Virus Evolution, 2019, 5(2): vez028

doi: 10.1093/ve/vez028
Resources

DISSEQT—DIStribution-based modeling of SEQuence space Time dynamics†

R. Henningsson,[1,2,3,4] G. Moratorio,[2,5] A.V. Bordería,[3] M. Vignuzzi,[2] and M. Fontes[3,6,7,8,*]

[1]The Centre for Mathematical Sciences, Lund University, Sweden, [2]Viral Populations and Pathogenesis Unit, Institut Pasteur, Paris, France, [3]The International Group for Data Analysis, Institut Pasteur, Paris, France, [4]Division of Clinical Genetics, Lund University, Sweden, [5]Laboratorio de Virología Molecular, Universidad de la República, Montevideo, Uruguay, [6]Department of Cancer Immunology, Genentech, South San Francisco, CA, USA, [7]The Center for Genomic Medicine, Rigshospitalet, Copenhagen, Denmark and [8]Persimune, The Centre of Excellence for Personalized Medicine, Copenhagen, Denmark

*Corresponding author: E-mail: fontes.magnus@gene.com

Abstract

Rapidly evolving microbes are a challenge to model because of the volatile, complex, and dynamic nature of their populations. We developed the DISSEQT pipeline (DIStribution-based SEQuence space Time dynamics) for analyzing, visualizing, and predicting the evolution of heterogeneous biological populations in multidimensional genetic space, suited for population-based modeling of deep sequencing and high-throughput data. The pipeline is openly available on GitHub (https://github.com/rasmushenningsson/DISSEQT.jl, accessed 23 June 2019) and Synapse (https://www.synapse.org/#!Synapse: syn11425758, accessed 23 June 2019), covering the entire workflow from read alignment to visualization of results. Our pipeline is centered around robust dimension and model reduction algorithms for analysis of genotypic data with additional capabilities for including phenotypic features to explore dynamic genotype–phenotype maps. We illustrate its utility and capacity with examples from evolving RNA virus populations, which present one of the highest degrees of genetic heterogeneity within a given population found in nature. Using our pipeline, we empirically reconstruct the evolutionary trajectories of evolving populations in sequence space and genotype–phenotype fitness landscapes. We show that while sequence space is vastly multidimensional, the relevant genetic space of evolving microbial populations is of intrinsically low dimension. In addition, evolutionary trajectories of these populations can be faithfully monitored to identify the key minority genotypes contributing most to evolution. Finally, we show that empirical fitness landscapes, when reconstructed to include minority variants, can predict phenotype from genotype with high accuracy.

Key words: multidimensional scaling; quasispecies; NGS; applied mathematics.

1. Introduction

Microbial infections, by viruses and bacteria, initially colonize their host as small, quite homogeneous populations, but short generation times and relatively high mutation rates quickly lead to large populations of high genetic diversity. It is well accepted that this diversity facilitates adaptation to the host through selection of variants from this pool of mutants, in response to environmental change (Domingo, Sheldon, and Perales 2012). With the advent of DNA sequencing, viruses and bacteria were the first organisms to be fully sequenced phage

†Special Issue Santa Fe Institute Workshop on Integrating Critical Phenomena and Multi-Scale Selection in Virus Evolution, supported by the NSF Rules of Life Program grant DEB-1830688.

MS2 in 1975 (Fiers et al. 1976); Haemophilus *influenzae* (Fleischmann et al. 1995); and Mycoplasma *genitalium* (Fraser et al. 1995) and the study of microbial evolution by phylogenetics has benefited from the hundreds to tens of thousands of consensus sequence genomes available for many microorganisms. More recently, High-throughput sequencing (HTS) technologies have added new depth to sequence data, capable of quantifying minority variants within the population that differ from the consensus sequence. For example, HTS studies of RNA viruses indicate that both experimental and clinical samples present hundreds to tens of thousands of low-frequency variants, constituting single nucleotide polymorphisms at nearly every nucleotide site along the genome (Acevedo, Brodsky, and Andino 2014; Bordería et al. 2015). Even before HTS, phenotypic differences between populations with the same consensus sequence have been observed and attributed to suspected differences in variant composition (Vignuzzi et al. 2006). However, characterization of these mutant 'swarms' has generally been limited to mean measures of overall diversity (e.g. Shannon entropy, mean variance, etc.). In a few cases, examples of mixed populations of single nucleotide variations were shown to contribute significantly to virus pathogenesis (Vignuzzi et al. 2006), fitness, and phenotype (Bordería et al. 2015; Xue et al. 2016), but focused on only a few variants. Since a mixed population can constitute an evolutionary stable strategy (ESS) (Smith and Price 1973; Reiter et al. 2015), the population might aim for an equilibrium where multiple variants coexist.

The rapidly expanding field of single-cell sequencing illustrates how the role of heterogeneity in general can be studied in more and more detail. The data is however complex and noisy, which presents new challenges in the development of algorithms and techniques for analysis, representation, and visualization (Bacher and Kendziorski 2016; Gawad, Koh, and Quake 2016; Svensson et al. 2016; Perkel 2017; Russell, Trapnell, and Bloom 2017). Although phylogenetic tools are well suited for understanding the evolutionary history of lineages and the relationships between lineages/individuals based on whole-genome consensus sequence data, they cannot take into account the variant composition hidden by the consensus. Higgins (1992) circumvents these issues by applying multidimensional scaling (MDS) for exploratory analysis, keeping distances between samples more in line with the measured quantities. PhyloMap (Zhang et al. 2011) superimposes phylogenetic trees on the MDS representation, trying to get the best from both worlds. Relying on consensus sequences only, these models are, however, not well suited for comparison of populations that might be identical at the consensus level but with key differences in the minority variant composition. Other tools are thus needed to adequately represent and visualize a microbial population in sequence space, focusing on where something is, rather than how it got there. Theoretical fitness landscape models, including Wright's (Wright 1932) and the NK landscapes (Kauffman and Weinberger 1989) of Kauffman using two parameters to model the landscape ruggedness paved the way for more recent advances where landscape models are (partially) based on empirical data. One approach is to study the impact of mutations at a few loci only (Whitlock and Bourguet 2000; Collins et al. 2004), thus artificially enforcing a low dimension of sequence space. To expand the fitness landscape analysis to a higher dimensional setting, Kouyos et al. (2012) utilized predictive models for *in vitro* fitness based on the amino acid sequence. For RNA viruses, the mathematical framework provided by the quasispecies theory has been used to describe the population dynamics of these pathogens (Biebricher and Eigen 2006). Seifert et al. (2015) assumed that viral populations reached mutation-selection equilibrium and applied the quasispecies equation to infer fitness values for the haplotypes in a swarm. However, it is generally accepted that mutation-selection balance is not reached throughout most stages of infection and under most experimental conditions.

Nevertheless, it is tempting to think a proper analysis of an evolving population may foretell whether and where the population will move in genotypic space (Stapleford et al. 2014). We believe that to attempt such an analysis, in an accurate and unbiased manner, we need to consider the mutant swarm (a population of closely related viral particles) in its entirety, rather than just the consensus sequence (the most frequent residue at each position) or other oversimplifications of the complex population structure. Minority variants (variant present in the swarm, but not in the consensus sequence) should be included in the analysis and not subjected to an arbitrary cutoff frequency. Viruses, the fastest mutators with small genomes, make ideal model organisms for studying short time-scale population dynamics and will thus be used to showcase the methods developed in this work.

Here, we present DISSEQT (DIStribution-based SEQuence space Time dynamics)—a pipeline for analyzing evolution of microbial populations. At the core is the ability to accurately represent the genetic heterogeneity of a mutant swarm, by modeling the population as a distribution over sequence space, thus making it possible to describe similarities and differences between populations down to the minority level, and to couple sequence space composition to phenotypic effects. We demonstrate the DISSEQT pipeline with examples from RNA virus evolution. First, we show how the DISSEQT sequence space model can uncover biologically relevant features. Second, we followed the evolutionary trajectories of longitudinal samples of experimentally evolved viral populations. Finally, by developing a fitness landscape model based on empirical fitness measurements, we demonstrate how phenotypic effects can be predicted from the population composition. Specifically, we show that the sequence space in which viral populations evolve are of relatively low dimension, and that biologically relevant signals can be readily captured and used to identify the key variants contributing most to phenotype. We confirm that minority variants contribute significantly to phenotype and must be taken into account for accuracy of genotype–phenotype prediction.

2. Results

2.1 Overview of the DISSEQT pipeline

The DISSEQT pipeline (Fig. 1, top panel) is designed for reproducibility and openness, from the ground up, using modern software solutions. The source code is openly available in GitHub and all software dependencies are open source. The software can either be installed locally or run directly from Docker images with all required software preinstalled. Running from Docker images simplifies setup and improves reproducibility since differences between local runtime environments are eliminated.

The overview described here is detailed in Section 4. The DISSEQT pipeline has three steps, serving different purposes. (1) Establishing a model for sequence space. (2) Reducing noise to make the model robust. (3) Visualization and phenotype prediction.

First, the raw reads for each sample (mutant swarm) were aligned iteratively until the consensus sequence converged and both automatic and manual quality controls were performed.

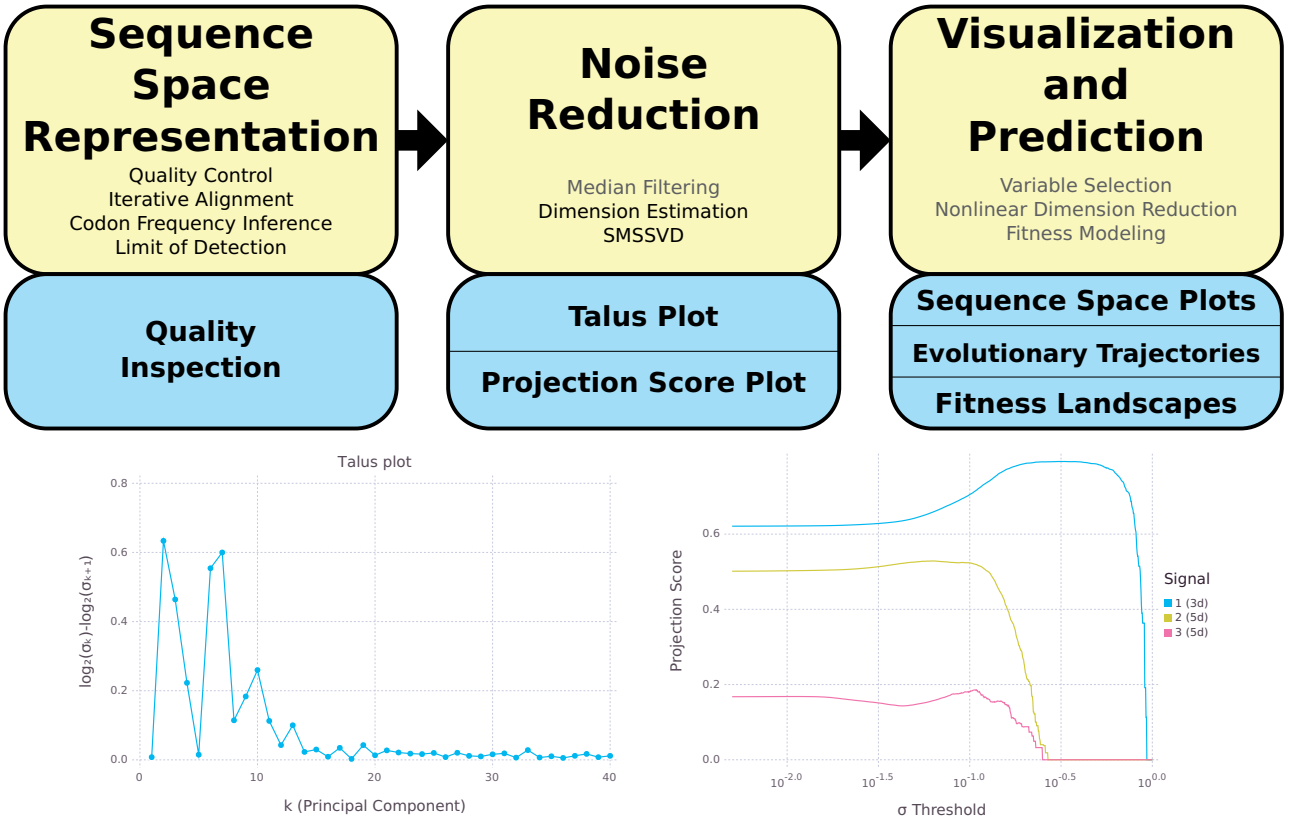

Figure 1. Top: The DISSEQT pipeline. The yellow boxes represent algorithms and data management. The blue boxes represent plots and other output. The analysis history of all results and plots can be traced back all the way to the raw input data. Steps that are only used in some analyses are displayed in gray text. *Sequence Space Representation*: Per sample raw sequencing data is passed through automatic quality control and aligned to a reference genome. Codon frequencies are inferred using quality scores in the aligned data and the limit of detection is estimated for each codon at each site. These are combined to form the sequence space representation. Consensus change reports and read coverage plots aid manual quality inspection. *Noise Reduction*: Median filtering along the time axis is used for time series data. Talus plots are used for dimension estimation and SMSSVD reduces the dimension robustly. *Visualization and Prediction*: Variable selection can be used for finding a small subset of explanatory variables. Nonlinear dimension reduction captures important features for low dimensional visualization of sequence space. Evolutionary trajectories are described in both sample and variable space. Fitness landscape models are used for visualization and prediction. Bottom left: *Talus plot* for the SynSyn data set. After thirteen dimensions, the Talus plot shows small variations around a low mean. Bottom right: *Projection Score Plot* for the SynSyn data set. SMSSVD finds three signals of Dimensions 3, 5, and 5 with different optima for variance filtering. Each curve displays the projection score of a signal as a function of the variance filtering threshold.

Maximum likelihood estimation was used to infer the codon frequencies for each position, using all reads overlapping that position, based on a multinomial model with noisy observations (due to sequencing errors). An initial sequence space representation was then constructed using the codon frequencies and a limit of detection estimated for each possible variant at each site, since sequencing error rates depend on the nucleotide context (Laehnemann, Borkhardt, and McHardy 2016) and the type of substitution (Fox et al. 2014). In this article, we focus on coding regions, which makes codons the natural basis for sequence space modeling, since they are closely connected to biological function and this choice does not impose any assumptions about the relative importance of synonymous versus nonsynonymous changes. All methods presented here are also applicable to non-coding regions, by basing the sequence space model on nucleotides rather than codons.

Second, a dimension estimate of the data was obtained by generating a Talus plot (Fig. 1, bottom left panel, and Supplementary Material), after which noise reduction was performed by SubMatrix Selection Singular Value Decomposition (SMSSVD) (Henningsson and Fontes 2019). SMSSVD is ideal for situations where complex data containing a very large number of variables have signals spread out over different (possibly overlapping) subsets of variables, with the goal of recovering all signals that can be detected, rather than only the strongest one.

Finally, the resulting sequence space representation was used for visualization and phenotype prediction. The evolutionary trajectories of viral populations were followed through time, using sparse methods to find low-frequency minority variants arising and driving the movement of the population in sequence space. Empirical fitness values were used to create fitness landscapes for prediction, using the population representations from Step 2, and for visualization, after an additional nonlinear dimension reduction step vital for getting a useful representation in 2d. The sequence space model created by the DISSEQT pipeline is also intended to be used as input to other software packages, e.g. for clustering and regression.

2.2 Generation of synthetic synonymous viral lineages with altered localization in sequence space and different minority variant compositions

Our goal was to develop and evaluate a pipeline that can capture the discrete signals within the mutant swarms in clinical

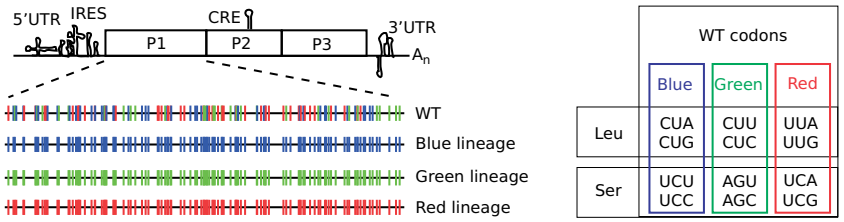

Figure 2. Right: Clusters of Leu/Ser codons according to different viral lineages. Color coding corresponds to synonymous codons used to genetic engineered each viral lineage 'Blue Lineage' (blue), 'Green Lineage' (green), or 'Red Lineage' (red). Left: Schematic of the Coxsackie virus genome indicating RNA structures required for replication (5′UTR, IRES, CRE, and 3′UTR) and the single open reading frame encoding capsid structural proteins (P1 region) and non-structural proteins (P2, P3 regions). The P1 region, in expanded view, shows 117 Ser/Leu codons for the wildtype (WT), blue, green, and red viral lineages.

or experimental samples—essentially, to monitor and analyze evolving populations before significant changes to consensus sequences occur. To do so, we generated samples of mutant swarms, bearing differences in minority variants. We used four genetically trackable virus populations that derived from the same infectious clone wild type Coxsackie virus B3. Within the capsid-coding region of wild type virus, 117 Serine and Leucine codons are represented by all six codons for each amino acid. We generated three additional synthetic synonymous (SynSyn) virus lineages (Fig. 2), some of which were previously published (Moratorio et al. 2017), in which these 117 codons were changed to belong exclusively to only one of three codon categories. These lineages were designed to retain the initial functional neutrality (i.e. same protein sequence), while occupying different starting points and potentially different trajectories in sequence space. Indeed, the differences in fitness values and phenotypes are small in comparison to the differences we observe within the lineages in the experiments described below (see Supplementary Fig. S1). However, the lineages should behave differently as mutations accumulate at these codons, by accessing different mutational neighborhoods with differing impacts of virus fitness.

Next, to introduce changes in minority variant composition without significantly altering consensus sequences of the mutant swarms, we evolved these virus populations in different conditions. Wild type and SynSyn viruses were serially passaged five times in triplicate in normal conditions, as well as in five different mutagenic conditions that are known to increase this virus's mutation rate (Beaucourt et al. 2011) three base analogs (ribavirin, 5-fluorouracil, and 5-azacytidine), amiloride, and Mn^{2+}. Low to moderate concentrations were used to accelerate evolution, while higher concentrations were employed to exacerbate fitness effects. We thus obtained 411 mutant swarms (301 passing strict quality controls) from these varied growth conditions, which were deep sequenced to obtain their entire variant compositions. Importantly, passaged samples in each lineage did not have significant consensus changes (in total across the samples, 144 substitutions at 4 different receptor binding sites and 35 substitutions at 12 other sites).

2.3 DISSEQT reveals that the sequence space occupied by evolving viral populations is of intrinsically low dimensionality

Theoretical sequence space is incredibly large, even for a small genome of length $n = 10,000$, the number of possible sequences is $4^n \approx 4 \cdot 10^{6020}$, so large that the number of atoms in the universe is miniscule in comparison. The number of sequences reachable within just $K = 10$ mutations, $\sum_{k=1}^{K} \binom{n}{k} k^3 \approx 1.6 \cdot 10^{38}$ is still vast, and it is unknown how much of sequence space is occupied by an evolving viral population. We generated Talus and Projection Score (Fontes and Soneson 2011) plots from the sequence data, which provide a visualization of how the contents of a data set spread out across different dimensions. These plots provide a qualitative estimate of the number of dimensions needed to capture all biologically relevant signals that stand out above the background noise. As shown in Fig. 1, bottom left panel, the Talus plot settles after thirteen dimensions, with small variations around a low mean, giving a dimension estimate of thirteen. In the Projection Score plot (Fig. 1, bottom right panel), SMSSVD has detected three signals, of dimensions 3, 5, and 5, where the variance filtering threshold for automatic noise reduction has been optimized for each signal.

Next, we examined which biological signals were captured in each dimension and whether incorporating minority variants in the analysis allows for improved monitoring of evolving populations, compared to consensus sequence analysis. To do this, sequence space representations of the mutant swarms were generated after noise reduction, where the final SMSSVD step decomposed the samples by principal components. Since almost no consensus changes occurred during the experiment, the principal components found patterns essentially related to differences in minority variants between mutant swarms. As shown in Fig. 3, the strongest signal, described by the first three principal components, clearly separates the samples in sequence space according to lineage (see Rows 1–3, above the diagonal, in Fig. 3). Importantly, further analysis of lower dimensions identified all biological treatments that were imposed on the viral populations. A complete separation in sequence space was observed for mutagenic treatment by 5-fluorouracil, ribavirin, and 5-azacytidine (see Rows 4, 5, and 7 below the diagonal, in Fig. 3), known to introduce specific nucleotide substitution biases. Even for treatment with Mn^{2+} and amiloride, which increase natural mutation rates without introducing nucleotide bias, a biological signal could be identified in most of the mutant swarms separating from other samples in Rows 9 and 11 (Fig. 3). Furthermore, these signals are detected despite the background noise and error introduced by the sample preparation and sequencing technology, which lies in even lower components. Finally, if the same analysis is performed using only each sample's consensus sequence, no patterns related to the mutagens is found (Supplementary Fig. S2). These results reveal an important feature of evolving mutant swarms: despite the fact that deep sequencing data is extremely high-dimensional, we showed here that evolving populations of RNA viruses, which present the highest mutation frequencies, are moving in an intrinsically low dimensional domain within sequence space. Indeed, all five of the biological pressures placed

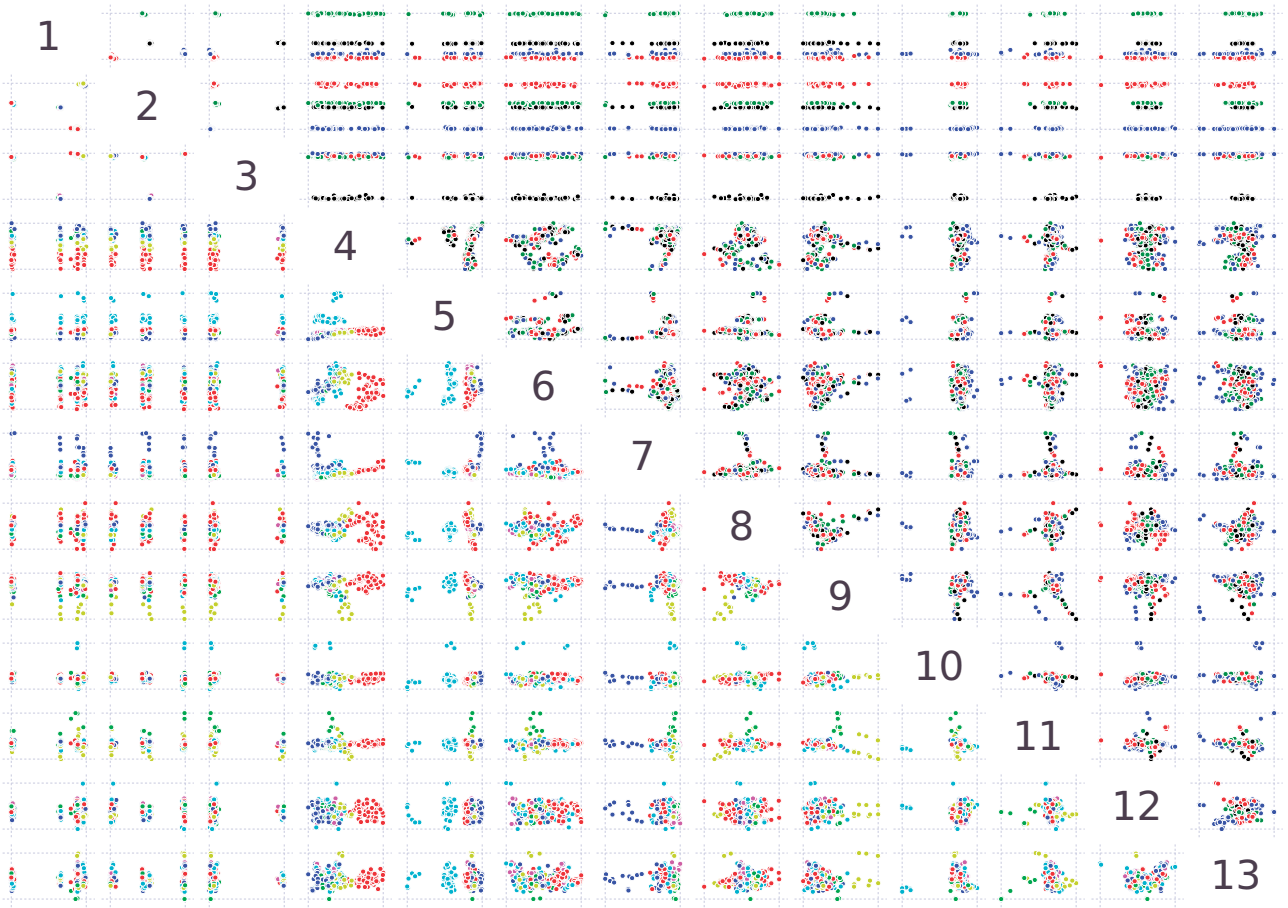

Figure 3. Pairwise scatter plots showing the first thirteen principal components in the analysis of the SynSyn data set plotted against each other. Plots above and below the diagonal are mirror images of each other. Each dot represents one viral population. Above the diagonal, samples are colored by lineage (black: 1, blue: 2, green: 3, red: 4) and below the diagonal, samples are colored by mutagen (red: 5-fluorouracil, light green: amiloride, blue: 5-azacytidine, yellow: Mn2+, cyan: ribavirin, Magenta: mock). All axes are rescaled to fill the plot area.

on these viral populations could be captured within the first thirteen components.

2.4 DISSEQT can monitor evolutionary trajectories and identify the minority variants involved in adaptation

Recently, we studied the adaptation of Coxsackie virus to a new cell line. Long term passages of experimentally evolved populations (120 generations per virus) were analyzed by deep sequencing. Lacking suitable computational tools, the original study focused on identifying variants in the structural protein-coding region of a wild type lineage that showed signs of positive selection in the final passages of adaptation (mutations appearing at >2% in more than one replicate, and only in the structural proteins known to be involved in adaptation to cell culture) (Bordería et al. 2015). In that study, we identified one consensus sequence change that occurred in all lineages during the first ten passages, followed by a cluster of minority variants that reached above 5 per cent in the last passage in each series. The data set however, contained whole-genome sequencing for three lineages of this virus: wild type, a higher replication fidelity lineage, and a lower fidelity lineage. Using DISSEQT, we could obtain a more complete picture by monitoring the evolutionary trajectories of three biological replicates per lineage (Fig. 4), without biasing toward nonsynonymous mutations in the structural protein region. The top panel gives an overview based on nonlinear dimension reduction, showing how the evolutionary trajectories of the replicates relate to each other. For each pair of replicates, the time of bifurcation was computed and this was extended to sample clusters using average linking hierarchical clustering. Before the time of bifurcation, the replicates are close in sequence space and follow the same evolutionary trajectory. The splits in the panel show when the bifurcations occur. All replicates shared the same starting point. Around Passage 4, the low fidelity replicates (yellow–orange) split from the others and shortly thereafter (around Passage 5) the wildtype replicates (magenta–purple) split from the higher fidelity replicates (green–cyan). These observations reflect what was expected, but could not be detected using classical approaches that monitored only a few positively selected alleles: that low fidelity, mutator strains generated more minority variants more rapidly compared to wildtype, and to high fidelity strains. The replicates within each lineage then followed similar trajectories until further bifurcating between Passages 7 and 19. As with the previous examples where lineages clustered together, these results also support the notion that although sequence space is theoretically huge, similar lineages will tend to travel along the same evolutionary trajectories during the initial periods of evolution.

While the above analysis gave information on rate and direction of evolution, it did not identify what minority variant component of each population contributed to adaptation

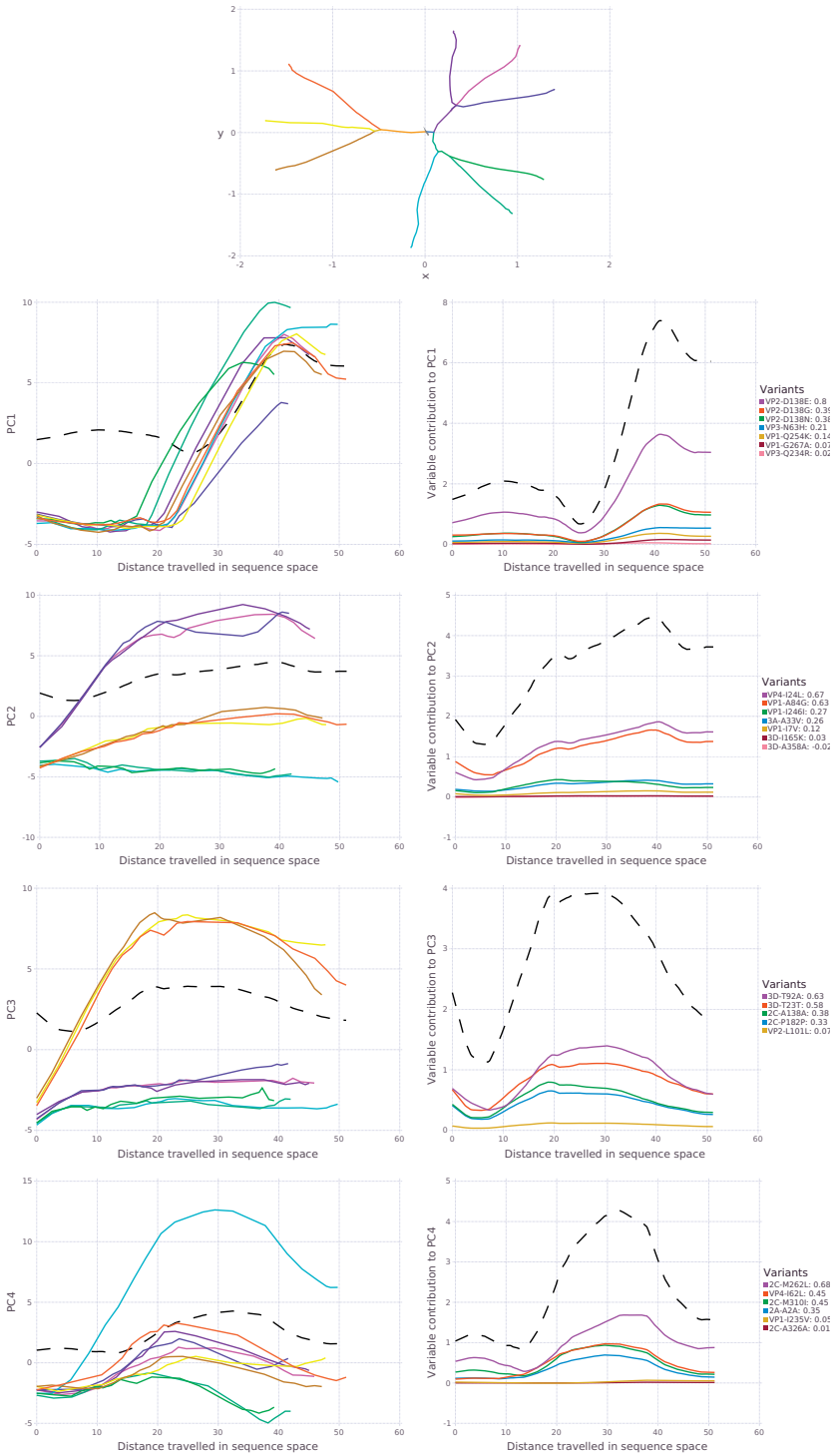

Figure 4. Top: Overview of the evolutionary trajectories of the nine replicates in the adaptability data set (Bordería et al. 2015), shown after nonlinear dimension reduction. Wild type (WT) replicates are shown in magenta–purple colors, replicates from the high fidelity lineage in green–cyan colors and replicates from the low fidelity lineage in yellow–orange colors. The starting point in sequence space is very close for all replicates. The splits indicate when the evolutionary trajectories bifurcate, i.e. when the replicates start to deviate from each other. Left column: Principal components for replicates as a function of arc length. Right column: Variable contributions as a function of arc length. Both columns: The dotted black line shows the total contribution to σ_k at s.

(Bordería et al. 2015) and the observed evolutionary signals. Thus, we broke the analysis down by component over time after variable selection, where principal components determined by SMSSVD followed by SPC (sparse principal components) (Witten, Tibshirani, and Hastie 2009) maps the trajectories of each replicate (Fig. 4, left panel), and identifies which variants contributed most to the signal in each component (right panel). The strongest signal in the first principal component captured time dynamics shared between all replicates regardless of lineage, which consisted of the amino acid residues in the

structural proteins responsible for adaptation to receptor usage (Bordería et al. 2015). The remaining components, however, identified several other mutations at sites that were missed by using the commonly used cutoff of 1–2 per cent minority variant frequency reflecting the expected sequencing error rate, and that could explain subtler phenotypic differences between lineages and between replicates. For wildtype, for example, two additional amino acid changes in the VP1 and VP4 structural proteins contributed most to these lineages' departure from others (principal component 2). Finally, the lower components (4 and onward) revealed variants that explain each replicate's divergence from others, including many variants in non-structural proteins such as the 2C (helicase) and 3C (protease) (Supplementary Fig. S3). Together, the results show that while low-frequency variants were identified at nearly every nucleotide site, the common biologically relevant signals arising during longer term evolution can be captured in relatively low dimension.

2.5 Visualization of evolution along an empirical fitness landscape

In RNA virus evolution, adaptation to new environments can often be attributed to single or few new mutations that become fixed in the population. Experimental evolution in the lab and convergent evolution in the field suggest that short term evolution may be of relatively low dimension, as supported by our findings. If so, then these initial movements in sequence space may be inherently predictable, provided a robust genotype–phenotype map could be generated. This connection between sequence space and fitness is most naturally illustrated as a fitness landscape, where fitness is shown as a function of location in sequence space. However, reconstructing such landscapes from empirical data has been challenging. To evaluate the ability of DISSEQT to correctly generate and visualize fitness landscapes, we first empirically measured the relative fitness of the wild type and SynSyn mutant swarms described above in a direct competition assay against a neutral, genetically marked competitor (Carrasco et al. 2007; Moratorio et al. 2017) (data available in Synapse). The visualization (Fig. 5, top panel) builds upon a 2d representation of sequence space, but using only the first two components from the SMSSVD representation is not sufficient since it ignores all other relevant signals in the data. Nonlinear dimension reduction by Isomap was used to distort sequence space such that the notion of closeness between mutant swarms is respected, taking all signals into account. Fitness was then added as the third dimension, interpolated by the Gaussian Kernel Smoother predictor (performance measured in Fig. 6, top panel). The figure shows the dynamics of the mutant swarms corresponding to each viral lineage evolving over time. The wild type lineage (black) occupied the centermost area of the landscape, surrounded by the other lineages. In general, wild type populations occupied high fitness regions of the landscape, with some variability. This observation confirmed that wild type virus is well adapted to the growth conditions used in these experiments, and should tolerate perturbations in the system, such as increases in mutational load. The green SynSyn lineage displayed the most dramatic fitness differences, reaching both very high and very low areas, whereas the blue SynSyn lineage showed a stable, plateau-like behavior without any significant drops in fitness. Finally, the red SynSyn lineage was stuck in an area of the fitness landscape without any fitness peaks. Indeed, the red lineage was shown to be attenuated *in vivo*, and unable to reach pathogenic outcomes available to wild type virus; while the blue lineage was shown to be more mutationally robust (Moratorio et al. 2017). Supplementary Fig. S4 shows the same fitness landscape, but with samples colored by mutagen. Importantly, the data show that 2d reconstruction of sequence space by nonlinear dimension reduction can adequately reconstruct a fitness landscape that captures the expected biological behavior of similar, yet different viral lineages.

2.6 Prediction of phenotype from genotype requires the input of minority variants

A prime goal in developing faithful representations of sequence space is the potential to assign phenotypes to known genotypes, and ultimately predict the phenotypes of new genotypes. For rapidly evolving populations, the presence of minority variants has been shown to contribute to phenotype, but this is not normally taken into account in genotype–phenotype mapping. Indeed, when the fitness landscape described above was reconstructed using only consensus sequence data, the landscape is considerably collapsed (Fig. 5, bottom panel).

We thus evaluated the relevance of our sequence space reconstructions (after noise reduction) in their ability predict virus fitness, a quantitative parameter often used to describe phenotype. The performance of different fitness models was compared (Fig. 6). Predicting fitness is inherently difficult. Thus, to get a baseline for the optimal performance that could be achieved, we used group-based predictors that rely on sample conditions, rather than deep sequencing data. The fine-grained group predictor using Lineage, Mutagen, and Dosage accurately described the sample conditions (Fig. 6, turquoise bar). In other words, when these three groupings are known for a sample, the prediction is over 69 per cent accurate. When only lineage and dose were considered, prediction was 36 per cent accurate, and if only lineage and mutagen were known, accuracy dropped to 11 per cent. For the landscape predictors based on the 2d Isomap, accuracy was 44 per cent. SMSSVD, on the other hand, which uses 13d reach predictability of 62 and 61 per cent from landscape or nearest neighbor predictors. The data revealed that while 2d Isomap performs well for visualization, prediction is best achieved when more components are incorporated. Importantly, when either Isomap or principal component analysis is performed solely on consensus sequences, prediction fails (2 and 14%, respectively). Furthermore, the performance of the SMSSVD predictors compared well to the predictor based on experimental conditions (Lineage, Mutagen, and Dose), the closest we have to a gold standard. In summary, the predictors based on our proposed sequence space representation vastly outperformed the consensus-based predictor. The data thus confirms that consensus sequencing of a viral population is not enough to accurately predict its phenotype.

3. Discussion

HTS is replacing more classic sequencing methods in microbiology, especially in studying RNA viruses, where every nucleotide can be easily covered with extreme depth. This has increased and renewed interest in better characterizing RNA virus populations to take into account their variability, particularly when trying to identify differences between clinical or experimental samples that have no significant differences in consensus sequence, yet present different phenotypes. Recent works show that indeed, most sites along a genome generate mutants at very low frequency. Following passage of poliovirus in cell

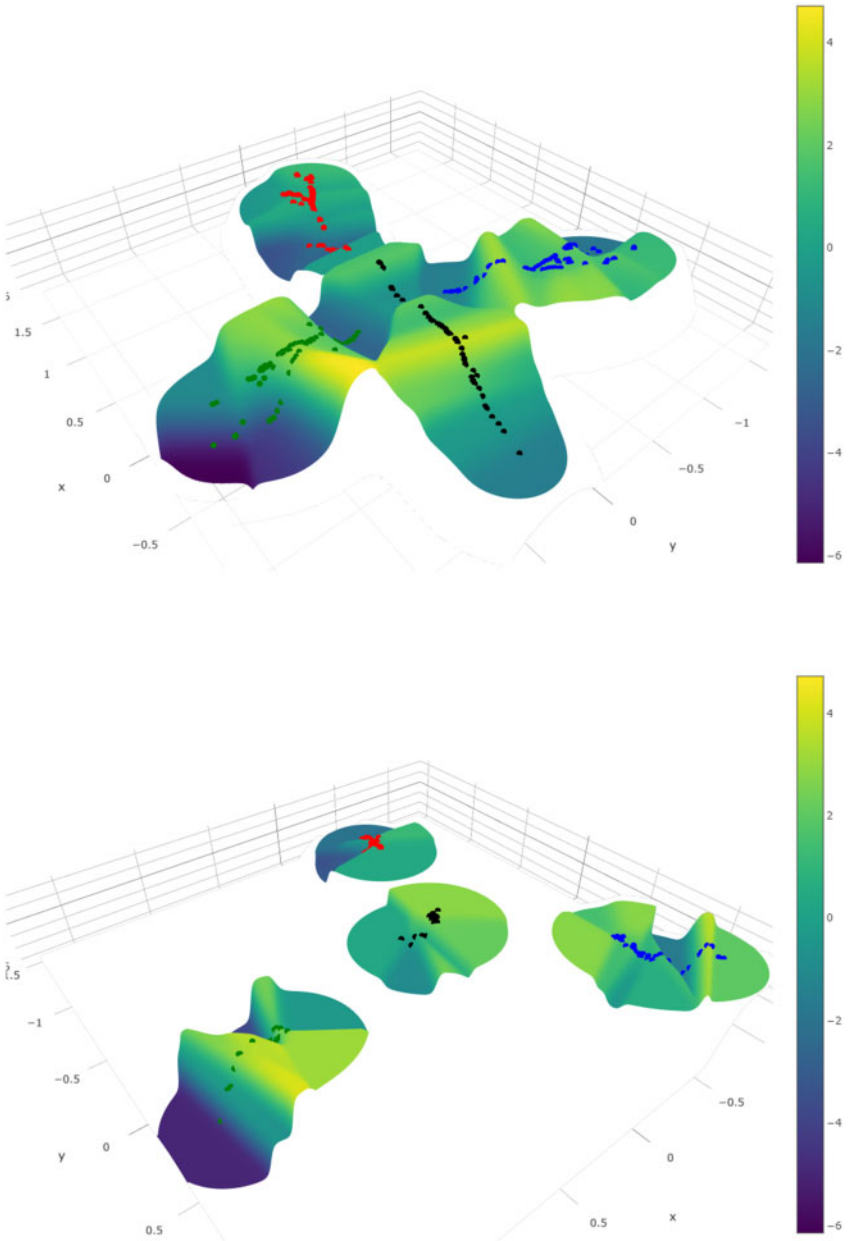

Figure 5. Top: Fitness landscape visualization of the SynSyn data set. Bottom: The same fitness landscape, constructed from consensus data only. Samples are colored by lineage (black: 1, blue: 2, green: 3, red: 4).

culture, Acevedo, Brodsky, and Andino (2014) identified an average of 16,500 variants, the equivalent of ~74 per cent of all possible variant alleles in each passaged sample. Similarly, the previous analysis of the Coxsackie virus B3 wildtype populations described in more detail here, identified variant alleles in 65–80 per cent of the sequenced regions (Bordería et al. 2015).

Despite the increasing accessibility of sequencing technology, we still lack the computational tools to use this data to its full potential. For instance, while an exhaustive list of variants can be generated per sample, to differentiate between similar, yet different, populations most studies have had to settle with using very basic mean measures such as Shannon entropy or mean variance. At best, these were followed up by a more targeted (and biased) focus on the few alleles suspected or known to be involved in the biological question being addressed.

A pre-existing obstacle to developing these tools was the uncertainty as to the size and dimensionality of sequence space actually occupied by evolving microbial populations. Mathematical sequence space is vast, even for the small genomes of RNA viruses. Theoretically, the high mutation rates of RNA viruses could reach a large amount of this space, questioning whether the evolution of these microbes could be inherently predictable. However, it is clear that biological constraints prohibit this from occurring, as most mutations will affect form or function and will not accumulate under strong purifying selection. *In vivo* and *in vitro* experimental evolution studies performed in independent replicates reveal that under a constant environment, the same set of mutations tends to emerge. This suggests that the sequence space available to a virus is indeed more limited, determined by its current genome sequence,

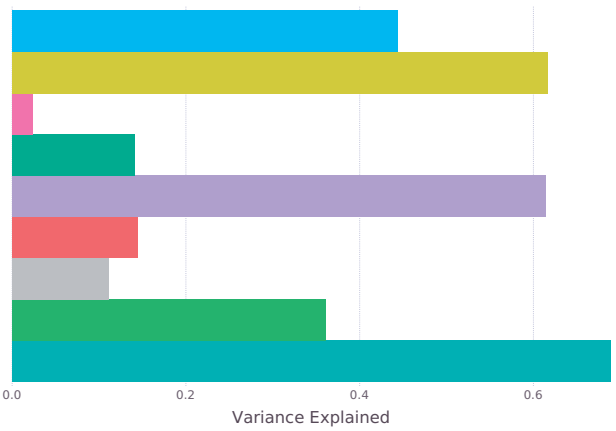

Figure 6. Comparison between different fitness predictors. Gaussian Kernel Smoother Predictors: Isomap 2d (blue), SMSSVD 13d (yellow), Consensus Isomap 2d (pink), and Consensus 13d (green). Nearest Neighbor Predictors: SMSSVD 13d (purple) and Consensus (red). Group Predictors: Lineage/Mutagen (gray), Lineage/Dose (light green), Lineage/Mutagen/Dose (turquoise).

raising the possibility that evolutionary trajectories may therefore be predicted at least in the very short term (the next one or few mutational steps). Fitness landscapes help understand what neighboring populations might represent distributions of genotypes of equal or increasing fitness and which regions define populations of lower fitness. Knowledge of fitness in the vicinity of a current population may help determine the most likely paths that will be taken during the evolution of the population. While this goal may seem lofty for large genomes, the small and highly constrained genomes of RNA viruses may be more amenable to such an exercise.

We have shown how the DISSEQT pipeline, using distribution-based modeling of complex, evolving viral populations, can uncover many different genotypic and phenotypic patterns without needing a priori hypotheses of which genetic alleles are to be studied. Importantly, the robust dimension reduction methods performed here have successfully separated biologically relevant signals from sequencing-related error and other noise, identifying key characteristics of the mutant swarm that drive evolution. This accentuates that global properties, like the 'shape' of the mutant swarm, are significant when trying to predict viral evolution. Sequencing error has long been an issue with characterizing microbial diversity and identifying true single nucleotide variants. Despite the presence of sequencing error, DISSEQT succeeded in finding structure in sequence space, made clear by the co-localization of populations subjected to similar environmental conditions and by accurate fitness predictions and fitness landscapes constructed on top of the sequence space representation.

Applied here, DISSEQT analysis has provided two key pieces of information regarding evolving RNA virus populations. First, that the biologically 'relevant' sequence space occupied by such populations is of intrinsically low dimension. In both data sets presented here, the SynSyn viruses that were manipulated to present discrete biological signals and the High-, Low- and Wildtype fidelity viruses evolving naturally to generate discrete differences in variant composition, the genetic signatures of biological interest were segregated and identified within an intrinsic space of very low dimension (10–20). Second and most importantly, we show that reliable prediction of phenotype from genotype requires the input of minority variants, underscoring the importance of studying RNA viruses, and perhaps other microbial organisms, as a population rather than as a single reference sequence.

At the core of our model is the representation of a population as a measure over a suitable genetic space. Using traditional bulk experimental techniques, averaging is performed already in the sampling and measuring steps of the management and analysis protocol; often resulting in relatively robust and normally (or log-normally) distributed data, well adapted for well-established statistical and machine learning analysis and visualization techniques.

Existing quasispecies reconstruction methods are fundamentally different from DISSEQT in that they will produce a small set of haplotypes and ignore the complete codon distribution at each locus. Haplotype reconstruction based on short sequence reads is a mathematically ill-conditioned problem in the sense that a small change in input data can give a completely different result and it is not feasible to reconstruct haplotypes that exist at low frequencies. The DISSEQT pipeline is designed to be able to handle situations where the signal-to-noise ratio is modest and is able to use information on low-frequency variants to get a more complete representation of the mutant swarms.

We have shown that by directly modeling and representing the distribution at each genetic locus of all measurable minority variants, followed by model reduction, we get low dimensional and robust models that capture the interaction between minority variants and, by coupling it with phenotypic measurements, make it possible to follow and predict trajectories in genotype–phenotype space. It opens up for extending the sequence space models presented in this work to situations where heterogeneity of populations can be hypothesized to be an important aspect that can be measured in a direct manner. In particular, data coming from single-cell sequencing have more variability, more artifacts, and often complex distributions (Bacher and Kendziorski 2016) and distribution-based modeling can be envisioned to be viable and provide a natural and biologically accurate representation of the data.

Cancer growth, fundamentally different in origin from viruses and bacteria, may still be usefully described in terms of similar evolutionary processes (Stratton, Campbell, and Futreal 2009). Larger genomes and frequent structural variation, such as chromosomal aberrations (Hanahan and Weinberg 2011) and fused genes (Lilljebjörn et al. 2016) does, however, make the situation more complex and further work is needed to adapt the sequence space modeling for these circumstances. The challenges lie in incorporating structural variants into to the underlying space in a way that preserves biological similarity and is feasible to infer from the data. A possible starting point for cancer data is to restrict the analysis to a chosen set of interesting genes that do not exhibit any structural variation, thus simplifying the collection of deep sequencing data and providing an easy fit to the sequence space models we propose.

4. Methods

4.1 Experimental setup

The experimental setup is described in detail in Moratorio et al. (2017) for the SynSyn data set and in Bordería et al. (2015) for the time series data on adaptation of Coxsackie virus to a new cell line.

4.2 Reproducible and traceable analysis

Traceability in the DISSEQT pipeline is provided by integration with the collaborative science platform Synapse. Every result

produced by DISSEQT can be traced back all the way to the original data files using the Synapse *provenance graph*, which describes the actions taken for every analysis step and connects input to output data. Sharing settings in Synapse makes it possible to open up the entire analysis to the public, but keeping sensitive data and unfinished analyses private if necessary. The analysis steps are self-contained in the sense that all data required to produce the output is downloaded from Synapse as needed. Hence, every analysis step can be reproduced locally by anyone executing the same actions. By changing parameters or making other changes, the impact of performing the analysis in a different manner can be investigated by others. Rerunning the entire analysis is also possible in this way. Furthermore, the analysis can be adapted to new data sets, such that the results can reproduced from new biological data.

4.3 Iterative alignment

For each mutant swarm, the sequenced reads were aligned to the reference genome of the lineage (WT or one of the SynSyn's) using BWA-MEM (Li 2013). The choice of alignment tool is not critical, but the same one should be used for all samples (mutant swarms) to get a consistent analysis. After alignment, the consensus sequence of the aligned sample is computed. If the consensus differs from the reference genome, the alignment starts over, now using the consensus as the new reference genome. This process is repeated until the consensus does not change. Iterative alignment combats an inherent problem that occurs when aligning to a reference genome—there will be a bias since reads that match the reference genome are easier to align, while reads that differ might be mapped incorrectly or cut off such that the variant is not included in the alignment. For changes at the consensus level, iterative alignment thus ensures that more reads are mapped correctly, allowing for a better frequency estimate. Even more important is that the ability to detect minority variants in the vicinity of consensus level changes is greatly improved, as the number of differences between reads containing the minority variant and the consensus will tend to be lower.

4.4 Quality control

Generating deep sequencing data is a complex procedure with many steps performed, both for the experiment itself and to prepare the data for sequencing. The DISSEQT pipeline provides several ways to evaluate the data to make sure that it is of high quality. Before alignment, adapters and poor quality bases are trimmed from the ends of reads using fastq-mcf (Aronesty 2011). At the end of the iterative alignment procedure, consensus sequences are automatically generated for all samples. It is expected that the consensus sequence will be more similar to the reference of the sample lineage, than to the reference of any other lineage used in the same sequencing run. If this is not the case, the sample is flagged as being mislabeled. Indels are also reported. Graphs showing the read coverage as a function of genome position are created. All samples from the same lineage and sequencing run are put in one graph, making it possible to identify problems with low read coverage for certain samples or genomic regions at a glance. Samples with a low mean read coverage can be removed automatically from downstream analysis. What threshold to use depends on the experimental setup, but we recommend keeping only samples with a mean read coverage above 1,000 for deep sequencing data. There are also tools in DISSEQT to remove samples that are suspected of being contaminated by other samples, identified by having a mixture of reads that are likely to originate from different reference genomes. The purpose of quality control is to validate that we are indeed studying what we set out to study. If a sample is showing unexpected patterns, in particular during quality control, we recommend that the aligned reads, the consensus sequence and any other measurements are inspected manually ensure that conclusions are not drawn from faulty data.

4.5 Haplotypes

Recovering the haplotype mix from a collection of short reads is a difficult, often ill-conditioned, and computationally intensive problem, but several software tools (Zagordi et al. 2011; Prosperi and Salemi 2012) are available (also see Beerenwinkel et al. 2012; McElroy, Thomas, and Luciani 2014) for overviews. The dominant haplotypes and their frequencies do not, however, completely characterize the viral population, another important aspect is how dispersed the individual viruses are around these central haplotypes. V-Phaser (Macalalad et al. 2012), V-Phaser 2 (Yang et al. 2013), and ShoRAH (Zagordi et al. 2011) find phased variants, pushing down the detection limit by assuming that real variants (at nearby loci) tend to co-vary, while errors do not. Unfortunately, V-Phaser and ShoRAH do not scale well for large data sets and V-Phaser 2 requires paired-end reads. For the reasons above, we chose the simpler and more robust path of making maximum likelihood (ML) estimates of the variant frequencies at each position, based on base quality data.

4.6 Sequence space representation

The genomic composition of microbial populations can be represented by a positive measure over a suitable space. Let Σ be an alphabet set, e.g. the set of nucleotides $\Sigma = \mathcal{N} := \{\mathrm{A},\ \mathrm{C},\ \mathrm{G},\ \mathrm{T}\}$, the set of codons $\Sigma = \mathcal{N} \times \mathcal{N} \times \mathcal{N}$ or the set of amino acids $\Sigma = \mathcal{A} = \{\mathrm{A}, \mathrm{R}, \mathrm{N},\ \ldots\}$. For the rest of this article, the set of codons will be used as the alphabet set, since the codons are closely connected to biological function and this choice does not impose any assumptions about the relative importance of synonymous versus nonsynonymous changes. The set of codons is the natural choice for coding regions, to analyze non-coding regions, the set of nucleotides could be used instead. Now define *sequence space* Σ^n as the set of sequences of length n over the alphabet Σ. Assuming that individual genomes in the population only differ by a finite number of point mutations (i.e. substitutions), the composition of the population is characterized by a positive measure over sequence space. The space of positive measures over sequence space will be denoted by $\mathcal{P}(\Sigma^n)$.

Inference of the population composition can be intractable from sequencing data due to short reads and/or high error rates. Let $P, Q \in \mathcal{P}(\Sigma^n)$ and define an equivalence relation such that $P \sim Q$ iff

$$P(C_i[x]) = \alpha Q(C_i[x]), \quad \forall x \in \Sigma,\ i \in \{1,\ 2, \ldots, n\},$$

for some constant $\alpha \in \mathbb{R}^+$, where

$$C_i[x] := \ \{s \in \Sigma^n;\ s_i = x\}$$

are the basic cylinder sets of Σ^n. Hence, P relates to Q if they have the same allele frequencies at all positions. Inference for the equivalence class $[P]$ from sequence data is possible even when P cannot be inferred since allele frequencies at different

positions can be estimated separately. The drawback is that minority variant linkage is lost, since we do not attempt single viron haplotype reconstruction. Each equivalence class $[P]$ is naturally represented by the frequency matrix $p \in \mathbb{R}^{n\times|\Sigma|}$ with $p_{i,x} = P(C_i[x])/P(\Sigma^n)$. Finally, the frequencies are transformed by $p \to \log_2(p+\alpha)$, where α denotes the limit of detection, to give minority variants higher impact in the model. The log transformation emphasizes relative differences in frequencies between variants instead of absolute differences in frequencies between variants.

4.7 Sequence space inference

Maximum likelihood estimation was used to infer the codon frequencies at any given position, using all reads overlapping that position, based on a multinomial model with noisy observations. At a given locus, let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{64})$ be the frequencies in the population for the sixty-four different codons, with $\theta_i \geq 0$ for all i and $\sum_i \theta_i = 1$. Consider a read and the fragment the read is sequenced from. Now let x be the observed codon in the read and z the unknown codon in the original fragment, then

$$P(x\,|\,|\boldsymbol{\theta}) = \sum_{z=1}^{64} P(x|z,\boldsymbol{\theta})P(z|\boldsymbol{\theta}) \;\; = \sum_{z=1}^{64} P(x|z)\theta_z.$$

We model $P(x|z)$ using the quality scores of the bases in the codon. If $\epsilon_1, \epsilon_2, \epsilon_3$ are the probabilities of a read error at bases 1, 2, and 3 in the codon and y^k is the base at position k in a codon y, then

$$P(x\,|\,|z) = \prod_{k=1}^{3}\left(\delta_{z^k}^{x^k}(1-\epsilon_k) + \left(1-\delta_{z^k}^{x^k}\right)\frac{\epsilon_k}{3}\right),$$

where ${\delta_a}^b$ is the Kronecker delta, the errors are thus assumed to be independent between bases in the codon and read errors are assumed to be equally likely to result in any of the other three bases. Assuming independent reads, the probability of the observations is

$$P(\boldsymbol{x}\,|\,|\boldsymbol{\theta}) = \prod_{i}^{N}\left(\sum_{z=1}^{64} P(x_i|z)\theta_z\right)$$

with observed codons $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ from reads 1 to N. The log-likelihood is thus

$$l(\boldsymbol{\theta};\boldsymbol{x}) = \sum_{i}^{N}\log\left(\sum_{z=1}^{64} P(x_i|z)\theta_z\right),$$

which is maximized numerically.

We noted that in our high read coverage data, bases with low-quality Phred scores tended to be biased toward certain nucleotide errors. Thus, we chose to not trust bases with a Phred score below 30. This was done by setting the ϵ_k of such nucleotides to 0.75, giving them no influence. Reads were excluded from the analysis if they caused the ML optimization problem to be underdetermined (e.g. when observing two reads with codons AAA and xAT respectively, where x means that the nucleotide is unknown, xAT is dropped since only the sum of the frequencies for AAT, CAT, GAT, and TAT can be determined).

4.8 Limit of detection

True minority variants can be hard to separate from sequencing errors. And in both cases, we expect the frequencies to be different depending on the nucleotide neighborhood and other factors (DePristo et al. 2011). A key difference is however that there are two sets of observations of the sequencing errors since the reads originating from the forward and reverse strands have different nucleotide neighborhoods for any given codon site. Indeed, for each sample, the codon frequencies from the two strands are expected to be approximately equal for true minority variants, something which is much less likely for sequencing errors. The differences in sequencing error behavior depending on the context thus lead us to estimate the limit of detection α separately for each locus and codon. For a given locus, the samples are grouped by run and consensus codon, to get similar sequencing errors across the samples in each group. Fix a codon and let $\boldsymbol{f}$ and $\boldsymbol{r}$ be two vectors where f_i and r_i are the inferred codon frequencies using reads from only the forward and reverse strands respectively, for sample i in the group. To limit the impact of sequencing errors on the downstream analysis, the transformed frequencies should be approximately equal, i.e. give a low value of the norm

$$\psi(\alpha) = ||\log_2(\boldsymbol{f}+\alpha\mathbb{1}) - \log_2(\boldsymbol{r}+\alpha\mathbb{1})||_{\text{RMS}}$$

where $\log_2$ acts elementwise and $\mathbb{1}$ is a vector of all ones. Now define the limit of detection

$$\alpha := \inf\left\{t \geq 0;\;\; \psi(t) \;\leq\; \log_2(1.5)\right\}.$$

The infimum exists since ψ is continuous and $\psi(t) \to 0$ as $t \to \infty$. The threshold $\log_2(1.5)$ is chosen such that if we have a single sample with $f_1 = x$ and $r_1 = 0$, then $\alpha = 2x$. Furthermore, ψ is a strictly decreasing function and α can thus be found by the bisection method or other root-finding methods. Finally we choose a conservative estimate of the limit of detection $\alpha_{c,x}$, for codon c at locus x, by taking the highest limit of detection estimated from the different sample groups $1, 2, \ldots, G$,

$$\alpha_{c,x} := \max\left\{10^{-3}, \alpha_{c,x}^{(1)}, \alpha_{c,x}^{(2)}, \ldots, \alpha_{c,x}^{(G)}\right\},$$

with upper indices denoting the sample group and where 10^{-3} is a commonly accepted lower limit of detection for sequencing data (Goodwin, McPherson, and McCombie 2016).

4.9 Dimension estimation using Talus plots

The Talus Plot provides a visualization of how the contents of a data set spread out across different dimensions and is designed to make it as easy as possible to make a qualitative estimate of the number of dimensions needed to capture all signals that stand out above the background noise. In Supplementary Material, we show how predictable aspects of the background noise can be used to discern signals from noise. In brief, when the Talus Plot has 'settled', with small variations around a low mean, then the noise can be expected to be dominant.

4.10 SMSSVD

SMSSVD (Henningsson and Fontes 2019) is a parameter-free dimension reduction technique designed for the reconstruction of multiple overlaid low-rank signals from a data matrix, corrupted by noise. It is ideal for exploratory analysis of complex

data, where different signals are spread out over different (possibly overlapping) subsets of variables, by limiting the influence of noise in variables that are not contributing to the signal. One of the major benefits of SMSSVD is its ability to detect signals with a low signal-to-noise ratio. SMSSVD shares many relevant properties with SVD, in particular orthogonality between components and the ability to extract variable loadings. The DISSEQT pipeline uses SMSSVD for noise reduction of the sequence space representation, since the number of variables is very large and we are trying to recover all signals that can be detected, not only the strongest one. Before applying SMSSVD, the data matrix is centered.

4.11 Fitness landscapes

Fitness landscapes, an important kind of genotype–phenotype map, are used to illustrate the connection between sequence composition and fitness of organisms. Here we show how a fitness landscape can be generated entirely from empirical data. Given a d-dimensional representation of sequence space, i.e. a set of sample points $\mathbf{x}^{(i)} \in \mathbb{R}^d$, $i = 1, \ldots, N$ with corresponding fitness values $y^{(i)} \in \mathbb{R}$, we want to reconstruct a surface $f : \mathbb{R}^d \to \mathbb{R}$ such that

$$f(\mathbf{x}^{(i)}) = y^{(i)}.$$

In practice however, we cannot expect a perfect fit of the surface. Differences in fitness between sample points that are close in the low dimensional representation will be difficult to capture. Furthermore, measurement noise will impact the reproducibility of the surface. To get a robust fitness landscape, we use a Gaussian Kernel Smoother (Hastie, Tibshirani, and Friedman 2009) and select the kernel width σ by cross-validation (repeated random subsampling). That is, we numerically find

$$\underset{\sigma}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{j=1}^{M} \left({f_{\text{train}}}^{(i)} \left(\mathbf{x}_{\text{test}}^{(i,j)}, \sigma \right) - y_{\text{test}}^{(i,j)} \right)^2,$$

where the data is randomly divided into a train and a test data set for each iteration i and

$$f_{\text{train}}^{(i)}(\mathbf{z}, \sigma) := \frac{\sum_{j=1}^{M} w_{\mathbf{z},\sigma}^{(i,j)} y_{\text{train}}^{(i,j)}}{\sum_{j=1}^{M} w_{\mathbf{z},\sigma}^{(i,j)}}, \text{ with } w_{\mathbf{z},\sigma}^{(i,j)} = e^{-\frac{\left\| \mathbf{z} - \mathbf{x}_{\text{train}}^{(i,j)} \right\|_2^2}{2\sigma^2}}.$$

4.12 Fitness evaluation

The Gaussian Kernel Smoother (fitness landscape) predictors are evaluated in comparison to other fitness predictors. Nearest Neighbor predictors uses the fitness of the closest sample in sequence space as the prediction and can be used for different sequence space models. In case of ties, the prediction is taken as the average over the tied samples. Group-based predictors use a predetermined grouping of the samples, predicting fitness as the average fitness among samples in the same group, and do not use sequence data at all.

Model accuracy for a predictor f is measured by fraction of variance explained,

$$1 - \frac{\sum_i \left(f(\mathbf{x}^{(i)}) - y^{(i)} \right)^2}{\sum_i \left(y^{(i)} - -y \right)^2},$$

where $\mathbf{x}^{(i)}$ is the representation of sample i used by the predictor, $y^{(i)}$ is the fitness of sample i, $-y$ the mean fitness over all samples and the second term in the expression is the variance of the residuals divided by the total variance. The models are evaluated by leave-one-out cross-validation. The kernel widths for the Gaussian Kernel Smoother predictors are estimated separately for each problem instance to avoid influence from the left-out sample.

4.13 Variable selection

We use SPC (Witten, Tibshirani, and Hastie 2009) for variable selection, after noise reduction by SMSSVD. SPC adds a variable-side L_1 (lasso) constraint to a formulation of SVD as an optimization problem, forcing sparsity by ensuring that many variables are 0 at the optima. The optimization problem is then solved for one component at a time, using an iterative algorithm. However, since the optimization problem is not necessarily convex, the algorithm might converge to local optima. To reduce the impact of this problem, and to ensure that the singular values are declining, we suggest an extension of the algorithm. It can be shown that if a component has a larger singular value than a previous one, then this solution is guaranteed to be a better starting guess for the optimization problem for the previous component. By rolling back and restarting the optimization at the previous component, we get closer to the globally optimal solution and make sure that the singular values are declining.

4.14 Nonlinear dimension reduction

By dimension reduction, we aim to identify the parts of sequence space that are explored by the samples. Linear dimension reduction techniques, like SMSSVD, are useful because they make very few assumptions about the structure of the data. Although there is no reason to believe that the underlying manifold is linear, the complexity that is necessary for biological systems is indeed often caused by nonlinearities, linear methods can still capture nonlinear patterns if the dimension is sufficiently high (Nash 1956). However, to get an informative visualization in just two or three dimensions, nonlinear dimension reduction is needed for complex data sets.

We apply Isomap (Tenenbaum, De Silva, and Langford 2000) to the data set after the noise reduction by SMSSVD (and the optional variable selection). Nonmetric MDS using Kruskal's stress criterion (Kruskal 1964) was used rather than classical MDS in the final step of the Isomap algorithm. This distorts the underlying space by expanding local structure that would otherwise be too small to notice, giving some importance to weaker signals in the data.

4.15 Time series

The evolution of a population over time is described by a curve $\boldsymbol{p}(t)$ in sequence space. In practice, we can only measure the values of a curve $\boldsymbol{p}(t)$ at discrete time points, and the measurements are subjected to noise. As the first step of noise reduction, a 3-point median filter over time is applied to the sequence space representation, to robustly reduce the impact of noise spikes. Following the noise reduction, the curve $\boldsymbol{p}(t)$ is reconstructed in the d-dimensional representation of sequence space, as a piecewise linear curve connecting the data points. Then, each curve is reparameterized by arc length s, starting at $s = 0$ for $t = 0$, since differences in mutation rates can cause the population to move at different speeds through sequence space.

The sequence space representation in terms of variables (variants) is time-invariant, but it is nevertheless important to

see how different parts of sequence space are explored as the replicates move. Let σ be the first singular value, with corresponding left and right singular vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, after dimension reduction of a matrix X by SMSSVD, SVD or SPC, then σ can be decomposed as a sum over variables and samples,

$$\sigma = \boldsymbol{u}^T X \boldsymbol{v} = \sum_{i,j} u_i X_{ij} v_j = \sum_{i,j} \sigma_{ij},$$

where $\sigma_{ij} := u_i X_{ij} v_j$ quantifies the importance of variable i and sample j for this component. By linear interpolation, this can be extended to $\sigma_j^{(r)}(s)$, for intermediate values of the curve parameter s for replicate r. The contribution of variable j at s is measured by $\sigma_j(s) := \sum_r \sigma_j^{(r)}(s)$ and $\sigma(s) := \sum_j \sigma_j(s)$ describes the importance of the first principal component at s. Plotting $\sigma(s)$ and $\sigma_j(s)$ along with the replicates, thus aid understanding of the dynamics. The definitions naturally extend to multiple components.

4.16 Bifurcations

We define the time of bifurcation $\beta(\boldsymbol{p}, \boldsymbol{q})$, between two curves $\boldsymbol{p}(t)$ and $\boldsymbol{q}(t)$ as the similarity measure

$$\beta(\boldsymbol{p}, \boldsymbol{q}) = \inf \{t : ||\boldsymbol{p}(t) - \boldsymbol{q}(t)||_2 \geq Bm\},$$

that is, the first point in time at which the distance between $\boldsymbol{p}(t)$ and $\boldsymbol{q}(t)$ is above a threshold. Here, B is a chosen threshold and m a normalization constant chosen to make the expression scale-invariant. If $\boldsymbol{p}^{(i)}(t)$ is defined for $t \in [0, T_i]$ and $T_{ij} := \min(T_i, T_j)$, then the mean distance over time between curves i and j is

$$m_{ij} = \frac{1}{T_{ij}} \int_0^{T_{ij}} ||\boldsymbol{p}^{(i)}(t) - \boldsymbol{p}^{(j)}(t)||_2 dt$$

and we let $m := \frac{1}{N(N-1)} \sum_{i \neq j} m_{ij}$, the mean over all pairs of the N curves. Average linking hierarchical clustering, based on the time of bifurcation similarity scores, naturally extends the concept to clusters of samples, giving recursive cluster splits and a cluster similarity score equal to the time of bifurcation at each split. For piecewise linear curves, m and $\beta\left(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(j)}\right)$ can be computed analytically.

Acknowledgements

This work is sponsored by the Defense Advanced Research Projects Agency INTERCEPT program managed by Dr Jim Gimlett, and administered though DARPA contract (HR00111720023): the content of the information does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

Data availability

Data is available through GitHub (https://github.com/rasmushenningsson/DISSEQT.jl, accessed 23 June 2019) and Synapse (https://www.synapse.org/#!Synapse: syn11425758, accessed 23 June 2019). Synapse DOI: 10.7303/syn11639899 and GitHub DOI: 10.5281/zenodo.3344557.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

References

Acevedo, A., Brodsky, L., and Andino, R. (2014) 'Mutational and Fitness Landscapes of an RNA Virus Revealed through Population Sequencing', *Nature*, 505: 686–90.

Aronesty, E. (2011) 'ea-utils: Command-Line Tools for Processing Biological Sequencing Data' <https://github.com/ExpressionAnalysis/ea-utils> accessed 23 June 2019.

Bacher, R., and Kendziorski, C. (2016) 'Design and Computational Analysis of Single-Cell RNA-Sequencing Experiments', *Genome Biology*, 17: 63.

Beaucourt, S. et al. (2011) 'Isolation of Fidelity Variants of RNA Viruses and Characterization of Virus Mutation Frequency', *Journal of Visualized Experiments*, 16. DOI: 10.3791/2953.

Beerenwinkel, N. et al. (2012) 'Challenges and Opportunities in Estimating Viral Genetic Diversity from Next-Generation Sequencing Data', *Frontiers in Microbiology*, 3: 329.

Biebricher, C., and Eigen, M. (2006) 'What Is a Quasispecies?', *Current Topics in Microbiology and Immunology*, 299: 1–31.

Bordería, A. et al. (2015) 'Group Selection and Contribution of Minority Variants during Virus Adaptation Determines Virus Fitness and Phenotype', *PLoS Pathogens*, 11: e1004838.

Carrasco, P. et al. (2007) 'A Real-Time RT-PCR Assay for Quantifying the Fitness of Tobacco Etch Virus in Competition Experiments', *Journal of Virological Methods*, 139: 181–8.

Collins, J. et al. (2004) 'Competitive Fitness of Nevirapine-Resistant Human Immunodeficiency Virus Type 1 Mutants', *Journal of Virology*, 78: 603–11.

DePristo, M. et al. (2011) 'A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data', *Nature Genetics*, 43: 491–8.

Domingo, E., Sheldon, J., and Perales, C. (2012) 'Viral Quasispecies Evolution', *Microbiology and Molecular Biology Reviews*, 76: 159–216.

Fiers, W. et al. (1976) 'Complete Nucleotide Sequence of Bacteriophage MS2 RNA: Primary and Secondary Structure of the Replicase Gene', *Nature*, 260: 500–7.

Fleischmann, R. et al. (1995) 'Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd.', *Science*, 269: 496–512.

Fontes, M., and Soneson, C. (2011) 'The Projection Score—An Evaluation Criterion for Variable Subset Selection in PCA Visualization', *BMC Bioinformatics*, 12: 307.

Fox, E. et al. (2014) 'Accuracy of Next Generation Sequencing Platforms', *Next Generation, Sequencing & Applications*, 1. DOI: 10.4172/jngsa.1000106.

Fraser, C. et al. (1995) 'The Minimal Gene Complement of *Mycoplasma genitalium*', *Science*, 270: 397–403.

Gawad, C., Koh, W., and Quake, S. (2016) 'Single-Cell Genome Sequencing: Current State of the Science', *Nature Reviews Genetics*, 17: 175–88.

Goodwin, S., McPherson, J., and McCombie, W. (2016) 'Coming of Age: Ten Years of Next-Generation Sequencing Technologies', *Nature Reviews Genetics*, 17: 333–51.

Hanahan, D., and Weinberg, R. (2011) 'Hallmarks of Cancer: The Next Generation', *Cell*, 144: 646–74.

Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning*, 2nd edn. Springer.

Henningsson, R., and Fontes, M. (2019) 'SMSSVD: SubMatrix Selection Singular Value Decomposition', *Bioinformatics*, 35: 478–86.

Higgins, D. (1992) 'Sequence Ordinations: A Multivariate Analysis Approach to Analysing Large Sequence Data Sets', *Computer Applications in the Biosciences*, 8: 15–22.

Kauffman, S., and Weinberger, E. (1989) 'The NK Model of Rugged Fitness Landscapes and Its Application to Maturation of the Immune Response', *Journal of Theoretical Biology*, 141: 211–45.

Kouyos, R. et al. (2012) 'Exploring the Complexity of the HIV-1 Fitness Landscape'. *PLOS Genetics*, 8: e1002551.

Kruskal, J. (1964) 'Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis', *Psychometrika*, 29: 1–27.

Laehnemann, D., Borkhardt, A., and McHardy, A. (2016) 'Denoising DNA Deep Sequencing Data—High-Throughput Sequencing Errors and Their Correction', *Briefings in Bioinformatics*, 17: 154–79.

Li, H. (2013) 'Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM', *arXiv 1303.3997*.

Lilljebjörn, H. et al. (2016) 'Identification of ETV6-RUNX1-Like and DUX4-Rearranged Subtypes in Paediatric B-Cell Precursor Acute Lymphoblastic Leukaemia', *Nature Communications*, 7: 11790.

Macalalad, A. et al. (2012) 'Highly Sensitive and Specific Detection of Rare Variants in Mixed Viral Populations from Massively Parallel Sequence Data', *PLoS Computational Biology*, 8: e1002417.

McElroy, K., Thomas, T., and Luciani, F. (2014) 'Deep Sequencing of Evolving Pathogen Populations: Applications, Errors, and Bioinformatic Solutions', *Microbial Informatics and Experimentation*, 4: 1.

Moratorio, G. et al. (2017) 'Attenuation of RNA Viruses by Redirecting Their Evolution in Sequence Space', *Nature Microbiology*, 2: 17088.

Nash, J. (1956) 'The Imbedding Problem for Riemannian Manifolds', *Annals of Mathematics*, 63: 20–63.

Perkel, J. (2017) 'Single-Cell Sequencing Made Simple', *Nature*, 547: 125–26.

Prosperi, M., and Salemi, M. (2012) 'QuRe: Software for Viral Quasispecies Reconstruction from Next-Generation Sequencing Data', *Bioinformatics*, 28: 132–3.

Reiter, J. et al. (2015) 'Biological Auctions with Multiple Rewards', *Proceedings: Biological Sciences*, 282: 20151041.

Russell, A., Trapnell, C., and Bloom, J. (2017) 'Extreme Heterogeneity of Influenza Virus Infection in Single Cells', *bioRxiv*, 193995.

Seifert, D. et al. (2015) 'A Framework for Inferring Fitness Landscapes of Patient-Derived Viruses Using Quasispecies Theory', *Genetics*, 199: 191–203.

Smith, J., and Price, G. (1973) 'The Logic of Animal Conflict', *Nature*, 246: 15–8.

Stapleford, K. et al. (2014) 'Emergence and Transmission of Arbovirus Evolutionary Intermediates with Epidemic Potential', *Cell Host & Microbe*, 15: 706–16.

Stratton, M., Campbell, P., and Futreal, P. (2009) 'The Cancer Genome', *Nature*, 458: 719–24.

Svensson, V. et al. (2016) 'Power Analysis of Single Cell RNA-Sequencing Experiments', *bioRxiv*, 73692.

Tenenbaum, J., De Silva, V., and Langford, J. (2000) 'A Global Geometric Framework for Nonlinear Dimensionality Reduction', *Science*, 290: 2319–23.

Vignuzzi, M. et al. (2006) 'Quasispecies Diversity Determines Pathogenesis through Cooperative Interactions in a Viral Population', *Nature*, 439: 344–8.

Whitlock, M., and Bourguet, D. (2000) 'Factors Affecting the Genetic Load in *Drosophila*: Synergistic Epistasis and Correlations among Fitness Components', *Evolution*, 54: 1654–60.

Witten, D., Tibshirani, R., and Hastie, T. (2009) 'A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis', *Biostatistics*, 10: 512–34.

Wright, S. (1932) 'The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution', *Proceedings of the 6th International Congress on Genetics*, 1: 356–66.

Xue, K. et al. (2016) 'Cooperation between Distinct Viral Variants Promotes Growth of H3N2 Influenza in Cell Culture', *eLife*, 5: e13974.

Yang, X. et al. (2013) 'V-Phaser 2: Variant Inference for Viral Populations', *BMC Genomics*, 14: 674.

Zagordi, O. et al. (2011) 'ShoRAH: Estimating the Genetic Diversity of a Mixed Sample from Next-Generation Sequencing Data', *BMC Bioinformatics*, 12: 119.

Zhang, J. et al. (2011) 'PhyloMap: An Algorithm for Visualizing Relationships of Large Sequence Data Sets and Its Application to the Influenza A Virus Genome', *BMC bioinformatics*, 12: 248.