Genomics 104 (2014) 417–430

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Beyond genome sequencing: Lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer

Jamie R. Blundell [a,*], Sasha F. Levy [b,c,**]

[a] Department of Applied Physics, Stanford University, Stanford, CA 94305, USA
[b] Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794-5252, USA
[c] Department of Biochemistry and Cellular Biology, Stony Brook University, Stony Brook, NY 11794-5215, USA

ARTICLE INFO

Article history:
Received 15 August 2014
Accepted 16 September 2014
Available online 28 September 2014

Keywords:
Evolution
Evolutionary dynamics
Lineage
Tracking
Tracing
Sequencing
Barcode

ABSTRACT

Evolving cellular communities, such as the gut microbiome, pathogenic infections, and cancer, consist of large populations of ~10^7–10^{14} cells. Because of their large population sizes, adaptation within these populations can be driven by many beneficial mutations that never rise above extremely low frequencies. Genome sequencing methods such as clonal, single cell, or whole population sequencing are poorly suited to detect these rare beneficial lineages, and, more generally, to characterize which mutations are most important to the population dynamics. Here, we introduce an alternative approach: high-resolution lineage tracking with DNA barcodes. In contrast to whole genome sequencing, lineage tracking can detect a beneficial mutation at an extremely low frequency within the population, and estimate its time of occurrence and fitness effect. Many lineage trajectories can be observed in parallel, allowing one to observe the population dynamics in exquisite detail. We describe some of the technical and analytical challenges to lineage tracking with DNA barcodes and discuss its applications to studies of evolution, infectious disease and cancer.

1. Introduction

1.1. The evolution of large asexual cell populations

The evolution of large asexual cell populations underlies many diseases, such those caused by bacteria, parasites and cancer. Biomedical research has historically focused on finding the mutational determinants of increased pathogenicity, cancer progression, or drug resistance — the specific beneficial mutations that result in increased fitness. Largely due to the rise of next generation sequencing technologies, these studies have been incredibly successful. In controlled systems, such as experimental evolution studies, identification of numerous beneficial mutational determinants has become a routine [1–7]. As we will discuss, finding the beneficial mutational determinants, or "drivers", in less well-controlled disease-relevant systems is significantly more challenging. Nevertheless, progress has been made here too [6,8–16]. In parallel, studies aimed at understanding how quickly mutations enter a population have been consistently revising our estimate of the beneficial mutation rate upward over time [17–30].

The net outcome of high beneficial mutation rates is that, even at relatively small population sizes (>10^6 cells), many beneficial mutations enter the population contemporaneously and compete with one another [1–3,31–33]. This process, known as clonal interference, blunts the impact of any one mutation on the population, and, more practically, makes evolution difficult to study [22,33–35]. By contrast with a regime of successive sweeps [22,31], whereby one beneficial mutation at a time enters and fixes within a population, in clonal interference, the evolutionary dynamics may be influenced by many adaptive mutations that exist at extremely low frequencies [31,36–38]. Furthermore, a full enumeration of all mutational determinants and their fitness effects on its own is not enough to provide full predictive insight into the adaptive process. This is because the rates at which each mutation enters the population and the population size are as important to the evolutionary dynamics as the mutations themselves (Box 1). Thus, studies aimed at measuring these rates, or, more generally, the dynamics by which many adaptive lineages compete within a population are of paramount importance.

Here, we discuss a powerful new experimental approach to study the evolutionary dynamics of large populations that relies on barcoding genomes with unique random DNA sequences that can be tracked over time. We first contrast the information that can be gained with this technique with that of genome sequencing, while highlighting its experimental and theoretical challenges. We will next briefly consider the potential synergies between barcode lineage tracking and genome sequencing. Finally, we will discuss the potential applications of barcode

* Corresponding author.
** Correspondence to: S.F. Levy, Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794-5252, USA.
E-mail addresses: jrb75@stanford.edu (J.R. Blundell), sasha.levy@laufercenter.org (S.F. Levy).

http://dx.doi.org/10.1016/j.ygeno.2014.09.005
0888-7543/

lineage tracking in studies of evolution, infection, drug resistance, and cancer.

Box 1
The importance of population size and mutation rate to each fitness effect.

Not all beneficial mutations have an equal impact on the evolutionary dynamics. While a mutation's fitness determines how rapidly a mutation expands once it enters the population, the waiting time for this event is determined by the product of its mutation rate and the population size. Imagine an extremely simplified environment and genotype where only two types of beneficial mutations are possible: one that occurs at a rate of 10^{-4} per cell per generation and causes a 1% increase in fitness, and another occurs at a rate of a 10^{-8} and causes a 10% increase in fitness. At small population sizes (e.g. 10^5) the large effect mutation does not matter because it will almost never occur (Fig. 1A). At very large population sizes (e.g. $>10^{10}$) the large effect mutation will almost certainly occur quickly, making the small effect mutations largely irrelevant (Fig. 1C). However, at intermediate sizes (e.g. 10^7–10^9), many small effect mutations will accumulate before a large effect mutation enters the population, with the waiting time for this event dependent on the population size (Fig. 1B). Thus, the mean fitness of the population will be driven first by many small effect mutations and then by a few large effect mutations. If the rates to each fitness class were to change, so would the population dynamics. In a more realistic scenario, where many fitness effects are possible, each occurring at a certain rate, the population dynamics will be controlled by a more complex interaction between these different fitness classes. Therefore, the evolutionary dynamics, of for example the rate at which antibiotic resistance emerges in a population, depends on the fitness effects, mutation rates to those fitness effects, and on the population size in a non-trivial way.

1.2. Too big data

The ideal evolution experiment to probe the dynamics of large populations would track the relative abundance of all genotypes in a population over time. Even in cells with small genomes (~10^6 bases) and in population sizes that are smaller than those typical of disease (~10^7 cells [39–42]), such an endeavor is impossible in practice: a single time point would require sequencing $>10^{13}$ bases at reasonably high coverage (~10×), at a cost of ~$15 million. The key difficulty is that the amount of sequencing required is a product of two large numbers: the population size, needed to achieve near single-cell frequency resolution, and the genome size, needed to track all sites in the genome. Clonal or population sequencing tracks all sites in the genome, generally at poor frequency resolution. Lineage tracking, which we discuss here, is of the reverse philosophy: maintain close to single-cell frequency resolution by ignoring (at least for the time-being) which sites in the genome are mutating. This approach may at first seem wrong-headed since it is often the specific identity of mutations that are of biological interest. However, lineage tracking can quantitatively measure properties of many mutations in parallel: their fitness effects and times of occurrence even while at very low frequencies in the population (Box 2). This could be particularly relevant to the evolution of drug resistance where there is growing evidence that, in some cases, it could be driven by a large number of low-abundance (and weak fitness-effect) mutations [43,44]. In addition, as we will discuss later, lineage tracking can be used in combination with genome sequencing to better identify the specific beneficial mutations that are most important to driving adaptation.

Box 2
The advantage of small lineages.

Because beneficial mutations are rare, a small lineage is unlikely to acquire a beneficial mutation, and even more unlikely to acquire two beneficial mutations contemporaneously. When a small lineage increases in frequency faster than is possible by drift, one can reasonably assume that the change in frequency is due to a single beneficial mutation that occurred within the lineage and expanded. The lineage trajectory can, in turn, be used to estimate properties of this beneficial mutation, such as its fitness effect and its time of occurrence (Fig. 2). If many beneficial small lineage trajectories can be observed, population parameters such as the mutation rate to each fitness effect, can be estimated. Single beneficial mutation estimates are impossible if lineage sizes are too large because multiple beneficial mutations will occur within that lineage contemporaneously and begin to grow (clonal interference within a lineage).

2. Limitations of clonal and population sequencing

We begin by discussing the limitations of genome sequencing methods for studying evolutionary dynamics. Our central thesis is that these methods, although useful for finding some of the mutational determinants of adaptation, lack the frequency resolution to quantitatively describe the dynamics. This is because low frequency mutations (<1%)

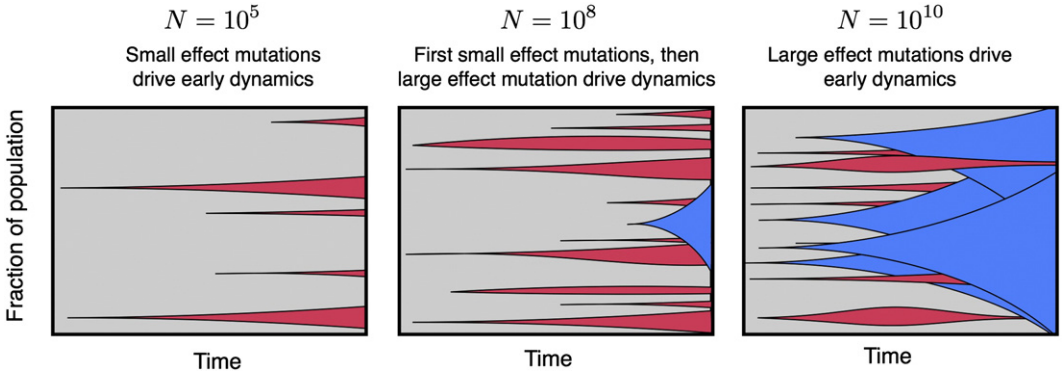

Fig. 1. (In Box 1). The impact of population size and mutation rate on the early evolutionary dynamics of a hypothetical population. (Left) At small population sizes, only commonly occurring, small fitness effect mutations (red) enter and drive the initial adaptation. (Middle) At large population sizes the dynamics are driven first by the commonly occurring small effect mutations, but later by the rarely occurring large effect mutations (blue). (Right) At very large population sizes, even the rarely occurring large effect mutations occur quickly and these drive the early dynamics.