Dimension Reduction and Signal Decomposition for Genotype–Phenotype Relations

Perby Henningsson, Rasmus

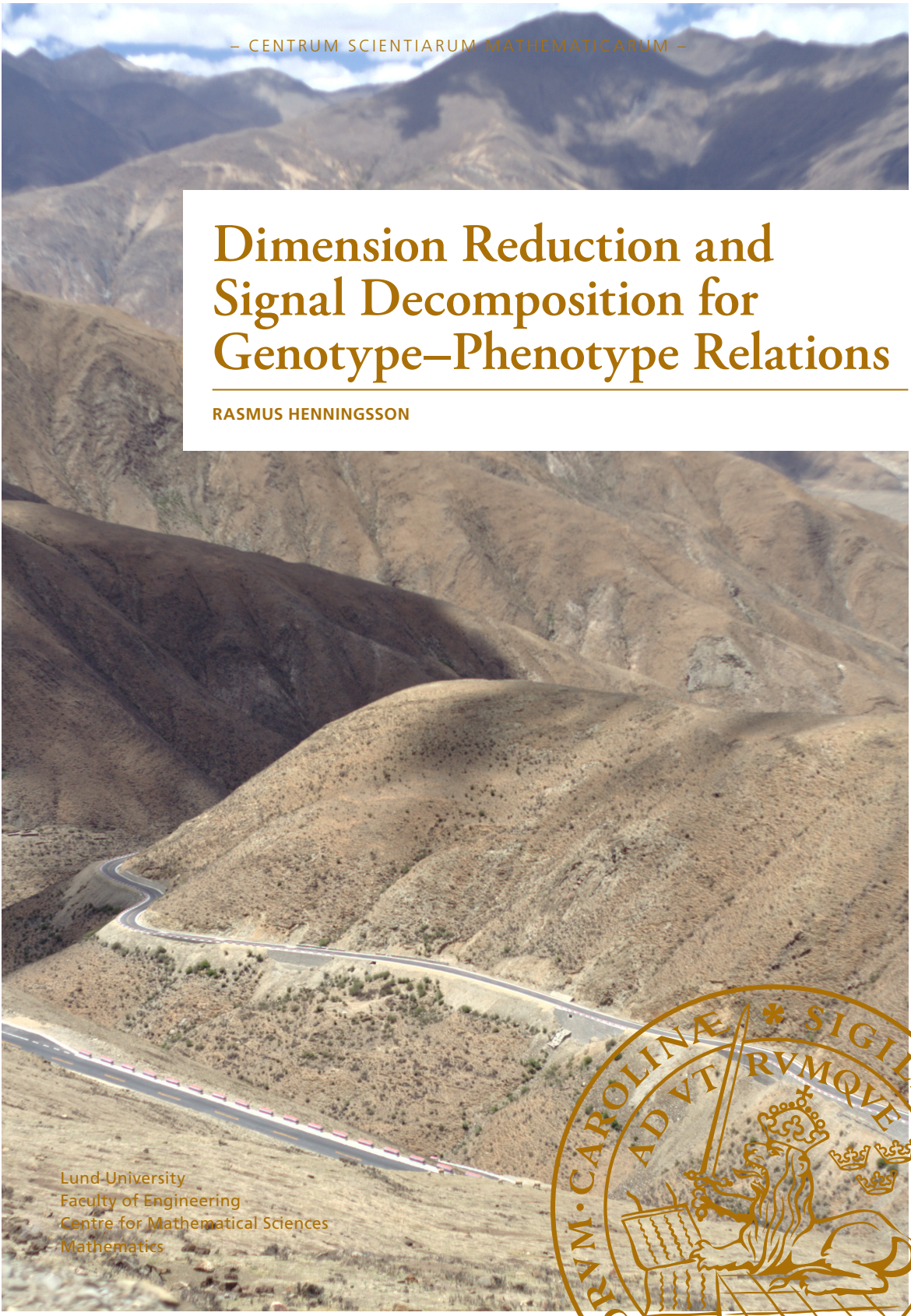
2017

Document Version:
Publisher's PDF, also known as Version of record

– CENTRUM SCIENTIARUM MATHEMATICARUM –

Dimension Reduction and Signal Decomposition for Genotype–Phenotype Relations

RASMUS HENNINGSSON



Lund University
Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Dimension Reduction and Signal Decomposition for Genotype–Phenotype Relations

Rasmus Henningsson

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Mathematics
Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden

http://www.maths.lth.se/

Doctoral Theses in Mathematical Sciences 2017:9
ISSN 1404-0034

ISBN 978-91-7753-479-2 (print)
ISBN 978-91-7753-480-8 (pdf)
LUTFMA-1065-2017

Printed in Sweden by MediaTryck, Lund 2017

ABSTRACT

Over the last few decades, DNA sequencing has developed from costing billions of dollars to get the complete sequence of the human genome, to being a routine procedure performed in labs all around the world. This has transformed the field of experimental biology since measurements can be done at a level of detail that was not possible before. Still, the relationship between genotype and low-level cellular processes on one hand, and high-level phenotypic traits on the other, tends to be very complex; measuring does not equal understanding. In the large data sets that are being gathered, it is often hard to uncover patterns that are truly meaningful, and not just arising by random chance.

In this work, we present novel methods for representing, exploring and visualizing genotype–phenotype data sets, with a particular focus on tracking changes driven by evolutionary processes as they occur. One challenge is to be able to quickly search for specific patterns in data coming from large genomes. We have adapted algorithms and data structures from the field of Information Retrieval, relying on inherent genomic structure to make efficient searches. In Paper I, we showcase these techniques with visualization of gene fusions in a study of paediatric B-cell precursor acute lymphoblastic leukaemia.

The complexity of biological processes, taken together with the fact that high-throughput measurements, such as DNA/RNA sequencing data, measure many different things at once, means that these data sets will often contain multiple overlaid signals. If data is collected in the field, rather than produced entirely under controlled

conditions in the lab, it is practically unavoidable. In Paper III, we present SMSSVD – SubMatrix Selection Singular Value Decomposition, a parameter-free unsupervised signal decomposition and dimension reduction method, particularly useful for data sets with many variables. By adaptively reducing the noise for each signal, SMSSVD creates a representation with many desirable properties inherited from the ordinary SVD, while being able to discover signals closer to the limit of detection.

In Paper II and Paper IV we describe models for representing genetically related but still heterogeneous microbial populations and show how the composition of the population determines the interaction with the host. The DISSEQT pipeline (DIStribution-based SEQuence space Time dynamics) developed in Paper IV, covers the entire workflow from read alignment to visualization of results. We model each population as a positive measure over sequence space and apply SMSSVD to get a robust representation. Using our model, we follow and visualize the evolutionary trajectories of the populations through time, highlighting important minority variants emerging. Finally, we demonstrate the relevance of our population model by showing that it can accurately predict the population fitness, whereas a model based on the consensus sequence fails.

POPULÄRVETENSKAPLIG SAMMANFATTNING

En av de grundläggande frågorna inom biologi är hur arvsmassan styr och påverkar biologiska processer. Att kartlägga dessa samband ger oss viktig kunskap och är ett led i utvecklingen av specifika behandlingar av så vitt skilda sjukdomar som olika virusinfektioner och cancerformer. Gemensamt för dessa två exempel är dock att de kännetecknas av evolution i 'miniformat'. Virus muterar väldigt snabbt, vilket skapar ett 'moln' av närbesläktade virus. I takt med att värdorganismens immunsystem upptäcker och attackerar virusen, sker så en kapplöpning, där vissa av virusvarianterna lyckas klara sig bättre och fortsätter att föröka sig. En cancer startar när en av kroppens egna celler drabbas av en eller flera mutationer som gör att den 'bryter sig loss' från det naturliga samspelet med andra celler och börjar dela sig i högre takt än normalt. Kroppen har olika skyddsmekanismer för att stoppa detta, men om mutationerna gör att cellen undviker dem, så växer cancern och fler och fler celler som delar de skadliga mutationerna skapas. I takt med att nya mutationer uppstår när cancern växer skapas en konkurrenssituation mellan de olika cellerna, där de som har egenskaper som gör att de kan föröka sig mer effektivt, gynnas på bekostnad av de andra.

De senaste decennierna har vi fått helt nya verktyg för att på bred front mäta vad som sker. Exempelvis kan vi nu med relativ enkelhet läsa av (sekvensera) en människas hela genom och representera det digitalt. För cancer innebär det att vi kan sekvensera tumörer för att karakterisera dem genetiskt och förstå cellpopulationens sammansättning. I vissa fall kan denna typ av mätningar vara avgörande för att ställa en precis diagnos och ge en fungerande behandling. På samma sätt har sekvenseringstekniken gjort det möjligt för oss att följa hur viruspopulationer utvecklas och i detalj följa evolutionen medan den sker.

Den här avhandlingen handlar om att ta fram metoder för att utforska, visualisera och i förlängningen förstå, de stora och komplexa datamängder som det handlar om. Ett återkommande problem är att det är svårt att se skogen för alla träd – det är helt enkelt för mycket data för att man ska kunna skapa sig en överblick om man försöker gå igenom den manuellt. Till den första artikeln som ligger till grund för avhandlingen utvecklade vi metoder för att söka efter angivna mönster i datamängder där de olika mätpunkterna motsvarar platser i ett genom. Vi visar också hur sökningarna kan användas för välja ut relevanta delar av genomet och visuellt åskådliggöra vad som skett där. Metoderna applicerades på data från patienter i en studie av fusionsgener (gener som uppstår när två helt olika gener slagits samman som följd av mutationer) i en viss typ av cancer.

I matematiken motsvarar varje egenskap vi mäter en dimension, och vi mäter idag ofta tusentals eller miljontals egenskaper, men för varje samband som finns mellan de olika egenskaperna reduceras den faktiska dimensionen, vilket möjliggör en rimlig överblick av insamlad data. Principalkomponentanalys är den klassiska metoden för att hitta sådana samband. Den är en mycket spridd och användbar metod för att förutsättningslöst leta efter mönster i stora datamängder. Hur den fungerar kan illustreras med ett exempel. Tänk er att vi håller en pepparkaksgris i handen och lyser på den med en ficklampa. Om vi håller grisen i rätt vinkel, kan vi från skuggan tydligt se grisens konturer, men vi kan också snurra på grisen så att skuggan inte blir mer än ett streck. Principalkomponentanalys hittar kort och gott den vinkel som bäst fångar objektets form. Korrespondensen mellan pepparkaksgrisen och dess skugga är inte perfekt, t.ex. syns inte eventuella ojämnheter på ytan i skuggan, men de huvudsakliga dragen kan fångas. I detta exempel har vi gått från ett tredimensionellt objekt till en tvådimensionell avbild, men matematiken bakom fungerar precis lika bra oavsett vilka dimensioner vi rör oss mellan. För de biologiska datamängder vi arbetar med är dimensionen väldigt mycket högre och principalkomponentanalys eller liknande metoder är ofta helt nödvändiga. Det går också att kvantifiera hur väl vi fångar objektets form beroende på hur många dimensioner vi behåller, vilket hjälper oss att förstå hur nära verkligheten den förenklade datamängden är.

Ett annat problem är att när vi gör många mätningar skapas det många oegentliga samband av ren slump. Om 10 000 personer slår fyra stycken 6-sidiga tärningar var, så kommer med stor sannolikhet flera av dem att slå fyra stycken 6:or*. Det betyder inte att dessa personer är ovanligt bra på att slå 6:or, utan bara

*I snitt fler än 7 stycken av dem.

att om vi gör något väldigt många gånger, så sker det osannolika ibland. I den tredje artikeln i avhandlingen har vi kombinerat denna observation med principalkomponentanalys för att skapa en ny metod som minskar störningarna från mätbruset. Vår metod är gjord för att hitta flera överlagrade 'signaler' i en datamängd och anpassar sig för att på bästa sätt minska bruspåverkan för varje enskild signal. Den är i första hand utvecklad för användning på biologiska datamängder, där vi ofta har många brusiga mätningar och olika oberoende effekter.

Att ett 'moln' av närbesläktade virus uppstår vid en virusinfektion har flera konsekvenser. De olika virusvarianterna konkurrerar med varandra, men det finns även exempel på när samarbete uppstår, vilket kan leda till att infektionen blir mer framgångsrik. Dessutom styr populationens sammanfattning vilka nya varianter som kan uppstå om ytterligare mutationer sker. I artikel två jämför vi olika virus som är fenotypiskt oskiljaktiga (kodar för identiska proteiner), men vars arvsmassa skiljer sig åt. De påverkas därmed på olika sätt av mutationer, vilket leder till att 'molnen' får olika sammansättning. Speciellt ser vi hur ett av virusen begränsas av sin arvsmassa och ger upphov till ett 'moln' med fler dysfunktionella virusvarianter.

Den fjärde artikeln vidareutvecklar grundidéerna från artikel två. Här tar vi fram allmängiltiga verktyg och metoder för att kunna analysera viruspopulationernas sammansättning och hur de utvecklas över tid. Baserat på sekvenseringsdata skapar vi en representation av varje viruspopulation och vilka varianter den innehåller. En mer robust modell skapas sedan genom användning av metoden från artikel tre med vilken inverkan av bruset kan minskas, utan att vi för den skull tar bort viktig information från datamängden. Modellen kan användas för att visualisera skillnader och likheter mellan olika viruspopulationer men också för att förutsäga egenskaper hos populationerna. Slutligen visar vi hur virusens fitness, deras förmåga att framgångsrikt fortplanta sig, beror på 'molnets' sammansättning.

PREFACE

The thesis is based on the four papers listed below.

H. Lilljebjörn, **R. Henningsson**, A. Hyrenius-Wittsten, L. Olsson, C. Orsmark-Pietras, S. von Palffy, M. Askmyr, M. Rissler, M. Schrappe, G. Cario, A. Castor, C. J. H. Pronk, M. Behrendtz, F. Mitelman, B. Johansson, K. Paulsson, A. K. Andersson, M. Fontes and T. Fioretos, "Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia", *Nature Communications* 7:11790 (2016).

G. Moratorio, **R. Henningsson**, C. Barbezange, L. Carrau, A. V. Bordería, H. Blanc, S. Beaucourt, E. Z. Poirier, T. Vallet, J. Boussier, B. C. Mounce, M. Fontes and M. Vignuzzi. "Attenuation of RNA viruses by redirecting their evolution in sequence space", *Nature Microbiology* 2, 17088 (2017).

R. Henningsson and M. Fontes, "SMSSVD – SubMatrix Selection Singular Value Decomposition", *ArXiv e-prints* (2017), arXiv:1710.08144 [stat.AP].

R. Henningsson, G. Moratorio, A. V. Bordería, M. Vignuzzi and M. Fontes, "DISSEQT – DIStribution based modeling of SEQuence Space Time dynamics", *In submission*.

The author has also contributed to the following papers.

R. Rodriguez-Roche, H. Blanc, A. V. Bordería, G. Díaz, **R. Henningsson**, D. Gonzalez, E. Santana, M. Alvarez, O. Castro, M. Fontes, M. Vignuzzi, and M.

G. Guzman, "Increasing clinical severity during a dengue virus type 3 Cuban epidemic: Deep sequencing of evolving viral populations", *Journal of Virology* 90.9 (2016): 4320-4333.

A. V. Bordería, O. Isakov, G. Moratorio, **R. Henningsson**, S. Agüera-González, L. Organtini, N. F. Gnädig, H. Blanc, A. Alcover, S. Hafenstein, M. Fontes, N. Shomron and M. Vignuzzi, "Group selection and contribution of minority variants during virus adaptation determines virus fitness and phenotype", *PLoS Pathogens* 11.5 (2015).

K. A. Stapleford, G. Moratorio, **R. Henningsson**, R. Chen, S. Matheus, A. Enfissi, D. Weissglas-Volkov, O. Isakov, H. Blanc, B. C. Mounce, M. Dupont-Rouzeyrol, N. Shomron, S. Weaver, M. Fontes, D. Rousset and M. Vignuzzi, "Whole-genome sequencing analysis from the chikungunya virus Caribbean outbreak reveals novel evolutionary genomic elements", *PLoS Neglected Tropical Diseases* 10.1 (2016).

ACKNOWLEDGEMENTS

During my PhD studies, I have had the luck to meet many wonderful people. People from different corners of the world and people with very different backgrounds. People from academia, but also people very distant from it. The impact they have had on me, is far greater than I would have imagined.

First, I would like to thank Magnus Fontes, my supervisor, for opening this world to me and for always providing a positive working environment. For every problem I faced during these years, ranging from mathematics to things entirely disparate, it was always easy to discuss with you and to find a path forward.

I want to express my thanks to all my brilliant colleagues at the Centre for Mathematical Sciences. Very few workplaces can offer discussions as interesting as the ones here, taking place during both working hours and breaks. Naming everyone I would like to thank would make for a very long list, but I must say that I have particularly appreciated the people in our usual lunch group, everyone at the 'fika' breaks and all active members of Lunds Matematiska Sällskap. My fellow PhD students have in many ways been my closest colleagues and among you all, I want to mention a few in particular, who really has meant a lot to me, Kerstin Johnsson, Sebastian Haner, Jiang Fangyuan, Jonas Wallin, Stefán Ingi Aðalbjörnsson, Viktor Larsson and Aleksis Pirinen.

Biology is complex and analyzing biological data is impossible without having the proper knowledge. At the Division of Clinical Genetics I was provided a gentle introduction to their field, by both Thoas Fioretos, who always took time to discuss with me, and Henrik Lilljebjörn, who could state everything in terms that made sense to someone with an engineering background. I have truly enjoyed working with you.

A substantial part of my PhD studies was spent at the Vignuzzi Lab at Institut Pasteur, Paris. Marco, thanks for making me a part of the lab. Gonzalo Moratorio,

we have had hours upon hours of fruitful scientific discussion, travelled the world to places warm and cold and watched countless football games, but most important is that I have gained a friend for life. A lab is nothing without its members, Stéphanie Beaucourt, Hervé Blanc, Enzo Poirier, Bryan Mounce, Kenny Stapleford, Lucia Carrau, Thomas Vallet, P.J. Hooikaas and everyone else I shared time with in the lab – I will miss you. I must also mention our friendly neighbors at the Saleh lab.

I was fortunate to be at Institut Pasteur at the same time as two amigos known as Gabriel Illanes and Jacob Bergstedt – the adventures will continue. Antonio Bordería has been involved in almost everything that I have taken part in at Pasteur and for this I am grateful. I'm lucky to have worked with you, J Boussier, Julie Le and the other members of IGDA.

I strive to make my research useful in practice and the feedback I have gotten from the people at Qlucore AB, most notably from Carl-Johan Ivarsson, Fredrik Nyberg, Mats Svensson and David Eklund, has helped me immensely. Together with everyone else at Qlucore, you create a great working atmosphere, and this is very important to me.

For long stretches of time, it has been difficult or downright impossible for me to find the time I wanted for family and close friends, but you have always been there when I needed you. I cannot describe how much this has meant to me. Whether I was in the need of discussing complicated work-related questions, or get away from them, I could always turn to my brother, Toivo.

When I started my PhD journey, I had no idea that it would take me to Paris. I certainly did not expect that by the end of it, I would feel at home in Paris, and even more so in Cixi and Shanghai. Yingdi, you are the reason I feel this way. Sharing life with you is so natural.

CONTENTS

INTRODUCTION

1 Genotype–Phenotype Maps

Long before the discoveries of DNA and genes, Gregor Mendel conducted a famous series of experiments showing how simple units of heredity can explain how certain traits are transferred from one generation to the next. He started with two varieties of pea plants that differed in one trait, for example one variant that had purple flowers and another that had white flowers. If self-pollinated, the offspring kept the same characteristic as the parent. He then created a new generation, denoted F_1, by cross-pollinating the plants such that every plant in F_1 had one parent with purple flowers and one with white flowers. All the pea plants in F_1 turned out to have purple flowers. However, when self-pollinating the plants in F_1, creating a new generation F_2, about three quarters of the plants had purple flowers and the rest had white flowers. Mendel then went on to explain this seemingly counter-intuitive result with a simple model of inheritance. Consider the variants (called *alleles*), B: purple flowers and b: white flowers. If every plant has two alleles, then the possible combinations (nowadays called the *genotype*) are BB, Bb and bb. In this case, B was *dominant*, meaning that plants with either BB or Bb had purple flowers while b was *recessive*, meaning that only plants with bb had white flowers. The quality of having either purple or white flowers is an example of what is called a phenotypic trait, or *phenotype* for short. When offspring is produced from two parents, then one allele is taken randomly, with equal chance, from each parent. The situation can thus be explained as follows. The original plants were either BB or bb, and self-pollination does not change this. However, with one parent from each group, the F_1 plants were all Bb. Hence they had purple flowers, but with a different genotype than the original plants with purple flowers. Finally, all genotypes are possible in the F_2 plants, whose parents are all Bb. Figure 1 shows

the possible outcomes. Since it's equally likely to inherit either allele from each parent, the likelihood for each genotype in F_2 is BB (25%), Bb (50%) and bb (25%), giving 75% chance of purple flowers and 25% chance of white flowers.

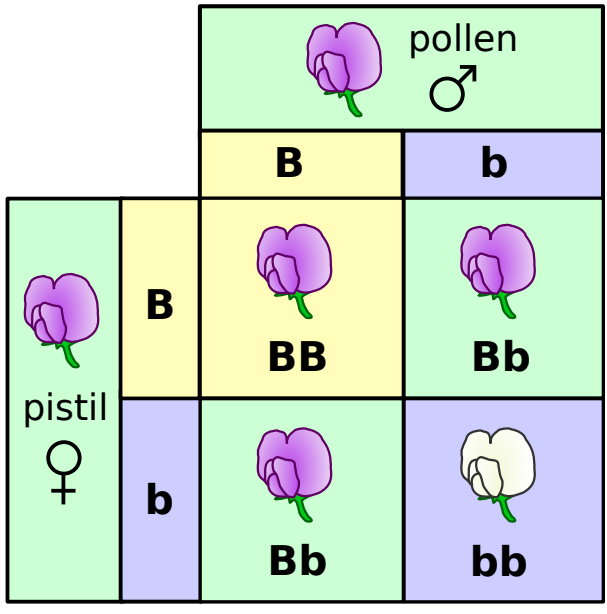


Figure 1: Possible genotypes and phenotypes of pea plants when both parents have the genotype Bb. Illustration by Madeleine Price Ball, CC0 license.

The situation described above is the canonical example of a *genotype–phenotype map*[1]. It highlights one of the most important properties of the map, it is not one-to-one; multiple genotypes result in the same phenotype. Depending on the context, the genotype can refer to the entire genetic makeup of an organism, the alleles present at a single loci, or anything in between. In a very broad definition, anything that is a consequence of the genotype can be considered a phenotype. Genotype–phenotype maps are great tools for reasoning about genetic causes for phenotypic traits, but it's necessary to understand that the relationship is highly complex behind the scenes. Epigenetics (heritable traits that are not encoded by DNA/RNA), gene regulation, environmental factors and randomness can all affect phenotype and the interactions are very different from case to case. As an extreme but illustrative example, the sex in Alligator mississippiensis is determined by the temperature of egg incubation[2]. Even in the deterministic case, the genotype–

phenotype mapping can be intricate. For instance, a phenotype might be affected by alleles at multiple *loci* (genomic positions), the interaction can take other forms than dominant/recessive, and different species can have different *ploidy* (the number of copies of each chromosome, which determines the number of alleles per loci in an individual).

A practical view of the genotype–phenotype map is often to consider multiple levels, or rather a network, of phenotypes. The first levels are directly related to cell processes. Starting from the genes that are expressed, the initial phenotypes are the (amounts of) different kinds of RNA, such as *transfer RNA* (tRNA) and *messenger RNA* (mRNA). mRNAs are in turn *translated*, creating proteins, which can also be viewed as phenotypes. The cell machinery consists of a complex network of regulatory systems with feedback loops, where the state and dynamics of these systems can be considered phenotypes, since they depend on the genotype. Through inter-cellular interactions, phenotypes form higher level networks, finally reaching phenotypes that are traits observable by unaided human senses, e.g. the color of a flower or the shape of limbs. Stable phenotypes are possible through robustness in the regulatory systems, and is crucial for biological life. (On the other hand, complete robustness would not result in any phenotypic variability at all, and thus no differences for selection to act upon.) The robustness makes it possible to, sometimes, consider the function that maps genotype to phenotype as a black box, knowing the input is enough to predict the output. In other cases, measuring both genotypes and lower-level phenotypes will make it easier to accurately predict higher-level phenotypes, since part of the uncertainty of the map can be removed.

2 Evolution

Evolution is a process that will take place if just a few basic prerequisites are met, and that can create complexity out of simplicity. It acts upon a population, where differences between individuals are inherited from one generation to the next and there is a source for new variations to appear. Furthermore, the differences must affect the ability for the individuals to reproduce and have successful offspring, a process called *selection*. In practice, selection often acts indirectly and requires survival and competing with the other individuals in the population for common resources. With these conditions in place, the population will evolve over time, as traits from individuals that reproduce successfully will spread throughout the

population. *Adaptation* occurs when a population becomes better suited for the current environment by evolving. It is not a necessary consequence of evolution, for instance, if the environment is harsh enough, the population might perish. *Fitness* is the abstract quality of an individual being successful in evolution, that is, to be able to reproduce and have offspring that is also able to reproduce and so on. It is often defined relative to the other individuals in a population.

There is nothing in the above description that limits evolution to the realms of biology. Evolutionary computing[3] applies evolutionary processes to problems as a kind of structured and automated trial and error. In brief, candidate solutions are assigned fitness values according to their ability to solve some kind of problem. New candidate solutions are created from previous successful ones and new variation is introduced randomly. Dawkins applied evolutionary theory to *memes* (ideas, behavior, etc.)[4], giving a framework to explain how some memes spread and change, while others do not.

In biology, the selective process is referred to as *natural selection*. When humans intervene by selecting which individuals in a population will reproduce, and thus choose the traits that will be passed on to future generations, it is called *artificial selection*. Since the beginning of agriculture, humans have adapted plants to suit our needs in this way. Animal breeding is another example.

Selection acts on the phenotype. Variation, inherited as genotypes (and epigenetic factors) are passed on, and in many cases recombined, from one generation to the next. Random mutations introduce new variation and ensures that the evolutionary process can continue. Even beneficial mutations are not guaranteed to spread and become common in a population[5]. At the individual level, randomness has a huge impact on selection, but on the population level, individual traits are better viewed as factors that affect the probability that an individual will reproduce.

2.1 Viruses

Viruses are biological entities that may or may not be considered as life, since they are unable to reproduce without a host cell. Alive or not, they are clearly subjected to evolution. Viruses are in fact great tools for studying evolution, since they tend to have short generation times, high mutation frequency and small genomes. Most of all, this is true for a group of viruses called positive-sense single-stranded RNA, or (+)ssRNA, viruses. As the name implies, their genomes are single-stranded RNA molecules. RNA molecules are less stable than their DNA counterparts and

viruses lack several error-correcting mechanisms that are present in other organisms, thus increasing mutation rates. Furthermore, their genomes also act as mRNA, and hence they can be translated directly into proteins by the cell machinery, without any intermediate steps. Taken together, these factors mean that the viruses can evolve rapidly, and their small genomes simplify data collection.

The quasispecies model[6], which is widely used for viruses, describes what happens in a population with asexual replication and high mutation rates. The key to the quasispecies model is the balance between selection and mutation. Imagine a population consisting of identical viruses, with a certain replication rate (fitness). During replication, mutations might occur, producing viruses that are not identical, but closely related to, the original. Now assume that these new viruses have a lower replication rate than the original viruses. What will happen as time passes? If the relative replication rate of the other viruses is low, and the mutation rate is low as well, we can expect that the vast majority of the population will be identical to the original viruses. A few viruses with lower fitness will appear, but since they replicate slowly, they will never reach any sizeable quantity. If instead the mutation rate is high, there will be a lot of low fitness viruses at any given point in time, forming a "cloud" around the central high fitness viruses. In the extreme, the high fitness viruses will not be produced at a rate high enough to sustain the population, ultimately leading to extinction. Quasispecies theory mainly deals with the scenario in-between, closely related viruses creating a heterogeneous cloud, a common situation with direct biological consequences. For instance, it allows viruses to adapt quicker to changes in the environment, since genotypic and phenotypic variety is already present in the cloud[7]. Another example is how interaction between viruses with genetic differences can determine pathogenesis[8]. Hence, we have every reason to believe that although a high mutation rate is a necessity of the chemistry of RNA replication, it is also an adaptation in itself, as the clouds formed can be critical to the survival of the viruses. In some sense, the size and shape of the quasispecies cloud can be thought of as a phenotype.

2.2 Cancer

In humans and other multicellular organisms that reproduce sexually, the ability of a single cell to spread its genes is not directly linked to the replication rate of the cell. Instead, the genes will spread if the individual the cell belongs to is able to reproduce. In this regime, cooperation between cells has been evolutionary successful, they are not in competition with each other, but will succeed or fail

together. One example of the cooperation is *apoptosis*, programmed cell death; a single cell sacrifices itself for the benefit of the organism as a whole. Apoptosis is an important process during embryonic development, and later in life to target cells that might become cancerous and to combat virus infections, to give two examples.

Cancer begins with a cell acquiring mutations that make it replicate faster and avoid mechanisms such as apoptosis[9]. Through cell division, a population of cells all sharing the original mutations is created. Diversity within the population is created as new mutations are introduced during cell division. The cells will compete with each other (and the non-cancerous cells) for resources, and successful cells will reproduce faster than the others. As we can see, the prerequisites for evolution are there[10]. A cancer is not beneficial for the host, and it's an evolutionary dead end – although technically possible, a cancer is extremely unlikely to spread outside of its host. That cancers still appear is just a consequence of the prerequisites for evolution being met.

Understanding cancer as an evolutionary process is one of the keys to a successful treatment. If a cancer patient is subjected to treatment, and a subpopulation of the cancer cells are more resistant to the treatment, then those cells will be favored by selection. Certain human genes, that normally have important functions, are more likely than other genes to cause cancer after a mutation has occurred. Around 140 such genes, called *oncogenes*, are known and a single tumor often have mutations in several of the oncogenes[11], which shows that there is some predictability to cancer. From similar starting points, the evolutionary processes driving cancer development can lead the cancers in different patients down similar paths. Knowing the genotype, and phenotypes such as gene expression levels, of a cancer, are thus critical for characterization and to find the appropriate treatment.

3 Measurements and Data

Analyzing genomic data is challenging. In part, because the data sets tend to be very large. The major obstacle is however that all genomic data needs to be interpreted in context, but figuring out the relevant context can be extremely difficult. The information carried by a genome is naturally described by one sequence of nucleotides per chromosome. (This is true for all life on Earth, including viruses, even though the organization of the genetic material could use other molecules than chromosomes.) Changes as small as the substitution of a single nucleotide

with another can have drastically different effects depending on where it happens. Many times, there are no observable phenotypic effects, but a change resulting in an amino acid change in a gene will for instance alter the protein coded by the gene, again with very different effects depending on the specific change and the role of the protein.

Genome sequencing [12] is an intricate process that makes it possible to digitalize information from DNA (or RNA) molecules and thus represent and store them in a computer. Due to technical limitations, the DNA molecules are fragmented into shorter pieces, suitable for the sequencers. The computer representation of a sequenced piece of DNA is called a *read*. After sequencing, the best matching location in a reference genome is found for each read, a process called *alignment*. This works well in general since the differences between individuals of the same species are few enough to make the problem tractable. The use of a reference genome also establishes a coordinate system for the genome and thus a basis for comparison between different samples. If no reference genome exists, *de novo* assembly can be used to piece the reads together, but the process is significantly more difficult than alignment, even more so when the reads are short. It is necessary to cover every part of the sequence multiple times, for several reasons. Since humans have two non-identical copies of each chromosome, data is needed from both to completely characterize the genotype. Sequencing is error-prone; gathering multiple independent measurements for each location is required for a statistically sound analysis. If a population (e.g. of viruses or of cancer cells) is sequenced, subpopulations can only be discovered and studied if there is enough data to infer the frequencies of different variants, and distinguish them from noise. Another examples is gene expression analysis by mRNA sequencing, where the number of reads in each gene is related to the expression level of the gene.

The human genome has a total length of over 3 billion nucleotides, many orders of magnitude larger than what is feasible to go through manually. The known structure of the genome, apart from being necessary for interpretation, can help narrowing down searches for different patterns to relevant regions. Many annotations, such as the location of genes, are publically available. A gene corresponds to an interval in the reference genome, and the exons (the parts of a gene coding for proteins) are intervals within the gene. We will now show how to efficiently search and make basic set-operations on huge collections of intervals. To our knowledge, this has not been described before.

4 Information Retrieval

The field of information retrieval deals with searching efficiently in large collections of data[13]. Most often, it is used for searching for text and phrases in a collection of documents (e.g. web pages), but the general principles can also be applied elsewhere. We will here adapt and expand some of the basic tools of (Boolean) information retrieval to the setting of searching in genomic data, relying on inherent genomic structure to make efficient searches.

A typical query in Boolean information retrieval is: find all documents containing both 'Paris' and 'Tourism'. The solution to this problem can be found by taking the intersection of two sorted lists of integers, given that the data is stored in a structured way. Assign a unique integer ID to each document, and assume that we have sorted lists of IDs of all documents containing 'Paris' and 'Tourism' respectively. The IDs of the documents containing both search terms are thus found by intersecting the lists. There are multiple algorithms for solving this problem, but the key insight is that not every element in the two lists must to be checked. As illustrated in Figure 2, if we can quickly find the first item in a list that is larger than or equal to a given integer, then we can skip over large stretches of document indices. In practice, this can be achieved by a binary search (half-interval search), or by augmenting the document lists with *skip pointers* that store information about what lies some steps ahead in the list. If one list is much shorter than the other, large portions of the longer list can be skipped.

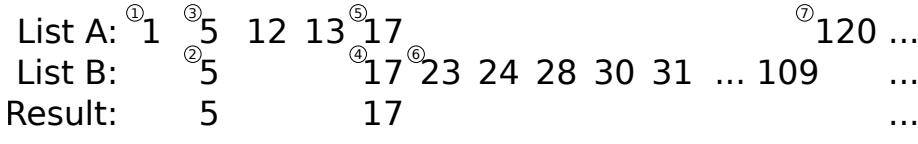


Figure 2: Illustration of a possible search scheme when intersecting two sorted lists of integers. Searches are performed by alternating between the two lists, looking for the first item in greater than or equal to the current element, until a match is found. The process is then repeated. The small circled numbers indicate the order in which the list elements are visited.

Searching in a genome is quite similar, but instead of searching through documents, we search among the coordinates that are given by the reference genome. The main difference is the relevance of proximity. Genomic features are often described by intervals (e.g. genes, promotor regions and CpG islands) or by unions of intervals (e.g. all exons in a gene and all genes in a specific pathway). Even single nucleotide variants that per definition affect a single position, must be in-

terpreted in the context of the surrounding genomic features. Hence, representing general features as sorted lists of intervals yields a natural and compact representation, in comparison to using sorted lists of integers, even if they mathematically can describe the same sets. To keep notation simple, we are here only considering genomes consisting of a single sequence (chromosome), but all concepts extend easily to a collection of sequences. We begin by noting that any subset A of the interval $[1, N]$ (the coordinates of a genomic sequence of length N) can be described uniquely by a sorted list of intervals. That is

$$A = \bigcup_{i=1}^{n} [a_i, b_i),$$

where $a_1 < b_1 < a_2 < b_2 < \ldots < a_n < b_n$. To adapt the algorithm outlined above to intervals, we will restate the search for the next item slightly, to give interval results. Given x, let

$$k \coloneqq \min\{i : \ x < b_i\}$$

then the first interval of A after x is given by

$$s(A, x) \coloneqq \begin{cases} [\max(a_k, x), b_k) & \text{if } k \text{ exists} \\ \emptyset & \text{otherwise} \end{cases}.$$

This is an extension of the case described above; if $[a, b) = s(A, x)$, then a is the first integer in A that is greater than or equal to x.

4.1 Searching in Intersections

Like before, searching in a sorted list of intervals can be performed by a binary search or using skip pointers. Algorithm 1 and Figure 3 shows how searching the intersection of A and B can be stated in terms of searches in A and B. The intersection of A and B can be computed by iteratively searching until the end has been reached. Searching in the union of A and B can implemented in a similar fashion, or be stated in terms of intersections and complements, since $A \cup B = (A^{\complement} \cap B^{\complement})^{\complement}$ and $s(A^{\complement}, x)$ is trivial to implement in terms of $s(A, x)$.

There is no need for A or B to be explicitly known in Algorithm 1. Composite queries, such as $(A \cup B) \cap C$ can be handled by searching in $A \cup B$ and C, where the former in turn requires searching in A and B. In this example, if C have few and short intervals, only small parts of $A \cup B$ are actually needed in the evaluation of $(A \cup B) \cap C$.

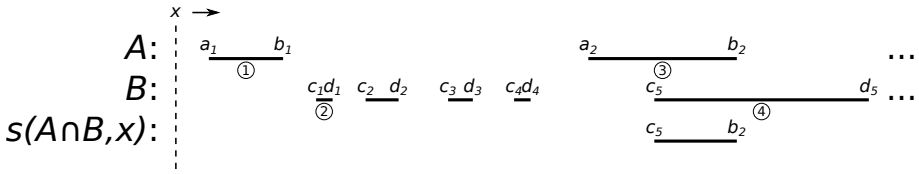


Figure 3: Illustration of the algorithm for searching in the intersection of two lists of intervals, A and B, starting at x and going right. The small circles indicates the order in which intervals are visited. Note how several intervals in B are skipped.

Algorithm 1 $s(A \cap B, x)$

Input: $A, B \subset [1, N]$, $x \in [1, N]$
Output: First interval in $A \cap B$ after x.

```
[a, b) ← s(A, x)
while [a, b) ≠ ∅ do
    [c, d) ← s(B, a)
    if [a, b) ∩ [c, d) ≠ ∅ then
        return [a, b) ∩ [c, d)
    end if
    [a, b) ← [c, d)
    swap(A, B)
end while
return ∅
```

4.2 Searching in Functions

The read coverage, the number of aligned reads covering each position in the reference genome, is one important factor to take into account when interpreting genomic data. In many cases, regions with low coverage are not of interest since they provide no reliable information. In exome sequencing, a technique for sequencing only the protein-coding parts of genes, the vast majority of the genome will have no or very few reads. Similarly, for mRNA sequencing, reads will also be aligned to the protein-coding parts of genes, but in this case the read coverage will be related to the expression level of the gene. A simple criteria for finding relevant parts of a genome is thus to search for regions where the read coverage is above a certain threshold.

A sketch of a data structure making efficient searches possible, in terms of both storage space and computation time, is shown in Figure 4. The underlying

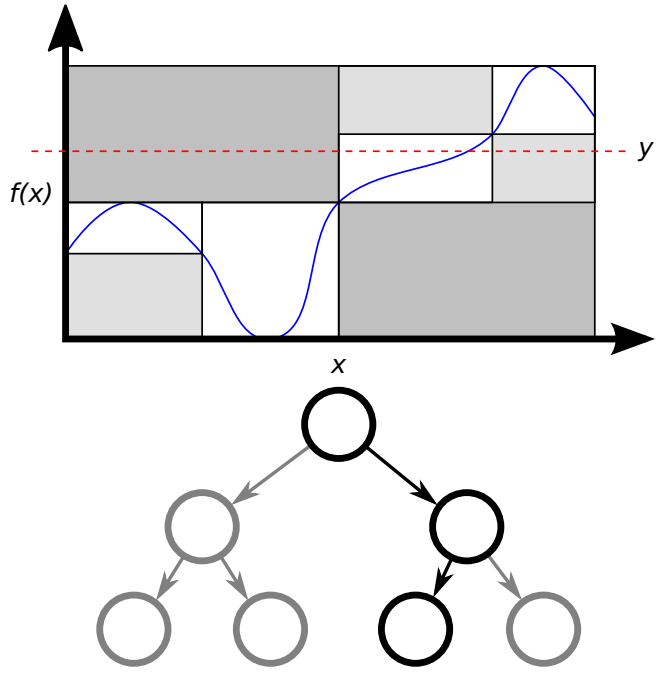


Figure 4: Sketch of a binary tree representing a function f (blue). The root node represents the entire function and is split into two child nodes at the vertical line at the center of the figure. Child nodes are split recursively (only three levels of the binary tree are shown). Since the minimum and maximum values of the function are stored for each node, entire nodes can be skipping during a search. Nodes that are visited when searching for the first point where the function value $f(x)$ if larger than a threshold y (red dashed line), are highlighted in black. The nodes that are skipped are shown in gray, and corresponding areas in the function graph are also gray.

assumption is that the read coverage tend to change quite slowly (the difference in read coverage from one position to the next is generally of low magnitude). The data is stored hierarchically in a binary tree. In a node $\mathcal{N}$, representing the interval $[\mathcal{N}_a, \mathcal{N}_b)$, we store the minimum value $\mathcal{N}_m$ and maximum value $\mathcal{N}_M$ of a function f (e.g. the read coverage) in that interval. The child nodes of $\mathcal{N}$, $\text{left}(\mathcal{N})$ and $\text{right}(\mathcal{N})$, correspond to the intervals $[\mathcal{N}_a, x)$ and $[x, \mathcal{N}_b)$ respectively, for some x such that $\mathcal{N}_a < x < \mathcal{N}_b$. The root node $\mathcal{R}$ represents the entire sequence $[1, N]$. Let

$$g(\mathcal{N}, y, x) \coloneqq \min\{z \in [\mathcal{N}_a, \mathcal{N}_b) : \ z \geq x \text{ and } f(z) > y\}$$

Algorithm 2 $g(\mathcal{P}, y, x)$

Input: $\mathcal{P}$ (Parent node), $y \in \mathbb{Z}$, $x \in [1, N]$
Output: First point $z \geq x$ such that $f(z) > y$.

```
C ← left(P)
if x < C_b and y < C_M then
    if y < C_m then
        return max(x, C_a)
    end if
    z ← g(C, y, x)
    if z ≠ ∞ then
        return z
    end if
end if
C ← right(P)
if x < C_b and y < C_M then
    if y < C_m then
        return max(x, C_a)
    end if
    z ← g(C, y, x)
    if z ≠ ∞ then
        return z
    end if
end if
return ∞
```

or ∞ if the minimum doesn't exist. That is, $g(\mathcal{R}, y, x)$ is the first point $z \geq x$ where the function value is larger than a threshold y. Algorithm 2 describes how g can be evaluated, using the binary tree. Similarly, l is defined as the first point where the function is less than or equal to y and the algorithm is practically identical. Searching for an interval $[a, b) = s(\{z : \ f(z) > y\}, x)$ is now done in two steps. First compute $a = g(\mathcal{R}, y, x)$ and then $b = l(\mathcal{R}, y, a)$. If $a = \infty$, there are no more intervals to be found and if $b = \infty$, the interval reaches the end of the sequence. The algorithmic complexity of searching for the next interval is logarithmic in the number of nodes in the binary tree. In practice, there is no need to subdivide nodes all the way until $\mathcal{N}_m = \mathcal{N}_M$. Instead, leaf nodes can store a

compact representation of the function on a short interval. Since the majority of nodes in a binary tree are leaf nodes, this can greatly reduce storage requirements, while still keeping the benefits of the search structure.

4.3 Searching in Dilated Sets

The search methods described above exploits the fact that neighboring genomic coordinates are functionally related. The neighborhood can also be explicitly used when formulating searches, e.g. to find genomic features that are close to each other and to give context around search results. Dilation (also known as Minkowski addition) is a fundamental operation in mathematical morphology that can be used to make a set expand into its neighborhood. In general, the dilation of A and B is defined as the set $\{a + b : \ a \in A, b \in B\}$, but we are only interested in the case when $B = [-d, d]$, a symmetric interval centered at the origin. The dilation of A and $[-d, d]$ will be denoted by $\mathrm{dilate}(A, d)$ and contains all points that are at distance less than or equal to d from some point in A, effectively growing A by d units. In particular, dilation by d units will make any 'gap' in A of length $2d$ or shorter disappear.

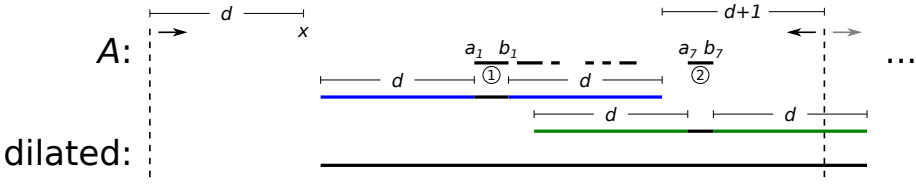


Figure 5: Illustration of the algorithm for searching in a dilated set. After finding the interval $[a_1, b_1)$ and extending it to $[a_1 - d, b_1 + d)$ by dilation (shown in blue), a backward search is performed to find the last interval in A that would connect with the first interval after dilation. The dilation of $[a_7, b_7)$ is shown in green and the dilation of A is shown in black at the bottom. Note how the contents of A between a_1 and b_7 will not affect the result, and that this region can thus be skipped when searching. The process is then repeated, a new forward search is performed (gray arrow), to see if the result should be extended further.

If d is small, such that the number of intervals in $\mathrm{dilate}(A, d)$ is the same as in A, then the algorithmic complexity of computing the dilation cannot be sublinear in the number of intervals. However, if d is larger, then not every interval of A must be visited, as shown in Figure 5. Since intervals that are no more than $2d$ apart will connect, the first search is followed by a backward search to find the last such interval, thus skipping any intervals in-between. Both forward and backward searches are necessary. Forward searches deal efficiently with intervals in A that

Algorithm 3 $s(\text{dilate}(A, d), x)$

Input: $\mathcal{A} \subset [1, N]$, $d \in \mathbb{N}$, $x \in [1, N]$
Output: First interval in $\text{dilate}(A, d)$ after x.

```
[a, b) ← dilate(s(A, x − d), d) ∩ [x, ∞)
if [a, b) = ∅ then
    return ∅
end if
loop
    z ← b + d + 1
    [u, v) ← dilate(r(A, z), d)
    if v = b then
        return [a, b) {Couldn't extend by backward search.}
    end if
    b ← v
    [u, v) ← dilate(s(A, z), d)
    if [u, v) = ∅ or u > b then
        return [a, b) {Couldn't extend by forward search.}
    end if
    b ← v
end loop
```

are much longer than d and the backward searches make it possible to skip many short intervals at a time. A complete implementation is given by Algorithm 3. In the implementation, note that the dilation of an empty set is empty, and that $\text{dilate}([a, b), d) = [a - d, b + d)$. The algorithm also relies on a reverse search $r(A, x)$ defined analogously to $s(A, x)$. Let

$$k \coloneqq \max\{i : \ x > a_i\}$$

then the last interval of A before x is given by

$$r(A, x) \coloneqq \begin{cases} [a_k, \min(x, b_k)) & \text{if } k \text{ exists} \\ \emptyset & \text{otherwise} \end{cases} .$$

5 Dimension Reduction

High throughput biological measurements are commonplace today in the study of genotypes and phenotypes. Often, the data can be put in form of a large data matrix X, where the element x_{ij} at row i and column j corresponds to a measurement of variable i for sample j. Examples of data sources include gene expression measurements using arrays or mRNA sequencing, metagenomics (sequencing of extremely diverse populations, e.g. gut microbiota and ecological samples)[14] and sequencing of populations of closely related organisms. After applying adequate normalization methods, such as TMM[15] and rlog[16] for mRNA sequencing data, assuming that the transformed variables are normally distributed is generally a reasonable assumption.

High throughput techniques make it possible to measure many variables in parallel. Since sample collection is often difficult and time-consuming, the number of variables will in general vastly outnumber the number of samples, which has both benefits and drawbacks. By measuring broadly, there is no need to tailor the measurements for a specific experiment, standard protocols can be followed. It also opens up for exploratory analysis, since there is no need to decide beforehand which variables are important. On the other hand, we can expect that most of the variables will only contribute to noise, making it easier to find spurious patterns. Furthermore, it complicates hypothesis testing[17].

5.1 Singular Value Decomposition

The Singular Value Decomposition (SVD) is an extremely useful tool in the study of matrices. It aids understanding by decomposing a matrix into three parts, accessible to interpretation and it can be used to create an optimal approximation of a matrix by a matrix of lower rank (useful for e.g. noise reduction). Using the SVD, any matrix X can be factorized as $X = U\Sigma V^T$, where $U^TU = I$, $V^TV = I$ and Σ is a diagonal matrix, with positive elements on the diagonal, ordered from largest to smallest. The diagonal elements of Σ are known as the (nonzero) *singular values* of X. Below, we will give a constructive proof, showing existence of the SVD, that can help understand the properties of SVD. For an alternative proof, see Golub and van Loan[18].

Definition 5.1. The *induced 2-norm* of a matrix X is

$$\|X\|_2 \coloneqq \sup_{\|\boldsymbol{v}\|_2=1} \|X\boldsymbol{v}\|_2 = \sup_{\boldsymbol{v}\neq 0} \frac{\|X\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2}. \tag{1}$$

Using that equality in the Cauchy-Schwarz inequality $\boldsymbol{x}^T\boldsymbol{y} \leq \|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2$ occurs iff $\boldsymbol{x}$ and $\boldsymbol{y}$ are equal (up to multiplication with a nonnegative scalar), we get the equivalent formulation

$$\|X\|_2 = \sup_{\boldsymbol{u},\boldsymbol{v}\neq 0} \frac{\boldsymbol{u}^T X\boldsymbol{v}}{\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2}. \tag{2}$$

(Consider the case $\boldsymbol{u} = X\boldsymbol{v}$.) It follows directly that the definition above is symmetric, $\|X\|_2 = \|X^T\|_2$.

Lemma 5.1. *If X is nonzero, there exists $\boldsymbol{v}$ that maximizes Equation 1 and any $\boldsymbol{v}$ that maximizes Equation 1 is orthogonal to the kernel of X.*

Proof. $\|X\boldsymbol{v}\|_2$ is a continuous function in $\boldsymbol{v}$ and $\|\boldsymbol{v}\|_2 = 1$ defines a compact set. Hence, there exists $\boldsymbol{v}$ that maximizes Equation 1. Now assume $\boldsymbol{v}$ is a maximizer of Equation 1. There exists a unique decomposition $\boldsymbol{v} = \boldsymbol{v}' + \boldsymbol{v}''$ such that $\boldsymbol{v}'$ is in the row space of X and $\boldsymbol{v}''$ is in the kernel of X. It follows that $\|X\boldsymbol{v}\|_2 = \|X\boldsymbol{v}' + X\boldsymbol{v}''\|_2 = \|X\boldsymbol{v}'\|_2$. Furthermore, $\|\boldsymbol{v}'\|_2 \leq \|\boldsymbol{v}' + \boldsymbol{v}''\|_2$ with equality iff $\boldsymbol{v}'' = 0$. Hence Equation 1 can only be maximized if $\boldsymbol{v}'' = 0$. □

Because of symmetry, it follows immediately that if $\boldsymbol{u}, \boldsymbol{v}$ maximizes Equation 2, then $\boldsymbol{u}$ is orthogonal to the cokernel of X.

Theorem 5.1. *Let $X \in \mathbb{R}^{P\times N}$ be a matrix of rank r. Then there exists $U \in \mathbb{R}^{P\times r}$, $\Sigma \in \mathbb{R}^{r\times r}$ and $V \in \mathbb{R}^{N\times r}$ such that $U^TU = I_r = V^TV$, Σ is a diagonal matrix with entries $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ and*

$$X = U\Sigma V^T.$$

Proof. Start with $X_1 \coloneqq X$ and take any $\boldsymbol{u}_1, \boldsymbol{v}_1$ that maximize Equation 2 for X_1. Without restriction, we can assume that $\|\boldsymbol{u}_1\|_2 = \|\boldsymbol{v}_1\|_2 = 1$. Let $\sigma_1 \coloneqq \|X_1\|_2 = \boldsymbol{u}_1^T X_1 \boldsymbol{v}_1$ and form

$$X_2 = X_1 - \boldsymbol{u}_1\sigma_1\boldsymbol{v}_1^T.$$

We will show that $r = \operatorname{rank}(X_1) = 1 + \operatorname{rank}(X_2)$, and several orthogonality properties. Let $\mathcal{Z} \coloneqq \{\boldsymbol{z}_i\}_{i=1}^{N-r}$ be a (possible empty) orthonormal basis of the kernel of X_1. For any $\boldsymbol{z} \in \mathcal{Z}$,

$$X_2\boldsymbol{z} = X_1\boldsymbol{z} - \boldsymbol{u}_1\sigma_1\boldsymbol{v}_1^T\boldsymbol{z} = 0 - 0 = 0,$$

since $\boldsymbol{v}_1$ is orthogonal to $\boldsymbol{z}$ by Lemma 5.1. From the lemma it also follows that $\boldsymbol{v}_1$ is in the row space of X_1 and hence there exists a set $\mathcal{W} \coloneqq \{\boldsymbol{w}_i\}_{i=1}^{r-1}$ that together with $\boldsymbol{v}_1$ form an orthonormal basis of the row space of X_1. For any $\boldsymbol{w} \in \mathcal{W}$

$$X_2\boldsymbol{w} = X_1\boldsymbol{w} - \boldsymbol{u}_1\sigma_1\boldsymbol{v}_1^T\boldsymbol{w} = X_1\boldsymbol{w} - 0 = X_1\boldsymbol{w},$$

since $\boldsymbol{v}_1$ is orthogonal to $\boldsymbol{w}$ by construction. Finally,

$$X_2\boldsymbol{v}_1 = X_1\boldsymbol{v}_1 - \boldsymbol{u}_1\sigma_1\boldsymbol{v}_1^T\boldsymbol{v}_1 = \boldsymbol{u}_1\sigma_1 - \boldsymbol{u}_1\sigma_1 = 0,$$

where $X_1\boldsymbol{v}_1 = \boldsymbol{u}_1\sigma_1$ follows from $\boldsymbol{u}_1 \propto X_1\boldsymbol{v}_1$ and $\sigma_1 = \boldsymbol{u}_1^T X_1 \boldsymbol{v}_1$.

That is, the elements of $\mathcal{Z}$ are in the kernel of both X_1 and X_2 and the elements of $\mathcal{W}$ are in the row space of both X_1 and X_2. To complete the orthonormal bases, we need to add $\boldsymbol{v}_1$ to the basis for the row space of X_1 and $\boldsymbol{v}_1$ to the basis of the kernel of X_2. This shows that $r = \operatorname{rank}(X_1) = 1 + \operatorname{rank}(X_2)$, since there is one more element in the basis of the row space for X_1.

We now repeat the process, forming $\boldsymbol{u}_2, \sigma_2, \boldsymbol{v}_2$ etc., until reaching X_{r+1} that has rank 0 and thus is zero. Note that $\boldsymbol{v}_i$ is in the kernel of every X_j such that $j > i$, showing that the all the $\boldsymbol{v}_i$'s are orthogonal. By symmetry, the $\boldsymbol{u}_i$'s are also orthogonal.

We have thus shown that

$$X = \sum_{i=1}^{r} \boldsymbol{u}_i\sigma_i\boldsymbol{v}_i^T = U\Sigma V^T,$$

where

$$\begin{aligned} U &\coloneqq \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_r \end{pmatrix}, \\ \Sigma &\coloneqq \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{pmatrix}, \\ V &\coloneqq \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_r \end{pmatrix}, \end{aligned}$$

and the columns of U and V are orthogonal, i.e. $U^TU = I_r = V^TV$.

Finally, assume that $\sigma_{i+1} > \sigma_i$ for some i. Then

$$\|X_i\|_2 \geq \boldsymbol{u}_{i+1}^T X_i \boldsymbol{v}_{i+1} = \boldsymbol{u}_{i+1}^T \left(X_{i+1} + \boldsymbol{u}_i \sigma_i \boldsymbol{v}_i^T\right) \boldsymbol{v}_{i+1} = \sigma_{i+1} > \sigma_i$$

which is a contradiction.

□

If the singular values are all different from each other, which they will be in practice if the data matrix has been subjected to noise, then the decomposition produced in the proof above is unique up to a common change in sign for each pair $\boldsymbol{u}_i, \boldsymbol{v}_i$. Indeed, it can be shown that this is the only possible decomposition of X into $U\Sigma V^T$, given the conditions on U, Σ and V.

Sometimes, it is convenient to let the columns of U and V form bases for $\mathbb{R}^P$ and $\mathbb{R}^N$ respectively. For V this is achieved by taking a basis of the kernel of X (cf. $\mathcal{Z}$ in the proof) and appending one column per element in the basis. By symmetry, the same is true for U. Finally, Σ is extended to a $P \times N$ matrix, by adding zeros. The same formula, $X = U\Sigma V^T$, still holds, but now $U^TU = UU^T = I_P$ and $V^TV = VV^T = I_N$.

We saw that $\|X\|_2 = \sigma_1$. Another matrix norm that is tightly connected to the SVD is the Frobenius norm.

Definition 5.2 (Frobenius norm)**.** The *Frobenius norm* of a matrix X is

$$\|X\|_F = \sqrt{\sum_{i,j} x_{ij}^2}.$$

Following the definition, we can state the square of the Frobenius norm as a sum of the squared vector norms of the columns of X, $\|X\|_F^2 = \sum_j \|X_{\cdot j}\|_2^2$. Since the vector norm $\|\cdot\|_2$ is invariant under rotation and reflection, it follows that the Frobenius norm is invariant under multiplication with orthonormal matrices, from both left and right (by symmetry). Hence

$$\|X\|_F = \|U\Sigma V^T\|_F = \|U^TU\Sigma V^TV\|_F = \|\Sigma\|_F = \sqrt{\sum_{i=1}^{r} \sigma_i^2},$$

where we have used the extended form of the SVD.

The *truncated* SVD is formed by keeping only the first k singular values of matrix X. Let $\Sigma_k \coloneqq \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k)$ and let U_k and V_k be the first k columns of U and V respectively. The Eckart-Young theorem[19] states that the truncated SVD gives the optimal rank-k approximation of a matrix.

Theorem 5.2 (Eckart-Young). *With the notation above,*

$$\inf_{\{Y\colon \operatorname{rank} Y \leq k\}} \|X - Y\|_F = \|X - U_k \Sigma_k V_k^T\|_F.$$

Note that $\|U_k \Sigma_k V_k^T\|_F^2 = \sum_{i=1}^{k} \sigma_i^2$ and $\|X - U_k \Sigma_k V_k^T\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2$.

5.2 Sparse Principal Components

Sparse Principal Components (SPC) is a special case of the Penalized Matrix Decomposition (PMD) framework by Witten et al[20]. When the number of variables is large, it can be hard to interpret the variable representation given by SVD, since they include all variables. Sparse methods such as SPC can simplify interpretation, by limiting the number of non-zero variables for each column of U. Witten et al. introduced additional constraints in the step for finding $\boldsymbol{u}, \boldsymbol{v}$ in the SVD procedure given above.

In the case of SPC, they solve the optimization problem:

$$\begin{aligned} \underset{\sigma, \boldsymbol{u}, \boldsymbol{v}}{\text{minimize}} \quad & \|X - \boldsymbol{u}\sigma\boldsymbol{v}^T\|_F^2 \\ \text{subject to} \quad & \|\boldsymbol{u}\|_2^2 = 1,\ \|\boldsymbol{v}\|_2^2 = 1,\ \|\boldsymbol{u}\|_1 \leq c,\ \sigma \geq 0 \end{aligned}$$

If $c \geq \sqrt{P}$, the solution coincides with the SVD, but if c is smaller, the lasso ($\|\cdot\|_1$) constraint will enforce sparse solutions, i.e. with many elements of $\boldsymbol{u}$ equal to 0. The smaller c, the higher degree of sparsity. The matrix updated is done like above, $X' \coloneqq X - \boldsymbol{u}\sigma\boldsymbol{v}^T$, and a constraint forcing $\boldsymbol{v}$ to be orthogonal to the previously computed columns is added to the minimization problem, since this is no longer guaranteed by the construction. For $\boldsymbol{u}$, no orthogonality constraint is added, which simplifies computations. As a result, the columns of U will not be orthogonal in general.

The minimization problem above is biconvex, that is, with $\boldsymbol{u}$ fixed it is convex in $\boldsymbol{v}$ and vice versa, and no closed form solution exists. Instead, it is solved by numerical iterations, alternatingly keeping $\boldsymbol{u}$ or $\boldsymbol{v}$ fixed, and solving for the other.

The objective function is not convex (in general), and the numerical optimization might get stuck at a local optima. We can expect this to happen more often when P is large and c is small. A consequence is that σ_{i+1} might be larger than σ_i, for some i, an undesirable property, since the goal is to represent the matrix as well as possible with a low-rank matrix. Witten et al. do not address this problem, but the following observation shows how the situation can be improved.

Let $\boldsymbol{u}_i, \boldsymbol{v}_i$ be local optimizers to the i'th optimization problem and suppose that $\sigma_{i+1} > \sigma_i$. Then

$$\begin{aligned}\boldsymbol{u}_{i+1}^T X_i \boldsymbol{v}_{i+1} &= \boldsymbol{u}_{i+1}^T \left(X_{i+1} + \boldsymbol{u}_i \sigma_i \boldsymbol{v}_i^T\right) \boldsymbol{v}_{i+1} \\ &= \boldsymbol{u}_{i+1}^T X_{i+1} \boldsymbol{v}_{i+1} + \sigma_i \boldsymbol{u}_{i+1}^T \boldsymbol{u}_i \boldsymbol{v}_i^T \boldsymbol{v}_{i+1} \\ &= \sigma_{i+1},\end{aligned}$$

where the second term disappears due to orthogonality of the $\boldsymbol{v}_i$'s. We see that $\boldsymbol{u}_{i+1}, \boldsymbol{v}_{i+1}$ is a better candidate solution to i'th optimization problem than $\boldsymbol{u}_i, \boldsymbol{v}_i$. Rolling back to the previous problem, now with $\boldsymbol{v}_{i+1}$ as the starting guess, we are thus guaranteed to reach a solution closer to the global optima. This procedure will ensure that the σ_i's are ordered from largest to smallest, and provide a better approximation of the matrix X.

6 Visualization

The purpose of measuring biological phenomena is to gain understanding of the underlying processes. It can be as easy as performing a simple experiment to test a well-formulated hypothesis, but the complexity of biological processes means that there is often a long road to travel before that point can be reached. In practice, we need to test vaguely stated hypotheses, search broadly for causes that could explain specific effects, or just describe what we see. However, the overwhelming amounts of data produced by high-throughput measurements are impossible to go through by hand. Finding suitable representations and developing methods for exploring and visualizing the data can greatly simplify all these tasks. Visualization of huge data sets requires a process for selecting what to show, to make it possible to spot overall patterns rather than get stuck in details. Done right, visualization can help the viewer get a qualitative understanding of the data, and hypotheses can be formed based on patterns that stand out. Visualization is about presenting a faithful view of the data, so that it can be interpreted by the human brain. It

is important to note that the method used to present the data influences which patterns can be discovered, and it's useful to have a basic understanding of the techniques used, in order to understand what we can see and what is hidden.

Information retrieval techniques can be used for visualization; fast searches make it possible to interactively explore the genome, find answers to questions about details and overall patterns, and gradually build an understanding of the data set. It facilitates working at different scales, and moving quickly between them. Genomic variants can for instance be viewed one by one, or together with other variants within the same gene or pathway.

Dimension reduction is a key instrument for visual presentation of complex data sets. A simple example of a data matrix X with 100 variables and 40 samples is shown in Figures 6 and 7. When viewing the entire data set (Figure 6), it is difficult to make out any patterns. Visualizing the complete data set by scatter plots is highly impractical, since the samples lie in 100-dimensional space. However, by finding a suitable rotation of the samples, an informative view of the data set can be shown using just a few dimensions. By Eckart-Young (Theorem 5.2), the best low-rank approximation of a matrix can be stated in terms of the SVD, $X = U\Sigma V^T$. Multiplying X with U^T from the left corresponds to rotation (and possibly mirroring) of the samples. It yields $U^T X = \Sigma V^T$, where the k'th column of $V\Sigma$ contains the coordinates of the samples for the k'th dimension in the new basis. This is most commonly referred to as Principal Component Analysis (PCA) and can equivalently be described as finding the directions of the data set that capture the most variance. In Figure 7, we compare four different ways to view the samples using four dimensions. Selecting a subset of the original variables, randomly or by choosing variables with large variance, is not enough to give an informative view. PCA is however able to reconstruct an almost optimal view, emphasizing that the view of the data set determines what patterns we can find.

Different dimension reduction methods give us different views into a data set and thus make it possible look for patterns of different kinds. Reducing the dimension of a data set results (almost surely) in a loss of information – a perfect representation using a lower number of dimensions cannot be expected. In Section 5, we described the loss functions for SVD (PCA) and SPC, showing how these methods penalize the loss of information in different ways. In both cases, the low-dimensional representation is constructed by minimization of the loss function. SPC aims to find a representation using a lower number of variables by constraining the loss function from SVD. This can increase interpretability and

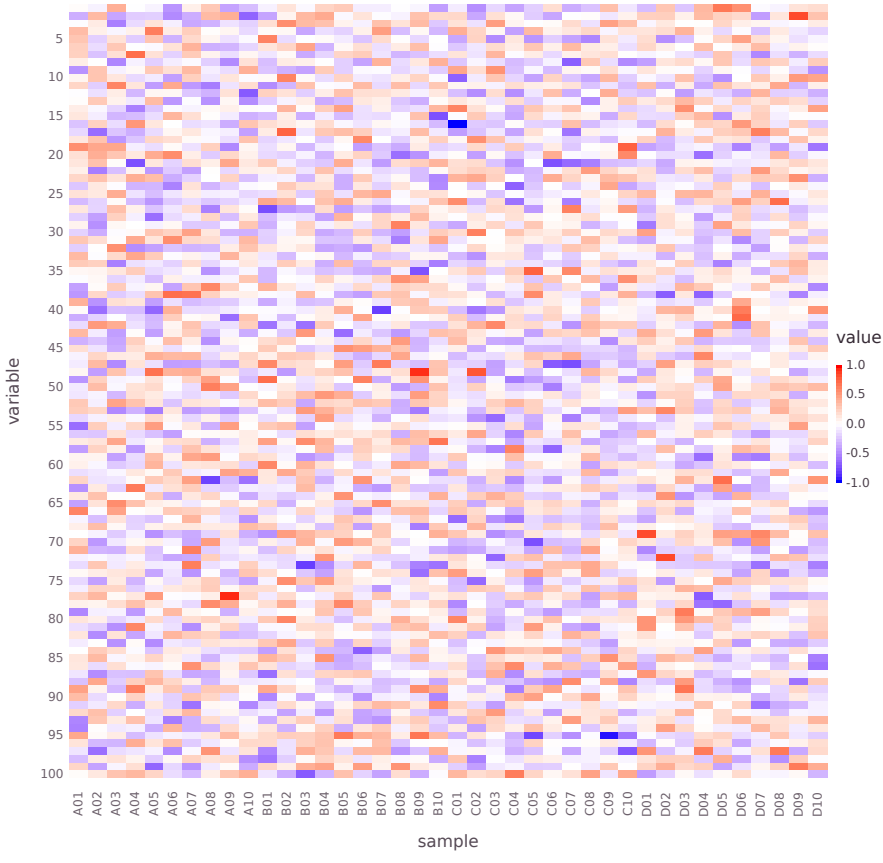


Figure 6: Heatmap showing a 100×40 data matrix for an example data set with 100 variables and 40 samples. Samples are generated from four classes (A–D) described by multivariate normal distributions with different means and identical covariance matrices σI. It is extremely difficult to discern any differences between the four classes.

can be motivated if we expect that most variables are irrelevant for the patterns we are looking for.

Multidimensional scaling (MDS)[21] is another method with strong similarities to PCA, but with a different starting point. Based on dissimilarity scores between all pairs of samples, the MDS loss function penalizes the difference between the dissimilarities and the distances in the low-dimensional representation.

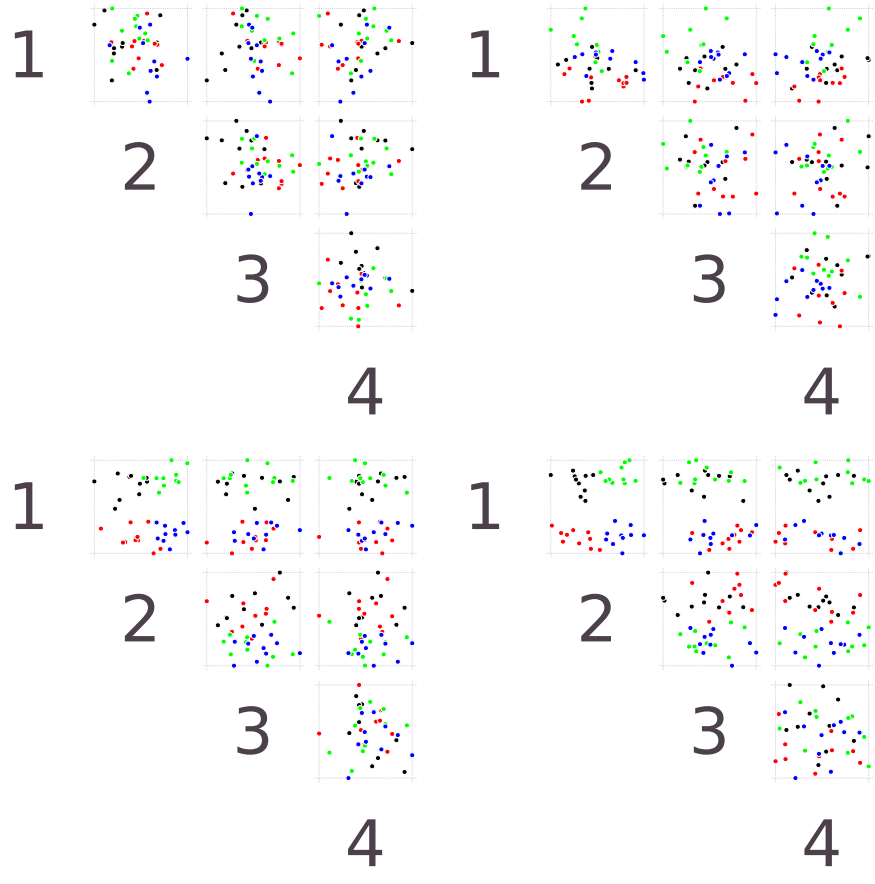


Figure 7: Four visualizations of the samples from the data set shown in Figure 6. The classes A–D are shown in black, red, green and blue respectively. Each of the four visualizations correspond to a different rotation of the samples, in high-dimensional space, and displaying the four first dimensions in pairwise plots. **Top left:** The first four variables (the first four rows of Figure 6). No apparent differences between the classes are seen. **Top right:** The four variables with highest variance are displayed. Some differences between the classes can be observed, but the distinction is not clear. **Bottom left:** A PCA (SVD) reconstruction is able to separate the four classes quite well in the first two dimensions. **Bottom right:** Ground truth, the means of the multivariate normal distributions for all four classes are in the plane spanned by dimension 1 and 2. Dimensions 3 and 4 give no additional information.

If the dissimilarity scores coincide with sample distances in Euclidean space, MDS is identical to PCA. The major benefit of visualization using MDS is the flexibility stemming from the fact that any way of measuring dissimilarities between samples can be used as input to the algorithm, in particular, no representation in Euclidean space (no data matrix) is needed. However, variable information is lost, since the input to MDS is stated entirely in terms of the samples.

Isomap[22] assumes that samples are constrained to a nonlinear manifold in high-dimensional space. In this setting, with some regularity assumptions for the manifold, the Euclidean distance in high-dimensional space between nearby samples is approximately equal to the geodesic distance on the manifold, but for far-apart samples, no such relationship can be assumed. Hence, Isomap constructs a neighborhood graph connecting nearby samples, with edge lengths equal to the Euclidean distance between samples such that the shortest distance between two samples in the graph is an approximation of the geodesic distance on the manifold. The graph distance are used as dissimilarity scores and a low-dimensional representation constructed by MDS. The Isomap model makes it possible to capture different structures at different scales and is thus in some sense able to fit more information into fewer dimensions, in comparison to PCA, which is very useful for visualizing complex data. A drawback is that it is very difficult to attach meaning to the different dimensions for Isomap, whereas the dimensions of PCA are linear combinations of the original variables.

7 Overview of Papers

Paper I

Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia

Henrik Lilljebjörn, Rasmus Henningsson, Axel Hyrenius-Wittsten, Linda Olsson, Christina Orsmark-Pietras, Sofia von Palffy, Maria Askmyr, Marianne Rissler, Martin Schrappe, Gunnar Cario, Anders Castor, Cornelis J.H. Pronk, Mikael Behrendtz, Felix Mitelman, Bertil Johansson, Kajsa Paulsson, Anna K. Andersson, Magnus Fontes and Thoas Fioretos

In this study, we delineate the fusion gene landscape in B-cell precursor acute lymphoblastic leukaemia (BCP ALL), finding new gene fusions of clinical importance, further stratifying patients. The paper shows how information retrieval

methods, adapted to a genomic setting, can be used to create informative visualization of fusion genes and other genomic data. In addition, we create an accurate subtype classifier based on RNA-seq measurements, by combining gene fusion detection with gene expression levels.

Author contributions: H.L. and T.F. conceived the project. H.L., A.H.-W., L.O., C.O.-P. M.A. and M.R. performed the experiments. H.L., R.H., A.H.-W., L.O., C.O.-P., S.v.P., F.M., B.J., K.P., A.K.A., M.F. and T.F. analysed the data. M.S., G.C., A.C. C.J.H.P. and M.B. provided samples and clinical data. H.L., K.P., A.K.A. and T.F. wrote the manuscript, which was reviewed and edited by the other co-authors.

Paper II

Attenuation of RNA viruses by redirecting their evolution in sequence space

Gonzalo Moratorio, Rasmus Henningsson, Cyril Barbezange, Lucia Carrau, Antonio V. Bordería, Hervé Blanc, Stephanie Beaucourt, Enzo Z. Poirier, Thomas Vallet, Jeremy Boussier, Bryan C. Mounce, Magnus Fontes and Marco Vignuzzi

The paper describes how changing the genomic composition of a virus, without changing the proteins it is coding for, control the formation of the viral population as mutations take place and how this affects the evolutionary potential of the virus. In particular, we show that the virus can be moved to a more detrimental region of sequence space, where mutations are more likely to cause Stop codons – effectively inactivating the virus, and that the virus is unable to escape from this region.

Author contributions: G.M., C.B. and M.V. designed the experiments. G.M., C.B., L.C., A.V.B., H.B., E.Z.P., S.B., T.V. and B.C.M. performed experiments. G.M., R.H., J.B., C.B., M.F. and M.V. analysed the data. G.M. and M.V. wrote the paper.

Paper III

SMSSVD – SubMatrix Selection Singular Value Decomposition

Rasmus Henningsson and Magnus Fontes

In this paper, we develop a method for signal decomposition of data matrices that is well suited for high-throughput biological data sets. Based on optimally chosen

variable subsets, we adaptively reduce the noise for each signal, thus improving the reconstruction accuracy and making it possible to uncover signals closer to the limit of detection. The method is simple to implement and results in orthogonal signals that are described using the original set of variables.

Author contributions: R.H. conceived the first version. R.H. and M.F. improved and developed the method. R.H. made the software implementation and evaluated the method on different data sets. R.H. wrote the paper, which was reviewed and edited by M.F.

Paper IV

DISSEQT – DIStribution based modeling of SEQuence Space Time dynamics

Rasmus Henningsson, Gonzalo Moratorio, Antonio V. Bordería, Marco Vignuzzi and Magnus Fontes

This paper defines a complete pipeline for analyzing, visualizing and predicting the evolutionary behavior of heterogeneous biological populations. We characterize each population by a positive measure over sequence space, which is inferred from deep sequencing data. SMSSVD is then used to create a robust and accurate model where populations can be compared to each other. We apply the pipeline to several viral population data sets and show that the composition of the population is crucial for understanding phenotypic effects – knowing just the consensus sequence is not enough.

Author contributions: G.M., M.V. and A.V.B. designed the experiments. G.M. and A.V.B. performed the experiments. R.H. and M.F. developed the pipeline and mathematical methods. R.H. made the software implementation and analyzed the data. R.H. wrote the paper, which was reviewed and edited by M.F., M.V. and G.M.

Bibliography

[1] Pere Alberch. From genes to phenotype: dynamical systems and evolvability. *Genetica*, 84(1):5–11, 1991.

[2] Mark WJ Ferguson and Ted Joanen. Temperature of egg incubation determines sex in Alligator mississippiensis. *Nature*, 296(5860):850–853, 1982.

[3] Agoston E Eiben and James E Smith. *Introduction to Evolutionary Computing*. Springer, second edition, 2015.

[4] Richard Dawkins. *The selfish gene*. Oxford university press, 4th revised edition edition, 2016.

[5] Philip J Gerrish and Richard E Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102:127, 1998.

[6] Christof K Biebricher and Manfred Eigen. What is a quasispecies? In *Quasispecies: Concept and Implications for Virology*, pages 1–31. Springer, 2006.

[7] Esteban Domingo, Verónica Martín, Celia Perales, Ana Grande-Pérez, Juan F García-Arriaza, and Armando Arias. Viruses as quasispecies: biological implications. In *Quasispecies: Concept and Implications for Virology*, pages 51–82. Springer, 2006.

[8] Marco Vignuzzi, Jeffrey K Stone, Jamie J Arnold, Craig E Cameron, and Raul Andino. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, 439(7074):344–348, 2006.

[9] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[10] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719, 2009.

[11] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, 2013.

[12] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016.

[13] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval. *An Introduction To Information Retrieval*, 2008.

[14] Torsten Thomas, Jack Gilbert, and Folker Meyer. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2(1):3, 2012.

[15] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

[16] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[17] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[18] Gene H Golub and Charles F Van Loan. *Matrix computations*. Johns Hopkins Univ. Press, third edition, 1996.

[19] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[20] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

[21] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, volume 2. Springer, 2009.

[22] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500): 2319–2323, 2000.

PAPERS

PAPER I

Nature Communications

Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia

Henrik Lilljebjörn, Rasmus Henningsson, Axel Hyrenius-Wittsten, Linda Olsson, Christina Orsmark-Pietras, Sofia von Palffy, Maria Askmyr, Marianne Rissler, Martin Schrappe, Gunnar Cario, Anders Castor, Cornelis J.H. Pronk, Mikael Behrendtz, Felix Mitelman, Bertil Johansson, Kajsa Paulsson, Anna K. Andersson, Magnus Fontes and Thoas Fioretos

Abstract

Fusion genes are potent driver mutations in cancer. In this study, we delineate the fusion gene landscape in a consecutive series of 195 paediatric B-cell precursor acute lymphoblastic leukaemia (BCP ALL). Using RNA sequencing, we find in-frame fusion genes in 127 (65%) cases, including 27 novel fusions. We describe a subtype characterized by recurrent *IGH-DUX4* or *ERG-DUX4* fusions, representing 4% of cases, leading to overexpression of *DUX4* and frequently co-occurring with intragenic *ERG* deletions. Furthermore, we identify a subtype characterized by an *ETV6-RUNX1*-like gene-expression profile and coexisting *ETV6* and *IKZF1* alterations. Thus, this study provides a detailed overview of fusion genes in paediatric BCP ALL and adds new pathogenetic insights, which may improve risk stratification and provide therapeutic options for this disease.

Paediatric B-cell precursor acute lymphoblastic leukaemia (BCP ALL), the most common childhood malignancy, is stratified into prognostically relevant genetic subgroups based on the presence of certain gene fusions and aneuploidies[1]. However, 25% of cases do not have any characteristic genetic aberrations at diagnosis, and the underlying driver events are unknown. For these cases, here denoted as 'B-other', the identification of pathogenetic changes will not only increase our understanding of the leukemogenic process, but may also be important in a clinical context, because such alterations can be used for improved risk classification and for targeted treatment. Recent genome-wide studies have provided critical pathogenetic insights into paediatric BCP ALL, including the identification of a dismal prognosis for cases with *IKZF1* deletions[2–5] and for cases with a 'Ph-like'[4–8] gene-expression signature similar to that of Philadelphia (Ph)-positive ALL. In addition, the mutational landscapes of BCP ALL subtypes defined by *ETV6-RUNX1*, *TCF3-PBX1*, *TCF3-HLF*, high hyperdiploidy (51–67 chromosomes), hypodiploidy (< 45 chromosomes) or *MLL* (also known as *KMT2A*) rearrangements have been delineated using high-resolution sequencing techniques[9–13]. These studies have almost exclusively been performed at the DNA level and no large-scale characterization of the gene-fusion landscape in paediatric BCP ALL has been reported to date. To gain a better understanding of the gene-fusion landscape of BCP ALL, we performed RNA sequencing (RNA-seq) in a population-based series of 195 paediatric (< 18 years of age) BCP ALL cases. We report that gene fusions are present in 65% of BCP ALL, and identify several new fusions and two novel subtypes; one characterized by recurrent *IGH-DUX4* or *ERG-DUX4* fusions and one characterized by an *ETV6-RUNX1*-like gene-expression profile, and coexisting *ETV6* and *IKZF1* alterations.

1 Results

1.1 Identified subtypes enable classification of 98% of cases

All 195 cases subjected to RNA-seq had previously been analysed by G-banding, fluorescent *in situ* hybridization (FISH) and molecular analyses for the detection of established genetic BCP ALL alterations as part of routine clinical diagnostics (Supplementary Fig. 1 and Supplementary Data 1). Using RNA-seq, we identified an in-frame fusion gene in 127/195 (65%) BCP ALL cases and out-of-frame fusions in 20/195 (10%) cases (Fig. 1 and Supplementary Data 2–4). Notably,

of the 68 cases lacking an in-frame fusion gene, the majority (64/68, 94%) were high-hyperdiploid ($n = 56$), hypodiploid ($n = 2$), Ph-like ($n = 3$), harboured a dic(9;20) ($n = 1$) or belonged to novel subtypes further described below ($n = 2$) (Fig. 1e and Supplementary Data 3). One subgroup, comprising 16% of B-other cases (4% of the entire BCP ALL cohort), harboured rearrangements of the double homeobox 4 (*DUX4*) gene and overlapped with a previously described group of patients with a homogenous gene-expression profile and frequent *ERG* deletions[6,14,15]. In addition, a new subtype, harbouring co-existing rearrangements of *ETV6* and *IKZF1* and associated with *ETV6-RUNX1*-like gene-expression pattern (3% of the cohort; 14% of B-other cases), was identified. Taken altogether, 98% of the BCP ALL cases could be classified into distinct genetic subtypes with a known underlying driver mutation or, less commonly, with a rare in-frame gene fusion (Figs 1f, 2, Supplementary Data 3), providing new insights and pathogenetic markers in BCP ALL.

1.2 DUX4-rearranged cases constitute a distinct BCP ALL subtype

Recurrent *DUX4* rearrangements were identified in 8/195 (4%) BCP ALL cases and were confined to B-other cases (8/50 cases, 16%; Figs 1a–c and 2, Supplementary Data 3). The rearrangements were either a fusion between *IGH* and *DUX4* (7/8 cases) or between *ERG* and *DUX4* (1 case). To confirm this and other findings within the B-other group, we performed RNA-seq of an independent validation cohort of 49 paediatric B-other cases that were negative for *BCR-ABL1*, *ETV6-RUNX1*, *TCF3-PBX1*, *MLL* rearrangements and high hyperdiploidy (Supplementary Data 5). This analysis revealed an additional 20 cases with *DUX4* rearrangements, resulting in a total of 26 cases with *IGH-DUX4* and 2 with *ERG-DUX4* across the 2 cohorts.

DUX4 encodes a homeobox-containing protein and is located within a subtelomeric D4Z4 repeat region on 4q and 10q. It is present in 11–100 copies on each allele, and is epigenetically silenced in somatic tissues. Loss of epigenetic silencing through shortening of the D4Z4 repeats leads to the degradation of muscle cells, and causes facioscapulohumeral muscular dystrophy[17,18].

To confirm the *DUX4* rearrangements at the genomic level, we performed mate-pair whole-genome sequencing (MP-WGS) in all eight cases in the discovery cohort, enabling powerful mapping of structural genomic rearrangements (Supplementary Data 6). These analyses confirmed the *DUX4* rearrangements at

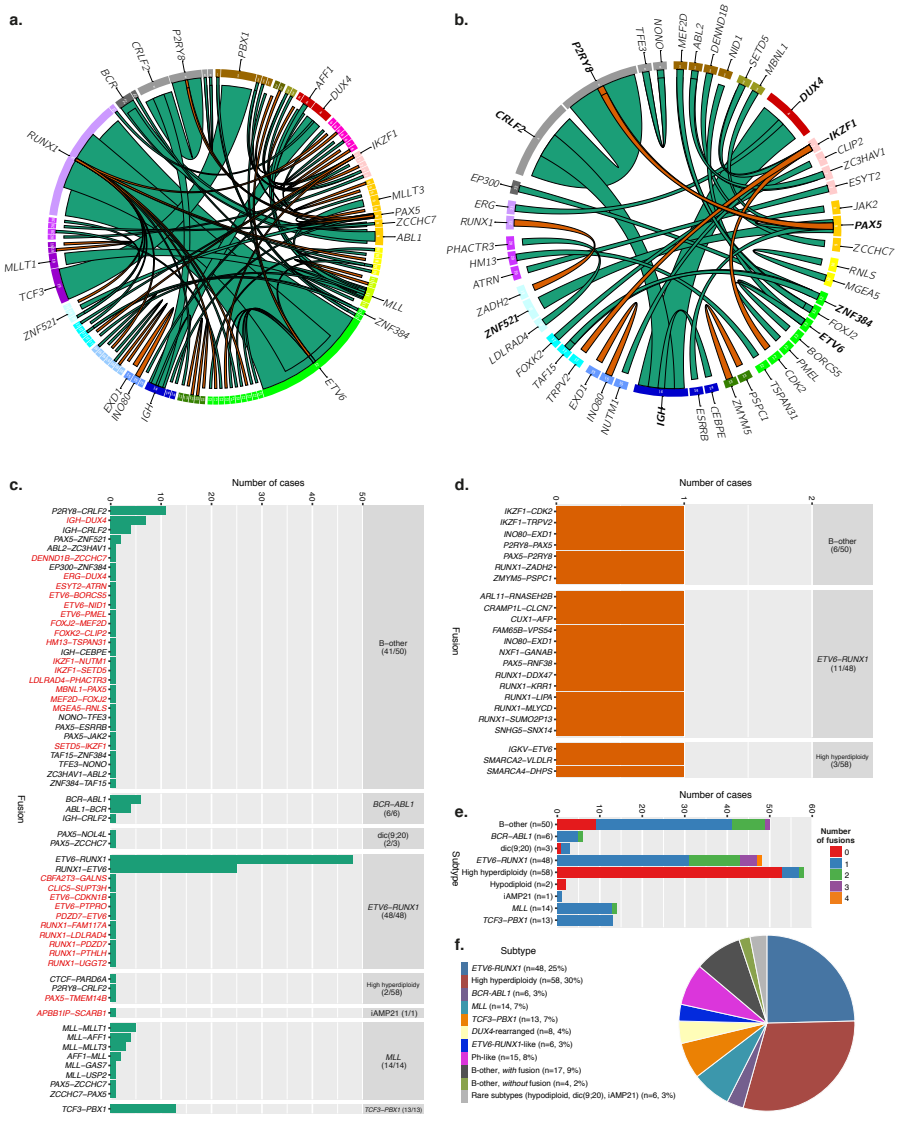


Figure 1: (Continued on the following page.)

Figure 1: **Overview of the gene fusions present in 195 paediatric BCP ALL cases in the discovery cohort.** (**a**) In-frame gene fusions (green) and out-of-frame gene fusions (orange) are illustrated using Circos [16]. Each ribbon has one end attached to the circle, indicating the 5′-partner gene of the fusion. The width of the ribbon is proportional to the number of detected fusions. Genes are arranged according to their genomic position (from chromosome 1–22 followed by X and Y) and chromosomes are marked in different colours. The gene symbol is denoted for genes involved in more than two unique fusions or in recurrent fusions. (**b**) In-frame gene fusions and out-of-frame gene fusions present in 50 B-other cases. The gene symbol for genes involved in more than two unique fusions or in recurrent fusions is indicated in bold. (**c**) The frequency of in-frame gene fusions by genetic subtype (indicated in the right column with the number of affected cases in parenthesis). Novel gene fusions are indicated in red ($n = 27$, reciprocal gene-fusion pairs counted as a single fusion) and previously described fusions are indicated in black ($n = 22$). (**d**) The frequency of out-of-frame gene fusions by genetic subtype (indicated in the right column with the number of affected cases in parenthesis). (**e**) Total number of gene fusions per case by genetic subtype (including both in-frame and out-of-frame fusions; reciprocal gene-fusion pairs counted as a single fusion). (**f**) Distribution of 195 BCP ALL cases within genetic subtypes defined by gene-expression profile and gene fusions detected by RNA-seq.

the DNA level in all cases, and revealed that the *IGH-DUX4* fusions resulted from insertions of a partial copy of *DUX4* into the *IGH* locus, including between 90–1,200bp upstream of *DUX4* and between 939 and 1,272bp of coding sequence from *DUX4* (Fig. 3 and Supplementary Fig. 2). Similarly, *ERG-DUX4* in case 75 was the result of an insertion of a partial copy of *DUX4* into intron 3 of *ERG*, containing 936bp of coding sequence of *DUX4* (Fig. 3j). Neither *IGH-DUX4* nor *ERG-DUX4* would give rise to a chimeric protein; instead, the rearrangements and expression pattern suggest that the relocation of *DUX4* induces its expression from regulatory regions of the partner gene (Fig. 3 and Supplementary Fig. 3). The full-length DUX4 protein consists of 424 amino acids, but 7 of the 8 genomically characterized cases expressed truncated *DUX4* transcripts encoding between 312 and 420 amino acids (Fig. 3). Only case 47 expressed the full coding length of *DUX4*. All variants, however, retained both homeobox domains of DUX4, thus preserving its DNA-binding capacity.

All *DUX4*-rearranged cases displayed a distinct overexpression of *DUX4* as determined by RNA-seq; in contrast, expression of this gene was significantly lower or absent in the other investigated 216 BCP ALL cases across the discovery and validation cohorts (Supplementary Fig. 3). Notably, all cases with *DUX4* rearrangements displayed a global gene-expression pattern matching that of a subgroup of BCP ALL cases previously reported to be associated with *ERG* deletions in 38–55% of cases (Supplementary Fig. 4) [6]. Conversely, all cases with this gene-expression profile had *DUX4* rearrangements and overexpression of *DUX4*,

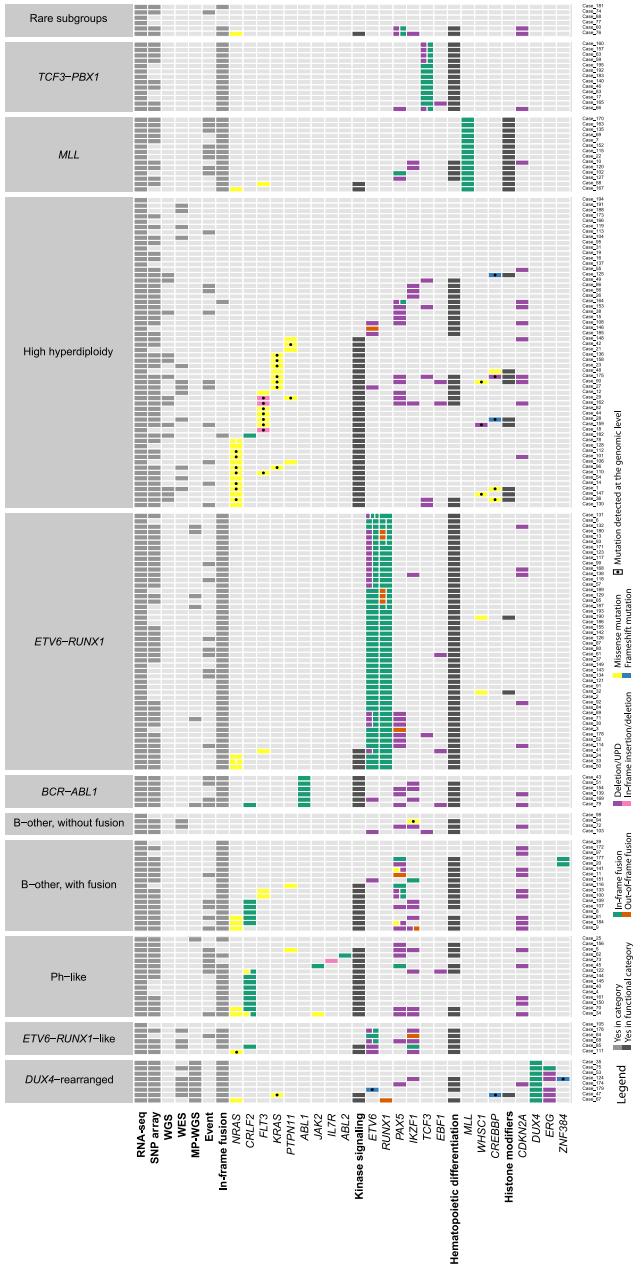


Figure 2: (Continued on the following page.)

Figure 2: **Genetic alterations present in 195 BCP ALL cases in the discovery cohort.** The cases are arranged according to genetic subtypes defined by gene-expression profile and gene fusions detected by RNA-seq, and were further characterized by SNP array, WGS, WES and MP-WGS. Genes recurrently altered in BCP ALL are arranged according to functional categories (kinase signalling, haematopoietic differentiation, histone modifiers and others). Events comprise induction failure and relapse.

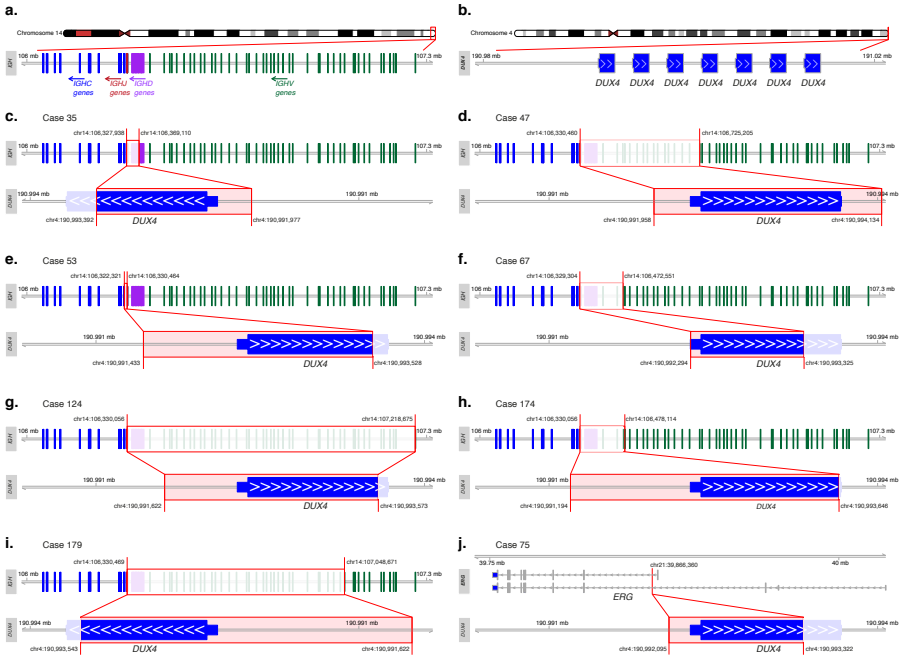


Figure 3: **DUX4 rearrangements in eight BCP ALL cases in the discovery cohort.** (**a**) Arrangement of immunoglobulin genes in the *IGH* locus. (**b**) Structure of the subtelomeric D4Z4 repeat region on 4q in the hg19 reference genome. This reference representation has seven repeats, each containing a *DUX4* gene. Healthy individuals have 11–100 repeats. (**c–i**) Structure of the *IGH-DUX4* rearrangements in (**c**) case 35, (**d**) case 47, (**e**) case 53, (**f**) case 67, (**g**) case 124, (**h**) case 174 and (**i**) case 179. (**j**) Structure of the *ERG-DUX4* rearrangement in case 75. All genomic coordinates are based on the human reference genome hg19. Because it is impossible to determine which *DUX4* repeat is involved in the rearrangement, the coordinates from the first *DUX4* repeat are represented in the figures.

indicating that the *DUX4* rearrangement is the founder event for this group (Supplementary Fig. 4). We determined the frequency of *ERG* deletions in *DUX4*-rearranged cases by MP-WGS in the discovery cohort and indirectly by ascertai-

ning truncated transcripts by RT-PCR[14] in the validation cohort. This revealed *ERG* deletions in 5/8 (63%) cases in the discovery cohort and in 10/20 (50%) cases in the validation cohort (Supplementary Fig. 5), supporting that the *DUX4*-rearranged subtype reported here is identical to the previously described subtype with a distinct gene-expression profile and frequent *ERG* deletions[6]. This group has consistently been associated with a favourable prognosis, both when defined by the distinct gene-expression profile[6], and when defined by the characteristic *ERG* deletions[14,15]. In the discovery cohort, we observed no relapses among the 8 *DUX4*-rearranged cases, while 4 of 20 cases (20%) experienced relapse in the validation cohort. With the identification of *DUX4* rearrangement as a new marker in BCP ALL, it will be interesting to ascertain its prognostic impact in larger, uniformly treated, cohorts. To characterize further the mutational landscape of *DUX4*-rearranged BCP ALL, whole-exome sequencing (WES) was performed in five *DUX4*-rearranged cases with matched constitutional samples available (Supplementary Data 6). These were found to harbour between 4 and 10 non-silent exome mutations each (Supplementary Data 7). The only recurrently mutated gene was the transcription factor *ZEB2*, with mutations in two cases (#75 and #174).

A borderline significance for cases with *DUX4* rearrangements being older than cases lacking such fusions (Mann-Whitney's two-sided test, $P = 0.051$; median age 6.5 versus 4 years) was seen in the discovery cohort. The median age at diagnosis of patient with *DUX4*-rearranged ALL in the combined cohorts was 8.5 years (range 2–15 years). Considering the pronounced age peak at 3–5 years for childhood BCP ALL in general[19], this indicates that *DUX4*-rearrangements are associated with older age, although this needs to be confirmed in larger cohorts. Interestingly, an association with older age has previously been described for cases with *ERG* deletions[14,15].

The complexity of the genomic region where *DUX4* is located is most likely the reason that *DUX4* fusions have not been previously discovered in BCP ALL. Our standard RNA-seq bioinformatics pipeline could only detect the rearrangement in 7 of 28 cases, whereas a guided analysis that identified RNA-seq reads that linked any region within 2kb of *DUX4* to the reads within the *IGH* locus identified the *IGH-DUX4* in an additional 19 cases (Supplementary Data 8). In the remaining two cases with *DUX4* overexpression, a fusion between *ERG* and *DUX4* was discovered by surveying the RNA-seq reads for regions similarly linked to the region surrounding *DUX4*. The aberrations were also not expected to be de-

tectable on the chromosomal level by either G-banding or FISH, due to the small sizes of the insertions. In line with this, G-banding results from the eight *DUX4*-rearranged cases in the discovery cohort showed normal karyotypes in four cases and unspecific changes in two cases; in two cases, G-banding analyses had failed (Supplementary Data 1). Taken altogether, RNA-seq followed by guided searches for *DUX4* chimeric transcripts is a reliable way to identify *DUX4* rearrangements, although both WES and MP-WGS allows the detection of the rearrangement at the genomic level (Supplementary Fig. 2).

1.3 ETV6-RUNX1-like gene expression in cases lacking the fusion

Gene-expression profiling based on the RNA-seq data showed that 6/50 (12%) B-other cases in the discovery cohort clustered with the *ETV6-RUNX1*-positive cases, despite lacking molecular evidence of this fusion by FISH, RT-PCR and RNA-seq (Fig. 4a–c, and Supplementary Table 1). The gene-expression similarities were further supported by gene set enrichment analysis (GSEA)[20] (Supplementary Figs 6–8). These six cases were thus denoted '*ETV6-RUNX1*-like ALL'. Interestingly, RNA-seq together with single-nucleotide polymorphism (SNP) array profiling revealed that five of the six *ETV6-RUNX1*-like cases harboured co-existing *ETV6* and *IKZF1* aberrations (Figs 2, 5; Supplementary Data 3 and Supplementary Table 1).

Specifically, case 64 contained an in-frame fusion between *ETV6* (at 12p13) and *PMEL* (at 12q13) together with an out-of-frame fusion between *IKZF1* (at 7p12) and *CDK2* (at 12q13) that lacked functional domains from *IKZF1* (Fig. 5a and Supplementary Table 1). Case 68 contained an in-frame fusion between *ETV6* and *BORCS5* (12p13) caused by a small deletion in 12p13, together with a deletion spanning the first exons of *IKZF1* (Fig. 5b and Supplementary Table 1). Case 85 contained an intragenic *ETV6* deletion and a t(3;7)(p25;p12) giving rise to in-frame reciprocal *SETD5-IKZF1* and *IKZF1-SETD5* fusions (Fig. 5c and Supplementary Table 1). Case 111 had interstitial deletions on both 7p and 12p, resulting in whole-gene deletions of *IKZF1* and *ETV6* (Fig. 5d and Supplementary Table 1). Case 176 carried a deletion of the entire 7p, including *IKZF1*, together with an in-frame fusion between *ETV6* and *NID1* (at 1q42) and an interstitial deletion on 12p removing the second *ETV6* allele (Fig. 5e and Supplementary Table 1). Finally, one case (#105) had no lesions affecting *ETV6* or *IKZF1* as detected by RNA-seq (analysis by SNP array was precluded due to lack

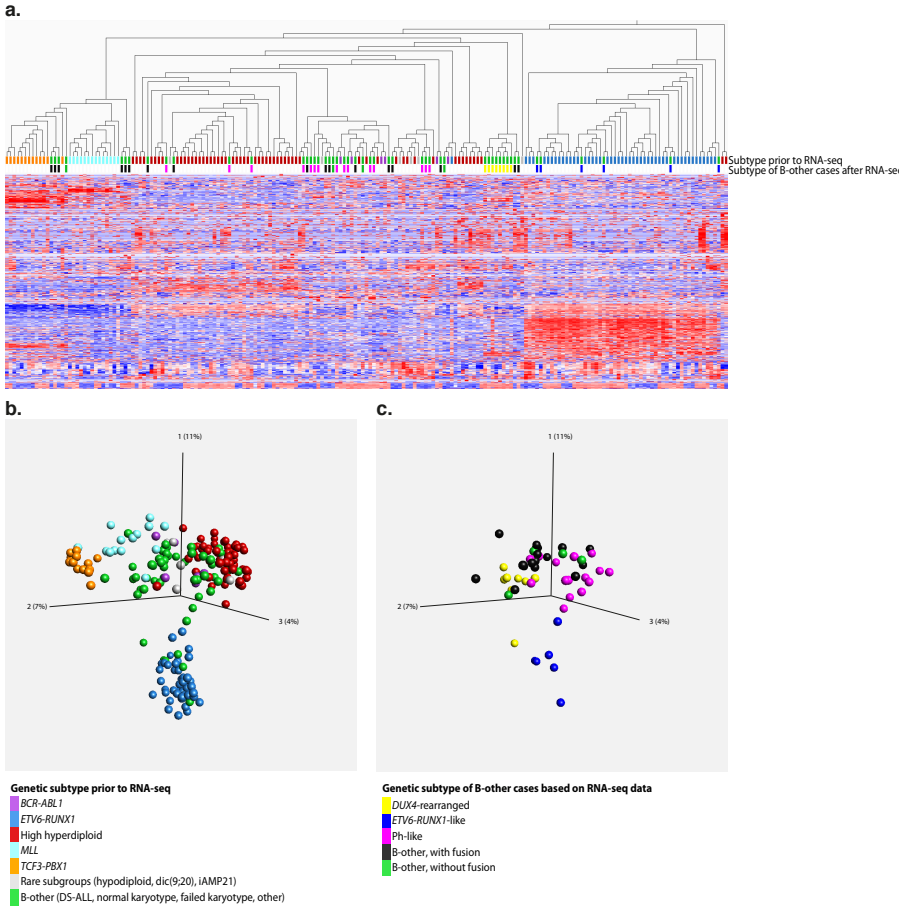


Figure 4: **Hierarchical clustering and principal component analyses of RNA-seq gene-expression data.** A variance threshold was set at standard deviation 0.285, retaining 638 variables. The colour-coding of BCP ALL subtypes, used throughout the figure, is indicated in the bottom. (**a**) Unsupervised hierarchical clustering analysis of 195 BCP ALL cases. Coloured boxes below the dendrogram indicate the subtype of each sample. The genetic subtype of B-other cases, based on the gene-expression and gene-fusion data, is indicated on the lower line. (**b**) Principal component analysis (PCA) of gene-expression data from all 195 BCP ALL cases. (**c**) PCA based on the data displayed in **b**, but only showing the 50 B-other cases colour-coded according to the genetic subtype based on the gene-expression and gene-fusion data. DS-ALL, Down's syndrome ALL; iAMP21, intrachromosomal amplification of chromosome 21.

of DNA). Thus, in total, genetic lesions affecting both *ETV6* and *IKZF1* were identified in all five cases where both RNA-seq and SNP array profiling could be performed (Fig. 5 and Supplementary Table 1). Combined lesions of *ETV6* and *IKZF1* were otherwise exceedingly rare outside of this group (3/152 (2%) cases with available SNP array data; $P < 0.001$, Fisher's exact test; Fig. 2). To characterize further *ETV6-RUNX1*-like ALL, we performed WES on four *ETV6-RUNX1*-like cases with available matched constitutional samples (cases 68, 85, 111 and 176; Supplementary Data 6 and 7). These cases carried between 3 and 29 non-silent exome mutations (with an allele frequency above 10%), but no gene was recurrently mutated.

RNA-seq of the independent validation cohort identified four additional cases with *ETV6-RUNX1*-like gene-expression profiles. Three of these harboured out-of-frame *ETV6* fusions (with *CREBBP* at 16p13, *BCL2L14* at 12p13 and *MSH6* at 2p16); in the fourth case, no fusion was detected (Supplementary Data 5). Unfortunately, no DNA was available for SNP array analyses, precluding a complete evaluation of deletions affecting *ETV6* or *IKZF1* in these cases.

We conclude that alterations of *ETV6*, either by the generation of alternative gene fusions, or, more rarely, *ETV6* deletions, in combination with *IKZF1* lesions, represent an alternative mechanism to elicit the same transcriptional perturbation as seen in classical *ETV6-RUNX1* fusion-positive cases. Interestingly, both *IKZF1* and *RUNX1* encode transcription factors important for B-cell maturation[21,22], and it is tempting to speculate that loss of IKZF1 may substitute for the altered function of RUNX1 in the ETV6-RUNX1 fusion protein. In line with this, we note that *IKZF1* deletions are rare in the *ETV6-RUNX1*-positive cases ($\sim 3\%$) in this and other cohorts[23,24].

While the small number of *ETV6-RUNX1*-like cases prohibited meaningful survival analyses, only two relapses were recorded among the ten *ETV6-RUNX1*-like cases in the combined discovery and validation cohort, indicating that the frequent *IKZF1* aberrations did not confer a dismal prognosis, as otherwise described for *IKZF1* deletions in BCP ALL[7,8]. However, further studies are warranted to evaluate the clinical impact of *IKZF1* deletions in *ETV6-RUNX1*-like BCP ALL.

1.4 In-frame gene fusions are present in most B-other cases

An in-frame fusion gene could be detected in 41/50 B-other cases (82%) in the population-based discovery cohort (Supplementary Data 3). The B-other cases could be subdivided into five non-overlapping groups: those with Ph-like ($n =$

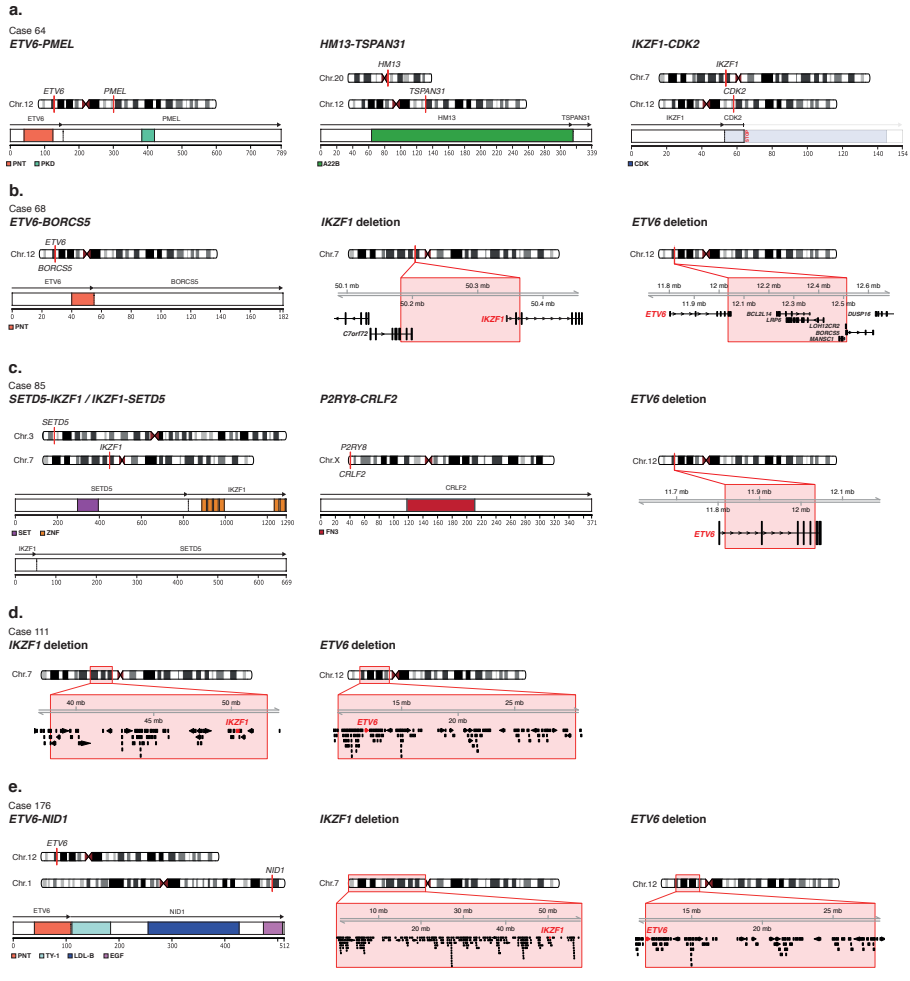


Figure 5: **Overview of aberrations in BCP ALL cases with ETV6-RUNX1-like gene-expression pattern.** The fusion genes are illustrated with a schematic overview of the chromosomal position of the genes involved in the fusion (top) and the retained protein domains together (bottom). Genomic deletions affecting *ETV6* and *IKZF1* are depicted with a red box indicating the deletion both at the chromosomal level and at the gene level, with *ETV6* and *IKZF1* in red. Illustrated protein domains: PNT, pointed domain; PKD, polycystic kidney disease domain; A22B, Peptidase A22B domain; CDK, protein kinase domain; SET, SET domain; ZNF, zinc-finger domain; FN3, fibronectin type-III domain; TY-1, thyroglobulin type-1 domain; LDL-B, LDL-receptor class B repeats; and EGF, EGF-like domain. (**a**) Gene fusions present in case 64. No SNP array data were available for this case. (Continued on the following page.)

Figure 5: *IKZF1-CDK2* is an out-of-frame fusion, with no functional domains from *CDK2* being included in the fusion protein. **(b)** Gene fusions and deletions present in case 68. The breakpoints of the *ETV6* deletion are within *ETV6* and *BORCS5*; likely representing the event that created the *ETV6-BORCS5* fusion gene. The breakpoints of the *IKZF1* deletion occur within the *C7orf72* and *IKZF1* genes, but RNA-seq data did not indicate the presence of a fusion transcript of these genes. **(c)** Gene fusions and deletions present in case 85. The *P2RY8-CRLF2* fusion does not contain any coding features from *P2RY8* but leads to overexpression of the entire coding region of *CRLF2*. **(d)** Deletions present in case 111. No gene fusions were detected in this case. **(e)** Gene fusions and deletions present in case 176.

15; Supplementary Fig. 9a) or *ETV6-RUNX1*-like ($n = 6$) gene-expression profiles, those with *DUX4* rearrangements ($n = 8$), and remaining cases with ($n = 17$) or without ($n = 4$) in-frame gene fusions ('B-other, with fusion' and 'B-other, without fusion', respectively, Fig. 2).

In agreement with previous descriptions of Ph-like BCP ALL, most cases (11/15, 73%) harboured gene fusions that deregulate the cytokine receptor CRLF2 (*P2RY8-CRLF2*, $n = 6$; and *IGH-CRLF2*, $n = 3$) or activate therapeutically targetable kinases (*ZC3HAV1-ABL2* in #62 and *PAX5-JAK2* in #45) (refs 7,8; Supplementary Data 3). In addition, RNA-seq data revealed mutations in the JAK-STAT pathway genes in 2/15 cases, 13% (Fig. 2 and Supplementary Data 3)[7,8].

Among the 17 cases in the 'B-other, with fusion' group, 11 cases (65%) harboured in-frame gene fusions previously described in BCP ALL[25]: *P2RY8-CRLF2* ($n = 4$), *PAX5-ZNF521* ($n = 2$), *EP300-ZNF384* (ref. 26) ($n = 1$), *IGH-CEBPE* ($n = 1$), *IGH-CRLF2* ($n = 1$), *PAX5-ESRRB* ($n = 1$) and *TAF15-ZNF384* ($n = 1$); in addition, a *NONO-TFE3* fusion gene, until now only reported in renal cell carcinoma[27,28], was found in a single case (#172; Fig. 1c and Supplementary Data 3). These fusions are likely genetic driver events in BCP ALL leukemogenesis. The importance of the novel in-frame gene fusions in the remaining five cases remains to be determined, but it is noteworthy that three had fusions (*DENND1B-ZCCHC7*, *MEF2D-FOXJ2*, *IKZF1-NUTM1*) involving genes recurrently rearranged in BCP ALL, namely *ZCCHC7*, *MEF2D*, *IKZF1* and *NUTM1* (ref. 25).

A high frequency of B-other cases from the validation cohort (36/49, 73%) also expressed an in-frame gene fusion. Using the same criteria as in the discovery cohort, the B-other cases in the validation cohort could be subdivided into *DUX4*-rearranged BCP ALL ($n = 20$), 'B-other, with fusion' ($n = 14$), 'B-other, without fusion' ($n = 7$), Ph-like ($n = 4$; Supplementary Fig. 9b) or

ETV6-RUNX1-like ($n = 4$; Supplementary Fig. 10). Within the 'B-other, with fusion' group, ten cases harboured in-frame fusions previously described in BCP ALL: *EP300-ZNF384* ($n = 3$), *PAX5-FOXP1* ($n = 2$), *P2RY8-CRLF2* ($n = 2$, one of these cases also harboured *PAX5-FOXP1*), *PAX5-DACH1* ($n = 1$), *PAX5-ETV6* ($n = 1$), *TCF3-HLF* ($n = 1$) and *TCF3-ZNF384* ($n = 1$). Three of the remaining cases had a novel in-frame *MEF2D-HNRNPUL1* gene fusion and one case expressed a novel *MED12-HOXA9*. Hence, the majority of cases in the 'B-other, with fusion' group express recurrent gene fusions. Further studies are required to establish if these can be further stratified into biologically and clinically meaningful subtypes. However, we note that cases with fusions affecting each of the genes *ZNF384*, and *MEF2D* formed distinct but separate expression clusters by unsupervised hierarchical clustering, thus outlining possible subtypes characterized by similar gene fusions (Supplementary Figs 4, 10).

1.5 Gene fusions in established genetic subgroups

Most of the established genetic BCP ALL subgroups are based on recurrent gene fusions such as *BCR-ABL1*, *ETV6-RUNX1*, *TCF3-PBX1*, and *MLL* fusions. In the discovery cohort, the presence of these fusions had been ascertained by routine diagnostic analyses. By RNA-seq we could confirm these known gene fusions or their reciprocal variants in 77/81 (95%) cases (Supplementary Fig. 11). This implies that the RNA-seq analysis provided a relatively complete overview of the entire fusion-gene landscape, including also the novel identified fusions. The four instances of known gene fusions that could not be confirmed by RNA-seq were presumably caused by low expression of the fusion or rearrangements too complex for the analysis pipeline to elucidate.

High-hyperdiploid cases showed a notable lack of fusion genes, with only 2/58 cases in the discovery cohort harbouring in-frame fusion genes (3%, $P < 0.001$, Fisher's exact test; an additional three cases carried out-of-frame fusions), in accordance with our recent findings from WGS[11] (Figs 1c,d and 2). It was also uncommon for cases with *BCR-ABL1* ($n = 6$), *TCF3-PBX1* ($n = 13$) and *MLL* fusions ($n = 14$) to have additional in-frame or out-of-frame gene fusions, the only examples being the in-frame fusions *IGH-CRLF2* (in case 79 with *BCR-ABL1*) and *ZCCHC7-PAX5* (in case 102 with *MLL-GAS7*; Fig. 1c and Supplementary Data 3). In contrast, among the *ETV6-RUNX1*-positive cases, 6/48 cases (13%) harboured in-frame fusions besides *ETV6-RUNX1* and its reciprocal variant, and 11/48 cases (23%) had out-of-frame fusions (Fig. 1c–e); the most

commonly affected genes were *ETV6* ($n = 3$) and *RUNX1* ($n = 10$). Two of the *ETV6* fusions (with *CDKN1B* at 12p13 in #6 and *PTPRO* at 12p12 in #131; Supplementary Data 9) were formed by deletions affecting the *ETV6* allele not taking part in in the *ETV6-RUNX1* fusion. All ten *RUNX1* fusions had *RUNX1* as the 5'-partner gene and occurred in *ETV6-RUNX1*-positive cases lacking the reciprocal *RUNX1-ETV6* transcript (Supplementary Fig. 11), suggesting that they arose together with the *ETV6-RUNX1* fusion through a three-way translocation.

To characterize further the *RUNX1* fusions at the genomic level, 5/10 *ETV6-RUNX1*-positive cases containing additional *RUNX1* fusions were analysed by MP-WGS. These analyses confirmed the *RUNX1* fusions at the DNA level (Supplementary Data 6, 9 and 10) and revealed that the genomic breakpoints were in close proximity to the *RUNX1* breakpoints in the *ETV6-RUNX1* fusion, consistent with the presence of complex translocations. Such complex translocations have previously been detected in the *ETV6-RUNX1*-positive cases by FISH and targeted sequencing[29,30]. Only one *RUNX1* fusion contained an undisrupted active domain from the partner gene; thus, the fusions typically resulted in disruption of the 3'-partner gene (Supplementary Fig. 12).

1.6 Fusion-gene network analysis

To ascertain the pattern of gene fusions in BCP ALL, we performed a fusion-gene network analysis[31] of the 58 unique in-frame gene fusions identified across the discovery and validation cohorts (Supplementary Fig. 13a). This analysis revealed that 15 genes (*BCR*, *CRLF2*, *DUX4*, *ETV6*, *IGH*, *IKZF1*, *JAK2*, *LDLRAD4*, *MEF2D*, *MLL*, *PAX5*, *RUNX1*, *TCF3*, *ZCCHC7* and *ZNF384*) were recurrently involved in chimeras (Supplementary Fig. 13a). A comparison with literature data[25] highlighted that the high frequencies of fusions involving *RUNX1*, *DUX4*, *IKZF1* and *LDLRAD4* were novel findings (Supplementary Fig. 13a,b). The *RUNX1* fusions were typically found in *ETV6-RUNX1*-positive cases, most likely arising through complex translocations as described above, and the *DUX4* fusions were identified in the novel BCP ALL subgroup described in this study.

IKZF1, encoding IKAROS, is known to be perturbed by deletions (15% of BCP ALL cases) and occasionally sequence mutations (2–6%; refs 5,32,33), but has previously never been described to fuse with other genes in BCP ALL. In the discovery cohort, the two in-frame fusions *SETD5-IKZF1* (#85 with *ETV6-RUNX1*-like gene expression) and *IKZF1-NUTM1* (#151, 'B-other, with fusion') retained functional domains from *IKZF1* (Supplementary Fig. 12e,f). *IKZF1-*

NUTM1 also contained essentially the entire coding region of *NUTM1*, akin to other *NUTM1* fusions in midline carcinoma[34] and in *MLL*-negative infant ALL[13]. The two out-of-frame fusions (*IKZF1-CDK2*, #64; and *IKZF1-TRPV2*, #9) contained no functional domains from *IKZF1* and, hence likely abolished the function of IKAROS. Thus, *IKZF1* fusions represent a novel mechanism for disrupting *IKZF1* in BCP ALL.

Aberrations in *LDLRAD4*, encoding a negative regulator of transforming growth factor-β signalling, have previously not been described in leukaemia. We identified two in-frame fusions involving this gene: *LDLRAD4-PHACTR3* (#70 with Ph-like gene expression) and *RUNX1-LDLRAD4* (#187 with *ETV6-RUNX1*). Both fusions retained the LDL-receptor class-A domain in the N-terminal region of the LDLRAD4 protein, whereas the SMAD interaction motif required for the regulation of transforming growth factor-β signalling was only retained in *RUNX1-LDLRAD4* (Supplementary Fig. 12a,d).

1.7 Intragenic splice variants and subtype classification

Somatic intragenic deletions are frequent in BCP ALL and result in the expression of truncated transcripts predicted to encode internally deleted proteins[35]. To investigate if we could identify truncated transcripts associated with the most common intragenic deletions in BCP ALL (*CDNK2A*, *PAX5*, *ETV6* and *IKZF1*)[35], we developed a novel relative splice junction quantification algorithm. This algorithm identified five truncated transcript variants affecting *ETV6*, *PAX5* and *IKZF1*, with a total of 25 (13%) BCP ALL cases in the discovery cohort harbouring at least one truncated transcript (Supplementary Fig. 14). Focal deletions concordant with the truncating transcripts were present in 15/20 cases (75%) with available SNP array data. In five cases, the truncated transcript occurred without evidence of a focal deletion, indicating either the presence of subclonal deletions below the detection level of the SNP array analysis or aberrant splicing caused by other mutational mechanisms.

Our detailed RNA-seq data also allowed analyses of splicing events occurring over the fusion breakpoints of the clinically important gene fusions *BCR-ABL1*, *ETV6-RUNX1*, *TCF3-PBX1* and *MLL* fusions, revealing a substantial heterogeneity in exon usage around the fusion breakpoints (Fig. 6a–d and Supplementary Figs 15–18); particularly for *ETV6-RUNX1* where the main variant joined exon 5 of *ETV6* with exon 2 of *RUNX1*, whereas alternative forms fused with exon 3 of *RUNX1* or a cryptic exon within intron 1 (Fig. 6b and Supplementary Fig. 16).

These were either observed together as splice variants or as single forms in individual cases; the alternative variants did not affect the runt domain of *RUNX1*. An alternative breakpoint joining exon 4 of *ETV6* with exon 2 of *RUNX1*, as has previously been described[36], was identified in 2/48 (4%) *ETV6-RUNX1*-positive cases (Supplementary Fig. 16).

Global gene-expression profiling by microarrays can discern between the genetic subtypes of BCP ALL, although with less than perfect accuracy[37–40]. We therefore constructed a classifier utilizing both gene-fusion and gene-expression data from RNA-seq. This classifier showed improved sensitivity (correctly classifying 180/195 cases, 92%) compared with a classifier based on the gene-expression data alone (correctly classifying 174/195 cases, 89%; Supplementary Fig. 19).

1.8 Mutational analysis

The mutational landscape of single-nucleotide variants in a larger number of genes has not been studied in an unselected series of BCP ALL. Because RNA-seq allows for the identification of expressed mutant alleles, we examined hotspot regions of 16 recurrently mutated genes in BCP ALL (representing 70% of all genes described to be mutated in more than 2 BCP ALL cases), ascertained in previous studies[9,11–13] or COSMIC[41] (Supplementary Data 11). This analysis revealed 56 mutations in 47 BCP ALLs, with genes in the RTK-RAS signalling pathway being the most commonly mutated: *NRAS* (23/195, 12%), *FLT3* (7/195, 4%), *PTPN11* (6/195, 3%) and *KRAS* (3/195, 2%; Supplementary Data 11). We also had genomic mutation data from 61 of the cases from WES ($n = 22$), WGS ($n = 12$), both WES and WGS ($n = 1$), or Sanger sequencing ($n = 26$; refs 11,42). We observed good concordance between hotspot mutations identified by RNA-seq and the genomic data, although some mutations in *KRAS* ($n = 5$) and *FLT3* ($n = 4$) observed at the DNA level escaped detection at the transcriptional level. The mutational spectra differed between subtypes, with *NRAS* and *KRAS* mutations being enriched in high-hyperdiploid cases[43], and *CRLF2*, *JAK2* and *IL7R* mutations in Ph-like ALL cases[7,8] (Fig. 2).

2 Discussion

Gene fusions are strong driver mutations in neoplasia, and have provided fundamental insights into the disease mechanisms involved in tumourigenesis. In

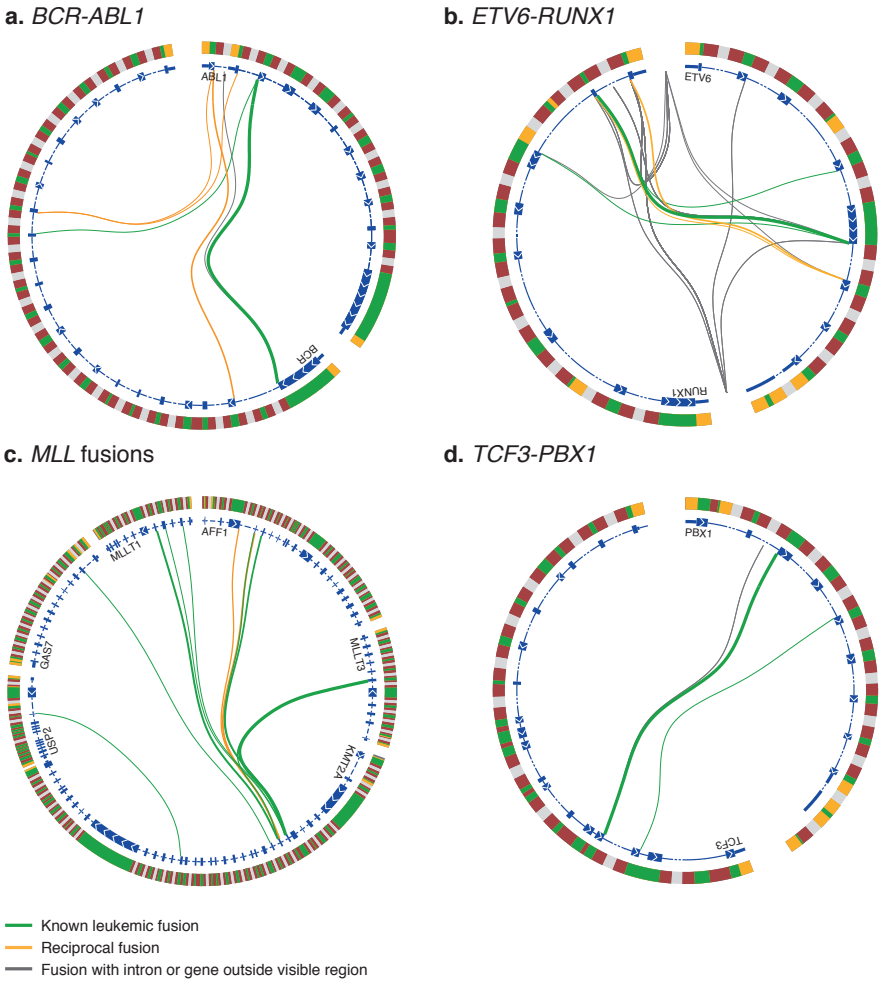


Figure 6: **Splice patterns over fusion breakpoints.** Illustration of all detected fusion breakpoints in BCP ALL cases with (**a**) *BCR-ABL1*, (**b**) *ETV6-RUNX1*, (**c**) *MLL* fusions, and (**d**) *TCF3-PBX1*. Genes are arranged clockwise by genomic position. The outer circle represents the genomic region encompassing the indicated genes. Yellow indicates untranslated regions, green indicates coding exons, and red and grey indicate intronic regions (the latter are not to scale). The inner circle represents one or two overlaid reference transcripts of the indicated gene. Coding exons are indicated by a thick line with white arrows indicating the direction of the gene, introns are indicated by a thin or dashed line and untranslated regions are indicated by a medium thick line. Connecting lines between transcripts illustrate fusion breakpoints detected by at least three (for *BCR-ABL1*, *ETV6-RUNX1* and *MLL* fusions) or ten reads (for *TCF3-PBX1*). Fusion breakpoints in individual BCP ALL cases are depicted in Supplementary Figs 15–18.

addition, they are increasingly used for diagnostic purposes, risk stratification and disease follow-up, and several chimeric proteins encoded by gene fusions serve as specific targets for treatment[31].

We here describe the gene-fusion landscape of paediatric BCP ALL, and show that the majority of cases (65%) express in-frame gene fusions, including most B-other cases (82%) previously described to lack specific genetic changes. The notable exception was high-hyperdiploid cases where only 3% of cases harboured an in-frame fusion gene. The low number of in-frame fusions in this group, however, highlights that the background level of gene-fusion generation in BCP ALL is low. Indeed, the median number of fusion genes per case in this study was 1 for all major subtypes (> 3 cases) apart from high-hyperdiploid cases, showing that additional fusion genes are rarely needed for leukemogenesis. In *ETV6-RUNX1*-positive cases, however, additional fusions of unclear pathogenetic importance were present in 35% of cases. These typically involved *ETV6* or *RUNX1*. For the latter gene the fusions were generated through three-way translocations also creating the *ETV6-RUNX1* fusion.

We demonstrate, for the first time, that 16% of B-other cases (4% of BCP ALL) harboured rearrangements involving the *DUX4* gene. The frequency of such rearrangements differed between the discovery and validation cohorts; something that could possibly be explained by the higher mean age of the latter (7.1 versus 6.1 years). However, the true incidence of *DUX4* rearrangements in childhood BCP ALL needs to be further assessed in larger patient cohorts. The rearrangements resulted in fusions between *IGH* and *DUX4*, or less commonly, *ERG* and *DUX4*, causing aberrant *DUX4* expression. *DUX4* has previously only been reported to be rearranged in round-cell sarcomas, forming a recurrent *CIC-DUX4* fusion gene. That fusion, however, only includes a small C-terminal part of DUX4, not including the two homeobox domains[44], and is therefore likely to be functionally different from the fusions described here. Notably, all cases with *DUX4* rearrangements described herein displayed a gene-expression signature matching that of a subgroup of BCP ALL reported to be associated with frequent *ERG* deletions[6]. *DUX4* encodes a transcription factor normally expressed in germ cells that regulates the expression of genes involved in germline and early stem cell development[18,45]. Hence, it is tempting to speculate that the aberrant expression of DUX4 in the rearranged cases cause activation of transcriptional programmes that normally are expressed during early stem cell development. In contrast to the *ERG* deletions, *DUX4* rearrangements were present in all cases with the characte-

ristic gene-expression pattern, implying that *DUX4* rearrangements constitute the founder event of this subtype and that *ERG* deletions are secondary cooperating events.

Global gene-expression profiling is a powerful tool to identify leukaemias with similar mutational backgrounds, as exemplified by the Ph-like subtype: such cases were initially identified as having gene-expression patterns similar to those of Ph-positive BCP ALL cases[4–6] and were only later identified as being characterized by genetic alterations that activate kinase or cytokine receptor signalling[7,8]. Within the B-other group we identified a second novel subtype, consisting of cases with a gene-expression profile similar to that of *ETV6-RUNX1*-positive cases but lacking this fusion gene. Instead, they harboured lesions affecting both *ETV6* and *IKZF1* in all cases with ascertainable data. We termed this subtype *ETV6-RUNX1*-like ALL. In contrast to cases with *ETV6-RUNX1*-like gene expression that have been reported in literature, but where cryptic *ETV6-RUNX1* were not excluded[46], we performed extensive genetic analyses to rule out a cryptic *ETV6-RUNX1* rearrangement. Thus, we propose that combined *ETV6* and *IKZF1* lesions together may activate similar transcriptional programmes as the ETV6-RUNX1 fusion protein.

The *DUX4*-rearranged and *ETV6-RUNX1*-like subtypes together with the well-established subgroup of Ph-like BCP ALL[4–8] accounted for 59 and 71% of B-other cases in the discovery and validation cohorts, respectively. Of the remaining B-other cases in the two cohorts, 74% expressed rare previously reported, or novel in-frame gene fusions, many of which contained genes with recurrent alterations in BCP ALL[25,41]. Again, these findings illustrate that paediatric BCP ALL, with the exception of the high-hyperdiploid, near-haploid and low-hypodiploid subgroups[11,12], is characterized by the presence of fusion genes. Because many gene fusions will be rare or even private, our study reinforces that RNA-seq may be a powerful tool for unbiased screening of fusion genes in a clinical setting with an unmatched power to detect novel but targetable gene fusions in BCP ALL[8,47].

In conclusion, this study provides a detailed view of the fusion gene landscape in paediatric BCP ALL, identifying several new gene fusions as well as distinct subgroups of BCP ALL. Apart from increasing our understanding of the pathogenesis of paediatric BCP ALL, this may help improve risk stratification and eventually increase the therapeutic options for this most common form of childhood malignancy.

3 Methods

3.1 Patients

Between January 1992 and January 2013, 283 paediatric (< 18 years) BCP ALL cases were analysed as part of clinical routine diagnostics at the Department of Clinical Genetics, University and Regional Laboratories Region Skåne, Lund, Sweden. Of these, RNA or material suitable for RNA extraction from bone marrow ($n = 171$) or peripheral blood ($n = 24$) taken at diagnosis was available from 195 (69%) cases, comprising the discovery cohort. The vast majority was treated according to the Nordic Society of Paediatric Haematology and Oncology (NOPHO) ALL 1992, 2000 or 2008 protocols[48]. There were no significant differences in gender or age distribution between cases where RNA-seq could or could not be performed; however, cases analysed by RNA-seq had higher white blood cell counts (median $9.85 \times 10^9 \mathrm{l}^{-1}$, range $0.9–802 \times 10^9 \mathrm{l}^{-1}$ versus median $5.5 \times 10^9 \mathrm{l}^{-1}$, range $0.8–121 \times 10^9 \mathrm{l}^{-1}$; $P = 0.003$; two-sided Mann-Whitney's U-test). The validation cohort consisted of 49 paediatric BCP ALL cases treated according to the Berlin-Frankfurt-Münster (BFM) 2000 protocol[49]. All cases in the validation cohort were tested for *BCR-ABL1*, *ETV6-RUNX1*, *TCF3-PBX1*, *MLL* rearrangements, and high hyperdiploidy in accordance with the treatment protocol, and were found negative for these aberrations. Informed consent was obtained according to the Declaration of Helsinki and the study was approved by the Ethics Committee of Lund University.

3.2 RNA sequencing

The cDNA sequencing libraries were prepared from poly-A selected RNA using the Truseq RNA library preparation kit v2 (Illumina) according to the manufacturer's instructions, but with a modified RNA fragmentation step lowering the incubation time at 94 °C from 8 min to 10 s to allow for longer RNA fragments. The cDNA libraries were sequenced using a HiScanSQ (Illumina) or NextSeq 500 (Illumina).

3.3 Gene-fusion detection

Gene fusions were detected by combining three methods. Novel fusions were detected by Chimerascan[50] (0.4.5) and TopHat-Fusion-post[51] (2.0.7), followed by

a custom filter strategy and validation by RT-PCR. Known fusion transcripts of *BCR-ABL1*, *ETV6-RUNX1*, *TCF3-PBX1* and *MLL* fusions were detected by aligning all reads to a reference consisting of the known fusion transcripts and normal transcript variants of the genes, and counting reads uniquely aligned to the fusion transcripts. *IGH-CRLF2* fusions were detected by identifying cases that had > 50 reads within a 65-kb region surrounding *CRLF2* paired to a read within the *IGH* locus; these fusions were then validated using FISH. The following filter strategy was used for selection of validation candidates from Chimerascan and TopHat-Fusion-post results: all fusions reported by Chimerascan to be supported by ten or more reads over the fusion junction or > 50 total reads, all fusions reported by TopHat-Fusion-post to be supported by > 15 reads covering the fusion junction, and remaining interchromosomal fusions detected by Chimerascan that were also detected by TopHat-Fusion-post were included for validation, unless: (1) Chimerascan annotated the event as 'Read through'; (2) the affected exons had $> 75\%$ of reads mapped with a quality score below five; (3) reads supporting the same fusion were detected (by either TopHat or Chimerascan) in one of 20 sorted normal bone marrow samples; (4) the fusion indicated rearrangement within an *IG* or *TCR* locus or involved two HLA genes (the latter were presumed to represent normal constitutional HLA variants); (5) the fusion involved two non-coding genes; or (6) the constituent genes were located less than 10kb apart.

In addition, remaining fusions detected by either Chimerascan or TopHat-Fusion-post were included if the fusion (1) was reciprocal to a fusion passing the above filters or a previously reported fusion in BCP ALL; or (2) contained one of the recurrently altered genes *ETV6*, *RUNX1*, *MLL*, *PAX5* or *IKZF1*.

3.4 Gene-expression analysis

The raw unfiltered RNA-seq reads were aligned to human reference genome hg19 using TopHat 2.0.7, with the parameters --fusion-search and --bowtie1 to enable fusion detection. Gene-expression values were calculated as fragments per kilobase of transcript per million reads (fpkm) using Cufflinks 2.2.0 (ref. 52). Hierarchical clustering and principal component analyses were performed using Qlucore Omics Explorer (v3.1; Qlucore, Lund, Sweden). In brief, the data were normalized to a mean of 0 and a variance of 1. Hierarchical clustering of both samples and variables was performed using Euclidean distance and average linkage.

3.5 Genomic sequencing analyses

For 11 cases in the discovery cohort, whole-exome libraries were prepared from diagnostic and follow-up samples using the Nextera Rapid Capture Exome Kit (Illumina) according to the manufacturer's instructions. Paired 2×151 bp reads were produced from the exome libraries using a NextSeq 500 (Illumina). The reads were aligned to human reference genome hg19 using BWA 0.7.9a (ref. 53) and PCR duplicate reads were filtered out using SAMBLASTER[54]. Somatic variant calling was performed using Strelka[55]. For 15 cases in the discovery cohort, MP-WGS libraries were prepared using the Nextera Mate Pair Library Preparation Kit (Illumina). Paired 2×76 bp reads were produced from the mate-pair libraries using a NextSeq 500 (Illumina). The reads were aligned to human reference genome hg19 using BWA 0.7.9a (ref. 53) and PCR duplicate reads were filtered out using SAMBLASTER[54]. For 24 high-hyperdiploid cases in the discovery cohort, extensive characterization using WES ($n = 11$), WGS ($n = 12$) or both ($n = 1$) has been previously described[11].

3.6 Identification of leukaemia-specific splice variants

Splicing differences between samples were characterized by ascertaining the relative frequencies of splice junction usage across all observed splice donor and acceptor sites, from reads aligned by TopHat. All intragenic splice junctions that were supported by at least 10 reads in at least one sample and that involved at least one annotated exon were included. For each splice donor or acceptor site, alternative splicing was quantified by measuring the fraction of reads supporting each observed splice junction containing that site. If a splice acceptor or donor site was not covered by any reads within a sample, the corresponding variables were treated as missing values and reconstructed as the average value of samples that had data for the site. From these data, all splice variants in *CDKN2A*, *PAX5*, *ETV6* and *IKZF1* that were not present in a reference transcript and not detected in one of 20 normal bone marrow populations (sorted from four donors) were included in the analysis.

3.7 Gene set enrichment analysis

GSEA was performed on gene-expression data obtained from the RNA-seq analysis, using Qlucore Omics Explorer (v3.1). Signal-to-noise ratio was used as ranking metrics for analysing curated gene-ontology gene sets (C5) acquired from the

Molecular Signatures Database (MSigDB). Gene sets with < 15 or > 500 genes were excluded. Enriched gene sets after 1,000 permutations at an false-discovery rate of < 0.25 and a nominal $P < 0.05$ were considered as significant.

3.8 Support vector machine classification

A classifier based on the gene-expression and gene-fusion data was created to categorize the samples into the subtypes *BCR-ABL1*, *ETV6-RUNX1*, high hyperdiploidy, *MLL*, *TCF3-PBX1*, and 'B-other + rare subgroups'. The subtypes were considered to be mutually exclusive. First, one-versus-all support vector machine classifiers[56] with linear kernels were created for all subtypes. They were based on the $\log_2$ transformed gene-expression data after variable selection by removal of variables with low variance across the samples. The threshold was set to a standard deviation of 0.29, resulting in 583/23,285 variables (2.5%) being used. Next, the classifiers were augmented by the detected gene fusions. If a gene fusion corresponding to one of the subgroups was found in a sample, it was classified as belonging to that subgroup regardless of the expression profile. These samples were treated as having $\pm\infty$ distance to the support vector machine classification hyperplane. Finally, a multiclass classifier was created from all the one-versus-all classifiers, by selecting the class that had the lowest signed distance between the sample and the classification hyperplane. The performance of the multiclass and all binary subgroup classifiers was evaluated by leave-one-out cross-validation.

3.9 RNA-Seq mutation calling

The raw unfiltered reads were aligned to human reference genome hg19 using STAR 2.4.0j (ref. 57). Putative mutations within hotspot regions of 16 genes were identified using VarScan 2.3.7 (ref. 58). The variants were annotated using Annovar[59] and known constitutional variants were excluded from the list.

3.10 RT-PCR and Sanger sequencing

For gene-fusion validation, primer3 was used to design primers for amplifying a region larger than 200bp covering the fusion breakpoint. Reverse transcription was performed using M-MLV (Thermo Fischer Scientific) and PCR was performed using Platinum Taq (Thermo Fischer Scientific). The PCR products were purified using Exosap-it (Affymetrix) or Qiaquick gel extraction kit (Qiagen) and

then Sanger sequenced by a commercial sequencing service provider. RT-PCR for detection of truncated *ERG* transcripts was performed using primers previously described[14]. Sanger sequencing of *FLT3*, *NRAS*, *KRAS* and *PTPN11* in 26 high-hyperdiploid cases in the discovery cohort was performed using primers described in Supplementary Table 2. This data has been published previously[42].

3.11 SNP array analysis

SNP array analysis was performed on DNA extracted from bone marrow or peripheral blood at diagnosis for 156 BCP ALL cases. The analysis was performed using HumanOmni1-Quad and Human1M-Duo array systems (Illumina) with data analysis using Genomestudio 2011.1 (Illumina). The SNP array data has been previously published[3].

3.12 Statistical methods

Two-sided P values were calculated using Fisher's exact test or Mann-Whitney's U-test. P-values of < 0.05 were considered statistically significant.

3.13 Data availability

RNA-seq and MP-WGS data have been deposited at the European Genome-phenome Archive (EGA), under the accession code EGAS00001001795. WES and WGS data are available for academic purposes by contacting the corresponding author, as the patient consent does not cover depositing data that can be used for large-scale determination of germline variants.

Acknowledgements

This work was supported by the Swedish Cancer Society, the Swedish Childhood Cancer Foundation, the Swedish Research Council, the Knut and Alice Wallenberg Foundation, the Inga-Britt and Arne Lundberg Foundation, the Gunnar Nilsson Cancer Foundation, the Medical Faculty of Lund University and Governmental Funding of Clinical Research within the National Health Service. We thank Birthe Fedders from the ALL-BFM laboratory and Andrea Biloglav from the Division of Clinical Genetics, Lund, for expert technical assistance.

Author contributions

H.L. and T.F. conceived the project. H.L., A.H.-W., L.O., C.O.-P. M.A. and M.R. performed the experiments. H.L., R.H., A.H.-W., L.O., C.O.-P., S.v.P., F.M., B.J., K.P., A.K.A., M.F. and T.F. analysed the data. M.S., G.C., A.C. C.J.H.P. and M.B. provided samples and clinical data. H.L., K.P., A.K.A. and T.F. wrote the manuscript, which was reviewed and edited by the other co-authors.

Additional information

Supplementary Information accompanies this paper at `http://www.nature.com/naturecommunications`

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at `http://npg.nature.com/reprintsandpermissions/`

How to cite this article: Lilljebjörn, H. *et al.* Identification of *ETV6-RUNX1*-like and *DUX4*-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nat. Commun.* 7:11790 doi: 10.1038/ncomms11790 (2016).

Bibliography

[1] Ching-Hon Pui and William E Evans. Treatment of acute lymphoblastic leukemia. *New England Journal of Medicine*, 354(2):166–178, 2006.

[2] Giovanni Martinelli, Ilaria Iacobucci, Clelia Tiziana Storlazzi, Marco Vignetti, Francesca Paoloni, Daniela Cilloni, Simona Soverini, Antonella Vitale, Sabina Chiaretti, Giuseppe Cimino, et al. IKZF1 (Ikaros) deletions in BCR-ABL1–positive acute lymphoblastic leukemia are associated with short disease-free survival and high rate of cumulative incidence of relapse: a GIMEMA AL WP report. *Journal of Clinical Oncology*, 27(31):5202–5207, 2009.

[3] Linda Olsson, Anders Castor, M Behrendtz, Andrea Biloglav, Erik Forestier, Kajsa Paulsson, and Bertil Johansson. Deletions of IKZF1 and SPRED1 are associated with poor prognosis in a population-based series of pediatric B-cell precursor acute lymphoblastic leukemia diagnosed between 1992 and 2011. *Leukemia*, 28(2):302, 2014.

[4] Monique L Den Boer, Marjon van Slegtenhorst, Renée X De Menezes, Meyling H Cheok, Jessica GCAM Buijs-Gladdines, Susan TCJM Peters, Laura JCM Van Zutven, H Berna Beverloo, Peter J Van der Spek, Gaby Escherich, et al. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. *Lancet Oncology*, 10(2):125–134, 2009.

[5] Charles G Mullighan, Xiaoping Su, Jinghui Zhang, Ina Radtke, Letha AA Phillips, Christopher B Miller, Jing Ma, Wei Liu, Cheng Cheng, Brenda A Schulman, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. *New England Journal of Medicine*, 360(5):470–480, 2009.

[6] Richard C Harvey, Charles G Mullighan, Xuefei Wang, Kevin K Dobbin, George S Davidson, Edward J Bedrick, I-Ming Chen, Susan R Atlas, Huining Kang, Kerem Ar, et al. Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome. *Blood*, 116(23):4874–4884, 2010.

[7] Kathryn G Roberts, Ryan D Morin, Jinghui Zhang, Martin Hirst, Yongjun Zhao, Xiaoping Su, Shann-Ching Chen, Debbie Payne-Turner, Michelle L Churchman, Richard C Harvey, et al. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. *Cancer Cell*, 22(2):153–166, 2012.

[8] Kathryn G Roberts, Yongjin Li, Debbie Payne-Turner, Richard C Harvey, Yung-Li Yang, Deqing Pei, Kelly McCastlain, Li Ding, Charles Lu, Guangchun Song, et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. *New England Journal of Medicine*, 371(11):1005–1015, 2014.

[9] Elli Papaemmanuil, Inmaculada Rapado, Yilong Li, Nicola E Potter, David C Wedge, Jose Tubio, Ludmil B Alexandrov, Peter Van Loo, Susanna L Cooke, John Marshall, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nature Genetics*, 46(2):116–125, 2014.

[10] Ute Fischer, Michael Forster, Anna Rinaldi, Thomas Risch, Stéphanie Sungalee, Hans-Jörg Warnatz, Beat Bornhauser, Michael Gombert, Christina Kratsch, Adrian M Stütz, et al. Genomics and drug profiling of fatal TCF3-HLF-positive acute lymphoblastic leukemia identifies recurrent mutation patterns and therapeutic options. *Nature Genetics*, 47(9):1020–1029, 2015.

[11] Kajsa Paulsson, Henrik Lilljebjörn, Andrea Biloglav, Linda Olsson, Marianne Rissler, Anders Castor, Gisela Barbany, Linda Fogelstrand, Ann Nordgren, Helene Sjögren, et al. The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Nature Genetics*, 47(6):672–676, 2015.

[12] Linda Holmfeldt, Lei Wei, Ernesto Diaz-Flores, Michael Walsh, Jinghui Zhang, Li Ding, Debbie Payne-Turner, Michelle Churchman, Anna Andersson, Shann-Ching Chen, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nature Genetics*, 45(3):242–252, 2013.

[13] Anna K Andersson, Jing Ma, Jianmin Wang, Xiang Chen, Amanda Larson Gedman, Jinjun Dang, Joy Nakitandwe, Linda Holmfeldt, Matthew Parker, John Easton, et al. The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. *Nature Genetics*, 47(4):330–337, 2015.

[14] Emmanuelle Clappier, Marie-Françoise Auclerc, Jérôme Rapion, Marleen Bakkus, Aurélie Caye, Ahlème Khemiri, Claudine Giroux, Lucie Hernandez, Emmanuelle Kabongo, Suvi Savola, et al. An intragenic ERG deletion is a marker of an oncogenic subtype of B-cell precursor acute lymphoblastic

leukemia with a favorable outcome despite frequent IKZF1 deletions. *Leukemia*, 28(1):70, 2014.

[15] M Zaliova, O Zimmermannova, P Dörge, C Eckert, A Möricke, M Zimmermann, J Stuchly, A Teigler-Schlegel, B Meissner, R Koehler, et al. ERG deletion is associated with CD2 and attenuates the negative impact of IKZF1 deletion in childhood acute lymphoblastic leukemia. *Leukemia*, 28(1):182, 2014.

[16] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9): 1639–1645, 2009.

[17] Richard JLF Lemmers, Patrick J Van Der Vliet, Rinse Klooster, Sabrina Sacconi, Pilar Camaño, Johannes G Dauwerse, Lauren Snider, Kirsten R Straasheijm, Gert Jan Van Ommen, George W Padberg, et al. A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science*, 329(5999): 1650–1653, 2010.

[18] Lauren Snider, Linda N Geng, Richard JLF Lemmers, Michael Kyba, Carol B Ware, Angelique M Nelson, Rabi Tawil, Galina N Filippova, Silvère M van der Maarel, Stephen J Tapscott, et al. Facioscapulohumeral dystrophy: incomplete suppression of a retrotransposed gene. *PLoS Genetics*, 6(10): e1001181, 2010.

[19] Ching-Hon Pui, Leslie L Robison, and A Thomas Look. Acute lymphoblastic leukaemia. *Lancet*, 371(9617):1030–1043, 2008.

[20] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[21] Birte Niebuhr, Neele Kriebitzsch, Meike Fischer, Kira Behrens, Thomas Günther, Malik Alawi, Ulla Bergholz, Ursula Müller, Susanne Roscher, Marion Ziegler, et al. Runx1 is essential at two stages of early murine B-cell development. *Blood*, 122(3):413–423, 2013.

[22] Tanja A Schwickert, Hiromi Tagoh, Sinan Gültekin, Aleksandar Dakic, Elin Axelsson, Martina Minnich, Anja Ebert, Barbara Werner, Mareike Roth, Luisa Cimmino, et al. Stage-specific control of early B cell development by the transcription factor Ikaros. *Nature Immunology*, 15(3):283–293, 2014.

[23] Petra Dörge, Barbara Meissner, Martin Zimmermann, Anja Möricke, André Schrauder, Jean-Pierre Bouquin, Denis Schewe, Jochen Harbott, Andrea Teigler-Schlegel, Richard Ratei, et al. IKZF1 deletion is an independent predictor of outcome in pediatric acute lymphoblastic leukemia treated according to the ALL-BFM 2000 protocol. *Haematologica*, 98(3):428–432, 2013.

[24] Arian van der Veer, Esmé Waanders, Rob Pieters, Marieke E Willemse, Simon V Van Reijmersdal, Lisa J Russell, Christine J Harrison, William E Evans, Vincent HJ van der Velden, Peter M Hoogerbrugge, et al. Independent prognostic value of BCR-ABL1-like signature and IKZF1 deletion, but not high CRLF2 expression, in children with B-cell precursor ALL. *Blood*, 122(15):2622–2629, 2013.

[25] Felix Mitelman, Bertil Johansson, and Fredrik Mertens (Eds.). Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer, 2016. URL `http://cgap.nci.nih.gov/Chromosomes/Mitelman`.

[26] Yoshihiro Gocho, Nobutaka Kiyokawa, Hitoshi Ichikawa, Kazuhiko Nakabayashi, Tomoo Osumi, Takeshi Ishibashi, Hiroo Ueno, Kazuki Terada, Keisuke Oboki, Hiromi Sakamoto, et al. A novel recurrent EP300-ZNF384 gene fusion in B-cell precursor acute lymphoblastic leukemia. *Leukemia*, 29 (12):2445, 2015.

[27] Jeremy Clark, Yong-Jie Lu, Sanjiv K Sidhar, Chris Parker, Sandra Gill, Damian Smedley, Rifat Hamoudi, W Marston Linehan, Janet Shipley, and Colin S Cooper. Fusion of splicing factor genes PSF and NonO (p54 nrb) to the TFE3 gene in papillary renal cell carcinoma. *Oncogene*, 15(18), 1997.

[28] Yusuke Sato, Tetsuichi Yoshizato, Yuichi Shiraishi, Shigekatsu Maekawa, Yusuke Okuno, Takumi Kamura, Teppei Shimamura, Aiko Sato-Otsubo, Genta Nagae, Hiromichi Suzuki, et al. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nature Genetics*, 45(8):860–867, 2013.

[29] Mary Martineau, G Reza Jalali, Kerry E Barber, Zoë J Broadfield, Kan Luk Cheung, John Lilleyman, Anthony V Moorman, Sue Richards, Hazel M Robinson, Fiona Ross, et al. ETV6/RUNX1 fusion at diagnosis and relapse: some prognostic indications. *Genes, Chromosomes and Cancer*, 43(1):54–71, 2005.

[30] Yanliang Jin, Xingwei Wang, Shaoyan Hu, Jingyan Tang, Benshang Li, and Yihuan Chai. Determination of ETV6-RUNX1 genomic breakpoint by next-generation sequencing. *Cancer Medicine*, 5(2):337–351, 2016.

[31] Fredrik Mertens, Bertil Johansson, Thoas Fioretos, and Felix Mitelman. The emerging complexity of gene fusions in cancer. *Nature Reviews Cancer*, 15 (6):371–381, 2015.

[32] Linda Olsson, Ferras Albitar, Anders Castor, Mikael Behrendtz, Andrea Biloglav, Kajsa Paulsson, and Bertil Johansson. Cooperative genetic changes in pediatric B-cell precursor acute lymphoblastic leukemia with deletions or mutations of IKZF1. *Genes, Chromosomes and Cancer*, 54(5):315–325, 2015.

[33] Linda Olsson and Bertil Johansson. Ikaros and leukaemia. *British Journal of Haematology*, 169(4):479–491, 2015.

[34] Panagis Filippakopoulos and Stefan Knapp. Targeting bromodomains: epigenetic readers of lysine acetylation. *Nature Reviews Drug Discovery*, 13(5): 337, 2014.

[35] Charles G Mullighan, Salil Goorha, Ina Radtke, Christopher B Miller, Elaine Coustan-Smith, James D Dalton, Kevin Girtman, Susan Mathew, Jing Ma, Stanley B Pounds, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*, 446(7137):758, 2007.

[36] Marketa Zaliova, Claus Meyer, Gunnar Cario, Martina Vaskova, Rolf Marschalek, Jan Stary, Jan Zuna, and Jan Trka. TEL/AML1-positive patients lacking TEL exon 5 resemble canonical TEL/AML1 cases. *Pediatric Blood & Cancer*, 56(2):217–225, 2011.

[37] Eng-Juh Yeoh, Mary E Ross, Sheila A Shurtleff, W Kent Williams, Divyen Patel, Rami Mahfouz, Fred G Behm, Susana C Raimondi, Mary V Relling,

Anami Patel, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143, 2002.

[38] Mary E Ross, Xiaodong Zhou, Guangchun Song, Sheila A Shurtleff, Kevin Girtman, W Kent Williams, Hsi-Che Liu, Rami Mahfouz, Susana C Raimondi, Noel Lenny, et al. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, 102(8):2951–2959, 2003.

[39] Frederik W Delft, Tony Bellotti, Zhiyuan Luo, Louise K Jones, Naina Patel, Olga Yiannikouris, Alex S Hill, Mike Hubank, Helena Kempski, Danielle Fletcher, et al. Prospective gene expression analysis accurately subtypes acute leukaemia in children and establishes a commonality between hyperdiploidy and t (12; 21) in acute lymphoblastic leukaemia. *British Journal of Haematology*, 130(1):26–35, 2005.

[40] Anna Andersson, Cecilia Ritz, David Lindgren, Patrik Edén, Carin Lassen, Jesper Heldrup, Tor Olofsson, Johan Råde, Magnus Fontes, A Porwit-Macdonald, et al. Microarray-based classification of a consecutive series of 121 childhood acute leukemias: prediction of leukemic and genetic subtype as well as of minimal residual disease status. *Leukemia*, 21(6):1198, 2007.

[41] Simon A Forbes, Gurpreet Bhamra, Sally Bamford, Elisabeth Dawson, Chai Yin Kok, Jody Clements, Andrew Menzies, Jon W Teague, P Andrew Futreal, and Michael R Stratton. The catalogue of somatic mutations in cancer (COSMIC), 2008.

[42] Kajsa Paulsson, Andrea Horvat, Bodil Strömbeck, Fredrik Nilsson, Jesper Heldrup, Mikael Behrendtz, Erik Forestier, Anna Andersson, Thoas Fioretos, and Bertil Johansson. Mutations of FLT3, NRAS, KRAS, and PTPN11 are frequent and possibly mutually exclusive in high hyperdiploid childhood acute lymphoblastic leukemia. *Genes, Chromosomes and Cancer*, 47(1):26–33, 2008.

[43] Marian Case, Elizabeth Matheson, Lynne Minto, Rosline Hassan, Christine J Harrison, Nick Bown, Simon Bailey, Josef Vormoor, Andrew G Hall, and Julie AE Irving. Mutation of genes affecting the RAS pathway is common in childhood acute lymphoblastic leukemia. *Cancer Research*, 68(16): 6803–6809, 2008.

[44] Miho Kawamura-Saito, Yukari Yamazaki, Keiko Kaneko, Noriyoshi Kawaguchi, Hiroaki Kanda, Hiroyuki Mukai, Takahiro Gotoh, Tohru Motoi, Masashi Fukayama, Hiroyuki Aburatani, et al. Fusion between CIC and DUX4 up-regulates PEA3 family genes in Ewing-like sarcomas with t (4; 19)(q35; q13) translocation. *Human Molecular Genetics*, 15(13):2125–2137, 2006.

[45] Linda N Geng, Zizhen Yao, Lauren Snider, Abraham P Fong, Jennifer N Cech, Janet M Young, Silvere M van der Maarel, Walter L Ruzzo, Robert C Gentleman, Rabi Tawil, et al. DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Developmental Cell*, 22(1):38–51, 2012.

[46] Kazutoshi Iijima, Hiroki Yoshihara, Kentaro Ohki, Motohiro Kato, Takashi Fukushima, Akira Kikuchi, Junichiro Fujimoto, Yasuhide Hayashi, Katsuyoshi Koh, Atsushi Manabe, et al. An analysis of Ph-like ALL in japanese patients. *Blood*, 2013.

[47] Henrik Lilljebjörn, Helena Ågerstam, Christina Orsmark-pietras, Marianne Rissler, Hans Ehrencrona, L Nilsson, Johan Richter, and Thoas Fioretos. RNA-seq identifies clinically relevant fusion genes in leukemia including a novel MEF2D/CSF1R fusion responsive to imatinib. *Leukemia*, 28(4):977, 2014.

[48] Kjeld Schmiegelow, Erik Forestier, Marit Hellebostad, Mats Heyman, Jon Kristinsson, Stefan Söderhäll, and Mervi Taskinen. Long-term results of NOPHO ALL-92 and ALL-2000 studies of childhood acute lymphoblastic leukemia. *Leukemia*, 24(2):345, 2010.

[49] Valentino Conter, Claus R Bartram, Maria Grazia Valsecchi, André Schrauder, Renate Panzer-Grümayer, Anja Möricke, Maurizio Aricò, Martin Zimmermann, Georg Mann, Giulio De Rossi, et al. Molecular response to treatment redefines all prognostic factors in children and adolescents with B-cell precursor acute lymphoblastic leukemia: results in 3184 patients of the AIEOP-BFM ALL 2000 study. *Blood*, 115(16):3206–3214, 2010.

[50] Matthew K Iyer, Arul M Chinnaiyan, and Christopher A Maher. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, 27(20):2903–2904, 2011.

[51] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14 (4):R36, 2013.

[52] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010.

[53] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv e-prints*, March 2013.

[54] Gregory G Faust and Ira M Hall. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30(17):2503–2505, 2014.

[55] Christopher T Saunders, Wendy SW Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012.

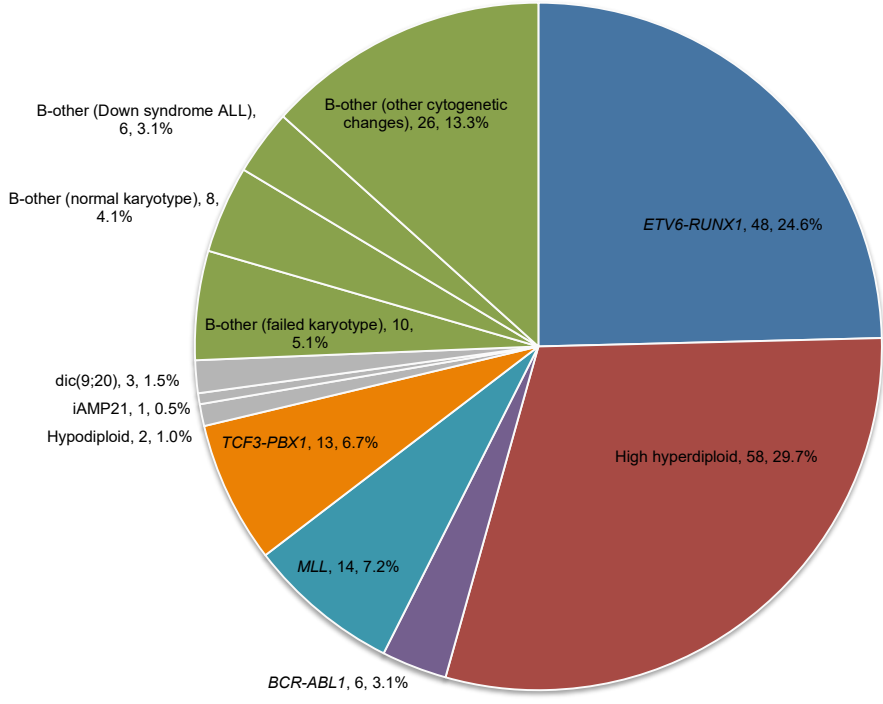
[56] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, volume 2. Springer, 2009.

[57] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[58] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, 2012.

[59] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, 2010.

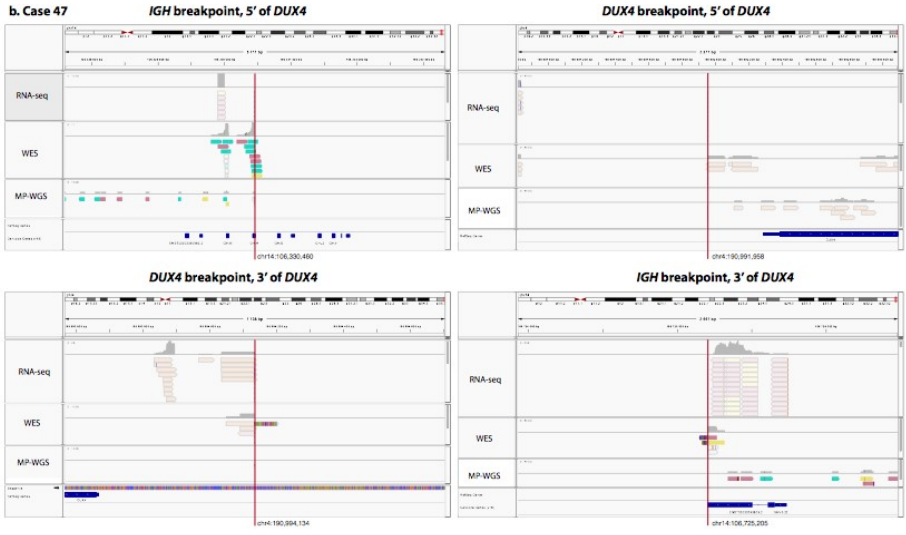
A Supplementary Figures



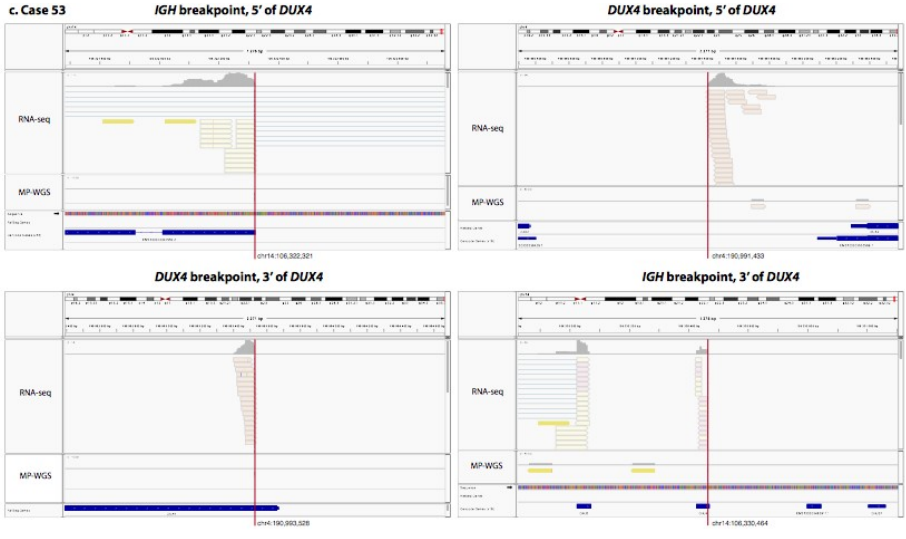
Supplementary Figure 1: **Distribution of the 195 BCP-ALL cases among thirteen genetic subtypes.** The subtypes "High hyperdiploid" (51-67 chromosomes) and "Hypodiploid" (<45 chromosomes) excludes cases positive for *BCR-ABL1*, *ETV6-RUNX1*, *MLL*, *TCF3-PBX1* and iAMP21. Of the 50 cases classified as B-other, 41 were negative for *BCR-ABL1*, *ETV6-RUNX1*, *MLL*, *TCF3-PBX1* and iAMP21, while nine cases had incomplete testing for one or more of these aberrations. Three cases originally classified as B-other were reclassified as *ETV6-RUNX1* or *MLL*, based on RNA-seq findings.

Supplementary Figure 2: **Identification of IGH-DUX4 breakpoints.** Reads from RNA-sequencing (RNA-seq), whole exome sequencing (WES), and mate pair whole genome sequencing (MP-WGS) that support a *IGH-DUX4* rearrangements are visualized using IGV [1] for the cases (**a**) #35, (**b**) #47, (**c**) #53, (**d**) #67, (**e**) #124, (**f**) #174, and (**g**) #179. Within each IGV window, the sequencing reads are illustrated below the chromosomal coordinates and include RNA-seq (top), WES (middle), and MP-WGS (bottom). WES data for breakpoint detection was available for the cases #35, #47, #174, and #179. The *DUX4* coding region or genes within *IGH* are indicated below the sequencing reads. The detected breakpoints are indicated by a red line. The visualization only contains read pairs consisting of one read within *IGH* and one read in the *DUX4*, and thus supporting a breakpoint involving these regions. Reads in the RNA-seq and WES libraries will be directed towards the breakpoint they support whereas reads in the MP-WGS library will be directed away from the breakpoint. The insert sizes were between 1-8 kb for MP-WGS libraries and between 200-400 for RNA-seq and WES libraries. The exact 5' breakpoints for case #35 and the exact 3' breakpoints for case #67 could not be conclusively determined from the visualized reads. The exact breakpoints for these cases were instead detected by guided assembly of all RNA-seq reads mapping to the *DUX4* region, followed by alignment of "overhang" regions. (Continued on the following pages.)

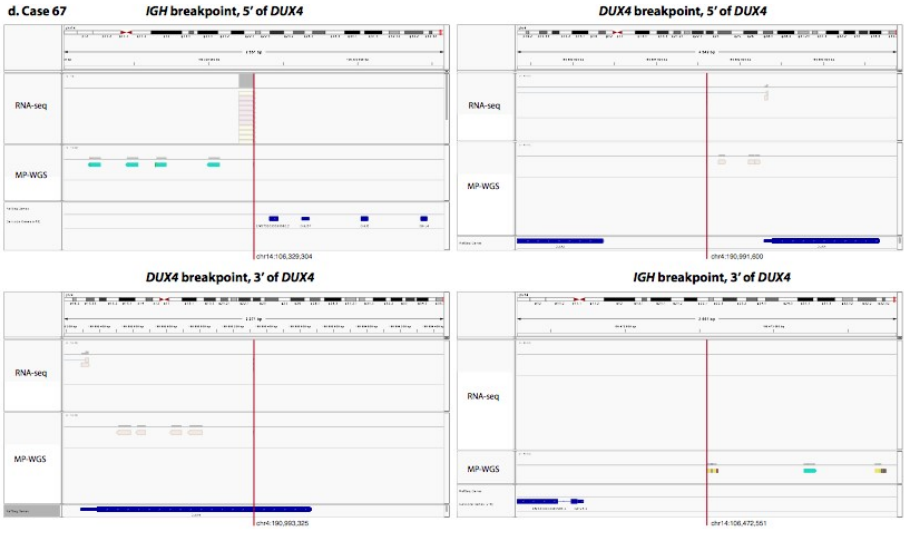
Supplementary Figure 2: **a.**



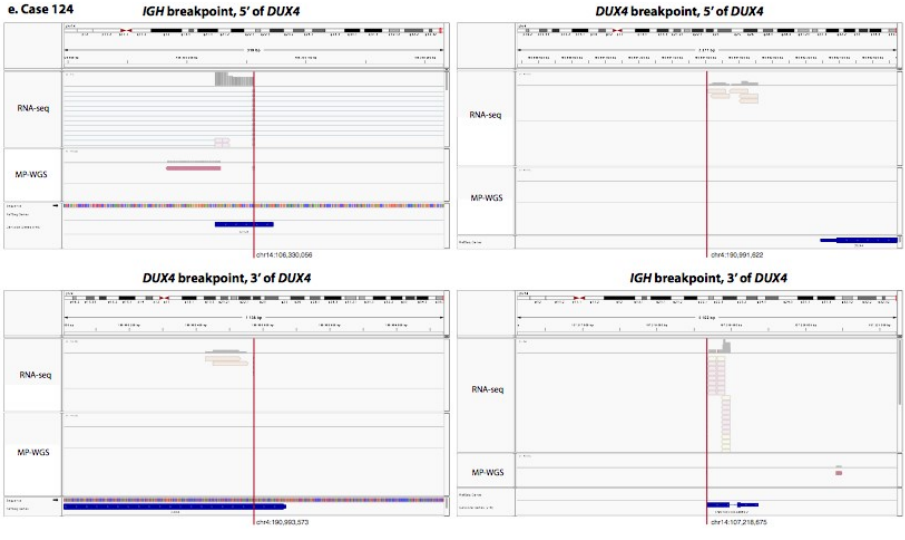
Supplementary Figure 2: **b.**



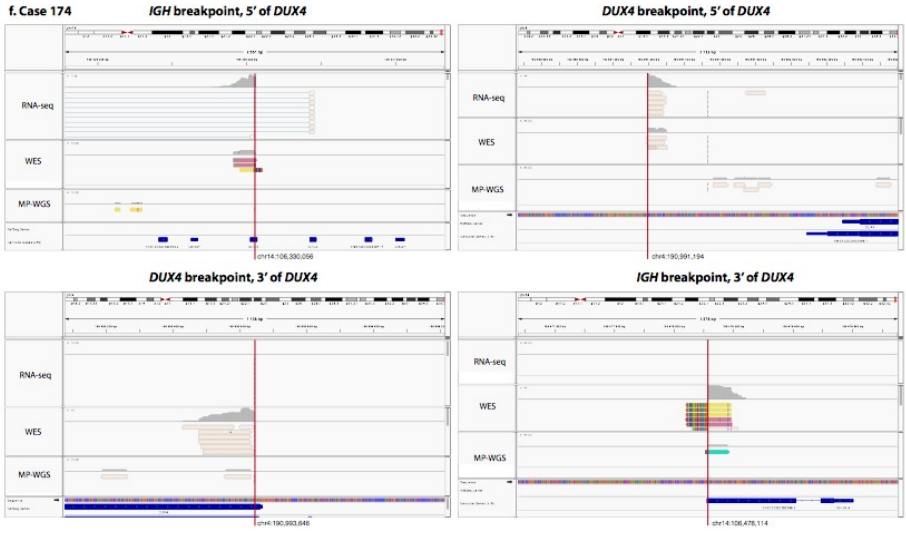
Supplementary Figure 2: **c.**



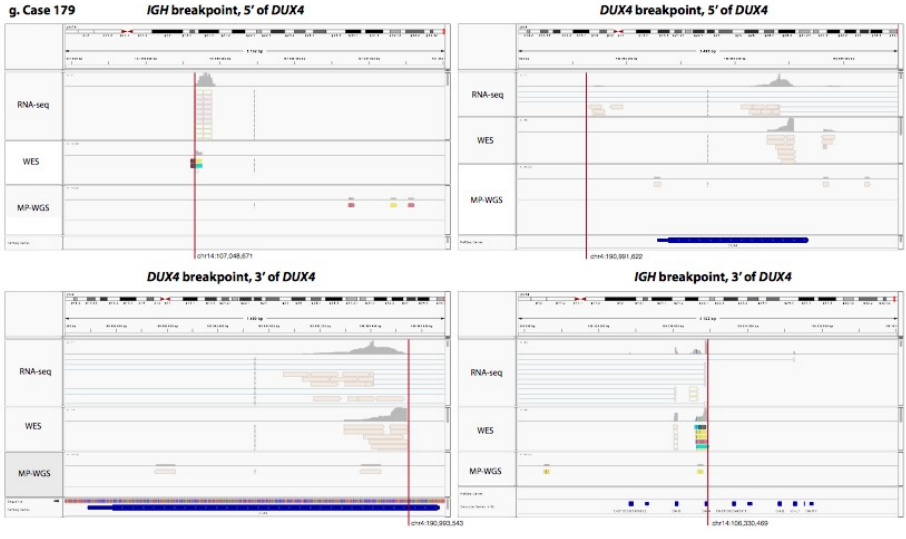
Supplementary Figure 2: **d.**



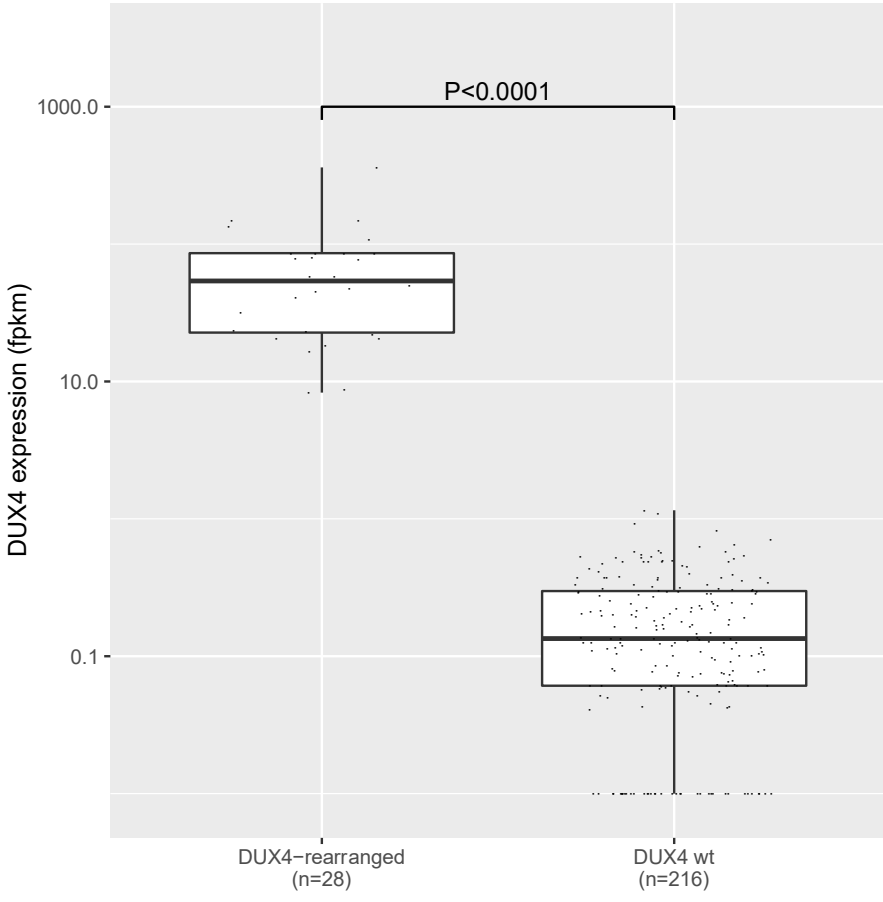
Supplementary Figure 2: **e.**



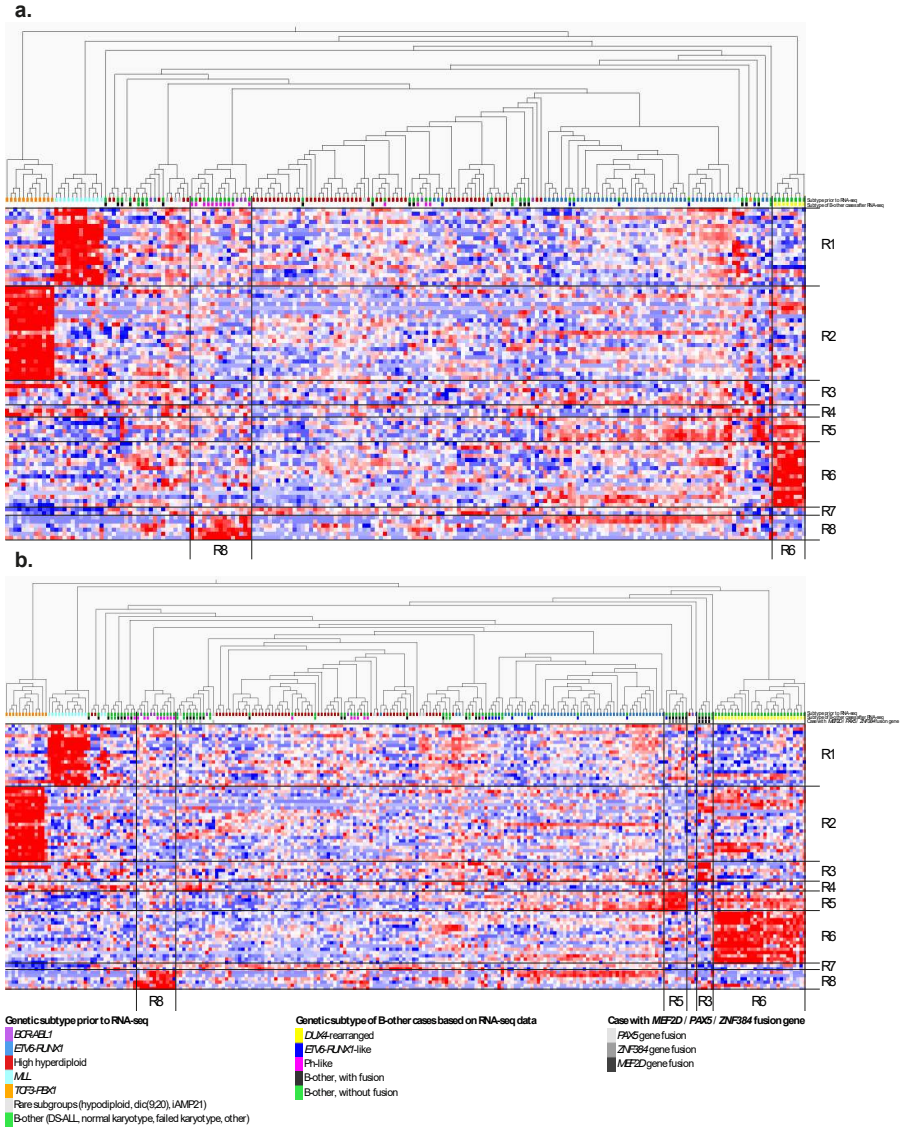
Supplementary Figure 2: **f.**



Supplementary Figure 2: **g.**

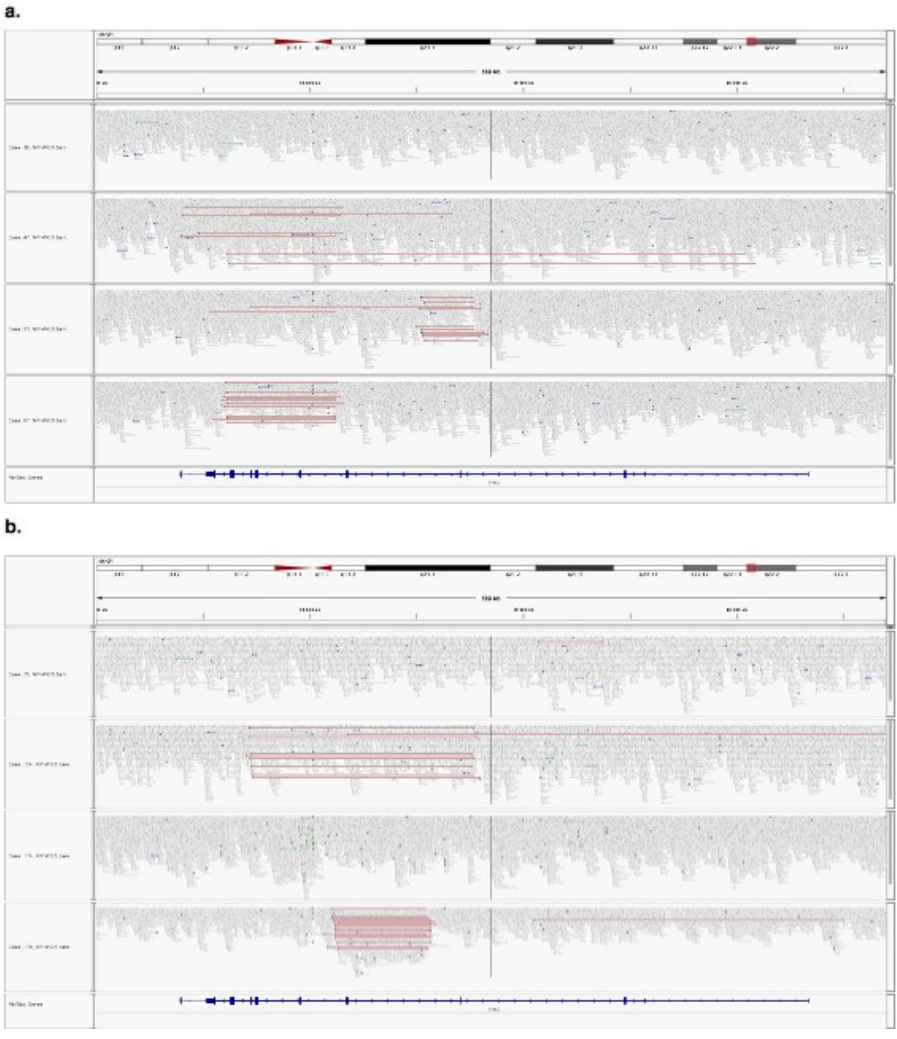


Supplementary Figure 3: **DUX4 expression in DUX4-rearranged cases.** *DUX4* gene expression in 244 BCP ALL cases with and without *DUX4* rearrangement from the combined discovery and validation cohorts. *DUX4* is significantly overexpressed in *DUX4*-rearranged cases. The boxes are defined by the first and third quartiles and whiskers extend from the boxes to the highest and lowest values. Two sided P-value calculated using Mann-Whitney U test.

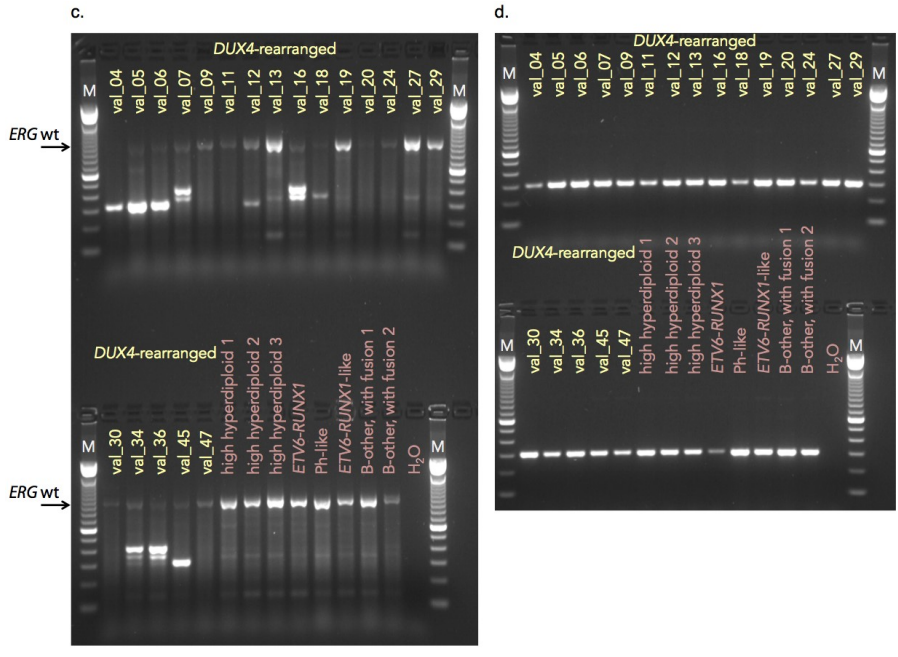


Supplementary Figure 4: (Continued on the following page.)

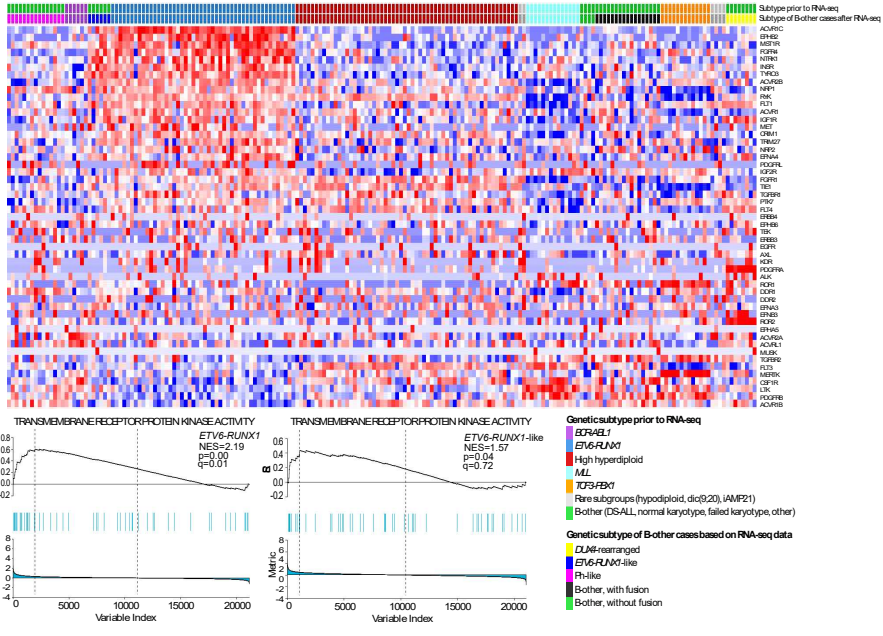
Supplementary Figure 4: **Hierarchical clustering based on the eight "ROSE" gene sets (R1-R8) described by Harvey et al[2].** (**a**) Hierarchical clustering of the 195 BCP ALL cases in the discovery cohort. Three *BCR-ABL1*-positive cases and 11 cases from the B-other group form a cluster with high expression of genes from the R8 gene set, indicating a Ph-like expression profile. In addition, all eight cases with *DUX4* rearrangement cluster together and exhibit high expression of the genes from the R6 gene set. (**b**) Hierarchical clustering of the 244 BCP ALL cases in the combined discovery and validation cohorts. All 28 *DUX4* rearranged cases cluster together and exhibit high expression of the genes from the R6 gene set. In addition, all six cases with *ZNF384* (*EP300-ZNF384*, $n = 4$; *TAF15-ZNF384*, $n = 1$, *TCF3-ZNF384*, $n = 1$) cluster together and exhibit high expression of the genes of the R5 gene set. All four cases with fusions involving *MEF2D* (*MEF2D-HNRNPUL1*, $n = 3$; *MEF2D-FOXJ2*, $n = 1$) cluster together and exhibit high expression of the genes of the R3 gene set.



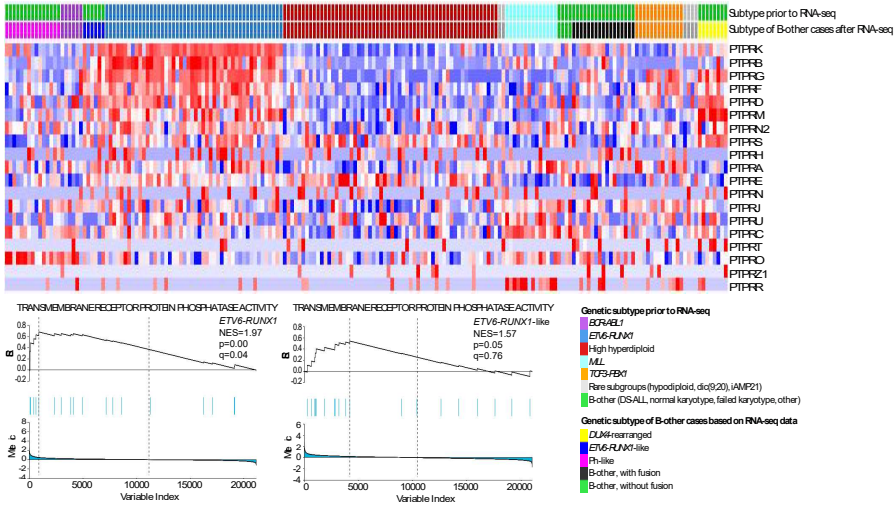
Supplementary Figure 5: **ERG deletions in DUX4-rearranged cases.** *ERG* deletions in *DUX4*-rearranged cases detected by mate pair whole genome sequencing (MP-WGS) and visualized using IGV[1]. (**a**) cases #35, #47, #53, and #67, and (**b**) cases #75, #124, #174, and #179. The insert sizes for MP-WGS were 1-8 kb. Read pairs mapped > 20 kb apart, indicating a deletion, are indicated in red. Read pairs indicating intragenic *ERG* deletions were detected in cases #47, #53, #67, #124, and #179. (Continued on the following page.)



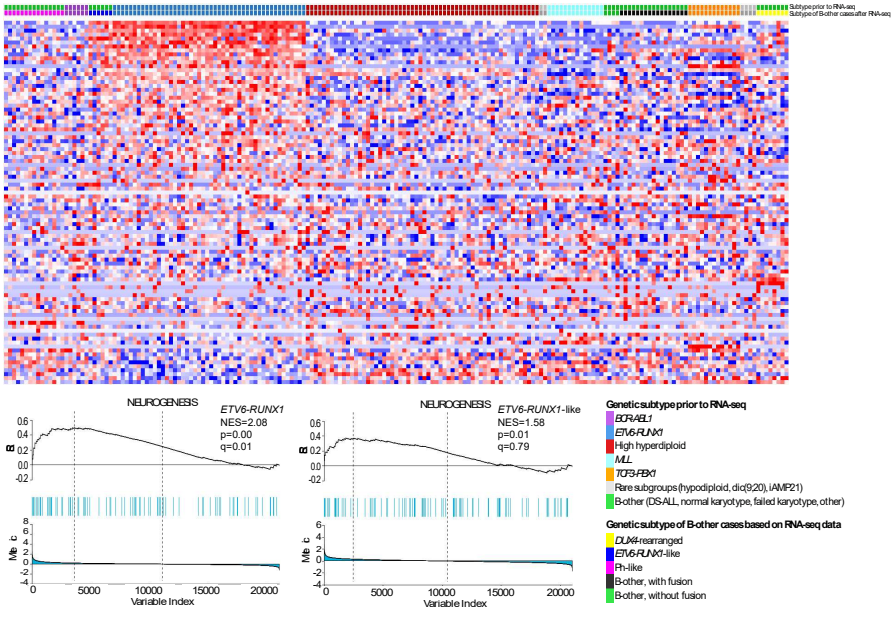
Supplementary Figure 5: **c-d.** *ERG* deletions in *DUX4*-rearranged cases in the validation cohort detected indirectly by RT-PCR. (**c**) RT-PCR of 20 BCP ALL cases with and eight BCP ALL cases without *DUX4* rearrangement. 10/20 cases with *DUX4* rearrangement express truncated *ERG* transcripts, indicating *ERG* deletions. (**d**) RT-PCR for *ABL1* verifying integrity of cDNA for all cases.



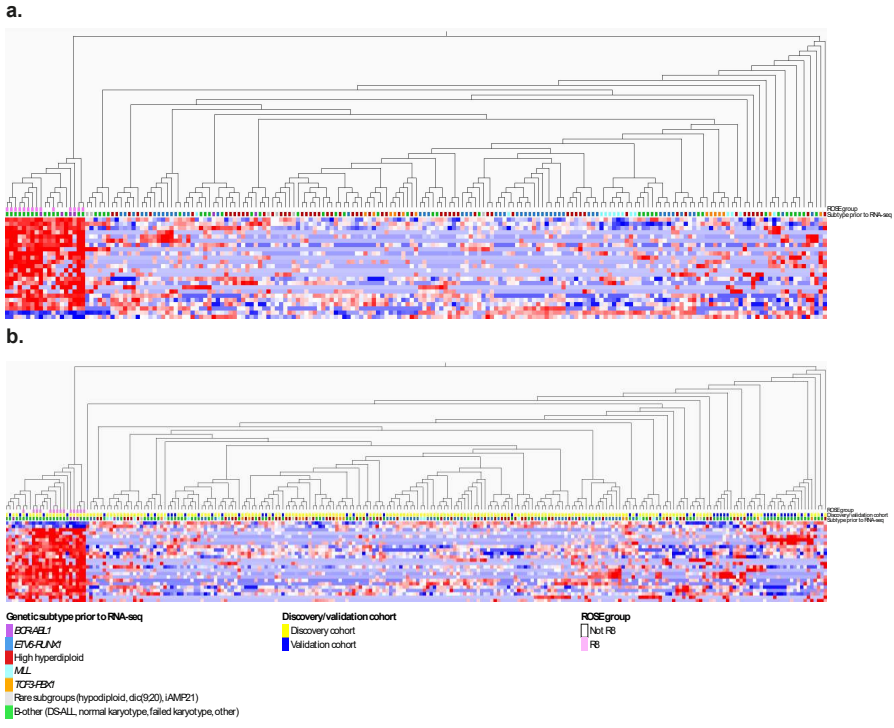
Supplementary Figure 6: **Expression of transmembrane receptor protein kinase activity genes in 195 BCP ALLs.** Expression profile of genes involved in transmembrane receptor protein kinase activity, highlighted by gene set enrichment analysis to be enriched in *ETV6-RUNX1*-positive cases.



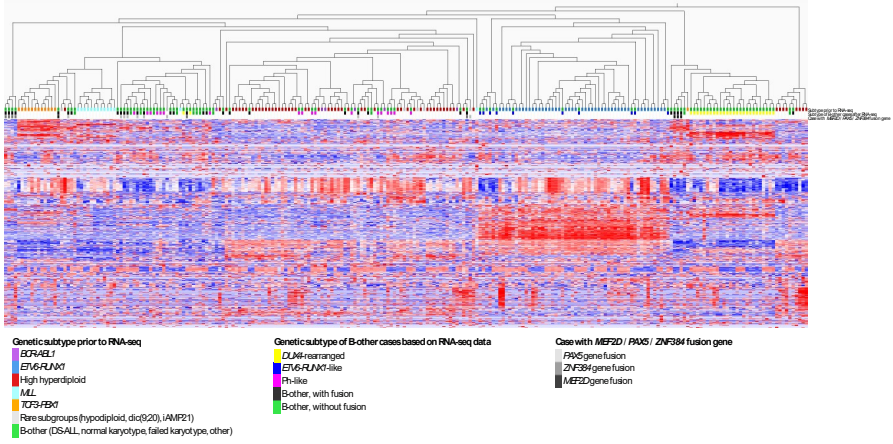
Supplementary Figure 7: **Expression of transmembrane receptor protein phosphatase activity genes in 195 BCP ALLs.** Expression profile of genes involved in transmembrane receptor protein phosphatase activity, highlighted by gene set enrichment analysis to be enriched in *ETV6-RUNX1*-positive cases.



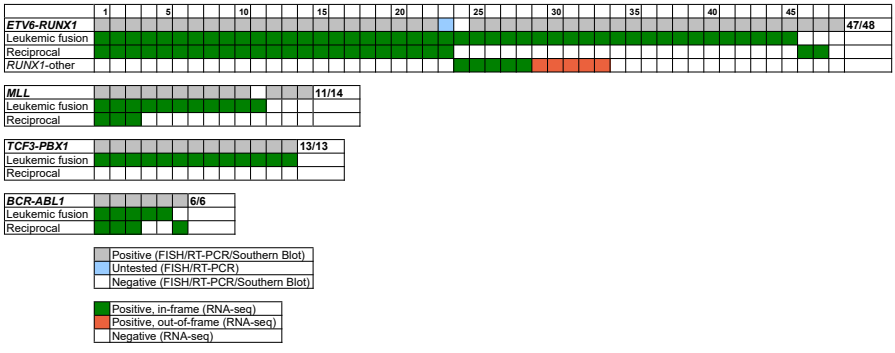
Supplementary Figure 8: **Expression of neurogenesis genes in 195 BCP ALLs.** Expression profile of genes involved in neurogenesis, highlighted by gene set enrichment analysis to be enriched in *ETV6-RUNX1*-positive cases.



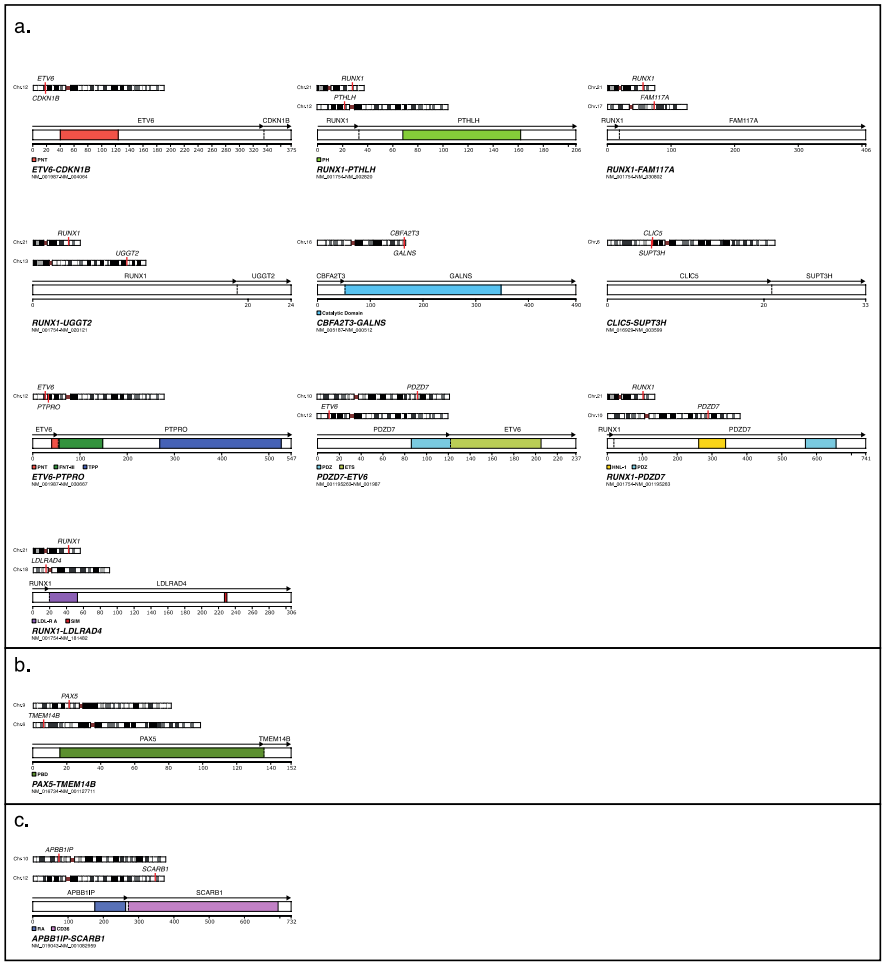
Supplementary Figure 9: **Identification of cases with Ph-like gene expression.** (**a**) Hierarchical clustering of 195 BCP ALL cases from the discovery cohort based on genes with a significantly ($P < 0.00001$, two-sided t-test) altered expression in the 11 B-other and three *BCR-ABL1*-positive cases from the R8-cluster (Supplementary Fig 4a). This analysis revealed five additional cases (four B-other and one *BCR-ABL1*-positive) with a similar gene expression profile. The 15 B-other cases in this cluster were labeled "Ph-like". (**b**) Hierarchical clustering of 244 BCP ALL cases from the combined discovery and validation cohorts based on the same genes as in (**a**). Five cases in the validation cohort cluster with the Ph-like cases. Of these, one case (case val_25) exhibited an *ETV6-RUNX1*-like gene expression profile and harbored an out of frame *ETV6-BCL2L14* gene fusion. The remaining four cases were labeled "Ph-like".



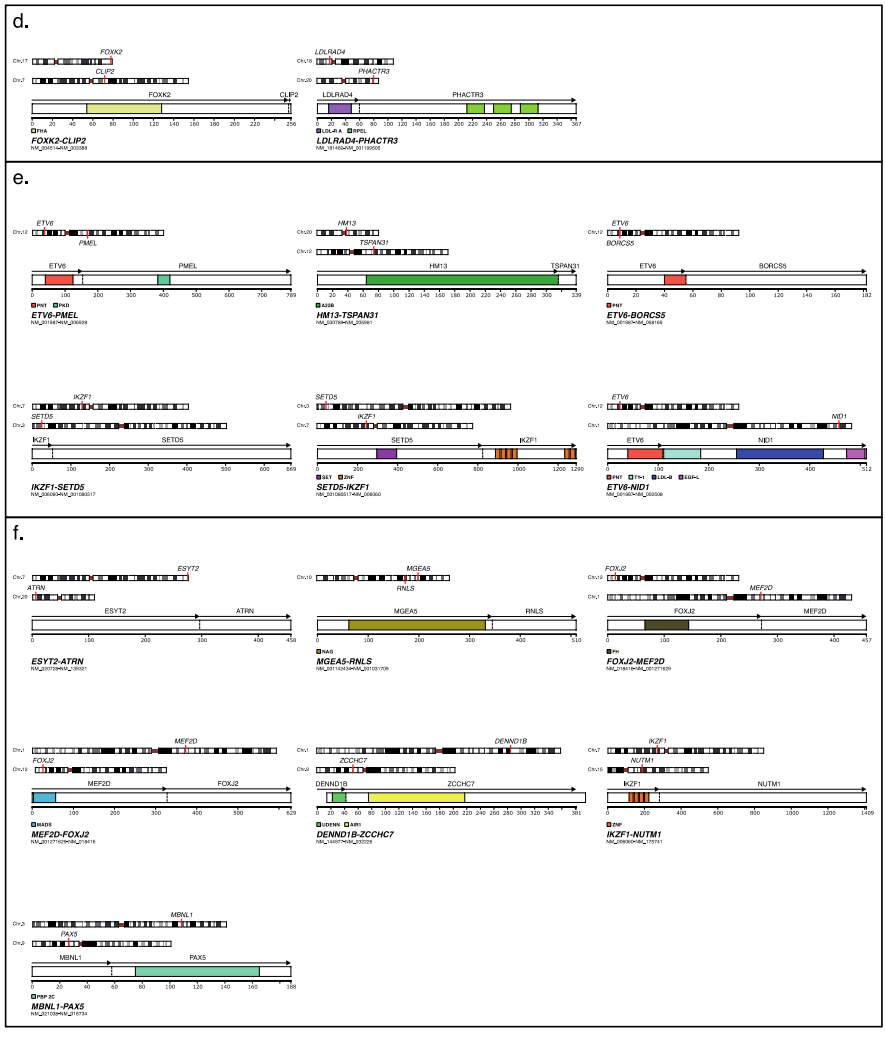
Supplementary Figure 10: **Unsupervised hierarchical clustering analysis of 244 BCP ALL cases in the combined discovery and validation cohorts.** The clustering is based on 631 genes that were retained after the variance threshold was set to 0.285. A total of 10 *ETV6-RUNX1*-negative cases cluster together with the *ETV6-RUNX1*-positive cases, 6 from the discovery cohort and 4 from the validation cohort.

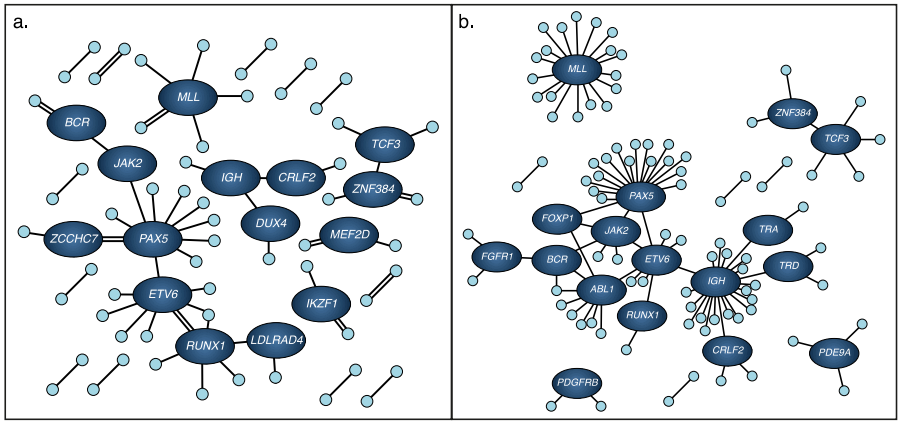


Supplementary Figure 11: **Concordance between fusion gene detection by RNA-seq and directed methods.** Concordance between fusion gene detection by RNA-seq and directed methods (RT-PCR, FISH, or Southern blot). *ETV6-RUNX1* or *RUNX1-ETV6* fusions could be detected by RNA-seq in 45/46 cases that had been positive by RT-PCR or FISH. In addition, *ETV6-RUNX1* fusions were detected in one case that had not been tested and in one case that had been negative by RT-PCR. The latter fusion had a rare breakpoint that was not detectable by the RT-PCR assay. *MLL* fusions were detected by RNA-seq in 10/13 cases that had been positive by FISH, RT-PCR, or Southern blot. In addition, one case that had been negative for *MLL*-rearrangements by FISH harbored an intrachromosomal *MLL*-USP2 fusion. *TCF3-PBX1* could be detected by RNA-seq in all 13 cases where the fusion had been detected by RT-PCR. *BCR-ABL1* or *ABL1-BCR* fusions could be detected in all cases that had been positive for *BCR-ABL1* by RT-PCR.

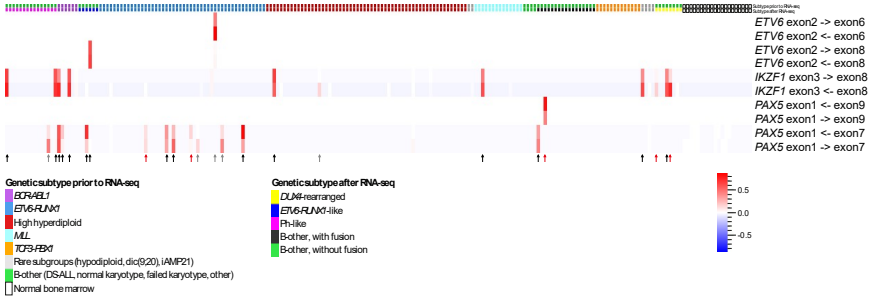


Supplementary Figure 12: **Gene positions and protein domains for the 25 novel fusion genes.** Gene positions and protein domains for the 25 novel fusion genes, divided by subtype: (**a**) *ETV6-RUNX1*-positive, (**b**) High hyperdiploid, (**c**) iAMP21, (**d**) Ph-like, (**e**) *ETV6-RUNX1*-like, and (**f**) B-other, with fusion. (Continued on the following page.)

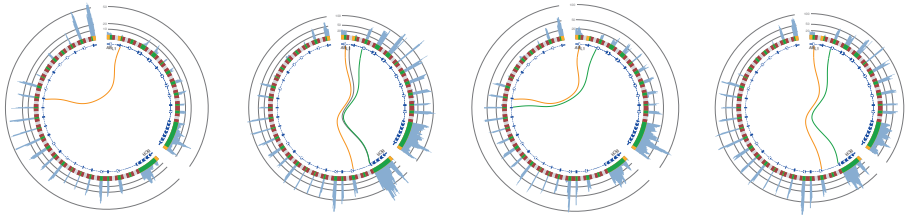

d.
FOXK2
CLIP2
FOXK2-CLIP2
LDLRAD4
PHACTR3
LDLRAD4-PHACTR3
e.
ETV6
PMEL
ETV6-PMEL
HMI3
TSPAN31
HM13-TSPAN31
ETV6
BORCS5
ETV6-BORCS5
IKZF1
SETD5
IKZF1-SETD5
SETD5
IKZF1
SETD5-IKZF1
ETV6
NID1
ETV6-NID1
f.
ESYT2
ATRN
ESYT2-ATRN
MGEA5
RNLS
MGEA5-RNLS
FOXJ2
MEF2D
FOXJ2-MEF2D
MEF2D
FOXJ2
MEF2D-FOXJ2
DENND1B
ZCCHC7
DENND1B-ZCCHC7
IKZF1
NUTM1
IKZF1-NUTM1
MBNL1
PAX5
MBNL1-PAX5



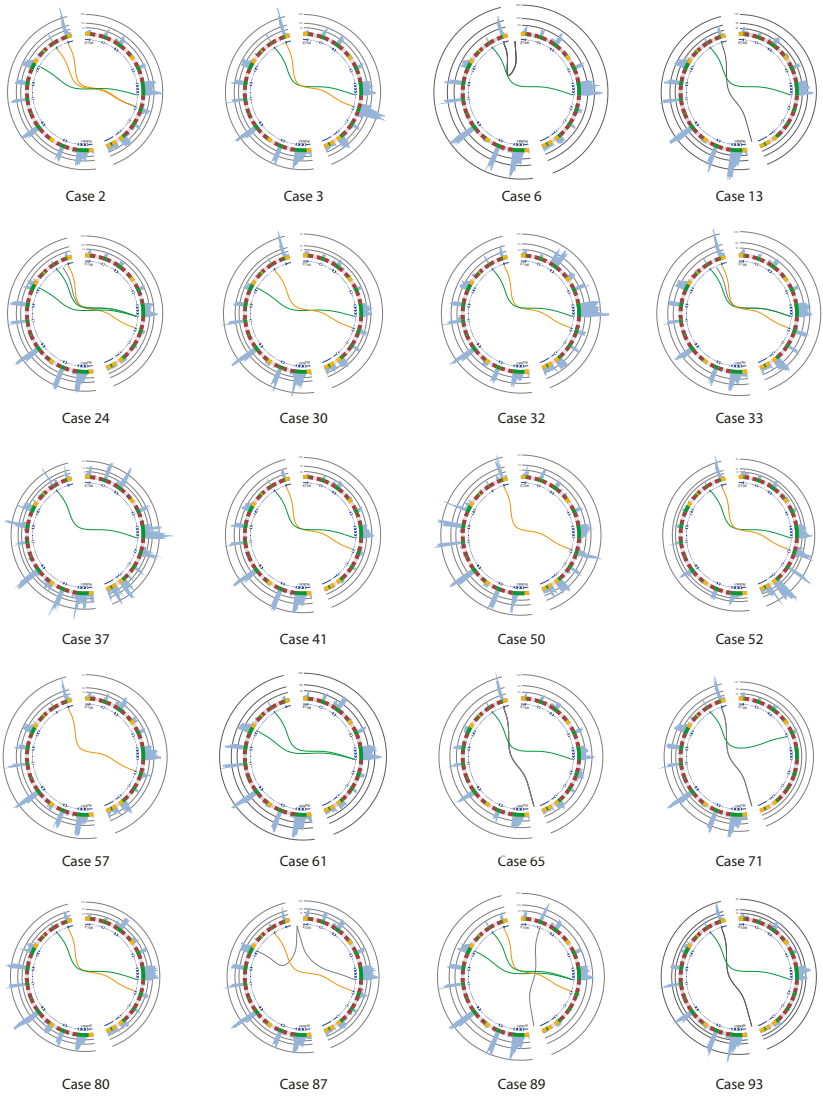
Supplementary Figure 13: **Gene fusion network analysis.** Gene fusion network analysis of (**a**) in-frame fusions detected in 244 BCP-ALL cases in the combined discovery and validation cohorts, and (**b**) fusions in BCP-ALL cases from literature data[3]. Genes with two or more fusion partners (from different cases) are indicated by their gene symbol. Only gene fusions discovered using guided methods (i.e. not deep sequencing) were included in the literature data.



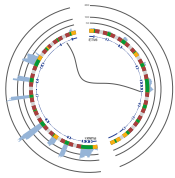
Supplementary Figure 14: **Leukemia specific splice variants in ETV6, PAX5, and IKZF1.** Heatmap illustrating the presence of leukemia specific splice variants in *ETV6*, *PAX5*, and *IKZF1*. All splice junctions not present in a reference transcript and not detected in 20 sorted bone marrow populations (from four individual donors) are depicted. *CDKN2A* was also analyzed, but no leukemia specific junctions were identified. Leukemia specific junctions in *ETV6*, *PAX5*, and *IKZF1* with a relative fraction above 0.1 was present in 25 cases, indicated by arrows. Black arrows indicate cases with splice junctions concordant with deletions detected by SNP array, red indicates discordance with SNP array data, and grey indicates cases with no SNP array data available.

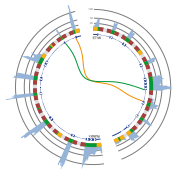


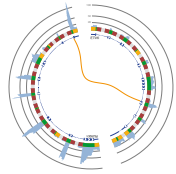
Supplementary Figure 15: **Fusion junctions detected in 4 BCR-ABL1-positive cases.** Genes are arranged clockwise by genomic position. The outer circle represents the genomic region encompassing the indicated genes. Yellow indicates untranslated regions, green indicates coding exons, and red and grey indicate intronic regions (the latter are not to scale). The inner circle represents one or two overlaid reference transcripts of the indicated gene. Coding exons are indicated by a thick line with white arrows indicating the direction of the gene, introns are indicated by a thin or dashed line, and untranslated regions are indicated by a medium thick line. Connecting lines between transcripts illustrate fusion breakpoints detected by at least three reads. Green lines indicate fusions breakpoints for *BCR-ABL1* and orange lines indicate fusion breakpoints for *ABL1-BCR*, as inferred from the position of the breakpoints.

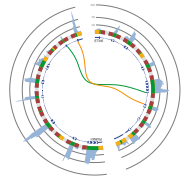


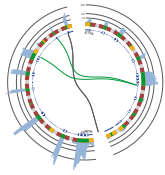
Supplementary Figure 16: **Fusion junctions detected in 44 ETV6-RUNX1-positive cases.** Connecting lines between transcripts illustrate fusion breakpoints detected by at least three reads. Green lines indicate fusions breakpoints for *ETV6-RUNX1* and orange lines indicate fusion breakpoints for *RUNX1-ETV6*, as inferred from the position of the breakpoints. Grey lines indicate fusion breakpoints involving an intron or a gene not depicted, regardless of the inferred fusion direction. (Continued on the following page.)

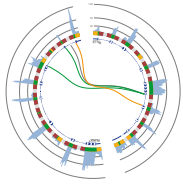

Case 99

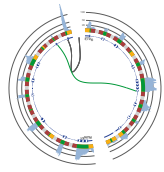

Case 114

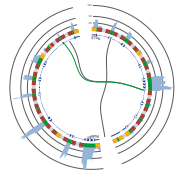

Case 117


Case 118


Case 123

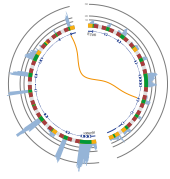

Case 126

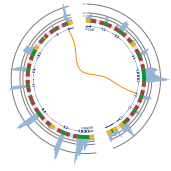

Case 129

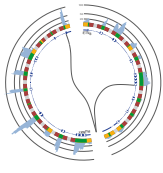

Case 131

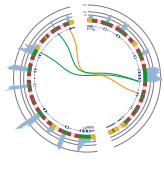

Case 132

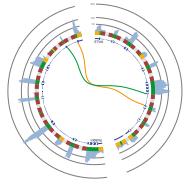

Case 134

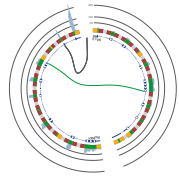

Case 138

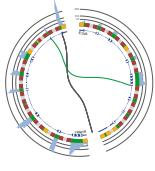

Case 142


Case 143


Case 149

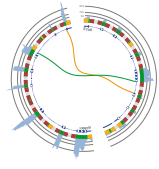

Case 155

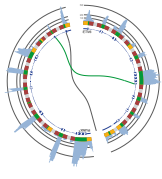

Case 168


Case 171

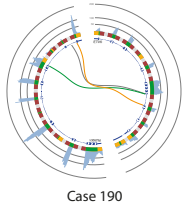

Case 178

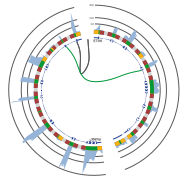

Case 180

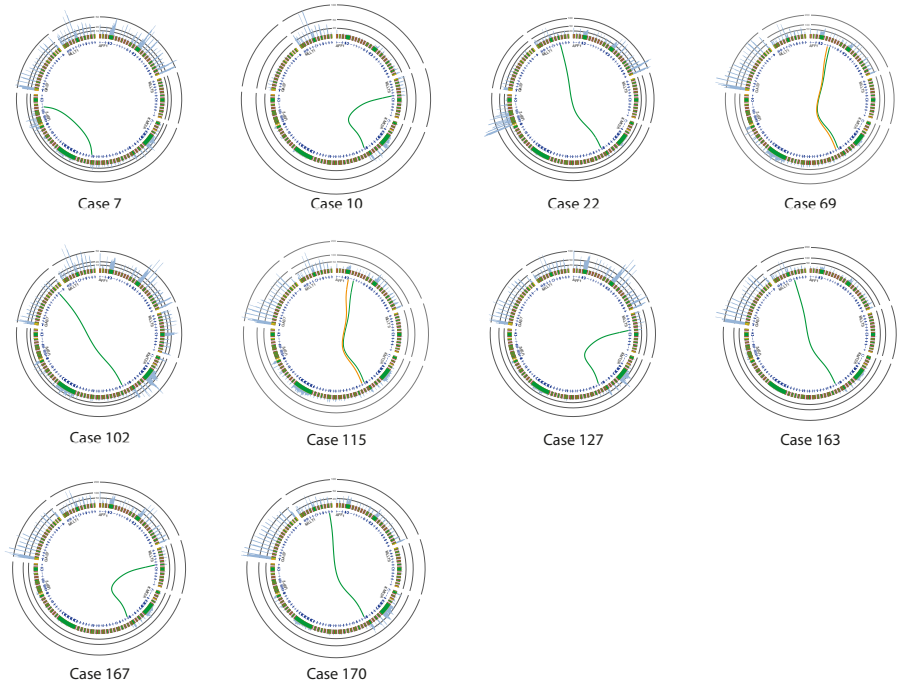

Case 186


Case 187

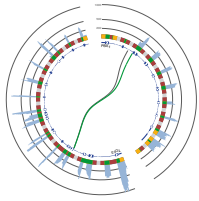

Case 189

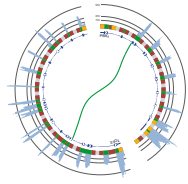

Case 190

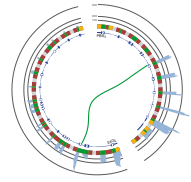

Case 193

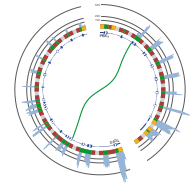


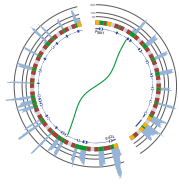
Supplementary Figure 17: **Fusion junctions detected in 10 cases positive for MLL-fusions.** Connecting lines between transcripts illustrate fusion breakpoints detected by at least three reads. Green lines indicate breakpoints where *MLL* is the 5′ partner and orange lines indicate fusion breakpoints where *MLL* is the 3′ partner, as inferred from the position of the breakpoints.

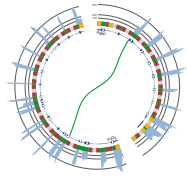

Case 17

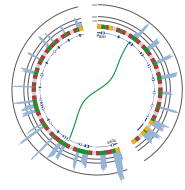

Case 46

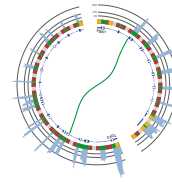

Case 59

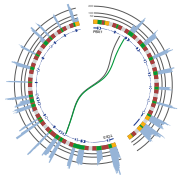

Case 63

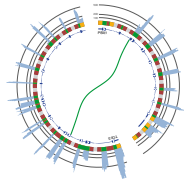

Case 66

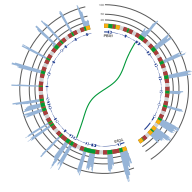

Case 83

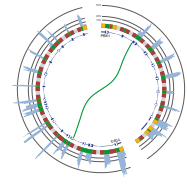

Case 140

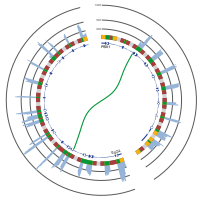

Case 157


Case 160

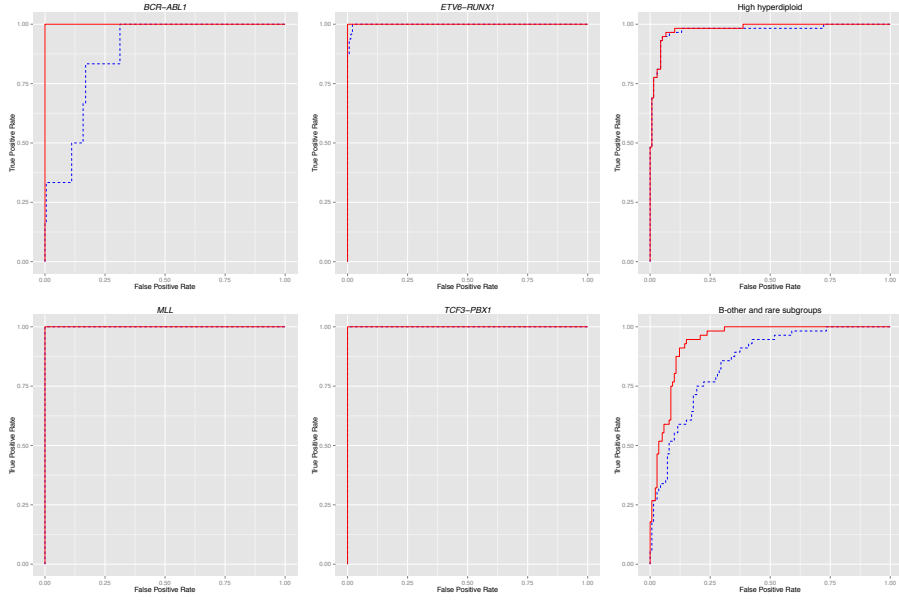

Case 165


Case 183


Case 192


Case 195

Supplementary Figure 18: **Fusion junctions detected in 13 TCF3-PBX1-positive cases.** Connecting lines between transcripts illustrate fusion breakpoints detected by at least ten reads. Green lines indicate fusions breakpoints for *TCFR3-PBX1*. Grey lines indicate fusion breakpoints involving an intron.



Supplementary Figure 19: **Classification of 195 BCP ALL cases from the discovery cohort based on gene expression data and gene fusion data.** True positive rates versus false positive rates plotted for classifiers based on 1) both gene fusion and gene expression data (red) and 2) only gene expression data (dashed blue). In total, 180/195 samples (92%) were correctly classified when utilizing both data sets, compared to 174/195 samples (89%) when only expression data was used.

Case	Gene Fusions (RNA-seq)	Deleted Genes (SNP-array)	Targeted ETV6-RUNX1 analysis		Karyotype
			FISH	RT-PCR	
Case 64	*ETV6-PMEL* *HM13-TSPAN31* *IKZF1-CDK2*	Not done	Negative for *ETV6-RUNX1* (wcp12)	Negative	46,XY,del(1)(q21),-7,-12,+mar, inc
Case 68	*ETV6-LOH12CR1*	*ETV6*, *IKZF1*, *PAX5*	Negative for *ETV6-RUNX1*	Negative	46,XX
Case 85	*P2RY8-CRLF2* *IKZF1-SETD5* *SETD5-IKZF1*	*ETV6*, *BTG1*	Not done	Negative	47,XY,t(3;7)(p25;p12),+21c,inc
Case 105	None	Not done	Not done	Negative	46,XY
Case 111	None	*ETV6*, *IKZF1*		Negative	45,XX,dic(7;12)(p11;p11)
Case 176	*ETV6-NID1*	*ETV6*, *IKZF1*		Negative	45,XY,?der(1)t(1;12)(q44;p13), dic(7;12)(p11;p11),del(12)(p13), -14,ins(14;?)(q24;?),+der(?)t(?;12) (?;p13) (partly based on FISH)

Supplementary Table 1: **Summary of genetic aberrations detected in six BCP ALL cases with ETV6-RUNX1-like gene expression profile.**

Analysis	Primer sequence
ERG deletion	Forward: 5'CTC CTC CAG CGA CTA TGG AC 3' Reverse: 5'GCG GCT GAG CTT ATC GTA GT 3'
FLT3 internal tandem duplication	Forward: 5´GCA ATT TAG GTA TGA AAG CCA GC 3´ Reverse: 5´CTT TCA GCA TTT TGA CGG CAA CC 3´
FLT3 activating mutations	Forward: 5´ATC ATC ATG GCC GCT CAC 3´ Reverse: 5´ GCA CTC AAA GGC CCC TAA CT 3´
NRAS exon 2 (codons 12 and 13)	Forward: 5'-GTACTGTAGATGTGGCTCGCCA-3' Reverse: 5'-GCCTCACCTCTATGGTGGGAT-3'
NRAS exon 3 (codon 61)	Forward: 5'-ACCCCCAGGATTCTTACAGAA-3' Reverse: 5'-GCCTGTCCTCATGTATTGGTCT-3'
KRAS exon 2 (codons 12 and 13)	Forward: 5'-TGTATTAACCTTATGTGTGACATGTTC-3' Reverse: 5'-CACCAGTAATATGCATATTAAAACAAG-3'
KRAS exon 3 (codon 61)	Forward: 5'-CTGTGTTTCTCCCTTCTCAGGATTC-3' Reverse: 5'-AAGAAAGCCCTCCCCAGTCCT-3'
PTPN11 exon 3	Forward: 5'-CCGACGTGGAAGATGAGATCTG-3' Reverse: 5'-CATACACAGACCGTCATGCATTTC-3'
PTPN11 exon 13	Forward: 5'-CTCTGAGTCCACTAAAAGTTGTGCAT-3' Reverse: 5'-AGCAAGAGAATGAGAATCCGCA-3'

Supplementary Table 2: **Primers used for detecting somatic mutations.**

Bibliography

[1] Helga Thorvaldsdóttir, James T Robinson, and Jill P Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 2013.

[2] Richard C Harvey, Charles G Mullighan, Xuefei Wang, Kevin K Dobbin, George S Davidson, Edward J Bedrick, I-Ming Chen, Susan R Atlas, Huining Kang, Kerem Ar, et al. Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome. *Blood*, 116(23):4874–4884, 2010.

[3] Felix Mitelman, Bertil Johansson, and Fredrik Mertens (Eds.). Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer, 2016. URL `http://cgap.nci.nih.gov/Chromosomes/Mitelman`.

PAPER II

Nature Microbiology

Attenuation of RNA viruses by redirecting their evolution in sequence space

Gonzalo Moratorio, Rasmus Henningsson, Cyril Barbezange, Lucia Carrau, Antonio V. Bordería, Hervé Blanc, Stephanie Beaucourt, Enzo Z. Poirier, Thomas Vallet, Jeremy Boussier, Bryan C. Mounce, Magnus Fontes and Marco Vignuzzi

Abstract

RNA viruses pose serious threats to human health. Their success relies on their capacity to generate genetic variability and, consequently, on their adaptive potential. We describe a strategy to attenuate RNA viruses by altering their evolutionary potential. We rationally altered the genomes of Coxsackie B3 and influenza A viruses to redirect their evolutionary trajectories towards detrimental regions in sequence space. Specifically, viral genomes were engineered to harbour more serine and leucine codons with nonsense mutation targets: codons that could generate Stop mutations after a single nucleotide substitution. Indeed, these viruses generated more Stop mutations both *in vitro* and *in vivo*, accompanied by significant losses in viral fitness. *In vivo*, the viruses were attenuated, generated high levels of neutralizing antibodies and protected against lethal challenge. Our study demonstrates that cornering viruses in 'risky' areas of sequence space may be implemented as a broad-spectrum vaccine strategy against RNA viruses.

Vaccines remain the most successful means of controlling morbidity and mortality caused by RNA viruses, yet relatively few viral vaccines exist. In recent years, several severe outbreaks have occurred: chikungunya and Zika viruses in the Americas, coronaviruses in the Middle East, and Ebola virus in West Africa. Consequently, there is a need for rationally designed and broadly applicable vaccine strategies. Given their high mutation rates, large population sizes and short generation times, RNA viruses can evolve rapidly, and strategies to control RNA viruses should take into account this adaptive potential. Due to their mutation rates, RNA viruses generate networks of closely related genetic variants, linked through mutation[1], that allow them to escape from selective pressures and adapt to different environments[2]. Ultradeep characterization of single nucleotide polymorphisms of viral populations reveals thousands of variants[3] heterogeneously distributed throughout the genome. This 'genetic architecture' suggests that certain mutations might be more or less accessible depending on the original nucleotide or codon, thereby defining different mutational neighbourhoods within the same sequence space[4]. In evolutionary biology, sequence space refers to every combination of a given sequence and theoretically is a vast multidimensional hypercube connecting all possible combinations. The localization of a virus population in sequence space, defined by its starting sequence, should then determine which mutational neighbourhoods are accessible. It is thus proposed that access to certain neighbourhoods will determine the potential of reaching beneficial mutations to facilitate adaptation[4,5].

However, even the best of mutational neighbourhoods is not without risk in terms of impact of mutation on fitness. Most studies addressing how organisms explore sequence space are theoretical and tested *in silico* using digital organisms[6,7]. Limited empirical data support the notion that the 'viable' sequence space is significantly smaller than the theoretical. For RNA viruses, most mutations are deleterious[8,9], and up to 40% of non-synonymous changes are lethal[3,10]. This is further evidenced in lethal mutagenesis, where antiviral treatment with mutagenic compounds leads to extinction[11,12]. Consequently, viruses have probably established a balance between generating beneficial mutations and tolerating detrimental ones. This trait, termed 'mutational robustness', is defined as 'phenotypic conservation in light of genetic variation'[13]. Given the biological constraints that limit the viable sequence space occupied by RNA viruses, we asked whether their capacity to explore sequence space (and, ultimately, their fitness) could be restricted by placing them closer to detrimental mutational neighbourhoods.

To address this question, we genetically engineered Coxsackie B3 and influenza A viruses to present altered synonymous codon architectures that would change their starting positions in sequence space and therefore limit their access to mutational neighbourhoods. Specifically, we rewired leucine and serine codons to constrain the expansion of viral populations towards detrimental mutational neighbourhoods, where the product of mutation would favour nonsense mutation targets resulting in Stop mutations. These engineered viruses were attenuated *in vivo*, with an increased number of Stop mutations in viral progeny. Animals immunized with these vaccine candidates were protected against lethal infection. We thus show that RNA viruses can be rationally attenuated by redirecting their evolutionary trajectories towards detrimental areas of sequence space.

1 Results

1.1 Reprogramming a viral genome to have enhanced proclivity for non-sense mutations

Our goal was to assess the effect of shifting the location of a virus in sequence space towards less 'hospitable' regions that increase its propensity to generate nonsense mutations. However, altering location in sequence space requires changes in nucleotide sequence, which can result in confounding factors such as changes in the amino-acid sequence or RNA structure, or the introduction of nucleotide and codon biases[14,15]. To minimize these factors, we introduced only synonymous changes, so that viral proteins retained the same amino-acid sequence and the same functions as wild-type virus. Furthermore, we only changed the codons for two amino acids with the highest codon redundancy (leucine and serine) to limit the overall change in nucleotide sequence to less than 5% and to focus on codons on which mutations would have the greatest impact on viability. Among the Leu and Ser codons, we defined a category termed '1-to-Stop', because point mutations on these codons could result in Stop mutations (Fig. 1a). We also defined a 'NoStop' category, which in contrast requires two nucleotide changes to reach Stop mutations. For Coxsackie virus B3 (CVB3), we targeted the P1 structural protein-coding region (Fig. 1a), which lacks RNA structural elements required for replication and translation. Because this viral RNA is translated into a single polyprotein that is cleaved into individual viral proteins, a Stop mutation appearing in this region will inactivate the virus. We thus generated CVB3 in which all 117

Ser/Leu codons in P1 were synonymously changed to either 1-to-Stop or NoStop categories. To investigate if this strategy can be applied broadly, we also targeted a very different RNA virus, the influenza A virus, which has a segmented, negative sense genome. In this case, two genomic segments were targeted independently: the PA polymerase subunit protein (111 Ser/Leu codons) or the haemagglutinin glycoprotein HA (94 Ser/Leu codons) (Fig. 1a).

Large-scale codon reshuffling may attenuate viruses by introducing a bias in codon pairs that are under-represented in the human genome[14], so we verified that this potentially confounding effect did not occur in our design. We compared our viruses (Fig. 1b) with those used by Coleman et al.[14] (PV-Min, PV-Max and PV-SD) to link codon pair bias and attenuation. These poliovirus (a related enterovirus) constructs were codon-shuffled in the same P1 region. The change in codon pair bias in our viruses was significantly less than for the PV-Min virus, which had 1,000-fold reduced replication[14]. Even PV-SD, which was not attenuated in their study, presented more bias than our constructs. The attenuation obtained from codon reshuffling can also result from dinucleotide bias, where higher CpG content decreases replication[15]. Compared to the dinucleotide frequency described for the highly attenuated echovirus 7 by Tulloch et al.[15] (CpG-high E7), no significant changes in CpG dinucleotide frequency were introduced in our 1-to-Stop and NoStop constructs (Fig. 1c). Subsequently, we investigated the production of infectious particles for these viruses at high (Fig. 1d,e) or low (Supplementary Fig. 1a,b) multiplicities of infection (MOI). All viruses grew to comparable final titres, although fewer infectious progeny were produced at some time points for the CVB3 1-to-Stop construct (Fig. 1d). To test whether this was due to defects in replication, we quantified RNA genome synthesis. All viruses generated the same amounts of RNA at every time point (Fig. 1f,g). We further confirmed these results in an *in vitro* RNA replication assay using replication complexes purified from infected cells, where yields for wild-type and 1-to-Stop viruses were equal (Supplementary Fig. 2). Thus, our data showed that, for both CVB3 and influenza A virus, the altered Leu/Ser codons had no discernible impact on RNA synthesis and replication kinetics. Finally, genetic and phenotypic stabilities were evaluated after 10 passages. No reversion in Ser/Leu altered positions was observed, and each virus retained its phenotype (growth titres, relative number of Stop mutations) (Supplementary Fig. 3).

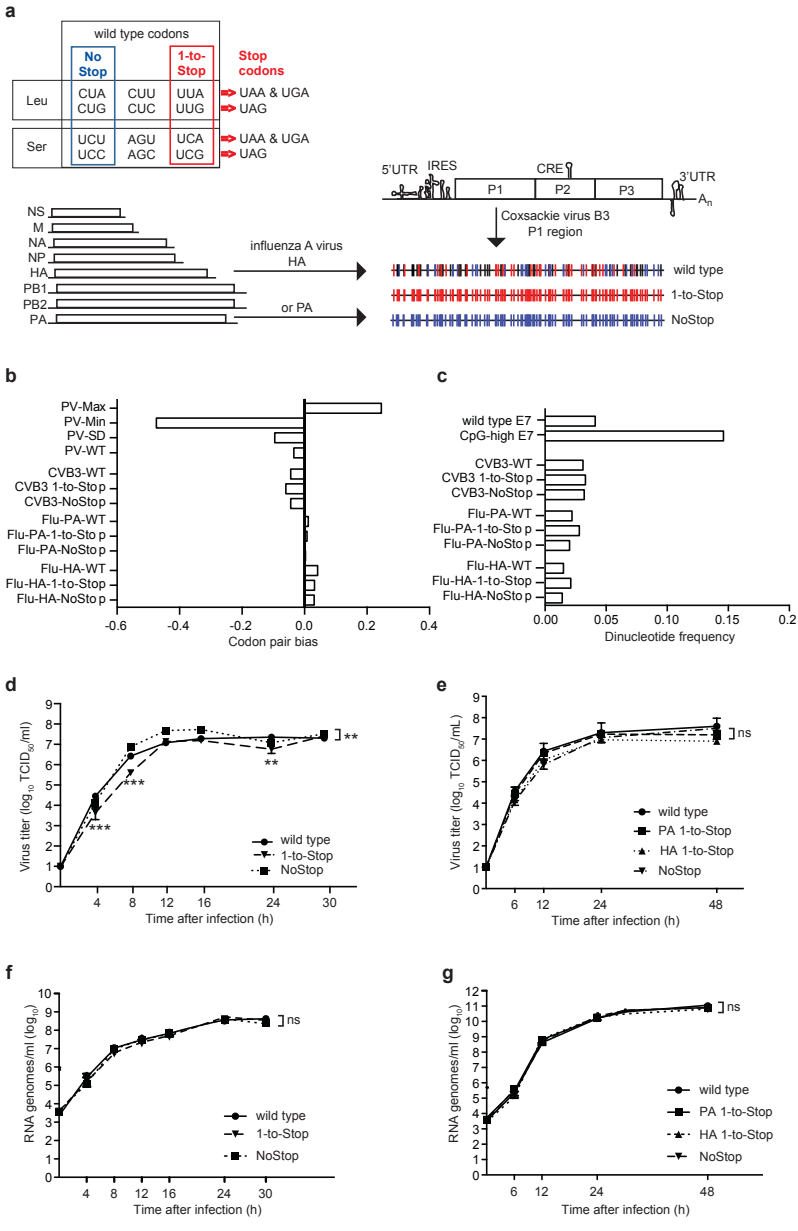


Figure 1: (Continued on the following page.)

Figure 1: **Construction of 1-to-Stop and NoStop RNA viruses. a**, Schematic representation of the six Leu and six Ser codons that are similarly represented in wild-type virus genomes, with codons belonging to the 1-to-Stop (red) and NoStop categories (blue) shown in colour. The stop mutations that can occur after a single point mutation are indicated. The Coxsackie virus B3 (CVB3) genome illustrates RNA structures required for replication (5'untranslated region (UTR), IRES, CRE and 3'UTR) and the single open reading frame encoding structural proteins (P1 region) and non-structural proteins (P2, P3 regions). The 117 Ser/Leu codons of the P1 region were altered to construct the 1-to-Stop viruses and NoStop viruses. The influenza A virus genome is composed of eight gene segments. The HA and PA genes were altered at 94 and 111 Ser/Leu codons, respectively, to generate the 1-to-Stop and NoStop viruses. A_n, polyA tail. **b**, Codon pair bias of wild-type, 1-to-Stop and NoStop CVB3 and influenza A viruses compared to previously published wild-type poliovirus (PV-WT) and constructs engineered to attenuate viruses through codon pair deoptimization: PV-Max, in which codon pairs over-represented in the human genome were maximized (no reduction in replication); PV-Min, in which codon pair bias was maximized by using codon pairs under-represented in the human genome (1,000-fold reduction in replication); and PV-SD, with randomly shuffled codons (no reduction in replication). **c**, CpG dinucleotide frequencies in wild-type, 1-to-Stop and NoStop CVB3 and influenza A viruses, relative to previously published wild-type Echovirus 7 (E7) and its high-CpG content construct shown to be attenuated. **d**, Production of infectious viral progeny over time of wild-type, 1-to-Stop and NoStop CVB3 viruses in HeLa cells infected at MOI=1. **e**, Production of infectious viral progeny over time of wild-type, 1-to-Stop PA and HA, and NoStop PA influenza A viruses in MDCK cells infected at MOI=1. **f**, Replication kinetics of wild-type, 1-to-Stop and NoStop CVB3 viruses in HeLa cells infected at MOI=1. **g**, Replication kinetics of wild type, 1-to-Stop PA and HA, and NoStop PA influenza A viruses in MDCK cells infected at MOI=1. In **d–g**, graphs show mean and s.e.m.; $n = 3$ per group. NS, non-significant; *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$ (two-way analysis of variance with a Bonferroni post-test, comparing wild type to each mutant).

1.2 1-to-Stop viruses have lower fitness and are highly sensitive to mutation

The relative fitness of wild-type, 1-to-Stop and NoStop viruses was measured under normal and mutagenic conditions against a neutral, genetically marked competitor[16,17]. For CVB3 (Fig. 2a), five mutagenic conditions were used: three base analogues (ribavirin, 5-fluorouracil (5-FU) and 5-azacytidine (5-AZC)), amiloride (which perturbs intracellular Mg^{2+} and Mn^{2+} concentrations, essential cofactors of the viral polymerase[18]) and Mn^{2+} itself, which increases the polymerase error rate. In all cases, the 1-to-Stop virus presented significantly lower fitness than wild-type CVB3, while the NoStop virus presented the same or higher fitness (Fig. 2a). Using an alternative method to evaluate viral fitness, we measured mean plaque size for each viral population treated with the three base analogues. The 1-to-Stop virus produced significantly smaller plaques, while the NoStop virus produced larger plaques (Fig. 2b). The progeny viruses obtained during mutagenic treatments were then deep-sequenced, and sequence data were computationally treated to re-

duce noise (see Methods) and mathematically modelled to estimate error (see Supplementary Section 'Mathematical assessment of background noise'). The number of reads presenting Stop mutations was then calculated. As expected, 1-to-Stop virus samples contained significantly more reads with Stop mutations than wild-type virus, whereas NoStop had significantly fewer (Fig. 2c). Finally, to further support that the 1-to-Stop viruses were sensitive to the increased mutational load on these codons, we coupled the 1-to-Stop virus with a high-fidelity polymerase [19]. This virus would have an intrinsically lower error rate to counter the extrinsically higher error resulting from mutagen treatment. Indeed, 1-to-Stop high-fidelity viruses recovered the wild-type virus phenotype (Supplementary Fig. 4).

For influenza A constructs, viruses were treated with three concentrations of ribavirin, 5-AZC or Mn^{2+} (Fig. 3a–c, left panels). Under each mutagenic condition, the 1-to-Stop PA and HA viruses were more sensitive to mutation, presenting lower titres than the wild type and their NoStop counterparts. In fact, 1-to-Stop viruses exhibited up to 50-fold reduction in viral titres with ribavirin treatment (Fig. 3a, left) and a 100-fold decrease after high concentrations of Mn^{2+} (Fig. 3c, left). We then quantified the number of Stop mutations in the mutagen-treated progeny viruses for each replicate and at each concentration. The 1-to-Stop PA populations presented a dose-dependent and significantly higher increase in the number of Stop mutations along the PA gene compared to wild-type virus (Fig. 3a–c, middle panels). As a control, the 1-to-Stop HA virus, which has a wild type-like PA sequence, did not present more Stop mutations in the PA gene. Instead, the 1-to-Stop HA virus presented more Stop mutations in the HA gene (Fig. 3a–c, right panels). Together, these data confirm for both viruses that relocalizing the position of an RNA virus in sequence space to increase its likelihood of generating non-sense mutations resulted in a higher sensitivity to increased mutational load.

1.3 1-to-Stop viruses are attenuated in vivo

To evaluate this attenuation strategy *in vivo*, mice were given a sublethal dose of wild-type, 1-to-Stop or NoStop CVB3, and virus titres were determined over seven days. Although the 1-to-Stop virus replicated with wild type-like kinetics during the first five days in most mice, it was no longer detectable in pancreata (Fig. 4a) and in most hearts (Fig. 4b) by day 7. Importantly, the NoStop CVB3 virus retained the same virulence phenotype as wild-type virus (Fig. 4a,b). By deep sequencing, we observed threefold more Stop mutations in 1-to-Stop virus in these tissues (Fig. 4c), as was observed in tissue culture (Fig. 2c,b and Supplementary

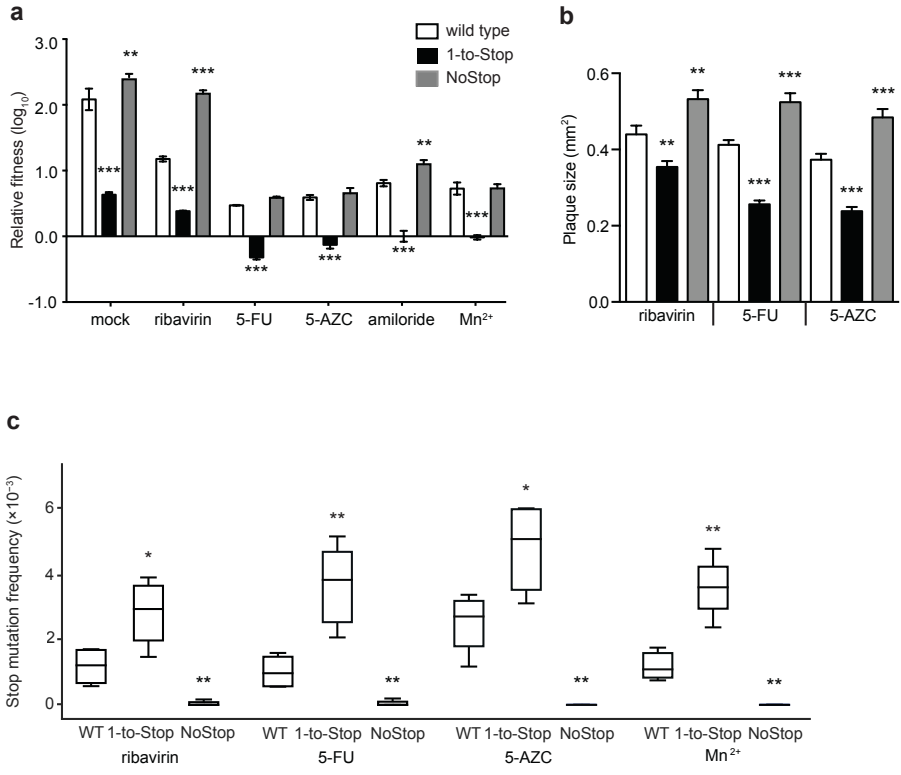


Figure 2: **The 1-to-Stop CVB3 virus is highly sensitive to mutation.** **a**, Relative fitness by a direct competition assay in the absence (mock) or presence of 200 μM ribavirin, 5-fluorouracil (5-FU), 5-azacytidine (5-AZC), amiloride or 1 mM manganese (Mn^{2+}). Wild type (white), 1-to-Stop (black) and NoStop (grey) competed against a neutral, marked wild-type CVB3. Graphs show mean and s.e.m.; $n = 6$ per group. $^{**}P < 0.01$, $^{***}P < 0.001$ (two-tailed unpaired t-test with Bonferroni correction, comparing wild type to each mutant). **b**, Viruses were grown in the presence of 200 μM of three different mutagens. Graphs show mean and s.e.m.; $n = 1{,}000$ per group. $^{**}P < 0.01$, $^{***}P < 0.001$ (Mann–Whitney test). **c**, Frequency of Stop mutations observed in deep-sequencing reads from wild type and 1-to-Stop populations passaged in 50 μM RNA mutagens: ribavirin, 5-FU, 5-AZC and 0.5 mM Mn^{2+}. Boxes show median and interquartile range, whiskers show range or 1.5 interquartile range in the case of outliers (indicated by dots); $n = 10$. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (two-tailed unpaired t-test with Bonferroni correction comparing wild type to each mutant).

Fig. 3c). The NoStop virus control, on the other hand, presented significantly fewer Stop mutations. To further evaluate the effect of mutation on each virus, we determined each population's specific infectivity (the ratio of total genomes versus infectious genomes produced) (Fig. 4d). Although all CVB3 viruses presented

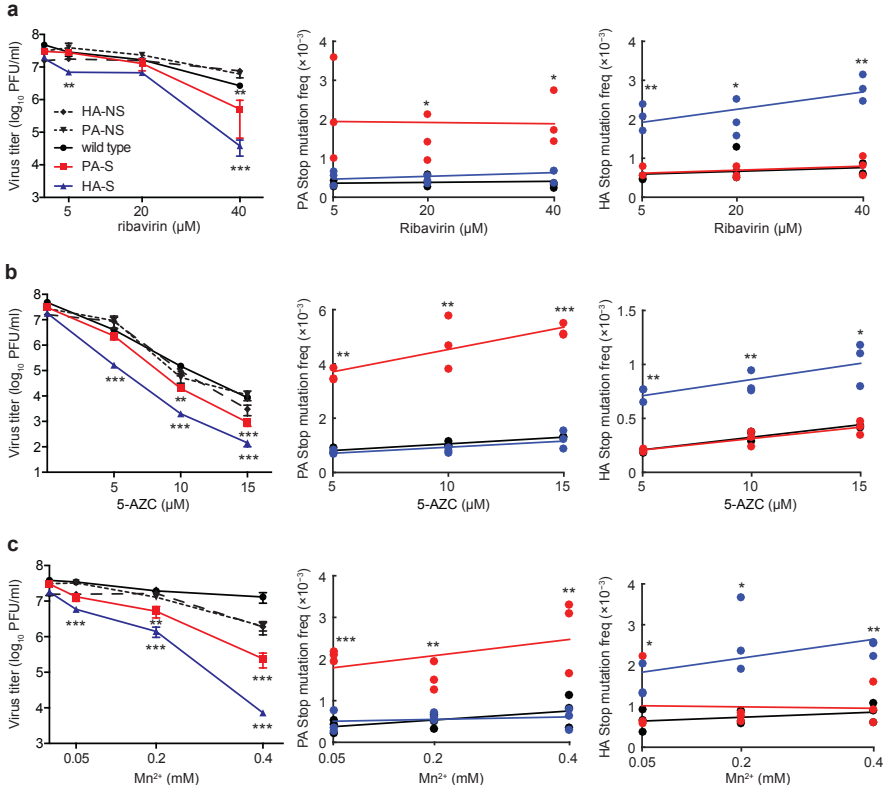


Figure 3: **The 1-to-Stop influenza A viruses are highly sensitive to mutation. a–c,** Left: sensitivity of wild-type, 1-to-Stop (PA-S and HA-S) and NoStop (PA-NS and HA-NS) influenza A viruses to increasing concentrations of ribavirin (**a**), 5-azacytidine (5-AZC) (**b**) and Mn^{2+} (**c**). Graphs show mean and s.e.m.; $n = 3$. $^{**}P < 0.01$, $^{***}P < 0.001$ (two-tailed unpaired t-test with Bonferroni correction, comparing wild type to each mutant). Middle: frequency of Stop mutations observed in deep-sequencing reads from wild-type (black) and 1-to-Stop (PA-S, red and HA-S, blue) progeny in the PA gene. Right: frequency of Stop mutations observed in deep-sequencing reads from wild-type (black) and 1-to-Stop (PA-S in red and HA-S in blue) progeny in the HA gene. Bars show mean and s.e.m.; $n = 3$ per group. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (two-tailed unpaired t-test with Bonferroni correction, comparing wild type to each mutant).

high infectivity at three days of infection, a significantly larger proportion of 1-to-Stop genomes were non-infectious at seven days.

For influenza, mice were infected intranasally with 1×10^5 plaque-forming units (p.f.u.) of wild type, 1-to-Stop PA or HA (PAS and HAS), or NoStop PA, and virus titres were determined in the lungs. Both 1-to-Stop viruses were attenuated

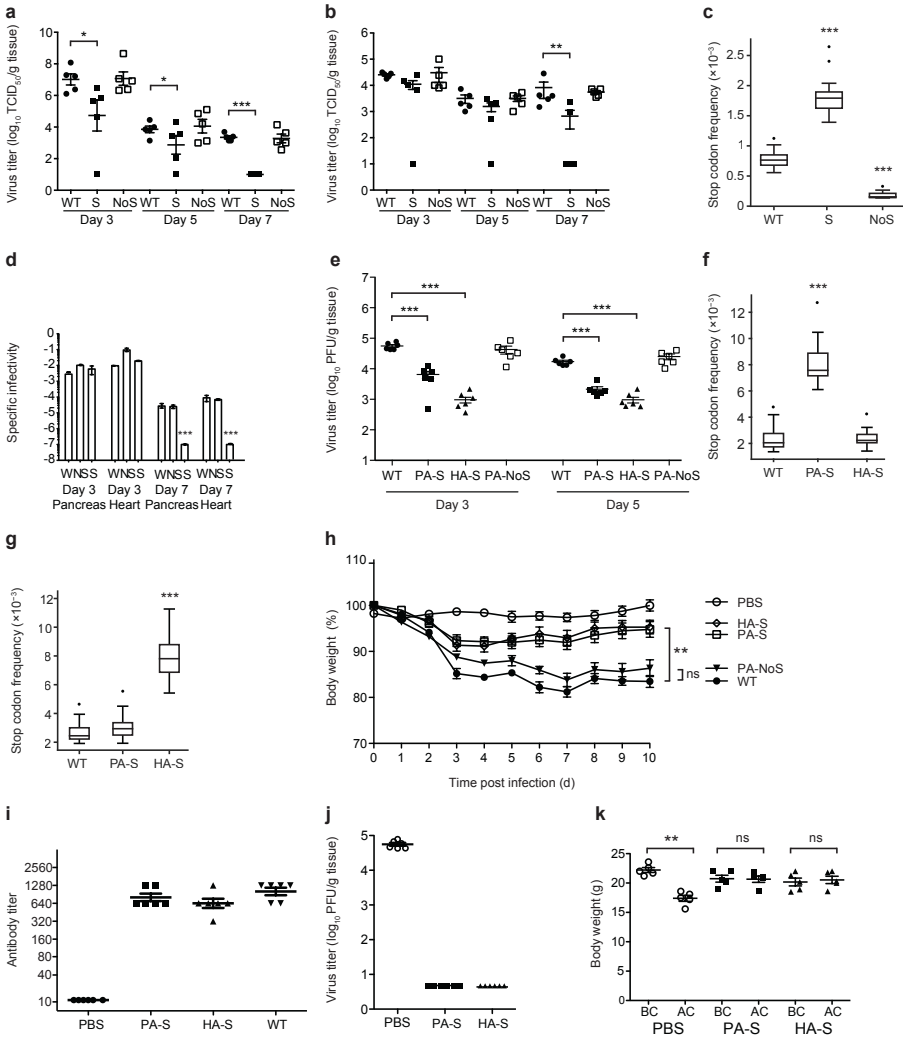


Figure 4: **The 1-to-Stop viruses are attenuated in vivo. a**,**b**, Virus titres in pancreata (**a**) and hearts (**b**) of mice infected with 1×10^5 $TCID_{50}$ of wild-type (WT), 1-to-Stop (S) and NoStop (NoS) CVB3. Bars show mean and s.e.m., and data are representative of three experiments. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$ (two-tailed unpaired t-test with Bonferroni correction). The limit of detection was 10^1 $TCID_{50}$ per ml. **c**, Frequency of Stop mutations in CVB3 populations from infected tissues (hearts and pancreata combined). Boxes show median and interquartile range, whiskers show range or 1.5 interquartile range in the case of outliers (indicated by dots); $n = 62$. $***P < 0.0001$ (two-tailed unpaired t-test with Bonferroni correction, comparing wild type to each mutant). (Continued on the following page.)

Figure 4: **d**, Specific infectivity ($TCID_{50}$ /RNA genomes) of CVB3 viruses from infected tissues. Bars show mean and s.e.m.; $n = 3$. ***$P < 0.001$ (two-tailed unpaired t-test with Bonferroni, comparing wild type to each mutant). **e**, Virus titres in lungs of mice infected with either wild-type, 1-to-Stop (PA-S and HA-S) or NoStop (PA-NoS) influenza A viruses. Bars show mean and s.e.m.; $n = 6$, and data are representative of three experiments. ***$P < 0.001$ (two-tailed unpaired t-test with Bonferroni correction). **f,g**, Frequency of Stop mutations in the PA (**f**) or HA (**g**) genes of influenza A viruses from infected lungs. Boxes show median and interquartile range, whiskers show range or 1.5 interquartile range in the case of outliers (indicated by dots); $n = 20$. ***$P < 0.001$ (two-tailed unpaired t-test with Bonferroni correction, comparing wild type to each mutant). **h**, Weight loss in mice infected with 1×10^5 $TCID_{50}$ of influenza variants. Graphs show mean and s.e.m.; $n = 6$. NS, non-significant; ***$P < 0.001$ (two-way analysis of variance). **i,j**, Mice were immunized with influenza 1-to-Stop variants (PA-S and HA-S), wild type or PBS. After 21 days, serum antibody titres were determined ($n = 6$) (**i**), mice were challenged with a lethal dose of virus (1×10^6 $TCID_{50}$) and lung titres were determined 4 days after challenge ($n = 6$) (**j**). Limit of detection < 10 p.f.u. g^{-1}. Bars show mean and s.e.m. **k**, Weights of immunized mice were compared before challenge (BC) and 14 days after challenge (AC) infection; $n = 5$ per group. NS, non-significant; **$P < 0.01$ (paired t-test with Bonferroni correction).

(10- to 50-fold reduction in titres), with a larger decrease for the HA construct (Fig. 4e). Once more, the NoStop virus control was as pathogenic as the wild-type virus. The number of Stop mutations present in progeny genomes in mouse lungs was quantified for the PA (Fig. 4f) and HA (Fig. 4g) genes. In both cases, the 1-to-Stop viruses presented a fourfold increase in Stop mutations compared with wild-type virus. Attenuation was further evaluated by monitoring daily weight loss (Fig. 4h). Mice infected with wild-type virus lost a mean of 12.5% of their weight by day 5, whereas those infected with 1-to-Stop variants lost 6.5% (HAS) and 7.5% (PAS).

1.4 1-to-Stop influenza viruses are immunogenic and protect against challenge

To investigate the vaccine potential of 1-to-Stop viruses, mice were immunized with either the 1-to-Stop viruses (HAS and PAS) or phosphate-buffered saline (PBS) and, after 21 days, were challenged with wild-type virus. All infected animals seroconverted, with antibody titres ranging from 320 to 1,280, as determined by haemagglutination inhibition assays (Fig. 4i). Similar titres were obtained from animals infected with wild-type virus. Following challenge infection, no virus was detected in the lungs of HAS- and PAS-immunized mice, compared to high titres in PBS-immunized mice (Fig. 4j). By 14 days after challenge, mice immunized with 1-to-Stop variants returned to normal weight, whereas PBS-immunized mice

only recovered 50% of the weight loss (Fig. 4k), a profile similar to naive mice infected with wild-type virus (Fig. 4h).

1.5 1-to-Stop virus coupled with a low-fidelity polymerase is optimally attenuated

Our results demonstrate that relocalizing a virus in an unfavourable region of sequence space, where a copy error has a higher likelihood of generating non-sense mutations, can attenuate viruses. Moreover, the treatment of these viruses with mutagens to extrinsically increase error rates resulted in even greater loss of infectivity. Previously, we have described CVB3 polymerase variants with intrinsically increased error rates that resemble mutagenic treatment[17]. We therefore engineered the low-fidelity viral polymerase I230F into the 1-to-Stop virus to generate a 'SpeedyStop' virus. We infected mice with wild-type, 1-to-Stop or SpeedyStop viruses and quantified virus titres in pancreata (Fig. 5a) and hearts (Fig. 5b). The degree of attenuation significantly increased for SpeedyStop virus compared to its normal-fidelity counterpart. Virus was undetectable in some mouse organs as early as three days after infection, and no longer detectable in any organ by day 7. Accordingly, a survival curve of mice receiving a lethal dose of wild type and equivalent doses of 1-to-Stop and SpeedyStop viruses revealed the latter to be completely attenuated (Fig. 5c). Finally, we deep-sequenced virus from infected tissues and confirmed that SpeedyStop presented a higher number of Stop mutations in sequencing reads than the other viruses (Fig. 5d).

1.6 1-to-Stop and SpeedyStop viruses induce high levels of neutralizing antibodies and protect against lethal challenge

To evaluate the immunogenicity and protective efficacy of the 1-to-Stop viruses, mice were immunized with 1×10^5 p.f.u. of each virus, or with PBS, and blood was collected after three weeks. Mice immunized with 1-to-Stop or SpeedyStop viruses produced high levels of antibody able to neutralize 1,000 p.f.u. of wild-type CVB3 (Fig. 5e). These same mice were challenged with a lethal dose (10 LD_{50}, that is, 10 times the dose lethal to 50% of animals tested) of wild-type CVB3. Most control mice succumbed to infection after eight days, whereas all of the 1-to-Stop or SpeedyStop immunized mice were protected (Fig. 5f).

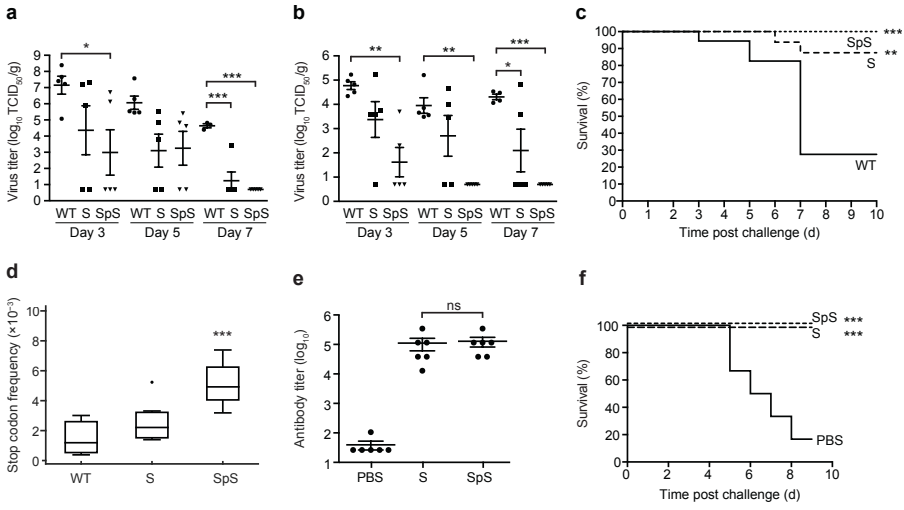


Figure 5: **Attenuation of 'SpeedyStop' virus, 1-to-Stop CVB3 coupled with a mutator polymerase. a**,**b**, Virus titres ($TCID_{50}$ per g) in pancreata (**a**) and hearts (**b**) of mice infected with 1×10^5 $TCID_{50}$ of wild type (WT), 1-to-Stop (S) or 1-to-Stop coupled with the low-fidelity polymerase mutation I230F, SpeedyStop (SpS) viruses. Bars show mean and s.e.m., and data are representative of three experiments. Day 7 values are set at the limit of detection. $*P < 0.05$, $***P < 0.001$ (two-tailed unpaired t-test with Bonferroni correction) **c**, Survival curve of mice infected with either 1×10^6 $TCID_{50}$ of wild-type (WT, solid line), 1-to-Stop (S, long dashes) or SpeedyStop (SpS, short dashes) viruses; $n = 17$. $***P < 0.001$ (Mantel–Cox test). **d**, Frequency of Stop mutations observed in deep-sequencing reads from wild-type (WT), 1-to-Stop (S) and SpeedyStop (SpS) CVB3 populations from infected tissues (hearts and pancreata combined). Boxes show median and interquartile range, whiskers show range or 1.5 interquartile range in the case of outliers, and individual dots indicate outliers; $n = 27$. $***P < 0.001$ (two-tailed unpaired t-test with Bonferroni correction, comparing wild type to each mutant). **e**, Neutralizing antibody titres (inverse dilution of sera able to neutralize 1,000 p.f.u. of wild type CVB3) in mice immunized with PBS, or 10^5 p.f.u. of 1-to-Stop (S) or SpeedyStop (SpS) viruses; $n = 6$. NS, non-significant (two-tailed unpaired t-test). **f**, Protection of mice in **e** immunized with PBS, 1-to-Stop (S, long dashes) or SpeedyStop (SpS, short dashes) and challenged with 10 LD_{50} of wild-type CVB3; $n = 6$. $***P < 0.001$ (Mantel–Cox test).

1.7 Empirical fitness distributions and landscape model

To further investigate how 1-to-Stop populations may be constrained in their exploration of sequence space and the fitness landscape, we measured the relative fitness of wild-type and 1-to-Stop CVB3 in tissue culture in biological triplicates under five mutagenic conditions, as well as under normal conditions. A range of low to high mutagenic conditions was used to accelerate evolution in sequence space and to increase the mutational load. Under normal conditions, both viruses

presented mostly positive fitness values (Fig. 6a). Under mutagenic conditions, a greater proportion of wild-type samples presented positive fitness compared to 1-to-Stop virus (Fig. 6a), and the relative decrease in fitness for 1-to-Stop virus was significant under every growth condition. In fact, only 3 populations out of 45 presented positive fitness values with respect to wild-type (Fig. 6b). We then characterized the mutant spectra of each sample by whole-genome deep sequencing to calculate the mean entropy of each population, as a general measure of sequence space exploration. For illustrative purposes, we generated a fitness landscape-like model based on the entropy values coupled with their empirical fitness values (Fig. 6c). The landscape reveals that for similar mean entropy values, 1-to-Stop populations consistently present a lower fitness and are unable to 'climb back up' to the fitness landscape occupied by wild-type virus.

2 Discussion

In this work, we present a strategy to attenuate viruses based on evolutionary principles, by restricting their evolutionary potential. Our experimental design is based on the rational relocalization of viral populations in sequence space to redirect them towards detrimental regions of the fitness landscape. Sequence space is a conceptual framework that can help monitor adaptive walks and evolutionary trajectories. For RNA viruses rapidly expanding in sequence space, emerging minority mutations can foretell the directionality of evolution well before the entire population shifts[2,20]. In theory, however, sequence space is immensely larger than the subregions occupied by biologically viable genotypes. For RNA viruses, these constraints derive from the compact nature of their genomes, and the likelihood that a mutation will hit an essential function is high. Indeed, the majority of mutations are deleterious, as evidenced in many studies that intrinsically or extrinsically increase mutation rates[3,10,21–23]. Despite this, viruses retain high mutation rates to facilitate adaptation. Consequently, mutational robustness has been suggested as an important counter-mechanism[24]. Although better characterized in theoretical and *in silico* work[25–27], some of the best experimental data have been obtained with non-coding RNA structures that assessed the retention of folding capacity or ribozyme activity in light of mutation[28]. Some evidence for mutational robustness has also been observed in RNA viruses[29–31]. A more recent study that analysed the large-scale codon reshuffling of poliovirus reported the impact of altering mutational robustness[4]. However, the extent of codon changes was such that robustness

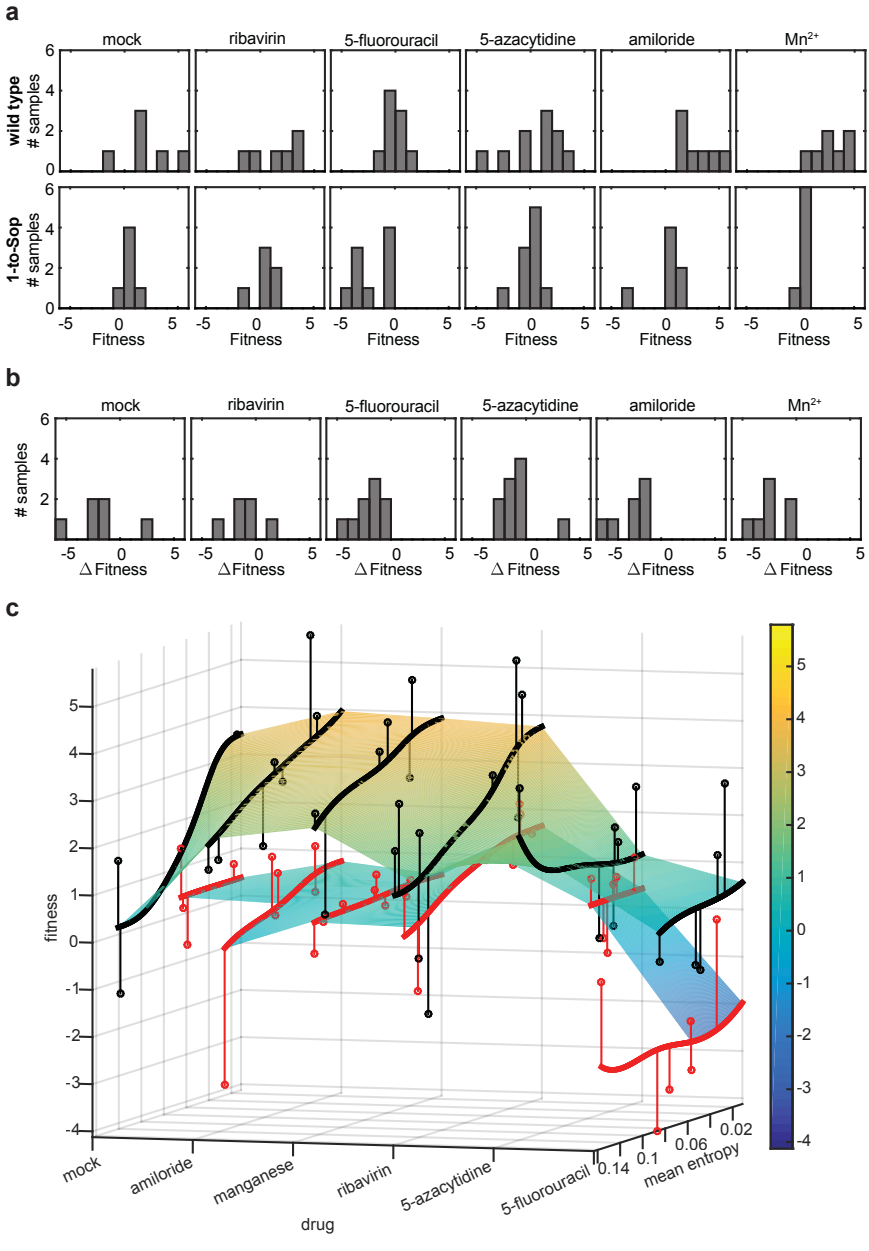


Figure 6: (Continued on the following page.)

Figure 6: **Re-localizing viruses to an inhospitable area of the fitness landscape. a**, Distribution of fitness values. The proportion (y axis, number of samples) of individual fitness values (x axis, log 10 fitness) of wild-type and 1-to-Stop populations derived from mock or mutagenic conditions. **b**, Relative change in fitness of 1-to-Stop compared to wild type under each growth condition. The differences between wild type and 1-to-Stop are significant ($P = 9.66 \times 10^{-8}$, two-tailed t-test, $n = 45$). **c**, Illustrative fitness landscape model based on empirical data. Wild-type (black) and 1-to-Stop (red) samples are shown as circles. Fitness values (log 10 fitness) are shown on the y axis and assigned colour according to the colour scale on the right, mutagenic conditions on the x axis and mean entropy calculations from whole-genome deep sequencing data on the z axis. Smoothed curves show trends in how mean entropy affects fitness for each construct and drug treatment. Lines connecting samples to the smoothed curves show deviations from the model. The smoothed curves are connected by linear interpolation to create sequence and fitness landscape surfaces, separately, for wild-type and 1-to-stop viruses.

could not be decoupled from other effects, such as CpG dinucleotide bias[15,32] or codon pair bias[14,33].

Redesigning the genomic architecture of an RNA virus, as performed here, relies on using its evolutionary potential to its own detriment. Other studies have recoded all amino acids to achieve viral attenuation, but we focused on two amino acids — leucine and serine — for two main reasons. The first was to minimize the confounding effects of manipulating the genome. In contrast to previous strategies[14,34], we targeted less than 5% of the genome. These viruses were designed to minimize or exclude the effects of codon deoptimization[35–38], codon pair deoptimization[14,33,34] and CpG/UpA dinucleotide bias[15,32,39]. We thus modified a minimum number and class of codons to retain as much as possible of the wild-type identity. However, it should be noted that although the aforementioned effects were minimized, we cannot be certain that some of these factors cannot, at least partially, account for some of the observed attenuation. The second reason for focusing only on leucine and serine was to take advantage of the unique properties of the Leu and Ser codons. These codons are the most redundant, with the greatest range of exploration of sequence space. This feature was addressed in the mathematical framework designed by Archetti[40] and based on McLachlan's chemical similarity matrix[41], which predicts the potential effect of point mutations over synonymous codons. These codons could be clustered into three groups, among which two were of particular interest for our design. The 1-to-Stop group presented the highest likelihood of changes into non-sense mutational targets (NSMTs). The other set contained the NoStop Ser/Leu codons, which are two mutations away from becoming a Stop codon, and made ideal controls with the same num-

ber of modifications.

By placing 1-to-Stop populations closer to detrimental mutational neighbourhoods, we conferred to them a lower mutational robustness that would be prone to the most detrimental type of mutation, the Stop mutation. Consequently, proximity to 'hostile' regions in sequence space may drive viruses to regions of lower fitness than a normal, more evolvable population[42]. It should be noted that, in addition to affecting mutational robustness, changes in adaptability and evolvability may also play into the attenuated phenotypes observed here. In addition, our findings experimentally support the notion that viruses may avoid volatile codons that neighbour Stop codons, as was proposed by Plotkin and Dushoff for influenza HA genes[43]. Moreover, deep sequencing supported the proposed mechanism of attenuation: an increase of three- to sixfold in Stop mutation frequencies. It should be noted that the absolute number of Stop mutations made during infection is probably higher than that captured at the moment of sampling, because such genomes cannot replicate and would thus be degraded and cleared.

Regarding *in vivo* studies, the 1-to-Stop viruses exhibited clear, attenuated phenotypes. We have demonstrated that a 1-to-Stop vaccine stock could be readily produced in cell culture with genetic and phenotypic stability, is both immunogenic and protective against lethal infection, and that coupling a mutator polymerase to this construct enhances attenuation without compromising immunogenicity. *In vivo* attenuation was also associated with higher frequencies of Stop mutations in target organs and a higher loss of infectivity. Additionally, translation of RNAs containing Stop codons would result in truncated proteins that may contribute to a better activation of the immune system[44,45]. Moreover, although non-viable genomes do not replicate, their presence could have further impact through defective interference[46,47] or through an adjuvant effect, as observed for defective interfering genomes[48,49].

It is important to emphasize that our measure of Stop mutations relies on deep sequencing technology that presents background noise. We took several measures to reduce this concern. From a mathematical standpoint, we modelled the noise in our samples and confirmed that our data were unlikely to be affected (see Supplementary Information). In addition, the background noise was further reduced by the maximum likelihood estimation of the frequencies, which takes base Phred quality scores into account. From an experimental and biological standpoint, we implemented many controls, such as the NoStop viral genomes that are as genetically modified as the 1-to-Stop genomes. We also increased sample sizes

and biological replicates, performed work in two different RNA viruses, intrinsically (fidelity variants) and extrinsically (mutagenic treatment) altered mutation pressure, and performed work in different cell types and animals. In total, 25 independent experiments, *in vitro* and *in vivo*, produced 420 sequencing samples to determine Stop mutations. In all cases, 1-to-Stop generated more Stop mutations than wild-type virus and NoStop controls. Nevertheless, one must consider the noise within the data and must not rely on the absolute values of the measures, but rather their relative values within each experiment with respect to the controls.

We developed this approach in two viral families, the Picornaviridae and Orthomyxoviridae, to cover the broad range of RNA virus biology. CVB3 is a non-enveloped, positive-sense, non-segmented, single-stranded RNA virus, whereas influenza A is an enveloped, negative-sense, segmented, single-stranded RNA virus. Finally, by modelling empirical data into an illustrative fitness landscape, we demonstrated that 1-to-Stop populations cannot escape the detrimental regions of sequence space associated with lower fitness. Our data suggest that this strategy could be broadly applied to potentially any RNA virus. In summary, our results are a proof of concept that viral genomes can be re-engineered to change their starting position in sequence space and redirect them towards detrimental mutational neighbourhoods, to generate 'suicidal', self-limiting vaccine strains.

3 Methods

3.1 Cells and viruses

HeLa and MDCK cells were obtained from the American Type Culture Collection without further authentication by our laboratory, but were confirmed to be mycoplasma-free. HeLa cells were maintained in DMEM medium with 10% new born calf serum (NBCS), and MDCK cells were maintained in MEM medium with 5% fetal calf serum (FCS). Wild-type Coxsackie virus B3 (Nancy strain), 1-to-Stop and No-Stop variants were generated from a pCB3-Nancy infectious cDNA plasmid. Wild-type influenza A virus (A/Paris/2590/2009 (H1N1pdm09)), 1-to-Stop and No-Stop variants were generated from bidirectional reverse genetics plasmids (provided by S. van der Werf at the Institut Pasteur). We generated 1-to-Stop and No-Stop viruses of Coxsackie and influenza A that bear 117 and 111/94 different synonymous codons, respectively, by *de novo* synthetic gene technology (Eurogentec). All newly generated DNA plasmids were Sanger sequenced in full

(GATC Biotech) to confirm each of the 117/111/94 positions. A detailed list of all codon changes introduced is provided in Supplementary Table 1. The low-fidelity 1-to-Stop virus, named SpeedyStop, was generated in CVB3 by insertion of the I230F mutation in the viral polymerase three-dimensional gene by site-directed mutagenesis of the 1-to-Stop CVB3 infectious clone.

3.2 Generation of Coxsackie virus stocks by in vitro transcription and transfection

CVB3 cDNA plasmids were linearized with Sal I. Linearized plasmids were purified with the Macherey-Nagel PCR purification kit. Linearized plasmid (5 µg) was *in vitro* transcribed using T7 RNA polymerase (Fermentas). Transcript (10 µg) was electroporated into HeLa cells, which were washed twice in PBS (w/o Ca^{2+} and Mg^{2+}) and resuspended in PBS (w/o Ca^{2+} and Mg^{2+}) at 1×10^7 cells per ml. Electroporation conditions were as follows: 0.4 mm cuvette, 25 mF, 700 V, maximum resistance, exponential decay in a Biorad GenePulser XCell electroporator. Cells were recovered in DMEM. A 500 µl volume of p0 virus stocks was used to infect fresh HeLa cell monolayers for three more passages. For each passage, virus was collected by three freeze-thaw cycles and clarified by spinning at 10,000 r.p.m. for 10 min. Three independent stocks were generated for each virus. Consensus sequencing of virus stocks used in downstream experiments confirmed the stability of the engineered mutations and did not detect any additional mutations across the genome.

3.3 Generation of influenza A virus stocks by reverse genetics

Using 35 mm plates and DMEM supplemented with 10% FCS, co-cultures of 293T (4×10^5 per well) and MDCK (3×10^5 per well) cells were transfected with the eight bidirectional plasmids both driving protein expression and directing vRNA template synthesis, using 0.5 µg of each plasmid and 18 µl of FUGENE HD (Roche). DNA and transfection reagents were first mixed, then incubated at room temperature for 15 min and finally added to cells, which were then incubated at 35 °C. Sixteen hours later, the DNA-transfection reagent mix was removed, cells were washed twice in DMEM, and 2 ml of DMEM containing 1 µg ml^{-1} of L-1-tosylamido-2-phenyl chloromethyl ketone treated trypsin (TPCK-trypsin, Sigma-Aldrich) was added. Cells were incubated at 35 °C for two more days, su-

pernatants were collected and clarified, and virus was titrated by median tissue culture infectious doses ($TCID_{50}$) as described below. Three independent stocks were generated for each virus. Consensus sequencing of virus stocks used in downstream experiments confirmed the stability of the engineered mutations and did not detect any additional mutations across the genome.

3.4 Genetic stability of viruses

To evaluate its genetic stability, all generated viruses were passaged ten times in HeLa cells (CVB3) or in MDCK cells (influenza A) at low MOI (0.01), and passages 1, 3, 5, 7 and 10 were sequenced.

3.5 Viral titres by $TCID_{50}$

Tenfold serial dilutions of virus were prepared in serum-free DMEM media. Dilutions were performed in 12 replicates, and 100 µl of dilution was transferred to 1×10^4 Vero-E6 (Coxsackie virus) or MDCK (influenza A virus) cells plated in 100 µl DMEM. After five days, living cell monolayers were fixed and stained with crystal violet 0.2%.

3.6 Viral titres by plaque assay

Vero-E6 (Coxsackie virus) or MDCK-SIAT (influenza A virus) cells were seeded into six-well plates and virus preparations were serially diluted (tenfold) in DMEM serum-free medium. Cells were washed twice with PBS and infected with 250 µl dilution for 30 min at 37 °C, after which a solid overlay comprising DMEM medium and 1% wt/vol agarose (Invitrogen) was added. Two days after infection, cells were fixed and stained with crystal violet 0.2%, and plaques were enumerated.

3.7 Replication kinetics and quantification of viral genomes

For growth kinetics, HeLa (Coxsackie virus) or MDCK (influenza A virus) cells were infected at an MOI of 1 or 0.1, frozen at different time points after infection, and later titred. Coxsackie viruses were collected by three freeze-thaw cycles, and influenza A viruses were collected in clarified supernatant. For real-time reverse transcription polymerase chain reaction (qRT-PCR) analysis of Coxsackie virus, total RNA was extracted by TRIzol reagent (Invitrogen) and purified. The TaqMan RNA-to-C_t one-step RT-PCR kit (Applied Biosystems) was used

to quantify viral RNA. Each 25 µl reaction contained 5 µl RNA, 100 µM each primer (forward 5'-GCATATGGTGATGATGTGATCGCTAGC-3' and reverse 5'-GGGGTACTGTTCATCTGCTCTAAA-3') and 25 pmol probe 5'-[6-Fam] GGTTACGGGCTGATCATG-3' in an ABI 7000 machine. Reverse transcription was performed at 50 °C for 30 min and 95 °C for 10 min, and was followed by 40 cycles at 95 °C for 15 s and 60 °C for 1 min. A standard curve ($y = -0.2837x + 12{,}611$, $R^2 = 0.99912$) was generated using *in vitro* transcribed genomic RNA. For influenza A virus, a similar Taqman methodology was used based on the WHO-recommended M-segment detection method. The qRT-PCR protocol consisted of an initial reverse transcription step (45 °C for 15 min), followed by an activation step of 3 min at 95 °C, 50 amplification cycles with 10 s at 95 °C, 10 s at 55 °C and 20 s at 72 °C and a final cooling step of 30 s at 40 °C. The primers used were forward: 5' CTT CTA ACC GAG GTC GAA ACG TA 3' and reverse: 5' GGT GAC AGG ATT GGT CTT GTC TTT A 3'. The probe was (HEX): 5' TCA GGC CCC CTC AAA GCC GAG 3'. The Ct values were converted into numbers of vRNA copies using a standard curve obtained with a serial dilution of a quantified synthetic M-segment RNA transcript.

3.8 Viral passages under mutagenic conditions

The mutagenic compounds (Sigma Aldrich) used were:

Ribavirin IUPAC 1-[(2*R*,3*R*,4*S*,5*R*)-3,4-dihydroxy-5-(hydroxy-methyl)oxolan-2-yl]-1*H*-1,2,4-triazole-3-carboxamide): 50, 100 and 200 µM for Coxsackie viruses and 5 and 20 µM for influenza A viruses; 5-Fluorouracil IUPAC 5-fluoro-1*H*-pyrimidine-2,4-dione: 50, 100 and 200 µM for Coxsackie viruses and 5 and 30 µM for influenza A viruses; 5-Azacitidine IUPAC 4-amino-1-*b*-D-ribofuranosyl-1,3,5-tria-zin-2(1*H*)-one: 50, 100 and 200 µM for Coxsackie viruses and 5 and 15 µM for influenza A viruses; Amiloride IUPAC 3,5-diamino-6-chloro-*N*-(diamino-methylene) pyrazine-2-carboxamide: 100 and 200 µM for Coxsackie viruses; Manganese (Mn^{2+}), 0.5 mM and 1 mM for Coxsackie viruses.

HeLa (Coxsackie virus B3) or MDCK (influenza A virus) cell monolayers in six-well plates were pretreated for 4 h with ribavirin, AZC, 5-FU, $MnCl_2$ or amiloride compounds at different concentrations. Cells were then infected at an MOI of 0.1 for Coxsackie and 0.001 for influenza A virus with passage 2 viruses. At 48 h post-infection, Coxsackie viruses were collected by three freeze-thaw cycles, and influenza A viruses were collected in clarified supernatant. Virus titres ($TCID_{50}$ or

plaque assay) were determined. The same procedure was repeated for five passages under each different mutagenic condition in three biological replicates, except for influenza A viruses, which were passaged only in low mutagenic conditions in ribavirin, 5-FU and 5-AZC.

3.9 Measurement of plaque size

Coxsackie virus plaque measurements were performed on subconfluent monolayers of 1×10^7 Vero-E6 cells in 10 cm dishes. To ensure non-overlapping plaques the amount of virus was determined empirically (40–70 per dish for Coxsackie). Each plate was scanned individually after 30 h post-infection at 300 d.p.i. Sixteen-bit image files were analysed using ImageJ. The same protocol was used to measure the plaque phenotype of mutagen pre-treated viral populations. Wild-type and 1-to-Stop viruses were submitted to high concentrations of ribavirin, 5-FU and AZC and time post-infection was increased to 40 h to better recover viral viability to perform plaque measures.

3.10 Highly quantitative direct competition assay for empirical fitness measures

For Coxsackie virus, relative fitness values were obtained by competing wild-type, NoStop and 1-to-Stop virus, obtained from different passages under each mutagen/compound, with a marked reference virus that contained four adjacent silent mutations in the polymerase region introduced by direct mutagenesis. Co-infections were performed in triplicate at an MOI of 0.01 using a 1:1 mixture of each variant with the reference virus. After 24 h, supernatants were collected and one volume of TRIzol reagent (Invitrogen) was added to extract the viral RNA. The proportion of each virus was determined by qRT-PCR on extracted RNA using a mixture of Taqman probes labelled with two different fluorescent reporter dyes. MGB_CVB3_WT detects wild-type and 1-to-Stop viruses with the sequence CGC**ATCG**TACCCATGG, and was labelled at the 5' end with a 6FAM dye; MGB_CVB3_Ref contains the four silent mutations, CGC**TAGC**TACCCATGG, and was labelled with a 5' VIC dye. Each 25 µl reaction contained 5 µl RNA, 900 nM of each primer (forward primer, 5'-GATCGCATATGGTGATGATGTGA-3'; reverse primer, 5'-AGCTTCAGCGAGTAAAGATGCA-3') and 150 nM of each probe. Using a known standard for the wild-type and reference virus during the qRT-PCR, we were able to calculate the RNA concentration for

each viral variant with high sensitivity. The relative fitness was determined by the method described in the work by Carrasco et al.[16,17], using the RNA determinations for each virus. Briefly, the formula $W = (R(t)/R(0))^{1/t}$ represents the fitness W of each mutant genotype relative to the common competitor reference sequence, where $R(0)$ and $R(t)$ represent the ratio of mutant to reference virus densities in the inoculation mixture and at t days post-inoculation (1 day in this case), respectively. The fitness of the normal wild type to reference virus was 1.019, indicating no significant differences in fitness caused by the silent mutations engineered in the reference virus (competitor).

3.11 In vitro replication assays in crude membranes

Three confluent T25 flasks of HeLa cells were infected at an MOI of 3 with wild-type or 1-to-Stop CVB3 viruses. After 16 h of infection, cells were trypsinized, collected and washed with ice-cold PBS, then resuspended in 1 ml swelling buffer made of 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 1.5 mM MgCl, one tablet of protease inhibitor (Complete Mini EDTA-free, Roche) diluted in autoclaved water. Cells were stored for 15 min on ice, then Dounce-homogenized with 30–40 strokes using a 7 ml Dounce All-Glass tissue grinder (Kimble-Chase). The unbroken cells and nuclei were removed by centrifugation at 500*g* for 5 min and the supernatant fraction was centrifuged at 12,000*g* for 10 min at 4 °C. The pellet was suspended in 1 ml of buffer (10 mM Tris hydrochloride pH 8.0, 10 mM NaCl, 15% glycerol). This wash step was repeated three times and the pellet was resuspended in 200 µl of storage buffer made of 250 mM sucrose, 10 mM Tris-HCl pH 7.4, 10 mM NaCl, and one tablet of protease inhibitor diluted in autoclaved water. Protein quantity was estimated by Bradford assay. Pellets were diluted at 10 mg ml^{-1}, aliquoted, and stored at -80 °C. To perform the *in vitro* replication assay, 25 µl of membrane extract were mixed with 25 µl replication solution made of 1× *in vitro* transcription buffer (SP6 mMESSAGE mMACHINE kit, Ambion), 10 mM dithiothreitol (Invitrogen), 10 µg ml^{-1} actinomycin D, 5 mM creatine phosphate, 25 µg ml^{-1} creatine phosphokinase, 1 mM ATP, 1 mM GTP, 1 mM CTP, 50 µM UTP, 1 µl RNAse out recombinant ribonuclease inhibitor (Invitrogen) in autoclaved water, 2 µg *in vitro* transcribed viral RNA and 20 µCi UTP [α-^{32}P] (PerkinElmer). Samples were incubated for 2 h at 37 °C. RNA extraction was performed using phenolchlorophorm, then samples were purified on Illustra MicroSpin S200 HR columns. Samples were run on a 1% agarose gel, dried in a gel-drier machine, and imaged using a Typhoon FLA9500 (GE Healthcare).

3.12 Mouse husbandry and ethics

Mice were kept in the Pasteur Institute animal facilities under Biosafety Level2 conditions, with water and food supplied *ad libitum*, and they were handled in accordance with the Animal Committee regulations of the Institut Pasteur in Paris, France, in accordance with the 2010/63 EU directive adopted on 22 September 2010 by the European Parliament and the European Union Council. Mouse protocols 2013-0101 and 2013-0021 were evaluated and approved by the Ethics Committee on Animal Experimentation CETEA no. 89 (Institut Pasteur), working under the French national Ministère de l'Enseignement supérieur et de la Recherche (MESR). All studies were carried out in BALB/c male mice (between five and six weeks old) from Charles River.

3.13 Coxsackie virus infections in vivo

Mice were infected intraperitoneally with 1×10^5 $TCID_{50}$ wild-type or 1-to-Stop viruses in 0.20 ml volumes. For tissue tropism studies, we collected whole organs (pancreas and heart) at 3, 5 and 7 days post-infection, and these were homogenized in PBS using a Precellys 24 tissue homogenizer (Bertin Technologies). Viral RNA was extracted using TRIzol reagent (Invitrogen). Full genome PCR, viral titres by $TCID_{50}$ and qRT-PCR were performed as described above. Survival curves were generated by injecting four-week-old mice ($n = 8$ mice per virus) with 5×10^6 $TCID_{50}$ of virus and monitoring morbidity and mortality for 10 days after infection. For protection studies, mice were immunized with PBS or 5×10^5 $TCID_{50}$ of 1-to-Stop or SpeedyStop virus. At 21 days after immunization, serum was collected to quantify the production of neutralizing antibodies. Mice were then challenged with 1×10^6 of wild-type virus (hyper-virulent strain 372V of Coxsackie virus B3) and survival was monitored over the following 10 days.

3.14 Neutralization assay

At 3 weeks after immunization, serum was collected, heat-inactivated at 56 °C for 30 min, and serially diluted in DMEM, and the CVB3 stock was diluted to a working concentration of 3×10^3 $TCID_{50}$. Neutralizing antibody titres were determined by $TCID_{50}$ reduction assay in Vero-E6 cells, and 50 µl of each diluted serum sample was mixed with 50 µl of CVB3 at a working concentration and added to 96-well plates for incubation at 37 °C for 2 h. Following incubation, eight

replicates of each dilution were used to infect 1×10^4 Vero-E6 cells seeded in a 96-well plate. At 6 days post-infection, the cells were observed under a microscope for the presence of cytopathic effect (CPE). Neutralization titres were determined as the highest serum dilution that could prevent CPE in more than 50% of cells.

3.15 Influenza virus infection in vivo

Mice were infected intra-nasally with 1×10^5 $TCID_{50}$ wild-type, 1-to-Stop or NoStop viruses as a 20 µl volume (diluted in PBS). Lungs were collected at three and five days post-infection and were homogenized in PBS using a Precellys 24 tissue homogenizer (Bertin Technologies). Infectious virus within homogenized tissues was titrated by plaque assay, and titres were expressed as p.f.u. per g organ (p.f.u. g^{-1}). Viral RNA was extracted using TRIzol reagent (Invitrogen). Virus genomic variability was evaluated by deep sequencing, as described in the following.

3.16 Serum antibody titre by haemagglutination inhibition assay

Mice were infected intra-nasally with 1×10^4 p.f.u. of wild-type, 1-to-Stop or NoStop viruses and bled for serum on day 21 post-infection. Antibody titres correspond to the maximum dilution able to inhibit agglutination of red blood cells in the presence of influenza virus under standardised conditions, as previously described[50].

3.17 Full genome analysis by deep sequencing

To estimate the population diversity of variants by deep sequencing, Coxsackie virus cDNA libraries were performed using the kit Maxima H Minus First Strand cDNA Synthesis (Thermofisher) and oligo dT as a primer from RNA extracted from virus generated in HeLa cells or different mouse organs. The viral genome was amplified using a high-fidelity polymerase (Phusion) to generate 1 amplicon of 7.2 kb in length (full-length genome). The primers and PCR were designed and optimized in the laboratory (5' GAAAACGCGGGGAGGGTCAAA 3' and 5' ACCCCCTCCCCCAACTGTAA 3'). For influenza A virus, the viral RNA genome was extracted from infected-cell supernatants (Macherey-Nagel), reverse-transcribed with an Accuscript High Fidelity 1st strand cDNA Synthesis kit (Agi-

lent) using 5'-AGCRAAAGCAGG-3' primer and amplified by PCR using a high-fidelity polymerase (Phusion). Eight PCRs were designed to cover the coding regions of the eight genomic segments (primer sequences are available upon request). For mouse organs, RNA was extracted with TRIzol reagent (Invitrogen) and PA and HA segments were targeted by PCR. The PCR products were purified and fragmented (Fragmentase), multiplexed, clustered on cBot for sequencing in GAIIX, or clustered and sequenced on NextSeq500, Illumina technology and analysed with established deep-sequencing data analysis tools and in-house scripts.

3.18 Codon frequencies

The sequenced reads for each sample were aligned to their respective reference genomes using BWA. Per-site codon frequencies were estimated for each sample by considering the reads covering the given site. Only nucleotides with Phred base quality scores ≥ 30 were used. The observed codon frequencies were modelled by a multinomial distribution, observed under noise. The noise model is given directly by the Phred base quality scores, which describe the probability of a read error at each nucleotide in each read. Finally, the ML (maximum likelihood) estimates of the codon frequencies were computed numerically using this model. Under perfect quality scores, the model would simplify to a multinomial model and each estimated codon frequency would correspond to the proportion of reads with that codon. However, with actual quality scores, the impact of read errors is reduced, because the model corrects for the read error rate. The mathematical model for background error is further described in the Supplementary Information.

3.19 Stop codons

Per-sample Stop codon frequencies were computed by summing the Stop codon frequencies over all modified codon sites, giving a number approximately equal to the frequency of viral genomes that have been rendered unviable by incorporating a Stop codon at one of the modified sites. The computations were done for all samples, and next-generation sequencing batch effects were avoided by only comparing samples obtained on the same sequencing runs. Box plots and linear regression plots were used to visualize the frequency distributions for relevant groups and covariates.

3.20 Fitness distribution graphs

Histograms were generated showing empirical fitness values with the samples grouped by construct and mutagenic conditions. The difference in fitness (ΔFitness) between pairs of wild-type and 1-to-Stop codons from the same experimental conditions were also computed and shown in histograms, again grouped by mutagen.

3.21 Entropy calculation from deep sequencing data

The entropy $-\sum_{i=1}^{64} \mathsf{P}(x_i)\log(\mathsf{P}(x_i))$, at a given codon site in a given sample, was computed directly from the ML estimates of the codon frequencies. Then, for each sample, the mean entropy was computed over all codon sites in the entire genome. Smoothed curves, capturing trends for the mapping of mean entropy to empirical fitness, were created for each construct and mutagen. The curves were then linearly interpolated between mutagens (roughly ordered by mutagen characteristics) to create an illustrative landscape surface.

3.22 Statistical methods

No statistical methods were used to predetermine the sample size. All experiments were performed three times and n values represent biological replicates. Equal variance was assumed. P values ≥ 0.05 were considered non-significant. For deep-sequencing analysis, when outliers were identified they were indicated in the figures and legends. For animal studies, mice were randomly allocated to different cages before experiments, and no mice were excluded from analyses. The investigator was blinded to group allocation when virus was titred from collected tissues.

3.23 Data availability

The data that support the findings of this study are available from the corresponding author upon request. In-house codes are also available at any time upon request to the authors.

Acknowledgements

The authors thank M. Declercq, M. Barbet, S. Van der Werf, C. Saleh and A. Pizzorno for assistance and discussions. This work was supported by an Institut Pasteur Innovative Ideas in Vaccinology – GPF Vaccinology grant, an ERC PoC grant no. 727758 and the Roux- Cantarini Fellowship (G.M.).

Author contributions

G.M., C.B. and M.V. designed the experiments. G.M., C.B., L.C., A.V.B., H.B., E.Z.P., S.B., T.V. and B.C.M. performed experiments. G.M., R.H., J.B., C.B., M.F. and M.V. analysed the data. G.M. and M.V. wrote the paper.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at `www.nature.com/reprints`.

Correspondence and requests for materials should be addressed to M.V.

How to cite this article: Moratorio, G. *et al.* Attenuation of RNAviruses by redirecting their evolution in sequence space. *Nat. Microbiol.* **2**, 17088 (2017).

Competing interests

The Institut Pasteur has filed a European Patent Application that covers the alteration of RNA virus sequence space and in particular the manipulation of codons to skew towards non-sense mutation targets such as Stop mutations, as a means of attenuating viruses. G.M., C.B. and M.V. are listed inventors (World Patent Application no. WO2016120412).

Bibliography

[1] Esteban Domingo and John J Holland. RNA virus mutations and fitness for survival. *Annual Reviews in Microbiology*, 51(1):151–178, 1997.

[2] Antonio V Bordería, Ofer Isakov, Gonzalo Moratorio, Rasmus Henningsson, Sonia Agüera-González, Lindsey Organtini, Nina F Gnädig, Hervé Blanc, Andrés Alcover, Susan Hafenstein, et al. Group Selection and Contribution of Minority Variants during Virus Adaptation Determines Virus Fitness and Phenotype. *PLoS Pathogens*, 11(5):e1004838, 2015.

[3] Ashley Acevedo, Leonid Brodsky, and Raul Andino. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature*, 505(7485):686, 2014.

[4] Adam S Lauring, Ashley Acevedo, Samantha B Cooper, and Raul Andino. Codon usage determines the mutational robustness, evolutionary capacity, and virulence of an RNA virus. *Cell Host & Microbe*, 12(5):623–632, 2012.

[5] Alex R Hall, Victoria F Griffiths, R Craig MacLean, and Nick Colegrave. Mutational neighbourhood and mutation supply rate constrain adaptation in Pseudomonas aeruginosa. *Proceedings of the Royal Society of London B: Biological Sciences*, 277(1681):643–650, 2010.

[6] Claus O Wilke. Adaptive evolution on neutral networks. *Bulletin of Mathematical Biology*, 63(4):715–730, 2001.

[7] Claus O Wilke, Jia Lan Wang, Charles Ofria, Richard E Lenski, and Christoph Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333, Jul 2001. ISSN 0028-0836. doi: 10.1038/35085569. URL `http://dx.doi.org/10.1038/35085569`.

[8] Rafael Sanjuán. Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365 (1548):1975–1982, 2010.

[9] José M Cuevas, Pilar Domingo-Calap, and Rafael Sanjuán. The fitness effects of synonymous mutations in DNA and RNA viruses. *Molecular Biology and Evolution*, 29(1):17–20, 2011.

[10] Rafael Sanjuán, Andrés Moya, and Santiago F Elena. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8396–8401, 2004.

[11] Jon P Anderson, Richard Daifuku, and Lawrence A Loeb. Viral error catastrophe by mutagenic nucleosides. *Annual Review of Microbiology*, 58:183–205, 2004.

[12] Celia Perales, Verónica Martín, and Esteban Domingo. Lethal mutagenesis of viruses. *Current Opinion in Virology*, 1(5):419–422, 2011.

[13] Santiago F Elena. RNA virus genetic robustness: possible causes and some consequences. *Current Opinion in Virology*, 2(5):525–530, 2012.

[14] J Robert Coleman, Dimitris Papamichail, Steven Skiena, Bruce Futcher, Eckard Wimmer, and Steffen Mueller. Virus attenuation by genome-scale changes in codon pair bias. *Science*, 320(5884):1784–1787, 2008.

[15] Fiona Tulloch, Nicky J Atkinson, David J Evans, Martin D Ryan, and Peter Simmonds. RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. *Elife*, 3:e04531, 2014.

[16] Purificacion Carrasco, José-Antonio Daròs, Patricia Agudelo-Romero, and Santiago F Elena. A real-time RT-PCR assay for quantifying the fitness of Tobacco etch virus in competition experiments. *Journal of Virological Methods*, 139(2):181–188, 2007.

[17] Nina F Gnädig, Stéphanie Beaucourt, Grace Campagnola, Antonio V Bordería, Marta Sanz-Ramos, Peng Gong, Hervé Blanc, Olve B Peersen, and Marco Vignuzzi. Coxsackievirus B3 mutator strains are attenuated in vivo. *Proceedings of the National Academy of Sciences*, 109(34):E2294–E2303, 2012.

[18] Laura I Levi, Nina F Gnädig, Stéphanie Beaucourt, Malia J McPherson, Bruno Baron, Jamie J Arnold, and Marco Vignuzzi. Fidelity variants of RNA dependent RNA polymerases uncover an indirect, mutagenic activity of amiloride compounds. *PLoS Pathogens*, 6(10):e1001163, 2010.

[19] Seth McDonald, Andrew Block, Stéphanie Beaucourt, Gonzalo Moratorio, Marco Vignuzzi, and Olve B Peersen. Design of a genetically stable high fidelity coxsackievirus B3 polymerase that attenuates virus growth in vivo. *Journal of Biological Chemistry*, 291(27):13999–14011, 2016.

[20] Ofer Isakov, Antonio V Bordería, David Golan, Amir Hamenahem, Gershon Celniker, Liron Yoffe, Hervé Blanc, Marco Vignuzzi, and Noam Shomron. Deep sequencing analysis of viral infection and evolution allows rapid and detailed characterization of viral mutant spectrum. *Bioinformatics*, 31(13): 2141–2150, 2015.

[21] Lawrence A Loeb, John M Essigmann, Farhad Kazazi, Jue Zhang, Karl D Rose, and James I Mullins. Lethal mutagenesis of hiv with mutagenic nucleoside analogs. *Proceedings of the National Academy of Sciences*, 96(4):1492–1497, 1999.

[22] Shane Crotty and Raul Andino. Implications of high RNA virus mutation rates: lethal mutagenesis and the antiviral drug ribavirin. *Microbes and Infection*, 4(13):1301–1307, 2002.

[23] Antonio V Bordería, Kathryn Rozen-Gagnon, and Marco Vignuzzi. Fidelity variants and RNA quasispecies. In *Quasispecies: From Theory to Experimental Systems*, pages 303–322. Springer, 2015.

[24] Adi Stern, Simone Bianco, Ming Te Yeh, Caroline Wright, Kristin Butcher, Chao Tang, Rasmus Nielsen, and Raul Andino. Costs and benefits of mutational robustness in RNA viruses. *Cell Reports*, 8(4):1026–1036, 2014.

[25] Claus O Wilke and Christoph Adami. Evolution of mutational robustness. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 522(1):3–11, 2003.

[26] Rafael Sanjuán, Javier Forment, and Santiago F Elena. In silico predicted robustness of viroid RNA secondary structures. II. interaction between mutation pairs. *Molecular Biology and Evolution*, 23(11):2123–2130, 2006.

[27] Rafael Sanjuán, Javier Forment, and Santiago F Elena. In silico predicted robustness of viroids RNA secondary structures. I. the effect of single mutations. *Molecular Biology and Evolution*, 23(7):1427–1436, 2006.

[28] Erik A Schultes and David P Bartel. One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, 289(5478):448–452, 2000.

[29] Rafael Sanjuán, José M Cuevas, Victoria Furió, Edward C Holmes, and Andrés Moya. Selection for robustness in mutagenized RNA viruses. *PLoS Genetics*, 3(6):e93, 2007.

[30] Jason D Graci, Nina F Gnädig, Jessica E Galarraga, Christian Castro, Marco Vignuzzi, and Craig E Cameron. Mutational robustness of an RNA virus influences sensitivity to lethal mutagenesis. *Journal of Virology*, 86(5):2869–2873, 2012.

[31] Francisco M Codoñer, José-Antonio Darós, Ricard V Solé, and Santiago F Elena. The fittest versus the flattest: experimental confirmation of the quasispecies effect with subviral pathogens. *PLoS Pathogens*, 2(12):e136, 2006.

[32] Nicky J Atkinson, Jeroen Witteveldt, David J Evans, and Peter Simmonds. The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. *Nucleic Acids Research*, 42(7):4527–4545, 2014.

[33] Cyril Le Nouën, Linda G Brock, Cindy Luongo, Thomas McCarty, Lijuan Yang, Masfique Mehedi, Eckard Wimmer, Steffen Mueller, Peter L Collins, Ursula J Buchholz, et al. Attenuation of human respiratory syncytial virus by genome-scale codon-pair deoptimization. *Proceedings of the National Academy of Sciences*, 111(36):13169–13174, 2014.

[34] Steffen Mueller, J Robert Coleman, Dimitris Papamichail, Charles B Ward, Anjaruwee Nimnual, Bruce Futcher, Steven Skiena, and Eckard Wimmer. Live attenuated influenza virus vaccines by computer-aided rational design. *Nature Biotechnology*, 28(7):723–726, 2010.

[35] Cara Carthel Burns, Jing Shaw, Ray Campagnoli, Jaume Jorba, Annelet Vincent, Jacqueline Quay, and Olen Kew. Modulation of poliovirus replicative

fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region. *Journal of Virology*, 80(7):3259–3272, 2006.

[36] Lluís Aragonès, Susana Guix, Enric Ribes, Albert Bosch, and Rosa M Pintó. Fine-tuning translation kinetics selection as the driving force of codon usage bias in the hepatitis A virus capsid. *PLoS Pathogens*, 6(3):e1000797, 2010.

[37] Antoine Nougairede, Lauriane De Fabritus, Fabien Aubry, Ernest A Gould, Edward C Holmes, and Xavier de Lamballerie. Random codon re-encoding induces stable reduction of replicative fitness of Chikungunya virus in primate and mosquito cells. *PLoS Pathogens*, 9(2):e1003172, 2013.

[38] Jia Meng, Sujin Lee, Anne L Hotard, and Martin L Moore. Refining the balance of attenuation and immunogenicity of respiratory syncytial virus by targeted codon deoptimization of virulence genes. *MBio*, 5(5):e01704–14, 2014.

[39] Eleanor Gaunt, Helen M Wise, Huayu Zhang, Lian N Lee, Nicky J Atkinson, Marlynne Quigg Nicol, Andrew J Highton, Paul Klenerman, Philippa M Beard, Bernadette M Dutia, et al. Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. *Elife*, 5:e12735, 2016.

[40] Marco Archetti. Genetic robustness at the codon level as a measure of selection. *Gene*, 443(1):64–69, 2009.

[41] Andrew D McLachlan. Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c_{551}. *Journal of Molecular Biology*, 61(2): 409–424, 1971.

[42] Christina L Burch and Chao Lin. Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature*, 406(6796):625, 2000.

[43] Joshua B Plotkin and Jonathan Dushoff. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proceedings of the National Academy of Sciences*, 100(12):7152–7157, 2003.

[44] Andrej Kovac, Norbert Zilka, Zuzana Kazmerova, Martin Cente, Monika Zilkova, and Michal Novak. Misfolded truncated protein τ induces innate

immune response via MAPK pathway. *The Journal of Immunology*, 187(5): 2732–2739, 2011.

[45] Mehran Dabaghian, Ali Mohammad Latifi, Majid Tebianian, Fariba Dabaghian, and Seyyed Mahmoud Ebrahimi. A truncated C-terminal fragment of Mycobacterium tuberculosis HSP70 enhances cell-mediated immune response and longevity of the total IgG to influenza A virus M2e protein in mice. *Antiviral Research*, 120:23–31, 2015.

[46] Scott Crowder and Karla Kirkegaard. Trans-dominant inhibition of RNA viral replication can slow growth of drug-resistant viruses. *Nature Genetics*, 37(7):701, 2005.

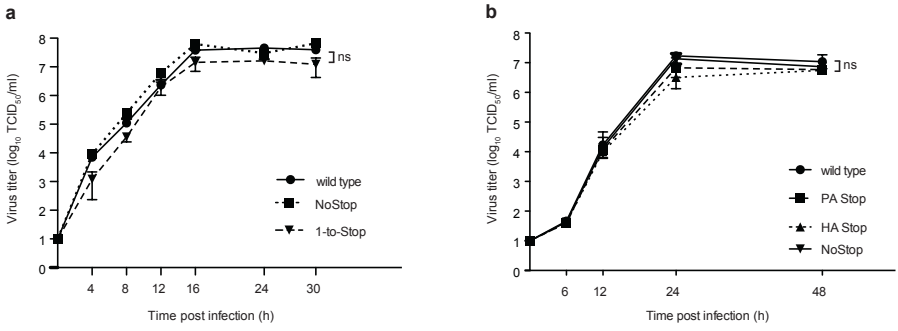
[47] Claudia González-López, Armando Arias, Nonia Pariente, Gema Gómez-Mariano, and Esteban Domingo. Preextinction viral RNA can interfere with infectivity. *Journal of Virology*, 78(7):3319–3324, 2004.

[48] Jacob S Yount, Thomas A Kraus, Curt M Horvath, Thomas M Moran, and Carolina B López. A novel role for viral-defective interfering particles in enhancing dendritic cell maturation. *The Journal of Immunology*, 177(7): 4503–4513, 2006.

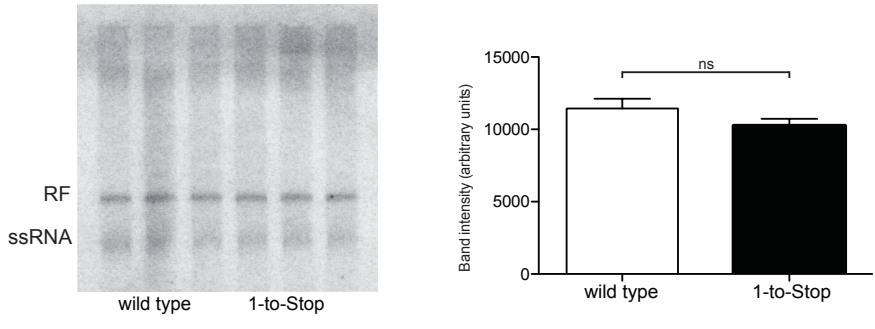
[49] Karla Tapia, Won-keun Kim, Yan Sun, Xiomara Mercado-López, Emily Dunay, Megan Wise, Michael Adu, and Carolina B López. Defective viral genomes arising in vivo provide critical danger signals for the triggering of lung antiviral immunity. *PLoS Pathogens*, 9(10):e1003703, 2013.

[50] Thierry Boge, Michel Rémigy, Sarah Vaudaine, Jérôme Tanguy, Raphaëlle Bourdet-Sicard, and Sylvie Van Der Werf. A probiotic fermented dairy drink improves antibody response to influenza vaccination in the elderly in two randomised controlled trials. *Vaccine*, 27(41):5677–5684, 2009.

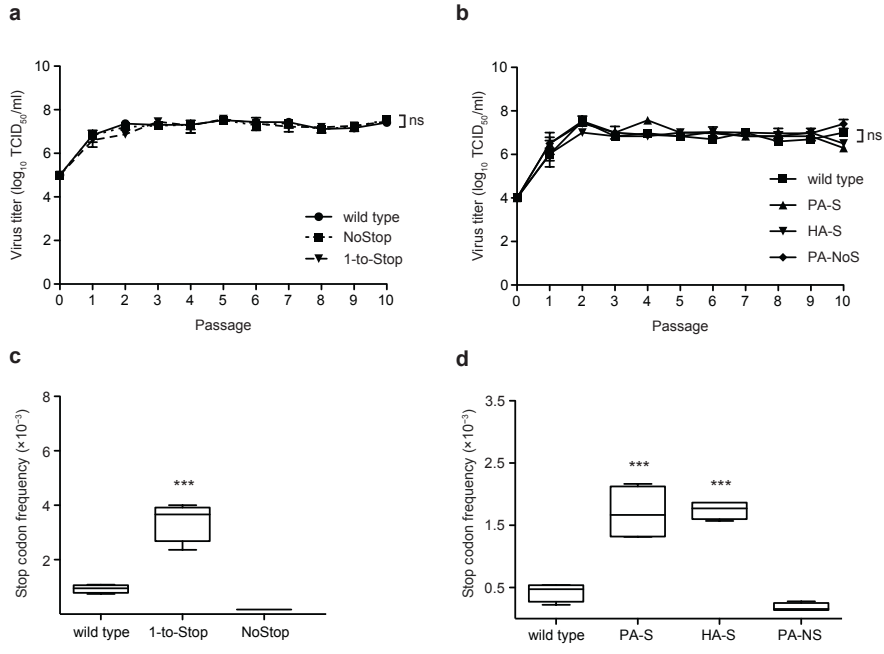
A Supplementary Figures



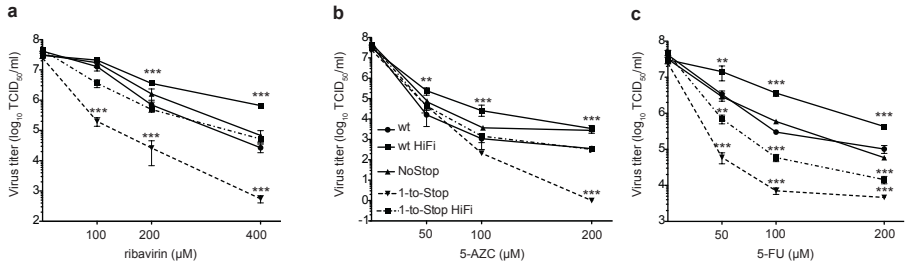
Supplementary Figure 1: (**a**) Replication kinetics of wild type, 1-to-Stop and NoStop Coxsackie virus B3 in HeLa cells infected at MOI 0.1. (**b**) Replication kinetics of wild type, 1-to-Stop PA and HA, and NoStop PA influenza A viruses in MDCK cells infected at MOI 0.1. Bars show mean and SEM; $n = 3$ per group. ns, non-significant (two-tailed unpaired t-test with Bonferroni correction comparing wild type to each mutant).



Supplementary Figure 2: **In vitro RNA replication assay.** Membranes containing replication complexes from HeLa cells infected with wild type or 1-to-Stop CVB3 viruses were purified and used for *in vitro* replication assays, by adding *in vitro* transcribed RNA, corresponding to wild type (3 samples) or 1-to Stop (3 samples) CVB3 viruses, and radiolabeled UTP. The Replicative Form (RF) and single stranded RNA (ssRNA) are visualized. Density of each band was determined by ImageJ. Bars show mean and SEM; $n = 3$ per group. ns, non-significant (two-tailed unpaired t-test).



Supplementary Figure 3: **Genetic and phenotypic stability of 1-to-Stop and NoStop viruses after serial passage in tissue culture.** (**a**, **b**) Virus titres over 10 serial passages. HeLa (**a**) or MDCK (**b**) cells were infected with CVB3 (**a**) or influenza A (**b**) variants at MOI 0.1. Virus titres were determined for each passage. Mean and SEM are shown; $n = 3$. ns, non-significant (two-way analysis of variance) (**c**,**d**) Frequency of Stop mutations observed in deep sequencing reads from wild type, 1-to-Stop, and NoStop variants from passage number 10 for CVB3 variants (**c**) or influenza A variants (**d**). Boxes show median and interquartile range, whiskers range or 1.5 interquartile range in case of outlier, individual dots indicate outliers; $n = 6$ per group. ***$P < 0.001$ (two-tailed unpaired t-test with Bonferroni correction, comparing wild type to each mutant).



Supplementary Figure 4: **1-to-Stop high-fidelity CVB3 recovers wild type phenotype in presence of mutagens.** Sensitivity of wild type (wt), wild type high-fidelity (wt HiFi), NoStop, 1-to-Stop and 1-to-Stop HiFi CVB3 viruses to increasing concentrations of **(a)** ribavirin, **(b)** 5-azacytidine (5-AZC) and **(c)** 5-fluorouracil (5-FU). Graphs show mean and SEM; $n = 3$ per group. $^{**}P < 0.01$, $^{***}P < 0.001$ (two-way analysis of variance with Bonferroni post test).

B Supplementary Methods

B.1 Mathematical assessment of stop codon background noise

The overall NGS error frequency in the context of our work, is not of interest *per se*; rather, only the fraction of errors that cause stop codons, since only those errors could affect our results. The question we ask here is whether more of fewer Stop mutations are observed at all of the altered Ser/Leu sites combined, rather than at individual sites. This increases our sample size by approximately 100-fold, thereby further increasing our statistical power. To model the frequency of 'false' stop codons due to sequencing error, we use a Poisson background noise model (a standard for estimating independent rare errors under minimal additional structural assumptions, which has already been used to model NGS error rates in the context of minority variant discovery) [1,2]. Since sequencing errors are nucleotide-context dependent [3], and the number of stop codons that can be reached by a single mutation is different for different 1-to-stop codons, we model y_{ij}, the number of observed stop codons for sample i at site j, as an observation of the random variable:

$$Y_{ij} \sim \mathsf{Po}\left(N_{ij}(\lambda_C + \mu_{ij})\right),$$

where N_{ij} is the total number of reads for the same sample and site, λ_C is the probability of observing an erroneous stop codon given the nucleotide context C, μ_{ij} is the true stop codon frequency for sample i at site j; and the sequencing errors

in different reads are assumed to be independent. The Poisson approximation is very good since N_{ij} is large and $\lambda_C+\mu_{ij}$ is small. Both the influence of nucleotide context and the number of different stop codons reachable within one mutation are captured in the λ_C parameter.

Let $\nu_{ij} \coloneqq N_{ij}(\lambda_C + \mu_{ij})$ and $P_\nu(x)$, the probability mass function of the Poisson distribution with rate ν. The log-likelihood is

$$l(\boldsymbol{\nu}; \boldsymbol{y}) = \sum_{i,j} \log P_{\nu_{ij}}(y_{ij}).$$

Hence, maximizing l is equivalent to maximizing each term separately, and the ML estimate for the Poisson distribution is given by $\hat{\nu}_{ij} = y_{ij}$. Reparameterizing in λ_C and $\boldsymbol{\mu}$, it follows that the ML estimate is achieved for any $\hat{\lambda}_C \in [0, \min_{ij} y_{ij}/N_{ij}]$, with $\hat{\mu}_{ij} = y_{ij}/N_{ij} - \hat{\lambda}_C$, since $\lambda_C \geq 0$ and $\mu_{ij} \geq 0$ for all i, j.

The log likelihood test statistic

$$D(\lambda_C^0) = 2\left(\max_{\lambda_C, \boldsymbol{\mu}} l(\lambda_C, \boldsymbol{\mu}) - \max_{\boldsymbol{\mu}} l(\lambda_C^0, \boldsymbol{\mu})\right),$$

measures the drop in log likelihood between the full model and a reduced model with λ_C fixed at λ_C^0 . By profile likelihood, a 95% confidence interval for λ_C consists of all λ_C^0 such that $D(\lambda_C^0) \leq \chi_1^2(0.95)$, where χ_1^2 is the quantile function for the χ^2 distribution with one degree of freedom, since the change in model order is one. Now, since $D(0) = 0$ and D is a decreasing function, the upper endpoint of the profile likelihood confidence interval for λ_C can be found by a binary search. Note how the ML estimate of each μ_{ij} in the reduced model is still given by $y_{ij}/N_{ij} - \hat{\lambda}_C$, but constrained such that $\mu_{ij} \geq 0$.

The model was thus applied to 5 (96-well plates each) sequencing runs, for a total of 420 samples. Only samples from the same run are compared, to ignore noise due to batch effect. Each codon site that had a 1-to-Stop codon in the wild type genome (and thus the identical codon, in the identical context, in the 1-to-Stop construct) was considered. The nucleotide contexts coincide with the 1-to-Stop codons (UUA, UUG, UCA and UCG), since a mutation that produces a stop codon must change the middle nucleotide in the codon. To produce a single confidence interval for the stop codon error rate λ for each sequencing run, the confidence intervals for the different nucleotide contexts, λ_{UUA}, λ_{UUG}, λ_{UCA}

and λ_{UCG}, were averaged with weights proportional to the number of codon sites for each context.

Sequencing Run	**Weighted 95% confidence interval for λ**
CVB3 invitro	$[0, 1.062 \cdot 10^{-6}]$
CVB3 invivo	$[0, 2.098 \cdot 10^{-6}]$
Flu HA invitro	$[0, 2.237 \cdot 10^{-6}]$
Flu PA invitro	$[0, 1.135 \cdot 10^{-7}]$
Flu HA invivo	$[0, 5.631 \cdot 10^{-6}]$
Flu PA invivo	$[0, 2.031 \cdot 10^{-7}]$

Thus, considering the 1-to-Stop codon sites, if the sequencing error rates were at the upper endpoints of the 95% confidence intervals, the contribution of sequencing errors to the stop codon frequency as it was computed (sum over all Leu/Ser) would be $\sim 10^{-6} \times 100 = 10^{-4}$ for the 1-to-Stop virus, which is still more than 10 times below the observed frequencies.

Bibliography

[1] Frank A Haight. Handbook of the Poisson distribution. 1967.

[2] Stéphanie Raymond, Florence Nicot, Nicolas Jeanne, Olivier Delfour, Romain Carcenac, Caroline Lefebvre, Michelle Cazabat, Karine Sauné, Pierre Delobel, and Jacques Izopet. Performance comparison of next-generation sequencing platforms for determining HIV-1 coreceptor use. *Scientific Reports*, 7, 2017.

[3] Matthijs RA Welkers, Marcel Jonges, Rienk E Jeeninga, Marion PG Koopmans, and Menno D de Jong. Improved detection of artifactual viral minority variants in high-throughput sequencing data. *Frontiers in Microbiology*, 5:804, 2015.

C Supplementary Tables

List of codon changes introduced in Coxsackie virus B3 and influenza A viruses. The table indicates the starting nucleotide position and identity of each codon in the Coxsackie virus B3 P1 region and the influenza A virus PA and HA genes, for wild type (**WT**), No-Stop (**NS**) and 1-to-Stop (**S**) viruses for both Serine (S) and Leucine (L).

Coxsackie virus B3					Influenza A virus HA gene					Influenza A virus PA gene				
Pos.	**WT**	**NS**	**S**	**AA**	**Pos.**	**WT**	**NS**	**S**	**AA**	**Pos.**	**WT**	**NS**	**S**	**AA**
789	TCA	TCT	TCA	S	45	CTA	CTA	TTA	L	70	CTT	CTA	TTA	L
822	CTG	CTG	TTG	L	54	CTG	CTG	TTG	L	148	TTG	CTG	TTG	L
831	AGC	TCC	TCG	S	57	CTA	CTA	TTA	L	169	TCG	TCC	TCG	S
840	TCC	TCC	TCG	S	90	TTA	CTA	TTA	L	202	TCA	TCT	TCA	S
885	TCC	TCC	TCG	S	117	TCA	TCT	TCA	S	217	TCT	TCT	TCA	S
891	TCA	TCT	TCA	S	141	CTA	CTA	TTA	L	235	CTA	CTA	TTA	L
963	TCA	TCT	TCA	S	168	TCT	TCT	TCG	S	238	TTG	CTG	TTG	L
966	CTA	CTA	TTA	L	177	CTT	CTA	TTA	L	301	AGT	TCT	TCA	S
975	CTC	CTG	TTG	L	180	CTA	CTA	TTA	L	340	CTT	CTA	TTA	L
981	TCC	TCC	TCG	S	204	CTA	CTA	TTA	L	349	TTG	CTG	TTG	L
1008	AGT	TCT	TCA	S	213	CTA	CTA	TTA	L	418	CTA	CTA	TTA	L
1023	TCA	TCT	TCA	S	231	TTG	CTG	TTG	L	442	TCT	TCT	TCA	S
1032	TTA	CTA	TTA	L	237	TTG	CTG	TTG	L	469	TCA	TCT	TCA	S
1041	TCC	TCC	TCG	S	267	CTG	CTG	TTG	L	511	CTT	CTA	TTA	L
1104	CTA	CTA	TTA	L	288	TCA	TCT	TCA	S	523	AGC	TCC	TCG	S
1113	AGT	TCT	TCA	S	291	CTC	CTG	TTG	L	547	CTT	CTA	TTA	L
1176	CTT	CTA	TTA	L	294	TCC	TCC	TCG	S	574	AGT	TCT	TCA	S
1182	TCT	TCT	TCA	S	303	AGC	TCC	TCG	S	580	AGT	TCT	TCA	S
1203	TCA	TCT	TCA	S	306	TCA	TCT	TCA	S	583	CTA	CTA	TTA	L
1224	CTG	CTG	TTG	L	312	TCC	TCC	TCG	S	592	TCC	TCC	TCG	S
1236	TTG	CTG	TTG	L	330	TCT	TCT	TCA	S	604	TCC	TCC	TCG	S
1239	TCG	TCC	TCG	S	333	AGT	TCT	TCA	S	664	CTT	CTA	TTA	L
1245	TTA	CTA	TTA	L	336	TCA	TCT	TCA	S	676	AGT	TCT	TCA	S
1251	CTG	CTG	TTG	L	384	CTA	CTA	TTA	L	679	CTC	CTG	TTG	L
1281	TTA	CTA	TTA	L	396	TTG	CTG	TTG	L	694	TCC	TCC	TCG	S
1323	TCT	TCT	TCA	S	399	AGC	TCC	TCG	S	697	AGC	TCC	TCG	S
1344	TTG	CTG	TTG	L	402	TCA	TCT	TCA	S	700	CTT	CTA	TTA	L
1347	CTA	CTA	TTA	L	408	TCA	TCT	TCA	S	760	CTT	CTA	TTA	L
1389	CTA	CTA	TTA	L	411	TCA	TCT	TCA	S	763	TCC	TCC	TCG	S
1404	TCC	TCC	TCG	S	444	AGT	TCT	TCA	S	772	TCA	TCT	TCA	S
1407	AGT	TCT	TCA	S	447	TCA	TCT	TCA	S	805	TTG	CTG	TTG	L
1416	TTG	CTG	TTG	L	465	TCG	TCC	TCG	S	826	CTC	CTG	TTG	L
1419	CTG	CTG	TTG	L	510	AGC	TCC	TCG	S	832	TTG	CTG	TTG	L
1464	TCC	TCC	TCG	S	525	TTA	CTA	TTA	L	847	CTT	CTA	TTA	L

Coxsackie virus B3					Influenza A virus HA gene					Influenza A virus PA gene				
Pos.	**WT**	**NS**	**S**	**AA**	**Pos.**	**WT**	**NS**	**S**	**AA**	**Pos.**	**WT**	**NS**	**S**	**AA**
1470	TCC	TCC	TCG	S	534	CTA	CTA	TTA	L	862	TCA	TCT	TCA	S
1479	TTG	CTG	TTG	L	552	TCA	TCT	TCA	S	871	CTG	CTG	TTG	L
1530	CTC	CTG	TTG	L	564	CTC	CTG	TTG	L	874	CTG	CTG	TTG	L
1560	CTA	CTA	TTA	L	567	AGC	TCC	TCG	S	886	CTG	CTG	TTG	L
1575	AGT	TCT	TCA	S	573	TCC	TCC	TCG	S	892	TTA	CTA	TTA	L
1605	AGT	TCT	TCA	S	603	CTC	CTG	TTG	L	895	AGT	TCT	TCA	S
1647	CTA	CTA	TTA	L	609	CTA	CTA	TTA	L	910	AGT	TCT	TCA	S
1671	CTA	CTA	TTA	L	630	TCT	TCT	TCA	S	934	CTA	CTA	TTA	L
1689	TCC	TCC	TCA	S	636	AGT	TCT	TCA	S	1027	CTC	CTG	TTG	L
1749	TTA	CTA	TTA	L	651	AGT	TCT	TCA	S	1048	CTA	CTA	TTA	L
1755	TTA	CTA	TTA	L	654	CTC	CTG	TTG	L	1057	CTA	CTA	TTA	L
1773	TTA	CTA	TTA	L	690	TCA	TCT	TCA	S	1114	AGC	TCC	TCG	S
1797	AGC	TCC	TCG	S	693	TCA	TCT	TCA	S	1120	TTG	CTG	TTG	L
1809	CTG	CTG	TTG	L	702	AGC	TCC	TCG	S	1132	CTC	CTG	TTG	L
1815	TCA	TCT	TCA	S	780	CTA	CTA	TTA	L	1192	CTT	CTA	TTA	L
1830	TCA	TCT	TCA	S	825	CTA	CTA	TTA	L	1207	AGT	TCT	TCA	S
1836	TCC	TCC	TCG	S	870	TCT	TCT	TCA	S	1228	TCT	TCT	TCA	S
1896	TTG	CTG	TTG	L	885	TCA	TCT	TCA	S	1231	CTA	CTA	TTA	L
1920	TCA	TCT	TCA	S	948	AGC	TCC	TCG	S	1237	AGC	TCC	TCG	S
1959	TCT	TCT	TCA	S	951	CTC	CTG	TTG	L	1273	TTG	CTG	TTG	L
1989	TCC	TCC	TCG	S	1008	AGC	TCC	TCG	S	1282	TCA	TCT	TCA	S
2001	TCT	TCT	TCA	S	1017	TTG	CTG	TTG	L	1285	AGC	TCC	TCG	S
2028	CTG	CTG	TTG	L	1023	CTG	CTG	TTG	L	1297	CTT	CTA	TTA	L
2043	TCG	TCC	TCG	S	1035	TTG	CTG	TTG	L	1342	AGC	TCC	TCG	S
2046	AGT	TCT	TCA	S	1050	TCT	TCT	TCA	S	1375	TCC	TCC	TCG	S
2055	AGT	TCT	TCA	S	1059	TCT	TCT	TCA	S	1429	TTG	CTG	TTG	L
2064	CTC	CTG	TTG	L	1068	CTA	CTA	TTA	L	1432	CTC	CTG	TTG	L
2067	CTA	CTA	TTA	L	1158	TCA	TCT	TCA	S	1441	TCC	TCC	TCG	S
2079	TTG	CTG	TTG	L	1176	CTG	CTG	TTG	L	1468	CTG	CTG	TTG	L
2100	TCA	TCT	TCA	S	1182	AGC	TCC	TCG	S	1483	AGC	TCC	TCG	S
2106	AGC	TCC	TCG	S	1224	TCT	TCT	TCA	S	1522	CTG	CTG	TTG	L
2115	CTT	CTA	TTA	L	1281	CTG	CTG	TTG	L	1549	TCT	TCT	TCA	S
2136	TCG	TCC	TCG	S	1302	TTA	CTA	TTA	L	1555	TTG	CTG	TTG	L
2160	CTT	CTA	TTA	L	1329	CTG	CTG	TTG	L	1588	AGT	TCT	TCA	S
2163	TTG	CTG	TTG	L	1356	CTG	CTG	TTG	L	1600	TCA	TCT	TCA	S
2172	TCA	TCT	TCA	S	1359	TTG	CTG	TTG	L	1603	CTC	CTG	TTG	L
2217	CTT	CTA	TTA	L	1365	CTA	CTA	TTA	L	1618	CTG	CTG	TTG	L
2247	CTA	CTA	TTA	L	1368	TTG	CTG	TTG	L	1651	CTT	CTA	TTA	L
2253	TCA	TCT	TCA	S	1386	TTG	CTG	TTG	L	1669	CTC	CTG	TTG	L
2256	AGT	TCT	TCA	S	1401	TCA	TCT	TCA	S	1672	TTG	CTG	TTG	L
2265	CTG	CTG	TTG	L	1416	TTA	CTA	TTA	L	1696	TCG	TCC	TCG	S
2283	AGC	TCC	TCG	S	1434	AGC	TCC	TCG	S	1711	CTA	CTA	TTA	L
2310	TCA	TCT	TCA	S	1440	CTA	CTA	TTA	L	1735	TCC	TCC	TCG	S

Coxsackie virus B3					Influenza A virus HA gene					Influenza A virus PA gene				
Pos.	**WT**	**NS**	**S**	**AA**	**Pos.**	**WT**	**NS**	**S**	**AA**	**Pos.**	**WT**	**NS**	**S**	**AA**
2385	AGC	TCC	TCG	S	1515	AGT	TCT	TCA	S	1777	CTT	CTA	TTA	L
2388	TCC	TCC	TCG	S	1551	TCA	TCT	TCA	S	1780	CTT	CTA	TTA	L
2412	TCA	TCT	TCA	S	1566	TTA	CTA	TTA	L	1786	TCT	TCT	TCA	S
2430	TCT	TCT	TCA	S	1596	CTG	CTG	TTG	L	1789	CTT	CTA	TTA	L
2439	CTA	CTA	TTA	L	1602	TCA	TCT	TCA	S	1804	AGC	TCC	TCG	S
2442	TTG	CTG	TTG	L	1623	TTG	CTG	TTG	L	1822	TCT	TCT	TCA	S
2463	TCG	TCC	TCG	S	1635	TCA	TCT	TCA	S	1825	TCT	TCT	TCA	S
2556	TCA	TCT	TCA	S	1647	AGT	TCT	TCA	S	1870	TCG	TCC	TCG	S
2574	CTC	CTG	TTG	L	1650	TCA	TCT	TCA	S	1894	TCA	TCT	TCA	S
2601	TCA	TCT	TCA	S	1653	TTG	CTG	TTG	L	1918	TCT	TCT	TCA	S
2655	TCA	TCT	TCA	S	1659	CTG	CTG	TTG	L	1942	TTA	CTA	TTA	L
2661	TCC	TCC	TCG	S	1668	TCC	TCC	TCG	S	1945	CTG	CTG	TTG	L
2667	TCA	TCT	TCA	S	1671	CTG	CTG	TTG	L	1954	TCT	TCT	TCA	S
2685	CTA	CTA	TTA	L	1683	AGT	TCT	TCA	S	1966	AGT	TCT	TCA	S
2694	TCA	TCT	TCA	S	1698	TCT	TCT	TCA	S	1969	CTA	CTA	TTA	L
2727	TCA	TCT	TCA	S	1707	TCT	TCT	TCA	S	1978	TCT	TCC	TCA	S
2757	TTA	CTA	TTA	L	1710	CTA	CTA	TTA	L	1987	CTT	CTA	TTA	L
2781	CTT	CTA	TTA	L						1999	TCG	TCC	TCG	S
2793	CTA	CTA	TTA	L						2008	TCG	TCC	TCG	S
2823	CTG	CTG	TTG	L						2017	TTG	CTG	TTG	L
2829	CTG	CTG	TTG	L						2020	CTT	CTA	TTA	L
2847	AGT	TCT	TCA	S						2023	CTC	CTG	TTG	L
2862	TCA	TCT	TCA	S						2038	CTT	CTA	TTA	L
2892	CTA	CTA	TTA	L						2050	CTG	CTG	TTG	L
2949	TCA	TCT	TCA	S						2071	CTT	CTA	TTA	L
2967	TCT	TCT	TCA	S						2080	CTA	CTA	TTA	L
2979	AGT	TCT	TCA	S						2104	CTG	CTG	TTG	L
3018	TCC	TCC	TCG	S						2125	TTG	CTG	TTG	L
3030	TTG	CTG	TTG	L						2128	CTT	CTA	TTA	L
3033	AGC	TCC	TCG	S						2137	TCT	TCT	TCA	S
3051	TCA	TCT	TCA	S						2149	TCC	TCC	TCG	S
3072	TCT	TCT	TCA	S						2155	CTC	CTG	TTG	L
3081	TCC	TCC	TCG	S						2167	CTG	CTG	TTG	L
3111	CTA	CTA	TTA	L										
3129	CTA	CTA	TTA	L										
3156	AGC	TCC	TCG	S										
3174	AGC	TCC	TCG	S										
3237	CTC	CTG	TTG	L										
3279	AGC	TCC	TCG	S										
3303	AGC	TCC	TCG	S										

PAPER III

arXiv

SMSSVD – SubMatrix Selection Singular Value Decomposition

Rasmus Henningsson and Magnus Fontes

Abstract

High throughput biomedical measurements normally capture multiple overlaid biologically relevant signals and often also signals representing different types of technical artefacts like e.g. batch effects. Signal identification and decomposition are accordingly main objectives in statistical biomedical modeling and data analysis. Existing methods, aimed at signal reconstruction and deconvolution, in general, are either supervised, contain parameters that need to be estimated or present other types of ad hoc features. We here introduce SubMatrix Selection SingularValue Decomposition (SMSSVD), a parameter-free unsupervised signal decomposition and dimension reduction method, designed to reduce noise, adaptively for each low-rank-signal in a given data matrix, and represent the signals in the data in a way that enable unbiased exploratory analysis and reconstruction of multiple overlaid signals, including identifying groups of variables that drive different signals.

The Submatrix Selection Singular Value Decomposition (SMSSVD) method produces a denoised signal decomposition from a given data matrix. The SMSSVD method guarantees orthogonality between signal components in a straightforward manner and it is designed to make automation possible. We illustrate SMSSVD by applying it to several real and synthetic datasets and compare its performance to golden standard methods like PCA

(Principal Component Analysis) and SPC (Sparse Principal Components, using Lasso constraints). The SMSSVD is computationally efficient and despite being a parameter-free method, in general, outperforms existing statistical learning methods.

A Julia implementation of SMSSVD is openly available on GitHub (https://github.com/rasmushenningsson/SMSSVD.jl).

1 Introduction

High throughput biomedical measurements, by design, normally capture multiple overlaid biologically relevant signals, but often also signals representing different types of biological and technical artefacts like e.g. batch effects. There exist different methods aimed at signal reconstruction and deconvolution of the resulting high dimensional and complex datasets, but these methods almost always contain parameters that need to be estimated or present other types of ad hoc features. Developed specifically for Omics data and more particularly gene expression data such methods include the gene shaving method [1], tree harvesting [2], supervised principal components [3] and amplified marginal eigenvector regression [4]. They employ widely different strategies do deal with the ubiquitous $P \gg N$ (many more variables than samples) problem in omics data. Gene Shaving uses the first principal component to iteratively guide variable selection towards progressively smaller nested subsets of correlated genes with large variances. An optimal subset size is then chosen using the 'gap statistic', a measure of how much better the subset is than what is expected by random chance. To find additional subsets (signals), each gene is first projected onto the orthogonal complement of the average gene in the current subset, and the whole process is repeated.

We here introduce SubMatrix Selection Singular Value Decomposition (SMSSVD), a parameter-free unsupervised dimension reduction technique primarily designed to reduce noise, adaptively for each low-rank-signal in a data matrix, and represent the data in a way that enable unbiased exploratory analysis and reconstruction of the multiple overlaid signals, including finding the variables that drive the different signals.

Our first observation for the theoretical foundation of SMSSVD is that the SVD of a linear map restricted to a hyperplane (linear subspace) share many properties with the SVD of the corresponding unrestricted linear map. Using this we show that, by iteratively choosing orthogonal hyperplanes based on criteria for op-

timal variable selection and concatenating the decompositions, we can construct a denoised decomposition of the data matrix. The SMSSVD method guarantees orthogonality between components in a straightforward manner and coincide with the SVD if no variable selection is applied. We illustrate the SMSSVD by applying it to several real and synthetic datasets and compare its performance to golden standard methods for unsupervised exploratory analysis: Classical PCA (Principal Component Analysis)[5] and the lasso or elastic net based methods like SPC (Sparse Principal Components)[6]. Just like PCA and SPC, SMSSVD is intended for use in wide range of situations, and no assumptions specific to gene expression analysis are made in the derivation of the method. The SMSSVD is computationally efficient and despite being a parameter-free method, in general, it outperforms or equals the performance of the golden standard methods. A Julia implementation of SMSSVD is openly available on GitHub.

2 Methods

Theorem 2.1. *Let* $X\big|_{\Pi} : \Pi \to X(\Pi)$ *be the restriction of a linear map* $X : \mathbb{R}^N \to \mathbb{R}^P$ *to a* d*-dimensional subspace* $\Pi \subset \mathbb{R}^N$ *such that* $\Pi \perp \ker X$*. Furthermore, let* $U\Sigma V^T = \sum_{i=1}^{d} \sigma_i U_{\cdot i} V_{\cdot i}^T$ *be the singular value decomposition of* $X\big|_{\Pi}$*. Then*

1. $V_{\cdot i} \perp \ker X, \ \forall i.$
2. $U_{\cdot i} \perp \operatorname{coker} X, \ \forall i.$
3. $XV = U\Sigma.$
4. $U^T X = \Sigma V^T + U^T X(I - VV^T).$
5. $(I - UU^T)X(I - VV^T) = (I - UU^T)X.$
6. $\operatorname{rank}(X) = d + \operatorname{rank}\big((I - UU^T)X\big).$

Remark. *In the statement of the theorem and in the proof below, we consider all vectors to belong to the full-dimensional spaces. In particular, we extend all vectors in subspaces of the full spaces with zero in the orthogonal complements.*

Proof. 1. The columns of V are an orthonormal basis of Π and thus orthogonal to $\ker X$. *2.* The columns of U are an orthonormal basis of $X(\Pi)$ and $X(\Pi) \perp$

coker X. *3.* $XV = X\big|_{\Pi}V = U\Sigma V^T V = U\Sigma$. *4.* Using *3* we get

$$\begin{aligned} U^T X &= U^T X V V^T + U^T X (I - VV^T) \\ &= \Sigma V^T + U^T X (I - VV^T). \end{aligned}$$

5. The statement follows from $(I - UU^T)XV = (I - UU^T)U\Sigma = 0$, where we have used that $U^T U = I$. *6.* Let $Y \coloneqq X(\Pi)$ and $Z \coloneqq \operatorname{im} X / X(\Pi)$ be the parts of the decomposition $\operatorname{im} X = Y \oplus Z$, which is possible since $Y \subset \operatorname{im} X$. The linear map $(I - UU^T)$ is orthogonal projection onto $X(\Pi)^\perp$ and thus maps $Y \to 0$ and $Z \to Z$. Since $\operatorname{rank} A = \dim(\operatorname{im} A)$, it follows immediately that $\operatorname{rank}(I - UU^T)X = \dim Z$ and that $\operatorname{rank} X = \dim Y + \dim Z = d + \dim Z$. □

Note that $V^T V$ is the orthogonal projection on Π and $U^T U$ is the orthogonal projection on $X(\Pi)$. If Π is spanned by the right singular vectors corresponding to the d largest singular values of X, then $U\Sigma V^T$ is the truncated SVD which by the *Eckhart-Young Theorem* is the closest rank d matrix to X in Frobenius and Spectral norms. Furthermore, if $\Pi = (\ker X)^\perp$, then $d = \operatorname{rank} X$ and $U\Sigma V^T$ is the SVD of X (without expanding U and V to orthonormal matrices). Also note that for these two cases, property *4* takes a simpler form, $U^T X = \Sigma V^T$ (symmetric to property *3*), but the residual $U^T X(I - VV^T)$ is nonzero in general.

Theorem 2.1 concerns the relationship between X and $U\Sigma V^T$ and shows that many important properties that hold for the (truncated) SVD are retained regardless how the subspace Π is chosen. The results from Theorem 2.1 are put into practice in this iterative algorithm. Let $X_1 \coloneqq X$ and repeat the following steps for $k = 1, 2, \ldots$

1. Choose Π_k.
2. Compute $U_k \Sigma_k V_k^T$ from $X_k\big|_{\Pi_k}$.
3. Let $X_{k+1} \coloneqq (I - U_k U_k^T) X_k$.

The iterations can continue as long as X_k is nonzero or until some other stopping

criteria is met. Finally, the results are concatenated:

$$U\Sigma V^T \coloneqq \begin{pmatrix} U_1 & U_2 & \dots & U_n \end{pmatrix} \begin{pmatrix} \Sigma_1 & & & \\ & \Sigma_2 & & \\ & & \ddots & \\ & & & \Sigma_n \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \\ \vdots \\ V_n^T \end{pmatrix}$$
$$= \sum_{k=1}^{n} U_k \Sigma_k V_k^T.$$

Orthogonality between columns within each U_k and V_k respectively follow immediately from the definition. Step 3 above, together with Theorem 2.1, property *2*, guarantees orthogonality between the columns of different U_k's, since the columns of U_k are in $\operatorname{coker} X_l$, for all $l > k$. Similarly, properties *5* and *1* of Theorem 2.1 imply orthogonality between the columns of different V_k's. That is, $U^T U = V^T V = I$. The diagonal entries of each Σ_k are decreasing, but the algorithm above does not ensure any structure between the blocks. In practice however, with each Π_k chosen to capture a strong signal in X_k, we can expect the SMS singular values to be decreasing, or at least close to decreasing.

The rank decreases by d_k in each iteration, that is $\operatorname{rank} X_k = d_k + \operatorname{rank} X_{k+1}$, which follows from property *6* in Theorem 2.1. This implies that $\operatorname{rank} U\Sigma V^T = \operatorname{rank} X$ if the iterations are run all the way until $X_k = 0$. In general, $U\Sigma V^T \neq X$, with equality iff the residual $U_k^T X_k (I - V_k V_k^T) = 0$ for all k. Indeed, if equality holds, then $U^T X - \Sigma V^T = 0$. Step 3 of the algorithm above now implies that $U_k^T X = U_k^T X_k$ and Theorem 2.1, property *4*, yields

$$U^T X - \Sigma V^T = \begin{pmatrix} U_1^T X_1 (I - V_1 V_1^T) \\ U_2^T X_2 (I - V_2 V_2^T) \\ \vdots \\ U_n^T X_n (I - V_n V_n^T) \end{pmatrix}.$$

To adaptively reduce noise, Π must depend on X. Our motivating example is to use Π for selecting a subset of the variables that are likely to be less influenced by noise. This is a special case of choosing Π after performing a linear transform of the variables, which is described in the following theorem:

Theorem 2.2. *Take a linear map $S : \mathbb{R}^L \to \mathbb{R}^P$ and an integer d such that $\operatorname{rank} S^T X \geq d$ and let $\tilde{U}\tilde{\Sigma}\tilde{V}^T$ be the rank d truncated SVD of $S^T X$. Furt-*

hermore let Π *be the subspace spanned by the columns of* $\tilde{V}$ *and let* $U\Sigma V^T$ *be the SVD of* $X\big|_\Pi$. *Then*

1. $\Pi \perp \ker X$.
2. $S^T U\Sigma V^T = \tilde{U}\tilde{\Sigma}\tilde{V}^T$.
3. $\{V_{\cdot 1}, V_{\cdot 2}, \ldots, V_{\cdot d}\}$ *and* $\{\tilde{V}_{\cdot 1}, \tilde{V}_{\cdot 2}, \ldots, \tilde{V}_{\cdot d}\}$ *are orthonormal bases of* Π.
4. $\{S^T U_{\cdot 1}, S^T U_{\cdot 2}, \ldots, S^T U_{\cdot d}\}$ *and* $\{\tilde{U}_{\cdot 1}, \tilde{U}_{\cdot 2}, \ldots, \tilde{U}_{\cdot d}\}$ *are bases of* $S^T X(\Pi)$.
5. $\|\Sigma\|_F \geq \frac{\|\tilde{\Sigma}\|_F}{\|S\|_2}$.
6. $U^T X = \Sigma V^T + U^T(I - SS^T)X(I - VV^T)$.

Proof. 1. The columns of $\tilde{V}$ are orthogonal to $\ker S^T X \supset \ker X$. *2.* $S^T U\Sigma V^T = S^T X\big|_\Pi = (S^T X)\big|_\Pi = \tilde{U}\tilde{\Sigma}\tilde{V}^T$. *3.* Follows immediately from the definitions. *4.* $\{\tilde{U}_{\cdot i}\}_{i=1}^d$ is a basis of $S^T X(\Pi)$. By property *2*, $\tilde{U} = S^T U\Sigma V^T \tilde{V}\tilde{\Sigma}^{-1}$, showing that $\{S^T U_{\cdot i}\}_{i=1}^d$ span $\{\tilde{U}_{\cdot i}\}_{i=1}^d$. Finally, since U and $\tilde{U}$ have the same rank, $\{U_{\cdot i}\}_{i=1}^d$ is also a basis of $S^T X(\Pi)$. *5.* For general matrices A and B, consider A acting on each column of B. We get

$$\|AB\|_F^2 = \sum_i \|AB_{\cdot i}\|_2^2 \leq \sum_i \|A\|_2^2 \|B_{\cdot i}\|_2^2 = \|A\|_2^2 \|B\|_F^2.$$

The result now follows from property *2*, with $A = S^T$ and $B = U\Sigma V^T$, since $\|AB\|_F = \|\tilde{U}\tilde{\Sigma}\tilde{V}^T\|_F = \|\tilde{\Sigma}\|_F$ and $\|B\|_F = \|\Sigma\|_F$. *6.* From Theorem 2.1, property *4*, we get $U^T X = \Sigma V^T + U^T X(I - VV^T)$. It remains to show that $U^T SS^T X(I - VV^T) = 0$. By property *4*, there exists a matrix Z such that $S^T U = \tilde{U}Z$ and

$$\begin{aligned} U^T SS^T X(I - VV^T) &= Z^T \tilde{U}^T S^T X(I - VV^T) \\ &= Z^T \tilde{\Sigma}\tilde{V}^T(I - \tilde{V}\tilde{V}^T) = 0, \end{aligned}$$

where $VV^T = \tilde{V}\tilde{V}^T$ because of property *3*. □

Corollary 2.1. *If* $S^T S = I$, *then* $\|\Sigma\|_F \geq \|\tilde{\Sigma}\|_F$.

Another way to interpret S is that SS^T defines a (possibly degenerate) inner product on the sample space, which is used to find Π. To see this, let $d = \operatorname{rank} S^T X$ so that $\tilde{U}\tilde{\Sigma}\tilde{V}^T = S^T X$ and $K := X^T S S^T X = \tilde{V}\tilde{\Sigma}^2\tilde{V}^T$, showing the well-known result that $\tilde{V}\tilde{\Sigma}^2\tilde{V}^T$ is an eigendecomposition of K, where $K_{ij} = \langle x_i, x_j \rangle := X_{\cdot i}^T S S^T X_{\cdot j}$ is the inner product of sample i and j. This naturally extends to kernel PCA, where K is defined by taking scalar products after an (implicit) mapping to a higher-dimensional space. Any method that results in a low-dimensional sample space representation can indeed be used, since Π is spanned by the columns of V by definition. We will not pursue these extensions here.

The Projection Score[7] provides a natural optimality criterion for S and d (and thus Π) needed in each iteration of the SMSSVD algorithm. It is a measure of how informative a specific variable subset is, when constructing a rank d approximation of a data matrix. A common application is to maximize the Projection Score over a sequence of variable subsets, where each subset consists of those variables that have a variance above a specific threshold. Using the notation from Theorem 2.2, the optimal variable subset describes a matrix S and the optimal low-rank approximation is $\tilde{U}\tilde{\Sigma}\tilde{V}^T$. Here S has exactly one element in each column equal to 1, at most one element in each row equal to 1 and all other elements equal to zero. Hence $S^T X$ corresponds to selecting a subset of the variables of a data matrix X and $S^T S = I$. In iteration k of the SMSSVD algorithm, we optimize the Projection Score jointly over the variance filtering threshold and the dimension, which gives both an optimal variable subset S_k and a simple dimension estimate d_k of the signal that was captured.

3 Results

The performance of SMSSVD is evaluated in comparison to SVD and SPC (Sparse Principal Components), a method similar to SVD, but with an additional lasso (L_1) constraint to achieve sparsity[6]. The methods are evaluated both for real data using three Gene Expression data sets and for synthetic data where the ground truth is known.

3.1 Gene Expression Data

Three Gene Expression data sets, two openly available with microarray data and one based on RNA-Seq available upon request from the original authors, were analyzed. Gene expression microarray profiles from a study of breast cancer[8] was previously used to evaluate SPC[6], but in contrast to their analysis, we use all 118 samples and all 22215 genes. Each sample was labeled as one of five breast cancer subtypes: 'basal-like', 'luminal A', 'luminal B', 'ERBB2', and 'normal breast-like'. In a study of pediatric Acute Lymphoblastic Leukemia (ALL), gene expression profiles were measured for 132 diagnostic samples[9]. The samples were labeled by prognostic leukemia subtype ('TEL-AML1', 'BCR-ABL', 'MLL', 'Hyperdiploid (>50)', 'E2A-PBX1', 'T-ALL' and 'Other'). Our final data set is from another pediatric ALL study, where gene expression profiling was done from RNA-Seq data for 195 samples[10]. The samples were aligned with Tophat2[11] and gene expression levels were normalized by TMM[12]. Only genes with a support of at least 10 reads in at least 2 samples were kept. The annotated subtypes in this data set were 'BCR-ABL1', 'ETV6-RUNX1', 'High hyperdiploid', 'MLL', 'TCF3-PBX1' and 'Other'. Here, 'Other' is a very diverse group containing everything that did not fit in first five categories. We thus present results both with and without this group included.

The ability to extract relevant information from the gene expression data sets was evaluated for each model by how well they could explain the subtypes, using the Akaike Information Criterion (AIC) for model scoring. Given the low-dimensional sample representations from SMSSVD, SVD or SPC (for different values of the sparsity parameter, c), a Gaussian Mixture Model was constructed by fitting one Multivariate Gaussian per subtype. The class priors were chosen proportional to the size of each subtype. The loglikelihood $l := \log \mathsf{P}(\mathbf{x}|\theta, M)$, where $\mathbf{x}$ are the subtype labels, M is the model and θ a vector of k fitted model parameters is used to compute the AIC $= 2k - 2l$. Figure 1 displays the AIC scores for the different models as a function of the model dimension. SMSSVD generally performs better than SVD, by a margin. Comparison with SPC is trickier, since the performance of SPC is determined by the sparsity parameter c and there is no simple objective way to choose c. However, SMSSVD compares well with SPC regardless of the value of the parameter.

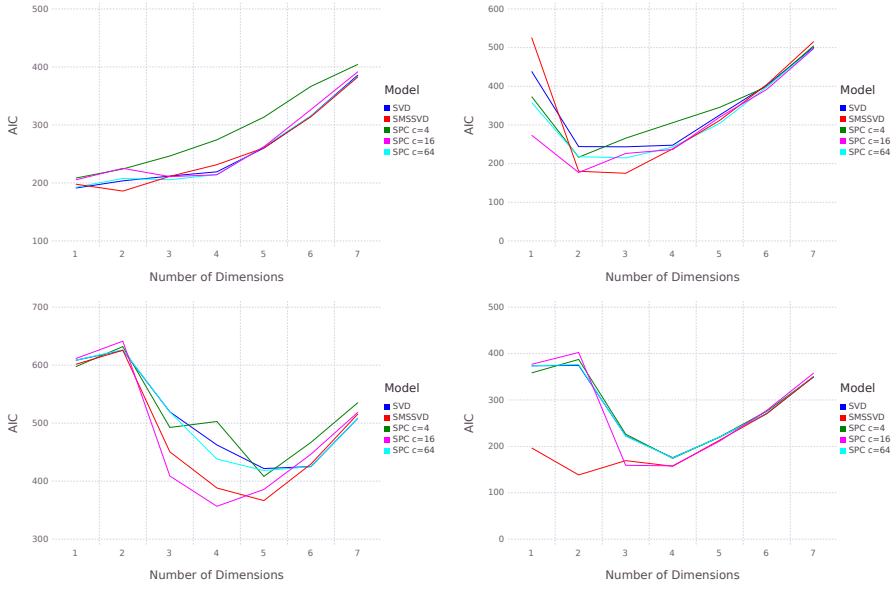


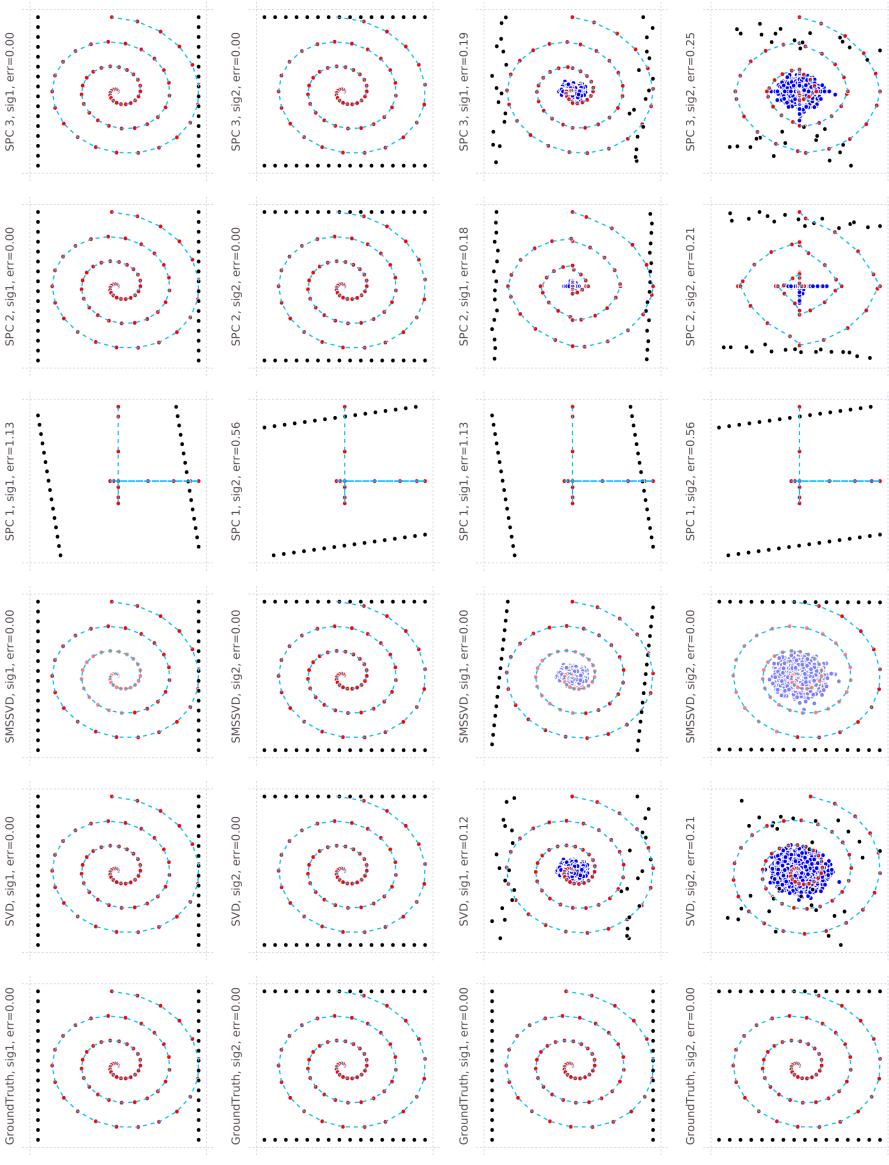
Figure 1: Evaluation of SMSSVD on different data sets, based on AIC scores when fitting a Gaussian Mixture Model to the subtypes. From top to bottom: A. Breast Cancer, B. Acute Lymphoblastic Leukemia (Microarray), C. Acute Lymphoblastic Leukemia (RNA-Seq), D. Acute Lymphoblastic Leukemia (RNA-Seq) with subtype 'Other' removed.

3.2 Synthetic Data

SMSSVD decomposes a matrix observed in noisy conditions as a series of orthogonal low-rank signals. The aim is to get a stable representation of the samples and then recover as much as possible of the variables, even for signals that are heavily corrupted by noise. To evaluate SMSSVD, we synthetically create a series of low-rank signals Y_k that are orthogonal (i.e. $Y_i^T Y_j = 0$ and $Y_i Y_j^T = 0$ for $i \neq j$) and that has a chosen level of sparsity on the variable side and try to recover the individual Y_k's from the observed matrix $X := \sum_k Y_k + \varepsilon$ where ε is a matrix and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij})$. To measure how well SMSSVD recovers the signals from the data, we look at each signal separately, considering only variables where the signal has support. Let $\text{err}(k)$ be the reconstruction error of signal k,

$$\text{err}(k) := \|R_k^T (Y_k - \hat{Y}_k)\|_F,$$

where $\hat{Y}_k$ is the reconstructed signal and R_k is defined such that multiplying with R_k^T from the left selects the variables (rows) where Y_k is nonzero.


SPC 3, sig1, err=0.00
SPC 3, sig2, err=0.00
SPC 3, sig1, err=0.19
SPC 3, sig2, err=0.25
SPC 2, sig1, err=0.00
SPC 2, sig2, err=0.00
SPC 2, sig1, err=0.18
SPC 2, sig2, err=0.21
SPC 1, sig1, err=1.13
SPC 1, sig2, err=0.56
SPC 1, sig1, err=1.13
SPC 1, sig2, err=0.56
SMSSVD, sig1, err=0.00
SMSSVD, sig2, err=0.00
SMSSVD, sig1, err=0.00
SMSSVD, sig2, err=0.00
SVD, sig1, err=0.00
SVD, sig2, err=0.00
SVD, sig1, err=0.12
SVD, sig2, err=0.21
GroundTruth, sig1, err=0.00
GroundTruth, sig2, err=0.00
GroundTruth, sig1, err=0.00
GroundTruth, sig2, err=0.00

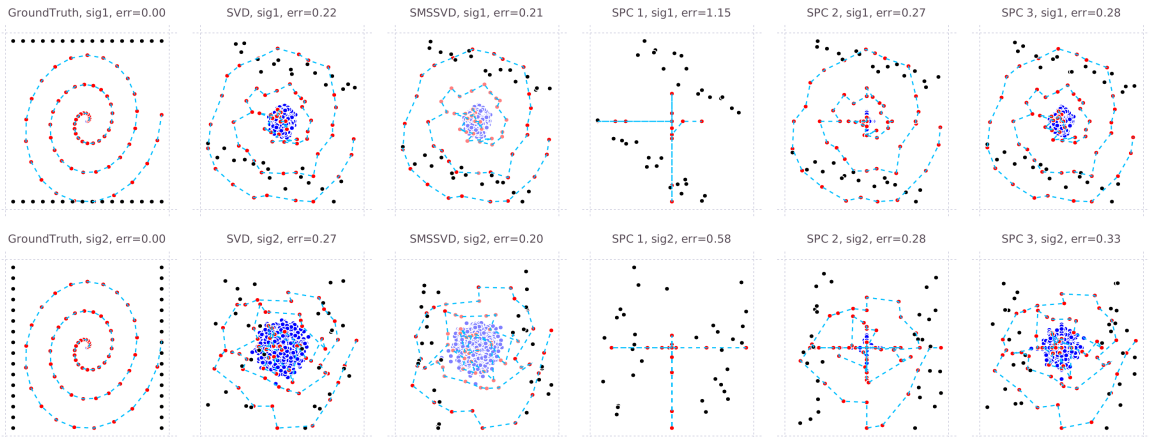


Figure 2: Two $2d$ signals with non-overlapping support for the variables are shown for no noise (Upper two rows), noise on non-signal variables only (Middle two rows) and for noise on all variables (Lower two rows). The reconstruction of the first signal is shown in the upper row and for the second signal in the lower row in each set. Different columns correspond to different methods, where SPC "1", "2" and "3" have regularization penalties of $c = 2$, 8 and 32 respectively, controlling the degree of sparsity. Samples are black, variables where the signal has support are red and other variables are blue. The variables in the support are connected with dashed lines, only to make it easier to spot how the variables are influenced by noise. For SMSSVD, variables selected by optimal variance filtering are shown in full color and other variables are shown in a whiter tone. Samples and variables are both scaled to fill the axes in each biplot. 32 samples and 5000 variables were used, of which each signal had support in 64 variables and the rest had noise only.

While SMSSVD is designed to find d-dimensional signals ($\hat{Y}_k \coloneqq U_k \Sigma_k V_k^T$), the same is not true for SVD and SPC. To test the ability to find the signals, rather than the ability to find them in the right order, the components are reordered using a algorithm that tries to minimize the total error by greedily matching the rank 1 matrices from the decomposition to signals Y_k, always picking the match that lowers the total error the most. The number of rank 1 matrices matched to each signal Y_k is equal to $\operatorname{rank} Y_k$. Note that with no noise present, SVD is guaranteed to always find the optimal decomposition.

The biplots in Figure 2 illustrate how SMSSVD works and how the signal reconstructions compares to other methods. If there is no noise, perfect decompositions are achieved by all methods apart from SPC with a high degree of sparsity. An artificial example where the noise is only added to the non-signal variables highlights that SMSSVD can still perfectly reconstruct both samples and signal variables, whereas the other methods display significant defects. Finally, when all variables are affected by noise, SMSSVD still get the best results.

Next, we created several data sets for a variety of conditions based on the parameters $N = 100$: Number of samples, P: Number of variables, L: Number of variables in the support of each signal, $K = 8$: number of signals and d: the rank of each signal. For each signal, we randomize matrices U_k and V_k, choose a diagonal matrix Σ_k and let $Y_k \coloneqq U_k \Sigma_k V_k^T$. For both V_k and U_k, each new column is created by sampling a vector of i.i.d. Gaussian random variables and projecting onto the orthogonal complement of the subspace spanned by previous columns (in current and previous signals). For U_k, we only consider the subspace spanned by L randomly selected variables. The result is then expanded by inserting zeros for the other $P - L$ variables. To complete the signal, let the i'th diagonal element of Σ_k, $(\Sigma_k)_{ii} \coloneqq 0.6^{k-1} 0.9^{i-1}$, such that there is a decline in the power between signals and within components of each signal. Finally, i.i.d. Gaussian noise is added to the data matrix. Figures 3, 4 and 5 show test results for data sets randomized in this way for different sets of parameters. SMSSVD is the only method that performs well over the whole set of parameters. The only situation where SMSSVD is consistently outperformed is by SVD for large L, and it is by a narrow margin. SMSSVD performs particularly well, in comparison to the other methods, in the difficult cases when the signal to noise ratio is low. SPC performance clearly depends on the regularization parameter which must be chosen differently in different situations. However, despite being a parameter-free method, SMSSVD outperforms SPC in most cases.

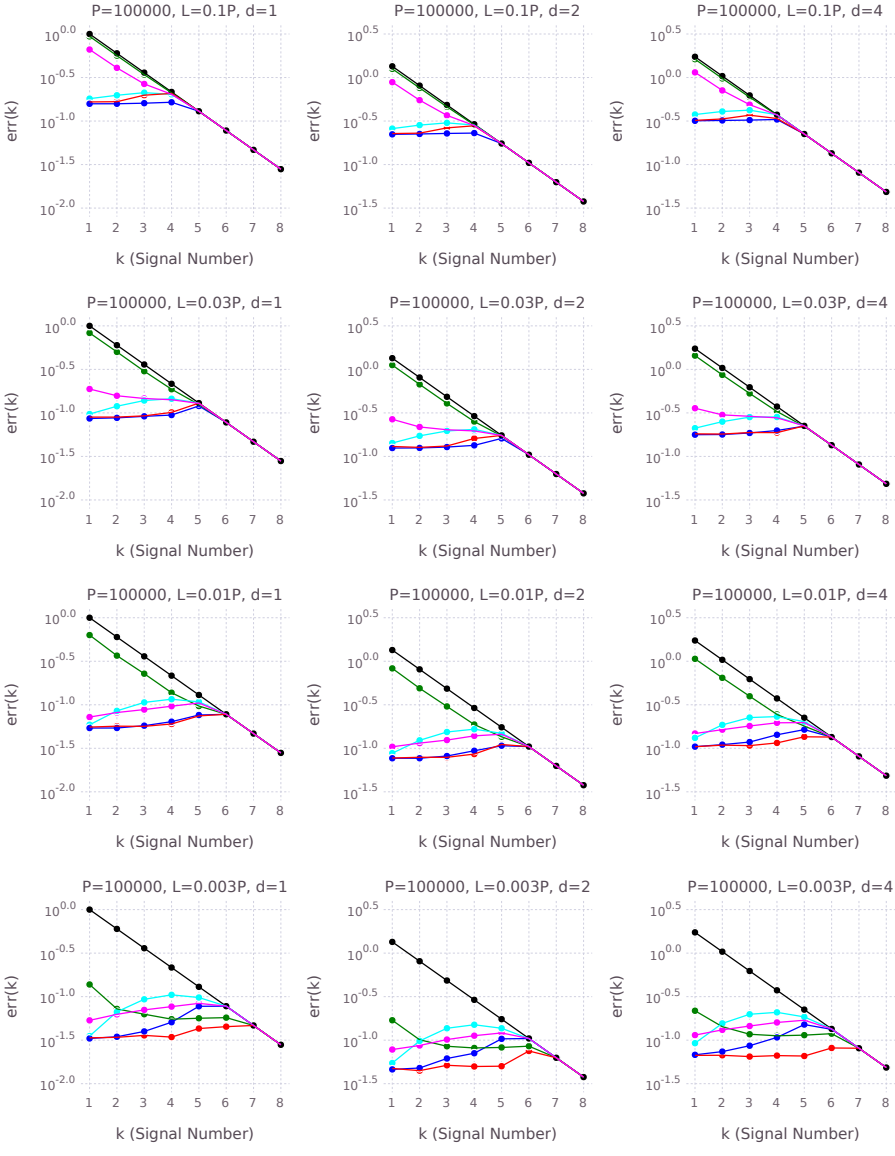


Figure 3: The reconstruction error, $\mathrm{err}(k)$, is shown for different conditions. The signal strength $\|Y_k\|_F$ (black) is shown for scale. The methods are: SVD (blue), SMSSVD (red) and SPC (green, magenta, cyan) with decreasing degree of sparsity (regularization parameters $c = 0.04\sqrt{P}$, $c = 0.12\sqrt{P}$ and $c = 0.36\sqrt{P}$ respectively). No errors larger than the signal strength are displayed as that indicates that a different signal has been found.

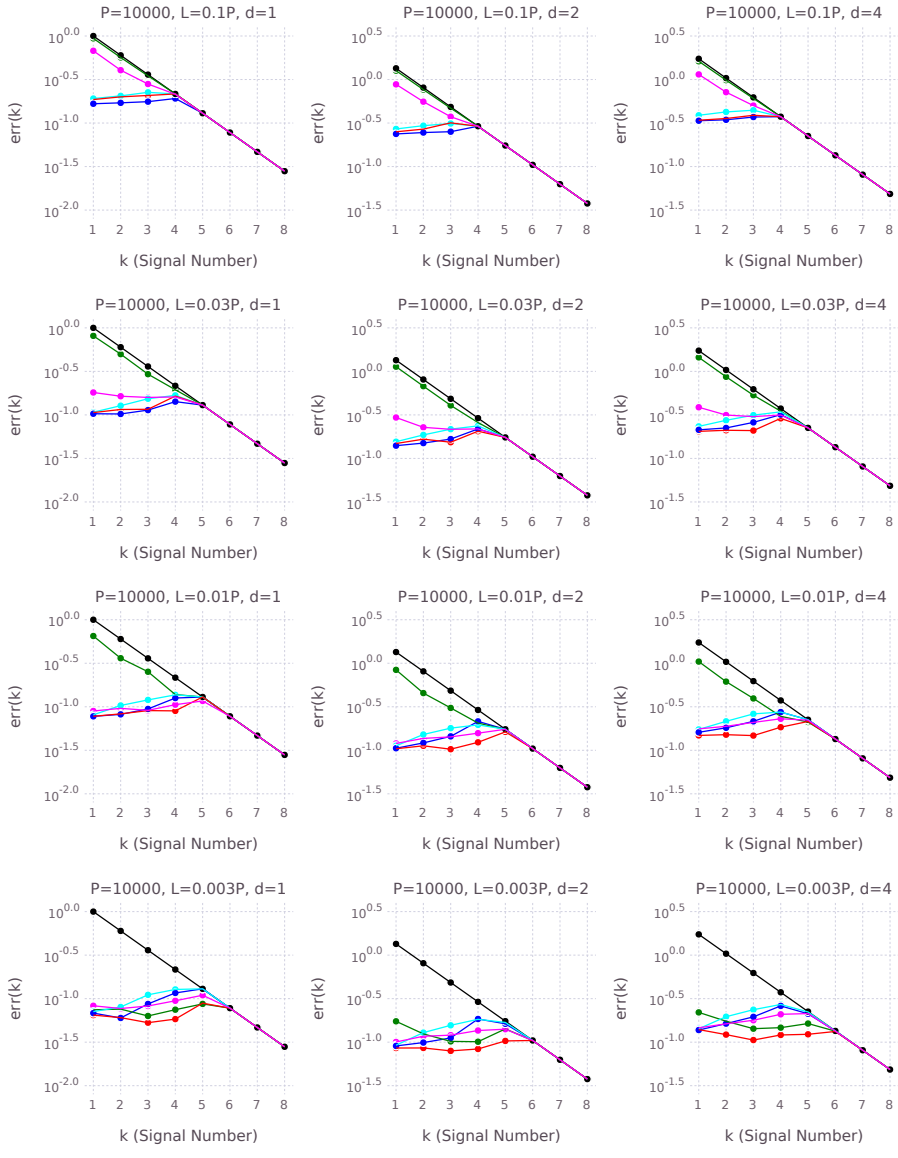


Figure 4: The reconstruction error, $\text{err}(k)$, is shown for different conditions. The signal strength $\|Y_k\|_F$ (black) is shown for scale. The methods are: SVD (blue), SMSSVD (red) and SPC (green, magenta, cyan) with decreasing degree of sparsity (regularization parameters $c = 0.04\sqrt{P}$, $c = 0.12\sqrt{P}$ and $c = 0.36\sqrt{P}$ respectively). No errors larger than the signal strength are displayed as that indicates that a different signal has been found.

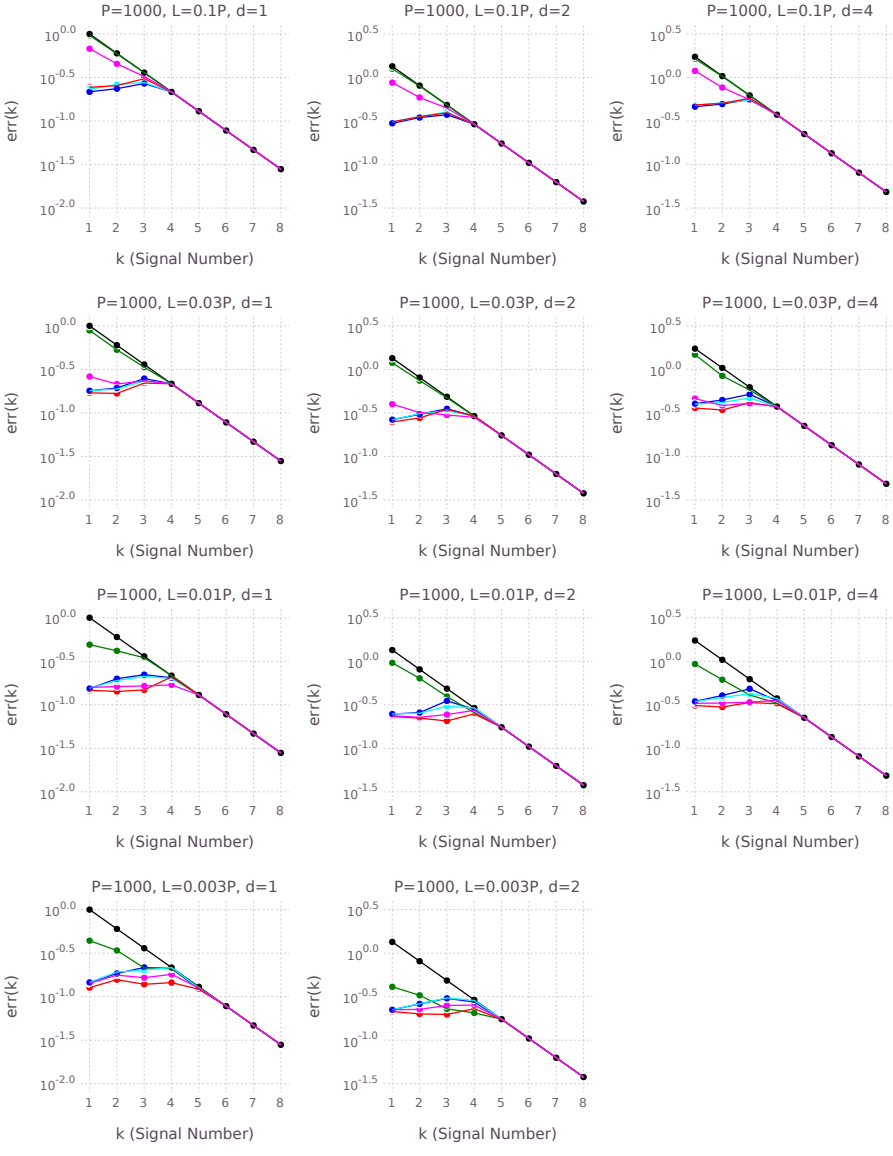


Figure 5: The reconstruction error, $\mathrm{err}(k)$, is shown for different conditions. The signal strength $\|Y_k\|_F$ (black) is shown for scale. The methods are: SVD (blue), SMSSVD (red) and SPC (green, magenta, cyan) with decreasing degree of sparsity (regularization parameters $c = 0.04\sqrt{P}$, $c = 0.12\sqrt{P}$ and $c = 0.36\sqrt{P}$ respectively). No errors larger than the signal strength are displayed as that indicates that a different signal has been found.

4 Discussion

We have presented SMSSVD, a dimension reduction technique designed for complex data sets with multiple overlaid signals observed in noisy conditions. When compared to other methods, over a wide range of conditions, SMSSVD performs equally well or better. SMSSVD excels in situations where $P \gg N$ (many more variables than samples) but most of the variables just contribute with noise, a very common situation for high throughput biological data. As a parameter-free method, SMSSVD requires no assumptions to be made of the level of sparsity. Indeed, SMSSVD can handle different signals within the same data set that exhibit very different levels of sparsity. Being parameter-free also makes SMSSVD suitable for automated pipelines, where few assumptions can be made about the data.

A common strategy when analyzing high dimensional data is to first apply PCA (SVD) to reduce the dimension to an intermediate number, high enough to give an accurate representation of the data set, but low enough to get rid of some noise and to speed up downstream computations (see e.g. van der Maaten et al.[13]). We argue that since SMSSVD can recover multiple overlaid signals and adaptively reduce the noise affecting each signal so that even signals with a lower signal to noise ratio can be found, it is very useful in this situation.

Our unique contribution is that we first solve a more suitable dimension reduction problem for robustly finding signals in a data set corrupted by noise and then map the result back to the original variables. We also show how this combination of steps gives SMSSVD many desirable properties, related to the SVD of both the full data matrix and of the smaller matrix from the variable selection step. Orthogonality between components is one of the cornerstones of SVD, but it is often difficult to satisfy the orthogonality conditions when other factors are taken into account. SPC does for instance give orthogonality for samples, but not for variables and the average genes of each subset in gene shaving are 'reasonable' uncorrelated. For SMSSVD, orthogonality follows immediately from the construction, simplifying interpretation and subsequent analysis steps. Theorem 2.2, property *2* highlights that the variables retained in the variable selection step are unaffected when the solution is expanded to the full set of variables. Hence, we can naturally view each signal from the point of view of the selected variables, or using all variables.

The variable selection step in the SMSSVD algorithm can be chosen freely. For exploratory analysis, optimizing the Projection Score based on variance filtering

is a natural and unbiased choice. Another option is to use Projection Score for response related filtering, e.g. ranking the variables by the absolute value of the t-statistic when performing a t-test between two groups of samples. The algorithm also has verbatim support for variable weighting, by choosing the S matrix as a diagonal matrix with a weight for each variable. Clearly this is a generalization of variable selection.

Kernel PCA, SPC, and other methods that give low-dimensional sample representations, but where the variable information is (partially) lost, can also be extended by SMSSVD (relying on Theorem 2.1 only), as long as a linear representation in the original variables can be considered meaningful. Apart from retrieving a variable-side representation, the SMSSVD algorithm also makes it possible to find multiple overlapping signals, by applying the dimension reduction method of interest as the first step of each SMSSVD iteration.

Acknowledgements

The authors would like to thank Thoas Fioretos and Henrik Lilljebjörn at the Department of Laboratory Medicine, Division of Clinical Genetics, Lund University, for giving us access to the RNA-seq data presented in Lilljebjörn et al. [10]

Bibliography

[1] Trevor Hastie, Robert Tibshirani, Michael B Eisen, Ash Alizadeh, Ronald Levy, Louis Staudt, Wing C Chan, David Botstein, and Patrick Brown. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1(2):research0003–1, 2000.

[2] Trevor Hastie, Robert Tibshirani, David Botstein, and Patrick Brown. Supervised harvesting of expression trees. *Genome Biology*, 2(1):research0003–1, 2001.

[3] Eric Bair and Robert Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, 2(4):e108, 2004.

[4] Lei Ding and Daniel J McDonald. Predicting phenotypes from microarrays using amplified, initially marginal, eigenvector regression. *Bioinformatics*, 33 (14):i350–i358, 2017.

[5] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

[6] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

[7] Magnus Fontes and Charlotte Soneson. The projection score – an evaluation criterion for variable subset selection in PCA visualization. *BMC Bioinformatics*, 12(1):307, 2011.

[8] Koei Chin, Sandy DeVries, Jane Fridlyand, Paul T Spellman, Ritu Roydasgupta, Wen-Lin Kuo, Anna Lapuk, Richard M Neve, Zuwei Qian, Tom Ryder, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541, 2006.

[9] Mary E Ross, Xiaodong Zhou, Guangchun Song, Sheila A Shurtleff, Kevin Girtman, W Kent Williams, Hsi-Che Liu, Rami Mahfouz, Susana C Raimondi, Noel Lenny, et al. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, 102(8):2951–2959, 2003.

[10] Henrik Lilljebjörn, Rasmus Henningsson, Axel Hyrenius-Wittsten, Linda Olsson, Christina Orsmark-Pietras, Sofia Von Palffy, Maria Askmyr, Marianne Rissler, Martin Schrappe, Gunnar Cario, et al. Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nature Communications*, 7, 2016.

[11] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14 (4):R36, 2013.

[12] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

[13] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

PAPER IV

In submission

DISSEQT – DIStribution based modeling of SEQuence Space Time dynamics

Rasmus Henningsson, Gonzalo Moratorio, Antonio V. Bordería, Marco Vignuzzi and Magnus Fontes

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. DISSEQT is openly available on GitHub* and Synapse†, covering the entire workflow from read alignment to visualization of results. DISSEQT 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 on of the highest degrees of population heterogeneity found in nature, making it natural to attempt a distribution-based model. Using DISSEQT, we empirically reconstruct the evolutionary trajectories of evolving populations in

*https://github.com/rasmushenningsson/DISSEQT.jl

†http://dx.doi.org/10.7303/syn11425758

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 and we are able to recover robust characteristics of the population distribution driving evolution. Faithfully monitoring the evolutionary trajectories, we can identify the key minority genotypes contributing most to the population characteristics. Finally, we show that empirical fitness landscapes, when reconstructed to include minority variants, can predict phenotype from genotype with high accuracy.

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 is through selection of variants from this pool of mutants, in response to environmental change. With the advent of DNA sequencing, viruses and bacteria were the first organisms to be fully sequenced (phage MS2 in 1975[1]; H.influenzae[2] and M.genitalium[3] in 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[4,5]. Even before HTS, phenotypic differences between populations with the same consensus sequence have been observed and attributed to suspected differences in variant composition. 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[6], fitness and phenotype[4,7], but focused on only a few variants. Since a mixed population can constitute an evolutionary stable strategy (ESS)[8,9], 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 [10–14]. 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 [15] circumvents these issues by applying multidimensional scaling (MDS) for exploratory analysis, keeping distances between samples more in line with the measured quantities. PhyloMap [16] 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 [17] and the NK landscapes [18] 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 [19,20], thus artificially enforcing a low dimension of sequence space. To expand the fitness landscape analysis to a higher dimensional setting, Kouyos et al. [21] 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 [22]. Seifert et al. [23] 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 and its variants composition may foretell whether and where the population will move in genotypic space [24], by looking at the dynamics of the population rather than a more static analysis at consensus level. Understanding how the population is developing in sequence space may help predict which directions it can go from there. Viruses, the fastest mutators with small genomes, make ideal model

organisms for studying short-term evolutionary processes 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 a distribution-based model designed to capture the heterogeneity of the populations which makes 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 and bacterial 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 microbial 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 (Figure 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 the Methods Section. 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 were aligned iteratively until the consensus

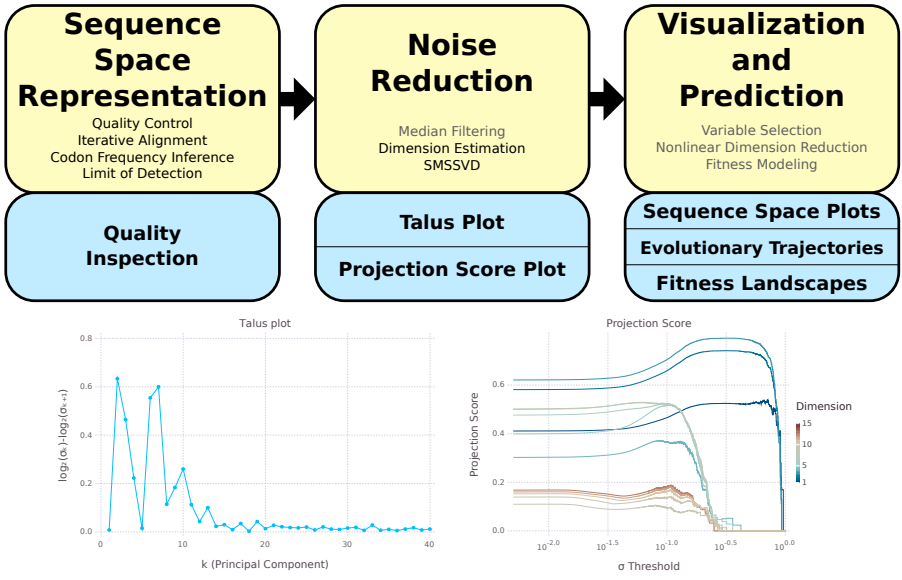


Figure 1: Top: The DISSEQT pipeline. The yellow boxes represent algorithms and data management. The blue boxes represent plots and other output. The analysis history of all results and plots can be traced back all the way to the raw input data. Steps that are only used in some analyses are displayed 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 13 dimensions, the Talus plot shows small variations around a low mean. Bottom right: **Projection Score Plot** for the SynSyn data set. SMSSVD finds 3 signals of dimensions 3, 5 and 5, with different optima for variance filtering. Each curve displays the projection score of the current signal as a function of the variance filtering threshold, as dimensions are progressively added to the model.

sequence converged and both automatic and manual quality controls were performed. 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. 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. 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 (Figure 1, bottom left panel, and Supplementary Methods), after which noise reduction was performed by SubMatrix Selection Singular Value Decomposition (SMSSVD)[25]. 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 variants arising and driving the movement in sequence space. Empirical fitness values were used to create fitness landscapes for prediction, using the representation 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 swarms of variants in clinical or experimental samples – essentially, to monitor and analyze evolving populations before significant changes to consensus sequences occur. To do so, we generated a collection of samples that would be representative of such populations, 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 (Figure 2), some of which were previously published[26], in which these

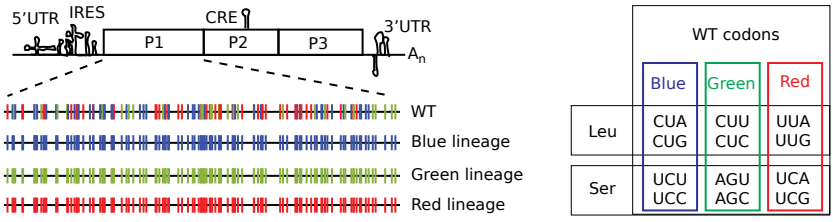


Figure 2: Right: Clusters of Leu/Ser codons according to different viral lineages. Color coding corresponds to synonymous codons used to genetic engineered each viral lineage 'Blue Lineage' (blue), 'Green Lineage' (green) or 'Red Lineage' (red). Left: Schematic of the Coxsackie virus genome indicating 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.

117 codons were changed to belong exclusively to only one of three codon categories. These lineages were designed to retain the initial functional neutrality (that is, 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 Extended Figure 1). 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 their consensus sequences, 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[27] three base analogues (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 microbial populations is of intrinsically low dimensionality

Theoretical sequence space is incredibly large, even for a small genome of length $n = 10000$, the number of possible sequences are $4^n \approx 4 \cdot 10^{6020}$, 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 much of sequence space is occupied by an evolving microbial population. We generated Talus and Projection Score[28] 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 Figure 1, bottom left panel, the Talus plot settles after 13 dimensions, with small variations around a low mean, giving a dimension estimate of 13. In the Projection Score plot (Figure 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 analysis of minority variants could better monitor the 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 Figure 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 Figure 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 Figure 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 (Figure 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, barely any biologically relevant signals are detected and no patterns related to the mutagens are found (Extended Figure 2). These results reveal an important feature of evolving microbial swarms: despite sequence space being of theoretically ultra-high dimension, we showed here that evolving microbial populations such as RNA viruses, which present the highest mutation frequencies, are of intrinsically low dimensionality. Indeed, all five of the biological pressures placed on these viral populations could be captured within the first 13 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)[4]. In that study, we identified one consensus sequence change that occurred in all lineages during the first 10 passages, followed by a cluster of minority variants that reached above 5% 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 (Figure 4), without biasing towards non-synonymous 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 repli-

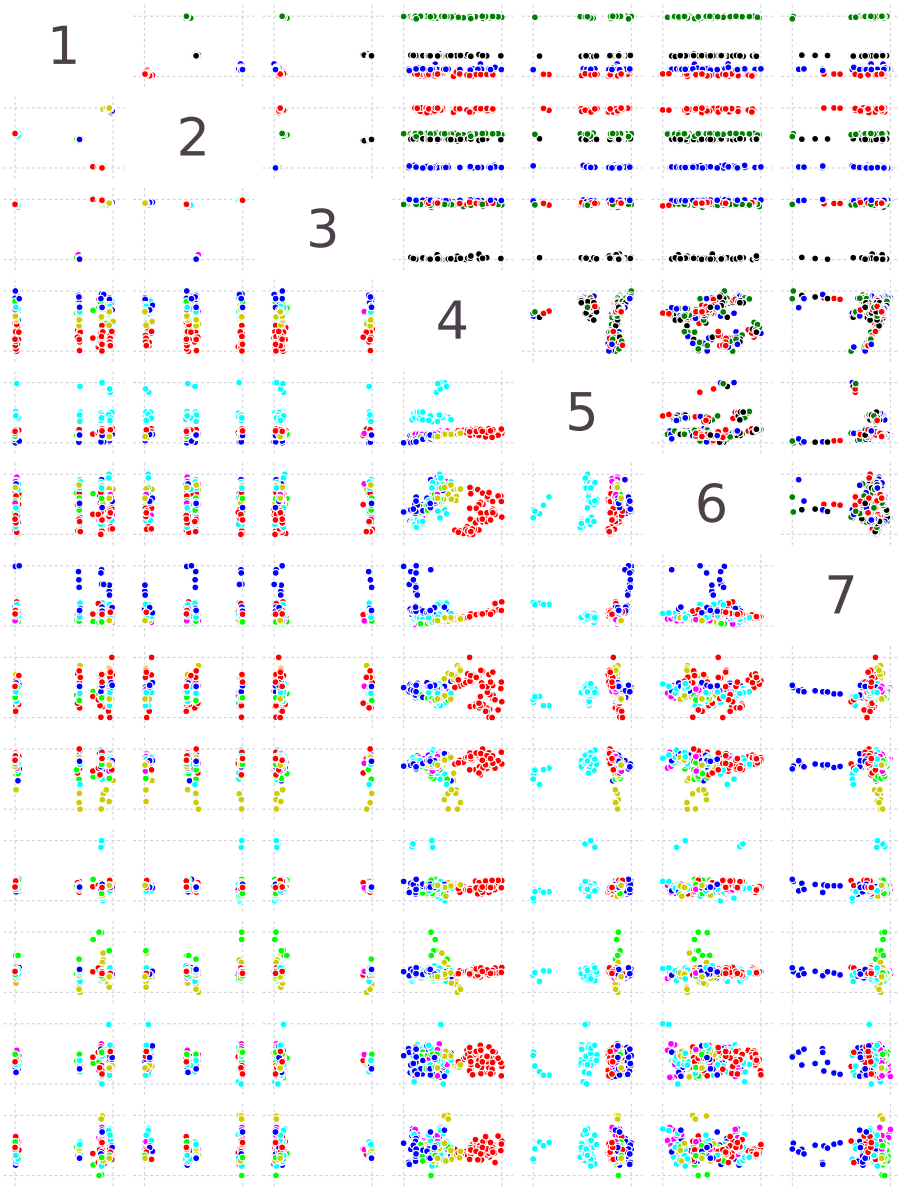


Figure 3: Pairwise scatter plots showing the first 13 principal components in the analysis of the SynSyn data set plotted against each other. Plots above and below the diagonal are mirror images of each other. Each dot represents one viral population. (Continued on the following page.)

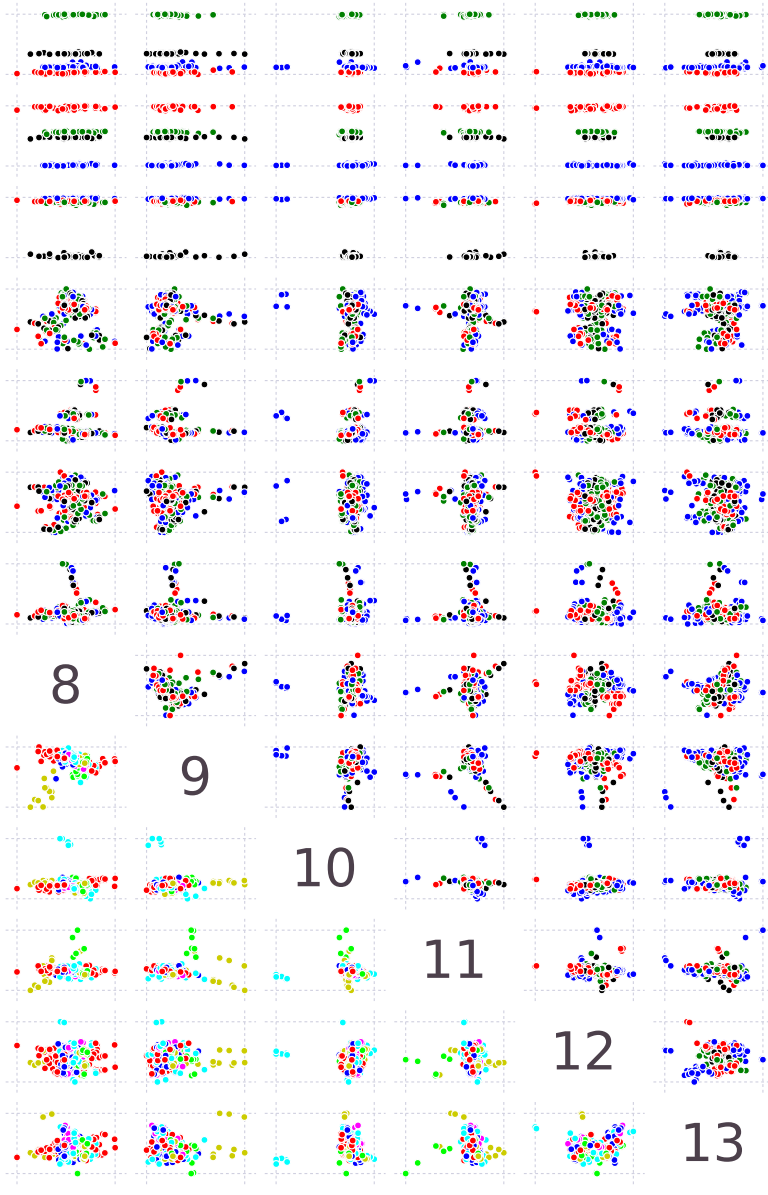


Figure 3: Above the diagonal, samples are colored by lineage (Black: 1, Blue: 2, Green: 3, Red: 4) and below the diagonal, samples are colored by mutagen (Red: 5-fluorouracil, Light green: amiloride, Blue: 5-azacytidine, Yellow: Mn^{2+}, Cyan: ribavirin, Magenta: mock). All axes are rescaled to fill the plot area.

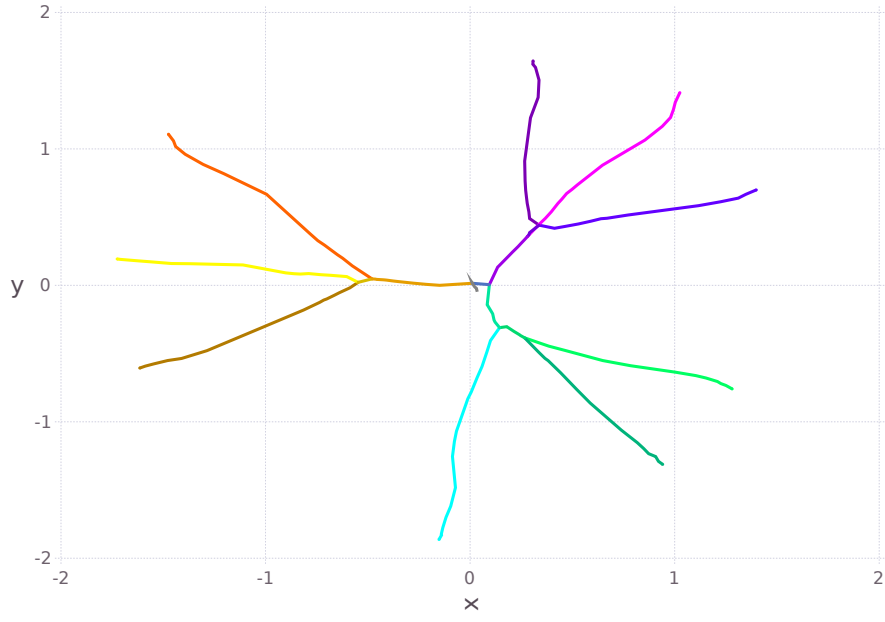


Figure 4: Above: Overview of the evolutionary trajectories of the 9 replicates in the adaptability data set[4], shown after nonlinear dimension reduction. WT replicates are shown in magenta-purple colors, replicates from the high fidelity lineage in green-cyan colors and replicates from the low fidelity 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. (Continued on the following page.)

cates (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 wild-type, and to high fidelity strains. The replicates within each lineage then followed similar trajectories until further bifurcating between passages 7–19. 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.

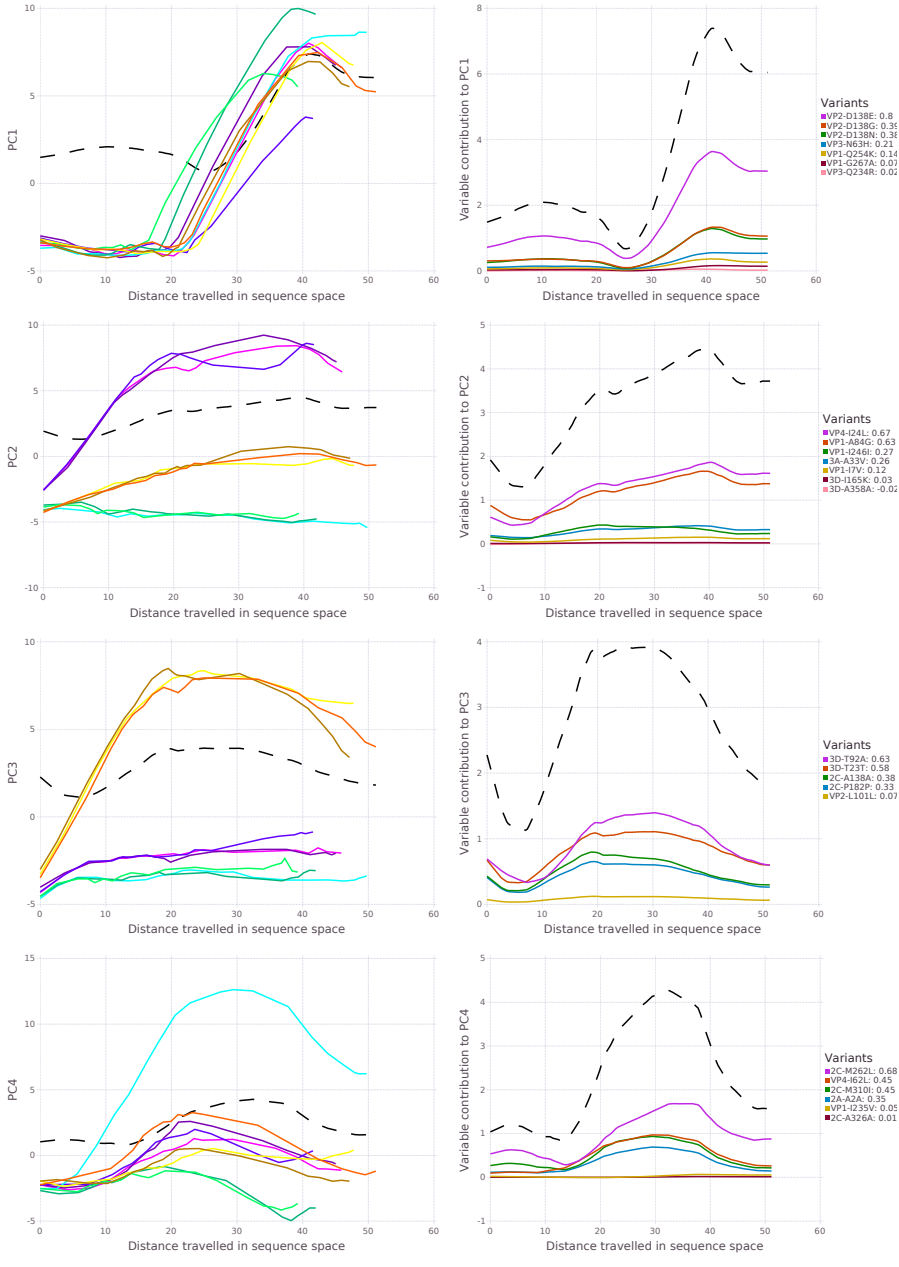


Figure 4: (Continued from the previous page.)

While the above analysis gave information on rate and direction of evolution, it did not identify what minority variant component of each population was responsible for adaptation 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 maps the trajectories of each replicate (left panels), and identifies which variants contributed most to the signal in each component (right panels). 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[4]. The remaining components, however, identified several other mutations at sites that were missed by using the classic cut-off of 1–2% minority variant frequency and that could explain differences 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 nonstructural proteins such as the 2C (helicase) and 3C (protease) (Extended Figure 3). 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 virus populations described above in a direct competition assay against a neutral,

genetically marked competitor[26,29] (data available in Synapse). The visualization (Figure 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 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 Figure 6, top panel). The figure shows the dynamics of viral clouds 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 with-out 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[26]. Extended Figure 4 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 (Figure 5, bottom panel).

We thus evaluated the relevance of our sequence space reconstructions (after

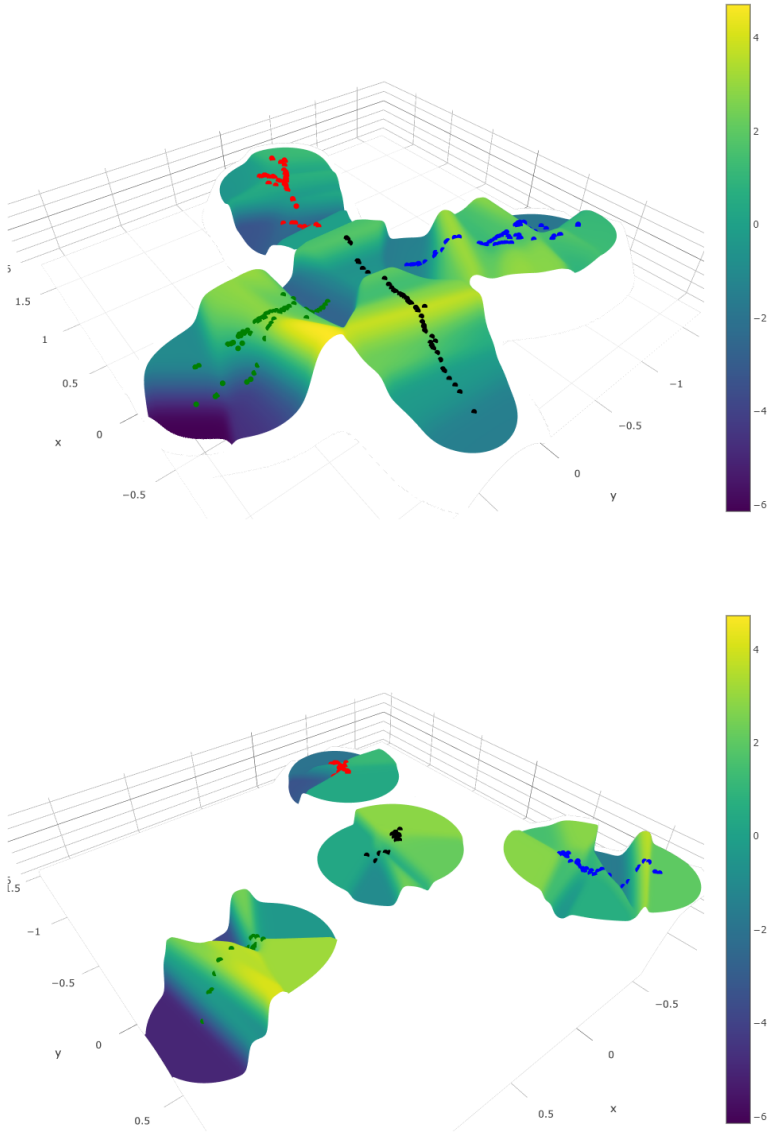


Figure 5: Top: Fitness landscape visualization of the SynSyn data set. Bottom: The same fitness landscape, constructed from consensus data only. Samples are colored by lineage (Black: 1, Blue: 2, Green: 3, Red: 4).

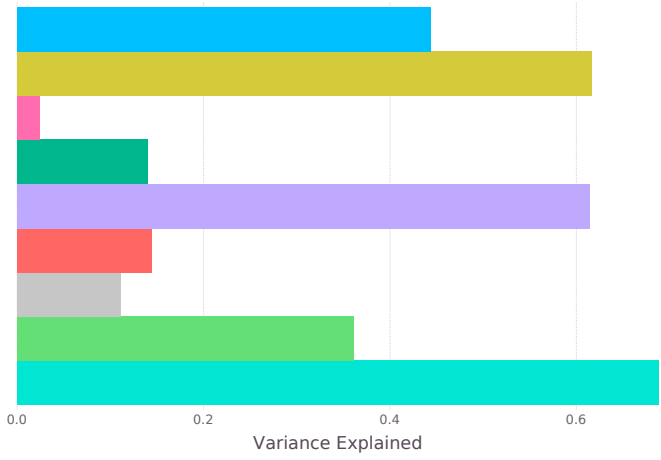


Figure 6: Comparison between different fitness predictors. Gaussian Kernel Smoother Predictors: Isomap 2d (Blue), SMSSVD 13d (Yellow), Consensus Isomap 2d (Pink) and Consensus 13d (Green). Nearest Neighbor Predictors: SMSSVD 13d (Purple) and Consensus (Red). Group Predictors: Lineage/Mutagen (Gray), Lineage/Dose (Light green), Lineage/Mutagen/Dose (Turquoise).

noise reduction) in their ability predict virus fitness, a quantitative parameter often used to describe phenotype. The performance of different fitness models was compared (Figure 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 (Figure 6, turquoise bar). In other words, when these three groupings are known for a sample, the prediction is over 69% accurate. When only lineage and dose were considered, prediction was 36% accurate, and if only lineage and mutagen were known, accuracy dropped to 11%. For the landscape predictors based on the 2d Isomap, accuracy was 44%. SMSSVD, on the other hand, which uses 13d reaches predictability of 62% and 61% 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 PCA is performed solely on consensus sequences, prediction fails (2% and 14%, respectively). Furthermore, the performance of the SMSSVD predictors compared well to the predictor ba-

sed 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 confirm that consensus sequencing of a viral population is not enough to understand its properties and cannot accurately predict its phenotype.

3 Discussion

High-throughput sequencing 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 culture, Acevedo et al.[5] identified an average of 16,500 variants, the equivalent of $\sim 74\%$ 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% of the sequenced regions[4].

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 tend to emerge. This suggests that the sequence space available to a virus is indeed more limited, determined by its current genome sequence, 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 microbial 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 quasispecies cloud that drive evolution. This accentuates that global properties, like the shape of the quasi-species cloud, are significant when trying to predict viral evolution. Sequencing error has long been an issue with characterizing microbial diversity and identifying true SNVs. 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.

We have shown that by directly modeling and representing the distribution at each genetic loci 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 [13] 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 [30]. Larger genomes and frequent structural variation, such as chromosomal aberrations [31] and fused genes [32] 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 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.

4.2 Iterative Alignment

Alignment of sequenced reads to a reference genome was done with BWA-MEM[33]. The choice of alignment tool is not critical, but the same one should be used for all samples 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 has converged. 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 variants at the majority 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 majority level variants is greatly improved, as the number of differences between reads containing the minority variant and the consensus will tend to be lower.

4.3 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[34]. 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 used for the initial alignment iteration, than to any other reference 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 in the same sequencing run (and using the same reference genome) are put in the same 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 1000 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.

4.4 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[35,36] are available, also see[37,38] 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[39], V-Phaser 2[40] and ShoRAH[35] 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 does 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.

4.5 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 doesn't 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 ordered 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 position can be estimated separately. The drawback is that minority variant linkage is lost. 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.

4.6 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 $\boldsymbol{\theta} = \left(\theta_1, \theta_2, \ldots, \theta_{64}\right)$ be the frequencies in the population for the 64 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

$$\begin{aligned}\mathsf{P}(x|\boldsymbol{\theta}) &= \sum_{z=1}^{64} \mathsf{P}(x|z,\boldsymbol{\theta})\mathsf{P}(z|\boldsymbol{\theta}) \\ &= \sum_{z=1}^{64} \mathsf{P}(x|z)\theta_z.\end{aligned}$$

We model $\mathsf{P}(x|z)$ 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

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

where δ_a^b is the Kroenecker 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 3 bases. Assuming independent reads, the probability of the observations is

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

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

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

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).

4.7 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[41]. 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 leads 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

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

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

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

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 $1, 2, \ldots, G$,

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

with upper indices denoting the sample group and where 10^{-3} is a commonly accepted lower limit of detection for sequencing data[42].

4.8 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 Methods, 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.

4.9 SMSSVD

SubMatrix Selection Singular Value Decomposition (SMSSVD)[25] 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.

4.10 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 $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$, $i = 1, \ldots, N$ with corresponding fitness values $y^{(i)} \in \mathbb{R}$, we want to reconstruct a surface $f : \mathbb{R}^d \to \mathbb{R}$ such that

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

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[43] and select the kernel width σ by cross-validation (repeated random subsampling). That is, we numerically find

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

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

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

4.11 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,

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

where $\boldsymbol{x}^{(i)}$ is the representation of sample i used by the predictor, $y^{(i)}$ is the fitness of sample i, $\bar{y}$ 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.

4.12 Variable Selection

We use SPC (Sparse Principal Components)[44] 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 a 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.

4.13 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[45]. However, to get an informative visualization in just two or three dimensions, nonlinear dimension reduction is needed for complex data sets.

We apply Isomap[46] to the data set after the noise reduction by SMSSVD (and the optional variable selection). Nonmetric multidimensional scaling using Kruskal's stress criterion[47] was used rather than classical multidimensional scaling 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.

4.14 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,

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

where $\sigma_{ij} \coloneqq \boldsymbol{u}_i X_{ij} \boldsymbol{v}_j$ 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 $\sigma_j(s) \coloneqq \sum_r \sigma_j^{(r)}(s)$ and $\sigma(s) \coloneqq \sum_j \sigma_j(s)$ 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.

4.15 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

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

that is, the first point in time at which the distance between $\boldsymbol{p}(t)$ and $\boldsymbol{q}$ 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 $T_{ij} \coloneqq \min(T_i, T_j)$, then the mean distance over time between curves i and j is

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

and we let $m \coloneqq \frac{1}{N(N-1)} \sum_{i \neq j} m_{ij}$, 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 $\beta(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(j)})$ can be computed analytically.

Bibliography

[1] Walter Fiers, Roland Contreras, Fred Duerinck, Guy Haegeman, Dirk Iserentant, Jozef Merregaert, W Min Jou, Francis Molemans, Alex Raeymaekers, A Van den Berghe, et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507, 1976.

[2] Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, Joseph M Merrick, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, pages 496–512, 1995.

[3] Claire M Fraser, Jeannine D Gocayne, Owen White, Mark D Adams, Rebecca A Clayton, Robert D Fleischmann, Carol J Bult, Anthony R Kerlavage, Granger Sutton, Jenny M Kelley, et al. The minimal gene complement of Mycoplasma genitalium. *Science*, pages 397–403, 1995.

[4] Antonio V Bordería, Ofer Isakov, Gonzalo Moratorio, Rasmus Henningsson, Sonia Agüera-González, Lindsey Organtini, Nina F Gnädig, Hervé Blanc, Andrés Alcover, Susan Hafenstein, et al. Group Selection and Contribution of Minority Variants during Virus Adaptation Determines Virus Fitness and Phenotype. *PLoS Pathogens*, 2015.

[5] Ashley Acevedo, Leonid Brodsky, and Raul Andino. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature*, 505(7485):686, 2014.

[6] Marco Vignuzzi, Jeffrey K Stone, Jamie J Arnold, Craig E Cameron, and Raul Andino. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, 439(7074):344–348, 2006.

[7] Katherine S Xue, Kathryn A Hooper, Anja R Ollodart, Adam S Dingens, and Jesse D Bloom. Cooperation between distinct viral variants promotes growth of H3N2 influenza in cell culture. *Elife*, 5:e13974, 2016.

[8] J Maynard Smith and George R Price. The logic of animal conflict. *Nature*, 246(5427):15–18, 1973.

[9] Johannes G Reiter, Ayush Kanodia, Raghav Gupta, Martin A Nowak, and Krishnendu Chatterjee. Biological auctions with multiple rewards. In *Proceedings of the Royal Society B: Biological Sciences*, volume 282, page 20151041. The Royal Society, 2015.

[10] Charles Gawad, Winston Koh, and Stephen R Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175, 2016.

[11] Jeffrey M Perkel. Single-cell sequencing made simple. *Nature*, 547(7661): 125, 2017.

[12] Valentine Svensson, Kedar N Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. Power Analysis of Single Cell RNA-Sequencing Experiments. *bioRxiv*, page 073692, 2016.

[13] Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*, 17(1):63, 2016.

[14] Alistair B Russell, Cole Trapnell, and Jesse D Bloom. Extreme heterogeneity of influenza virus infection in single cells. *bioRxiv*, page 193995, 2017.

[15] Desmond G Higgins. Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets. *Computer Applications in the Biosciences: CABIOS*, 8(1):15–22, 1992.

[16] Jiajie Zhang, Amir M Mamlouk, Thomas Martinetz, Suhua Chang, Jing Wang, and Rolf Hilgenfeld. PhyloMap: an algorithm for visualizing relationships of large sequence data sets and its application to the influenza A virus genome. *BMC Bioinformatics*, 12(1):248, 2011.

[17] Sewall Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proceedings of the Sixth International Congress of Genetics*, volume 1, 1932.

[18] Stuart A Kauffman and Edward D Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, 141(2):211–245, 1989.

[19] Michael C Whitlock and Denis Bourguet. Factors affecting the genetic load in Drosophila: synergistic epistasis and correlations among fitness components. *Evolution*, 54(5):1654–1660, 2000.

[20] Jennifer A Collins, M Gregory Thompson, Elijah Paintsil, Melisa Ricketts, Joanna Gedzior, and Louis Alexander. Competitive fitness of nevirapine-resistant human immunodeficiency virus type 1 mutants. *Journal of Virology*, 78(2):603–611, 2004.

[21] Roger D Kouyos, Gabriel E Leventhal, Trevor Hinkley, Mojgan Haddad, Jeannette M Whitcomb, Christos J Petropoulos, and Sebastian Bonhoeffer. Exploring the complexity of the HIV-1 fitness landscape. *PLoS Genetics*, 2012.

[22] Christof K Biebricher and Manfred Eigen. What is a quasispecies? In *Quasispecies: Concept and Implications for Virology*, pages 1–31. Springer, 2006.

[23] David Seifert, Francesca Di Giallonardo, Karin J Metzner, Huldrych F Günthard, and Niko Beerenwinkel. A framework for inferring fitness landscapes of patient-derived viruses using quasispecies theory. *Genetics*, 199(1): 191–203, 2015.

[24] Kenneth A Stapleford, Lark L Coffey, Sreyrath Lay, Antonio V Bordería, Veasna Duong, Ofer Isakov, Kathryn Rozen-Gagnon, Camilo Arias-Goeta, Herve Blanc, Stéphanie Beaucourt, et al. Emergence and transmission of arbovirus evolutionary intermediates with epidemic potential. *Cell Host & Microbe*, 15(6):706–716, 2014.

[25] Rasmus Henningsson and Magnus Fontes. SMSSVD – SubMatrix Selection Singular Value Decomposition. *ArXiv e-prints*, October 2017.

[26] Gonzalo Moratorio, Rasmus Henningsson, Cyril Barbezange, Lucia Carrau, Antonio V Bordería, Hervé Blanc, Stephanie Beaucourt, Enzo Z Poirier, Thomas Vallet, Jeremy Boussier, et al. Attenuation of RNA viruses by redirecting their evolution in sequence space. *Nature Microbiology*, 2:17088, 2017.

[27] Stéphanie Beaucourt, Antonio V Bordería, Lark L Coffey, Nina F Gnädig, Marta Sanz-Ramos, Yasnee Beeharry, and Marco Vignuzzi. Isolation of fidelity variants of RNA viruses and characterization of virus mutation frequency. *Journal of Visualized Experiments: JoVE*, (52), 2011.

[28] Magnus Fontes and Charlotte Soneson. The projection score – an evaluation criterion for variable subset selection in PCA visualization. *BMC Bioinformatics*, 12(1):307, 2011.

[29] Purificacion Carrasco, José-Antonio Daròs, Patricia Agudelo-Romero, and Santiago F Elena. A real-time RT-PCR assay for quantifying the fitness of Tobacco etch virus in competition experiments. *Journal of Virological Methods*, 139(2):181–188, 2007.

[30] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719, 2009.

[31] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[32] Henrik Lilljebjörn, Rasmus Henningsson, Axel Hyrenius-Wittsten, Linda Olsson, Christina Orsmark-Pietras, Sofia Von Palffy, Maria Askmyr, Marianne Rissler, Martin Schrappe, Gunnar Cario, et al. Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nature Communications*, 7, 2016.

[33] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv e-prints*, March 2013.

[34] Erik Aronesty. ea-utils: Command-line tools for processing biological sequencing data, 2011. URL `https://github.com/ExpressionAnalysis/ea-utils`.

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

[36] Mattia CF Prosperi and Marco Salemi. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics*, 28(1): 132–133, 2012.

[37] Kerensa McElroy, Torsten Thomas, and Fabio Luciani. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microbial Informatics and Experimentation*, 4(1):1, 2014.

[38] Niko Beerenwinkel, Huldrych F Günthard, Volker Roth, and Karin J Metzner. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology*, 3, 2012.

[39] Alexander R Macalalad, Michael C Zody, Patrick Charlebois, Niall J Lennon, Ruchi M Newman, Christine M Malboeuf, Elizabeth M Ryan, Christian L Boutwell, Karen A Power, Doug E Brackney, et al. Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. *PLoS Computational Biology*, 8(3):e1002417, 2012.

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

[41] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, 2011.

[42] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016.

[43] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, volume 2. Springer, 2009.

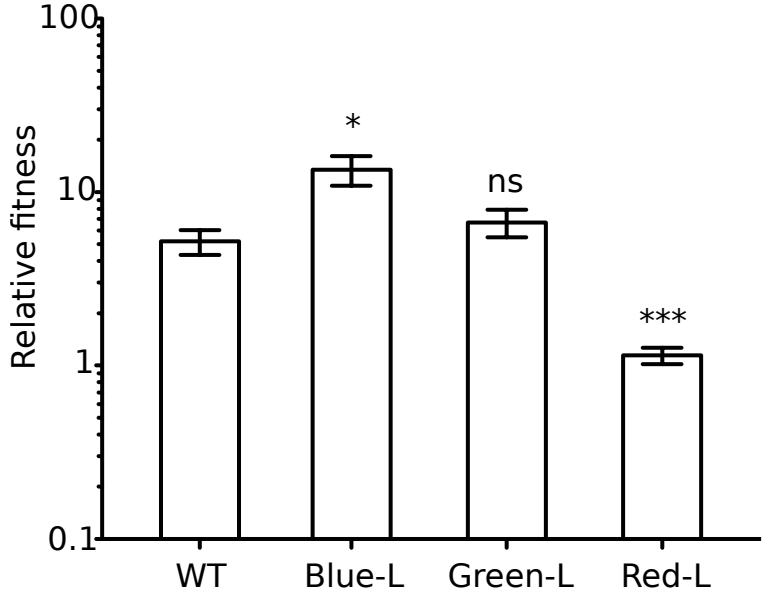
[44] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

[45] John Nash. The imbedding problem for Riemannian manifolds. *Annals of Mathematics*, pages 20–63, 1956.

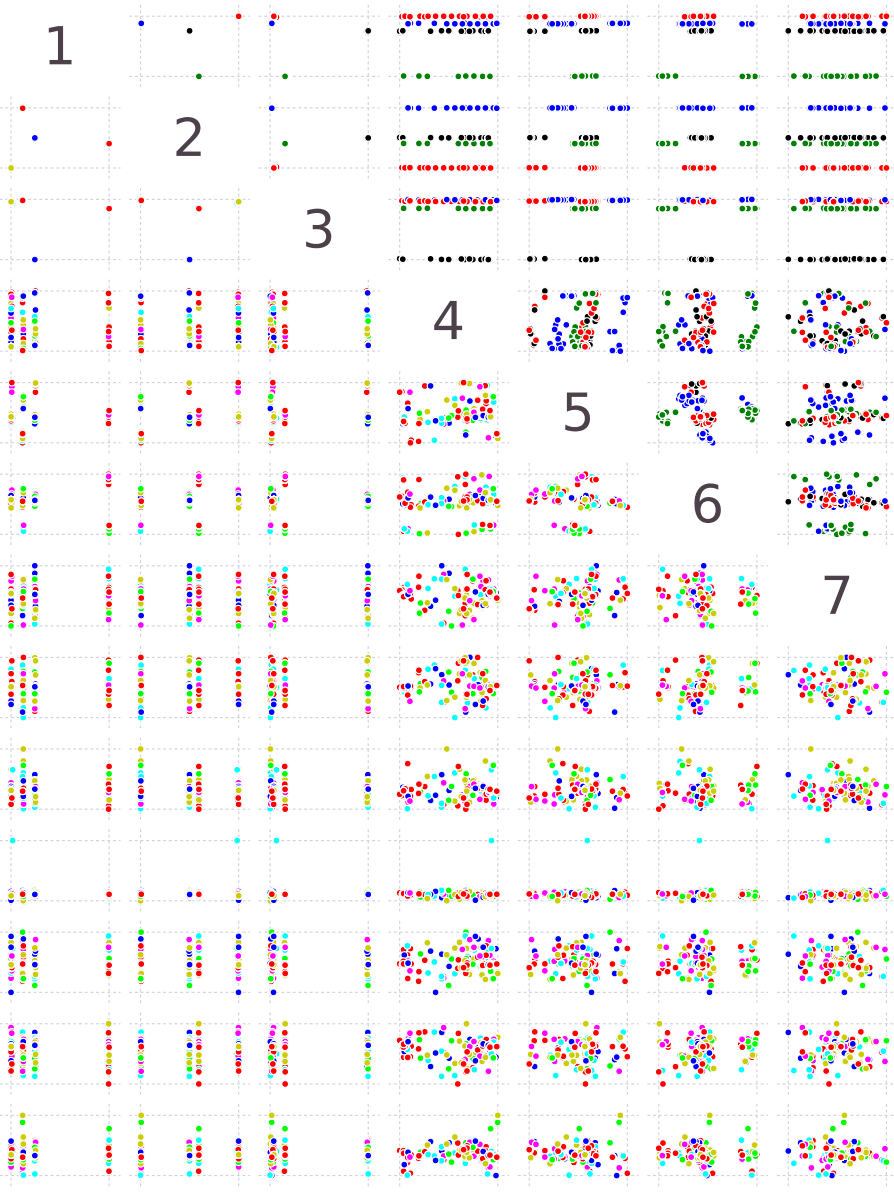
[46] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500): 2319–2323, 2000.

[47] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

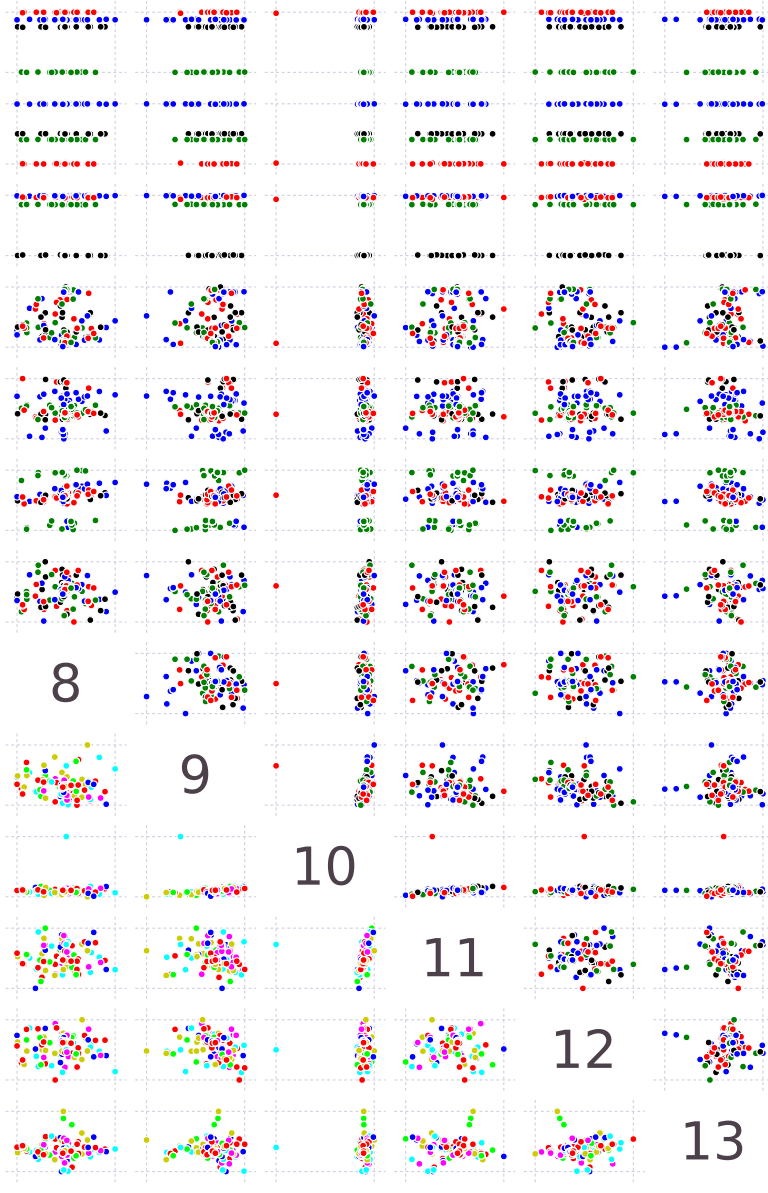
A Supplementary Figures



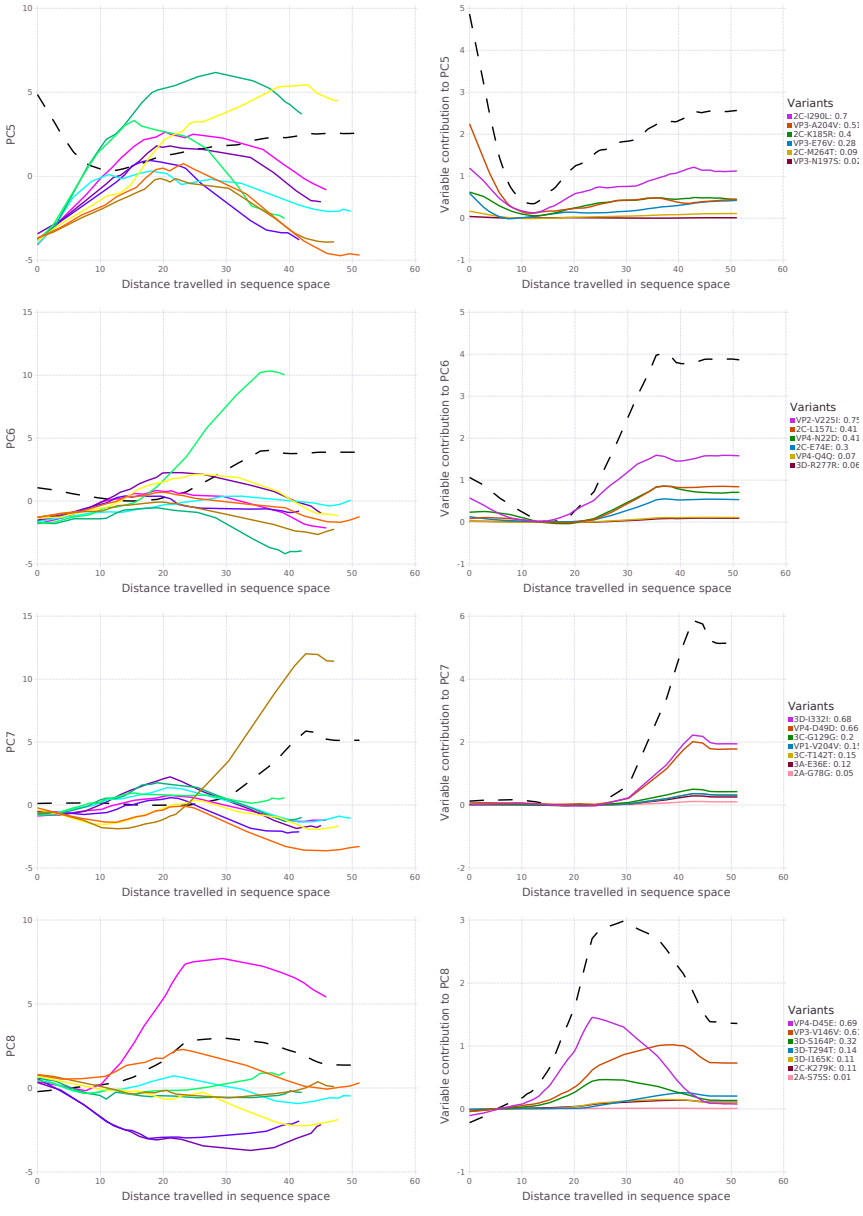
Supplementary Figure 1: Relative fitness by direct competition assay. Wild type (WT), Blue Lineage (Blue-L), Green Lineage (Green-L) and Red Lineage (Red-L). Mean and SEM are shown, $n = 6$, two-tailed unpaired t-test with Bonferroni correction. ns, not significant, $p = 0.334$; * $p = 0.013$; *** $p = 0.0008$.



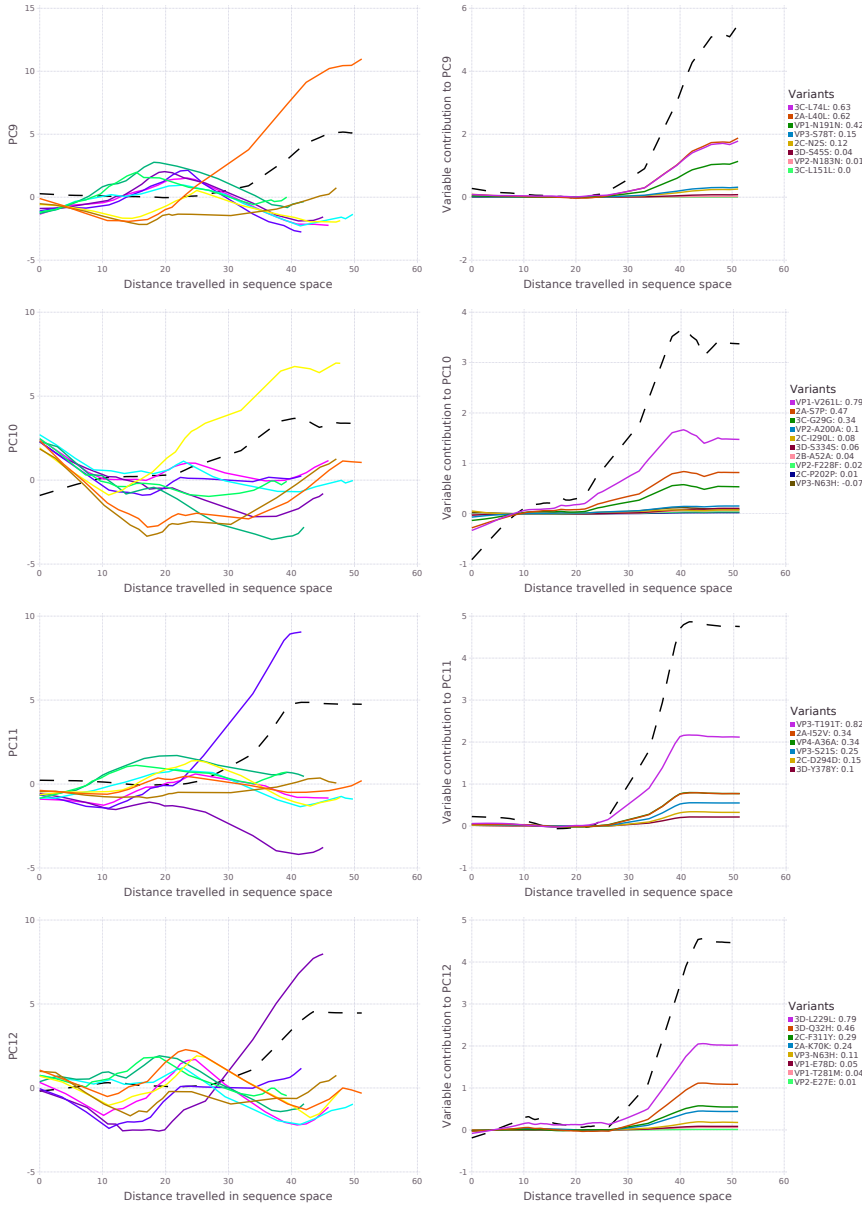
Supplementary Figure 2: Pairwise scatter plots showing the first 13 principal components in the analysis of the SynSyn data set plotted against each other, using consensus information only for the sequence space model. Plots above and below the diagonal are mirror images of each other. Each dot represents one viral population. (Continued on the following page.)



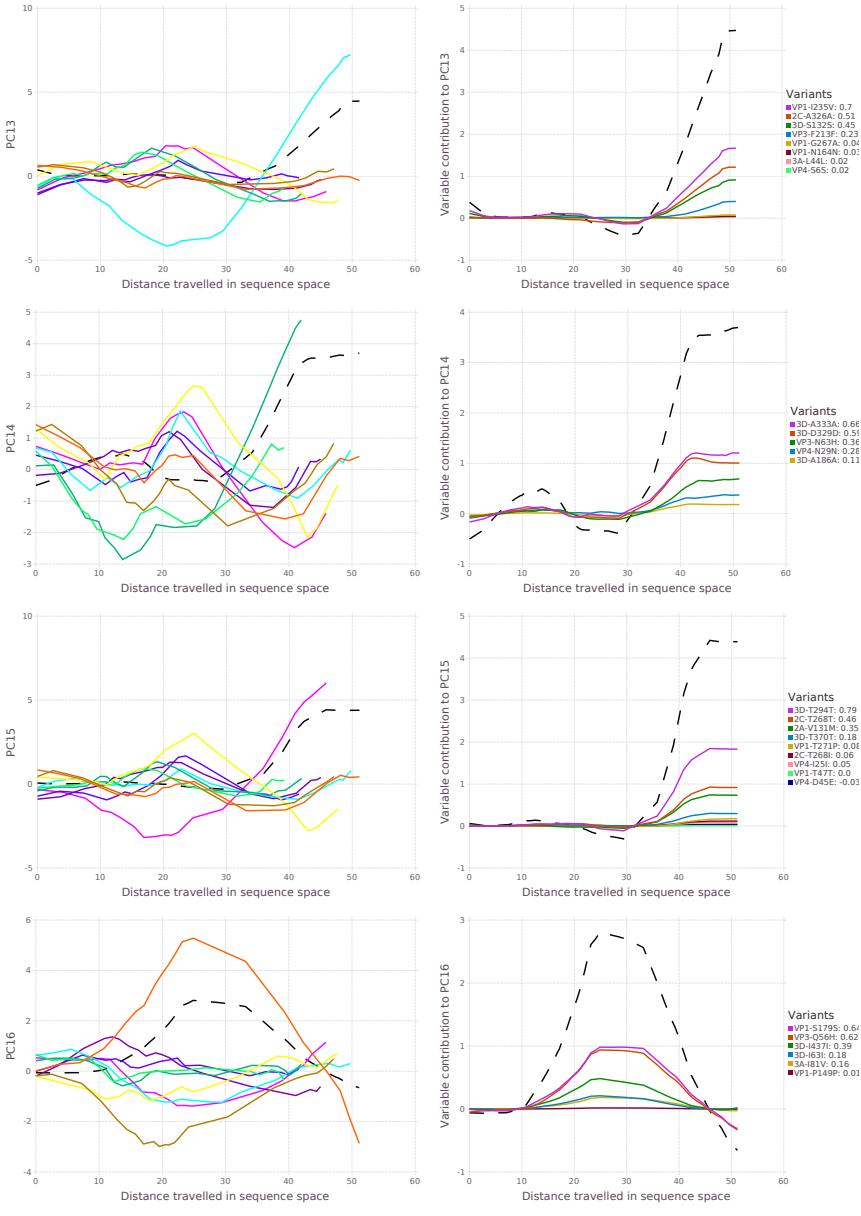
Supplementary Figure 2: Above the diagonal, samples are colored by lineage (Black: 1, Blue: 2, Green: 3, Red: 4) and below the diagonal, samples are colored by mutagen (Red: 5-fluorouracil, Light green: amiloride, Blue: 5-azacytidine, Yellow: Mn^{2+}, Cyan: ribavirin, Magenta: mock). All axes are rescaled to fill the plot area.



Supplementary Figure 3: **A.** Components 5–8. 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.

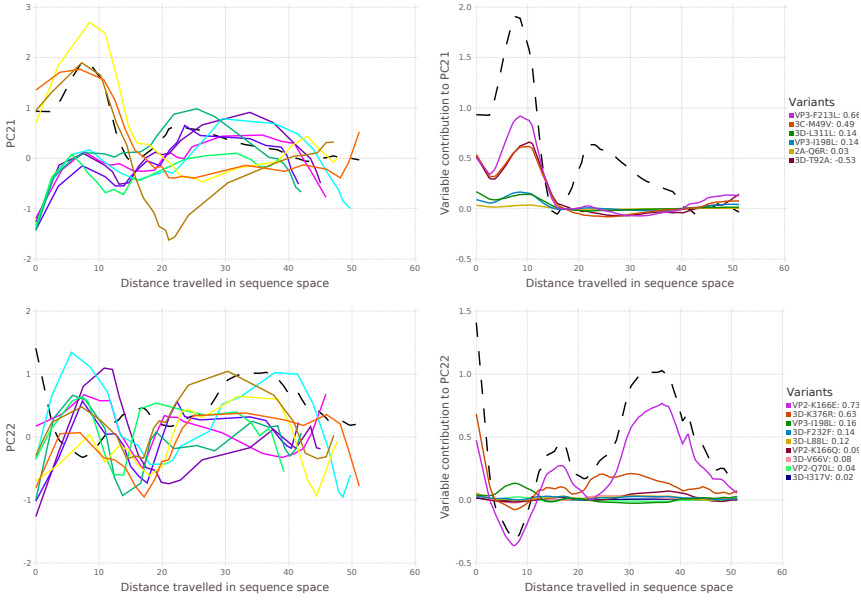


Supplementary Figure 3: **B.** Components 9–12. Left column: Principal components for replicates as a function of arc length. Right column: Variable contributions as a function of arc length. Both columns: The dotted black line shows the total contribution to σ_k at s.

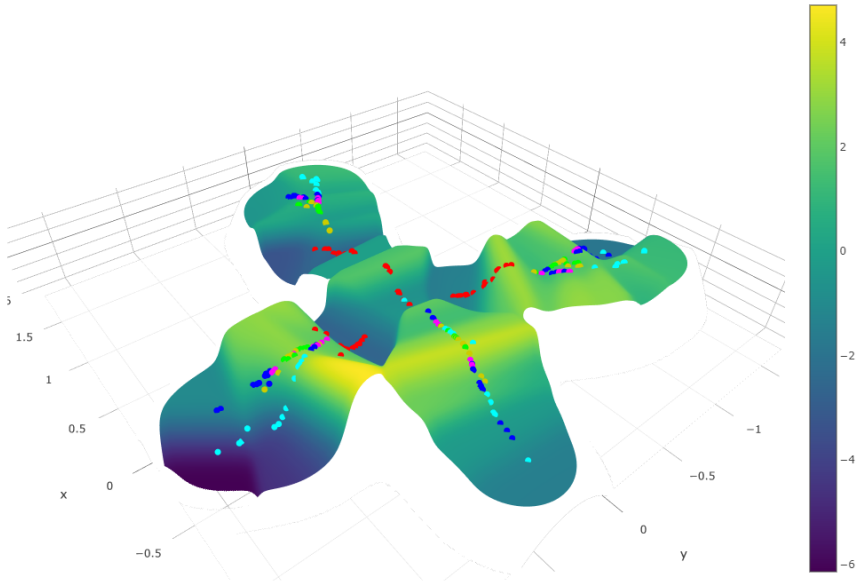


Supplementary Figure 3: **C.** Components 13–16. Left column: Principal components for replicates as a function of arc length. Right column: Variable contributions as a function of arc length. Both columns: The dotted black line shows the total contribution to σ_k at s.

Supplementary Figure 3: **D.** Components 17–20. 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.



Supplementary Figure 3: **E.** Components 21–22. 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.



Supplementary Figure 4: Fitness Landscape visualization of the SynSyn data set. Samples are colored by mutagen (Red: 5-fluorouracil, Light green: amiloride, Blue: 5-azacytidine, Yellow: Mn^{2+}, Cyan: ribavirin, Magenta: mock).

B Supplementary Methods

B.1 The Talus Plot

We will introduce the Talus Plot, a simple qualitative method for estimating the intrinsic dimension of a data set observed in noisy conditions. There already exists a plethora of methods for estimating the intrinsic dimension and one reason is that the problem is often inherently ill-posed. First, since the data set might be stretched out more along certain dimensions than others, the intrinsic dimension depends on the scale we are interested in. A narrow cylinder will for instance look like a line when zoomed out. Second, dimensions containing finer details of the data set might be masked or heavily corrupted by noise, making it impossible to make a clear-cut decision. The Talus Plot is designed to make it as easy as possible to find a dimension estimate that includes all dimensions that are detectable above the background noise. In particular, if the Talus plot gives a dimension estimate d, and the r ($r \leq d$) leading dimensions are removed, the new estimate will be $d - r$.

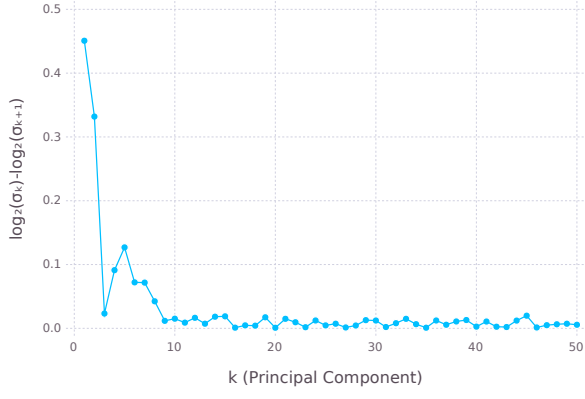


Figure 1: Talus Plot for a matrix X. The data matrix $X \in \mathbb{R}^{1000\times1000}$ is a rank 8 matrix with declining singular values with i.i.d. Gaussian noise added to each element. After $k = 8$, the Talus Plot shows small variation around a low mean, which is the expected behavior of a data matrix containing only noise.

In the Talus Plot, $\log(\sigma_k) - \log(\sigma_{k+1})$ is plotted as a function of the dimension k, where σ_k is the k'th singular value of the data matrix. An example is shown in Figure 1. Clearly, the Talus Plot is a close relative to the Scree plot that displays

σ_k^2 as a function of k. Taking the logarithm puts the singular values at a more reasonable scale where the background noise can be examined. As we will see, when noise dominates, the values in the Talus Plot will show small variations around a low mean. Thus, we find a dimension estimate using the Talus Plot by finding the breaking point at which the values start to exhibit this predictable behavior. To further understand the properties of the Talus Plot, we need some random matrix theory.

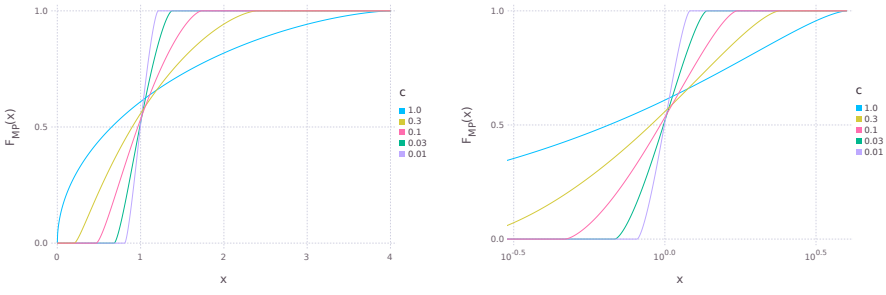


Figure 2: Cumulative distribution function of the Marchenko-Pastur distribution, for a few different values of c. Linear scale (left) and logarithmic scale (right) of the eigenvalues. Note that in the logarithmic scale, the CDF is well approximated by a linear function from the median ($y = 0.5$) and up, regardless of c.

Let $G(P, N)$ be the distribution of $P \times N$ random matrices with independent $\mathcal{N}(0, 1)$ elements. The distribution of the singular values of $A \sim G(P, N)$ (or equivalently, the eigenvalues of $A^T A$ since the k'th eigenvalue $\lambda_k = \sigma_k^2$) have been extensively studied in the literature [1]. The marginal distribution of unordered eigenvalues of $\frac{1}{P} A^T A$ converges asymptotically to the Marchenko-Pastur distribution [2] when $N, P \to \infty$ such that $\frac{N}{P} \to c \in [0, 1]$. The probability density function of the Marchenko-Pastur distribution is

$$f_{\text{MP}}(x) = \frac{\sqrt{(c_+ - x)(x - c_-)}}{2\pi c x},$$

where $c_\pm = (1 \pm \sqrt{c})^2$, the domain of $f_{\text{MP}}(x)$ is $[c_-, c_+]$ and we have assumed

that $N \leq P$. The corresponding cumulative distribution function is

$$F_{\mathrm{MP}}(x) = \frac{1}{2} + x f_{\mathrm{MP}}(x) + \frac{(1+c)\arcsin\left(\frac{x-1-c}{2\sqrt{c}}\right) - (1-c)\arcsin\left(\frac{(1+c)x-(1-c)^2}{2x\sqrt{c}}\right)}{2\pi c},$$

which is shown in Figure 2 for some different values of c. Notice how taking the logarithm of the eigenvalues makes it possible to find a good approximation of F_{MP} using a simple linear model, at least for the upper half of the eigenvalues/singular values. The behavior of the rest of the eigenvalues will clearly not be of interest when detecting where the background noise starts to dominate. A few realizations are shown in Figure 3, exemplifying the linear trend.

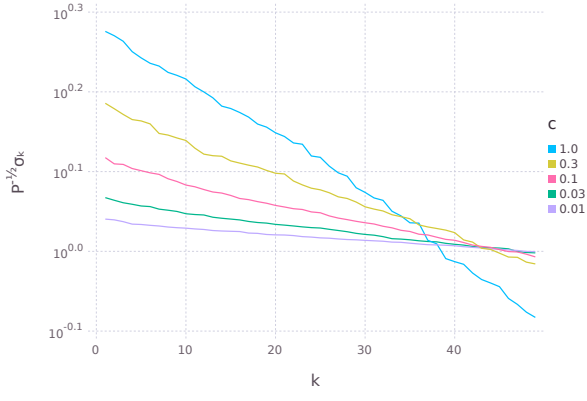


Figure 3: The singular values of A (scaled by $P^{-1/2}$), for a few different values of c. Each graph is a random realization. $A \sim G(P, N)$, $N = 99$ and $c = N/P$. The first 50 singular values are shown.

The effects of the linear decline are clearly visible in the Talus Plots (see Figure 1), as the base level around which the values deviate. To understand the deviations, we will look at the random variables $\kappa_k := F_{\mathrm{MP}}(\lambda_k)$ that take values in $[0, 1]$. Note that if one λ is randomly selected with equal probability from the set of random variables $\{\lambda_k;\ k = 1, \ldots, N\}$, then $F_{\mathrm{MP}}(\lambda)$ is by definition (asymptotically) uniformly distributed on $[0, 1]$. The transformation has the effect of removing trends (linear or other), but keeping the dependence structure between

the variables. The frequency contents of $\kappa_k - \kappa_{k+1}$ is shown in Figure 4, where the amplitude has been computed as the average over many realizations. Curiously close to a semi-circle, the amplitude peaks for the highest frequencies. Looking back at Figure 1, this corresponds well with the high-frequency oscillations that are visible in the singular values generated by the background noise.

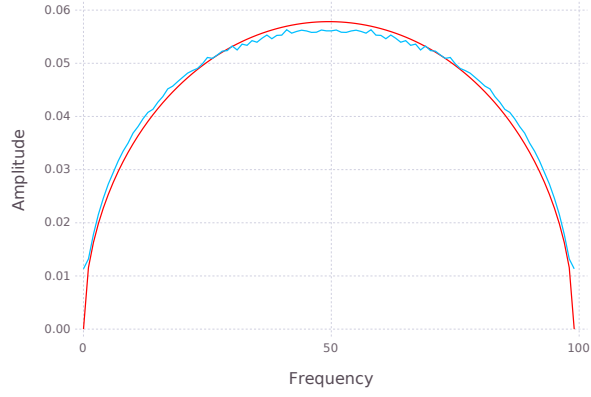


Figure 4: The absolute value of the coefficients in the discrete Fourier transform of $\kappa_k - \kappa_{k+1}$, computed as the average over 10000 random matrices. $P = 10000$ variables and $N = 1000$ samples. The shape is close to a semi-circle (red).

Bibliography

[1] Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, 9(4):543–560, 1988.

[2] Владимир А Марченко and Леонид А Пастур. Распределение собственных значений в некоторых ансамблях случайных матриц. *Математический сборник*, 72(4):507–536, 1967.

Printed by Media-Tryck, Lund 2017 NORDIC ECOLABEL, 3041 0903

Doctoral Theses in Mathematical Sciences 2017:9

ISBN 978-91-7753-479-2
ISSN 1404-0034