ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

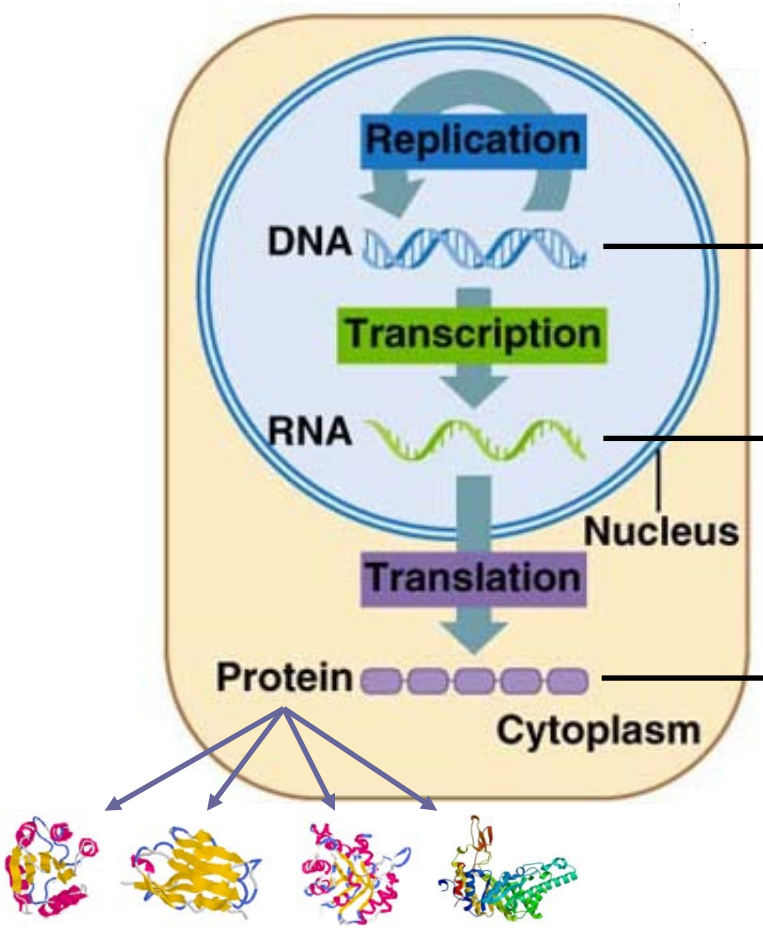
Genetic diversity of HIV

Niko Beerenwinkel

Outline

- HIV infection
- Quasispecies equation
- Measuring viral genetic diversity

The central dogma of molecular biology



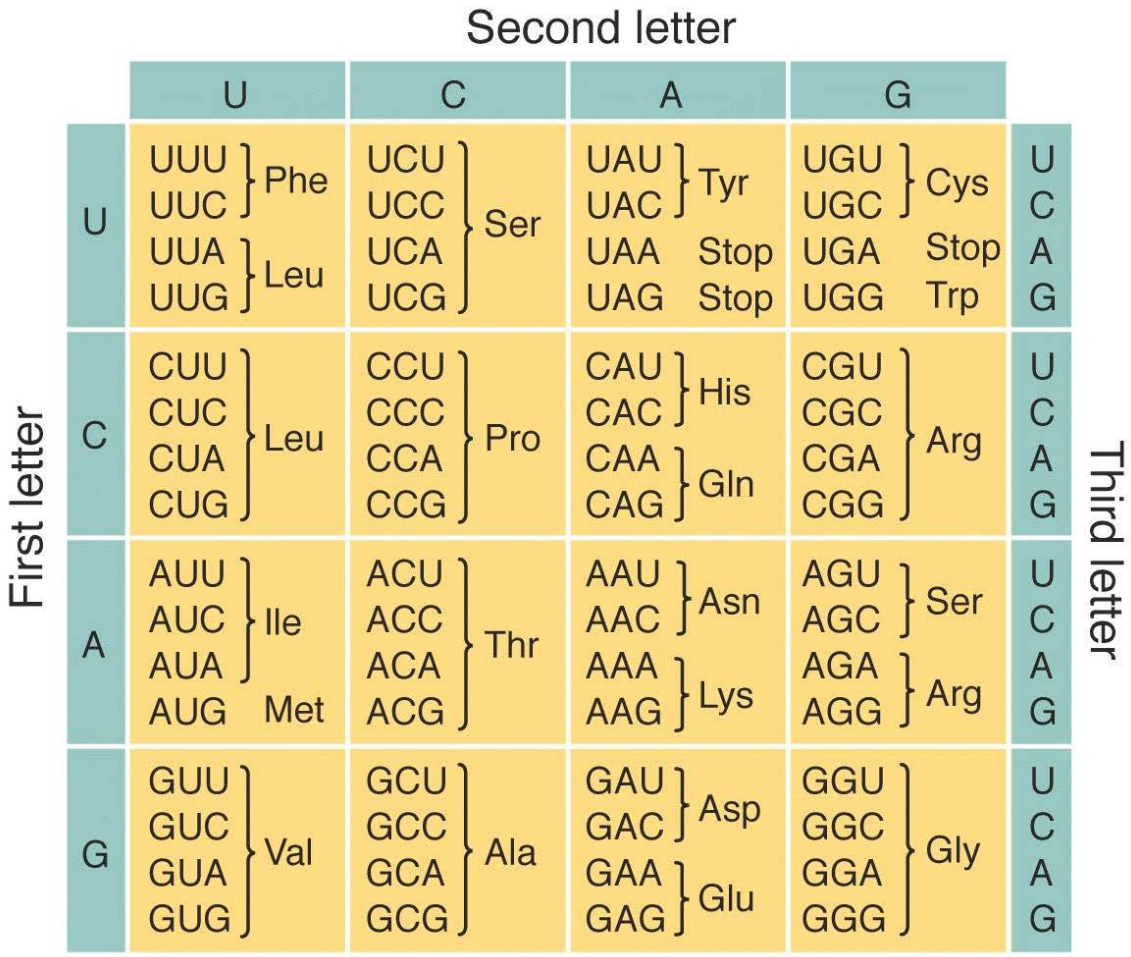
CCT CAG GTC ACT CTT …

CCU CAG GUC ACU CUU …

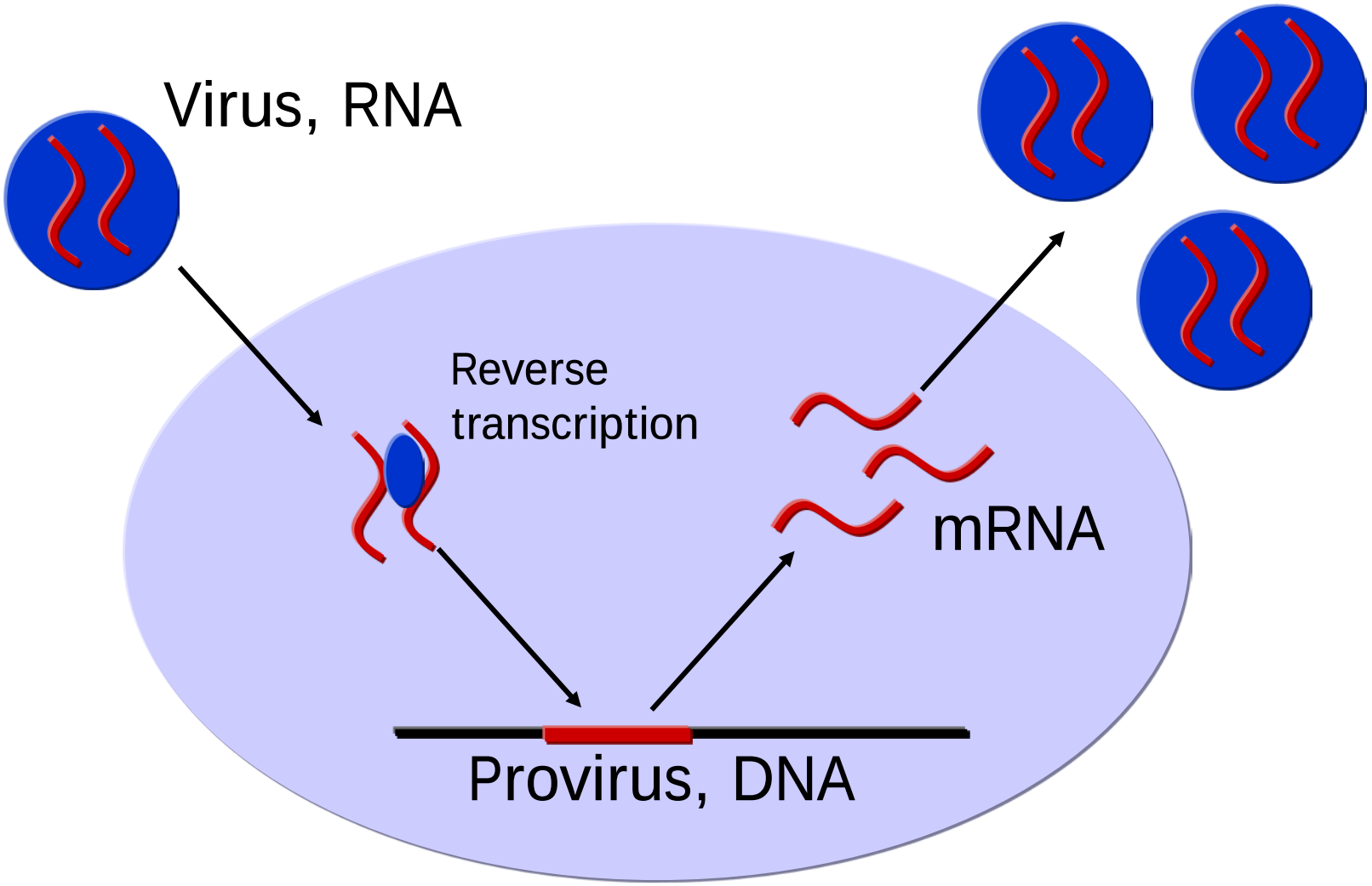
Pro Gln Val Thr Leu …

P Q V T L

The genetic code



HIV infection



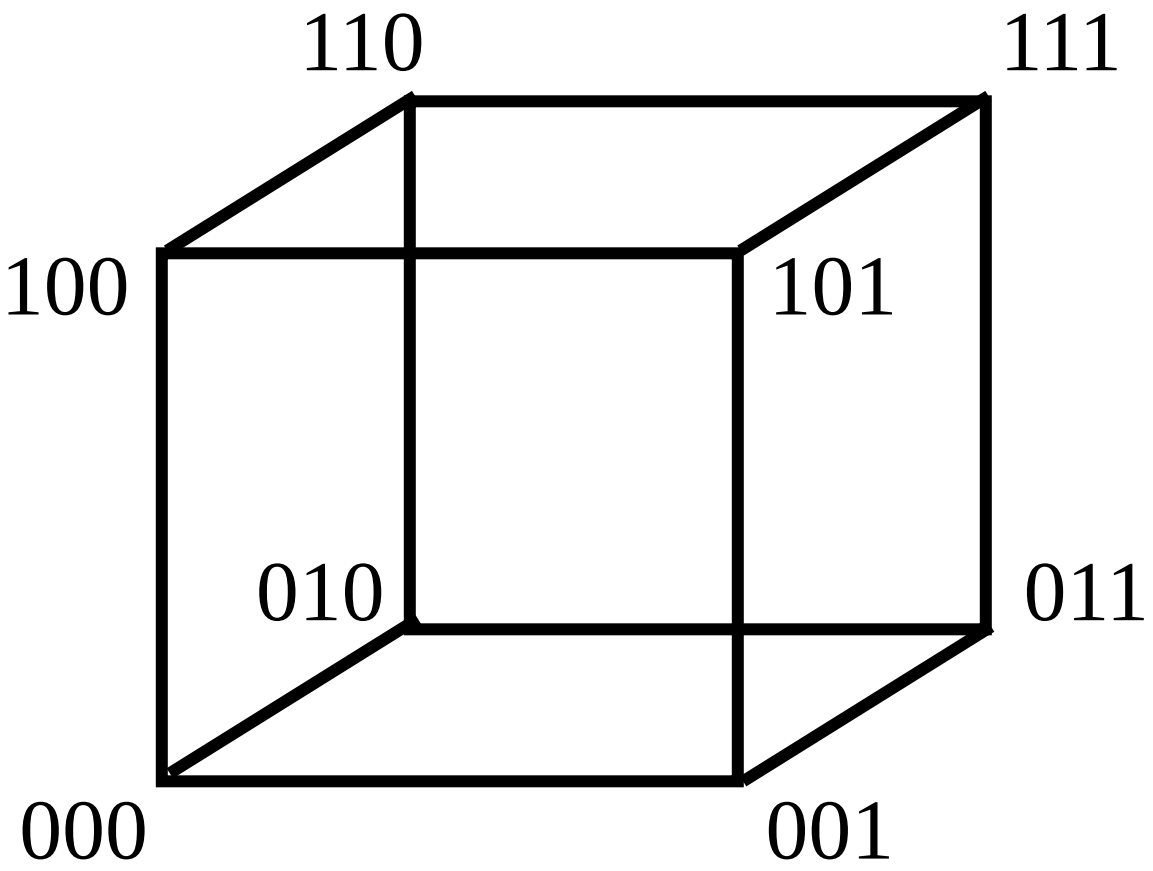
Parameters of HIV dynamics

- Short genome, 10,000 bases
- Very high mutation rate (no proof-reading of reverse transcription), $3 \cdot 10^{-5}$ per base per replication
- Large population sizes
- Short generation time
- Exposed to strong selective forces (immune response, antiretroviral therapy)

Sequence space

- Alphabet, $\mathcal{A}$
  - DNA: A, C, G, T
  - Protein: Ala, Arg, Asn, Asp, Cys, Glu, Gln, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val
  - Binary: 0, 1
- A *sequence* (or genome, or genotype if haploid populations are considered) is an element of $\mathcal{A}^L = \{(a_1, \ldots, a_L) | a_i \in \mathcal{A}\}$
- We set $\mathcal{A}^\star = \cup_{L \geq 0} \mathcal{A}^L$
- Genome sizes L range from 10^4 to 10^{11}, but we often focus on a subset of loci (and characters).

Binary sequences of length L = 3

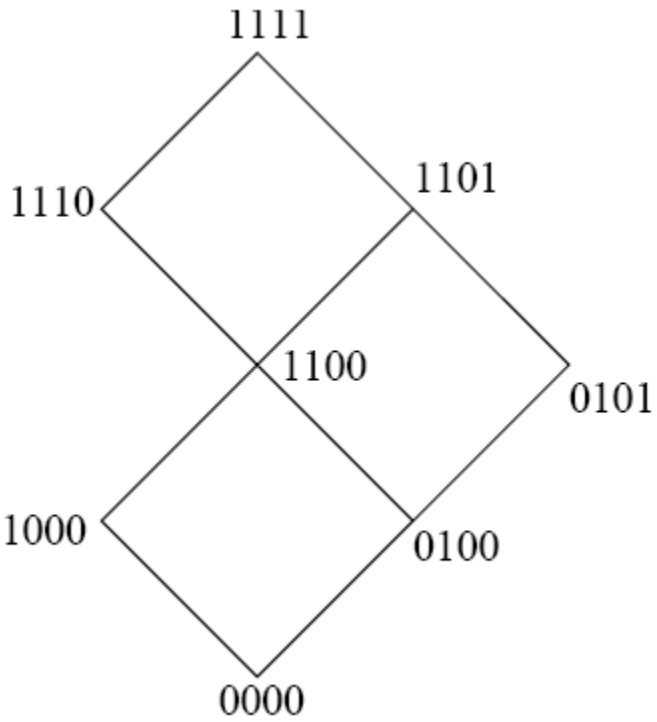


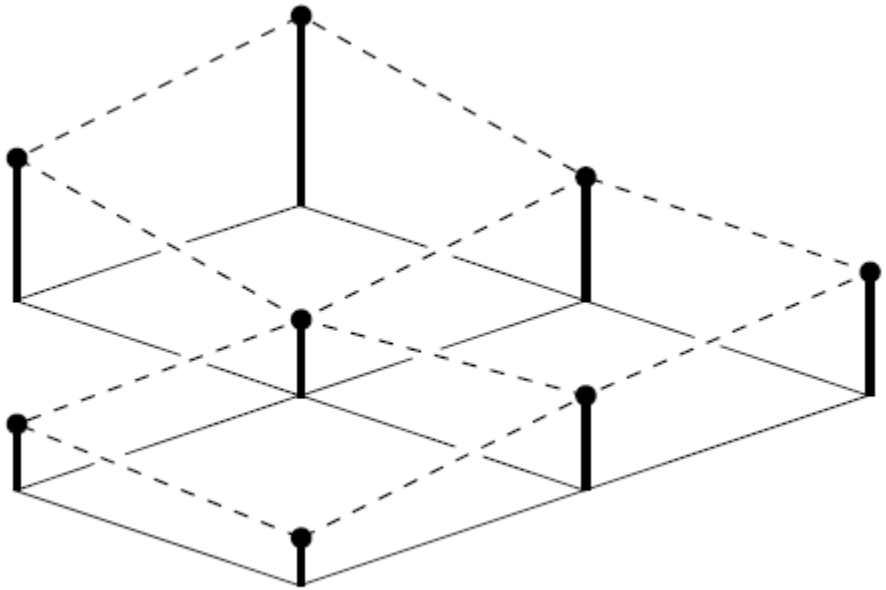
Geometry of sequence space

- The dimension of sequence space is $|\mathcal{A}|^L$.
- Distance is measured by the Hamming distance, or Manhattan distance, i.e., the number of mismatches.
- In this metric, all sequences are close to each other: the distance between any pair of sequences is bounded by L.
- For example, if $\mathcal{A}$ = {0,1} and L = 1000, then the dimension is $2^{1000} \approx 10^{301}$, but the maximum length of a shortest path from any one sequence to another is 10^3.
- Evolution is a trajectory in sequence space.
- We will often call a subset $\mathcal{G} \subset \mathcal{A}^\star$ the genotype space.

Fitness landscapes

- A fitness landscape is a mapping $f : \mathcal{G} \to \mathbb{R}$.





Genotype → phenotype → fitness

- More precisely, $f : \mathcal{G} \to \mathcal{P} \to \mathbb{R}$ because fitness depends on the phenotype of the organism (size, shape, behavior,metabolism, …).
- Phenotype space is more complex (and less well defined). Phenotypes can be discrete or continuous.
- In general, fitness depends on how the individual interacts with the ecological environment and with other individuals in the population.
- In general, fitness is difficult (often impossible) to measure.
- For some viruses (e.g., HIV) and bacteria (e.g., E. coli), there are in vitro assays for measuring fitness.

The quasispecies equation

- We consider (geno-)types i = 0, 1, 2, …, n.
- Let x(t) = (x_0(t), …, x_n(t)) denote the genotype frequencies.
- Let Q = (q_{ij}) be a mutation matrix, f = (f_0, …, f_n) a fitness landscape, and $\phi = x \cdot f$ the average fitness of the population.
- The quasispecies equation is

$$\dot{x}_i = \sum_{j=0}^{n} x_j f_j q_{ji} - \phi x_i, \qquad i = 0, \ldots, n$$

Properties of the quasispecies equation

$$\dot{x}_i = \sum_{j=0}^{n} x_j f_j q_{ji} - \phi x_i, \qquad i = 0, \ldots, n$$

- The term “species” refers to chemical species (RNA molecule types), rather than biological species.
- If replication is error free, Q = I, then we recover the selection equation (“survival of the fittest”).
- Generically, if Q is irreducible, there is a single, globally stable, equilibrium x^* in the interior of the probability simplex.
- In general, the equilibrium x^* does not maximize the average fitness ϕ. Mutation reduces population fitness.

Solving the quasispecies equation (1/2)

- The change of variables

$$X_i(t) = x_i(t)e^{\psi(t)} \quad \text{with } \psi(t) = \int_0^t \phi(s)ds$$

leads to the linear ODE system

$$\dot{X}_i = \sum_{j=0}^{n} X_j f_j q_{ji}$$

of absolute abundances X_i, because

$$X = \sum_{j=0}^{n} X_j = e^{\psi} \quad \text{and thus} \quad x_i = X_i/X$$

- The total population size X grows exponentially at rate ϕ,

$$\dot{X} = \dot{\psi}e^{\psi} = \phi X$$

Solving the quasispecies equation (2/2)

- With the mutation-selection matrix $W = (w_{ij}) = (f_j q_{ji})$ the quasispecies equation can be written in vector notation as

$$\dot{x} = xW - \phi x$$

- The equilibrium x^* is the solution of the eigenvalue problem

$$xW = \phi x$$

- The average fitness ϕ is the largest eigenvalue of the matrix W, and x^* is the corresponding (normalized) left eigenvector.

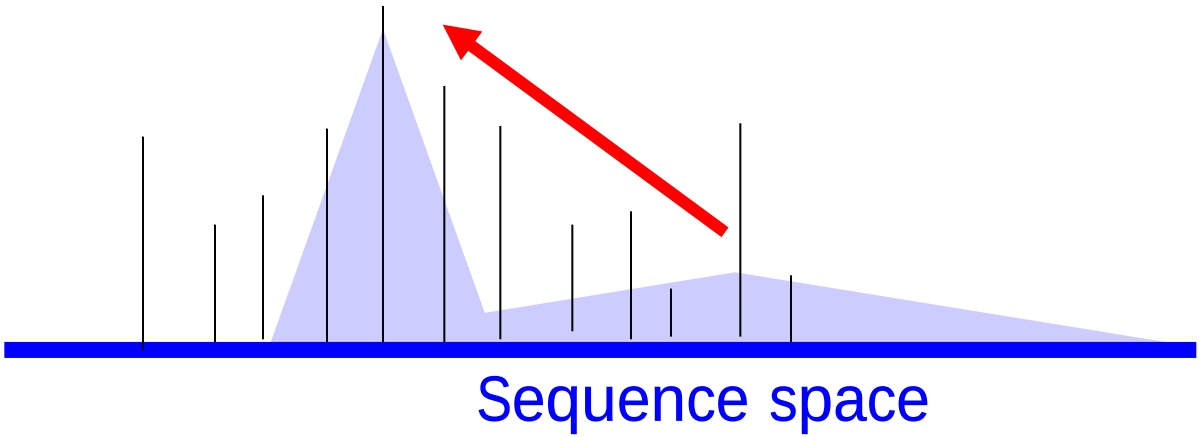
Binary sequences

- Consider binary sequences of fixed length L.
- We assume that only point mutations occur in this population (ignoring insertion, deletions, recombination), that they are independent and occur at the same rate u.
- $q_{ij} = u^{d(i,j)}(1-u)^{L-d(i,j)}, \quad d(i,j)$ Hamming distance
- For HIV, $L = 10^4$, $u = 3 \cdot 10^{-5}$. Thus:
  - The whole genome is copied error-free with prob. $(1-u)^L = 0.74$.
  - The copy has one (arbitrary) error with prob. $Lu(1-u)^{L-1} = 0.22$.
  - The copy has a particular error with prob. $u(1-u)^{L-1} = 2.2 \cdot 10^{-5}$.
  - With 10^9 new viruses per day, each one-point mutant occurs 22,000 times each day!

0000110011000110
0000110011000110
0000110011100110
0000110011000110
0000110011100110
0000100011000110
0100110011001110
0000110011000110
1000110011000010
0000110111000110
0000110011000110
0001110011000100
0000110011000110
….

Adaptation

- Adaptation of a population is localization in sequence space at a local maximum of the fitness landscape.



Adaptation of a quasispecies

- The equilibrium x^* of the quasispecies equation is a *mutation-selection balance* of the population: selection drives individuals towards a local peak of the fitness landscape, while mutation generates types of lower fitness.
- Localization is no longer possible, if mutation removes the fittest types faster than they are selected.
- There should be a necessary condition on the mutation rate and the selection strength for adaptation to occur.
- This condition is called the *error threshold*, and it exists for many, but not all, fitness landscapes.
- We look at the simplest possible, interesting case.

A simplified quasispecies equation

We consider binary sequences of lenght L. Type 0, the sequence $0 \ldots 0$ called wild type or master sequence, has fitness $f_0 > 1$. All other types have fitness 1.

The wild type is copied without error with probability $q = (1-u)^L$. We denote by x_0 the frequency of the wild type and by x_1 the sum of the frequencies of all other types. Ignoring back mutations to the wild type, the quasispecies equation becomes

$$x_0' = x_0(f_0 q - \phi) \qquad x_1' = x_0 f_0 (1-q) + x_1 - \phi x_1$$

or equivalently,

$$x_0' = x_0 \left[f_0 q - 1 - x_0(f_0 - 1) \right]$$

The equilibrium is

$$x_0^* = \begin{cases} \frac{f_0 q - 1}{f_0 - 1}, & \text{if } f_0 q < 1 \\ 0, & \text{if } f_0 q > 1 \end{cases}$$

Calculating the error threshold

The error threshold is given by

$$f_0 q > 1$$

For small mutation rates, $u \ll 1$, and moderate fitness values, $\log f_0 \approx 1$, we have

$$\log(f_0 q) = 1 + L \log(1-u) \approx 1 - Lu > \log 1 = 0$$

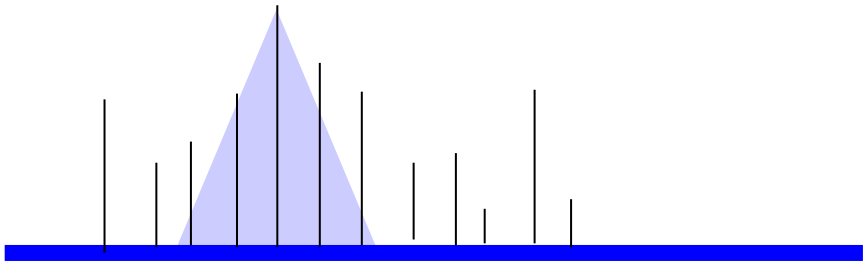
or

$$uL < 1$$

The critical mutation rate is $u_c = 1/L$. We call uL the genomic mutation rate. It is the expected total number of mutations per replication.

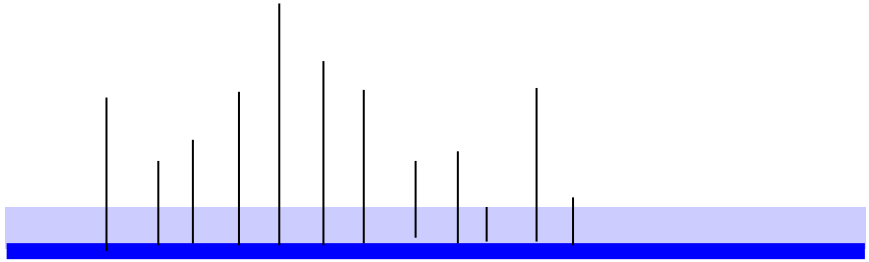
Most organisms for which both u and L are known live below the error threshold.
Many RNA viruses seem to live near the error threshold.
For HIV, we find $uL = 3 \cdot 10^{-5} \times 10^4 = 0.3$.

The error threshold



Sequence space

$uL < 1$

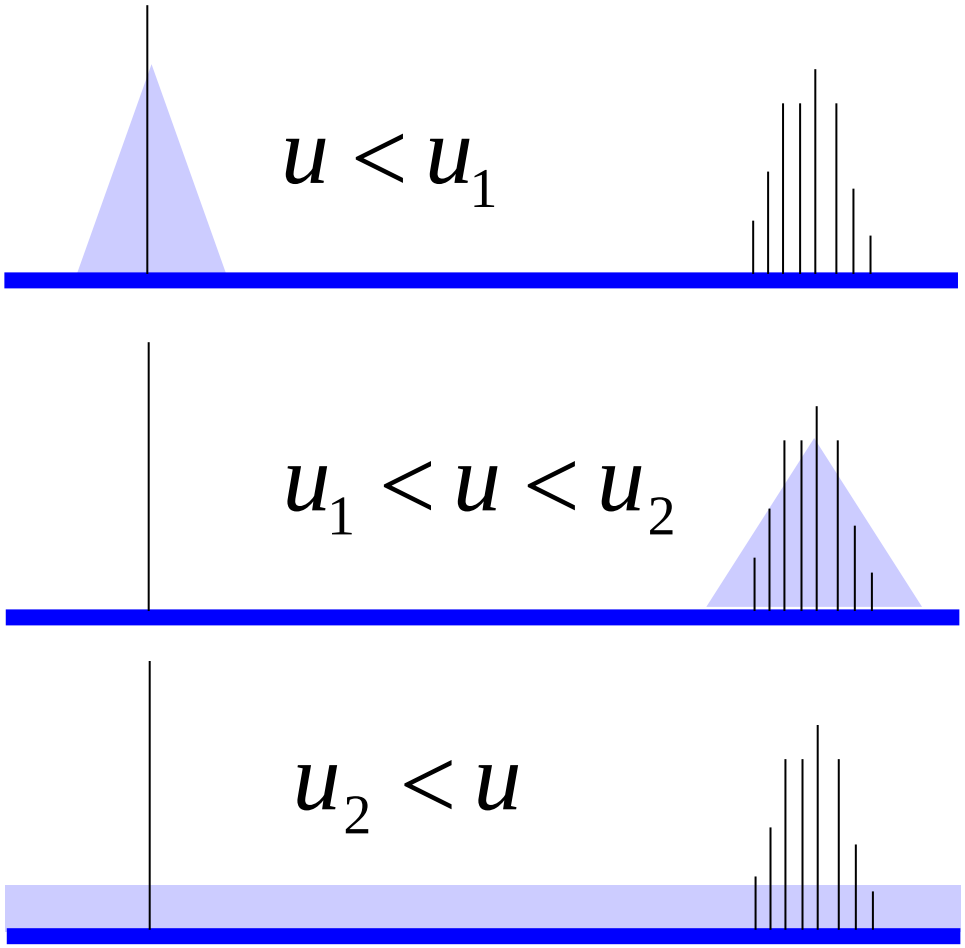


Sequence space

$uL > 1$

- The case uL > 1, is known as *mutational meltdown*.
- Some antiretroviral drugs seem to work by increasing u.

Selection of the quasispecies



- Selection can favor low and wide peaks over high and narrow peaks.
- For any given mutation rate, the equilibrium distribution (the quasispecies) with maximum average fitness is selected.

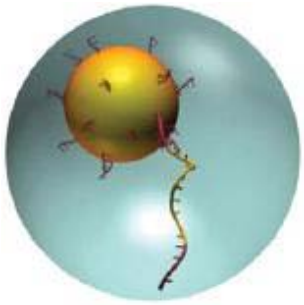
Measuring viral diversity

- How can we measure viral genetic diversity?

- Sanger sequencing (not very sensitive).
- Clonal Sanger sequencing (labor-intensive).
- Allele-specific PCR (restricted to few known mutations).

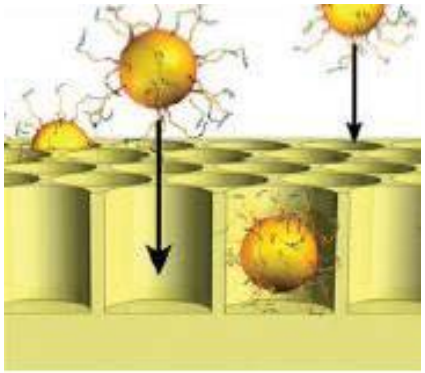
- Next-generation sequencing (NGS) technologies
  - 454/Roche
  - Solexa/Illumina
  - ABI SOLID
  - …

Pyrosequencing (454/Roche)

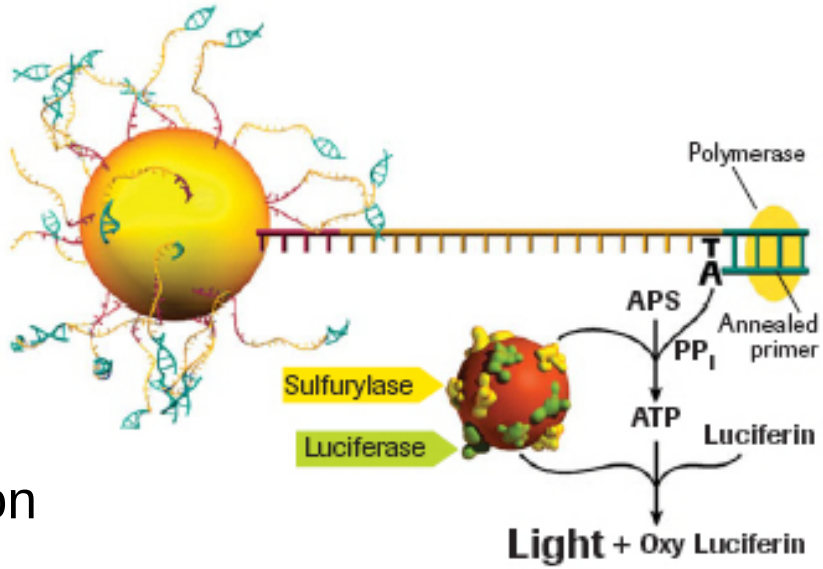


DNA fragments are attached to beads

fragments are amplified

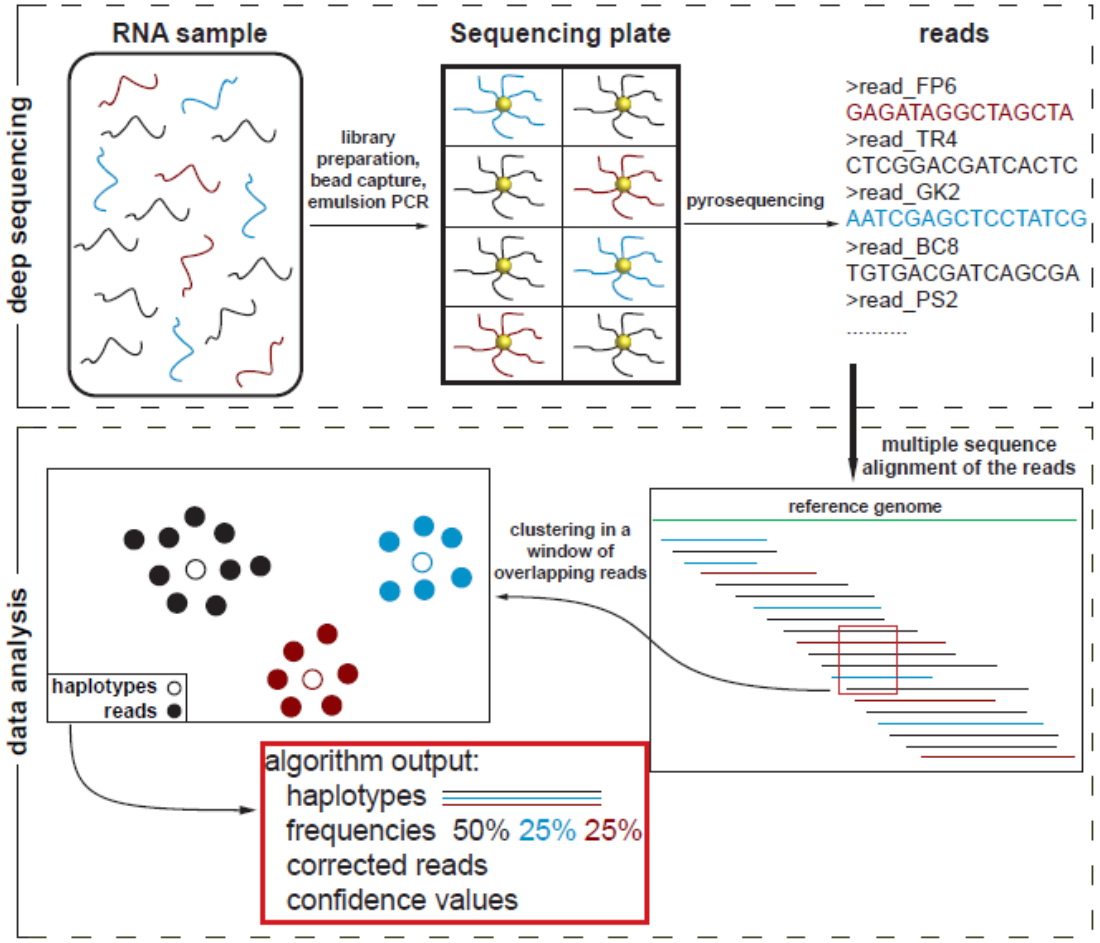


single beads in ~100.000 wells



DNA is read by detecting light emission associated with base incorporation

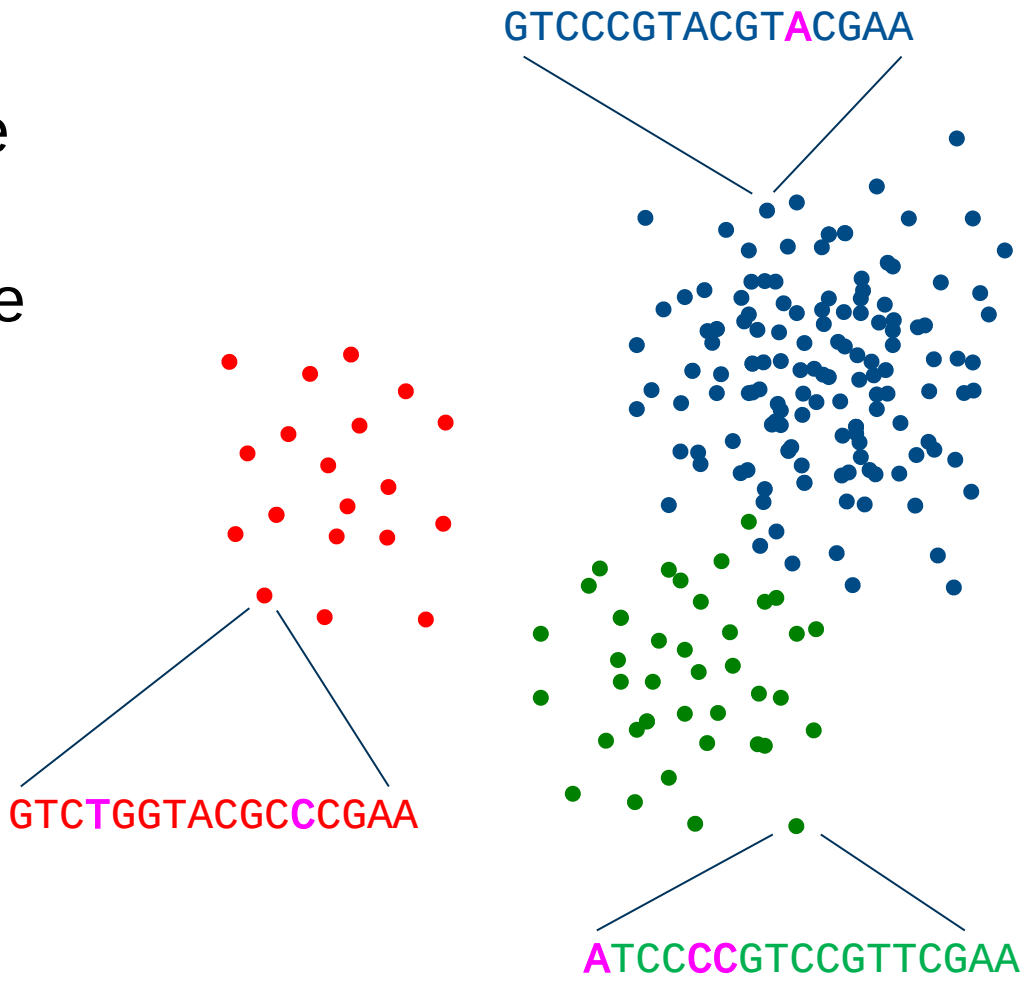
Sequencing a genetically diverse sample



Local haplotype reconstruction

Probabilistic clustering

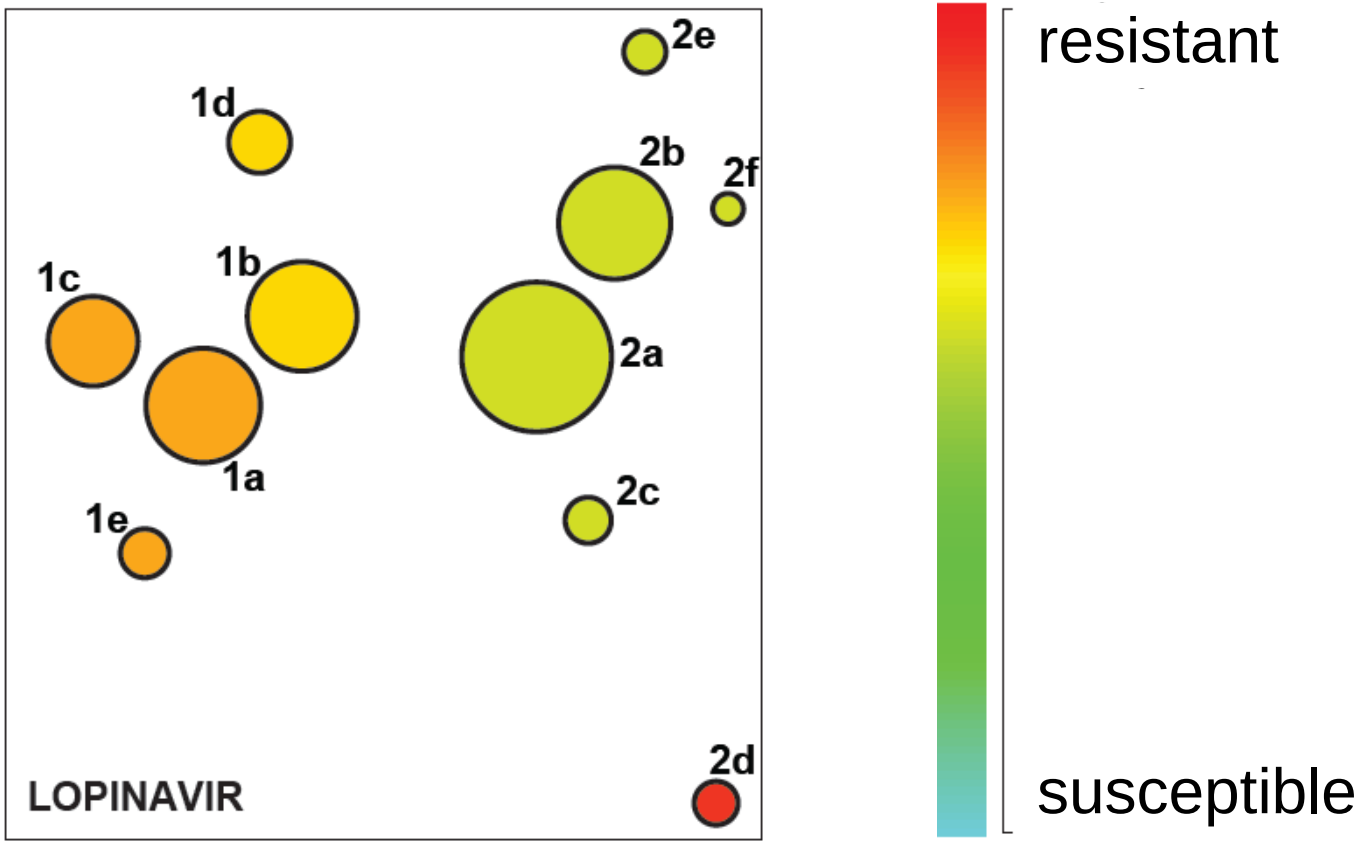
- Cluster = all reads originating from one haplotype
- Cluster center = true haplotype



Protease (amino acid) quasispecies distribution

Frequency		
Sanger	**Pyro**	**Mutations**
52.3	19.3	M46I, I54V, G73I, I84V, L90M
12.3	19.0	M46I, I54V, G73S, I84V, L90M
9.2	9.4	M46I
6.2	5.6	
4.6	7.1	M46I, I54V, G73S, L90M
4.6	1.9	M46I, I54V, G73I, L90M
3.1	5.8	M46I, G73I, I84V, L90M
3.1	0.0	L33F, M46I, I54V, G73S, I84V, L90M
1.5	1.9	M46I, L90M
1.5	0.0	L33F, M46I, I54V, G73I, I84V, L90M
1.5	0.0	M46I, I54V, G73N, I84V, L90M
0.0	4.9	M46I, G73I
0.0	4.7	M46I, I84V
0.0	4.0	M46I, G73S, I84V, L90M
0.0	3.1	M46I, I54V, G73S, V82I, I84V, L90M
0.0	2.9	M46I, I54V, G73I, I84V
0.0	2.9	M46I, I50V, I54V, G73I, I84V, L90M
0.0	2.0	I84V
0.0	1.4	I54V, G73I, I84V, L90M
0.0	1.2	M46I, I50V, I54V, G73S, L90M
0.0	1.1	M46I, I54V
0.0	1.0	M46I, I50V, I84V, L90M
0.0	0.5	M46I, I54V, G73I
0.0	0.3	G73S, I84V, L90M

Predicted drug resistance spectrum of viral quasispecies obtained from two clinical samples



Summary

- Evolution can be thought of as a time trajectory in sequence space.
- The quasispecies equation describes a mutation-selection balance.
- It predicts an error threshold.
- The quasispecies distribution can be found by solving a linear eigenvalue problem.
- Quasispecies theory has been applied to HIV.
- Next-generation sequencing technologies allow for (locally) estimating the population structure in genetically heterogeneous samples.