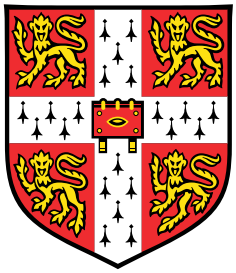
Molecular evolution of biological sequences



Ignacio Vázquez García

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Christ's College January 2017

En memoria de Felicidad Cirugeda Marzo.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Ignacio Vázquez García
January 2017

Acknowledgements

Many people have had a profound effect on my journey as a scientist so far. During my time in Cambridge I have been privileged to interact with and learn from the best. I am highly indebted to two communities, the Sanger Institute and the Department of Applied Mathematics and Theoretical Physics, which are an outstanding environment to develop as a scientist.

The work in this thesis could not have been accomplished without the generous help of a number of people. I would like to thank Ville Mustonen for welcoming me into his group and for the patient manner in which he has supervised this work and guided me through research. I have been fortunate to work closely with Andrej Fischer, Chris Illingworth and Daniel Kunz, all of whom have made a mark on this dissertation. The members of my thesis committee, Richard Durbin and Simon Tavaré, have provided a very valuable source of advice and fair critique. Both of them have also enabled the Mathematical Genomics and Medicine Programme to become an exceptional crossroads of disciplines, where I have met many brilliant colleagues: Hong Ge, Jürgen Jänes, Vagheesh Narasimhan and many others. Kati Sexton, Liz McIntyre, Annabel Smith and Christina Hedberg-Delouka provided very attentive administrative support, for which I am grateful. I also wish to thank my thesis examiners, Aylwyn Scally and Jürg Bähler, for their constructive and generous suggestions in the presentation and content of this dissertation.

A great many thanks go to all my collaborators, whose names certainly pepper this work and without whom it would not have been possible. In particular, I would like to thank Francisco Salinas, Johan Hallin, Jing Li and Gianni Liti from the Institute for Research on Cancer and Aging of Nice (France), as well as Elisa Alonso-Pérez and Jonas Warringer from the University of Gothenburg (Sweden). Equally, Leyla Bustamante and Julian Rayner from the Malaria Programme at the Sanger Institute have been wonderful collaborators. I also owe thanks to Fabio Puddu, Israel Salguero and Steve Jackson (Gurdon Institute) and Cristina Rada (MRC Laboratory of Molecular Biology) for enlightening discussions. The Quantitative Genomics meetings at the Sanger Institute have been an invaluable forum for

discussion, and I would like to thank members of the Cancer Genome Project who attended for their input. I have also had the opportunity to collaborate with many of them and others as part of the ICGC PCAWG consortium, including Stefan Dentro, Ignaty Leshchiner, Maxime Tarabichi, Quaid Morris, Peter Van Loo and David Wedge. I have also been extremely fortunate to spend time in various research groups as part of our programme, and also to travel to remote corners of the world where I have met many brilliant scientists. I wish to thank Sarah Teichmann (Sanger Institute) for hosting me in her laboratory as part of our rotations. I also express my gratitude to Boris Shraiman (University of California, Santa Barbara) and Michael Elowitz (Caltech), who were a source of inspiration for new ideas while in Santa Barbara.

My stay in Cambridge has been enriched by truly great friendships. I am especially thankful to Magda Reis, who has always provided a steady compass while steering through uncharted waters. Thanks for standing by me during every struggle and every success. Erik Garrison, Paloma Navarro and Manuela Carrasquilla have been great new companions in this journey, as well as old friends Christoph Aymanns and Ludomir Garreau whom I have not seen as often as we would have liked. And Lia Chappell, who has saved me from disaster on countless occasions. I would also like to thank all my friends at Christ's and everyone who has been involved with the Cambridge University Handball Club. Finally, I cannot be grateful enough to my family for their unwavering support and for teaching me the important things in life: to Concha, Ignacio and Carlos.

The work presented in this thesis has been supported by the Wellcome Trust as part of the PhD programme in Mathematical Genomics and Medicine at the University of Cambridge, by the Sanger Early Career Innovation Award, by Fundación Ibercaja, and partially supported by the National Science Foundation during a research stay at the Kavli Institute for Theoretical Physics (University of California, Santa Barbara).

Abstract

Evolution is an ubiquitous feature of living systems. The genetic composition of a population changes in response to the primary evolutionary forces: mutation, selection and genetic drift. Organisms undergoing rapid adaptation acquire multiple mutations that are physically linked in the genome, so their fates are mutually dependent and selection only acts on these loci in their entirety. This aspect has been largely overlooked in the study of asexual or somatic evolution and plays a major role in the evolution of bacterial and viral infections and cancer.

In this thesis, we put forward a theoretical description for a minimal model of evolutionary dynamics to identify driver mutations, which carry a large positive fitness effect, among passenger mutations that hitchhike on successful genomes. We examine the effect this mode of selection has on genomic patterns of variation to infer the location of driver mutations and estimate their selection coefficient from time series of mutation frequencies. We then present a probabilistic model to reconstruct genotypically distinct lineages in mixed cell populations from DNA sequencing. This method uses Hidden Markov Models for the deconvolution of genetically diverse populations and can be applied to clonal admixtures of genomes in any asexual population, from evolving pathogens to the somatic evolution of cancer.

To understand the effects of selection on rapidly adapting populations, we constructed sequence ensembles in a recombinant library of budding yeast (*S. cerevisiae*). Using DNA sequencing, we characterised the directed evolution of these populations under selective inhibition of rate-limiting steps of the cell cycle. We observed recurrent patterns of adaptive mutations and characterised common mutational processes, but the spectrum of mutations at the molecular level remained stochastic. Finally, we investigated the effect of genetic variation on the fate of new mutations, which gives rise to complex evolutionary dynamics. We demonstrate that the fitness variance of the population can set a selective threshold on new mutations, setting a limit to the efficiency of selection.

In summary, we combined statistical analyses of genomic sequences, mathematical models of evolutionary dynamics and experiments in molecular evolution to advance our understanding of rapid adaptation. Our results open new avenues in our understanding of population dynamics that can be translated to a range of biological systems.

Table of contents

List of figures

List of tables

Glossary

Allele One of two or more alternative forms of a gene or DNA sequence.

Amino acid One of the building blocks of proteins. Twenty different amino acids are commonly found in proteins (e.g., arginine, serine, leucine, etc.).

Base The alphabet used by DNA molecules as their building block. The bases in DNA are adenine (`A`), cytosine (`C`), guanine (`G`) and thymine (`T`).

Beneficial mutation An advantageous or beneficial mutation increases fitness. It would therefore be subject to positive selection.

Codon Three adjacent bases in DNA and messenger RNA that encode an amino acid in a protein.

Deleterious mutation A deleterious mutation decreases fitness. It would therefore be under negative selection.

Diploid A cell or an organism having two copies of the genome.

Driver mutation A mutation that gives a fitness advantage to the cells that carry it, via the modifications in phenotype that it causes.

Epistasis Context-dependent fitness effects due to genetic interactions between mutations at different sites in the genome.

Fitness The relative ability of an individual (or genome) to produce surviving offspring in a population.

Fixation A process whereby one variant becomes more frequent in a population until all other variants go extinct. Once this variant fixes, it is found in all individuals of a population.

Frameshift mutation A mutation in a coding region that adds or subtracts a number of bases that is not a multiple of three, causing the triplet reading frame to be shifted.

Gene Part of a DNA sequence that encodes a functional RNA molecule or a protein.

Genetic code A near-universal code to convert a set of three adjacent nucleotides (or *codon*) to one of the amino acids. Three these codons are also devoted to starting or stopping the process of translation. The code is redundant, with $4^3 = 64$ possible codons specifying only 20 amino acids plus the start and stop signals.

Genetic drift Stochastic change of allele frequencies through generations, with characteristic time N.

Genotype Combination of allelic states across one or more chromosomes at a polymorphic site in the genome.

Germline mutation A heritable mutation present in the cells that are responsible for passing genetic information from one generation to the next.

Haploid A cell or an organism that has one copy of the genome.

Haplotype Combination of allelic states in the same chromosome.

Heterozygous Cell or organism carrying different alleles in each chromosome copy.

Homozygous Cell or organism carrying the same allele in each chromosome copy.

Linkage (dis)equilibrium Absence (presence) of correlations between mutations in the genome, which determines how likely two alleles are to be co-inherited.

Locus An entity of the genome which can be a single base, a gene or a chromosomal region.

Meiosis A type of cell division in which one cell goes through stages of DNA replication followed by two consecutive rounds of cell division, to produce four potential daughter cell. It is this last step that assures the generation of genetic diversity in sexual reproduction. Meiosis reduces the chromosome number per cell by half.

Mitosis A type of cell division in which one cell duplicates and divides the genetic material, creating two identical daughter cells with the same number of chromosomes as the parent. Mitosis plays a role in cellular reproduction, growth, repair and asexual reproduction.

Mutation The process by which random nucleotide changes are introduced in individuals, with characteristic time $\frac{1}{\mu}$. These changes in DNA can be single-base substitutions, insertions, deletions, or large-scale rearrangements.

Neutral mutation A neutral mutation is one that does not change the fitness.

Non-synonymous mutation Mutations which lead to a change in the amino acid sequence. They include missense and nonsense mutations. Missense mutations can introduce amino acids with similar chemical properties (e.g., hydrophobicity or charge) which may have little effect; or they may introduce amino acids with different properties that are likely to be deleterious. Nonsense mutations result in a stop codon and are likely to be deleterious.

Nucleotide Monomers that conform the DNA and RNA polymers.

Open reading frame Reading frames correspond to one of three possible ways that a DNA sequence can be divided into consecutive triplets. An open-reading frame (ORF) is a reading frame with a start codon.

Passenger mutation A mutation assumed to have neutral impact on cell phenotype.

Polymorphism The existence of two or more genetic variants (e.g. DNA sequences, proteins, chromosomes) or two or more phenotypic variants in a population. A polymorphic variant is found in a fraction of individuals within a population.

Recombination Exchange of DNA between the two copies in a chromosomal pair. This is the process by which multi-allelic genomes are mixed during meiosis.

Selection Deterministic frequency changes of alleles with fitness difference ΔF.

Somatic mutation A non-heritable mutation arising in cells that are not passed on to the next generation.

Subclone Maximal set of cells carrying the same arbitrary set of mutations.

Synonymous mutation A nucleotide change with no effect on the protein sequence, due to the degeneracy of the genetic code. Synonymous mutations are often assumed to be neutral.

Transition Interchanges of two-ring purines (`A<>G`) or of one-ring pyrimidines (`C<>T`).

Transversion Interchanges of purine (`C`, `T`) for pyrimidine bases (`A`, `G`), which therefore involve exchange of one-ring and two-ring structures.

Chapter 1

Introduction

'El tiempo es la materia de la que he sido creado.'

— Jorge Luis Borges,
'Nueva refutación del tiempo' (1946)

1.1 Information processing in the cell

Living matter has the capacity to self-replicate, which sets it apart from its inanimate counterpart. *Omne vivum ex vivo* – the notion that 'all life comes from life' – was promulgated by Louis Pasteur when he experimentally disproved the spontaneous generation of life [1], a view that had been widely held since Ancient Greece. Cells, which are the elementary constituents of living matter, contain a complete instruction set to make a quasi identical copy of themselves. Deoxyribonucleic acid (DNA) is the informational molecule in a cell's genome that stores and copies this information to build a new organism. The instructions contained in the genome are passed on in an unbroken thread that stretches for billions of years, from primordial cells to us and every other organism on Earth.

This thesis will focus on the study of the evolutionary dynamics of rapid adaptation at the molecular level. It will combine statistical analyses of genomic sequences, mathematical models of evolutionary dynamics and experiments in molecular evolution. We will first introduce the main constituents that process information in the cell, with a particular emphasis on DNA as our main molecule of interest. We will then present the evolutionary forces that can alter the information content in DNA. We shall then discuss molecular genetic techniques that can be used to study these processes, followed by an introduction to population

genetic inference from DNA sequences. Finally, we summarise each of the chapters and briefly outline their inter-relationships.

1.1.1 Genetic basis of inheritance

The study of genetics deals with the basic principles that govern how information in the genome is interpreted to make the components of a cell, and how it provides the means to transmit this information. Genetics began as an investigation into the transmission of variation between individuals, such as differences in the colour of pea seeds and fly eyes [2]. In doing so it was able to successfully identify the factors of heredity, or how information is passed on from generation to generation. From these studies, geneticists inferred the existence of genes and many of their properties [3, 4]. But genes remained mathematical abstractions and the apparently unchanging character of heritable traits was taken as fact. Yet this posed a problem for any molecular description of the inheritance of traits, as the inherently probabilistic nature of the fundamental laws of physics and chemistry would violate their stability and invariance of traits. Erwin Schrödinger noted this in relation to the protruding 'Habsburg lip', a dominant trait that had been passed down through generations in this royal family [5]. He speculated that inheritance must be based on an 'aperiodic solid', a suggestion that anticipated DNA as the carrier of genetic information. This prescient view was proven right by the demonstration that genetic information for virulence could be transferred between bacteria, which prompted the identification of DNA as the information-carrying molecule [6, 7]. The

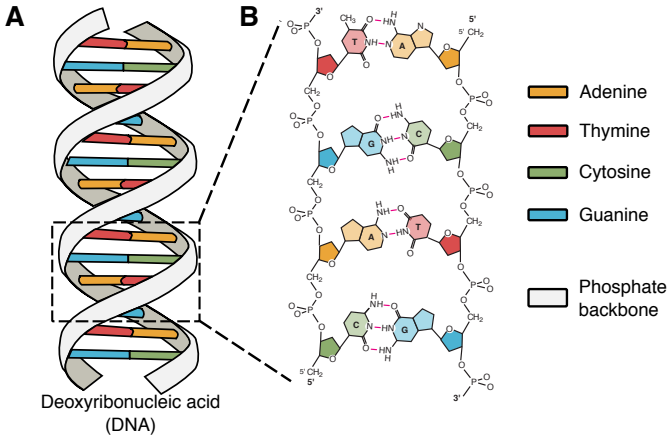


Fig. 1.1 **(A)** Deoxyribonucleic acid (DNA) has a double-helix structure. **(B)** Base pairing in the Watson-Crick structure of DNA. Hydrogen bonding between bases enables the correct pairings. The sugar and phosphate groups are identical for all bases, which form the outer backbones of the helix. Images adapted from Wikimedia Commons.

field of molecular biology then rapidly developed with Francis Crick and James Watson's discovery of the double-helix structure of DNA [8], which explained crucial observations of crystallographic X-ray diffraction patterns from DNA fibres that Rosalind Franklin and Maurice Wilkins had acquired (Fig. 1.1A). The structure of DNA itself had extraordinary explanatory power, as it told us how genetic information is copied within a cell and transmitted from generation to generation. This led to the discovery of a near-universal genetic code, which links the sequence of DNA to the structure of proteins [9], and ultimately to the realisation that DNA encodes ribonucleic acids (RNA), which in turn encode proteins [10, 11].

1.1.2 Biophysical properties of the genome

The structure of DNA consists of two complementary chains twisted around each other, forming a double helix. Genetic information resides in the linear order of nucleotides along a strand of DNA, divided into genes. The four-base DNA code specifying the amino acid sequence of a protein is copied, or *transcribed*, into RNA. During *translation*, the four-base code of messenger RNA (mRNA) is decoded into the 20-amino acid alphabet of proteins by reading three bases at a time. The genetic code assigns meanings to an unbounded number of nucleotide sequences, mapping each triplet of bases to a codon. A specific protein is then produced by the transcription of DNA into mRNA and the subsequent translation of mRNA in the ribosome. Finally, for DNA to copy itself as the cell is preparing to divide, it must undergo accurate *replication* through DNA polymerase-mediated complementary pairing of adenine (`A`) with thymine (`T`) and of guanine (`G`) with cystosine (`C`) as shown in Figure 1.1B.

DNA sequences contain a complete description of the organism and provide insights into the conservation and variation in proteins known to have important functional roles in the cell. We share thousands of individual proteins with other eukaryotes, hundreds of protein complexes and the majority of cellular organelles [12]. The essential genes and networks needed for cell division have revealed the conservation of the transcription, translation and replication machinery across the tree of life [12].

1.2 Evolutionary forces

Molecular biology has supplied a large number of novel tools that, together with genetics, have enabled researchers to address questions regarding the nature of how information is transmitted, retained and modified [13]. We have begun to elucidate how organisms devel-

oped by exploring the evolution of genomes and the origins of diversity at the molecular level. However, the state of biology from today's vantage point may appear comparable to chemistry before the periodic table. Even though we have read most human genes and some regularities have been found, there is a major lack of unifying principles.

Is there room for any such principles in modern biology? To answer this, we will briefly review the main processes that can modify the information content of DNA: mutation, selection, recombination and genetic drift. We will discuss the need for a quantitative description of how evolutionary forces alter a population through changes in allele and genotype frequencies. In particular, we will put this in the context of mathematical models in population genetics, which should give us a sense of the evolutionary changes that can occur on different timescales, of what is common and what is extremely unlikely.

1.2.1 Sources of genetic variation: mutation and recombination

The stable biophysical and chemical properties of DNA make it an ideal carrier of genetic information. How many errors are made each time a genome is copied? DNA must replicate itself with extreme fidelity, and the reliability of the copying process is limited by thermal noise. We know that the dominant energetic contribution towards the stability of DNA comes from hydrogen bonds between bases in complementary base pairs, which is of the order $\Delta E \sim 10 k_B T$ [14]. Therefore, the probability of an incorrect base pairing should be, according to the Boltzmann distribution, $e^{-\Delta E/k_B T} \sim 10^{-4}$. Considering that a typical gene is a thousand nucleotides long, then replication of the DNA would introduce roughly one mutation in every tenth protein [15]. However, genomes can be copied with almost no mistakes and most organisms have per-genome mutation rates in the range of $10^{-8} - 10^{-12}$ [16, 17]. We saw that the flow of genomic information from DNA into functional proteins by transcription results in the synthesis of mRNA, and the subsequent process of translation of that mRNA into the string of amino acids that make up a protein. Organisms can withstand errors in transcription ($10^{-5} - 10^{-4}$) [18, 19] and translation ($10^{-4} - 10^{-3}$) [20], but the archival copy in DNA has to remain very stable in comparison. The accuracy in DNA replication can only be explained by a process known as proofreading, whereby the polymerase checks the newly added base before adding the next one [12]. There are also additional safeguards that fix incorrect bases after replication, like the base excision repair system which repairs `T·G` mismatches and damaged bases, mismatch excision repair that corrects other mismatches and small insertions and deletions, or nucleotide excision which repairs chemical adducts that distort normal DNA shape [12]. Other DNA lesions that are not repaired

by these mechanisms (e.g., double-strand breaks) can be corrected by two systems, namely homologous recombination and non-homologous end-joining [12].

Recombination itself is the other evolutionary force that introduces new genetic variation. Homologous recombination takes place in sexual reproduction during meiosis, and enables the exploration of completely new regions of the genotype space by bringing together the genomes of different individuals. While alleles located on different chromosomes randomly segregate in every meiosis, those linked on the same chromosome do not. In most species there are one to two crossover events per chromosome per replication [21]. Two genes on a chromosome that have a 1% probability of crossover per generation are defined to be at a distance of one centimorgan, or cM. In humans, the average rate of recombination is about 1 cM per 1 Mb [21]. The distribution of recombination events in the genome is far from uniform, with hot spots near chromosome telomeres and cold spots towards their centromeres [22, 23]. Recombination events can disrupt the co-inheritance of alleles in the genome. The association between alleles at different loci can be estimated by a quantity known as 'linkage disequilibrium'. Over time, a haplotype block will begin to be broken down by recombination and the haplotype frequency will decrease, which translates as a decay in linkage disequilibrium.

1.2.2 Eliminating diversity: genetic drift

Genetic variation arises stochastically by mutation and can have idiosyncratic effects: the order and timing of mutations can be different in different individuals, setting populations off on different futures. Reproductive fluctuations – also known as genetic drift – are a stochastic evolutionary force that influences whether a mutation will be kept or lost in the population. Logically, the magnitude of genetic drift is related to the size of the population. When genetic drift plays a dominant role in the fixation of new mutations, it can be shown that the rate of substitution is then equal to the rate of mutation. This is the underlying principle behind the 'molecular clock' hypothesis, which states that the number of substitutions between two organisms should scale linearly with time [24]. A noteworthy example of its exactitude is the α-globin protein – one of the amino acid chains that make up haemoglobin – which accumulates one amino acid change every 6 million years [24]. We will see now that genetic drift stands as a noisy background of neutral (or nearly neutral) mutations that can obscure the effects of natural selection.

1.2.3 Shaping genetic variation: natural selection

Most mutations in a gene misspell the protein it encodes. Very few of these changes are beneficial to the organism. The majority of changes are innocuous or mildly harmful, some lethal. Selection is a deterministic evolutionary force that acts on these changes in a characteristic timescale which is inversely proportional to their effect on fitness. Without the knowledge at the time that genes exist or how they bring change, Charles Darwin and Alfred Wallace described that the selection of whole organisms will direct changes in any self-replicating unit of living matter [25]. As a result, we now know that mutations that are passed on through generations will define the genetic makeup of the population. In turn, whilst not affecting long-term evolution, selection will also be at play all the way down to individuals, to cells and molecules [26]. We will now explore which new possibilities can be directed by selection on different timescales and organisational levels.

Within populations, differential reproductive success alters the structure of gene pools which gives rise to evolution. It has typically been thought that selection only acts over very long timescales: if the selective pressure is very weak, such as the effect of translational optimisation on codon usage bias, it can take billions of years [27]. However, recent findings suggest it can also occur on very short timescales: for instance, major changes in the evolution of the seasonal influenza virus can take place within just a few years and be selected for, resulting in vaccines needing frequent updates (Fig. 1.2B) [28]. What is yet more extraordinary, the sequence diversity of influenza viruses pales in comparison to certain pathogens like HIV [29]. In a single individual who has been infected for 6 years, the turnover rate of the HIV genome is equivalent to all influenza viruses around the globe in one entire season, and to many millions of years in humans.

At the cellular level, clonal evolution can take place when some cells with faster rate of replication can be selected for in a cell population or tissue. This process takes place throughout our lifetimes. Cancer, which is the end-state of development and ageing, also progresses under Darwinian evolution driven by somatic mutations and clonal expansions [30–32]. In the interim process, malignant cells can evolve major genomic aberrations (e.g., whole-genome duplications) up to 20 years before a cancer is detected [33]. This suggests that there is much room left to explore with regards to how the genomes in our cells, and the genetic makeup of all organisms, evolve throughout their lifetime [34].

Natural selection is also operating within cells. For example, it plays a role on timescales of milliseconds in directing different unfolded copies of the same protein through multiple paths from a denatured state to a unique folded structure [39]. Similarly, receptor molecules

of the immune system undergo random processes of mutation and recombination to maintain a very diverse repertoire – each specialised in recognising specific pathogens – and are directed by clonal selection to mount a response within hours, e.g., to vaccination [36, 40–42].

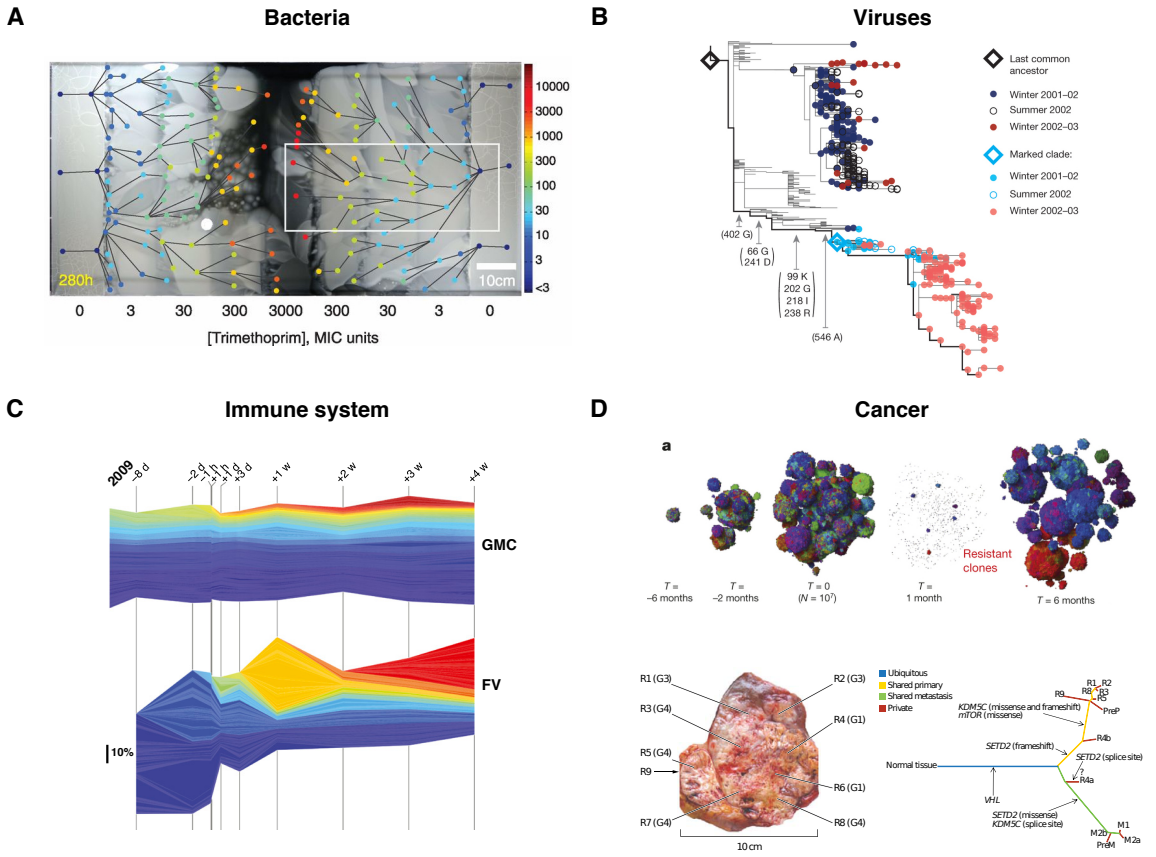


Fig. 1.2 Examples of evolutionary dynamics, genetic diversity and selection. **(A)** Bacteria: *E. coli* evolving in a spatial gradient of the antibiotic trimethoprim can adapt to a $\sim 10^3$-fold increase in concentration after 11 days. **(B)** Viruses: Seasonal influenza viruses circulating around the globe are under constant selective pressure to adapt to the host's immune system. **(C)** Immune system: The immune repertoire measured by sequencing before and after vaccination against the seasonal influenza virus. The temporal evolution of clones in the immune repertoire of two individuals shows very different responses over a timescale of a few weeks. **(D)** Cancer: A spatial model of tumour growth before and after chemotherapy selection (top). Genetic heterogeneity in space and time is often observed in real tumours (bottom). The phylogenetic relationships between spatially separated biopsy samples from renal-cell carcinomas recapitulate significant intratumour heterogeneity. Images adapted from Baym et al. [35], Łuksza and Lässig [28], Laserson et al. [36], Waclaw et al. [37] and Gerlinger et al. [38].

1.2.4 Genetic interactions and genetic linkage

How many mutations typically contribute to adaptive dynamics? A biological function can rarely be attributed to a single gene. One of the rare cases is haemoglobin, which is the sole carrier of gas molecules into the bloodstream [43]. Mutations in haemoglobin can convert a glutamic acid codon (GAG) to a valine codon (GTG), leading to a malformation of blood cells that causes sickle-cell anaemia [44]. On the other hand, if a biological function can be attributed to many genes, the effects of mutations in different genes can be context-independent (additive) and all mutational paths will lead to the same functional genotype. In other words, if the topology of the fitness landscape is smooth, there will not be any historical contingency or 'memory' of intermediate states in the mutational path. Conversely, when the effects of mutations are context-dependent (epistatic), evolutionary outcomes may also become path-dependent. If a combination of mutations is beneficial, but intermediate mutants are deleterious compared to the wild-type, epistasis will cause the rate of evolution to slow down. This scenario often arises, for example, in the emergence of bacterial or viral resistance to drugs [45, 46]. To determine which scenario is at play, it is clear that with a genome of size L, characterising additive and epistatic interactions in all 2^L combinations of mutations is unattainable: this figure exceeds the number of atoms in our Universe by many orders of magnitude. Rather than characterising the effect of every possible mutation, following the genetic and phenotypic trajectories of population ensembles can reveal the statistical properties of the underlying fitness landscape, e.g., its 'ruggedness'. It may then be within reach to have a theoretical description of the statistical structure of typical population trajectories.

1.3 Molecular genetic techniques

In this thesis, we hope to bridge what the biophysical constraints are on these basic molecular processes at the molecular level (the microscopic dynamics of the genome) and their biological relevance (the macroscopic contribution to the organism's fitness). Many attempts to do this thus far have been saddled with unrealistic expectations limited by technological advances, but we have reasons to be optimistic with our current abilities to 'read' and more recently 'write' information in the genome. Recent advances in experimental technologies give us great insight into the functioning of biological systems both at the molecular and intra-cellular levels, as well as the level of large scale functional systems in the organism [13]. We will now discuss these developments, focusing on the ability to make quantitative measurements of the constitutive elements of cells and populations, and establish links to their

function. We will also highlight opportunities to measure and modify population genetic parameters like the population size, mutation rate, selective constraints, or recombination rate.

1.3.1 DNA, RNA and protein sequencing

The recent development of genome sequencing techniques has enabled a revolution in the study of biological systems [47]. While the current state-of-the-art reflects otherwise, the ability to sequence proteins and RNA preceded DNA sequencing. Fred Sanger was the first to determine the ordered amino acid sequence of the protein insulin in 1952, by fragmenting the protein into polypeptide chains, separating the pieces by chromatography and electrophoresis, and completing the sequence by comparing overlaps between fragments [48]. RNA came next, and in 1965 Robert Holley and colleagues deciphered the nucleotide sequence of alanine transfer RNA (tRNA) by a similar process [49]. Researchers then steadily added to the number of ribosomal and tRNA sequences, which led to the first complete protein-coding gene – the coat protein of bacteriophage virus MS2 – being sequenced, followed by its complete RNA genome [50, 51]. Researchers began to adapt these methods in order to sequence DNA. Two methods subsequently transformed the field in 1977, independently showing that one could determine the nucleotide order in a DNA sequence by measuring distances from a radioactive label to positions of specific bases in the DNA molecule. These were the 'chemical cleavage' method developed by Maxam and Gilbert [52], and the 'chain termination' method by Sanger, Nicklen and Coulson [53], which demonstrated that DNA fragments could be separated on the basis of size using electrophoresis, and the order of bases inferred. The 'chain-termination' method, commonly known as Sanger sequencing, enabled the first DNA genome to be sequenced: the 5 kb-long genome of bacteriophage virus $\phi X174$ [54]. Sanger sequencing continues to be used to date, generating read-lengths of up to $\sim$1 kb, and per-base error rates as low as 1×10^{-5} [55].

This method for DNA sequencing was put to immediate use with a strategy known as shotgun sequencing, which is based on sequencing random clones and hierarchically assembling them using the overlaps between DNA fragments [56]. A number of improvements to Sanger sequencing were made which involved replacing radiolabelling with fluorescence-based detection and automation through capillary-based electrophoresis [57]. The automation of laboratory protocols and computational methods for sequence analysis optimised sequencing to read $\sim$1 kb per day, and the generation of sequence data began growing exponentially. With a yearly doubling in capacity in the 1990s, several genomes could be suc-

cessfully completed including the first free-living organism (*H. influenzae*, 2 Mb) and several model organisms (*E. coli*, 4.6 Mb; *S. cerevisiae*, 12 Mb; *C. elegans*, 100 Mb; *A. thaliana*, 135 Mb) which were among the first to be sequenced [58–62]. A vision to sequence the human genome – which is ∼3.2 Gb long – came to fruition by integrating automated Sanger sequencing with a strategy to clone large fragments of the human genome into bacterial artificial chromosomes. This monumental undertaking culminated in the sequencing of the first human genome in 2001 [63, 64]. However, it still demanded onerous colony picking and plasmid preparation.

Breakthroughs in scale only came with alternative sequencing techniques that had been explored in parallel and did not involve electrophoresis. Rather than using *in vivo* bacterial cloning, these new high-throughput approaches amplified the DNA templates to be sequenced *in vitro*. They roughly fall into two categories: sequencing by ligation and sequencing by synthesis. In sequencing by ligation, a DNA fragment of interest is hybridised to a probe sequence that is bound to a fluorescent label, and is then ligated to an adjacent oligonucleotide for imaging [55]. The emission spectrum of the fluorescent label indicates the identity of the base complementary to specific positions in the probe. In sequencing by synthesis, an enzyme drives the synthesis (e.g., a polymerase or a ligase) and a signal, such as fluorescence or a change in ionic concentration, identifies the incorporation of a nucleotide into an elongating strand [55]. These high-throughput sequencing methods have eventually superseded Sanger sequencing in most applications. Massive parallelisation is facilitated by the creation of many millions of individual sequencing-by-ligation or sequencing-by-synthesis reaction centres, thus reading millions of molecules in parallel. This has ultimately enabled order-of-magnitude improvements in scale, progressing from sequencing 250 bp to 250 Gb per day [65]. The genomes of a large number of organisms have been sequenced, including ourselves [66] and our near cousins [67], going all the way back to unicellular organisms. More widely, high-throughput DNA sequencing is enabling a wide range of applications to quantify different biological phenomena (e.g., genetic variation, RNA transcription, protein-DNA interactions and chromosome conformation) [47]. In recent years, new single-molecule sequencing technologies are surpassing the limitations of previous short-read high-throughput techniques, enabling long read lengths and real-time sequencing [68].

1.3.2 Genome engineering and directed evolution

Once we have characterised sequence variants in the genome, a principal aim is their functional analysis in order to establish genotype-phenotype relationships (e.g., which variants affect gene regulation or protein function). Generating a library of mutants that uniformly represents all possible nucleotide or amino acid substitutions should reveal the distribution of effect sizes of mutations within the space of all possible sequence variants, but doing this efficiently remains challenging [69]. Cells can be randomly mutagenised by chemical and physical agents (e.g., alkylating agents or ionising radiation), which are nonspecific and can introduce substitutions at single bases, or small insertions or deletions. Random mutagenesis has been commonly used in traditional genetic screens. On the other hand, site-directed mutagenesis was developed to alter DNA in a non-random, targeted fashion by synthesising a short DNA primer which contains the desired mutation and can hybridise with the DNA in the gene of interest, using phage- or polymerase chain reaction (PCR)-mediated methods [70]. In recent years, a panoply of new methods have been developed that use engineered nucleases for genome editing with greater efficiency and precision, e.g., CRISPR/Cas9 [71–73].

In addition, recombination can enable access to distant parts of genotype space, and it can be used to randomise entire genomes of organisms with sufficient sequence homology. *In vitro* recombination methods such as DNA shuffling have been very successful at reassorting mutations in individual proteins to access beneficial combinations of mutants [74], but they can only sample local regions of a fitness landscape. To study the global picture, *in vivo* recombination is an approach to generate libraries of genetically and phenotypically distinct segregants by crossing divergent organisms that maintain sequence homology [75, 76]. Knowledge of phylogenetic information is thus essential to generate a diverse, genome-wide library that samples a functionally enriched space of the fitness landscape [77].

1.3.3 Screening and selection strategies

How can we measure the distribution of effect sizes on fitness within the space of sequence variants? Functional analysis of variants in regulatory elements and proteins encoded in the genome provide details into sequence-function relationships [78]. Isolating functional variants by selection has long been used in the study of single-gene function, making it possible to characterise the local neighbourhood of a given point in genome space. Genetic screens on single-gene evolution are typically carried out directly on gene products *in vitro*, and aim to determine the spectrum of effect sizes of individual protein residues [79]. To apply

functional selections directly to populations of molecules, genotype and phenotype must be linked [77]. For example, selection for binding affinity can be achieved by capturing protein library members with desired binding properties to an immobilised target [80]. An alternative approach that can only be used for proteins with enzymatic activity links protein activity to organismal survival as a basis for selection [81–83]. This method has successfully been used to evolve enzymes that neutralise or export antibiotics [84], or increase viral infectivity [85, 86]. New paradigms, such as continuous evolution, are going beyond the generation and screening of mutant libraries in a single cycle, and repeatedly performing evolution cycles of mutation and selection without manual intervention [87, 88]. Continuous evolution can markedly increase the number of steps in the sequence space that can be explored in the search for optimal variants.

However, at the genomic scale, carrying out saturation mutagenesis of the genotype space or combinatorial reconstruction based on reconstructed intermediates is not easily feasible. The number of possible sequences for a nucleotide sequence of length L will grow exponentially as 4^L. Thus already at the scale of single genes, a complete combinatorial library of each possible 25-mer would contain roughly 10^{15} unique molecules. Constructing and screening a complete library at this scale is therefore out of reach, and it is no surprise that the global structure of fitness landscapes remains elusive. To address this, we would like to map the relationship between genotype and fitness at the level of the genome in an unbiased manner, by screening for successful clones in libraries with vast genetic diversity created by recombination. We will try to address this question of repeatability of evolutionary trajectories under conditions of strong selection and rapid adaptation (e.g., with antimicrobial drugs), which is rarely asked about biological systems since natural populations can hardly be replicated with well-defined initial conditions. Given the sources of randomness inherent to biological systems, it will be necessary to study the outcomes for replicate evolutions of the same population and repetitions of the same experiment. In terms of genotype and phenotype dynamics, it will also be key to identify the relevant degrees of freedom to measure.

1.4 Population genetic inference and modelling

Can we gain an understanding about evolutionary processes from reading the genome? Can we understand molecular evolution as a dynamical process in terms of quantitative principles? These questions have been theoretically addressed in the field of population genetics for the past century. Population genetic theory gives us an intuition for the relative contributions to the diversity in a population of evolutionary forces like mutation, selection, recom-

bination and genetic drift. More generally, population genetics has been able to reconcile the macroevolutionary changes described by Darwinian evolution with mounting evidence from molecular genetics for cumulative microevolutionary effects. The prevailing assumption in population genetic inference has been that beneficial mutations occur seldom enough that they can be modelled independently. This picture – developed by Fisher, Haldane, Wright and others [89, 90] – is the foundation of our current inference methods for genomic data. In conventional statistical inference, we normally record many independent data points to sample a space of low dimensions, and analytical formulations for the inference using all available information are often possible. With population genetic data, typically we only observe a single draw from the evolutionary process in a high-dimensional space. Moreover, analytical formulations of the inference usually cannot be derived from all available data.

Population genetic inference was initially motivated by the first techniques of protein sequencing which enabled biologists to glimpse the extent of protein-level variation [91, 92], spurring the development of the neutral [93, 94] and quasi-neutral theories of molecular evolution [95]. Later, the introduction of DNA sequencing technology allowed for inferences to be drawn from nucleotide-level variation [96]. These theories simulate the evolution of haplotypes forward in time and have enabled the pursuit of two related goals: (i) to describe the distribution of genetic diversity that enables comparisons between populations, and (ii) to infer how current genetic diversity evolved. Another class of methods premised on the neutral assumption, known as coalescent theory, aims to describe backward in time how variants sampled from a population may have originated from a common ancestor [97, 98]. Coalescent theory focuses on genealogical descriptions of the ancestry of a certain haplotype. Based on these methods, we can begin to reconstruct events that happened in the past, for example, estimating when two organisms diverged and the rate at which they did so [99, 100]. The action of selection can also be inferred according to the neutral theory: site-specific amino acid preferences can be learned from sequencing data by comparing patterns of synonymous and non-synonymous mutations at the codon level [96, 101, 102]. Other selection tests are based on the frequency of variant sites, since selective forces cause appreciable changes in allele frequencies over many generations [103, 104].

All inference considered so far assumes that all polymorphic sites in the population are the result of a single segregating mutation. This framework in population genetics generally holds true in small populations, where the fixation or extinction of a mutation is on its own merits. Nevertheless, with the availability of DNA sequences we are now learning that rare, large-effect mutations are the exception rather than the rule [105]. Multiple

mutations are normally present in large populations, so that the fate of each mutation depends on every other mutation in the genome [106]. This is particularly true for rapidly adapting populations, such as microbial populations, immune cells, or cancer cells. Models must therefore account for the fact that population dynamics will depend on many interacting components of the genome and is thus a complex many-body problem. Based on this observation, mathematical models of microbial evolution from population genetics and epidemiological theory are starting to describe within- and between-host evolution of bacterial and viral pathogens [107–109], as well as the emergence of drug-resistance mutations [46]. Carcinogenesis has also been regarded as an evolutionary process driven by stepwise somatic mutations and clonal expansions [30] involving as many as ten new traits [31]. Although tumour evolution is commonly understood as a sequential process of activation of oncogenes, which stimulate cell proliferation, and deactivation of tumour suppressor genes, which limit proliferation, recent evidence suggests that most of these events are often not successive but simultaneous [33, 110]. With the availability of sequence data of cancer cells, population genetic models of asexual evolution have been applied to tumour initiation [111], progression [112] and drug resistance [113, 114]. These models of evolutionary dynamics can be deterministic or stochastic, and assume either well-mixed or structured populations [115]. Plenty of work remains to be done in this area, where a solid foundation for these models should contain a general description of the dynamics of natural selection in genetically diverse populations that replaces the neutral theory as the null model [106].

Rapid evolution and adaptation pose major challenges to public health on many fronts, and being able to forecast the course of evolution is critical to anticipate the emergence of new microbial pathogens, improve our response to pandemics and forestall antimicrobial and chemotherapy resistance [46, 116]. Population genetic inferences are already helping us to retrospectively understand how evolution has shaped the genomes of present-day organisms, for example by revealing genetic mechanisms for antibiotic production and resistance that already existed before there was multicellular life [117], or by explaining why penicillin-resistant dysentery existed 10 years before the discovery of penicillin [118]. The evolutionary principles that generate these (and not other) genomes constrain the space of possibilities in a practical sense, which may help us predict the evolution of a population forward in time [119]. Evolutionary forecasting has already been demonstrated to work in certain contexts, such as predicting the evolution of the seasonal influenza virus within a time horizon of several months, up to a year [28, 120]. To formulate theories that can be applied to microbial evolution, cancer evolution or drug resistance, the challenge before us is to take the study of evolution from its largely retrospective and qualitative state to a field

with understanding built up from observations, controlled experiments, phenomenology, and quantitative theory [121].

1.5 Thesis outline

As an orientation for the reader, we provide a brief thesis roadmap. We begin, in **Chapter 2**, by giving a first-principles introduction to minimal models of evolutionary dynamics. We present the challenge of discerning 'driver' mutations under selection from hitchhiking 'passengers'. From the perspective of theoretical population genetics, a problem that must be addressed to make headway is formalising how the complex dynamics of the system at large emerge from the laws describing individual mutations and their interactions. We show using a minimal multi-locus model of sequence evolution that we can discern scenarios where changes in allele frequency are caused by natural selection at specified (or at linked) loci. This inference method utilises time-resolved polymorphism data to obtain maximum-likelihood estimates of the locus under selection and its selective advantage.

In **Chapter 3**, we lay the foundation to characterise populations that consist of multiple, genotypically distinct cell populations by genome sequencing. We formulate this as an inverse problem where one would like to computationally reconstruct the genomic structure of the population from the observed set of sequence reads. We introduce Hidden Markov Models (HMMs) and show how they can be used to model sequences of a mixed population of related cells. We show how this approach is powerful to reconstruct subclonal copy-number profiles, genotypes and population frequencies, and how it can be applied to clonal admixtures of genomes in any asexual population, from evolving pathogens to the somatic evolution of tumours.

In **Chapter 4**, we present an experimental test for different sequence ensembles with minimal-to-maximal genetic diversity in a long-term selection experiment with budding yeast (*S. cerevisiae*). The objective is to characterise a complex fitness landscape carrying out unbiased sampling of the genotype space with an ensemble of genetically homogeneous and heterogeneous populations, observing their evolutionary trajectories by DNA sequencing as an approach to deduce the selective constraints to adaptation. The operating principle is that biologically relevant states of genomes will be enriched in the mutation neighbourhood accessible around the local fitness peaks. Thus, the kinds of selective constraints we impose on growth-limiting processes (e.g. with antimicrobial drugs) should bias and expose these states.

In **Chapter 5**, we test how genetic variation can affect the fate of new mutations. The broad goal is to record time histories of mutations throughout a population by DNA sequencing as an approach to deduce the effect of extant genetic backgrounds on new mutations. Coupled with time-resolved phenotyping, this experiment permits both the detection of functionally relevant mutations that rapidly confer antimicrobial resistance, and the dynamic observation of adapting populations with appropriate temporal resolution. Here, we show that selection can indeed drive changes to the fitness distribution within a population, that we can observe the mutations affecting the fitness distribution, and that we can decompose the contribution of pre-existing and *de novo* genetic variation to fitness.

In **Chapter 6**, we present the analysis of mutational processes under selective constraints which affect the fidelity of genome replication. We characterise the spectrum of mutational processes that are active in the experiment presented in Chapter 4. We apply an expectation-maximisation (EM) algorithm to identify distinct mutational processes that are caused by endogenous and exogenous DNA damage.

This thesis contains work that has been reported or will appear in the following peer-reviewed publications and pre-prints:

- I. Vázquez-García, F. Salinas, J. Li, A. Fischer, B. Barré, J. Hallin, A. Bergström, E. Alonso-Perez, J. Warringer, V. Mustonen*, and G. Liti*, Clonal heterogeneity influences the fate of new adaptive mutations, *Cell Reports* **21**, no. 3 (2017), pp. 732–744. [*equal contribution]
- I. Vázquez-García, V. Mustonen, and G. Liti, Principles of systems biology, no. 22, *Cell Systems* **5**, no. 4 (2017), pp. 305–309.
- J. Li, I. Vázquez-García, K. Persson, A. González-Seviné, J.-X. Yue, B. Barré, M. N. Hall, A. D. Long, J. Warringer, V. Mustonen, and G. Liti, Patterns of selection reveal shared mutational targets over short and long evolutionary timescales, *bioRxiv* **229419** (2017), *under review*.
- A. Fischer, I. Vázquez-García, and V. Mustonen, The value of monitoring to control evolving populations, *Proc. Natl. Acad. Sci. U.S.A.* **112**, no. 4 (2015), pp. 1007–1012.
- A. Fischer, I. Vázquez-García, C. J. Illingworth, and V. Mustonen, High-definition reconstruction of clonal composition in cancer, *Cell Reports* **7**, no. 5 (2014), pp. 1740–1752.
- I. Vázquez-García, E. Alonso-Perez, J. Li, J. Hallin, M. C. Reis, G. Liti, J. Warringer, and V. Mustonen, Clonal diversity accelerates the evolution of antimicrobial resistance, *in preparation*.
- I. Vázquez-García, F. Puddu, S. P. Jackson, and V. Mustonen, Mutational processes caused by genome replication stress, *in preparation*.

Part of this thesis has been devoted to apply these concepts and methods to the Pan-Cancer Analysis of Whole Genomes (PCAWG) project, as part of the International Cancer

Genome Consortium (ICGC).[1] This consortium has generated a dataset which comprises whole-genome sequencing of tumour-normal pairs from 2,663 donors and covers 38 cancer types [122]. In this context, we have studied mutational processes in the cancer genome, related to work reported in Chapter 6. We have also applied methods described in Chapter 3 to study selection and subclonal heterogeneity on a pan-cancer scale. Finally, we have taken part in a benchmark exercise for subclonal reconstruction methods,[2] and we use these datasets as proof-of-principle in Chapter 3.

- P. Campbell, G. Getz, J. M. Stuart, J. O. Korbel, and L. Stein for the ICGC PCAWG Initiative, Pan-cancer analysis of whole genomes, *bioRxiv* **162784** (2017), *under review*.
- M. Gerstung*, C. Jolly*, I. Leshchiner*, S. C. Dentro*, [10 auth.], I. Vázquez-García, [27 auth.], P. T. Spellman*, D. C. Wedge*, P. Van Loo* for the Evolution and Heterogeneity Working Group – ICGC PCAWG Initiative, The evolutionary history of 2,658 cancers, *bioRxiv* **161562** (2017), *under review*. [*equal contribution]
- S. C. Dentro*, I. Leshchiner*, K. Haase*, M. Tarabichi*, J. Wintersinger*, A. Deshwar*, K. Yu*, Y. Rubanova*, G. Mcintyre*, I. Vázquez-García, [28 auth.], W. Wang*, Q. D. Morris*, D. C. Wedge*, P. Van Loo* for the Evolution and Heterogeneity Working Group – ICGC PCAWG Initiative, Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types, *bioRxiv* **312041** (2018), *under review*. [*equal contribution]
- A. Salcedo*, M. Tarabichi*, S. M. Espiritu*, A. G. Deshwar*, [17 auth.], DREAM SMC-Het Participants, [5 auth.], K. Ellrott*, D. C. Wedge*, Q. D. Morris*, P. Van Loo*, P. C. Boutros*. Creating standards for evaluating tumour subclonal reconstruction, *bioRxiv* **310425** (2018), *under review*. [*equal contribution]

[1] ICGC Pan-Cancer Analysis of Whole Genomes (PCAWG) project [https://dcc.icgc.org].

[2] ICGC-TCGA-DREAM Somatic Mutation Calling Challenge for Tumour Heterogeneity (SMC-Het) [https://www.synapse.org/#!Synapse:syn2813581].

Chapter 2

Minimal models of evolutionary dynamics

2.1 Introduction

The interplay of deterministic and stochastic evolutionary forces on population dynamics can be understood mathematically. In this chapter, we will first introduce a minimal model with mutation, selection and genetic drift. Although these are not the only evolutionary forces, the three of them suffice to gain a heuristic understanding of the probability of fixation or extinction of a mutation. We work up to it with a set of reduced models, starting with the 'single-locus' model (Section 2.3). Section 2.4 argues that the main feature neglected by the single-locus model is linkage between multiple sites in the genome, which is key to describe the dynamics of rapidly adapting populations. To redress this shortcoming, Section 2.4 introduces a 'multi-locus' model, which we extend to formulate a deterministic description of driver-passenger dynamics. We use this model to discern drivers from hitchhiking passengers and to estimate their fitness effect.[1] All throughout this chapter, we will develop some intuitions for the rate at which a population can adapt or decline; and for which mutations will fix and which will go extinct. This will help us determine what should or should not surprise us and what ought to be able to happen in a given evolutionary timescale.

This work was carried out in collaboration with V. Mustonen (V.M.) at the Wellcome Trust Sanger Institute (Cambridge, UK).[2]

[1] The computational methods reported in this chapter are available from the GitHub code repository [https://github.com/ivazquez/PhD-thesis/tree/master/Chapter2].

[2] I.V.-G. and V.M. formulated the minimal model of driver-passenger dynamics; I.V.-G. implemented the simulations; I.V.-G. and V.M. interpreted the results.

2.2 Population genetics of rapid adaptation

We first focus on relating the state of the genome to 'quantitative traits', which are phenotypic characteristics of individuals. We will focus on fitness as a general quantitative trait which describes the reproductive success of an individual, but our exposition is directly applicable to any quantitative phenotype. Suppose that the state of the genome of each individual – its genotype – can be represented by a vector $\boldsymbol{g} = (g_1, g_2, \dots, g_L)$ of length L, where $g_i \in \{0, 1\}$. We can define a function of the genotype $\boldsymbol{g}$, which may describe a quantity like fitness, or any other quantitative trait. A fitness landscape is a useful metaphor for a map from the high-dimensional space of genotypes to a low-dimensional space of reproductive success. Following the notation by Neher and Shraiman [123], we can decompose the fitness function $F(\boldsymbol{g})$ by summing the contributions from single sites in the genome, pairs of sites and higher-order terms,

$$F(\boldsymbol{g}) = \langle F \rangle + \sum_{i=1}^{L} \sigma_i g_i + \sum_{i<j}^{L} \sigma_{ij} g_i g_j + \sum_{i<j<k}^{L} \sigma_{ijk} g_i g_j g_k + \dots \tag{2.1}$$

We will focus on the simplest models of evolution that are solely additive, i.e., they only include contributions up to the first term of the sum. Equation (2.1) is then simplified as an additive fitness function:

$$F(\boldsymbol{g}) = \langle F \rangle + \sum_{i=1}^{L} \sigma_i g_i \tag{2.2}$$

A mutation conferring a fitness difference $\Delta F_g = F_g - \langle F \rangle$ with respect to the mean fitness of the population will be subject to selection. Therefore, the genetic contributions of a mutation to fitness will set the timescales of evolutionary changes in the genome. As shown in Figure 2.1, strong-effect mutations will confer a large selective advantage or disadvantage and will behave largely deterministically, dominated by selection. Such genotype changes can be studied experimentally (e.g., in a genetic screen). On the other hand, weak-effect mutations will be dominated by drift which can only be observed in timescales of the order $\mathcal{O}(N)$ generations. In this thesis, we will focus on strong-effect mutations that are directly accessible experimentally.

Most often we can only observe the distribution of quantitative traits $P(F_g, t)$ (or fitness distribution) of the population at time t, not the distribution of genotypes $P(\boldsymbol{g}, t)$. Changes to the fitness distribution can be described by considering the probability of finding an individ-

ual with fitness in the interval $[F, F + \delta F]$. By projecting from an individual's genotype to the related fitness interval, it can be easily shown that the dynamics of the fitness distribution are given by the rate of change in fitness of genotype $\boldsymbol{g}$ with respect to the average fitness of the population [123]. In the simplest case with selection but no mutation or recombination, this is described by

$$\frac{\partial}{\partial t} P(F_g, t) = \left(F_g - \langle F \rangle\right) P(F_g, t) \tag{2.3}$$

Integrating over the fitness F gives us that the change in the average fitness (or rate of adaptation) is directly proportional to the fitness variance, first derived by Fisher [89]

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle F \rangle = \left\langle \left(F_g - \langle F \rangle\right)^2 \right\rangle = \sigma_F^2 \tag{2.4}$$

where σ_F^2 is the fitness variance of the population. Equation (2.3) has a Gaussian travelling wave solution, indicating that the abundance of fitter individuals will rise since they will grow more rapidly, and the less fit individuals will go extinct [124, 125]. As a result, the mean of this fitness distribution $\langle F \rangle$ shifts and the variance σ_F^2 decreases. This is commonly known as Fisher's fundamental theorem of natural selection [89], which describes the asexual case with selection. To prevent adaptation from stalling during asexual evolution, new variation has to be constantly introduced by mutation, for which case it was recently shown that the

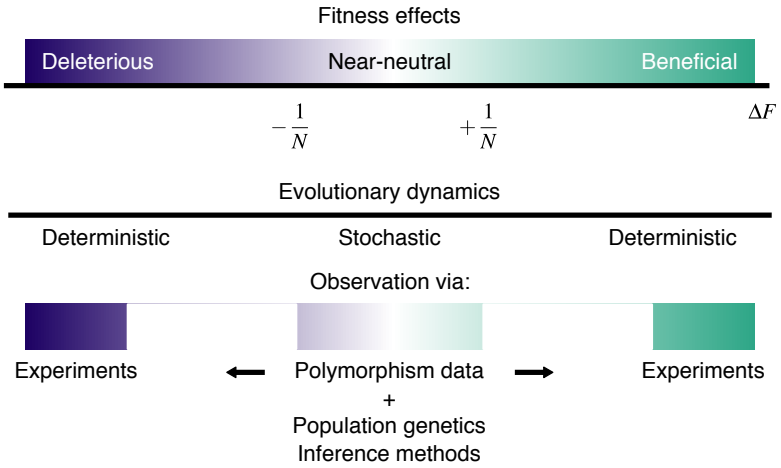


Fig. 2.1 The fitness effects of mutations, ΔF, set the timescale for selection. The dynamics of strongly beneficial or deleterious alleles are typically dominated by the deterministic forces of selection, which can be measured experimentally over short timescales. Mutations with weak or neutral fitness effects of order $\mathcal{O}\left(\pm\frac{1}{N}\right)$ are dominated by genetic drift and can be studied using genetic polymorphism data over long timescales.

fitness distribution will be driven by the stochastic tip of new mutations [125]. This theorem can also be extended to the sexual case [123], which demonstrates that recombination limits the effects of selection to the additive component of the fitness variance, as non-additive fitness components are lost due to the disruption of beneficial combinations of alleles which are reshuffled by recombination. Note that we also assume that the fitness function $F(\boldsymbol{g})$ remains constant over time. This formalism has also been generalised elsewhere to account for time-dependent fitness landscapes (see Mustonen and Lässig [126]).

2.3 Single-locus dynamics

From one generation to the next, different evolutionary forces modify the distribution of genotypes $P(\boldsymbol{g}, t)$ in the population. To bridge between the dynamics of the genotype distribution and the coarse-grained dynamics of traits like fitness, we would like to start with the most basic evolutionary model that includes the effects of mutation, selection and genetic drift.[1] This is commonly known as a *single-locus* model, since it approximates the evolution of a genome of length $L = 1$. This model assumes a haploid population of finite size with N individuals, each of which has a genotype g that can have one of several states (or alleles).

The two best known models of stochastic evolutionary dynamics that describe the changes to the distribution of genotypes $P(g, t)$ are the Wright-Fisher and Moran models [89, 127]. The Moran model, which describes a birth-death process for a population of infinite size, is analytically tractable but difficult to simulate. On the other hand, the Wright-Fisher model describes a population of constant size N as a discrete Markov process that is evolved by sampling the previous generation with replacement. We will focus on the Wright-Fisher model but we should not worry about the specific details of each, as the two can be shown to be equivalent if they are recasted using a system-size expansion (see Van Kampen [128] and Gardiner [129]).

Suppose we have n_g individuals with genotype g in a population of size $N = \sum_g n_g$. The number of individuals n_g is fully specified by $P(n_g, t)$ at each time t. In other words, $P(n_g, t)$ is a probability distribution which describes the changes to the number of individuals over

[1] The terms *diffusion* and *drift* are used to describe two opposite forces in population genetics and in the physics literature on stochastic processes. To avoid any confusion, we make use of the population genetics convention: drift refers to stochastic fluctuations in offspring number that follow an unbiased random walk in the space of mutation frequencies.

time. If time is measured in discrete generations, the master equation is then

$$\frac{\partial}{\partial t} P(n_g, t) = \sum_{n'_g} W\left(n_g \mid n'_g\right) P(n'_g, t) - W\left(n'_g \mid n_g\right) P(n_g, t) \tag{2.5}$$

where $n_g = 1, \ldots, N$. The first term on the right hand side is the probability to transition from n_g individuals with genotype g in the previous generation to n'_g individuals in the current generation, and the second term is the probability to go from n'_g in the current generation to n_g in the previous generation. If there are any changes to the genotype of individuals between generations s.t. $n_g \to n'_g$, then for the total population size N to remain constant there must be an equivalent flux between $n'_g \to n_g$.

For Wright-Fisher evolution, the distribution of genotypes in the current generation is going to depend on the distribution of genotypes in the previous generation. Therefore, the transition probability between n_g and n'_g is going to depend on the number of ways N individuals can be partitioned into sets of genotypes. For any partition, the probability of drawing a number of individuals n_g with a certain genotype will depend on the frequency of that genotype in the previous generation, $x_g = \frac{n_g}{N}$. For simplicity, we will limit ourselves to the single-locus, two-allele case, so that the genotype can only correspond to one of two alleles ($g = 0$ and $g = 1$). The transition matrix W can then be expressed as the probability under binomial sampling to draw $n'_g = N x'_g$ individuals with genotype g in the current generation from the previous generation,

$$W\left(n'_g \mid n_g\right) = \binom{N}{n'_g} \left[x_g + \frac{1}{N} f(x_g)\right]^{n'_g} \left[1 - x_g - \frac{1}{N} f(x_g)\right]^{N - n'_g}. \tag{2.6}$$

The basic formulation with $f(x_g) = 0$ only describes the process of neutral evolution under random genetic drift, with binomial resampling of the population in each generation out of a population of fixed size $N = n_0 + n_1$. The function $f(x_g)$ can incorporate the action of other evolutionary forces like mutation or selection. Mutations can be incorporated in $f(x_g)$ by introducing a probability that an individual with genotype $g = 0$ gives rise to an offspring with genotype $g = 1$ at a rate μ (and vice versa). The effects of selection can be incorporated into the Wright-Fisher model by changing the transition probabilities $W\left(n'_g \mid n_g\right)$ with a weighted function $f(x_g) \propto e^{F_g - \langle F \rangle}$, such that frequency changes will be biased by the fitness differential. Instead of each new generation being created by random sampling of alleles from the previous generation, we use weighted sampling of alleles from the previous generation to reflect the selective differences between alleles.

We can then rewrite the transition matrix by defining the frequency of either of the two alleles in the population, $x_0 = \frac{n_0}{N}$ or $x_1 = \frac{n_1}{N}$. It suffices to keep track of the frequency x of one of the two alleles (e.g., $g = 1$), as the sum of the number of individuals with each allele is constrained by the total population size, s.t. $x_1 = 1 - x_0$. The changes in frequency δx between consecutive generations cannot be very large, so that the transition matrix can be recasted as $W\left(n \,\middle|\, n'\right) \to W\left(x + \delta x \,\middle|\, x'\right)$. With this change of variables, it can be shown that this transition probability is a Gaussian [123, 130], such that the first moment of this distribution is given by the mean number $\langle n' \rangle = Nx$. The second moment is the variance $\langle n'^2 \rangle - \langle n' \rangle^2 = Nx(1-x)$, so the noise will scale as $\frac{x(1-x)}{N}$ and will vanish in an infinite population as $N \to \infty$.

From the master equation in Equation (2.5), we are interested in the large-N limit in order to understand the qualitatively different behaviours of the system. In this limit, we arrive at a diffusion approximation of the master equation [131]:

$$\frac{\partial}{\partial t} P(x,\, t) = \Big[\underbrace{\frac{\partial}{\partial x} \mu\,(1-2x)}_{\text{mutation}} + \underbrace{\frac{\partial}{\partial x} \sigma x(1-x)}_{\text{selection}} + \underbrace{\frac{1}{2N} \frac{\partial^2}{\partial x^2} x(1-x)}_{\text{genetic drift}} \Big] P(x,\, t) \tag{2.7}$$

where $P(x,\, t)$ is the probability of finding n copies of genotype $g = 1$ at time t. The first term describes the influx of new mutations at rate μ. The second term is the deterministic effect of selection shifting the probability distribution according to the selection coefficient σ. The third term accounts for number fluctuations in reproduction between generations and is often referred to as genetic drift. The fact that fluctuations scale as $\frac{1}{\sqrt{N}}$ indicates that the smaller the population, the more pronounced the role of fluctuations is. In the physics literature, this is commonly known as the Fokker-Planck equation, and it is the continuum limit of well-known classes of models like the Wright-Fisher model or the Moran model. There is no closed-form solution for the time evolution of $P(x,\, t)$, but we can find results for the equilibrium distribution.

Ignoring the mutation terms in Equation (2.7), we focus on the selection and genetic drift terms which perform a biased random walk. In the limit of $t \to \infty$, Kimura showed that the probability of fixation is [132],

$$P_{\text{fix}} = \frac{1 - e^{-2N\sigma x}}{1 - e^{-2N\sigma}}. \tag{2.8}$$

It becomes clear then that the fixation probability P_{fix} increases as a function of $N\sigma$ and will tend to 0 or 1 at frequencies $x = 0$ and $x = 1$, respectively.

2.3.1 Dynamics of neutral mutations

Now we can study the dynamics of the probability distribution of genotypes. Suppose that at some initial time we have a sharply peaked distribution $P(x, 0) = \delta(x - x_0)$, where $\delta(x)$ is the Dirac delta function. What is the probability of finding that mutation at frequency x later? The frequency $x(t)$ performs a random walk due to the fluctuations in offspring number caused by genetic drift. We notice that the genetic drift term vanishes when $x = 0$ or $x = 1$. This just tells us that, in the absence of new mutations, the mutation will be either lost ($x = 0$) or it will have taken over the entire population ($x = 1$). Both of these are absorbing states, so once the mutation is lost it cannot be regained. If we start with a sharp distribution of alleles at frequency x, either the wild-type or the mutant allele will eventually disappear from the population. So the asymptotic states have to be either extinct or fixed, and the probability associated with both of these cases is described by the Kimura formula in Equation (2.8).

2.3.2 Dynamics of mutations under selection

How do the effects of selection alter the prospects of a mutation? Suppose that mutants with genotype $g = 1$ have a factor $1 + \sigma$ more offspring per generation than the wild-type, with genotype $g = 0$. For simplicity, we will assume that all mutations have the same fitness effect σ. In the $N \to \infty$ limit, we drop the genetic drift term completely for now (i.e., make a deterministic approximation). The frequency x of genotype g will then evolve as

$$\frac{dx}{dt} = \sigma x(1 - x)\,. \tag{2.9}$$

If we solve this ordinary differential equation for a time interval Δt, the solution is a logistic function

$$x(t + \Delta t) = \frac{x(t)\, e^{\sigma \Delta t}}{1 - x(t) + x(t)\, e^{\sigma \Delta t}}, \tag{2.10}$$

Therefore, it takes roughly $\tau = \frac{1}{\sigma}$ generations for selection to change the frequency of a mutation. This is effectively the only regime that can be accessed to study strong-effect mutations by current-day sequencing technologies, which can typically resolve a mutation frequency down to $x = 0.01$, still in the deterministic regime.

Nevertheless, number fluctuations can still dramatically alter the evolutionary dynamics in the deterministic scenario. As with the neutral case, either allele 0 or 1 will eventually become fixed in this model. In the limit of $t \to \infty$, we already mentioned that the probability

of fixation P_{fix} goes to either 0 or 1 at the boundaries according to Equation (2.8). We can divide the solutions into three regimes, with the following piecewise form given by Durrett [133]:

(i) While the beneficial allele 1 is rare, the number of 1's can be approximated to take τ_1 generations to establish:

$$\tau_1 = \frac{1}{\sigma} \tag{2.11}$$

To fix, the allele needs to reach size N, and below frequency $x = \frac{1}{\sigma}$ it is just neutral.

(ii) While the frequency of allele 1 is in the range $x \in \left\{\frac{1}{N}, 1 - \frac{1}{N}\right\}$, there is very little randomness and it follows the logistic solution of the differential equation we showed above in Equation (2.10). It will then take

$$\tau_2 = \frac{1}{\sigma} \log(N\sigma) \tag{2.12}$$

generations for the favourable allele 1 to transit from its entry into the population ($x = \frac{1}{N}$) to near fixation ($x = 1 - \frac{1}{N}$).

(iii) While the beneficial allele 1 is nearing fixation – or, equally – the deleterious allele 0 is rare, the number of 0's can be approximated to take τ_3 generations to go extinct and consequently for the allele 1 to fix:

$$\tau_3 = \frac{1}{\sigma} \tag{2.13}$$

Adding up all the contributions, a mutation under selection will normally take $\tau_{\text{fix}} \sim \frac{1}{\sigma} \ln(N\sigma)$ generations to fix in the population.

If mutations are rare – suppose that a mutation occurs in a single individual – we will first wait $\frac{1}{N\mu}$ generations for the mutations to happen, and then most of the time the mutation will go to extinct. The probability that the mutation will not go extinct is proportional to the selective advantage of the mutation, σ. The relative strength of selection and genetic drift will therefore depend on the typical fixation time of mutations by drift (N) and the typical timescale of selection $\left(\frac{1}{\sigma}\right)$. All in all, we can summarise the expected fixation probabilities

we have discussed for the cases of beneficial and deleterious mutations:

$$\text{Beneficial mutations:} \quad P_{\text{fix}}(\sigma, x) \sim \begin{cases} 1 & \text{for} \quad x \gtrsim \frac{1}{N\sigma} \\ N\sigma & \text{for} \quad x \lesssim \frac{1}{N\sigma} \end{cases}$$

$$\text{Deleterious mutations:} \quad P_{\text{fix}}(\sigma, x) \sim \begin{cases} 0 & \text{for} \quad 1-x \gtrsim \frac{1}{N\sigma} \\ N\sigma(1-x) & \text{for} \quad 1-x \lesssim \frac{1}{N\sigma} \end{cases}$$

Finally, in the single-locus scenario one can extend this analysis to accelerate or delay the loss of a mutation in a population. In a published work outside the scope of this thesis [114], we have introduced a control u in the selection term, s.t. $\sigma \to \sigma + u$. This control can be modified to keep a finite population polymorphic under Wright-Fisher evolution by influencing the selective difference between two alleles. We have derived the optimal strategy for the didactic example of a control task of maintaining an initial polymorphism in the frequency range $0 < x < 1$ for as long as possible by linearly changing the selection coefficient instantaneously, in response to and as a function of $x(t)$. We refer the reader to this study for further details [114].

2.4 Multi-locus dynamics

So far, we discussed the single-locus model as an instructive scenario with many exact analytical solutions. However, genomes almost always contain more than one variable locus so we must account for the presence of other mutations. Rather than deriving a generalisation of the single-locus dynamics for L loci, we focus on the simplest case for a *multi-locus* model of the genome. The $L = 2$ case is more complex than the $L = 1$ case. Mathematically, the frequency $\boldsymbol{x}$ itself is a vector and we now have three degrees of freedom instead of one. We will first restrict our attention to asexual evolution and then come back to address the effects of recombination during sexual evolution.

We consider the simplest possible model involving multiple loci: a two-locus model with two possible alleles at each locus, in which the driver mutation at locus i is significantly more advantageous than any passenger at locus j, i.e., $\sigma_i \gg \sigma_j$ for all j. We refer to the two alleles at one locus as $a \in \{0, 1\}$, and the alleles at the other locus as $b \in \{0, 1\}$. There are four quantities of interest, namely the frequencies of these four combinations which must add to 1. Therefore, only three of these quantities are independent: the frequency x_i^1 of allele 1_a (with the frequency x_i^0 of allele 0_a given by $x_i^0 = 1 - x_i^1$), the frequency x_j^1 of allele 1_b (with

$x_j^0 = 1 - x_j^1$), and the correlations between pairs of loci, $C_{ij} = \left\langle x_i x_j \right\rangle - \langle x_i \rangle \left\langle x_j \right\rangle$, also known as 'linkage disequilibrium'.

According to our model, the dynamics of passenger mutations are then fully specified by the strength of the local driver, with no need to specify the exact pairwise interactions between every possible pair. The effect of the driver locus on all passenger loci j is therefore given by:

$$x_j^b(t) = \sum_{a \in \{0,1\}} x_i^a(t) \frac{x_{ij}^{ab}(t_0)}{x_i^a(t_0)} \quad \text{for} \quad j \neq i \tag{2.14}$$

where the two-locus haplotype frequency is $x_{ij}^{ab}(t_0) = x_i^a(t_0) x_j^b(t_0) + (-1)^{a+b} C_{ij}$.

We will now relax the condition that mutations are physically linked in the genome. In addition to the basic elements we already saw, recombination will be an additional evolutionary force that becomes especially important when there are many variable loci in a population. To incorporate recombination, we must account for the pairwise linkage structure in a population. We define r as the probability of recombination between two loci, or equivalently $r = \rho \Delta_{ij}$ as a function of the distance Δ_{ij} between them in base pairs (bp) and the recombination rate ρ in units of $\text{bp}^{-1}\text{gen}^{-1}$. Adding a superscript to indicate the generation number, the two-locus haplotype frequency evolves as

$$x_{ij}^g = (1 - \rho \Delta_{ij}) x_{ij}^{g-1} + \rho \Delta_{ij} x_i^{g-1} x_j^{g-1} \tag{2.15}$$

During asexual evolution, when a driver mutation arises on a locus i it will be linked to all other passenger loci j in the genome as part of the same haplotype. Over several generations of sexual recombination, the frequency of the driver mutation may increase but recombination will introduce this allele into other haplotypes. As a result, linkage disequilibrium C_{ij} will decay as a function of time [134]. We can parameterise C_{ij} in terms of the recombination rate, so for N_c generations of sexual recombination we can track the decay of linkage disequilibrium by relating the current value $C_{ij}(t)$ to the initial value $C_{ij}(t_0)$, s.t. $C_{ij}(t) = (1 - \rho)^{N_c} C_{ij}(t_0)$.

2.5 Inference of selection from sequence data

Suppose that we sample a population history at multiple sites in the genome by sequencing, which is given by a vector $\boldsymbol{n}_i = \left(n_1, n_2, \ldots, n_L\right)$ with loci defined at $i \in \{1 < \ldots < L\}$. Tak-

ing measurements at several time points, $t \in \{t_0 < \ldots < T\}$, we find that $n_i(t)$ individuals carry allele 1 at locus i at time t, out of a total of $N_i(t)$ individuals in the population. Thus, we observe a time series of allele frequencies $\boldsymbol{x}_i(t) = \left(\frac{n_1}{N_1}, \frac{n_2}{N_2}, \ldots, \frac{n_L}{N_L}\right)_t$. For example, for locus $i = 7$ in this time series we may observe changes in allele frequency over time: $x_7(t_0) = \frac{17}{40}$, $x_7(t_1) = \frac{21}{37}$, $x_7(t_2) = \frac{29}{44}$, and so on. At time t_0 we detect 17 out of 40 reads reporting allele 1 and the remaining 23 reads report allele 0, and the prevalence of allele 1 increases thereafter. Can temporal fluctuations in the observed frequency of this allele be explained by sampling noise, or did they arise due to selection either at a specified or at another, fully linked, locus? Under the neutral hypothesis, changes in allele frequency are caused only by genetic drift and sampling, i.e., the selection coefficient σ acting on allele 1 is fixed to be zero. Under the alternative hypothesis, our model explains the frequency trajectories with a non-zero selection coefficient σ.

2.5.1 Maximum likelihood estimation

Given we have observed a population history $\boldsymbol{n}_i(t)$, we can model a sequence of draws of the mutant allele at locus i according to a binomial distribution at each time point t. Then, we can define a probability for the time series between time t_0 and T as

$$P\left(\boldsymbol{n}_i \mid \theta\right) = \prod_{t=t_0}^{T} \mathrm{Bin}\left(n_i(t) \,\middle|\, N_i(t), \theta\right) = \prod_{t=t_0}^{T} \binom{N_i(t)}{n_i(t)} \left[1 - x_i(t)\right]^{N_i(t)} x_i^{N_i(t)-n_i(t)}, \tag{2.16}$$

where $N_i(t)$ denotes the total number of draws, i.e., the sequence read depth at locus i at time t.

To obtain the total probability of these observations given an explicit evolutionary model, we consider a state-space model defined in Equations 2.4 and 2.10 with a set of model parameters θ,

$$\log P\left(\boldsymbol{n}_{1:L} \mid \theta\right) = \log P\left(\boldsymbol{n}_i \,\middle|\, \theta_i^{\mathrm{dri.}}\right) + \sum_{j \neq i} \log P\left(\boldsymbol{n}_j \,\middle|\, \theta_i^{\mathrm{dri.}}, \theta_j^{\mathrm{pass.}}\right) \tag{2.17}$$

We showed in the previous section that the set of frequencies of all possible allelic combinations at two loci is well described by the frequency of one of the alleles at both loci and their two-point correlation (i.e., linkage disequilibrium), so that the model is fully specified by $\theta \in \left\{\theta_i^{\mathrm{dri.}}, \theta^{\mathrm{pass.}}\right\}$, where $\theta_i^{\mathrm{dri.}} \in \left\{\sigma_i, \rho_i, x_i^a(t_0)\right\}$ and $\theta_j^{\mathrm{pass.}} \in \left\{x_j^b(t_0)\right\}$.

From this we can write down the log-likelihood $\mathcal{L}$ of model θ given the full trajectory,

$$\begin{aligned}\mathcal{L}(\theta|\boldsymbol{n}_{1:L}) &= \sum_{1\le i<j\le L} \log P\left(\theta_i^{\text{dri.}}, \theta_j^{\text{pass.}} \,\middle|\, \boldsymbol{n}_j\right) \\ &= \log P\left(\theta_i^{\text{dri.}} \,\middle|\, \boldsymbol{n}_i\right) + \sum_{1\le i<j\le L} \log P\left(\theta_i^{\text{dri.}}, \theta_j^{\text{pass.}} \,\middle|\, \boldsymbol{n}_j\right)\end{aligned}$$

Since we only consider one driver mutation i that influences every passenger trajectory j, the log-likelihood has a clean factorisation into two terms. Given the driver trajectory, one can independently maximise likelihoods of the passenger trajectories, which only have their initial conditions and their linkage to the driver as free parameters. We can work with the log-likelihood function which is now a function of four variables, i.e., $\theta = \left\{\sigma_i, \rho, x_i^a(t_0), x_i^b(t_0)\right\}$ (we will refer to this as the 'unconstrained' objective function). Critically, if the recombination rate ρ is known, then we could simply weight the terms in the log-likelihood function by the linkage disequilibrium C_{ij} at each sampled generation. We incorporate this knowledge of the pairwise linkage structure for the case in which the recombination rate is known at t_0. In this 'constrained' case, one can then treat recombination analogously to other fixed parameters, such that the objective function becomes $\theta = \left\{\sigma_i, x_i^a(t_0), x_i^b(t_0)\right\}$.

As a result, the model parameters θ of the model can be found by choosing a value $\hat{\theta}$ that maximises the log-likelihood, i.e., $\hat{\theta} = \text{argmax}_\theta\, \mathcal{L}(\theta|\boldsymbol{n}_{1:L})$. The maximisation is not analytically tractable as the log-likelihood gradient cannot be computed and setting it to zero does not result in a closed-form solution. For a low-dimensional vector, doing an exhaustive search over a region of $\hat{\theta}$ would produce the largest likelihood. However, we do not know if $\hat{\theta}$ lies in a subspace, thus it is not practical in most cases and becomes very computationally expensive for multidimensional optimisation. Instead, local optima can be computed efficiently using the Nelder-Mead simplex algorithm. For this we need the numerical allele frequencies to be smooth functions of the parameters for a given set of realizations of the noise. We therefore use the GSL library to find maximum-likelihood estimates of the model parameters $\hat{\theta}$ using the simplex [135].

2.5.2 Simulation

We use the Wright-Fisher model to create a simulation ensemble that can be used to test the inference method. The expected genotype frequencies are calculated every generation after mutation, selection, and recombination, and then the population is resampled from this genotype distribution. We carry out simulations with a fixed population size of $N = 10^4$

individuals. Every individual in the population has a genotype $\boldsymbol{g} = \{g_1, \ldots, g_L\}$ composed of $L = 10^4$ sites. We outline the steps required to update the population.

The update rule when and individual acquires a mutation at locus i is defined by

$$P(\boldsymbol{g}) \leftarrow \left(1 - \sum_{i=0}^{L-1} \mu\right) P(\boldsymbol{g}) + \sum_{i=0}^{L-1} \mu P^*(\boldsymbol{g}) \tag{2.18}$$

where μ denotes the mutation rate. The first term introduces the loss of mutations to a neighbouring genotype by modifying one locus i at a time. The second term represents the gain of a mutation at locus i, keeping the state of other unmutated loci unchanged. This modified distribution of genotypes with a mutation at locus i is indicated by $P^*(\boldsymbol{g})$.

We assume that there is selection on a single site, which we call a driver. We also assume all other mutations to be neutral and refer to those as passengers. Each generation of N individuals is sampled with replacement from the previous generation, where genotypes at the driver locus i are weighted according to their fitness and sampled with probability

$$P(\boldsymbol{g}) \leftarrow \frac{e^{F(\boldsymbol{g})}}{\frac{1}{N}\sum_{i=1}^{N} e^{F(\boldsymbol{g}_i)}} P(\boldsymbol{g}) \tag{2.19}$$

We will assume that selection is constant across time, i.e., $F(\boldsymbol{g}) = \text{const}$.

We explore the effect of variation in the recombination rate by randomly sampling a set of chromosomes from individuals in the previous generation. Assuming that some individuals undergo mating and recombination (with probability r) and others do not (with probability $1 - r$), the update rule for one recombination event to occur in one generation is

$$P(\boldsymbol{g}) \leftarrow (1 - r)P(\boldsymbol{g}) + rP^*(\boldsymbol{g}) \tag{2.20}$$

where the first term describes the distribution of non-recombined genomes, $P(\boldsymbol{g})$. The second term includes the distribution of recombinants, $P^*(\boldsymbol{g})$, specifying the set of recombined genomes drawn from the previous generation. This includes the contribution from maternal and paternal genomes inherited by the progeny, $P^*(\boldsymbol{g}^m)$ and $P^*(\boldsymbol{g}^p)$. This is weighted by the probability $R(\gamma)$ of a certain inheritance pattern γ specifying which of the loci are maternally or paternally inherited, such that the distribution of recombinants is

$$P^*(\boldsymbol{g}) \equiv \sum_{\gamma} \sum_{g} R(\gamma)\, P^*(\boldsymbol{g}^m)\, P^*(\boldsymbol{g}^p) \tag{2.21}$$

This update step can be generalised to multiple generations by pairing up individuals with this mating scheme to produce recombinant offspring. However, if we were to track the complete genotype distribution $P(\boldsymbol{g})$ we would need to consider all possible pairs of parents and all possible arrangements to combine their genetic information. Although we will not need to track the full distribution, Zanini and Neher [136] have proposed a method for efficient simulation of changes to the genotype distribution with arbitrary recombination patterns in large populations, exploiting redundancies in the recombination step using Fourier decomposition.

In our simulations, to create the founder population we first generated two random genotypes which diverge at $L = 10^4$ loci by using mutation update rules for 200 generations, with a mutation rate $\mu = 1 \times 10^{-10}$ bp^{-1}gen^{-1}. We then applied the recombination update rules starting with the two random genotypes, allowing for linkage between the alleles with a uniform recombination rate $\rho = 1 \times 10^{-6}$ bp^{-1}gen^{-1}. Between generations, we draw the number of recombination events in each pair of genomes from a Poisson distribution with rate ρ such that events are randomly distributed across the chromosomes. The resulting haploid recombinant genomes form two new individuals in the next generation. As a result, mutations stemming from each of the diverged founders are normally distributed, with a mean frequency $x = \frac{N}{2}$ in the population. A simulation ensemble is then built by evolving a population of size $N = 10^4$ with selection update rules for 27 generations. We set one driver mutation to be under selection and assume all other mutations are passengers. To incorporate the effects of noisy sampling in sequencing, allele frequencies are estimated at each time point by sampling a limited number of individuals. We draw binomial random variables for each locus with an uniform sample size of 100 reads, reflecting the sequencing coverage. The simulation parameters are summarised in Table 2.1.

Table 2.1 Simulation parameters for driver-passenger inference.

Variable	Symbol	Value
population size	N	10^4
number of sites	L	10^4
mutation rate	μ	1×10^{-10} bp^{-1}gen^{-1}
selection coefficient	σ	$\{\pm 0.05, \pm 0.1, \pm 0.2, \pm 0.3\}$
recombination rate	ρ	1×10^{-6} bp^{-1}gen^{-1}

Given these simulated observations, we then solve the time evolution for frequencies under this scenario in Equation (2.10). Figure 2.2 shows typical frequency trajectories of driver-passenger dynamics during a selective sweep, for drivers of different strength and a range of selection coefficients. The driver mutation increases in frequency with time, at a rate which is proportional to its selection coefficient. Based on these frequency trajectories we

can extract maximum-likelihood estimates of the unknown driver location and the selection coefficient. To test the unconstrained and constrained driver-passenger models, we will now examine the performance of the inference procedure in simulations under the Wright-Fisher model.

2.5.3 Localisation of drivers under selection

Firstly, we investigated the effect of changing the selective advantage of the favoured allele while the recombination rate was kept fixed (Fig. 2.3A). As expected, the error decreases when we use our knowledge of the recombination map, although for low σ, the expected error remains more or less similar (Fig. 2.3A). For large σ, the estimator begins to perform poorly because the variance of $\hat{\sigma}$ and the bias both increase. This is partly due to the finite resolution of sampling, as drivers with large σ become rapidly fixed and may not be captured by intermediate times. It is the observations at intermediate allele frequencies that give us precision in our estimates. For instance, changing the sampling frequency for $\sigma = 0.05$ from every 9 generations to every generation would make virtually no difference to the error.

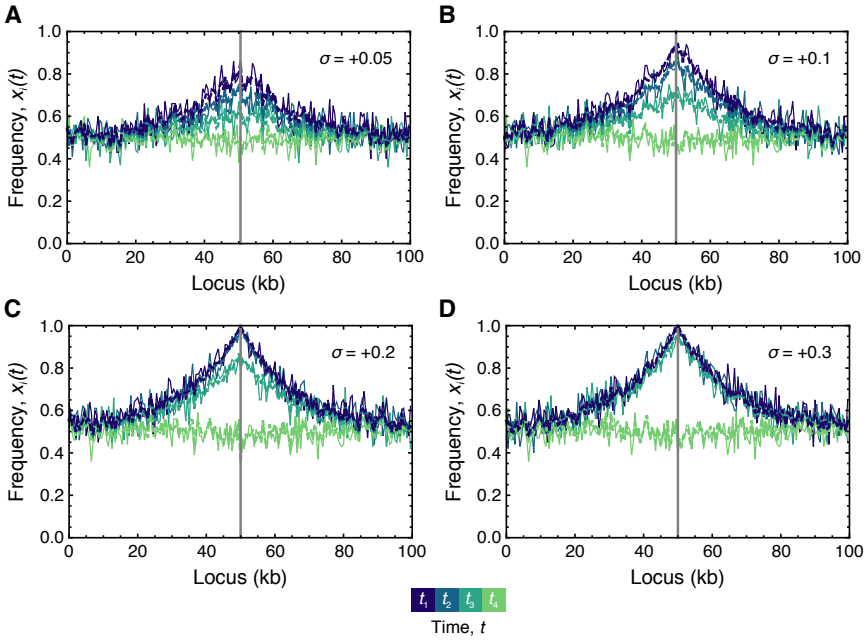


Fig. 2.2 Frequency trajectories described by the driver-passenger model. Time series of allele frequencies $x_i(t)$ are shown along the y-axis for locus positions with index i along the x-axis. In each panel, colours denote time points of each time series $t = (t_1, t_2, t_3, t_4)$. Solid lines indicate the true allele frequency x_i and dashed lines indicate the inferred allele frequency $\hat{x}_i$. Panels **(A-D)** show a range of scenarios with different selection coefficients of the driver, $\sigma \in \{+0.05, +0.1, +0.2, +0.3\}$. Vertical lines show the location of the driver.

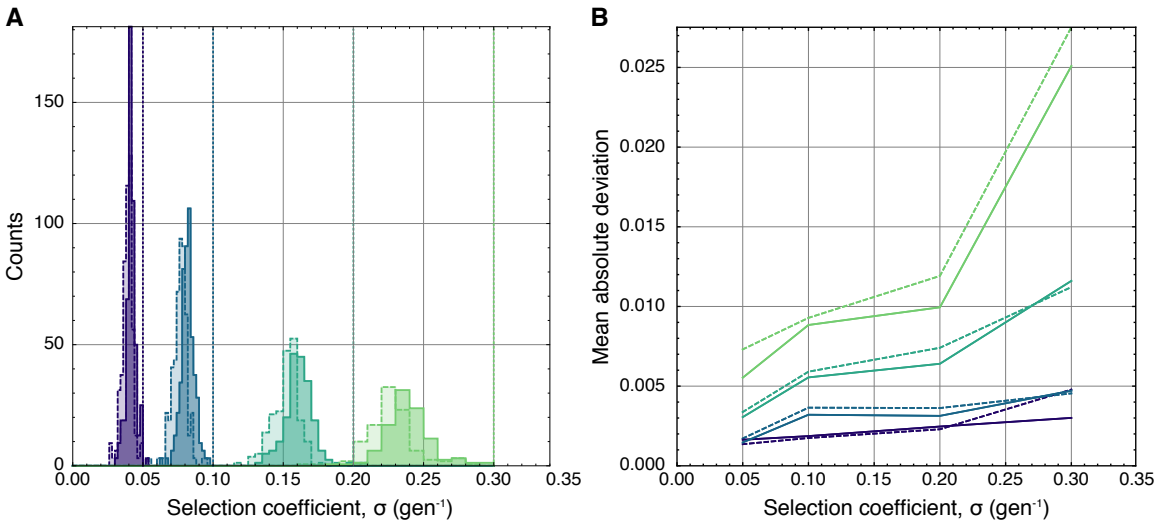


Fig. 2.3 Maximum likelihood estimates of the selection coefficient $\hat{\sigma}_i$. Solid (dashed) lines indicate datasets where the recombination rate was fixed (learned) using the constrained (unconstrained) driver-passenger model. **(A)** Histograms of the inferred selection coefficients $\hat{\sigma}$ for 160 simulations with true selection coefficients $\sigma \in \{\pm 0.05, \pm 0.1, \pm 0.2, \pm 0.3\}$. Dotted vertical lines show the true values. We combined the results across simulations with the same absolute value, $|\sigma|$, so each of histogram shows 40 simulations. **(B)** Mean absolute error of estimates of $\hat{\sigma}_i$. Each point is the mean of 40 independent simulations.

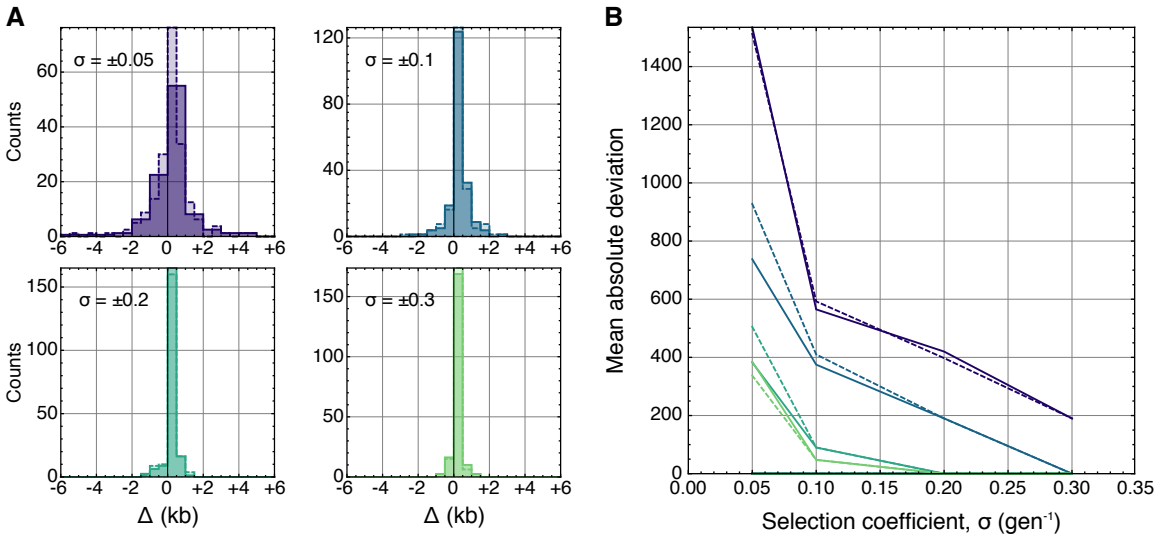


Fig. 2.4 Maximum likelihood estimates of the driver location $\hat{d}_i$. The maximum likelihood estimates $\hat{d}_i(n_i)$ of d_i is shown as $\Delta = \hat{d}_i - d_i$. Solid (dashed) lines indicate datasets where the recombination rate was fixed (learned) using the constrained (unconstrained) driver-passenger model. **(A)** Histograms of the results of 160 simulations varying the selection coefficient $\sigma \in \{\pm 0.05, \pm 0.1, \pm 0.2, \pm 0.3\}$. **(B)** Mean absolute error of estimates of $\hat{d}_i$ as a function of σ. Each point is the mean of 40 independent simulations.

Secondly, we also examined the effect that genomic correlations caused by linkage have on the error of maximum likelihood estimates of the driver location $\hat{d}_i$ (Fig. 2.4A). As σ increases, the estimates of the driver location become sharply peaked around the true value and the expected error in $\hat{d}_i$ decreases (Fig. 2.4B). As we expected, the driver mutation affects neighbouring sites up to a characteristic distance, and the effect of uniform recombination on multiple passengers in the same region is self-averaging over short distances. We have also observed in additional simulations that given a linkage structure with non-uniform recombination, the effect of drivers on the frequency of neutral passengers becomes asymmetric to each side of the driver locus during hitchhiking. Without knowledge of the correlation structure in the inference, there is a resulting bias in the error estimate of driver locations due to a trade-off in the frequency of passenger alleles if linkage is kept fixed.

Overall, the local correlation structure is one of the main determinants of the accuracy of the maximum-likelihood estimates. Provided we have a sufficiently large population for the deterministic approximation to hold and informative observations at intermediate allele frequencies, we can significantly improve our predictions with knowledge of the fine-scale linkage map, which will help us correctly decide whether a marker changed in frequency hitchhiking with a nearby driver or just due to statistical noise.

2.6 Summary

In this chapter we introduced minimal models of evolutionary dynamics to understand qualitatively different scenarios of population dynamics that arise due to evolutionary forces like mutation, selection, genetic drift and recombination. Here, special attention has been paid to the deterministic regimes in the single- and multi-locus cases to obtain insights into the temporal evolution of mutation frequencies through a population. We have implemented a model for driver and passenger mutations on the basis of a deterministic approximation for the growth of the driver allele during a selective sweep. This enables us to quantify the selective dynamics of multiple mutations in the genome despite the mathematical difficulties that arise from the microscopic details of multi-loci statistics. This model for the inference of fitness parameters compares the statistics of mutation histories to neutral expectations. We showed that this approach can discern driver mutations directly under selection from hitchhiking passengers using simulations. We will test the validity of our model in Chapter 4 as a means to estimate the selective effects of mutations and localise drivers under selection in a controlled evolution experiment.

Following from this work, extensions which relax certain assumptions of the driver-passenger model would be useful to improve our understanding of rapidly adapting populations. Firstly, mutations in different parts the genome do not have identical fitness effects and thus selective forces should be parameterised by a distribution of fitness effects. This should allow better interpretation of co-occurring drivers at different loci which are associated with a common quantitative trait. Secondly, the driver-passenger model currently only predicts fixation events for mutations under selection, but it does not address that many adaptive mutations will only sweep to intermediate frequencies. Incomplete sweeps and co-existing subpopulations are not only common in our laboratory evolution experiments (see Chapter 5) but have also been reported for chemotherapy-resistant tumour subpopulations [137] or artemisinin-resistant malaria strains [138]. Future work should aim to accommodate incomplete sweeps by extending the driver-passenger model beyond additive fitness effects and incorporating epistatic interactions between multiple loci.

Chapter 3

Probabilistic reconstruction of subclonal heterogeneity

3.1 Introduction

As we saw in earlier chapters, mutations are physically linked in the genome during asexual or somatic evolution. Their fates are therefore mutually dependent and selection can only act on these sets of loci in their entirety. At the genomic level, these correlations leave a large imprint on the data. Although this may be a curse when trying to distinguish driver mutations from passengers, correlations can be exploited to reconstruct the clonal lineages in a population. In this chapter, we introduce a probabilistic inference method to infer the clonal composition of a population when selection is sufficiently strong to amplify fit genotypes. We first review the molecular technologies that can be used to characterise the composition of genotypes in a population. We focus on high-throughput DNA sequencing of a mixed-cell population, where one can formulate the computational reconstruction of the genomic structure of the population as an inverse problem. Hidden Markov Models (HMM) are presented as powerful tools for probabilistic modelling, and specifically to model 'hidden' subclonal states belonging to different lineages in a mixed population. We first develop a general-purpose HMM for filtering and noise reduction of discrete observations from DNA sequence reads. We then discuss how state estimation with HMMs can combine correlated information from different sources of genetic variation to identify the number of subclones and their populations fractions, and to identify regions with subclone-specific mutations and copy-number aberrations.[1] The algorithmic performance is tested on simulated and real datasets. Finally, we discuss how this method can help understand temporally- or spatially-resolved genetic heterogeneity in a range of systems.

[1] The computational methods reported in this chapter are available from the GitHub code repository [https://github.com/ivazquez/PhD-thesis/tree/master/Chapter3].

The work reported in this chapter was carried out in collaboration with A. Fischer (A.F.), C. Illingworth (C.I.), V. Mustonen (V.M.) and S. Dentro (S.D.) at the Wellcome Trust Sanger Institute (Cambridge, UK), M. Tarabichi (M.T.) at the Francis Crick Institute (London, UK) and I. Leshchiner at the Broad Institute (Cambridge, MA).[1]

3.2 Molecular technologies for subclonal reconstruction

Mapping the lineage relations among the cells of an organism has been the holy grail of many fields of biology: from stem cell research, to developmental biology, cancer biology, or immunology [139]. Cell lineages can reveal the sequences of events like cell division, migration, or apoptosis that lead from the zygote to an adult organism. Nevertheless, complete cell lineage trees have only been reconstructed for simple organisms such as the model worm, *C. elegans* [140].

The accumulation of germline mutations in the genomes of individuals maintains a record of our shared evolutionary history over millions of years. Just as well, somatic genetic variation can mark and identify subpopulations of cells, and it can potentially be used to map the cell lineage of a complete organism [141]. Similarly, cells can also accumulate epigenetic changes – such as DNA methylation or histone modifications – that also serve as a record of the evolutionary history of different subpopulations [142].

Over time, subpopulations of cells may arise and expand driven by new beneficial mutations, depleting the pool of genetic diversity. Or they may decay and be outcompeted by other subpopulations. As shown in Figure 3.1A, these subpopulations can be reconstructed from whole-population, whole-genome sequencing data based on their private and shared genetic variants and their prevalence in a population [76, 143]. Recently, the advent of whole-genome, single-cell sequencing is offering the possibility to measure clonal genotypes and prevalences directly [144–147]. Nevertheless, several technical sources of noise still result in missing data, mostly caused by unintended measurements of doublets of cells and the failure to detect both alleles at heterozygous loci [148]. Furthermore, even if we had 'perfect', noise-free methods to sequence DNA from every single cell, we may still be unable to recover the complete lineage tree of all cells from naturally occurring somatic mutations. Unlike species, which have had sufficient time to accumulate many informative mutations,

[1] A.F., I.V.-G., C.I. and V.M. formulated the models; A.F. led the implementation of the software, with contributions from I.V.-G. and V.M; I.V.-G., C.I. and V.M. extended this algorithm to reconstruct subclonality in populations of microbes, parasites or viruses; I.V.-G. and V.M. applied this method to simulated data (generated by A.F., I.V.-G. and V.M.) and real data (generated by S.D., M.T. and I.L.).

individual cells accumulate mutations rarely and randomly, and thus do not guarantee that using mutations as lineage markers can distinguish every cell.

The crux to solving this problem may be complemented in the future by prospective approaches that can carry out lineage tracing forward in time [141]. For instance, lineage tracing using inducible genetic labelling has been a common approach in transgenic models, e.g., with the Cre-Lox system [149, 150]. In this system, a transient drug pulse can control the Cre recombinase to be sequestered into the nucleus of the cell, ultimately leading to the transcription of a fluorescent reporter. As a result, one can confer the expression of a hereditary label on targeted cells. A very promising alternative is lineage tracing by molecular

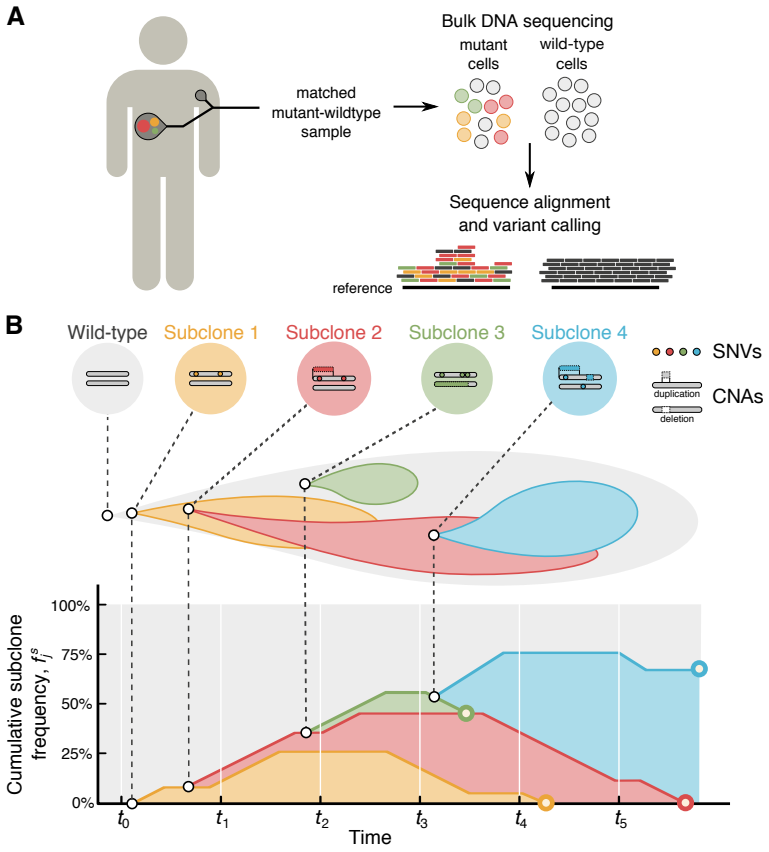


Fig. 3.1 Reconstruction of clonal evolution using genome sequencing. **(A)** Whole-genome, whole-population DNA sequencing of cells in matched wild-type/mutant samples. Copy-number aberrations (CNA), single-nucleotide variants (SNV) and other mutations can be identified from the sequence data, by aligning the sequence reads to a reference genome. **(B)** Schematic view of subclonal diversification. In this example, mutations in daughter cells of a single founder cell (left) diverge into subclones (reflected by different colours). A point mutation occurs early on with a subsequent gain of a chromosome arm and a short deletion at a later stage, each followed by clonal expansion (subclones 1, 2, and 4). A short-lived lineage arises independently and goes extinct (subclone 3).

recording, which relies on the continuous and controlled generation of stochastic variation at fixed locations in the genome [151–154]. This can be achieved by arrays of CRISPR/Cas9 target sites which are progressively edited over many cell divisions, each of which acts as a barcode. Lineage trees can then be reconstructed, since edited barcode sequences are related to one another by shared mutations and can be read by sequencing. Hypothetically, if we wanted to trace the full development of a human embryo from 1 to $\sim 3.7 \times 10^{13}$ cells in an adult, we can estimate the minimum number of barcodes that would yield useful information on the complete cell lineage tree. These recent approaches have experimentally shown to be capable of storing $\sim$5 bits of information per locus, which is sufficient to distinguish 2^5 different cells [155, 156]. To uniquely identify the $\sim 3.7 \times 10^{13}$ cells in a human will require at least 10 such barcodes per cell, such that $(2^5)^{10} > 3.7 \times 10^{13}$. This is very promising compared to the complexity of clonal tree partitions that can be currently recovered from whole-population, whole-genome sequencing, As we will see in this chapter, with the current state-of-the-art we are only able to reconstruct up to 2 or 3 subpopulations of cells. By comparison, encoding the cellular identity of such few subpopulations corresponds to an information content of only 2 bits.

3.3 Reconstruction of subclonal heterogeneity

Our aim will be to characterise subclones using whole-population, whole-genome sequencing, which does not yield direct information on long-range haplotypes when applied to mixed cell populations. Here and throughout this thesis, we employ the term 'subclone' to refer to subpopulations which comprise a group of cells carrying the same set of mutations. Only few of these subclones expand in a population thus becoming detectable by whole-population, whole-genome sequencing. We now discuss approaches to analyse data produced by this experimental technique and to infer the clonal composition of a population. We will focus on the following question: given a mixed sample of the genomes of cells that have accrued mutations, can we reconstruct their evolutionary history? This problem consists of three parts: (i) identification of subclones, (ii) reconstruction of subclone-specific profiles, and (iii) inference of evolutionary relationships between subclones.

To keep a broad scope, we adopt a general terminology that defines two compartments in a population: a *mutant* compartment (used interchangeably to refer to tumour cells undergoing somatic evolution, mutant cells that are drug-resistant, etc.); and a *wild-type* compartment (to refer to the normal tissue surrounding tumour cells, or to the ancestral population that is drug-sensitive, etc.). Our working definition of a subpopulation – or subclone – will

be the maximal set of cells carrying the same arbitrary set of mutations in the mutant compartment. The standard convention is to refer to a mutation as *clonal* or *fixed* if it appears in all cells of a population. If it only appears in a fraction of the cells, it is typically referred to as *subclonal* or *polymorphic*.

To reconstruct the subclone dynamics in a cell population from sequence data, we would like to formulate the problem under the following assumptions:

(i) Cells evolve somatically or asexually by evolutionary forces like mutation, selection or genetic drift and, crucially, without recombination. This ensures that there are long-range correlations along their genome, which can, in principle, be reconstructed from short DNA sequences.

(ii) The population consists of a mixture of subclones, i.e., groups of genetically identical cells. The total number of subclones N_c is unknown. The relative subclone frequency f_j^s of subclone j in sample s of the population is also unknown. The number of 'macroscopic' subclones – those which can be reconstructed from real data – is small.

(iii) Each subclone carries a unique genotype and a unique copy-number profile, both of which are unknown.

(iv) There is a distinct wild-type compartment of the population which differs from the macroscopic subclones, e.g., by having a different set of genotypes. The fraction of the wild-type compartment is also unknown.

(v) When several samples are jointly analysed, the same subclonal populations are assumed to be present in all samples. However, their frequencies in some of the samples can be zero.

3.3.1 Data types

The clonal composition of a population sample can be inferred from whole-genome sequencing data using two different kinds of information. Firstly, the profile of copy-number changes, if subclones are indeed defined by their copy-number profiles. Secondly, the number of reads reporting mutations, which can distinguish between subclones even if there are no copy-number changes in the mutant compartment of the population. On this basis, we can distinguish three different data types: copy-number aberrations (CNA), B-allele frequencies (BAF) and single-nucleotide variants (SNV).

Copy-number aberrations (CNA) are gains or losses of chromosomes that are acquired *de novo* by the mutant compartment. A deletion in the genome will translate as a drop in read depth in the mutant cells compared to the wild-type, with fewer reads or no reads in the deleted region. Conversely, duplications or amplifications of the genome equate to more aligned sequences in the mutant with respect to the wild-type.

B-allele frequencies (BAF) report the abundance of pre-existing single-nucleotide variants which are heterozygous in the wild-type sample (e.g., germline variants in humans) and display allelic imbalances in the mutant sample (e.g., loss of the only remaining wild-type copy of a tumour suppressor gene). The loss of one of the chromosome copies can be due to *de novo* copy-number changes or copy-neutral loss-of-heterozygosity (LOH).

Single-nucleotide variants (SNV) are *de novo* point mutations, small insertions or deletions which can arise at any time in the evolution between the wild-type and the mutant subpopulations (e.g., somatic mutations from the fertilised egg to the development of a tumour).

We would like to perform the inferences jointly across all data types, which can greatly improve the evidence for one of several competing solutions. However, the resulting clonal decompositions need not be the same given different data types. As Figure 3.2 shows, two subclones with identical *de novo* mutations can still have different copy-number profiles (orange and red subclones). Hence, we would like to perform an integrative analysis to attain a clonal decomposition jointly at the level of CNAs, BAFs, and SNVs. The method should provide inferences of the number of subclones detected in the sample, their population frequencies across time and/or space, and subclone-specific posterior probabilities of copy-number profiles as well as pre-existing and *de novo* variant genotypes.

3.3.2 Computational methods

Several computational methods have been developed to reconstruct clonal composition from whole-population, whole-genome sequencing data. These methods mostly focus on inferring the clonal composition of a cancer cell population, and they vary according to type of input data and their assumptions about phylogenetic processes. Regarding the integration of data layers, some methods aim at clustering SNVs (e.g., PyClone, PhyloSub) [157, 158], while others incorporate CNA data in their inference (e.g., ABSOLUTE, THetA, TITAN) [159–161].

Sequencing a cell population enables the detection of SNVs and their allele frequencies, which can be used for subclonal reconstruction. In order to estimate the frequency of mutant cells carrying the SNV, these allele frequencies need to be corrected if they occur in

regions with CNA or LOH, and also must be corrected for the presence of wild-type cells. Because the genome in which a certain SNV arose is unknown, SNVs are clustered into sets of mutations according to their estimated allele frequency. Bayesian mixture models have been commonly used to cluster SNVs based on a tree stick-breaking process, estimating the number of mixture clusters together with their frequencies and densities [157, 162]. To infer a tree phylogeny between SNV clusters, most methods make two simplifying assumptions: (i) no mutation occurs twice in the course of evolution ('infinite sites' assumption), and (ii) no mutation is lost (no back mutations) [163]. Computational methods have implemented these assumptions to infer tree phylogenies. However, most of these methods have limitations, focusing exclusively on SNV variants in regions of the genome which are copy-number neutral

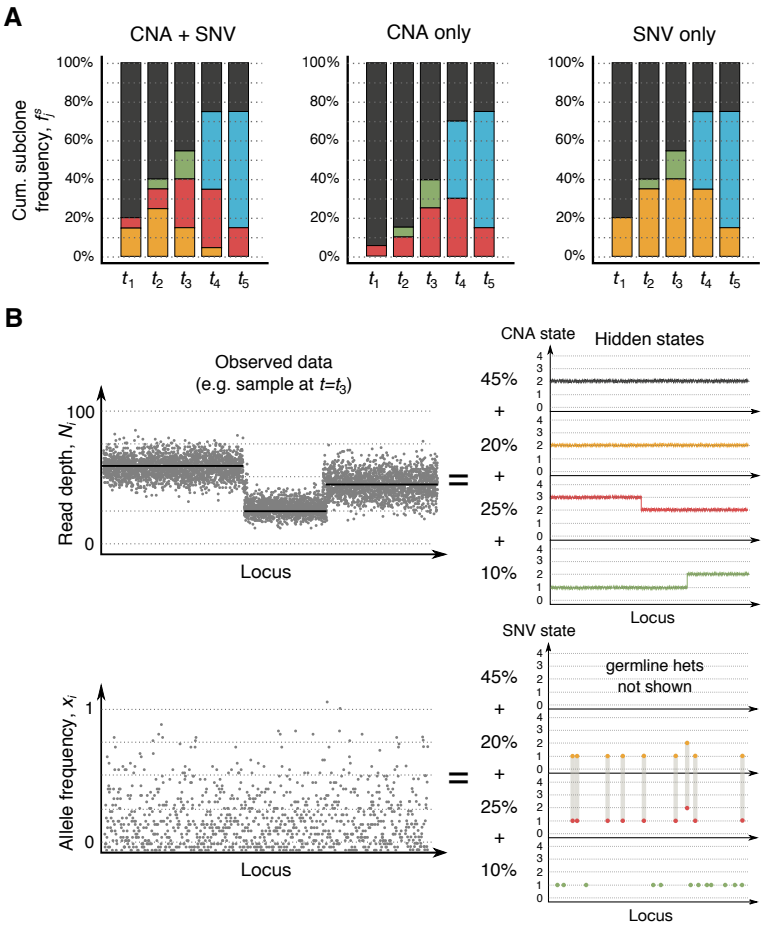


Fig. 3.2 Schematic of a statistical model of subclonal heterogeneity. **(A)** In the left column, the demarcation of the clonal lineages using CNA and SNV information is shown. Middle and right columns show different decompositions using only one of these data types. **(B)** DNA sequencing of a cell population in a mutant sample and a matched wild-type. The different data layers (left) can be used to infer the underlying population structure (right); vertical lines highlight shared SNVs.

and free from allelic imbalances. Even methods which can incorporate local copy-number information into the SNV inference assume that all copy-number events are clonal [157].

Patterns of CNA and BAF are also informative for subclonal reconstruction. They can also be identified by sequencing a cell population, based on segmentation of the genome-wide read depth profile (CNA) in the mutant sample, and the allele counts in the mutant sample of loci that are heterozygous in the wild-type sample (BAF). CNA segmentation is typically obtained from normalised read counts, which are defined as the ratio between the local DNA copy number in a mixture of mutant and wild-type cells. BAF segmentation is derived from the ratio between the minor allele and the total allele count. However, most CNA-based algorithms work for the limit case of a fully clonal population [160], or consider all loci independently rather than modelling the actual CNA events, an assumption which is ill-suited as multiple loci are affected by CNA and are not independent. These events can span megabases, and computational methods have addressed this by aggregating statistical strength across adjacent genomic loci induced by segmental CNA and LOH [159–161].

While several methods have used probabilistic approaches in this context [161], these have tried to account for noisy data but they do not jointly infer individual subclonal fractions, subclone genotypes and subclone copy-number profiles, often using only one of the available data types. SNV frequency data only weakly constrain the set of possible clonal structures. Equally, CNA data alone can be consistent with multiple clonal compositions, which can only be disambiguated by BAF or SNV data. To solve this identifiability issue, we will aim to integrate all these data types in a joint probabilistic inference.

3.4 Hidden Markov Models

To tackle this problem, we first introduce probabilistic modelling and inference. Consider the state x_i of a genome of length L that can be described by a Markov chain shown in Figure 3.3A. The state of the Markov chain is directly visible to us observers, and thus the only unknown parameters are the state transition probabilities, $P\left(x_i \,\middle|\, x_{i-1}\right)$. However, we normally do not observe a sequence of states $x_{1:L}$ directly in the real world. Instead, we observe an indicator of the true state when we measure a set of observations $X_{1:L}$. Markov chains are not able to account for what our belief of the true state of the genome is, given these observations. Furthermore, Markov chains cannot by themselves easily account for long correlations in the data. Figure 3.3A is a first-order Markov chain, which assumes that x_{i-1} contains everything we need to know about the entire previous history, $x_{1:i-2}$. We can add a dependency from x_{i-2} to x_i in a second-order Markov chain, but this will still

be inadequate to describe long-range correlations in the observations (e.g., in read depth). Building ever higher order Markov chains is not feasible, as the number of parameters will blow up. As a result, Markov chains can only accommodate short-range correlations in the genome.

An alternative probabilistic model that addresses these shortcomings is the Hidden Markov Model (HMM), which assumes there is an underlying process with a hidden state x_i that can be modelled by a first-order Markov chain, but we only get noisy observations X_i of this process (Fig. 3.3B). HMMs are a widely used probabilistic model for sequential and time-series data [164, 165]. This is a probabilistic approach, which has the advantages that it will provide us with uncertainty estimates of our inference and it will enable us to integrate observations from related samples and from multiple data layers.

A HMM consists of two components: an emission model and a transition model (or propagator). The observations X_i are described by an emission probability, $P\left(X_i \,\middle|\, x_i, \theta\right)$, that takes the current hidden state x_i as a parameter and also depends on global parameters θ. This emission probability describes the noise in our observations. The hidden states may take discrete or continuous values. Changes to the hidden state along the genome are fully determined by the transition probability (or propagator), $P\left(x_{i+1} \,\middle|\, x_i, \theta\right)$, which usually will also depend on global model parameters θ. The transition model describes what

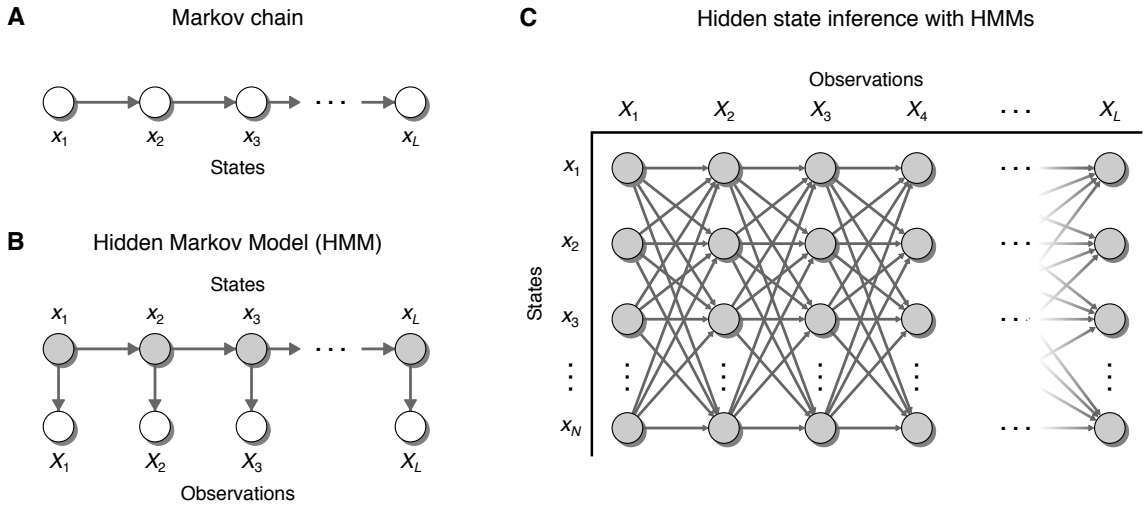


Fig. 3.3 Schematic diagrams of a Markov chain and a Hidden Markov Model. The arrows in the diagrams indicate conditional dependencies. Based on Rabiner and Juang [164]. **(A)** First-order Markov chain along a genome of length L. The states of the chain x_i are shown by white circles. **(B)** First-order HMM. The observations X_i are shown by white circles, and the hidden states of the chain x_i are shown by grey circles. **(C)** For hidden state inference, the trellis of hidden states can take values x_i corresponding to observation X_i at genomic position i ($i = 1, \ldots, L$).

the probability of our current state x_i is, given the previous state x_{i-1}. The joint probability distribution over all states is then

$$P(x,\, X) = P(x_1) \prod_{i=1}^{L-1} P\left(x_{i+1} \,\middle|\, x_i\right) \prod_{i'=1}^{L} P\left(X_{i'} \,\middle|\, x_{i'}\right). \tag{3.1}$$

The first product term denotes the state transitions of the Markov model. The second product term denotes the probability of the observations given the states. The equivalent graphical model is shown in Figure 3.3B.

Now consider the problem of reconstructing the hidden Markov chain $x_{1:L} \in \mathbf{R}$ from a sequence of randomly generated emissions $X_{1:L}$ (Fig. 3.3C). To do this, we would ideally like to estimate the joint distribution $P(x \mid X)$ given all the observations X throughout our current history. Normally, however, it is sufficient to determine the probability for the current state given all observations, $P\left(x_i \mid X\right)$, which is simpler. We will define two quantities: firstly, $\alpha_i(x_i)$ which is the joint probability of the current state and all previous observations; and secondly, $\beta_i(x_i)$, which is the conditional probability of all future observations along the sequence. These will be a key part of our inference procedure, since the product of these two quantities is proportional to the conditional probability $P\left(x_i \mid X\right)$ after normalising, i.e., $\alpha_i(x_i)\, \beta_i(x_i) \propto P\left(x_i \mid X\right)$.

Firstly, we will use HMMs for **data filtering** and state estimation. This consists of filtering observations X, which are typically discrete and very noisy. We would like to calculate the posterior probability distribution $P\left(x_i = x \,\middle|\, X_{1:L}, \theta\right)$ of the system to be at hidden state x at locus i in the genome, given all the observations. Once we know the posterior, we will estimate the global parameters θ, which are unknown. We refer to this step as 'filterHD', and it is presented in Section 3.5.

Secondly, we will define HMMs for **subclonal reconstruction**. The aim will be to relate the filtered data signal to the subclonal properties of genomes by assuming there is a hidden state x that is a property of the subclone (e.g., the number of chromosome copies or the number of mutated copies in a subclone). The hidden state is not directly accessible and must be inferred from filtered measurements of observables X, while accounting for errors in the measurement process. We present the subclonal reconstruction step in Section 3.6, which we call 'cloneHD'.

3.5 Continuous state-space HMM for data filtering

We will work through a toy example of asexual or somatic evolution to present the algorithm. Figure 3.4 shows a simulated example for two subclones ($N_c = 2$) of size $f_j^s \in \{0.54, 0.16\}$. At first glance, the three data layers we consider – CNA, BAF and SNV profiles – seem very hard to interpret. The noise makes it even harder, so can we somehow remove the noise? Our first task is to reduce the noise in the raw signals of mutation counts and read depth counts. We will then combine these filtered estimates in Section 3.6 to build the best overall model of the clonal composition, integrating relevant information from related samples or from sources of variation coming from multiple data layers. In order to estimate the hidden state given a set of measurements of a raw observable we will use a HMM design similar to a Kalman filter [166]. Unlike the Kalman filter, HMMs do not necessarily assume knowing how to compute the transition probabilities, and they will also give us the probability distribution for the hidden state [167, 168].

3.5.1 Emission models

To deal with the filtering problem, we need a generative emission model that accounts for noise in the observed data which is tractable enough to solve the inference. The emission models that we will consider here are the binomial model, the Poisson model, and their overdispersed counterparts, the beta-binomial and negative binomial model, respectively. These emission models must be able to simulate datasets of mutation counts or read depth signal which are comparable to real data.

Binomial model

Firstly, we consider a binomial emission model which can describe *de novo* single-nucleotide variants (see Figure 3.4C). For the binomial model, our observations $X = \{n_i, N_i\}_{i=1\dots L}$ are the number of reads reporting the variant allele $n_{1:L}$ out of $N_{1:L}$ reads, such that $n_i \in \{0, \dots, N_i\}$. The sequence of sample depths $N_{1:L}$ can be the same for all i, or they can follow their own Markov chain. Given these observations, the emission model describes the success probability $x_i \in [0, 1]$ of observing n_i out of N_i reads reporting the variant allele at locus i (i.e., the number of successful binomial trials), which follows the binomial probability mass function:

$$n_i \sim \text{Bin}\left(n_i \,\middle|\, N_i, x_i\right) = \binom{N_i}{n_i} x_i^{n_i} \left(1 - x_i\right)^{N_i - n_i}. \tag{3.2}$$

The observations can often be overdispersed in real data due to measurement noise. We can account for overdispersion using the beta-binomial distribution:

$$n_i \sim \text{Beta-Bin}\left(n_i \,\middle|\, N_i, x_i, C\right) = \binom{N_i}{n_i} \frac{\text{Beta}\left(n_i + C\,x_i, N_i - n_i + C\,(1-x_i)\right)}{\text{Beta}\left(C\,x_i, C\,(1-x_i)\right)} \tag{3.3}$$

where $\text{Beta}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ and $\Gamma(z) = \int_0^\infty \mathrm{d}t\; t^{z-1}\; e^{-t}$ is the gamma function. The shape parameter $C > 0$ has the property that in the limit $C \to \infty$ we recover the binomial distribution, since the beta distribution is a conjugate of the binomial distribution.

To account for random errors in SNV detection against a reference, we can incorporate an error term in the emission model in Equation (3.2). For the binomial model (also beta-binomial), random emissions with rate ε are drawn from the uniform distribution in the range $[0, N_i]$:

$$n_i \sim (1-\varepsilon)\,\text{Bin}\left(n_i \,\middle|\, N_i, x_i\right) + \varepsilon\,\text{Unif}_{[0,N_i]}. \tag{3.4}$$

Each of the mutant allele observations can either be a true positive mutation as described by the first term, or a false positive emitted by the second term. Depending on the data source it may be necessary to include this random noise emission channel.

Poisson model

The second layer of information is given by copy-number aberrations. Given a set of read depth observations $X = \left\{N_i\right\}_{i=1\ldots L}$, where each observation corresponds to the median sequencing depth in a window i, we assume that they follow a Poisson process (see Figure 3.4A). Then for each window, the number of events N_i can be described by a Poisson rate $x_i \in \mathbf{R}_+$. The variation in the number of DNA sequences aligning in a region of the genome is not exactly Poisson, but the approximation is very good. The probability that we obtain N_i counts in window i is given by the Poisson distribution, so that the emission model is:

$$N_i \sim \text{Pois}\left(N_i \,\middle|\, x_i\right) = \frac{x_i^{N_i} e^{-x_i}}{N_i!}. \tag{3.5}$$

The read depth observations may be overdispersed due to measurement noise, or resulting from the product of random variables as sequencing read depth results from DNA replication, which is the product of randomly fluctuating rate constants. To account for overdispersion,

we can use the negative binomial distribution:

$$N_i \sim \mathrm{NB}\left(N_i \,\middle|\, x_i, C\right) = \frac{\Gamma(N_i + C)}{\Gamma(C)\,\Gamma(N_i + 1)} \left(\frac{C}{x_i + C}\right)^{C} \left(\frac{x_i}{x_i + C}\right)^{N_i} \tag{3.6}$$

where Γ is the gamma function defined above. In the limit $C \to \infty$, the negative binomial distribution approaches the Poisson distribution.

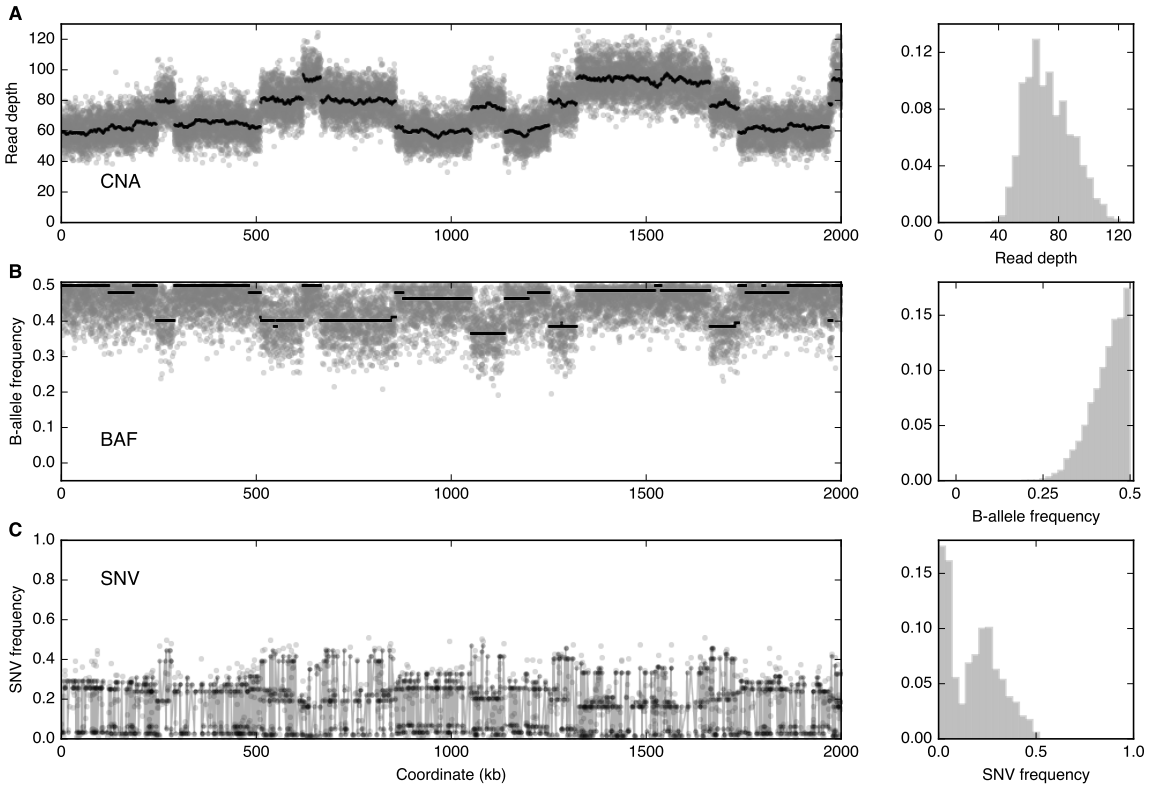


Fig. 3.4 Simulated example of CNA, BAF and SNV profiles in the presence of two subclones ($N_c = 2$) of size $f_j^s \in \{0.54, 0.16\}$. A full breakdown of the simulation parameters is given in Table 3.1. Left panels show the raw data counts. The posterior mean emission rates are shown as solid black lines. Right panels show the density distribution for each track. **(A)** CNA: Read-depth track from a mixture of a wild-type subpopulation plus two mutant subclones that can be used to infer subclone-specific copy-number profiles. Here the mean number of reads per chromosome copy (or mass M^s) is $M^s = 15$ (see Section 3.6.1). **(B)** BAF: B-allele counts for pre-existing mutations in the mixture help to decide between balanced and unbalanced copy-number changes. **(C)** SNV: *De novo* mutation counts.

Symmetric binomial model

Finally, heterozygous single-nucleotide variants that are pre-existing in all cells (e.g., in the germline) may be affected by *de novo* copy-number aberrations (see Figure 3.4B). Gains or losses of chromosomes or chromosomal regions that are imbalanced will split the observed frequencies, commonly referred to as B-allele frequencies. B-allele frequency observations, $X = \left\{n_i, N_i\right\}_{i=1\ldots L}$, correspond to the sequence of sample depths $N_{1:L}$ and the number of reads reporting the minor variant allele $n_{1:L}$ out of $N_{1:L}$ reads, where n_i and $(N_i - n_i)$ are considered identical observations. These integer observations can be described by a symmetric binomial emission model with success probability $x_i \in [0, 1]$,

$$n_i \sim \text{Bin}\left(n_i \,\middle|\, N_i, x_i\right) + \text{Bin}\left(N_i - n_i \,\middle|\, N_i, x_i\right). \tag{3.7}$$

If the data is overdispersed, the probability can take the form of a symmetric beta-binomial distribution

$$n_i \sim \text{Beta-Bin}\left(n_i \,\middle|\, N_i, x_i\right) + \text{Beta-Bin}\left(N_i - n_i \,\middle|\, N_i, x_i\right), \tag{3.8}$$

with shape parameter C (Equation (3.3)).

To summarise, our observations for SNV data are the number of reads reporting a *de novo* variant allele n_i out of a total of N_i reads, s.t. $X = \left\{n_i, N_i\right\}_{i=1\ldots L}$. The SNV emission models are the binomial and beta-binomial processes, defined by the success probability $x_i \in [0, 1]$ which will be the hidden state of our HMM. For CNA data, our observations are the total number of reads per window, s.t. $X = \left\{N_i\right\}_{i=1\ldots L}$. The CNA emission models are Poisson and negative binomial processes, which are defined for hidden rates $x_i \in \mathbf{R}_+$. For BAF, our observations are the number of reads reporting a pre-existing variant allele n_i out of a total of N_i reads, s.t. $X = \left\{n_i, N_i\right\}_{i=1\ldots L}$. The BAF emission models are symmetric binomial or symmetric beta-binomial, both defined by success probability $x_i \in [0, 1]$.

3.5.2 Transition models

The transition models are used to model correlated processes and segment the CNA and BAF data in a probabilistic fashion, allowing the HMMs to make state transitions only at potential jump sites that are informative about the subclonal structure of the population.

A transition model is used to determine, given current state x_i, what the probability of the next state x_{i+1} is. Having observed typical CNA and BAF datasets, there is one desideratum

for a suitable transition propagator: it must be able to model phenomena where the state does not change for a certain stretch of the genome, and can also eventually transition to other states. This transition model between states x_i and x_{i+1} will have the form

$$x_i \to x_{i+1} \sim (1 - \pi_{i,i+1}) \times \text{stay} + \pi_{i,i+1} \times \text{jump}\,. \tag{3.9}$$

Here, $\pi_{i,i+1}$ is the conditional probability of a jump between two loci i and $i+1$. At a given position i, there is a probability p (per base) to 'jump' to a new state, or a probability $(1-p)$ to 'stay' in the current state of the genome.

Piecewise-constant propagator

One possible transition model is a piecewise-constant propagator, which is able to account for jumps in the observations that separate regions with long-range correlations. A piecewise-constant state sequence can be generated by

$$P\left(x_{i+1} \,\middle|\, x_i, \theta\right) = \left(1 - \pi_{i,i+1}\right)\ \delta\left(x_{i+1} - x_i\right) + \pi_{i,i+1}\ P_0(x_i \mid \theta), \tag{3.10}$$

where $\pi_{i,i+1}$ is the probability that the system has jumped between positions i and $i+1$. This probability is defined as $\pi_{i,i+1} \equiv 1 - (1-p)^{\Delta_{i,i+1}}$, where $\Delta_{i,i+1}$ is the distance between loci i and $i+1$ (in bases). The Markov chain can perform two types of transitions between states: jumping to a new state drawn randomly from a given proposal distribution P_0 with probability p (per base), or otherwise remaining in its current state with probability $(1-p)$ according to $\delta(x_{i+1} - x_i)$, which is the Dirac delta function. In this scenario, the only parameter θ that needs to be learned is the jump probability p.

Jump-diffusion propagator

We can generalise the propagator of the piecewise-constant process with a jump-diffusion propagator. The 'jump' component can account for abrupt changes, while the 'diffusion' component can accommodate smooth, short-range correlations in the signal that may arise due to sequence biases

$$P\left(x_{i+1} \,\middle|\, x_i,\ \theta\right) = \left(1 - \pi_{i,i+1}\right)\ \mathcal{N}\left(x_{i+1} - x_i, \sigma\sqrt{\Delta_{i,i+1}}\right) + \pi_{i,i+1}\ P_0\left(x_i \mid \theta\right) \tag{3.11}$$

where the first term is a normal distribution with mean $x_{i+1} - x_i$ and standard deviation $\sigma\sqrt{\Delta_{i,i+1}}$, and allows the hidden state x to diffuse with diffusion constant σ, with probability

$(1 - p)$. According to the second term, the hidden state can jump to a new value randomly drawn from P_0 with jump probability p (per base). The uniform distribution can be used as our proposal distribution $P_0(x \mid \theta)$. The two parameters $\theta = (p, \sigma)$ need to be learned simultaneously, and both p and σ relate to the stiffness of the hidden state, with $p = 0$ corresponding to a pure diffusion process and $\sigma = 0$ to a pure jump process.

3.5.3 Forward-backward algorithm

For correlated data, such as CNA and BAF, the total data likelihood can be efficiently computed via the forward-backward algorithm for HMMs [164]. A generalisation of this algorithm is needed if the hidden variable x_i is a continuous rather than a discrete state label. Then the current information about x_i is encoded in a continuous probability distribution, instead of a finite vector. The forward algorithm takes then the form of a Bayesian online-learning iteration, where new observations X_i are incorporated *online* (as they 'arrive') into a posterior distribution for x_i. The forward iteration is described in two steps to be performed for each locus in turn [164].

$$\text{Predict:} \quad P\left(x_i \,\middle|\, X_{1:i-1}\right) = \int \mathrm{d}x_{i-1}\, P\left(x_i \,\middle|\, x_{i-1}, \theta\right)\, P\left(x_{i-1} \,\middle|\, X_{1:i-1}\right) \tag{3.12}$$

$$\text{Update:} \quad P\left(x_i \,\middle|\, X_{1:i}\right) = \frac{P\left(X_i \,\middle|\, x_i\right) P\left(x_i \,\middle|\, X_{1:i-1}\right)}{\int \mathrm{d}x\; P\left(X_i \mid x\right) P\left(x \,\middle|\, X_{1:i-1}\right)} \tag{3.13}$$

The result of the 'predict' step can be interpreted as a *prior* distribution for the locus i. For $i = 1$, we can use the proposal distribution $P_0(x)$. The 'update' step is a Bayesian computation of the *posterior* distribution, given the current observation X_i. The backward form of this iteration is performed analogously.

The forward conditional distribution is computed in the forward 'update' step (for α). This step involves computing α one step ahead to act as the new prior for locus i, and then absorbing the observed data from locus i using Bayes' rule [164]. Subsequently, the backward conditional distribution is computed in the backward 'predict' step (for β). If we have already computed $\beta_i(x)$, it can be shown that $\beta_{i-1}(x)$ follows by recursion [169] and we get the marginals

$$\alpha_i(x) \equiv P\left(x_i = x \,\middle|\, X_{1:i}\right) \quad \text{and} \quad \beta_i(x) \equiv P\left(x_i = x \,\middle|\, X_{i+1:L}\right). \tag{3.14}$$

Having computed the forward and backward calculations, we can combine them to get the posterior distribution $\gamma_i(x) \propto \alpha_i(x)\beta_i(x)$. The final result of the forward-backward algorithm

is the posterior distribution $\gamma_i(x)$ after normalising,

$$\gamma_i(x) \equiv P\left(x_i = x \,\middle|\, X_{1:L}\right) = \frac{\alpha_i(x)\,\beta_i(x)}{\int \mathrm{d}y\; \alpha_i(y)\,\beta_i(y)}. \tag{3.15}$$

The forward-backward step can be calculated by a $\mathcal{O}(N^2 L)$ pass through the data, as it involves a $N \times N$ multiplication at every step [169]. Once this is done the resulting distributions can be used to iteratively update the transition probabilities and observables to then estimate the hidden state and vice versa. Going back and forth iteratively, this design is just the expectation-maximisation (EM) algorithm which we will also encounter in Chapter 6.

Once we have computed the posterior probability $\gamma_i(x)$, there is a range of possibilities. For example, we can do inference to detect changes in the hidden state with any of the transition models we proposed earlier. With the piecewise-constant or the jump-diffusion models, we can calculate the posterior probability $\hat{\pi}_{i,i+1} \equiv P\left(x_i \neq x_{i+1} \mid X\right)$ that at least one jump has actually occurred between loci i and $i+1$. To this end, we compare the two transition probabilities to go from x_i to x_{i+1} either by diffusion or by a jump:

$$\hat{\pi}_{i,i+1} = \frac{\int \mathrm{d}x_i \int \mathrm{d}x_{i+1}\; \alpha(x_i)\, \frac{\pi_{i,i+1}}{b-a}\, P\left(X_{i+1} \,\middle|\, x_{i+1}\right)\, \beta(x_{i+1})}{\int \mathrm{d}x_i \int \mathrm{d}x_{i+1}\; \alpha(x_i) \left[\frac{\pi_{i,i+1}}{b-a} + (1-\pi_{i,i+1})\mathcal{N}\left(x_{i+1} - x_i, \sigma\sqrt{\Delta}\right)\right]\, P\left(X_{i+1} \,\middle|\, x_{i+1}\right)\, \beta(x_{i+1})}. \tag{3.16}$$

Here, a and b are the lower and upper boundary conditions of the propagator, respectively. These need to be chosen carefully to approximate the continuous distribution $P(x \mid X)$ on a finite grid.

3.5.4 Total data likelihood and parameter learning

The global parameters θ need to be estimated jointly. We can learn the parameters numerically by maximising the total log-likelihood of the model given the data, s.t. $\hat{\theta} \equiv \mathrm{argmax}_\theta\; \mathcal{L}(X|\theta)$. These global parameters θ include the jump probability p, the diffusion constant σ, the error rate ε and the shape parameter C.

As is standard in HMM calculations, we need to aggregate the log-transformed normalisation terms in each forward 'update' step,

$$\mathcal{L}(X|\theta) = \sum_{i=1}^{L} \log \int \mathrm{d}x_i\; P\left(X_i \,\middle|\, x_i, \theta\right) P\left(x_i \,\middle|\, X_{1:i-1}, \theta\right). \tag{3.17}$$

For correlated models, this total log-likelihood is obtained automatically during the forward algorithm. This form of the log-likelihood function becomes simpler for uncorrelated process, where the prior distribution for x is the same at all loci.

This data filtering step is going to help efficiently carrying out the subclonal reconstruction. Setting a minimum jump probability, we can determine the loci where transitions between subclonal states are supported by the data. If this threshold is set too high some true transitions might be missed, leading to incorrect reconstructions. If it is set too low, too many segments will be introduced, slowing the algorithm down. A minimum jump probability of 1% or greater is typically a good compromise, as shown by Figure 3.5 for CNA and BAF profiles. Practically, this entails first filtering the wild-type and mutant CNA and BAF datasets independently, with $p > 0$ and $\sigma > 0$. This yields the posterior mean $\hat{x}_i$ of the wild-type emission rate, and the total log-likelihood and global parameters for the mutant. Assuming that the bias field χ_i for the wild-type is shared by the mutant sample, the segmentation can then be re-run for the mutant CNA or BAF datasets together with the bias field.

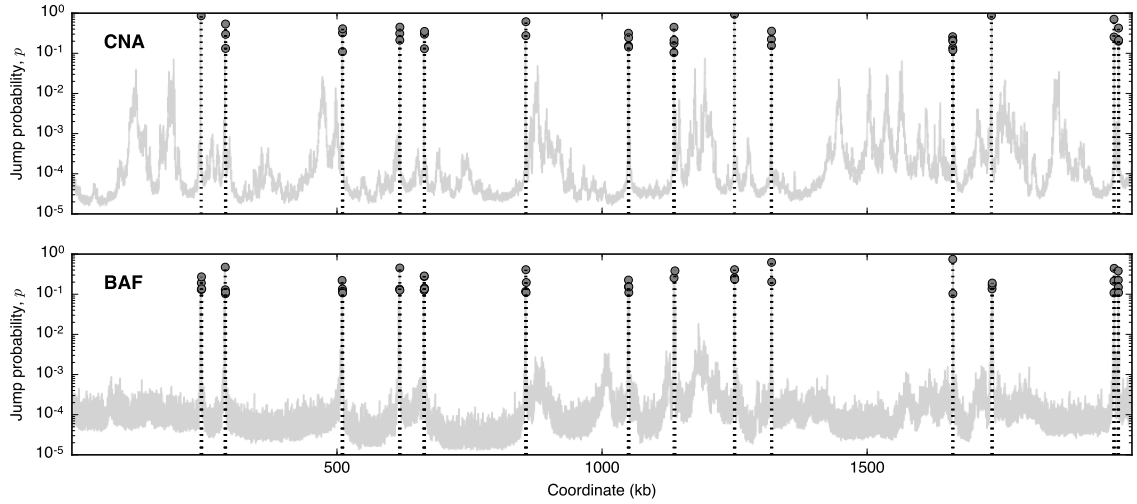


Fig. 3.5 Global parameters of the Hidden Markov Model for data filtering. The jump probability p per base (y-axis) is shown in logarithmic scale by genomic position (x-axis). CNA jumps are shown on the top panel and BAF jumps on the bottom panel. Loci with jump probability $p > 0.01$, which are indicated by pins, coincide with the location of the jumps.

To sum up, we have used continuous state-space HMMs to: (i) learn the global parameters $(p, \sigma, \varepsilon, C)$; (ii) get the posterior distribution $\gamma_i(x) = P\left(x_i \mid X\right)$ at each locus; and (iii) determine the posterior jump probability for each transition. But the hidden state x in the data filtering step has told us nothing so far about subclonality.

3.6 Discrete state-space HMM for subclonal reconstruction

We will now define a new HMM to carry out the subclonal reconstruction. The properties of each subclone's genome (copy-number state, genotype) can be encoded in hidden variables $x = \{x_i\}_{i=1\dots L}$ defined at loci i along the genome of length L. The hidden variables are properties of each of the subclones and can be either *correlated* along their genome – constituting a Markov chain – or *uncorrelated* – representing a point process. Typically, hidden variables are real or integer, and usually positive. These hidden variables are directly related to the set of directly observed data $X = \{X_i\}_{i=1\dots L}$, where X_i is usually an integer (e.g., read depth for CNA data or number of variant alleles for SNV data).

3.6.1 State space

So what is the connection between the data emission rate defined in our HMM for data filtering and subclonal genomes? What is the hidden state in the new HMM for subclonal reconstruction? We now give formal definitions of the hidden states and the global parameters of our model for subclonal population structure.

The **hidden states** for each of the subclones include their total and minor copy numbers and their SNV genotype:

- *Total copy number.* The number of chromosome copies c_i across N_c subclones is:

$$\boldsymbol{c}_i = \{c_{ij}\}_{j=1\dots N_c} = \left(c_{i1}, c_{i2} \dots, c_{iN_c}\right) \quad \text{where} \quad i = 1, \dots, L$$

  Increasing the state space by increasing the maximum number of copies $c^{\max}$ substantially scales up model complexity. For N_c subclones, the number of copy-number states is $(c^{\max} + 1)^{N_c}$. However, gains of 5 or more copies are rare, hence we will typically limit the state space to a small copy-number spectrum, e.g., $c^{\max} = 5$.

- *Minor copy number.* The number of copies of the minor allele b_i across N_c subclones, defined at heterozygous loci, is:

$$\boldsymbol{b}_i = \{b_{ij}\}_{j=1\dots N_c} = \left(b_{i1}, b_{i2} \dots, b_{iN_c}\right) \quad \text{where} \quad b_{ij} \leq c_{ij}$$

  The number of mutated copies b_{ij} in subclone j is therefore bounded by the total copy number c_{ij} at each locus i.

- *Genotype.* The SNV genotypes of the subclones g_{ij} represent how many copies of a mutation are present in locus i of subclone j. The maximum number of mutated copies for each g_{ij} is the total copy number c_{ij}.

$$\boldsymbol{g}_i = \left\{g_{ij}\right\}_{j=1\ldots N_c} = \left(g_{i1}, g_{i2} \ldots, g_{iN_c}\right) \quad \text{where} \quad g_{ij} \in \left\{0, \ldots, c_{ij}\right\}$$

  We note that the maximum copy number limit does not have to be the same as the c_{ij} described earlier (see Section 3.7 for a discussion on CNA-based and SNV-based subclone reconstruction and copy number).

The **global parameters** $\left\{M^s, \boldsymbol{f}^s\right\}$ are jointly inferred:

- *Mass.* The mass M^s, or sequencing yield per chromosome copy, is the gauge that relates copy number to the mean sequencing depth.

- *Subclonal fraction.* Evolutionary changes in population composition across samples (e.g., in time or space) are indicated by different fractions of subclonal cells f_j^s, which are defined as:

$$\boldsymbol{f}^s = \left(f_1^s \ldots, f_{N_c}^s\right) \quad \text{where} \quad s = 1, \ldots, N_s$$

  The total fraction of subclonal cells in sample s, also referred to as purity, is then $F^s = \sum_{j=1}^{n} f_j^s \leq 1$. We will use the notation f_j^s to denote the subclone fractions, but depending on the data layers used, the partition of different subpopulations need not coincide (see Figure 3.2). We will return to discuss the self-consistency issue of CNA-based and SNV-based subclone fractions in Section 3.7.

We note that the joint inference of both mass M^s and subclonal fractions f_j^s can give rise to degenerate solutions using only CNA data. This is especially critical if not all copy-number states are occupied in a subclone. As an example, assume a fully clonal population ($N_c = 1$) with $F^s = 0.7$ and a copy-number profile visiting states $c_{i1} \in (1, 2, 3)$. There is also an alternative explanation for this scenario with a purity of $F^s = 0.35$ and $c_{i1} \in (0, 2, 4)$, resulting in exactly the same likelihood. In general, there can be several degenerate explanations of the CNA data that trade higher mass for smaller subclone fractions at different copy numbers or vice versa. In the presence of these degeneracies, the mass M^s is a critical gauging parameter that needs to be reliably estimated.

3.6.2 Inference of clonal composition from copy number

As a first step, the read depth data of mutant samples can be analysed with a jump-diffusion Poisson filter (or negative binomial), defined in Section 3.5. This accomplishes two tasks: firstly, the global parameters p (jump probability), C (shape parameter) and ϵ (random error rate) can be estimated and fixed hereafter; secondly, if a matched wild-type sample without a mutant component is available, the mean sequencing depth of the wild-type sample will not be constant along the genome, but will show modulations due to sequencing and mapping effects. This can be incorporated into a non-trivial weight function χ_i, which we refer to as the bias field (with $\langle \chi_i \rangle = 1$).

Suppose the reads come from a mixture of wild-type cells and cells from N_c subclones with fractions $\left\{ f_j^s \right\}_{j=1\ldots N_c}$, such that $F^s \equiv \sum_{j=1}^{N_c} f_j^s \leq 1$ is the clonal purity of the sample. We denote with N_i the average number of reads at locus i. A locus needs to be large enough such that N_i and N_{i+1} can be considered statistically independent (e.g., 1 kb for read lengths of 125 bases). An emission model that describes our observations N_i and explicitly incorporates the clonal composition takes the following form:

$$N_i^s \sim \text{Pois}\left(N_i^s \,\middle|\, \chi_i \, M^s \, \langle c \rangle_i^s \right) \quad \text{with} \quad \langle c \rangle_i^s \equiv c_0 (1 - F^s) + \sum_{j=1}^{N_c} c_{ij} \, f_j^s \tag{3.18}$$

Here, the parameter M^s is the mean mass – or sequencing yield – per chromosome copy of DNA, and is a function of sequencing depth. The hidden variables $\left\{ c_{ij} \right\} \in \mathbf{N}_0^{N_c}$ are discrete, and they correspond to the copy-number states of each of the N_c subclones. Each follows a pure jump process along the genome, with jump probability p per base. The function χ_i captures the modulation of the read depth profile that affects wild-type and mutant DNA alike and reflects mappability issues, biases due to origins of replication or other effects across the genome. We denote with $\langle c \rangle_i^s$ the mean copy number of locus i, which is a weighted average between the fraction of wild-type cells with copy number c_0 and the fraction of mutant subclones with their respective copy-number profiles c_{ij}. The normal copy number for human DNA is diploid ($c_0 = 2$) for the autosomes and female sex chromosomes, or haploid ($c_0 = 1$) for male sex chromosomes X and Y. For yeast DNA, normal copy number is typically haploid ($c_0 = 1$) or diploid ($c_0 = 2$).

As we mentioned earlier, CNA and BAF data are probabilistically segmented during data filtering, allowing the subclonal inference HMMs to make state transitions only at potential jump sites. At segmented sites with jump probability $\hat{\pi}_i$, the transition probability between

states $\boldsymbol{c}$ and $\boldsymbol{c}'$ is:

$$P\left(\boldsymbol{c}_i' \,\middle|\, \boldsymbol{c}_{i-1}\right) = \hat{\pi}_i W_{\boldsymbol{c}',\boldsymbol{c}} + (1-\hat{\pi}_i) I \tag{3.19}$$

where $W_{\boldsymbol{c}',\boldsymbol{c}}$ is the transition matrix between states $\boldsymbol{c}$ and $\boldsymbol{c}'$, and I is the identity matrix. In the construction of the transition matrix $W_{\boldsymbol{c}',\boldsymbol{c}}$ we limit the transitions between states c and c' to a single state change, allowing both nested and independent events to occur in multiple subclones and avoiding compensatory copy-number changes. For example, in a population with $N_c = 3$ subclones, if the copy-number state of the subclones is $\boldsymbol{c} = (2,2,1)$ at position i, then a transition to a new copy-number state at $i+1$ may only take values $\boldsymbol{c}' = (2,3,1)$, $\boldsymbol{c}' = (3,2,1)$ or $\boldsymbol{c}' = (3,3,1)$. This constraint forces the Markov chain, represented in the transition matrix $W_{\boldsymbol{c}',\boldsymbol{c}}$, to be restricted to diagonal matrix entries and to one banded diagonal.

Does it work for simulated data? This is a minimum requirement for any model inference algorithm. With simulated data, we know the hidden state to compare against, and in the

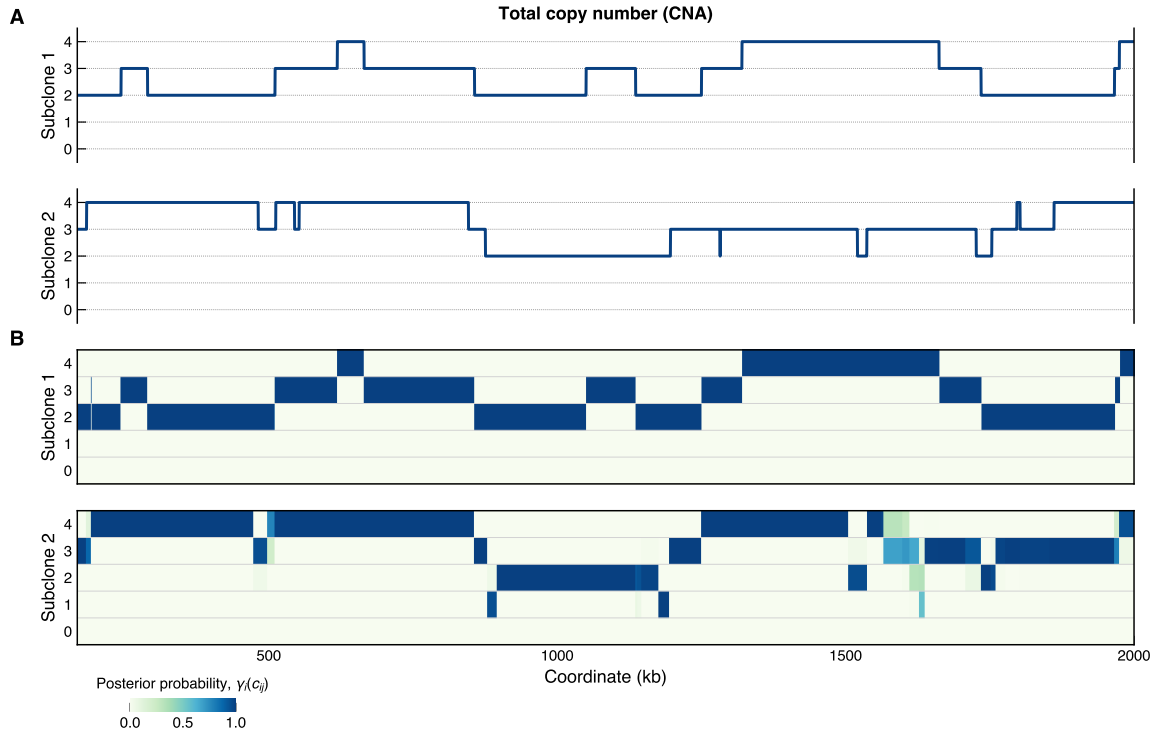


Fig. 3.6 Subclone-specific total copy number. This example shows the reconstruction of total copy number in a simulated example with two subclones (see Figure 3.4). **(A)** True subclonal total copy number states, c_{ij}. **(B)** The posterior probability $\gamma_i(\boldsymbol{c}_i)$ for each subclone i is shown, with subclone 1 (top) and subclone 2 (bottom). The inferred copy number states closely match the true profiles.

limit of infinite data size we should be able to recover the true state. Figure 3.6 shows the true and inferred total copy number states for our simulated example with $N_c = 2$ subclones. The inferred subclone-specific profiles closely follow the true profiles.

3.6.3 Inference of clonal composition from minor allele imbalances

The second piece of information used to infer clonal composition comes from heterozygous SNVs in diploid (or more generally polyploid) chromosomes ($c_0 > 1$). The emission model for observations (n_i^s, N_i^s) is

$$n_i^s \sim \text{Bin}\left(n_i^s \,\middle|\, N_i^s, x_i^s\right) + \text{Bin}\left(N_i^s - n_i^s \,\middle|\, N_i^s, x_i^s\right) \tag{3.20}$$

$$\text{where} \quad x_i^s \equiv \frac{\langle b \rangle_i^s}{\langle c \rangle_i^s}, \quad \text{and} \quad \langle b \rangle_i^s \equiv (1 - F^s)\, b_0 + \sum_{j=1}^{N_c} b_{ij}\, f_j^s \tag{3.21}$$

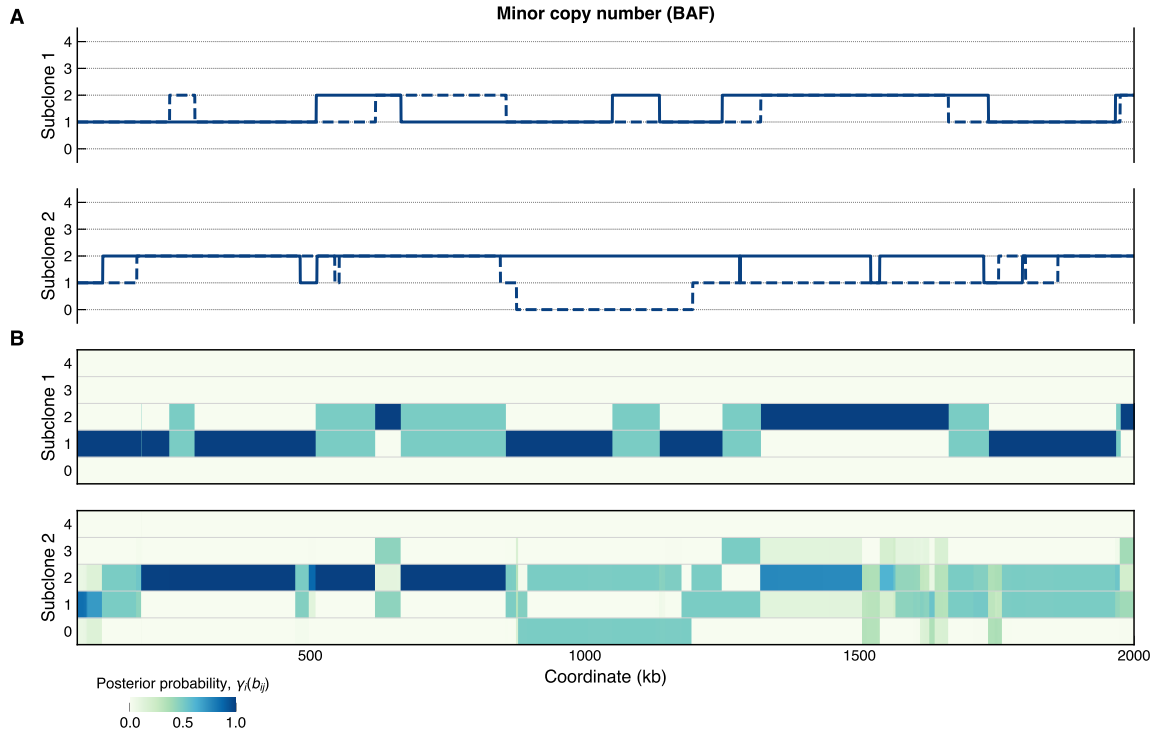


Fig. 3.7 Subclone-specific minor copy number. This example shows the reconstruction of minor copy number in a simulated example with two subclones (see Figure 3.4). **(A)** True subclonal minor copy number states, b_{ij}. Copies inherited from each parent are shown as either solid or dashed lines. **(B)** The posterior probability $\gamma_i(\boldsymbol{b}_i)$ for each subclone i is shown, with subclone 1 (top) and subclone 2 (bottom). The inferred copy number states closely match the true profiles.

where n_i out of N_i reads reporting a variant allele at locus i stem either from the maternal or the paternal chromosome. The mean number of copies of the minor allele at locus i is $\langle b \rangle_i^s$. If there are no aberrations, both the maternal and paternal copies are kept ($b_0 = 1$).

The transition model is defined by the propagator:

$$P\left(\boldsymbol{b}_i' \,\middle|\, \boldsymbol{b}_{i-1}\right) = \hat{\pi}_i W_{b',b} + (1 - \hat{\pi}_i) I \tag{3.22}$$

where $W_{b',b}$ is the transition matrix between states $\boldsymbol{b}$ and $\boldsymbol{b}'$, and I is the identity matrix. Unlike with CNA data, the transition matrix $W_{b',b}$ is not restricted and thus the Markov chain is fully mixing.

We now compare the subclone-specific posterior probability of the minor copy number reconstruction to the true states in our simulated example (Fig. 3.7). The true minor copy number is drawn in Figure 3.7A as a solid line, showing that the inferred estimates of the minor copy number states are almost equal to the true values. We can recapitulate the correct number of breakpoints and their location.

3.6.4 Inference of clonal composition from point mutations

A third, orthogonal piece of information to be used for the subclone inference are the point mutations found across the genome and their frequencies in the population. Most of the mutations are near-neutral 'passengers', but all these passengers carry information about the 'drivers' that cause the clonal compositions to change. Harnessing that information can help identify likely 'drivers' as mutations whose trajectory is most compatible with the change in clonal composition inferred globally.

Uncorrelated genotypes

Consider observations that report a mutation found in n_i out of N_i reads. This set of mutated loci indexed by i is detected with respect to a reference genome used for alignment. The sequencing reads originate again from a mixture of wild-type cells and mutant cells. We assume that the number of alternate reads n_i at locus i is described by the following emission

model:

$$x_i^s \approx \frac{n_i^s}{N_i^s}, \quad \text{with} \quad n_i^s \sim \text{Bin}\left(n_i^s \,|\, N_i^s,\, x_i^s\right) \tag{3.23}$$

$$\text{where} \quad x_i^s \equiv \frac{\langle g \rangle_i^s}{\langle c \rangle_i^s}, \quad \text{with} \quad \langle g \rangle_i^s \equiv (1 - F^s) g_0 + \sum_{j=1}^{N_c} g_{ij}\, f_j^s \tag{3.24}$$

where each mutation i has a true variant allele frequency x_i. Here, the subclone genotypes $g_{ij} \in \{0, \ldots, c_{ij}\}$ specify how many copies of the mutation are present in each subclone, and they are unknown. The maximum number of allele for each g_{ij} is the copy number c_{ij} of locus i in subclone j. On the numerator of our definition of x_i^s, the mean subclone genotype $\langle g \rangle_i^s$ is the average number of mutated copies at locus i in sample s, averaged over all subclones. By definition, the wild-type compartment has no SNVs ($g_0 = 0$) and this fraction of the population does not contribute to $\langle g \rangle_i^s$. This term can therefore be interpreted as the mean allele frequency of false-positive *de novo* SNVs. Each of the mutant subclone genotypes g_{ij}

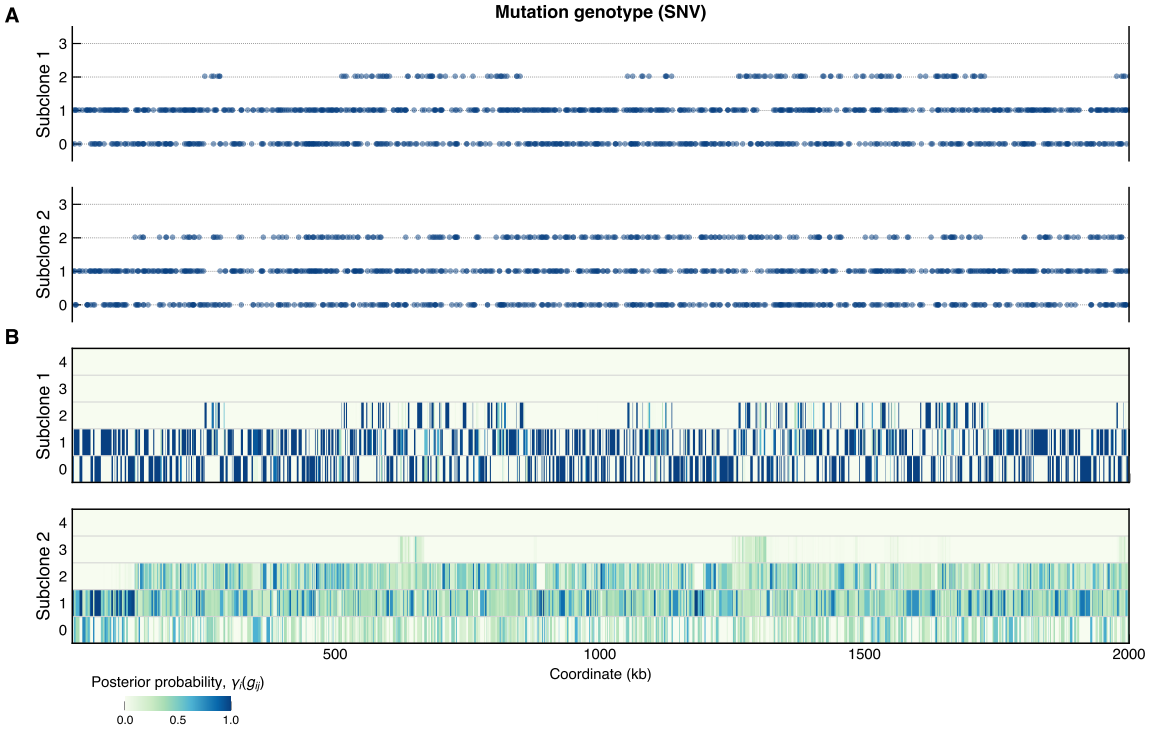


Fig. 3.8 Subclone-specific single-nucleotide variants. This example shows the reconstruction of subclonal genotypes in a simulated example with two subclones (see Figure 3.4). **(A)** True subclonal genotypes, g_{ij}. **(B)** The posterior probability $\gamma_i(\boldsymbol{g}_i)$ for each subclone i is shown, with subclone 1 (top) and subclone 2 (bottom).

contributes to $\langle g \rangle_i^s$, weighted by their frequency f_j^s in the population. On the denominator, the mean total copy number $\langle c \rangle_i^s$ is the number of chromosome copies, averaged over all subclones.

For the simulated example, Figure 3.8 shows the subclone-specific posterior probability of SNV genotypes. The largest subclone of size $f_j^s = 0.52$ offers the least challenge, as expected, and the majority of genotypes are correctly assigned. A greater challenge is posed by the small subclone of size $f_j^s = 0.07$, where the assignment of subclonal SNVs is more uncertain as indicated by the spread of the posterior distribution. The algorithm yields correct maximum likelihood estimates $\hat{f}_j^s$ in at least 90% of cases of the 20 optimisation trials obtained for each series, irrespective of starting position.

Correlated genotypes

For the sake of generality, we consider the case where mutations in the genome of a cell are physically linked and take place as a correlated process. This scenario may arise in organisms that can reproduce both sexually and asexually (e.g., yeast in the laboratory, or malaria parasites in the wild). During asexual evolution, each lineage can independently accumulate uncorrelated SNVs along the genome, and in the sexual phase long SNV haplotypes from different individuals can be brought together. This will translate into long-range correlations and sudden jumps in the SNV allele frequency. In this scenario, we can assume that the number of mutant reads n_i^s still follows a binomial distribution. The only substantial difference to the uncorrelated case is that here the genotype of a particular subclone is persistent across large regions of the genome, reflecting the haplotype structure of the population. Altogether, SNV data can then be modelled with persistence along the genome with a transition model

$$P\left(\boldsymbol{g}_i' \,\middle|\, \boldsymbol{g}_{i-1}\right) = \hat{\pi}_i W_{g',g} + (1 - \hat{\pi}_i) I \tag{3.25}$$

where $W_{g',g}$ is the transition matrix between states $\boldsymbol{g}$ and $\boldsymbol{g}'$, and I is the identity matrix. If the genotypes g are correlated along the genome, the total log-likelihood is again computed via the forward algorithm with respect to an independent jump process for each of the clonal genotypes.

3.7 Complexity and model selection

To integrate the data across multiple layers in a single sample, we must obtain the total log-likelihood of all observations. The iterative forward-algorithm provides the total log-

likelihood function of a set of observations. For correlated variables, such as CNA

$$\mathcal{L}_{\text{CNA}}(\boldsymbol{X}, \boldsymbol{f}, \boldsymbol{M}) = \sum_{i=1}^{L} \log \sum_{\boldsymbol{c}_i} \left[\prod_{s=1}^{N_s} P\left(N_i^s \mid \boldsymbol{c}_i, \boldsymbol{f}^s, M^s\right) \right] P\left(\boldsymbol{c}_i \mid \boldsymbol{X}_{1:i-1}, \boldsymbol{f}, \boldsymbol{M}\right) \tag{3.26}$$

$$\text{with} \quad \boldsymbol{f} \equiv \left\{f_j^s\right\}_{j=1\ldots N_c}, \quad \boldsymbol{M} \equiv \left\{M^s\right\}, \quad \boldsymbol{X} \equiv \left\{N_i^s\right\}_{i=1\ldots L}, \quad \boldsymbol{c}_i \equiv \left\{c_{ij}\right\}_{j=1\ldots N_c} \tag{3.27}$$

where the log-likelihood aggregates the evidence from samples $s = 1, \ldots, N_s$. The BAF log-likelihood $\mathcal{L}_{\text{BAF}}$ takes a similar form. In contrast, the last term in the sum of the SNV log-likelihood $\mathcal{L}_{\text{SNV}}$ – without persistence along the genome – is just the prior expectation for the genotype distribution $P\left(\boldsymbol{g}_i \mid \boldsymbol{X}_{1:i-1}\right) = P_0(\boldsymbol{g})$. We note that SNV-based frequencies of subclones, which we define as $\boldsymbol{f}^s_{\text{SNV}}$, reflect subclones that are defined by SNV allele frequencies only. These are to be distinguished from CNA-based subclone frequencies $\boldsymbol{f}^s_{\text{CNA}}$ that describe subclones defined by their copy-number profile. In general, the $\boldsymbol{f}^s_{\text{SNV}}$ and $\boldsymbol{f}^s_{\text{CNA}}$ fractions do not have to be identical. A single CNA-subclone might consist of two or more SNV-subclones and vice versa. It all depends on which mutations are drivers and which are passengers, and whether they are SNVs, CNAs or both, that will define the dynamic changes to the subclone fractions.

In a HMM, the main issue for model selection is the number of states. In our case, the total number of HMM states grows as $(c^{\max} + 1)^{N_c}$. By Occam's razor, we must choose the number of subclones N_c and the maximum total copy number $c^{\max}$ conservatively: as small as possible, but as large as necessary. The total log-likelihood is the objective function that we use to find the set of $\boldsymbol{f}^s$ with the highest statistical support, i.e., start with $\boldsymbol{f}^s = \{\}$, then find the best solutions $\hat{\boldsymbol{f}}^s$ for $\boldsymbol{f}^s = \{f_1^s\}$, $\boldsymbol{f}^s = \{f_1^s, f_2^s\}$ and so on. There are standard heuristics to find N_c and $c^{\max}$ systematically, e.g., the Bayesian Information Criterion (BIC), that compare goodness-of-fit with model complexity

$$\text{BIC} = 2\left(\mathcal{L}_{\text{CNA}} + \mathcal{L}_{\text{BAF}} + \mathcal{L}_{\text{SNV}}\right) - k \log\left(L_{\text{CNA}} + L_{\text{BAF}} + L_{\text{SNV}}\right), \tag{3.28}$$

where k introduces a penalty term for model complexity, s.t. $k \equiv (c^{\max}+1)^{N_c} + N_s(N_c+1)$, and L is the number of observations in each dataset (CNA, BAF, SNV). The first term in k penalises for the maximum number of copy-number states available, $c^{\max}$. The second term penalises for an increasing number of global parameters to be learned with the introduction of new subclones across N_s samples. As a guiding principle for model selection, BIC is a metric which tends to penalise complex models more heavily as the size of the genotype

space rapidly scales with the number of subclones, giving preference to simpler models. Using grid search to determine the number of subclones N_c, we have derived a heuristic value that requires an improvement of 50 log-likelihood units in BIC for the introduction of a new subclone.

We may also analyse multiple time-resolved or spatially-resolved samples N_s, which share the same N_c subclonal populations, albeit at possibly different frequencies f^s and maybe sampled at different sequencing depths so that masses M^s differ. The basic assumption is that the samples share subclones with the same copy-number profile $\{c_{ij}\}$ or SNV genotype $\{g_{ij}\}$, but with different $\left(f^s, M^s\right)$, leading to different observations $\left(n_i^s, N_i^s\right)$. The likelihood above is then the product over all N_s samples. In the context of cancer, these samples can originate from different focal points of a solid tumour or from primary and metastatic tumour or even from time-resolved samples. In the context of a drug-resistant microbial population, samples may be spatially distributed around the colony or may also be time-resolved. In general, joint analysis of several related samples helps with the clonal inference as we will demonstrate in Section 3.8.1.

3.7.1 Prior distributions

If we have some *a priori* knowledge into the system, it is straightforward to build this into HMMs. Certain subclonal genome states may be inter-dependent, such that posterior estimates of one HMM can be used as an informative prior for another HMM. Specifically, the CNA posterior distribution can capture long correlations in the total number of available genotype states and is an informative prior to the two other HMMs (BAF and SNV). The CNA posterior distribution can therefore be incorporated as a geometric prior to BAF and SNV.

The BAF genotype prior is informed by the posterior for the total copy number state $\gamma_i(\boldsymbol{c}_i)$, which ensures consistency across different data types:

$$P\left(b_{ij} \,\middle|\, \gamma_i(\boldsymbol{c}_i)\right) = \sum_{c_{ij}=0}^{c^{\max}} P\left(b_{ij} \,\middle|\, c_{ij}\right) \gamma_{ij}(c_{ij})$$

$$\text{where} \quad P\left(b \mid c\right) \equiv \begin{cases} p^{\left|b-\frac{c}{2}\right|} & \text{for} \quad 0 < p \leq 1;\ b \leq c \\ 0 & \text{otherwise} \end{cases}$$

This prior assumes that not every genotype b is equally likely. Certain combinations of aberrant chromosome numbers derived from the wild-type require a greater number of in-

termediate steps than others. For instance, with a maternal and a paternal copy of each chromosome in humans, the normal 1:1 configuration can give rise to 2:1 in a single duplication, whereas 0:3 would require one loss and two gains. The total prior probability for genotype combinations $b_i = \left\{b_{ij}\right\}_{j=1\ldots N_c}$ is then the product of above priors across subclones.

If the SNV genotypes g are uncorrelated along the genome, we can maximise the total log-likelihood in Equation (3.17) with a suitable SNV genotype prior. This prior can be uniform or it can favour particular genotypes, e.g., mutations in one copy only ($g_{ij} = 1$). The SNV inference can be informed by the posterior for the total copy number state $\gamma_i(\boldsymbol{c}_i)$ at each locus.

$$P\left(g_{ij} \,\middle|\, \gamma_i(\boldsymbol{c}_i)\right) = \sum_{c_{ij}=0}^{c^{\max}} P\left(g_{ij} \,\middle|\, c_{ij}\right) \gamma_{ij}(c_{ij})$$

$$\text{where} \quad P\left(g \mid c\right) \propto \begin{cases} p^g & \text{for} \quad g \leq c \\ 0 & \text{otherwise} \end{cases}$$

The zero genotype state $\boldsymbol{g} = 000\ldots$, where no single subclone has a mutation, corresponds to the SNV prior probability $P\left(\boldsymbol{g}_i = 0 \,\middle|\, \gamma_i(\boldsymbol{c}_i)\right)$ that a subclonal mutation is a false positive.

To build an informative SNV genotype prior we must take into account the lineage relations between subclones. Therefore, we treat the relation between subclones as a rooted tree, where the wild-type clone is taken as the root, from which the clonal mutations of the mutants are derived (Fig. 3.9). Each node in the tree represents a cluster of mutations. On the way from this root to the terminal nodes, we assume that a cell may acquire a new mutation exactly once. Each of the connecting edges then represents mutations that distinguish the child from the parent node. In the subtree below this state switch, the genotype always remains mutated. This is commonly referred to as the 'infinite sites' assumption, because mutations are assumed to only occur once.

Given the underlying tree phylogeny, we would like to determine the prior probability for a SNV to belong to one of the tree clusters. The aforementioned constraints on the tree phylogeny restrict the SNV genotype prior and thus the family of lineage trees that are possible. We can describe the SNV genotype prior $P\left(g_{ij} \,\middle|\, \gamma_i(\boldsymbol{c}_i)\right)$ as the adjacency matrix of the tree. A tree phylogeny that complies with the 'infinite sites' assumption requires, firstly, that for each child subclone i there exists a subclone j, such that $g_j \subseteq g_i$ and $\sum_{k=1}^{N_c}(g_{ik} - g_{jk}) = 1$. Secondly, the diagonal of the adjacency matrix must be equal to 1, i.e., $g_{ii} = 1$ for all $i \in N_c$. Independently of these constraints on subclonal genotype transitions, the SNV inference can still be informed by the mean total copy number for any value of $g \leq c$. We implemented

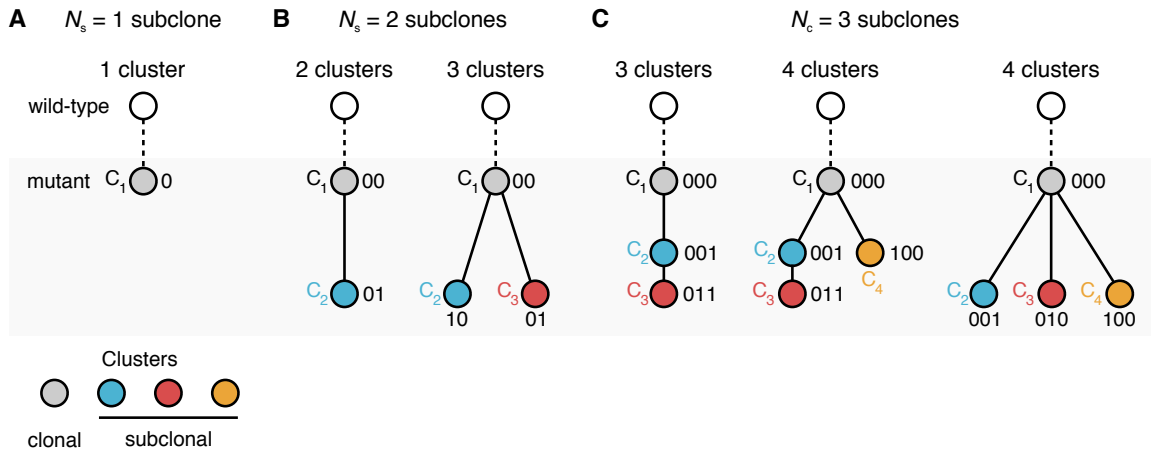


Fig. 3.9 Mapping between subclone SNV genotypes and SNV clusters. For each combination of N_c subclones, there is an equivalent number of unique SNV genotype clusters N_g. Each node in a tree is a cluster, and mutations are acquired along the edges of the tree, from the root (top) to the leaves (bottom). Each node is labelled by the cluster genotype. The ancestral node linking to the clonal cluster is also shown. SNV genotype priors that are compliant with the 'infinite sites' assumption constrain the set of tree topologies. The examples show families of trees compliant with this assumption relating **(A)** $N_c = 1$, **(B)** $N_c = 2$, and **(C)** $N_c = 3$ subclones, each of which contains one or several mutation clusters. $N_c = 1$ corresponds to a scenario where there are no subclones in the sample, so all mutations belong to the clonal cluster denoted as a grey node. The $N_c = 2$ scenario has a single subclone corresponding to one clonal and one or several subclonal clusters. The $N_c = 3$ scenario has two subclones corresponding to either one or two subclonal clusters (red and blue) describing linear or branched evolution. For example, with $N_c = 2$ subclones, genotype $g = 00$ denotes all clonal mutations, $g = 10$ mutations that are private to the blue cluster and $g = 01$ mutations that are private to the red cluster.

the set of SNV genotype priors shown in Figure 3.9. Resolving the complete space of tree topologies is currently limited by the resolution of the data, but it may be possible to implement a generative model of tree structures in the future, instead of constraints on the prior, which would require an integration over all possible trees.

3.8 Reconstruction performance on simulated and real data

We would like to evaluate the performance of our algorithm in reconstructing three features: (i) the total number of subclones, (ii) their subclonal frequency, and (iii) the accuracy of posterior estimates of subclonal genotypes. The benchmark will consist of a simulated dataset that incorporates multiple data layers and sampling, and a real dataset which covers a broad range of subclonal population structures. We consider two main performance measures for the simulated dataset: normalised error in subclone copy number and SNV genotypes, and

absolute error in subclone frequencies per sample. We defined the normalised error as the amount of posterior probability per locus not assigned to its true state. For instance, normalised errors close to zero mean that the algorithm has correctly reconstructed the copy-number profile or SNV genotypes for a subclone. For the real dataset, we will use the root mean squared error as our metric, as we will only be comparing assigned SNV genotype states rather than the full posterior distribution.

3.8.1 Benchmark with multiple data layers and sampling

We assessed the ability of our algorithm to recover these features of interest from simulated datasets over a range of plausible parameters. We first generated subclone genotypes by starting from a wild-type diploid founder that contains a random set of loci mutated in one copy, serving as pre-existing heterozygous variants (for B-allele counts). The second set of mutations, the *de novo* mutations (SNVs), were randomly distributed across both chromosome copies. Thirdly, the genome of the mutant subclone probabilistically acquired copy-number gains and losses of chromosomes (CNAs) chosen randomly. This subclone was then replicated to seed two subclones that independently underwent four cycles of similar dynamics of mutation and copy-number changes. The dynamics resulted in datasets where between 7-27 breakpoints across the genome had a total copy number change, $\sim 2,500$ loci contained somatic SNVs and $\sim 2,300$ loci contained B-allele variants (Table 3.1).

Table 3.1 Simulated benchmark dataset for subclonal reconstruction. This benchmark dataset comprises 100 samples, each with a wild-type and two mutant subclones. The notation used here is extensively explained in Section 3.6. Similar parameter values are used in the simulated examples in this chapter.

Data	Variable	Symbol	Value
	number of subclones	N_c	2
	mass	M^s	$\{15, 30, 120\}$
	purity	F^s	0.7
	subclonal fractions	f_j^s	$\{0.54, 0.16\}$
	random error rate	ϵ	0
CNA	number of segments	L_{CNA}	$7-27$
	copy number (wild-type)	c_0	2
	copy number (mutant)	c_{ij}	$0-4$
	shape parameter	C	100
	jump probability (or stiffness)	p	1×10^{-5}
BAF	number of sites	L_{BAF}	$\sim 2,300$
	shape parameter	C	900
	diffusion constant	σ	5×10^{-4}
	jump probability (or stiffness)	p	1×10^{-6}
SNV	number of sites	L_{SNV}	$\sim 2,500$
	shape parameter	C	200

For each parameter set, we simulated a 1 Mb region with $L = 20,000$ observations and mass $M^s = \{15, 30, 120\}$ which are the average number of reads per locus, per chromosome copy (Table 3.1). We drew Poisson random numbers with the rate being the sequencing depth multiplied by the locus-specific copy number of the cell mixture. This read depth was used to draw a variant allele (B-allele or *de novo* SNV) count with the true variant fraction from a binomial distribution at that locus. We compute maximum likelihood estimates using different numbers of subclones. Our choice of jump probability for simulated data is set to $p = 4 \times 10^{-5}$ per base.

In Figures 3.10A and 3.10B, we show that the subclone copy-number states and somatic SNV genotypes are successfully reconstructed from the simulated data, with errors in the inferred states close to the minimum achievable given the noise level. As expected, the performance increases when more samples are used (time points in the simulations). Similarly, inferring the clonal composition according to multiple data layers (e.g., CNA+BAF, CNA+BAF+SNV) consistently outperforms single data layers (CNA or SNV only). Although the normalised error in the copy number and SNV genotype posterior probabilities is a useful indicator for the overall performance of subclonal reconstruction, it has some limitations. For example, even when subclone frequencies are given, substantial uncertainty about the hidden state remains at low sequencing depths or given a small number of samples. This is especially the case for the *de novo* SNV genotype state, which in general has no persistence along the genome.

Figure 3.10C shows that the mean absolute error per sample between the true frequencies and the inferred ones is small and decreases as a function of the number of samples and with the addition of data types. This result exemplifies how closely the underlying subclonal dynamics can be learned, e.g., in time or space. Inferences where the mass was not accurately captured often show poor accuracy because the solution differs from the correct copy-number profile by an overall shift, typically by one copy.

3.8.2 Benchmark with diverse subclonal structures

Our benchmark on simulated data shows that cloneHD can successfully incorporate information across multiple data types and across correlated samples. However, it may not reflect the idiosyncratic features of real data. Furthermore, we would like to evaluate the robustness and fidelity of the algorithm as a function of parameters like the number of subclones N_c, the sample mass M^s, the sample purity F^s or the composition of subclonal frequencies f_j^s. We assessed the performance of the algorithm on an ensemble of synthetically-derived sam-

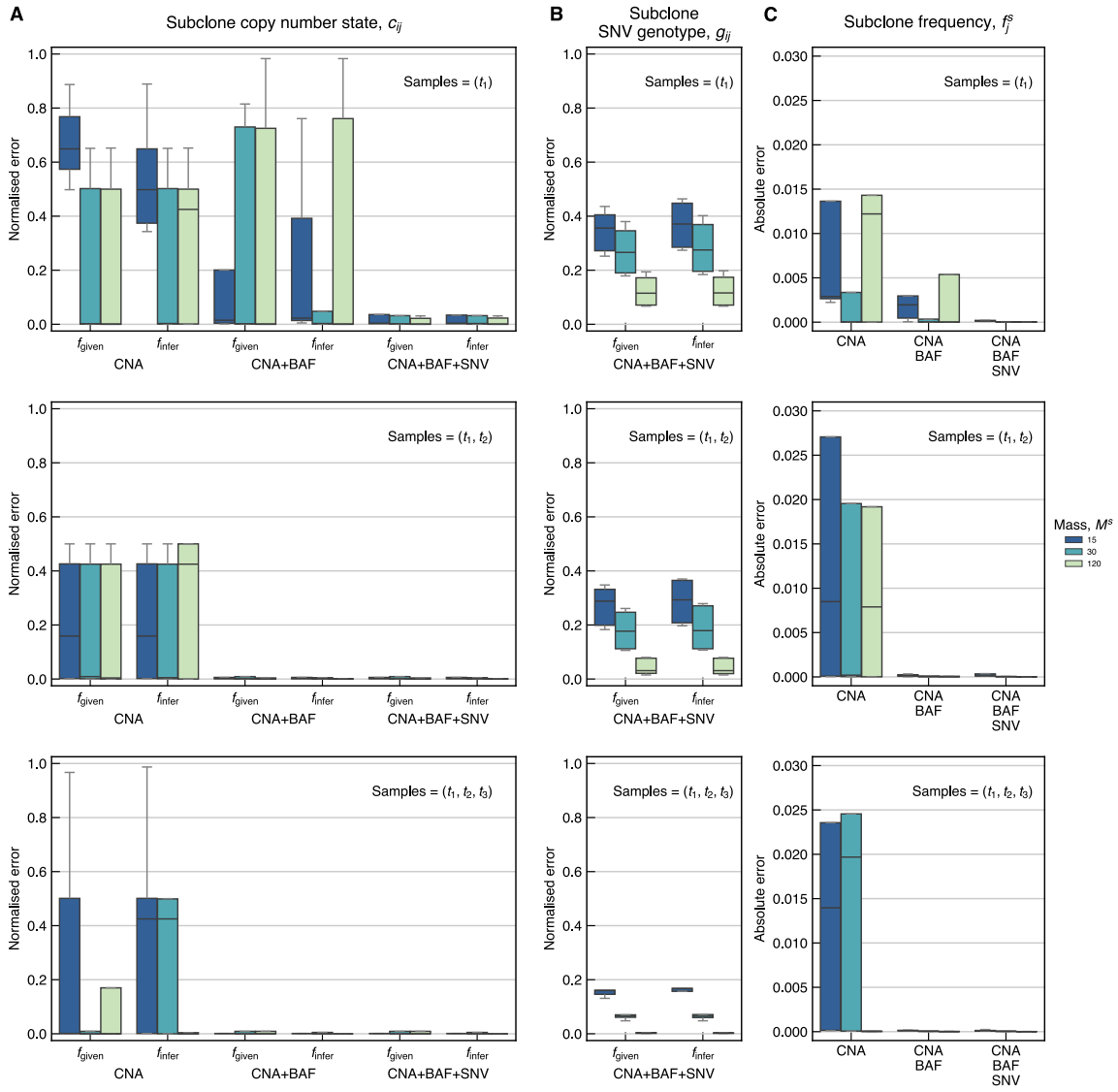


Fig. 3.10 Benchmark of reconstruction fidelity with simulated data. Subclonal inference of 100 simulated samples with $N_c = 2$ subclones and a purity $F^s = 0.7$ demonstrate a strong performance in the reconstruction of subclonal copy-number profiles, genotypes, and frequencies. Parameter values of this benchmark dataset are listed in Table 3.1. In the boxplots, horizontal black lines denote median values, areas show upper and lower quartiles and whiskers denote data within 1.5× of the interquartile range. Outlier points outside this range are not shown for clarity. **(A, B)** Normalised error in (A) copy-number state and (B) SNV genotype as a function of the number of samples (t_1, $t_1 - t_2$, $t_1 - t_3$) and the data types used in the inference (CNA, CNA+BAF, CNA+BAF+SNV), including the case where the true frequencies are given (denoted f_{given}) and inferred (denoted f_{infer}). Using more data types (e.g., CNA+BAF instead of CNA only) and using more samples each help achieve a higher performance. **(C)** Error between true and inferred subclone frequencies f_j^s, averaged over time points. The inference of subclone frequencies is increasingly more accurate with increasing number of samples.

ples with real mutation profiles, copy-number profiles, subclonal structures and evolutionary history reflecting situations in real biological tumours across various cancer types. A dataset generated as part of the ICGC Pan-Cancer Analysis of Whole Genomes (PCAWG) project is used, consisting of 965 tumour-normal pairs simulated by S. Dentro (Wellcome Trust Sanger Institute, Cambridge, UK), M. Tarabichi (Francis Crick Institute, London, UK) and I. Leshchiner (Broad Institute, Cambridge, MA). This dataset has been created as a blind test and has a ground truth, so it can be used to identify shortfalls in subclonal reconstruction algorithms and determine the conditions that affect their performance. These samples contain sequencing data from normal and tumour samples, with spiked-in germline and somatic variants, representing both clonal and subclonal single-nucleotide substitutions, small insertions, deletions, structural variants and copy-number changes (Table 3.2). A range of tumour purity and ploidy values are featured, as well as a range of mutational burdens and mutation types. We will briefly summarise how the samples have been simulated, having discussed this with the organisers of this benchmark after analysing these samples and evaluating our performance.

Firstly, the SNV cluster frequencies and the number of mutations per cluster are randomly generated. A clonal cluster was included in each simulated sample. The number of subclones in each sample is chosen randomly from the set $N_c \in \{0, 1, 2, 3, 4\}$ for samples with linear evolution, and $N_c \in \{2, 3, 4\}$ for samples with branched evolution. Cluster frequencies were chosen to be at least $f_j^s > 0.1$, at a distance of 0.1 away from the nearest cluster. The number of mutations L_{SNV} in each sample was chosen from a uniform distribution, then drawing the fraction of mutations per cluster from this distribution and randomly assigning mutations to clusters according to these fractions (see Table 3.2). Secondly, CNA profiles were chosen from samples in the ICGC PCAWG dataset. Segments with subclonal copy-number state were rounded to the nearest clonal copy-number state. These segments were then assigned

Table 3.2 Real benchmark dataset for subclonal reconstruction. The dataset includes 965 samples. The notation used here is extensively explained in Section 3.6. Each sample has a wild-type subpopulation and one or several mutant subpopulations.

Data	Variable	Symbol	Value
	number of subclones	N_c	$0-4$
	number of genotype clusters	N_g	$0-7$
	mass	M^s	$3.8-24.2$
	purity	F^s	$0.16-1.00$
	subclonal fractions	f_j^s	$0.10-1.00$
CNA	number of segments	L_{CNA}	$4\times10^3-5\times10^4$
	copy number (wild-type)	c_0	2
	copy number (mutant)	c_{ij}	$0-6$
SNV	number of sites	L_{SNV}	$1\times10^2-1\times10^5$

randomly to subclones in the same frequency space as the SNV clusters. For each sample, purity was drawn from the distribution of consensus purities across all samples in the ICGC PCAWG dataset.

Secondly, a beta-binomial model was used to simulate read depth profiles for each sample. To do this, all normal coverage profiles from the ICGC PCAWG dataset were fitted to a beta-binomial distribution. For each simulated sample, a mean read depth was chosen by sampling from the distribution of mean depths of all tumour samples in the ICGC PCAWG dataset. Read depth at each site N_i in the simulation was then randomly sampled from a beta-binomial distribution using the fitted parameters, and scaled accordingly to account for local copy-number changes. The true SNV frequency, SNV subclone genotype, and copy-number state at the site of the mutation were used to calculate the expected SNV frequency. Finally, the measured mutation count n_i for each SNV site was chosen from a binomial distribution according to the read depth N_i and the expected SNV frequency x_i at the locus.

This simulated dataset is designed to test the subclonal reconstruction performance on SNV clustering. Given that the majority of methods use SNV data with a fixed CNA segmentation, we used the consensus copy number and consensus purity as priors for cloneHD SNV clustering. We used the mean total copy number $\langle c \rangle_i^s$ and the available copy number $c^{\max}$ per

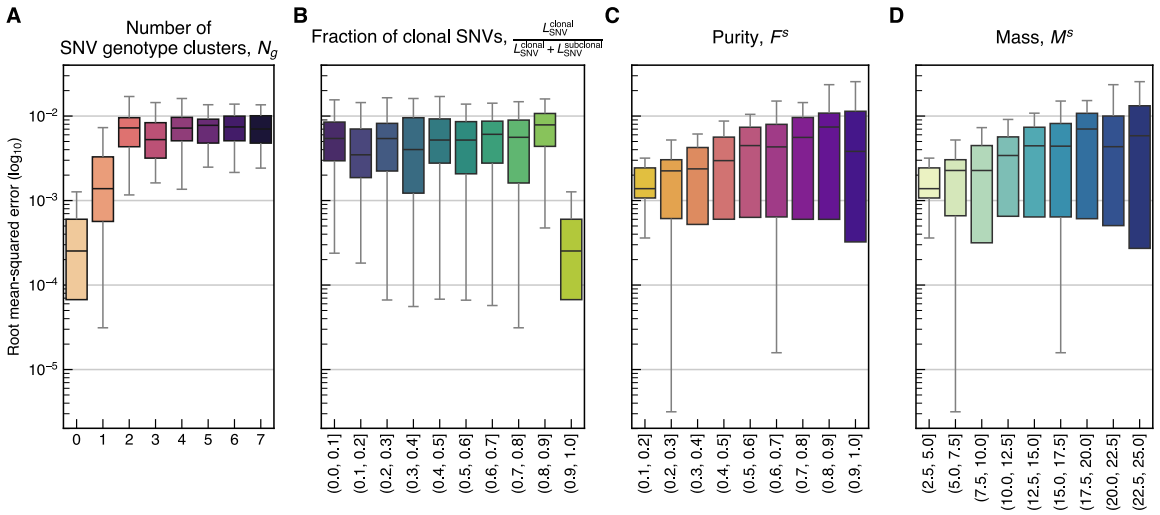


Fig. 3.11 Benchmark of reconstruction fidelity with real data. Subclonal inference of 965 synthetically-derived real samples. Each panel shows the root mean squared error between the true and inferred SNV genotype assignments in logarithmic scale (y-axis) as a function of **(A)** the number of subclones N_c, **(B)** the fraction of clonal SNVs $\frac{L_{\text{SNV}}^{\text{clonal}}}{L_{\text{SNV}}^{\text{clonal}}+L_{\text{SNV}}^{\text{subclonal}}}$, **(C)** the sample purity F^s and **(D)** the sample mass M^s (x-axis). Continuous parameters in panels (B-D) are split by quantiles (shown by colours). Parameter values of this benchmark dataset are listed in Table 3.2.

locus as informative priors to the HMM. We mapped our solutions from a representation of subclone genotypes to clusters. To remove clusters that are not supported by sufficient high-confidence SNVs, we required that the sum over the SNV posterior probability of clusters at 0.9 or higher probability corresponded to a minimum of 10 high-confidence SNVs assigned to each cluster. Similarly, other methods participated in this benchmarking exercise, including: BayClone [170], CCube, CliP, CTPsingle [171], DPclust, Phylogic, PhyloWGS [172], PyClone [157], Sclust, and SVclone [173]. Unpublished methods are described in our ICGC PCAWG manuscript [110].

To evaluate how well the different methods perform at assigning SNVs to individual clusters, we compared the number of SNVs assigned to each cluster in a sample by individual methods to the true SNV cluster assignment, using the root mean squared error between the true and inferred number of SNVs per cluster. A total of 965 samples have been scored, which we report for cloneHD in Figure 3.12. Several properties of each sample reflect the complexity of the population structure: the number of subclones N_c, the sample purity F^s and the fraction of clonal SNVs $\frac{L_{\text{SNV}}^{\text{clonal}}}{L_{\text{SNV}}^{\text{clonal}}+L_{\text{SNV}}^{\text{subclonal}}}$. The sample mass M^s reflects the resolution of the data. cloneHD performs well across a wide range of samples. As shown in Figure 3.11A, we are able to correctly assign up to 3 distinct subclone SNV genotype clusters. However, the accuracy drops for 4 or more clusters. These clusters cannot be distinguished since the frequency spectrum becomes denser as the number of subclone genotype states increases. The fraction of clonal SNVs in a population reflects the extent of subclonal heterogeneity within a sample (Fig. 3.11B), showing that SNVs are correctly assigned to clusters in samples that are mostly clonal. Those samples where the algorithm was too conservative and missed small subclones show poor scores because the solution wrongly assigns those SNV genotypes to a different subclone. Figures 3.11C and 3.11D show how the assignment of clonal and subclonal SNVs to clusters depends on the sample purity F^s and the sample mass M^s. Reconstructions are accurate when the sample purity and mass are high since SNVs have a higher representation in the data signal, although there is a wide variance in the accuracy which reflects the Poisson noise for coverage. However, samples with low purity and mass exhibit a monotonic error decrease, which is due to the fact that SNV callers do not detect variants supported by few reads (e.g., >3 reads). As a result, the lower end of the binomial distribution is undersampled, which causes all SNVs to be biased towards higher frequencies, thus resulting in SNV genotype clusters shifting up in frequency.

To determine the parameter range where methods provide robust estimates, we must identify when they begin to fail with increasing population complexity and lower resolution in the data. In Figure 3.12A, we show the normalised error between the true and inferred

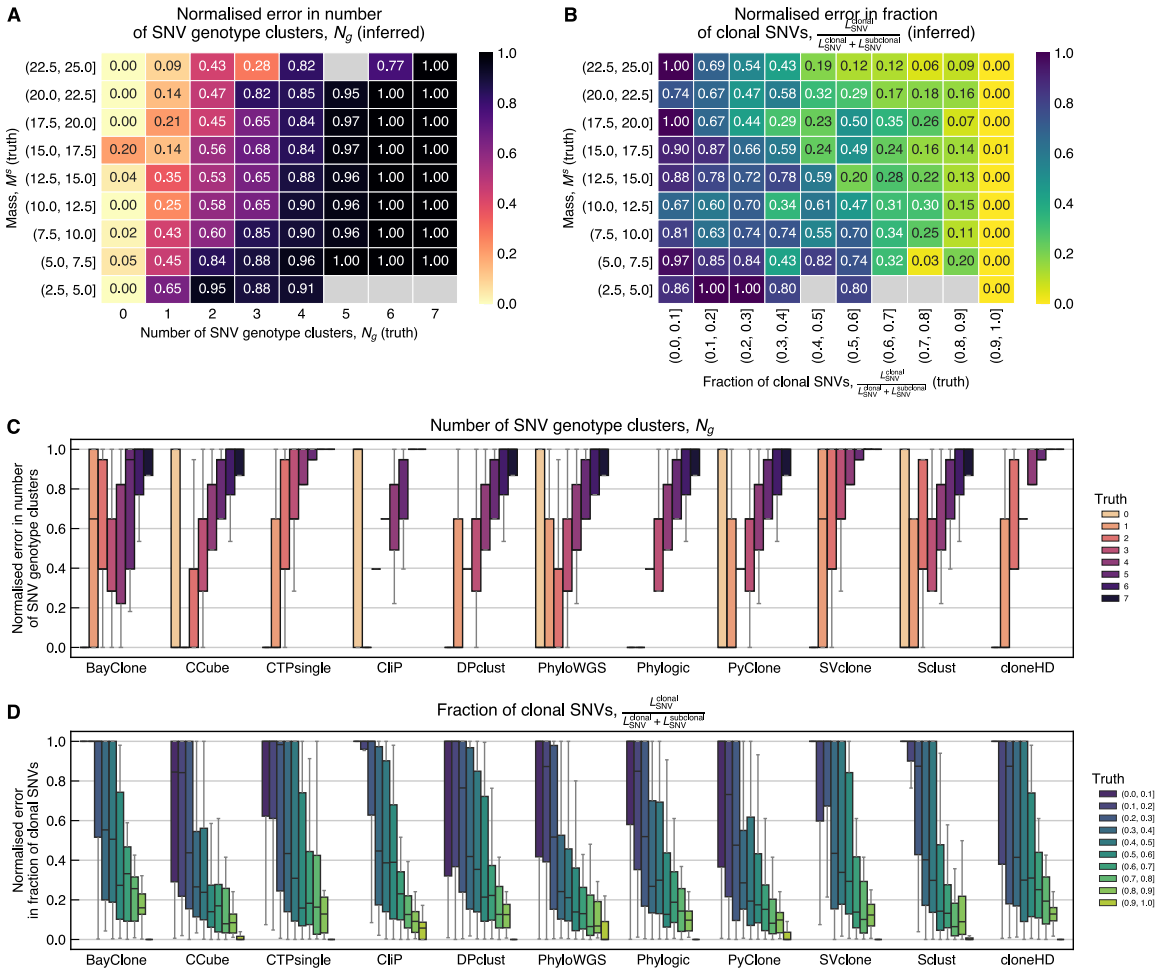


Fig. 3.12 Performance comparison of subclonal reconstruction methods on the real data benchmark (965 samples). Comparison between inferred SNV genotype assignments of individual methods and the true SNV genotype, as a function of the fraction of the number of SNV genotype clusters N_g and the mass M^s, measured relative to the ground truth. The comparison is shown for cloneHD and 10 methods developed by other research groups. Parameter values of this benchmark dataset are listed in Table 3.2. **(A, B)** Normalised difference between the true and inferred number of subclones and the true and inferred fraction of clonal SNVs. The normalised error is indicated by the colour maps. (A) Each matrix element shows the average of the normalised difference between true and inferred number of SNV genotype clusters across sets of samples, grouped by the true number of SNV genotype clusters of each sample (x-axis), and the true mass binned by quantiles (y-axis). (B) Each matrix element shows the average of the normalised difference between true and inferred fraction of clonal SNVs across sets of samples, grouped by the true fraction of clonal SNVs of each sample binned by quantiles (x-axis), and the true mass binned by quantiles (y-axis). Combinations with missing values are indicated by grey matrix elements. **(C, D)** In each panel, the methods are shown along the x-axis and normalised metrics of inferred parameters are plotted on the y-axis. Colours indicate true parameters. In the boxplots, horizontal black lines denote median values, areas show upper and lower quartiles and whiskers denote data within 1.5× of the interquartile range. Outlier points outside this range are not shown for clarity. (C) Comparison of the normalised error in the inferred number of SNV genotype clusters (y-axis), as a function of the true number of SNV genotype clusters (shown by colours). (D) Comparison of the normalised error in the inferred fraction of clonal SNVs (y-axis), measured relative to the true fraction of clonal SNVs, split by quantiles (shown by colours).

number of SNV genotype clusters for cloneHD as a function of these parameters. Errors inferring the correct number of SNV genotype clusters increase with greater population complexity and lower sample mass, which suggests a minimum sample mass of $M^s = 10$ and a maximum of $N_g = 4$ clusters that can be resolved with this data. In Figure 3.12B, the fraction of clonal SNVs reflects the ability of cloneHD to correctly group SNVs into clusters. There are typically no errors in identifying the clonal cluster and only few errors in correctly assigning SNVs to subclonal clusters in samples with >50% clonal SNVs and a sample mass of $M^s > 10$. However, cloneHD fails to identify the clonal cluster in samples with fewer clonal SNVs and lower resolution.

Overall, cloneHD is representative of the performance of other individual methods on this dataset. When looking at the absolute performance of all methods, most can resolve up to three clusters, and the precision drops for four or more as overlapping clusters with SNVs of similar frequency are difficult to resolve (Fig. 3.12C). Methods like cloneHD or DPclust are typically conservative in introducing new clusters, compared to other methods which consistently overfit the number of SNV genotype clusters (e.g., Phylogic). Samples with fully clonal or a large majority of clonal SNVs can be correctly reconstructed by most methods. Several of these methods have highest agreements when correlating the inferred fraction of clonal SNVs, including cloneHD, DPclust, PyClone or Sclust. In samples with fewer than ∼30% clonal SNVs, most methods assign mutations to the wrong cluster (Fig. 3.12D).

3.9 Summary

This chapter presented a probabilistic algorithm for modelling clonal admixtures of genomes in any asexually or somatically evolving population. First, we introduced Hidden Markov Models followed by a presentation of their use for data filtering and inference with DNA sequence data of mixed cell populations. We have shown the great benefit of performing a simultaneous analysis using several available data types (i.e., read depths, B-allele counts and SNV counts). This integrative approach resolves potential degeneracies in subclonal solutions that would otherwise be equivalent. This is further alleviated by observing the subclones at various correlated points in time or space, when subclonal frequencies differ. We showed that this framework can be used to identify and reconstruct subclonal genome states from single samples, and systematically track how populations respond to selection using time-series data. Our simulated examples provide robust estimates of the subclone fractions and subclonal copy-number and genotype states. Real benchmark datasets also highlighted common challenges that the subclonal reconstruction problem poses. Sequenc-

ing of engineered cell-line mixtures may also serve as a good benchmark in future studies (see e.g. Farahani et al. [174]). While we have identified many interesting signatures of subclonal heterogeneity and set a precedent for methods that have since incorporated CNAs and SNVs in a joint inference (see e.g., Deshwar et al. [172]), there are several important limitations to our method. Firstly, the algorithm must strike a balance between a model complexity that explains all the signal visible in the data, while not overfitting with a model that is too flexible and may find spurious solutions. Secondly, phylogenetic relations between subclones are not jointly inferred with the clonal composition, but determined *a posteriori*.

This algorithm opens many avenues for future work. Now that we move towards a new paradigm of ascribing quantitative fitness attributes to genotypes under selection, a necessary first step is to accurately measure the abundance of genetically distinct lineages. Much research remains to be done in understanding clonal dynamics in microbial evolution (e.g., immune evasion, therapy resistance) or cancer evolution (e.g., initiation, progression, localisation of origins of metastasis or prevention of drug resistance). In this respect, we have applied this method to whole-population, whole-genome sequencing data of 2,655 tumour-normal pairs corresponding to 38 cancer types, generated by the ICGC Pan-Cancer Analysis of Whole Genomes project [33, 110, 122]. Given the primordial role of positive selection in the evolution of clonal and subclonal drivers that we have observed in this dataset, further work is needed to characterise the full spectrum of subclones and their lineage relations, which will reveal the fitness distribution within tumours. Resolving the majority of small subclones that comprise the tail of the clone size distribution in any complex population will probably require alternative approaches that enable frequency measurements close to single-cell resolution, e.g., direct isolation and sequencing of single cells, targeted sequencing, or continuous barcode integration and sequencing to track individual lineages. As we will see in Chapter 5, the fitness distribution will ultimately set the stage for the evolutionary dynamics that ensue. Therefore, future work should incorporate integrative analyses of clonal genotype and fitness to build a unified view of the selective constraints on asexually or somatically evolving genomes.

Chapter 4

Population diversity and the rate of clonal evolution

4.1 Introduction

Earlier in this thesis, we investigated minimal models of evolutionary dynamics, with and without noise, and in small and large populations. Despite their differences, these systems share generic properties. We also demonstrated that DNA sequencing technologies are capable of detecting multiple genotypically distinct subclones in a population of cells. These observations suggest that testable quantitative theory is ideally complemented by targeted experiments. In this chapter, we investigate the evolutionary dynamics of populations adapting to antimicrobial drugs as a model system of rapid adaptation, to determine if they acquire mutations of independent effects (additive), or are constrained by the path-dependent effects of interactions with other mutations (epistatic). We used directed evolution in budding yeast (*S. cerevisiae*) to quantify the contingency of evolutionary trajectories, by finding differences in the adaptability of genetically heterogeneous and homogeneous populations created from a recombinant library of randomised genomes. We focus on finding the genomic determinants of evolutionary processes, particularly focusing on the action of selection in this chapter, and on mutational processes in Chapter 6. We will aim to determine whether genetic diversity enables or hinders our ability to predict which mutations will fix and to predict outcomes like the fitness increase.[1]

This work has been carried out in collaboration with V. Mustonen (V.M.) at the Wellcome Trust Sanger Institute (Cambridge, UK), E. Alonso-Pérez (E.A.-P.), J. Hallin (J.H.)

[1] Data analyses related to this chapter are available from the GitHub code repository [https://github.com/ivazquez/PhD-thesis/tree/master/Chapter4].

and J. Warringer (J.W.) at the University of Gothenburg (Sweden), and J. Li (J.L.) at the Institute for Research on Cancer and Aging of Nice (France).[1]

4.2 Genomic constraints on adaptation

Probing the dynamics of genome evolution in cellular communities, such as pathogenic infections, the gut microbiome, and cancer, is starting to improve our understanding of the role of genetic diversity in disease progression [29, 175] and drug resistance [116]. Within populations, extant genetic diversity is known to negatively impact viral pathogenesis [176] and recrudesce the severity of bacterial infections [177, 178], and plays an important and quantitatively predictable role in treatment failures for HIV [179, 180]. Genetic heterogeneity is also known to be of prognostic value for cancer progression [181] and cancer drug resistance [182]. At the phenotypic level, these systems share in common that their rate of adaptation is expected to be proportional to the fitness variability in the population, which is described by population genetic theory we discussed in Chapter 2. At the genetic level, the evolutionary success of a new beneficial mutation in a heterogeneous population will be influenced by the net fitness effect of all mutations in a cell where it randomly occurs. Cell-to-cell genetic and phenotypic heterogeneity may thus impede accurate predictions of the impact of a driver mutation before occurring in a specific cell. To understand the predictability of evolutionary outcomes at the genetic and phenotypic levels, we will first recap our current knowledge of biological systems under ongoing clonal selection.

Regularities between adapting populations have been commonly observed at the phenotypic level [183], but the underlying genotypic process differs as a function of the spectrum of escape mutations and can depend strongly on history. At the level of individual genes, the stepwise acquisition of single mutations has proven to be effective at improving a function that already exists and is accessible through a set of intermediate genotypes [83]. For example, mutations in the TEM-1 β-lactamase gene confer a wide spectrum of resistance to antibiotics used in the clinic, like penicillins, cephamycins and cephalosporins. Combinatorial libraries are enabling the synthetic construction of all mutational trajectories of TEM-1 and other genes, to assess their complete fitness landscapes [45, 184]. These studies have

[1] I.V.-G. designed the experiments, I.V.-G., E.A.-P. and J.H. carried out the experiments, E.A.-P. and J.H. maintained the populations by high-throughput pinning, I.V.-G. monitored population growth using transmissive scanning, E.A.-P. and I.V.-G. stored the sample record, J.L. extracted DNA for sequencing, I.V.-G. prepared sequencing libraries in collaboration with the sequencing pipeline at the Wellcome Trust Sanger Institute; I.V.-G. developed theory, implemented computational methods and analysed data; and I.V.-G. and V.M. interpreted the results.

drawn attention to the seemingly general property that only few mutational paths within a protein are accessible. This entrenchment and repeatability of mutational trajectories paints a promising picture, since it suggests they may be predictable. Indeed, by evolving single enzymes in the laboratory, numerous studies have shown it possible to predict TEM-1 evolutionary trajectories observed in the wild or in a clinical setting [81, 82]. Most mutational hotspots found in clinical isolates can be independently identified in the laboratory using directed evolution, which also correctly predicts mutant combinations in subsequent evolutionary paths [82].

Nonetheless, most functions involve many genes and pathways and require longer mutational jumps, with intermediate states that are neutral or deleterious. To pave the way towards a predictive theory of adaptation at the genomic scale, we must begin to characterise the properties of mutational networks beyond single genes, connecting the genomes of cells in a population. At the phenotype level, coarse-grained models try to describe the adaptive dynamics based on observations which suggest that phenotype-fitness maps are smooth, governed by 'macroscopic' epistasis [183]. These models suggest a relation of diminishing returns when new mutations arise, whereby a partially resistant genetic background will require few mutations to confer complete resistance to a drug and its mutational path will be short, while a sensitive genetic background will need to traverse a longer mutational path. In contrast, genotype-fitness maps are high-dimensional and their structure is typically very rugged, dominated by 'microscopic' epistasis [83]. The systematic interrogation of convergent genotypes allows us to probe microscopic interactions in these fitness landscapes [185]. The statistical properties of these mutational paths will be informative in modelling evolutionary outcomes [186].

Controlled time course experiments provide the means to study the balance between evolutionary forces like mutation, selection and genetic drift at the genome scale [187–189] and test the predictability of evolutionary outcomes [119]. In general, stochastic events limit predictability. For instance, a low mutation supply limits the time horizon for prediction, as waiting times to the next mutation are highly stochastic. However, neither the supply of driver mutations nor elimination by drift are limiting factors in rapidly adapting populations. Deterministic processes like selection prune the space of evolutionary paths, giving rise to recurrent mutational patterns and revealing a preferential order for interacting mutations [45, 190]. Despite the occurrence of genetic heterogeneity, only a small fraction of possible resistance genotypes appears to be frequently accessed, implying a degree of evolutionary predictability. Fluctuations in the strength of selection occur over time and space, such as during antibiotic dosing schedules [191, 192] or in spatial drug gradients [35, 193–195].

We would like to determine which of these characteristics work to the advantage of predicting adaptive genotypes, which may permit more effective pre-emptive interventions [196]. The simultaneous emergence of multiple beneficial mutations has been repeatedly observed in laboratory evolution of microbial and viral populations, suggesting that the systems under study most likely evolve under clonal interference [29, 188]. Therefore, we hypothesise that two factors play a key role in predictably accelerating resistance evolution: (i) the coexistence of genetic diversity in a population, which reduces the number of intermediate genotypes to reach a fitness peak; and (ii) the rise of clones driven by temporal and spatial gradients, which can accelerate the selection of resistant mutants by evading competition in higher concentrations and leveraging dosage increases for new mutants to enter the next selective window.

4.3 Experimental design

To examine the role of genetic diversity in clonal evolution, we combined experimental techniques of recombinant crossing, high-throughput robotic pinning, directed evolution and DNA sequencing to simultaneously adapt genetically heterogeneous and genetically homogeneous populations to multiple environments, and measure the genotypes of an ensemble of populations before and after selection (Fig. 4.1). In our experiment, we will aim to generate diversity between divergent genotypes in order to uniformly sample a fitness landscape that can be explored by directed evolution. We will focus on selective constraints imposed by antimicrobial drugs at several rate-limiting steps in the cell cycle.

Combinatorial library design

To generate the progenitor populations, we begin with two strains of budding yeast which have diverged over millions of years (*divergence* phase), that are randomly mated by meiotic recombination to generate a large pool of recombinant mosaic haplotypes (*crossing* phase), followed by isolation of single-haplotype clones (*isolation* phase). All founder populations are then evolved by selecting a fraction of the population under stress without severe bottlenecks (*selection* phase).

In the **divergence phase**, parental strains are derived from a West African strain (DBVPG6044; *MAT*α, *ura3::KanMX*, *lys2::URA3*, *ho::HphMX*) isolated from palm wine and a North American strain (YPS128; *MAT*a, *ura3::KanMX*, *ho::HphMX*) isolated from oak tree. Hereafter we refer to these strains as WA and NA, respectively. These strains are se-

lected from two distant lineages and feature 52,466 single-nucleotide differences uniformly distributed across the genome (Fig. 4.1A, left panel). Since naturally occurring deleterious mutations have been selected against over long evolutionary timescales, the parental genotypes are enriched for functional diversity which is not readily accessible using other techniques, such as random or site-directed mutagenesis.

During the **crossing phase**, our collaborators carried out 12 rounds of random mating and sporulation (meiosis) between WA and NA. The cross population (WAxNA) consists of $10^7 - 10^8$ unique haplotypes, with a pre-existing variant segregating every 230 bp on average (Fig. 4.1A, centre panel). This design results in the frequency spectrum of background mutations to be normally distributed, so that pre-existing variants are established and do not need to overcome genetic drift. We refer to the parental genotype of each individual in the cross as its genetic background.

In the **isolation phase**, a subset of segregants are randomly selected from the WAxNA F_{12} cross using single-cell bottlenecks (Fig. 4.1A, right panel). These segregants span a range of fitness in multiple environments (Fig. 4.1B, right panel).

To investigate the repeatability of mutational paths and test for path-dependent effects of selection we evolved replicate populations of the three aforementioned types of founders, shown in Figure 4.1B: divergent genotypes, a recombinant cross derived by random mating between these two divergent genotypes, and segregant isolates from this cross. To distinguish between replicates of founder populations, we label each population by founder genotype f, by the wild-type ploidy c_0, mating type m, and indexed by biological replicate r. Each of the founders is characterised by different degrees of within-population diversity – measured by the average Hamming distance between any two cells in the population, $\bar{\Delta}$ – and by the number of unique genomes (or subclones) in the population, N_c. According to these two metrics, we distinguish the following properties in the three types of founder populations:

(i) Divergent genotypes ($\bar{\Delta} = 0$, $N_c = 1$): Two wild, diverged strains of budding yeast (*S. cerevisiae*), whose genotypes differ at L sites. They comprise and ensemble of populations that are each labelled by $f \in \{\text{WA}, \text{NA}, \text{WA/NA}\}$, $c_0 \in \{1, 2\}$, $m \in \{a, \alpha, a/\alpha\}$ and $r \in \{1, \ldots, r_i, n_r = 10\}$.

(ii) Recombinant cross ($\bar{\Delta} \simeq 3.1 \times 10^4$, $N_c = 10^7$): A recombinant cross of the two diverged strains (i), where the genotypes in the population are assembled from the available alleles more or less at random such that any two individuals differ at approximately $2 \sum_{i=1}^{L} x_i(1 - x_i)$ sites (x_i being the allele frequency at locus i). Each

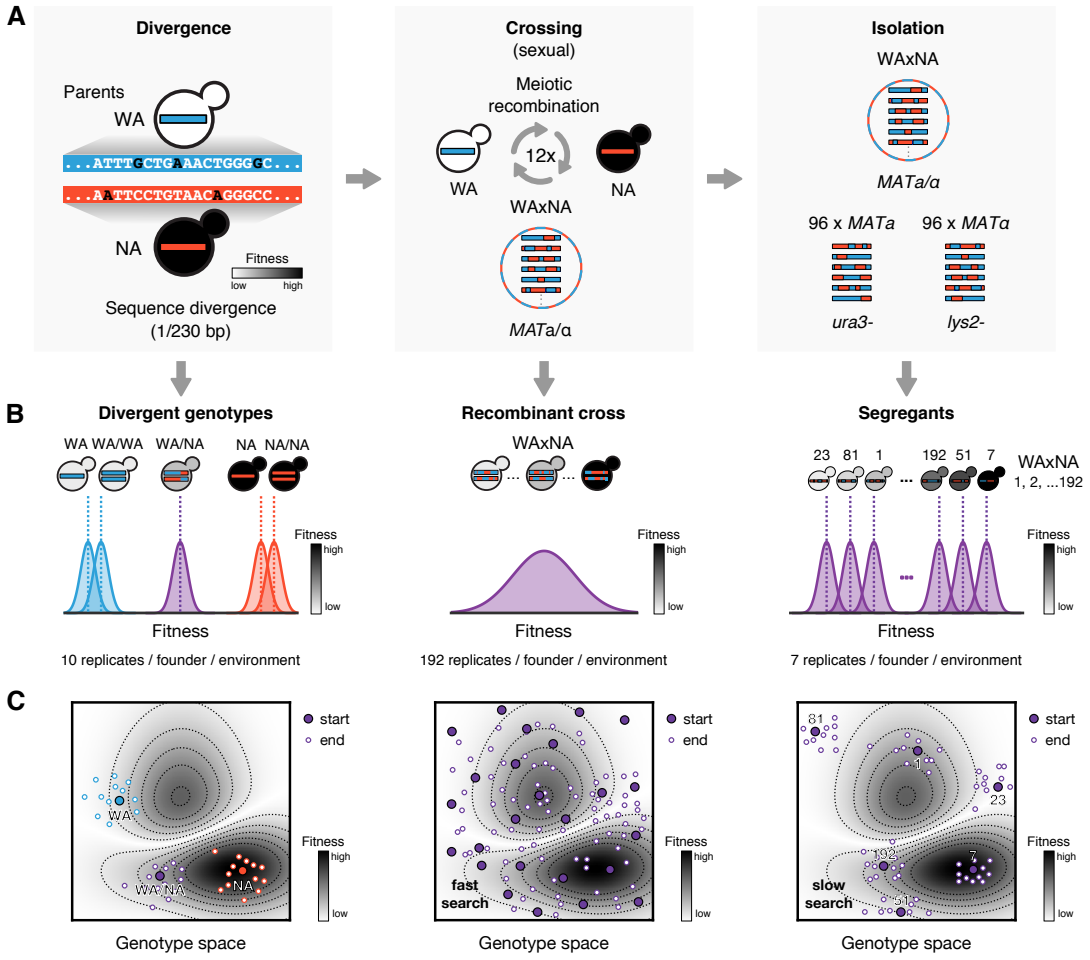


Fig. 4.1 Schematic outline of the combinatorial library design in budding yeast (*S. cerevisiae*). **(A)** The founder genotypes are derived in three different stages: divergence, crossing and isolation. Left panel: Two wild strains of budding yeast – West African (WA) and North American (NA) – have diverged for millions of years and differ at 1/230 bp in the genome. Haploid (WA, NA) and diploid versions (WA/WA, NA/NA) are used, as well as their hybrid (WA/NA). Centre panel: A recombinant cross of these two diverged genotypes, WAxNA, consists of 10^7–10^8 unique haplotypes. Right panel: We randomly selected 192 segregants from the WAxNA cross (96 × *MAT*a and 96 × *MAT*α). **(B)** Fitness distribution of the founder populations in an archetypal environment. Left panel: Diverged parental genotypes display extreme fitness differences. Haploid strains are typically fitter than diploid strains. Centre panel: Individuals from the recombinant cross inherit a phenotypic continuum that is normally distributed. Right panel: Isolate segregants from the recombinant library are chosen to span a range of fitness and start at different locations in the fitness landscape. **(C)** Schematic projection of evolutionary trajectories on sequence space. The colour gradient and the contour lines depict a fitness landscape with one local maximum (light grey) and one global maximum (dark grey). Large circles indicate the projection of founder genotypes before selection and small circles show an ensemble of replicate trajectories after selection. Strains are coloured according to their background genotype (WA: ●, NA: ●, WAxNA: ●). Left panel: Diverged genotypes that are close to a global fitness peak will require few mutations to adapt (e.g., NA), whereas those further down the slope will need to acquire more mutations (e.g., WA). Centre panel: The maximal diversity of the recombinant cross enables a fast search of sequence space towards a fitness optimum, with the fittest backgrounds requiring few intermediate genotypes. Right panel: Those same genotypes evolved independently (e.g., 7, 23, 51, 81, 192) will carry out a slower search towards a fitness optimum.

population is labelled by $f \in \{\text{WAxNA}\}$; $c_0 \in \{2\}$, $m \in \{a/\alpha\}$ and $r \in \{1, \ldots, r_i, n_r = 192\}$.

(iii) Segregants ($\bar{\Delta} = 0$, $N_c = 1$): Segregants from the recombinant cross (ii), where each population condenses into a single genotype identical to each other individual in the population, and different at $2\sum_{i=1}^{L} x_i(1 - x_i)$ sites compared to other segregants of the recombinant cross. Each population can take values $f \in \{\text{WAxNA-1}, \ldots, \text{WAxNA-192}\}$, $c_0 \in \{1, 2\}$, $m \in \{a, \alpha\}$ and $r \in \{1, \ldots, r_i, n_r = 7\}$.

When we project the genotype or genotypes in these populations on a schematic fitness landscape, we expect that divergent genotypes should be close to either peaks or troughs of the landscape (Fig. 4.1A). Populations with maximal founder diversity should have greater accessibility to fitness peaks and will require shorter paths to adapt (Fig. 4.1B). Conversely, populations with minimal founder diversity may become committed to a genotypic fate early on and thus could become entrenched at local fitness maxima (Fig. 4.1C). Qualitative features of the topography of the fitness landscape – such as its ruggedness – will influence these outcomes, which we will aim to infer from the parallel evolution of individual evolutionary trajectories. Since the number of genotypes grows exponentially with the number of loci L, we will always have $N_c \ll 2^L$, even when N_c itself is large. The large majority of genotypes in this space will therefore be unoccupied, but the coverage of the genotype space should be uniform.

Parallelised high-throughput pinning and selection

During the **selection phase**, we carry out perturbation protocols on 5,760 replicate populations of the three types of founders mentioned above, with $N = 1,152$ populations in each of five environments. We designed a randomised plate layout in 1,536-pin plate format, which is constructed from twelve 96-well plates using the Singer Instruments RoToR HDA robot (Fig. 4.2A). We maintained one in every four positions empty in each plate to control for cross-contamination (Fig. 4.2B). All populations were grown in 1,536-pin polystyrene plates (Singer Instruments, SBS-format PlusPlates). Each cycle lasted for 72 h (∼30 generations) after which the populations were transferred to new plates using transfer pads (Singer Instruments, RePads 1536 Short). Population sizes oscillate between 10^5 and 10^7 individuals in boom-bust growth cycles, so new mutations are expected every cycle.

All founders were evolved for 31 cycles (∼930 generations) subject to constant and fluctuating environments, imposed by inhibition of nucleotide synthesis (with hydroxyurea –

HU) and cellular growth (with rapamycin – RM). Hydroxyurea (also referred to as hydroxycarbamide) inhibits DNA synthesis (Fig. 4.2B). Rapamycin (also known as everolimus or sirolimus) decreases the rate of protein synthesis and the production of ribosomes, tRNAs and translation factors, halting cell growth. Each of these chemical inhibitors imposes fundamental growth trade-offs: with hydroxyurea, the cell will have to compromise polymerisation and excision of errors during DNA replication; with rapamycin, we expect a trade-off between the cell extending its replicative lifespan on the one hand, and immediate growth and cell division on the other. In addition to targeting rate-limiting steps of the cell cycle, these chemical inhibitors cover two of the most common modes of action of antibiotics and chemotherapy drugs.

In the constant environments, the concentration of selective inhibitors was maintained over time (Fig. 4.2A), both in hydroxyurea (HU-C: 2.5 mg ml^{-1}) and rapamycin (RM-C: 0.1 µg ml^{-1}). The dynamic environments impose temporal selection gradients that define different environmental epochs, with two-fold additive increments in hydroxyurea (HU-D: 2.5, 5.0, 7.5, 10, 12.5 and 15 mg ml^{-1}) and two-fold multiplicative increments in rapamycin (RM-D: 0.1, 0.2, 0.4, 0.8, 1.6 and 3.2 µg ml^{-1}). Each epoch lasts 5 cycles, or ∼150 generations. The control environment is composed of yeast nitrogen base (YNB) at 6.7 g l^{-1} (0.67%), glucose at 2%, agar at 2% and complete supplemented media (CSM) at 790 mg l^{-1} (0.079%), which is a ready-made mixture of all essential amino acids. We will generally refer to this base medium in the control environment as SC, and any stress environments contain the same mixture and the corresponding drug. Environments were maintained in 5 independent plates, with 1,152 populations per environment propagated in parallel. We carried out two independent runs of the experiment, each lasting for 31 cycles (93 days). The drug concentrations were chosen based on the dose response of the WA and NA divergent genotypes. We selected concentrations that maximised the differential growth between the two strains in each environment, resulting in a 10-fold difference between them. The solutions are kept at room temperature in glass bottles, wrapped with aluminium foil to avoid light-induced degradation. Temperature is kept constant at 30 ± 0.5°C using a thermostat. To maintain pH constant, we used a buffer solution of succinic acid, ammonium sulphate ($(NH_4)_2SO_4$) and sodium hydroxide (NaOH). This buffer solution maintains pH 5.8 throughout each cycle.

To maintain a record of the experiment that can be recalled, we stored a record of populations in glycerol. We kept complete records of cycles 0, 5, 10, 15, 20, 25, and 30 in 96-well format, as well as partial records of targeted populations of interest in every cycle. Each of the 1,536-pin experimental plates was separated into twelve 96-well plates, by pinning from

agar to liquid medium using the Singer RoToR HDA robot. After 2 days of incubation at 30 ˚C, we added 100 µl of 30% glycerol solution and samples were stored at −40 ˚C.

Genome sequencing

In order to make statistical inferences about the fitness landscape, we selected an ensemble of populations to carry out whole-genome, whole-population sequencing for comprehensive detection of many relevant classes of mutations, including base substitutions, small insertions and deletions, and copy-number aberrations. Out of 1,152 populations in each environment, we targeted the ancestral (or 'wild-type') populations and evolved (or 'mutant') populations for sequencing in cycles 0 and 30, respectively. The sequence data encompass all five environments evenly, as summarised in Table 4.1. This dataset includes 7 out of 7 replicates for the divergent genotypes (WA, NA, WA/WA, NA/NA and WA/NA), 96 out 96 replicates for two independent WAxNA recombinant crosses, and 5 out of 5 replicates for 25 out of 192 isolate segregants from the cross. In all, 1,178 out of 5,760 ancestral-evolved pairs had their whole genome sequenced. To avoid systematic biases, all population replicates taken for sequencing were matched across environments and sampled from the same plate positions. This minimises systematic biases due to spatial gradients and nutrient availability. On average, 5.2% of populations went extinct, the majority of them in the HU-D environment (Fig. 4.2A). This translated in fewer replicate populations of certain genetic backgrounds being sampled (WA and WA/WA in HU-D: 23.6% extinct).

Table 4.1 Summary of populations analysed by whole-genome sequencing. All founder populations at $t = 0$ days were sequenced before deriving replicate lines (labelled by §). Out of the total number of replicates in our experiment ('total'), we aimed to sequence a number of populations at $t = 93$ days ('seq.'). This included 7 out of 7 replicates for each of the divergent genotypes (WA, NA, WA/WA, NA/NA and WA/NA), 96 out 96 replicates for two independent recombinant crosses, and 5 out of 5 replicates for 25 out of 192 segregants from these crosses.

			Cycle 0 (0 days)		Cycle 30 (93 days)									
			All environments		HU-C		HU-D		RM-C		RM-D		SC	
Type	Founder	Ploidy	Total	Seq.	Total	Seq.	Total	Seq.	Total	Seq.	Total	Seq.	Total	Seq.
Divergent	WA	haploid	140	140§	28	11	28	1	28	11	28	11	28	11
		diploid	70	70§	13	5	13	1	13	5	13	5	13	5
	NA	haploid	140	140§	28	8	28	9	28	10	28	10	28	10
		diploid	65	65§	13	5	13	5	13	5	13	5	13	5
	WA/NA	diploid	65	65§	13	5	13	4	13	5	13	5	13	5
Recombinant	WAxNA (cross)	diploid	480	480§	96	96	96	93	96	96	96	96	96	96
Segregant	WAxNA (seg.)	haploid	4325	4325§	865	115	865	65	865	120	865	120	865	120
		diploid	475	475§	95	0	95	0	95	0	95	0	95	0
		Total	5760	5760	1152	245	1152	178	1152	252	1152	252	1152	252

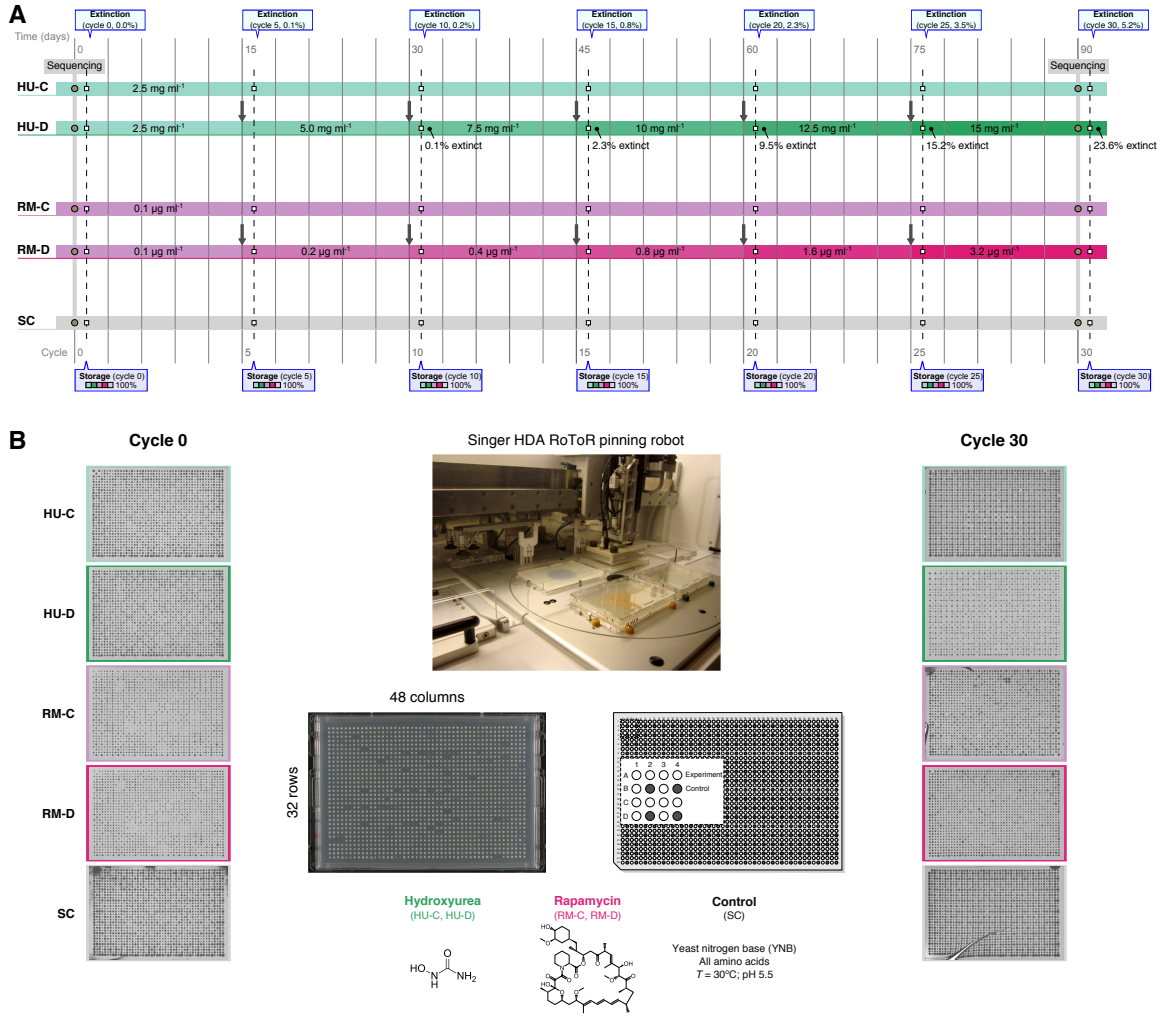


Fig. 4.2 Schematic outline of the experiment. **(A)** Timeline of the experiment: 5,760 replicate populations of budding yeast (*S. cerevisiae*) were grown in parallel for 93 days (31 cycles). Replicate populations of the founder genotypes were arranged in a master plate layout and propagated in five environments: hydroxyurea – constant (HU-C: ●), hydroxyurea – dynamic (HU-D: ●), rapamycin – constant (RM-C: ●), rapamycin – dynamic (RM-D: ●), and control (SC: ●). In the constant environments, the concentration of selective inhibitors was maintained over time, at 2.5 mg ml^{-1} for HU-C and 0.1 µg ml^{-1} for RM-C. In dynamic environments the concentration was increased in each environmental epoch, each one lasting for 5 cycles (black arrow). Cycles are delimited by thin vertical lines and environmental epochs by thick lines. Contamination tests were carried out every epoch, and a complete record of all populations was also kept every epoch, freezing all populations in glycerol and storing them at −40 °C (white squares). Whole-population, whole-genome sequencing was carried out in cycles 0 and 30 for an ensemble of populations (grey circles). **(B)** A Singer RoToR HDA pinning robot was used to maintain colonies growing on solid agar in 3-day cycles (top). The master plate layout is arranged in 1,536-pin format (bottom), with populations in 3 out of 4 colony positions (1,152 per plate) and the fourth colony kept as a spatial control (384 per plate). Time-lapse images of the 1,536-pin experimental plates before and after selection (left: cycle 0; right: cycle 30).

Genomic DNA was extracted from the samples using the 'Yeast MasterPure' kit (Epicentre, USA). Sequencing libraries were made starting from 50–500 ng of DNA using the Nextera XT library preparation kit, without selection by fragment size. The Illumina HiSeq 2500 instrument was used to sequence the DNA libraries with 2×125 bp reads. The samples were sequenced using Illumina TruSeq SBS v4 chemistry at the Wellcome Trust Sanger Institute.[1] The median number of reads per population, per base was 54× (quartiles 50×–59×, maximum 170×). The median insert size between pairs of reads was 250 bp.

4.4 Sequence analysis

To create the alignment and detect sequence variants, the pipeline works as follows: (i) sequence reads were aligned to the yeast reference genome R64-1-1 using the Burrows-Wheeler aligner [197]; (ii) PCR duplicates were removed; (iii) pre-existing single-nucleotide variants and *de novo* single-nucleotide variants, small insertions and deletions were called in ancestral-evolved pairs using MuTect2 [198]; and (iv) *de novo* copy-number gains or losses were detected in ancestral-evolved pairs (see methods in Chapter 3).

4.4.1 Single-nucleotide variants, insertions and deletions

To detect single-nucleotide variants (SNVs), and short insertions and deletions (indels), we analysed genome sequences of the founder samples and derived filters discriminating between pre-existing and *de novo* mutations.

To classify mutations as **pre-existent**, we identified sites where the WA and NA genotypes differ, which should therefore comprise a complete set of variants segregating in the cross. We limited the set of sites to single-base differences between WA and NA. For every segregating nucleotide position in the aligned sequences of a given population, we determined the number of times each base occurred at that position and the total number of reads at these loci, (n_i^t, N_i^t) at $t = 0$ days. The counts were polarised to report the WA allele at each locus, as neither the WA nor the NA strains are the reference genome. We will study the role of extant genetic variation on new mutations in Chapter 5 and we will focus on newly acquired mutations for the rest of this chapter.

We identified a high-confidence set of ***de novo*** SNVs and indels across all populations. To evaluate directly whether the events were acquired or were present in the founder pop-

[1] The sequence data are available from the European Nucleotide Archive, with study accession no. PRJEB4645.

ulations, we determined the mutation counts and the total number of reads covering these loci, (n_i^t, N_i^t), at $t = 0$ and $t = 93$ days. To avoid false positive variant calls due to sequencing errors or mutations that should be classified as pre-existent instead, we excluded sites fulfilling the following criteria: (i) variants that were monomorphic in both of the parents, (ii) heterozygous loci in either parental haploid strain, which would indicate copy-number variation, and (iii) variants with a mutation frequency $x \leq 0.1$.

These variants define datasets of mutation counts $\{g_{ij}\}_{j=1\ldots N_p}$ for a set of N_p populations at multiple loci in the genome, s.t. $i = 1 \ldots L$. A total of $L = 52,466$ pre-existing base substitutions and $L = 11,601$ *de novo* mutations were observed in $N_p = 1,178$ ancestral-evolved pairs. Over 84% of *de novo* mutations detected were supported by at least 5 mutant reads and over 92% by at least 20 reads in total. We estimated the mutation frequency at time t, $x_i^t = \frac{n_i}{\sum_j n_j}$, where n_i is the number of mutated sequences with mutation i present in the population, divided by the total number of sequences aligned at this position, $N_i = \sum_j n_j$. We must remember that the true and observed frequencies of a mutation can differ because only a limited subset of cells is sequenced at every position. This is particularly important at low frequencies, where the error in our estimates is much larger. Assuming that sampling is the only source of noise, we saw in Chapter 2 that the variance associated with a mutation with frequency x_i is $\Delta x_i^2 = \frac{x_i(1-x_i)}{N_i}$ where N_i is the coverage at position i. We therefore expect frequency errors to scale as $x^{-1/2}$ and will be particularly pronounced at low frequencies, where every additional read can significantly impact the reported frequency. In addition to this, sequencing errors are estimated at 1% per base pair. We have therefore limited our observations to mutations at frequencies above $x > 0.1$.

4.4.2 Copy-number aberrations

We detected copy-number aberrations (CNAs) by segmenting the read depth as a function of genomic position. As we described in Chapter 3, we modelled the average number of reads N_i at locus i using Poisson emissions. We define a locus large enough such that N_i and N_{i+1} can be considered statistically independent (e.g., 1 kb for read lengths of 125 bp). We learned the global parameters using the Hidden Markov Model for fuzzy segmentation and data filtering defined in Section 3.4, learning a bias field χ_i and estimating the jump probability per unit length ($p = 1 \times 10^{-8}$) and the diffusion constant ($\sigma = 1.0 \times 10^{-4}$). This takes into account persistence and correlations along the genome and allows for jumps if there are emerging subclones in the populations. We carried out bias correction between the wild-type and mutant populations by: (i) filtering wild-type and mutant CNA without bias

field, but with full jump-diffusion, i.e., finite jump probability ($p > 0$) and diffusion constant ($\sigma > 0$), and (ii) filtering wild-type and mutant with bias field (namely the wild-type posterior mean) and no diffusion ($\sigma = 0$). If the total log-likelihoods of these two steps are comparable then the bias field is shared, which normally held true for most populations. However, we observed biological variability in read depth near origins of replication in HU-C and HU-D, due to the fact that this inhibitor can arrest the cell cycle and cause cells to become synchronised at G_1 or S phase. We therefore distinguished between 'asynchronous' and 'synchronous' CNA bias fields by clustering the mean bias field $\langle \chi \rangle_i$ between populations.

We then performed the cloneHD inference to learn the subclonal structure in CNA mode, incrementally adding subclones from 0 to 4, with up to $c^{\max} = 5$ chromosome copies and accounting for the bias field. To avoid multiple local optima in the numerical optimisation, we provided initial values of the parameters and ran 20 trials and 10 restarts. We require a total log-likelihood gain greater than 20,000 units as a cut-off for the inclusion of an additional subclone. Subclone-specific copy numbers can then be estimated from the mean of the posterior probability for total copy number, $\gamma(c_{ij})$. The total copy number in subclone i can take values $c_{ij} \in \{0, 1, \ldots, c^{\max}\}$. We obtained a set of copy-number profiles $\{c_{ij}\}_{j=1\ldots N_p}$, where $i = 1, \ldots, L$ indexes a set of L segments with normal ($c_i = c_0$) or aberrant copy number ($c_i \neq c_0$). To maximise the fidelity of our copy-number reconstruction we only considered the major subclone in each population. Dropping the subclone index, the set of copy-number profiles of N_p populations is then also labelled by $j = 1, \ldots, N_p$. To find chromosome-level aberrations, we determined the total copy number difference between evolved and ancestral populations and calculated the mean total copy number, $\langle c_i - c_0 \rangle_{i \in C}$, for each chromosome C, rounding to the nearest integer. Copy-number aberrations that are part of a subclone are considered to be fixed in the population according to their subclone fraction ($F^s > 0.99$), otherwise they are considered polymorphic.

4.5 Variability in clonal evolution across populations

We first set out to investigate the influence of founder genotypes on the mutation rate and on clonal diversity, as we expect these to play a major role in the variation of mutation rates and hence on the acquisition of drug resistance. By counting differences of a population with respect to the reference, we can devise a way to score individual populations. For each of the $N_p = 1,178$ populations, the genome sequence of population p is denoted by $\boldsymbol{g}^p = \{g_1^p, \ldots, g_L^p\}$,where L is the length of the genome, i.e., the total number of loci. We

can therefore define the pairwise mismatch between any two populations as:

$$s^n\left(\boldsymbol{g}^{p_1}, \boldsymbol{g}^{p_2}\right) = \sum_{i,j=1}^{L} \left[g_{ij}^{p_1}(1 - g_{ij}^{p_2}) + (1 - g_{ij}^{p_1}) g_{ij}^{p_2} \right] \tag{4.1}$$

which is the Hamming distance between two genotype vectors $\boldsymbol{g}^{p_1}$ and $\boldsymbol{g}^{p_2}$ (i.e., the sum of pairwise allele differences across all loci). The mismatch is 0 if $\boldsymbol{g}^{p_1}$ and $\boldsymbol{g}^{p_2}$ are equal, and is positive otherwise.

The sequence of each population p can then be assigned a distance $s^n\left(\boldsymbol{g}^{p}, \boldsymbol{g}^{\text{ref}}\right)$ away from the wild-type reference genome $\boldsymbol{g}^{\text{ref}}$. As we described in Section 4.4 on sequence analysis, this is a 'consensus' reference that includes all segregating sites in the recombinant cross,

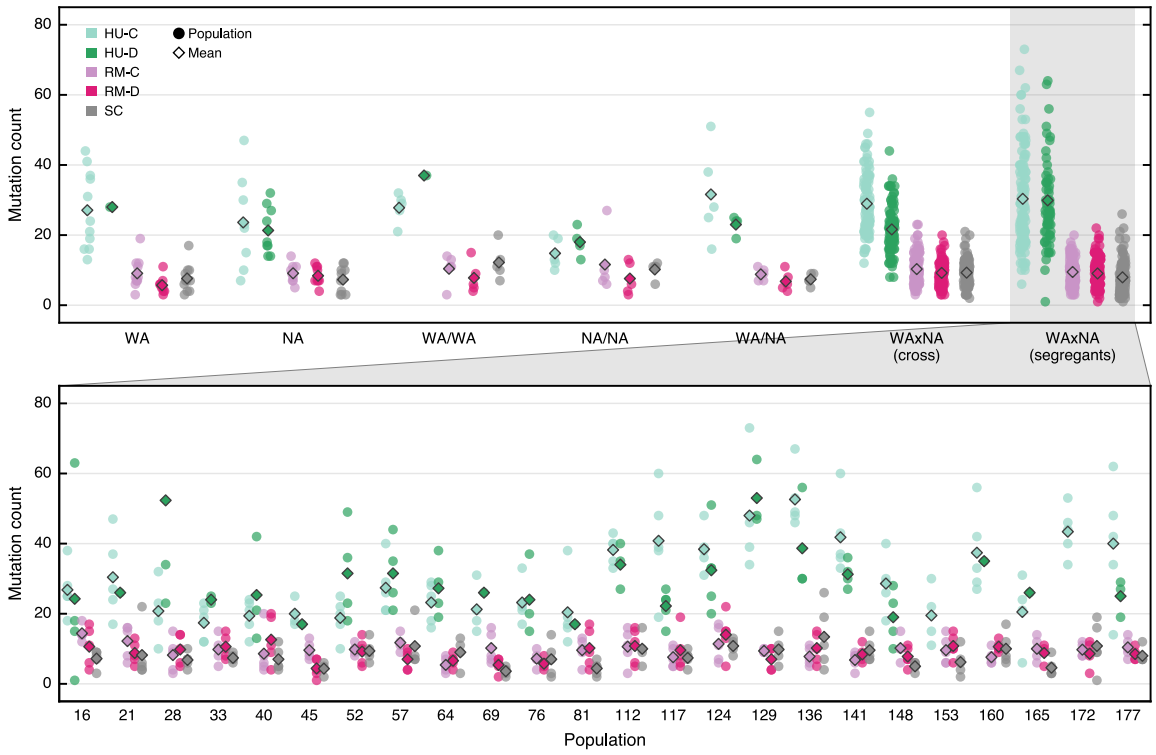


Fig. 4.3 Variation in mutation rate observed across populations. Each dot corresponds to one replicate population in one of five environments (HU-C: ●, HU-D: ●, RM-C: ●, RM-D: ● and SC: ●). The vertical position indicates the total number of mutations per population (circles) and the mean mutation count per founder (diamonds). Note that the y-axis is limited and does not show the mutation count of one outlier population (WAxNA-28 in HU-D: 100 mutations). Top panel: From left to right, the distribution of founder-specific mutation counts is shown for the divergent genotypes (WA, NA, WA/WA, NA/NA, WA/NA), the recombinant cross and segregant founders (see Figure 4.1 for a description). The counts are displayed with jitter along the horizontal axis for clarity. Bottom panel: Zoomed inset of mutations in the segregant isolates, broken down by individual founder genotypes.

such that any variation acquired before the selection phase is factored out. We therefore have a way to estimate the total number of mutations X^p acquired by a population after 30 cycles, which is given by the number of mismatches with respect to its ancestral reference, $X^p = s^n\left(\boldsymbol{g}^p, \boldsymbol{g}^{\text{ref}}\right)$. This distance metric can be equally be used for copy-number gains or losses by replacing $\boldsymbol{g}^p \rightarrow \boldsymbol{c}^p$, where distances are measured according to the number of chromosome-level copy-number aberrations, and gains or losses of a chromosome are considered as different alleles. We will make use of both metrics in the next sections.

Analysis of the founder genotypes reveals that the mean mutation rate varied by more than 2-fold across founders. There is a strong dependence between the identity of the founder and the number of mutations acquired by replicate populations derived from that founder (Fig. 4.3). The least and most mutated founders were all segregant isolates, ranging from 10 to 23 mutations per population (WAxNA-69 and WAxNA-129, respectively), which suggests that there may be heritable differences between segregants in the efficiency of endogenous repair processes.

4.5.1 Functional impact of mutations

To assess the functional relevance of mutations, we recall that the majority of mutations are expected to be, on average, neutral or mildly deleterious. We can distinguish adaptive mutations from neutral or deleterious mutations, as non-beneficial mutations should not arise and fix independently as frequently as adaptive mutations. Single-base substitutions are 10 times more frequent than insertions and deletions (see Section 4.5.2), and there is an enrichment of base substitutions in HU-C and HU-D. On average, populations acquired three times as many non-synonymous mutations (changing an amino acid) compared to synonymous mutations (not changing an amino acid), shown in Figure 4.4B. The functional impact of mutations can be estimated by combining evolutionary conservation and protein-domain information and may help us distinguish driver from passenger mutations, particularly since it will highlight substitutions that disrupt evolutionarily conserved residues in proteins (Fig. 4.4C).

In Equation (4.1) we defined a metric for the number of pairwise mismatches between the genomes of two populations, $s^n\left(\boldsymbol{g}^{p_1}, \boldsymbol{g}^{p_2}\right)$. Measuring the mismatches of population p with respect to the reference genome, we can define a score $S(\boldsymbol{g})$ over populations. This score can be defined at the level of individual nucleotides $S^n(\boldsymbol{g})$, genes $S^g(\boldsymbol{g})$, or pathways

$S^p(\boldsymbol{g})$. At the nucleotide level,

$$S^n(\boldsymbol{g}) \equiv \sum_{i=1}^{L_n} \sum_{j=1}^{N_p} s_i^n(g_{ij}). \tag{4.2}$$

where L_n is the number of uniquely mutated bases, N_p is the number of population genomes, and g_{ij} indicates the variant genotype of locus i in population j. We can form a distribution of scores $P(S(\boldsymbol{g}))$ over population sequences $\underline{\boldsymbol{g}} = \{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{N_p}\}$, i.e., g_i denotes the nucleotide at position i in a given population and it maps to a coordinate in the alignment.

Each mutation can change the coding sequence of a gene $g \in \{1, \ldots, L\}$, where $L = 5,148$ is the total number of coding genes in yeast. The nucleotide position for a mutated base maps to the corresponding gene coordinates. This enables us to define gene-level observables by adding the nucleotide-level counts per gene

$$S^g(\boldsymbol{g}) \equiv \sum_{i=1}^{L_g} \sum_{j=1}^{N_p} s_i^g(g_{ij}) = \sum_{j=1}^{N_p} \frac{1}{l_j^g} \sum_{i \in g} s_i^n(g_{ij}) \tag{4.3}$$

where L_g is the number of uniquely mutated genes, and $j = 1, \ldots, N_p$. We must scale the score by the gene length, l^g, measured in nucleotides. This will enable us to compare genes of varying length.

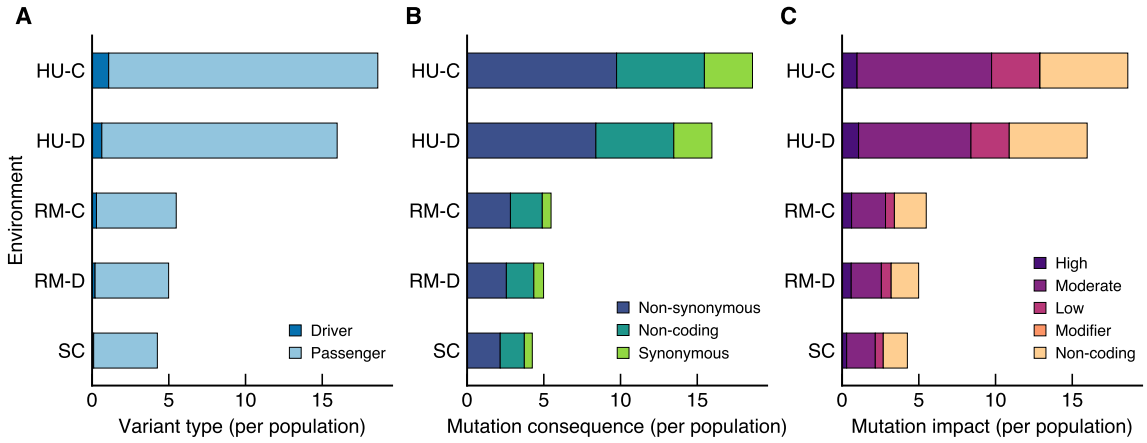


Fig. 4.4 Functional impact of mutations. **(A)** Average number of SNVs, small insertions and deletions per population. **(B)** Number of synonymous and non-synonymous mutations in coding regions, and mutations in non-coding regions, averaged per population. **(C)** The functional effect of mutations is estimated by combining evolutionary conservation and protein-domain information.

At higher levels of biological coarse graining, we can also define pathway-level observables by aggregating the gene-level counts. The mutated base is aligned in a gene sequence associated with a certain pathway

$$S^p(\boldsymbol{g}) \equiv \sum_{i=1}^{L_p}\sum_{j=1}^{N_p} s_i^p(g_{ij}) = \sum_{j=1}^{N_p}\frac{1}{l_j^p}\sum_{i\in p} s_i^g(g_{ij}) \tag{4.4}$$

where L_p is the number of uniquely mutated pathways, and $j = 1,\ldots,N_p$. In the same vein as gene-level observables, the score needs to be scaled by the mutable target size of the pathway, defined by the total mutable gene lengths of the pathway members, $l^p = \sum_{g\in p} l^g$, found in the Gene Ontology database [199, 200].

This is an additive scoring system, which corresponds to the assumption that mutations at different sites have occurred independently. With the count scores in place, we would like to know what the level of functional convergence is. We can form a null sequence set $\boldsymbol{g}^*$ assuming that a nucleotide occurs at random with probability $p(g_i^*)$, so the probability of the whole sequence is $P(S(\boldsymbol{g})) = \prod_{i=1}^{L} p(g_i^*)$. We can permute the index of the individuals j by randomising $\boldsymbol{g}$ at every position i, $g_{ij} \to g_{ip[j]}$, where $p[j]$ is a random permutation of indices $j \in [1, N_p]$. Finally, we can project the set $\boldsymbol{g}^*$ to form a distribution $P^*(S(\boldsymbol{g}))$.

The cumulative distribution of scores provides evidence for parallel evolution, with populations accumulating similar mutations in parallel (Fig. 4.5). This trend is occurring primarily at the gene and pathway levels, but not at the nucleotide level. By comparing the cumulative difference of the bins $C(S) = \sum_{S_0}^{S} P(S(\boldsymbol{g})) - P^*(S(\boldsymbol{g}))$, we can identify measurable differences between the observed and the expected distribution of count scores (Kolmogorov-Smirnov test, $P < 0.05$). In Figure 4.5B, we can observe a significant deviation of the observed gene-level count score away from the null distribution. Clearly, mutations are not random with regards to the count score at the gene level. As expected, mutations with a large negative effect and likely strongly deleterious are suppressed. If the score were uncorrelated to the true fitness contribution of the mutation, then the data would be on a vertical line. At the pathway level, the target space of our scoring model increases very rapidly with the number of mutated genes. Unlike the nucleotide-to-gene map which is one-to-one, gene-to-pathway assignments are many-to-many. This makes any conclusions about the degree of parallelism across pathways conditional on the number of members of each pathway that determine the null probability distribution.

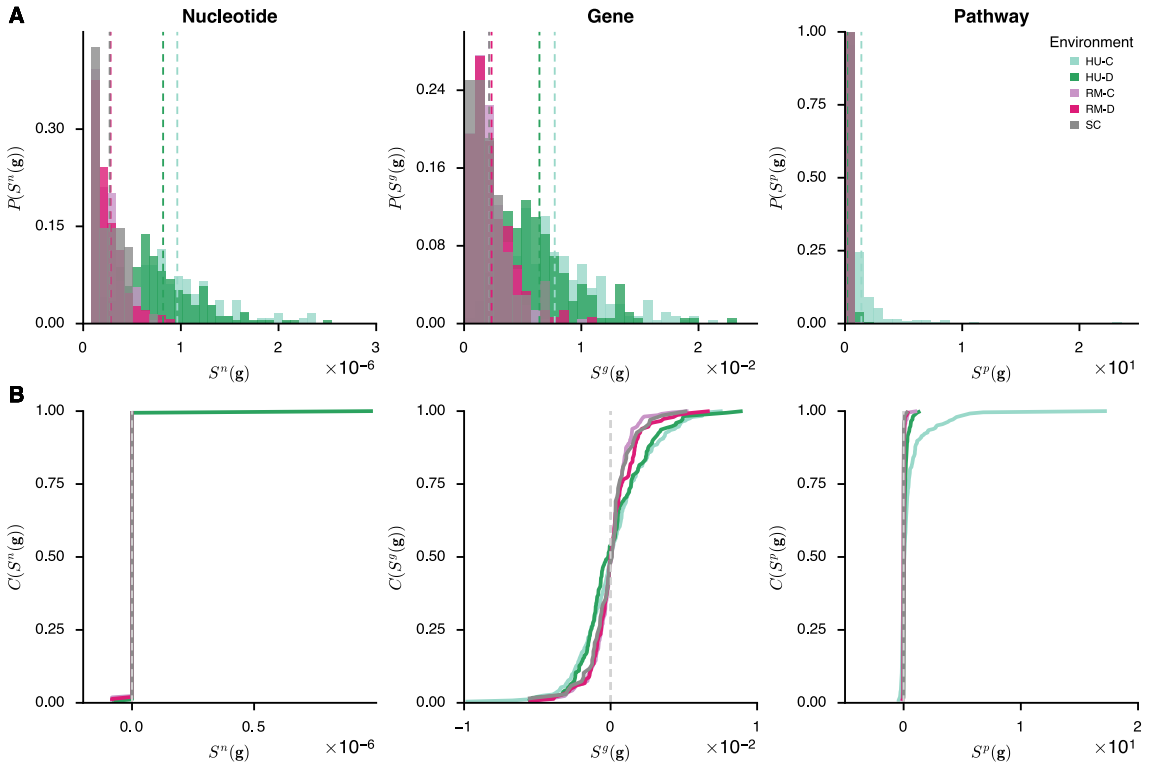


Fig. 4.5 (A) Probability distribution function of count scores for nucleotide-, gene- and pathway-level observables by environment (HU-C: ●, HU-D: ●, RM-C: ●, RM-D: ● and SC: ●). The vertical lines show the mean of each distribution. **(B)** The ranked count score, $S(\boldsymbol{g})$, (x-axis) is shown against the empirical cumulative distribution of the difference between the observed and null count scores, $C(S(\boldsymbol{g}))$ (y-axis). From left to right, the cumulative distribution of the bins, $C(S(\boldsymbol{g})) = \sum_{S_0}^{S} P(S(\boldsymbol{g})) - P^*(S(\boldsymbol{g}))$, is shown at the nucleotide, gene and pathway levels. This statistic compares the distribution of scores, $P(S(\boldsymbol{g}))$, to the null distribution expected by chance, $P^*(S(\boldsymbol{g}))$, obtained by permuting the populations. If the score is uncorrelated to the true recurrence of mutations, then the data will be on the flat vertical line.

4.5.2 Recurrence of mutations

Our scoring system suggests that gene-level observables are a useful 'atomic' unit to understand the repeatability of molecular changes. Of the 11,601 mutations we identified, 7,484 fall within coding regions. Amongst them, we can hope to distinguish driver mutations from neutral or deleterious passenger mutations by their recurrence, as non-beneficial mutations should not independently arise and fix at the same loci as frequently as driver mutations. We shall, however, keep a sceptic eye on such a gene-centric view that may lead us to ascribe intrinsic functions to a gene solely based on recurrence. We will postpone characterising potential interactions between genes to Section 4.6.2. We expect that a large number of replicate populations will increase our statistical power both to detect true driver genes under selec-

tion (sensitivity) and to distinguish them from the background of hitchhiking passengers (specificity).

If we count the number of times each gene is recurrently hit, 1,246 genes are mutated in more than one population and there is a clear excess of parallel mutations: 886 are present in two populations, 243 are present in three populations and 117 in four or more populations. We would like to test these observations against the null hypothesis that those genes recurrently mutated happen by chance. This model is equivalent to the case of comparing biased and unbiased multinomial sampling of balls of different colour from an urn. Out of k possible genes, we draw n mutations with replacement according to the multinomial distribution, where n is the number of observed coding mutations in each population. Figure 4.6 compares the observed and expected number of genes mutated k times. If these mutations are distributed over the 5,148 coding genes in yeast, we would expect only 57 genes mutated four or more populations and 15 genes mutated in five populations or above.

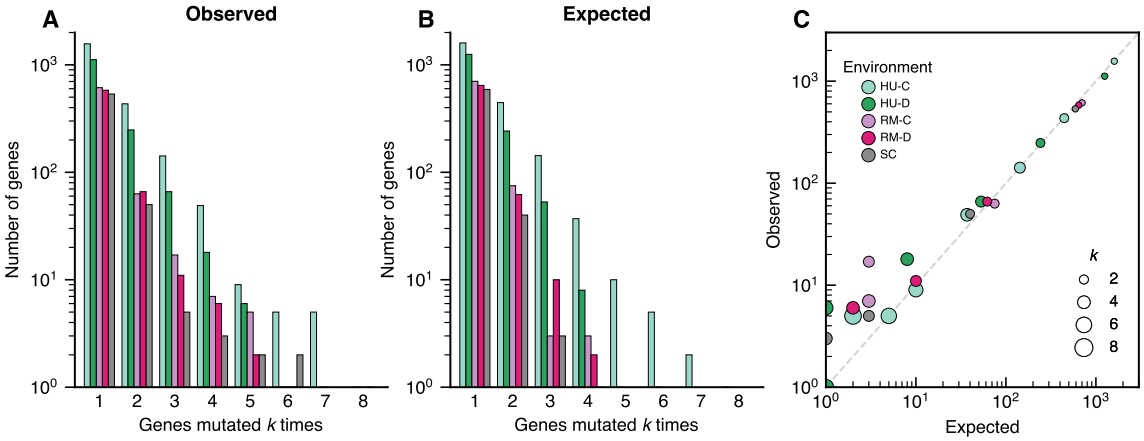


Fig. 4.6 Statistics of mutations in coding regions, showing the number of genes independently mutated in each environment (HU-C: ●, HU-D: ●, RM-C: ●, RM-D: ● and SC: ●). **(A, B)** The bar counts indicate the number of genes that are mutated k times in independent populations. Null expectations are generated under a multinomial distribution, distributing the total number of mutations per environment among all yeast open-reading frames. **(C)** Observed vs. expected on the basis of the null distribution. Circle size indicates the number of genes mutated k times in independent populations.

Multi-hit point mutations

We focus on multi-hit genes which are independently mutated in several populations and are putatively beneficial. Figure 4.7 shows the number of mutation hits observed in recurrently mutated genes. In total, there were 481 coding mutations across 117 genes independently mutated in four or more populations.

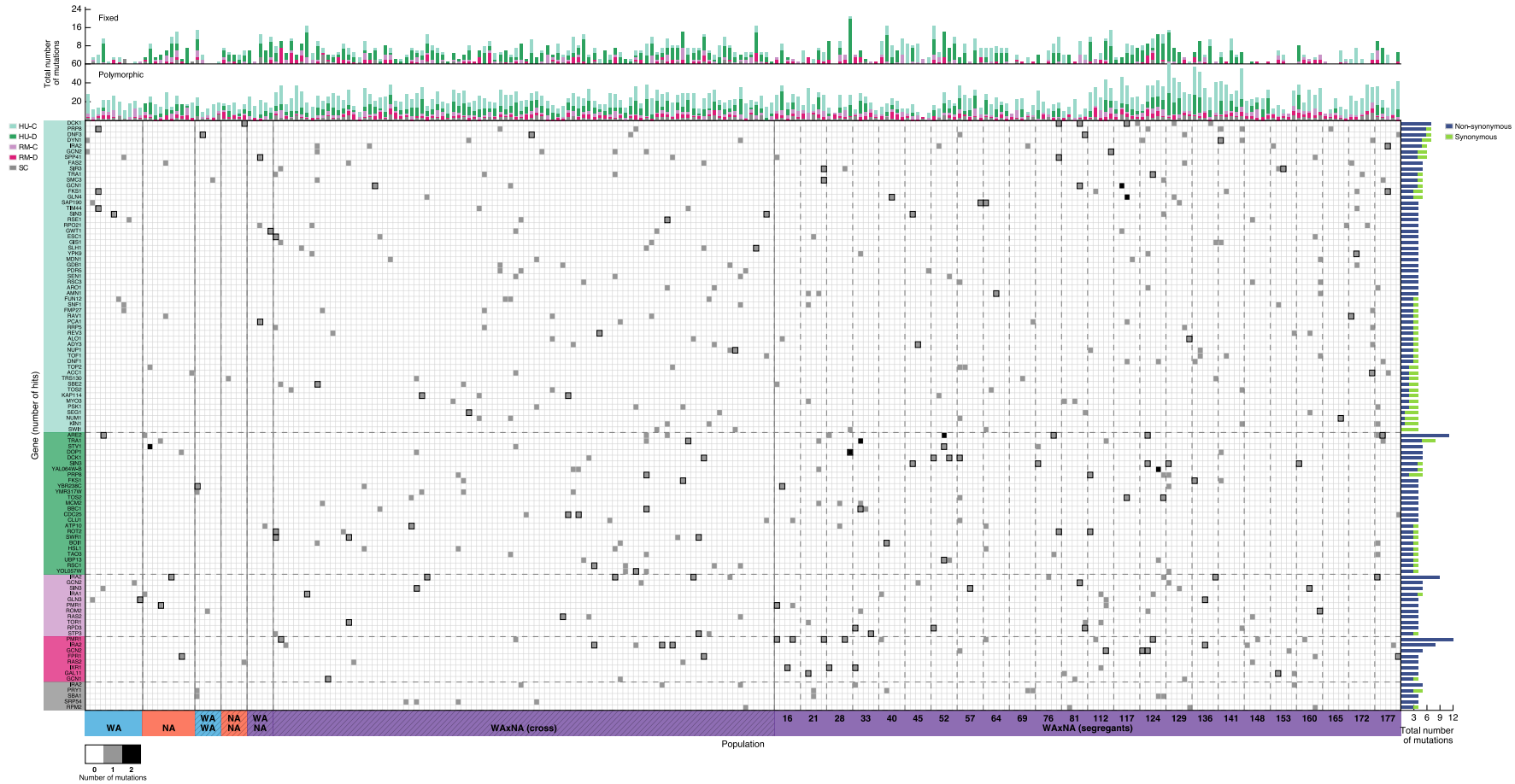


Fig. 4.7 Sequence-level evolution in constant and time-dependent fitness landscapes, showing multi-hit genes recurrently mutated in ≥4 populations. Middle: The data matrix shows the count score $s_i^g(g_{ij})$, with the number of mutation hits found in population p at each recurrently mutated gene. The colour spectrum corresponds to the estimated mutant frequency. Rows are ordered along the left axis by the number of recurrent mutations observed in each gene, as represented by the vertical scale which is delimited by environment (dashed lines; HU-C: ●, HU-D: ●, RM-C: ●, RM-D: ● and SC: ●). Columns are ordered along the bottom axis, delimited by the type of population (solid lines; WA: ●, NA: ●, WAxNA: ●) and the founder genotype (dashed lines). The wild-type copy number of the founder is shown on the bottom labels ($c_0 = 1$, solid; $c_0 = 2$, hatched). Boxed elements indicate a fixed mutation in a given population ($x > 0.49$ in haploids and $x > 0.99$ in diploids). Top: The stacked bars above the data matrix show the total number of mutations per population, broken down by fixed and polymorphic mutations, and coloured by environment. Right: The bars to the right of the data matrix report the total number of synonymous (green) and non-synonymous mutations (blue) in genes recurrently mutated in each environment. The distribution and consequences of mutations in these recurrently mutated genes are shown in Figure 4.8.

Firstly, we observed 362 coding mutations across 74 recurrently hit genes under inhibition of nucleotide synthesis by hydroxyurea. Overall, mutations in HU-C and HU-D appear to be generally good at delaying the cell cycle, slowing growth and allowing enough time for DNA repair before replication. We give a brief overview, as the spectrum of recurrent mutations is vast. Two of the most commonly mutated genes were *DCK1* and *PRP8*. *DCK1* is a guanine nucleotide exchange factor which is mutated in 12 populations. Each of these mutations were highly penetrant, going to fixation in 8 out of 12 cases, and typically affected both the conserved catalytic region of Dck1p and other domains. We also observed 11 missense variants in the protein kinase Tra1p, which is an essential gene involved in DNA repair. Two of these mutations occurred in the C-terminal region that is related to phosphatidylinositol 3-kinases [201]. Missense variants in *ARE2* were also recurrent and much more common in HU-D than HU-C (11 vs. 2 mutations). Finally, recurrent missense variantss in the *AMN1* gene may affect the exit of mitosis. Similar amino acid changes to one of the mutated residues we observe (D377H) have been reported to cause widespread gene expression changes in naturally occurring variants [202, 203].

Secondly, there were 97 coding mutations in RM-C and RM-D that were recurrently found across 15 genes, particularly targeting components of the target-of-rapamycin (TOR) pathway. SNVs introduced stop codons or frameshift deletions in *FPR1*, which is a binding partner of rapamycin that together in a complex inhibits the TOR pathway. There were 5 loss-of-function mutations of the nutrient-responsive kinase Tor1p that have been shown to increase replicative lifespan [204]. In Chapter 5, we will report validations of *FPR1* and *TOR1* performed by F. Salinas (Institute for Research on Cancer and Aging of Nice, France) that confirm their role as driver genes. Putative loss-of-function mutations in *PMR1* were the most frequent, introducing stop codons or frameshifts in the cation-transporting ATPase domain (Fig. 4.8B). *PMR1* is part of the TOR pathway and is a negative regulator of *TORC1* [205]. Therefore, mutations in *PMR1* most likely decrease or abolish the activity of the gene, and do so in a dose-dependent fashion as they were more prevalent in RM-D than RM-C (12 vs. 4 mutations). This protein is required for calcium and magnesium transport, and 13 out of 16 mutations disrupted either the Ca^{2+} or Mn^{2+}-transporting domains [206]. The Gln3p phosphatase, which is phosphorylated by the TOR kinases Tor1p and Tor2p, was mutated in 5 populations, and is known to induce rapamycin resistance [207]. Other serine/threonine protein kinases that play a role in sensing amino acid deficiency were recurrently mutated (Gcn1p, Gcn2p). Notably, mutations in these protein kinases may delay protein synthesis, as they phosphorylate the necessary factor to initiate protein translation (elF2). Finally, both members of the Rpd3-Sin3 histone deacetylase complex – which is

required for adaptation to nutrient limitation – had recurrent loss-of-function mutations that have been shown to yield them unable to carry out rapamycin-induced repression of ribosomal protein genes [208]. This may be an alternative adaptive strategy to changes in TOR signalling.

Certain genes were ubiquitously mutated in all environments, even in the absence of stress (*IRA1*, *IRA2* and *RAS2*), all of which are part of the Ras/PKA pathway and are known to regulate cell growth in response to glucose availability [75, 188, 209]. These genes are also recurrently mutated in RM-C and RM-D as inhibition of the TOR proteins by rapamycin leads to a transcriptional response similar to nutrient deprivation.

The set of recurrent mutations differs between pairs of founders (Fig. 4.7, top panel). Divergent genotypes typically follow different paths, with WA having on average 0.5 fixed

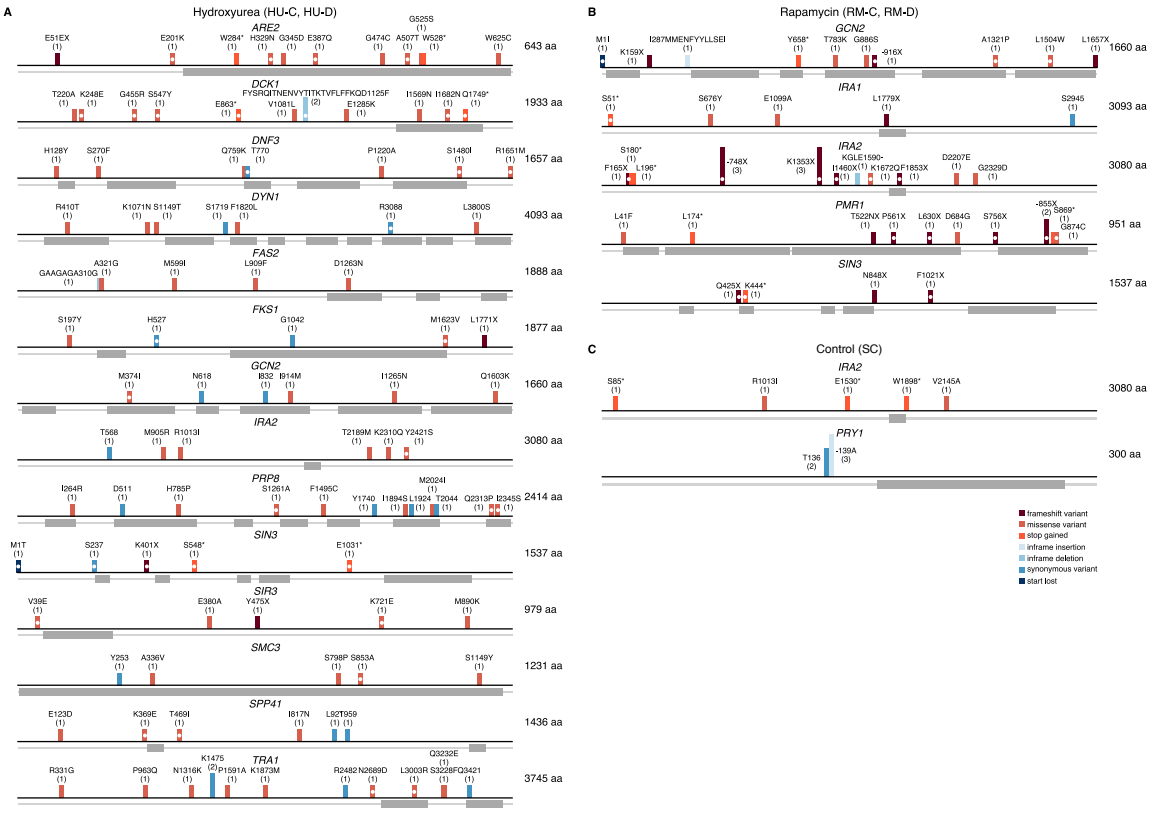


Fig. 4.8 Distribution and consequences of mutations in recurrently mutated genes. We observed 481 coding mutations across 89 genes recurrently hit in ≥5 populations in **(A)** hydroxyurea (HU-C, HU-D), **(B)** rapamycin (RM-C, RM-D) and **(C)** the control environment (SC). The bar height indicates the number of substitutions observed per codon, which is also shown in brackets. Red colours indicate non-synonymous changes; blue colours indicate synonymous changes. Fixed mutations are highlighted by white dots. The gene diagrams show the location of encoded protein domains from start to stop codon. The domain locations are annotated from Pfam release 30.0 [http://pfam.xfam.org].

mutations per population involving 8 distinct genes, while NA have on average 1.2 fixed mutations in 4 distinct genes that do not overlap with recurrent mutations in WA. Replicates of the recombinant cross are similar amongst themselves, having on average 1.1-1.5 drivers per sample distributed over 33 distinct genes. This suggests that the vast number of unique founder sequences covering the fitness landscape yield many paths accessible to adapt. Segregants however are quite variable, with the least mutated founder (WAxNA-69) having on average 0.4 fixed mutations per population involving 3 distinct multi-hit genes, while the most mutated founder (WAxNA-129) has on average 1.2 fixed mutations across 16 distinct multi-hit genes. Certain segregant genotypes show a degree of entrenchment in genotype space. Two replicates of WAxNA-117 in HU-C each acquired two missense variants in the same gene: *GCN1* (P582L, D1018Y) and *GLN4* (R375P, T492), respectively. This was even more frequent in HU-D, suggesting that dynamic regimes may limit the number of positively selected paths available: e.g., two non-synonymous *ARE2* mutations in WAxNA-52 (G525S, W625C), and similarly for *TRA1* in WAxNA-33 (K1475, S3228F), *STV1* in NA (L230*, D502H), *DOP1* in WAxNA-28 (2 × G1681V) and *YAL064W-B* in WAxNA-124 (V77, P78S).

Multi-hit copy-number aberrations

To determine the role of copy-number aberrations, we reconstructed the subclone-specific total copy number (see Section 4.4.2) and calculated the mean copy-number difference per chromosome, $\langle c_i - c_0 \rangle_{i \in C}$, rounded to the nearest integer. We observed recurrent copy-number aberrations across replicate populations. Figure 4.9 shows the number of copy-number gains and losses observed in each chromosome. There were chromosomal gains in 401 out of 632 haploids (64%) and 381 out of 547 diploids (67%). Bias correction successfully accounted for the technical variation in read depth of more than 90% of populations, most of which divide asynchronously. However, the biological variation in read depth in synchronous populations have characteristic biases near origins of replication, which together with the technical noise compromised the subclonal copy-number reconstruction. As a result, some copy-number losses may be false positives, particularly in HU-C and SC.

Gains of chromosomes VI and X were particularly prevalent in HU-C and HU-D (`n>2n`, `n>3n`, `2n>3n` and `2n>4n`). In addition, there were frequent gains in chromosomes VIII and XII in RM-C and RM-D (`n>2n` and `2n>3n`). These aneuploidies can provide a simultaneous increase in the number of copies of putative driver genes (and thus their expression dosage). Chromosomal losses were rarer, never found in haploids and only observed in 202 out of 547 diploids (37%).

Fig. 4.9 Sequence-level evolution in constant and time-dependent fitness landscapes, showing subclone-specific chromosomal aberrations. Middle: The data matrix shows subclone-specific copy-number changes per chromosome C found in a population, $\langle c_i - c_0 \rangle_{i \in C}$. The colour scale indicates the number of copy gains (light to dark blue: +1, +2, +3) or losses (light to dark red: -1, -2, -3). Rows are ordered along the left axis by the total number of copy gains and losses per chromosome, as represented by the vertical scale which is delimited by environment (dashed lines; HU-C: ●, HU-D: ●, RM-C: ●, RM-D: ● and SC: ●). Columns are ordered along the bottom axis, delimited by the type of population (solid lines; WA: ●, NA: ●, WAxNA: ●) and the founder genotype (dashed lines). The wild-type copy number of the founder is shown on the bottom labels ($c_0 = 1$, solid; $c_0 = 2$, hatched). Boxed elements indicate a fixed copy-number aberration in a given population (total subclone fraction $F^s > 0.99$). Top: The bars above the data matrix indicate a classification of copy-number changes according to whether they are gains or losses. The stacked bars show the total number of chromosome gains and losses observed in each replicate population. Right: The bars to the right of the data matrix show the total number of chromosomal gains (blue) or losses (red) per chromosome, counted across all replicate populations in each environment.

The set of recurrent copy-number aberrations also differs between pairs of founders, just like we saw it was the case with mutations (Fig. 4.9, top panel). Divergent genotypes typically follow starkly different paths. Copy-number gains of chromosomes VIII and XII appear to be mutually exclusive and specific to WA and NA backgrounds, respectively. The genomic structure may therefore constrain the mutational patterns that a population can acquire, thus limiting or opening new evolutionary paths that it can follow.

4.5.3 Fixation of mutations and genetic hitchhiking

To analyse the role of natural selection in shaping the global mutation frequencies, we can examine the consequences of a selective process acting differentially on genes across the genome using an evolutionary model. The frequency of mutations in a population is still a 'static' measure, so we will use this to show how we can extract information on the fate of mutations from ensembles of subclones across populations.

To quantify the extent to which the evolutionary dynamics is dominated by selective sweeps, we can divide mutations between those acquired before the last selective sweep, which are shared by all cells within the population (fixed), and new variants that occurred after the emergence of the common ancestor (polymorphic). Fixed mutations are expected to occur in ~100% of sequences in haploid individuals, and in ~50% of diploid individuals. Overall, we identified 2,325 mutations that fixed in any one population (13% of all detected mutations), with an average of 3.4 ± 2.9 fixed mutations and 8.1 ± 6.9 polymorphic mutations per population. Fixed mutations were likely acquired either before or during the most recent complete selective sweep. This suggests that there is massive hitchhiking of multiple mutations in one or several lineages (Fig. 4.10A). A population acquires a driver, then acquires passengers in the meantime before the next driver arrives. Given the extent of clustering of mutation fixations, however, beneficial mutations may often co-hitchhike with other drivers, particularly in HU-C (3.6 ± 2.6 fixations) and HU-D (6.0 ± 3.7 fixations). Conversely, only 1 to 2 mutations are simultaneously fixed in most populations in RM-C (2.0 ± 1.2), RM-D (2.0 ± 1.5) and SC (1.7 ± 1.4). These rare drivers reside in genomes alongside 4.5 passenger mutations on average. The number of fixation events per population is non-Poisson in HU-C and HU-D, which suggests that clonal interference is playing a major role and many beneficial mutations will be wasted to competition with other lineages (see e.g., Strelkowa and Lässig [210] for a similar observation).

Non-synonymous mutations in multi-hit genes are more likely to fix in the population (118/401, 29%) compared to other non-synonymous mutations (1,346/7,003, 19%). Syn-

onymous variants also fix more often than expected under neutrality, suggesting that many neutral mutations hitchhike with driver mutations that eventually fix. Paradoxically, populations which appear to be adapting faster in HU-C and HU-D based on the number of non-synonymous nucleotide fixations may also be accumulating significant deleterious loads of hitchhiking passengers. To quantify the overrepresentation of non-synonymous mutations among diverse positions where mutations are under strong positive selection, we binned the fraction of mutated reads in bins of size $[x - \delta x, x + \delta x]$, where $\delta x = 0.1$. The contrasting behaviour of non-synonymous mutations compared to synonymous and non-coding mutations can be observed from the histogram of mutation abundance – also known as the frequency spectrum – shown in Figure 4.11 for all replicate populations of a given ploidy in each environment. Most mutant clones detected in our study are relatively small, with a median frequency of 29%. However, the range of clone sizes is wide, spanning from one hundred thousand ($x \simeq 0.2$) up to several million cells ($x \simeq 1.0$). Figure 4.11A shows that the clone size distribution is heavy tailed, although our data provides a distorted picture owing to low sensitivity to small clones. While the spectra agree for frequencies below 20%, non-synonymous mutations are strongly overrepresented at higher frequencies, especially in RM-C and RM-D (Fisher's exact test at $x = 0.5$, $P < 10^{-4}$). This corroborates the interpretation that sweeping non-synonymous mutations more often 'drag' linked synonymous mutations to fixation in hydroxyurea than in rapamycin.

Idiosyncratic, population-specific mutations can have stronger effects in hydroxyurea than recurrently mutated genes, of which there are less than 10 that reach fixation (Fig. 4.11B).

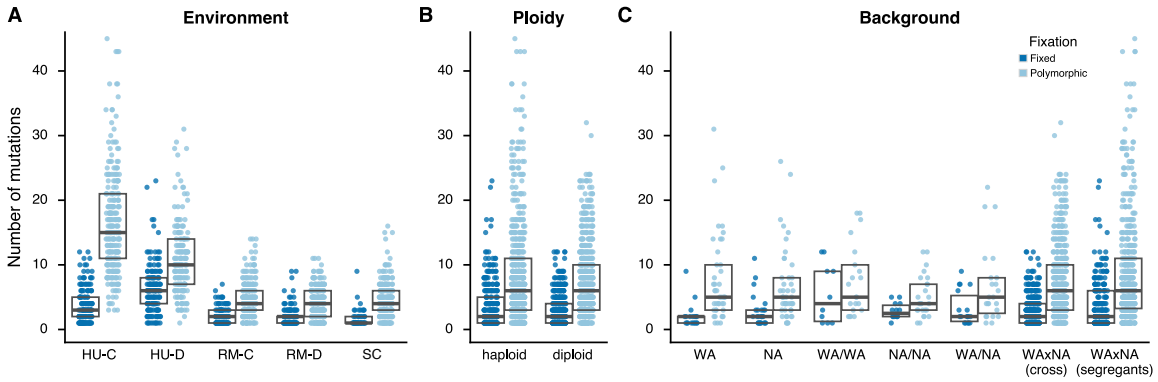


Fig. 4.10 Number of fixed and polymorphic mutations per population, broken down by **(A)** environment, **(B)** founder ploidy and **(C)** founder genotype (see Figure 4.1). Mutations found at frequency $x > 0.99$ in haploids and $x > 0.49$ in diploids are considered to be fixed (●), otherwise they are considered polymorphic (●). The counts are displayed with jitter along the horizontal axis for clarity. Mean and 25%/75% percentiles are shown, with means as thick horizontal black lines and the inter-quartile range delimited by thin lines.

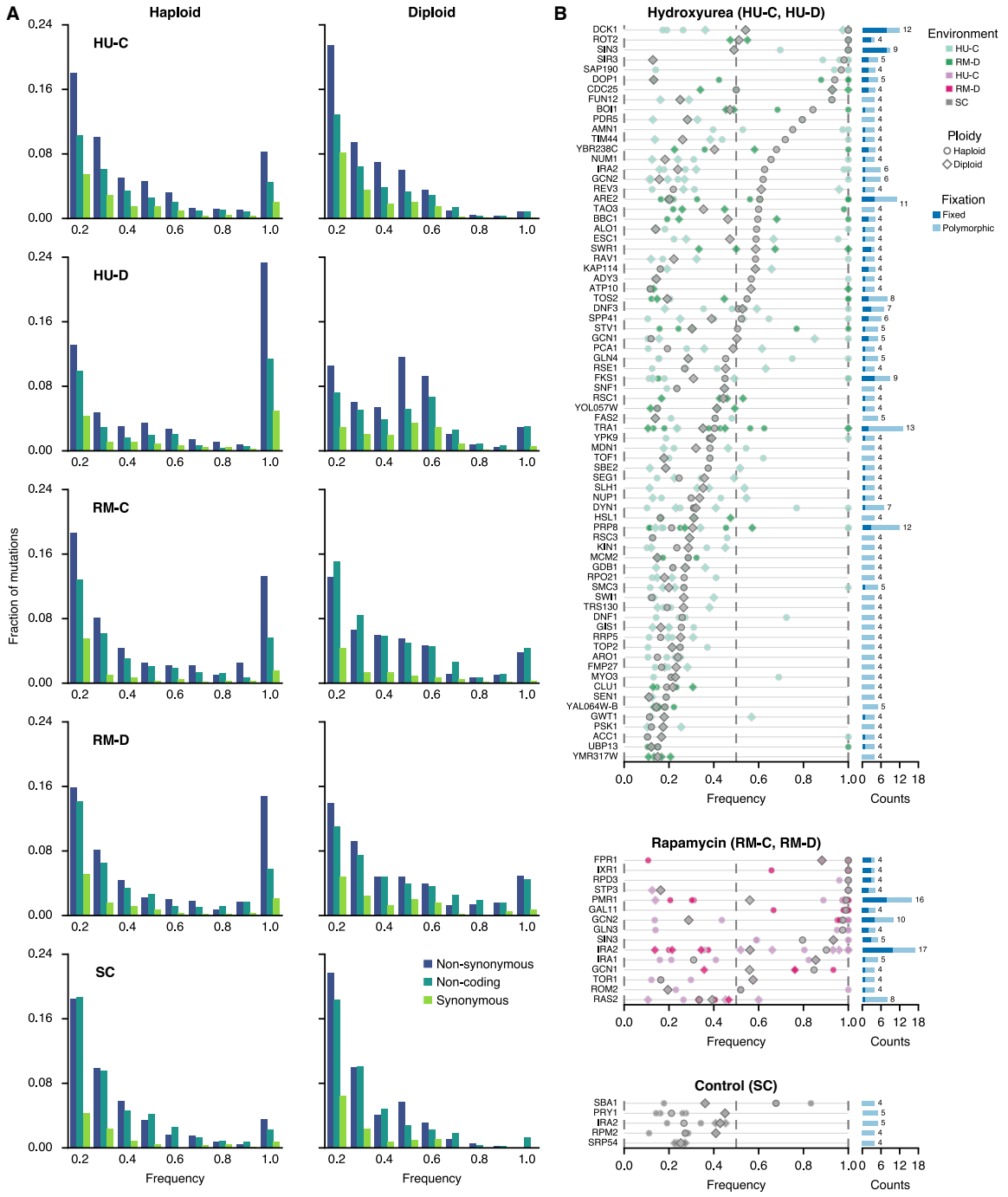


Fig. 4.11 Frequency spectrum of classes of mutations according to their variant allele frequency, broken down by environment and the ploidy of the founder. The analysis is limited to mutant clones above frequency $x > 0.1$. **(A)** Mutations are binned by their variant allele frequency x in bins of size $\delta x = 0.1$ (x-axis), shown against the normalised fraction of mutations per bin (y-axis). The binned counts are coloured by the mutation effect (non-synonymous: ●, non-coding: ●, synonymous: ●). **(B)** Frequency of recurrent coding mutations in genes mutated in ≥4 populations (described in Figure 4.7). The frequency x is shown for replicate populations that acquired a mutant variant of a gene (x-axis); mutated genes are grouped together by environment and sorted by the median mutation frequency (y-axis), which is indicated by grey markers. Mutations in haploid founders are expected to fix at frequency $x = 1.0$ (circles) and at $x = 0.5$ in diploid founders (diamonds). The mutation counts are aggregated on the right margin, coloured by fixation status (fixed: ●, polymorphic: ●).

This suggests that there is no simple relationship between the frequency of mutations and their selective advantage in conditions of nucleotide deprivation. In contrast, many fewer genes are recurrently mutated in rapamycin, yet the probability of fixation is much higher. Members of the TOR pathway, viz., *FPR1*, *TOR1* or *PMR1*, are part of a reduced group of 8 genes that reach fixation more than 50% of times (Fig. 4.11B).

4.6 Parallelism and co-occurrence of mutations

Thus far, our scoring for putatively selected mutations assumed the simplest additive model that a single mutation confers a fitness advantage by itself across founder genotypes. Genes do not act on their own, so the alternative hypotheses are that multiple mutations interact epistatically. Correlated behaviour in mutations at different positions in the genome carries information about functional and structural constraints acting at these positions.

4.6.1 Divergence of populations

Understanding the correlation structure between populations is crucial to capture the properties of the fitness landscape. To begin with a naïve examination of the genotype space, we compute a matrix (Π) of similarity between pairs of populations, such that Π_{p_1,p_2} gives the fraction of amino acids that are common between the populations p_1 and p_2. This information can be extracted by calculating the number of similar mutations found between each pair (p_1, p_2) of populations. The number of shared mutations in a subset of populations is counted based on the gene-level sequences. Given a set P of populations, we define a correlation matrix between pairs of populations (p_1, p_2) as

$$\Pi^{(P)}_{p_1 p_2} = \left\langle X_{p_1,g} X_{p_2,g} \right\rangle_{g\in G} - \left\langle X_{p_1,g} \right\rangle_{g\in G} \left\langle X_{p_2,g} \right\rangle_{g\in G} . \tag{4.5}$$

The histogram shows a monotonically decreasing distribution with groups of populations showing mean pairwise identities at recurrently hit loci ranging from 0.1 to greater than 0.5 (Fig. 4.14A). This pattern suggests that the majority of founders are distant in genotype space, but certain genotypes in rapamycin form multifarious clusters. We can further examine this assertion by direct visualisation of the sequence correlation matrix averaged by founder, $\Pi^{(F)}_{f_1 f_2}$ (Fig. 4.14B). Several unique mutations are shared between different populations, suggesting that they may occur by convergent evolution. For instance, WAxNA-16 and WAxNA-112 acquired similar SNVs and indels in rapamycin (Fig. 4.12B). These mutations do not stem from a common source, as we are only considering newly acquired mutations.

However, it is also possible that these mutations may be statistically more frequent and likely to arise in two populations independently. On the contrary, divergent genotypes WA and NA were less likely to converge upon the same solution, as evidenced by the sparse off-diagonal elements in Figure 4.12B.

Given two populations p_1 and p_2, each with a genome with L sites, we defined $\Pi^{(P)}_{p_1 p_2}$ as the number of pairwise similarities between them. For N_p populations, the average number of pairwise mutations is

$$\begin{aligned}\Pi^{(P)} &= \binom{N}{2}^{-1} \sum_{\{p_1,p_2\}} \Pi^{(P)}_{p_1 p_2} \\ &= \frac{2}{N(N-1)} \sum_{p_1} \sum_{p_2 > p_1} \Pi^{(P)}_{p_1 p_2}\end{aligned}$$

where the sum is over all pairs $(p_1, p_2) \in \{1, 2, \ldots, N_p\}$ evaluated over the upper triangular matrix of $\Pi^{(P)}_{p_1 p_2}$, the diagonal should be ignored ($p_1 \neq p_2$) and the sum normalised by the total number of pairwise combinations.

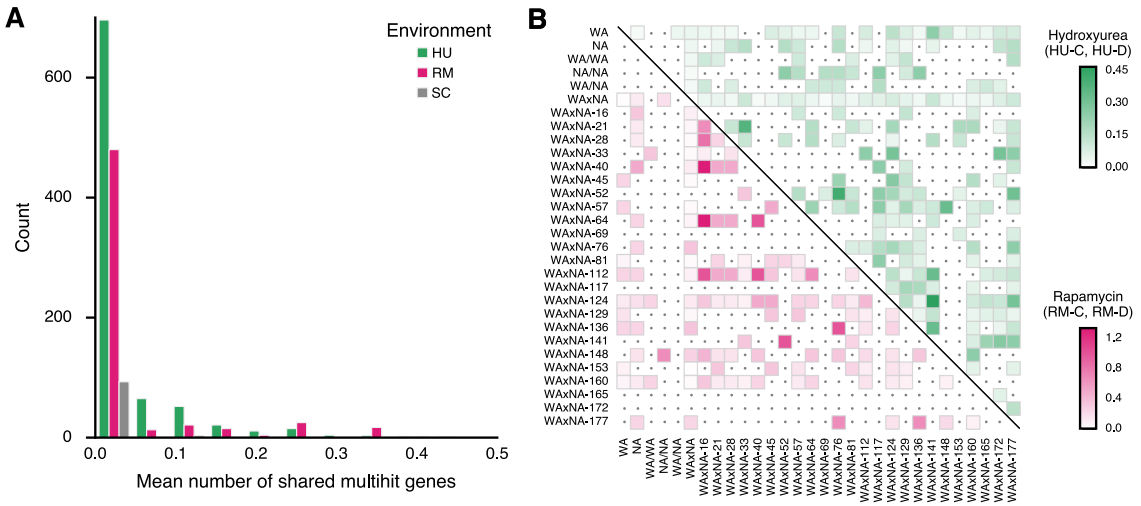


Fig. 4.12 Pairwise similarity $\Pi^{(F)}_{f_1 f_2}$ between any two founder genotypes at recurrently mutated loci found in ≥4 populations. **(A)** Marginal count of the average number of multi-hit genes mutated in populations derived from any two founders, grouped by environment. **(B)** The heatmap shows the average number of mutations shared by two populations descended from the founders indicated in the row and column headers. The colour gradient represents the mean number of shared mutations in multi-hit genes. The upper triangular matrix shows convergence for populations in hydroxyurea (HU-C and HU-D: ●) and the lower triangular matrix in rapamycin (RM-C and RM-D: ●). Convergence in the control environment is not shown, as only 17 pairs of founders shared patterns of mutated genes.

We would like to know whether two populations derived from the same founder are more likely to follow the same evolutionary path than two populations coming from different founders. From our definition of $\Pi^{(P)}_{p_1 p_2}$, the average number of pairwise similarities in mutated genes between any two populations descended from the same founder is

$$\begin{aligned}\Pi^{(F)} &= \sum_{f_1} \binom{N_{f_1}}{2}^{-1} \sum_{\{f_1,f_2\}} \Pi^{(F)}_{f_1 f_2} \\ &= \sum_{f_1} \frac{2}{N_{f_1}(N_{f_1}-1)} \sum_{f_1>f_2} \sum_{\{p_1,p_2\}\in F} \Pi^{(P)}_{p_1 p_2}\end{aligned}$$

where we sum over the block-diagonal elements of the upper triangular matrix that correspond to populations (p_1, p_2) derived from the same founder F. The normalisation factor is limited to the total number of pairwise combinations for each founder F.

We calculated the null distribution of $\Pi^{(P)}$ by randomly distributing the total number of putatively functional mutations found in each population across all yeast open-reading frames using the multinomial distribution. Since we want to compare the degree of convergence between and within founders, the null distribution of $\Pi^{(F)}$ can be obtained by random permutation of the founder labels.

There is a significant degree of convergence between replicate populations at the gene level (Fig. 4.13A). Across founders, populations have more mutations in common than is expected by chance, especially in hydroxyurea ($\Pi^{(P)} = 3.01 \times 10^{-1}$) and rapamycin ($\Pi^{(P)} = 8.28 \times 10^{-2}$), and less so in the control environment ($\Pi^{(P)} = 3.19 \times 10^{-2}$). In contrast, the null distribution for founder-specific convergence depends on the degree of overall convergence and is dominated by the large number of replicates of the recombinant cross. These founders form a 'genotype cloud' in sequence space and we only detect a small set of clones that eventually sweep in the population, without an evident influence of founder identity. It will be necessary to assess founder-specific convergence excluding these populations. As it stands, founder-specific convergence is low and most replicate populations derived from the same founder appear to take largely independent paths with only few exceptions that we noted earlier (Fig. 4.13B).

4.6.2 Correlations between mutations

Finally, we compare mutation patterns across different genes. We build a correlation matrix such that $\Pi^{(G)}_{g_1 g_2}$ contains the number of times gene g_1 has been found mutated in the same population as gene g_2. Given a set of genes G, we define a correlation matrix between pairs

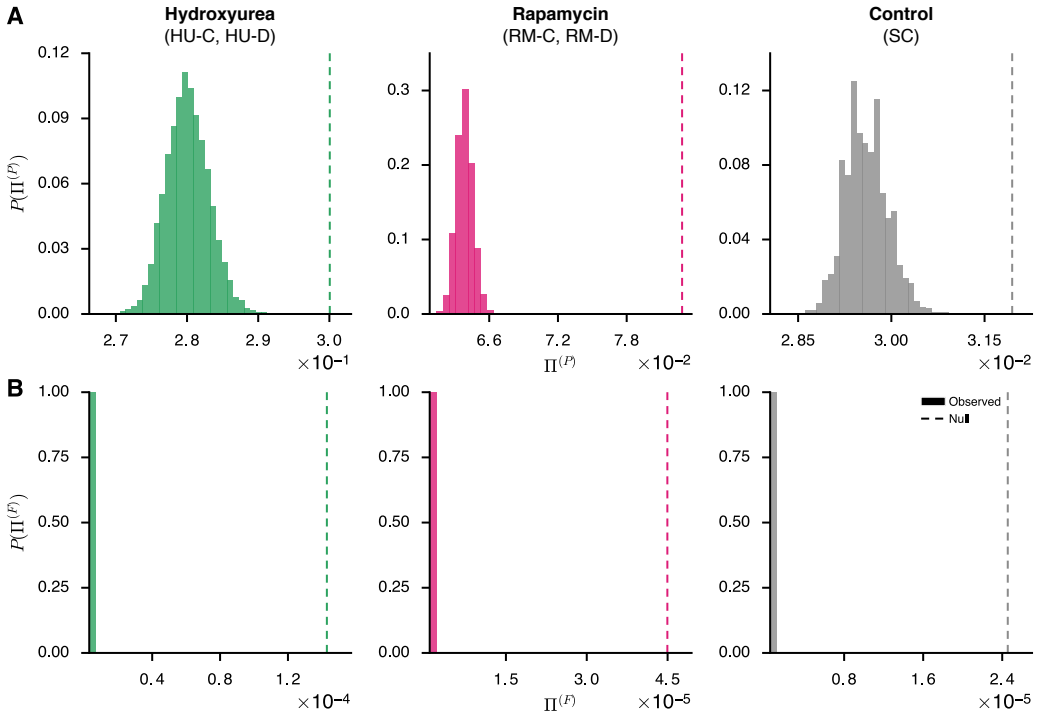


Fig. 4.13 Degree of parallelism and convergence at the gene level as measured by **(A)** the average number of similar genes mutated between populations, $\Pi^{(P)}$, and **(B)** the average number of similar genes mutated between populations derived from the same founder, $\Pi^{(F)}$. Vertical lines indicate the observed values of $\Pi^{(P)}$ and $\Pi^{(F)}$ in our dataset across environments (HU: ●, RM: ●, SC: ●). The histograms show the null distributions for these metrics, estimated from 10^4 randomisations. Note that the $\Pi^{(F)}$ null distributions are conditional on the observed degree of overall convergence.

of genes (g_1, g_2) as

$$\Gamma^{(G)}_{g_1 g_2} = \left\langle X_{p,g_1} X_{p,g_2} \right\rangle_{p \in P} - \left\langle X_{p,g_1} \right\rangle_{p \in P} \left\langle X_{p,g_2} \right\rangle_{p \in P}, \tag{4.6}$$

where the first term represents the joint frequency of having mutant alleles in genes g_1 and g_2. The second term is the product of the average mutation frequency of each gene independently. Averages are made over all genes g in the set G under consideration. We do not want to count the same gene pair twice, s.t. $\Gamma_{g_1 g_2} = \Gamma_{g_2 g_1}$ so we build a triangular matrix under the constraint that $g_1 > g_2$. Ideally, we would like to discover pathways solely by analysing gene groups and use the topology of gene interactions given enough samples, but we cannot attempt this without prior information. Considering that 3% of genes (117/3,594) are hit four times or more, only few genes will be informative to detect genetic interactions, so G is taken to be the set of environment-specific multi-hit genes in ≥4 populations rather than all mutated genes.

Similarly, to build a correlation matrix of copy-number aberrations, $\Pi^{(C)}_{c_1 c_2}$ contains the number of times chromosome c_1 has been found aberrated in the same population as chromosome c_2. Given a set of chromosomes C, we define a correlation matrix between pairs of chromosomes (c_1, c_2) as

$$\Gamma^{(C)}_{c_1 c_2} = \left\langle X_{p,c_1} X_{p,c_2} \right\rangle_{p \in P} - \left\langle X_{p,c_1} \right\rangle_{p \in P} \left\langle X_{p,c_2} \right\rangle_{p \in P}, \tag{4.7}$$

with similar properties as the gene correlation matrix. To detect co-occurring copy-number aberrations, we consider all pairwise interactions between chromosomes in set C.

Inspection of the correlation matrix clearly indicates that few positions show significant correlation to other loci elsewhere in the genome (Fig. 4.14B). Amongst them, we expect that pairs of genes that tend to be co-mutated most likely operate in different pathways. Conversely, pairs of genes exhibiting mutual exclusivity may be functionally redundant, potentially because they act in the same pathway. We therefore compared, for each pair of driver mutations, the extent to which their sets of target genes overlapped. The number of mutations acquired per population is scarce to determine the complete spectrum of adap-

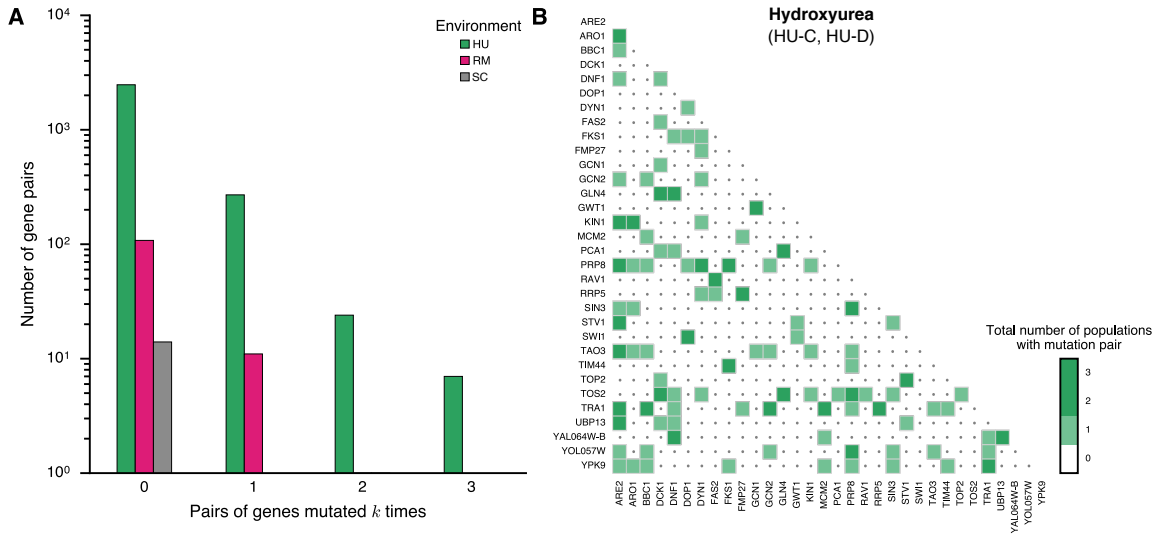


Fig. 4.14 Pairwise sequence similarity between **(A, B)** any two recurrently mutated genes. Correlations between any two recurrently mutated genes – mutated in ≥4 populations – are measured by the correlation matrix, $\Gamma^{(G)}_{g_1 g_2}$ (Equation (4.6)). **(A)** Total number of co-mutated pairs observed k times (x-axis) and the frequency of each in the dataset (y-axis), grouped by environment (HU: ●, RM: ●, SC: ●). **(B)** The heatmap shows the total number of pairwise mutations between two multi-hit genes in hydroxyurea, indicated by the discrete scale on the colour bar. The triangular matrix shows convergence for populations in hydroxyurea (HU-C and HU-D: ●). Convergence of multi-hit genes in rapamycin is not shown.

tive double mutants. However, a small subset of genes indicate preferential co-mutation, while no patterns of mutual exclusivity can be resolved (Fig. 4.14A). Six gene pairs are independently co-mutated three or more times in HU-C and HU-D: *ARE2-ARO1*, *ARE2-TAO3*, *BBC1-TRA1*, *DCK1-GLN4*, *MCM2-TRA1*, *PRP8-SIN3* and *PRP8-YOL057W*. For chromosome-level copy-number changes, we observed chromosomes VIII and XII to be frequently co-amplified in rapamycin (data not shown). These pervasive chromosomal gains may act as transient responses to stress [211]. It remains to be shown whether these correlation patterns are functionally significant, and whether we can distinguish them from correlations that could arise due to limited sampling of populations or due to kinship between founder genotypes.

4.7 Summary

In this chapter, we have presented a comprehensive portrait of evolutionary dynamics of the genome by directed evolution and selection, using budding yeast as a model organism. We focused on understanding the selective constraints under inhibition with antimicrobial drugs which impose trade-offs at several rate-limiting steps in the cell cycle. To explore a broad spectrum of evolutionary outcomes, selection starts with a pool of sequences derived from two diverged genotypes that are random at one in every 230 bp. This allows for a maximally unbiased sampling of genotype space. We can sample over 10^7 genome sequences, and after several cycles of selection, one can recover the descendant clones of one or several functional genomes from the founder population. The 'swarm' of genotypes in each of the founders is designed so the coverage of genotype space ranges from uniformly sampled ($N_c = 10^7$) to vanishingly small ($N_c = 1$).

We observed recurrent mutation patterns across populations by characterising genetic variants and tracking the fate and dynamics of mutations over a period of 93 days. To identify driver mutations under selection from hitchhiking passengers, we compared independent realisations that followed parallel adaptive paths to a given selection pressure. Parallel genotypic paths showed convergence at the level of genes, suggesting that gene-level observables act as functional units of the genome and providing a sensible sequence model. The spectrum of parallel adaptive mutations in hydroxyurea was highly variable between replicate populations, comparable to the substantial variability observed with other antimicrobial drugs (e.g., chloramphenicol and doxycycline) [192]. Conversely, rapamycin resistance reproducibly arose by repeatable mutational paths, similar to other drugs with a narrow spectrum of escape mutations (e.g., trimethoprim and ciprofloxacin) [192].

Replicate populations derived from the same founder systematically acquired parallel mutations in similar genes, suggesting that they explore a neighbouring genotype space common to all of them. The rate at which they explore the genotype space differs based on the ensemble diversity. On average, the length of mutational paths required by genetically diverse populations to adapt was shorter than in genetically homogeneous populations, though very unpredictable between replicate populations. Conversely, those same genotypes in isolation typically became entrenched in a local fitness maximum and required similar numbers of mutations to adapt, which demarcates the limits to the predictability of outcomes at the sequence level. This suggests that co-existing genotypes in a mixed population, rather than isolated genotypes, accelerate the search for adaptive mutations in a fitness landscape.

In summary, we present the first comprehensive characterisation of the role of clonal heterogeneity in the acquisition of antimicrobial resistance, and of the influence of drug-dosing schedules on the evolutionary dynamics of escape mutations. Understanding the ramifications of genetic heterogeneity is one of the most pressing issues in the treatment with antimicrobial and chemotherapy drugs. We have shown that controlled evolutionary experiments in model systems are an important way to build a coarse-grained quantitative description of complex systems that otherwise may not be tractable experimentally. Methods that can track the evolution of macroscopic subclones and estimate their fitness advantage are already being used to understand microbial and cancer evolution [212, 213]. For instance, profiling tumours for pre-existing drug-resistant subclones which have been correlated with worse outcomes is already improving estimates of survival times before anti-EGFR therapy [214, 215]. Future studies should strive to make accurate and falsifiable predictions of evolutionary dynamics over short timeframes in a range of rapidly adapting populations.

Chapter 5

Dynamics of selective sweeps and the fate of new mutations

5.1 Introduction

The presence of genetic variation conditions the fate of new mutations. In this chapter we analyse how the interaction between existing and acquired mutations gives rise to complex evolutionary dynamics from experimental tests on directed evolution in budding yeast (*S. cerevisiae*). We first quantify the time evolution of pre-existing and new mutations and measure the changes to the fitness distribution. We use the results to parameterise the characteristic timescale of selection and to distinguish driver and passenger mutations based on the theory presented in Chapter 2. We discuss the emergence of macroscopic subclones driven by new mutations and the role that genomic instability plays in accelerating sequence evolution and broadening the fitness distribution. We then compare the candidate driver mutations predicted by the driver-passenger model with the fitness effects of genetic constructs engineered by our collaborators. We end the chapter on the development of an experimental test to decompose the fitness contributions of pre-existing and *de novo* mutations using a recombinant sequence ensemble. Based on the decomposition of fitness components we demonstrate that the underlying fitness variance can set a selective threshold on new mutations, confirming recent results from population genetic theory.[1]

This work has been done in collaboration with V. Mustonen (V.M.) and A. Fischer (A.F.) at the Wellcome Trust Sanger Institute (Cambridge, UK), F. Salinas (F.S.), J. Li (J.L.) and G. Liti (G.L.) from the Institute for Research in Cancer and Aging of Nice (France), and

[1] Data analyses related to this chapter are available from the GitHub code repository [https://github.com/ivazquez/PhD-thesis/tree/master/Chapter5].

E. Alonso-Perez (E.A.-P.) and J. Warringer (J.W.) from the University of Gothenburg (Sweden).[1]

5.2 Maintenance and loss of genetic variation

Consider the prototypical scenario that arises when individuals in a population acquire heritable genetic or non-genetic changes to adapt and thrive in a new environment. Since the seminal findings by Salvador Luria and Max Delbrück that phage-resistant bacteria can acquire adaptive mutations prior to selection [216], measuring the fitness effects and dynamics of mutations has been key to map the principles of evolutionary adaptation. The focus has typically been on characterising few mutations at a time under the implicit assumption that beneficial mutations are rare, treating pre-existing and acquired mutations separately. However, many mutations are often simultaneously present in a population, which result in fitness differences between individuals that selection can act upon.

Given that mutations in asexual populations are physically linked in the genome, the fates of pre-existing and *de novo* mutations are mutually dependent and selection can only act on these sets of variants in their entirety. Genome evolution experiments on isogenic populations have revealed both adaptive sweeps and pervasive clonal competition in large populations where the mutation supply is high. This phenomenon, known as clonal interference, takes place as mutations in different individuals cannot recombine via sexual reproduction and is now relatively well understood both experimentally and theoretically [188, 217, 218]. Experiments on populations with extensive genetic variation have demonstrated that beneficial mutations expand in a repeatable way [75]. However, the role of *de novo* mutations has been negligible in these experiments, either because they were too short or related to the selective constraints used. As we showed in Chapter 4, it is becoming clear from experimental studies that extant genetic variants are sufficient to steer the fate of populations into different evolutionary paths. A study which was able to anticipate new mutations found that one or few genetic variants were sufficient to affect the fate of subsequent beneficial mutations, hinting that the joint dynamics of new mutations have to be considered in the light of pre-existing variation [105].

[1] I.V.-G., V.M. and G.L. designed research; F.S. and J.L. maintained the cell culture and extracted DNA for sequencing, F.S. engineered the genetic constructs for validation, J.H. and I.V.-G. constructed the recombinant library, J.L. carried out the Luria-Delbrück fluctuation assay, E.A.-P. and I.V.-G. recorded phenotypic measurements, I.V.-G., V.M. and A.F. developed theory, implemented computational methods and analysed data.

The ensuing interaction between existing and subsequent mutations has been theoretically considered under different population genetic scenarios which we discussed in Chapter 2. A key theoretical prediction is that a new beneficial mutation will only establish if it has a selective advantage greater than a characteristic value that depends on the underlying fitness distribution [219, 220]. However, this is an important hypothesis that remains to be tested: namely, whether genetic diversity can change the evolutionary fate of new adaptive mutations by limiting the number of backgrounds where they can still outcompete the fittest extant individuals. Understanding the impact of genetic heterogeneity on adaptive dynamics is particularly urgent as recent findings indicate that it can play a major role in the development of resistant bacterial infections [177] and in cancer recurrence [38, 137].

We can delineate two lines of enquiry into this hypothesis: (i) To what extent can the adaptive response be attributed to genetic variation already present in a population and how much to acquired? (ii) How do the aggregate effects of pre-existing variation influence the fate of new mutations?

5.3 Experimental design

To address these questions, we investigated the interaction between pre-existing (or background) genetic variation and new mutations in a population of diploid cells with unique combinations of alleles under selection for antimicrobial resistance. The cells originate from two diverged *S. cerevisiae* strains (WA and NA) and their recombinant cross (WAxNA), which were already introduced in Section 4.3. Starting from WA, NA and WAxNA founders, we asexually evolved populations of $\sim 10^7$ cells in serial batch culture with inhibitors of nucleotide synthesis (hydroxyurea) and of cellular growth (rapamycin) at concentrations impeding, but not ending, cell proliferation (Fig. 5.1). We derived replicate lines of WA, NA (2 each in hydroxyurea and rapamycin) and WAxNA (6 in hydroxyurea, 8 in rapamycin and 4 in a control environment), propagating them for 32 days in 48-hour cycles ($\sim$200 generations). We monitored evolutionary changes by whole-genome sequencing of populations after 2, 4, 8, 16 and 32 days, as well as clonal isolates at 0 and 32 days. Finally, we measured the rate of growth at the initial and final time point for a subset of populations, and quantified the relative fitness contributions of background and *de novo* variation using a genetic cross.

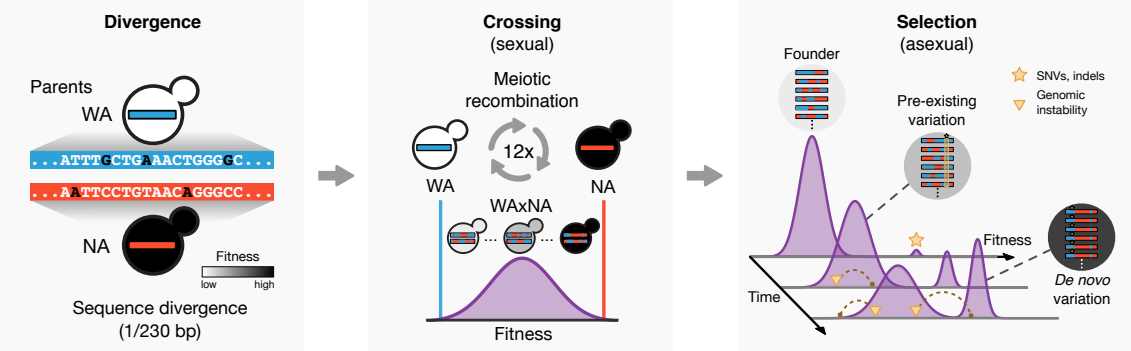


Fig. 5.1 Schematic diagram of the divergence, crossing and selection phases of the experiment. Two diverged budding yeast (*S. cerevisiae*) lines (WA and NA) were crossed for twelve rounds, generating a large ancestral population of unique haplotypes (see Figure 4.1 for a full description). These diploid cells were asexually evolved for 32 days in stress and control environments and their adaptation was studied by whole-population sequencing, isolate sequencing and phenotyping. Populations evolved resistant macroscopic subclones driven by individual cells with beneficial genetic backgrounds (i.e., parental allele configurations) and by beneficial *de novo* mutations that provided a resistance phenotype.

Genome sequencing

We followed the evolution of these populations over the course of the experiment using whole-genome sequencing. Whole-population sequencing was performed after $t = (0, 2, 4, 8, 16, 32)$ days, and single-cell derived clones were also sequenced at $t = 0$ days and $t = 32$ days (Table 5.1). Genomic DNA was extracted from the samples using the 'Yeast MasterPure' kit (Epicentre, USA). The samples were sequenced with Illumina TruSeq SBS v4 chemistry, using paired-end sequencing on Illumina HiSeq 2000/2500 at the Wellcome Trust Sanger Institute.[1] WA and NA populations are labelled by their background, the environment in the selection phase and the selection replicate, e.g., NA RM 1. WAxNA populations are labelled by background, number of crossing rounds, cross replicate, selection environment and selection replicate, e.g., WAxNA F12 2 HU 1. Time series samples are labelled from T0 to T32 and isolate clones carry a suffix, e.g., C1, C2, etc. Read alignment and detection of pre-existing and *de novo* variation were carried out as described in Section 4.4.

[1] Sequence data are available from the European Nucleotide Archive, under study accession no. PRJEB2608 for the parental strains and the ancestral individuals, and study accession no. PRJEB4645 for the time-resolved populations. Mutation calls for the two aforementioned datasets are available under study accession no. PRJEB13491. Instructions to download the data and sample labels are available on the GitHub code repository [https://github.com/ivazquez/PhD-thesis/tree/master/Chapter5].

Table 5.1 Summary of populations and clonal isolates analysed by whole-genome sequencing. The best-fit number of subclones N_c are shown together with the total clonal fraction, $F^t = \sum_{j=1}^{n} f_j^t$, after 32 days of selection. Per population, the union set of driver mutations found by whole-population and clone genome sequencing is shown. The genotypes of driver mutations found in clonal isolates were validated by Sanger sequencing (labelled by §). WA/WA populations in hydroxyurea did not survive beyond 4 days of selection (labelled by †).

Time	**Background**	**Cross**		**Selection**			**Clonality**		**Drivers**
		Gen.	Rep.	Environment	Rep.	Isolates	N_c	F^t	
0 days	WA/WA	–	–	YPD	–	–	–	–	
	NA/NA	–	–	YPD	–	–	–	–	
	WAxNA	F_{12}	1	YPD	–	C1–C96	–	–	
			2	YPD	–	C1–C96	–	–	
2–32 days	WA/WA	–	–	YPD+HU	1†	–	–	–	
					2†	–	–	–	
				YPD+RM	1	–	–	–	*TOR1* W2038L§
					2	–	–	–	*TOR1* F2045L§
	NA/NA	–	–	YPD+HU	1	–	–	–	*RNR4* R34I§, K114M§
					2	–	–	–	*RNR4* R34I§, K114M§
				YPD+RM	1	–	–	–	*FPR1* K11fs§; *TOR1* S1972R, W2038L§
					2	–	–	–	*FPR1* M1I§; *TOR1* S1972I§
	WAxNA	F_2	1	YPD+RM	1	–	2	0.74	*TOR1* W2038L
					2	–	1	0.10	
				YPD	1	–	0	–	
		F_{12}	1	YPD+HU	1	C1–C2	2	0.58	*RNR4* R34G§, R34I§
					2	C1–C2	1	0.20	*RNR4* R34I§
					3	C1–C6	2	0.65	*RNR2* Y169H§; chr. II LOH
				YPD+RM	1	C1–C3	3	0.85	*CTF8*^{NA}; *FPR1* W66*§, W66S
					2	C1–C6	2	0.20	*CTF8*^{NA}; *FPR1* W66S; *TOR1* W2038L§
					3	C1–C3	2	0.72	*CTF8*^{NA}; *FPR1* W66*§; *TOR1* S1972I
					4	–	2	0.81	*CTF8*^{NA}; *FPR1* W66*§
				YPD	1	–	0	–	
					2	–	0	–	
			2	YPD+HU	1	C1–C2	2	0.63	*RNR4* R34G§, R34I§
					2	C1–C4	2	0.32	*RNR2* N151H, T206I§; *RNR4* R34I§
					3	C1–C6	2	0.34	*RNR2* E154G§; *RNR4* R34I§
				YPD+RM	1	C1–C3	4	0.93	*CTF8*^{NA}; *FPR1* W66S, W66*§
					2	C1–C6	1	0.10	*CTF8*^{NA}; *TOR1* W2038C§
					3	C1	1	0.10	*CTF8*^{NA}; *FPR1* S102R
					4	–	1	0.11	*CTF8*^{NA}; *FPR1* S102R
				YPD	1	–	0	–	
					2	–	0	–	

Fitness measurements

We carried out phenotyping for a series of experiments: (i) to follow the evolution of the fitness distribution, (ii) to reconstruct the fitness components using a genetic cross, and (iii) to validate the fitness effects of driver mutations with engineered genetic constructs.[1] We used a high-resolution scanning platform to carry out the measurements, monitoring growth in a 1,536-colony design on solid agar medium [221]. Solid media plates designed for use with

[1] Phenotype data are available from the GitHub code repository, including instructions to download the data and sample labels [https://github.com/ivazquez/PhD-thesis/tree/master/Chapter5].

the Singer RoToR HDA robot (Singer Ltd) were used throughout the experiment. Casting was performed on a levelled surface, drying for ∼1 day. We distributed samples over 1,152 positions across each plate in a randomised layout, keeping every fourth position for 384 controls used for removal of spatial bias. We performed transmissive scanning at 600 dpi using an 8-bit grey scale, capturing four plates per image. Plates were fixed by custom-made acrylic glass fixtures. Pixel intensities were normalised and standardised across instruments using transmissive scale calibration targets (Kodak Professional Q-60 Color Input Target, Kodak Company, USA). We maintained a high-humidity environment at 30 ˚C during measurements, keeping scanners covered in custom-made boxes to avoid light influx and minimise evaporation.

Experiments were run for 3 days and scans were continuously performed every 20 minutes. Each image stack is then processed in a two-pass analysis: (i) during image acquisition, positions in each image are matched to the fixed calibration model using the fixture orientation markers, allowing detection and annotation of plates and transmissive scale calibration strips; and (ii) after image acquisition, the entire image stack is segmented to identify the location of the plate, the transmissive scale calibration strip, and the colony positions based on pinning format. Differences in pixel intensity are converted to population size estimates $N(t)$ by calibration to independent cell number estimates (spectrometer and flow cytometry). Based on these, we obtained growth curves in physical units.

Raw measurements of population size were smoothed in a two-step procedure: (i) a median filter identified and removed local spikes in each curve; and (ii) a Gaussian filter reduced the influence of any local noise remaining. Since we expect a population to double in size during the average time taken to progress through the cell cycle, we use an exponential growth model defined as $N(t) = N(0)\, e^{\lambda t}$, where λ is the absolute growth rate. If the time that has passed is exactly the doubling time τ, it is trivial to show that within this time span the growth rate can be rewritten as $\lambda = \frac{\ln 2}{\tau}$. It follows that λ can then be estimated from the linear fit of any two log-transformed measurements of $N(t)$ in exponential phase, according to $\lambda = \frac{1}{t_f - t_i} \ln \frac{n(t_f)}{n(t_i)}$. To account for systematic errors caused by spatial variation within plates, the absolute growth rate was rescaled by taking the log-transformed difference between the observed estimate and the neighbouring control populations, i.e., the relative growth rate at position (i, j) is then $\lambda_{ij} \to \log \frac{\lambda_{ij}}{\lambda_{ij}^{\text{norm}}}$.

5.4 Timescales of selection

Two regimes of selection became readily apparent in both sequence and phenotype. Initially, there were local changes in the frequency of parental alleles under selection (Fig. 5.2). Over time, macroscopic subclonal populations arose and expanded, depleting the pool of genetic diversity. These successful genotypes persisted in time, manifested by broad jumps in the allele frequency visible across the genome (Fig. 5.2). But what drives these clonal expansions? Is it the founder haplotypes themselves, *de novo* mutations relegating the parental variation to the role of passengers, or their combined action?

5.4.1 Selective effects on pre-existing variation

To determine the adaptive value of background variation, we identified regions where local allele frequencies changed over the time course of the selection experiments. We observed patterns of selective sweeps when a 'driver' allele with a significant fitness advantage starts to gain in frequency due to the selective pressure (Fig. 5.2). Therefore, frequency changes over time indicate that positive (negative) selection is acting on beneficial (deleterious) background alleles. This movement also causes allele frequency changes at nearby loci containing 'passenger' alleles that are genetically linked with the driver, in a process called genetic hitchhiking (see Section 2.4).

To discern drivers and passengers, we consider a model of a population evolving in a regime of strong selection, where there is a favoured allele (driver) at locus i, and a set of linked passengers. We have presented the formulation of this model of multi-locus evolution in Section 2.4, which can be be used to analyse selection acting on pre-existing genetic variation generated by the recombinant cross in this experiment. Genetic drift plays a negligible role for allele frequency changes in the selection phase as the population size ($\sim 10^7$ cells) is much larger than its duration ($\sim$200 generations). Furthermore, the frequency spectrum of background mutations is normally distributed, so that pre-existing variants are already established and do not need to overcome genetic drift. Therefore, we can assume that the allele frequencies evolve deterministically and the remaining noise is due to sampling caused by finite sequencing depth. A selective sweep is then well approximated by a model of the frequency x_i^{WA} of the WA allele at locus i which satisfies the logistic equation (see Equation (2.10)). For a given driver locus and a set of passengers the model is fully specified by the strength of selection, the pairwise linkage structure, and allele frequencies at both driver and passenger loci at $t = 0$ days. We learn the free parameters of the model using maximum

likelihood with a binomial noise model accounting for sequencing noise. Furthermore, the scan can be performed systematically across loci by fixing pairwise linkage according to a recombination map of the cross (see Illingworth et al. [222]) and the initial allele frequencies at $t = 0$ days.

We performed a systematic driver scan including passengers within variable window sizes $\{\pm 2\,\text{kb}, 5\,\text{kb}, 10\,\text{kb}, 30\,\text{kb}, 50\,\text{kb}\}$. Emerging subclones result in global allele frequency changes that supersede the local signal, which is the hallmark of selection acting on pre-existing variation. In consequence, we only considered time points when populations had not yet become clonal, up to $t = 4$ days. For each scan we selected the top 200 loci (out of 52,466) and then required that a given window was identified to be among the top scoring

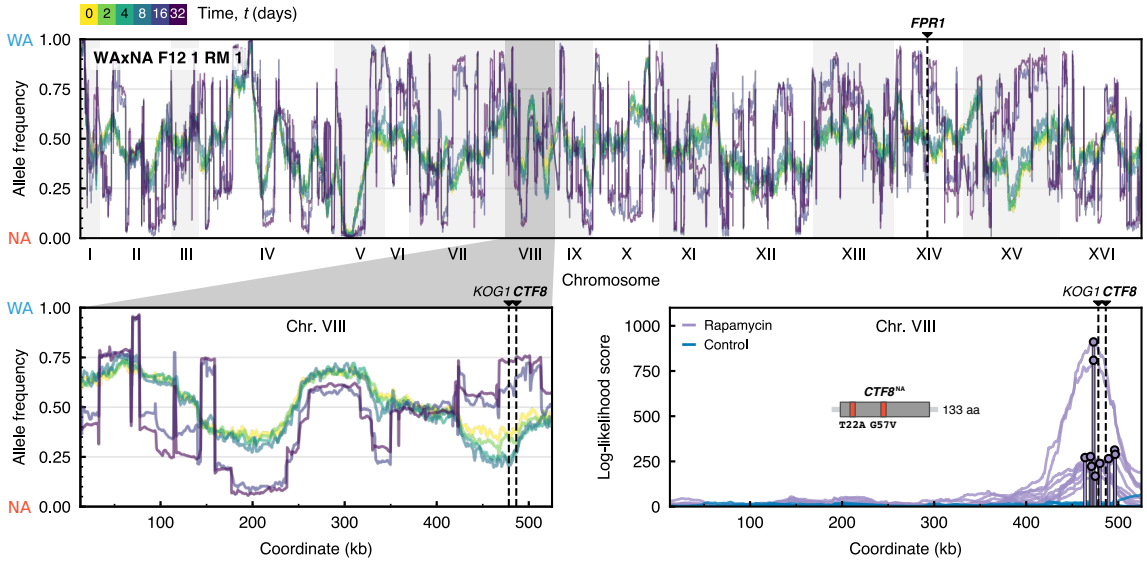


Fig. 5.2 Genome-wide allele frequency of pre-existing parental variants after $t = (0, 2, 4, 8, 16, 32)$ days, measured by whole-population sequencing for a representative population in rapamycin. Top panel: Chromosomes are shown on the x-axis; the frequency of the WA allele at locus i, x_i^{WA}, is shown on the y-axis. The reciprocal frequency of the NA allele is equivalent since $x_i^{\text{NA}} = 1 - x_i^{\text{WA}}$. Bottom left panel: Zoomed inset of the shaded region shows allele frequency changes in chromosome VIII during selection in rapamycin. Early time points 2, 4 and 8 show localised allele frequency changes at 460–490 kb due to a beneficial NA allele sweeping with hitchhiking passengers. Late time points 16 and 32 show abrupt jumps between successive loci that reflect the parental haplotype of emerging subclone(s). These long-range correlations can alter the frequency of parental alleles independently of their fitness value. In case of a fully clonal population, allele frequencies at 0, 0.5 and 1.0 would correspond to the background genotypes NA/NA, WA/NA, and WA/WA of a diploid clone that reached fixation. Bottom right panel: We tested a model where each allele is proposed to be a driver under selection, with linked passenger alleles also changing in frequency by genetic hitchhiking. Top log-likelihood scores are shown for all populations in this region of interest. We validated the $CTF8^{\text{NA}}$ allele to be strongly beneficial for rapamycin resistance (Fig. 5.10).

ones in at least two populations. The remaining windows were merged if their passenger loci overlapped. Finally, we required that the region was not identified among those scoring highly in the control environment.

The scan identified a region of interest for rapamycin resistance, found in chromosome VIII (460–490 kb). The signal is visible in all rapamycin populations but not in the control. However, we were not able to localise it fully due to a low recombination rate in this region and possibly also caused by the presence of multiple drivers. The region as a whole has strong support across populations to contain one or several beneficial NA allele(s) in rapamycin, albeit we cannot statistically map the signal more finely. Our collaborators followed up two candidate genes in the region (*CTF8* and *KOG1*), and we validated *CTF8* to have a resistance phenotype (see Section 5.5.2). Carrying the $CTF8^{NA}$ allele confers a 36% growth rate advantage over the $CTF8^{WA}$ allele. *KOG1*, which falls within the same region and is a subunit of the TORC1 complex, differs by seven missense mutations between the parents. However, reciprocal hemizygous deletions only revealed a modest fitness difference between WA and NA sequences of *KOG1*. We did not find events that replicated across all populations in hydroxyurea.

5.4.2 Diversity and clonal selection

To reconstruct clonal expansions in the WAxNA populations we used background genetic variants as markers. Using the probabilistic reconstruction algorithm we presented in Chapter 3, we inferred the subclonal genotypes and their frequency in the populations, both of which are unknown *a priori*. We identified jumps in correlated genotypes using the data filtering method on ancestral and evolved SNV data, using the posterior mean allele frequencies of the ancestral population to act as a bulk component for the inference (see Section 3.5). We then carried out subclonal reconstruction with cloneHD in SNV mode, as visual inspection did not reveal copy-number aberrations from whole-population sequencing. For each population, we systematically tried 0–4 subclones and determined the total data likelihoods under each model. The number of subclones per population are summarised in Table 5.1, together with the time evolution of subclone frequencies (see Figure 5.3 for a set of representative populations). We required a log-likelihood gain greater than 20,000 units for the inclusion of an additional subclone. This is necessary as the bulk component of the population can also change throughout the experiment. This conservative cut-off only allows genome-wide signals to be associated with a subclone. It prevents other solutions being

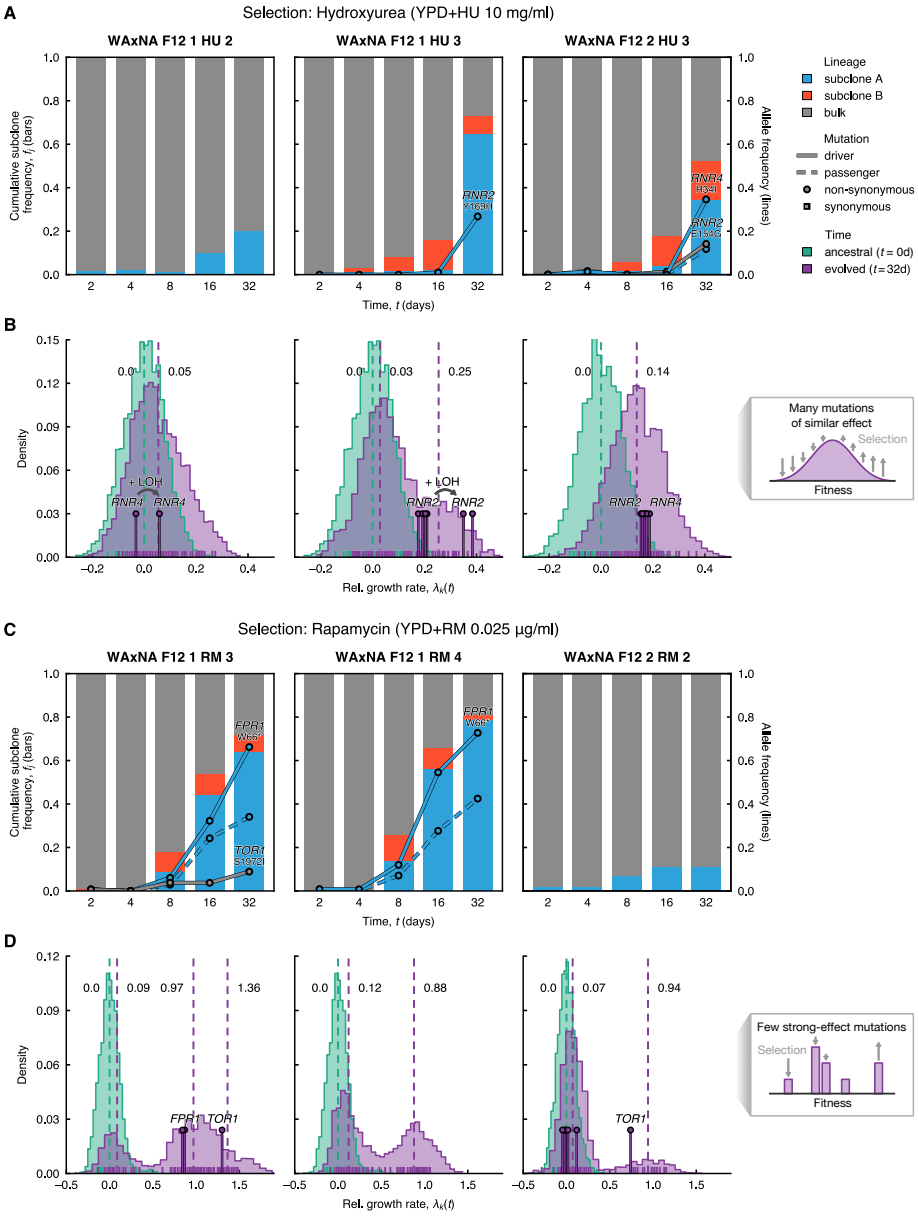


Fig. 5.3 Reconstruction of subclonal dynamics. Competing subclones evolved in hydroxyurea and rapamycin experienced a variety of fates. **(A, C)** Time is on the x-axis, starting after crossing when the population has no macroscopic subclones. Cumulative haplotype frequency of subclones (bars) and allele frequency of *de novo* mutants (lines) are on the y-axis. Most commonly, selective sweeps were observed where a spontaneous mutation arose and increased in frequency. Driver mutations are solid lines and passenger mutations are dashed lines, coloured by subclone assignment; circles and squares denote non-synonymous and synonymous mutations, respectively. **(B, D)** Variability in intra-population growth rate, λ, estimated by random sampling of 96 individuals at initial (t = 0 days, ●) and final (t = 32 days, ●) time points (n = 32 technical replicates per individual). Mean growth rates by individual are shown at the foot of the histogram (Fig. 5.4). The posterior means of the distribution modes fitted by a Gaussian mixture model are indicated as dashed lines. The fitter individuals (pins) carry driver mutations, detected by targeted sampling and sequencing. The insets on the right-hand side depict a schematic of the fitness distribution in two limit cases: if there are many mutations of similar effect, the fitness wave will be smooth and unimodal; if only few mutations of large effect exist, the fitness distribution will become multimodal.

favoured that would introduce artifactual subclones with suitable genotypes to improve fits in regions where selection acts on the bulk (see Section 3.7).

We found at least one subclone under selection in all WAxNA populations, but none in the control environment (Fig. 5.3). Clonal competition was prevalent with two or more expanding subclones in 12 out of 16 WAxNA populations. No population became fully clonal during the experiment, with subclone frequencies stabilising after 16 days in several rapamycin populations. To ascertain the expansion of subclones throughout the experiment, we determined the allele frequency of *de novo* mutations in WA, NA and WAxNA populations during the selection phase from whole-population sequencing. We found that these mutations typically did not reach detectable frequency (i.e., between 1–5%) until more than 8 days had passed, with steady increases thereafter (Fig. 5.3).

5.4.3 Fitness distribution and population averaging

To characterise the fitness of cells in a heterogeneous population with multiple subclones, i.e., where several haplotypes may be present, we must measure the growth properties of an ensemble of cells. With an ensemble method, we will typically measure the population average. However, since we found subclones co-existing, these may be found in states that are far from the population mean. Hence, we determined the intra-population growth rate of the populations at the start and the end of the selection phase (Figs. 5.3 and 5.4). We established this by phenotyping 96 randomly isolated individuals from 3 populations per environment at 0 and 32 days, as well as the 44 sequenced individuals at 32 days. For each population where we sampled n_k isogenic individuals, we estimated the probability distribution $P(\lambda(t))$ of the growth rate λ at time t. With an ensemble of $n_k = 96$ individuals per time point we took $n = 32$ replicate measurements per individual. The replicates were measured in two independent runs, evenly distributed over 16 experimental plates which were initiated from a single pre-culture plate and run in 4 scanners, all in parallel.

We modelled the time-dependent probability distribution of the data, $\{\lambda_n(t)\}_{n=1}^{n_k}$, as a mixture model of normal distributions for each time point t,

$$P(\lambda(t)) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\lambda(t)|\mu_k,\ \sigma_k^2\right) \tag{5.1}$$

where K is the number of components. We can interpret the mixing coefficients, π_k, as the bulk and multiple clonal components. We determined the fraction of cells in the fitter, faster clonal state(s) and the slower, bulk component by fitting a mixture of normal distributions

with $K \in \{1, 2, 3\}$ components. There are $2K + 1$ fitting parameters, which are learned by maximising the likelihood function $P(\lambda(t))$: the component means $\{\mu_k\}$ and variances $\{\sigma_k^2\}$, and the relative weights between them. In multimodal populations, the weights are in good agreement with the average of two consecutive inflection points surrounding the trough between the bulk and the clonal subpopulations (Fig. 5.4).

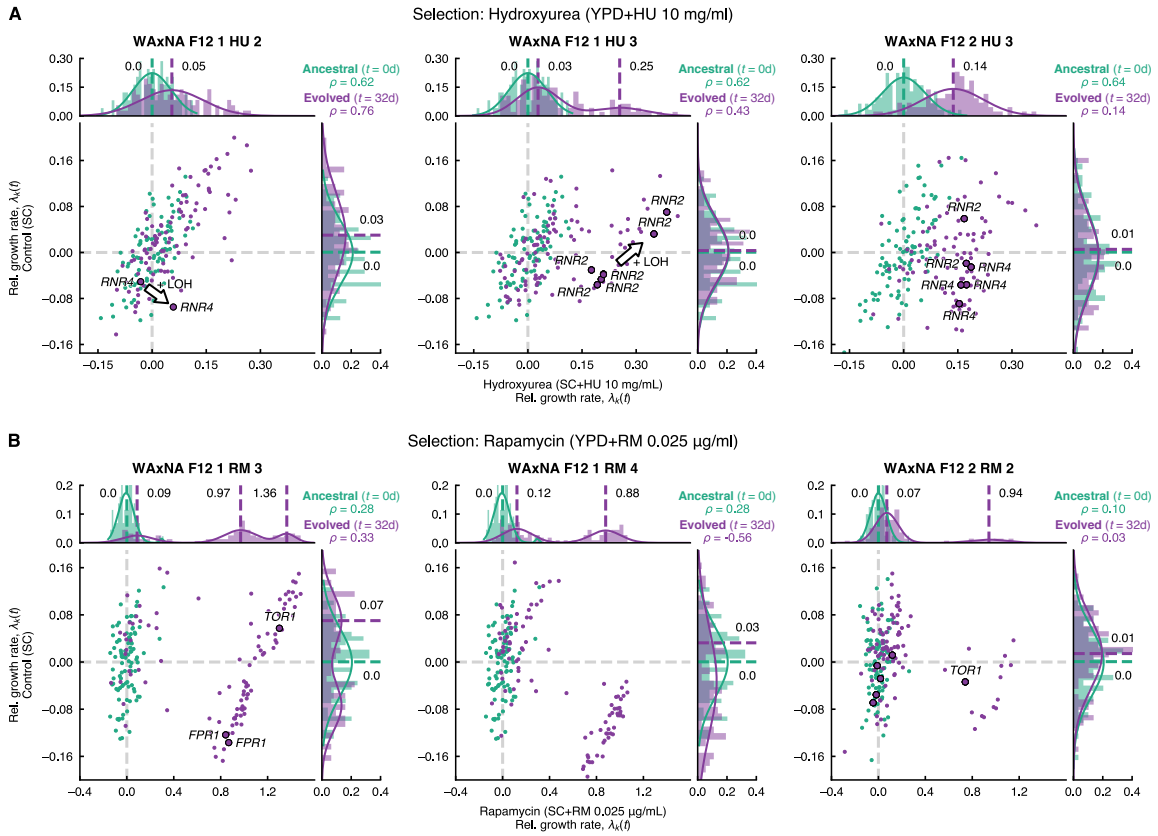


Fig. 5.4 Variability in intra-population growth rate and fitness correlations. Fitness correlations of ancestral and evolved populations across environments, estimated by random sampling of individuals at initial ($t = 0$ days, ●) and final time points ($t = 32$ days, ●), before and after selection in **(A)** hydroxyurea and **(B)** rapamycin. The mean growth rate λ per individual k is shown, averaging over $n = 32$ technical replicates per individual (see Figure 5.3B and D). Mean growth rate λ_k in the stress environment (x-axis) compared to the control environment (y-axis). Using a Gaussian mixture model, we found the posterior probability of the mixture modes of the the best-fit mixture (solid lines). The posterior means of the distribution modes are indicated as dashed lines. The fitter individuals carry driver mutations, as determined by targeted sampling and sequencing. Spearman's rank correlation, ρ, between the growth rate of isolates in the stress and control environments are shown on the top right of each panel. Positive rank correlations between stress and control environments are the most common. Negative rank correlations indicate an average fitness cost of driver mutations in the absence of the drug, e.g., in *FPR1*.

Clonal expansions were evident from changes to the fitness distribution of cells. In rapamycin selection, the phenotype distribution became multimodal after 32 days, reflecting the fitness of subclones substantially improving with respect to the mean fitness of the bulk population (Fig. 5.3D). The clonal subpopulations divided on average twice as fast as the ancestral population. Sequenced isolates with driver mutations in *FPR1* and *TOR1* were on the leading edge of the phenotype distribution, far ahead of the bulk. Furthermore, the bulk component showed a 10% average improvement, possibly due to selection of beneficial genetic backgrounds. Conversely, bimodality was only detected in one population in hydroxyurea selection (WAxNA F12 1 HU 3), where the clonal peak grew 25% faster on average compared to the ancestral, and the bulk 7% faster on average across all populations (Fig. 5.3B). Isolates with *RNR2* driver mutations fell onto the leading edge of the fitness distribution. These six isolates originated from the same expanding subclone and two of them had 13% faster growth rate than the remaining four, although they all shared the same heterozygous *RNR2* driver mutation. In both of these clones, we found a large region in chromosome II to have undergone loss-of-heterozygosity (LOH), offering a putative genetic cause for their growth advantage (Fig. 5.6A). Finally, to understand how the fitness of a typical population changes across environments we characterised the fitness correlations of ancestral and evolved clones, with and without stress (Fig. 5.4). The rank order in clone fitness did not change significantly in the absence of stress, except for a strong average fitness cost of driver mutations in *FPR1*.

5.5 Driver mutations and ongoing diversification

To genetically characterise the subclonal variation before and after selection, we isolated and sequenced 192 clones drawn from WAxNA populations at $t = 0$ days (Fig. 5.5) and 44 clones at $t = 32$ days (Fig. 5.6). Overall, we identified 91 SNVs and indels in 173 ancestral haploid isolates and 140 point mutations in 44 evolved diploid isolates. We detected 82 SNVs and 1 insertion across 22 evolved isolates in hydroxyurea (range 1–8 per isolate), containing 10 adaptive mutations in *RNR2* and 12 in *RNR4* (Fig. 5.6A). There were 56 SNVs and 1 deletion across 22 evolved isolates in rapamycin (range 0–6 per isolate), which contained 8 adaptive mutations in *FPR1* and 5 in *TOR1* (Fig. 5.6B). 33 out of 36 mutations detected in WAxNA populations by whole-population sequencing could also be found in clonal isolates. All *de novo* driver mutations found by clone sequencing were confirmed by targeted Sanger sequencing. Assuming the ancestral genomes to have passed through ~150 generations during the crossing phase, we estimated a point mutation rate $\mu_{\mathrm{SNV+indel}} = 2.89\times10^{-10}$

per base per generation; and similarly for evolved genomes going through ~200 generations in the selection phase ($\mu_{\text{SNV+indel}} = 1.33 \times 10^{-9}$ bp^{-1}gen^{-1}). We detected two instances of cross-contamination between populations, so the derived events in clones isolated from these populations are valid to estimate the mutation rate but should not be counted to have arisen independently.

From population and isolate sequence data, we observed 19 recurrent *de novo* mutations in the ribonucleotide reductase subunits *RNR2* and *RNR4* during hydroxyurea selection and in the rapamycin targets *FPR1* and *TOR1* during rapamycin selection (Fig. 5.6). Each of

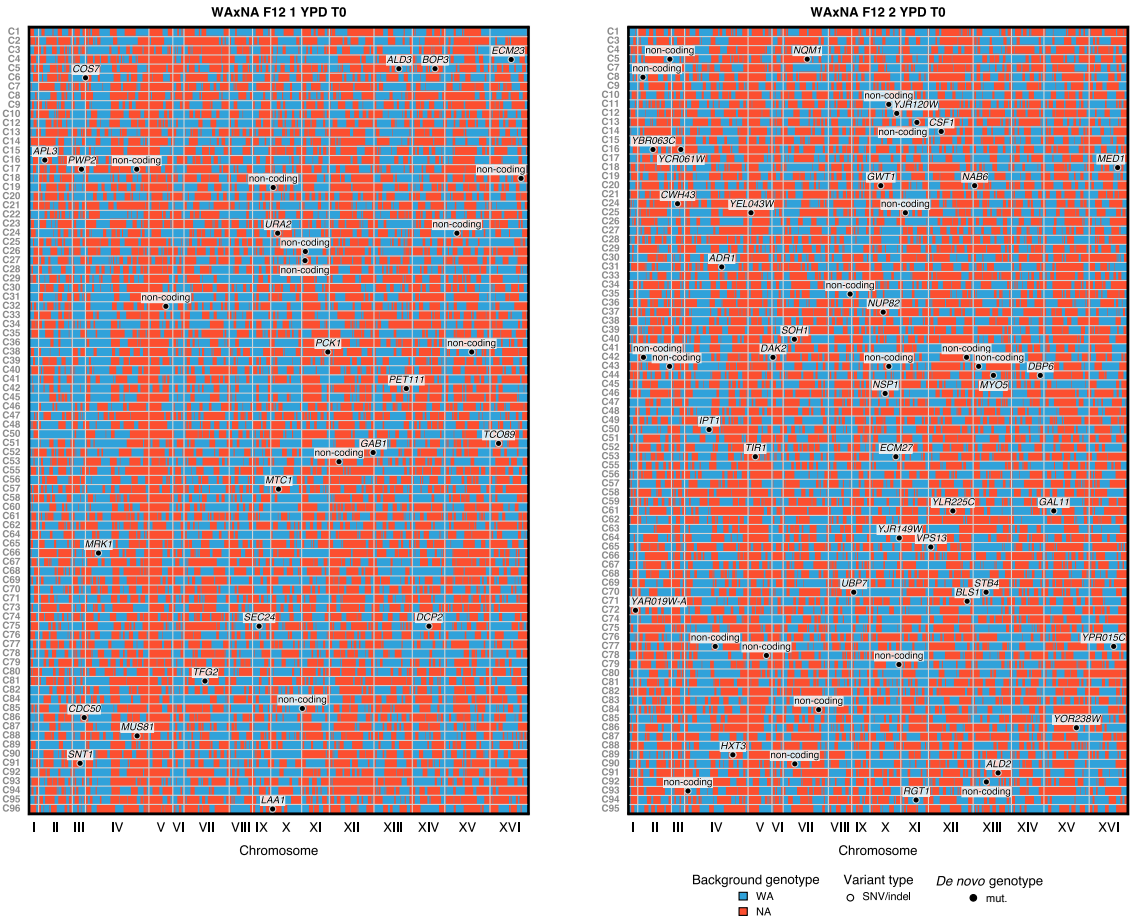


Fig. 5.5 Genetic heterogeneity in sequences of ancestral clones. Sequences of ancestral haploid clones sampled from the WAxNA F_{12} founder populations, which were obtained by bulk crossing between the WA and NA parents. Pre-existing and *de novo* SNVs and indels were detected by whole-genome sequencing in single-cell clones derived from ancestral populations at $t = 0$ days. Chromosomes are shown on the x-axis; clone isolates are listed on the left. WA (●) and NA (●) represent haploid genotypes. Individual cells with unique background genotypes carry private *de novo* SNVs and indels (circles). A copy-number gain of chromosome IX (n>2n) was also found in clone C50 of WAxNA F12 2 YPD T0 (not shown).

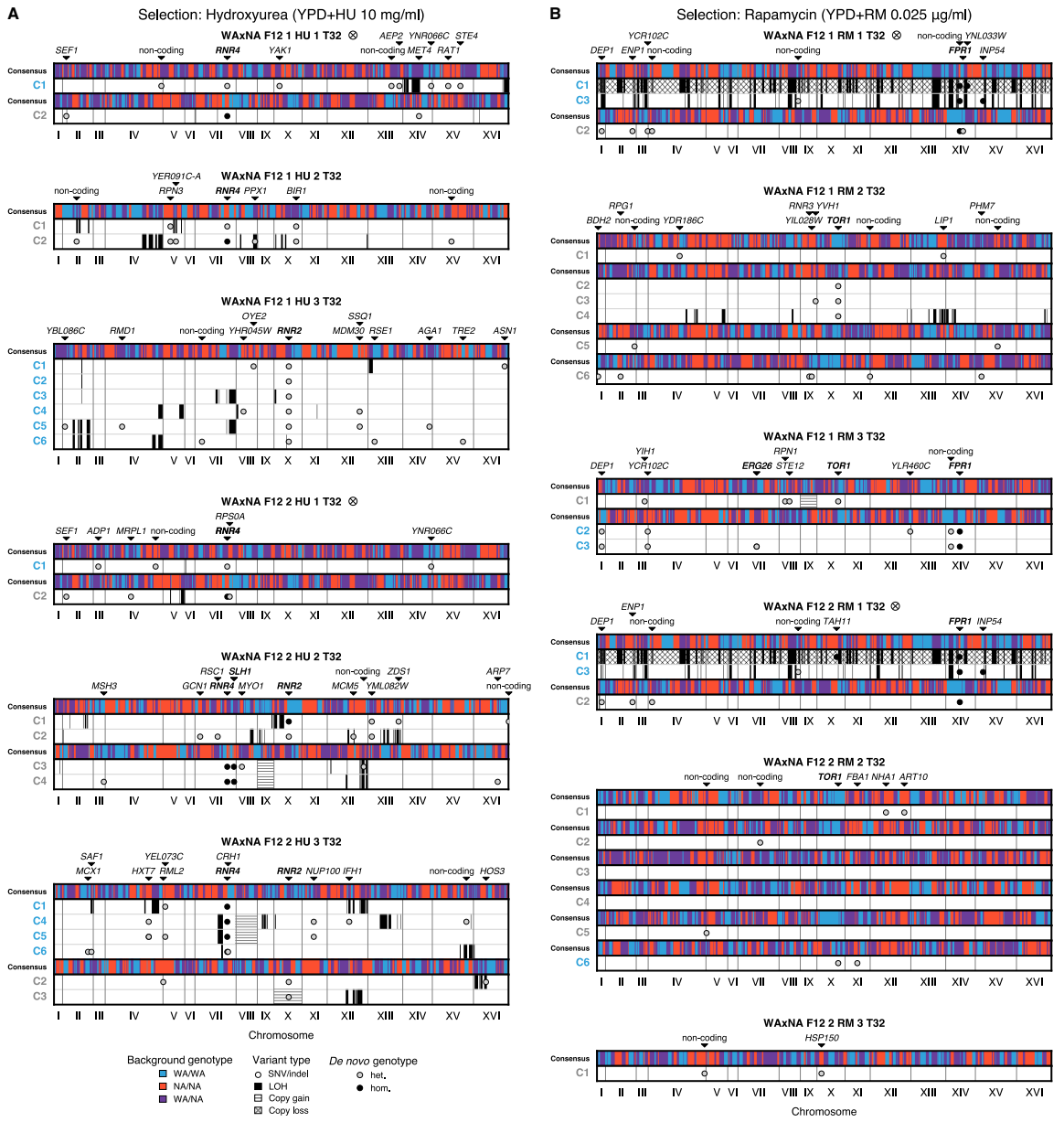


Fig. 5.6 Genomic instability in sequences of evolved clones. Sequences of evolved diploid clones sampled from WAxNA F_{12} populations. SNVs, indels and chromosome-level aberrations were detected by whole-genome sequencing in single-cell clones derived from evolved populations, after $t = 32$ days in **(A)** hydroxurea or **(B)** rapamycin (Table 5.1). Chromosomes are shown on the x-axis; clone isolates are listed on the left, coloured by lineage (see Figure 5.3). The consensus shows the majority genotype across population isolates with sequence identity greater than 80%. WA/WA (●) and NA/NA (●) represent homozygous diploid genotypes and WA/NA (●) represents a heterozygous genotype. Individual cells with shared background genotype carry *de novo* SNVs and indels (circles), *de novo* mis-segregations with loss-of-heterozygosity (solid segments) and *de novo* gains or losses in copy number (hatched segments). Driver and passenger mutations are listed along the bottom (drivers are in boldface). Populations marked by ⊗ indicate cross-contamination during the selection phase, but any derived events are independent.

these driver mutations had a drug-resistant growth rate phenotype (Figs. 5.8, 5.9 and 5.10) and carried a private background of ∼31,000 passenger mutations on average, compared to other sequenced isolates. All *FPR1* mutations were homozygous and likely to inactivate the gene or inhibit its expression. In contrast, *TOR1* mutations were heterozygous while we found *RNR2* and *RNR4* mutations in both heterozygous and homozygous state. The variant allele fractions of these mutations mirrored the inferred subclonal dynamics (Fig. 5.3A and C). All driver mutations occurred in highly conserved functional domains: 3 out of 4 unique variants in *RNR2* (N151H, E154G and Y169H) and 2 out of 3 unique variants in *RNR4* (R34G/I) mapped to a conserved domain of the ribonucleotide reductase small chain. *FPR1* mutations occurred at codon W66, either introducing a premature stop codon – truncating the residue required for rapamycin binding (Y89) – or changing to serine. All five driver SNVs in *TOR1* (S1972I/R, W2038L/C and F2045L) mapped to the FKBP12-rapamycin-binding (FRB) domain, which is ∼100 aa long, so they most likely disrupt drug binding (Fig. 5.6B).

To identify copy-number aberrations from clone sequencing, we segmented the coverage depth as a function of genomic position (see Sections 3.4 and 4.4.2). We found one copy-number gain (`n>2n`) of chromosome IX in ancestral haploid isolates. Evolved diploid isolates accrued copy-number gains (`2n>3n`) in chromosomes VIII, IX and X in hydroxyurea and chromosome IX in rapamycin, as well as a whole-genome copy loss (`2n>n`) in rapamycin (Fig. 5.6). In contrast to the recurrent point mutations, this evidence is inconclusive about whether they are adaptive, but we found that several of these gains are repeatedly acquired across a large ensemble of replicate populations in Chapter 4.

Using heterozygous genetic variants as markers, we could detect mis-segregation of chromosomes leading to loss-of-heterozygosity (LOH). The presence or absence of the WA or the NA allele provides a robust signal of heterozygosity or LOH that is not affected by sampling noise in coverage. We used the Hidden Markov Model for correlated SNVs introduced in Section 3.6.4 to genotype the sequenced isolate samples at segregating sites. Firstly, we used the sequences of haploid individuals from the ancestral population, drawn before the last round of crossing, to create *in silico* diploid genomes and calculate the length distribution of homozygous segments. Similarly, we measured the length distribution of homozygous segments from evolved isolate genomes. We observed a significant increase of long homozygosity tracts in the evolved clones – a hallmark of LOH (Fig. 5.7A). Secondly, we directly counted LOH events in populations using multiple sequenced isolates from the same expanding subclone. We grouped isolate sequences by subclone lineage, requiring at least 80% genotype similarity to belong to the same lineage. In hydroxyurea, this resulted in 22 isolates stemming from 8 clonal lineages, with more than a single isolate each. In rapamycin,

22 isolates were assigned to 4 clonal lineages, with more than a single isolate each. For each clonal lineage, we inferred its ancestral genotype. In case of a locus with a unique genotype across all isolates we assigned this to be the ancestral state. In all other cases we inferred the ancestral state to be heterozygous, as lost alleles cannot be regained. We then annotated all the isolates from each clone for LOH events. Figure 5.6 shows the inferred ancestral genotypes and the derived SNVs, indels, LOH events and copy-number variants, grouped by population and clonal background. To determine the rate of LOH events, we counted the number of independent events within a chromosome that have led to the gain or loss of the ancestral allele in the evolved isolate sequences. This estimate is challenging given the ancestral states contain both homozygous and heterozygous loci, so that the precise end points of individual LOH events are uncertain. To obtain a lower bound, we counted whether any isolate had undergone LOH affecting ≥10 consecutive background variants, for each chromosome in each clone. We found 48 events in hydroxyurea and 24 events in rapamycin (6 per genome per clone). We excluded two haploid individuals from this counting as well as from the length distribution of homozygosity tracts in Figure 5.7A.

To exemplify the interaction of genomic instability with pre-existing and *de novo* variation, inspection of *de novo* mutations in the WAxNA F12 1 HU 3 population shows that one *RNR2* mutation spans six isolates, being part of an expanding subclone (Fig. 5.6A). These isolates have further diversified by acquiring passenger mutations and undergoing LOH. Clones C5 and C6 grow faster than the other four and share a large LOH event in chromosome II that is not present in the other isolates, possibly providing the growth advantage and broadening the fitness distribution (Fig. 5.3B). An alternative route to homozygosity was observed in a single clone found to be haploid (clone C1 in WAxNA F12 2 RM 1) and therefore homozygous genome-wide. This haploid clone is closely related to a diploid clone (C3) from the same population and both clones share the same *FPR1* W66* *de novo* mutation (Fig. 5.6B). These data are consistent with the appearance of the *FPR1* heterozygous mutation in an ancestral diploid clone that took two independent routes – focal LOH or meiosis – to unveil the recessive driver mutation.

5.5.1 Luria-Delbrück fluctuation assay

We compared our genome-wide estimates of the point mutation and LOH rates based on the mutation counts in clone genome sequences to locus-specific measurements of the LOH rate using a fluctuation test. Our collaborators performed a fluctuation assay to determine the LOH rate by following the loss of a heterozygous *URA3* marker that results in 5-FOA-

resistant colonies. In all strains tested the *URA3* gene was deleted from its native location in chromosome V and inserted in the *lys2* locus (*lys2::URA3*) in chromosome II (~470 kb). The strains were first plated in *URA* dropout medium and then streaked for single colonies in stress and control environments. Colonies were grown for 3 days at 30 ˚C. Cells were resuspended in water and cell concentration was measured by flow cytometry to obtain a correct dilution factor in the subsequent plating. Cells from each replicate were plated in YPDA to determine total number of colony-forming units and 5-FOA plates (1 g l^{-1}) to count colonies that are *URA3*-defective. The loss of the *URA3* marker was confirmed by diagnostic PCR. Four replicates per experiment were used to determine the rate.

We fitted the fluctuation data to a model of the Luria-Delbrück distribution. Based on the fluctuation test, LOH rate can be estimated by $\mu_{\mathrm{LOH}} = \frac{m}{N}$, where N is the average number of cells per culture. To determine the mean number of LOH events m, we used the probability generating function of the Luria-Delbrück distribution defined by Hamon and Ycart [223].

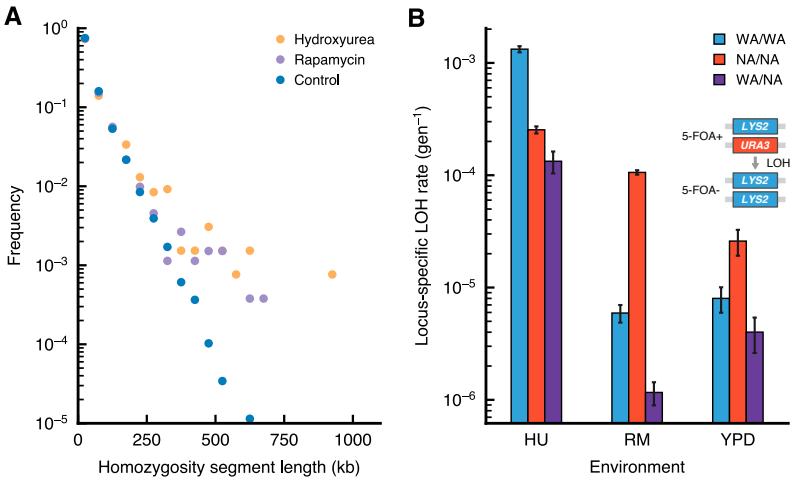


Fig. 5.7 Pervasive genomic instability. **(A)** The length distribution of homozygous segments, in bins corresponding to 50-kb increments, shows an excess of long homozygosity tracts above 300 kb in hydroxyurea and rapamycin (Kolmogorov-Smirnov test, $P < 0.01$). Ancestral haploid isolates are used to compare a set of *in silico* diploid genomes to evolved diploid isolates. Only unrelated isolate backgrounds were included. **(B)** Background- and environment-dependent rates of loss-of-heterozygosity were measured in a fluctuation assay by loss of the *URA3* marker. Resistant colonies growing in 5-fluororotic acid (5-FOA+) indicate loss of the marker. Based on the number of 5-FOA+ colony-forming units (c.f.u.), the mean number of LOH events are estimated using the empirical probability-generating function of the Luria-Delbrück distribution (Section 5.5.1). The locus-specific LOH rates are shown, given by the mean number of LOH events divided by the total number of cells in YPD. Error bars denote the upper and lower 95% confidence intervals. LOH rates were elevated in hydroxyurea compared with the control environment and manifested background-dependent effects between the parents and their hybrid.

In the control environment, we observed a rate of $\mu_{\text{LOH}} = 2.59 \times 10^{-5}$ per generation in the NA background, consistent with previous reports [224]. We observed an intermediate rate in the WA background ($\mu_{\text{LOH}} = 8.01 \times 10^{-6}\,\text{gen}^{-1}$) and the WAxNA F_1 hybrid had an approximately ten-fold lower rate ($\mu_{\text{LOH}} = 4.01 \times 10^{-6}\,\text{gen}^{-1}$). These data indicate that LOH rates can vary between genetic backgrounds. The stress environments themselves also have an active role in accelerating genome evolution by genomic instability. There was a sharp increase of LOH rates when colonies were grown in hydroxyurea, irrespective of the background tested. This finding is consistent with previous studies in the laboratory strain S288C reporting that replication stress promotes recombinogenic DNA damage [224]. We also observed a background-dependent increase in LOH rate in the presence of rapamycin, especially in the NA founder. Our estimates of the point mutation rate based on the mutation counts in ancestral and evolved clones ($\sim 10^{-10}\,\text{bp}^{-1}\,\text{gen}^{-1}$) and of the LOH rate based on the fluctuation assay ($\sim 10^{-5}\,\text{gen}^{-1}$), suggest that any recessive genes will be likely to lose the wild-type allele by LOH. Given that the LOH rate is much higher than the point mutation rate and it typically affects large regions (100-1,000 kb, see Figure 5.7A), recessive mutations can feasibly be 'rescued' by LOH.

5.5.2 Validation of candidate driver mutations

To test candidate driver mutations, our collaborators engineered hemizygous strains to compare allelic differences in driver genes with pre-existing and newly acquired mutations (Fig. 5.8). They also engineered gene deletions of driver genes to confirm whether their knockouts are beneficial. We performed $n = 64$ replicate measurements of each construct in two independent runs, which were initiated from a single pre-culture plate, evenly distributed over 16 experimental plates and simultaneously run in 4 scanners. The growth rate of each of these strains λ is shown in Figures 5.9 and 5.10, labelled by genetic background b and genotype g.

Reciprocal hemizygosity tests in ancestral hybrids confirmed background-dependent effects in *CTF8*, with strong positive selection on the NA allele as predicted by our model of driver-passenger dynamics (Fig. 5.10). *KOG1*, which is a component of the TOR signalling pathway, did not show any allelic differences, but deleting either copy caused haploinsufficiency in rapamycin. No allelic differences were observed for *DEP1*, *INP54* and *YNR066C*, which are confirmed as passengers. Our collaborators also deleted either the wild-type or the mutated allele of evolved mutant clones, generating pairs of clones identical throughout the genome except for the candidate driver mutation. The four genes harbouring *de novo* driver

mutations do not appear to show allelic differences between the two parental backgrounds as shown by the reciprocal hemizygosity test (Fig. 5.10).

Deleting one copy of *RNR2* in WA and NA diploids and sporulating these strains resulted in tetrads with two viable spores and two unviable *rnr2*Δ mutants, indicating that this gene is essential in both backgrounds. *RNR2* is also essential in the laboratory S288C background, consistent with its key role in catalysing the rate-limiting step of dNTP synthesis. Furthermore, the heterozygous deletions of *RNR2* diploids show strong haploinsufficiency for hydroxyurea resistance (Fig. 5.9). In contrast to its interaction partner, *RNR4* is not essential in the laboratory background. However, deletion of this gene in diploid WA and NA

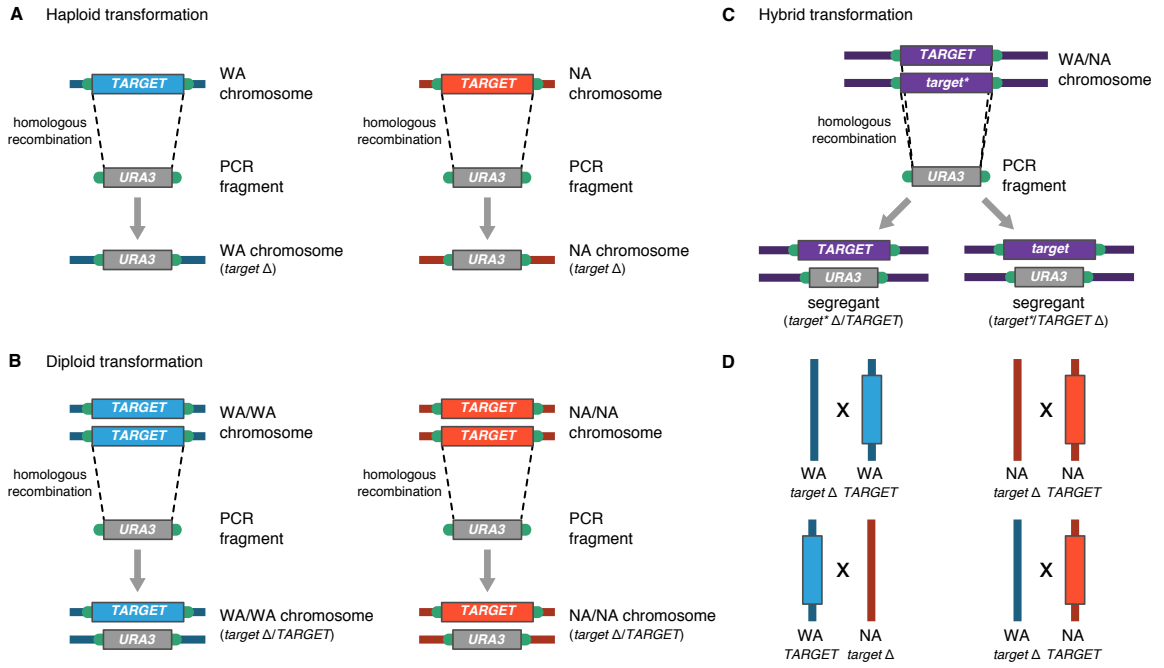


Fig. 5.8 Strategy for engineered genetic constructs. Gene deletions were introduced by homologous recombination between the terminals of the PCR product and the corresponding genomic sequence where the gene to be deleted ('target') is encoded. Blue and red lines indicate WA and NA chromosomes, respectively. Flanking regions in green indicate two different homologous sequences targeted for recombination, which are 30-40 bp long in budding yeast. **(A)** Genes of interest were individually deleted in both WA and NA haploids, resulting in *rnr4*Δ, *fpr1*Δ and *tor1*Δ strains in both parental backgrounds. **(B)** A similar strategy was used to delete genes in WA and NA homozygous diploids. *RNR2* and *RNR4* were only deleted in one allele while there is the wild-type gene remaining in the other allele. **(C)** Evolved segregants with *de novo* mutations were isolated from the WAxNA F_{12} populations. Using the same strategy, *RNR2* or *TOR1* mutants could be rid of either the wild-type allele or the mutated allele. **(D)** The strain constructed in (A) was crossed with the parental strain with wild-type gene to obtain strains with deleted genes in WA and NA homozygous diploids and WA/NA hybrid.

backgrounds proved it to be essential in the WA background. The NA strain is viable after deletion, though with severe growth defects. Diploid hemizygous strains for *RNR4* deletions in both backgrounds show increased sensitivity due to dosage effects (Fig. 5.9). *FPR1* and *TOR1* are not essential genes and our collaborators could perform deletions in both haploids and diploids. *FPR1* directly binds rapamycin, inhibiting the TOR pathway, and its deletion is highly penetrant (Fig. 5.10). Deletion of one copy of *FPR1* does not increase the growth rate in rapamycin, confirming that this gene is recessive. This is consistent with our observations that all *FPR1* mutations observed in this experiment are homozygous, and we found most of them typically reach fixation in the experiments with a large ensemble of replicate populations reported in Chapter 4 (see Figure 4.10). In contrast, *TOR1* deletion results in

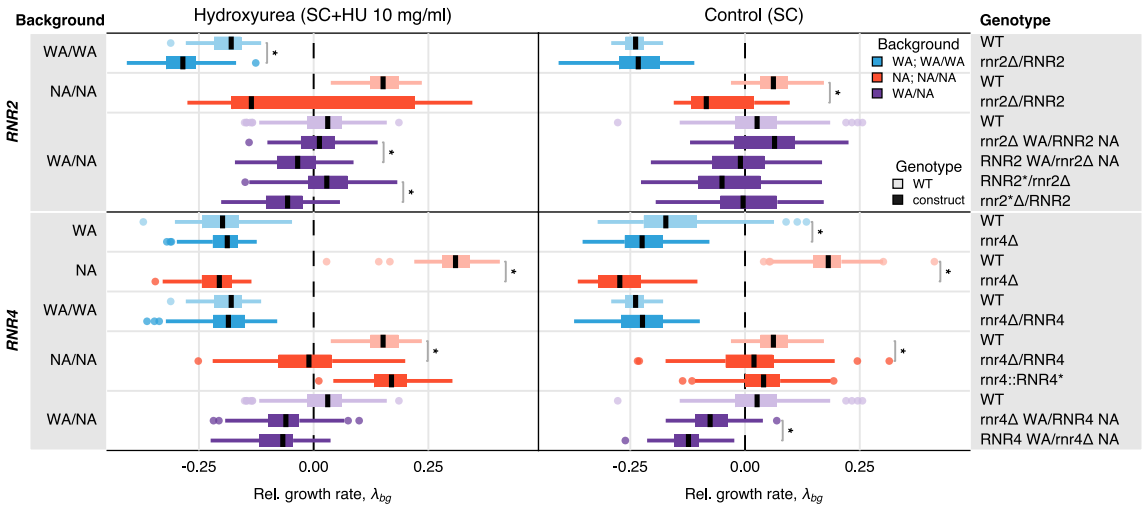


Fig. 5.9 Validation tests for driver mutations in hydroxyurea, measured in SC+HU (left panel) and SC (right panel). The relative growth rate, λ_{bg}, of each construct is shown for $n_r = 64$ measurement replicates. Genetic constructs are grouped by candidate gene and by background of the construct, where the background b can be WA, NA (haploid); WA/WA, NA/NA (diploid); WA/NA (hybrid), and the genotype g can be wild-type for the gene, deleted or hemizygous. Relative growth rates are normalised with respect to the mean population growth rate $\langle \lambda_k \rangle_{t=0}$ at $t = 0$ days (see Figures 5.3B and 5.4A). Medians and 25%/75% percentiles are shown for each genetic construct, with medians as horizontal lines and outliers highlighted. The colour of each of the boxes reflects the background (WA and WA/WA, ●; NA and NA/NA, ●; WA/NA, ●). Lighter shades indicate a wild-type (WT) control for a specific background and darker shades are the candidate strains. For a given background, we compared deletion strains against their respective WT control (e.g., *rnr4*Δ vs WT in WA background) and hemizygous strains against the equivalent hemizygous strain where the opposite copy has been deleted (e.g., *rnr4*Δ WA/*RNR4* NA vs *RNR4* WA/*rnr4*Δ NA in WA/NA background). To test statistical significance we used a non-parametric Wilcoxon rank-sum test. Significance tests between two strains with $P < 10^{-4}$ are highlighted with an asterisk. *RNR2* and *RNR4* are confirmed as driver genes for hydroxyurea resistance.

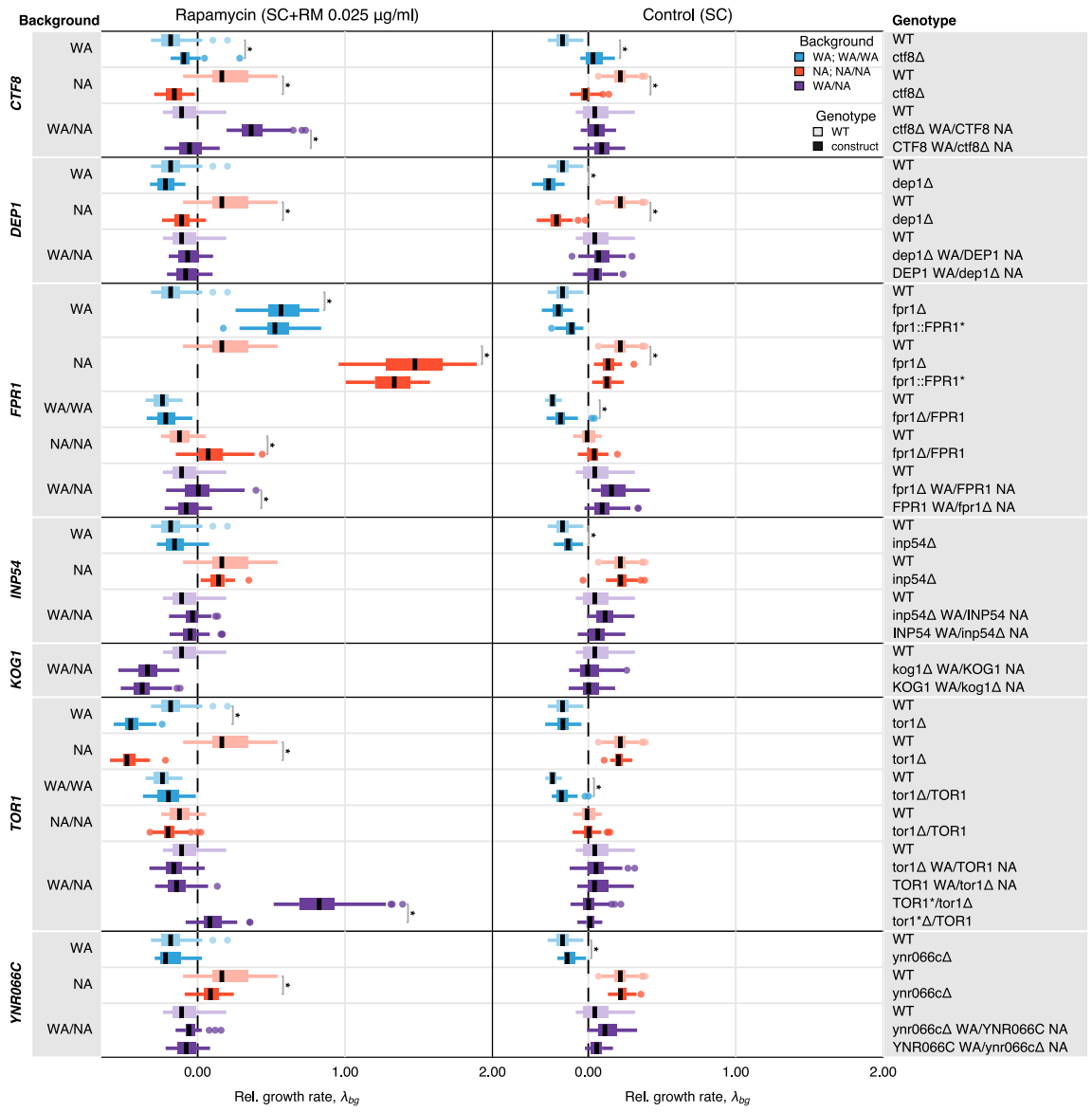


Fig. 5.10 Validation tests for driver and passenger mutations in rapamycin, measured in SC+RM (left panel) and SC (right panel). The relative growth rate, λ_{bg}, of each construct is shown for $n_r = 64$ measurement replicates. Genetic constructs are grouped by candidate gene and by background of the construct, where the background b can be WA, NA (haploid); WA/WA, NA/NA (diploid); WA/NA (hybrid), and the genotype g can be wild-type for the gene, deleted or hemizygous. Relative growth rates are normalised with respect to the mean population growth rate $\langle \lambda_k \rangle_{t=0}$ at $t = 0$ days (see Figures 5.3D and 5.4B). Medians and 25%/75% percentiles are shown for each genetic construct, with medians as horizontal lines and outliers highlighted. The colour of each of the boxes reflects the background (WA and WA/WA, ●; NA and NA/NA, ●; WA/NA, ●). Lighter shades indicate a wild-type (WT) control for a specific background and darker shades are the candidate strains. For a given background, we compared deletion strains against their respective WT control (e.g., *fpr1*Δ vs WT in WA background) and hemizygous strains against the equivalent hemizygous strain where the opposite copy has been deleted (e.g., *fpr1*Δ WA/*FPR1* NA vs *FPR1* WA/*fpr1*Δ NA in WA/NA background). To test statistical significance we used a non-parametric Wilcoxon rank-sum test. Significance tests between two strains with $P < 10^{-4}$ are highlighted with an asterisk. *FPR1* and *TOR1* are confirmed as driver genes for rapamycin resistance.

high sensitivity to rapamycin and a single deleted copy does not alter the drug response (Fig. 5.10).

5.6 Ensemble measurements of fitness effects

Finally, we sought to partition and quantify the individual fitness contributions of pre-existing and *de novo* genetic variation. The genotype space is extremely vast, but we can uniformly sample a representative ensemble to reconstruct a fraction of the genetic backgrounds where beneficial mutations could have arisen. To this end, we designed a genetic cross where background and *de novo* variants are re-shuffled to create new combinations (Fig. 5.11A). We randomly isolated diploids from both ancestral and evolved populations, sporulated these and determined whether the derived haploids contained wild-type or mutated *RNR2*, *RNR4*, *FPR1* and *TOR1* alleles. We then crossed haploids to create a large array of diploid hybrids where all genotypes (+/+, +/–, –/–) for each of these genes exist in an ensemble of backgrounds, thus recreating a large fraction of the genotype space conditioned on the presence or absence of driver mutations. By measuring the growth rates of both haploid spores and diploid hybrids, we can estimate and partition the variation in fitness contributed by the background genotype and *de novo* genotypes using a linear mixed model (Figs. 5.12 and 5.15).

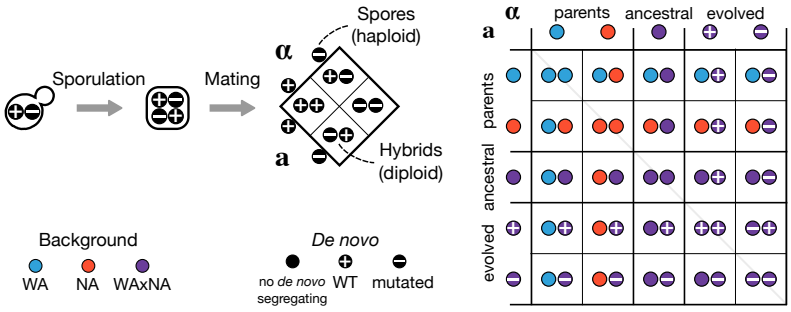


Fig. 5.11 Experimental outline of ensemble fitness measurements using a recombinant library. To determine fitness effects of background variation and *de novo* mutations in hydroxyurea (*RNR2*, *RNR4*) and rapamycin (*FPR1*, *TOR1*), we isolated individuals from ancestral and evolved populations. From these diploid cells, we sporulated and selected haploid segregants of each mating type. Spores with mutations in *RNR2*, *RNR4* and *TOR1* were genotyped to test if they carry the wild-type or mutated allele. We crossed the *MAT*a and *MAT*α versions to create hybrids (48×48 in hydroxyurea and 56×56 in rapamycin). Independent segregants were used to measure biological variability of ancestral and evolved backgrounds.

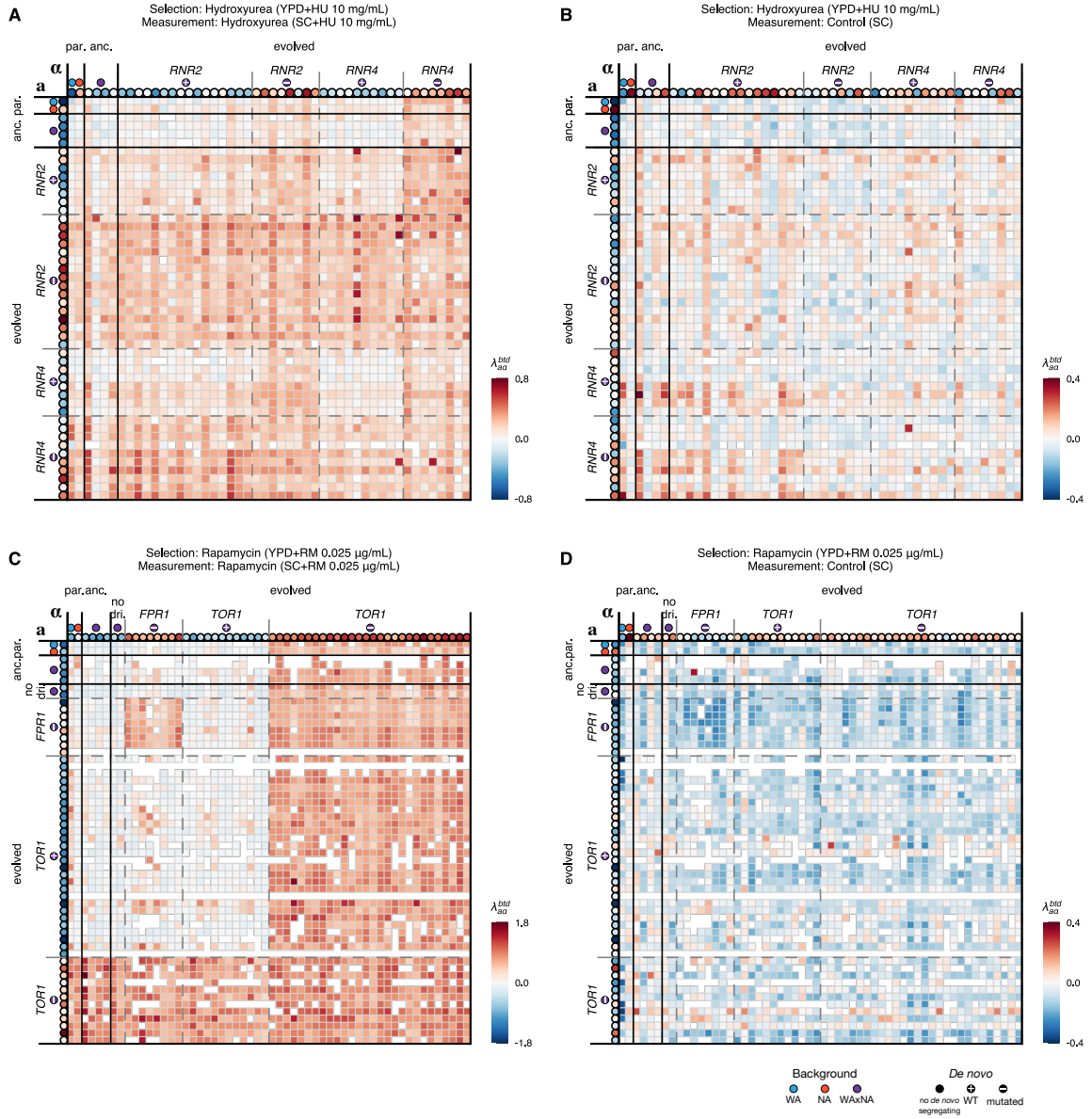


Fig. 5.12 Fitness contribution of genetic background and *de novo* mutations. Given an ensemble of n_b unique genetic backgrounds ($n_b = 48$ in hydroxyurea and $n_b = 56$ in rapamycin), we constructed a matrix of size $n_b \times n_b$ where every unique haploid background is crossed against itself and all other haploid backgrounds, and the two must be of opposite mating type (*MAT*a or *MAT*α). Each matrix element is labelled by background genotype, b; *de novo* genotype, d; time, t; and auxotrophy, x. Measurements of relative growth rates of spores $\lambda^{btd}_{\{a,\alpha\}}$ and hybrids $\lambda^{btd}_{a\alpha}$ are shown, normalised with respect to the ancestral WAxNA crosses. Measurements are taken in **(A)** SC+HU 10 mg ml^{-1} and **(B)** SC; **(C)** SC+RM 0.025 µg ml^{-1} and **(D)** SC, respectively. The colour scale for all matrices to the right of each panel indicates the relative fold-change with respect to the ancestral WAxNA crosses. White boxes indicate missing data due to mating inefficiency and slow growth. All panels follow the legend in Figure 5.11.

5.6.1 Genetic cross

The genetic cross included the parents, ancestral and evolved isolates. Our collaborators derived haploid lines by sporulation on KAc medium from the ancestral and evolved clones. Only tetrads with four viable spores were chosen for continuation in the experiment. Spores were genotyped for mating type (*MAT*a, *MAT*α) using tester strains and for auxotrophies (*ura3*, *lys2*) by plating on dropout medium. We chose spores from tetrad configurations with the mating marker co-segregating as *MAT*a, *ura3* or *MAT*α, *lys2*, allowing a systematic cross between all strains of opposite mating type. Therefore, the WA and NA haploid parents were used in *MAT*a, *ura3* and *MAT*α, *lys2* configurations. Eight ancestral haploid segregants (4 *MAT*α, *lys2* and 4 *MAT*a, *ura3*) were randomly isolated from the ancestral population. For the hydroxyurea environment, we probed individually beneficial *de novo* mutations in *RNR2* (Y169H) and *RNR4* (R34I). The *RNR2* mutant was isolated from WAxNA F12 1 HU 3 (clone C3) and the *RNR4* mutant from WAxNA F12 2 HU 1 (clone C1) at $t = 32$ days. For rapamycin, three evolved clones isolated at $t = 32$ days were used: one clone with no identifiable driver from WAxNA F12 2 RM 2 (clone C1), a homozygous *FPR1* mutant (W66*) from WAxNA F12 2 RM 1 (clone C3); and a heterozygous *TOR1* mutant (W2038L) from WAxNA F12 1 RM 2 (clone C3). Our collaborators then determined whether each spore inherited the wild-type or the mutated allele by Sanger sequencing of the candidate gene.

A genetic cross of size 48×48 in hydroxyurea yielded 2,304 hybrids, and 56×56 in rapamycin, giving 3,136 hybrids. Our collaborators performed the genetic cross using the Singer RoToR HDA robot on YPDA plates. Subsequently, the hybrid populations were grown for two rounds on minimal medium to ensure colonies of solely diploid cells and avoid haploid leakage. A small number of crosses were not successful due to mating inefficiency or slow growth (56 in hydroxyurea and 654 in rapamycin), leaving a total of 2,248 and 2,482 hybrids, respectively. This was due to mistyping of the mating locus in one *FPR1* spore and three *TOR1* spores, which were excluded together with their derived hybrids. Phenotypic measurements of the crosses were carried out using the high-throughput method of yeast colony growth described in Section 5.3.

5.6.2 Fitness effects of pre-existing and *de novo* variation

We obtained a set of measurements for the growth rate λ of individuals, each of which has a unique combination of background genotype b, *de novo* genotype d, sampling time t and auxotrophy x. Every haploid genome being crossed is an independent background indexed

by $b_{\{a,\alpha\}} = \{1, 2, \ldots, n_b\}$ ($n_b = 48$ in HU and $n_b = 56$ in RM, either a or α), such that reshuffled diploid hybrids are parameterised by $b_{a\alpha}$. Genetic backgrounds are sampled before the cross (parents), before selection starts at $t = 0$ (ancestral) or after $t = 32$ days (evolved), and are labelled by $t_{\{a,\alpha\}} = \{1, 2, \ldots, n_t\}$ ($n_t = 2$ for the parents; $n_t = 4$ at $t = 0$; $n_t = 42$ in HU and $n_t = 46$ in RM at $t = 32$). We denote *de novo* genotypes by $d_{\{a,\alpha\}} = \{1, 2, \ldots, n_d\}$ ($n_d = 12$ for *RNR2*; $n_d = 9$ for *RNR4*; $n_d = 1$ without driver; $n_d = 4$ for *FPR1*, $n_d = 20$ for *TOR1*). Haploid spores are auxotroph and segregate with the mating locus, such that $x_{\{a,\alpha\}} \in$ {*ura3-*, *lys2-*}, whereas diploid hybrids do not have amino acid deficiencies. To estimate the measurement error, we carried out n_r replicate measurements of each unique

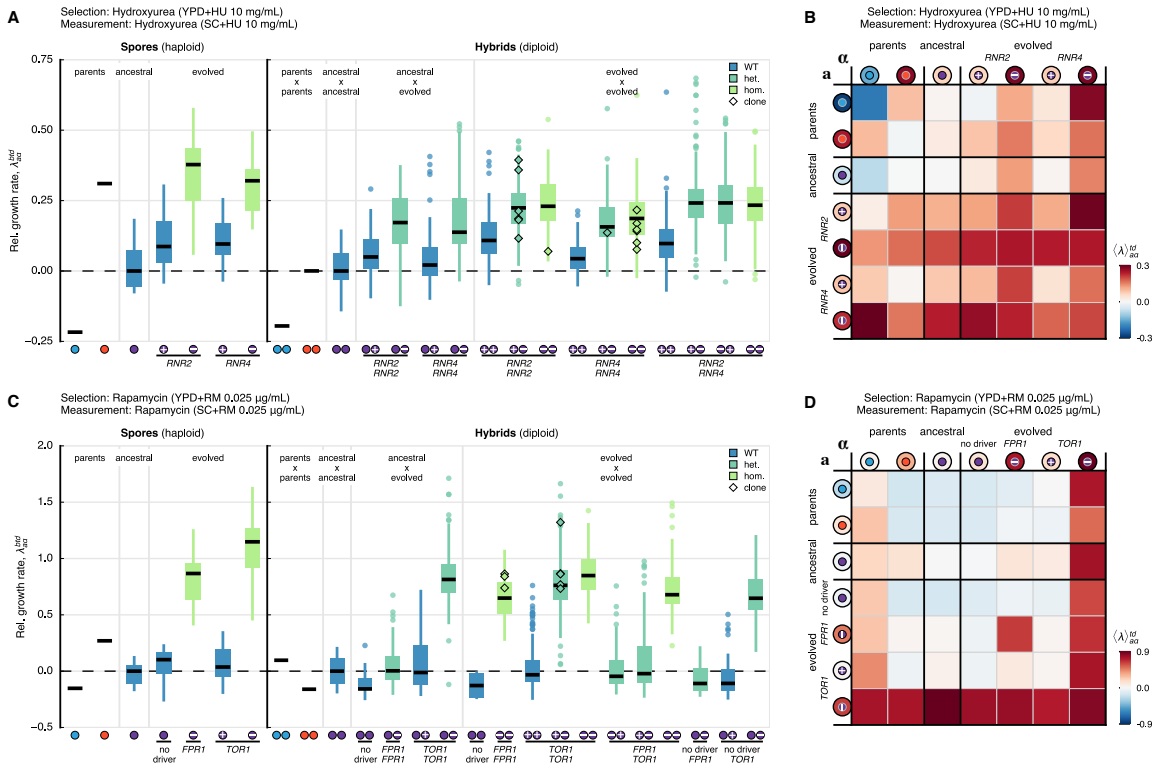


Fig. 5.13 Ensemble-averaged fitness effects of genetic background and *de novo* mutations. **(A, C)** Relative growth rate, λ, measured with respect to the ancestral population for multiple combinations (background genotype, b; *de novo* genotype, d; time of sampling during the selection phase, t; auxotrophy, x) and averaged over measurement replicates. Medians and 25%/75% percentiles across groups are shown, with medians as horizontal black lines and coloured by *de novo* genotype [wild-type +/+ (●); heterozygote +/– (●); homozygote –/– (●)]. Outliers (circles) and isolated, selected clones with matching genotypes (diamonds) are highlighted. **(B, D)** Ensemble average of the relative growth rate of spores, $\langle\lambda\rangle^{td}_{\{a,\alpha\}}$, and hybrids, $\langle\lambda\rangle^{td}_{a\alpha}$, measured in (B) hydroxyurea and (D) rapamycin. The colour scale for all matrices is shown at the right and indicates the difference in the ensemble average with respect to the ancestral WAxNA crosses. All panels follow the legend in Figure 5.11. An extended version of the figure with all combinations can be found in Figure 5.12.

spore ($n_r = 12$ in HU and $n_r = 6$ in RM) and of each hybrid genotype combination ($n_r = 3$). Replicates were initiated from the same pre-culture plate, evenly distributed over 32 plates and run in 4 scanners, all in parallel.

The data matrix shows the fitness effect of every *de novo* genotype d at each background b sampled at time t, averaged over measurement replicates and measured relative to the ancestral population (Fig. 5.12). Based on these measurements, we observed that *de novo* mutations are beneficial, yet their associations to genetic backgrounds have idiosyncratic effects. The effects of *de novo* mutations are mediated by background fitness as evidenced by the large phenotypic variance. Genetic crosses between different backgrounds need not give rise to a 'symmetric' phenotype, as we only enforce 2:2 segregation for the mating locus *MAT*a/α. Whilst background variants will co-segregate with the mating locus, *de novo* mutations need not.

To examine the average fitness effects of functional genotypes in hydroxyurea (*RNR2*, *RNR4*) or rapamycin (*FPR1*, *TOR1*), we calculated an ensemble average of the growth rate λ. The ensemble average $\langle\lambda\rangle$ is either taken over single spore backgrounds $b_{\{a,\alpha\}}$ or pairs of hybrid backgrounds $b_{a\alpha}$ with different degrees of relatedness,

$$\langle\lambda\rangle^{td}_{\{a,\alpha\}} = \frac{1}{n_b}\sum_{b=1}^{n_b}\lambda^{btd}_{\{a,\alpha\}} \quad \text{and} \quad \langle\lambda\rangle^{td}_{a\alpha} = \frac{1}{n_b}\sum_{b=1}^{n_b}\lambda^{btd}_{a\alpha}\,, \tag{5.2}$$

where $\langle\cdots\rangle$ denotes the mean over genetic backgrounds. The ensemble average over backgrounds shows that the mean effect of *RNR2*, *RNR4* and *TOR1* mutations is fully dominant and highly penetrant regardless of the background (Fig. 5.13B and D). In contrast and as we already observed, *FPR1* mutants are recessive and only increase growth rate when homozygous, again irrespective of the background (Fig. 5.13D). Recombinants with *RNR2* and *RNR4* mutations show epistatic interactions, consistent with the products encoded by these genes which are known to interact as subunits of the same protein complex (Fig. 5.13A). After conditioning for *RNR2*, *RNR4*, *FPR1* and *TOR1* driver mutation status, a large fraction of the phenotypic variance still remained, reflecting the effect of the genetic backgrounds in which they emerged (Fig. 5.13A and C).

5.6.3 Variance decomposition

We would like to partition the variation in fitness contributed by background and *de novo* driver mutations using linear mixed models. To model genetic backgrounds containing beneficial mutations we need to describe how likely a phenotype is in the presence or absence

of any mutation. We restrict our model to pairs of individuals that are not closely related to avoid spurious correlations by population structure, so we retain ancestral and evolved individuals and exclude the parents. We are interested in the aggregate effect across all mutations within a spore or hybrid rather than the effects of individual variants. As the data represents a finite sample from the distribution of all possible genetic backgrounds, the background contribution to the phenotype is naturally modelled as a random-effect term (i.e., individual genetic backgrounds are drawn at random from a population, and the variance of the underlying distribution is to be inferred). In addition, other systematic effects that potentially contribute to fitness are modelled as fixed-effect terms: (i) time t when the individual was sampled, i.e., at $t = 0$ (ancestral) or $t = 32$ days (evolved); (ii) *de novo* driver mutation status d of the individual, e.g., *FPR1* driver mutation in homozygous state; and (iii) auxotrophy, denoted by x, e.g., *ura3-* or *lys2-*. We implemented four nested linear mixed models outlined below.

Model 1

We first considered a model where we only account for the background, without other effects. This means that the observed growth rate λ_b for a background b conditioned on the random effect taking a value β_b is distributed as

$$\lambda_b\big|_{\mathcal{B}=\beta_b} \sim \mathcal{N}\left(\beta_0 + \beta_b x_b, \sigma_\epsilon^2\right), \tag{5.3}$$

where β_0 is a shared constant baseline per background that must be inferred, σ_ϵ^2 represents measurement noise, x_b is an element from the model design matrix (here 1 for each b as they all are assigned a value). Finally, the background growth rate is distributed as $\mathcal{B} \sim \mathcal{N}\left(0, \Sigma^2\right)$ and its variance Σ^2 is a model parameter to be inferred. We note that for each background b we have multiple measurement replicates of λ_b. Altogether, Model 1 has three modelling parameters, β_0, Σ^2 and σ_ϵ^2.

Models 2, 3 and 4

Model 2 includes the same factors as Model 1, but the time of sampling t is nested as a fixed effect. Model 3 also accounts for *de novo* driver mutation status denoted by d. In addition, Model 4 includes a fixed effect accounting for amino acid deficiencies (or auxotrophy), denoted by x. Altogether the growth rate λ_{btdx}, conditioned on the random effect taking a value

β_b, is distributed as

$$\lambda_{btdx}\big|_{\mathcal{B}=\beta_b} \sim \mathcal{N}\left(\beta_0 + \underbrace{\beta_b x_b}_{\text{random}} + \underbrace{\beta_t x_t + \beta_d x_d + \beta_x x_x}_{\text{fixed}}, \sigma_\epsilon^2\right), \tag{5.4}$$

where β_t, β_d, β_x are fixed-effect terms to be inferred and x_t, x_d, x_x are elements of the model design matrix. Compared to Model 1, Models 2, 3 and 4 have extra parameters β_t, β_d, and β_x. The number of free parameters depends on how many unique levels each factor contains, e.g., how many driver mutations are sampled in the experiment.

The likelihood for a data vector λ given the full model (Model 4) can then be written as

$$\begin{aligned} P(\lambda \mid \text{model}) &= P\left(\lambda \,\middle|\, \beta_0, \beta_t, \beta_d, \beta_x, \Sigma^2, \sigma_\epsilon^2\right) \\ &= \prod_{1 \le a < \alpha \le n_b} \prod_{r=1}^{n_r} \int P\left(\lambda_{btdx} \,\middle|\, \beta_b, \beta_0, \beta_t, \beta_d, \beta_x, \Sigma^2, \sigma_\epsilon^2\right) \times P\left(\beta_b \,\middle|\, \Sigma^2\right) d\beta_b \end{aligned}$$

where the integrand is the product of the probability density given by Equation (5.4) and the posterior distribution over the random effects.

Next, we applied all four models to the fitness measurements of the genetic cross: a genetic cross based on hydroxyurea selection, measured in hydroxyurea and a control environment; and a genetic cross based on rapamycin selection, measured in rapamycin and a control environment, both for spores and hybrids. We fitted each model using restricted maximum likelihood. Using Akaike's Information Criterion (AIC) for model selection, Model 4 scored highest across all environments apart from those selected and measured in hydroxyurea, where both spores and hybrids supported Model 3. We compared the fitted and observed values and in all cases the fits were good, as shown in Figure 5.14 for Model 4.

We can assess the overall goodness-of-fit of the models by the proportion of variance explained. In particular, we would like to know the contribution of various model components to the overall fit, and to do so we obtain separate measures for the partial contributions of fixed and random effects, $r^2 = \frac{\sigma_F^2 + \sigma_R^2}{\sigma_F^2 + \sigma_R^2 + \sigma_\epsilon^2}$, where σ_R^2 is the variance contribution by random effects, any incremental fixed effect contributes additively to the fixed-effect variance, s.t. $\sigma_F^2 = (\beta_t x_t + \beta_d x_d + \beta_x x_x)$, and r^2 represents the proportion of variance explained by the fixed and random effects combined. Dropping the σ_R^2 term from the numerator, we can evaluate r^2 and the fixed-effects variance r_F^2 for linear mixed models, as described by Gelman and Hill [225], and estimate the background contribution to the variance by $r^2 - r_F^2$. Then to further delineate the fixed-effect variances to individual contributions, we used the

simpler models and their estimated r^2_F. Estimates of the variance components are shown in Figure 5.15. We note that modelling the background component using fixed effects instead leads to a variance decomposition that is nearly identical to that with linear mixed models as described here. However, we note that modelling the background as a fixed effect leads to a large number of parameters (one extra parameter per background) and thus describing the background by random effects is a better model for the data.

As shown in Figure 5.15, background genetic variation accounted for an estimated 51% of the growth rate variance under hydroxurea exposure, more than twice the estimated 23% contributed by *RNR2* and *RNR4 de novo* mutations. Furthermore, these driver mutations have landed on genetic backgrounds much fitter than average in the ancestral fitness distribution, as denoted by the estimated 7% explained by the time of sampling. Both of these results directly imply that moderate-effect *de novo* mutations must arise on favourable genetic backgrounds to give rise to macroscopic subclones. In contrast, under rapamycin exposure,

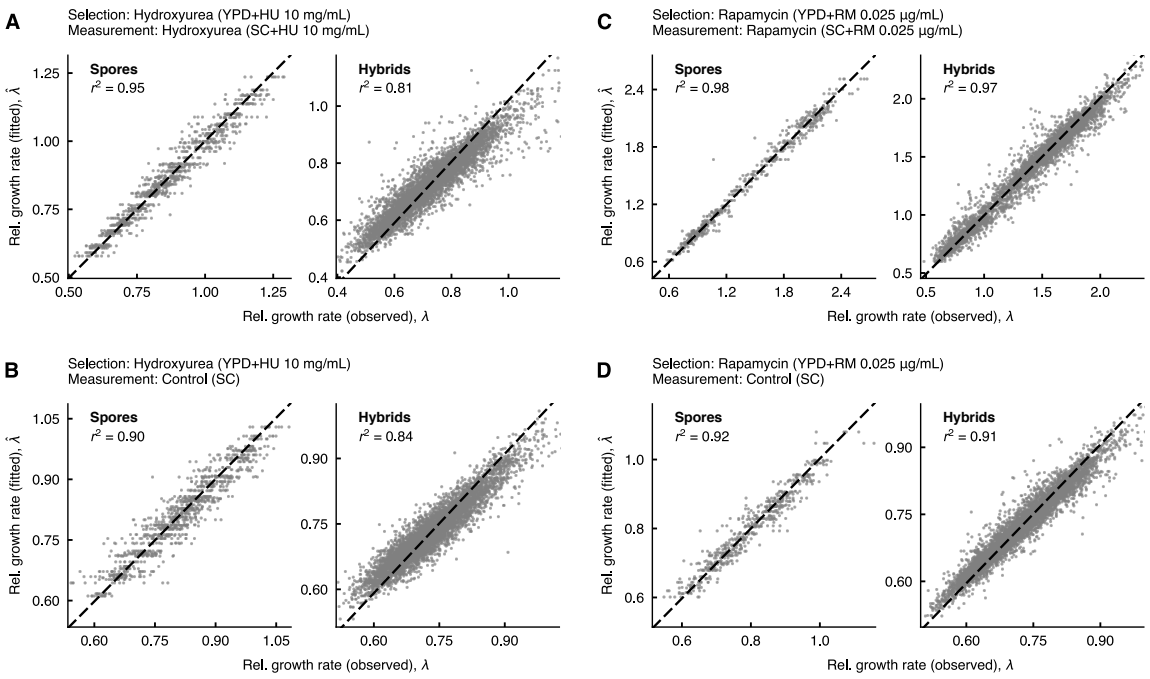


Fig. 5.14 Hierarchical analysis of variance in the genetic cross using linear mixed models. We model the growth rate of spores, $\lambda^{btd}_{\{a,\alpha\}}$, and hybrids, $\lambda^{btd}_{a\alpha}$, as a function of background genotype b, *de novo* genotype d, sampling time during selection t, and auxotrophy x. Relative growth rates are accurately fitted by this model (Model 4). Measurements are taken in SC+HU 10 mg ml^{-1} and SC only for populations selected in hydroxyurea **(A, B)**; SC+RM 0.025 µg ml^{-1} and SC only for populations selected in rapamycin **(C, D)**. The scatter shows a set of measurements λ (x-axis) against the fitted rates $\hat{\lambda}$ (y-axis). The total variance explained, r^2, is separately computed for spores and hybrids by environment.

the pre-existing genetic variation accounted for only 22% of the variance, much less than the 70% attributed to *FPR1* and *TOR1* mutations. Such large-effect mutations can expand in a vast majority of backgrounds, explaining how they can almost entirely surpass the bulk of the fitness distribution (Fig. 5.3D). Taken together, these results are consistent with the aggregation of small-effect, pre-existing variants which can condition the fate of new mutations in both selection environments.

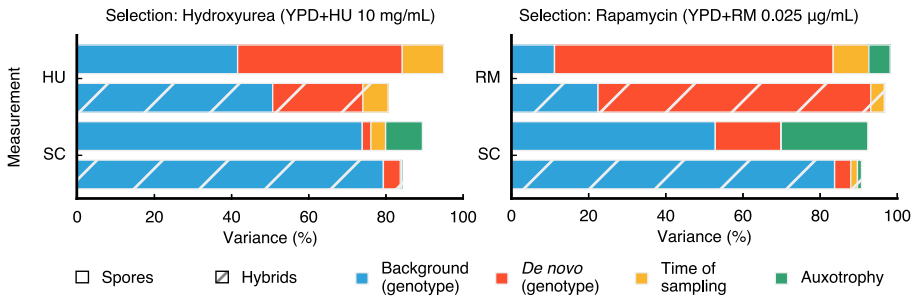


Fig. 5.15 Variance decomposition of the growth rate of spores (solid) and hybrids (hatched) that can be attributed to effects of background and *de novo* genotypes, to the time of sampling during selection and to any potential auxotrophies. Estimates of variance components are obtained by fitting the linear mixed models using restricted maximum likelihood (see Figure 5.14).

5.7 Summary

In this chapter, we carried out an experimental test on the evolutionary dynamics of new mutations in the presence of genetic variation and compared our results with theoretical predictions. To address this, we evolved genetically diverse populations of budding yeast (*S. cerevisiae*) consisting of $\sim 10^7$ diploid cells with unique haplotype combinations. We studied the asexual evolution of these populations under selective inhibition with antimicrobial drugs by time-resolved whole-genome sequencing and phenotyping. All populations underwent clonal expansions driven by *de novo* mutations, but remained genetically and phenotypically diverse. Despite the genetic diversity of the founder cells, we observed recurrent adaptive mutations. However, the founding fitness variance limited the scope for adaptive mutations to expand. The clones exhibited continued evolution by widespread genomic instability, rendering recessive *de novo* mutations homozygous and refining pre-existing variation. Finally, we decomposed the fitness contributions of pre-existing and *de novo* mutations by creating a large recombinant library of adaptive mutations in an ensemble of genetic backgrounds. Both pre-existing and *de novo* mutations substantially contributed to fitness. We find that the

relative fitness of pre-existing variants sets a selective threshold for new adaptive mutations as predicted by population genetic theory (discussed earlier in Chapter 2).

Our findings are relevant to understand the clonal evolution of populations with extensive pre-existing variation, such as microbial infections or cancer, which easily reach sizes of billions of cells. A quantitative link can be made to the asexual evolution of pathogens and somatic evolution in cancer: $\sim 31,000$ loci are expected to be heterozygous in any single cell of the recombinant cross. This is of the order of the typical mutation load in bacterial infections or cancer. The typical number of available escape mutations to antibiotics or chemotherapy drugs in both of these systems is limited, and it is comparable to the balance we observe between the number of drivers and passengers. Qualitatively, our results suggest that measuring driver mutation fitness with respect to the background distribution will be of key importance in our understanding of clonal evolution.

Chapter 6

Genome-wide biases in the mutational spectrum

6.1 Introduction

We have provided evidence in Chapter 4 of recurrent patterns of selection at the molecular level. Despite these regularities, we have shown that different founders typically follow different mutational paths in the genotype space. To establish whether these patterns are influenced by the activity of specific mutational processes, we aim for a systematic and robust characterisation of the spectrum of mutations under the rate-limiting conditions we imposed by chemical inhibition. In this chapter, we would therefore like to investigate the effects that fluctuations in the rate-limiting components of nucleotide synthesis have on mutagenesis. Before describing the results of our experiments, we briefly review potential constraints in fundamental biochemical networks that we perturb by chemical inhibition. We will characterise mutational processes by the type of mutations observed, their local sequence context, their timing and their distribution along the genome. We then represent each as a linear combination of mutational processes, using expectation-maximisation to reduce the dimensionality of the features. For this purpose, we only consider the presence or absence of a mutation and not the size of its clone or the mutation frequency. We are able to reconstruct signatures of an endogenous mutational process and a hydroxyurea-specific process, in line with its effect in reducing the deoxynucleoside triphosphate (dNTP) pool size which are the building blocks needed for DNA repair.[1]

This project has been carried out in collaboration with V. Mustonen (V.M.) at the Wellcome Trust Sanger Institute (Cambridge, UK).[2]

[1] Data analyses related to this chapter are available from the GitHub code repository [https://github.com/ivazquez/PhD-thesis/tree/master/Chapter6].

[2] I.V.-G. designed research, implemented computational methods based on previous work [226] and analysed data; I.V.-G. and V.M. interpreted results.

6.2 Mutational processes

The survival of all organisms requires faithful DNA replication to avoid deleterious mutations. New mutations in a cell's genome can arise due to endogenous processes, like stochastic errors in the DNA replication machinery, enzymatic changes to DNA or defective DNA repair [227, 228]. Similarly, exogenous DNA damage can be caused by mutagen exposure from sources like ionising radiation (e.g., X-rays or gamma rays) [229], ultraviolet radiation [230], tobacco smoke [231] or aristolochic acid [232, 233]. These major challenges to genomic integrity can lead to cellular transformation and carcinogenesis [234]. Endogenous and exogenous mechanisms of DNA damage have been studied in model organisms such as yeast [235–237] or worms [238, 239] and also in human cells [240]. In Chapter 4, we presented an experiment where we characterised the mutation dynamics of populations under exposure to antimicrobial drugs. Specifically, we focused on inhibitors of nucleotide synthesis for their highly specific roles as the limiting steps that determine the rate of DNA replication. In an environment with limited nucleotides there will be trade-offs between the rate of growth and the fidelity of DNA replication. Furthermore, it has been suggested that alterations in nucleotide pools or expression of editing deaminases cause cells to become unable to accurately replicate and repair their DNA [241, 242]. A similar scenario may arise with drugs that inhibit RNA transcription or protein translation, where the accuracy and kinetics will be affected by the availability of RNA polymerases or of different tRNAs.

In this chapter, we aim to characterise the spectrum of mutations under balanced and imbalanced levels of nucleotides. To begin with, we must understand the rate-limiting steps that impose constraints on genome replication. Firstly, the rate of replication is limited by nucleotide synthesis because the length of the cell cycle must be longer than the time taken by the cell to synthesise its new genome. DNA replication typically proceeds at 200 base pairs per second [243]. Simple prokaryotes like viruses have kilobase-long genomes that can be replicated serially. Bacteria, however, have megabase-long genomes, and they resolve the dilemma of replicating their genome in time for cell division by employing nested replication forks for parallel processing. Eukaryotes can also successfully decouple genome length and doubling time thanks to their usage of multiple DNA replication start sites. Yeast replicates its 12 Mb genome once every 90 minutes, needing approximately 500 origins of replication and over 10 replication forks coexisting in a single cell [244]. By inhibiting nucleotide synthesis as a selective constraint, we will test whether cells that have very long overall cell cycle times can maintain short DNA replication times, or instead if fitter cells that grow faster also shorten their replication time. A question that follows is therefore whether repeatable

substitution patterns can be observed in different selective environments, and if so whether they involve similar or different mutational biases.

6.2.1 Variation in mutation rate

In Chapter 4, we described an experiment where we propagated 5,760 populations for 93 days under selective inhibition with antimicrobial drugs. Populations were subject to constant and dynamic environments, at constant and increasing drug concentrations imposed by inhibition of nucleotide synthesis (with hydroxyurea – HU-C and HU-D) and cellular growth (with rapamycin – RM-C and RM-D). Control populations were propagated in parallel in a control environment (SC). We detected pre-existing and *de novo* SNVs and indels across 1,178 ancestral-evolved genome pairs sampled at 0 and 93 days (~930 generations). We aggregated the mutation counts $\{g_i\}_{j=1\ldots N_p}$ for this set of $N_p = 1,178$ populations at multiple loci in the genome. We aggregated the total number of mutations X^p acquired by a population after 93 days, which is given by the number of mismatches with respect to its ancestral reference, $X^p = s^n\left(\boldsymbol{g}^p, \boldsymbol{g}^{\text{ref}}\right)$. A total of $L = 52,466$ pre-existing base substitutions and $L = 11,601$ *de novo* mutations were observed. We will focus on the spectrum of *de novo* mutations which have been acquired under these selective pressures.

Now that we have an estimate of the number of mutations each population has acquired, we can try to understand how mutations occur using a Poisson process. The defining feature of a Poisson process is that each event (mutation arrival) is independent of all others. If you consider a single cell, then the DNA will acquire mutations at a rate μ which, for the moment, we assume to be uniform for all loci. Now each mutation event is independent of

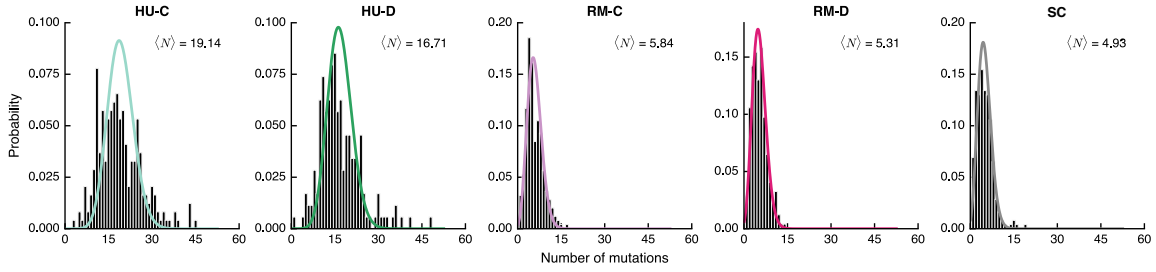


Fig. 6.1 Modelling mutation rate as a Poisson process. Histograms of the number of mutation counts per population X^p are experimentally measured. The coloured curves represent the Poisson distributions estimated by a method of maximum likelihood using the histogram data. Environments are ordered by their mean mutation rate, with the highest rate (left) induced by HU-C and the lowest (right) found in the control environment. The total number of mutations per population typically varies by 4-fold between lowest and highest, and by 5-fold within the two most mutagenic environments (HU-C and HU-D).

the previous event and in a very small interval of time the chance of two or more mutations is negligible. Hence, the number of mutations in an individual for a fixed time window (T) follows a Poisson distribution. The probability that X^p substitutions occur in population p over time T is

$$X^p \sim \text{Pois}\left(X^p \mid \mu T\right) = \frac{(\mu T)^{X^p} e^{-\mu T}}{X^p!}, \tag{6.1}$$

where μ is the mutation rate per site per unit time. Maximum-likelihood estimates of the mutation rate can then be obtained for the duration of the experiment ($T = 93$ days).

We first set out to investigate the influence of the rate-limiting environments on mutagenesis, as we expect these to play a major role in the variation of mutation rates. A large number of populations have a high rate of mutation, with almost one mutation occurring every cycle: there is an average of 19.14 events per population in HU-C, and 16.71 events per population in HU-D (Fig. 6.1). This vastly exceeds that of populations in RM-C (5.84), RM-D (5.31) and SC (4.93). The observations are consistent with the expected mean and variance of a Poisson process, which are both equal to μ. Since we know the number of sites at risk to be mutated in the whole genome (L) – which we approximate to be $L \simeq 1.23 \times 10^8$ bases in the original 'consensus' genome – then we can calculate a per-site mutation rate. The average per-site mutation rates are in the range $0.43 - 1.67 \times 10^{-10}$ from the least to the most mutagenic environment (SC and HU-C, respectively).

6.2.2 Spectrum of single-nucleotide variants

There are well-known systematic differences in the rates of transition and transversion mutations. To characterise the mutation spectrum of transitions and transversions, we count nucleotide substitutions strand-symmetrically by the pyrimidine of the mutated Watson-Crick base pair (i.e., `C`·`G` and `A`·`T`), of which there are 6 different types. In this way C-to-T (written `C>T`) and `T>C` transition mutations, and `C>A`, `C>G`, `T>A`, and `T>G` transversion mutations, encompass all possible SNVs. These are equivalent, respectively, to `G>A` and `A>G` transition mutations, and `G>T`, `T>A`, `A>T` and `A>C` transversion mutations. We refer to these as mononucleotide mutation channels, defined by one of six possible unique base pair changes ($N_c = 6$).

The observed data, X_c^p, is the number of mutations in channel c and population p of size $12 \times N_p$, since we split each of the $N_c = 6$ mononucleotide channels across coding and non-coding strands. Overall, the prevalence of different mutations is markedly uneven: `C>T` and `T>C` transitions are particularly prominent. Transitions are more than twice as frequent

as transversions, even though two possible transversions exist to every transition. This is commonly observed across the tree of life since transitions preserve the chemical structure of bases in DNA, whereas transversions do not [12].

6.2.3 Sequence context

We will now focus on the mutational spectra from different populations by considering the immediate sequence context of the mutated base. Context-dependent biases have been measured in multiple species, and particularly well known are the mutation hotspots caused by CpG methylation in vertebrates. Given our definition of mononucleotide mutation channels on the Watson-Crick base pair (C>A, C>G, C>T, T>A, T>C, T>G), each channel has four possible $5'$ and four $3'$ neighbouring nucleotides. A classification of trinucleotide mutation channels can therefore be defined by substitution class and sequence context immediately $5'$ and $3'$ to the mutated base ($N_c = 6 \times 4 \times 4 = 96$). In this notation we use the linear sequence of bases along the $5'$-to-$3'$ direction, e.g., a C>T mutation flanked by a $5'$ guanine and a $3'$ thymine occurs at GpCpT.

Each mutation was mapped to one of $N_c = 96$ trinucleotide mutation channels, defined by one of six possible unique base pair changes and one of the sixteen different trinucleotide sequence contexts. The mutation counts form a matrix X_c^p of size $96 \times N_p$ (Fig. 6.3B). We can also estimate the frequency of sites at risk in the genome – the mutation opportunity – which incorporates the genome frequencies of the triplets where each mutation type can occur (Fig. 6.3A).

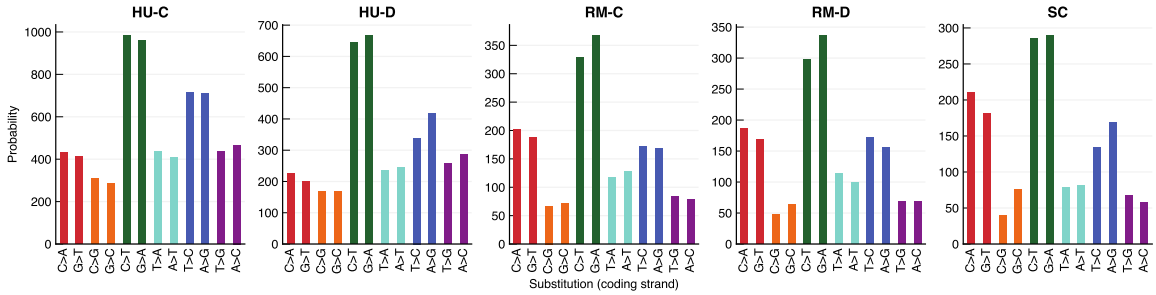


Fig. 6.2 Breakdown of base substitutions by nucleotide change. Each colour indicates the number of mutations for one of the six classes of nucleotide substitutions. The total counts for each mononucleotide mutation channel ($N_c = 6$) are split between those in the coding (untranscribed) and the non-coding (transcribed) strand.

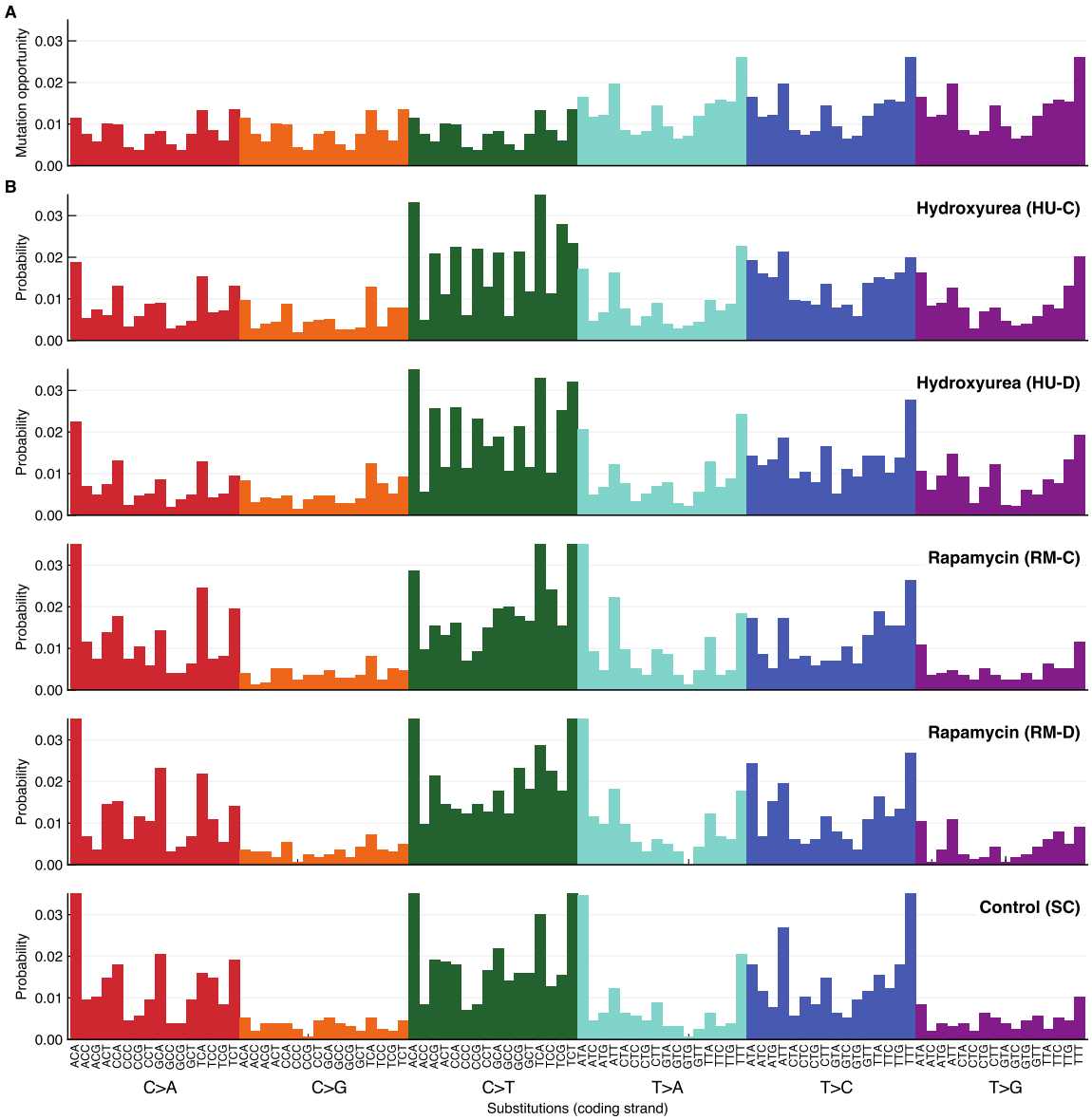


Fig. 6.3 Sequence context of single-nucleotide substitutions across trinucleotide mutation channels ($N_c = 96$). **(A)** Relative usage of each site at risk to be mutated. The bar height indicates the frequency of trinucleotides in the genome where each substitution type can occur, which we refer to as mutation opportunities. **(B)** Each bar shows the total number of mutations observed across replicate populations for a class of single-nucleotide substitutions in a given trinucleotide context. Each of the environments is shown from top to bottom (HU-C, HU-D, RM-C, RM-D, SC).

6.3 Inference of mutational processes

The spectrum of mutations we have observed is highly variable and very rich in environment-specific, founder-specific and context-dependent features. To this extent, we would like to reconstruct features from sequence data that may resolve which fundamental mutational processes are behind them. Firstly, we would like to infer context-specific mutational features (or mutational 'signatures'), which we define by μ_{sc} ($s = 1, \ldots, N_s$, $c = 1, \ldots, N_c$), with $\sum_{c=1}^{N_c} \mu_{sc} = 1$ [226]. Secondly, we would like to estimate the number of mutations in each sample that are associated with a mutational signature, also known as the 'activity' or 'exposure'. We define the activity of process s in population p as x_s^p.

Earlier we described the combined observations of channel- and population-specific mutation counts by the matrix X_c^p ($c = 1, \ldots, N_c$, $p = 1, \ldots, N_p$). Fischer et al. [226] recently proposed a probabilistic approach to infer mutational signatures that can account for the noise in mutation counts, partly due to their stochastic origin. Assuming the mutations to arise independently, they describe the probability of observing the vector X_c^p of mutation counts in population p across channels using a Poisson model

$$P\left(X^p \mid x^p, \omega^p, \mu\right) = \prod_{c=1}^{N_c} \text{Pois}\left(X_c^p \,\middle|\, \sum_{s=1}^{N_s} x_s^p \mu_{sc} \omega_c^p\right) \tag{6.2}$$

where x_s^p is the activity of a signature s in population p, μ_{sc} is the probability of signature s to generate a mutation in channel c, ω_c^p is the mutation opportunity in channel c, and N_s is the number of mutational signatures.

We used the expectation-maximisation (EM) algorithm implemented by Fischer et al. [226][1] to identify specific mutational signatures. EM decomposes the matrix of mutation counts into a set of sparse factors putatively representing different mutational processes, and into population-specific activities for each process. The mutational signatures are calculated from the relative frequency of mutations at the 96 triplets defined by the mutated base and each flanking base on either side. The hidden states are the activities of signature s in population p, x_s^p.

The expectation-maximisation algorithm can be stated in two steps. We begin with an initial estimate for μ that respects the normalisation of mutation counts, $\sum_{c=1}^{N_c} \mu_{sc} = 1$. In the E-step, given the observed data X^p and the current best estimate μ we find all the hidden

[1] Expectation-maximisation implemented in the EMu software [226].

data x. In the M-step, we update all N_s signatures $\mu^{(k)} \to \mu^{(k+1)}$.

$$\text{E-step:} \qquad \hat{x}^p = \underset{x}{\text{argmax}} \log P\left(X^p \,\middle|\, x, \omega^p, \mu^{(k)}\right) \tag{6.3}$$

$$\text{M-step:} \quad \mu^{(k+1)} = \underset{\mu}{\text{argmax}} \sum_{p=1}^{N_p} \log P\left(X^p \mid \hat{x}^p, \omega^p, \mu\right) \tag{6.4}$$

After convergence to an estimate of the mutational spectra $\hat{\mu}$, we can then evaluate the log-likelihood, integrating out the hidden data using a saddle-point approximation [226]. To select the correct number of mutational processes given the data, we note that increasing the number of signatures N_s typically increases the data likelihood $P(X \mid \hat{\mu})$ and provides a better explanation of the data. Fischer et al. [226] propose using an information criterion like BIC to penalise for model complexity,

$$\text{BIC} = 2 \log P(X \mid \mu) - N_s(N_c - 1) \log N_p. \tag{6.5}$$

where N_s is the number of mutational signatures, N_c is the number of channels and N_p is the number of populations analysed.

The results of the EM algorithm – model parameters (spectra μ) and hidden data (activities x) – allow to probabilistically assign mutations to processes. A global inference can be first carried out by aggregating counts for populations across all environments to estimate the global activities:

$$\tilde{X}^p_{sc} = \frac{N_s \hat{\mu}_{sc} \omega^p_c}{\sum_{i=1}^{N_s} (\hat{\mu}\omega^p)_i}. \tag{6.6}$$

Here, the choice of initial entries of the μ_{sc} matrix is set to the observation means. The choice of initial guess for μ_{sc} must respect the normalisation of the channels. The initial weights of the global activities $x^{p,g}_s$ of signature s in population p are uniform.

We can then use global activities $\{\hat{x}^{p,g}_s\}$ as an informative prior for the local inference of the activity of each signature in each population:

$$\tilde{X}^{p,l}_{sc} = \frac{N_s \hat{x}^{p,g}_s \hat{\mu}_{sc} \omega^{p,l}_c}{\sum_{b=1}^{N_s} \hat{x}^{p,g}_s (\hat{\mu}\omega^{p,l})_b} \tag{6.7}$$

giving us the locally inferred activities, $\{\hat{x}_s^{p,l}\}$. The stability of our solutions was confirmed by comparing the assignment of mutational signatures from the local inference to the global inference across all environments, per environment or per founder.

The inference identified two uncorrelated mutational signatures (Fig. 6.4). Signature 1 corresponds to an endogenous process or processes that are active in all populations. This signature is characterised by C>T substitutions at ApCpA and TpCpT trinucleotides, as well as C>A and T>A substitutions at ApTpA (Fig. 6.4, top panel). Signature 2 is a hydroxyurea-specific signature. This mutational process is characterised by C>T mutations at ApCpN and GpCpN, and T>A and T>G mutations at TpTpT (Fig. 6.4, bottom panel). Signatures in which mutations have different target bases, e.g., C>T and T>G in Signature 2, may inevitably try to split into multiple signatures because different nucleotide pools are involved for the repair of each base and relative dNTP abundance will also fluctuate in time. Additional signatures are not supported by the data, however, as the number of counts per trinucleotide channel is finite.

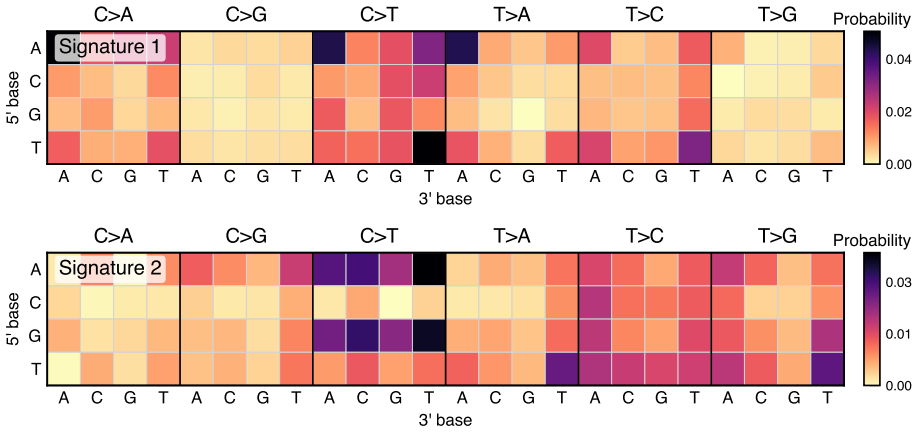


Fig. 6.4 The context-dependent mutation spectrum supports two signatures. Each of the tiles shows one of the 6 base substitutions, with 16 trinucleotide channels sorted by the flanking 3′ base (x-axis) and 5′ base (y-axis). The colour scale shows the probability $\hat{\mu}$. Signature 1 is characterised by C>T mutations at ApCpA and TpCpT, C>A at ApCpA and T>A at ApTpA. Signature 2 is characterised by C>T mutations at ApCpN and GpCpN, and T>A and T>G mutations at TpTpT.

Signature 1 has an average baseline activity of 8 mutations per population (Fig. 6.5A). Signature 2 contributes 16 mutations on average in HU-C and 12 mutations in HU-D. There is an inverse relationship between the dose of hydroxyurea and the number of base substitutions. There were minor differences in the total mutation load depending on the founder genotype, largely related to uracil or lysine deficiencies of the segregants (Fig. 6.5B). The activity of each mutational process, $\hat{x}_s^p$, indicates that the relative contributions of each of

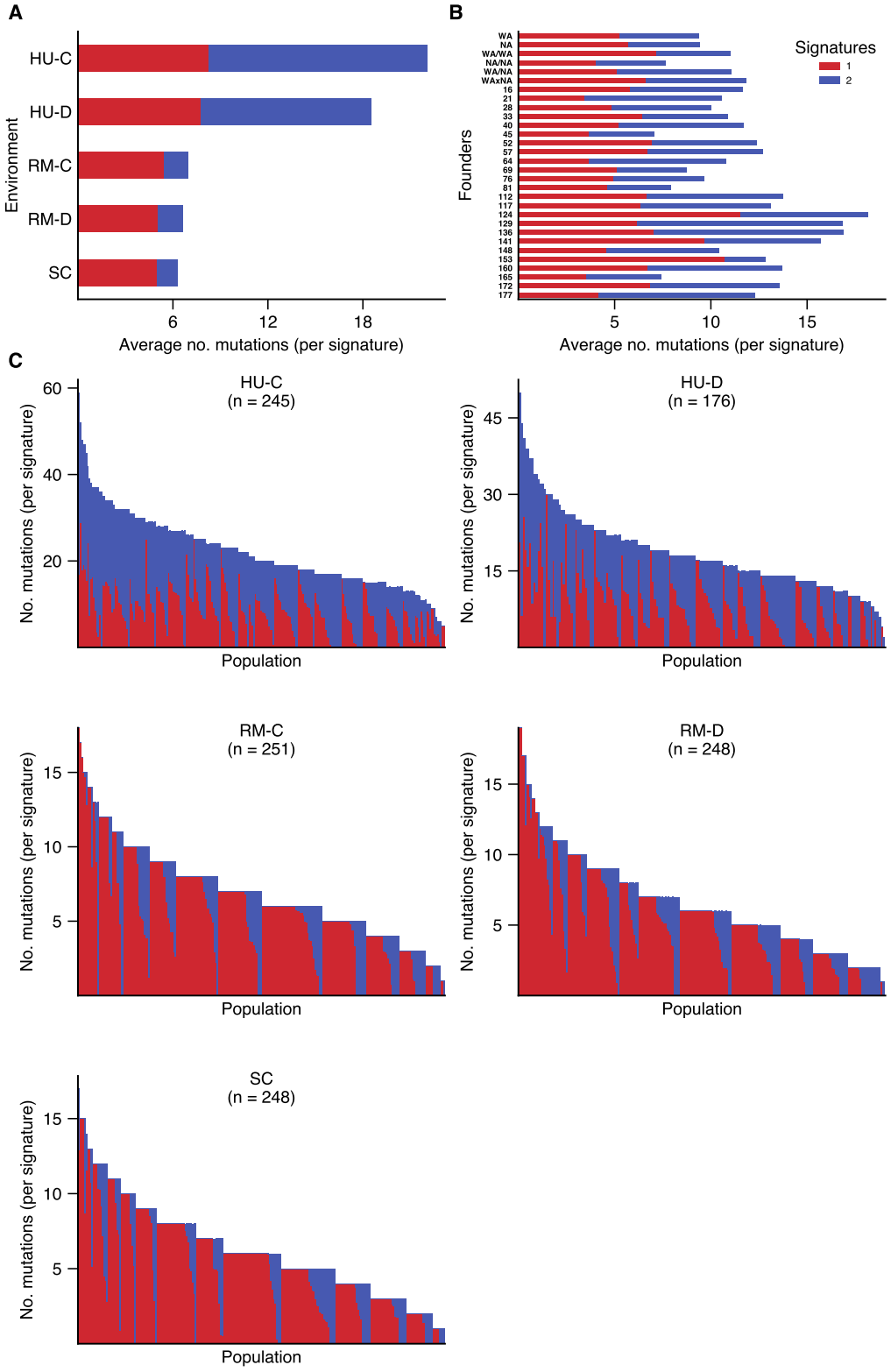


Fig. 6.5 Activity of mutational signatures. **(A)** The average activity of mutational signatures by environment, $\frac{1}{N_e}\sum_{p\in E}\hat{x}_s^p$, shown by the bar length along the horizontal axis. **(B)** The average activity of each mutational signature calculated by founder, $\frac{1}{N_f}\sum_{p\in F}\hat{x}_s^p$. Founder genotypes are ordered from top to bottom, indicated by the labels on the left margin and coloured by the type of background. **(C)** The activity $\hat{x}_s^p$ of each mutational process shows their contribution in each population. The absolute contribution of different mutation signatures is depicted by stacked bars and ordered by the total number of mutations.

the signatures across replicates is similar when more than one signature is active. Even though Signature 2 is hydroxyurea-specific, several mutations can be attributed to this process in other environments which is an artifact caused by finite-size effects of the total number of counts per channel when carrying out the optimisation without sparsity constraints (Fig. 6.5C).

6.4 Mechanistic models of mutagenesis

We would like to determine whether we can associate certain mutational processes with the failure of endogenous repair mechanisms or the activity of exogenous mutagens. Two different mutational processes could have the same mutational signature, which means we should be cautious in ascribing a single mutational process to a mutational signature. Signature 1 reflects an endogenous process or processes of DNA damage that is ubiquitous across all replicate populations. This may be attributed to deamination, which occurs spontaneously in all DNA bases that contain primary amines albeit at markedly different rates (Fig. 6.6A). The signature profile is consistent with two common deamination reactions: (i) `C>T` transitions arising through direct replication over uracils generated by cytidine deamination; and (ii) `C>A` transversions due to replication over abasic sites formed after uracil excision by uracil-DNA glycosylase (Fig. 6.6A). Similar processes have been observed in model organisms (e.g., yeast [245] or worms [238]) and in humans [228, 246]. However, we note that the baseline rate of `C>T` transitions at `CpG` dinucleotides – the most common form of deamination in vertebrates – is low, consistent with yeast lacking DNA methylation (Cristina Rada, personal communication).

The mutations described by Signature 2 are due to an exogenous process involving nucleotide depletion by hydroxyurea. Free nucleotides are normally collected by DNA polymerase and attached by complementary base pairing, moving in the $5'$-to-$3'$ direction (polymerisation). Polymerases also carry out $3'$-to-$5'$ error correction and are capable of excising misincorporated bases (proofreading). Depleting the cell of dNTPs inhibits nucleotide polymerisation and can result in stalled replication forks. We now work through the rationale that this mutational signature occurs when holding open the excision-repair induced gaps by inhibiting nucleotide polymerisation, which increases competition between the polymerisation and proofreading activities of replication complexes. Interestingly, models of polymerisation and excision which assume that both these processes take place in *equilibrium* through a series of reversible steps subsume that, as dNTP abundance becomes rate-limiting, more rapid synthesis should result in decreased fidelity which fails to explain our observations.

Instead, *out-of-equilibrium* models account for the fact that nucleotide incorporation is irreversible [247, 248] and predict that:

(i) During polymerisation, if a 'wrong' nucleotide is present at sufficiently high concentrations to compete with the 'right' nucleotide the misincorporation rate at a non-Watson-Crick base pair can increase.

(ii) During proofreading, a high concentration of the next nucleotide promotes its incorporation, thereby favouring polymerisation and blocking the 3′-exonucleolytic proofreading of a previously misincorporated nucleotide (known as the 'next-nucleotide' effect).

To distinguish between these two mechanistic models (Fig. 6.6B), we must take toll of our observations thus far. Firstly, we already described that mutation rates are higher with lower doses of hydroxyurea, which is only plausible by the competition of correct and incorrect nucleotides at high abundance of dNTPs. Secondly, the shift in mutation spectrum

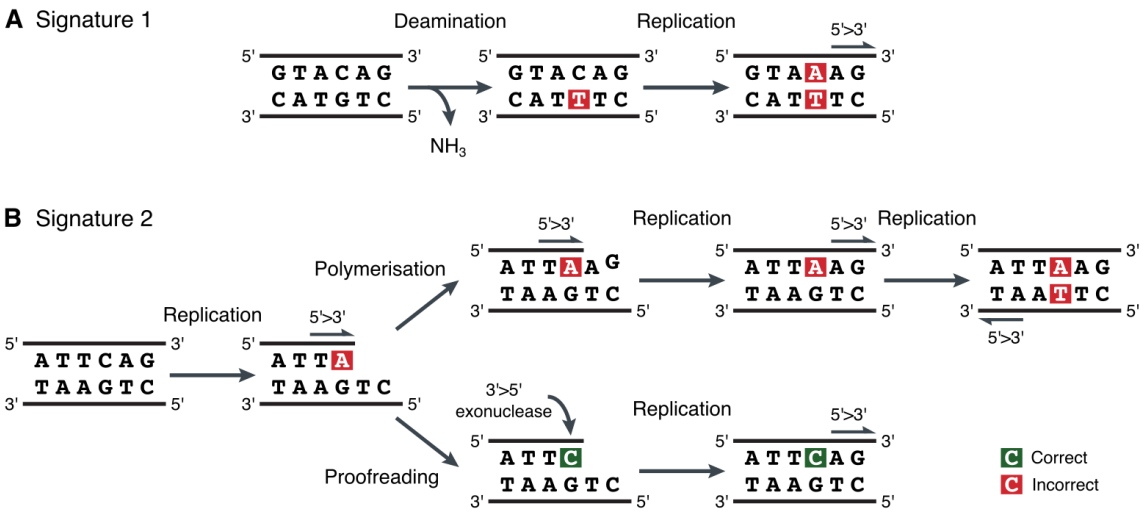


Fig. 6.6 Mechanistic models of mutagenesis. **(A)** Signature 1 is compounded of a number of endogenous DNA damage processes caused by spontaneous deamination. This signature is common across taxa and gives rise to characteristic C>T and C>A mutations (see Figure 6.4). **(B)** Signature 2 occurs due to spontaneous mutagenesis in conditions of nucleotide depletion. During DNA replication, replicative DNA polymerases in budding yeast carry out 5′-to-3′ polymerisation and 3′-to-5′ proofreading. The competition between the polymerisation and proofreading activities of polymerase can lead to the misincorporation of nucleotides and impair the ability to proofread flanking nucleotides – known as the 'next-nucleotide' effect –. This is clearly evidenced by an elevated mutation load under nucleotide depletion in HU-C and HU-D, as well as an excess of C>T mutations at ApCpN and GpCpN trinucleotide contexts (see Figures 6.4 and 6.5).

displayed by Signature 2 is influenced by the flanking nucleotides – those 3′ to the misincorporated base being present in excess – suggesting the existence of a proofreading function as part of the replication complex, which has been previously described *in vitro* [249] and we observe *in vivo*. The most plausible mechanistic model for this mutational signature is therefore determined by the competition between the polymerisation and proofreading activities of replication complexes. This trade-off is dominant over the ultimate effect of nucleotide depletion: fork stalling leading to the formation of double-strand breaks, in which both strands in the double helix are severed. Double-strand breaks can cause genomic rearrangements that are particularly hazardous to the cell and are most likely selected against. Furthermore, we already characterised single-cell clones in Chapter 5 where we observed genomic rearrangements that enable the selection of beneficial mutant alleles by loss of heterozygosity. We do not discard that fork stalling may be more common in HU-D than HU-C due to the dose-dependent duration of cell-cycle arrest, but the mechanistic evidence for competition between polymerisation and proofreading should remain invariant.

6.5 Summary

The rate of genetic mutations, mainly made by DNA polymerase, is typically much lower than the error rates of RNA polymerase or the ribosome. In this chapter, we tackled the challenge of systematically detecting and quantifying errors in DNA replication under nucleotide pool imbalances using budding yeast. The error rate we observe by DNA sequencing is around 10^{-10} per base on average, but it varies considerably between nucleotides and in a manner that depends on the neighbouring bases. Treating yeast with drugs that alter the nucleotide balance impaired DNA proofreading, exposing specific patterns of errors induced in each nucleotide position. The endogenous substitution patterns we found in budding yeast show significant similarity to those of deamination observed in other organisms, while the exogenous patterns caused by nucleotide depletion suggest a common chemical basis for errors under genome replication stress mediated by the ‘next-nucleotide’ effect. Finally, we proposed a kinetic proofreading model based on basic thermodynamics that may be able to explain some of the error patterns observed. Future studies should aim to characterise whether substitutions under impaired replication fidelity tend to occur in positions that are less evolutionarily conserved, which may indicate that organisms have evolved mechanisms to minimise deleterious substitutions with detrimental phenotypic effects.

Chapter 7

Epilogue

In this thesis, we investigated theoretical and experimental approaches to understanding the evolutionary dynamics of rapidly adapting populations. We discussed two approaches to understanding biological systems: one of them described populations as dynamical stochastic systems that process and transmit information, whilst a data-driven approach could build a description of the genetic diversity in a population using DNA sequencing. We argued briefly that the lessons learned from simple model systems such as budding yeast (*S. cerevisiae*) can help us understand the role of the evolutionary forces of mutation, selection and genetic drift. By applying these methods to the measurement of the adaptive responses in cell populations under rate-limiting conditions, we have learned a number of new insights regarding the nature of selection in populations with extensive genetic variation that we summarise below. Finally, we present open questions and possibilities for future consideration.

Models of evolutionary dynamics distinguish driver and passenger mutations

In Chapter 2, we first introduced a minimal model with mutation, selection and genetic drift. We started with a reduced single-locus model that gave us a heuristic understanding of the probability of fixation or extinction of a neutral mutation or a mutation under selection. The complex dynamics of the system that emerge from multiple mutations in the genome quickly become intractable due to the microscopic details of multi-locus statistics. However, we could implement a multi-locus model for genetic hitchhiking using a deterministic approximation for the motion of a driver mutation during a selective sweep. This model can successfully discern mutations directly under selection from hitchhiker mutations. We compared this to experimental data for a selective sweep described in Chapter 5 and we can localise a validated driver mutation under selection, as predicted by our theory.

Rapidly evolving populations represent a challenging modelling problem, with complex and noisy dynamics. Despite the success of the driver-and-passengers model in explaining empirical data, the assumption that mutations have identical effects everywhere in the genome should be reconsidered. The model currently predicts fixation events, but it does not address that many adaptive mutations will initially only sweep to intermediate frequencies. Incomplete sweeps are common in our experiments and have been observed in many populations [250]. Accommodating incomplete sweeps will require extending the driver-and-passengers model beyond additive effects. Finally, we reported elsewhere that for the stochastic single-locus model, one can extend this analysis to include a control that can accelerate or delay the loss of a mutation in a population [114]. We derived optimal adaptive controls in this scenario that could potentially be experimentally tested for model-based control of population dynamics. Demonstrating that the evolution of biological systems can be predictably controlled would be a significant milestone with far-reaching implications, both for population genetic theory and for the development of adaptive therapies that halt the evolution of antimicrobial and chemotherapy resistance.

Subclonal population structure can be inferred by whole-genome sequencing

In Chapter 3, we addressed a data-driven problem to enable the computational reconstruction of the clonal composition of a population using DNA sequencing. We first reviewed recent efforts to quantify the genetic diversity of mixed populations of cells by means of bulk sequencing, single-cell sequencing and lineage tracing techniques. Before addressing the problem of inferring subclonality from DNA sequences, we provided an overview of the current state-of-the-art of algorithms addressing this problem, and we gave a brief introduction to probabilistic modelling and Hidden Markov Models (HMM). As a first step, we presented a general-purpose probabilistic filtering algorithm for one-dimensional discrete data, similar in spirit to a Kalman filter. It is a continuous state-space HMM with Poisson or binomial emissions and a jump-diffusion propagator. It can be used for scale-free smoothing, fuzzy data segmentation and filtering of DNA sequencing datasets. Secondly, we presented a HMM with a discrete state-space to reconstruct the subclonal structure of a population, and we showed we can exploit correlations in the data to discern macroscopic subclones in a mixed population. Our statistical algorithms allow us to accurately estimate the number of subclonal populations, their fractions in the sample, and the subclone-specific total copy number profiles, B-allele status and SNV genotypes with high resolution. Two key ingredients underlie our approach. Firstly, we exploited multiple layers of genetic variation to reconstruct the clonal composition: single-nucleotide variants (both pre-existing and

de novo) and copy-number aberrations. This integrative approach overcomes the degeneracy of subclonal solutions, whereby the same clonal composition may otherwise be generated by multiples of the hidden genomic state at different scaled population fractions. Secondly, the algorithm can exploit the information found in spatially- or temporally-resolved samples, taking into account correlations along the genome caused by events such as copy-number changes. Finally, we evaluated the algorithmic performance on simulated datasets and synthetically-derived samples from human cancer. Using these datasets, we have shown through quantitative comparisons of methods that computational inference of both subclonal structure and correct assignment of mutations to subclones are highly dependent on incorporating the copy number state of the allele being measured. Our implementation for filtering and inference is scalable and can be applied to clonal admixtures of genomes in any asexual population, from evolving pathogens to the somatic evolution of tumours.

Several new lines of research are ripe to directly benefit from these methods: the algorithm is ideally suited to infer the clonal composition of a population when selection is sufficiently strong to amplify fit genotypes. One major line of work being currently pursued is characterising clonal and subclonal driver genes under selection in pan-cancer datasets based on the subclonal inference [122]. This is revealing that subclonal drivers are as frequently found to be under selection as clonal ones [33, 110]. As was shown in the exposition of the algorithm, the uncertainty in assigning subclonal states reflected the inherent limits of inference with the resolution of current technologies. This is a data-derived issue, as the current state-of-the-art in DNA sequencing has limitations to resolve the long tail of the clone size distribution, which is key to understand the role of selection in evolutionary dynamics. Clearly, future models for sequencing data with resolution at the single-cell level should be able to achieve much greater reconstruction depth by incorporating the lineage relations directly in the inference. These methods may be complemented by lineage tracing techniques that can be used for molecular recording at temporal resolutions that can trace every cell division [154–156, 251].

Path-dependent effects of selection on the genetic diversity of a population

In Chapter 4, we used a recombinant library design of randomised genomes of budding yeast to study the nature of contingency effects and selection on genetic variation. We carried out directed evolution and DNA sequencing of this library to simultaneously adapt genetically heterogeneous and homogeneous populations to an environment and measure the genotypes of an ensemble of populations. We imposed selection by inhibition of different rate-limiting steps of the cell cycle, like replication or translation. We wanted to determine whether a

'cloud' of recombinant genotypes accelerates evolution compared to isolate segregants. The average length of mutational paths required by genetically diverse populations was very predictable. Conversely, those same genotypes in isolation could sometimes become entrenched in a local fitness maximum and require many more mutations to adapt, which sets limits to the predictability of outcomes at the sequence level. We have provided evidence of recurrent patterns of mutations at the molecular level, but the outcomes at the genetic level are stochastic. In particular, we have shown that different founders follow different mutational paths in the genotype space [183]. The pool of adaptive mutations is common to most closely related genotypes, but some rare mutations can be seen to dramatically change the trajectory of a few populations.

Overall, we have gained insights into the role of chance and selection and how they influence the sequence evolution of an ensemble of population trajectories. Regarding the patterns of recurrence at the molecular level, we could extend our scoring method to estimate the fitness effects of the mutations by forming an evolutionary model with an explicit fitness function. This can be done by projecting sequences to scores as we did, and then drawing uncorrelated nucleotides to obtain a null distribution $P_0(S)$ to define a function $Q(S) \propto P_0(S)e^{2NF(S)}$ and estimate fitness differences from the data (see e.g., Moses and Durbin [101] and Fischer et al. [102]). Of particular interest will be conditional mutants, whose fitness effect may only be expressed in extreme stress environments. Using arguments from mutation-selection balance, we may be able to localise these by estimating the fitness cost of these mutations in different environments (see e.g., Zanini et al. [109]). Finally, an outstanding challenge at the population level is to test the value of genetic diversity as a statistical indicator for population robustness. Whilst we have found that a set of random genotypes generated by recombination can efficiently explore the genotype space, it remains to be seen whether genetic diversity can ensure survival and prevent population collapse during clonal evolution. In addition to the genome sequence measurements, a complete record of population size changes has been kept by real-time monitoring throughout the experiment using high-throughput scanning. These real-time measurements could be used to evaluate genetic diversity as an early-warning indicator for the propensity of a population to thrive or collapse.

Genetic variation sets a selective threshold on the fate of new mutations

In Chapter 5, we studied the evolutionary dynamics of populations containing extensive fitness variability in yeast. As we discussed in Chapter 2, genetically diverse populations are expected to form a travelling fitness wave, with the mean fitness increasing at a rate that is

proportional to its fitness variance [125, 252]. We therefore set out to test a key theoretical result that predicts the existence of a threshold selective advantage above which the fate of a new mutation becomes decoupled from the background it lands on [219, 220]. Our results show that large populations can readily find beneficial *de novo* mutations, but their adaptive trajectories are simultaneously shaped by pre-existing and *de novo* variation with overlapping timescales. We observed a balance between the loss of diversity due to selection, and active diversification mechanisms that partially re-established and refined existing variants. The background genotypes were continuously re-configured by genomic instability, diversifying the expanding subpopulations. Measurements of the fitness distribution revealed two different outcomes of selection: if many mutations had comparable fitness effects as in hydroxyurea, the fitness distribution remained smooth; on the contrary when few large-effect mutations were available, such as mutations in the TOR pathway in rapamycin, the fitness distribution became multimodal. We also carried out ensemble-averaged fitness measurements of a recombinant library of pre-existing and *de novo* mutations. We found that large-effect mutations, such as those in the TOR pathway, confer resistance to rapamycin regardless of the genetic background where they arise. These mutations were of sufficient magnitude to surpass the bulk of the fitness distribution and can be interpreted to be above the selective threshold. Conversely, the pre-existing fitness variance influenced the fate of *de novo* drivers like *RNR2* and *RNR4* mutations, which needed to land on a favourable background to be competitive. These results confirm the theoretical prediction that new mutations are expected to be successful only if they land on a favourable background or if they are beneficial enough to escape from the bulk dynamics by their own merits.

Taken together, our findings can help to understand the asexual or somatic evolution of large populations with extensive genetic variation. Bacterial infections and cancer, which easily reach sizes of billions of cells, host a comparable mutation load before any selective treatment is applied. For example, the number of pre-existing variants in this model system is comparable to the typical number of somatic mutations accrued before treatment during carcinogenesis, which varies between $10^2 - 10^5$ depending on the cancer type [253]; and it is also comparable to the genetic diversity in bacterial communities (e.g. in cystic fibrosis patients [254]). The existence of a selective threshold demonstrated by our combinatorial strategy of background averaging shows that populations can be found at the limit cases above and below the selective threshold, suggesting that that these dynamics may represent a general mode of adaptation. Indeed, several biological systems have already been characterised to operate at the edge of the two regimes: large-effect mutations being amplified on well-adapted background genotypes have been observed in laboratory populations [105] and

in the wild, e.g. in the seasonal influenza virus [28]. Critically, it follows from these observations that predicting the outcome of selection will hinge on characterising the background fitness variance and on finding a common framework to describe the selective potential of a population [255]. These may be necessary requisites to eventually rationalise the design of therapies in the treatment of bacterial and viral infections or cancer. Overall, we hope these results will encourage new theoretical and empirical investigations of the complex interplay of selection simultaneously acting on pre-existing and *de novo* genetic variation, and of the role of genomic instability continuously moulding the genomes in a population.

Genome-wide signatures of endogenous and exogenous mutational processes

In Chapter 6, we aimed to characterise the spectrum of mutations under selective constraints which affect the fidelity of genome replication. Firstly, we characterised two mutational processes observed in the experiments in Chapter 4: an endogenous process of spontaneous deamination active across all populations, and an environment-specific mutational process caused by the competition of the polymerisation and proofreading of DNA under nucleotide depletion. Supporting this exogenous process, we observed genomes to be more accurately replicated at low substrate concentration of dNTPs, becoming *less* efficient at *higher* absolute concentrations. The mutational spectrum also revealed a consistent pattern of the so-called 'next-nucleotide' effect: that is, the probability of excision of the last residue by a polymerase depended not only on the base sitting opposite but also on the rate of incorporation of the next nucleotide. These two observations can only be parsimoniously reconciled by the competition between the polymerisation and proofreading activities of DNA replication complexes [247, 248]. The 'next-nucleotide' effect was first observed *in vitro* [249, 256] and, to our knowledge, this is the first observation *in vivo*.

Regarding the mutational processes that have been characterised, a genomic readout provides future opportunities for the detection of mutagenic chemicals with much higher sensitivity than current assays, such as the Ames test [257]. We can now also measure the cooperative kinetics of DNA polymerisation and proofreading by sequencing. This is a particularly interesting avenue, as nucleotide pool imbalances are known to play a role in oncogenic activation in humans when the Rb-E2F pathway is activated, which can be caused by the cellular oncogene cyclin E or HPV viral oncogenes [241]. This suggests that the isolation of mutagenic signatures in yeast could potentially be used to detect the activity of mutational processes of different origin but similar molecular basis.

In summary, we used theoretical, computational and experimental methods to study the evolutionary dynamics of rapid adaptation, combining statistical analyses of genomic sequences, mathematical models of evolutionary dynamics and experiments in molecular evolution. We covered a wide range of different topics, contributing to the general understanding of structure of genotypic and phenotypic variability in population dynamics that can be translated to a range of biological systems, from microbes, viruses or parasites, to cellular populations, such as tumours.

References

[1] L. Pasteur. *Mémoire sur les corpuscules organisés qui existent dans l'atmosphère: examen de la doctrine des générations spontanées*. 1862.

[2] G. Mendel. Versuche über Pflanzenhybriden, *Verhandlungen des naturforschenden Vereins Brünn* **4** (1866), pp. 3–47.

[3] T. H. Morgan et al. *The mechanism of Mendelian heredity*. Henry Holt and Company, 1915.

[4] N. W. Timofeeff-Ressovky, K. G. Zimmer, and M. Delbrück. "Über die Natur der Genmutation und der Genstruktur", *Nachrichten Göttingen*. Weidmannsche Buchhandlung, 1935.

[5] E. Schrödinger. *What is life?* Cambridge University Press, 1944, pp. 90–165.

[6] O. T. Avery, C. M. MacLeod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types, *J. Exp. Med.* **79**, no. 6 (1944), pp. 137–158.

[7] A. D. Hershey and M. Chase. Independent functions of viral protein and nucleic acid in growth of bacteriophage, *J. Gen. Physiol.* (1952), pp. 39–56.

[8] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids, *Nature* **171**, no. 4356 (1953), pp. 737–738.

[9] M. W. Nirenberg and J. H. Matthaei. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides, *Proc. Natl. Acad. Sci. U.S.A.* **47**, no. 10 (1961), pp. 1588–1602.

[10] F. H. Crick. On protein synthesis, *Symp. Soc. Exp. Biol.* **12** (1958), pp. 138–163.

[11] F. H. Crick et al. The general nature of the genetic code for proteins, *Nature* **192** (1961), pp. 1227–1232.

[12] H. Lodish et al. *Molecular cell biology*. 6th ed. W. H. Freeman and Company, 2008.

[13] L. H. Hartwell et al. From molecular to modular cell biology, *Nature* **402**, no. 6761 (1999), pp. 47–52.

[14] R. Phillips et al. *Physical biology of the cell*. 2nd ed. Garland Science, 2013.

[15] W. Bialek. *Biophysics: Searching for principles*. 1st ed. Princeton University Press, 2012.

[16] J. W. Drake et al. Rates of spontaneous mutation, *Genetics* **148**, no. 4 (1998), pp. 1667–1686.

[17] T. A. Kunkel and K. Bebenek. DNA replication fidelity, *Annu. Rev. Biochem.* **69** (2000), pp. 497–529.

[18] J.-F. Gout et al. Large-scale detection of in vivo transcription errors, *Proc. Natl. Acad. Sci. U.S.A.* **110**, no. 46 (2013), pp. 18584–18589.

[19] M. Imashimizu et al. Direct assessment of transcription fidelity by high-resolution RNA sequencing, *Nucleic Acids Res.* **41**, no. 19 (2013), pp. 9090–9104.

[20] H. S. Zaher and R. Green. Fidelity at the molecular level: Lessons from protein synthesis, *Cell* **136**, no. 4 (2009), pp. 746–762.

[21] M. A. Jobling, M. E. Hurles, and C. Tyler-Smith. *Human evolutionary genetics: Origins, peoples and disease*. 1st ed. Garland Science, 2004.

[22] S. Myers et al. A fine-scale map of recombination rates and hotspots across the human genome, *Science* **310**, no. 5746 (2005), pp. 321–324.

[23] J. Wang et al. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm, *Cell* **150**, no. 2 (2012), pp. 402–412.

[24] E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins, *Evol. Genes Proteins* (1965), pp. 97–166.

[25] C. Darwin. *On the origin of species*. John Murray, 1859.

[26] S. Leibler and E. Kussell. Individual histories and selection in heterogeneous populations, *Proc. Natl. Acad. Sci. U.S.A.* **107**, no. 29 (2010), pp. 13183–13188.

[27] J. B. Plotkin and G. Kudla. Synonymous but not the same: The causes and consequences of codon bias, *Nat. Rev. Genet.* **12**, no. 1 (2011), pp. 32–42.

[28] M. Łuksza and M. Lässig. A predictive fitness model for influenza, *Nature* **507**, no. 7490 (2014), pp. 57–61.

[29] F. Zanini et al. Population genomics of intrapatient HIV-1 evolution, *eLife* **4** (2015), pp. 1–26.

[30] P. C. Nowell. The clonal evolution of tumor cell populations, *Science* **194**, no. 4260 (1976), pp. 23–28.

[31] D. Hanahan and R. A. Weinberg. The hallmarks of cancer, *Cell* **100**, no. 1 (2000), pp. 57–70.

[32] C. A. Klein. Selection and adaptation during metastatic cancer progression, *Nature* **501**, no. 7467 (2013), pp. 365–372.

[33] M. Gerstung et al. The evolutionary history of 2,658 cancers, *bioRxiv* **161562** (2017).

[34] M. Greaves and C. C. Maley. Clonal evolution in cancer, *Nature* **481**, no. 7381 (2012), pp. 306–313.

[35] M. Baym et al. Spatiotemporal microbial evolution on antibiotic landscapes, *Science* **353**, no. 6304 (2016), pp. 1147–1151.

[36] U. Laserson et al. High-resolution antibody dynamics of vaccine-induced immune responses, *Proc. Natl. Acad. Sci. U.S.A.* **111**, no. 13 (2014), pp. 4928–4933.

[37] B. Waclaw et al. Spatial model predicts dispersal and cell turnover cause reduced intra-tumor heterogeneity, *Nature* **525**, no. 7568 (2015), pp. 261–267.

[38] M. Gerlinger et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing, *N. Engl. J. Med.* **366**, no. 10 (2012), pp. 883–892.

[39] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels, *Nat. Struct. Biol.* **4**, no. 1 (1997), pp. 10–19.

[40] J. Lederberg. Genes and antibodies, *Science* **129**, no. 3364 (1959), pp. 1649–1653.

[41] J. A. Weinstein et al. High-throughput sequencing of the zebrafish antibody repertoire, *Science* **324**, no. 5928 (2009), pp. 807–810.

[42] F. Horns et al. Signatures of selection in the human antibody repertoire: selective sweeps, competing subclones, and neutral drift, *bioRxiv* **145052** (2017).

[43] R. Milo et al. The relationship between evolutionary and physiological variation in hemoglobin, *Proc. Natl. Acad. Sci. U.S.A.* **104**, no. 43 (2007), pp. 16998–17003.

[44] V. M. Ingram. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin, *Nature* **178**, no. 4537 (1956), pp. 792–794.

[45] D. M. Weinreich et al. Darwinian evolution can follow only very few mutational paths to fitter proteins, *Science* **312**, no. 5770 (2006), pp. 111–114.

[46] P. A. zur Wiesch et al. Population biological principles of drug-resistance evolution in infectious diseases, *Lancet Infect. Dis.* **11**, no. 3 (2011), pp. 236–247.

[47] J. Shendure et al. DNA sequencing at 40: past, present and future, *Nature* **550**, no. 7676 (2017), pp. 345–353.

[48] F. Sanger. The arrangement of amino acids in proteins, *Adv. Protein Chem.* **7**, no. C (1952), pp. 1–67.

[49] R. W. Holley et al. Structure of a ribonucleic acid, *Science* **147**, no. 3664 (1965), pp. 1462–1465.

[50] W. M. Jou et al. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein, *Nature* **237**, no. 5350 (1972), pp. 82–88.

[51] W. Fiers et al. Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene, *Nature* **260**, no. 5551 (1976), pp. 500–507.

[52] A. M. Maxam and W. Gilbert. A new method for sequencing DNA, *Proc. Natl. Acad. Sci. U.S.A.* **74**, no. 2 (1977), pp. 560–564.

[53] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors, *Proc. Natl. Acad. Sci. U.S.A.* **74**, no. 12 (1977), pp. 5463–5467.

[54] F. Sanger et al. Nucleotide sequence of bacteriophage ϕX174 DNA, *Nature* **265**, no. 5596 (1977), pp. 687–695.

[55] J. Shendure and H. Ji. Next-generation DNA sequencing, *Nat. Biotechnol.* **26**, no. 10 (2008), pp. 1135–1145.

[56] R. Staden. A strategy of DNA sequencing employing computer programs, *Nucleic Acids Res.* **6**, no. 7 (1979), pp. 2601–2610.

[57] L. M. Smith et al. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, no. 6071 (1986), pp. 674–679.

[58] R. D. Fleischmann et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd, *Science* **269**, no. 5223 (1995), pp. 496–512.

[59] F. R. Blattner et al. The complete genome sequence of Escherichia coli K-12, *Science* **277**, no. 5331 (1997), pp. 1453–1462.

[60] A. Goffeau et al. Life with 6000 genes, *Science* **274**, no. 5287 (1995), pp. 546–567.

[61] The C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: A platform for investigating biology, *Science* **282**, no. 5396 (1998), pp. 2012–2018.

[62] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana, *Nature* **408**, no. 6814 (2000), pp. 796–815.

[63] E. S. Lander et al. Initial sequencing and analysis of the human genome, *Nature* **409**, no. 6822 (2001), pp. 860–921.

[64] J. C. Venter et al. The sequence of the human genome, *Science* **291**, no. 5507 (2001), pp. 1304–1351.

[65] D. R. Bentley et al. Accurate whole human genome sequencing using reversible terminator chemistry, *Nature* **456**, no. 7218 (2008), pp. 53–59.

[66] 1000 Genomes Project Consortium. A global reference for human genetic variation, *Nature* **526**, no. 7571 (2015), pp. 68–74.

[67] A. Scally et al. Insights into hominid evolution from the gorilla genome sequence, *Nature* **483**, no. 7388 (2012), pp. 169–175.

[68] A. H. Laszlo et al. Decoding long nanopore sequencing reads of natural DNA, *Nat. Biotechnol.* **32**, no. 8 (2014), pp. 829–833.

[69] M. Gasperini, L. Starita, and J. Shendure. The power of multiplexed functional analysis of genetic variants, *Nat. Protoc.* **11**, no. 10 (2016), pp. 1782–1787.

[70] F. Storici, L. K. Lewis, and M. A. Resnick. In vivo site-directed mutagenesis using oligonucleotides. *Nat. Biotechnol.* **19**, no. 8 (2001), pp. 773–776.

[71] M. Jinek et al. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity, *Science* **337**, no. 6096 (2012), pp. 816–822.

[72] L. Cong et al. Multiplex genome engineering using CRISPR/Cas systems, **339**, no. 6121 (2013), pp. 1766–1769.

[73] L. Yang et al. RNA-guided human genome engineering via Cas9, **339**, no. 6121 (2013), pp. 823–827.

[74] W. P. Stemmer. Rapid evolution of a protein in vitro by DNA shuffling, *Nature* **370**, no. 6488 (1994), pp. 389–391.

[75] L. Parts et al. Revealing the genetic structure of a trait by sequencing a population under selection, *Genome Res.* **21**, no. 7 (2011), pp. 1131–1138.

[76] I. Vázquez-García et al. Clonal heterogeneity influences the fate of new adaptive mutations, *Cell Rep.* **21**, no. 3 (2017), pp. 732–744.

[77] M. S. Packer and D. R. Liu. Methods for the directed evolution of proteins, *Nat. Rev. Genet.* **16**, no. 7 (2015), pp. 379–394.

[78] J. B. Kinney et al. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence, *Proc. Natl. Acad. Sci. U.S.A.* **107**, no. 20 (2010), pp. 9158–9163.

[79] R. N. McLaughlin Jr et al. The spatial architecture of protein function and adaptation, *Nature* **491**, no. 7422 (2012), pp. 138–142.

[80] D. M. Fowler et al. High-resolution mapping of protein sequence-function relationships, *Nat. Methods* **7**, no. 9 (2010), pp. 741–746.

[81] A. Crameri et al. DNA shuffling of a family of genes from diverse species accelerates directed evolution, *Nature* **391**, no. 6664 (1998), pp. 288–291.

[82] M. C. Orencia et al. Predicting the emergence of antibiotic resistance by directed evolution and structural analysis, *Nat. Struct. Biol.* **8**, no. 3 (2001), pp. 238–242.

[83] A. C. Palmer et al. Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes, *Nat. Commun.* **6**, no. 7385 (2015), pp. 1–8.

[84] M. A. Stiffler, D. R. Hekstra, and R. Ranganathan. Evolvability as a function of purifying selection in TEM-1 β-lactamase, *Cell* **160**, no. 5 (2015), pp. 882–892.

[85] R. T. Hietpas, J. D. Jensen, and D. N. A. Bolon. Experimental illumination of a fitness landscape, *Proc. Natl. Acad. Sci. U.S.A.* **108**, no. 19 (2011), pp. 7896–7901.

[86] B. Thyagarajan and J. D. Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin, *eLife* **3** (2014), pp. 1–26.

[87] K. M. Esvelt, J. C. Carlson, and D. R. Liu. A system for the continuous directed evolution of biomolecules, *Nature* **472**, no. 7344 (2011), pp. 499–503.

[88] B. C. Dickinson et al. Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. *Proc. Natl. Acad. Sci. U.S.A.* **110**, no. 22 (2013), pp. 9007–12.

[89] R. A. Fisher. *The genetical theory of natural selection*. Clarendon Press, 1930.

[90] S. Wright. Evolution in Mendelian populations, *Genetics* **16**, no. 97 (1930), pp. 97–159.

[91] R. C. Lewontin and J. L. Hubby. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics* **54**, no. 2 (1966), pp. 595–609.

[92] H. Harris. Enzyme polymorphisms in man, *Proc. R. Soc. London* no. 324 (1971), pp. 301–313.

[93] M. Kimura. Evolutionary rate at the molecular level, *Nature* **217**, no. 5129 (1968), pp. 624–626.

[94] J. L. King and T. H. Jukes. Non-Darwinian evolution, *Science* **164**, no. 1 (1969), pp. 788–798.

[95] T. Ohta. Slightly deleterious mutant substitutions in evolution, *Nature* **246**, no. 5428 (1973), pp. 96–98.

[96] M. Kreitman. Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster, *Nature* **304**, no. 5925 (1983), pp. 412–417.

[97] J. Kingman. The coalescent, *Stochastic Processes and their Applications* **13**, no. 3 (1982), pp. 235–248.

[98] S. Tavaré et al. Inferring coalescence times from DNA sequence data, *Genetics* **145**, no. 2 (1997), pp. 505–518.

[99] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences, *Nature* **475**, no. 7357 (2011), pp. 493–496.

[100] S. Schiffels and R. Durbin. Inferring human population size and separation history from multiple genome sequences, *Nat. Genet.* **46**, no. 8 (2014), pp. 919–925.

[101] A. M. Moses and R. Durbin. Inferring selection on amino acid preference in protein domains, *Mol. Biol. Evol.* **26**, no. 3 (2009), pp. 527–536.

[102] A. Fischer, C. Greenman, and V. Mustonen. Germline fitness-based scoring of cancer mutations, *Genetics* **188**, no. 2 (2011), pp. 383–393.

[103] F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, *Genetics* **123**, no. 3 (1989), pp. 585–595.

[104] J. C. Fay and C. I. Wu. Hitchhiking under positive Darwinian selection, *Genetics* **155**, no. 3 (2000), pp. 1405–1413.

[105] G. I. Lang, D. Botstein, and M. M. Desai. Genetic variation and the fate of beneficial mutations in asexual populations, *Genetics* **188**, no. 3 (2011), pp. 647–661.

[106] R. A. Neher and O. Hallatschek. Genealogies of rapidly adapting populations, *Proc. Natl. Acad. Sci. U.S.A.* **110**, no. 2 (2013), pp. 437–442.

[107] B. T. Grenfell et al. Unifying the epidemiological and evolutionary dynamics of pathogens, *Science* **303**, no. 5656 (2004), pp. 327–332.

[108] D. Seifert et al. A framework for inferring fitness landscapes of patient-derived viruses using quasispecies theory, *Genetics* **199**, no. 1 (2015), pp. 191–203.

[109] F. Zanini et al. In vivo mutation rates and the landscape of fitness costs of HIV-1, *Virus Evol.* **3**, no. 1 (2017).

[110] S. C. Dentro et al. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types, *bioRxiv* **312041** (2018).

[111] N. Beerenwinkel et al. Genetic progression and the waiting time to cancer, *PLoS Comput. Biol.* **3**, no. 11 (2007), pp. 2239–2246.

[112] I. Bozic et al. Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci. U.S.A.* **107**, no. 43 (2010), pp. 18545–18550.

[113] I. Bozic et al. Evolutionary dynamics of cancer in response to targeted combination therapy, *eLife* **2013**, no. 2 (2013), pp. 1–15.

[114] A. Fischer, I. Vázquez-García, and V. Mustonen. The value of monitoring to control evolving populations, *Proc. Natl. Acad. Sci. U.S.A.* **112**, no. 4 (2015), pp. 1007–1012.

[115] N. Beerenwinkel et al. Cancer evolution: Mathematical models and computational inference, *Syst. Biol.* **64**, no. 1 (2015), e1–e25.

[116] A. C. Palmer and R. Kishony. Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance, *Nat. Rev. Genet.* **14**, no. 4 (2013), pp. 243–248.

[117] V. M. D'Costa et al. Antibiotic resistance is ancient, *Nature* **477**, no. 7365 (2011), pp. 457–461.

[118] K. S. Baker et al. The extant World War 1 dysentery bacillus NCTC1: A genomic analysis, *Lancet* **384**, no. 9955 (2014), pp. 1691–1697.

[119] M. Lässig, V. Mustonen, and A. M. Walczak. Predicting evolution, *Nat. Ecol. Evol.* **1**, no. 3 (2017), pp. 1–9.

[120] R. A. Neher, C. A. Russell, and B. I. Shraiman. Predicting evolution from the shape of genealogical trees, *eLife* **3** (2014), e03568.

[121] M. M. Desai. Statistical questions in experimental evolution, *J. Stat. Mech. Theory Exp.* **2013**, no. 1 (2013), P01003.

[122] P. J. Campbell et al. Pan-cancer analysis of whole genomes, *bioRxiv* **162784** (2017).

[123] R. A. Neher and B. I. Shraiman. Statistical genetics and evolution of quantitative traits, *Rev. Mod. Phys.* **83**, no. 4 (2011), pp. 1283–1300.

[124] I. M. Rouzine, J. Wakeley, and J. M. Coffin. The solitary wave of asexual evolution, *Proc. Natl. Acad. Sci. U.S.A.* **100**, no. 2 (2003), pp. 587–592.

[125] M. M. Desai, D. S. Fisher, and A. W. Murray. The speed of evolution and maintenance of variation in asexual populations, *Curr. Biol.* **17**, no. 5 (2007), pp. 385–394.

[126] V. Mustonen and M. Lässig. Fitness flux and ubiquity of adaptive evolution, *Proc. Natl. Acad. Sci. U.S.A.* **107**, no. 9 (2010), pp. 4248–53.

[127] P. A. P. Moran. Random processes in genetics, *Proc. Camb. Philol. Soc.* **54** (1958), p. 60.

[128] N. Van Kampen. *Stochastic processes in physics and chemistry*. 3rd ed. Elsevier, 2007.

[129] C. Gardiner. *Stochastic methods: a handbook for the natural and social sciences*. 4th ed. Springer-Verlag, 2009.

[130] A. Fischer. "Minimal models of evolution: germline fitness effects of cancer mutations and stochastic tunneling under strong recombination". PhD thesis. Universität zu Köln, 2011.

[131] W. J. Ewens. *Mathematical population genetics*. 2nd ed. Springer-Verlag, 2004.

[132] M. Kimura. Stochastic processes and distribution of gene frequencies under natural selection, *Cold Spring Harb Symp Quant Biol* **20** (1955), pp. 33–53.

[133] R. Durrett. *Probability models for DNA sequence evolution*. 2nd ed. Springer-Verlag, 2008.

[134] C. J. R. Illingworth et al. Quantifying selection acting on a complex trait using allele frequency time series data, *Mol. Biol. Evol.* **29**, no. 4 (2012), pp. 1187–1197.

[135] Free Software Foundation. *GNU Scientific Library Reference Manual*. 3rd ed. Network Theory Ltd, 2009.

[136] F. Zanini and R. A. Neher. FFPopSim: An efficient forward simulation package for the evolution of large populations, *Bioinformatics* **28**, no. 24 (2012), pp. 3332–3333.

[137] D. A. Landau et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia, *Cell* **152**, no. 4 (2013), pp. 714–726.

[138] O. Miotto et al. Genetic architecture of artemisinin-resistant Plasmodium falciparum, *Nat. Genet.* **47**, no. 3 (2015), pp. 226–234.

[139] D. Frumkin et al. Genomic variability within an organism exposes its cell lineage tree, *PLoS Comput. Biol.* **1**, no. 5 (2005), pp. 382–394.

[140] J. E. Sulston et al. The embryonic cell lineage of the nematode Caenorhabditis elegans, *Dev. Biol.* **100**, no. 1 (1983), pp. 64–119.

[141] M. B. Woodworth, K. M. Girskis, and C. A. Walsh. Building a lineage from single cells: genetic techniques for cell lineage tracking, *Nat. Rev. Genet.* **18**, no. 4 (2017), pp. 230–244.

[142] K. D. Siegmund et al. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc. Natl. Acad. Sci. U.S.A.* **106**, no. 12 (2009), pp. 4828–4833.

[143] A. Fischer et al. High-definition reconstruction of clonal composition in cancer, *Cell Rep.* **7**, no. 5 (2014), pp. 1740–1752.

[144] N. Navin et al. Tumour evolution inferred by single-cell sequencing, *Nature* **472**, no. 7341 (2011), pp. 90–94.

[145] N. E. Potter et al. Single-cell mutational profiling and clonal phylogeny in cancer, *Genome Res.* **23**, no. 12 (2013), pp. 2115–2125.

[146] E. Shapiro, T. Biezuner, and S. Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science, *Nat. Rev. Genet.* **14**, no. 9 (2013), pp. 618–630.

[147] M. A. Lodato et al. Somatic mutation in single human neurons tracks developmental and transcriptional history, *Science* **350**, no. 6256 (2015), pp. 94–98.

[148] A. Roth et al. Clonal genotype and population structure inference from single-cell tumor sequencing, *Nat. Methods* **13**, no. 7 (2016), pp. 573–576.

[149] A. M. Klein et al. Stochastic fate of p53-mutant epidermal progenitor cells is tilted toward proliferation by UV B during preneoplasia. *Proc. Natl. Acad. Sci. U.S.A.* **107**, no. 1 (2010), pp. 270–275.

[150] C. Blanpain and B. D. Simons. Unravelling stem cell dynamics by lineage tracing, *Nat. Rev. Mol. Cell Biol.* **14**, no. 8 (2013), pp. 489–502.

[151] A. McKenna et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing, *Science* **353**, no. 6298 (2016), aaf7907.

[152] J. P. Junker et al. Massively parallel whole-organism lineage tracing using CRISPR/Cas9 induced genetic scars, *bioRxiv* **056499** (2016).

[153] K. L. Frieda et al. Synthetic recording and in situ readout of lineage information in single cells, *Nature* **541**, no. 7635 (2017), pp. 107–111.

[154] R. Kalhor, P. Mali, and G. M. Church. Rapidly evolving homing CRISPR barcodes, *Nat. Methods* **14**, no. 2 (2017), pp. 195–200.

[155] N. Roquet et al. Synthetic recombinase-based state machines in living cells, *Science* **353**, no. 6297 (2016), aad8559.

[156] S. L. Shipman et al. Molecular recordings by directed CRISPR spacer acquisition, *Science* **353**, no. 6298 (2016), aaf1175.

[157] A. Roth et al. PyClone: Statistical inference of clonal population structure in cancer, *Nat. Methods* **11**, no. 4 (2014), pp. 396–398.

[158] W. Jiao et al. Inferring clonal evolution of tumors from single nucleotide somatic mutations, *BMC Bioinformatics* **15**, no. 3 (2014), pp. 1–16.

[159] S. L. Carter et al. Absolute quantification of somatic DNA alterations in human cancer, *Nat. Biotechnol.* **30**, no. 5 (2012), pp. 413–421.

[160] L. Oesper, A. Mahmoody, and B. J. Raphael. THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data, *Genome Biol.* **14**, no. 7 (1910), pp. 1–41.

[161] G. Ha et al. TITAN: Inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data, *Genome Res.* **24**, no. 11 (2014), pp. 1881–1893.

[162] L. Oesper, G. Satas, and B. J. Raphael. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data, *Bioinformatics* **30**, no. 24 (2014), pp. 3532–3540.

[163] S. Nik-Zainal et al. The life history of 21 breast cancers, *Cell* **149**, no. 5 (2012), pp. 994–1007.

[164] L. Rabiner and B. Juang. An introduction to Hidden Markov Models, *IEEE ASSP Mag.* **3** (1986), pp. 4–16.

[165] R. Durbin et al. *Biological sequence analysis*. 1st ed. Cambridge University Press, 1998.

[166] R. E. Kalman. A new approach to linear filtering and prediction problems, *J. Basic Eng.* **82**, no. 1 (1960), pp. 35–45.

[167] G. Welch and G. Bishop. *An introduction to the Kalman Filter*. Tech. rep. University of North Carolina at Chapel Hill, 1995, pp. 1–16.

[168] M. S. Arulampalam et al. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, *IEEE Trans. Signal Process.* **50**, no. 2 (2002), pp. 174–188.

[169] K. P. Murphy. *Machine learning: A probabilistic perspective*. 1st ed. MIT Press, 2012.

[170] S. Sengupta et al. BayClone: Bayesian nonparametric inference of tumor subclones using NGS data, *Proc. Pacific Symp. Biocomput.* (2015), pp. 467–478.

[171] N. Donmez et al. Clonality inference from single tumor samples using low-coverage sequence data, *J. Comput. Biol.* **24**, no. 6 (2017), pp. 515–523.

[172] A. G. Deshwar et al. PhyloWGS: Reconstructing subclonal composition and evolution from whole genome sequencing of tumors, *Genome Biol.* **16**, no. 35 (2015), pp. 1–20.

[173] M. Cmero et al. SVclone: Inferring structural variant cancer cell fraction, *bioRxiv* **172486** (2017).

[174] H. Farahani et al. Engineered in-vitro cell line mixtures and robust evaluation of computational methods for clonal decomposition and longitudinal dynamics in cancer, *Sci. Rep.* **7**, no. 13467 (2017), pp. 1–13.

[175] A. Stern et al. The evolutionary pathway to virulence of an RNA virus, *Cell* **169**, no. 1 (2017), pp. 35–46.

[176] M. Vignuzzi et al. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population, *Nature* **439**, no. 7074 (2006), pp. 344–348.

[177] T. D. Lieberman et al. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat. Genet.* **46**, no. 1 (2014), pp. 82–87.

[178] X. Didelot et al. Within-host evolution of bacterial pathogens, *Nat. Rev. Microbiol.* **14**, no. 3 (2016), pp. 150–162.

[179] R. Paredes et al. Pre-existing minority drug-resistant HIV-1 variants, adherence, and risk of antiretroviral treatment failure, *J. Infect. Dis.* **201**, no. 5 (2010), pp. 662–671.

[180] P. S. Pennings. Standing genetic variation and the evolution of drug resistance in HIV, *PLoS Comput. Biol.* **8**, no. 6 (2012), e1002527.

[181] C. C. Maley et al. Genetic clonal diversity predicts progression to esophageal adenocarcinoma, *Nat. Genet.* **38**, no. 4 (2006), pp. 468–73.

[182] N. McGranahan and C. Swanton. Clonal heterogeneity and tumor evolution: past, present, and the future, *Cell* **168**, no. 4 (2017), pp. 613–628.

[183] S. Kryazhimskiy et al. Global epistasis makes adaptation predictable despite sequence-level stochasticity, *Science* **344**, no. 6191 (2014), pp. 1519–1522.

[184] J. V. Rodrigues et al. Biophysical principles predict fitness landscapes of drug resistance, *Proc. Natl. Acad. Sci. U.S.A.* **113**, no. 13 (2016), E1964–E1964.

[185] O. Tenaillon et al. The molecular diversity of adaptive convergence, *Science* **335**, no. 6067 (2012), pp. 457–461.

[186] C. Josephides and P. S. Swain. Predicting metabolic adaptation from networks of mutational paths, *Nat. Commun.* **8**, no. 1 (2017), pp. 1–15.

[187] D. R. Hekstra and S. Leibler. Contingency and statistical laws in replicate microbial closed ecosystems, *Cell* **149**, no. 5 (2012), pp. 1164–1173.

[188] G. I. Lang et al. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations, *Nature* **500**, no. 7464 (2013), pp. 571–574.

[189] N. J. Cira, M. T. Pearce, and S. R. Quake. Neutral and niche dynamics in a synthetic microbial community, *bioRxiv* **107896** (2017).

[190] J. R. Meyer et al. Repeatability and contingency in the evolution of a key innovation in phage lambda, *Science* **335**, no. 6067 (2012), pp. 428–432.

[191] M. Hegreness et al. Accelerated evolution of resistance in multidrug environments, *Proc. Natl. Acad. Sci. U.S.A.* **105**, no. 37 (2008), pp. 13977–13981.

[192] E. Toprak et al. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection, *Nat. Genet.* **44**, no. 1 (2012), pp. 101–105.

[193] Q. Zhang et al. Acceleration of emergence of bacterial antibiotic resistance in connected microenvironments, *Science* **333**, no. 6050 (2011), pp. 1764–1767.

[194] R. Hermsen, J. B. Deris, and T. Hwa. On the rapidity of antibiotic resistance evolution facilitated by a concentration gradient, *Proc. Natl. Acad. Sci. U.S.A.* **109**, no. 27 (2012), pp. 10775–10780.

[195] P. Greulich, B. Waclaw, and R. J. Allen. Mutational pathway determines whether drug gradients accelerate evolution of drug-resistant cells, *Phys. Rev. Lett.* **109**, no. 8 (2012), pp. 1–5.

[196] K. A. Lipinski et al. Cancer evolution and the limits of predictability in precision cancer medicine, *Trends in Cancer* **2**, no. 1 (2016), pp. 49–63.

[197] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics* **25**, no. 14 (2009), pp. 1754–1760.

[198] K. Cibulskis et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples, *Nat. Biotechnol.* **31**, no. 3 (2013), pp. 213–219.

[199] The Gene Ontology Consortium. Gene Ontology: Tool for the unification of biology, *Nat. Genet.* **25**, no. 1 (2000), pp. 25–29.

[200] The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources, *Nucleic Acids Res.* **45**, no. D1 (2017), pp. D331–D338.

[201] A. Saleh et al. Tra1p is a component of the yeast Ada·Spt transcriptional regulatory complexes, *J. Biol. Chem.* **273**, no. 41 (1998), pp. 26559–26565.

[202] G. Yvert et al. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat. Genet.* **35**, no. 1 (2003), pp. 57–64.

[203] J. Ronald et al. Local regulatory variation in Saccharomyces cerevisiae, *PLoS Genet.* **1**, no. 2 (2005), pp. 213–222.

[204] M. Kaeberlein et al. Regulation of yeast replicative life span by TOR and Sch9 in response to nutrients. *Science* **310**, no. 5751 (2005), pp. 1193–1196.

[205] G. Devasahayam et al. Pmr1, a Golgi Ca2+/Mn2+-ATPase, is a regulator of the target of rapamycin (TOR) signaling pathway in yeast, *Proc. Natl. Acad. Sci. U.S.A.* **103**, no. 47 (2006), pp. 17840–17845.

[206] G. Devasahayam, D. J. Burke, and T. W. Sturgill. Golgi manganese transport is required for rapamycin signaling in Saccharomyces cerevisiae, *Genetics* **177**, no. 1 (2007), pp. 231–238.

[207] M. E. Cardenas et al. The TOR signaling cascade regulates gene expression in response to nutrients, *Genes Dev.* **13**, no. 24 (1999), pp. 3271–3279.

[208] J. R. Rohde and M. E. Cardenas. The Tor pathway regulates gene expression by linking nutrient sensing to histone acetylation, *Mol. Cell. Biol.* **23**, no. 2 (2003), pp. 629–635.

[209] S. Venkataram et al. Development of a comprehensive genotype-to-fitness map of adaptation-driving mutations in yeast, *Cell* **166**, no. 6 (2016), pp. 1585–1596.

[210] N. Strelkowa and M. Lässig. Clonal interference in the evolution of influenza, *Genetics* **192**, no. 2 (2012), pp. 671–682.

[211] A. H. Yona et al. Chromosomal duplication is a transient evolutionary solution to stress, *Proc. Natl. Acad. Sci. U.S.A.* **109**, no. 51 (2012), pp. 21010–21015.

[212] S. F. Levy et al. Quantitative evolutionary dynamics using high-resolution lineage tracking, *Nature* **519**, no. 7542 (2015), pp. 181–186.

[213] Z. N. Rogers et al. A quantitative and multiplexed approach to uncover the fitness landscape of tumor suppression in vivo, *Nat. Methods* **14**, no. 7 (2017), pp. 737–742.

[214] K. Y. Su et al. Pretreatment Epidermal Growth Factor Receptor (EGFR) T790M mutation predicts shorter EGFR tyrosine kinase inhibitor response duration in patients with non-small-cell lung cancer, *J. Clin. Oncol.* **30**, no. 4 (2012), pp. 433–440.

[215] P. Laurent-Puig et al. Clinical relevance of KRAS-mutated subclones detected with picodroplet digital PCR in advanced colorectal cancer treated with Anti-EGFR therapy, *Clin. Cancer Res.* **21**, no. 5 (2015), pp. 1087–1097.

[216] S. E. Luria and M. Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, no. 6 (1943), pp. 491–511.

[217] P. J. Gerrish and R. E. Lenski. The fate of competing beneficial mutations in an asexual population, *Genetica* **102-103**, no. 1-6 (1998), pp. 127–144.

[218] R. A. Neher. Genetic draft, selective interference, and population genetics of rapid adaptation, *Annu. Rev. Ecol. Evol. Syst.* **44**, no. 1 (2013), pp. 195–215.

[219] S. Schiffels et al. Emergent neutrality in adaptive asexual evolution, *Genetics* **189**, no. 4 (2011), pp. 1361–1375.

[220] B. H. Good et al. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations, *Proc. Natl. Acad. Sci. U.S.A.* **109**, no. 13 (2012), pp. 4950–4955.

[221] M. Zackrisson et al. Scan-o-matic: High-resolution microbial phenomics at a massive scale, *G3 Genes|Genomes|Genetics* **6**, no. 9 (2016), pp. 3003–3014.

[222] C. J. R. Illingworth et al. Inferring genome-wide recombination landscapes from advanced intercross lines: application to yeast crosses, *PLoS One* **8**, no. 5 (2013), pp. 1–10.

[223] A. Hamon and B. Ycart. Statistics for the Luria-Delbrück distribution, *Electron. J. Stat.* **6** (2012), pp. 1251–1272.

[224] M. A. Barbera and T. D. Petes. Selection and analysis of spontaneous reciprocal mitotic cross-overs in Saccharomyces cerevisiae, *Proc. Natl. Acad. Sci. U.S.A.* **103**, no. 34 (2006), pp. 12819–12824.

[225] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. 1st ed. Cambridge University Press, 2006.

[226] A. Fischer et al. EMu: probabilistic inference of mutational processes and their localization in the cancer genome, *Genome Biol.* **14**, no. 4 (2013), R39.

[227] A. A. Larrea et al. Genome-wide model for the normal eukaryotic DNA replication fork, *Proc. Natl. Acad. Sci. U.S.A.* **107**, no. 41 (2010), pp. 17674–17679.

[228] L. B. Alexandrov et al. Signatures of mutational processes in human cancer, *Nature* **500**, no. 7463 (2013), pp. 415–421.

[229] S. Behjati et al. Mutational signatures of ionizing radiation in second malignancies, *Nat. Commun.* **7**, no. 12605 (2016), pp. 1–8.

[230] E. D. Pleasance et al. A comprehensive catalogue of somatic mutations from a human cancer genome, *Nature* **463** (2010), pp. 191–197.

[231] L. B. Alexandrov et al. Mutational signatures associated with tobacco smoking in human cancer, *Science* **354**, no. 6312 (2016), pp. 618–622.

[232] M. L. Hoang et al. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing, *Sci. Transl. Med.* **5**, no. 197 (2013), pp. 1–10.

[233] S. L. Poon et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool, *Sci. Transl. Med.* **5**, no. 197 (2013), pp. 1–10.

[234] S. P. Jackson and J. Bartek. The DNA-damage response in human biology and disease, *Nature* **461**, no. 7267 (2009), pp. 1071–1078.

[235] A. Serero et al. Mutational landscape of yeast mutator strains, *Proc. Natl. Acad. Sci. U.S.A.* **111**, no. 5 (2014), pp. 1897–1902.

[236] P. C. Stirling et al. Genome destabilizing mutator alleles drive specific mutational trajectories in Saccharomyces cerevisiae, *Genetics* **196**, no. 2 (2014), pp. 403–412.

[237] Y. O. Zhu et al. Precise estimates of mutation rate and spectrum in yeast, *Proc. Natl. Acad. Sci. U.S.A.* **111**, no. 22 (2014), E2310–E2318.

[238] B. Meier et al. C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency, *Genome Res.* **24**, no. 10 (2014), pp. 1624–1636.

[239] B. Meier et al. Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers, *bioRxiv* **149153** (2017).

[240] J. V. Forment et al. Genome-wide genetic screening with chemically mutagenized haploid embryonic stem cells, *Nat. Chem. Biol.* **13**, no. 1 (2017), pp. 12–14.

[241] A. C. Bester et al. Nucleotide deficiency promotes genomic instability in early stages of cancer development, *Cell* **145**, no. 3 (2011), pp. 435–446.

[242] C. Swanton et al. APOBEC enzymes: Mutagenic fuel for cancer evolution and heterogeneity, *Cancer Discov.* **5**, no. 7 (2015), pp. 704–712.

[243] J.-B. Lee et al. DNA primase acts as a molecular brake in DNA replication, *Nature* **439**, no. 7076 (2006), pp. 621–624.

[244] C. A. Nieduszynski et al. OriDB: A DNA replication origin database, *Nucleic Acids Res.* **35** (2007), pp. 40–46.

[245] M. Lynch et al. A genome-wide view of the spectrum of spontaneous mutations in yeast, *Proc. Natl. Acad. Sci. U.S.A.* **105**, no. 27 (2008), pp. 9272–9277.

[246] R. Rahbari et al. Timing, rates and spectra of human germline mutation, *Nat. Genet.* **48**, no. 2 (2015), pp. 126–133.

[247] J. J. Hopfield. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity, *Proc. Natl. Acad. Sci. U.S.A.* **71**, no. 10 (1974), pp. 4135–4139.

[248] J. Ninio. Kinetic amplification of enzyme discrimination, *Biochimie* **57**, no. 5 (1975), pp. 587–595.

[249] T. A. Kunkel, L. A. Loeb, and M. F. Goodman. On the fidelity of DNA replication, *J. Biol. Chem.* **259**, no. 3 (1984), pp. 1539–1545.

[250] M. K. Burke et al. Genome-wide analysis of a long-term evolution experiment with Drosophila, *Nature* **467**, no. 7315 (2010), pp. 587–590.

[251] S. Hormoz et al. Inferring cell state transition dynamics from lineage trees and endpoint single-cell measurements, *Cell Syst.* **3**, no. 5 (2016), pp. 419–433.

[252] I. M. Rouzine and J. M. Coffin. Evolution of human immunodeficiency virus under selection and weak recombination, *Genetics* **170**, no. 1 (2005), pp. 7–18.

[253] M. S. Lawrence et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes, *Nature* **499**, no. 7457 (2013), pp. 214–218.

[254] T. D. Lieberman et al. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures, *Nat. Genet.* **46**, no. 1 (2014), pp. 82–87.

[255] S. Boyer et al. Hierarchy and extremes in selections from pools of randomized proteins, *Proc. Natl. Acad. Sci. U.S.A.* **113**, no. 13 (2016), pp. 3482–3487.

[256] A. R. Fersht. Fidelity of replication of phage ϕX174 DNA by DNA polymerase III holoenzyme: Spontaneous mutation by misincorporation, *Proc. Natl. Acad. Sci. U.S.A.* **76**, no. 10 (1979), pp. 4946–4950.

[257] B. N. Ames et al. Carcinogens are mutagens: A simple test system combining liver homogenates for activation and bacteria for detection, *Proc. Natl. Acad. Sci. U.S.A.* **70**, no. 8 (1973), pp. 2281–2285.