or diploid organisms, viral particles are completely dismantled after infecting host cells, while thousands of progenies are produced from a single copy of the parental genetic material [31,32]. The mathematical language to describe the development of cellular lineage needs to be reconsidered for most viruses [33,34].

In recent years, we and other groups have developed robust and convenient methodology of viral genetic barcodes. Early methods like TRIP (Thousands of Reporters Integrated in Parallel) or B-HIVE (Barcoded HIV ensembles) inserted random genetic barcodes into the retroviral genomes, and established cellular clones to study the transcription activity of proviral DNA on different genomic positions [38,39]. The workflows and the library scales

are similar to cellular barcodes and only characterized a single step of the viral life cycle, but they established the concept of viral genetic barcodes and provided some analysis tools for it. Later, different groups constructed replication competent barcoded virus libraries on HIV-1, SIV, and Influenza A virus [40-43]. At the same time, the advancement of high throughput sequencing and related algorithms enabled rapid sequencing, mapping, clustering and counting of barcode data [44-48]. Because of the high complexity and high mutation rate of the viral genetic barcodes, specific tools and pipelines to reduce sequencing errors and provide accurate barcode calling were invented [49-51]. At the same time, mathematical frameworks for barcode designs were developed [52-54]. Modern barcode libraries have been optimized to tolerate mutations and sequencing errors, while not interfering with the biological functions of the virus. All these inventions and improvements brought us to a point where we could apply the viral genetic barcodes to solve many biological problems.

2. Applications of the viral genetic barcodes

2.1. Quantify the latent reservoir

The biggest barrier to HIV-1 cure is the persistent latent reservoir. The virus integrates into the host genome and can stay dormant for decades before rebound. The latent cells are transcriptionally and translationally identical to uninfected cells without external stimulation. The future therapeutic methods for HIV-1 cure rely on accurately monitoring the size of the latent reservoir [55]. Various techniques investigate the size of the latent reservoir. Quantitative viral outgrowth assay (QVOA) is the most used reference for the replication competent reservoir [56,57]. It measures the number of viruses reactivated after *in vitro* T cell activation. But it often underestimates the latent reservoir size because no

drug can activate all infected T cells. Near full-length proviral amplification and intact proviral DNA assay (IPDA) infers the virus replication capacity by its sequence [58,59]. But these methods tend to overestimate the replication competent provirus because many mutations have unknown functions and may lead to a defective virus. The latent reservoir size can also be estimated by sequencing the spontaneous mutations on the viral genome and constructing the phylogenetic tree [60-62]. But it is affected heavily by sampling errors and cannot accurately calculate the reservoir size.

We have developed a genetically barcoded HIV-1 library that can actively replicate in the humanized mice model and undergo latency [43,66]. By sequencing the barcoded virion RNA, we accurately quantified the number of latent viral clones that can be reactivated after certain treatment. The library can be used as a benchmark to evaluate many experimental therapies targeting the latency reservoir, e.g. the latency reversal agents or latency promoting agents [67,68].

2.2. Describe the characteristics of the latent reservoir

The size of HIV-1 latent reservoir within viral suppressed patients is mainly maintained by T cell clonal expansion [69,70]. Describing the development history of the latent reservoir and characterizing the phenotypic features of clonally expanded host T cells help us better target infected cells and design therapies. Because the cells harboring a proviral clone share the same proviral integration site, sequencing the sites and infer the abundance of each site help characterize the population structure of the latent reservoir [63-65]. However, integration site sequencing alone cannot distinguish replication competent provirus with the defective proviruses, thus overestimates the latent reservoir size. Many multi-omics methods were developed to characterize the clonal expansion and virus

reactivation in the latent reservoir. Simultaneous TCR, Integration site and Provirus sequencing (STIP-seq) used flow cytometry to isolate reactivated provirus and sequenced the integration site with virus mutations [71]. It identified the virus production from proviruses integrated near cancer related genes. But it relies on *ex vivo* reactivation of the proviruses and cannot characterize the cellular status *in vivo*. HIV-1 SortSeq using viral RNA probes isolated latently infected cells and profiled their transcriptional features [72]. It is also restricted by *ex vivo* reactivation. Expanded CRISPR-compatible cellular indexing of transcriptomes and epitopes sequencing (ECCITE-seq) reveals the transcriptomic features of latently infected cells and trace the clonal dynamics using the lymphocyte VDJ sequence [73,74]. But it needs to be further optimized for HIV-1 to capture low abundance viral RNA or proviral DNA. Parallel HIV-1 RNA, integration site, and proviral sequencing (PRIP-seq) used microwell and multiple displacement genome amplification to sequence integration sites and cellular transcription features simultaneously [75]. But the relatively low throughput restricts its application in many scenarios. Focused interrogation of cells by nucleic acid detection and sequencing (FIND-seq) used hydrogel microfluidics and nucleic acid cytometry to measure latently infected cells with their transcriptome at unperturbed status [77,78]. They identified that clonally expanded latent cells usually have a feature of suppressed virus transcription. Phenotypic and proviral sequencing (PheP-seq) achieved a similar conclusion by integrating the multi-omics data from the same patients [76].

In this thesis, we will present a more convenient and high-throughput method using viral genetic barcodes in a humanized mice model to study the features of clonally expanded latent cells.

2.3. Quantify population bottlenecks

The disease severity and prognosis of many viruses is affected by the initial amount of infection [79-81]. For latently infected viruses, the amount of viral load set point when disease progresses to latency will determine the length of the latency period [82]. It is important to estimate the population size of virus life history, especially at these bottleneck stages [83]. Many phylogenetic methods have been proposed to infer the population history and estimate the bottleneck size. But genetic barcodes provide a straightforward quantification of virus population bottlenecks during transmission, tissue dissemination, latency establishment and other processes [40,42,66]. For a uniformly distributed barcode library, the number of barcodes observed after the bottleneck event indicates the number of individuals that survived through the bottleneck. Moreover, many experimental therapies, such as broad neutralizing antibodies or vaccines that elicit tissue resident T cells, can reduce the infection bottleneck size and protect at the site of infection [86-88]. Barcoded viruses can benchmark the efficiency of these therapies.

2.4. Label mutations

Viral genetic barcodes can also be used outside of complex organisms, just serving as a powerful tool to label single virions. Many studies have used genetic barcodes to label different mutants, facilitating the sequencing analysis [89,90]. On many occasions, the mutation itself can serve as a genetic barcode, and can directly be associated with phenotypes [91,92]. But an intentionally designed long genetic barcode can help in the situations where mutants are too long to be sequenced or the screening process may disturb the original mutations.

3. High throughput fitness profiling

Understanding the clonal history can be complicated in the context of the ever-changing nature of the virus genetic materials. In the real world, new viral clones continuously emerge from the existing clones [28-30]. For most populations, the mutation rate and the adaptation process has achieved an equilibrium where most viruses can never reach the genotype that fits the environment perfectly [93]. In some scenarios, the number of spontaneous deleterious mutations outweighs the natural selection process, leading to a population crush [94]. Virus population genetics only makes sense in the light of understanding the functions of the mutations.

The functions of the viral mutations can be generalized to the ability to adapt to various environments. The ability of adaptation is a fundamental concept in evolutionary biology, termed fitness [33]. A genotype has higher fitness means it can adapt to the environment better. All genotypes and their fitness form a hyperspace called the fitness landscape. Like a real landscape, genotypes march on evolutionary paths by accumulating mutations, and travel across valleys or hills of lower or higher fitness [95].

The fitness profiling is a method that can efficiently characterize the whole fitness landscape [96,97]. It pools a library of viral mutants and lets them compete in certain conditions. The frequency change after the competition was quantified by the next generation sequencing. And the fitness of all mutations is inferred from the frequency change.

Fitness profiling data can be used to explain the evolutionary history of virus populations, or to identify new functions of known proteins [90,98-102]. Here I will introduce two of its applications involved in our work.

4. Applications of the high throughput fitness profiling

4.1. Quantify genetic interactions among mutations

Evolutionary pathways are not simply accumulating mutations. They are restricted by the topography of the fitness landscape [104]. Positive genetic interactions rescue the deleterious mutations, help viruses to gain new mutations [105]. While negative genetic interactions create barriers to acquiring new mutations [106]. Previous works have observed these constraints at the molecular level by constructing individual mutants [101,103], or inferring from phylogenetic data[110]. Recently, people used deep mutational scanning to profile the local genetic interactions in GFP, tRNA and other proteins [107-109]. Our group constructed combinatory mutant libraries on HIV-1 protease and characterized the high order interactions among drug resistant mutations [91]. We found the intensity of positive genetic interactions increased with the number of drug resistant mutations accumulated. This indicates HIV-1 protease has a rugged fitness landscape on drug resistant mutations, stabilizing the genotypes even in the absence of protease inhibitors. Other groups also used the high throughput fitness profiling platform to characterize positive or negative interactions on other parts of the HIV-1 genome or on other viruses [110-113].

4.2. Surveil evolutionary history

In human history, new viruses emerged frequently and left significant impacts on society [116]. Understanding the evolutionary history of the viruses help us prevent future cross-species spillover and predict possible new strains that are evolving in human or animal populations. Phylogeny trees and molecular clocks are well-established methods to infer the viral evolutionary history [114,115]. They are getting more accurate with the plethora of sequencing data available during the global SARS-CoV-2 pandemic [117,118]. However,

these methods cannot explain the function of emerging mutations, thus falling short in predicting the trend of evolution.

High throughput fitness profiling can annotate a large amount of epidemiological sequencing data, explaining the function of new mutations [90]. We constructed a SARS-CoV-2 nucleocapsid mutant library that covers all possible single amino acid substitutions. We quantified the stability of all mutants and annotated the public viral sequencing database. Our findings showed SARS-CoV-2 is gradually increasing its stability during the early transmission period in the human species. Other groups also used the similar method to explain the evolutionary history of SARS-CoV-2 [119,120].

5. References

1. Valdivia, R. H., Hromockyj, A. E., Monack, D., Ramakrishnan, L., & Falkow, S. (1996). Applications for green fluorescent protein (GFP) in the study of host pathogen interactions. *Gene*, *173*(1), 47-52.

2. Fitch, W. M., & Margoliash, E. (1967). Construction of phylogenetic trees: a method based on mutation distances as estimated from cytochrome c sequences is of general applicability. *Science*, *155*(3760), 279-284.

3. Orkin, S. H. (1986). Reverse genetics and human disease. *Cell*, *47*(6), 845-850.

4. Schepers, K., Swart, E., van Heijst, J. W., Gerlach, C., Castrucci, M., Sie, D., ... & Schumacher, T. N. (2008). Dissecting T cell lineage relationships by cellular barcoding. *The Journal of experimental medicine*, *205*(10), 2309-2318.

5. Gerrits, A., Dykstra, B., Kalmykowa, O. J., Klauke, K., Verovskaya, E., Broekhuis, M. J., ... & Bystrykh, L. V. (2010). Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood, The Journal of the American Society of Hematology*, *115*(13), 2610-2618.

6. Lu, R., Neff, N. F., Quake, S. R., & Weissman, I. L. (2011). Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nature biotechnology*, *29*(10), 928-933.

7. Suryawanshi, G. W., Arokium, H., Kim, S., Khamaikawin, W., Lin, S., Shimizu, S., ... & Chen, I. S. (2021). Longitudinal clonal tracking in humanized mice reveals sustained polyclonal repopulation of gene-modified human-HSPC despite vector integration bias. *Stem cell research & therapy*, *12*(1), 1-20.

8. Wu, C., Li, B., Lu, R., Koelle, S. J., Yang, Y., Jares, A., ... & Dunbar, C. E. (2014). Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. *Cell Stem Cell*, *14*(4), 486-499.

9. Verovskaya, E., Broekhuis, M. J., Zwart, E., Ritsema, M., van Os, R., de Haan, G., & Bystrykh, L. V. (2013). Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. *Blood, The Journal of the American Society of Hematology*, *122*(4), 523-532.

10. Verovskaya, E., Broekhuis, M. J., Zwart, E., Weersing, E., Ritsema, M., Bosman, L. J., ... & Bystrykh, L. V. (2014). Asymmetry in skeletal distribution of mouse hematopoietic stem cell clones and their equilibration by mobilizing cytokines. *Journal of Experimental Medicine*, *211*(3), 487-497.

11. Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., ... & Fire, A. Z. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science translational medicine*, *1*(12), 12ra23-12ra23.

12. Levy, S. F., Blundell, J. R., Venkataram, S., Petrov, D. A., Fisher, D. S., & Sherlock, G. (2015). Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, *519*(7542), 181-186.

13. Blundell, J. R., & Levy, S. F. (2014). Beyond genome sequencing: lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer. *Genomics*, *104*(6), 417-430.

14. Flynn, J. M., Chain, F. J., Schoen, D. J., & Cristescu, M. E. (2017). Spontaneous mutation accumulation in Daphnia pulex in selection-free vs. competitive environments. *Molecular Biology and Evolution*, *34*(1), 160-173.

15. Cornils, K., Bartholomae, C. C., Thielecke, L., Lange, C., Arens, A., Glauche, I., ... & Fehse, B. (2013). Comparative clonal analysis of reconstitution kinetics after transplantation of hematopoietic stem cells gene marked with a lentiviral SIN or a γ-retroviral LTR vector. *Experimental hematology*, *41*(1), 28-38.

16. Brugman, M. H., Suerth, J. D., Rothe, M., Suerbaum, S., Schambach, A., Modlich, U., ... & Baum, C. (2013). Evaluating a ligation-mediated PCR and pyrosequencing method for the detection of clonal contribution in polyclonal retrovirally transduced samples. *Human Gene Therapy Methods*, *24*(2), 68-79.

17. Biasco, L., Pellin, D., Scala, S., Dionisio, F., Basso-Ricci, L., Leonardelli, L., ... & Aiuti, A. (2016). In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases. *Cell stem cell*, *19*(1), 107-119.

18. Sun, J., Ramos, A., Chapman, B., Johnnidis, J. B., Le, L., Ho, Y. J., ... & Camargo, F. D. (2014). Clonal dynamics of native haematopoiesis. *Nature*, *514*(7522), 322-327.

19. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., ... & Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, *560*(7719), 494-498.

20. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., ... & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, *32*(4), 381-386.

21. Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., ... & Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, *19*, 1-16.

22. Raj, B., Gagnon, J. A., & Schier, A. F. (2018). Large-scale reconstruction of cell lineages using single-cell readout of transcriptomes and CRISPR–Cas9 barcodes by scGESTALT. *Nature protocols*, *13*(11), 2685-2713.

23. Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., & Junker, J. P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nature biotechnology*, *36*(5), 469-473.

24. Sender, R., Bar-On, Y. M., Gleizer, S., Bernshtein, B., Flamholz, A., Phillips, R., & Milo, R. (2021). The total number and mass of SARS-CoV-2 virions. *Proceedings of the National Academy of Sciences*, *118*(25), e2024815118.

25. Lumby, C. K., Zhao, L., Breuer, J., & Illingworth, C. J. (2020). A large effective population size for established within-host influenza virus infection. *Elife*, *9*, e56915.

26. Hodel, F., Patxot, M., Snäkä, T., & Ciuffi, A. (2016). HIV-1 latent reservoir: size matters. *Future virology*, *11*(12), 785-794.

27. Abdel-Mohsen, M., Richman, D., Siliciano, R. F., Nussenzweig, M. C., Howell, B. J., Martinez-Picado, J., ... & Montaner, L. J. (2020). Recommendations for measuring HIV reservoir size in cure-directed clinical trials. *Nature medicine*, *26*(9), 1339-1350.

28. Cuevas, J. M., Geller, R., Garijo, R., López-Aldeguer, J., & Sanjuán, R. (2015). Extremely high mutation rate of HIV-1 in vivo. *PLoS biology*, *13*(9), e1002251.

29. Ribeiro, R. M., Li, H., Wang, S., Stoddard, M. B., Learn, G. H., Korber, B. T., ... & Perelson, A. S. (2012). Quantifying the diversification of hepatitis C virus (HCV) during primary infection: estimates of the in vivo mutation rate.

30. Nobusawa, E., & Sato, K. (2006). Comparison of the mutation rates of human influenza A and B viruses. *Journal of virology*, *80*(7), 3675-3678.

31. Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., & Ho, D. D. (1996). HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, *271*(5255), 1582-1586.

32. Kissler, S. M., Fauver, J. R., Mack, C., Olesen, S. W., Tai, C., Shiue, K. Y., ... & Grad, Y. H. (2021). Viral dynamics of acute SARS-CoV-2 infection and applications to diagnostic and public health strategies. *PLoS biology*, *19*(7), e3001333.

33. Fisher, R. A. (1999). *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press.

34. Nowak, M., & May, R. M. (2000). *Virus dynamics: mathematical principles of immunology and virology: mathematical principles of immunology and virology*. Oxford University Press, UK.

35. Neher, R. A., & Leitner, T. (2010). Recombination rate and selection strength in HIV intra-patient evolution. *PLoS computational biology*, *6*(1), e1000660.

36. Zhou, Y., & Holmes, E. C. (2007). Bayesian estimates of the evolutionary rate and age of hepatitis B virus. *Journal of molecular evolution*, *65*, 197-205.

37. Froissart, R., Roze, D., Uzest, M., Galibert, L., Blanc, S., & Michalakis, Y. (2005). Recombination every day: abundant recombination in a virus during a single multi-cellular host infection. *PLoS biology*, *3*(3), e89.

38. Akhtar, W., de Jong, J., Pindyurin, A. V., Pagie, L., Meuleman, W., de Ridder, J., ... & van Steensel, B. (2013). Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell*, *154*(4), 914-927.

39. Chen, H. C., Martinez, J. P., Zorita, E., Meyerhans, A., & Filion, G. J. (2017). Position effects influence HIV latency reversal. *Nature structural & molecular biology*, *24*(1), 47-54.

40. Varble, A., Albrecht, R. A., Backes, S., Crumiller, M., Bouvier, N. M., Sachs, D., & García-Sastre, A. (2014). Influenza A virus transmission bottlenecks are defined by infection route and recipient host. *Cell host & microbe*, *16*(5), 691-700.

41. Fennessey, C. M., Pinkevych, M., Immonen, T. T., Reynaldi, A., Venturi, V., Nadella, P., ... & Keele, B. F. (2017). Genetically-barcoded SIV facilitates enumeration of rebound variants and estimation of reactivation rates in nonhuman primates following interruption of suppressive antiretroviral therapy. *PLoS pathogens*, *13*(5), e1006359.

42. Amato, K. A., Haddock III, L. A., Braun, K. M., Meliopoulos, V., Livingston, B., Honce, R., ... & Mehle, A. (2022). Influenza A virus undergoes compartmentalized replication in vivo dominated by stochastic bottlenecks. *Nature Communications*, *13*(1), 3416.

43. Marsden, M. D., Zhang, T. H., Du, Y., Dimapasoc, M., Soliman, M. S., Wu, X., ... & Zack, J. A. (2020). Tracking HIV rebound following latency reversal using barcoded HIV. *Cell Reports Medicine*, *1*(9), 100162.

44. Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular cell*, *58*(4), 586-597.

45. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.

46. Langdon, W. B. (2015). Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData mining*, *8*(1), 1-7.

47. Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150-3152.

48. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272.

49. Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research*, *27*(3), 491-499.

50. Blois, S., Goetz, B. M., Bull, J. J., & Sullivan, C. S. (2022). Interpreting and de-noising genetically engineered barcodes in a DNA virus. *PLOS Computational Biology*, *18*(11), e1010131.

51. Zhang, T. H., Wu, N. C., & Sun, R. (2016). A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing. *BMC genomics*, *17*(1), 1-9.

52. Thielecke, L., Aranyossy, T., Dahl, A., Tiwari, R., Roeder, I., Geiger, H., ... & Cornils, K. (2017). Limitations and challenges of genetic barcode quantification. *Scientific Reports*, *7*(1), 43249.

53. Xu, Q., Schlabach, M. R., Hannon, G. J., & Elledge, S. J. (2009). Design of 240,000 orthogonal 25mer DNA barcode probes. *Proceedings of the National Academy of Sciences*, *106*(7), 2289-2294.

54. Cornils, K., Thielecke, L., Hüser, S., Forgber, M., Thomaschewski, M., Kleist, N., ... & Fehse, B. (2014). Multiplexing clonality: combining RGB marking and genetic barcoding. *Nucleic acids research*, *42*(7), e56-e56.

55. Sengupta, S., & Siliciano, R. F. (2018). Targeting the latent reservoir for HIV-1. *Immunity*, *48*(5), 872-895.

56. Wong, J. K., Hezareh, M., Gunthard, H. F., Havlir, D. V., Ignacio, C. C., Spina, C. A., & Richman, D. D. (1997). Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science*, *278*(5341), 1291-1295.

57. Finzi, D., Hermankova, M., Pierson, T., Carruth, L. M., Buck, C., Chaisson, R. E., ... & Siliciano, R. F. (1997). Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science*, *278*(5341), 1295-1300.

58. Bruner, K. M., Murray, A. J., Pollack, R. A., Soliman, M. G., Laskey, S. B., Capoferri, A. A., ... & Siliciano, R. F. (2016). Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nature medicine*, *22*(9), 1043-1049.

59. Ho, Y. C., Shan, L., Hosmane, N. N., Wang, J., Laskey, S. B., Rosenbloom, D. I., ... & Siliciano, R. F. (2013). Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell*, *155*(3), 540-551.

60. Bandera, A., Gori, A., Clerici, M., & Sironi, M. (2019). Phylogenies in ART: HIV reservoirs, HIV latency and drug resistance. *Current opinion in pharmacology*, *48*, 24-32.

61. Mok, H. P., Norton, N. J., Hirst, J. C., Fun, A., Bandara, M., Wills, M. R., & Lever, A. M. (2018). No evidence of ongoing evolution in replication competent latent HIV-1 in a patient followed up for two years. *Scientific Reports*, *8*(1), 2639.

62. Brodin, J., Zanini, F., Thebo, L., Lanz, C., Bratt, G., Neher, R. A., & Albert, J. (2016). Establishment and stability of the latent HIV-1 DNA reservoir. *Elife*, *5*, e18889.

63. Sherrill-Mix, S., Lewinski, M. K., Famiglietti, M., Bosque, A., Malani, N., Ocwieja, K. E., ... & Bushman, F. D. (2013). HIV latency and integration site placement in five cell-based models. *Retrovirology*, *10*, 1-14.

64. Schröder, A. R., Shinn, P., Chen, H., Berry, C., Ecker, J. R., & Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, *110*(4), 521-529.

65. Cohn, L. B., Silva, I. T., Oliveira, T. Y., Rosales, R. A., Parrish, E. H., Learn, G. H., ... & Nussenzweig, M. C. (2015). HIV-1 integration landscape during latent and active infection. *Cell*, *160*(3), 420-432.

66. Kim, J. T., Zhang, T. H., Carmona, C., Lee, B., Seet, C. S., Kostelny, M., ... & Zack, J. A. (2022). Latency reversal plus natural killer cells diminish HIV reservoir in vivo. *Nature communications*, *13*(1), 121.

67. Bashiri, K., Rezaei, N., Nasi, M., & Cossarizza, A. (2018). The role of latency reversal agents in the cure of HIV: a review of current data. *Immunology letters*, *196*, 135-139.

68. Ahlenstiel, C. L., Symonds, G., Kent, S. J., & Kelleher, A. D. (2020). Block and lock HIV cure strategies to control the latent reservoir. *Frontiers in Cellular and Infection Microbiology*, *10*, 424.

69. Yeh, Yang-Hui Jimmy, et al. "The clonal expansion dynamics of the hiv-1 reservoir: Mechanisms of integration site-dependent proliferation and hiv-1 persistence." *Viruses* 13.9 (2021): 1858.

70. Liu, R., Simonetti, F. R., & Ho, Y. C. (2020). The forces driving clonal expansion of the HIV-1 latent reservoir. *Virology journal*, *17*(1), 1-13.

71. Cole, B., Lambrechts, L., Gantner, P., Noppe, Y., Bonine, N., Witkowski, W., ... & Vandekerckhove, L. (2021). In-depth single-cell analysis of translation-competent HIV-1 reservoirs identifies cellular sources of plasma viremia. *Nature Communications*, *12*(1), 3727.

72. Liu, R., Yeh, Y. H. J., Varabyou, A., Collora, J. A., Sherrill-Mix, S., Talbot Jr, C. C., ... & Ho, Y. C. (2020). Single-cell transcriptional landscapes reveal HIV-1–driven aberrant host gene transcription as a potential therapeutic target. *Science translational medicine*, *12*(543), eaaz0802.

73. Collora, J. A., Liu, R., Pinto-Santini, D., Ravindra, N., Ganoza, C., Lama, J. R., ... & Ho, Y. C. (2022). Single-cell multiomics reveals persistence of HIV-1 in expanded cytotoxic T cell clones. *Immunity*, *55*(6), 1013-1031.

74. Mimitou, E. P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., ... & Smibert, P. (2019). Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature methods*, *16*(5), 409-412.

75. Einkauf, K. B., Osborn, M. R., Gao, C., Sun, W., Sun, X., Lian, X., ... & Lichterfeld, M. (2022). Parallel analysis of transcription, integration, and sequence of single HIV-1 proviruses. *Cell*, *185*(2), 266-282.

76. Sun, W., Gao, C., Hartana, C. A., Osborn, M. R., Einkauf, K. B., Lian, X., ... & Lichterfeld, M. (2023). Phenotypic signatures of immune selection in HIV-1 reservoir cells. *Nature*, 1-9.

77. Clark, I. C., Mudvari, P., Thaploo, S., Smith, S., Abu-Laban, M., Hamouda, M., ... & Boritz, E. A. (2023). HIV silencing and cell survival signatures in infected T cell reservoirs. *Nature*, 1-8.

78. Clark, I. C., Wheeler, M. A., Lee, H. G., Li, Z., Sanmarco, L. M., Thaploo, S., ... & Abate, A. R. (2023). Identification of astrocyte regulators by nucleic acid cytometry. *Nature*, 1-3.

79. Khosroshahi, H. T., & Mardomi, A. (2021). The initial infectious dose of SARS-CoV-2 and the severity of the disease: possible impact on the incubation period. *Future Virology*, *16*(5), 369-373.

80. Marois, I., Cloutier, A., Garneau, É., & Richter, M. V. (2012). Initial infectious dose dictates the innate, adaptive, and memory responses to influenza in the respiratory tract. *Journal of leukocyte biology*, *92*(1), 107-121.

81. Ward, R. L., Bernstein, D. I., Young, E. C., Sherwood, J. R., Knowlton, D. R., & Schiff, G. M. (1986). Human rotavirus studies in volunteers: determination of infectious dose and serological response to infection. *Journal of Infectious Diseases*, *154*(5), 871-880.

82. Fraser, C., Hollingsworth, T. D., Chapman, R., de Wolf, F., & Hanage, W. P. (2007). Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. *Proceedings of the National Academy of Sciences*, *104*(44), 17441-17446.

83. Kariuki, S. M., Selhorst, P., Ariën, K. K., & Dorfman, J. R. (2017). The HIV-1 transmission bottleneck. *Retrovirology*, *14*, 1-19.

84. Sagulenko, P., Puller, V., & Neher, R. A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus evolution*, *4*(1), vex042.

85. Lundgren, E., Romero-Severson, E., Albert, J., & Leitner, T. (2022). Combining biomarker and virus phylogenetic models improves HIV-1 epidemiological source identification. *PLOS Computational Biology*, *18*(8), e1009741.

86. Julg, B., Tartaglia, L. J., Keele, B. F., Wagh, K., Pegu, A., Sok, D., ... & Barouch, D. H. (2017). Broadly neutralizing antibodies targeting the HIV-1 envelope V2 apex confer protection against a clade C SHIV challenge. *Science translational medicine*, *9*(406), eaal1321.

87. Gardner, M. R., Fellinger, C. H., Kattenhorn, L. M., Davis-Gardner, M. E., Weber, J. A., Alfant, B., ... & Farzan, M. (2019). AAV-delivered eCD4-Ig protects rhesus macaques from high-dose SIVmac239 challenges. *Science translational medicine*, *11*(502), eaau5409.

88. Hansen, S. G., Marshall, E. E., Malouli, D., Ventura, A. B., Hughes, C. M., Ainslie, E., ... & Picker, L. J. (2019). A live-attenuated RhCMV/SIV vaccine shows long-term efficacy against heterologous SIV challenge. *Science translational medicine*, *11*(501), eaaw2607.

89. Muñoz-Moreno, R., Martínez-Romero, C., Blanco-Melo, D., Forst, C. V., Nachbagauer, R., Benitez, A. A., ... & García-Sastre, A. (2019). Viral fitness landscapes in diverse host species reveal multiple evolutionary lines for the NS1 gene of influenza A viruses. *Cell reports*, *29*(12), 3997-4009.

90. Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H., Dingens, A. S., ... & Bloom, J. D. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *cell*, *182*(5), 1295-1310.

91. Zhang, T. H., Dai, L., Barton, J. P., Du, Y., Tan, Y., Pang, W., ... & Sun, R. (2020). Predominance of positive epistasis among drug resistance-associated mutations in HIV-1 protease. *PLoS genetics*, *16*(10), e1009009.

92. Du, Y., Zhang, T. H., Dai, L., Zheng, X., Gorin, A. M., Oishi, J., ... & Sun, R. (2017). Effects of mutations on replicative fitness and major histocompatibility complex class I binding affinity are among the determinants underlying cytotoxic-T-lymphocyte escape of HIV-1 gag epitopes. *MBio*, *8*(6), e01050-17.

93. Haigh, J. (1978). The accumulation of deleterious genes in a population—Muller's ratchet. *Theoretical population biology*, *14*(2), 251-267.

94. Chao, L. (1990). Fitness of RNA virus decreased by Muller's ratchet. *Nature*, *348*(6300), 454-455.

95. Pitzer, E., & Affenzeller, M. (2012). A comprehensive survey on fitness landscape analysis. *Recent advances in intelligent engineering systems*, 161-191.

96. Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nature methods*, *11*(8), 801-807.

97. Wu, N. C., Young, A. P., Al-Mawsawi, L. Q., Olson, C. A., Feng, J., Qi, H., ... & Sun, R. (2014). High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. *Scientific reports*, *4*(1), 1-8.

98. Wu, N. C., Young, A. P., Al-Mawsawi, L. Q., Olson, C. A., Feng, J., Qi, H., ... & Sun, R. (2014). High-throughput identification of loss-of-function mutations for anti-interferon activity in the influenza A virus NS segment. *Journal of virology*, *88*(17), 10157-10164.

99. Wu, N. C., Du, Y., Le, S., Young, A. P., Zhang, T. H., Wang, Y., ... & Sun, R. (2016). Coupling high-throughput genetics with phylogenetic information reveals an epistatic interaction on the influenza A virus M segment. *BMC genomics*, *17*(1), 1-15.

100. Hanning, K. R., Minot, M., Warrender, A. K., Kelton, W., & Reddy, S. T. (2022). Deep mutational scanning for therapeutic antibody engineering. *Trends in Pharmacological Sciences*, *43*(2), 123-135.

101. Starr, T. N., & Thornton, J. W. (2016). Epistasis in protein evolution. *Protein science*, *25*(7), 1204-1218.

102. Du, Y., Xin, L., Shi, Y., Zhang, T. H., Wu, N. C., Dai, L., ... & Sun, R. (2018). Genome-wide identification of interferon-sensitive mutations enables influenza vaccine design. *Science*, *359*(6373), 290-296.

103. Ortlund, E. A., Bridgham, J. T., Redinbo, M. R., & Thornton, J. W. (2007). Crystal structure of an ancient protein: evolution by conformational epistasis. *science*, *317*(5844), 1544-1548.

104. Fragata, I., Blanckaert, A., Louro, M. A. D., Liberles, D. A., & Bank, C. (2019). Evolution in the light of fitness landscape theory. *Trends in ecology & evolution*, *34*(1), 69-82.

105. Trindade, S., Sousa, A., Xavier, K. B., Dionisio, F., Ferreira, M. G., & Gordo, I. (2009). Positive epistasis drives the acquisition of multidrug resistance. *PLoS genetics*, *5*(7), e1000578.

106. Bank, C., Hietpas, R. T., Jensen, J. D., & Bolon, D. N. (2015). A systematic survey of an intragenic epistatic landscape. *Molecular biology and evolution*, *32*(1), 229-238.

107. Olson, C. A., Wu, N. C., & Sun, R. (2014). A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current biology*, *24*(22), 2643-2651.

108. Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., ... & Kondrashov, F. A. (2016). Local fitness landscape of the green fluorescent protein. *Nature*, *533*(7603), 397-401.

109. Li, C., Qian, W., Maclean, C. J., & Zhang, J. (2016). The fitness landscape of a tRNA gene. *Science*, *352*(6287), 837-840.

110. Lyons, D. M., & Lauring, A. S. (2018). Mutation and epistasis in influenza virus evolution. *Viruses*, *10*(8), 407.

111. Sanjuán, R., Cuevas, J. M., Moya, A., & Elena, S. F. (2005). Epistasis and the adaptability of an RNA virus. *Genetics*, *170*(3), 1001-1008.

112. Fernandes, J. D., Faust, T. B., Strauli, N. B., Smith, C., Crosby, D. C., Nakamura, R. L., ... & Frankel, A. D. (2016). Functional segregation of overlapping genes in HIV. *Cell*, *167*(7), 1762-1773.

113. Gordon, D. E., Watson, A., Roguev, A., Zheng, S., Jang, G. M., Kane, J., ... & Krogan, N. J. (2020). A quantitative genetic interaction map of HIV infection. *Molecular cell*, *78*(2), 197-209.

114. Kannan, S. K., & Warnow, T. J. (1994). Inferring evolutionary history from DNA sequences. *SIAM Journal on Computing*, *23*(4), 713-737.

115. Ho, S. Y. (2014). The changing face of the molecular evolutionary clock. *Trends in Ecology & Evolution*, *29*(9), 496-503.

116. Wasik, B. R., de Wit, E., Munster, V., Lloyd-Smith, J. O., Martinez-Sobrido, L., & Parrish, C. R. (2019). Onward transmission of viruses: how do viruses emerge to cause epidemics after spillover?. *Philosophical Transactions of the Royal Society B*, *374*(1782), 20190017.

117. Zelenova, M., Ivanova, A., Semyonov, S., & Gankin, Y. (2021). Analysis of 329,942 SARS-CoV-2 records retrieved from GISAID database. *Computers in Biology and Medicine*, *139*, 104981.

118. McBroome, J., Thornlow, B., Hinrichs, A. S., Kramer, A., De Maio, N., Goldman, N., ... & Turakhia, Y. (2021). A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. *Molecular biology and evolution*, *38*(12), 5819-5824.

119. Ouyang, W. O., Tan, T. J., Lei, R., Song, G., Kieffer, C., Andrabi, R., ... & Wu, N. C. (2022). Probing the biophysical constraints of SARS-CoV-2 spike N-terminal domain using deep mutational scanning. *Science Advances*, *8*(47), eadd7221.

120. Frank, F., Keen, M. M., Rao, A., Bassit, L., Liu, X., Bowers, H. B., ... & Ortlund, E. A. (2022). Deep mutational scanning identifies SARS-CoV-2 Nucleocapsid escape mutations of currently available rapid antigen tests. *Cell*, *185*(19), 3603-3616.

Chapter 2

Predominance of positive epistasis among resistance-associated mutations in HIV-1 protease

1. Abstract

Drug-resistant mutations often have deleterious impacts on replication fitness, posing a fitness cost that can only be overcome by compensatory mutations. However, the role of fitness cost in the evolution of drug resistance has often been overlooked in clinical studies or *in vitro* selection experiments, as these observations only capture the outcome of drug selection. In this study, we systematically profile the fitness landscape of resistance-associated sites in HIV-1 protease using deep mutational scanning. We construct a mutant library covering combinations of mutations at 11 sites in HIV-1 protease, all of which are associated with resistance to protease inhibitors in clinic. Using deep sequencing, we quantify the fitness of thousands of HIV-1 protease mutants after multiple cycles of replication in human T cells. Although the majority of resistance-associated mutations have deleterious effects on viral replication, we find that epistasis among resistance-associated mutations is predominantly positive. Furthermore, our fitness data are consistent with genetic interactions inferred directly from HIV sequence data of patients. Fitness valleys formed by strong positive epistasis reduce the likelihood of reversal of drug resistance mutations. Overall, our results support the view that strong compensatory effects are involved in the emergence of clinically observed resistance mutations and provide insights to understanding fitness barriers in the evolution and reversion of drug resistance.

2. Author Summary

Antiretroviral drugs have achieved great success in controlling the HIV pandemic. However, the therapy fails sometimes owing to the low drug adherence and/or the

emergence of resistance associated mutations on viral genome. The persistence of drug resistance poses challenges in using antiretroviral drugs for long term control or pre-exposure prophylaxis. To understand the mechanisms of resistance evolution and persistence, we profiled the replication fitness of over 1000 HIV-1 mutants with combinations of resistance associated mutations on its protease gene. We found that although resistance associated mutations greatly reduce replication fitness, they interact positively to alleviate the mutational load. These genetic interactions, termed epistasis, increase the ruggedness along the evolution paths, restricting resistance associated mutations from reversal. Our data support the clinical observations that drug resistance mutations tend to persist even when antiretroviral drug is discontinued.

3. Introduction

Antibiotics and antiviral drugs have achieved great success in recent history[1]. However, therapeutic failure may occur due to low adherence and the emergence of drug resistance[2,3]. The increasing amount of drug resistant pathogens is a global threat to public health[4-11]. The genetic barrier to drug resistance, defined as the number of mutations needed to acquire resistance, is a major determining factor of treatment outcomes[12-14]. Another important but often overlooked aspect of drug resistance is the fitness barrier[15-17]. Resistance associated mutations (RAMs) in pathogen proteins may decrease enzymatic activities, interfere with molecular interactions, or destabilize the protein structure[18-22]. Because of the impaired replication capacity without drug selection, drug-resistant mutants cannot normally outcompete wild-type or establish in the population[23-25]. However, drug-resistant mutants can sometimes reach substantial frequency in the population. Fluctuating drug concentrations may create time windows

when drug-resistant mutants replicate better than wild-type virus[26]. Moreover, compensatory mutations can rescue the impaired replication capacity of mutants and stabilize drug resistance[27-29]. Thus, comprehensive quantification of the fitness landscape is needed to predict the evolution of drug resistance[30,31].

Epistasis, i.e., genetic interactions between mutations, is prevalent in molecular evolution[32-34]. Negative epistasis decreases fitness of the double mutant, posing constraints on gaining multiple mutations[35,36]. It plays an important role in shaping the local fitness landscape[37]. Positive epistasis increases replication capacity of the double mutant, facilitating pathogens to acquire and maintain drug resistance[38-40]. Positive epistasis may create a fitness valley that prevents drug resistant mutations from reversal[41]. Collectively, positive and negative epistasis determine the topography of the fitness landscape[42] and the course of drug resistance evolution[32]. Empirical studies on the genetic interactions between RAMs, especially in high-order mutants, are still rare[43,44].

HIV-1 protease inhibitors are important components of combination antiretroviral therapy[45] that target HIV-1 protease enzymatic activity[46,47]. Second-generation protease inhibitors have extremely high binding affinity to viral protein[48]. Resistance to them typically requires more mutations than resistance to first-generation protease inhibitors and other antiretroviral drugs[49,50]. For example, mutation K103N on reverse transcriptase is sufficient to confer HIV-1 nevirapine (NVP) resistance[51], while more than 4 *de novo* mutations are needed for protease inhibitor Darunavir (DRV) resistance[52]. Protease inhibitor-resistant viruses with multiple RAMs also have significantly reduced fitness[53,54]. HIV-1 gained RAMs on protease during sub-optimal protease inhibitor therapy[55]. Most resistance mutations directly affect the binding affinity between HIV-1 protease and the

inhibitor, but they are likely to be deleterious because they also reduce binding to the native substrate of HIV-1 protease. To compensate the deleterious effect, some other RAMs stabilize HIV-1 protease, allowing drug-resistant virus to replicate as efficiently as its parental wild-type virus[27,56]. The compensatory effects between pairs of RAMs have been studied in several studies and are available on the Stanford HIV drug resistance database[57-61]. Meanwhile, reversals of protease inhibitor resistance-associated mutations were rarely seen clinically, even when therapy was interrupted[62] or when mutant virus infected drug-naïve patients[63,64]. These observations indicate that epistasis may be important for the evolution of protease inhibitor resistance. Recent analyses of sequence co-variation in drug-targeted HIV Pol proteins (protease, reverse transcriptase and integrase) and co-evolutionary Potts model provide evidence that epistasis plays an important role in drug resistance. Despite being disfavored in the wild-type background, primary resistance mutations can become entrenched by the complex mutation patterns which arise in response to drug therapy[65,66].

Here, we present a quantitative high-throughput genetics approach to study the fitness distribution and epistasis of HIV-1 protease inhibitor RAMs[67,68]. Combining these data with clinical data and fitness models, we found that positive epistasis was predominant and especially enriched among RAMs, and prevalent along drug resistance evolutionary paths. Our results suggest that fitness hills created by epistasis result in barriers that entrench RAMs, and thus drug-resistant viruses are unlikely to revert after transmission to drug-naïve patients or discontinuation of anti-retroviral drug treatment.

4. Results

4.1. Fitness profiling of RAMs in HIV protease

To study the interactions among RAMs in HIV protease, we constructed a library of virus mutants that covers combinations of amino acid substitutions at 11 resistance-associated sites in HIV protease (Fig 1A, Table 1, $2^9 \times 3^2 = 4608$ genotypes). To ensure sufficient coverage, we harvested more than 30000 colonies after transforming *E. coli*. These sites have been annotated as major drug resistance sites in Stanford Drug Resistance Database[69], and all have been shown to be strongly associated with drug resistance. In our mutant library, 9 sites have one amino acid substitution and the other 2 sites have 2 amino acid substitutions (Fig 1A, Table 1). 2736 out of 4608 possible genotypes (59.38%) were covered in the plasmid library.

We quantified the relative fitness of mutants using high-throughput fitness profiling (Fig 1B, See Material and Methods for details). We performed 3 independent transfection experiments to validate the reproducibility of fitness profiling. 20 million 293T cells were transfected and 50 million T cells were infected in each experiment. For each biological replicate, relative fitness was calculated independently. The Pearson's correlation coefficients of single, double and triple mutations between replicates range from 0.80 to 0.82 (Fig 1C and S1 Fig). After filtering out mutants with low frequency or low reproducibility among replicates of input virus libraries (see Material and Methods for details), we were able to estimate the relative fitness of 1219 genotypes. The fitnesses of all single mutants, and more than 70% of double and triple mutants, were quantified (S2 Fig).

To validate the quantification of relative fitness, we conducted competition experiments with individually constructed protease mutants. We performed two sets of validation experiments. For the first set, we packaged the mutant virus and wild-type virus independently and mixed them in pairs for head-to-head competition. The frequency of the mutant virus and wild-type virus were quantified by deep sequencing and the relative fitness was calculated in the same way as we did in library screening. A total of 7 mutants were constructed and validated. For the second set of experiments, we mixed all 7 single mutants with wild-type virus in competition experiments. The relative fitness was defined in the same way. The fitness measured in validation experiments was highly correlated with the fitness in library screening (Fig 1D, $R = 0.84$ for each independent validation, Pearson's correlation test). In addition, we compared the selection coefficients of HIV-1 protease mutants measured in an independent study by Boucher et al[71]. and the relative fitness values in our experiment (Fig 1E, S2 Table). The experimental results from two studies show a good correlation (Pearson's correlation coefficient is 0.79), supporting the reliability of our experimental methods.

4.2. Positive epistasis rescues the mutational load of RAMs

We first looked at fitness effect of RAMs. In our definition, a mutant virus of relative fitness -1 means that the relative frequency of this mutant drops 10 fold after infection in cell culture. All single mutations were deleterious to virus replication (Fig 2A). The relative fitness of single mutants ranged from -2.33 (V82F) to -0.19 (L90M). This is consistent with previous reports that randomly introduced mutations were mostly deleterious to protease enzymatic activity or HIV-1 replication capacity[72-74]. Random mutagenesis in other viruses also revealed a lack of beneficial mutations in well-adapted systems[75-77]. RAMs in

particular were also reported to be deleterious to virus replication. They may destabilize viral protein, affect enzymatic activities or impact other protein-protein interactions[78].

We then analyzed epistasis between all pairs of RAMs. Previous studies have shown the prevalence of epistasis among pairs of random mutations or spontaneously accumulated mutations[79]. However, studies focused on the epistasis among drug resistance mutations are still limited[30,39,72,75,80]. Based on the fitness effect of single RAMs, we predicted the relative fitness of double mutants with the assumption that no epistasis existed among any two single mutations (i.e., the predicted relative fitness of a double mutant was the sum of those of two single mutants) (Fig 2B). Surprisingly, the observed relative fitness of most double mutants was significantly higher than the predicted values ($p = 2.2 \times 10^{-6}$, two-sided Wilcoxon rank sum test), suggesting that positive epistasis is prevalent among RAMs (Fig 2B inset). Pairwise epistasis between two RAMs is quantified as $\varepsilon_{i,j} = f_{i,j} - f_i - f_j$, f_i represents the relative fitness of mutants i. The distribution of epistasis ranged from -0.69 (M46I and L90M) to 2.34 (L76V and V82F) and 86.6% of pairwise interactions between RAMs are positive.

We also analyzed the extent of epistasis among high-order mutants. We observed a trend that relative fitness decreased as the order of mutants increased (S3 Fig). This is consistent with previous reports that mutational load restricted virus replication capacity[81,82]. To better quantify the fitness cost of multiple mutations, we calculated the frequency of viable mutants by different thresholds, $f > -2$ or $f > -4$. The frequency of viable mutant virus decreased as the number of mutations increased (Fig 2C), consistent with previous observations in HIV-1 and other RNA viruses[83-86]. We then predicted the relative fitness of high-order mutants by summing the relative fitness of corresponding

single mutants. We observed more viable mutants than would be predicted without epistasis (Fig 2C). This indicated pervasive positive epistasis rescued high-order mutants from lethal relative fitness, which is consistent with other clinical observations in protease inhibitor resistant virus. As a result, positive epistasis partially relieved HIV-1 mutational load and allowed viruses to explore more sequence space.

4.3. Enrichment of positive epistasis among RAMs

There are two possible explanations for the observed positive epistasis among RAMs of HIV protease. The first hypothesis is that all mutations in HIV protease tend to interact positively. The second hypothesis is that epistasis among random mutations in HIV protease is on average zero, but positive epistasis is enriched among RAMs. We introduced the Potts model to test our hypotheses, while simultaneously testing whether our finding of prevalent positive epistasis among RAMs carries over to the clinical setting. Potts models, originally developed in statistical physics, have been employed previously to use the population-level frequencies and correlations between different mutations to estimate their fitness effects[87-90]. In the Potts model, the probability of observing a genotype $\vec{A} = \{A_1, A_2, \dots, A_{99}\}$ is given by equations in Fig 3A. Here the $A_i, i \in \{1,2, \dots ,99\}$ are variables that represent the amino acid at site i on each of the 99 sites of protease. Two sets of Potts parameters, fields $h_i(A_i)$ and couplings $J_{ij}(A_i, A_j)$, give the statistical energy $E(\vec{A})$, which is negatively correlated with fitness. These parameters are estimated in order to reproduce the frequencies and correlations between mutations that are observed in the data. The fields $h_i(A_i)$ represent the fitness effect of amino acids A_i at sites i alone, while the couplings $J_{ij}(A_i, A_j)$ describe epistatic interactions between amino

acids A_i at site i and A_j at site j. For both the couplings and the fields, positive parameter values correspond to beneficial effects on fitness, while negative values correspond to deleterious fitness effects. We applied a maximum entropy method[91] to an alignment of 20911 HIV-1 clade B protease sequences from drug naïve patients, obtained from the Los Alamos National Laboratory HIV sequence database (hiv.lanl.gov, accessed 24 March 2017) to calculate these two sets of Potts parameters.

Then we calculated $E(\vec{A})$ for all mutants in our protease library. We found that the Potts energy for single, double or triple mutants ($\Delta E = E_{mut} - E_{WT}$) is significantly correlated with the relative fitness we measured in our screening ($\rho = -0.46$, $p = 1.2 \times 10^{-14}$, Spearman's correlation test, Fig 3B). The correlation was lower than previous analysis in HIV-1 Gag and Env region[88,90]. This may be due in part to strong phylogenetic bias on the inferred Potts parameters, because protease is highly conserved. It is also possible that epistatic interactions with cleavage sites on other parts of the HIV-1 genome and complicated anti-innate immunity functions of protease obscure the effects of individual mutations on replicative fitness *in vitro*[57,59,92].

The Potts couplings $J_{ij}(A_i, A_j)$ give the contribution of pairwise epistatic interactions between amino acids A_i and A_j at sites i and j, respectively. We compared the couplings among RAMs and among all other possible mutations on protease (Fig 3C). Couplings of other protease mutations clustered near 0, while those of RAMs are significantly more positive than that of other mutations (D = 0.22, $p = 2.1 \times 10^{-07}$, two-sided K-S test). Moreover, $J_{ij}(A_i, A_j)$ among RAMs were also more positive than those between RAMs and other residues (D = 0.22, $p = 5.1 \times 10^{-07}$, two-sided K-S test). Although the fields

$h_i(A_i)$ of RAMs are more negative than other mutations, the difference is not significant (Fig 3D, D = 0.25, $p = 0.20$, two-sided K-S test). We note that the magnitude and the variation of field parameters is much larger than that of coupling parameters (Fig 3C-D). The Interquartile Range (IQR, i.e. the middle 50%) of field parameters is 3.55, while the IQR of coupling parameters is 0.15. The standard deviation of field parameters is 2.29, while the standard deviation of coupling parameters is 0.37. Overall, analysis based on the Potts model is consistent with our experimental results that positive epistasis is enriched among RAMs, and lends support to our second hypothesis that epistasis among random mutations in HIV protease is on average zero.

4.4. Implications of positive epistasis in evolution

To study the role of epistasis in evolution, we analyzed the evolutionary pathways covering all genotypes with up to 4 amino acid substitutions from the wild-type virus (13 single mutants, 67 double mutants, 176 triple mutants and 290 quadruple mutants) (Fig 4A). Mutants are linked if they differ by one amino acid substitution.

We have found that all 13 RAMs are deleterious on the wild-type background (Fig 2A). However, the fitness effect of a single RAM becomes less deleterious on genetic backgrounds where other RAMs have been fixed (S4 Fig). Following the generalized definition of epistasis proposed by Shah et al.[93], we define trajectory-based epistasis $\varepsilon_{M,j}$ that measures the deviation of the fitness effect if the order of mutations were reversed. $\varepsilon_{M,j} = f_{M,j} - f_M - f_j$, where f_M and f_j represent the relative fitness of background M and single mutant j [94]. For example, mutation j can be deleterious on the wild-type background but beneficial on another genetic background that mutation i has been fixed.

Trajectory-based epistasis is calculated for each amino acid substitution and averaged over genetic backgrounds with a certain Hamming distance to the wild-type (Fig 4B). For all RAMs profiled in this study, we find that trajectory-based epistasis is overall positive and increases steadily with the number of substitutions, i.e., the fitness contribution of a specific amino acid substitution becomes more positive if more RAMs have been fixed. Our results are consistent with previous analyses of sequence co-variation in HIV-1 protease[65,66], where inferred epistastic interactions among mutations at PI resistance associated sites lead to entrenchment of primary drug resistance mutations. In this study, we combine the analyses of co-variation (Potts model) with comprehensive experimental fitness data of HIV-1 protease mutants (including a large number of higher-order mutants) to provide direct evidence of positive epistasis among RAMs of second-generation PIs.

We tested the hypothesis that positive epistasis prevented resistance associated genotypes from reverting to wild-type[95,96]. Although RAMs incurred significant fitness cost, some drug resistant mutants would not revert to wild-type after transmitting to a drug naïve patient. We quantified the frequency of accessible evolutionary pathways between mutants and wild-type in our experimentally measured fitness landscape of HIV protease RAMs. A reversal path is defined to be accessible if and only if the virus fitness increases monotonically along the path. For example, quadruple mutant V32I_M46I_I54L_V82F has many paths to revert to wild-type (Fig 4A). Among them, reversing V32I, I54L, V82F and M46I in order is an accessible path (Fig 4A, red line). On the contrary, reversing I54L, V82F, M46I and V32I is not an accessible path because there are 2 steps with decreasing fitness (Fig 4A, blue line). We found that among double mutants, 44 have two accessible reversal paths to the wild type, 20 have only one accessible reversal path, and

interestingly 3 of them have none. These 3 mutants (I50V_T74P, M46I_I54M and L76V_V82F) represent local fitness peaks and the reversal to wild-type is blocked by a fitness valley. We found that the number of accessible reversal paths decreased with the accumulation of RAMs (Fig 4C). This indicates that protease mutants become less likely to revert to wild-type as the number of RAMs increases. Our results are consistent with clinical observations that protease inhibitor resistance associated mutations seldom reverted even when therapies were interrupted or drug-naïve patients were infected. The difficulty of reversal also explains the rising frequency of drug resistant HIV-1 viruses in acute phase patients.

5. Discussion

In this study, we systematically quantified the fitness effect of RAMs of HIV-1 protease. While all RAMs reduced the virus replication fitness, pervasive positive epistasis among RAMs alleviated the fitness cost substantially. Moreover, we analyzed the HIV sequence data from patients by the Potts model. We found the statistical energy inferred from HIV sequences *in vivo* correlated well with the replication fitness measured *in vitro*. Based on our fitness data and the mutational couplings inferred by the Potts model, we showed that positive epistasis is enriched among RAMs of HIV-1 protease, in both local fitness landscape and evolutionary paths. Finally, we studied the role of epistasis in evolutionary pathways. We found that positive epistasis among RAMs entrenches drug resistance and blocks the reversal paths to wild-type virus, which has important implications for the design of anti-retroviral therapies. Through this project, we also established a high-throughput platform to quantify the genetic interactions among a group of mutations. Another independent study profiled the fitness effect of all single amino acid change on

HIV protease. The data showed significant correlation with our study (Figure 1E, Pearson's correlation coefficient (R) is 0.79).

There are a few limitations of this study. Firstly, we only measured the fitness effect of RAMs in the absence of protease inhibitors. We are not able to quantify drug resistance of RAMs because protease inhibitors block multiple rounds of virus infection and prevent us from accurate examination of mutant frequency under drug selection. Also, we did not sequence other genes of HIV-1. HIV-1 mutates rapidly due to low fidelity of reverse transcriptase [97,98]. There might be compensatory mutations occurring on other proteins that rescued the protease RAMs. Secondly, the correlation between our validation experiments and high-throughput screening experiments was less than the correlation observed in similar experiments in bacteria and yeast [99,100]. The correlation between Potts energy and experimental fitness is also lower than previous reports on Gag and Env regions. Mechanistic difference between logistic growth and viral growth may complicate the quantification of viral fitness [101]. Direct measurement of viral frequency may not linearly correlate to the probability of replication [102]. Moreover, we tested a large number of higher-order mutants (i.e. multiple mutations from the wild-type virus). Our experimental dataset not only contains clinically observed genotypes but also combinations of mutations that was not observed in patients, which are highly deleterious and may suffer from higher experimental errors. If we exclude higher-order mutants and very deleterious genotypes (S5 Fig), the Spearman's correlation between fitness and Potts energy is higher ($\rho = -0.54$, compared to $\rho = -0.46$ in Figure 3B). Thirdly, we did not cover all clinically observed polymorphism, given the bottlenecks in virus library screening. We chose to prioritize for RAMs of second-generation protease inhibitors

Darunavir (DRV) and Tipranavir (TPV), which are considered to have high genetic barriers (i.e. multiple RAMs are involved in the emergence and reversal of drug resistance). According to Stanford Drug Resistance Database , the RAMs that we chose contribute to the resistance to DRV and TPV (S2 Table). The only exception is L90M, which is frequently found in drug resistant viruses. The RAMs and the combinatorial genotypes in our library are prevalent in patients and documented in Stanford Drug Resistance Databases (Table 1). Future work could be extended to cover more clinically observed polymorphism in HIV-1 protease and other drug-targeted proteins. Finally, the correlation between Potts energy and experimental fitness is confounded by many factors, like different selection pressures *in vivo* and *in vitro*, or phylogenetic bias. Nonetheless, we observe moderate but statistically significant correlation between the coupling parameters in the Potts model and the experimental epistasis (S6 Fig, Spearman's correlation test, $p = 6.8 \times 10^{-3}$). We note that the coupling parameters in the Potts model and the experimental measure of epistasis (calculated for WT genetic background) are conceptually different, representing Fourier coefficients and Taylor coefficients of the fitness landscape[103]. Our findings are consistent with the literature that Potts model couplings are strongly associated with contact residues in the three-dimensional structure of protein families [104,105]. We tested a series of different statistical models, including the binary (Ising) model inferred via ACE, the Potts model inferred via pseudo-likelihood maximization (a popular approach to analyzing sequence data from protein families), and the Potts model inferred via ACE, to examine the epistatic effects among drug resistance mutations (S7 Fig). We found that the Potts model inferred via ACE is the best choice to analyze epistasis in our study.

Statistical models suggest a pervasive negative distribution of fitness effect for single mutations on HIV-1 [31,88,106]. Previous models also predicted the entrenchment of deleterious RAMs by positive epistasis. This dataset provides a unique chance to experimentally test these statistical hypotheses. The predominance of positive epistasis is also observed in HIV-1 and in other organisms[30,39,107]. However, they either relied on naturally-occurring resistant clones or indirectly activating gene functions. This report is the first dataset to systematically quantify the epistasis among functional residues in HIV-1 drug resistance evolution, without the bias of drug selection and *in vivo* evolution. Overall, our results are important for understanding drug resistance evolution. We found positive epistasis plays a critical role in HIV-1 gaining and maintaining drug resistance. Epistasis makes the fitness landscape rugged, preventing RAMs from reversion to wild-type, even when antiviral therapy is interrupted or virus transmits to a healthy individual[95,108].

Positive epistasis involves many kinds of molecular mechanisms. We find that the relative fitness of single mutants is not a significant factor of positive epistasis. We compared h_i in the Potts model for all RAMs and other single mutants. They were not significantly different ($p = 0.20$, K-S test). Physical distance between residues is a significant factor contributing to positive epistasis. The physical distances between these residues were significantly less than those between any two random residues on HIV-1 protease (D = 0.32, $p = 3.9 \times 10^{-10}$, two-sided K-S test, S8 Fig), suggesting that physical contact among RAMs might contribute to the observed positive epistasis. Notably, their average distance was more than 10 Å, indicating most of them did not have direct contact. Some mutations may have structurally stabilizing effect to other residues. We used FoldX and

Rosetta to predict the folding free energy ($\Delta\Delta G$) as a quantification of protein stability for all mutants in our library (S8 Fig) [109,110]. We notice that mutation V82F contributed to the positive epistasis on many genetic backgrounds (Fig 4B), but it did not contribute much to the stabilizing effect. Thus, structurally stabilizing effects cannot fully explain the predominance of positive epistasis observed in this study. Future studies on the structure and function of HIV-1 protease mutants will help elucidate the molecular mechanisms underlying the interactions among RAMs.

6. Material and Methods

6.1. Plasmid library construction

HIV-1 RAMs were picked according to their prevalence in protease inhibitor treated patients. We chose 11 residues with 13 mutations to construct a combination of HIV-1 protease mutant library (Table 1).

We used a ligation-PCR method to construct the library on NL4-3 backbone, which is an infectious subtype B strain. All possible combinations of these 13 mutations are $2^9 \times 3^2 = 4608$ genotypes. The mutagenesis region spanned 243 nucleotides on HIV-1 genome. We split the region into 5 oligonucleotides and ligate them in order by T4 ligase (from New England BioLabs). The sequence of oligonucleotides are shown in S3 Table. After each ligation, we recovered the product by PCR and used restriction enzyme BsaI-HF (from New England BioLabs) to generate a sticky end for the next step ligation.

After making the 243-nucleotide mutagenesis fragment, we PCR amplified the upstream and downstream regions near this fragment and used overlap extension PCR to ligate

them together. We then cloned it into full length HIV-1 NL4-3 background. We harvested more than 30,000 *E. coli* colonies to ensure sufficient coverage of the library complexity.

6.2. Virus production

The plasmid DNA was purified by HiPure Plasmid Midi Prep Kit (from Thermo Fisher Scientific). To produce virus, we used 16 μg plasmid DNA and 40 μL lipofectamine 2000 (from Thermo Fisher Scientific) to transfect 2×10^7 293T cells, in 3 independent biological replicates. We changed media 12 hours post transfection. The supernatant was harvested 48 hours post transfection, labeled as input virus and frozen at −80 °C. We harvested 40mL viruses from each transfection. Virus was quantified by p24 antigen ELISA kit (from PerkinElmer).

6.3. Library screening

CEM cells were cultured in RMPI 1640 (from Corning) with 10% FBS (from Corning). To passage library in T cells, we added 25 mL viruses and 120 μg polybrene to 50 million CEM cells. We achieved 10 ng p24 (10^8 physical viral particles) for every million CEM cells during infection. We washed cells and completely changed media 6 hours post infection. We supplemented the cells with fresh media 3 days post infection and harvested supernatant 6 days post infection. We centrifuged supernatant at $500 \times g$ for 3 minutes to remove the cells and cell debris. The rest of supernatant was frozen at −80 °C.

In summary, we carefully controlled the experiment scales to ensure the library complexity was maintained in every step. Briefly, we harvested $> 3 \times 10^4$ *E. coli* colonies during bacteria transformation, which ensured ~ 6-fold coverage of the expected complexity (4608 genotypes). We then transfected 2×10^7 HEK 293T cells with 16 μg plasmid library

to package infectious viruses. We used 25 mL viruses (500 ng p24, $\sim 5 \times 10^9$ viral particles) to infect 2×10^7 million CEM cells for each biological replicate.

6.4. Sequencing library preparation

We used QIAamp viral RNA mini kit (from QIAGEN) to extract virus RNA from supernatant. We then used DNase I (from Thermo Fisher Scientific) to remove the residual DNA. We used random hexamer and SuperScript III (from Thermo Fisher Scientific) to synthesize cDNA. The virus genome copy number was quantified by qPCR. The qPCR primers are 5'--3' and 5'--3'.

At least 2×10^5 copies of viral genome were used to make sequencing libraries. We PCR amplified the mutagenesis regions using the following primers: 5'--3' and 5'--3'. We then used BpmI (from New England BioLabs) to cleave the primers and ligate the sequencing adapter to the amplicon. We used PE250 program on Illumina MiSeq platform to sequence the amplicon.

6.5. Calculation of fitness and epistasis

We used custom python codes to map the sequencing reads to reference NL4-3 genome. Mutations were called if both forward and reverse reads have the same mutation and phred quality scores are both above 30. All codes are available on https://github.com/Tian-hao/protease-inhibitor. All data were deposited in SRA (short read archive) database under accession PRJNA546460. For each replicate of the virus library from the transfected 293T cells, we reached 4.45×10^5 to 6.05×10^5 sequencing depth. We filtered out the genotypes with frequency fewer than 5×10^{-5} in any biological

replicate and the genotypes whose frequency differ more than 10 folds between any two biological replicates.

Relative fitness $f_{m,r}$ of mutant m in experiment r (biological replicates) was defined as Equation 1.

$$f_{m,r} = \log_{10}\left(\frac{F_{m,r,output}}{F_{m,r,input}} / \frac{F_{WT,r,output}}{F_{WT,r,input}}\right)$$

$F_{m,r,input}$ is the frequency of mutant m before screening. $F_{m,r,output}$ is the frequency of mutant m after passaging. $F_{WT,r,input}$ is the frequency of wild-type virus before screening. $F_{WT,r,output}$ is the frequency of wild-type virus after passaging.

The relative fitness f_m was defined as the average of 3 biological replicates (Equation 2). However, if relative fitness was missing in one replicate, we only average the other two replicates. The relative fitness value of all mutants was shown in S1 Table.

$$f_m = \sum_{t=1}^{R} f_{m,r} \, / R$$

, where R is the number of biological replicates.

Pairwise epistasis $\varepsilon_{i,j}$ between mutant i and mutant j was defined as:

$$\varepsilon_{i,j} = f_{i,j} - f_i - f_j$$

, where $f_{i,j}$ refers to the relative fitness of double mutant i and j.

Trajectory-based epistasis $\varepsilon_{M,j}$ between a multi-mutation genotype M and another genotype differ by one mutation j was defined as:

$$\varepsilon_{M,j} = f_{M,j} - f_M - f_j$$

6.6. Potts model

Data used to infer parameters for the Potts model were downloaded from the Los Alamos National Laboratory HIV sequence database, as described in the main text. Sequences were processed as previously described [111]. Briefly, we first removed insertions relative to the HXB2 reference sequence. We also excluded sequences labeled as "problematic" in the database, and sequences with gaps or ambiguous amino acids present at >5% of residues were removed. Remaining ambiguous amino acids were imputed using simple mean imputation.

Each sequence in the multiple sequence alignment (MSA) is represented as a vector of variables $\vec{A} = \{A_1, A_2, \ldots, A_N\}$, where $N = 99$ is the length of the sequence. Each of the A_i represents a (set of) amino acid(s) present at residue i in the protein sequence. To choose the amino acids at each site that would be explicitly represented in the model, we first computed the frequency $p_i^*(A)$ of each amino acid A at each site i in the MSA. To compute these frequencies, we weighted the sequences such that the weight of all sequences from each unique patient was equal to one, thereby avoiding overcounting in cases where many sequences were isolated from a single individual. We then explicitly modeled the q_i most frequently observed amino acids at each site that collectively capture at least 90% of the Shannon entropy of the distribution of amino acids at that site . All remaining, rarely observed amino acids were grouped together into a single aggregate state. For these data, this choice resulted in an average of three explicitly modeled states at each site (minimum of 2, maximum of 6).

The Potts model is a probabilistic model for the 'compressed' sequences $\vec{A}$, where the probability of observing a sequence $\vec{A}$ is

$$\begin{aligned} P(\vec{A}) &= \frac{1}{Z} e^{-E(\vec{A})}, \\ E(\vec{A}) &= -\sum_{i=1}^{m} h_i(A_i) - \sum_{i=1}^{m} \sum_{j=i+1}^{m} J_{ij}(A_i, A_j). \end{aligned}$$

Here the normalizing factor

$$Z = \sum_{\vec{A}} e^{-E(\vec{A})}$$

ensures that the probability distribution is normalized. We used ACE to infer the set of Potts fields $h_i(A_i)$ and couplings $J_{ij}(A_i, A_j)$ that result in average frequencies and correlations between amino acids in the model [eq:potts-supp] that match the frequencies $p_i^*(A_i)$ and correlations $p_{ij}^*(A_i, A_j)$ observed in the data. We used a regularization strength of $\gamma = 7 \times 10^{-5}$ in the inference, which is roughly equal to one divided by the number of unique patients from which the sequence data were obtained. We used "consensus gauge," where the fields and couplings for the most frequent residue at each site in the protein are set to zero. We confirmed that the parameters inferred by ACE resulted in a Potts model that accurately recovered the correlations present in the data.

6.7. Validation experiments

We constructed 7 single mutants by site-directed mutagenesis. The primers used this experiment are listed in S3 Table. We used overlap-extension PCR to amplify the fragment with mutated nucleotides. We ligated the fragment with NL4-3 backbone using

ApaI and SbfI. We transformed competent *E.coli* and picked single colonies. We sequenced the protease region of plasmids to make sure there is only desired mutant in this region. 7 mutants were L10F, I47V, T74P, L76V, V82F, V82T, L90M.

We produced mutant viruses in 293T cells, mixed them with wild-type and infected CEM cells. The frequencies of mutant virus before and after infection were quantified by deep sequencing. We did 2 biological replicates with each validation method. For validation 1, we pairwisely mixed the mutant and wild-type virus oor competition. For validation 2, we mixed all 7 mutants and wild-type virus.

6.8. Protein stability prediction

Mutants' stability was predicted using either FoldX or Rosetta. For FoldX, we used the protease structure (PDB: 3S85) as reference and repaired the structure using the RepairPDB function. The free energy of the mutants was computed by using the BuildModel function under default parameters. For Rosetta analysis, we used the protease crystal structure (PDB: 6DGX) as reference and score function ddg_monomer to evaluate the effect of mutations. Each mutant was evaluated 10 times and the average score was used as $\Delta\Delta G$.

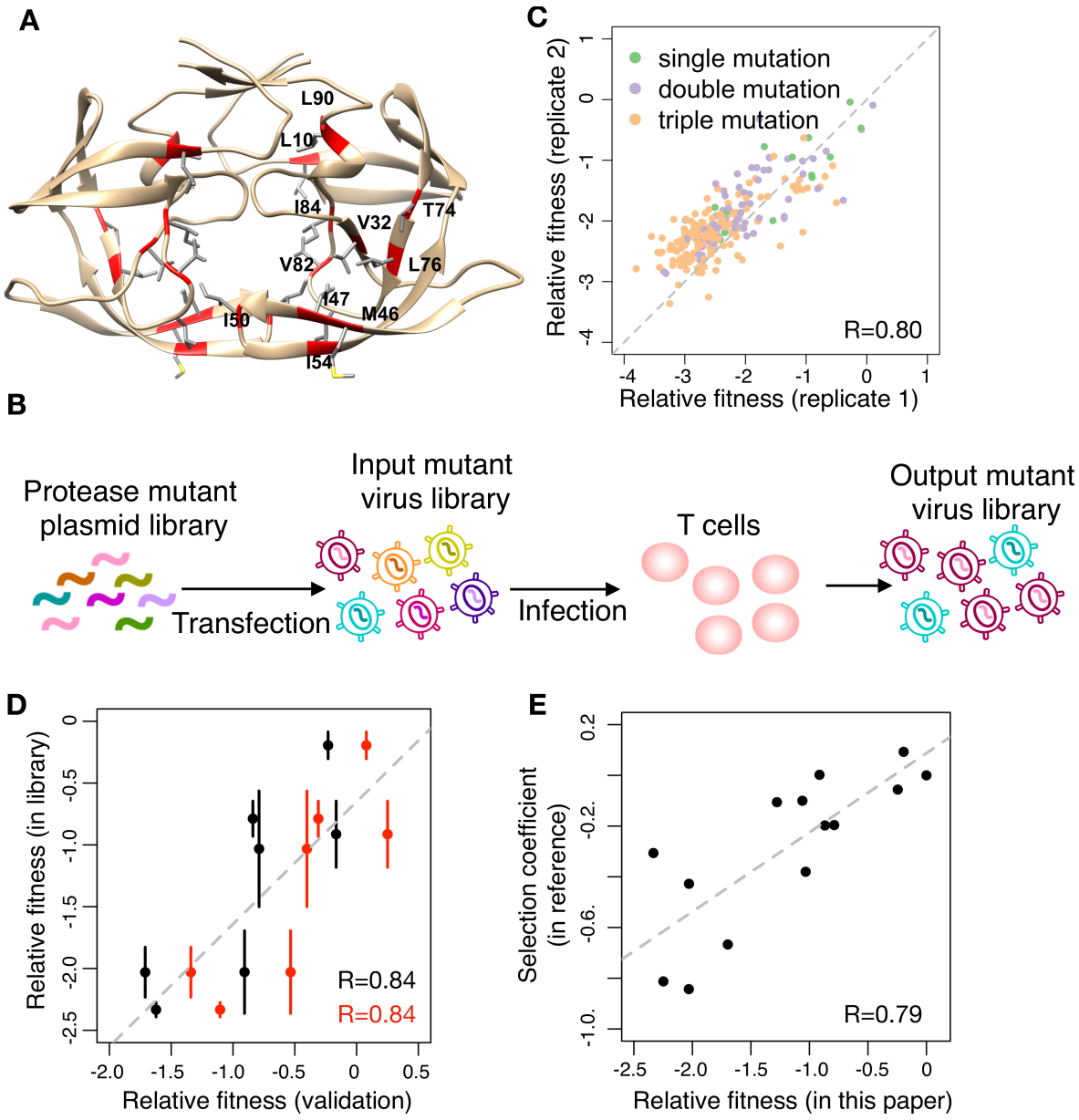


Figure 2-1. High-throughput fitness profiling of combinatorial HIV-1 protease mutant library.

(A) The structure of protease dimer (PDB: 4LL3). The side chains of selected resistance associated residues are shown. (B) Workflow of the fitness profiling. Protease mutations were introduced into NL4-3 background. T cells were infected by the mutant virus library. The frequency of mutants before (input library) and after (output library) selection was deep sequenced. (C) The

correlation of relative fitness between two biological replicates. Pearson correlation coefficient (R) is 0.80. (D) Two independent validation experiments were performed. We constructed 7 protease single mutant plasmids and recovered viruses independently. We mixed each mutant virus with wild-type virus (validation 1, black dots) and passaged in T cells for 6 days. We also mixed all 7 mutant viruses together with wild-type (validation 2, red dots) and infected T cells for 6 days. The relative fitness of each mutant was quantified by the same means as that in the library. Pearson correlation coefficients (R) for validation 1 and validation 2 are both 0.84. Error bar is standard deviation ($n = 3$). (E) The correlation of relative fitness in this study with the experimental selection coefficients in 71. Pearson correlation coefficients (R) is 0.79.

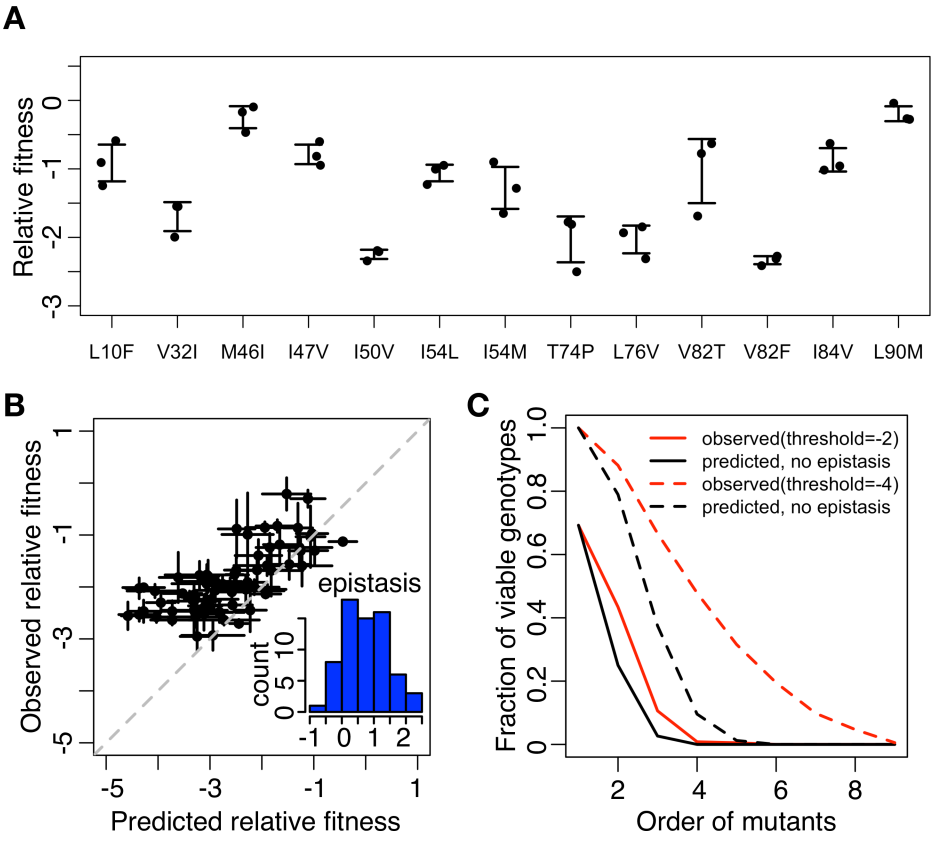
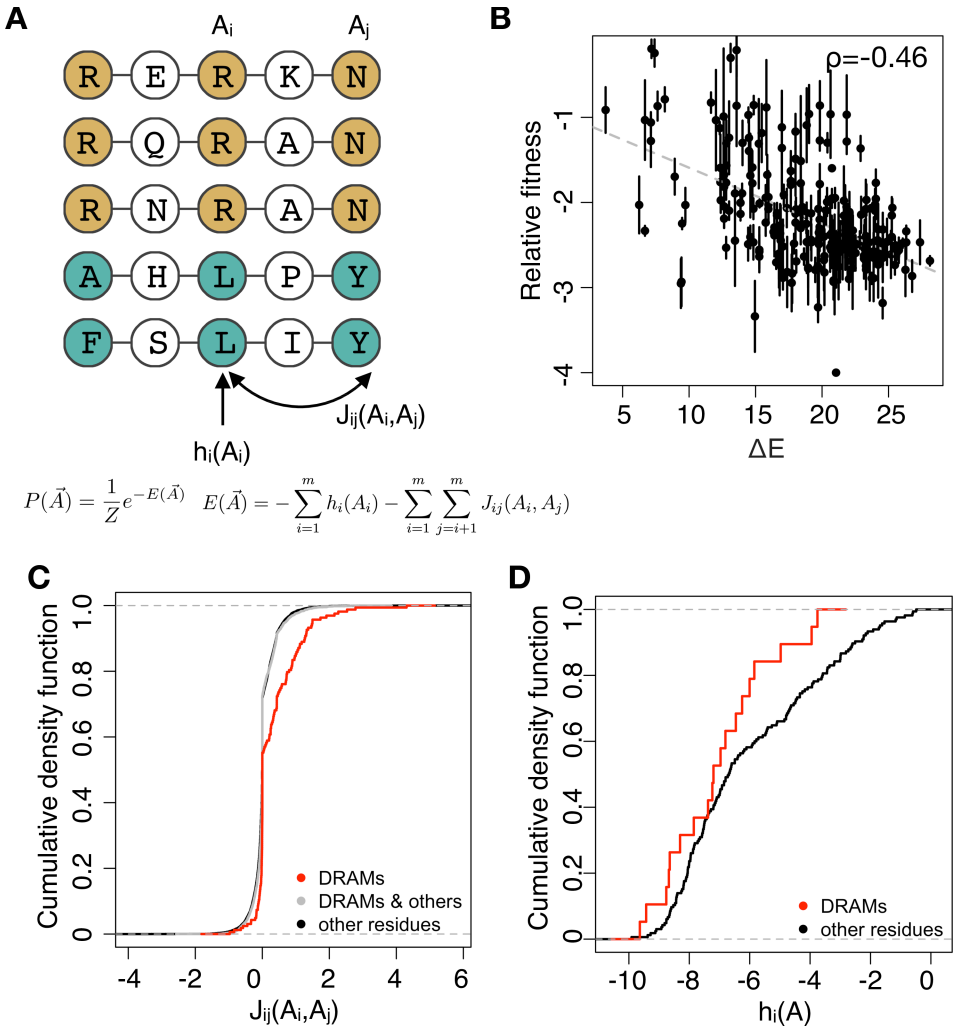


Figure 2-2. Positive epistasis is enriched among RAMs.

(A) Relative fitness of single mutants. Error bar is standard deviation (n=3). (B) The predicted relative fitness and observed relative fitness of double mutants. The predicted relative fitness was the sum of that of the two single mutants. Inset, the distribution of epistasis between double mutants. Error bar is standard deviation (n=3). (C)The predicted and observed fraction of viable mutants. A mutant was defined as viable if its relative fitness is higher than -4 (dashed line) or -2(solid line).



$$P(\vec{A}) = \frac{1}{Z} e^{-E(\vec{A})} \quad E(\vec{A}) = -\sum_{i=1}^{m} h_i(A_i) - \sum_{i=1}^{m} \sum_{j=i+1}^{m} J_{ij}(A_i, A_j)$$

Figure 2-3. Positive epistasis rescues the mutational load of RAMs.

(A) The conceptual graph of Potts model. Potts model uses the probability of mutations occurring with other mutations to estimate the statistical energy. h_i is the field parameter while J_ij is the coupling parameter. (B) The correlation of Potts energy(ΔE=E_mut-E_WT) and relative fitness of mutants with lower than 4 RAMs. Spearman correlation coefficient (ρ) is -0.46. (C)The cumulative density function of coupling parameters of RAMs and all other mutations. Coupling parameters between RAMs are more positive positive than those between RAMs and others (D=0.22, p=2.1×10^(-7), two-sided K-S test) and those between other residues(D=0.22, p=5.1×10^(-7), two-sided K-S test). (D) The cumulative density function of field parameters of RAMs and all other

mutations. Field parameters of RAMs and other residues are not significantly different (D=0.25, p=0.20, two-sided K-S test).

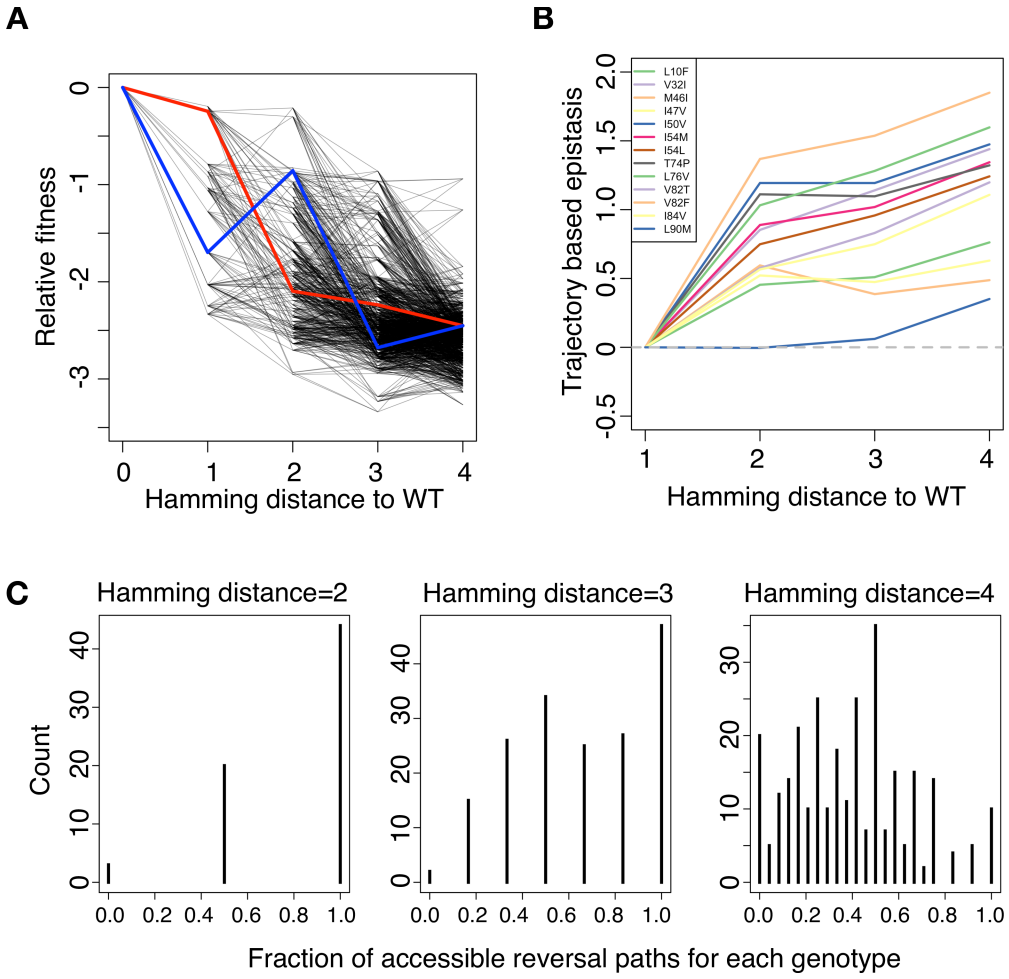
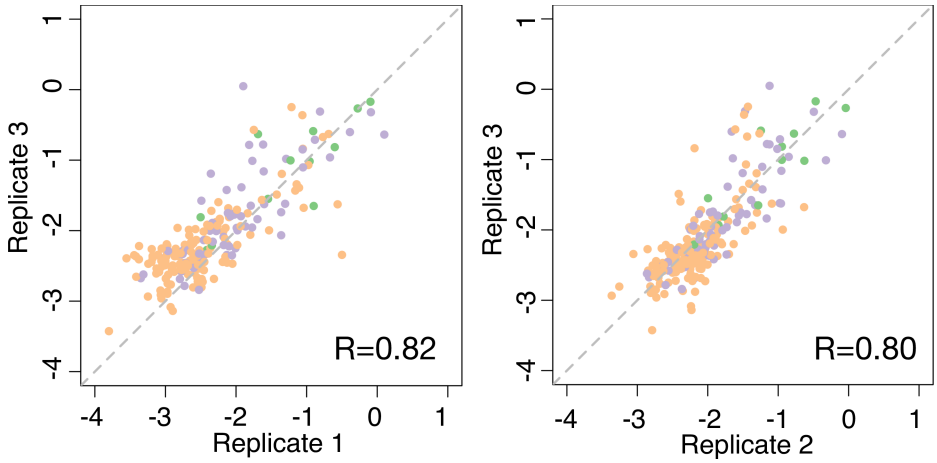


Figure 2-4. Ruggedness in fitness landscapes prevents RAMs from reversion to wild-type.

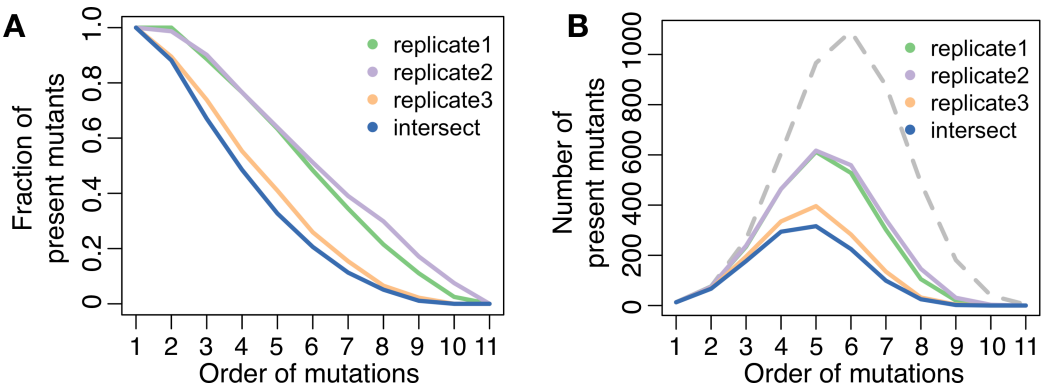
(A) Fitness with possible evolutionary trajectories. Mutants are linked if they only have one residue difference. Red line represents an accessible path that a quadruple mutant can take and reverse to wild-type. Blue line represents an inaccessible reversal path to wild-type for that mutant. (B) Trajectory-based epistasis is calculated for each amino acid substitution and averaged over genetic backgrounds with a certain Hamming distance to the wild-type. The fitness effect of a single mutation becomes less deleterious on genetic backgrounds where other RAMs have been

fixed. (C) The distribution of accessible paths for all genotypes with a certain hamming distance to wild type.



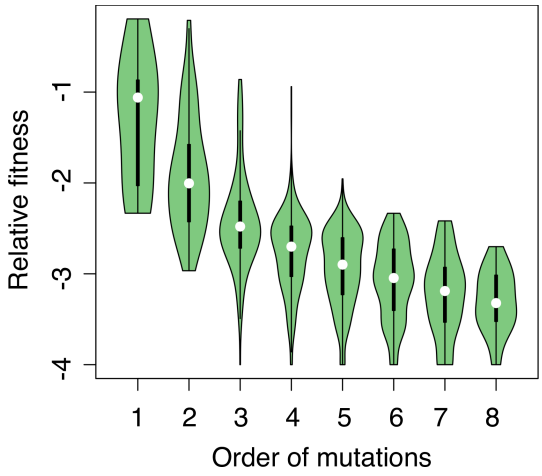
Supplementary Figure 2-1. The correlation of relative fitness among biological replicates.

All single mutants, double mutants and triple mutants are shown. R stands for Pearson correlation coefficient.

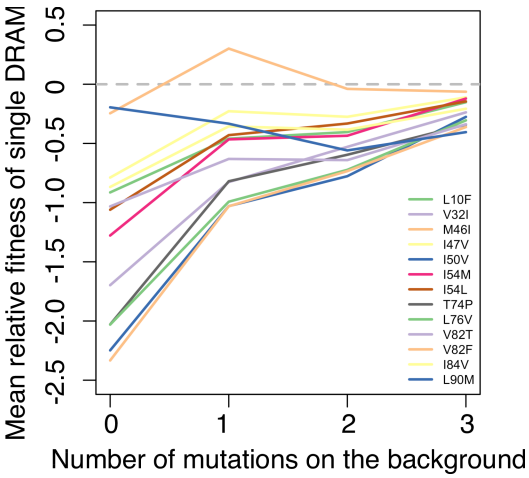


Supplementary Figure 2-2. Coverage of protease mutant library.

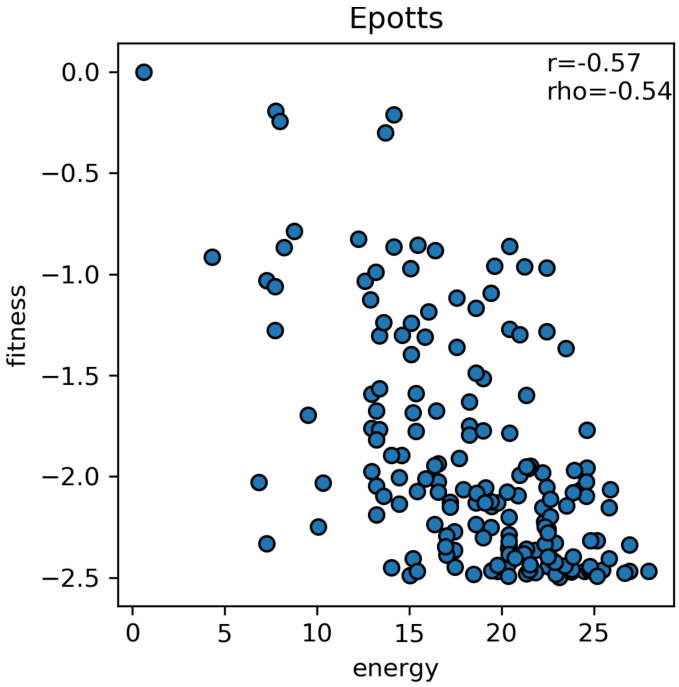
(A) Fraction of expected protease mutants in each transfection virus library. (B) Number of mutants in each transfection virus library. Dashed line represents the number of all possible combinations of mutations.



Supplementary Figure 2-3. Relative fitness of different order of mutations.

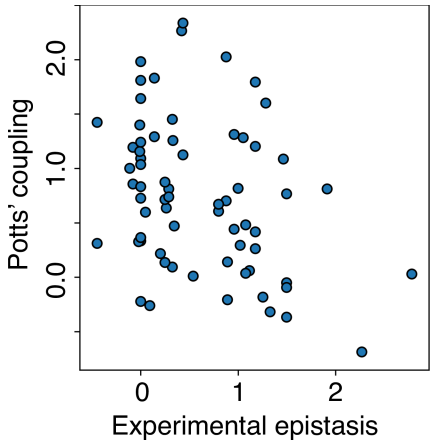


Supplementary Figure 2-4. Relative fitness of single RAMs on different genetic backgrounds.



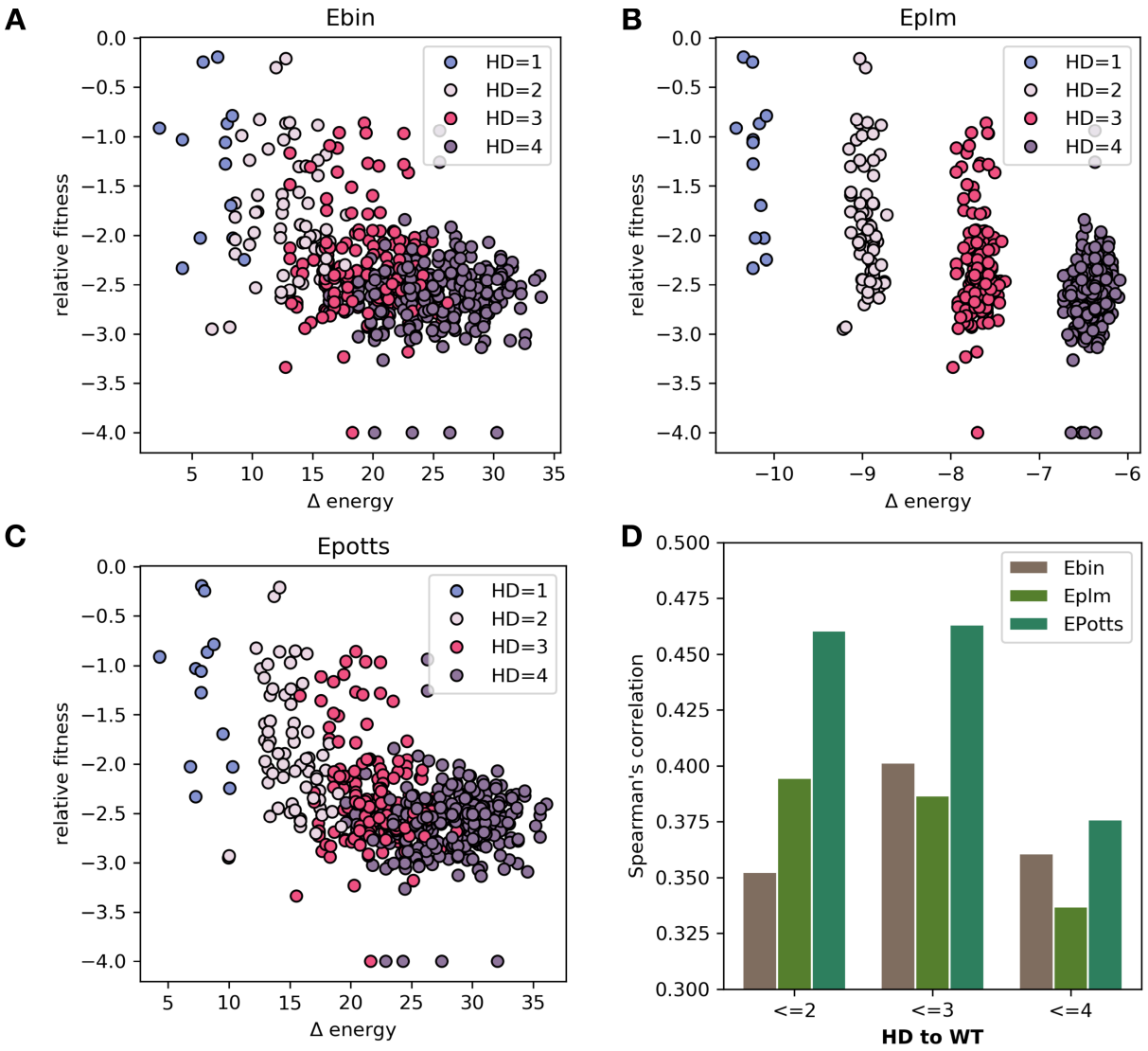
Supplementary Figure 2-5. Correlation between Potts energy and relative fitness for low order mutants.

Mutants with relative fitness higher than -2.5 and numbers of mutations lower than 4 is shown. The Pearson's correlation coefficient is -0.57. The Spearman's correlation coefficient is -0.54.



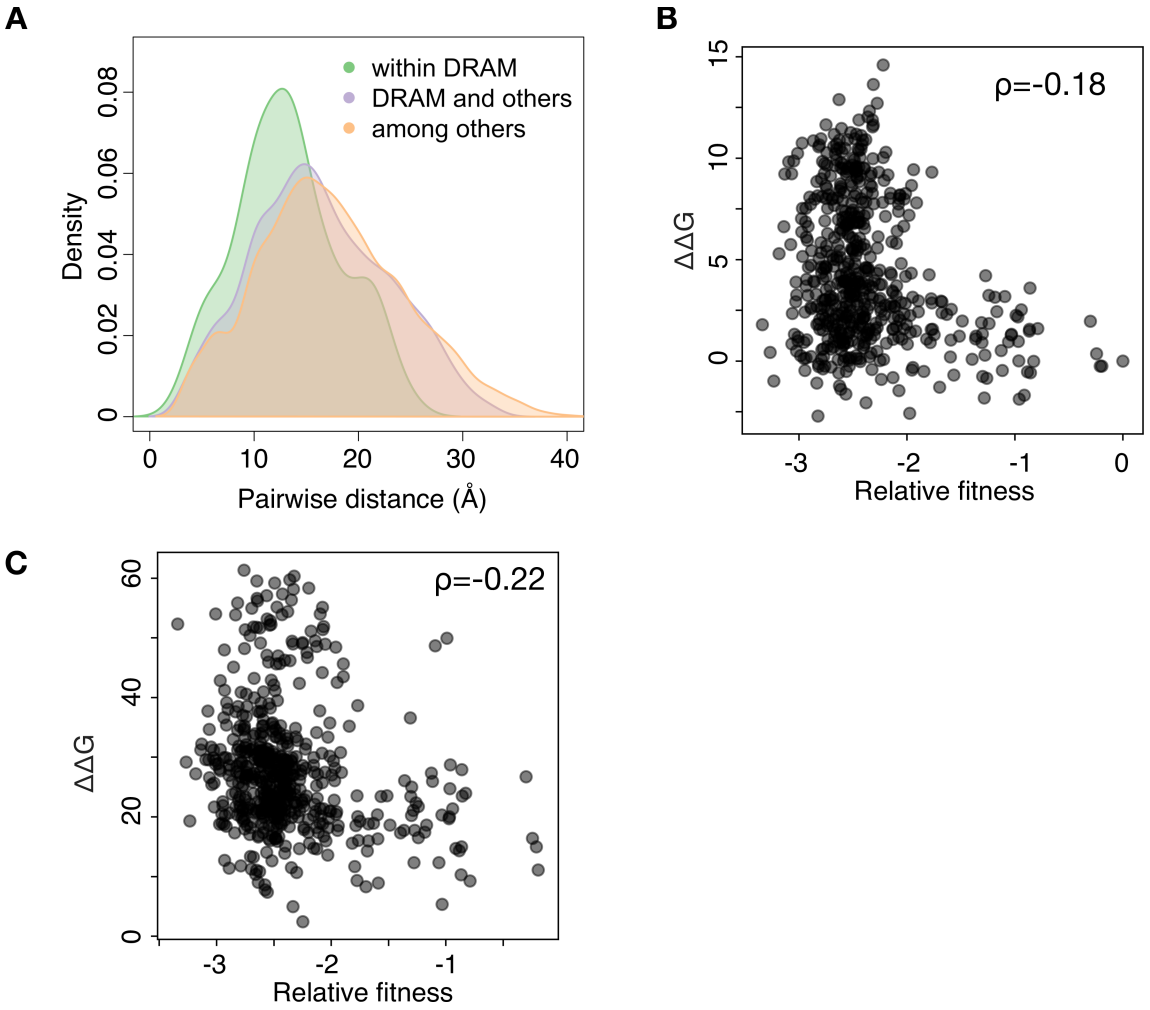
Supplementary Figure 2-6. The correlation between Potts' coupling parameters with experimental epistasis.

The pairwise epistasis between all RAMs in our library was compared with Potts' coupling parameters. The Spearman's correlation coefficient is -0.33. The p value for the Spearman's correlation coefficient is 6.8×10^(-3).



Supplementary Figure 2-7. Correlation between relative fitness and different statistical models.

(A, B & C). The correlation between relative fitness with (A, bin) binary (Ising) model inferred via ACE, (B, plm) the Potts model inferred via pseudo-likelihood maximization, or (C, potts) the Potts model inferred via ACE. (D) Spearman's correlation coefficients for different models. Mutants were classified according to their HD to wild-type. HD, hamming distance.



Supplementary Figure 2-8. Structure insights on resistance associated mutations.

(A) Distribution of pairwise distance among resistance associated residues and other residues. The distance between the C-α of two residues was shown. (B & C) Correlation between mutants' relative fitness and protein stability (ΔΔG). ΔΔG is predicted by FoldX (B) or Rosetta (C). The correlation coefficients were calculated for mutants with lower than 5 mutations. ρ stands for Spearman's correlation coefficient.

Table 2-1. List of protease inhibitor resistance associated mutations covered in the library.

Residue number	Consensus	Mutation	Prevalence in clinical dataset[a]	Occurrence in *in vitro* dataset[b]
10	L	F	1.54%	10.20%
32	V	I	1.37%	7.53%
46	M	I	4.32%	22.19%
47	I	V	0.88%	4.36%
50	I	V	0.30%	1.85%
54	I	L	0.68%	4.92%
54	I	M	0.48%	3.02%
74	T	P	0.37%	2.15%
76	L	V	0.46%	2.92%
82	V	T	0.64%	4.05%
82	V	F	0.33%	1.54%
84	I	V	3.00%	17.12%
90	L	M	7.71%	31.78%

[a] From 148840 subtype B protease sequences in Los Alamos Database.

[b] From 1951 isolates tested in PhenoSense assay.

7. Reference

1. Palella Jr FJ, Delaney KM, Moorman AC, Loveless MO, Fuhrer J, Satten GA, et al. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. New England Journal of Medicine. 1998;338(13):853–860.

2. Maggiolo F, Airoldi M, Kleinloog HD, Callegaro A, Ravasio V, Arici C, et al. Effect of adherence to HAART on virologic outcome and on the selection of resistance-conferring mutations in NNRTI-or PI-treated patients. HIV Clinical Trials. 2007;8(5):282–292.

3. Shafer RW. Rationale and Uses of a Public HIV Drug-Resistance Database. The Journal of Infectious Diseases. 2006;194(Supplement_1):S51–S58.

4. Lin J, Nishino K, Roberts MC, Tolmasky M, Aminov RI, Zhang L. Mechanisms of antibiotic resistance. Frontiers in Microbiology. 2015;6:34.

5. Lontok E, Harrington P, Howe A, Kieffer T, Lennerstrand J, Lenz O, et al. Hepatitis C virus drug resistance–associated substitutions: state of the art summary. Hepatology. 2015;62(5):1623–1632.

6. McKimm-Breschkin JL. Resistance of influenza viruses to neuraminidase inhibitors—a review. Antiviral Research. 2000;47(1):1–17.

7. Alexander BD, Perfect JR. Antifungal resistance trends towards the year 2000. Drugs. 1997;54(5):657–678.

8. Kontoyiannis DP, Lewis RE. Antifungal drug resistance of pathogenic fungi. The Lancet. 2002;359(9312):1135–1144.

9. on Antimicrobial Resistance R. Tackling drug-resistant infections globally: final report and recommendations. Review on Antimicrobial Resistance; 2016.

10. Forum WE. The Global Risks Report 2018, 13th Edition. World Economic Forum; 2018.

11. Blair JM, Webber MA, Baylay AJ, Ogbolu DO, Piddock LJ. Molecular mechanisms of antibiotic resistance. Nature Reviews Microbiology. 2015;13(1):42.

12. Altmann A, Beerenwinkel N, Sing T, Savenkov I, Däumer M, Kaiser R, et al. Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. Antiviral Therapy. 2007;12(2):169.

13. Brenner BG, Wainberg MA. Clinical benefit of dolutegravir in HIV-1 management related to the high genetic barrier to drug resistance. Virus Research. 2017;239:1–9.

14. Deforche K, Cozzi-Lepri A, Theys K, Clotet B, Camacho RJ, Kjaer J, et al. Modelled in vivo HIV fitness under drug selective pressure and estimated genetic barrier towards resistance are predictive for virological response. Antiviral Therapy. 2008;13(3):399.

15. Devereux HL, Emery VC, Johnson MA, Loveday C. Replicative fitness in vivo of HIV-1 variants with multiple drug resistance-associated mutations. Journal of Medical Virology. 2001;65(2):218–224.

16. Andersson DI, Levin BR. The biological cost of antibiotic resistance. Current Opinion in Microbiology. 1999;2(5):489–493.

17. Andersson DI, Hughes D. Antibiotic resistance and its cost: is it possible to reverse resistance? Nature Reviews Microbiology. 2010;8(4):260.

18. Götte M. The distinct contributions of fitness and genetic barrier to the development of antiviral drug resistance. Current Opinion in Virology. 2012;2(5):644–650.

19. Mesplède T, Quashie PK, Osman N, Han Y, Singhroy DN, Lie Y, et al. Viral fitness cost prevents HIV-1 from evading dolutegravir drug pressure. Retrovirology. 2013;10(1):22.

20. Sibley CH, Hyde JE, Sims PF, Plowe CV, Kublin JG, Mberu EK, et al. Pyrimethamine–sulfadoxine resistance in Plasmodium falciparum: what next? Trends in Parasitology. 2001;17(12):570–571.

21. Zhou J, Price AJ, Halambage UD, James LC, Aiken C. HIV-1 resistance to the capsid-targeting inhibitor PF74 results in altered dependence on host factors required for virus nuclear entry. Journal of Virology. 2015;89(17):9068–9079.

22. Piana S, Carloni P, Rothlisberger U. Drug resistance in HIV-1 protease: flexibility-assisted mechanism of compensatory mutations. Protein Science. 2002;11(10):2393–2402.

23. Deeks SG, Wrin T, Liegler T, Hoh R, Hayden M, Barbour JD, et al. Virologic and immunologic consequences of discontinuing combination antiretroviral-drug therapy in HIV-infected patients with detectable viremia. New England Journal of Medicine. 2001;344(7):472–480.

24. Frost SD, Nijhuis M, Schuurman R, Boucher CA, Brown AJL. Evolution of lamivudine resistance in human immunodeficiency virus type 1-infected individuals: the relative roles of drift and selection. Journal of Virology. 2000;74(14):6262–6268.

25. Deeks SG, Hoh R, Neilands TB, Liegler T, Aweeka F, Petropoulos CJ, et al. Interruption of treatment with individual therapeutic drug classes in adults with multidrug-resistant HIV-1 infection. Journal of Infectious Diseases. 2005;192(9):1537–1544.

26. Rosenbloom DI, Hill AL, Rabi SA, Siliciano RF, Nowak MA. Antiretroviral dynamics determines HIV evolution and predicts therapy outcome. Nature Medicine. 2012;18(9):1378.

27. Nijhuis M, Schuurman R, De Jong D, Erickson J, Gustchina E, Albert J, et al. . Aids. 1999;13(17):2349–2359. doi:10.1097/00002030-199912030-00006.

28. zur Wiesch PS, Engelstädter J, Bonhoeffer S. Compensation of fitness costs and reversibility of antibiotic resistance mutations. Antimicrobial Agents and Chemotherapy. 2010;54(5):2085–2095.

29. Maisnier-Patin S, Andersson DI. Adaptation to the deleterious effects of antimicrobial drug resistance mutations by compensatory evolution. Research in Microbiology. 2004;155(5):360–369.

30. Bonhoeffer S, Chappey C, Parkin NT, Whitcomb JM, Petropoulos CJ. . Science. 2004;306(5701):1547–1550. doi:10.1126/science.1101786.

31. Hinkley T, Martins J, Chappey C, Haddad M, Stawiski E, Whitcomb JM, et al. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. Nature Genetics. 2011;43(5):487.

32. Starr TN, Thornton JW. Epistasis in protein evolution. Protein Science. 2016;25(7):1204–1218.

33. Michalakis Y, Roze D. Epistasis in RNA viruses. Science. 2004;306(5701):1492–1493.

34. Parera M, Perez-Alvarez N, Clotet B, Martínez MA. Epistasis among deleterious mutations in the HIV-1 protease. Journal of Molecular Biology. 2009;392(2):243–250.

35. Olson CA, Wu NC, Sun R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Current Biology. 2014;24(22):2643–2651.

36. Bank C, Hietpas RT, Jensen JD, Bolon DN. A systematic survey of an intragenic epistatic landscape. Molecular Biology and Evolution. 2014;32(1):229–238.

37. Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, et al. Local fitness landscape of the green fluorescent protein. Nature. 2016;533(7603):397.

38. Borrell S, Gagneux S. Strain diversity, epistasis and the evolution of drug resistance in Mycobacterium tuberculosis. Clinical Microbiology and Infection. 2011;17(6):815–820.

39. Trindade S, Sousa A, Xavier KB, Dionisio F, Ferreira MG, Gordo I. Positive epistasis drives the acquisition of multidrug resistance. PLoS Genetics. 2009;5(7):e1000578.

40. Silva RF, Mendonça SC, Carvalho LM, Reis AM, Gordo I, Trindade S, et al. Pervasive sign epistasis between conjugative plasmids and drug-resistance chromosomal mutations. PLoS Genetics. 2011;7(7):e1002181.

41. Yang WL, Kouyos RD, Böni J, Yerly S, Klimkait T, Aubert V, et al. Persistence of transmitted HIV-1 drug resistance mutations associated with fitness costs and viral genetic backgrounds. PLoS Pathogens. 2015;11(3):e1004722.

42. Fragata I, Blanckaert A, Louro MAD, Liberles DA, Bank C. Evolution in the light of fitness landscape theory. Trends in Ecology & Evolution. 2018;.

43. Cong Me, Heneine W, García-Lerma JG. The fitness cost of mutations associated with human immunodeficiency virus type 1 drug resistance is modulated by mutational interactions. Journal of Virology. 2007;81(6):3037–3041.

44. Martinez-Picado J, Savara AV, Sutton L, Richard T. Replicative fitness of protease inhibitor-resistant mutants of human immunodeficiency virus type 1. Journal of Virology. 1999;73(5):3744–3752.

45. Lv Z, Chu Y, Wang Y. HIV protease inhibitors: a review of molecular selectivity and toxicity. HIV/AIDS (Auckland, NZ). 2015;7:95.

46. Strack PR, Frey MW, Rizzo CJ, Cordova B, George HJ, Meade R, et al. Apoptosis mediated by HIV protease is preceded by cleavage of Bcl-2. Proceedings of the National Academy of Sciences. 1996;93(18):9571–9576.

47. Gougeon ML. Cell death and immunity: apoptosis as an HIV strategy to escape immune attack. Nature Reviews Immunology. 2003;3(5):392.

48. Velazquez-Campoy A, Kiso Y, Freire E. The binding energetics of first-and second-generation HIV-1 protease inhibitors: implications for drug design. Archives of Biochemistry and Biophysics. 2001;390(2):169–175.

49. Harrigan PR, Hogg RS, Dong WW, Yip B, Wynhoven B, Woodward J, et al. Predictors of HIV drug-resistance mutations in a large antiretroviral-naive cohort initiating triple antiretroviral therapy. The Journal of Infectious Diseases. 2005;191(3):339–347.

50. Lu Z. Second generation HIV protease inhibitors against resistant virus. Expert opinion on drug discovery. 2008;3(7):775–786.

51. Eshleman SH, Jones D, Galovich J, Paxinos EE, Petropoulos CJ, Jackson JB, et al. Phenotypic drug resistance patterns in subtype A HIV-1 clones with nonnucleoside reverse transcriptase resistance mutations. AIDS Research & Human Retroviruses. 2006;22(3):289–293.

52. De Meyer S, Vangeneugden T, Van Baelen B, De Paepe E, Van Marck H, Picchio G, et al. Resistance profile of darunavir: combined 24-week results from the POWER trials. AIDS Research and Human Retroviruses. 2008;24(3):379–388.

53. Barbour JD, Wrin T, Grant RM, Martin JN, Segal MR, Petropoulos CJ, et al. Evolution of phenotypic drug susceptibility and viral replication capacity during long-term virologic failure of protease inhibitor therapy in human immunodeficiency virus-infected adults. Journal of Virology. 2002;76(21):11104–11112.

54. Stoddart CA, Liegler TJ, Mammano F, Linquist-Stepps VD, Hayden MS, Deeks SG, et al. Impaired replication of protease inhibitor-resistant HIV-1 in human thymus. Nature Medicine. 2001;7(6):712.

55. Bangsberg DR, Moss AR, Deeks SG. Paradoxes of adherence and drug resistance to HIV antiretroviral therapy. Journal of Antimicrobial Chemotherapy. 2004;53(5):696–699.

56. Condra JH, Schleif WA, Blahy OM, Gabryelski LJ, Graham DJ, Quintero J, et al.. In vivo emergence of HIV-1 variants resistant to multiple protease inhibitors; 1995.

57. Dam E, Quercia R, Glass B, Descamps D, Launay O, Duval X, et al. Gag mutations strongly contribute to HIV-1 resistance to protease inhibitors in highly drug-experienced patients besides compensating for fitness loss. PLoS pathogens. 2009;5(3).

58. Chang MW, Torbett BE. Accessory mutations maintain stability in drug-resistant HIV-1 protease. Journal of molecular biology. 2011;410(4):756–760.

59. Robinson LH, Myers RE, Snowden BW, Tisdale M, Blair ED. HIV type 1 protease cleavage site mutations and viral fitness: implications for drug susceptibility phenotyping assays. AIDS research and human retroviruses. 2000;16(12):1149–1156.

60. Flynn WF, Chang MW, Tan Z, Oliveira G, Yuan J, Okulicz JF, et al. Deep sequencing of protease inhibitor resistant HIV patient isolates reveals patterns of correlated mutations in Gag and protease. PLoS computational biology. 2015;11(4).

61. Rhee SY, Taylor J, Fessel WJ, Kaufman D, Towner W, Troia P, et al. HIV-1 protease mutations and protease inhibitor cross-resistance. Antimicrobial Agents and Chemotherapy. 2010;54(10):4253–4261.

62. Brenner BG, Routy JP, Petrella M, Moisi D, Oliveira M, Detorio M, et al. Persistence and fitness of multidrug-resistant human immunodeficiency virus type 1 acquired in primary infection. Journal of Virology. 2002;76(4):1753–1761.

63. Johnson JA, Li JF, Wei X, Lipscomb J, Irlbeck D, Craig C, et al. Minority HIV-1 drug resistance mutations are present in antiretroviral treatment–naïve populations and associate with reduced treatment efficacy. PLoS Medicine. 2008;5(7):e158.

64. Barbour JD, Hecht FM, Wrin T, Liegler TJ, Ramstead CA, Busch MP, et al. Persistence of primary drug resistance among recently HIV-1 infected adults. Aids. 2004;18(12):1683–1689.

65. Flynn WF, Haldane A, Torbett BE, Levy RM. Inference of epistatic effects leading to entrenchment and drug resistance in HIV-1 protease. Molecular biology and evolution. 2017;34(6):1291–1306.

66. Biswas A, Haldane A, Arnold E, Levy RM. Epistasis and entrenchment of drug resistance in HIV-1 subtype B. eLife. 2019;8.

67. Qi H, Olson CA, Wu NC, Ke R, Loverdo C, Chu V, et al. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis C viral fitness and drug sensitivity. PLoS Pathogens. 2014;10(4):e1004064.

68. Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. Adaptation in protein fitness landscapes is facilitated by indirect paths. Elife. 2016;5:e16965.

69. Rhee SY, Liu T, Ravela J, Gonzales MJ, Shafer RW. Distribution of human immunodeficiency virus type 1 protease and reverse transcriptase mutation patterns in 4,183 persons undergoing genotypic resistance testing. Antimicrobial Agents and Chemotherapy. 2004;48(8):3122–3126.

70. HIV Databases;. Available from: http://www.hiv.lanl.gov/.

71. Boucher JI, Whitfield TW, Dauphin A, Nachum G, Hollins III C, Zeldovich KB, et al. Constrained mutational sampling of amino acids in HIV-1 protease evolution. Molecular biology and evolution. 2019;36(4):798–810.

72. Parera M, Fernandez G, Clotet B, Martinez MA. HIV-1 protease catalytic efficiency effects caused by random single amino acid substitutions. Molecular Biology and Evolution. 2006;24(2):382–387.

73. Du Y, Zhang TH, Dai L, Zheng X, Gorin AM, Oishi J, et al. Effects of mutations on replicative fitness and major histocompatibility complex class I binding affinity are among the determinants underlying cytotoxic-T-lymphocyte escape of HIV-1 gag epitopes. mBio. 2017;8(6):e01050–17.

74. Al-Mawsawi LQ, Wu NC, Olson CA, Shi VC, Qi H, Zheng X, et al. High-throughput profiling of point mutations across the HIV-1 genome. Retrovirology. 2014;11(1):124.

75. Sanjuan R, Moya A, Elena SF. . Proceedings of the National Academy of Sciences. 2004;101(43):15376–15379. doi:10.1073/pnas.0404125101.

76. Silander OK, Tenaillon O, Chao L. Understanding the evolutionary fate of finite populations: the dynamics of mutational effects. PLoS Biology. 2007;5(4):e94.

77. Dai L, Du Y, Qi H, Wu NC, Lloyd-Smith JO, Sun R. Quantifying the evolutionary potential and constraints of a drug-targeted viral protein. bioRxiv. 2016; p. 078428.

78. Parera M, Martinez MA. Strong epistatic interactions within a single protein. Molecular Biology and Evolution. 2014;31(6):1546–1553.

79. Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. Negative epistasis between beneficial mutations in an evolving bacterial population. Science. 2011;332(6034):1193–1196.

80. Elena SF. Little evidence for synergism among deleterious mutations in a nonsegmented RNA virus. Journal of Molecular Evolution. 1999;49(5):703–707.

81. Goepfert PA, Lumm W, Farmer P, Matthews P, Prendergast A, Carlson JM, et al. Transmission of HIV-1 Gag immune escape mutations is associated with reduced viral load in linked recipients. Journal of Experimental Medicine. 2008;205(5):1009–1017.

82. Sierra S, Dávila M, Lowenstein PR, Domingo E. Response of foot-and-mouth disease virus to increased mutagenesis: influence of viral load and fitness in loss of infectivity. Journal of Virology. 2000;74(18):8316–8323.

83. Zhang H, Yang B, Pomerantz RJ, Zhang C, Arunachalam SC, Gao L. The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. Nature. 2003;424(6944):94.

84. Harris RS, Bishop KN, Sheehy AM, Craig HM, Petersen-Mahrt SK, Watt IN, et al. DNA deamination mediates innate immunity to retroviral infection. Cell. 2003;113(6):803–809.

85. Mangeat B, Turelli P, Caron G, Friedli M, Perrin L, Trono D. Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. Nature. 2003;424(6944):99.

86. Crotty S, Cameron CE, Andino R. RNA virus error catastrophe: direct molecular test by using ribavirin. Proceedings of the National Academy of Sciences. 2001;98(12):6895–6900.

87. Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. Immunity. 2013;38(3):606–617.

88. Mann JK, Barton JP, Ferguson AL, Omarjee S, Walker BD, Chakraborty A, et al. The Fitness Landscape of HIV-1 Gag: Advanced Modeling Approaches and Validation of Model Predictions by In Vitro Testing. PLOS Computational Biology. 2014;10(8):1–11. doi:10.1371/journal.pcbi.1003776.

89. Butler TC, Barton JP, Kardar M, Chakraborty AK. Identification of drug resistance mutations in HIV from constraints on natural evolution. Physical Review E. 2016;93(2):022412.

90. Louie RH, Kaczorowski KJ, Barton JP, Chakraborty AK, McKay MR. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. Proceedings of the National Academy of Sciences. 2018;115(4):E564–E573.

91. Barton JP, De Leonardis E, Coucke A, Cocco S. ACE: adaptive cluster expansion for maximum entropy graphical model inference. Bioinformatics. 2016;32(20):3089–3097.

92. Solis M, Nakhaei P, Jalalirad M, Lacoste J, Douville R, Arguello M, et al. RIG-I-mediated antiviral signaling is inhibited in HIV-1 infection by a protease-mediated sequestration of RIG-I. Journal of Virology. 2011;85(3):1224–1236.

93. Shah P, McCandlish DM, Plotkin JB. Contingency and entrenchment in protein evolution under purifying selection. Proceedings of the National Academy of Sciences. 2015;112(25):E3226–E3235.

94. Draghi JA, Plotkin JB. Selection biases the prevalence and type of epistasis along adaptive trajectories. Evolution. 2013;67(11):3120–3131.

95. Kitayimbwa JM, Mugisha JYT, Saenz RA. . Theoretical Population Biology. 2016;112:33–42. doi:10.1016/j.tpb.2016.08.001.

96. Wensing AJ, van de Vijver D, Angarano G, Åsjö B, Balotta C, Boeri E, et al. . The Journal of Infectious Diseases. 2005;192(6):958–966. doi:10.1086/432916.

97. Roberts JD, Bebenek K, Kunkel TA. The accuracy of reverse transcriptase from HIV-1. Science. 1988;242(4882):1171–1173.

98. Cuevas JM, Geller R, Garijo R, López-Aldeguer J, Sanjuán R. Extremely high mutation rate of HIV-1 in vivo. PLoS Biology. 2015;13(9):e1002251.

99. Han TX, Xu XY, Zhang MJ, Peng X, Du LL. Global fitness profiling of fission yeast deletion strains by barcode sequencing. Genome biology. 2010;11(6):R60.

100. Van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nature methods. 2009;6(10):767.

101. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science. 1996;271(5255):1582–1586.

102. Fernandes JD, Faust TB, Strauli NB, Smith C, Crosby DC, Nakamura RL, et al. Functional segregation of overlapping genes in HIV. Cell. 2016;167(7):1762–1773.

103. Weinberger ED. Fourier and Taylor series on fitness landscapes. Biological cybernetics. 1991;65(5):321–330.

104. Uguzzoni G, Lovis SJ, Oteri F, Schug A, Szurmant H, Weigt M. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. Proceedings of the National Academy of Sciences. 2017;114(13):E2662–E2671.

105. Levy RM, Haldane A, Flynn WF. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. Current opinion in structural biology. 2017;43:55–62.

106. Chen L, Perlina A, Lee CJ. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. Journal of Virology. 2004;78(7):3722–3732.

107. He X, Qian W, Wang Z, Li Y, Zhang J. Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. Nature genetics. 2010;42(3):272.

108. Yerly S, Kaiser L, Race E, Bru JP, Clavel F, Perrin L. Transmission of antiretroviral-drug-resistant HIV-1 variants. The Lancet. 1999;354(9180):729 – 733. doi:https://doi.org/10.1016/S0140-6736(98)12262-6.

109. Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. Nature. 2006;444(7121):929.

110. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins: Structure, Function, and Bioinformatics. 2011;79(3):830–838.

111. Barton JP, Goonetilleke N, Butler TC, Walker BD, McMichael AJ, Chakraborty AK. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nature Communications. 2016;7:11660.

Chapter 3

Increasing abundance of SARS-CoV-2 N protein during transmission in human revealed by a high-throughput mutagenesis screening

1. Abstract

The continuous evolution of SARS-CoV-2 in human population adds great uncertainties in COVID-19 control. Even though mutations on N protein are prevalent in emerging viral variants, their effects were not well studied. To understand the evolutionary potential of the N protein, we quantified the mutational effects of all possible substitutions and found a trend of increasing N protein abundance during SARS-CoV-2 evolution in human.

2. Introduction

Since the outbreak of COVID19 in late 2019, SARS-CoV-2, the causative viral pathogen, continuously expands its population and spread to almost all regions of the world. New variants of SARS-CoV-2 were frequently identified, some of which result in advantage in transmission and quickly become dominant in collected sequences[1-3]. Mutations on the S protein were the most extensively studied, as interactions of the S protein with cellular host receptors greatly influence the viral infectivity[4], in addition to S protein being the target of antibody-based COVID19 therapeutics as well as the antigen for the majority of the approved vaccines. Nevertheless, many mutations appear on other regions of the SARS-CoV-2 genome, and the implication of these mutations on virulence and transmission are largely unknown[5]. N protein is one of the most abundant viral proteins in SARS-CoV-2. This highly conserved protein has important functions in multiple steps of viral life cycles, and contains multiple epitopes that can elicit T cell and antibody responses for viral control[6-9]. Therefore, N protein has been proposed to be the target of T cell-based vaccines that can provide additional layers of protection. However, current

studies are lacking in comprehensively understanding the constrains or impact of N protein evolution.

3. Results

3.1 Profile the mutational effect on the N protein

Given the importance of N protein, it has been proposed that the availability of N protein is a rate-limiting step in the life cycle of some coronaviruses [10,11]. To test whether this is the case for SARS-CoV-2, we overexpressed N protein under an inducible promoter in Vero E6 cells before SARS-CoV-2 viral infection, and observed an increase of viral titer at 24- and 48-hours post-infection (Figure 1A). This led us to the hypothesis that the evolution of SARS-CoV-2 may be constrained by abundance of the N protein. To profile the impact of each codon substitution in N protein abundance, we developed a flow cytometry-based high-throughput screening platform (Extended Data Figure 1A). Lentivirus-based libraries with all possible single amino acid substitutions on SARS-CoV-2 N were constructed. The viral protein was expressed in a human lung epithelial cell line (A549) and cells were sorted based on the fluorescence conjugated antibody against the FLAG-tag, indicative of the abundance of N protein in cells. Four populations with different N protein abundance were deep-sequenced separately and the inferred fluorescence intensity (mutational effect) of each mutant was calculated according to its frequency in different populations. The mutational effect (*ME*) was defined by the following equation: $ME = \frac{\sum_i^4 f_i \times i}{\sum_i^4 f_i}$, where f_i stands for the frequency of each mutation in the population i. We observed a high reproducibility among 3 independent experiments (Extended Data Figure 1B), and a clear separation between the synonymous mutations and the nonsense mutations (Extended Data Figure 1C). The mutational effects of a total of 7707

substitutions were quantified (Extended Data Table1). We found that a broad range of mutations are beneficial or neutral in SARS-CoV-2 N protein (Figure 1B). Overall, 1.89% (133) missense mutations increased the abundance (2 standard deviation higher than the abundance of silent mutations) and 6.68% (469) mutations decreased its abundance (2 standard deviation lower than the abundance of silent mutations). We also noted different regions have different tolerance to mutations (Figure 1C). The N-arm and linker regions are more tolerant to mutations, while other patches, such as residues 72-76 and residues 112-117, are less tolerant to mutations, indicating that these regions could serve as potential vaccine targets.

To further demonstrate the accuracy of our dataset, we randomly picked 5 destabilizing and 3 neutral mutations, and individually measured the mutant protein abundance by Western blot. Protein abundance detected by Western blot correlated significantly with the profiling result (Figure 1E, quantified in Extended Data Figure 2C). Mapping the mutational effect onto the available N terminal domain (NTD) and C terminal domain (CTD) structures, we observed that the residues on the surface of the N protein were more tolerant to mutations (Extended Data Figure 1E, quantified in Extended Data Figure 1F). Protein thermo-stability was one of the main determining factors of protein abundance. Therefore, we calculated the correlation between the first-principle predicted protein stability and our experiment data for protein abundance. The destabilizing effect of all possible substitutions on the NTD and CTD of the N protein were calculated using Rosetta with the corresponding crystallography structures (PDB ID: 7cdz and 7ce0). Consistent with previous reports[12], mutations with large destabilizing effects were mostly deleterious (Figure 1D). These results demonstrated the accuracy of our dataset that

generated a comprehensive annotation of the impact of each amino acid substitution on N protein abundance.

3.2. Estimate the impact of N protein mutations on natural occurring variants.

With a comprehensive mutational effect dataset, we seek to test whether N protein abundance constrain viral evolution by analyzing the sequences of naturally occurring mutant viruses. We retrieved 237,226 SARS-CoV-2 whole genome sequences from GISAID. Mutations in the genome region encoding N protein were accumulating since early 2020 (Extended Data Figure 2A). Consistent with our profiling data, the linker region showed the highest mutation occurrence (Figure 2A). There are 86 double mutants that can be experimentally examined in our dataset. We found the mutational effect of 91% of these mutants is close to the sum-up of single mutations (within 33% standard deviation, Extended Data Figure 1D). This finding validated the usage of our dataset to infer the mutational effect of naturally occurring N variants. We found that the N protein abundance of natural occurring variants were significantly correlated with the time of sampling (Figure 2B), suggesting that a directional evolution on N protein abundance had taken place. Moreover, we plotted the correlation coefficient between the N mutation monthly occurrence and our abundance dataset (Figure 2C, Extended Data Figure 2B), which showed a positive correlation as early as mid-2020, indicating an early onset of the directional evolution. To confirm this finding, we selected several prevalent natural N variants and assessed their abundance in Vero-E6 cells by Western blot (Figure 2D). We found many of them with elevated protein abundance as compared to the initial N protein. For example, variant D3L_R203K_G204R_S235F N derived from dominant lineage B.1.1.7, showed increased abundance by 2.1-fold in cells. Our data suggest that the N

protein abundance per se, even in the absence of infection, is increasing in the real world due to naturally occurring mutations and may contribute to SARS-CoV-2 evolution.

4. Discussion

In summary, we provided a comprehensive and accurate dataset of SARS-CoV-2 N mutant abundance. Our dataset suggests that new N variants have impacts on viral replication and evolution. This analysis could also be applied to the deep mutational scanning datasets on other viral proteins[13]. The receptor binding domain of the S protein showed an increasing trend of ACE2 binding capacity recently (Extended Data Figure 2D). These trends indicate different time and direction of the selection pressures on SARS-CoV-2. Understanding the direction of evolution can help us better design anti-viral drugs and vaccines to interrupt the viral life cycle. Moreover, it has been shown in several cases that elevated protein abundance plays important roles in evolution by enabling new or improved functions[14]. Future experiments are needed to directly test different protein functions of the new N variants. Lastly, our data suggest the importance of surveillance and functional testing of other SARS-CoV-2 proteins, in addition to S, for effective pandemic control.

5. Methods

5.1. N protein mutant library construction

SARS-CoV-2 N gene was divided into 6 sub-libraries, each containing 71 target residues. An oligonucleotide containing NNK (N stands for any nucleotide, K stands for guanine and thymine) was synthesized to introduce all possible single amino acid substitutions to a certain residue. For each sub-library, 71 mutated oligonucleotides were synthesized. A total of 418 oligonucleotides were synthesized to cover all residues on the N gene. The

sequences of all the oligonucleotides can be found in the Extended Data Table 2. The mutated oligonucleotides were then ligated with the wild-type oligonucleotides by PCR to create a full-length N gene fragment. The N gene fragments were cloned into a lentiviral vector with a Tet-on promoter and FLAG-tag on the C-terminal of the open reading frame. For each sub-library, more than 30,000 *E. coli* clones were harvested to ensure the sufficient coverage (>10-fold) of each mutant. The colonies were scraped from the plates and purified by the Invitrogen HiPure Plasmid purification kit. We used lipofectamine 2000 to transfect each plasmid library with lentiviral packaging plasmid (PAX2) and envelope plasmids (pVSV-G). Two million 293T cells were used for each library to conserve library complexity. The lentiviral library was harvested 48 hours after transfection. DNase I (40ng/mL) and $MgSO_4$ (1mM) were added to the library to remove residual plasmid DNA from the supernatant. Three independent transfections were performed to ensure reproducibility.

For each sub-library, 20 million A549 cells were transduced with 100uL lentiviral library. The multiplicity of infection (MOI) is ~0.1. This reduces the possibility of superinfection. Puromycin was added to A549 cells and maintained for 7 days. Three independent transductions and selection were carried out. The remaining cells were used as the mutant N protein-expressing cell libraries in the subsequent experiments.

5.2. Flow cytometry

Doxycycline (1ug/mL) was added to the cells 24 hours before staining. The cells were fixed by 2% Formaldehyde and penetrated by eBioscience Intracellular Fixation and Permeabilization buffer. The cells were stained with PE/Cyanine7 anti-FLAG tag antibody (200ng/mL) at room temperature for 30 minutes. The cells were then washed twice and

loaded to the sorting machine. A total of 10 million cells were sorted for each sub-library and each replicate. Every sample was sorted into 4 populations according to the PE/Cyanine7 intensity.

5.3. Sequencing library preparation

The total DNA was extracted from each sorted cell population. The corresponding mutated regions were enriched by PCR. The products were ligated with NEB sequencing adapters and subject to the next-generation sequencing. The sequencing libraries were run on the Illumina NovaSeq6000 platform with paired-end 250 bp read settings. A total of 800 million reads were sequenced. All raw data can be found on NCBI short read archive under the accession ID PRJNA740111.

5.4. Data analysis

The amplicons for each sub-library were retrieved by comparing reads with PCR primers. If both primers were mapped in a pair of reads, and the difference between primers and the mapped regions is smaller than 3 nucleotides, the read pairs were identified as the N gene. The forward and the reverse reads were then compared with the initial variant individually. If a mutation was observed in both read directions, it will be identified as a true mutation. Read pairs were then translated according to the N gene open reading frame. The read pair was discarded if it has more than 10 amino acid substitutions. The occurrence of each type of mutation was quantified by the count of corresponding read pairs. The frequency is normalized to the average frequency of all silent mutations with the sub-library. The mutational effect of each single amino acid substitution and double mutants were calculated as described in the main text.

The protein stability of single mutations on NTD and CTD (ΔΔG) was predicted by PyRosetta. The PDB files 7CDZ and 7CE0 were cleaned and trimmed to single-chain atoms. Then, all side chains were repacked and minimized using the score function ddg_monomer. Then we introduced all possible substitutions within the structures and repacked all atoms within 8Å using linmin mover. The new structure was scored again and the difference between the new score and the score of the initial variant was used as ΔΔG. The procedure was repeated 10 times and the mean results were used in the correlation analysis.

The aligned sequences of naturally occurring variants were downloaded from the GISAID. All sequences were translated according to the reference genome WIV04. Sequences with truncated N gene or undetermined nucleotides were discarded. Sequences with unclear sampling time annotation were also discarded. Mutations were called by comparing the sequence with WIV04. The mutational effect on a naturally observed N protein variant was predicted as the product of mutational effects of all included single mutations. If a substitution is not observed in our profiling data, its mutational effect was set to 1.

All statistics were done by SciPy. Protein structures were colored and visualized by Chimera. All custom codes are available upon request.

5.5. Western blot

Representative mutations were constructed on the lentiviral vector mentioned above. The lentiviruses with different nucleocapsid mutations were packaged and used for infecting Vero-E6 cells with MOI = 0.1. After 7 days of puromycin selection, cells were treated with Doxycycline (1ug/mL) and subject to protein extraction. Cells were lysed by RIPA buffer

for 10 minutes on ice. Cell debris and supernatant were collected and denatured in the LDS sample buffer. 10uL cell lysate was loaded on the SDS-PAGE gel and transferred to the PVDF membrane. The membrane was stained by the anti-actin antibody and anti-FLAG tag antibody for 1 hour. The HRP conjugated secondary antibody was also incubated for 1 hour. The membrane was visualized using Radiance Q HRP substrate for quantitative western.

5.6. Infection

Vero-E6 cells expressing N protein were plated in 12-well plates and induced by Doxycycline (1ug/mL). 100uL virus was added to each well and the supernatant was collected every day after the infection for 2 days. The virus genome was extracted using Qiagen Viral RNA mini kit and quantified by RT-qPCR.

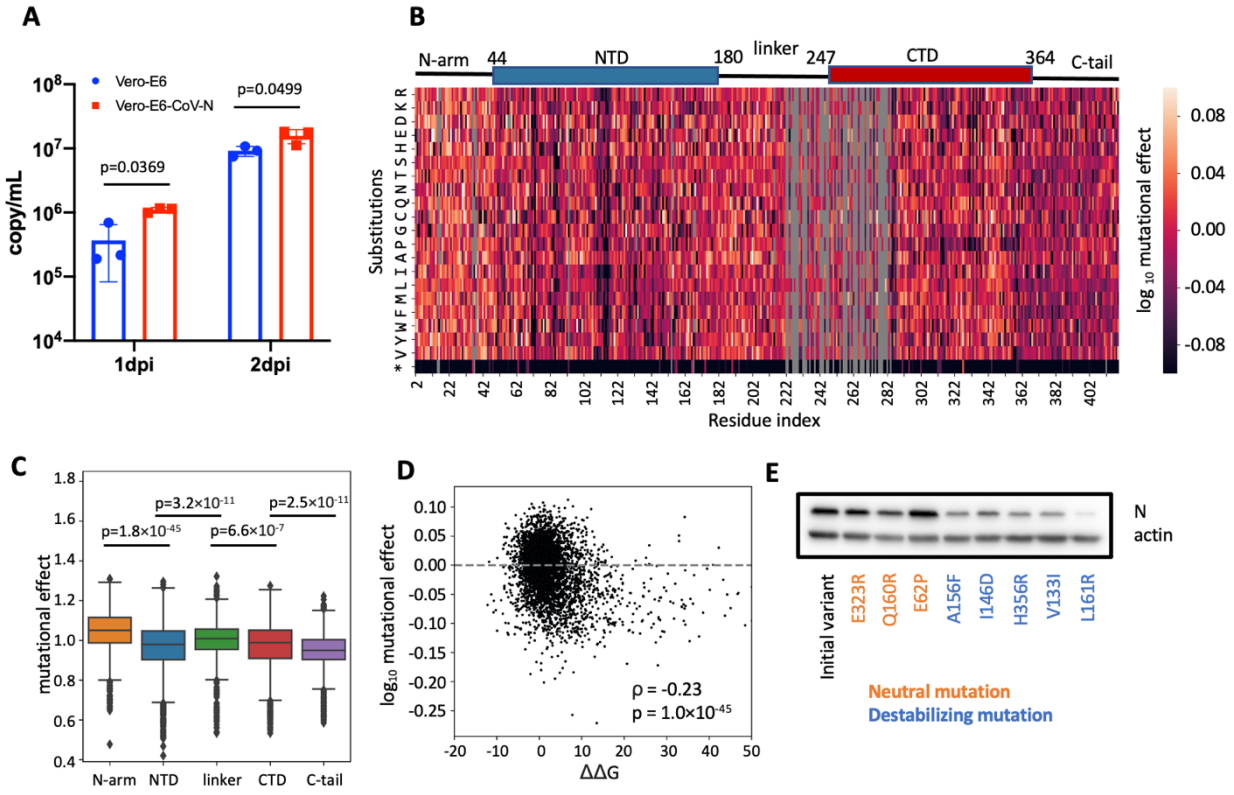


Figure 3-1. High-throughput profiling of the mutational effect on SARS-CoV-2 N protein.

A) Viral replication in Vero-E6 and Vero-E6-CoV-N cells. The cells were infected 1 day after induction of N protein expression. B) Heatmap showing the mutational effect of all possible single amino acid substitutions. C) Boxplot showing the median mutational effect in each domain. D) *In silico* validation of the profiling result. All possible single amino acid substitutions on the N protein NTD (PDB ID: 7cdz) and CTD (PDB ID: 7ce0) were modelled and their impact on protein stability were inferred by Rosetta. E) Western blot

validation of the N protein abundance in transduced Vero-E6 cells. NTD, N-terminal domain. CTD, C-terminal domain.

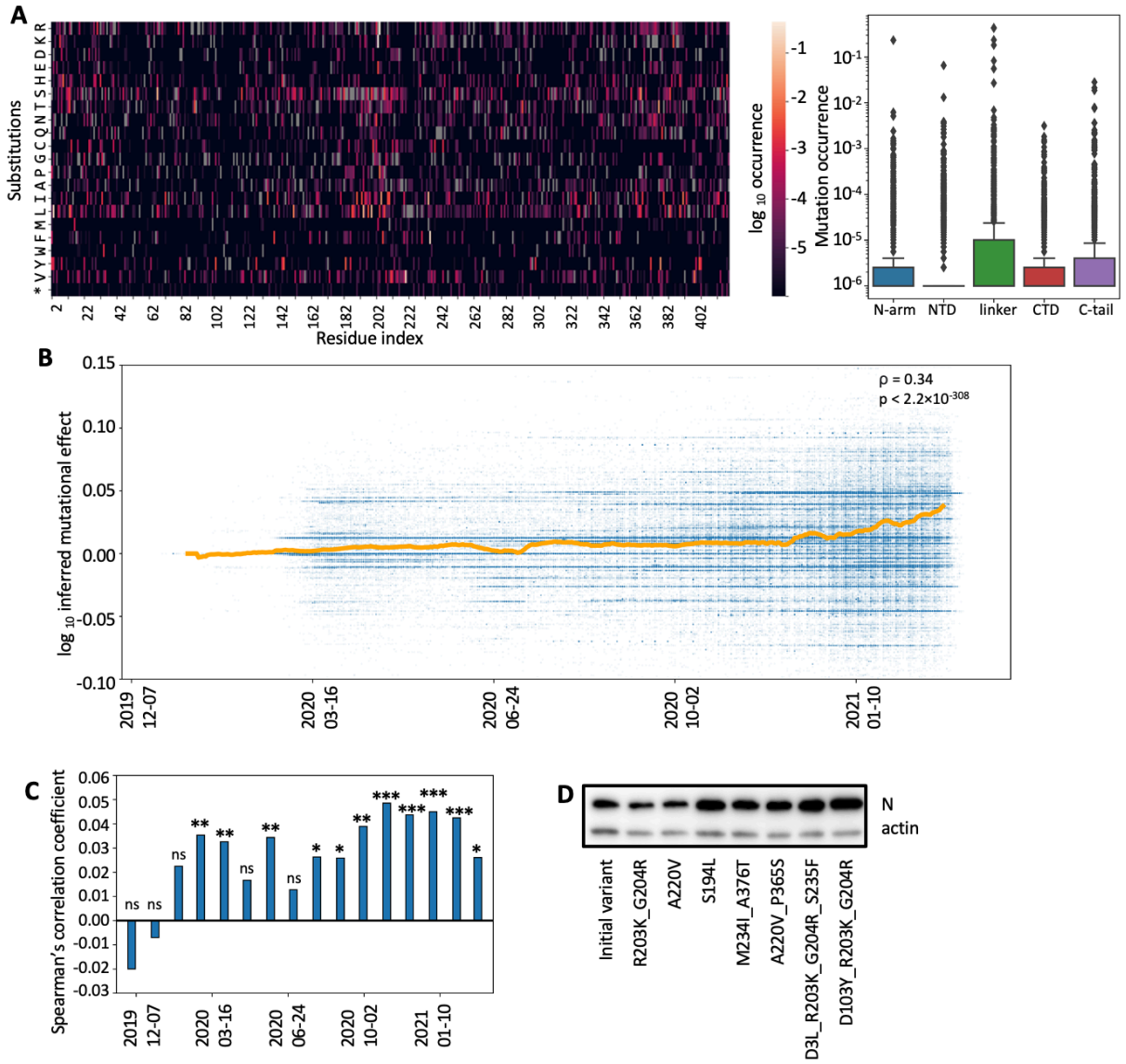
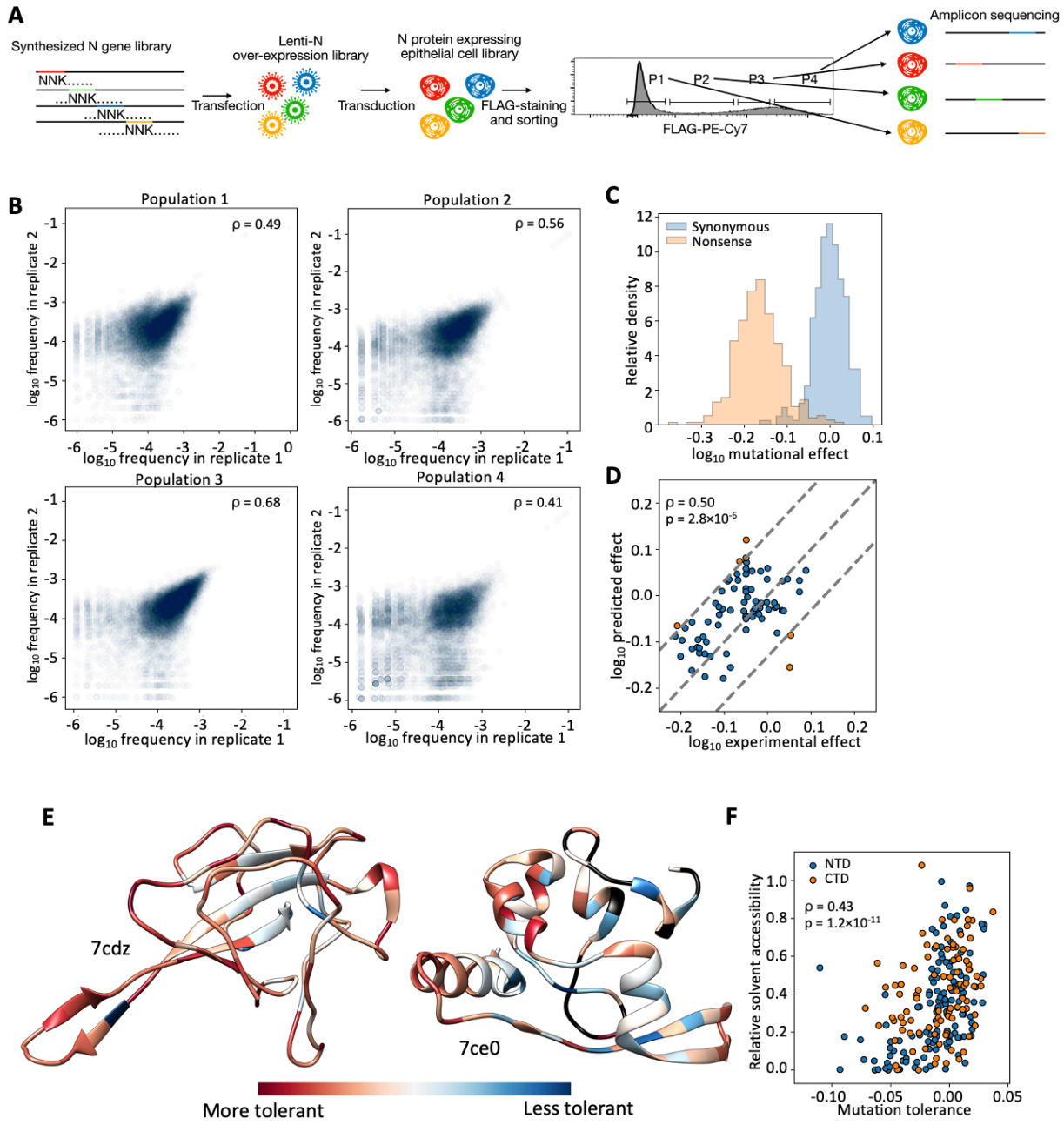


Figure 3-2. Naturally occurring mutant variants have higher N protein abundance.

A) Heatmap showing the frequencies of naturally occurring mutations in the GISAID database. Boxplot showing the median and max mutation occurrence in each domain. B) Inferred total mutational effect of all naturally occurring variants in the GISAID database. ρ is the Spearman's correlation coefficient. C) Monthly correlation between mutational

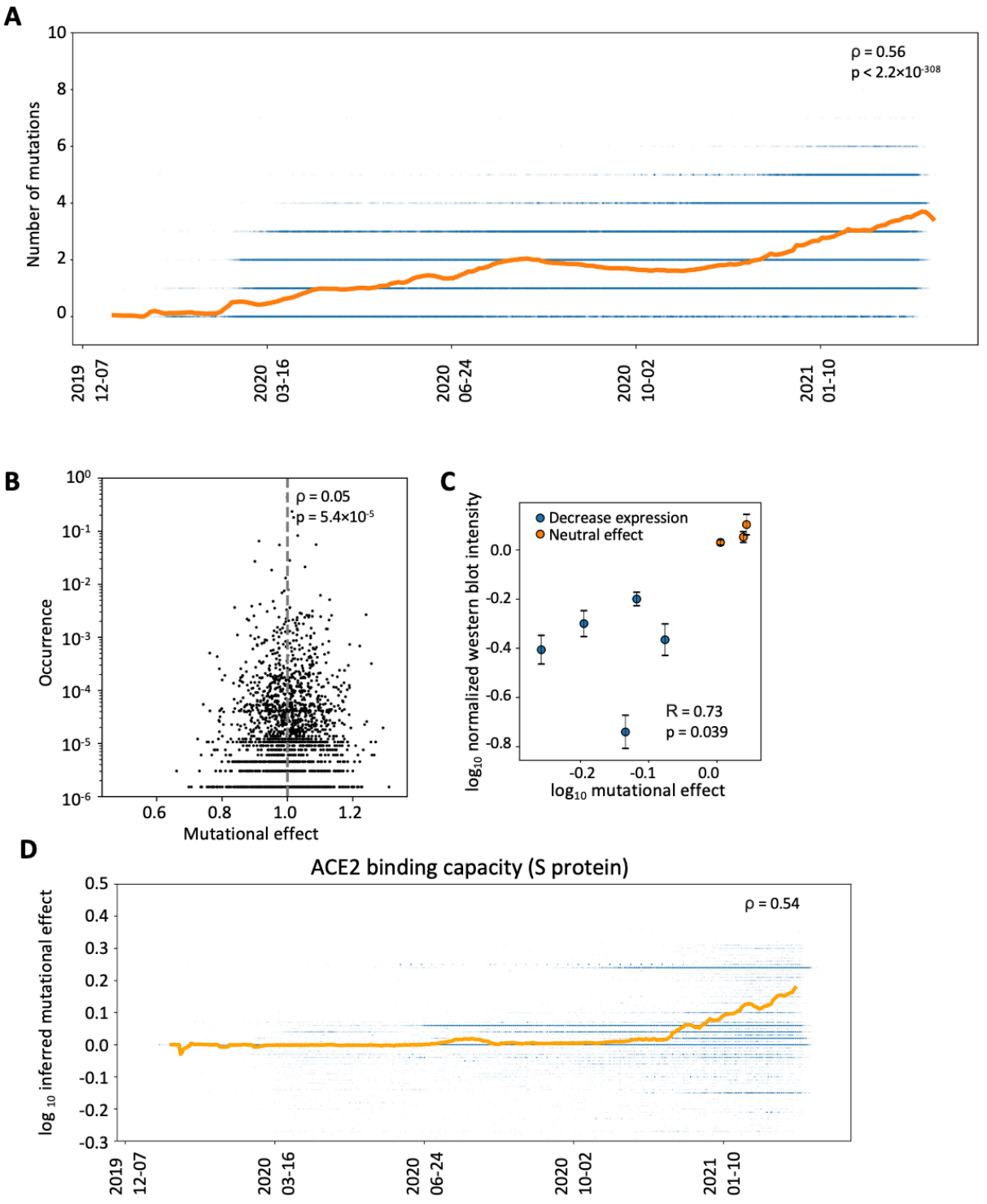
effects and its natural occurrence. D) Individual validation of the mutational effect on N protein abundance in transduced Vero-E6 cells.



Supplementary Figure 3-1. Quality of the mutational effect profiling.

A) Diagram showing the workflow of mutational effect profiling. B) Scatter plot showing the correlation of mutation frequency in each sorted population. Two biological replicates were shown. Rho is the Spearman's correlation coefficient. C) Histogram showing the

distribution of mutational effect for synonymous mutations and nonsense mutations. D) The correlation between experimental mutational effect and inferred mutational effect for high-order mutants. The inferred mutational effect is the summation of all corresponding single mutations' effects. E) Mutation tolerance (mean log mutational effect of all substitutions) of each residue. F) The correlation between the mutation tolerance and the relative solvent accessibility.


A
Number of mutations
ρ = 0.56
p < 2.2×10-308
2019 12-07
2020 03-16
2020 06-24
2020 10-02
2021 01-10
B
Occurrence
ρ = 0.05
p = 5.4×10-5
Mutational effect
C
Decrease expression
Neutral effect
log10 normalized western blot intensity
R = 0.73
p = 0.039
log10 mutational effect
D
ACE2 binding capacity (S protein)
log 10 inferred mutational effect
ρ = 0.54

Supplementary Figure 3-2. Characteristics of naturally occurring mutations.

A) Natural SARS-CoV-2 variants accumulated mutations on the N gene over time. B) Scatter plot showing the correlation between experimental mutational effect and natural occurrence. C) Correlation between mutational effect in the high-throughput screening and the protein abundance quantified by the Western blot. Intensities from 3 independent experiments were averaged. D) Inferred ACE2 binding capacity for S protein variants with mutant receptor binding domain.

6. Reference

1. Davies, Nicholas G., et al. "Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England." *Science* (2021).

2. Davies, Nicholas G., et al. "Increased mortality in community-tested cases of SARS-CoV-2 lineage B. 1.1. 7." *Nature* 593.7858 (2021): 270-274.

3. Plante, Jessica A., et al. "Spike mutation D614G alters SARS-CoV-2 fitness." *Nature* 592.7852 (2021): 116-121.

4. Starr, Tyler N., et al. "Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding." *Cell* 182.5 (2020): 1295-1310.

5. Wu, Aiping, et al. "One year of SARS-CoV-2 evolution." *Cell Host & Microbe* 29.4 (2021): 503-507.

6. Grifoni, Alba, et al. "Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals." *Cell* 181.7 (2020): 1489-1501.

7. Le Bert, Nina, et al. "SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls." *Nature* 584.7821 (2020): 457-462.

8. Shrock, Ellen, et al. "Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity." *Science* 370.6520 (2020).

9. Long, Quan-Xin, et al. "Antibody responses to SARS-CoV-2 in patients with COVID-19." *Nature medicine* 26.6 (2020): 845-848.

10. Cong, Yingying, et al. "Nucleocapsid protein recruitment to replication-transcription complexes plays a crucial role in coronaviral life cycle." *Journal of virology* 94.4 (2020).

11. Hurst, Kelley R., et al. "An interaction between the nucleocapsid protein and a component of the replicase-transcriptase complex is crucial for the infectivity of coronavirus genomic RNA." *Journal of virology* 84.19 (2010): 10276-10288.

12. Leman, Julia Koehler, et al. "Macromolecular modeling and design in Rosetta: recent methods and frameworks." *Nature methods* 17.7 (2020): 665-680.

13. Starr, Tyler N., et al. "Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding." *Cell* 182.5 (2020): 1295-1310.

14. Bloom, Jesse D., et al. "Protein stability promotes evolvability." *Proceedings of the National Academy of Sciences* 103.15 (2006): 5869-5874.

Chapter 4

Genetic barcoded HIV-1 revealed the correlation between integration, transcription and splicing

1. Abstract

The position of the HIV-1 integration site is a critical factor that impacts both the prognosis of the disease and the responsiveness of the virus to latency reversal agents. This is due to the fact that the local chromatin accessibility to the host transcription machinery varies in different genomic positions, affecting virus transcription. To investigate this phenomenon, we employed a viral genetic barcode system and a novel single cell sequencing method to simultaneously sequence the viral integration site and viral mRNA splicing. Our findings demonstrate that the position of the integration site plays a crucial role in determining the fate of the provirus by affecting viral mRNA splicing. Additionally, our study reveals the impact of different latency reversal agents on viral mRNA splicing activity at the single cell level. These results provide new insights into the mechanisms of HIV-1 latency and the potential for developing new therapies that target specific integration sites to effectively treat HIV-1 infection.

2. Introduction

2.1. The position of HIV-1 integration site affect the fate of the virus

HIV-1 latency is characterized by devoid of viral gene transcription and translation. The fate of the provirus is not only affected by the cellular status but also by the starting state of the proviral transcription and stochastic effect[1]. In latently infected patients, different types of CD4+ T cells have different amounts of proviruses. Latent viruses were mostly found in central memory T cells and transitional memory T cells. While naive T cells and effector memory T cells contribute less to the latent reservoir[2]. The activation status of the infected cells also affects HIV-1 transcription. For example, TNFα signal will activate

virus transcription by recruiting NF-κB heterodimer to the viral promoter[3]. However, chronic HIV-1 infection also leads to chronic immune activation, which induces inhibitory pathways of T cell activation and viral transcription, such as PD1 and CTLA4[4].

HIV-1 transcription is also affected by its intrinsic regulation program. An actively transcribing provirus will maintain its activity even if the host cell enters a resting state[5]. This is achieved by a positive feedback loop regulating by viral encoded transcription factor tat. If the basal activity of HIV-1 promoter passes a certain threshold, it will express a small amount of tat, which can recruit pTEF-b and destine the virus to an active state[6].

The position of the provirus integrated in the host genome also affects the viral transcription[7]. HIV-1 with higher transcription activity is more frequently found in host gene transcription active regions[8,9]. This may be explained by the high order nucleus architectures which distribute transcription apparatus unevenly, or by the dynamic positioning of the histones from nearby genes. For example, proviruses near the nuclear membrane have higher transcription activities[10]. Besides, if the flanking regions of the virus integration site have an opposite oriented host gene, the viral transcription will also be reduced due to transcriptional interference[11,12].

The responsiveness of the HIV-1 transcription to latency reversal agents (LRAs) may also have positional effects. LRAs like HDAC inhibitors act on the 3 nucleosomes located at the HIV-1 LTR promoter[13]. PKC agonists also induce histone acetylation at the same region[14]. However, the initial status of the nucleosomes and the availability of histone modification enzymes is affected by the high order nucleus architectures. As a result, LRAs with different mechanisms will reactivate different groups of proviruses[9].

The HIV-1 integration sites are not completely random. They are found more frequently in transcription active regions[15] and splicing frequent regions[16]. The HIV-1 preintegration complex is guided to the integration site by the host factor LEDGF, which is a chromatin associated protein enriched at the H3K36 trimethylation sites[17]. This mechanism ensures HIV-1 provirus has access to an abundance of host transcription and splicing machinery. Meanwhile, proviruses from different integration sites still have magnitudes of variations in transcription activity[6]. This allows HIV-1 to have massive intrahost diversity in terms of replication capacity. Thus some proviruses will enter dormancy, leading to a long, persistent, but sometimes reversible latency[18].

2.2. HIV-1 life cycle is tightly regulated by alternative splicing

The splicing activities of retroviruses are under delicate balance. Unwanted or inaccurate splicing leads to unproductive life cycles, as observed in human endogenous retroviruses[19]. Cryptic splicing resulting in non-functional ORF was also observed in HIV-1[20]. On the other hand, over-splicing leads to insufficient expression of viral structural genes[21], inhibiting virus replication. While inhibiting viral specific splicing via SF2/ASF also inhibited retroviral replication[22].

The balance between different splicing sites' activity is achieved by cis-acting splicing elements. Many exonic splicing silencers (ESSs), exonic splicing enhancers (ESEs) and intronic splicing silencers (ISSs) can recruit splicing factors and hnRNP to the splicing sites, adjusting the relative splicing activities[23]. Moreover, viruses can also produce RNA binding proteins with nuclear export signals that hijacks the CRM1 or NXF1 pathways. This allows unspliced RNA to be exported to cytoplasm before the spliceosome has a chance to process it[24].

Compared to a classic promoter-based gene regulation system, splicing is more economical for a virus with a smaller genome and higher mutation rates. The simplest splicing site can be achieved by a dinucleotide AG and another dinucleotide GU flanking the intron sequence. A cis-acting enhancer or silencer is around 3-8 nucleotides. A total of 20 to 30 nucleotides can faithfully produce the RNA isoform. A viral encoded RNA export gene is around 300 nucleotides long[25]. Some retroviruses do not have their own RNA export protein. Instead, a cis-acting RNA exporting sequence for unspliced RNA is only a few tandem repeats of 15 nucleotides[26]. In contrast, eukaryotic promoters are around 100 to 1000 nucleotides long. A viral encoded transcriptional factor gene can be 300 to 3000 nucleotides long. Some herpesviruses need more than 15 transcription factors to regulate their expression programs[27].

Hijacking host spliceosomes for temporal regulation of viral gene expression is also very responsive to the host environment. Taking HIV-1 as an example, the fate of the provirus is decided by the amount of multi-spliced RNA isoforms, which is affected by the availability of host transcription and splicing factors[28]. On the other hand, HIV-1 enters latency if the host cell is not actively transcribing. The viral RNA splicing sites serve as a sensor of the host cells' activities.

2.3. The challenges in profiling HIV-1 alternative splicing.

HIV-1 transcribes from its LTR promoter. There are 4 major splicing donor sites and 7 major splicing acceptor sites on its mRNA. After transcription, the mRNA was spliced to generate different 5' UTR and express different viral genes. There are 3 major types of splicing products, including the multi-spliced RNA (msRNA), which expresses tat, rev and

nef; the single spliced RNA (ssRNA), which expresses vif, vpr, vpu and env; the unspliced RNA (usRNA), which expresses gag/pol and serves as the viral genomic RNA.

HIV-1 utilizes the human spliceosome for its own splicing. When the viral transcription begins, the amount of host spliceosome outnumbers the viral RNA, so most viral RNA is spliced into msRNA and transported to the cytoplasm for translation. The early stage of the HIV-1 life cycle is characterized by the expression of genes on the msRNA. The viral protein rev is an RNA export protein that specifically binds to the unspliced viral RNA and sheds them from the host spliceosome. While tat is a transcription factor that binds to viral LTR and significantly promotes viral transcription. With the accumulation of rev and tat, more ssRNA and usRNA was exported to cytoplasm for translation. HIV-1 expresses more ssRNA and usRNA at the later stage of its life cycle.

The delicate control of the HIV-1 splicing has been characterized by gel electrophoresis and next-generation sequencing [29-31]. However, all these researches were performed in bulk samples. To fully understand the fate of the provirus, it is critical to profile the HIV-1 mRNA splicing at the single cell level. Current single cell mRNA sequencing methods failed to efficiently capture HIV-1 mRNA sequencing due to following reasons [32]. Firstly, droplet-based methods add cellular barcodes at 3' end or 5' end of the mRNA. But the HIV-1 splicing occurs after D1 site and before A7 site. It is more than 1000 bp away from the 3' end. So the high throughput sequencer cannot read the cellular barcode and the splicing site simultaneously. Secondly, microwell-based methods add a unique cellular barcode in each well and support full length mRNA amplification. But it cannot efficiently target HIV-1 infected cells or HIV-1 mRNA, leading to a high cost of scaling up. Lastly, 3rd generation single molecule sequencing can efficiently sequence full length mRNA.

But HIV-1 splicing sites are adjacent and compact. The high error rate of the 3rd generation sequencing cannot faithfully capture low frequency splicing events.

In this paper, we will introduce a method to efficiently sequence HIV-1 splicing at single cell level, using viral barcodes. And combining this method with other multi-omics methods, we will show how the position of the HIV-1 integration affects virus transcription and splicing.

3. Results

3.1. Barcode-integration site linkage sequencing revealed the positional effect of HIV-1 transcription

A 21-nucleotide random sequence was inserted into HIV-1 NL4-3 strain to create a genetically barcoded viral library (Figure 1A). More than 1 million colonies were harvested during library construction. This ensures the probability of two cells being infected with the virus of the same barcode is less than 5% if less than 0.1 million cells were infected (Figure S1A). We sequenced the barcode library to confirm its diversity. The average number of nucleotide differences between any two barcodes was 12, so we can easily distinguish barcodes from the next generation sequencing data (Figure S1B). The barcode frequency of each virus stays nearly the same after producing the virus, indicating that the barcode's impact on the virus replication is negligible (Figure S1C). The frequency of most barcodes was uniformly distributed (Figure S1D). In summary, the barcode library can be used to label individual provirus in vitro.

To measure the integration site and the transcriptional activity of each provirus at the same time, we used the barcoded library to infect primary CD4+ T cells and separate genomic DNA and RNA using the silica membrane (Figure S2A). A protease inhibitor was

added to the viral culture to prevent multiple rounds of infection, so each infected cell will have a unique barcode. We quantified the viral transcription by sequencing the barcode region and counting the relative abundance of each barcode (Figure 1C, 1D). We observed the number of actively transcribing barcode viruses was not changing significantly during the first 48 hours post infection. But the average transcriptional activity of each virus was increasing by more than 10-fold. Consistent with previous studies, the increment of transcription was not uniform on all cells, but only a proportion of cells were actively transcribing, while the rest remained a low activity.

We then developed a sequencing protocol to sequence the proviral integration site and the barcode sequence simultaneously (Figure 1B). Briefly, we ligated a L-shape adapter on the randomly fragmented cellular genome. A 21-nucleotide unique molecular identifier (UMI) was included in the adapter to label each proviral molecule. Multiple steps of semi-nested PCR enriched the fragments including the integration junction. The PCR product was enzymatically digested and self-circularized to bring the barcode and the integration site together. Then the short fragments containing UMI, barcode and the integration site were amplified for high-throughput sequencing. This allows us to link the integration site and viral transcription of each cell. The provirus was divided into three categories according to their sequences near the integration junction (Figure S2B). If there is no sequence after the integration junction, the provirus is divided into unintegrated. It is an intermediate step of virus reverse transcription and integration. The half life of this molecule conformation is short. If the sequence maps to the HIV-1 reference genome, the provirus is considered to be auto-integrated. This is a kind of integration byproduct that the virus did not successfully find the host DNA to integrate and accidentally

destroyed its own genome. If the sequence maps to the human reference genome, the provirus is considered as integrated.

In the first 48 hours after infection, the frequency of integrated provirus is gradually increasing (Figure 2A). Among those proviruses, the frequency of actively transcribing proviruses is only increasing for the integrated form (Figure 2B). This is consistent with the previous knowledge that HIV-1 has to be integrated for the following transcription steps. The relative transcriptional activity of integrated provirus was also ~100 times higher than the unintegrated and the auto-integrated forms (Figure 2C).

Then we analyzed the positions of the integration sites. Consistent with previous publications, HIV-1 integration took place on all chromosomes (Figure 3A), but it favored the transcriptionally active region (Figure 3B). Moreover, the integration sites were observed more frequently near the active histone markers but less frequently near the repressive histone markers (Figure S3A). The proviruses were also more likely to be integrated in the short interspersed nuclear elements (SINEs), but less likely to be in other endogenous retrovirus LTR regions (Figure S3B).

We then looked at the transcriptional activity of each provirus. We found the viruses integrated in host gene regions had significantly higher transcriptional activity than those integrated in the intergenic regions (Figure 3C). However, the transcriptional activity of the cell and the activity of the host gene being integrated was negatively correlated (Figure 3D). This is probably due to the promoter occlusion resulting from the hyper active host gene depleting nearby transcriptional machineries. The distribution of actively transcribed provirus and inactive transcribing provirus were also correlated with nearby histone modifications (Figure 3E). The actively transcribed proviruses were closer to

active histone markers than the inactive proviruses, and vice versa. The transcriptional activity of proviruses inside of repeat region SINEs were also higher than the proviruses outside SINEs. All these data suggest that local genomic structure and transcriptional activity are important for the proviral transcription.

3.2. Barcode - alternative splicing linkage sequencing profiled the abundance of different viral genes

To quantify the exact composition of viral mRNA in each infected cell, we developed a linkage method to sequence the barcode and the splicing junction at the same time (Figure 4A). HIV-1 has 3 major classes of splicing products: the unspliced form, the single spliced form and the multi-spliced form. All splicing events took place at the 5' of the gene coding region, resulting in different 5' UTRs (Figure 4B). We designed the reverse transcription primers to add a UMI to each mRNA molecule immediately after the barcode sequence. Then we amplified the whole HIV-1 transcripts because all of them share the same leading sequence at the 5' end before the first splicing donor site. The PCR product was enzymatically digested and circularized to bring the 5' UTR and the barcode region close. Lastly, we used 3 different sets of primers to amplify the barcode and 5' UTR region. The 3 sets of primers were designed for 3 major splicing classes. Our methods captured more than 80 different splicing forms, including all the well characterized major forms shown in Figure 4B. The sequences of the 5' UTR can be used to infer the translation product. In Figure 4C, we show the abundance of different viral mRNA in a single cell. We recovered 28 different types of mRNA isoforms for all 7 genes just in one cell. The unspliced form of viral RNA is the most abundant form in this cell.

We then looked at the relationship between viral transcription and splicing activity by linking the splicing sequencing result with the RNA amplicon quantification result. We calculated the diversity of viral mRNA in each cell (Figure 5A). We found that for each cell, the viral mRNA diversity positively correlated with the total transcriptional activity of the virus. This indicates more viral transcription will generate more types of viral RNA. However, the whole population showed a different trend on a temporal scale (Figure 5B). Previously we showed that the viral transcription increased over time because a portion of viruses' activity increased (Figure 1C). But the average mRNA diversity of each virus was decreasing over time. This indicates as the fate of the virus is decided, the transcriptional program will be more committed to a certain set of viral mRNA, thus the diversity is decreasing in a temporal scale.

We then reconstructed the dynamics of each viral ORF for each cell (Figure 5C). We found the frequency of Tat, Vif, Vpr and Vpu/Env per cell were increasing over time while Rev/Nef per cell were decreasing. This is consistent with the previous knowledge of the 3 step transcription program. The early viral genes, like Rev and Nef, are heavily spliced and are decreasing over time. The middle genes, like Vif, Vpr and Vpu only need the splicing of D1-A4 intron and are expressed as the second group. The late genes are the full length viral mRNA that can express structural proteins and also serve as the genomic RNA. We also observed the variation of the genomic RNA, Vif, Vpr, and Tat are decreasing over time, indicating a process of fate decision.

We calculated the frequencies of ORFs with the total viral transcriptional activity. We found at the early time point, Tat, Vif and Vpr frequency were negatively correlated with total viral RNA (Figure 5D). This indicates the function of Tat on viral transcription. Rev is

the only viral gene that is positively correlated with the within cell diversity of viral mRNA, indicating an important role of RNA export in regulating mRNA diversity (Figure 5E).

The abundance of viral different genes were correlated (Figure 6A). There were 3 obvious groups of viral genes. Genomic RNA, Vif, Vpr and Tat are positively correlated and are usually expressed together. They indicate an active transcription program that leads to production of virions. Vpu/Env is negatively correlated with the major group, it indicates the last set of viral structural genes, regulating the cellular environment for pyroptosis and virion budding, inhibiting the synthesis of cellular genes. Nef is negatively correlated with all other genes. It should be considered as a dormant state of the virus' fate, where almost all viral genes were not expressed. Only a few viral mRNA molecules were heavily sliced into Nef mRNA. Rev is not correlated with any of the other genes. It has an important role in fate decisions and regulates the diversity of viral mRNA (Figure 6B).

The viral gene expression could be summarized using principal component analysis (PCA) and plotted in 2 dimensional (Figure S4A). The fate decision process could be visualized by inferring a pseudotime based on gene expression using SlingShot (Figure S4B). The pseudotime correlated well with the real sampling time.

Lastly, we linked the viral gene expression with the position of its integration site using the barcode sequences (Figure 6C). We found proviruses expressed more Tat if they were close to active histone marker H4K20me1. Rev is more frequent if the integration site is close to the active marker H3K4m1. However, Nef is more frequent if the integration site is close to the repressive marker H3K27me3. These data suggest that the position of integration sites affect viral transcription by differentially regulating its gene expression.

3.3. The effect of latency reversal agents was associated with the position of the integration site

Latency reversal agents (LRAs) are promising therapies to eliminate HIV-1 latent reservoir and achieve a complete cure. There are two major categories of LRAs, the HDAC inhibitor and the PKC agonist. To evaluate their effect on the position of the integration site and the viral RNA splicing, we used SAHA (HDAC inhibitor) and Bryostatin (PKC agonist) to treat the barcoded cells infected resting CD4+ T cells and sequence the viral integration and splicing 24 hours after the treatment. We used CD3/CD28 beads to mimic the TCR activation signal and serve as a positive control of T cell activation. All treatments substantially increased the viral transcription (Figure 7A). While TCR activation has the highest amount of virus induction, the effect of LRAs are more obvious on the unspliced viral RNA. We then looked at the transcriptional activity of each provirus by counting the barcode abundance in the viral RNA (Figure 7B). We found none of the treatments can activate the transcription of all proviruses. They were only activating a proportion of proviruses or increasing the transcriptional activity of already active proviruses. SAHA treatment activated more viruses than Bryostatin, while the average transcriptional activity of provirus activated by Bryostatin was stronger (Figure S5A, S5B). SAHA remodels the histone modifications in the nucleus and may expose the beforehand dormant provirus. While Bryostatin induces the T cell activation signal and recruits many transcription factors into the nucleus. There are many crosstalks between two pathways, they eventually all lead to T cell activation and provirus transcription, but they may have preference of different provirus according to the position of their integration sites. We found both drugs activated provirus in gene regions and intergenic regions (Figure 7C).

But SAHA was more effective for provirus in intergenic regions. We tested if the odds of virus activation is associated with nearby histone modifications. Both drugs had a preference of proviruses near the active histone markers. But SAHA did not omit the provirus near the repressive marker H3K9me3 (Figure 7D).

We also sequenced the viral gene expression for each provirus. We found that within cell diversity of viral mRNA was decreasing after the LRAs treatment (Figure 8A). This indicates LRAs initiated the fate decision program and increased the total transcriptional activity. We did PCA visualization of all proviruses according to their viral gene expression. We found TCR activation resulted in a significant upregulation of viral genomic RNA, and the phenotype of Byrostatin treatment was closer to the TCR than SAHA was. The viruses treated by SAHA or combination group had higher amounts of Vif and Tat, indicating that SAHA treatment was slow in activating the viruses and the activated viruses were phenotypically heterogeneous. The data indicates we should fine-tune the treatment time of different LRAs while designing the combination therapies.

4. Discussion

In this paper, we developed a new method to simultaneously sequence the HIV-1 integration, transcription and alternative splicing at the single cell level. The integration site of the virus can determine the accessibility of the host transcription machinery, thereby affecting the level and pattern of viral gene expression, and ultimately the outcome of the infection. Meanwhile, the alternative splicing of the viral mRNA generates different forms of viral proteins, which have diverse biological functions and modulate the host immune response to the virus. We found the position of the HIV-1 integration site may affect the viral transcription by differentially regulating the abundance of different viral

genes. This is achieved because the transcription and splicing machinery are not uniformly distributed in the host nucleus, and different intensity of splicing can lead to different viral gene expression patterns, affecting virus fate decision. Studying the relationship between the integration site and alternative splicing can provide a more comprehensive understanding of the complex mechanisms of HIV-1 pathogenesis and may lead to the development of new treatment strategies that can specifically target and manipulate these processes. Testing drugs facilitating or inhibiting virus splicing together with LRAs may result in better coverage and efficiency of latency reversal.

Compared with current single cell multi-omics technologies, our method achieved the throughput of ~0.1 million cells with the cost of ~100 dollars. This is 50 fold higher than throughout and 50 fold cheaper than most commercial solutions. This allowed us to profile dozens of conditions easily. Thus, this method can be widely used for drug development and in vitro screenings.

The method also has shortcomings. It cannot characterize the phenotype of the infected cells. Additional cellular barcodes may be needed for this purpose. Future studies can combine this method with droplet based single cell labeling or long read sequencing to generate a more comprehensive dataset for HIV-1 multi-omics.

5. Methods

5.1. Generation of barcoded HIV-1 library

The design of the barcoded virus is shown in Figure 1A. Two fragments were PCR amplified using following primers: makeBC_F2(GGCTTGGAAAGGATTTTGCTATAANNCNNCNNCNNCNNCNNCNNCTATAAGATGGGTGGCAAGTGGTC) and makeBC_R2(GCTCCATGTTTTTCTAGGTC),

makeBC_F1(CAGATCCATTCGATTAGTGAAC) and makeBC_R1(TTATAGCAAAATCCTTTCCAAGCC). The PCR generated 2 products of 343-bp and 176-bp respectively. The products were then purified and eluted in TE buffer. The 2 fragments have a 24- bp overlapping region, so they were assembled together using the NEB HiFi Assembly kit. The 495-bp assembled fragment was then amplified again using primers makeBC_F1 and makeBC_R2. The fragment was digested by restriction enzymes BamHI-HF and XhoI, and then purified. The pNL4-3 vector was also digested by these 2 enzymes and was purified by agarose gel electrophoresis. One ug of the insert fragment and 5ug of the vector was assembled using NEB HiFi Assembly kit. The assembled DNA was purified by ethanol precipitation with Pellet Paint NF co-precipitant. One ug of the purified DNA was transformed into the MegaX E. coli electrocompetent cells and plated to 20 15cm agar plates. The plates were cultured at 37°C overnight. More than 0.5 million colonies were scratched from the surface of the plates. One mg plasmid DNA was extracted from the bacteria pellet. The barcode region on the plasmid was confirmed by Sanger sequencing.

20 ug of plasmid DNA was transfected into 20 million 293T cells using the Calcium phosphate transfection reagent. The virus library was harvested 48 hours after transfection. DNase I (40ng/mL) and MgSO4 (1mM) were added to the library to remove residual plasmid DNA from the supernatant. The barcoded virus library was aliquoted and frozen at -80°C for future use.

5.2. Virus library infection

100 million HIV-1 free PBMC was obtained from anonymous donors. We use CD4 microbeads to isolate CD4+ cells from PBMC. The cells were then counted on a

hemocytometer. The cells were cultured in RPMI-1640 with 10%FBS, 1% PenStrep, 10mM HEPES, 1mM sodium pyruvate, 0.1mg/mL Normocin, and 10U/mL recombinant h-IL2 with 5 million cells per mL density. 10uL CD3/CD28 beads were added to the culture for every million cells. 72 hours post activation, the CD3/CD28 beads were removed from the culture by a magnet and the cells were counted again. Virus worth 100ng p24 was mixed with 1 million CD4+ cells in the culture media and spinoculated for 90 minutes at 3000 rpm. After spinoculation, the cells were washed by fresh media and cultured with 10U/mL h-IL2 and 100nM Darunavir. The infected cells were harvested every 12 hours for subsequent assays.

5.3. Sequencing library preparation

Total nucleic acid was purified using Qiagen DNA/RNA purification kit. The RNA was reverse transcribed using RT primer (CAAGTGCCTAGATCCTCGAGNNNNNNNNNNNNNNNNNNNNCACTTGCCACCCATCTTATA) and purified by Invitrogen PCR clean-up kit. Here 21 consecutive random nucleotides serve as a unique molecular identifier (UMI) for each RNA molecule. For every infection, 3 types of the sequencing library were constructed. 1) Barcode amplicon library. 2) Barcode - integration site linkage library. 3) Barcode - splicing site linkage library.

For the amplicon library, the barcode region was amplified using primers CAAGTGCCTAGATCCTCGAG and GGCTTGGAAAGGATTTTGCTATAA. More than 10 million copies of viral cDNA were used as the template. This ensured sufficient coverage of most barcodes. The PCR product was confirmed using gel electrophoresis and purified by the PCR clean-up kit. I then used NEBNext Ultra II DNA library prep kit

to make pair-end sequencing libraries. Around 50 million reads were sequenced for each sample.

The workflow of barcode integration site linkage sequencing was shown in Figure 1B. I used HinP1I (NEB) to digest the infected host genomic DNA. 63% of genomic fragments were at the length of 25 bp to 3000 bp, this ensures high circularization rate in the following steps. The digested DNA was then purified by PureLink PCR clean-up kit (Invitrogen). UltraII End-repair Module (NEB) prepared the DNA for ligation. A custom adaptor was annealed in the TE buffer (10mM Tris-HCl, 0.1mM EDTA, pH8.0). The sequence of the adapter's reverse strand is TTGAGGTTTGCAGTTG. It has a 5 prime modification of a phosphorylation group, which facilitates TA ligation with the genome fragments. The 3 prime amino modification blocks the polymerase from adding nucleotides at its downstream, maintaining the L-shape conformation of the adapter. 3 consecutive phosphorothioate bonds at the 3 prime end to stabilize the adapter, preventing it from enzymatic degradation. The forward strand of the adapter is ACCATCAACCCCGAATTCNNNNNNNNNNNNNNCAACTGCAAACCTCAAT. It anneals with the reverse strand and contains a 14-nucleotide UMI. 50pmol adapter was ligated to 1μg of fragmented genomic DNA. All ligated products were purified and amplified using 4 rounds of semi-nested PCR. All PCRs used the same reverse primer sequence: ACCATCAACCCCGAATTC. But the forward primer sequences anneal to different parts of the HIV-1 genome to increase the PCR specificity. They are in the order of F4 (AGTGAACGGATCCTTAGCACTTAT), F3 (CTCCTACAGTATTGGAGTCAGG), F2 (AGCCATAGCAGTAGCTGAGG) and F1 (GTACTCGAATTCGGGCTTGGAAAGGATTTTGCTATAA). All forward primers contain

3 consecutive phosphorothioate bonds at the 3 prime end, preventing the exonuclease activity of the polymerase, increasing the PCR specificity. Primer F3 and F2 are used with the reverse primer containing the 5 primer phosphorylation modification, to enable lambda exonuclease digestion after PCR, which can eliminate the product of unspecific amplification. The final PCR product was purified and digested by EcoRI-HF (NEB). This created two sticky ends on the DNA. 100 ng DNA was purified and subject to self ligation in a 100μL reaction. The reaction used 2 units of T4 ligase (Invitrogen) at room temperature for 4 hours. The ligation efficiency was confirmed by quantitative PCR using primers ivF (ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGTCAGTGTGGAAAATCTCT), ivR (GAGTTCAGACGTGTGCTCTTCCGATCTTTTTGACCACTTGCCACCCAT) and synthetic standard templates. One thousand to 10 thousand copies of DNA per uL can be circularized. One third of the ligation product was used as the PCR template for the inverse PCR, using the same primers as the quantitative PCR. Phosphorothioate bond modification was used to increase PCR specificity. One tenth of the product was then subject to the final round of PCR, which adds Illumina sequencing adapters to the library. The primers are AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNACACTCTTTCCCTACACGAC and CAAGCAGAAGACGGCATACGAGATNNNNNNNNNNGTGACTGGAGTTCAGACGTGTGC. N stands for indexing sequence to distinguish different samples. Around 10 million reads were sequenced for each library.

The workflow of the barcode - alternative splicing linkage sequencing was shown in Figure 4A. The near full length viral cDNA was amplified using primers CAAGTGCCTAGATCCTCGAG and ATCGATCTCGAGGCACGGCAAGAGGCGAGGG. Three consecutive phosphorothioate bonds at the end of the primers increased the PCR specificity. The amplified product was purified and digested by XhoI (NEB) for 2 hours. 1ng digested DNA was self ligated in a 100uL ligation system. The reaction used 2 units of T4 ligase (Invitrogen) at room temperature for 4 hours. The ligation product was divided into 3 portions and subjected to 3 PCR. They all used the same forward primer (ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCTTGGAAAGGATTTTGCTATAA) which targets the upstream of the barcode region. But they use different reverse primers annealed to the downstream of the major splicing acceptor sites of the mRNA isoform families. For unspliced RNA, the reverse primer (GAGTTCAGACGTGTGCTCTTCCGATCTCTAGTCAAAATTTTTGGCGTACTCAC) anneals to the beginning of intron 1. For single spliced RNA, the reverse primer (GAGTTCAGACGTGTGCTCTTCCGATCTTCGTCGCTGTCTCCGCTTCT) anneals to the downstream of the A5 site. For multi-spliced RNA, the reverse primer (GAGTTCAGACGTGTGCTCTTCCGATCTCCCTCGGGATTGGGAGGTGG) anneals to the downstream of the A7 site. The PCR extension step only takes 20 seconds, so primers anneal to the downstream of the intended region would not have time for amplification and only 5UTR regions can be amplified. The cycle number was determined by the Ct number of a test run quantitative PCR, which only used 2uL of the ligation product. Finally, a 10-rounds PCR added illumina sequencing adapters and

indexes on the end of the library. Around 10 million reads were sequenced for each library.

All libraries were sequenced on the Illumina NovaSeq6000 platform using the PE150 setting.

5.4. Sequencing data analysis

The sequencing data were all analyzed by custom python codes.

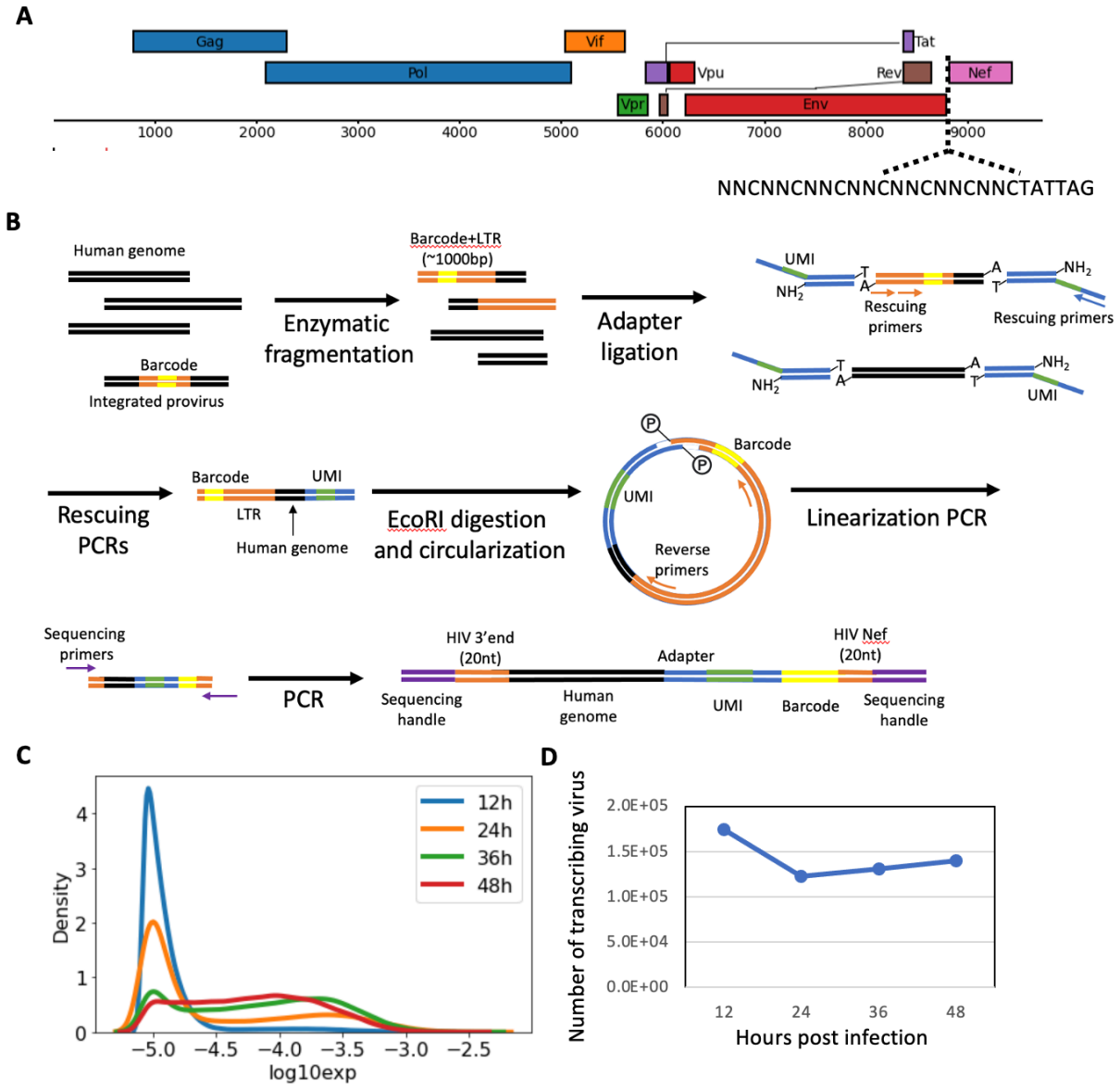


Figure 4-1. Barcode - integration site linkage sequencing.

A) Construction of the barcoded HIV-1 library.

B) Barcode - integration site linkage workflow.

C) Distribution of single viral transcription activity.

D) Count of actively transcribing barcodes.

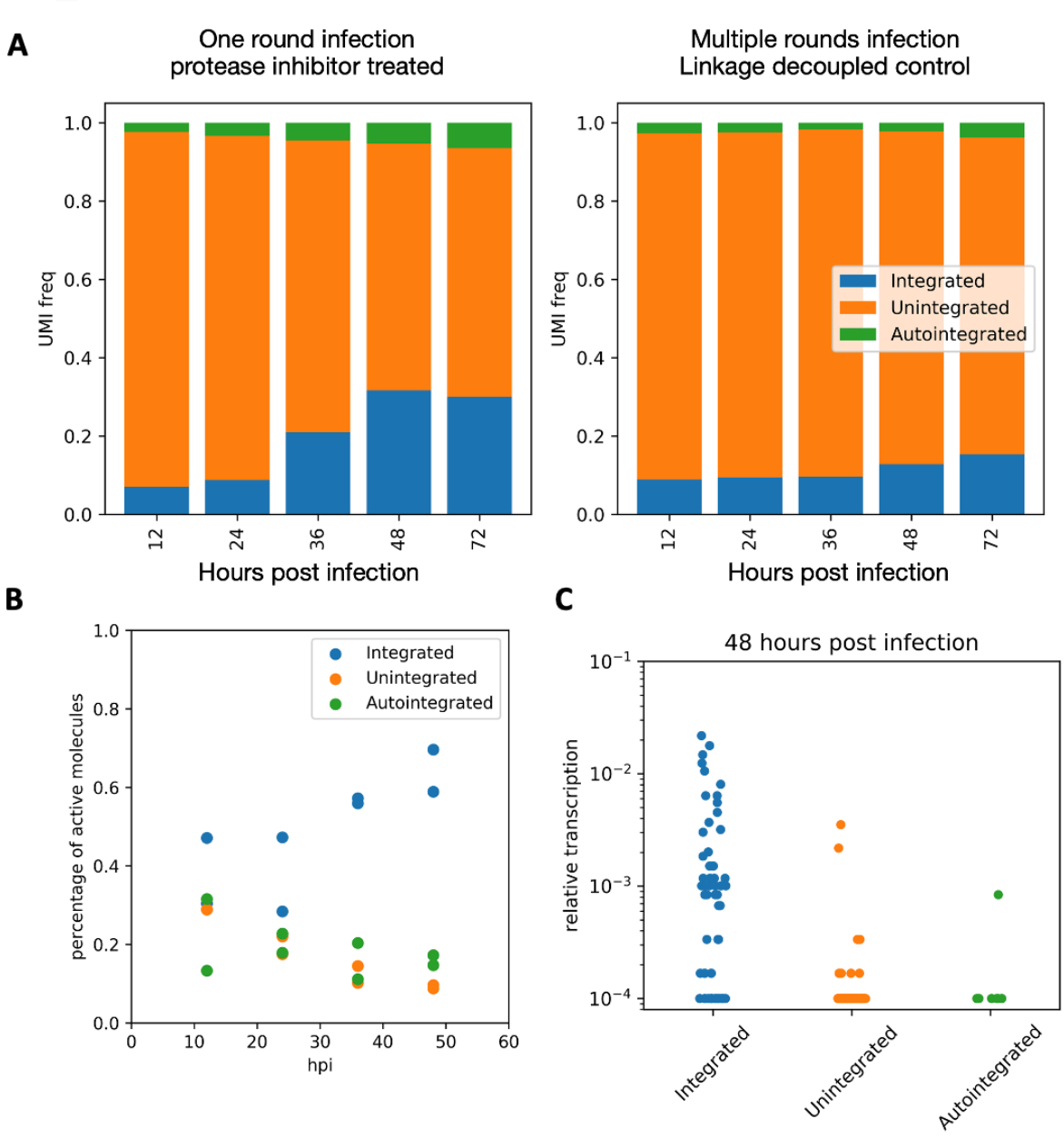


Figure 4-2. Transcriptional activity of different proviral conformations.

A) Abundance of different proviral DNA.

B) Percentage of active provirus at different time points.

C) The relative transcription activity of different proviral conformations.

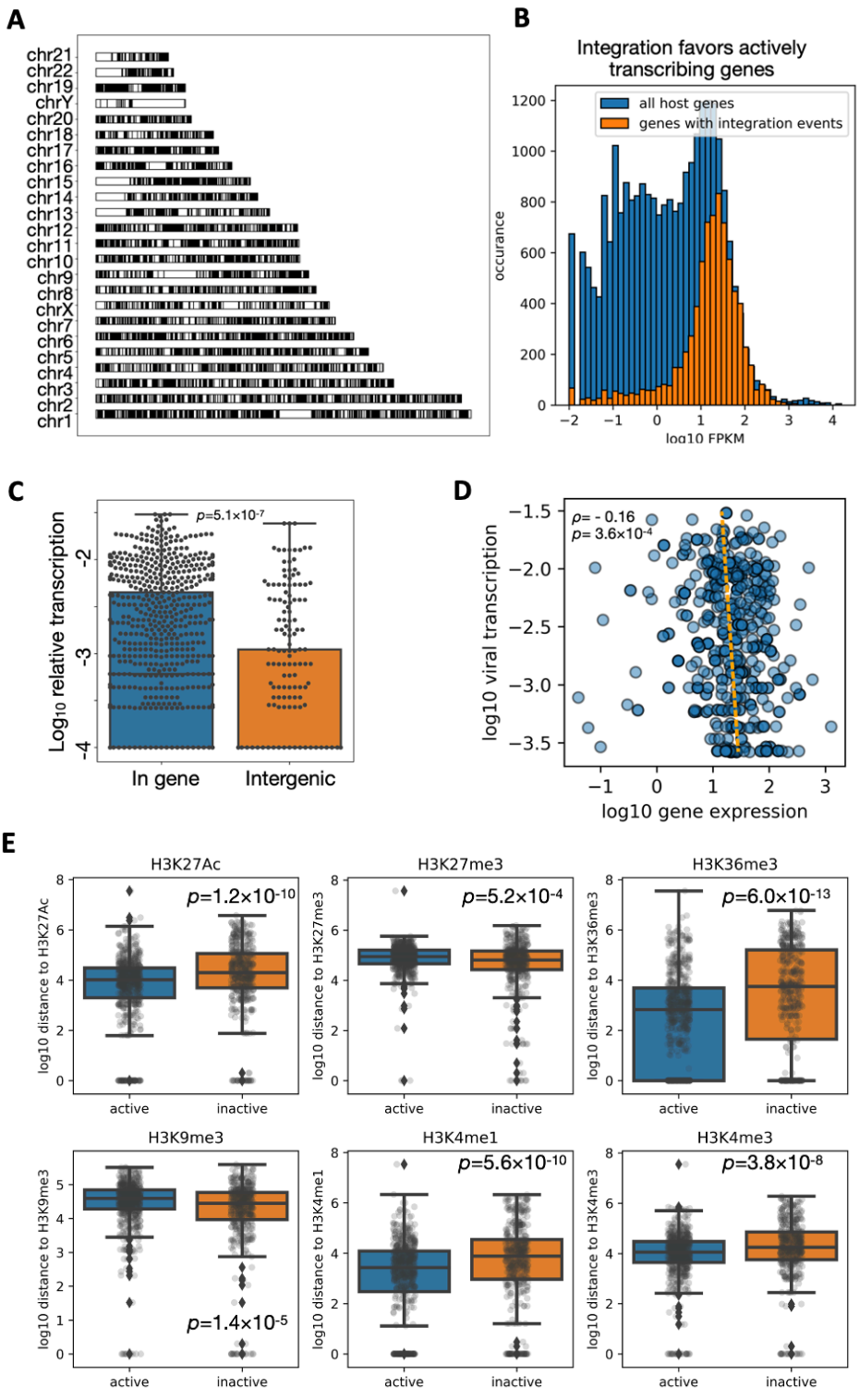


Figure 4-3. Positional effect of the HIV-1 integration site.

A) The distribution of HIV-1 integration sites in this study.

B) The transcriptional activity of the host genes that have proviruses integrated in them.

C) The transcriptional activity of proviruses within or outside the gene regions.

D) The correlation between the provirus transcription activity and the integrated host gene transcription activity.

E) The distance to nearby histone modifications for active or inactive provirus.

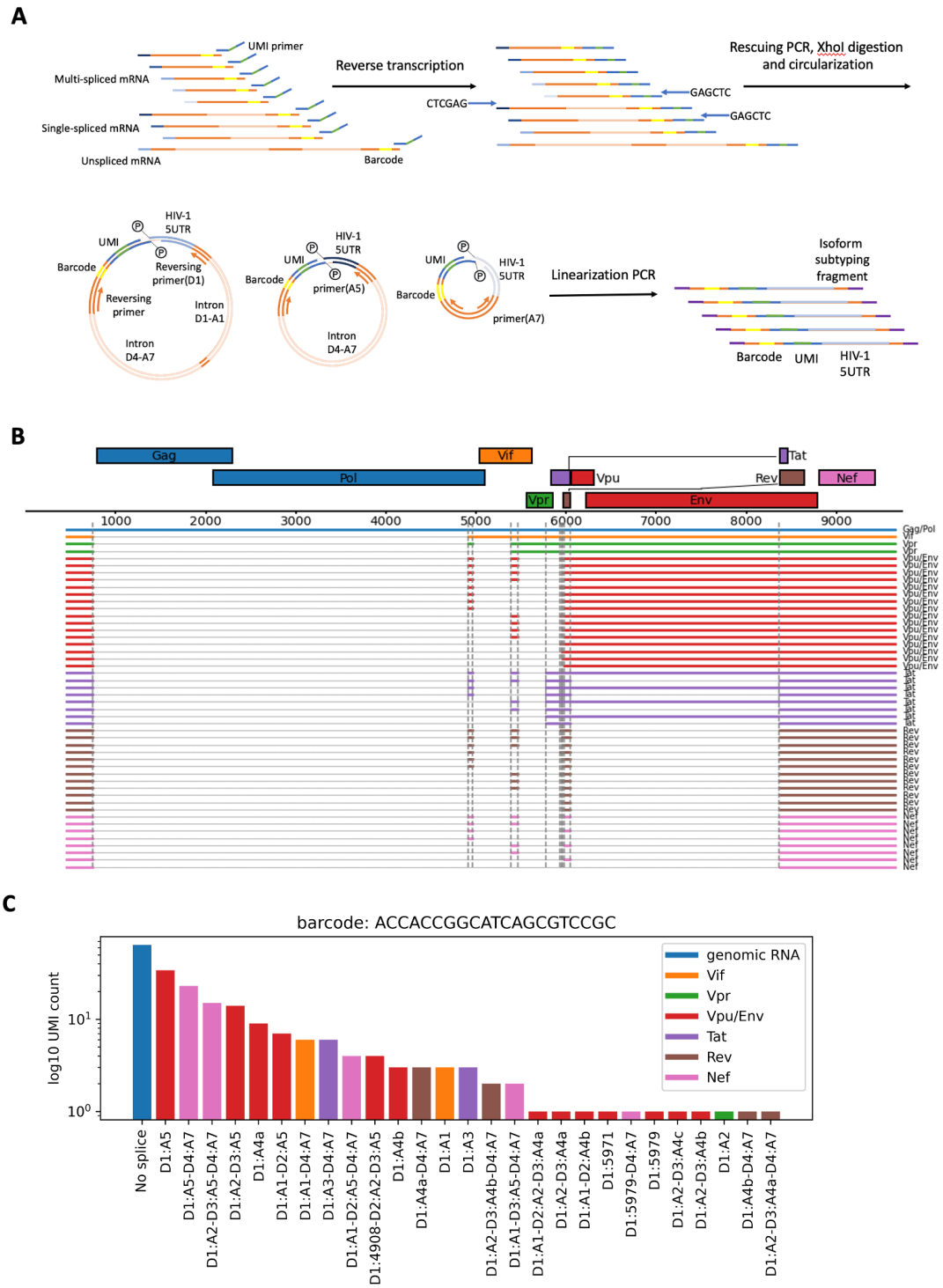


Figure 4-4. Barcode - alternative splicing linkage sequencing.

A) Barcode - alternative splicing linkage workflow.

B) The main forms of HIV-1 mRNA isoforms.

C) An example of viral mRNA isoforms within one cell.

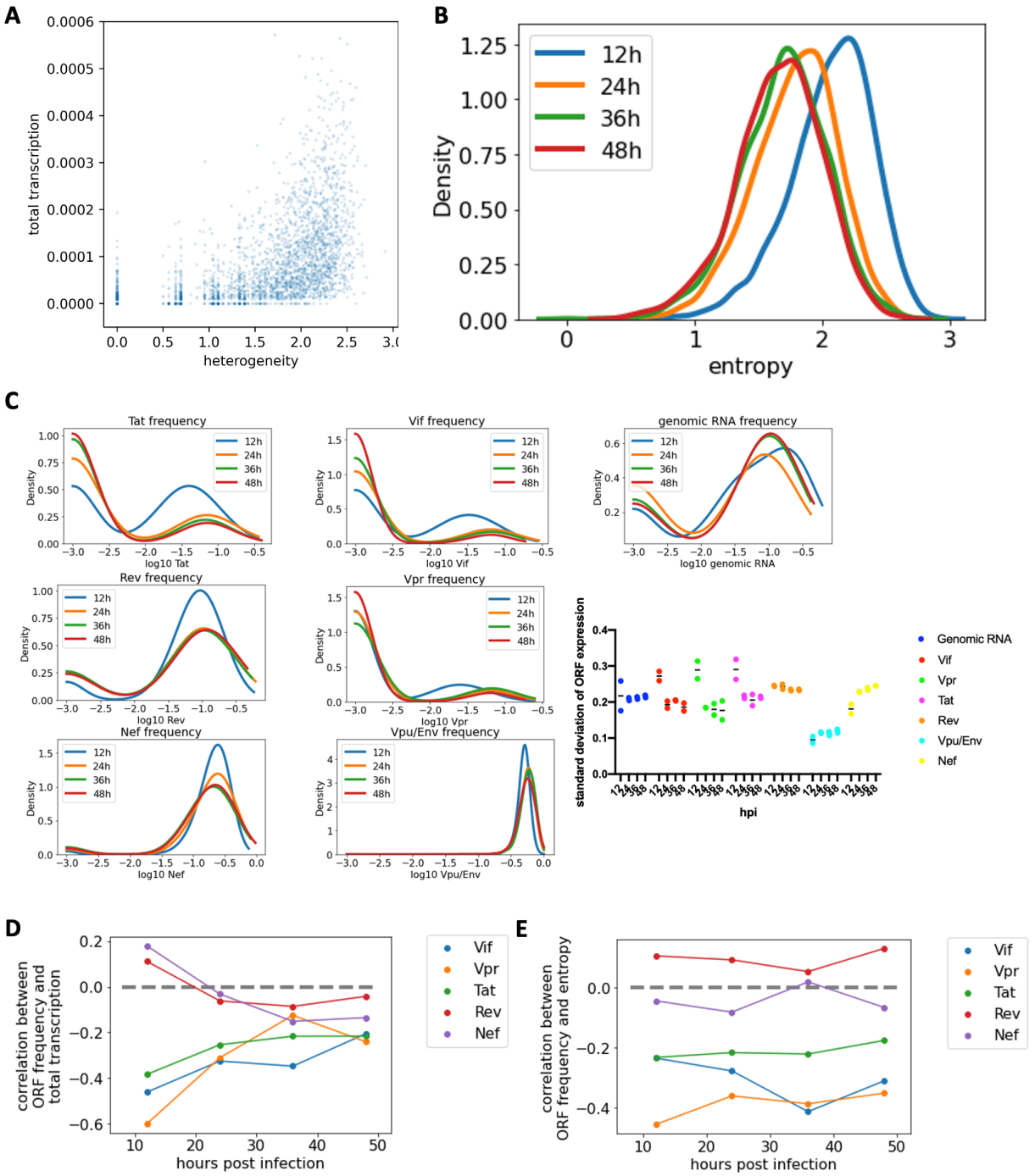


Figure 4-5. Single cell viral gene expression analysis.

A) The correlation between viral gene expression and mRNA heterogeneity.

B) The distribution of mRNA heterogeneity over time.

C) The distribution of viral gene abundance over time and the variation summary.

D) The correlation between viral gene abundance and total viral transcription.

E) The correlation between viral gene abundance and mRNA heterogeneity.

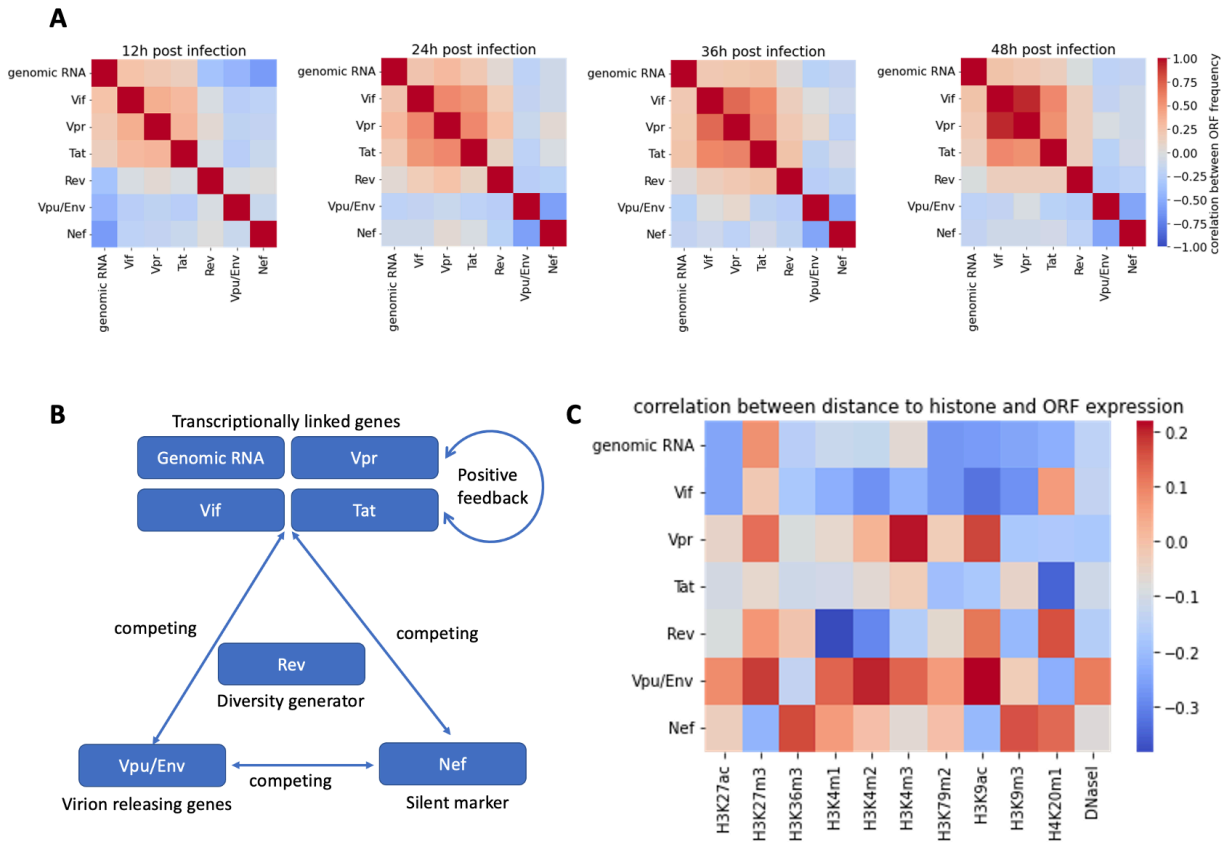


Figure 4-6. Correlation between viral gene expression.

A) The heatmap showing the correlation coefficients between different viral ORFs.

B) The model of HIV-1 viral gene transcription regulation.

C) The relationship between virus integration sites and viral gene expression.

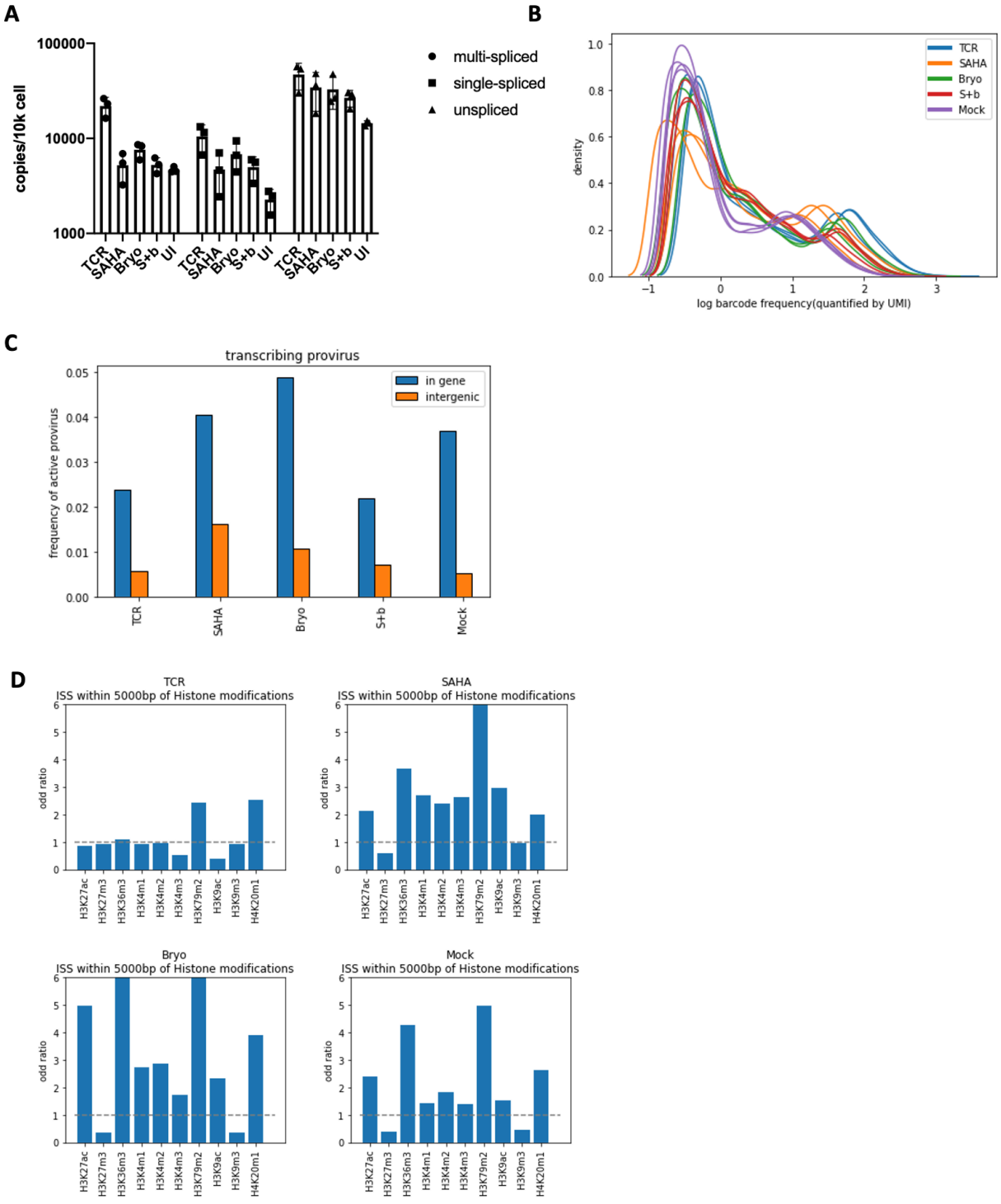


Figure 4-7. Effect of LRAs on different provirus.

A) The copy number of viral mRNA after LRA treatment.

B) The distribution of viral RNA abundance in each cell.

C) The frequency of active provirus in different genomic regions.

D) The odds of actively transcribing provirus near certain histone markers.

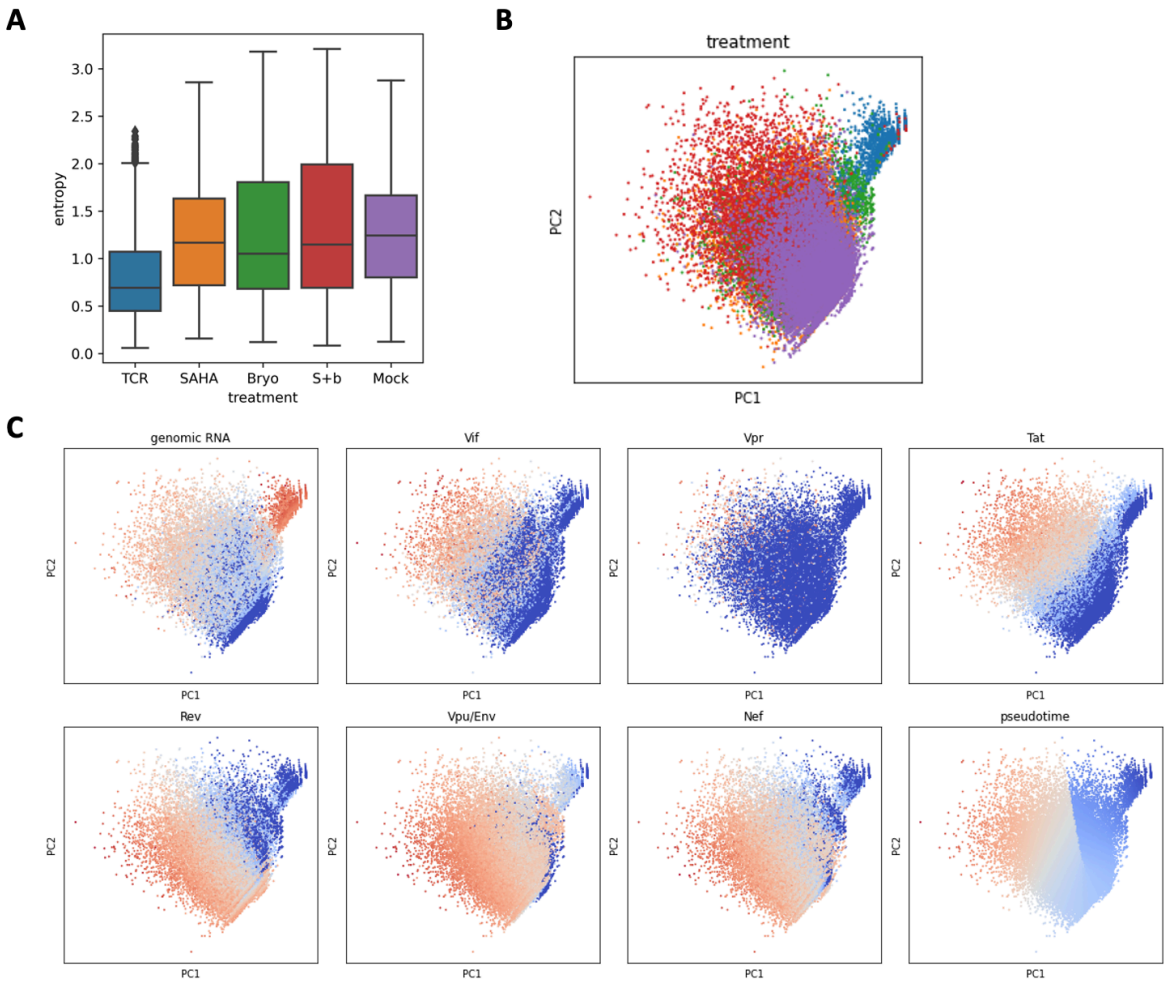
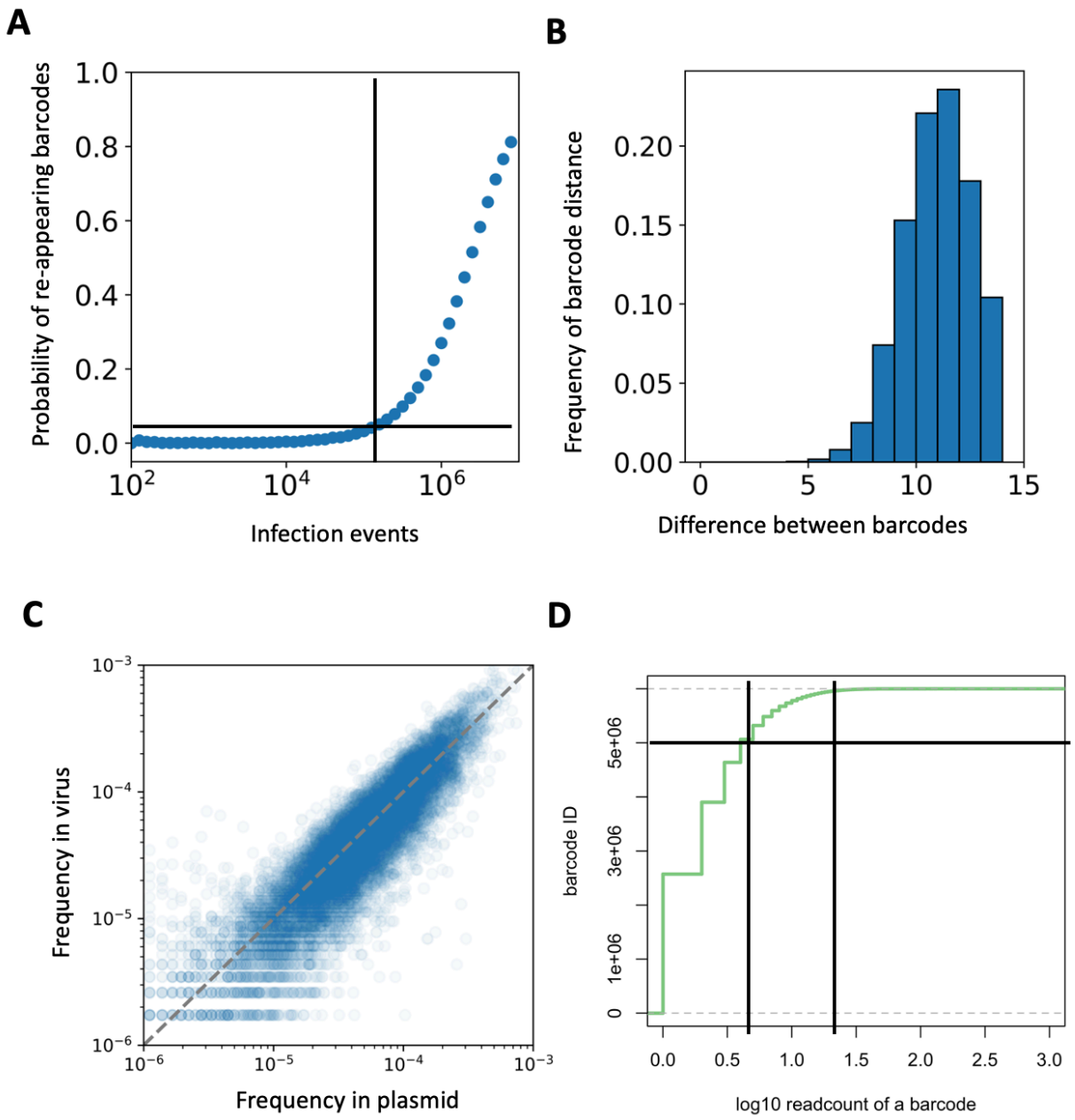


Figure 4-8. LRAs affect viral gene expression.

A) The single cell mRNA heterogeneity in different treated groups.

B) The PCA visualization of virus gene expression, colored by treatment.

C) The PCA visualization of virus gene expression, colored by genes.



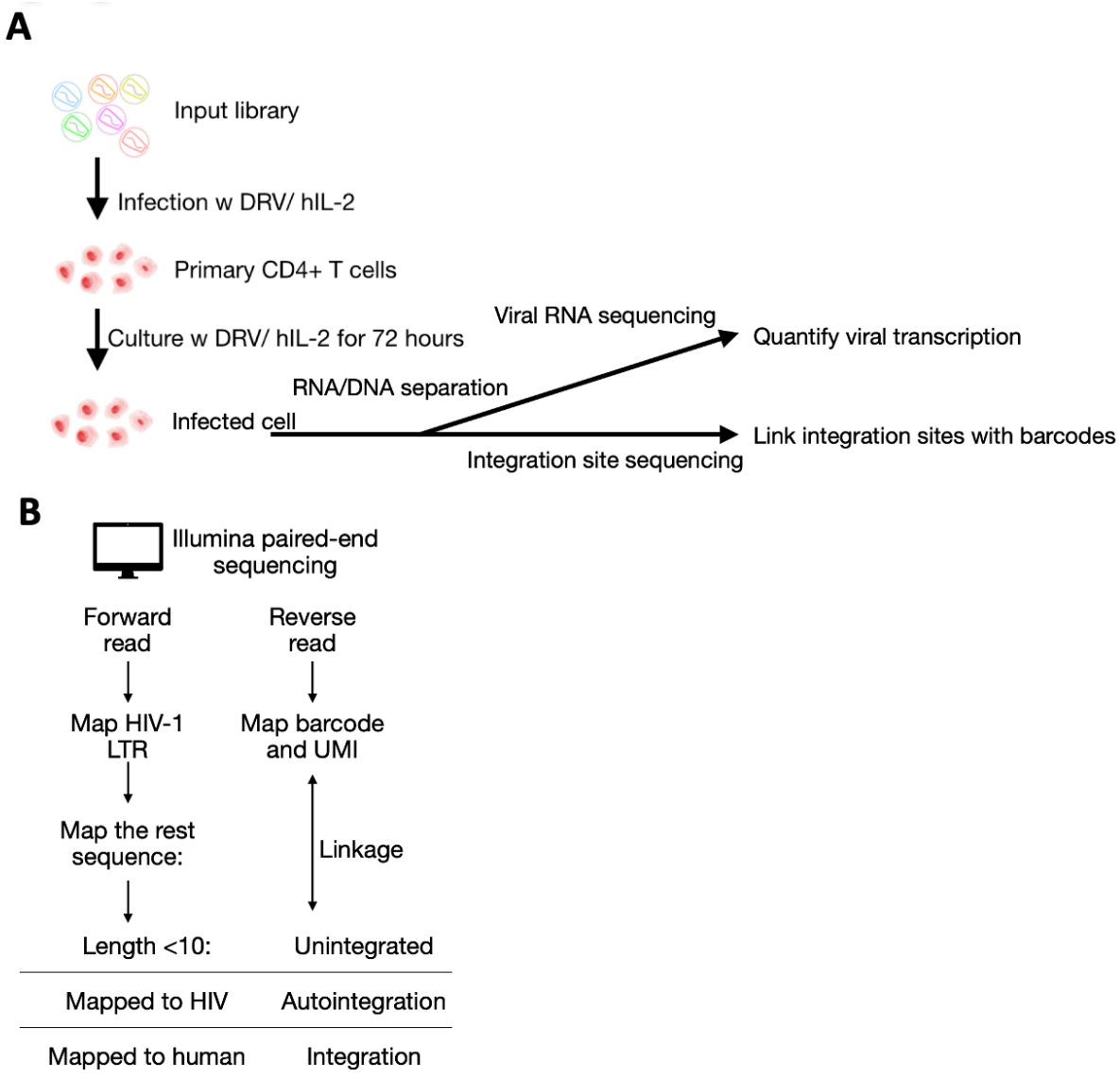
Supplementary Figure 4-1. The quality of the library.

A) The probability of two cells sharing the same barcode with different infection size.

B) The hamming distance between two barcodes.

C) The barcode frequency in plasmid and in viruses.

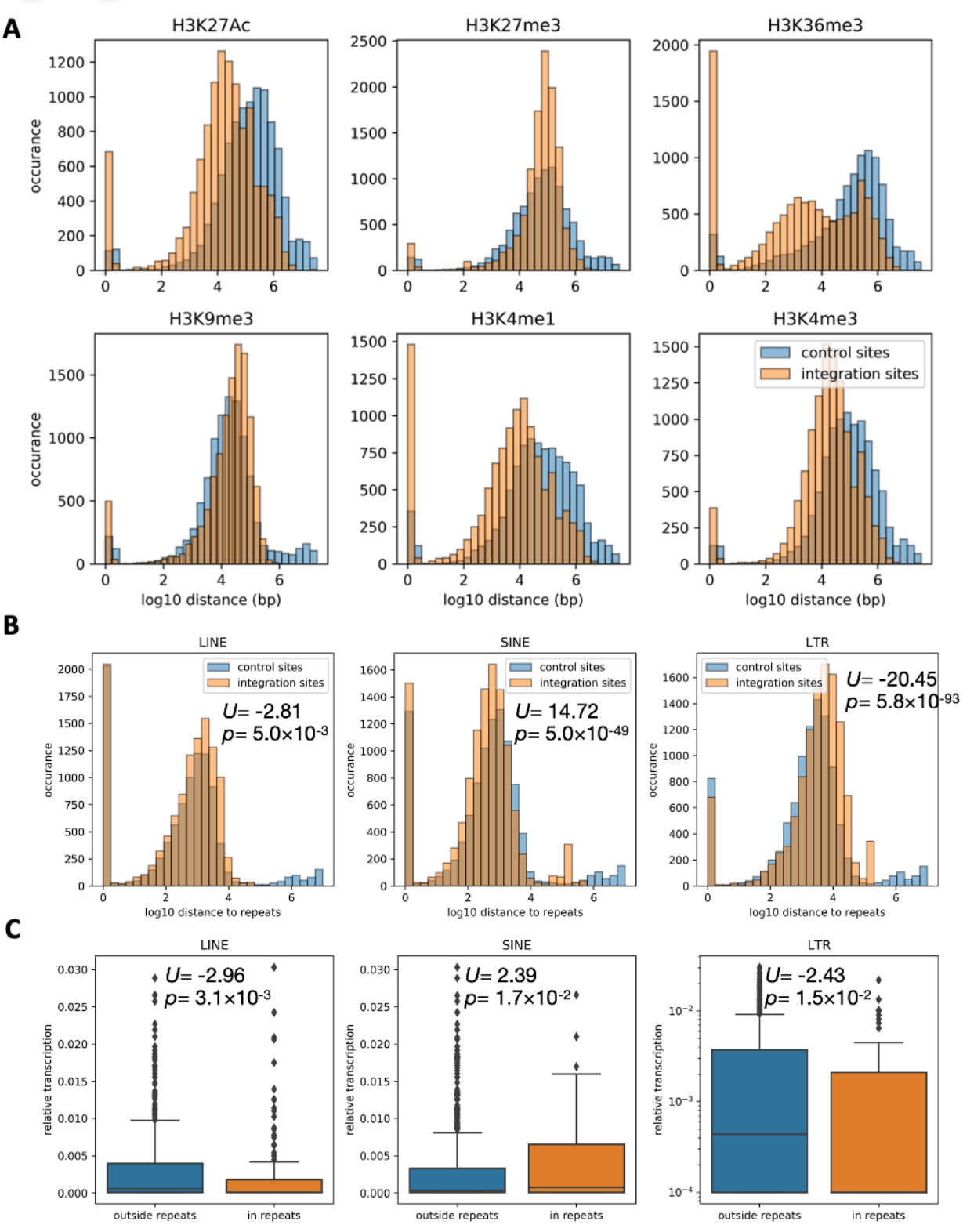
D) The cumulative density function of barcode occurrence.



Supplementary Figure 4-2. Workflow diagram of the linkage sequencing.

A) Experiment design for the barcode integration site linkage sequencing.

B) Data analysis pipeline.

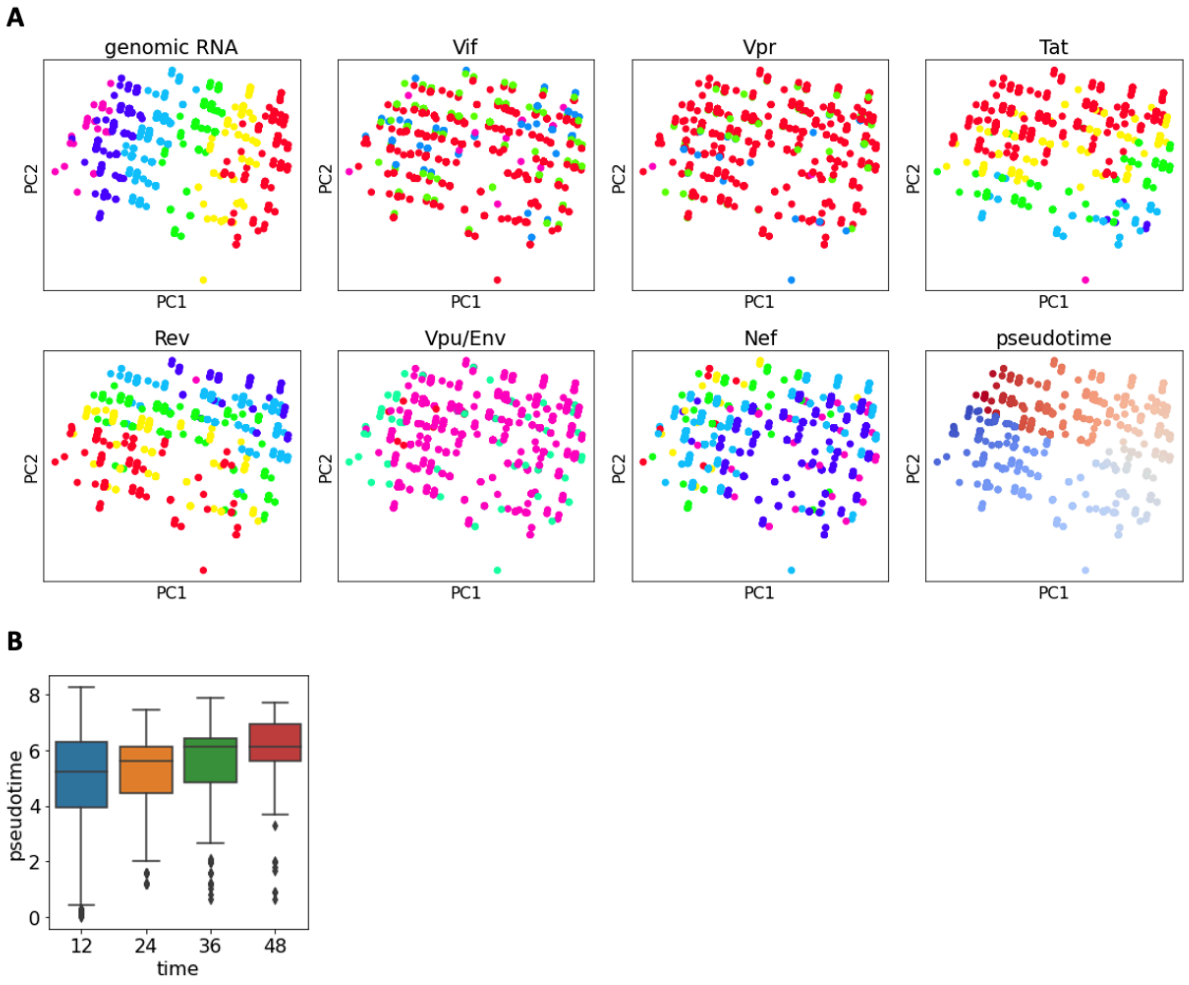

A
H3K27Ac
H3K27me3
H3K36me3
H3K9me3
H3K4me1
H3K4me3
occurance
log10 distance (bp)
control sites
integration sites
B
LINE
SINE
LTR
U= -2.81
p= 5.0×10-3
U= 14.72
p= 5.0×10-49
U= -20.45
p= 5.8×10-93
log10 distance to repeats
C
U= -2.96
p= 3.1×10-3
U= 2.39
p= 1.7×10-2
U= -2.43
p= 1.5×10-2
relative transcription
outside repeats
in repeats

Supplementary Figure 4-3. Distribution of integration site.

A) Distribution of integration site near histone markers.

B) Distribution of integration site near repeat regions.

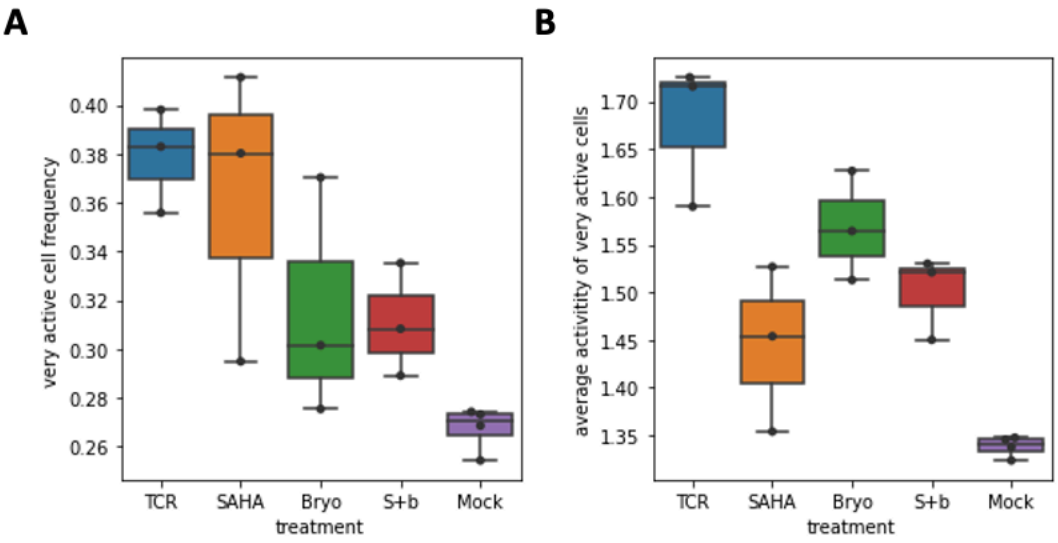
C) Transcriptional activity of provirus near repeat regions.



Supplementary Figure 4-4. The PCA visualization of virus gene expression.

A) The PCA visualization of virus gene expression, colored by genes.

B) The pseudotime inference of each cell.



Supplementary Figure 4-5. Transcriptional activity of LRAs treated cells.

A) Frequency of active provirus.

B) Mean transcriptional activity of the active provirus.

6. References

1. Dahabieh, M. S., Battivelli, E., & Verdin, E. (2015). Understanding HIV latency: the road to an HIV cure. *Annual review of medicine*, *66*, 407-421.
2. Chomont, N., El-Far, M., Ancuta, P., Trautmann, L., Procopio, F. A., Yassine-Diab, B., ... & Sékaly, R. P. (2009). HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nature medicine*, *15*(8), 893-900.
3. Poli, G., Kinter, A., Justement, J. S., Kehrl, J. H., Bressler, P., Stanley, S., & Fauci, A. S. (1990). Tumor necrosis factor alpha functions in an autocrine manner in the induction of human immunodeficiency virus expression. *Proceedings of the National Academy of Sciences*, *87*(2), 782-785.
4. Day, C. L., Kaufmann, D. E., Kiepiela, P., Brown, J. A., Moodley, E. S., Reddy, S., ... & Walker, B. D. (2006). PD-1 expression on HIV-specific T cells is associated with T-cell exhaustion and disease progression. *Nature*, *443*(7109), 350-354.
5. Razooky, B. S., Pai, A., Aull, K., Rouzine, I. M., & Weinberger, L. S. (2015). A hardwired HIV latency program. *Cell*, *160*(5), 990-1001.
6. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P., & Schaffer, D. V. (2005). Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell*, *122*(2), 169-182.
7. Jordan, A., Defechereux, P., & Verdin, E. (2001). The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *The EMBO journal*, *20*(7), 1726-1738.

8. Battivelli, E., Dahabieh, M. S., Abdel-Mohsen, M., Svensson, J. P., Da Silva, I. T., Cohn, L. B., ... & Verdin, E. (2018). Distinct chromatin functional states correlate with HIV latency reactivation in infected primary CD4+ T cells. *Elife*, *7*, e34655.
9. Chen, H. C., Martinez, J. P., Zorita, E., Meyerhans, A., & Filion, G. J. (2017). Position effects influence HIV latency reversal. *Nature structural & molecular biology*, *24*(1), 47-54.
10. Marini, B., Kertesz-Farkas, A., Ali, H., Lucic, B., Lisek, K., Manganaro, L., ... & Lusic, M. (2015). Nuclear architecture dictates HIV-1 integration site selection. *Nature*, *521*(7551), 227-231.
11. Han, Y., Lin, Y. B., An, W., Xu, J., Yang, H. C., O'Connell, K., ... & Siliciano, R. F. (2008). Orientation-dependent regulation of integrated HIV-1 expression by host gene transcriptional readthrough. *Cell host & microbe*, *4*(2), 134-146.
12. Shan, L., Yang, H. C., Rabi, S. A., Bravo, H. C., Shroff, N. S., Irizarry, R. A., ... & Siliciano, R. F. (2011). Influence of host gene transcription level and orientation on HIV-1 latency in a primary-cell model. *Journal of virology*, *85*(11), 5384-5393.
13. Van Lint, C., Emiliani, S., Ott, M., & Verdin, E. (1996). Transcriptional activation and chromatin remodeling of the HIV-1 promoter in response to histone acetylation. *The EMBO journal*, *15*(5), 1112-1120.
14. Lusic, M., Marcello, A., Cereseto, A., & Giacca, M. (2003). Regulation of HIV-1 gene expression by histone acetylation and factor recruitment at the LTR promoter. *The EMBO journal*, *22*(24), 6550-6561.

15. Schröder, A. R., Shinn, P., Chen, H., Berry, C., Ecker, J. R., & Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, *110*(4), 521-529.
16. Singh, P. K., Plumb, M. R., Ferris, A. L., Iben, J. R., Wu, X., Fadel, H. J., ... & Levin, H. L. (2015). LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes & development*, *29*(21), 2287-2297.
17. Pradeepa, M. M., Sutherland, H. G., Ule, J., Grimes, G. R., & Bickmore, W. A. (2012). Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS genetics*, *8*(5), e1002717.
18. Wang, Z., Simonetti, F. R., Siliciano, R. F., & Laird, G. M. (2018). Measuring replication competent HIV-1: advances and challenges in defining the latent reservoir. *Retrovirology*, *15*(1), 1-9.
19. Goodier, J. L. (2016). Restricting retrotransposons: a review. *Mobile DNA*, *7*(1), 1-30.
20. Ocwieja, K. E., Sherrill-Mix, S., Mukherjee, R., Custers-Allen, R., David, P., Brown, M., ... & Bushman, F. D. (2012). Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic acids research*, *40*(20), 10345-10355.
21. Sherrill-Mix, S., Ocwieja, K. E., & Bushman, F. D. (2015). Gene activity in primary T cells infected with HIV89. 6: intron retention and induction of genomic repeats. *Retrovirology*, *12*(1), 1-19.
22. Bakkour, N., Lin, Y. L., Maire, S., Ayadi, L., Mahuteau-Betzer, F., Nguyen, C. H., ... & Tazi, J. (2007). Small-molecule inhibition of HIV pre-mRNA splicing as a novel antiretroviral therapy to overcome drug resistance. *PLoS pathogens*, *3*(10), e159.

23. Emery, A., & Swanstrom, R. (2021). HIV-1: to splice or not to splice, that is the question. *Viruses*, *13*(2), 181.
24. Bodem, J. (2011). Regulation of foamy viral transcription and RNA export. *Advances in virus research*, *81*, 1-31.
25. Truman, C. T. S., Järvelin, A., Davis, I., & Castello, A. (2020). HIV Rev-isited. *Open Biology*, *10*(12), 200320.
26. Zolotukhin, A. S., Michalowski, D., Smulevitch, S., & Felber, B. K. (2001). Retroviral constitutive transport element evolved from cellular TAP (NXF1)-binding sequences. *Journal of Virology*, *75*(12), 5567-5575.
27. Liu, X., Hong, T., Parameswaran, S., Ernst, K., Marazzi, I., Weirauch, M. T., & Bass, J. I. F. (2020). Human virus transcriptional regulators. *Cell*, *182*(1), 24-37.
28. Yukl, S. A., Kaiser, P., Kim, P., Telwatte, S., Joshi, S. K., Vu, M., ... & Wong, J. K. (2018). HIV latency in isolated patient CD4+ T cells may be due to blocks in HIV transcriptional elongation, completion, and splicing. *Science translational medicine*, *10*(430), eaap9927.
29. Nguyen Quang, N., Goudey, S., Ségéral, E., Mohammad, A., Lemoine, S., Blugeon, C., ... & Gallois-Montbrun, S. (2020). Dynamic nanopore long-read sequencing analysis of HIV-1 splicing events during the early steps of infection. *Retrovirology*, *17*(1), 1-24.
30. Emery, A., Zhou, S., Pollom, E., & Swanstrom, R. (2017). Characterizing HIV-1 splicing by using next-generation sequencing. *Journal of virology*, *91*(6), e02515-16.

31. Purcell, D. F., and MALCOLM A. Martin. "Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity." *Journal of virology* 67.11 (1993): 6365-6378.
32. Arzalluz-Luque, Á., & Conesa, A. (2018). Single-cell RNAseq for the study of isoforms—how is that possible?. *Genome biology*, *19*(1), 1-19.

Chapter 5

Barcoded HIV-1 reveals the proviral transcription in clonally expanded T cells

1. Abstract

Clonal expansion of the infected T cells is the major reason HIV-1 latency reservoir can persist for a long time. The clonally expanding T cells have a transcriptomic feature of virus suppression and cell replication. A direct and high-throughput quantification of HIV-1 transcription activity in these cells are needed. Here we introduced a new sequencing technique to simultaneously measure virus integration site, virus transcription and host cell clonal expansion, enabling us to trace virus and host cell clonal dynamics *in vivo*. We found quantitative negative correlation between virus transcription and host cell expansion. The conformation of the provirus and the position of the integration sites both significantly affect these activities.

2. Introduction

The barrier to HIV-1 cure is the persistence of the latent reservoir. Although modern anti-retroviral therapy (ART) can control the viral load of HIV-1 patient at an undetectable level and greatly improve the prognosis, the drugs do not eliminate the proviruses or the infected cells, leaving a stable reservoir of dormant viruses. The virus replication will rebound when the therapy is interrupted or drug resistant virus emerges[1]. The proviruses can even reactivate at low level despite of good adherence of ART[2], posing more risk on expanding the latent reservoir. Reducing the size of the latent reservoir is the central regime of the future HIV-1 therapies. The half-life of the latent reservoir in ART-treated patients is ~44 months[3]. On the one hand, the reservoir size is decreasing due to cytotoxic effect of HIV-1 activation and inflammatory pyroptosis of the infected T cells[4]. On the other hand, the antigen specific T cell clonal expansion, homeostatic T cell replication, HIV-1 mediated oncogenic effect and sporadic virus replication all lead to the

increase of the reservoir size[5]. Understanding how these biological processes affect the dynamics of latent reservoir is the key to design therapies for HIV-1 cure.

Various techniques investigate the development history of the latent reservoir. Quantitative viral outgrowth assay (QVOA) is the most used reference for the replication competent reservoir size[6]. Digital droplet PCR and quantitative PCR targeting different parts of the HIV-1 genome are faster alternative methods to it[7,8]. But these methods lack the resolution to distinguish different proviral lineages and are hard to explain the causes of proviral clonal expansion. The high mutation rate of HIV-1 reverse transcription introduces ~1 mutation every round of the viral life cycle[9]. The mutations serve as a natural barcode of the proviral genome. With the full-length HIV-1 proviral sequencing, the proviral clonal history can be inferred from the abundance of different virus variants[10]. The integration site can also be considered as a unique proviral lineage barcode, because the chances of two random integration events occur at the same position on the human genome is nearly zero[11]. The abundance of the certain proviral clone can be inferred from the types of truncated DNA fragments during sequencing library preparation[12]. Recent advancement in single cell multi-omics enables provirus' mutations, integration sites and T cell phenotypes to be measured simultaneously[13-17], improves the sensitivity and accuracy of measuring proviral lineages in the latent reservoir.

Our group introduces synthetic genetic barcodes on HIV-1 genome to efficiently trace viral lineages and measure latent reservoir diversity in animal models[18,19]. Here, we presented a new method that can simultaneously sequence the proviral integration site and the viral barcode, and quantify the absolute number of each provirus clone. Combining it with our previous sequencing method for barcode abundance in viral RNA,

we can reconstruct the detailed population structure of the latent reservoir. The abundance of the integration site indicates the clonal expansion of the provirus with the host T cell. The proviral barcode records the history of virus establishing infection and re-seeding the latent reservoir. The RNA viral barcode revealed the replication activity of the corresponding proviral lineage. In humanized mice model, we found T cell clonal expansion for both active provirus and inactive provirus, but was more frequent in inactive viruses. We also found the activity of both T cell clonal expansion and viral transcription is affected by the position of the proviral integration.

3. Results

3.1. Integration site and barcode linkage sequencing reveals the structure of provirus population.

A 21-nucleotide genetic barcode was inserted downstream of the Env of an R5-tropic HIV-1 vector (Figure 1A). Every 3 nucleotide is a cytosine to avoid unwanted start codon. A short fragment of Nef Kozak sequence was repeated to keep the barcode from affecting Nef translation. The barcoded HIV-1 replicated at the same level of the parental strain in primary CD4+ cells, indicating the barcode does not affect the virus replication capacity (Figure S1A). ~100 thousand clones was harvested during library construction, minimizing the probability of two proviral clones have identical barcodes in vivo. The frequency of barcode was uniform (Figure S1B). There are ~10 nucleotide difference between two randomly selected barcodes, enabling easy identification of the barcodes using deep sequencing (Figure S1C). The frequency of the barcode stays identical after virus packaging in HEK293T cells and passaging in primary CD4+ T cells (Figure S1D,

S1E), indicating the barcode sequences pose no selection advantage to the virus replication.

To understand the clonal dynamics of proviruses in vivo, we used this library in humanized mice. we intravenously injected barcoded NFNSX and monitored acute viremia by qRT-PCR in plasma samples for 4 weeks. The injected virus library has ~100 thousand unique barcodes, while only a few hundred HIV-1 virions can establish infection in one humanized mouse. This ensures each virus lineage will be represented by different barcode in one mouse. After 4 weeks, we sacrificed 5 mice and extracted viral RNA and proviral DNA from plasma, spleen, bone marrow and the implant. The rest mice were administered ART comprised of raltegravir (RAL), emtricitabine (FTC), and tenofovir disoproxil fumarate (TDF) in the animal feed for 6 weeks. At the end of the suppressed period, we sacrificed 7 mice and assayed their virus population. ART was then interrupted for the rest 8 mice. After virus rebound in these mice, we harvested their organs (Figure 1B).

We developed a novel sequencing technique to profile virus barcode and the integration site simultaneously (Figure 1C, Figure S2A, See Methods for details). It uses a unique molecular identifier (UMI) to link the barcode sequence and the integration site, while quantifying the absolute number of proviral molecules. The quantification correlated with standard proviral qPCR within the viral load range of 4 logs (ρ=0.6127, p=2.886×10^(-6), Spearman's correlation test, Figure 1D). The number of proviruses quantified by UMI is ~10 fold higher than qPCR because the sequencing library was prepared with multiple sensitive nested PCR steps. The detection limit of this method is 2 copies per μg DNA (Figure 1E). We sequenced a total of 292,552 proviral molecules in 3 organs of 20 mice.

The proviruses were classified into integrated, linear, circular and auto-integrated according to the sequence linked to the UMI (Figure S2B, S2C). Integrated, circular and auto-integrated provirus decreased 14∼22 fold during latency. While linear provirus decreased ∼148 fold during latency (Figure 1F). This indicates linear provirus is a short-lived form produced during active viral replication, while other proviral forms are more persistent and important in the latent reservoir[20].

3.2. Quantifying the events of T cell clonal expansion and virus reseeding.

We then estimated the frequency of virus replication and T cell clonal expansion as we rationalised above. The count of integration site per barcode represents the number of re-seeding event of a viral lineage subtracting the number of extinction events of the provirus (infected T cell) lineage. The count of integration site per barcode is significantly reduced during latency (Figure 2A, p=1.60×10^(-8), ranksums test), suggesting a massive extinction of provirus lineages. The count increased during rebound phase (Figure 2A, p=9.47×10^(-3), ranksums test), because the viral population expanded and resulted in many new re-seeding events. The per mouse average also supported this conclusion (Figure 2B).

The count of UMI per integration site represents the number of infected T cell clonal expansion events of a provirus. If the UMI count of an integration site is more than 1, the corresponding T cell has undergone at least one clonal expansion events. If the UMI count equals to 1, the infected T cells could either be not expanding or not sampled. We found the count of T cells with UMI count more than 1 is significantly higher in acute phase than that in the latency phase (Figure 2C, p=0.039, ranksums test). This is consistent with the observation in human where extensive T cell activation and expansion took place

during the acute phase[21]. We then focused on the T cells that have undergone clonal expansion and found the average UMI per integration site were not significantly different (Figure 2D, $p>0.05$, ranksums test). But the count of T cell clones with clonal expansion were significantly higher in acute phase than that in the latency (Figure 2E, $p=0.011$, ranksums test). It is more T cell clones that were activated during the acute phase but not the intensity of activation was stronger. The analysis including all T cells also supported this conclusion (Figure S3).

We analyzed the virus population in spleen, bone marrow and the implant. The total number of virus lineage, infected T cell lineage and different forms of proviruses were not significantly different among organs (Figure S4). The overlap frequency of the viral barcode among organs indicates the exchange of virus clones among different local reservoirs, while the overlap frequency of the integration sites represents the T cell migration events among reservoirs (Figure 3A). We observed both events in all three sampling time point. But the frequency of viral clone exchange were significantly lower during latency (Figure 3B, $p=4.91\times10^{(-3)}$, ranksums test), while the frequency of T cell migration were not significantly lower (Figure 3C, $p>0.05$, ranksums test). This suggested during acute and rebound phase, virions in the circulation may be the key component of viral population exchange among local reservoirs.

3.3. Analysis of the non-transcribing proviruses.

There are 10∼355 barcodes in the DNA samples of one mouse, representing the number of viral clones successfully established the infection and was detected. The number of clones was not significantly reduced during ART treatment or after rebound, indicating ART alone cannot efficiently reduce the proviral reservoir diversity in humanized mice

(Figure 4A). We also sequenced the barcodes in the viral RNA of corresponding organs. If a barcode was only observed in DNA but not RNA, it suggests all proviruses of that viral lineage was not transcribing. If a barcode was observed in both DNA and RNA, it suggests at least one provirus of that viral lineage is transcribing. We found the percentage of viral lineage being transcribing during rebound than that during the acute phase (Figure 4B, p=3.41×10^(-3), ranksums test). ART treatment significantly reduced the number of actively transcribing viral lineages, indicating accumulation of defective proviruses during latency. The number of integration sites per barcode of the active virus lineage were significantly higher that of the inactive viruses (Figure 4C, p=5.90×10^(-10), ranksums test), implying actively transcribing viruses lead to more re-seeding events.

We analyzed the distribution of actively transcribing proviruses. We found the frequency of active transcribing barcode are not significantly different among organs. But in the spleen, the acute phase has significantly more active proviruses than the rebound phase (Figure 4D, p=0.013, ranksums test). This indicates the elimination of active viruses during ART is efficient in the spleen. We also tested how different DNA conformation affect virus transcription. The number of viral lineage (barcode count) in the form of circular DNA is 3.65 fold more than integrated and linear forms (Figure S5A, p=2.66×10^(-6), ranksums test). The number of DNA molecules in the circular form is also 3.57 fold more than the others (Figure S5B, p=9.57×10^(-4), ranksums test). But the number of viral lineage observed in RNA has an opposite trend (Figure 4E). We found significantly less viral RNA barcodes from circular forms of proviruses (p=2.33×10^(-6), ranksums test). This proves circular form is a dead-end of the virus life cycle.

3.4. The positional effect of integration site on virus replication and T cell clonal expansion

We tested the correlation between viral transcription and T cell clonal expansion. As previously defined, we classified 19514 integrated proviruses in acute and rebound phase into expanded host T cells and unexpanded host T cells according to their UMI count. 27.41% clonal expanded proviruses were not transcribing, while 6.96% unexpanded proviruses were not transcribing. This suggests viral transcription is less frequent in clonally expanded host T cells (Figure 5A, OR=0.25, p=1.91×10^(-197), Fisher's exact test). The analysis counting T cell clones supports the same conclusion (Figure S6A, OR=0.59,p=2.80×10^(-6), Fisher's exact test). We focus on the T cell clones that have clonally expanded. We found the UMI count of each proviruses were also significantly higher in the non-transcribing proviruses (Figure 5B, p=6.18×10^(-4), ranksums test), indicating the number of expansion events were more frequent in non-transcribing proviruses. The analysis on all T cell clones also supports the same conclusion (Figure S6B, p=0.019, ranksums test). We then focused on the host cells that have viral RNA transcription. We found relative viral RNA abundance was significantly higher in unexpanded T cells (Figure 5C, p=1.51×10^(-3), ranksums test). The analysis on all T cell clones is affirmative (Figure S6C, p=6.70×10^(-7), ranksums test). These data suggest the negative correlation between viral transcription and T cell clonal expansion are both qualitative and quantitative.

Among 15305 unique integrated proviruses clones in this study, 9547 integrated into human genes. Compared to a randomly generated control, the provirus in our dataset are 1.17 fold more likely to be found in the gene region (p=1.86×10^(-59), Fisher's exact test).

And the distance from the integration site to the nearest genes were closer than that of a random control (Figure 6A, p=2.23×10^(-247), chi-square test). We also compared the distribution of the integration sites to annotated genome regulation elements and some histone modifications with well-defined functions (Figure 6B). We found the integration sites were enriched in host transcription regulators and active histone markers. Notably some suppressive histone markers were also enriched near the integration sites. HIV-1 integration is guided towards active host gene regions by LEDGF complex but in vivo selection in humanized mice favors the non-transcribing provirus, hence the integration site near inactive regions were also enriched.

We then asked if the T cell clonal expansion activity is affected by the genomic features near the integration site (Figure 6C). And we noticed clonally expanded host cells are more likely to have provirus integrated near active histone markers and the enhancer regions. Provirus integrated in gene regions were also more likely to expand. We focus on the T cells that have expanded and found proviruses integrated in gene regions have more expansion events (Figure 6D, p=0.014, ranksums test). This suggests integration near active host genes may interrupt with host gene functions and leads to more uncontrolled T cell activation or expansion. We also calculated the correlation between the integration sites and viral transcription (Figure 6E). We found actively transcribing viruses were not enriched in gene region or regulation elements, but more likely to locate near the suppressive histone marker H3K27me3. This contrasts common knowledge in vitro and suggests ART in vivo eliminates replication competent viruses near active host gene regions and select for viruses integrated near suppressive regions.

4. Discussion

We profiled the clonal expansion, transcription activity and integration sites of 2,726 viral lineages, 15,305 proviral clones and 292,552 proviral molecules. Our data shows HIV-1 transcription is less frequent and weaker in clonally expanded cells. The activity of T cell expansion and viral transcription is affected by the conformation of the proviral DNA and the position of the integration sites. Comparing to previous studies, this study has following progress. It is the first high-throughput and quantitative dataset in humanized mice to profile the integration sites at different stages of the disease progression. Secondly, it uses integration site sequencing data to distinguish different proviral DNA conformations. The analysis pipeline could be easily adapted to many other integration site sequencing datasets. Thirdly, it introduced UMI to quantify the clonal expansion of provirus, which is more accurate and has a wider range of quantification than the method using the length of DNA ends from random fragmentation. Moreover, the synthetic genetic barcode on the viral genome allows accurate tracing of viral lineages despite of mutations and re-seeding. Lastly, comparing with previous viral mutation and integration site parallel sequencing[17,22], this method increased the throughput for over one thousand-fold, making it possible to be applied in various conditions. This study established a series of methodology to generate and analyze a multi-dimensional dataset of HIV-1 integration, transcription and host cell clonal expansion.

We are aware that some of the analysis in this study could achieve higher statistical power if we can classify the virus transcription activity more accurately. Because a virus lineage can infect multiple host T cells, we are not certain if a provirus is transcribing viral RNA even if we observe the corresponding barcode in the RNA samples. The barcode

sequence in the viral RNA could be transcribed from any of the provirus clones containing the same barcode. We are only certain about the inactive provirus, which has no corresponding barcode in RNA samples, because viral RNA sequencing is very sensitive and reproducible[19]. This inaccurate classification of active provirus reduced the significance of our statistical tests. The classification on expanded and unexpanded T cells has a similar problem. We are only certain that a cell has expanded if the UMI count is more than 1. But for the cells with UMI count equal to 1, it could either be a dormant cell or we did not sample the other molecules from the same cell lineage. This confusion will not change the conclusions of our study, but requires us to sequence a large number of viruses to achieve statistical significance.

Many *in vitro* studies showed HIV-1 transcription is affected by local genomic features, including histone modifications, 3D nucleus structure, host promoter and enhancers[23]. Active local gene transcription and splicing facilitate viral transcription. But clinical data showed HIV-1 proviruses located mainly near inactive genomic regions in elite controllers[24]. Our data explained that the in vivo selection reduced the number of transcribing proviruses during latency, resulting in the less active provirus clone being amplified by T cell clonal expansion. We observed many proviruses integrated within cancer related genes or genes affecting T cell activation. HIV-1 integration may inhibit local gene expression by interrupting its transcription and splicing. While viral promoter may recruit transcription machinery and facilitate downstream gene expression. We observed expanded T cell clones with integration site near oncogenic genes, like EZH2 and MECOM, as well as tumor suppressor genes like CREBBP and SMARCA4. However, the exact mechanisms leading to clonal expansion remains to be studied by

transcriptomic methods. Future works can combine our technique with single cell multi-omics to elucidate the fate of the infected T cells.

5. Methods

5.1. Barcoded HIV-1 library construction

The design of the barcoded virus is shown in Figure 1A. Two fragments covering upstream and downstream of the barcode region were PCR amplified using following primers: makeBC_F2 (GGAAAGGGCTTTGCTATAAGNNCNNCNNCNNCNNCNNCNNCTATAAGATGGGTG GCAAGTGGTCAA) and makeBC_R2 (TGCAGACCCTGCACTCCATG), makeBC_F1 (GGAGTGGAAGCCATAATAAGAATT) and makeBC_R1 (CTTATAGCAAAGCCCTTTCCAAG). The products were then purified and eluted in TE buffer. The 2 fragments have 20bp overlapping region, then they were assembled together using NEB HiFi Assembly kit. The assembled fragment was amplified again using primers makeBC_F1 and makeBC_R2. The fragment was digested by restriction enzyme NcoI-HF and EcoRI-HF, and purified using PureLink PCR clean-up kit (Invitrogen). The NFNSX vector was also digested by these 2 enzymes and was purified by agarose gel electrophoresis. One μg of the insert fragment and 5μg of the vector was assembled using NEB HiFi Assembly kit. The assembled DNA was purified by ethanol precipitation with Pellet Paint NF co-precipitant (EMD millipore). One μg of the purified DNA was transformed into the MegaX E. coli electrocompetent cells (Invitrogen) and plated to 20 15cm agar plates. The plates were cultured at 37°C overnight. More than 0.5 million colonies were scratched from the surface of the plates. One mg plasmid DNA was

extracted from the bacteria pellet. The barcode region on the plasmid was confirmed by Sanger sequencing. 20 μg of plasmid DNA was transfected into 20 million 293T cells using the Calcium phosphate transfection reagent (Takara). The virus library was harvested 48 hours after transfection. DNaseI (40ng/mL) and (1) were added to the library to remove residual plasmid DNA from the supernatant. The barcoded virus library was aliquoted and frozen at -80°C for future use.

5.2. Sequencing library preparation

The barcode - integration site linkage sequencing library was prepared as the workflow shown in Figure S2A. Mouse DNA and RNA was extracted using Allprep DNA/RNA Mini kit (Qiagen). One μg DNA was subject to enzymatic fragmentation using HinP1I (NEB). The digested DNA was then purified by PureLink PCR clean-up kit (Invitrogen). UltraII End-repair Module (NEB) prepared the DNA for ligation. A custom adaptor was annealed in the TE buffer. The sequence of the adapter's reverse strand is TTGAGGTTTGCAGTTG. It has a 5-prime modification of a phosphorylation group, which facilitates TA ligation with the genome fragments. The 3 prime amino modification blocks the polymerase from adding nucleotides at its downstream, maintaining the L-shape conformation of the adapter. 3 consecutive phosphorothioate bonds at the 3 prime end stabilize the adapter, preventing it from enzymatic degradation. The forward strand of the adapter is ACCATCAACCCCGAATTCNNNNNNNNNNNNNNCAACTGCAAACCTCAAT. It anneals with the reverse strand and contains a 14-nucleotide UMI. 50pmol adapter was ligated to 1μg of fragmented genomic DNA. All ligated product was purified and amplified using 4 rounds of semi-nested PCR. All PCRs used the same reverse primer sequence: ACCATCAACCCCGAATTC. But the forward primer sequences anneal to different part of

HIV-1 genome to increase the PCR specificity. They are in the order of F4 (GCTACCACCGCTTGAGAGAC), F3 (CTCTTGACTGTAACGAGGATTG), F2 (GAACTTCTGGGACGCAGGG) and F1 (GTACTCGAATTCAGGGCTTGGAAAGGGCTTTG). All forward primers contain 3 consecutive phosphorothioate bonds at the 3-prime end, preventing the exonuclease activity of the polymerase, increasing the PCR specificity. Primer F3 and F2 are used with the reverse primer containing the 5-primer phosphorylation modification, to enable lambda exonuclease digestion after PCR, which can eliminate the product of unspecific amplification. The final PCR product was purified and digested by EcoRI-HF (NEB). This created two sticky ends on the DNA. 100ng DNA was purified and subject to self-ligation in a 100μL reaction. The reaction used 2 units of T4 ligase (Invitrogen) in room temperature for 4 hours. The ligation efficiency was confirmed by quantitative PCR using primers ivF (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTAGTCAGTGTGGAAAATCTCT), ivR (ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTGACCACTTGCCACCCAT) and synthetic standard templates. One third of the ligation product was used as the PCR template for the inverse PCR, using the same primers as the quantitative PCR. Phosphorothioate bond modification was used to increase PCR specificity. One tenth of the product was then subject to the final round of PCR, which adds Illumina sequencing adapters to the library. The primers are AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNACACTCTTTCCCTACA CGAC and

CAAGCAGAAGACGGCATACGAGATNNNNNNNNNNGTGACTGGAGTTCAGACGTGT GC. N stands for indexing sequence to distinguish different samples.

The RNA was reverse transcribed using RT primer AGGATGAGTGTCAGCCAGTGNNNNNNNNNN-NNNNNCATTCTTTCCCTTACAGTAGACC and purified by PureLink PCR clean-up kit (Invitrogen). Here 21 consecutive random nucleotides serve as a UMI for each RNA molecule. The barcode region was amplified using primers amp_F1 (TTGGTGGAATCTCCTACAGTATTGG) and amp_R (AGGATGAGTGTCAGCCAGTG). The PCR product was amplified again using primers amp_F2 (TACAAGAATAAGACAGGGCTTGG) and amp_R. The final product was confirmed using gel electrophoresis and purified by the PCR clean-up kit. We then use NEBNext Ultra II DNA library prep kit to make pair-end sequencing libraries. All libraries were mixed and purified for Illumina NovaSeq6000 PE150 sequencing. 10 million reads were retrieved for each sample.

5.3. Data Analysis

The data analysis pipeline was summarized in Figure S2B and S2C. For the barcode - integration site linkage sequencing, the barcode and UMI was extracted by mapping their flanking sequences. The sequence downstream of the HIV-1 LTR and upstream of the L-shape adapter was extracted as provirus integration site. If the sequence maps to the plasmid of NFNSX, it is discarded as contamination. if the sequence is less than 10 nucleotides, it is considered as linear unintegrated provirus. The rest sequences were aligned with human genome hg38 Ensemble release 108 or the NFNSX genome by bowtie2[25], and classified as integrated or auto-integrated. If the sequence maps

immediately downstream of the HIV-1 LTR, it is classified as circular. We then identified the true UMIs from the sequencing errors by counting the occurrence of the UMIs[26]. The count of the true UMI should follow a normal distribution while the sequencing errors were Poisson distributed. We set the threshold of calling true UMI for each sample to the separation point of the bi-modal count distribution. Then we assigned the most commonly observed barcode and integration site for each UMI. With the help of UMI, we identified the barcode and integration site for each provirus molecule.

The barcode and UMI in RNA samples were also retrieved by mapping their flanking sequences. The true UMI was identified by its count distribution. The occurrence of barcode in RNA was quantified by counting UMI. An extra clustering step was carried out for barcode to reduce sequencing errors[27]. Custom codes for mapping and counting were available upon requests.

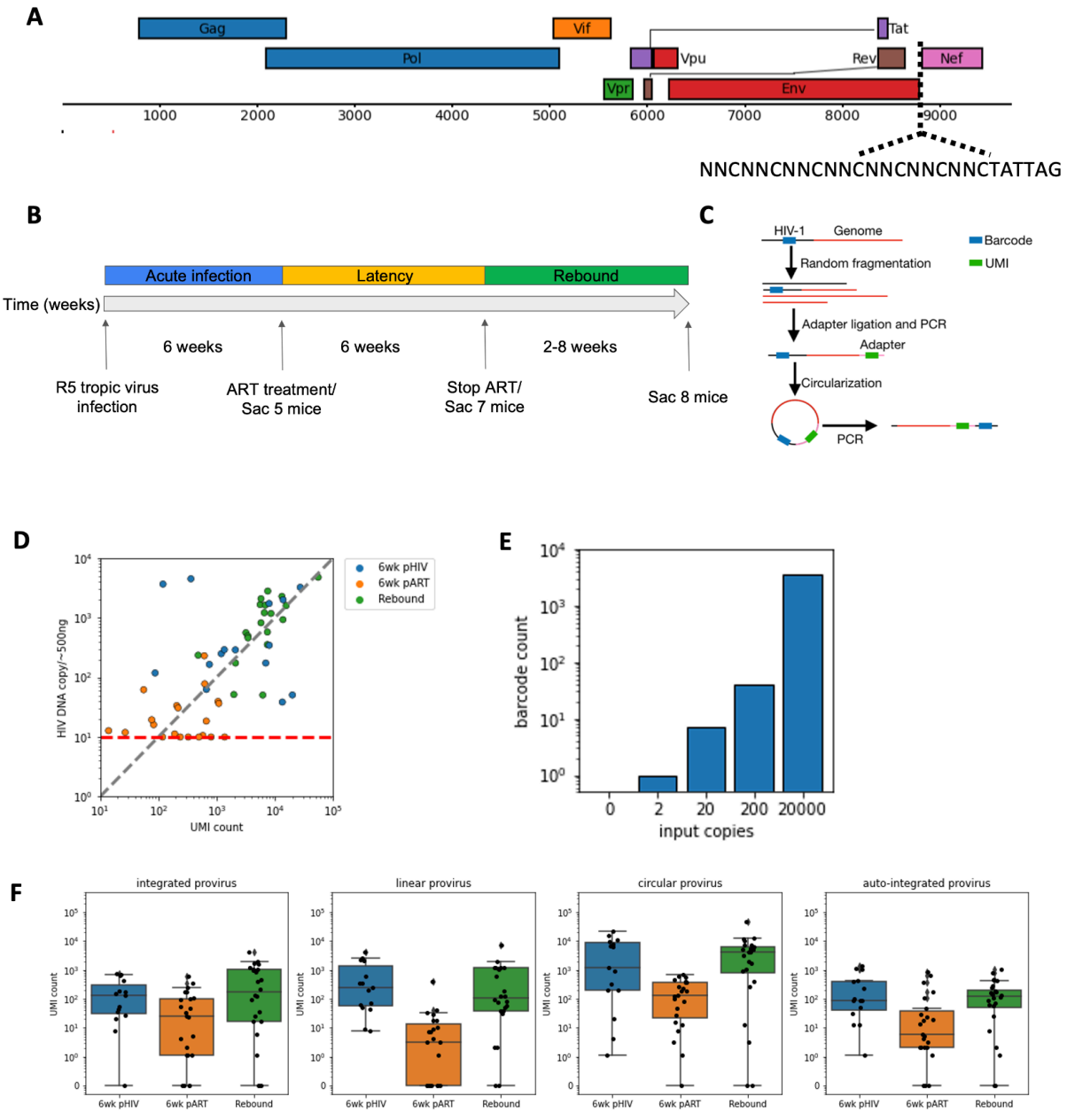


Figure 5-1. The barcode - integration site sequencing accurately measures the latent reservoir in humanized mice.

A) The design of genetically barcoded NFNSX. A 21-nt barcode was inserted between Env and Nef. B) Schematic representation of the mouse experiment. Acute infection of

20 mice with barcoded NFNSX for 6 weeks. ART for 6 weeks. Rebound for 2-8 weeks. C) Workflow of the barcode - integration site linkage sequencing. Briefly, genomic DNA was randomly fragmented and ligated with a UMI labelled L-shape adapter. Then we used a series of semi-nested PCR to amplify the integration junction. The amplified product was circularized to bring barcode and the integration site together. Lastly, we used PCR to amplify the barcode, UMI and integration site region and append the sequencing adapter. D) The correlation between proviral DNA qPCR and viral load quantified by UMI. Each dot represents an organ. E) The detection limit of the barcode - integration site linkage sequencing. We mixed different copies of plasmid of the barcoded virus with mouse genomic DNA and run the sequencing protocol. Because the barcode library has a complexity of ~100 thousand, the possibility of having two plasmid molecule with identical barcode is negligible. The number of barcodes recovered from the sequencing data is $15\% \sim 50\%$. F) The number of different conformations of the provirus in different sampling time. Integrated, linear, circular and auto-integrated provirus was classified according to the sequence attached to the UMI.

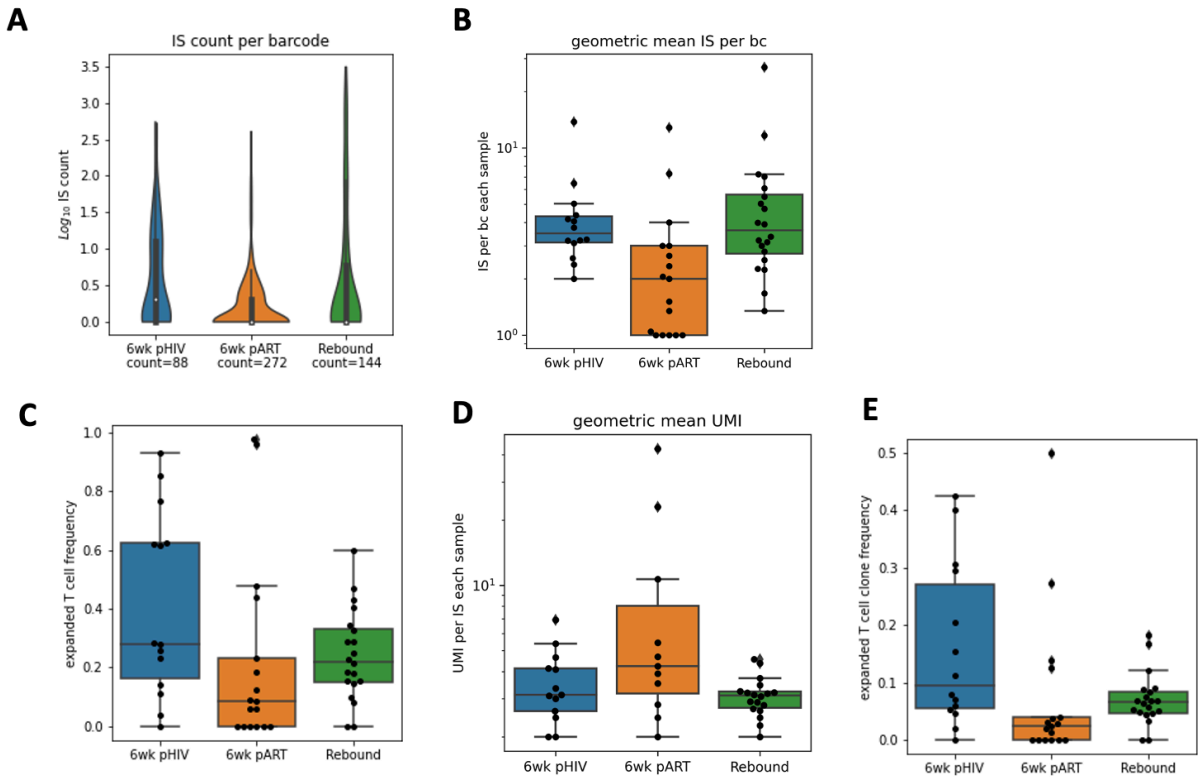
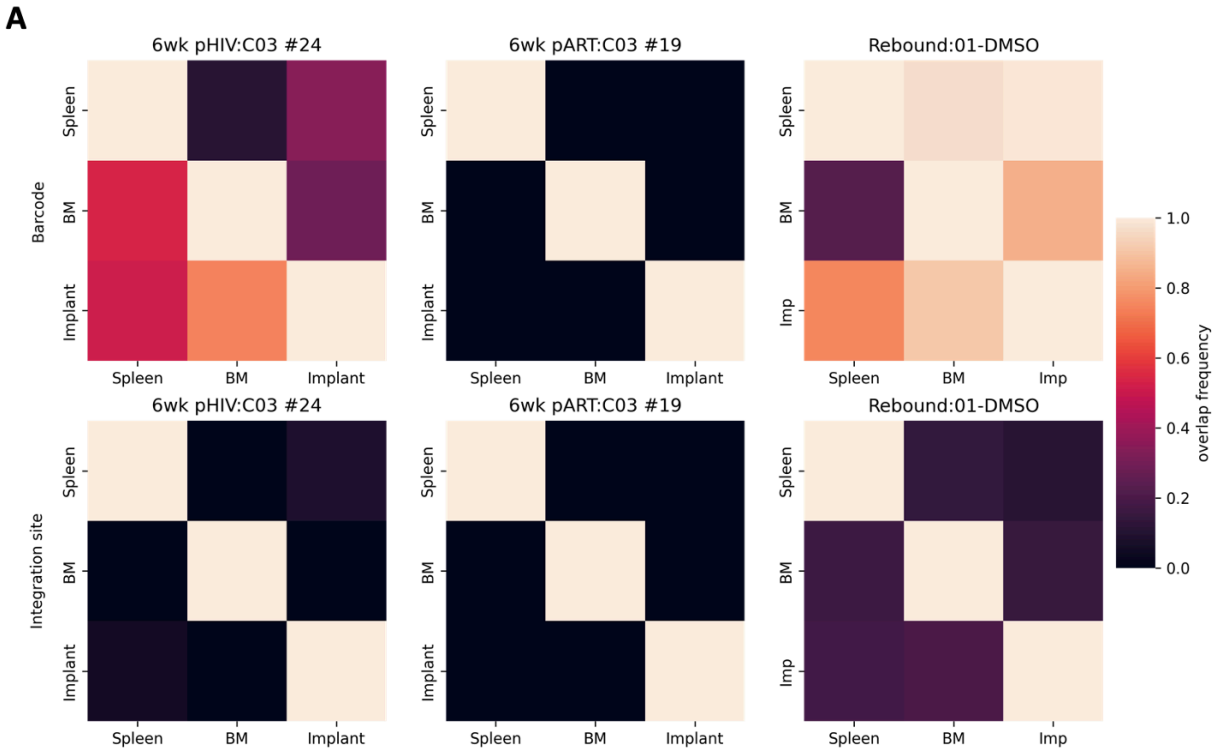
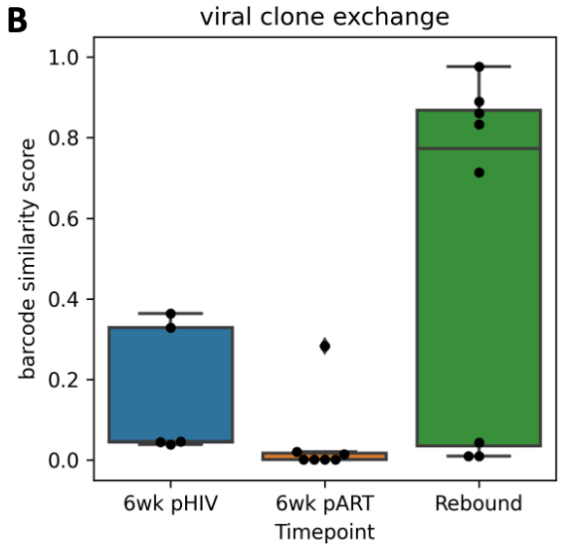


Figure 5-2. Estimating the number of virus re-seeding and T cell clonal expansion.

A) The number of integration sites per barcode. The value represents the number of re-seeding events per viral clone. Only integrated provirus was counted. B) Geometric mean of the number of integration sites per barcode. Each dot represents an organ sample. C) Relative frequency of T cells (proviral molecules) that have clonally expanded. Clonally expanded T cells was identified if the integration site has more than 1 UMI. D) Geometric mean of the number of UMI per integration site. The value represents the number of clonally expansion events per T cell clone. E) Relative frequency of T cell clones that have clonally expanded.


A
6wk pHIV:C03 #24
6wk pART:C03 #19
Rebound:01-DMSO
Barcode
Integration site
Spleen
BM
Implant
Imp
overlap frequency
1.0
0.8
0.6
0.4
0.2
0.0


B
viral clone exchange
barcode similarity score
6wk pHIV
6wk pART
Rebound
Timepoint

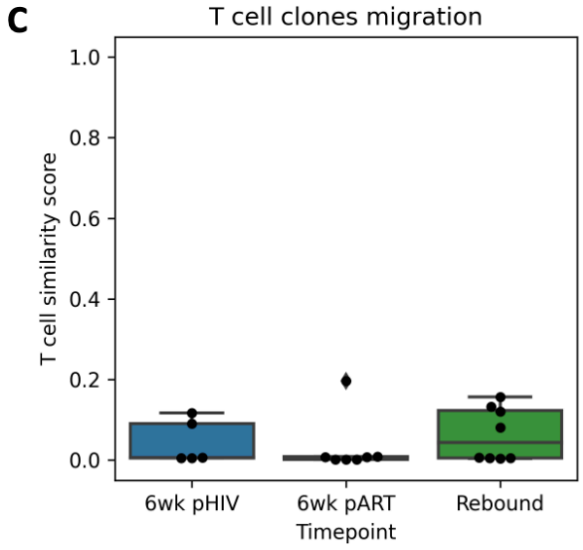

C
T cell clones migration
T cell similarity score
6wk pHIV
6wk pART
Rebound
Timepoint

Figure 5-3. Virus exchange and T cell migration among organs.

A) Heatmaps showing barcode and integration site overlap among organs. One representative mouse from each sampling time was selected. The overlap frequency between organ A (horizontal label) and organ B (vertical label) is the frequency of virus/cell in organ A also observed in organ B. B) The barcode similarity score at different time point. The similarity score is defined as the average overlap frequency among any two organs. Each dot represents one mouse. C) The integration site similarity score at different time point.

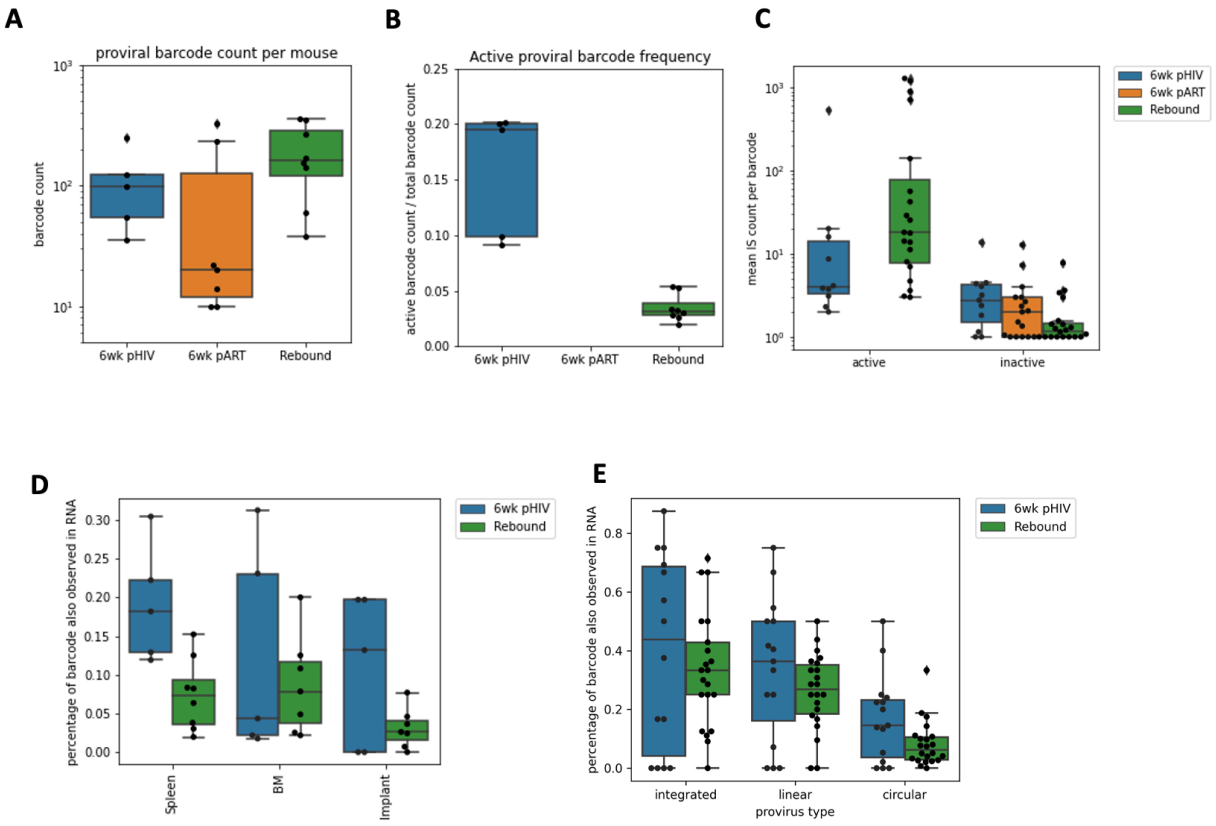


Figure 5-4. The transcriptional activity of proviruses.

A) The total number of proviral barcode in each mouse. B) The relative frequency of actively transcribing proviruses. Actively transcribing proviruses was defined if the proviral barcode is also observed in RNA. C) Geometric mean of the number of integration site per barcode for active provirus and inactive provirus. Each dot represents an organ. D) The relative frequency of active provirus in different organs. E) The relative frequency of active provirus as different provirus conformations.

A

Cell count	Expanded T cell	Unexpanded T cell
Transcribing viral RNA	5126	12138
Not transcribing viral RNA	1405	845

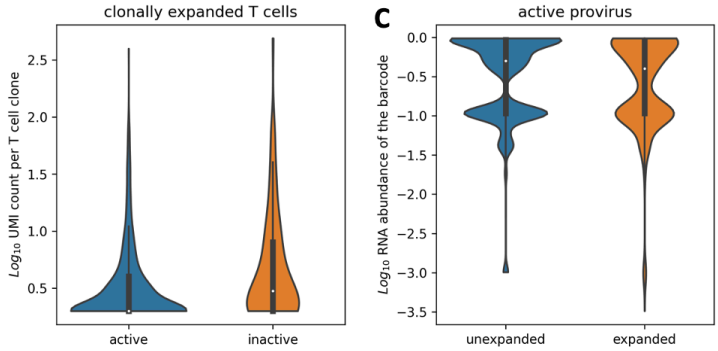
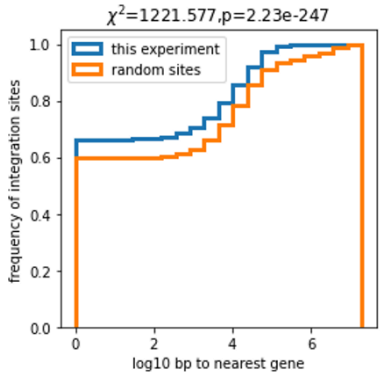


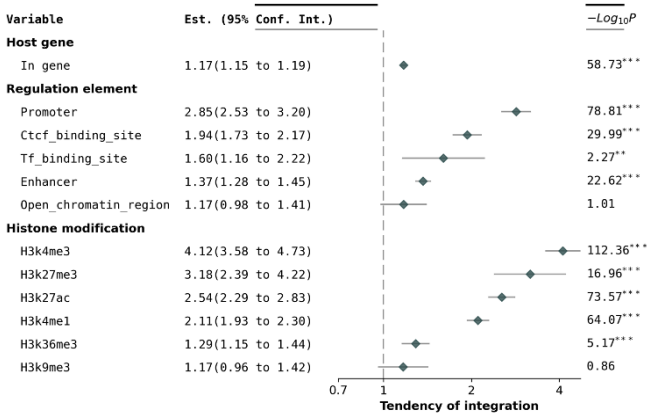
Figure 5-5. The correlation between T cell expansion and proviral transcription.

A) The contingency table of T cell clonal expansion and provirus transcription. The number of cells in each catalogue was listed. B) The number of clonal expansion events of actively transcribing provirus and inactive provirus. The analysis focuses on the T cells that already have clonal expansion. C) The relative proviral transcription activity of clonally expanded T cells and unexpanded T cells. Only actively transcribing provirus was included in the analysis.

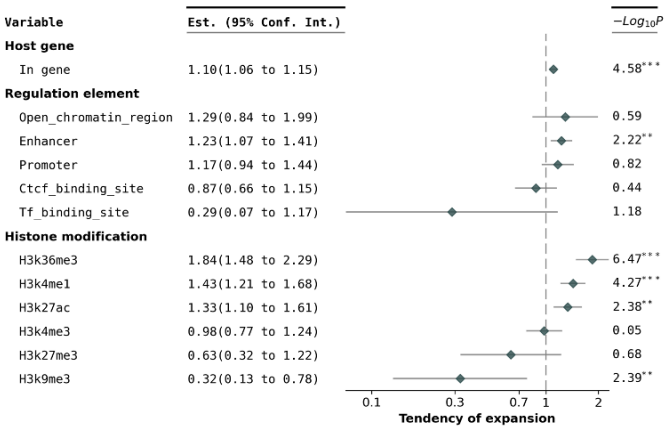
A


χ^2=1221.577,p=2.23e-247
this experiment
random sites
frequency of integration sites
log10 bp to nearest gene

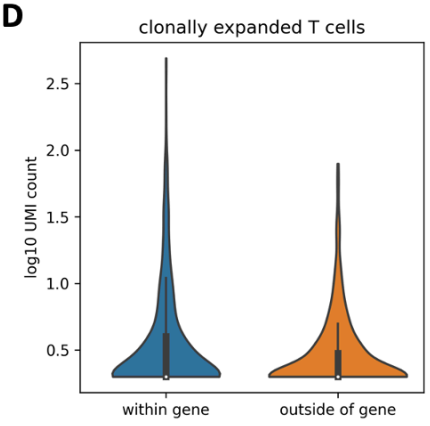
B


Variable
Est. (95% Conf. Int.)
$-Log_{10}P$
Host gene
In gene 1.17(1.15 to 1.19) 58.73***
Regulation element
Promoter 2.85(2.53 to 3.20) 78.81***
Ctcf_binding_site 1.94(1.73 to 2.17) 29.99***
Tf_binding_site 1.60(1.16 to 2.22) 2.27**
Enhancer 1.37(1.28 to 1.45) 22.62***
Open_chromatin_region 1.17(0.98 to 1.41) 1.01
Histone modification
H3k4me3 4.12(3.58 to 4.73) 112.36***
H3k27me3 3.18(2.39 to 4.22) 16.96***
H3k27ac 2.54(2.29 to 2.83) 73.57***
H3k4me1 2.11(1.93 to 2.30) 64.07***
H3k36me3 1.29(1.15 to 1.44) 5.17***
H3k9me3 1.17(0.96 to 1.42) 0.86
Tendency of integration

C


Variable
Est. (95% Conf. Int.)
$-Log_{10}P$
Host gene
In gene 1.10(1.06 to 1.15) 4.58***
Regulation element
Open_chromatin_region 1.29(0.84 to 1.99) 0.59
Enhancer 1.23(1.07 to 1.41) 2.22**
Promoter 1.17(0.94 to 1.44) 0.82
Ctcf_binding_site 0.87(0.66 to 1.15) 0.44
Tf_binding_site 0.29(0.07 to 1.17) 1.18
Histone modification
H3k36me3 1.84(1.48 to 2.29) 6.47***
H3k4me1 1.43(1.21 to 1.68) 4.27***
H3k27ac 1.33(1.10 to 1.61) 2.38**
H3k4me3 0.98(0.77 to 1.24) 0.05
H3k27me3 0.63(0.32 to 1.22) 0.68
H3k9me3 0.32(0.13 to 0.78) 2.39**
Tendency of expansion

D


clonally expanded T cells
log10 UMI count
within gene
outside of gene

E

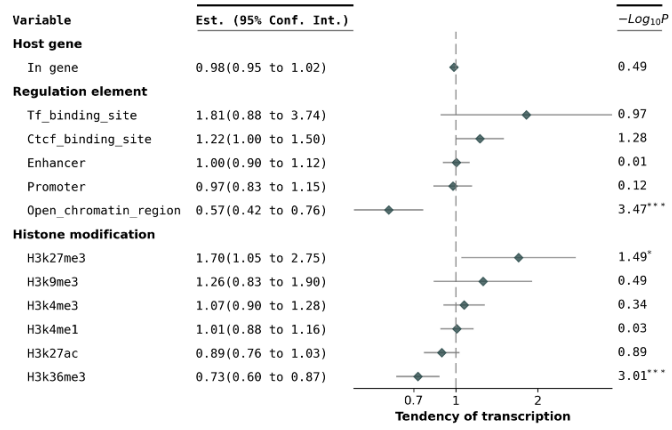
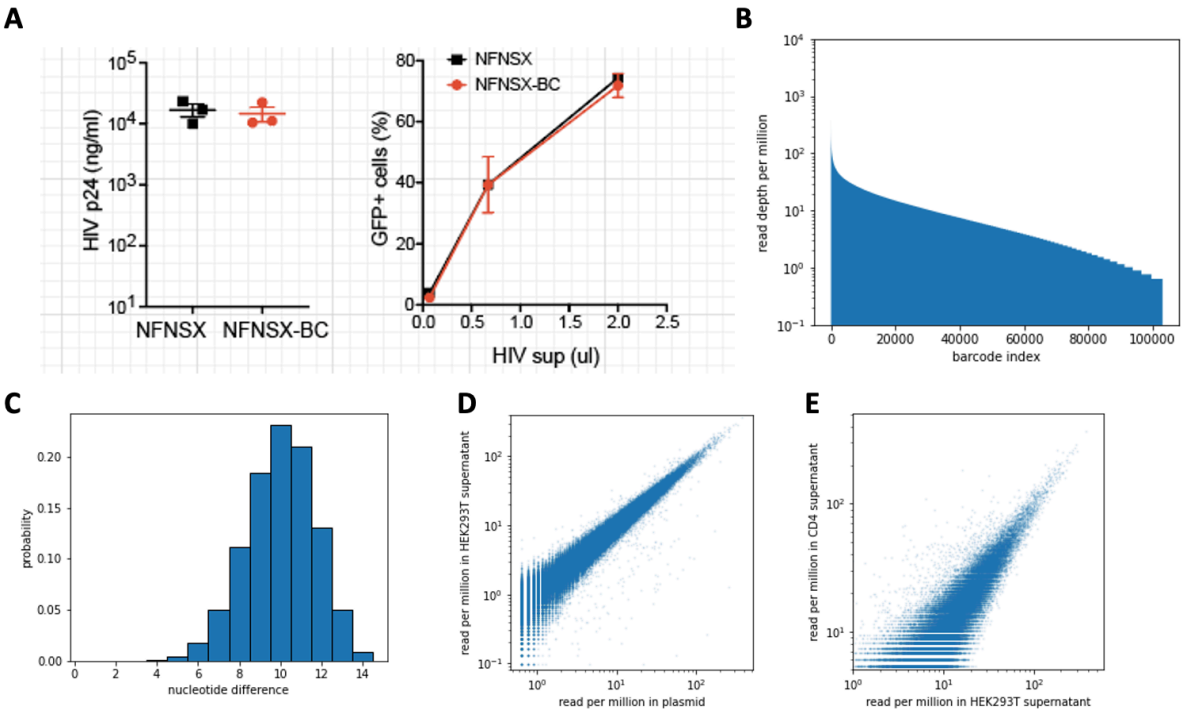

Variable
Est. (95% Conf. Int.)
$-Log_{10}P$
Host gene
In gene 0.98(0.95 to 1.02) 0.49
Regulation element
Tf_binding_site 1.81(0.88 to 3.74) 0.97
Ctcf_binding_site 1.22(1.00 to 1.50) 1.28
Enhancer 1.00(0.90 to 1.12) 0.01
Promoter 0.97(0.83 to 1.15) 0.12
Open_chromatin_region 0.57(0.42 to 0.76) 3.47***
Histone modification
H3k27me3 1.70(1.05 to 2.75) 1.49*
H3k9me3 1.26(0.83 to 1.90) 0.49
H3k4me3 1.07(0.90 to 1.28) 0.34
H3k4me1 1.01(0.88 to 1.16) 0.03
H3k27ac 0.89(0.76 to 1.03) 0.89
H3k36me3 0.73(0.60 to 0.87) 3.01***
Tendency of transcription

Figure 5-6. The positional effect of integration sites on T cell clonal expansion and proviral transcription.

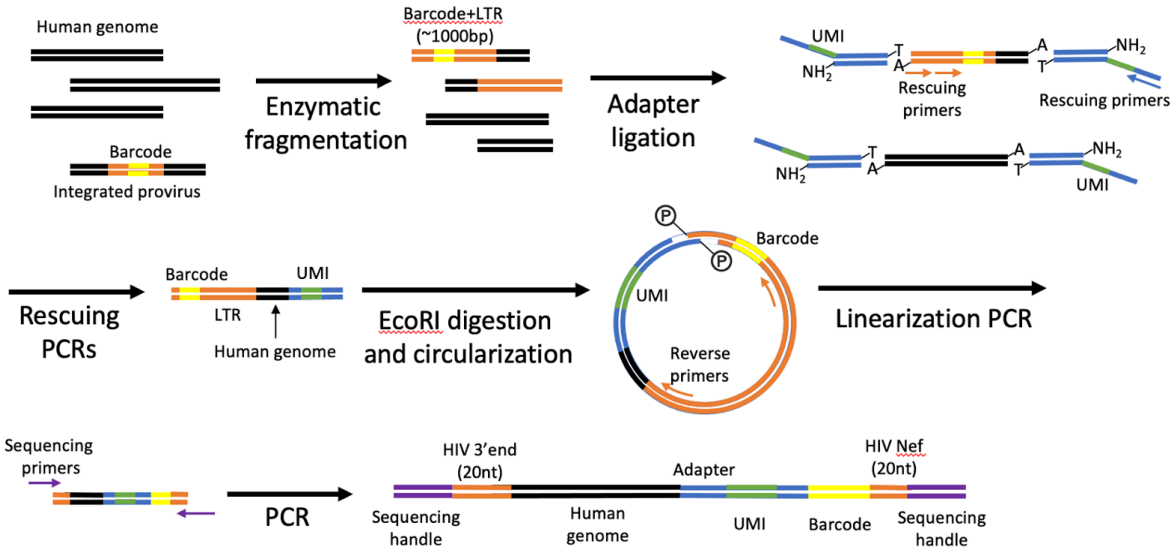
A) The distance to nearest host gene of the integration sites in this study. A randomly generated list with the same number of integration sites were used as a control. B) The correlation between the integration sites and nearby genomic features. The coefficient plot shows the odds ratio and p value of the Fisher's exact test, comparing our dataset with the randomly generated control. C) The correlation between T cell clonal expansion and nearby genomic features of the integration sites. The coefficient plot shows the odds ratio and p value of the Fisher's exact test, comparing expanded and unexpanded T cells. D) The number of T cell clonal expansion events of provirus integrated within host gene or in the intergenic regions. Only the T cells underwent clonal expansion were included in the analysis. E) The correlation between provirus transcription and nearby genomic features of the integration sites. The coefficient plot shows the odds ratio and p value of the Fisher's exact test, comparing actively transcribing provirus and inactive provirus.



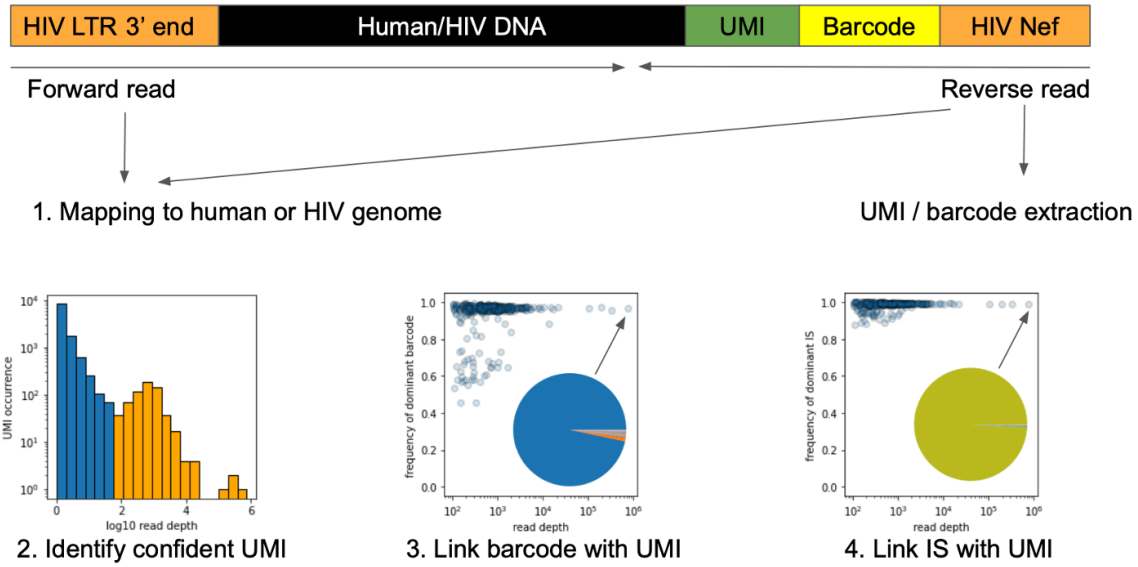
Supplementary Figure 5-1. The quality of the genetically barcoded HIV-1 library.

A) The replication capacity of the barcoded virus. B) The read depth distribution of the barcodes in the plasmid library. C) The distribution of hamming distance between any two barcode sequences. D) The correlation between barcode frequency from the plasmid and that from the virus stock produced in HEK293T. E) The correlation between barcode frequency from the virus stock produced in HEK293T and that from infected primary CD4+ cells.

A



B

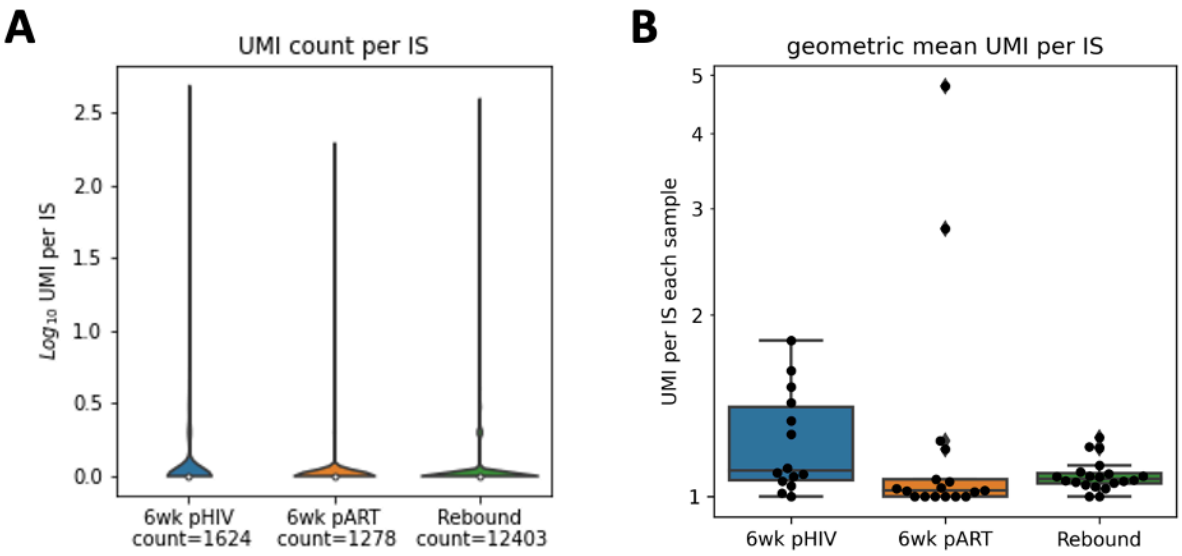


C

Provirus classification	DNA sequence
Integrated	Mapped to reference human genome (hg38)
Linear	Shorter than 10bp
Circular	Mapped to HIV-1_{NFNSX}, immediate downstream of 5'LTR
Auto-integrated	Mapped to HIV-1_{NFNSX}, but not circular

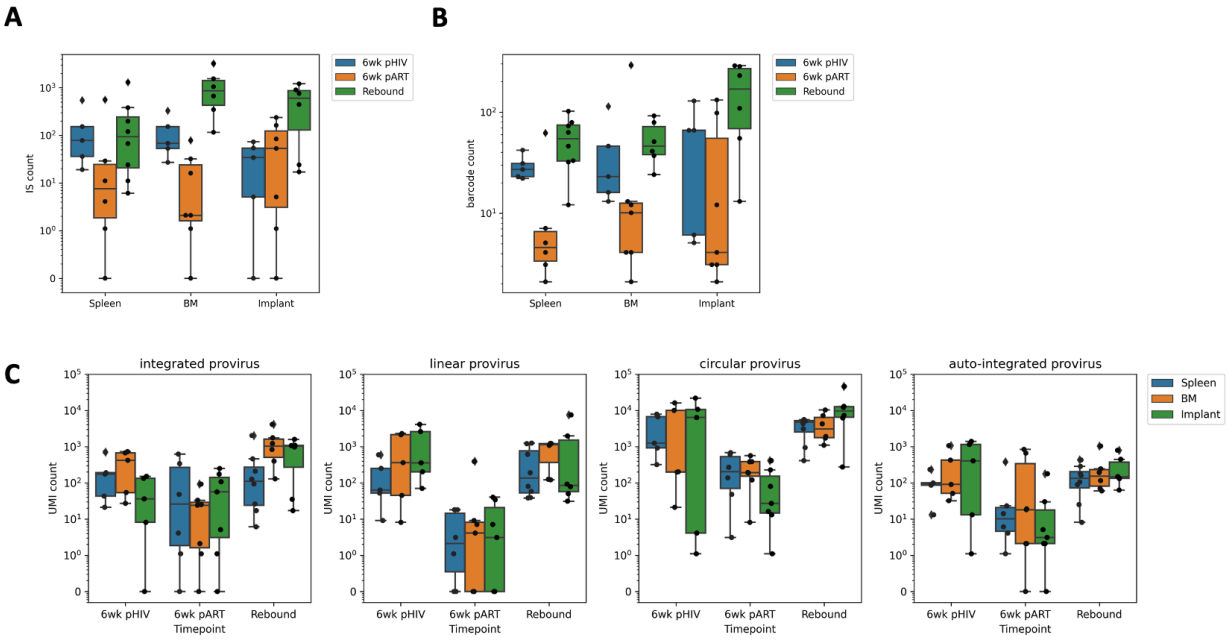
Supplementary Figure 5-2. The workflow of barcode - integration site linkage sequencing.

A) The detailed workflow of barcode - integration site sequencing library preparation. B) The data analysis pipeline for integration site mapping, UMI calling and linkage assignment. C) The criteria of distinguishing different proviral DNA conformations. See the Methods section for more details.



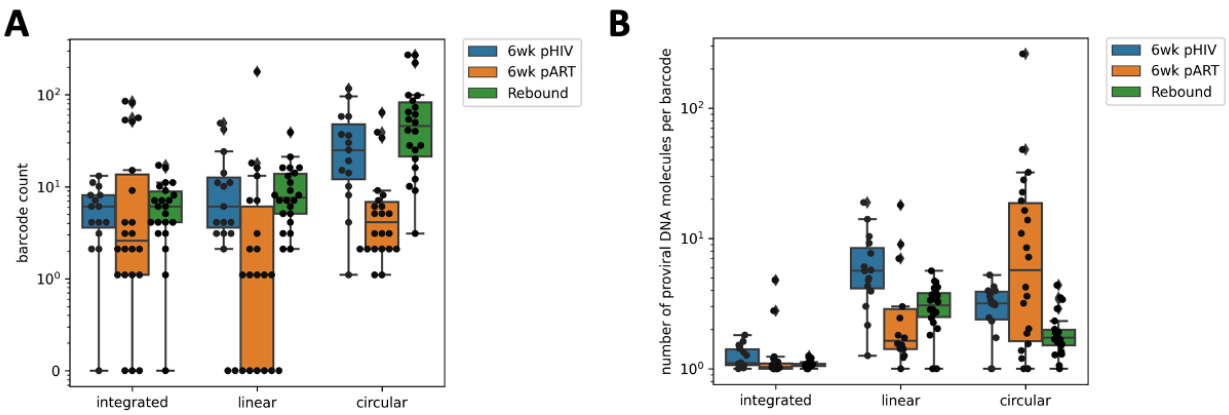
Supplementary Figure 5-3. The clonal expansion of infected T cells.

A) The number of UMI per integration site at different time point. It represents the number of clonal expansion events for each infected T cell. B) The geometric mean of the number of UMI per integration site at different time point.



Supplementary Figure 5-4. The virus population in different organs.

A) The number of integration sites in different organs. B) The number of barcodes in different organs. C) The number of proviral DNA molecules as different conformations in different organs.

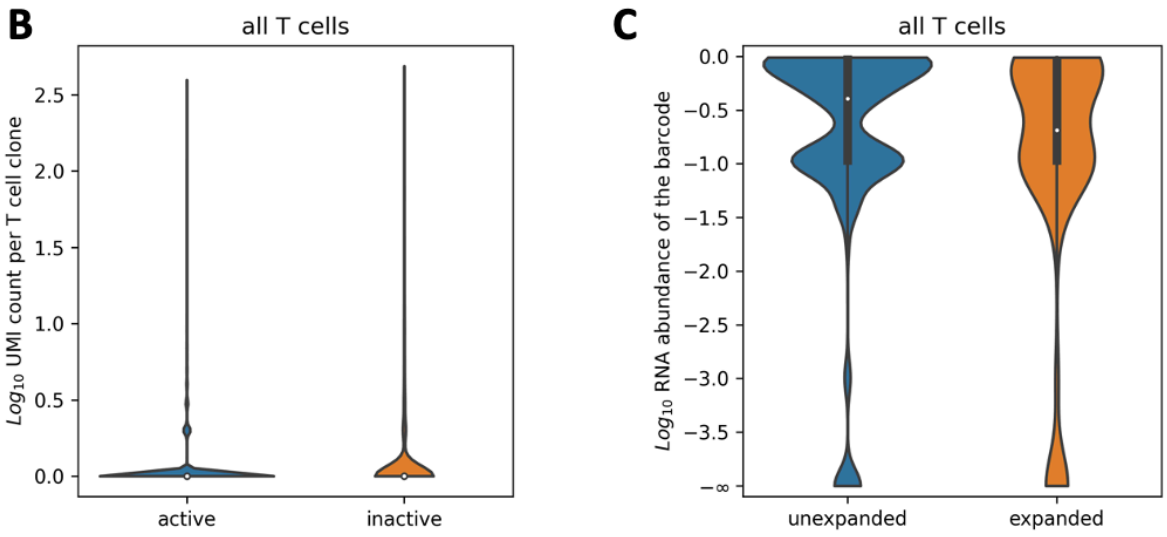


Supplementary Figure 5-5. The activity of proviral DNA in different conformations.

A) The number of barcodes in different conformations. B) The number of proviral DNA molecules per barcode as different conformations. Each dots represents an organ.

A

Clone count	Expanded T cell	Unexpanded T cell
Active provirus	934	12138
Inactive provirus	110	845



Supplementary Figure 5-6. The correlation between T cell expansion and proviral transcription.

A) The contingency table of T cell clonal expansion and provirus transcription. The number of clones in each catalogue was listed. B) The number of clonal expansion events of actively transcribing provirus and inactive provirus. The analysis uses all cells, including expanded and unexpanded T cells. C) The relative proviral transcription activity of clonally

expanded T cells and unexpanded T cells. All provirus was included in the analysis, including transcribing and inactive ones.

6. References

1. B Nachega, J. et al. HIV treatment adherence, drug resistance, virologic failure: evolving concepts. Infect. Disord. Targets (Formerly Curr. Drug Targets-Infectious Disord. 11, 167–174 (2011).
2. Halvas, E. K. et al. Hiv-1 viremia not suppressible by antiretroviral therapy can originate from large t cell clones producing infectious virus. The J. clinical investigation 130, 5847–5857 (2020).
3. Siliciano, J. D. et al. Long-term follow-up studies confirm the stability of the latent reservoir for hiv-1 in resting cd4+ t cells. Nat. medicine 9, 727–728 (2003).
4. Doitsh, G. et al. Cell death by pyroptosis drives cd4 t-cell depletion in hiv-1 infection. Nature 505, 509–514 (2014).
5. Yeh, Y.-H. J., Yang, K., Razmi, A. & Ho, Y.-C. The clonal expansion dynamics of the hiv-1 reservoir: Mechanisms of integration site-dependent proliferation and hiv-1 persistence. Viruses 13, 1858 (2021).
6. Wong, J. K. et al. Recovery of replication-competent hiv despite prolonged suppression of plasma viremia. Science 278, 1291–1295 (1997).
7. Bruner, K. M. et al. A quantitative approach for measuring the reservoir of latent hiv-1 proviruses. Nature 566, 120–125 (2019).
8. Gaebler, C. et al. Combination of quadruplex qpcr and next-generation sequencing for qualitative and quantitative analysis of the hiv-1 latent reservoir. J. Exp. Medicine 216, 2253–2264 (2019).
9. Roberts, J. D., Bebenek, K. & Kunkel, T. A. The accuracy of reverse transcriptase from hiv-1. Science 242, 1171–1173 (1988).

10. Ho, Y.-C. et al. Replication-competent noninduced proviruses in the latent reservoir increase barrier to hiv-1 cure. Cell 155, 540–551 (2013).
11. Wagner, T. A. et al. Proliferation of cells with hiv integrated into cancer genes contributes to persistent infection. Science 345, 570–573 (2014).
12. Berry, C. C. et al. Estimating abundances of retroviral insertion sites from dna fragment length data. Bioinformatics 28, 755–762 (2012).
13. Liu, R. et al. Single-cell transcriptional landscapes reveal hiv-1–driven aberrant host gene transcription as a potential therapeutic target. Sci. translational medicine 12, eaaz0802 (2020).
14. Einkauf, K. B. et al. Parallel analysis of transcription, integration, and sequence of single hiv-1 proviruses. Cell 185, 266–282 (2022).
15. Collora, J. A. et al. Single-cell multiomics reveals persistence of hiv-1 in expanded cytotoxic t cell clones. Immunity (2022).
16. Clark, I. C. et al. Hiv silencing and cell survival signatures in infected t cell reservoirs. Nature 1–8 (2023).
17. Sun, W. et al. Phenotypic signatures of immune selection in hiv-1 reservoir cells. Nature 1–9 (2023).
18. Marsden, M. D. et al. Tracking hiv rebound following latency reversal using barcoded hiv. Cell Reports Medicine 1, 100162 (2020).
19. Kim, J. T. et al. Latency reversal plus natural killer cells diminish hiv reservoir in vivo. Nat.communications 13, 1–14 (2022).
20. Sloan, R. D. & Wainberg, M. A. The role of unintegrated dna in hiv infection. Retrovirology 8, 1–15(2011).

21. Butz, E. A. & Bevan, M. J. Massive expansion of antigen-specific cd8+ t cells during an acute virus infection. Immunity 8, 167–175 (1998).

22. Lian, X. et al. Signatures of immune selection in intact and defective proviruses distinguish hiv-1 elite controllers. Sci. translational medicine 13, eabl4097 (2021).

23. Chen, H.-C., Martinez, J. P., Zorita, E., Meyerhans, A. & Filion, G. J. Position effects influence hiv latency reversal. Nat. structural & molecular biology 24, 47–54 (2017).

24. Jiang, C. et al. Distinct viral reservoirs in individuals with spontaneous control of hiv-1. Nature 585, 261–267 (2020).

25. Langdon, W. B. Performance of genetic programming optimised bowtie2 on genome comparison and analytic testing (gcat) benchmarks. BioData mining 8, 1–7 (2015).

26. Fennessey, C. M. et al. Genetically-barcoded siv facilitates enumeration of rebound variants and esti- mation of reactivation rates in nonhuman primates following interruption of suppressive antiretroviral therapy. PLoS pathogens 13, e1006359 (2017).

27. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. Cd-hit: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152 (2012).

Chapter 6

A benchmark study on error-correction by read-pairing and tag-clustering in amplicon-based deep sequencing

1. Abstract

Background

The high error rate of next generation sequencing (NGS) restricts some of its applications, such as monitoring virus mutations and detecting rare mutations in tumors. There are two commonly employed sequencing library preparation strategies to improve sequencing accuracy by correcting sequencing errors: read-pairing method and tag-clustering method (i.e. primer ID or UID). Here, we constructed a homogeneous library from a single clone, and compared the variant calling accuracy of these error-correction methods.

Result

We comprehensively described the strengths and pitfalls of these methods. We found that both read-pairing and tag-clustering methods significantly decreased sequencing error rate. While the read-pairing method was more effective than the tag-clustering method at correcting insertion and deletion errors, it was not as effective as the tag-clustering method at correcting substitution errors. In addition, we observed that when the read quality was poor, the tag-clustering method led to huge coverage loss. We also tested the effect of applying quality score filtering to the error-correction methods and demonstrated that quality score filtering was able to impose a minor, yet statistically significant improvement to the error-correction methods tested in this study.

Conclusion

Our study provides a benchmark for researchers to select suitable error-correction methods based on the goal of the experiment by balancing the trade-off between sequencing cost (i.e. sequencing coverage requirement) and detection sensitivity.

2. Background

Next-generation sequencing is being widely used in biomedical research. Several sequencing technologies, such as chained ligation (SOLiD), pyrosequencing (454), reversible dye (Illumina), fluorescent nucleotides (PacBio), and ion semiconductor (Ion Torrent) have been developed and commercialized. While different technologies have their own features (e.g. long read-length for PacBio and high output for Illumina), high sequencing error rate is a common problem for all existing next generation sequencing platforms. The high error rate significantly impedes the application of these technologies to detect rare variants in genetically heterogeneous populations.

To resolve the problems associated with the high error rate, experimental methods have been developed for distinguishing real mutations from sequencing errors. One such method is to take advantage of the paired-end feature of Illumina sequencing by removing the inconsistent forward and reverse read pairs [1–5]. Another common approach is to use nucleotide tags [6–12]. Although variations of sequencing library prepration method using nucleotide tags have been proposed, the underlying philosophy is the same. Briefly, a highly heterogeneous pool of random oligonucleotides (also known as tags or Primer IDs) is assigned to the individual nucleic acid molecules to label the original template copy. Subsequently, the same tag would be observed in different reads. This can be considered as resampling of the same original DNA template. By comparing the sequence reads that share the same tag, a corrected consensus sequence can be generated, and stochastic sequencing errors can be distinguished from real mutations. Recently, another innovative approach, known as circle sequencing [13], has been developed. With a similar design to tag-clustering methods, circle sequencing allows each

DNA template to be read multiple times on a single read. These sequencing error-correction methods have been successfully applied to detect rare mutations in heterogeneous cancer tissues [14], mixed microbe populations [15], and viral quasispecies [10].

In this study, a highly uniform plasmid template from a single bacteria clone was sequenced. We applied the read-pairing correction method, as well as tag-clustering correction method to the same template. We systematically compared the error profiles and sequencing coverage of different methods to describe the pros and cons of each strategy.

3. Results

3.1. Experimental design

To compare the efficiency of different error-correction methods, the sequencing library was prepared from a clonal plasmid carrying the protein G antibody interacting domain (Fig. 1). An 88 bp region of this domain was amplified through PCR. The sequence is shown in Additional file 1: Figure S1. The length of the target region in this study was similar to the read-length being used in amplicon-based deep sequencing cancer studies [16, 17]. The target region contained 54.5 % GCs. In comparison, the average GC content of human genes ranges from 34 % to 66 % [18]. Therefore, the properties of the target region in this study resembled that of the sequences of interest in other applications.

The target region was first amplified by PCR. A tag, comprising eight random nucleotides “N”, was included in both forward and reverse primers. Thus, a total of 16 random nucleotides were present in the resultant PCR product. The complexity of the tags was ~ 4×109 per sample. Around 6×106 tagged molecules were then amplified to generate

identical copies of each tagged molecule. The product from this second PCR was subjected to deep sequencing on the Illumina HiSeq 2500 platform. In this study, two technical replicates from the same clone were included. We were expecting ~5 copies per tagged molecule to be sequenced, with ~30 million sequencing reads in total. This experimental design allowed us to perform two independent error-correction approaches, namely read-pairing consensus and tag-clustering consensus. Read-pairing consensus, which was based on the sequence identities of the forward and reverse reads, was used to filter out read pairs that were unmatched. Tag-clustering consensus was 1) to group the reads by the tag sequence, and 2) to filter out groups that carried reads with different sequence identities. Based on these two error-correction approaches, we compared the results from four types of analyses: Scheme 1: Raw reads; Scheme 2: Read-pairing consensus; Scheme 3: Tag-clustering consensus; Scheme 4: Combined consensus (read-pairing consensus, followed with tag-clustering consensus).

3.2. Error rate profiling

In this study, sequencing errors were categorized into four types namely transition (A↔G and C↔T), transversion (A↔C, A↔T, G↔C, and G↔T), insertion and deletion.

In the raw sequencing data, all four error types were identified. They distributed with a peak at 10^{-4} per nt and a long tail to 10^{-2} per nt (Fig. 2 a, Scheme 1 forward and reverse). The error rate was not normally distributed (Additional file 2: Figure S2, $p < 2.2 \times 10^{-16}$, Shapiro-Wilk normality test). The transition rate had a median of 3.3×10^{-4} per nt and a mean of 1.5×10^{-3} per nt. The transversion rate had a median of 5.7×10^{-4} per nt and a mean of 3.1×10^{-3} per nt, which was ~2-fold higher than transition rate. The rates of insertion and deletion errors were not normally distributed either. The rates of insertions

and deletions were 10-fold lower than that of substitutions (i.e. transition and transversion), confirming that the insertion and deletion errors in Illumina platform were relatively low [19]. The insertion rate had a median of 3.2×10^{-5} per nt and a mean of 2.9×10^{-4} per nt, while the deletion rate had a median of 1.3×10^{-4} per nt and a mean of 5.3×10^{-4} per nt.

All error-correction schemes improved the sequencing results significantly. But different schemes showed different advantages for correcting different error types (Fig. 2 a). Read-pairing consensus (Scheme 2) significantly reduced insertion and deletion rates by ~100-fold ($p= 9.6\times10^{-60}$, Wilcoxon signed-rank test). In contrast, transition and transversion rates were only reduced by no more than 10-fold ($p= 2.0\times10^{-59}$, Wilcoxon signed-rank test). Tag-clustering consensus (Scheme 3) reduced substitution error rates ~20 fold ($p= 3.9\times10^{-58}$, Wilcoxon signed-rank test), but the decrease in insertion and deletion rates was only significant at the middle region of the sequencing reads ($p= 9.6\times10^{-60}$, Wilcoxon signed-rank test).

Read-pairing consensus showed significantly lower insertion and deletion rates than tag-clustering consensus ($p= 8.0\times10^{-53}$, Wilcoxon signed-rank test), while transition and transversion rates were lower in tag-clustering consensus than that in read-pairing consensus ($p= 2.4\times10^{-12}$, Wilcoxon signed-rank test). Combined consensus performed the best for both substitution rates ($p = 1.5\times10^{-38}$, Wilcoxon signed-rank test) and insertion and deletion rates ($p = 2.9\times10^{-25}$, Wilcoxon signed-rank test). The medians for all four categories of errors in different analysis scheme were shown in Fig. 2 b. In conclusion, the tag-clustering correction method was very effective for substitution errors,

but not for insertion and deletion errors. In contrast, the read-pairing method was very effective for insertion and deletion errors, but not for substitution errors.

In the unfiltered dataset, the error rate of reverse reads was ~3 times higher than that of forward reads ($p = 1.0 \times 10^{-91}$, Wilcoxon signed-rank test). This is likely due to a lower quality of reverse reads, which resulted from oxidation during the sequencing run [20]. Notably, there were some high rate errors in the reverse reads, marked as blue arrows in Fig. 2 a. At position 57, the transversion error rate was as high as 12.4 %. In the raw sequencing reads, this position often displayed as 'N', which resulted from poor base-calling quality during the sequencing run. After tag-clustering correction, this error was significantly decreased, but was still at 3.4 %. Although our analysis showed that tag-clustering consensus performed better than read-pairing consensus in handling substitution errors, this advantage was not seen in this particular case, which implied the low robustness of tag-clustering method. In conclusion, high quality reads are necessary for avoiding erroneous results from tag-clustering scheme and achieving effective information utilization.

Notably, there were some real mutations in the templates that may arise from potential sources, including mutation accumulation during bacteria clonal formation, PCR procedures, and cross contamination of single mutant samples. Those mutations were buried in the unfiltered dataset but were easily identified after error correction, as indicated by the red arrows in Fig. 2 a. The frequencies of real mutations did not change significantly before and after error-correction. This result showed the necessity of error-correction methods for detecting low frequency variants.

3.3. Reproducibility

To confirm the reproducibility of our result, we compared two technical replicates from the same template. All four categories of errors were highly correlated between the technical replicates (Fig. 3 a). The high correlation between the error profiles of the raw data implied a sequence-specific error pattern for Illumina sequencing platform [21]. This correlation remained high after error-correction, suggesting that the error-correction methods retained the sequence-specific error patterns.

The prevalence of sequence-specific errors was also evident in the correlation between the forward reads and reverse reads (Fig. 3 b). Even for the exact same batch of templates, error patterns between forward reads and reverse reads differed dramatically, as shown by the low correction coefficient. The correlation remained low after tag correction, implying its weakness at correcting sequence-specific errors.

To further examine the error reproducibility, we did a linear regression for the different schemes (Additional file 3: Figure S3). We used the results from the combined consensus to approximate the true mutation rates. According to the previous conclusion, the rates of real mutations remain similar after error-correction, which mapped on the diagonal lines of Additional file 3: Figure S3a. But the sequencing errors were reduced significantly using combined consensus which mapped on the up-left panel of Additional file 3: Figure S3a. Thus, most observed insertions and deletions were due to sequencing errors. However, most observed substitutions comprise both sequencing errors and mutations from the templates.

3.4. Quality score and coverage loss

Coverage loss was one of the major concerns in using the error-correction methods. We counted the read number after each error-correction schemes (Fig. 4 a). The coverage of read-pairing correction was 42 % of the raw sequencing data, which was similar to the ideal 50 % loss. Forward reads of tag-clustering correction reached a coverage of 12 % (20 % in the ideal case), while the reverse reads had only 0.4 %. Combined consensus had 6 % coverage of the original data (ideally 10 %). Therefore, our study has shown that using correction methods increases the sequencing cost per nucleotide ~2.4 fold (1/0.42 $\approx$ 2.4) for read-pairing correction, ~8.3 fold (1/0.12 $\approx$ 8.3) for tag-clustering method (based on forward reads), and ~17 fold (1/0.06 $\approx$ 17) for combined consensus. There was a significant trade-off between detection sensitivity and coverage. Researchers needs to consider the balance between coverage loss and detection limit when choosing a suitable error-correction method.

4. Discussion

Over the last decade, next-generation sequencing has become a popular technique in biomedical research due to its increasing throughput and decreasing cost. Illumina sequencing platform is the most widely used next generation sequencing platform, having two shortcomings: high error rate and short read-length. While Illumina has been increasing its read-length through the recent development of MiSeq platform, the error rate remains at ~0.1 % to 1 % per nt. This error rate may be negligible in certain applications that only require the information of consensus sequence, such as cellular genome sequencing and transcriptome profiling. However, such error rate will significantly impede those applications that require the detection of rare mutations.

Consequently, different experimental approaches have been implemented to overcome this drawback [4–8, 10, 11, 13, 24]. In general, these approaches sacrifice read coverage for a higher sensitivity. Thus, error-correction indirectly increases the per nucleotide cost of sequencing. Therefore, the type of error-correction method should be selected based on the desired sensitivity to minimize the sequencing cost. Here, we proposed several guidelines for choosing an error-correction method, for Illumina HiSeq platform.

1. Error-correction methods should be applied if the required detection limit is lower than 1%.
2. Read-pairing method is sufficient for detecting variants with frequencies higher than 0.1%, and is effective for detecting rare insertions and deletions.
3. Tag-clustering method is necessary for detecting variants with frequencies lower than 0.1%. However, extra depth and high-quality data is needed for carrying out tag-clustering method.
4. Coupling tag-clustering method and read-pairing method is recommended.

We notice that tag-clustering error-correction methods could not avoid certain types of errors. We propose several reasons. Firstly, the sequencing platforms use the first few nucleotides to estimate the parameters for phasing correction. The sequence of tags could induce systematic errors. The templates with the same tags would have the same error in this phasing process [21]. Secondly, the templates with tags were all sequenced at the same time. Thus, the buffer quality could result in quality drop at the same position of all reads, which could make tags unable to correct the errors. Thirdly, tags were not amplified or sampled evenly during library preparation. The DNA polymerase had bias for

certain primers. In this study, we achieved a polynomial distribution of tags (Additional file 4: Figure S4), which reduced the third systematic error. But tag region itself generated bias.

There are some caveats that limit the power of this study. Firstly, random nucleotide tags were added to the template by PCR. Thus, errors that emerged during the PCR steps cannot be corrected. Such errors should exist here despite a high fidelity DNA polymerase was being used to minimize the PCR errors. The true mutations are therefore comprised of mutations in the original templates (within clone variation), and PCR induced errors. Moreover, there may be cross-contamination from other experiments being performed in the lab that involved mutagenesis. Sampling during plasmid extraction, template amplification, and dilution will also add to the heterogeneity of the templates. In short, the true mutation rate of the sequencing template is not known in this study, which prevents us from precisely quantifying the error rate in each error correction scheme.

While not being addressed in this study, there are numerous computational error-correction methods being developed [25–28]. Most, if not all, of these computational approaches were developed to handle raw sequencing reads. While this study indicates that read filtering based on quality score may only slightly improve the sensitivity, it is unknown whether the sensitivity for deep sequencing may benefit further from combining experimental approach and computational approach. Benchmarking for such integrative error-correction strategy is needed to be done in the future.

Amplicon sequencing is becoming a more popular approach in various research fields because of its high sequencing coverage of a target region of interest. Amplicon sequencing has been widely used in cancer research for diagnosis and disease

monitoring purposes [16, 17, 29, 30]. In addition, amplicon sequencing on 16S rDNA gene and other conserved regions is commonly used to characterize the genetic structure of microbe communities [31–33]. Nonetheless, depending on the specific goal, different studies may investigate different genetic regions of interest from different sources of specimens, and employ different sequencing platforms with different read-lengths. In the future, the performance of error-correction strategies should also be evaluated with the consideration of additional parameters, such as samples with extreme GC contents and various degree of genetic diversity, and the usage of other sequencing platforms.

5. Methods

5.1. Sequencing library preparation

The target sequence was a synthetic construct of protein G on the pCR-Blunt vector [34] (Additional file 1: Figure S1a). Clonal protein G sequencing template was amplified by PCR using primer pair (replicate 1): 5'-CTA CAC GAC GCT CTT CCG ATC TNN NN A CAN NNN AGT ACG CTA ACG ACA ACG G-3' and 5'-TGC TGA ACC GCT CTT CCG ATC TNN NNA CAN NNN TCG GAT CCT CCG GAT TCG G-3', or primer pair (replicate 2): 5'-CTA CAC GAC GCT CTT CCG ATC TNN NN G TGN NNN AGT ACG CTA ACG ACA ACG G-3' and 5'-TGC TGA ACC GCT CTT CCG ATC TNN NNG TGN NNN TCG GAT CCT CCG GAT TCG G-3'. The underlined nucleotides were served as distinguishing replicate 1 and 2. The eight randomized nucleotides, 4 Ns from each of the forward and reverse primer were served as the tag for error-correction. The entire amplified region (including the primer annealing region) on protein G was 5'-AGT ACG CTA ACG ACA ACG GTG TCG ACG GTG AAT GGA CCT ACG ACG ACG CTA CCA AAA CCT TCA CGG TTA CCG AAT CCG GAG GAT CCG A-3'. The condition of this first PCR was as

follow: 2 mins at 95 °C, then 18 three-step cycles of 20 seconds at 95 °C, 15 seconds at 58 °C, and 20 seconds at 68 °C, and a 1 min final extension at 68 °C. The PCR product was purified using PureLink PCR Purification Kit (Life Technologies, Carlsbad, CA). For each sample, ~6 million copies of the purified PCR product were used for the second PCR. The second PCR was performed using primer pair: 5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG-3' and 5'-CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG-3'. The condition of the second PCR was the same as that of the first PCR, except 22 cycles were performed instead of 18. All PCRs were performed using KOD DNA polymerase (EMD Millipore, Billerica, MA) with 1.5 mM MgSO4, 0.2 mM of each dNTP (dATP, dCTP, dGTP, and dTTP) and 0.5 μM each of the forward and reverse primers. The resultant product was sequenced by Illumina HiSeq 2500 platform.

5.2. Data analysis

Illumina HiSeq paired-end reads were demultiplexed using the three bp barcode on both forward read and reverse read. The first 12 bp of the read was identified as a tag. For downstream analysis of sequencing error, this 12 bp region was trimmed. As a result, only 88 bp was processed for calculating error rate. After the dataset being processed by the indicated error-correction scheme, pairwise local alignment against the reference protein G sequence was performed. The alignment was carried out using pairwise2 function in the Biopython package [35]. The alignment scoring was as follow: 1 for identical, –1 for mismatching, –1 for gap opening, –0.5 for gap extending. All downstream analyses were performed by custom python scripts.

Error-correction Scheme 1 (no error-correction)

Errors were called from the raw read. No pairing or quality score filtering was applied on the dataset.

Error-correction Scheme 2 (read-pairing)

Pairing was performed by comparing the nucleotide sequence of the trimmed foward read and trimmed reverse read (88 bp in both cases). Only those read pairs with a reverse complementary match were used for downstream analysis.

Error-correction Scheme 3 (tag-clustering)

The tags for the forward read and reverse read were combined and used for grouping reads as described [8]. Briefly, reads that shared the same tag were grouped together as a read group. Read grouping was performed independently for forward read and reverse read. Read groups with a size of less than three reads were discarded. A read group was considered as a real read if all reads in the read group were identical. Otherwise, the read group would be discarded.

Error-correction Scheme 4 (read-pairing and tag-clustering)

First, read-pairing was performed as described in Scheme 2. The paired reads were then subjected to tag grouping as described in Scheme 3. Of note, under this scheme, read grouping was performed on the paired read instead of independently on forward read and reverse read.

5.3. Availability of supporting data

Raw sequencing data have been submitted to the NIH Short Read Archive (SRA) under accession number: BioProject PRJNA293914. Custom scripts for data analyzing and plotting were deposited in https://github.com/Tian-hao/errorcorrection.

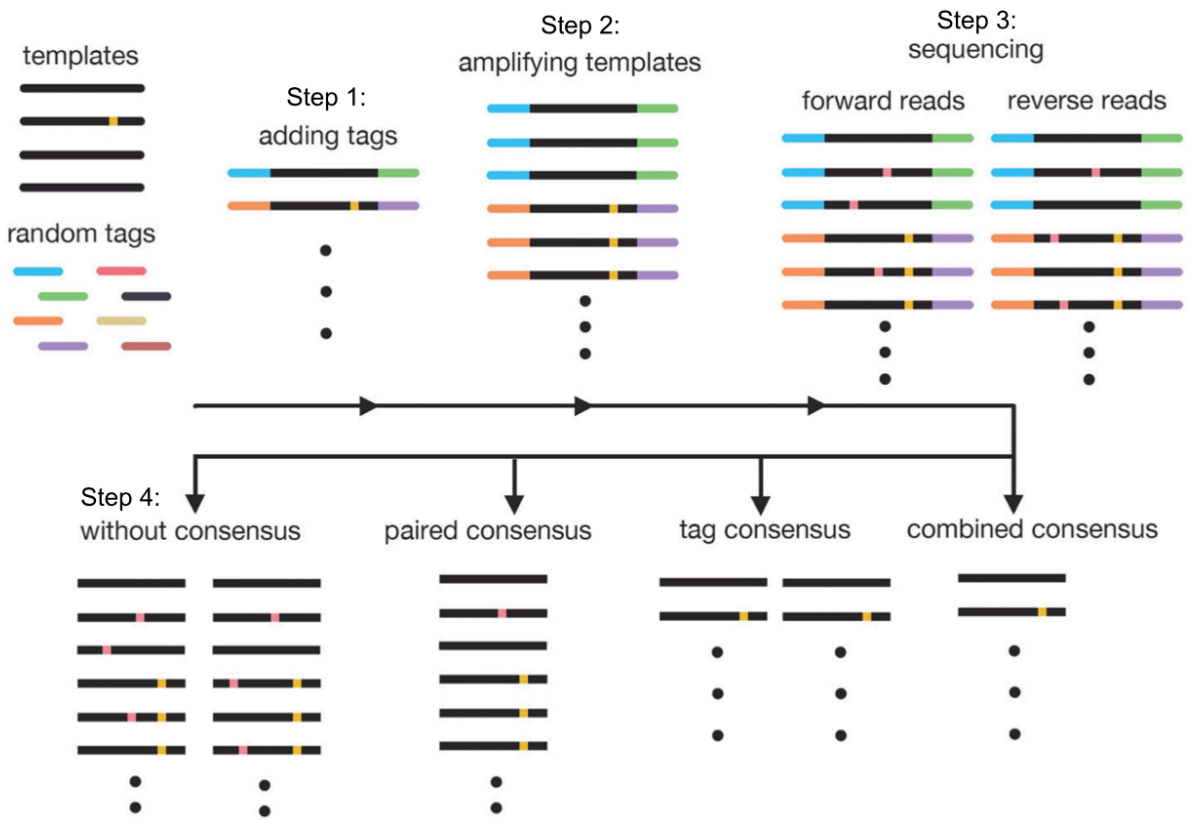


Figure 6-1. Schematic representation of the experimental design.

To compare the efficiency of different error-correction methods, we generated the sequencing library in the following steps. Step 1: Linking tags to the templates. Step 2: Amplifying templates with paired end sequencing adapter. Step 3: Sequencing the library on Illumina Hiseq platform. After sequencing, we compared the efficiency of different error-correction methods. Paired-end consensus was to filter out the pairs of reads that were not identical. Tag consensus was to filter out groups of reads that were with same tags but not identical. Combined consensus used both methods for filtering. The real low frequency variants are indicated as yellow dots. And the sequencing errors are indicated as pink dots.

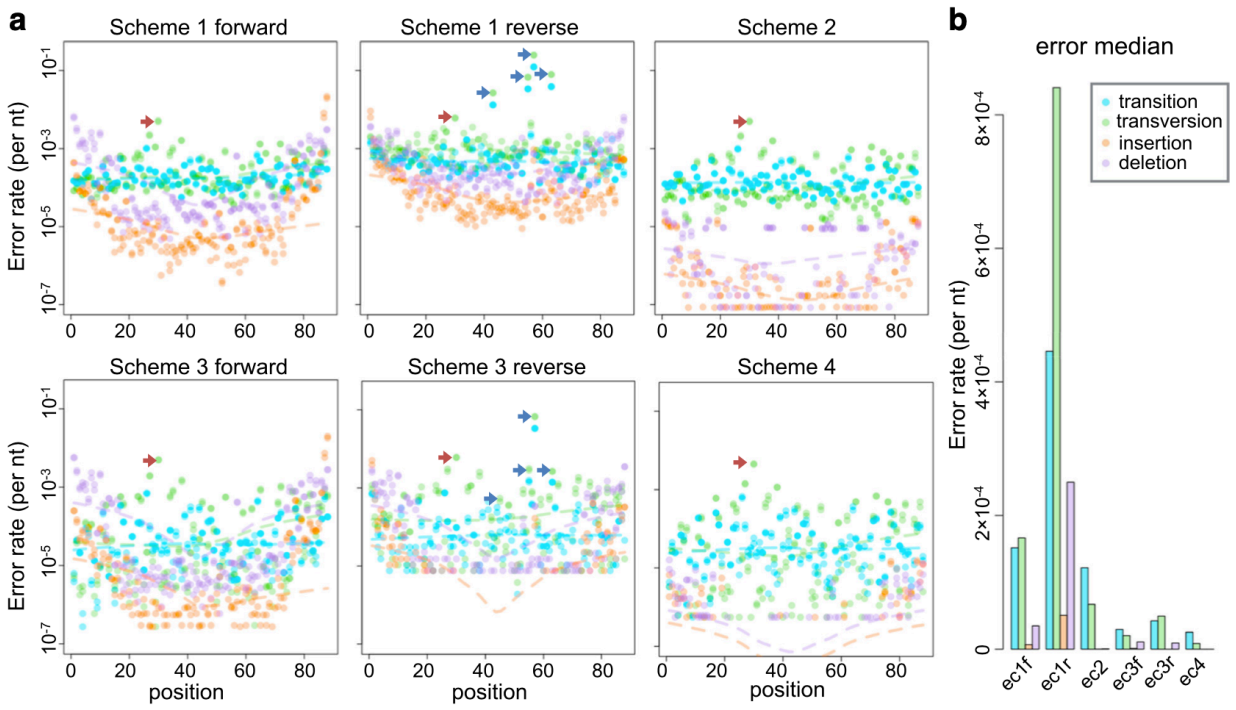


Figure 6-2. Error rates in different error-correction methods.

a Detailed profiling of error rate on every nucleotide. Every dot represents the observed error rate on a certain nucleotide. Blue, green, orange and purple represent transition, transversion, insertion and deletion, respectively. The dashed lines represent the value of local regression. Blue arrows indicate some high rate errors. Red arrows indicate a highly possible real mutation. Two technical replicates are plotted on the same subgraph. **b** Barplot of medians of different error-correction schemes. The labels, ec1f, ec1r, ec2, ec3f, ec3r, and ec4 represent Scheme 1 forward reads, Scheme 1 reverse reads, Scheme 2, Scheme 3 forward reads, Scheme 3 reverse read, Scheme 4, respectively.

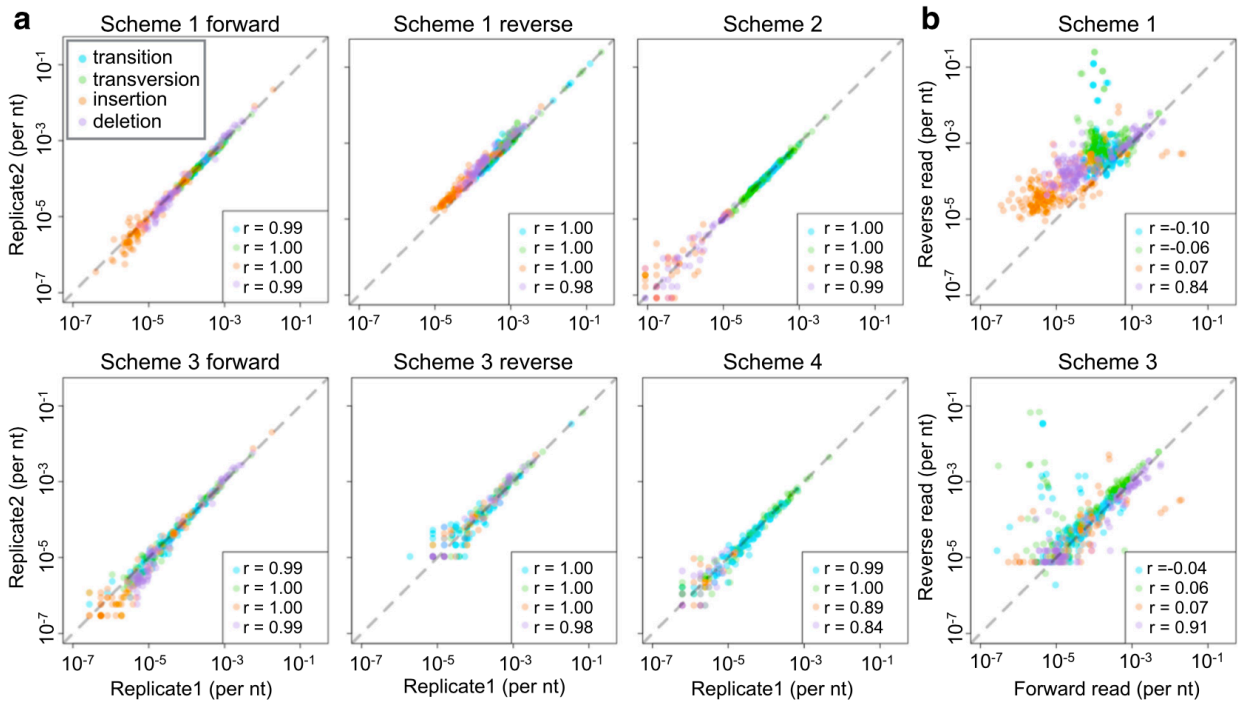


Figure 6-3. Error reproducibility.

a The error rate correlation between two technical replicates. Every dot represents a certain position on the target templates. Values on x-axis and y-axis represent error rate at replicate 1 and replicate 2 respectively. **b** The error rate correlation between forward and reverse reads. Every dot represents a certain position on the target templates. Values on x-axis and y-axis represent error rate at forward reads and reverse reads respectively. r is Pearson's correlation coefficient. The dashed lines are references of complete reproducibility.

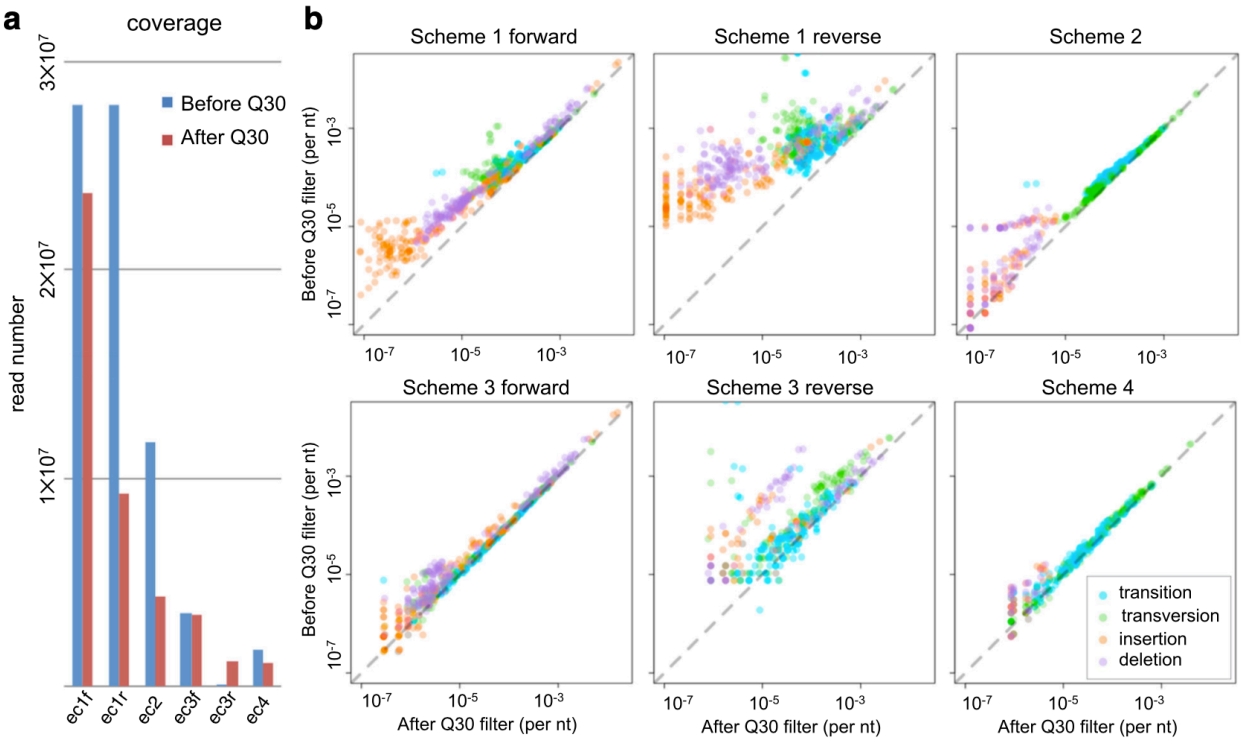


Figure 6-4. The effect of quality score and coverage.

a Barplot of coverage in different error-correction schemes, before and after quality score filtering. The labels, ec1f, ec1r, ec2, ec3f, ec3r, and ec4 represent Scheme 1 forward reads, Scheme 1 reverse reads, Scheme 2, Scheme 3 forward reads, Scheme 3 reverse read, Scheme 4, respectively. **b** The errors rate correlation between original data and quality score filtered data. The dashed lines represent complete identical error rates before and after quality score filtering.

a

Forward primer

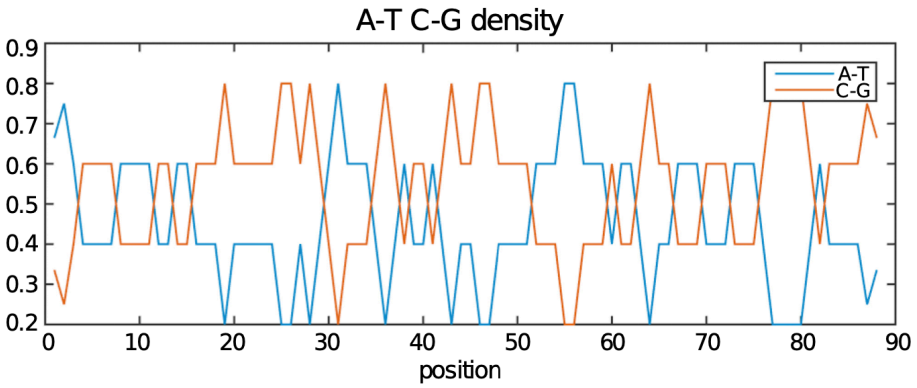
5′-AGTACGCTAACGACAACGG-3′

AGTACGCTAACGACAACGGTGTCGACGGTGAATGGACCTACGACGACGCTACCAAAACCTTCACGGTTACCGAATCCGGAGGATCCGA

5′-GGCTTAGGCCTCCTAGGCT-3′

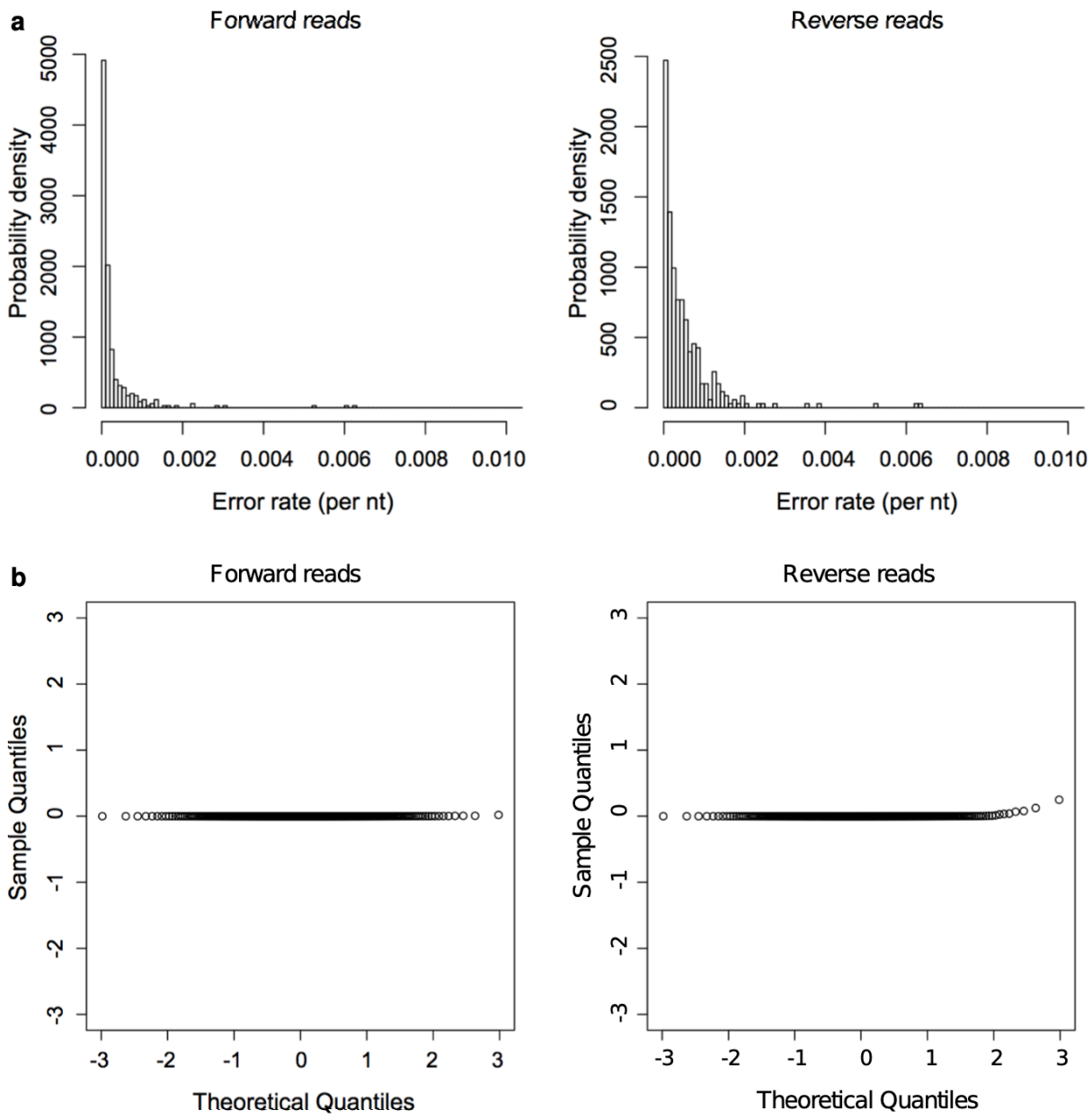
Reverse primer

b



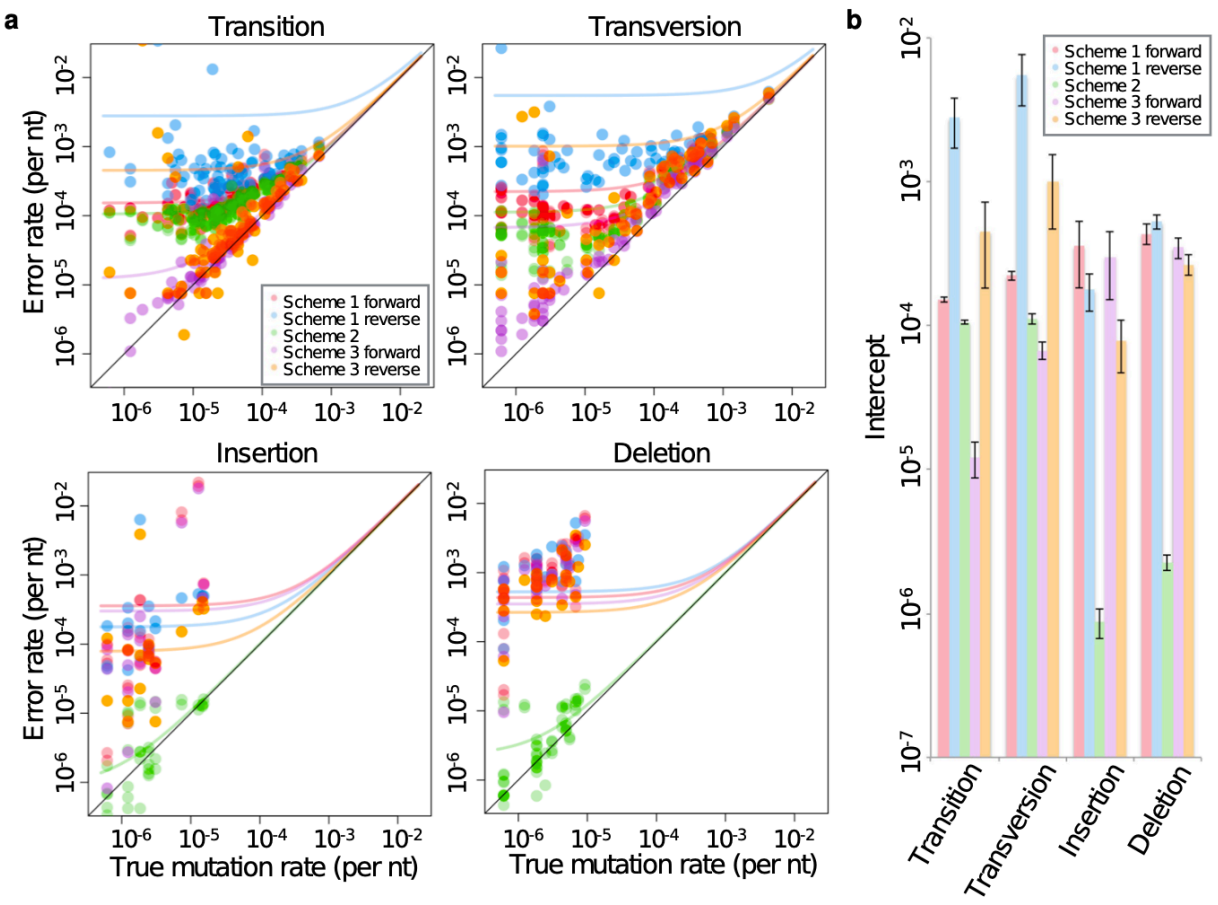
Supplementary Figure 6-1. Sequence properties of protein G.

(a) The sequence of 88 bp template was shown in DRuMS color schemes. The overlapping region of target sequence and forward primer or reverse primer was shown. (b) The A-T C-G density plot along the target sequence. Matlab nucleotide sequence analysis toolbox was used to plot this figure.



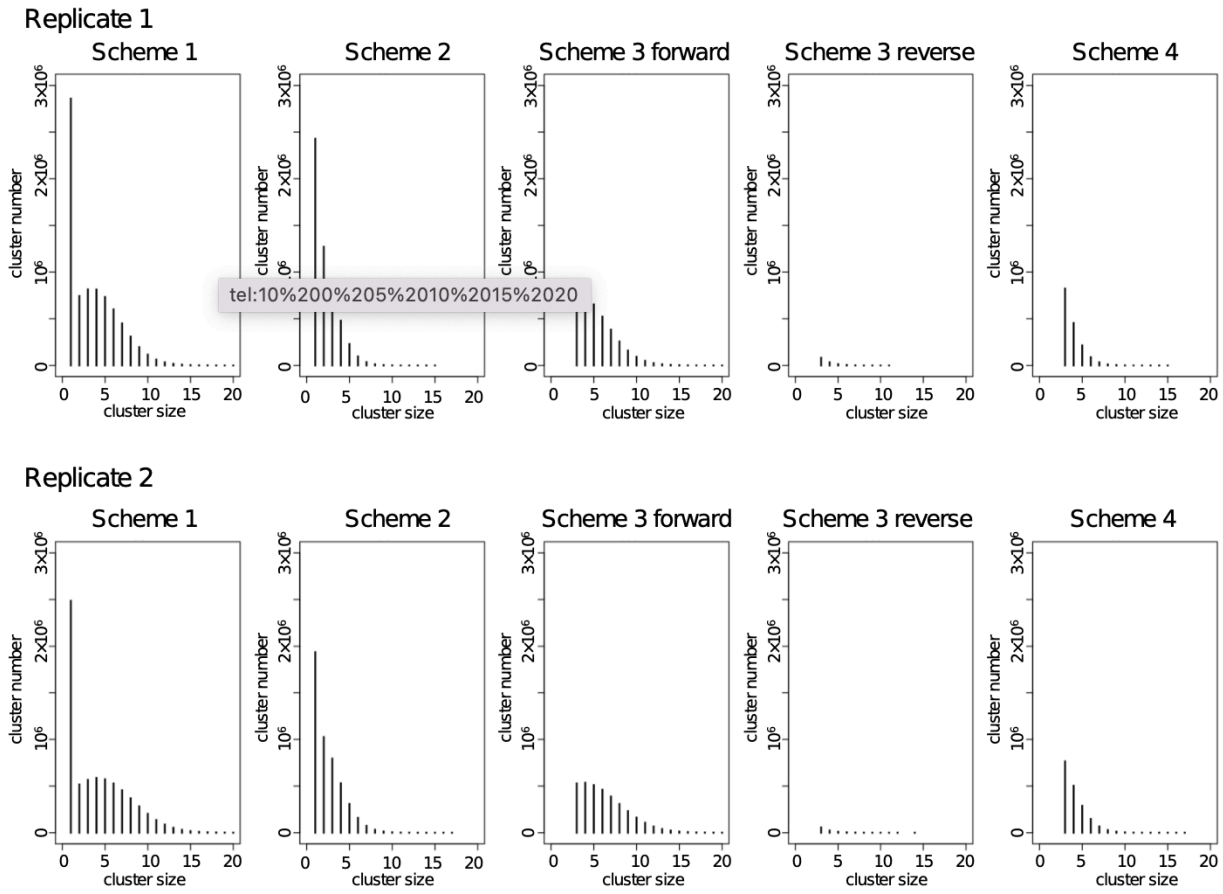
Supplementary Figure 6-2. Error rates distribution in the original dataset.

(a) The histogram of error rates. The error rates of four types of errors on all nucleotides were counted. (b) Normal Q-Q plot of error rate distribution. Sample quantiles showed great deviation from normal distribution.



Supplementary Figure 6-3. Error rate correlation among different error-correction schemes.

(a) Linear regression between true mutations and different error-correction methods. The model y~x+a was adapted to do regression. Every dot represents a position on the target sequence and the values on x-axis and y-axis represent error rates of combined consensus and certain consensus, respectively. Colored lines are regression result. (b) Barplot of the intercepts a from the linear regression. Error bar is standard error. The colors represents different error-correction schemes, which are labeled in the graph.



Supplementary Figure 6-4. Tag distribution in different error-correction schemes.

Tags are random nucleotides for readout consensus, comprising 8 nucleotides from each direction of reads. Every bar represents the number of tags that appeared certain times. Scheme 1 means the tag distribution in the original dataset.

6. References

1. Bloom JD. An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol Biol Evol. 2014;31:1956–78.

2. Thyagarajan B, Bloom JD. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. elife. 2014;3:e03300.

3. Qi H, Olson CA, Wu NC, Ke R, Loverdo C, Chu V, et al. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis c viral fitness and drug sensitivity. PLoS Pathog. 2014;10: e1004064.

4. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, et al. High-resolution mapping of protein sequence-function relationships. Nat Methods. 2010;7:741–6.

5. Pan L, Shah AN, Phelps IG, Doherty D, Johnson EA, Moens CB. Rapid identification and recovery of enu-induced mutations with next-generation sequencing and paired-end low-error analysis. BMC Genomics. 2015;16:1263.

6. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. 2011;108:9530–5.

7. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci USA. 2012;109:14508–13.

8. Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, Qi H, et al. High-throughput profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. Sci Rep. 2014;4:4942.

9. Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, Qi H, et al. High- throughput identification of loss-of-function mutations for anti-interferon activity in the influenza a virus ns segment. J Virol. 2014;88:10157–64.

10. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. Accurate sampling and deep sequencing of the hiv-1 protease gene using a primer id. Proc Natl Acad Sci U S A. 2011;108:20166–71.

11. Gout JF, Thomas WK, Smith Z, Okamoto K, Lynch M. Large-scale detection of in vivo transcription errors. Proc Natl Acad Sci USA. 2013;110: 18584–9.

12. Brodin J, Hedskog C, Heddini A, Benard E, Neher RA, Mild M, et al. Challenges with using primer IDs to improve accuracy of next generation sequencing. PloS One. 2015;10:e0119123.

13. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, et al. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proc Natl Acad Sci U S A. 2013;110: 19872–7.

14. Narayan A, Carriero NJ, Gettinger SN, Kluytenaar J, Kozak KR, Yock TI, et al. Ultrasensitive measurement of hotspot mutations in tumor dna in blood using error-suppressed multiplexed deep sequencing. Cancer Res. 2012;72:3492–8.

15. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013;493:45–50.

16. Hadd AG, Houghton J, Choudhary A, Sah S, Chen L, Marko AC, et al. Targeted, high-depth, next-generation sequencing of cancer genes in formalin-fixed, paraffin-embedded and fine-needle aspiration tumor specimens. J Mol Diagn. 2013;15:234–47.

17. Beadling C, Neff TL, Heinrich MC, Rhodes K, Thornton M, Leamon J, et al. Combining highly multiplexed pcr with semiconductor-based sequencing for rapid cancer genotyping. J Mol Diagn. 2013;15:171–6.

18. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860–921.

19. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. Genome Biol. 2011;12:R112.

20. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative dna damage during sample preparation. Nucleic Acids Research. 2013;41:e67.

21. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y,

et al. Sequence-specific error profile of illumina sequencers. Nucleic Acids Research. 2013;39:e90.

22. Rosen MJ, Davison M, Bhaya D, Fisher DS. Microbial diversity. fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science (New York, NY). 2015;348: 1019–23.

23. Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, et al. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. Cell systems. 2015;1:72–87.

24. Zhou S, Jones C, Mieczkowski P, Swanstrom R. Primer ID validates template sampling depth and greatly reduces the error rate of Next-Generation sequencing of HIV-1 genomic RNA populations. J Virol. 2015;89:8540–55.

25. Mohiyuddin M, Mu JC, Li J, Asadi NB, Gerstein MB, Abyzov A, et al. Metasv: an accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics. 2015;31:2741–4.

26. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. Snver: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res. 2011;39:e132.

27. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. Nat Methods. 2013;10:57–9.

28. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the illumina hiseq and miseq platforms. ISME J. 2012;6:1621–4.

29. De Leeneer K, Hellemans J, De Schrijver J, Baetens M, Poppe B,

Van Criekinge W, et al. Massive parallel amplicon sequencing of the breast cancer genes brca1 and brca2: opportunities, challenges, and limitations. Hum Mutat. 2011;32:335–44.

30. Forshew T, Murtaza M, Parkinson C, Gale D, Tsui DW, Kaper F, et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Sci Translational Med. 2012;4: 136ra68.

31. Consortium HMP. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.

32. Tonge DP, Pashley CH, Gant TW. Amplicon-based metagenomic analysis of mixed fungal samples using proton release amplicon sequencing. PloS One. 2014;9:e93849.

33. de Boer P, Caspers M, Sanders J, Kemperman R, Wijman J, Lommerse G, et al. Amplicon sequencing for the quantification of spoilage microbiota in complex foods including bacterial spores. Microbiome. 2015;3:30.

34. Olson CA, Wu NC, Sun R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Curr Biol. 2014;24:2643–51.

35. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics (Oxford England). 2009;25: 1422–3.

Chapter 7

A mechanistic model for virus dynamics in high-throughput viral fitness profiling

1. Abstract

High-throughput viral fitness profiling is an accurate way to measure the replication capacity of a swarm of viral strains. It uses high-throughput sequencing to quantify virus frequency after bulk competition. Its data provide insights on genetic interactions (epistasis) and fitness landscapes. Here we described a mechanistic model of virus dynamics in these experiments. We examined the robustness of screening readout (relative fitness) in different time point and different populations. We found competition allowed more accurate quantification of virus replication capacity than individual growth did. The model also suggests epistasis of relative fitness level could rise without interactions on phenotypic level, explains previous observations of predominance of non-specific epistasis.

2. Introduction

High-throughput fitness profiling (HFP), also known as deep mutational scanning (DMS), is a powerful tool in generating comprehensive genotypic-phenotypic data. It can discover mutants with novel phenotype[1-3], characterize the interactions between mutations[4], and reveal the properties of fitness landscape[5]. It not only explores mechanisms of protein evolution[6], but also helps explain virus evolution[7,8] and make novel vaccines[9,10]. HFP couples a screening process with deep sequencing to quantify the enrichment of species. The relative enrichment can be transformed into the fitness of interest. For virus, the screening is naturally the process of virus growth. Virus fitness is correlated with the relative enrichment of virus frequency. However, virus growth is in a host-pathogen system, where resource is limited. The virus frequency measured at the end of screening does not strictly correlate to higher replication capacity. To quantitatively study the

relationship of virus replication capacity and its frequency dynamics during screening, we introduced a Beretta-Kuang model for bacteria-phage population dynamics to the HFP system[11].

Fitness landscape is an abstraction of linking genotypes to reproductive success[12]. HFP provides unprecedented datasets to characterize fitness landscape[13]. However, HFP only measures relative frequencies of variants. The relationship between relative frequencies and fitness needs to be examined. The distribution of epistasis determines the shape and ruggedness of fitness landscape, which restricts evolutionary pathways and adaptation potentials[14]. Negative and non-specific epistasis are frequently observed in many HFP datasets[5,6,15,16]. It is speculated the threshold effect of mutations destabilizing proteins contributes to the predominance of negative epistasis[17-19]. In this paper, we show that both positive epistasis and negative epistasis can rise through viral dynamics, even without any genotypic or phenotypic level interactions.

3. Result

3.1. Virus replication capacity can be estimated from virus frequency in high-throughput fitness profiling.

The model of virus population dynamics was derived from Nowak and May's classic model[20] and adapted Beretta and Kuang's modification on *in vitro* host cell replication rate[11]. We introduced K strains of viruses into the model and rewrite the model in the form of equation 1-3 (Figure 1A).

$$\frac{\mathrm{d}T}{\mathrm{d}t} = r\left(1 - \frac{(T + \sum_k I_k)}{C}\right)T - \sum_k i_k V_k T$$

$$\frac{\mathrm{d}I_k}{\mathrm{d}t} = i_k V_k T - d_k I_k$$

$$\frac{\mathrm{d}V_k}{\mathrm{d}t} = p_k I_k - s_k V_k - i_k V_k T$$

T denotes the density of uninfected cells. I_k denotes the density of cells that are infected by the k-th strain. V_k denotes the density of k-th strain virus. The parameters are explained in Figure 1A. i_k represents the infectivity of virus. $i_k V_k T$ is rate of healthy cells being infected by k-th strain virus. r is the replicate rate of healthy cells. $\left(1 - \left((T + \sum_k I_k)/C\right)\right)$ is the density correction factor for cell replication rate. d_k represents the toxicity of virus. $d_k I_k$ is the rate of cell death which has been infected by k-th virus. p_k describes the productivity of virus. $p_k I_k$ is the rate of virus production by cells infected by k-th strain. s_k represents virus stability. Virus degradation rate is $s_k V_k$.

The model has following prerequisites that may be different from actual scenario. Firstly, viruses don't infect the same cell. This is common during early phase of infection when virus concentration is low. It is also true for some viruses which can down-regulate its receptor after infection, e.g. HIV-1[21]. Secondly, viruses release from infected cell with a constant rate. This approximation is reasonable for lytic viruses which has short latent phase. Thirdly, cells replicate in a Logistic manner. It resembles most of cell culture conditions as long as cell debris and toxic metabolites do not accumulate in culturing media.

We initiated simulation using parameters from previous HIV-1 dynamic models[22]. We randomized virus infectivity i_k, productivity p_k, toxicity d_k and stability s_k around wild-type virus using following distribution[23] (Figure S1A):

$$P(s) = \lambda e^{-\lambda|s|}(1 + e^{2Ns})^{-1}$$

N were tuned so that 10.1% to 13.6% of the four parameters are more beneficial than those of wild-type. The amount of input virus is normally distributed (Figure S1A). We used 1000 virus strains as input. This is the approximate library size of many HFP project[3,7,9]. We simulated 7 days of infection, which is similar time scale of HIV-1 HFP experiments.

The simulation result was shown in Figure 1B. Viral frequencies showed different distribution from input parameters (Figure S1B). The percentage of mutants that are more abundant than wild-type dropped from 49.1% to 1.1% after 7 days of screening, which indicates most strains had lower replication capacity than wild-type virus. To quantitatively describe the replication capacity, we use following equation to calculate relative fitness ($RF_{k,t}$) of viral strain k at timepoint t.

$$RF_{k,t} = \frac{f_{k,t}/f_{k,t_0}}{f_{k_0,t}/f_{k_0,t_0}}$$

The $f_{k,t}$ represents frequency of viral strain k at timepoint t. k_0 represents strain wild-type. The distribution of RF resembles the distribution of frequencies (Figure S1B). RF at the final time point (7days post infection) did not completely correlate with model parameters s_k (stability), d_k (toxicity), p_k (productivity), or i_k (infectivity) because these parameters only depicted part of viral life cycle (Figure S1C). However, RF correlated strongly with the effective replication number $R0$ (Figure 1C). $R0$ is defined below:

$$R0_k = \frac{(p_k/d_k - 1)i_k C}{s_k}$$

We then calculated RF at every time point (Figure S2A). The correlation of RF with $R0$ increased at the beginning of screening and plateaued at 2 days post infection (Figure 1C). Notably, RF's correlation coefficient increased faster than that of frequencies. The

absolute value of correlation coefficients between RF and other model parameters also plateaued fast (Figure S1D). These indicated RF was a more sensitive method to characterize virus replication capacity, comparing to virus frequencies.

Selection coefficient(s) is another commonly used parameter in experimental evolution to characterize the capacity of reproduction and competition. s is usually defined as the derivatives of relative frequencies. s in different experimental systems can be compared directly because it is independent of population density[24]. We calculated the dynamics of s in our simulation (Figure S2B). $s_{k,t}$ of viral strain k at time point t is defined as following:

$$s_{k,t} = \frac{\mathrm{d}\log f_k}{\mathrm{d}t} - \frac{\mathrm{d}\log f_{k_0}}{\mathrm{d}t}$$

s and RF correlated significantly at early time of screening but then correlation coefficient decreased (Figure S2C). This is because the overall intensity of competition decreased at the late time of screening when available uninfected cells were few. When the whole population is collapsing, virus selection coefficient no longer served as an accurate parameter to characterize virus replication capacity. However, RF was still robust at late time point. Moreover, the distribution of RF widened at late time point (Figure S2D), which ensured more confident measurement during actual experiments. The distribution of s is more restricted at the end of screening. To validate this, we constructed a mutagenesis HIV-1 library, which contained 1755 Gag mutants. We passaged the virus library in THP1 cells for 7 days and deep-sequenced the population every two days. We calculated RF and s for each mutant at day 3, day5 and day 7 (Figure S2E). RF and s were correlated for many mutants. However, in later time points, some mutants' s returned to 0, while RF are still widely distributed.

3.2. The robustness of relative fitness is restricted by population structure and experiment setups.

RF only correlated with $R0$ at some situations. We tuned model parameters and used different sets of input to simulate screening experiments. We changed the scale of virus infectivity(i_k) and productivity(p_k). The scale is defined as reciprocal of λ in equation 4. We simulated 2500 combinations of infectivity's scale and productivity's scale (Figure 2A). RF correlated with $R0$ strongly independent of the scale of infectivity. However, it is more vulnerable to the scale of productivity. The limitation of RF indicates that viral fitness profiling experiment is not suitable for libraries that accumulated too many mutations that fitness effect is extreme.

We compared the traditional way of measuring viral replication capacity with high-throughput fitness profiling. We used same dataset of viral replication parameters and simulated viral growth independently using Nowak and May's model (Figure 2B). All simulations started with exactly same number of infected cells. We calculated the correlation of viral load and replication parameters. The correlation coefficients were high only when most viral mutants were growing at log phase, which is difficult to capture accurately during actual experiments. And different mutants may have different length of log phase. The correlation between viral load and replication capacity ($R0$) dropped at later time points. We also simulated binary competition experiments of every viral strain with wild-type virus. Virus frequency correlated with $R0$ at most stages of virus replication. This indicates pairwise competition is an accurate and robust way to measure virus replication capacity.

To validate this point, we inserted fluorescent proteins into HIV-1 genome. We used mVenus and mCherry as two fluoresent marker and constructed capsid mutation N74D on them. N74D mutants induced higher interferon response than wild-type virus in monocytes and may result in less replication (Unpublished data). By mixing wild-type virus and mutants, we are able to trace the frequency of infected cells by different virus for 7 days (Figure S3). Two viruses peaked at the same time and N74D mutant was weaker than wild-type virus no matter which fluorescent protein it is linked to.

HFP provides a synchronization effects of virus replication dynamics. Different strains with different replication capacities peaked at the same time during infection (Figure 1B). This greatly helped accurate measurement of a biologically relevant parameter, relative fitness. Traditional measurement with independent viral culture cannot characterize virus replication with such resolution and confidence. HFP also has a normalization effect. Randomness in the amount of virus input was cancelled out during competition (Figure S1C, S1D). This greatly decreased labors required in fitness quantification.

3.3 Non-specific epistasis can be explained by the viral dynamic model.

Many theories described the advantages and disadvantages of genetic interactions on viral evolution. Beneficial mutations can function synergistically so that the fitness effect of double mutations is larger the multiplicativity of fitness effect by two single mutations. Deleterious mutations can function antagonizingly to relieve the fitness effect. Both types of interactions are defined positive epistasis. On the other hand, if double mutation results in lower fitness effect than the multiplicativity of single mutations does, the genetic interaction between these mutations are defined negative epistasis. Positive epistasis can help virus gain new mutations and accelerate its evolution. While negative epistasis is

stabilize the wild-type genotype by removing deleterious mutations. Both positive and negative epistasis is observed virus populations. However, the type of epistasis may differ in different systems and different ways of measuring viral fitness[18,26-28]. The multiplicablity or addibility of measured viral fitness needs to be proved before any conclusions on epistasis is drawn.

Predominance of negative epistasis can be explained by the threshold effects of protein stability. A deleterious mutation may destabilize the protein while not destroying the protein while the summation of two deleterious mutation may completely dissemble the protein and diminish its function. The concave function of protein stability (physical property) to protein function (biological property) explains the prevalence of negative epistasis on protein function level. Similarly, the concaveness of $R0$ to RF function affects the type of epistasis on RF level.

We simulated the dynamics of a viral population with uniformly distributed infectivity, and calculated RF at the end of infection (Figure S4A). $R0$ was linearly correlated with infectivity (i_k) while RF was a convex function of infectivity. This also resulted in a convex projection from $R0$ to RF. The projection remains nonlinear when we changed productivity (p_k, Figure S4B), toxicity (t_k, Figure S4C) and stability (s_k, Figure S4D). The shape is convex for infectivity and productivity and concave for toxicity. This indicated positive epistasis when the variation of infectivity and productivity predominated and negative epistasis when the variation of toxicity predominated.

To furthur prove this hypothesis, we simulated the population dynamics with all replication parameters randomized. We generated two single mutants' libraries, each with all four parameters randomly sampled. The size of single mutants' library was 30. We then

combined two libraries and generated double mutants with all possible combinations of four parameters. The size of double mutants library was $30 \times 30 = 900$. The replication parameters were simply products of those from single mutants. With different randomization trials, we find the function of $R0$ to RF could either be concave or convex (Figure 3A, B). We then calculated expected $R0$ and RF simply by multiplying $R0$ and RF of single mutants. Expected $R0$ was linearly correlated with $R0$ while expected RF was not. If doubled mutants had measured RF larger than expected RF, they were defined positive epistasis in HFP experiments (Figure 3A) and vice versa (Figure 3B). Both conditions could happen when there were no $R0$ epistasis.

We profiled an Influenza library to validate the point. We constructed a library with 59 single mutants and 736 double mutants and passaged the virus in A549 cell with or without interferon. Interferon inhibit influenza replication by decreasing its productivity (p_k) and toxicity(d_k). We calculated the expected RF of all double mutants by multiplexing RF of corresponding single mutants. The same library showed predominant negative epistasis without interferon treatment but predominant positive epistasis with interferon (Figure 3C).

4.Discussion

The non-linearity among genotypes and fitness has long been centric to fitness landscape theory. Besides the frequently discussed aspects of genotype-phenotype interactions and physical-biological property interactions, we argued that the nonlinear relationship between phenotype to fitness is also important to shape fitness landscape. This relationship could be furthur investigated in diploid populations and haploid populations

with Logistic population dynamics models. The fitness landscape in protein evolution could also be reviewed in modeling K_d-frequency relationship.

HFP provides unique datasets of mutants' fitness. However, fitness is a simplification of replication capacity. Many other factors will affect the relationship between phenotype and evolution outcomes. Extra cares are needed in adapting HFP datasets to evolutionary analysis. This paper not only used simulated datasets but also carried out a few HFP experiments. However, more profiling methods are needed to accurately and massively quantify different aspects of virus replication capacity.

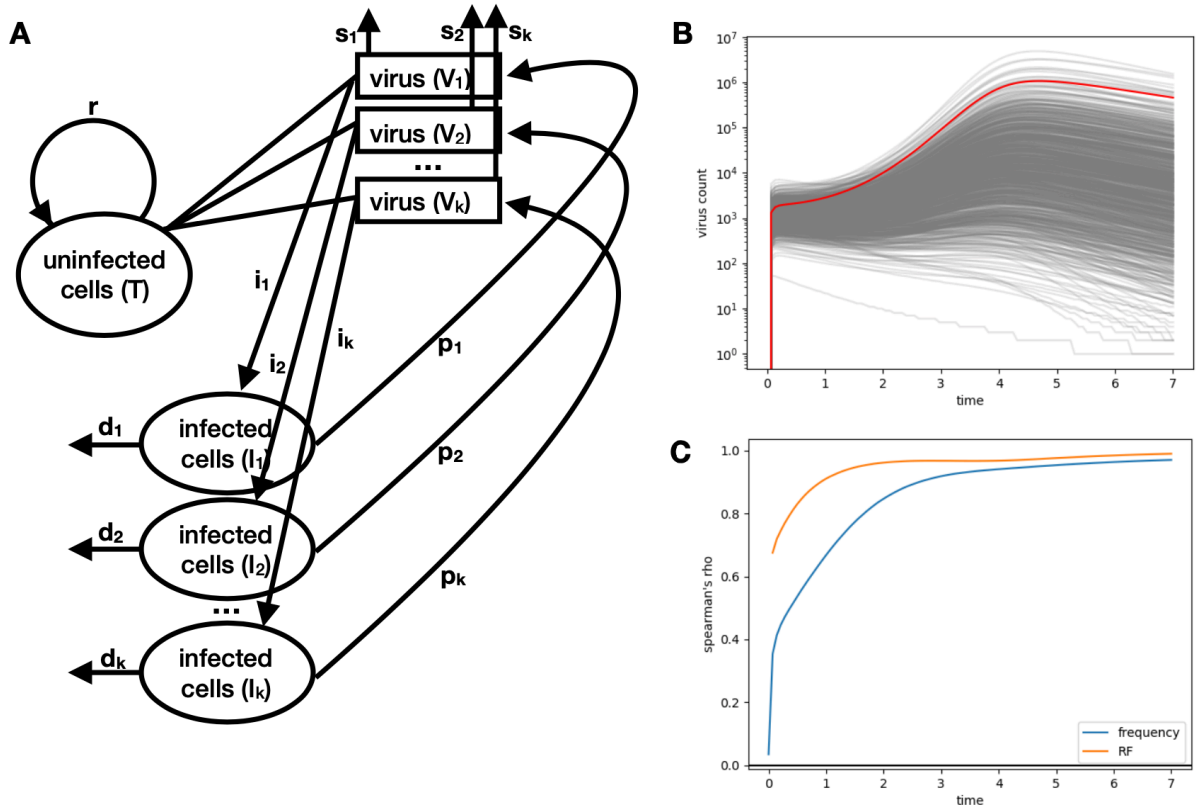


Figure 7-1. Characterization of high-throughput fitness profiling viral dynamics model.

A) Diagram of the variables and reactions tracked by the model. Model parameters are described in manuscript. B) Temporal dynamics of virus count in a simulation. C) The dynamics of correlation between effective replication capacity (R0) and frequency (blue line) or relative fitness (orange line).

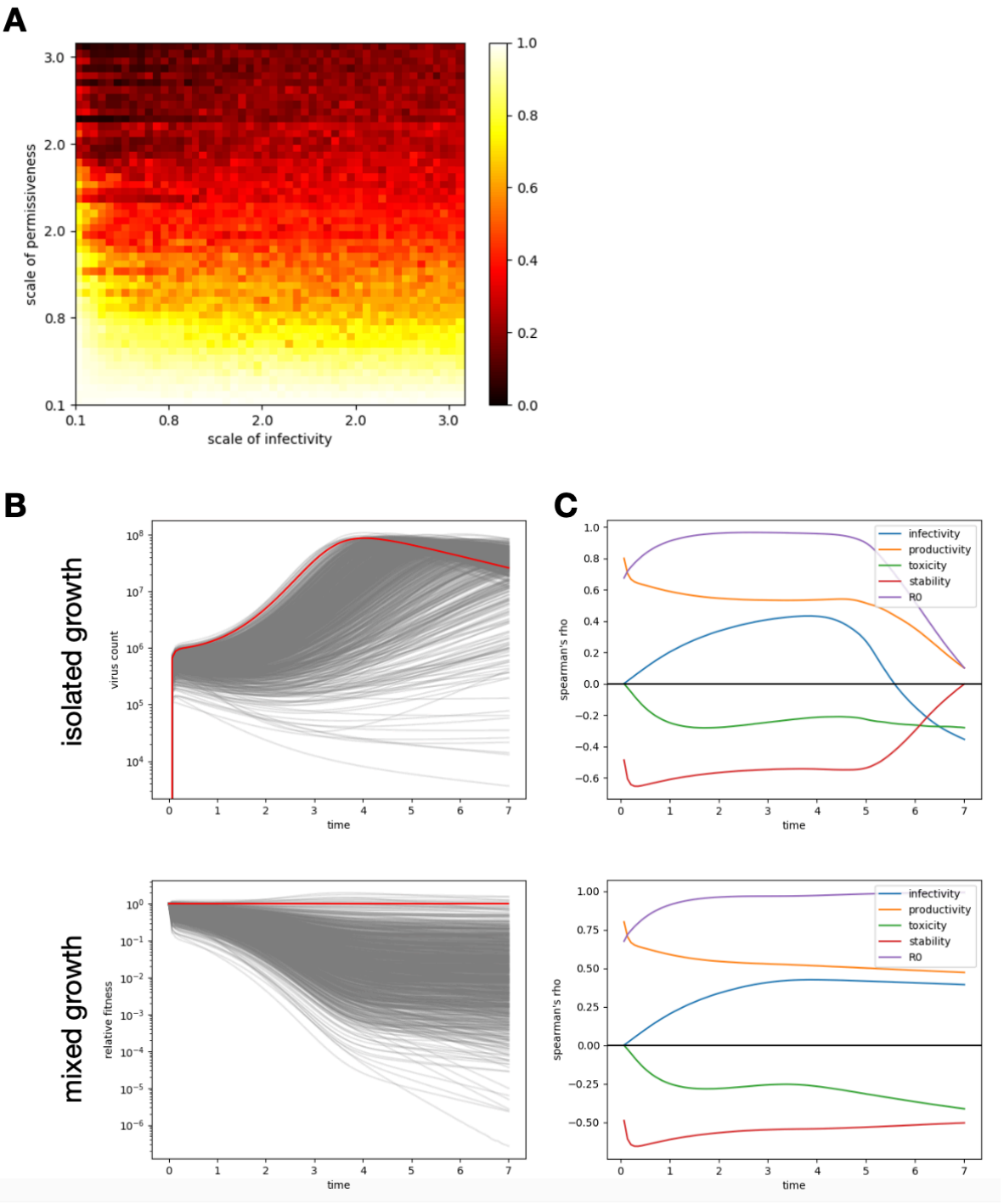

A
scale of permissiveness
scale of infectivity
B
isolated growth
virus count
time
mixed growth
relative fitness
time
C
infectivity
productivity
toxicity
stability
R0
spearman's rho
time

Figure 7-2. The effect of population structure and experiment procedures on the robustness of relative fitness.

A) Scale of replication capacity distribution affects correlation between R0 and relative fitness. B) Simulated dynamics of independent virus growth and mixed virus growth. C)The dynamics of correlation between model parameters and virus frequencies.

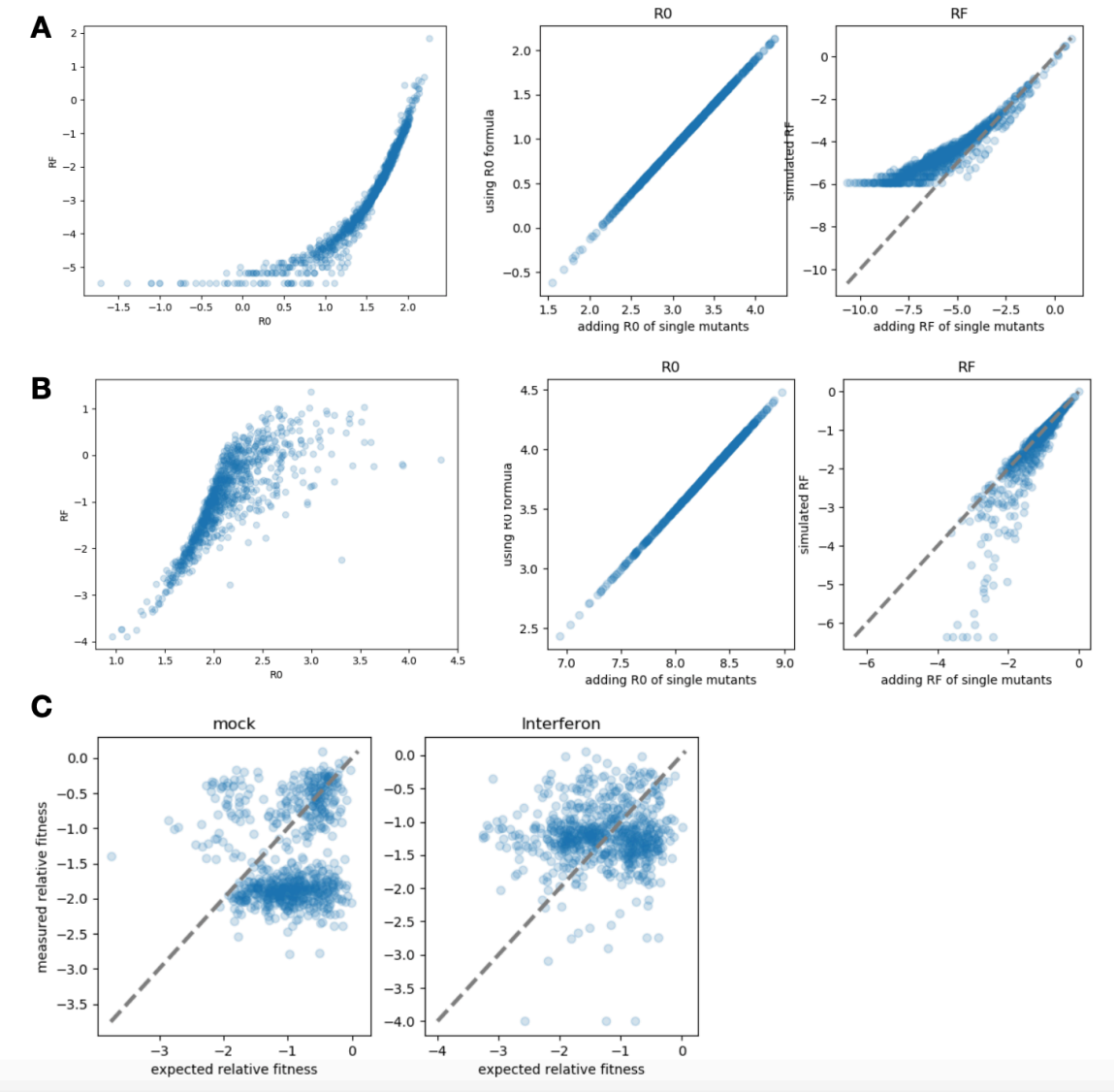
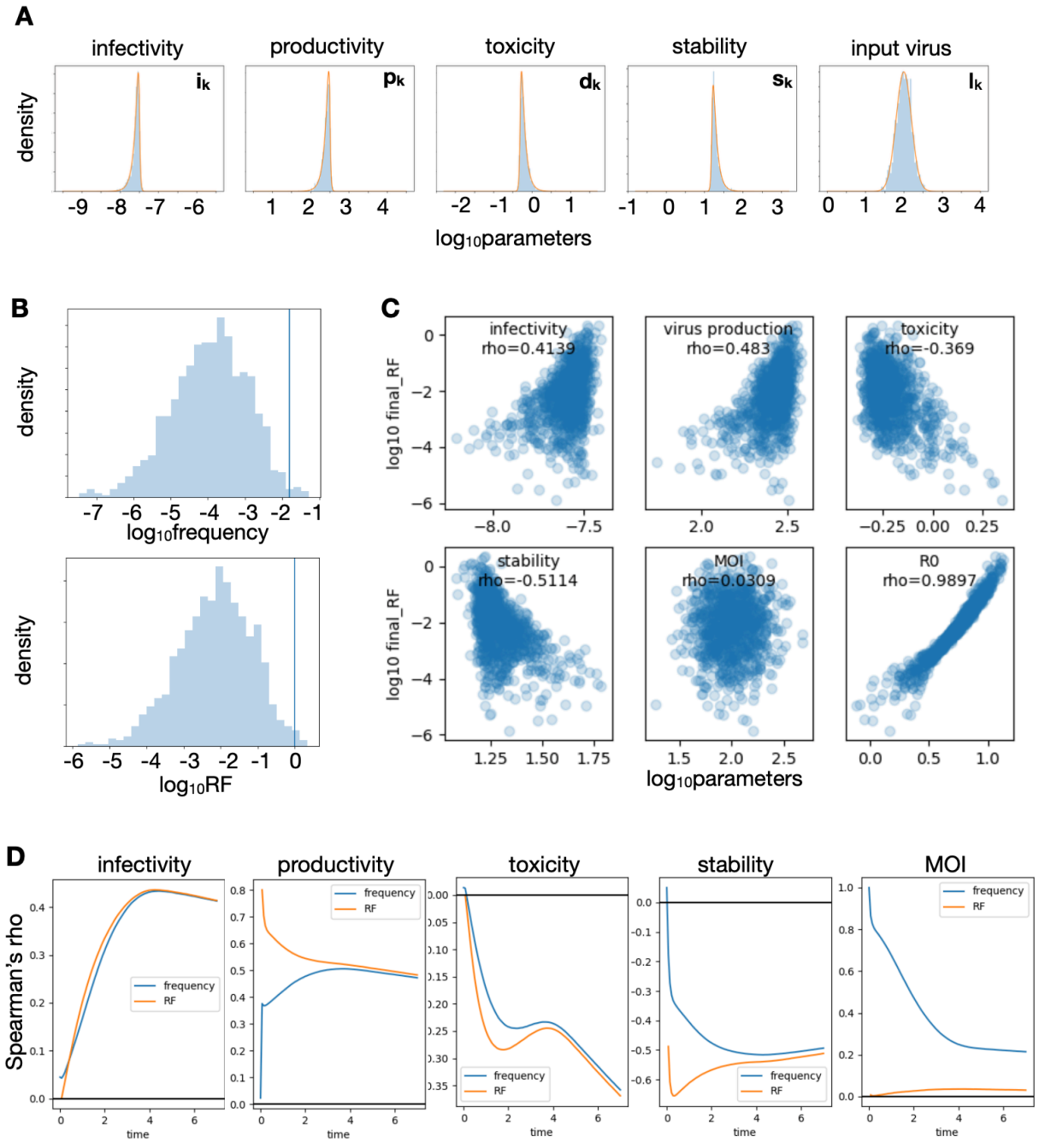


Figure 7-3. Construction of epistasis without phenotypic interactions.

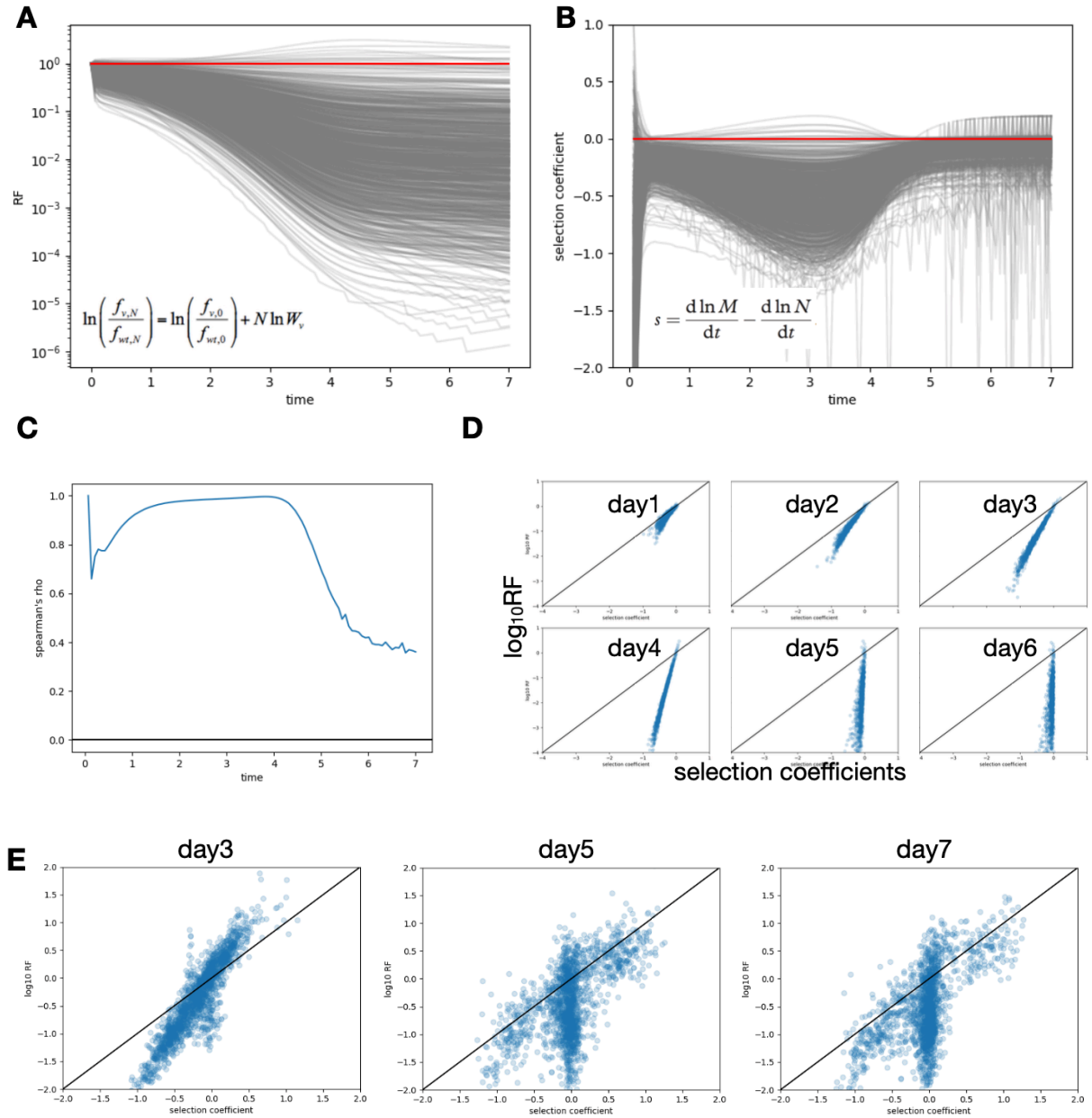
A) The projection from R0 to relative fitness, in a population with linearly composed replication parameters. Virus productivity is the parameters with widest distribution. B) The projection from R0 to relative fitness, in a population with linearly composed replication parameters. Virus toxicity is the parameters with widest distribution. C) Epistasis of an Influenza mutagenesis library under different conditions.



Supplementary Figure 7-1. Characterization of high-throughput fitness profiling viral dynamics model.

A) Distribution of model parameters. B) Frequency and relative fitness distribution at the end time of simulation. Blue line is the value of wild-type. C) Correlation of model input

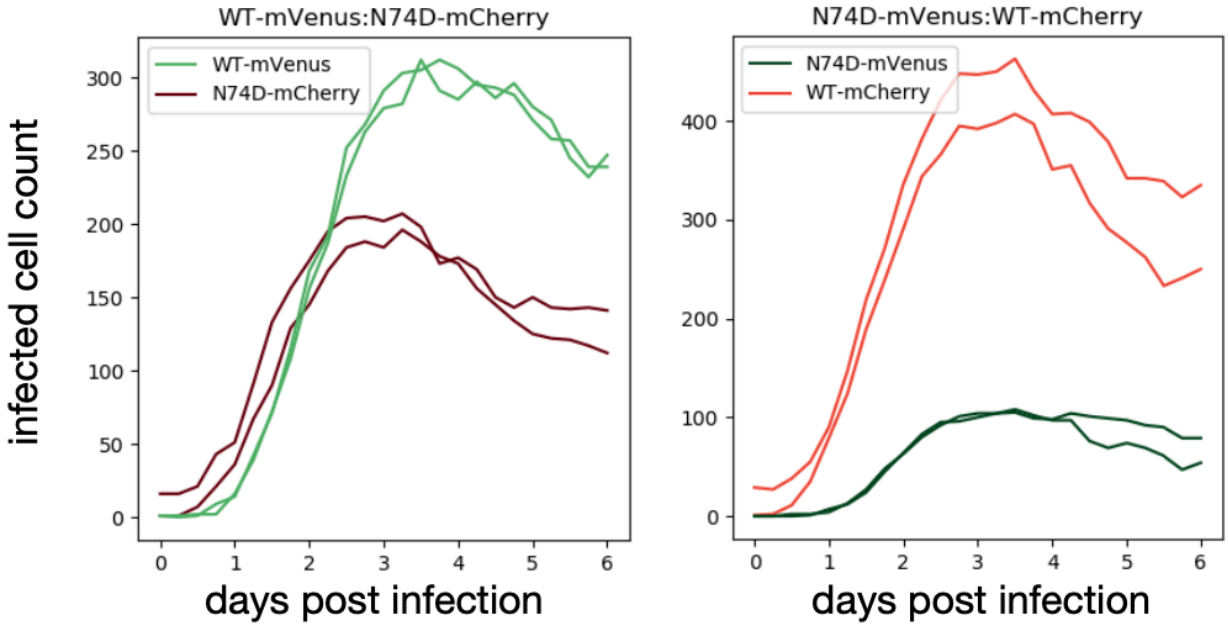
parameters and final relative fitness. R0 is effective replication capacity. D) The dynamics of correlation between model parameters and frequency (blue line) or relative fitness (orange line).



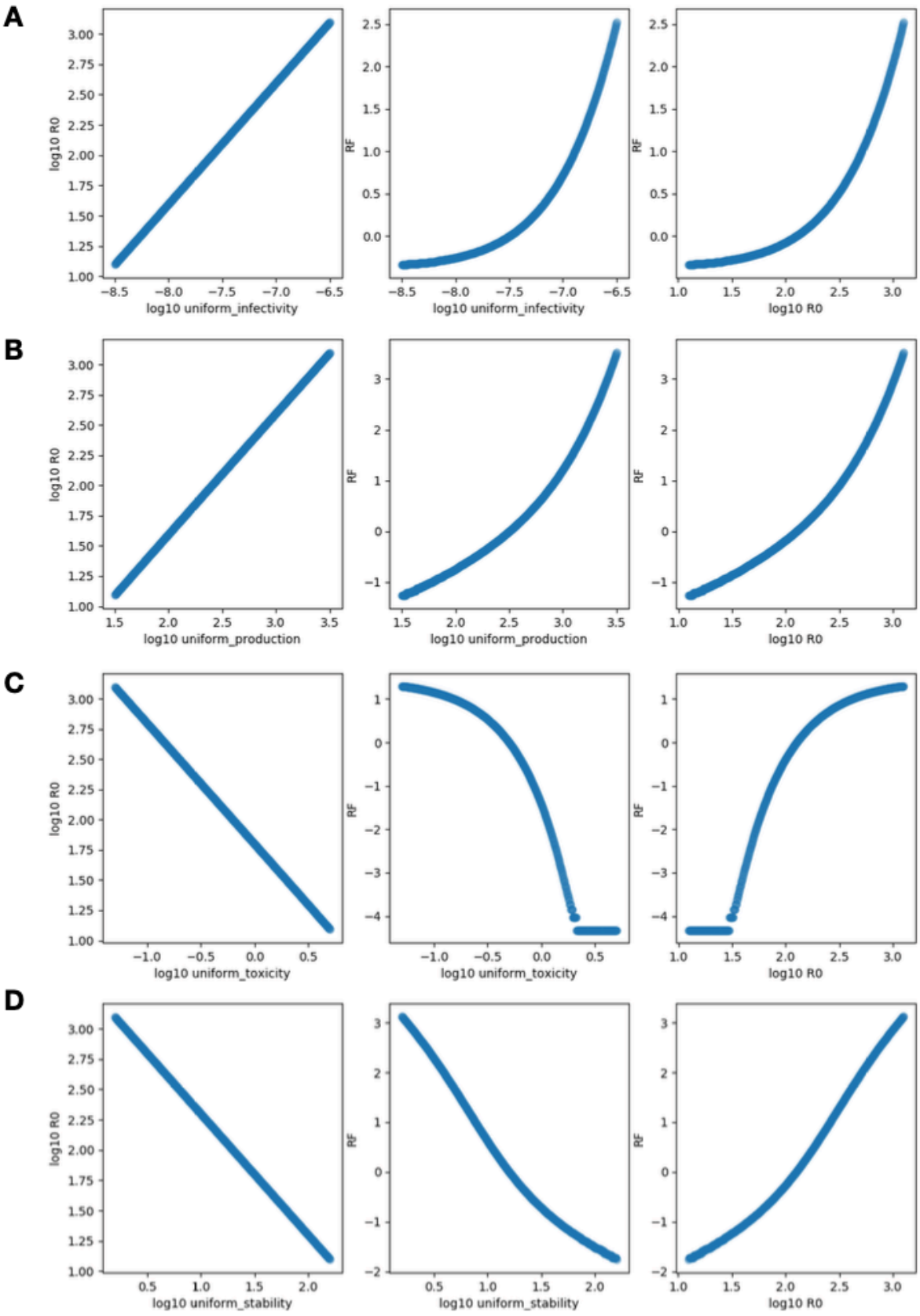
Supplementary Figure 7-2. Correlation analysis of relative fitness and selection coefficient.

A) Relative fitness dynamics in one simulation. Relative fitness was defined as the ratio of mutant frequency after and before screening. B) selection coefficient dynamics in the same simulation. Selection coefficient is defined as the derivative of relative fitness. C)

Correlation dynamics of selection coefficients and relative fitness. D) The correlation of selection coefficients and relative fitness at 6 selected time point. E) The correlation of selection coefficients and relative fitness in an experimental dataset.



Supplementary Figure 7-3. Competition of fluorescent virus.


A
log10 R0
log10 uniform_infectivity
RF
log10 uniform_infectivity
RF
log10 R0
B
log10 R0
log10 uniform_production
RF
log10 uniform_production
RF
log10 R0
C
log10 R0
log10 uniform_toxicity
RF
log10 uniform_toxicity
RF
log10 R0
D
log10 R0
log10 uniform_stability
RF
log10 uniform_stability
RF
log10 R0

Supplementary Figure 7-4. The relationship between replication capacity parameters and relative fitness.

A) The projection from infectivity to R0 and relative fitness, in a population with uniformly distributed infectivity. B) The projection from productivity to R0 and relative fitness, in a population with uniformly distributed productivity. C) The projection from toxicity to R0 and relative fitness, in a population with uniformly distributed toxicity. D) The projection from stability to R0 and relative fitness, in a population with uniformly distributed stability.

5. References

1. Qi, H. et al. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis c viral fitness and drug sensitivity. PLoS pathogens 10, e1004064 (2014).

2. Wu, N. C. et al. Functional constraint profiling of a viral protein reveals discordance of evolutionary conservation and functionality. PLoS genetics 11, e1005310 (2015).

3. Gong, D. et al. High-throughput fitness profiling of zika virus e protein reveals different roles for glycosylation during infection of mammalian and mosquito cells. iScience 1, 97–111 (2018).

4. Wu,N.C., Dai,L., Olson,C.A., LloydSmith,J.O.& Sun, R.Adaptation. in protein fitness landscapes is facilitated by indirect paths. Elife 5, e16965 (2016).

5. Sarkisyan, K. S. et al. Local fitness landscape of the green fluorescent protein. Nature 533, 397 (2016).

6. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Curr. Biol. 24, 2643–2651 (2014).

7. Wu, N. C. et al. A structural explanation for the low effectiveness of the seasonal influenza h3n2 vaccine. PLoS pathogens 13, e1006682 (2017).

8. Dingens, A. S. et al. Complete functional mapping of infection-and vaccine-elicited antibodies against the fusion peptide of hiv. PLoS pathogens 14, e1007159 (2018).

9. Du, Y. et al. Genome-wide identification of interferon-sensitive mutations enables influenza vaccine design. Science 359, 290–296 (2018).

10. Wang, L. et al. Generation of a live attenuated influenza vaccine that elicits broad protection in mice and ferrets. Cell host & microbe 21, 334–343 (2017).

11. Beretta, E. & Kuang, Y. Modeling and analysis of a marine bacteriophage infection. Math. Biosci. 149, 57–76 (1998).

12. Fragata, I., Blanckaert, A., Louro, M. A. D., Liberles, D. A. & Bank, C. Evolution in the light of fitness landscape theory. Trends ecology & evolution (2018).

13. Blanco, C., Janzen, E., Pressman, A., Saha, R. & Chen, I. A. Molecular fitness landscapes from high-coverage sequence profiling. Annu. review biophysics (2019).

14. Aguilar-Rodríguez, J., Payne, J. L. & Wagner, A. A thousand empirical adaptive landscapes and their navigability. Nat. ecology & evolution 1, 0045 (2017).

15. Jacquier, H. et al. Capturing the mutational landscape of the beta-lactamase tem-1. Proc. Natl. Acad. Sci. 110, 13067–13072 (2013).

16. Bendixsen,D.P., Østman,B. & Hayden,E.J. Negative epistasis in experimental rna fitness landscapes. J. molecular evolution 85, 159–168 (2017).

17. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. Protein Sci. 25, 1204–1218 (2016).

18. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an. influenza protein. Elife 2, e00631 (2013).

19. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. Proc. Natl. Acad. Sci. 103, 5869–5874 (2006).

20. Nowak, M., Nowak, M. A. & May, R. M. Virus dynamics: mathematical principles of immunology and virology (Oxford university press, 2000).

21. Greenberg, M., DeTulleo, L., Rapoport, I., Skowronski, J. & Kirchhausen, T. A dileucine motif in hiv-1 nef is essential for sorting into clathrin-coated pits and for downregulation of cd4. Curr. Biol. 8, 1239–S3 (1998).

22. Hill, A. L., Rosenbloom, D. I., Nowak, M. A. & Siliciano, R. F. Insight into treatment of hiv infection from viral dynamics models. Immunol. reviews 285, 9–25 (2018).

23. Rice, D. P., Good, B. H. & Desai, M. M. The evolutionarily stable distribution of fitness effects. Genetics genetics–114 (2015).

24. Metz, J. A., Nisbet, R. M. & Geritz, S. A. How should we define 'fitness' for general ecological scenarios? Trends Ecol. & Evol. 7, 198–202 (1992).

25. Phillips, P. C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat. Rev. Genet. 9, 855 (2008).

26. Bonhoeffer, S., Chappey, C., Parkin, N. T., Whitcomb, J. M. & Petropoulos, C. J. Evidence for positive epistasis in hiv-1. Science 306, 1547–1550 (2004).

27. Hinkley, T. et al. A systems analysis of mutational effects in hiv-1 protease and reverse transcriptase. Nat. genetics 43, 487 (2011).

28. Parera, M., Perez-Alvarez, N., Clotet, B. & Martínez, M. A. Epistasis among deleterious mutations in the hiv-1 protease. J. molecular biology 392, 243–250 (2009).

Chapter 8

Concluding Remarks

In this dissertation, I covered my work on virus genetic barcodes and high throughput fitness profiling systems. By inserting a genetic barcode into HIV-1, we successfully tracked the dynamics of the latent reservoir in humanized mice. It also allowed us to study how the position of HIV-1 integration sites can affect the virus transcription and T cell clonal expansion. Moreover, we used it to label single cells' behavior in vitro and achieved the first virus alternative splicing sequencing at the single cell resolution. The high throughput fitness profiling system enabled us to measure the fitness effect of a large number of mutations. The dissertation covered its application in studying the genetic interactions between drug resistance associated mutations and in tracking the evolutionary trend of viruses in the real world. We also applied this platform in many other projects, such as discovering new protein protein interactions, searching for the conservative epitopes for vaccine design, explaining the key residue for cross-species transmission. Both methods are at their infant period of development. There are numerous potential applications for them.

Future applications for the fitness profiling system

The pandemic of SARS-CoV-2 alerted people to the importance of rapid production and optimization of pharmaceutical interventions. New escape mutations keep emerging from current vaccines and test strips. The research and development teams need to iterate endlessly after the emergence of new mutants. Fitness profiling platforms can rapidly simulate the mutation and selection scenario in the real world and predict the future escape mutations. The platform has the potential to be integrated in all major production pipelines for test kit, vaccines, antiviral drugs and immunotherapies.

One obstacle for the wide application of this method is the lack of a good screening system. For SARS-CoV-2, the antibodies and antiviral drugs can be produced without a viral reverse genetic system. But the fitness profiling system is more robust with a reverse genetic system to produce desired mutations. That is the reason we designed a virus-free screening system for SARS-CoV-2 N protein. Based on flow cytometry and cell display, we set up a pipeline to screen any structural and functional viral proteins in the future.

The fitness profiling system can make more use if combined with cellular libraries. For example, the mutant virus library can be screened in a CRISPR KO cell library and the growth of each virus in each type of mutant cell can be quantified by ECCITE-seq. By multiplexing another library to the virus library, we can do thousands of fitness profiling experiments at the same time.

With the development of genome editing, future libraries can also be constructed directly on human genomes and make all possible mutations in situ. Combining a cellular fitness profiling system with a viral fitness profiling system will help us to understand host-pathogen interactions in terms of the arm race evolution between them.

Future applications of the viral genetic barcodes

The rapid advancement of long sequencing reads and single-cell sequencing methods have significantly facilitated the application of viral genetic barcodes. In vivo, these tools allow for the profiling of spontaneous mutations along with barcode sequences, enabling a comprehensive understanding of the clonal dynamics of viruses, including the emergence, fixation, and extinction of mutations. By studying the evolutionary history of each clone, observing convergent and divergent evolution, and testing hypotheses on

local competition, researchers can accurately investigate the underlying mechanisms of viral evolution. In vitro, these methods enable the sequencing of viral transcription and integration, as well as the profiling of single-cell proteomic and transcriptomic features of host cells. This provides a complete picture of how the virus's life history affects cellular activities, as well as how cells regulate or inhibit viral replication.

Despite these advances, there are still some obstacles that need to be overcome. Currently, long-read sequencing methods such as PacBio or Nanopores are limited in terms of accuracy and throughput. Thus, the cost of reconstructing mutations on a full-length viral genome remains prohibitively high for large virus populations. However, the continuous development of third-generation sequencing is expected to bring the cost down to an acceptable level in the next few years. Another challenge is the limited ability of single-cell sequencing techniques, including microwell and droplet-based methods, to enrich viral DNA and sequence the genome. However, the integration of hydrogel-based in situ PCR methods holds promise for overcoming this obstacle. I anticipate an explosion of similar method development in the next few years as the scientific community continues to push the boundaries of this exciting field.

The viral genetic barcodes system also holds great potential for integration into the antiviral therapy development pipeline. By describing how therapies interrupt the virus population and create bottlenecks, we have used this system to evaluate latency reversal agents and cell therapies. In the future, this technology may be applied to vaccine studies, gene editing therapies, and small-molecule antiviral drugs.