Two problems in microbial genome evolution: diversification fronts and the evolution of compositional bias.

Kalin Vetsigian & Nigel Goldenfeld

K. Vetsigian and N. Goldenfeld Global divergence of microbial genome sequences mediated by propagating fronts, PNAS 102, 7332-7337 (2005).

K. Vetsigian and N. Goldenfeld. Genome rhetoric and the emergence of compositional bias. *Proc. Natl. Acad. Sci. (USA)* **106**, 215-220 (2009)

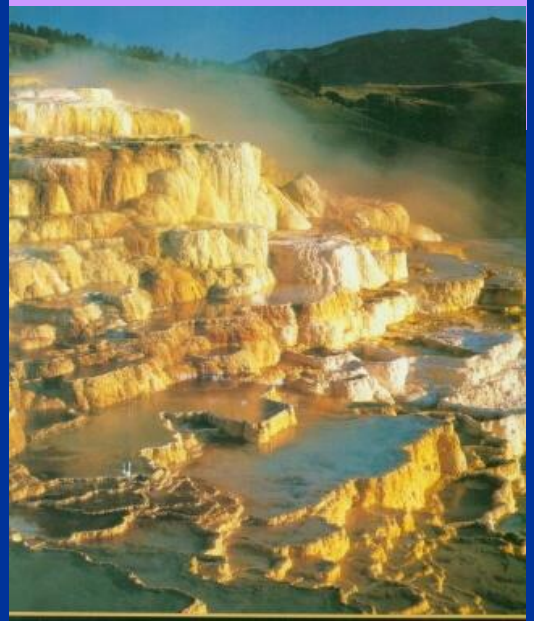
Overview

Exploring dynamical systems from the genome to the environment

Organisms and their environment – genomic ecology

Ecological genomics of diverse environments

Geothermal hot sprints
Gastrointestinal microbiomes



Organization at the genome level – emergent states of life

Horizontal gene transfer
Microbial "speciation"

Evolution of cells and hierarchical order

Origin of genetic code

Systems biology of Microbes and Viruses

Lysogeny
Prophage induction

Transposon dynamics
Genome rearrangements

The failure of reductionism

- Steven Jay Gould, New York Times, Feb 19, 2001
  - “Homo sapiens possesses between 30,000 and 40,000 genes... In other words, our bodies develop under the directing influence of only half again as many genes as the tiny roundworm ....”
  - “The collapse of the doctrine of one gene for one protein, and one direction of causal flow from basic codes to elaborate totality, marks the failure of reductionism for the complex system that we call biology.”
  - “First, the key to complexity is not more genes, but more combinations and interactions generated by fewer units of code — and many of these interactions (as emergent properties, to use the technical jargon) must be explained at the level of their appearance, for they cannot be predicted from the separate underlying parts alone.”

Outline

- Part 1: global diversification of microbial genomes through the interplay between recombination and point mutation

- Part 2: compositional bias in microbial genomes. How feedback between resource allocation and template-directed synthesis leads to multistability of evolved microbial genomes

Part 1: Diversification fronts

Speciation without selection

Global sequence divergence in microbes and (now) eukaryotes

K. Vetsigian and N. Goldenfeld. Global divergence of microbial genome sequences mediated by propagating fronts. *Proc. Natl. Acad. Sci. (USA)*., 102, 7332–7337 (2005)

Bacterial evolution in the presence of genetic exchange

- Bacteria are not asexual; yet they are not sexual…
- Their evolution is communal
- No clear framework. Many open questions:
  - Are there bacterial species?
  - Is their history captured by a tree of life?
  - How to classify different modes of evolution?

Mechanism for global diversification even in the absence of selection and ecological barriers to exchange

Outline

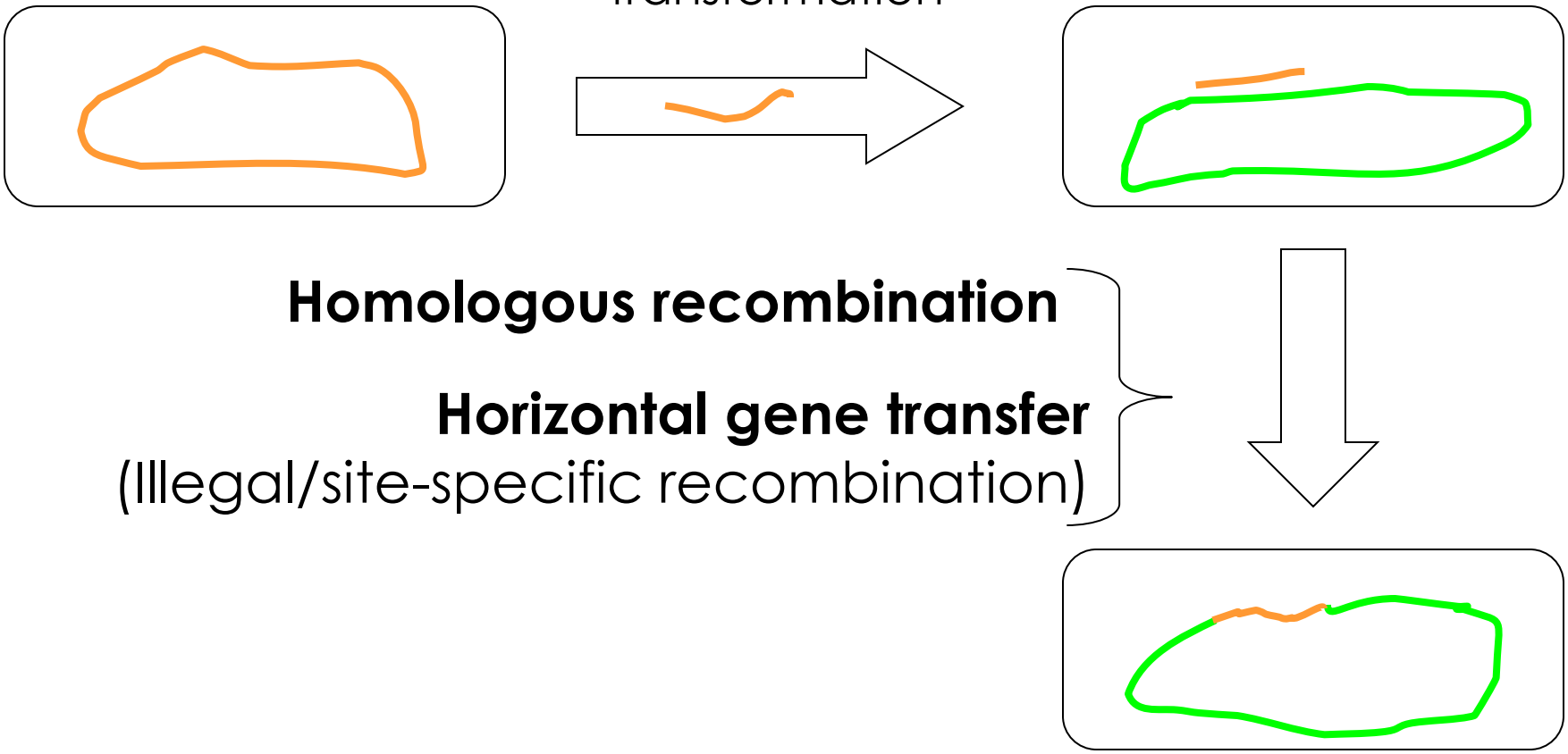
- Homologous recombination
- Model interactions
- Emergent property of communal evolution: **propagation of diversification fronts**
- **Classification of bacteria** based on properties relevant for their communal evolution
- Experimental evidence
- Biological consequences
  - Mechanism for speciation
  - Dynamical barrier to rearrangements and HGT

Genetic exchange

donor cell

- conjugation
- transduction
- transformation

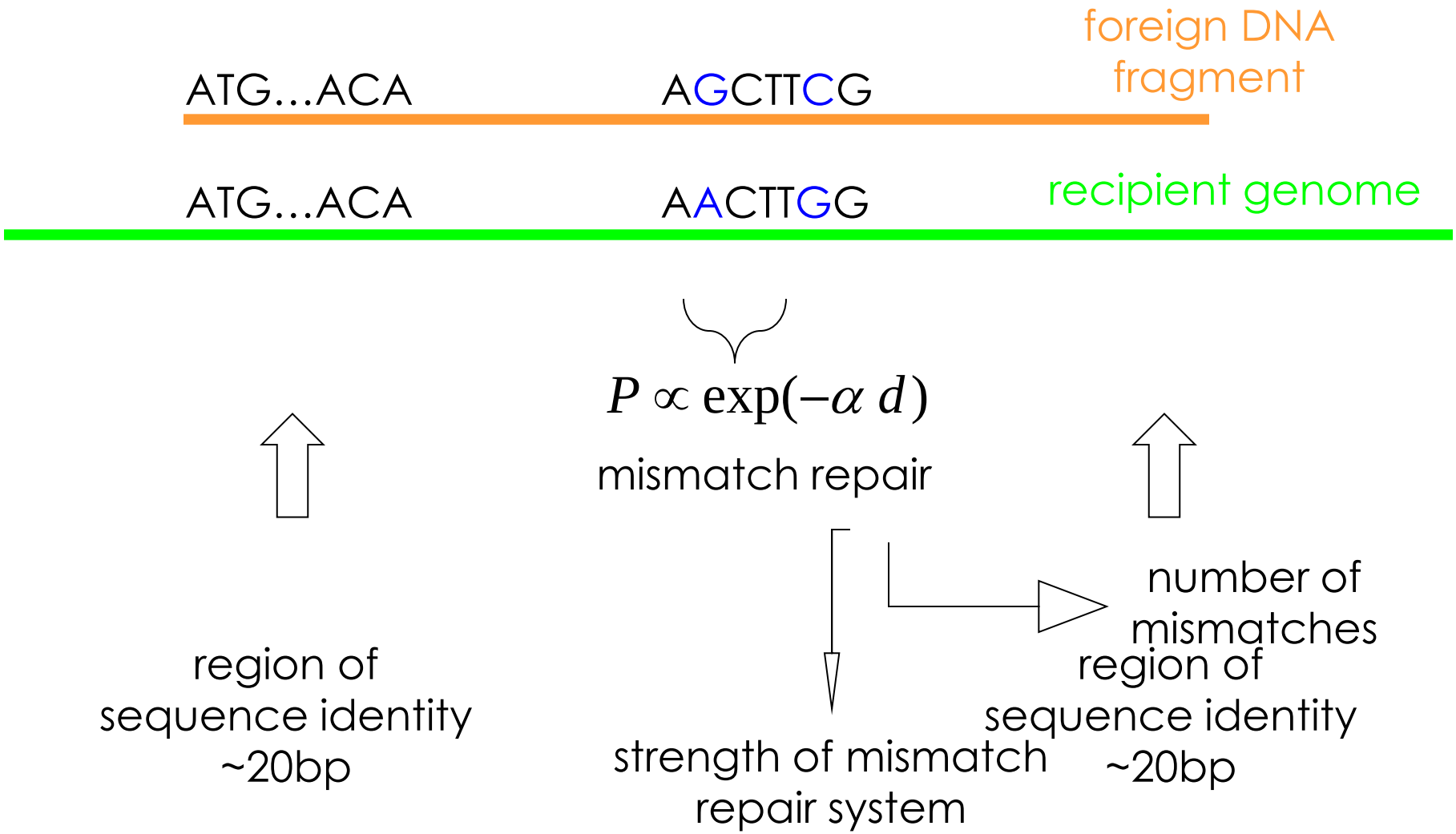
recipient cell



Homologous recombination

Horizontal gene transfer
(Illegal/site-specific recombination)

Homologous recombination



Model for genome evolution

- Homologous recombination
- Point mutations

Frequent (almost neutral)

model explicitly

- Genome rearrangements
- Horizontal gene transfer

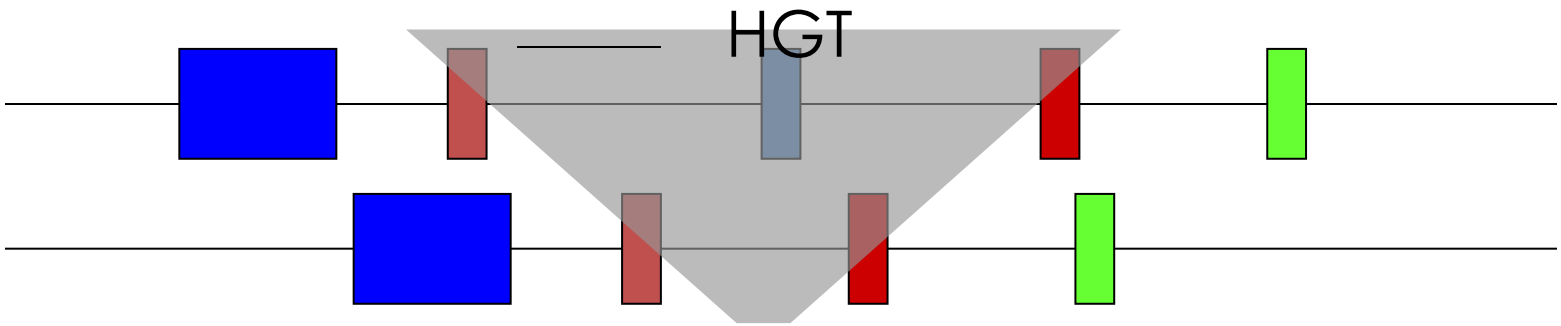
(apparently) occasional

Dynamically suppressed

incorporate through initial conditions

HGT inhibits recombination

- Observation: HGT can inhibit recombination locally†



- Suggests: Global genetic isolation requires accumulation of hundreds of HGT islands (Lawrence 2002)
  - Speciation is a gradual process

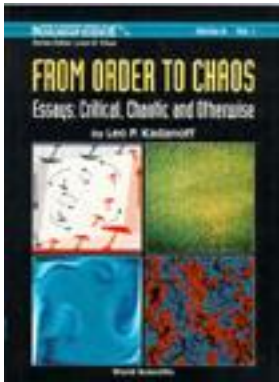
Our work: Modelling the dynamical effects of HGT changes conclusion

Modeling the interplay between recombination and point mutation

Purpose of computer simulation

- At that time there was a great national push toward understanding the dynamics of urban development. Jay Forrester of MIT had developed a computer model of urban change, which took a very simplified view of a urban society … and then used the output of that model to prescribe social policy. I did not like the policy prescribed.
- Consequently, I set out to use the modeling tools that Forrester had developed to reach conclusions which were more to my liking. … Our first result was that while not changing the model at all, we could reach opposite conclusions from that of the Forrester group.
- We went on to build other models which more accurately recorded our own prejudices and points of view. But after a while, the point we had made began to sink in. If these models really represented little more than we could say in words, why not leave out the computer?
- The construction of this sort of computer model seemed to be a rather pointless endeavor. For this reason, and others, I moved away from urban studies.

Leo P. Kadanoff 1993



Purpose of computer simulation

- Moral of this story:
  - Either compute to get numerical information that verbal arguments cannot address
  - Or compute to find emergent phenomena: an outcome of the dynamics that is not mandatory, and usually collective
    - Example: the Hamiltonian of a fluid and a solid are identical, but only for low temperatures, can there be a non-zero shear modulus

Model of neutral evolution

- *N* circular genomes of length *L*; alphabet of size *n*

☞ ***Fitness does not depend on genome sequence***

- Point mutations
  - each letter changes with rate *m* – Poisson process

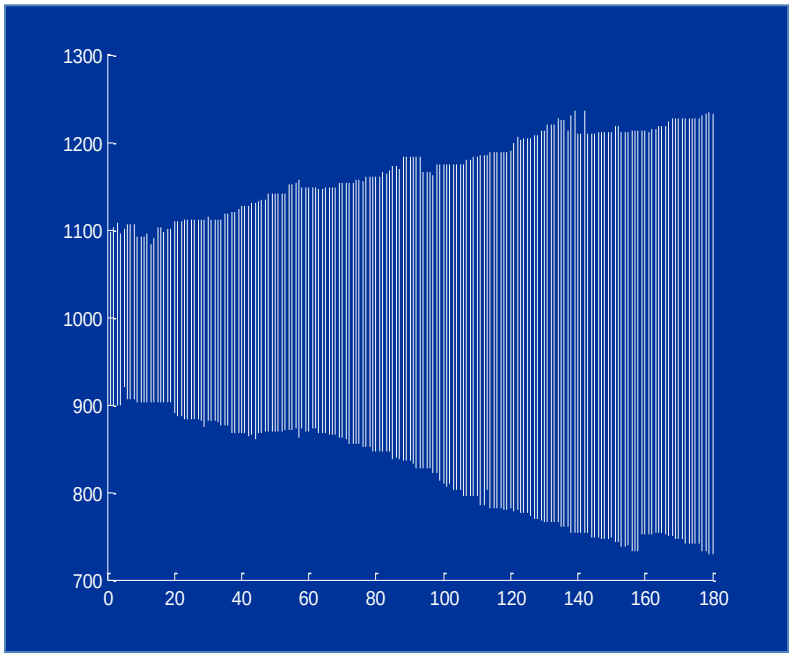
- Homologous recombination
  - fragment size fixed, *F*, or drawn from distribution
  - recombination attempts are made with rate *r*
  - require sequence identity of size *M* at both ends
  - probability of incorporation is $\exp(-\alpha\, d)$

Observations about the model

- Interplay between opposing tendencies
  - point mutations **increase** genome differences
  - recombinations **decrease** them
- Uniform and diverged phases
  - mutations **weak**: all genomes are similar
  - mutations **strong**: all genomes different (in a neutral way)
- The uniform phase is metastable
  - genome differences inhibit recombination

Simulate consequences of HGT

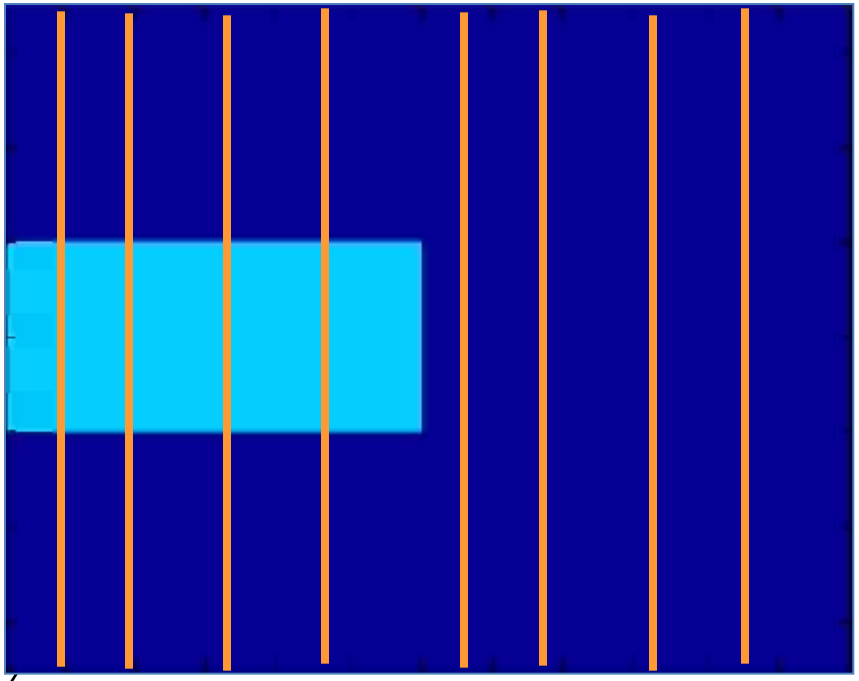
Growth of region of sequence diversity



Position along the chromosome

time

Initial condition



DNA sequence

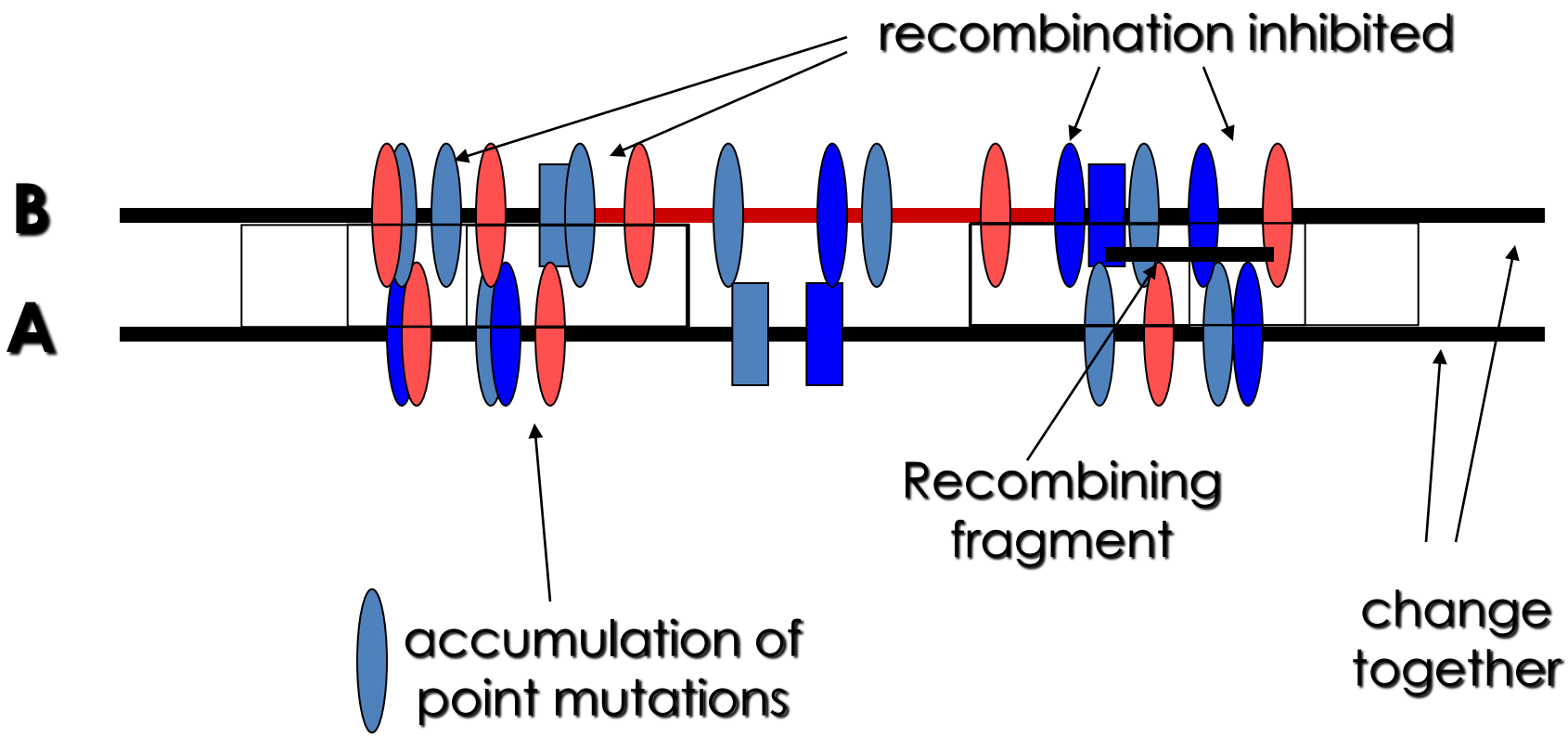
niche 1

niche 2

organisms

region where genomes from niche 1 and niche 2 are different

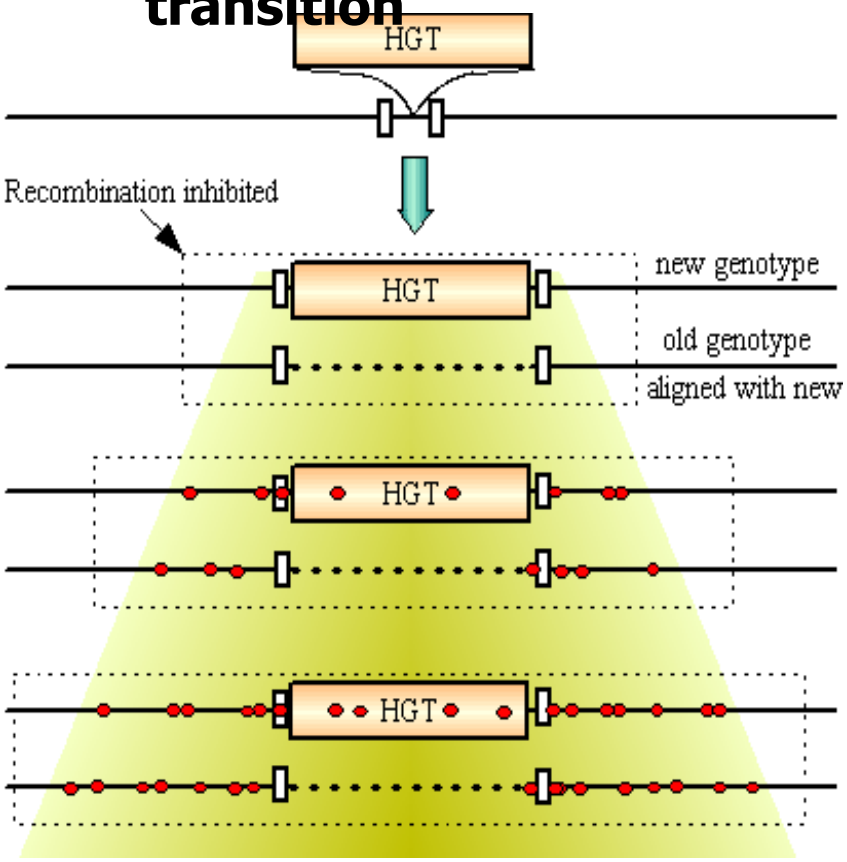
Front propagation mechanism

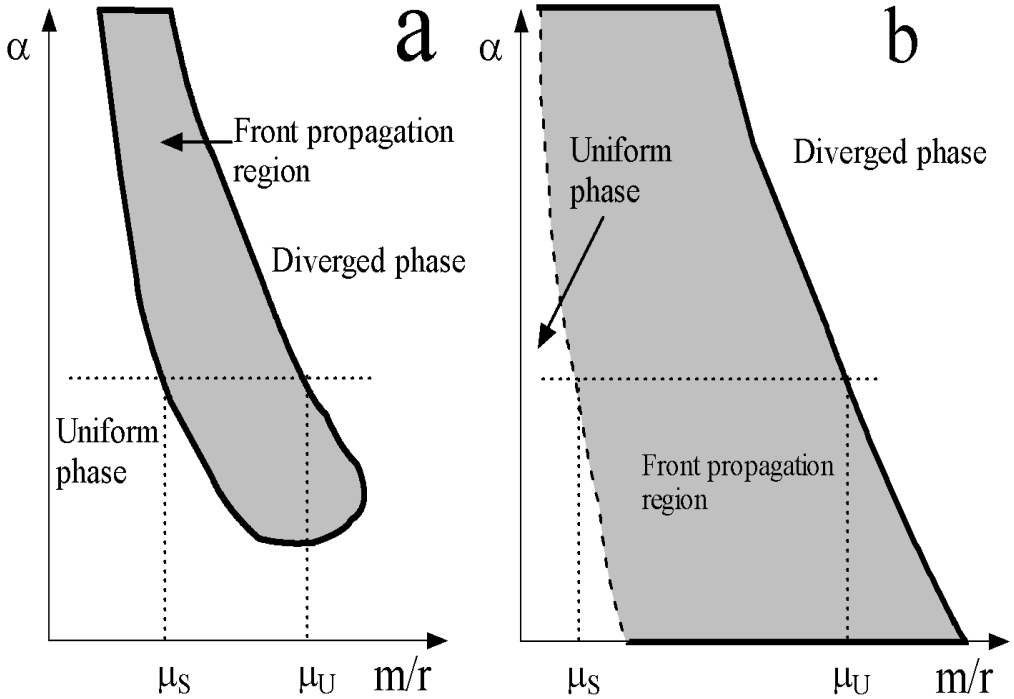


☞ **Region of sequence divergence expands over the entire genomes.**

Front propagation in microbial genomes

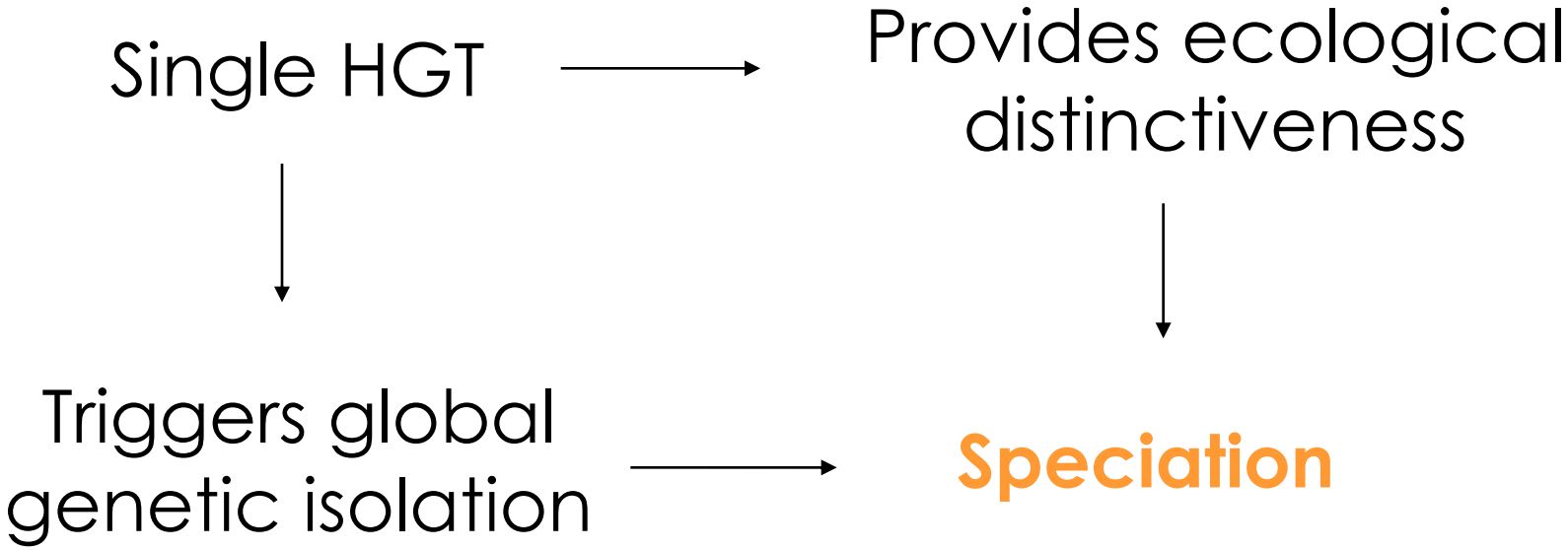
Competition between mutation and recombination → phase transition





Speciation in absence of selection is possible!
Mechanism is observed in *Bacillus* …

Mechanism for Speciation



Back to simulations:

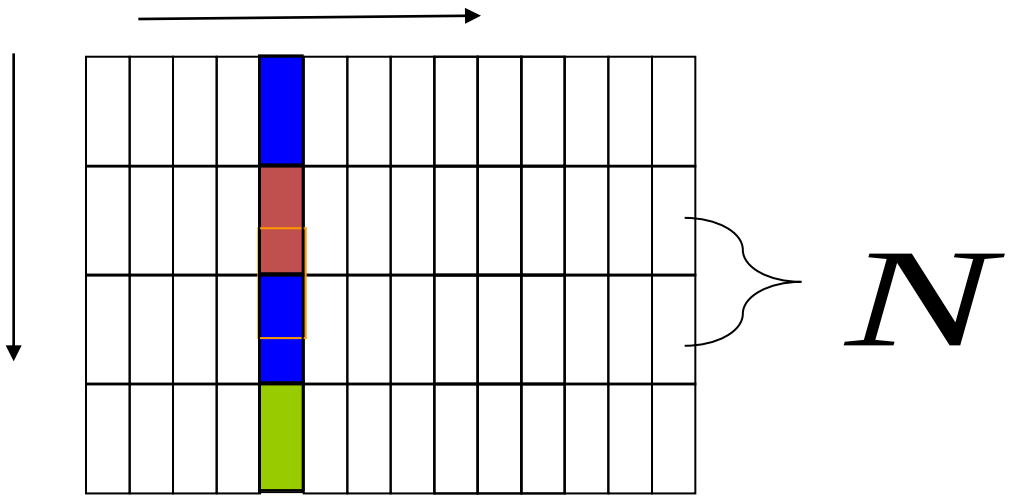
When do we see diversification fronts?

Construct phase diagram

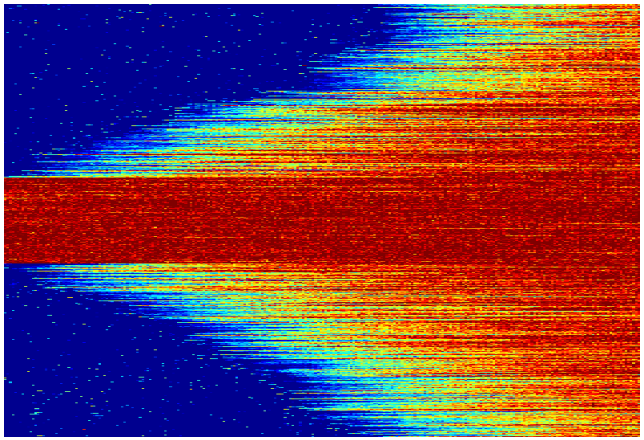
⇓

Classification of Bacteria

Order Parameter



Visualize evolution



time

Average difference across the population at position x

$$\psi(x) = \frac{n}{n-1}\frac{1}{N(N-1)}\sum_{i,j}\left(1-\delta_{A_{i,x},A_{j,x}}\right)$$

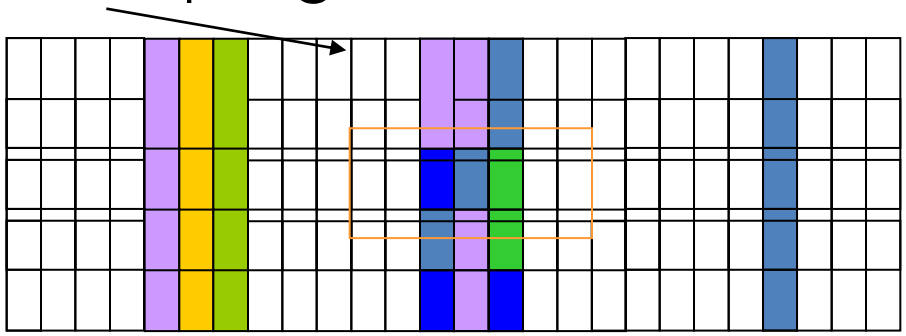
$\psi = 1$ diverged

$\psi = 0$ uniform

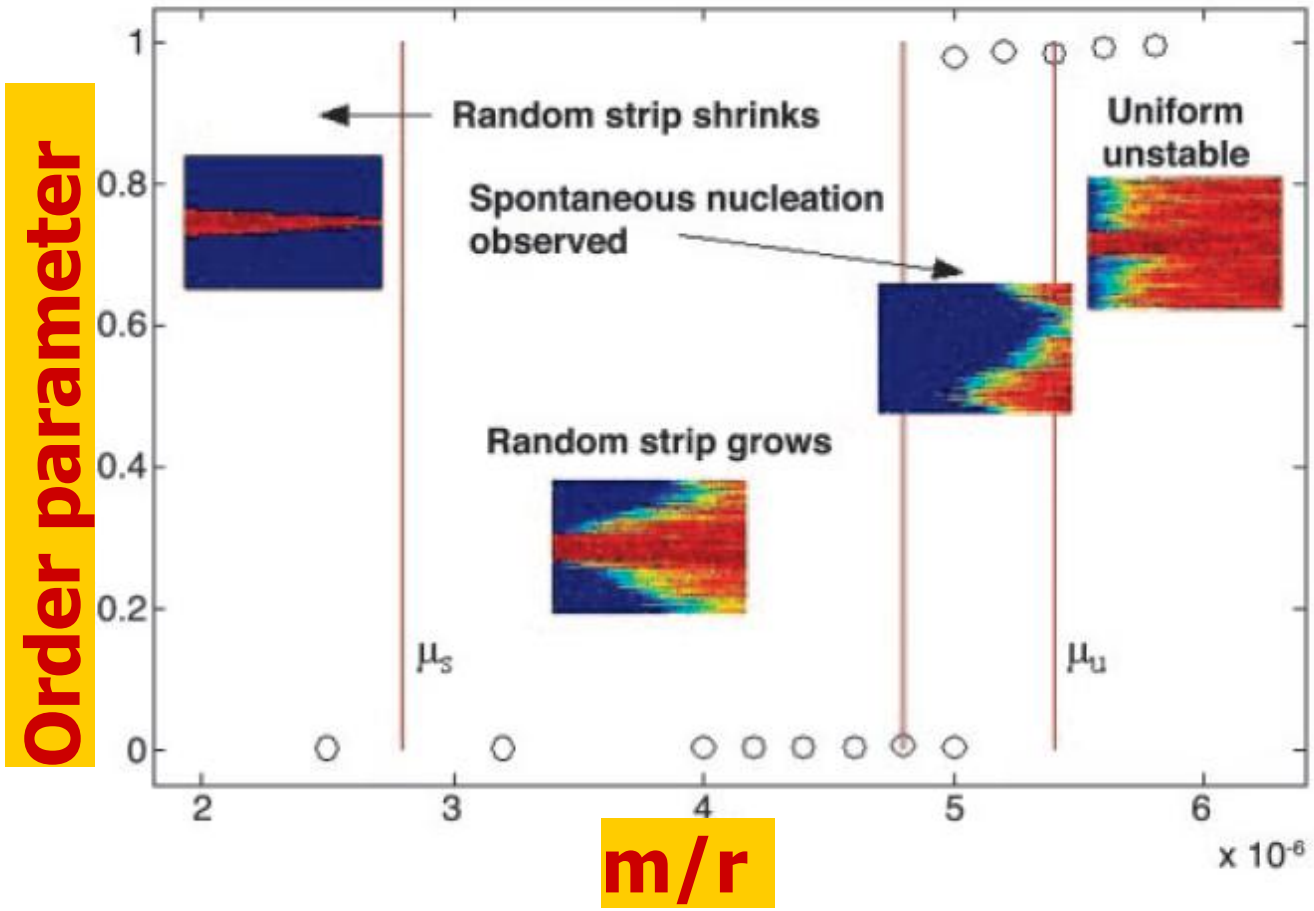
Different models and initial conditions

- Three models:
  - Class 0: Sequence identity not required
  - Class 1: Sequence identity at only one end required
  - Class 2: Sequence identity at both ends required

- Two initial conditions
  - Uniform – all genomes are the same
  - Random strip – genomes same except for a strip



Phase diagram



m=mutation rate

r=recombination rate

Fig. 3. Starting from a uniform state, the order parameter equilibrates to values close to 0 or 1 in model I with $\alpha = 0.4$, $F = 500$, $M = 10$, $L = 10{,}000$, $N = 20$, and $n = 2$, indicating the existence of distinct uniform and diverged phases. The inset figures depict the genome population for the indicated value of m/r, as a function of time. The vertical axis represents position along the genome and the color scale indicates the value of the order parameter (blue denoting uniform phase, red denoting diverged phase), whereas the horizontal axis is simulation time. For $\mu_s < \mu < \mu_u$, the random strip triggers a diversification front. For μ close to μ_u, spontaneous nucleation is possible.

Results for the one end model

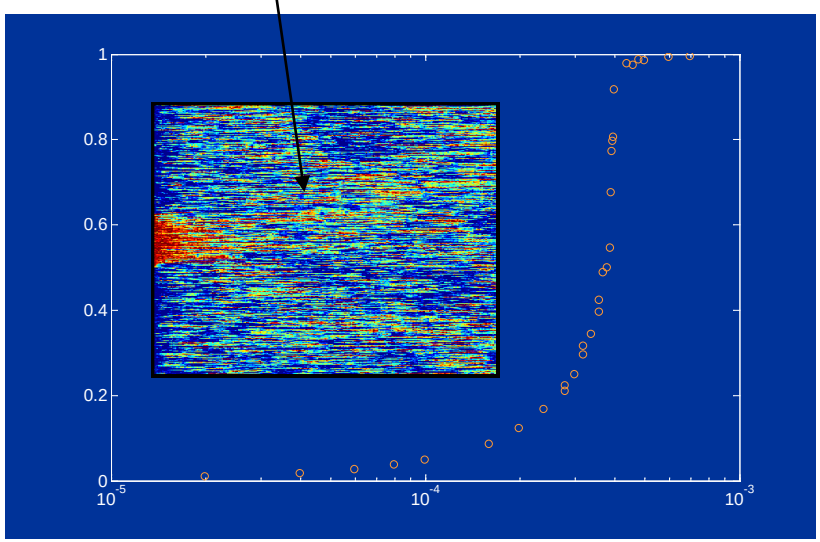
$\alpha = 0$ ← Strength of mismatch repair → $\alpha = 0.4$

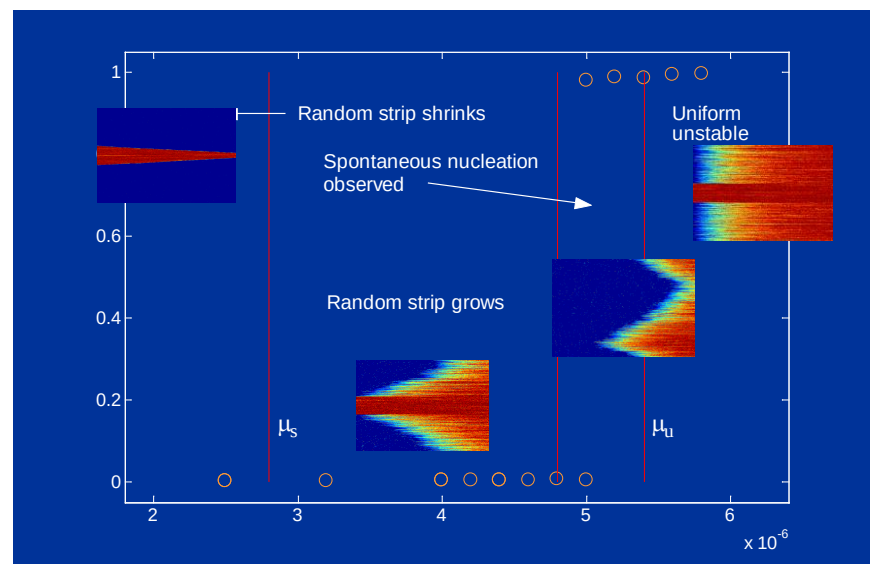
- No distinct phases
- Uniform and random strip relax to the same OP value
- **No front propagation**

- Two distinct phases
- Uniform and random strip relax to different OP value
- **Front propagation region**

Dissolving random strip

Equilibrated order parameter

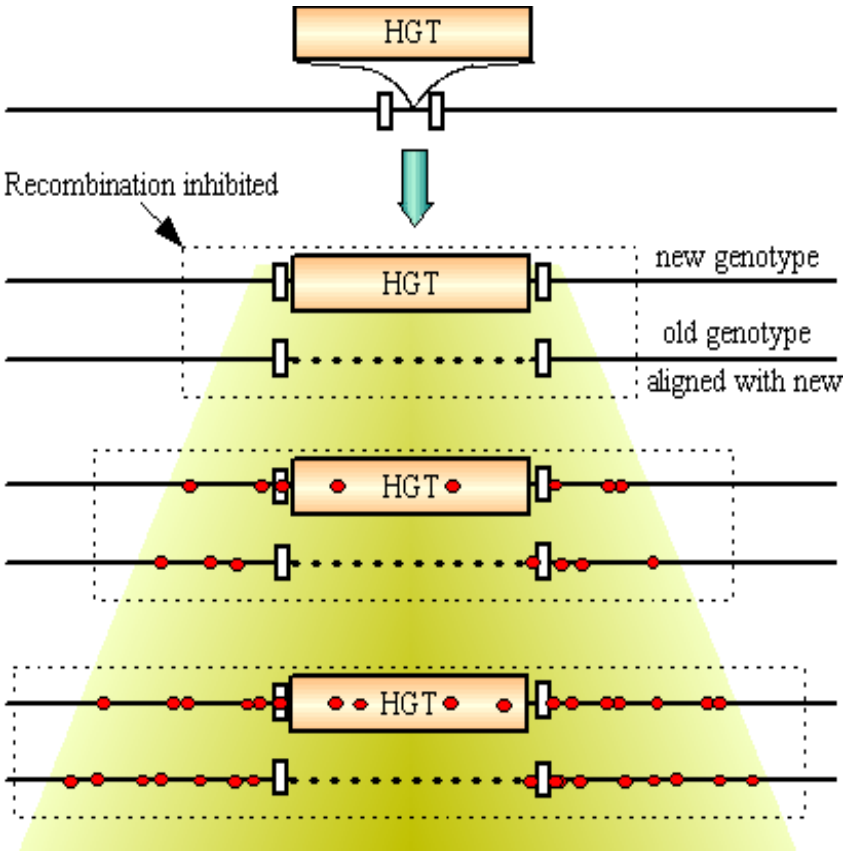


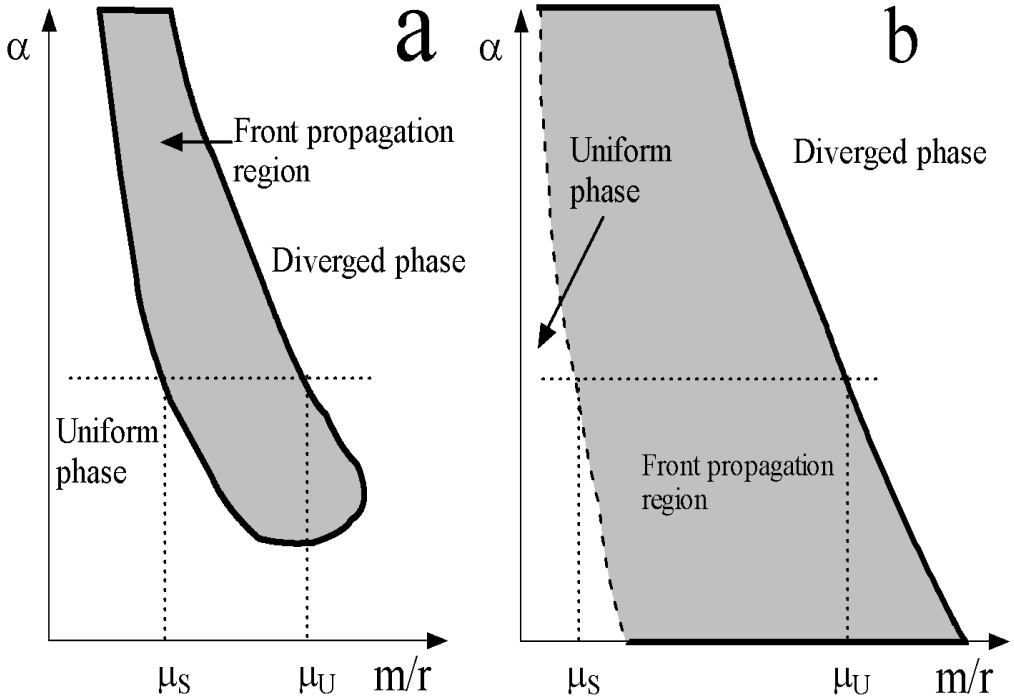


m/r ← Ratio of mut and rec rates → **m/r**

Front propagation in microbial genomes

Competition between mutation and recombination → phase transition





Speciation in absence of selection is possible!
Mechanism is observed in *Bacillus* …

Classify bacterial modes of evolution based on homologous recombination details

- Bacteria requiring <u>one end</u> sequence identity
  - Diversification fronts unlikely
  - Partial genetic isolation stable (no "species")
- Bacteria requiring <u>two end</u> sequence identity
  - Diversification fronts likely
  - Partial genetic isolation unstable – leads to global isolation (well defined "species")

Refine classification by learning more about the details of homologous recombination and examining their relevance

Current knowledge of homologous recombination mechanisms

- Universally true (so far): Probability of recombination exponentially decreases with sequence divergence
- Experimentally determined differences:
  - ***E. coli*** – one end sequence identity; strong mismatch repair
  - ***Bacillus*** – two end sequence identity; weak mismatch repair
  - ***Streptococcus pneumoniae*** – intermediate in strength mismatch repair, easily saturated

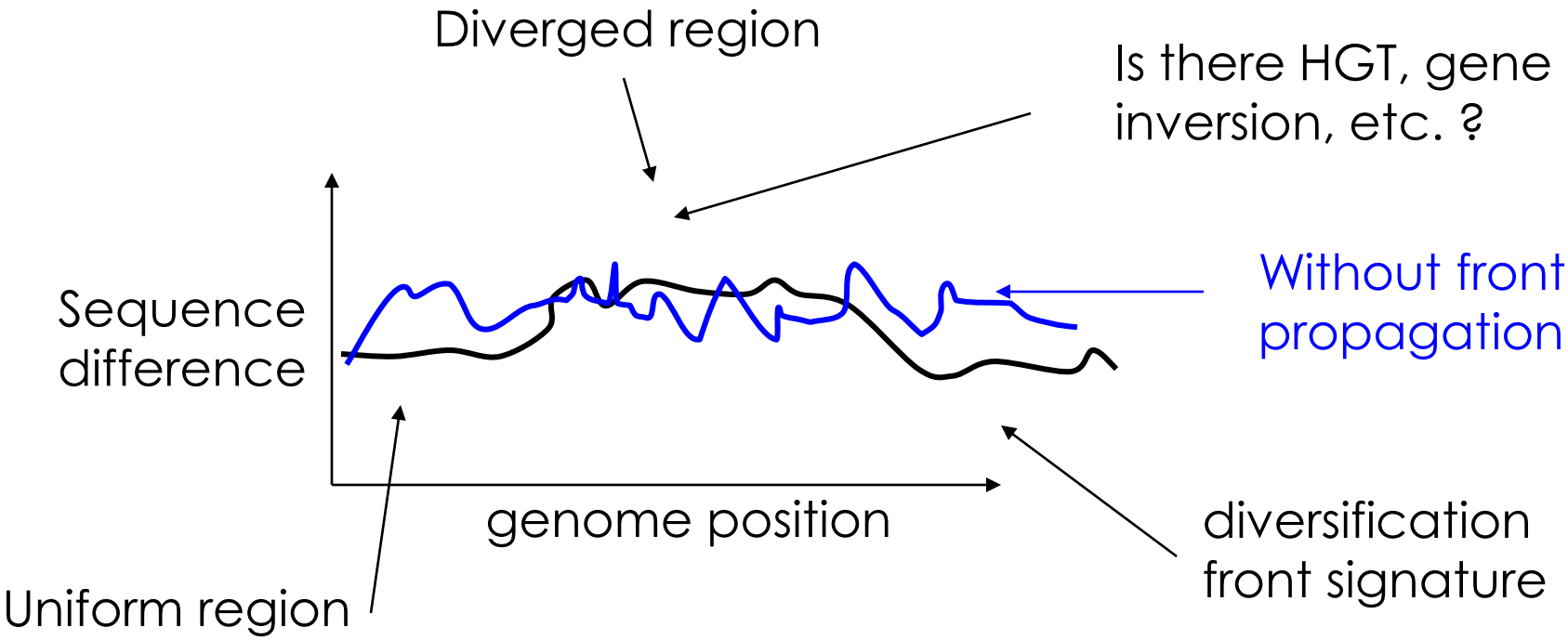
Comparative genomics: is there a signature of the diversification front?

Evidence from genome data

- Seek diversification fronts where expected by model
  - ***Bacillus cereus* group**
  - **Sequence identity at both ends**
  - Several closely related genomes sequenced, highly collinear
  - Genetic exchange likely: plasmids, sequence independent DNA uptake
- Compare with bacteria for which fronts are not expected
  - ***Buchnera aphidicola***
  - Intracellular symbiont
  - no RecA gene
- Diversification front signature
  - Step-like pattern of sequence difference along genomes
  - Fat-tailed distribution of maximal exact match lengths

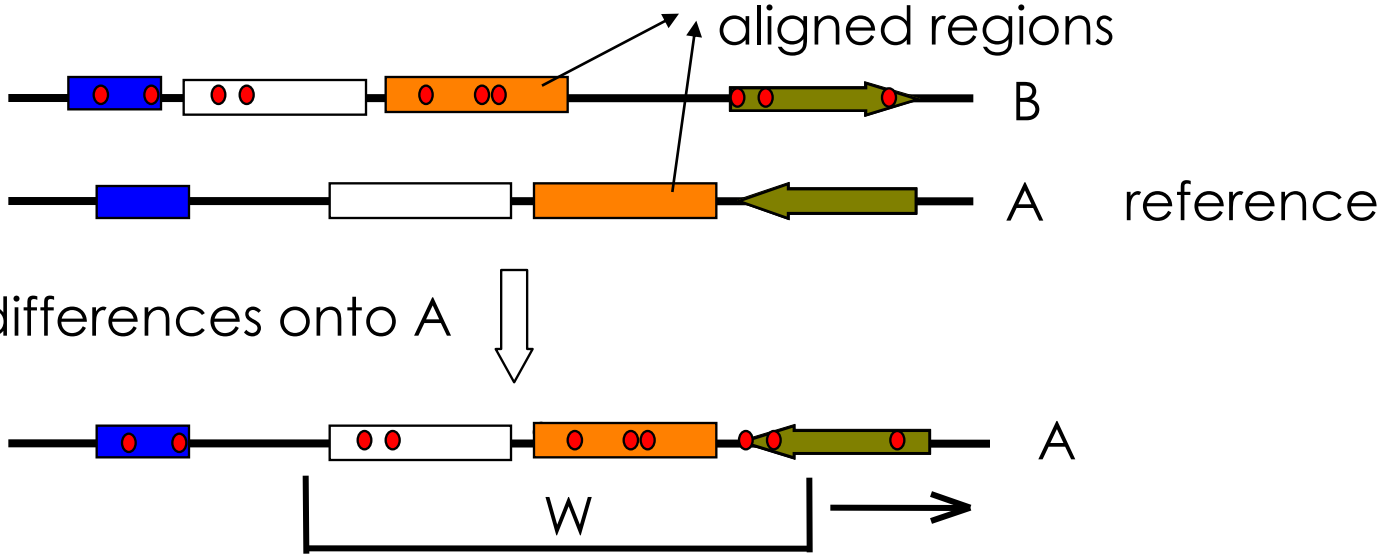
Step-like difference pattern

- Compare closely related bacteria with almost colinear genomes



Constructing difference profiles

- Global alignments – MUMMER (TIGR)
- Coarse graining



Map the differences onto A

Record the number of differences M and the number of positions in aligned regions K

if K > f W

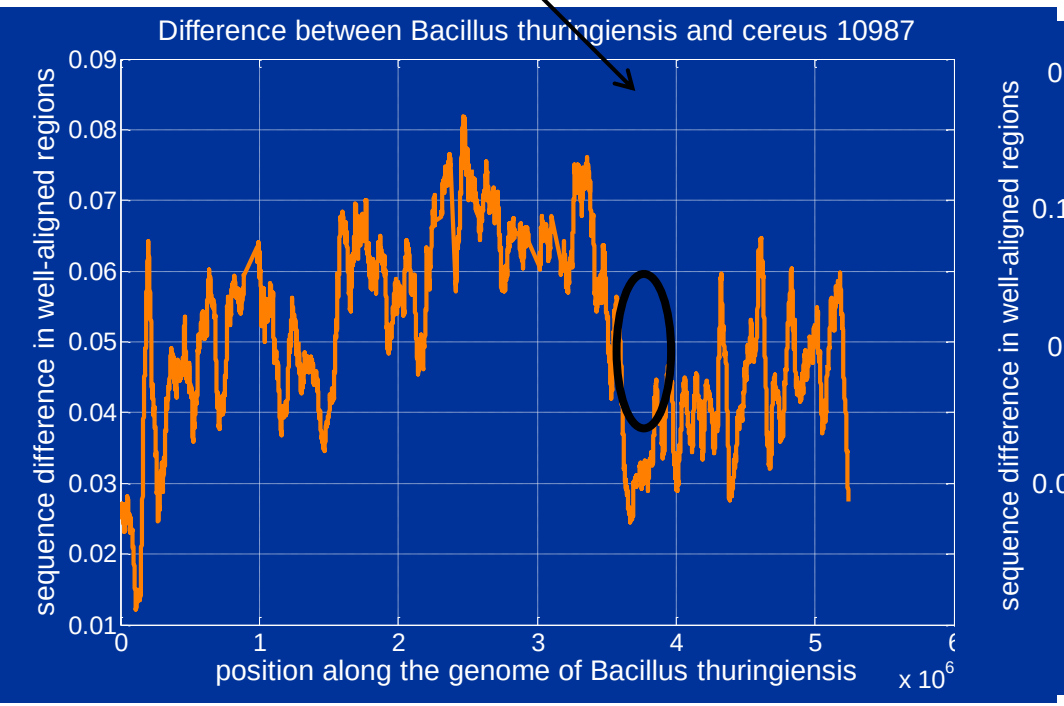
Divergence = M/K

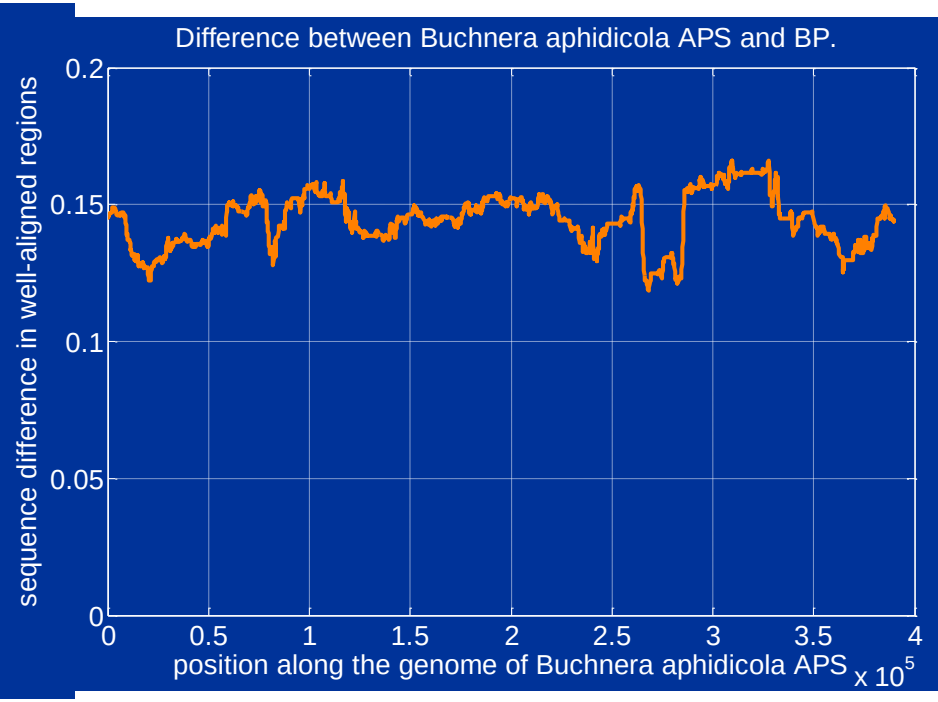
A step in the difference profile?

Bacillus cereus group

Buchnera aphidicola

candidate for diversification front





Alternative explanations

- Not involving homologous recombination:
  - varying point mutation rates (for whatever reason)
  - varying structural features. Example: protein density
- Varying homologous recombination rates
  - because of a diversification front
  - varying density of genome features that inhibit recombination

Discriminate by looking at:

1) **distribution of maximal exact match lengths**
2) **different components:**
   synonymous, non-synonymous, intergene
3) **partially randomized data sets**

Distribution of maximal exact match lengths

ACGTGCTACCGGACTCG—GTCAUGTAGCATGTA
AAGTGGTACCGGACTCGAGTCAUGTCGCATGTT

$L_1 = 3$ $L_1 = 11$ $L_1 = 7$ $L_1 = 6$

Poisson distributed random mutations

↓

Exponential distribution of maximal exact match lengths

↓

STD/mean ≈ 1

Statistically significant deviations of STD/mean from 1 indicate clustering

recombination

~~structural~~

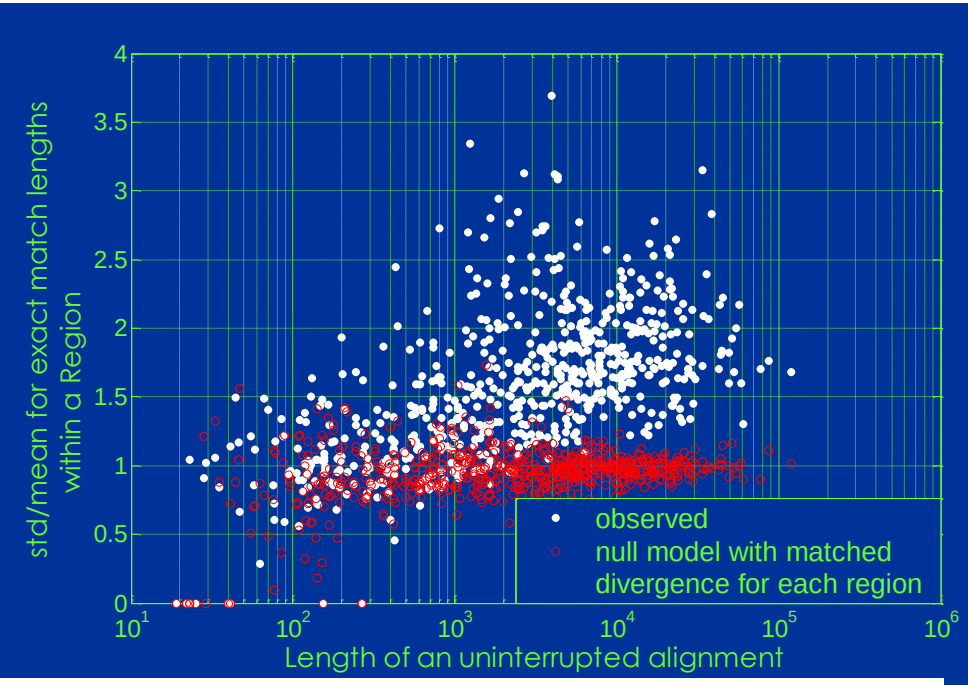
By looking at different components separately

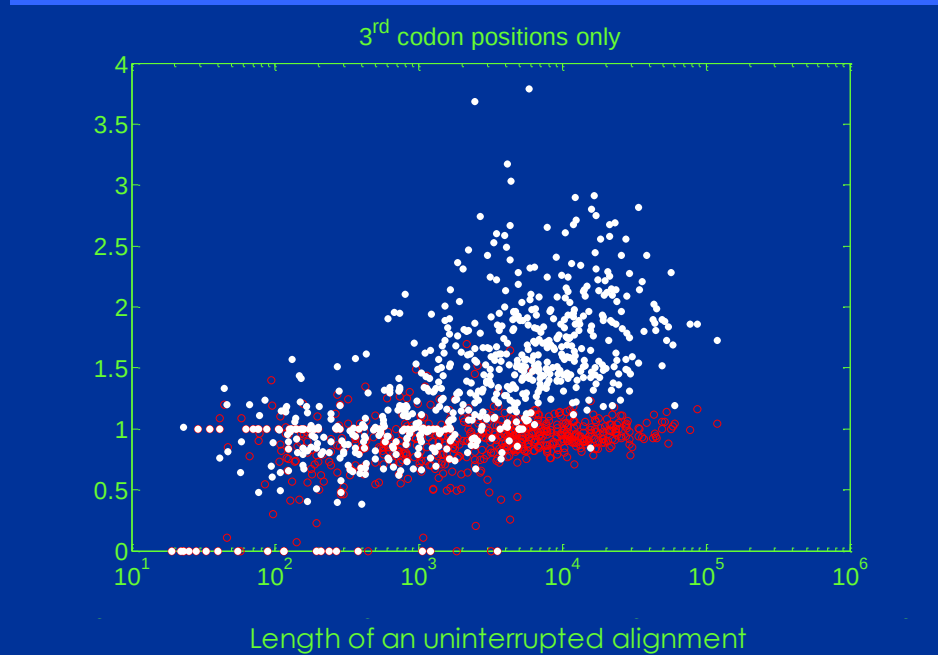
Distribution of exact matches

The positive correlation with alignment length is consistent with inhibition of recombination by adjacent non-aligned regions

<u>Clustering due to gene-intergene structure</u>?

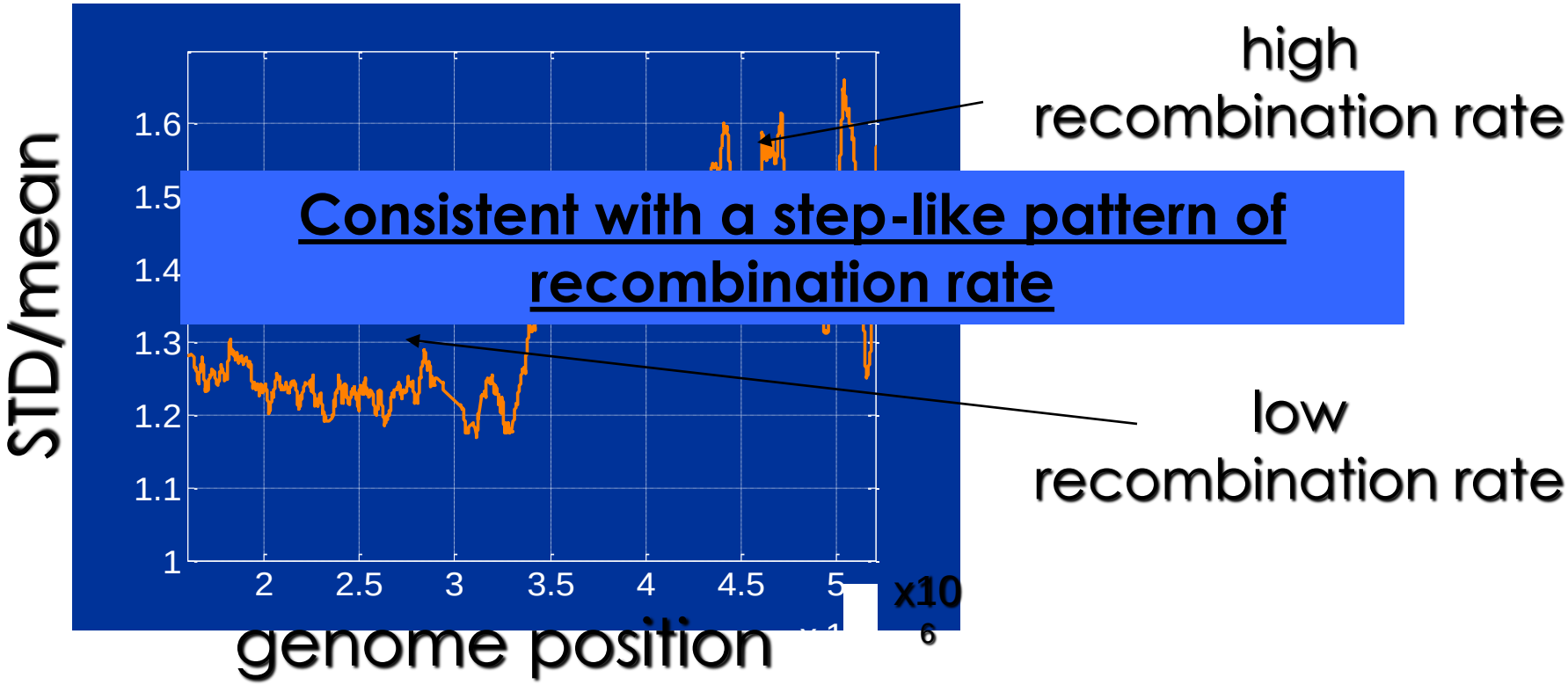
No, STD/mean deviation also present in the synonymous component.





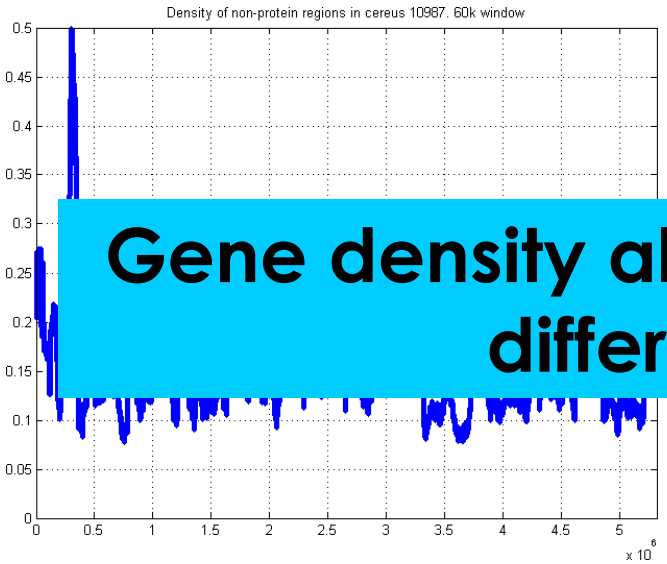
Step-like STD/mean profile for Bacillus cereus

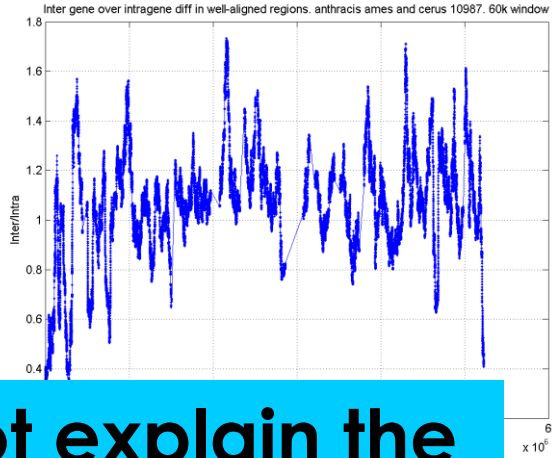
- The step pattern is not a result of a varying mutation rate
- Step-pattern disappears in a null model with matched divergence of each gene – not a result of distribution of genes with different evolutionary rates



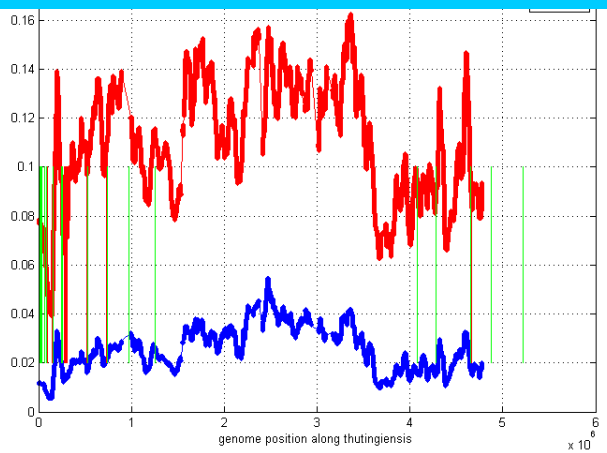
Gene density

- Hypothesis: Intergene regions have higher mutation rates **and the step pattern is a result of systematic variation of gene d**



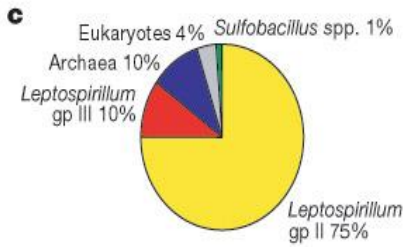


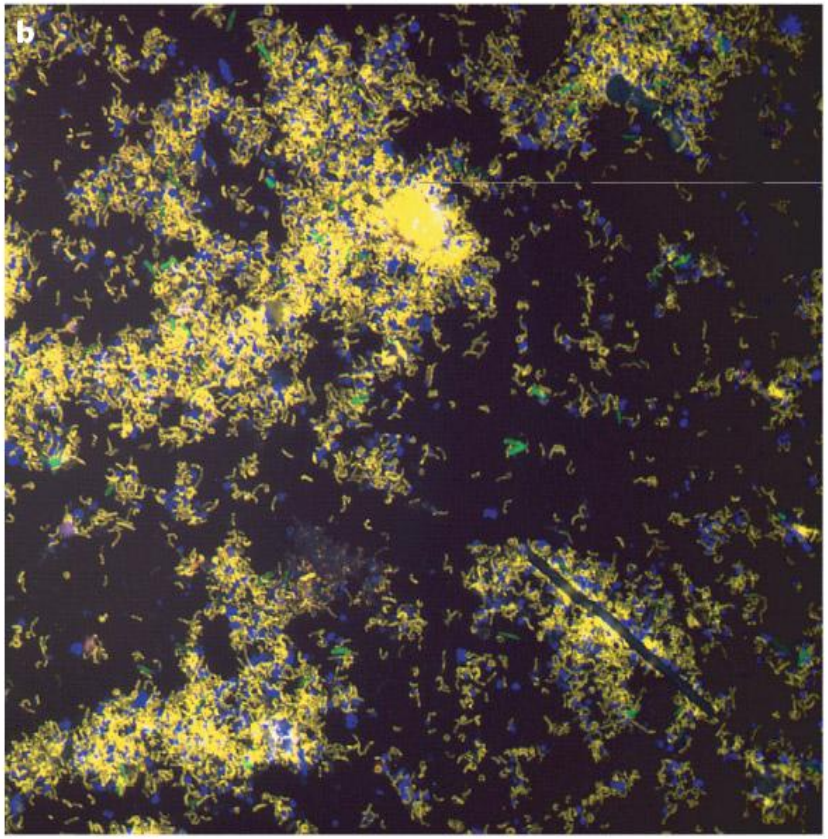
Gene density alone cannot explain the difference pattern



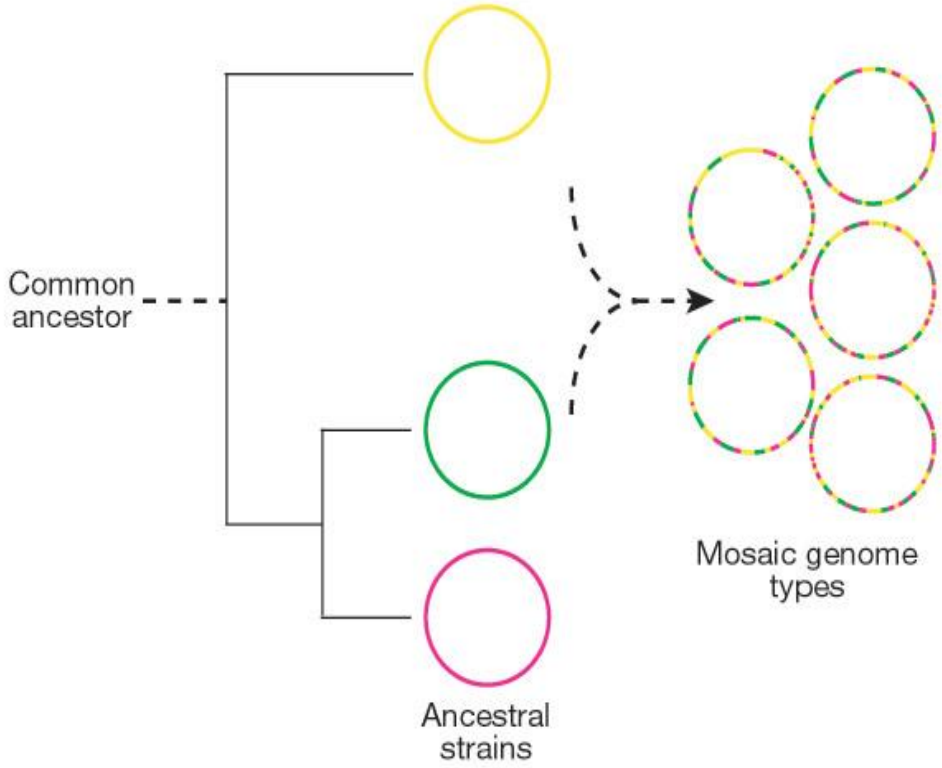
Recent study of extreme environment

- Tyson et al (2004) examined acid mine drainage.
  - Abandoned mine with FeS_2
  - Microbe interactions -> sulphuric acid (pH ~ 0.5)
  - Extreme environment -> low diversity
    - 2 abundant organisms (*Ferroplasma* and *Leptospirillum*)
    - Whole genome reconstruction achievable by shotgun sequencing





Mosaicism in microbial genomes



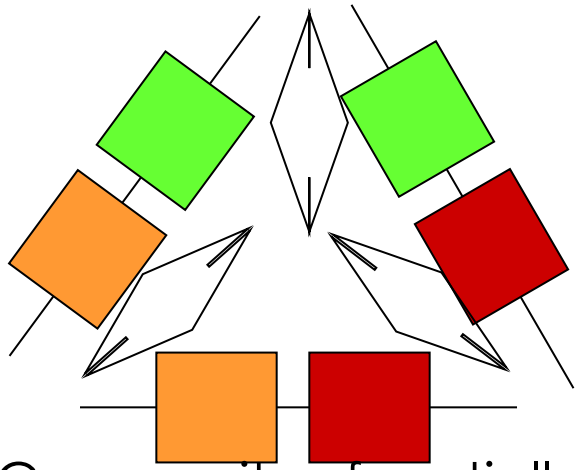
- Mosaic structure of genome suggests microbes swapped intact large quantities of genetic material, from three closely-related ancestral genomes
- Could it be the other way round? Could the mosaic structure and the "ancestral genomes" actually have descended from a homogeneous community?

  Spontaneous speciation?
- Ecology and evolution become intertwined

Evolutionary consequences of diversification fronts

Are there bacterial species?

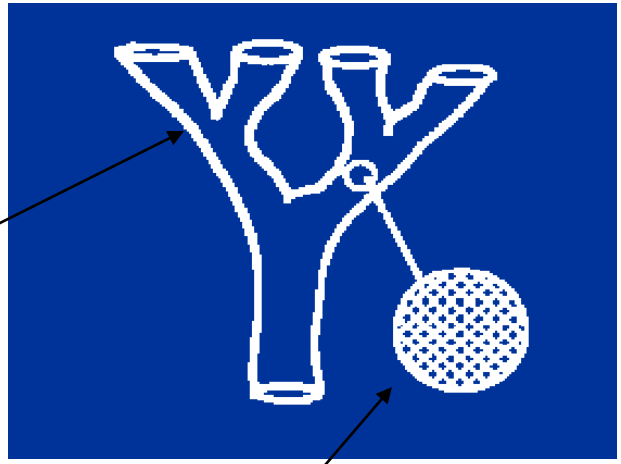
- Species genetic isolation is a global property
- Local isolation is unstable because of fronts



Community of partially isolated organisms

Diversification fronts tend to partition a bacterial community into globally isolated groups

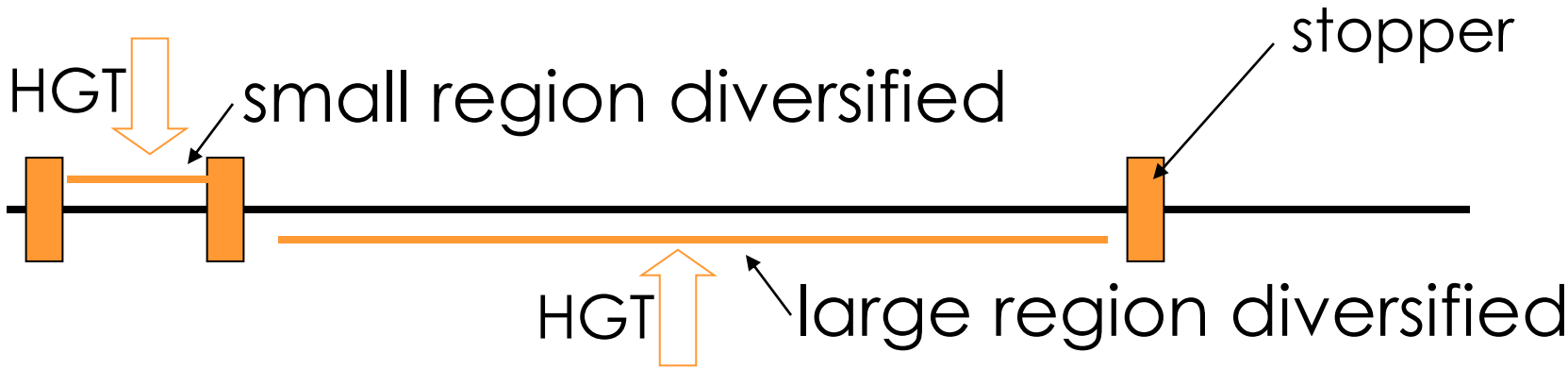
- **Mechanism for speciation**
  - **Ecological distinctiveness + local isolation → global isolation**



recombination

Front Stoppers

- Sufficiently long highly conserved genome regions stop the diversification fronts
- Candidates for stoppers:
  - rRNA operons
  - Overlapping genes
  - Very highly expressed genes/high codon bias
- **Are evolutionary successful HGT islands/rearrangement break points preferentially located near stoppers?**



Conclusion from this study

- Predict that horizontal transfer events can initiate or nucleate diversification fronts leading to speciation that propagate along microbial genomes over evolutionary time
  - Whether this occurs depends upon the way in which alien DNA is incorporated into the chromosome
- Fronts observed where theory predicts and not observed in microbes where theory predicts they should not arise
- Predict a mosaic structure of genomes that is observed and is otherwise puzzling

Diversification fronts in eukaryotes

Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes

Dacheng Tian[1]*, Qiang Wang[1]*, Pengfei Zhang[1], Hitoshi Araki[1,2], Sihai Yang[1], Martin Kreitman[3], Thomas Nagylaki[3], Richard Hudson[3], Joy Bergelson[1,3] & Jian-Qun Chen[1]

Mutation hotspots are commonly observed in genomic sequences and certain human disease loci[1–7], but general mechanisms for their formation remain elusive[7–11]. Here we investigate the distribution of single-nucleotide changes around insertions/deletions (indels) in six independent genome comparisons, including primates, rodents, fruitfly, rice and yeast. In each of these genomic comparisons, nucleotide divergence (D) is substantially elevated surrounding indels and decreases monotonically to near-background levels over several hundred bases. D is significantly correlated with both size and abundance of nearby indels. In comparisons of closely related species, derived nucleotide substitutions surrounding indels occur in significantly greater numbers in the lineage containing the indel than in the one containing the ancestral (non-indel) allele; the same holds within species for single-nucleotide mutations surrounding polymorphic indels. We propose that heterozygosity for an indel is mutagenic to surrounding sequences, and use yeast genome-wide polymorphism data to estimate the increase in mutation rate. The consistency of these patterns within and between species suggests that indel-associated substitution is a general mutational mechanism.

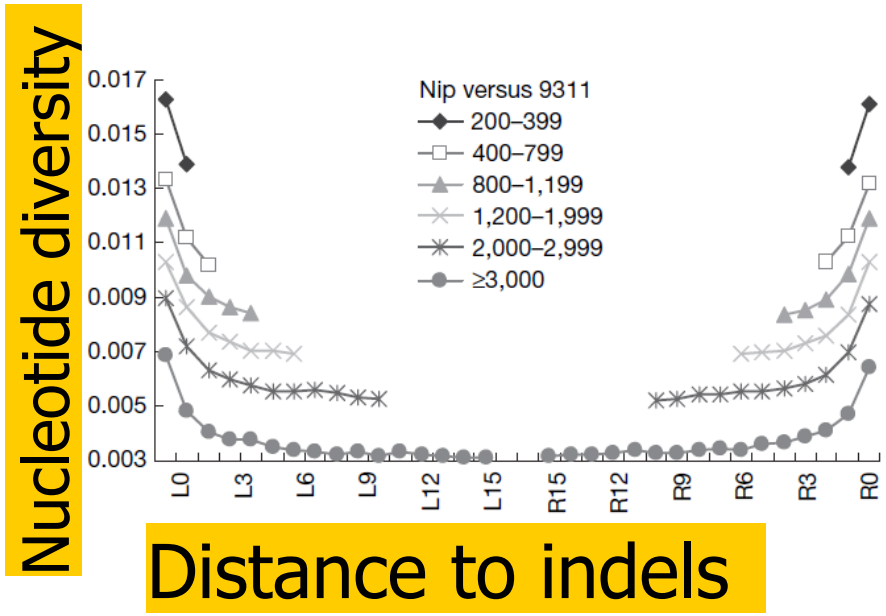


Figure 2 | Relationships between nucleotide divergence (D) and the distance to indels (d_1) as a function of indel interval length (d_2). The lines from top to bottom represent the length of intervals (d_2): 200–399, 400–799, …, ≥3,000 bp for the comparison between human and chimpanzee (**a**) and two rice lines (**b**).

Indels better tolerated in non-coding regions, so may yield a high rate of evolution for gene expression, leading to speciation and phenotype evolution

Ectopic recombination

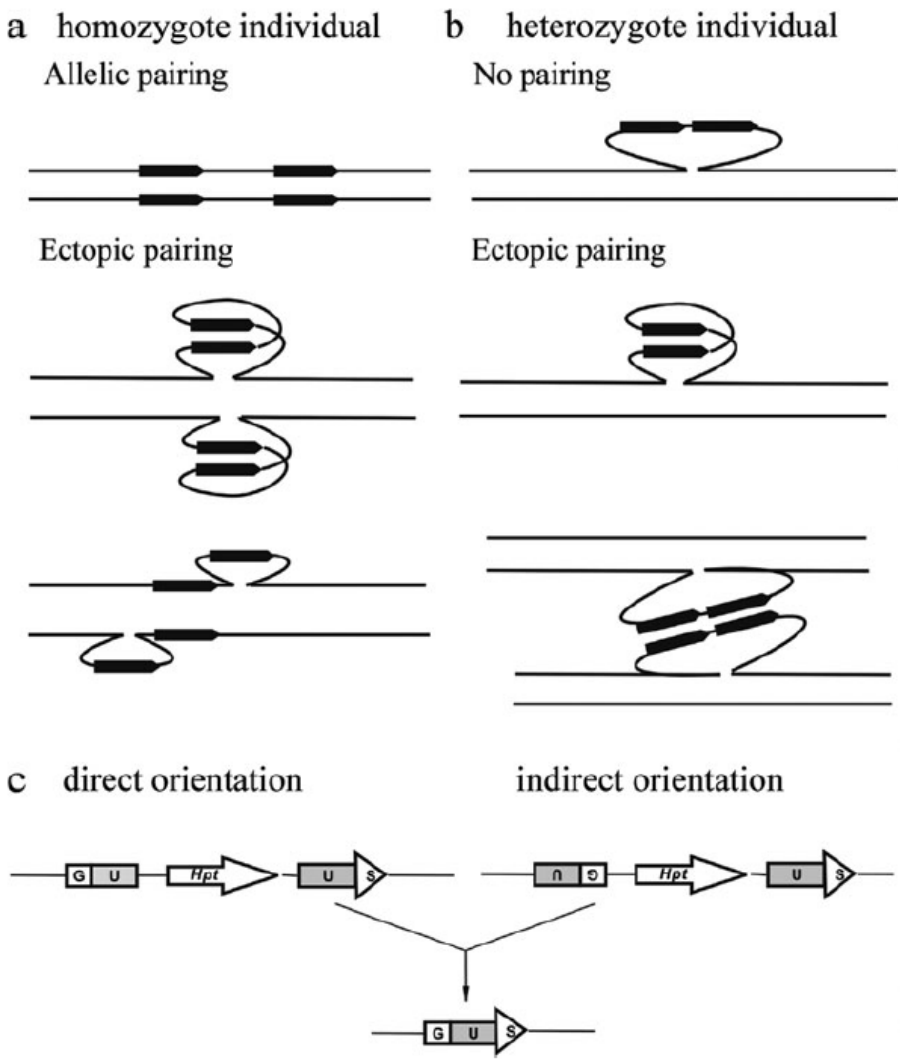


FIG. 1.—Schematic representation of a pairing model during meiosis in noninsertion DNA (*a*, symmetric sequences), insertion (*b*, asymmetric), and recombination substrates (*c*) integrated in transgenic lines. The 2 parallel lines represent 2 homologous chromosomes, and the gray boxes depict the homologous sequences. The ectopic pairing, defined as nonallelic pairing, is showed either in the same (center left and right) or in the different chromosomes (bottom left and right between homologous and nonhomologous chromosomes, respectively). In (*c*), the T-DNA of "G," "U," "S," and Hpt represent part of GUS (uidA gene) and a marker gene (hygromycin phosphotransferase gene), and the arrows denote orientation of the genes. The recombination substrates carry the partially overlapping uidA region. The intact GUS genes can be restored via all types of ectopic paring displayed above (Gherbi et al. 2001).

- Ectopic (non-allelic pairing) recombination increased by 14-fold with insertion present
- Insertions impact chromosome instability, genome variation and evolution

Sun et al. 2008

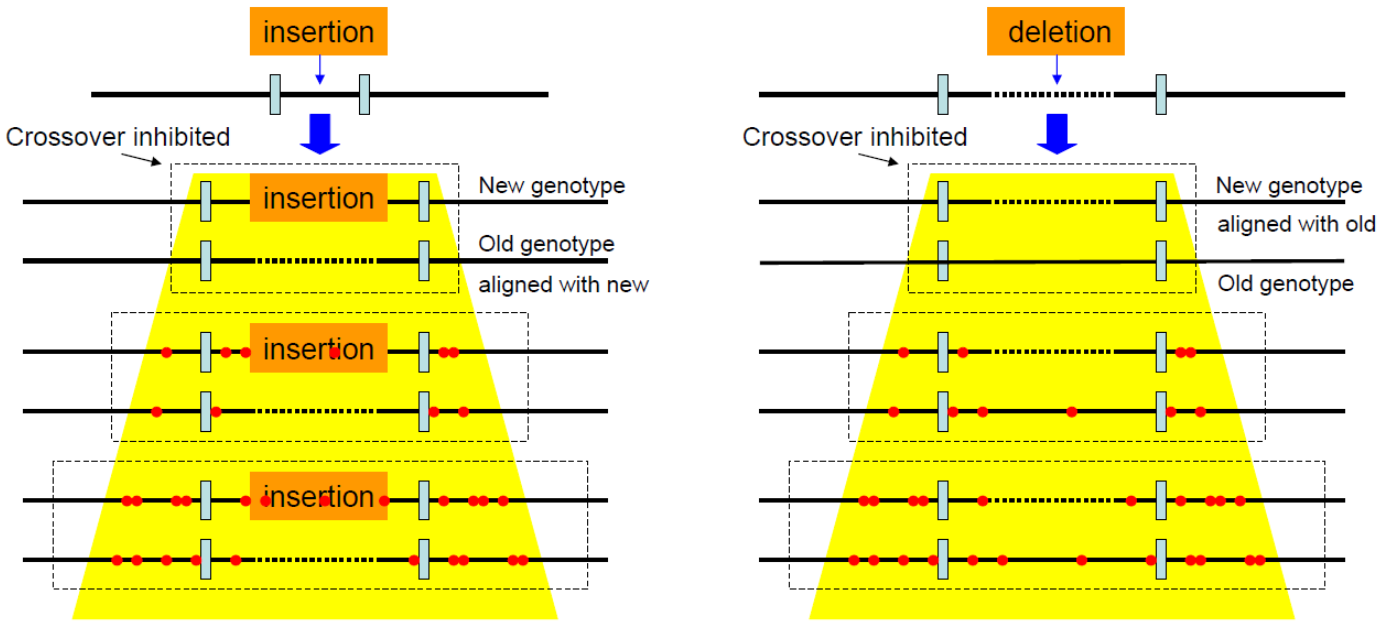
Insertion DNA Promotes Ectopic Recombination during Meiosis in *Arabidopsis*

Xiaoqin Sun,[1] Yuanli Zhang,*[1] Sihai Yang,* Jian-Qun Chen,* Barbara Hohn,† and Dacheng Tian**

*State Key Laboratory of Pharmaceutical Biotechnology, Plant Molecular Institute, Nanjing University, Nanjing, China; and †Friedrich Miescher Institute, Basel, Switzerland

Nucleotide insertion/deletions are common polymorphisms in living organisms; however, little is known about their genetic behavior during meiosis. Here, the recombination frequency (RF) of isogenic strains of transgenic *Arabidopsis thaliana*, that differ in the presence or absence of an insertion, was compared. We screened over 6 million seedlings and found that during meiosis the unpaired DNA insertions paired with ectopic homologues demonstrated a 13.8 times higher RF than that of noninsertion DNA. The direct measurement of recombination events provided the first evidence that a large piece of insertion DNA had a unique genetic behavior during meiosis. This pattern was consistently observed in different lines varying in overlapping sequence, construct orientation, chromosome location, and crossing direction. We suggest that higher ectopic recombination is promoted by DNA insertions and that this mechanism exists commonly in plants. Therefore, insertion DNA plays a nontrivial role in shaping genetic variation, chromosome instability, and genome evolution.

Diversification fronts in eukaryotes



- Indels common in eukaryotes: e.g. 20% rice genome
- Mutation rate enhanced near indels (Tian et al. 2008)
  - Genome symmetry broken → polymorphism
  - Genetic isolation observed
- Indels promote ectopic recombination during meiosis

Outline

- Part 1: global diversification of microbial genomes through the interplay between recombination and point mutation

- Part 2: compositional bias in microbial genomes. How feedback between resource allocation and template-directed synthesis leads to multistability of evolved microbial genomes

Nonlinear genome dynamics over evolutionary time

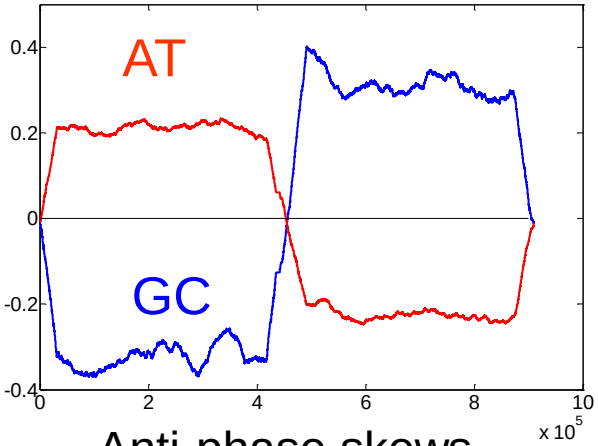
Breakdown of mutation-drift-selection

K. Vetsigian and N. Goldenfeld. Genome rhetoric and the emergence of compositional bias. *Proc. Natl. Acad. Sci. (USA)* (2008)

The puzzle of genome bias

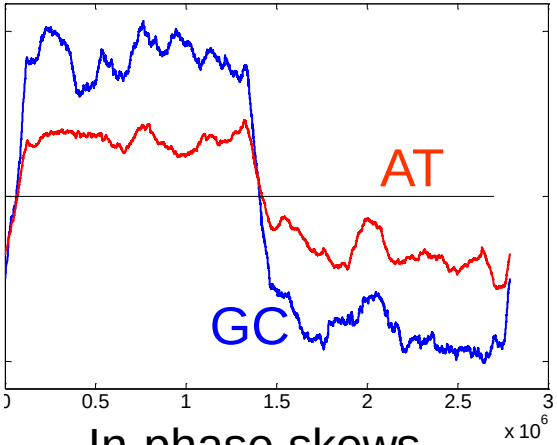
- **GC content** (Ex.: *Streptomyces coelicolor* **72 %,** *Arcobacter butzleri* **27%)**
- **Codon usage**
- **AT skew** = (A-T)/(A+T), **GC skew** = (G-C)/(G+C)

Borrelia burgdorferi

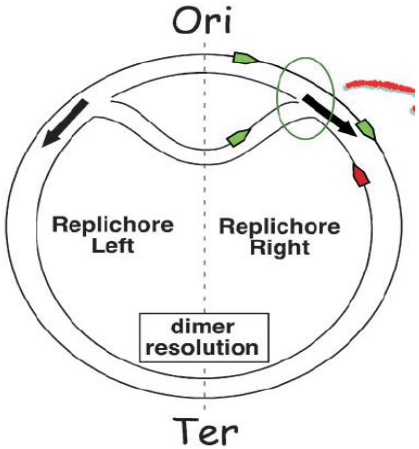


Anti-phase skews

Clostridium tetani



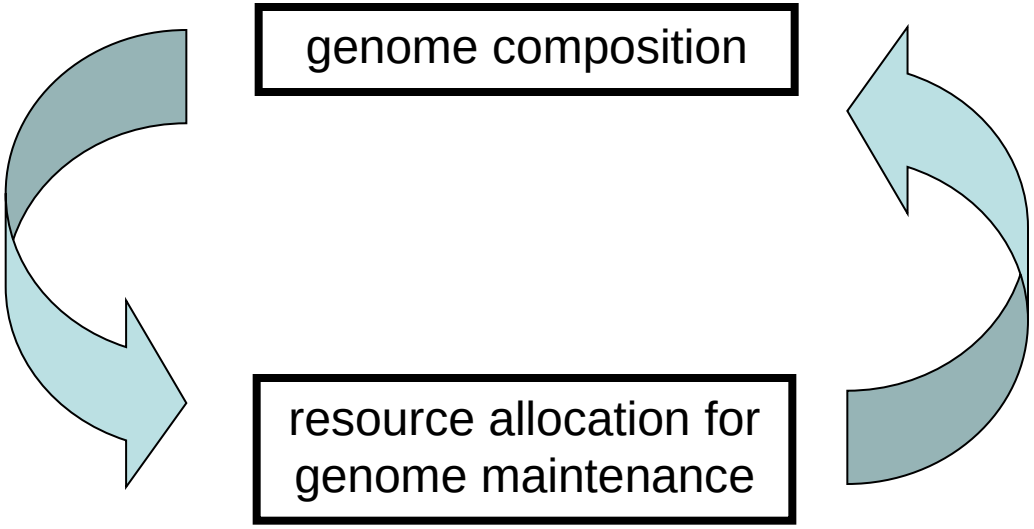
In-phase skews



- All possible combinations of positive and negative skews are observed
- Proposed biological explanations only lead to one sign of the skews

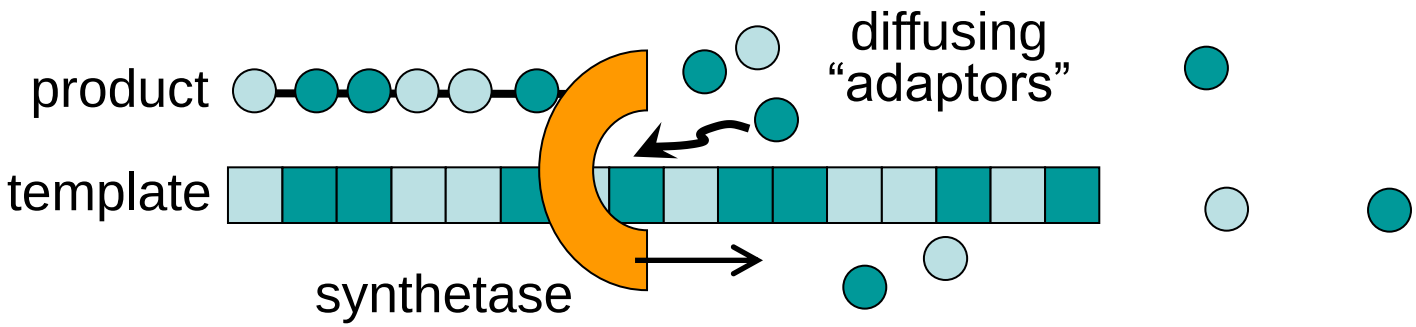
Emergent genome editing:
Systems biology meets evolution

System level evolutionary dynamics of genome *maintenance* processes directs genome *evolution*



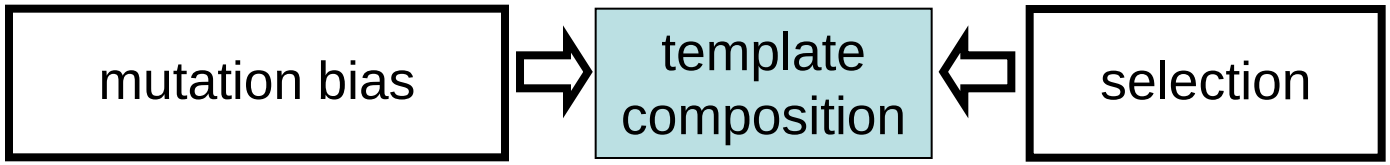
Template-directed synthesis:

the least common denominator of replication, transcription and translation



process	template letters	adaptors	synthetase	encoding rule
translation	codons	tRNAs	ribosome	Genetic code
transcription	A,G,C,T	NTPs	RNA polymerase	Watson-Crick pairing
replication	A,G,C,T	dNTPs	DNA polymerase	Watson-Crick pairing

Mutation-selection theory of template-directed synthesis

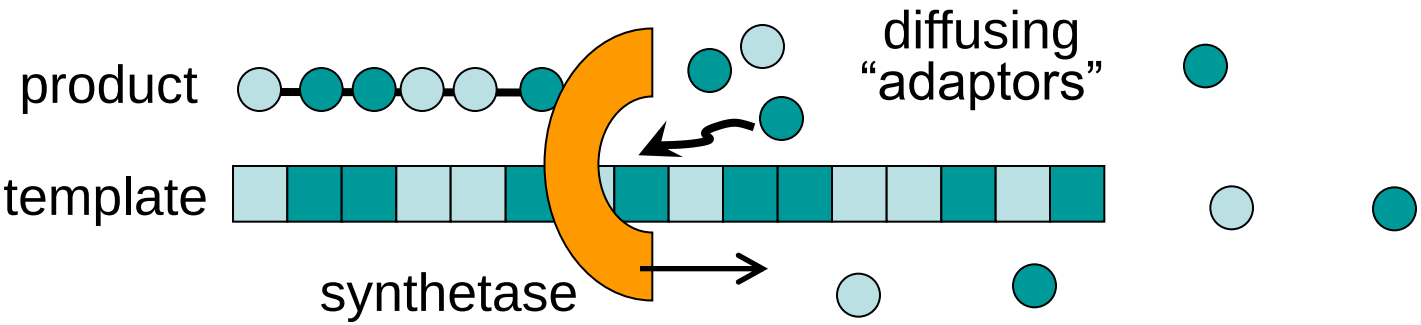


- Example: Different tRNA species and abundances in different organisms *select* for different codon usage
  - More abundant tRNAs select for their codons (to optimize translational efficiency)

- But why are mutation and selection different in the first place (given the universal function + structure of information processing)?

- Such a model requires different *ad hoc* mechanisms for each bias

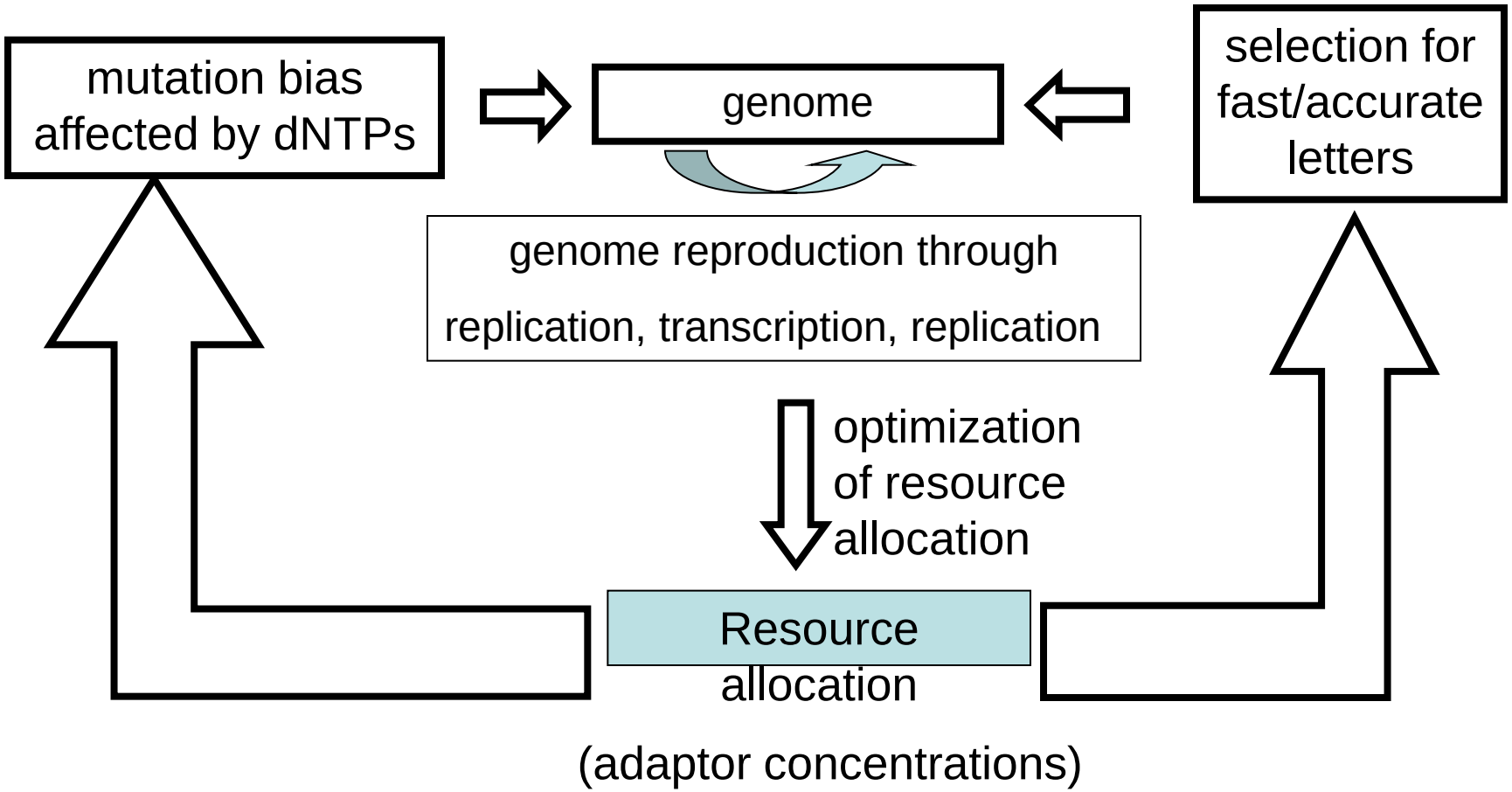
Speed and accuracy are affected by the adaptor concentrations



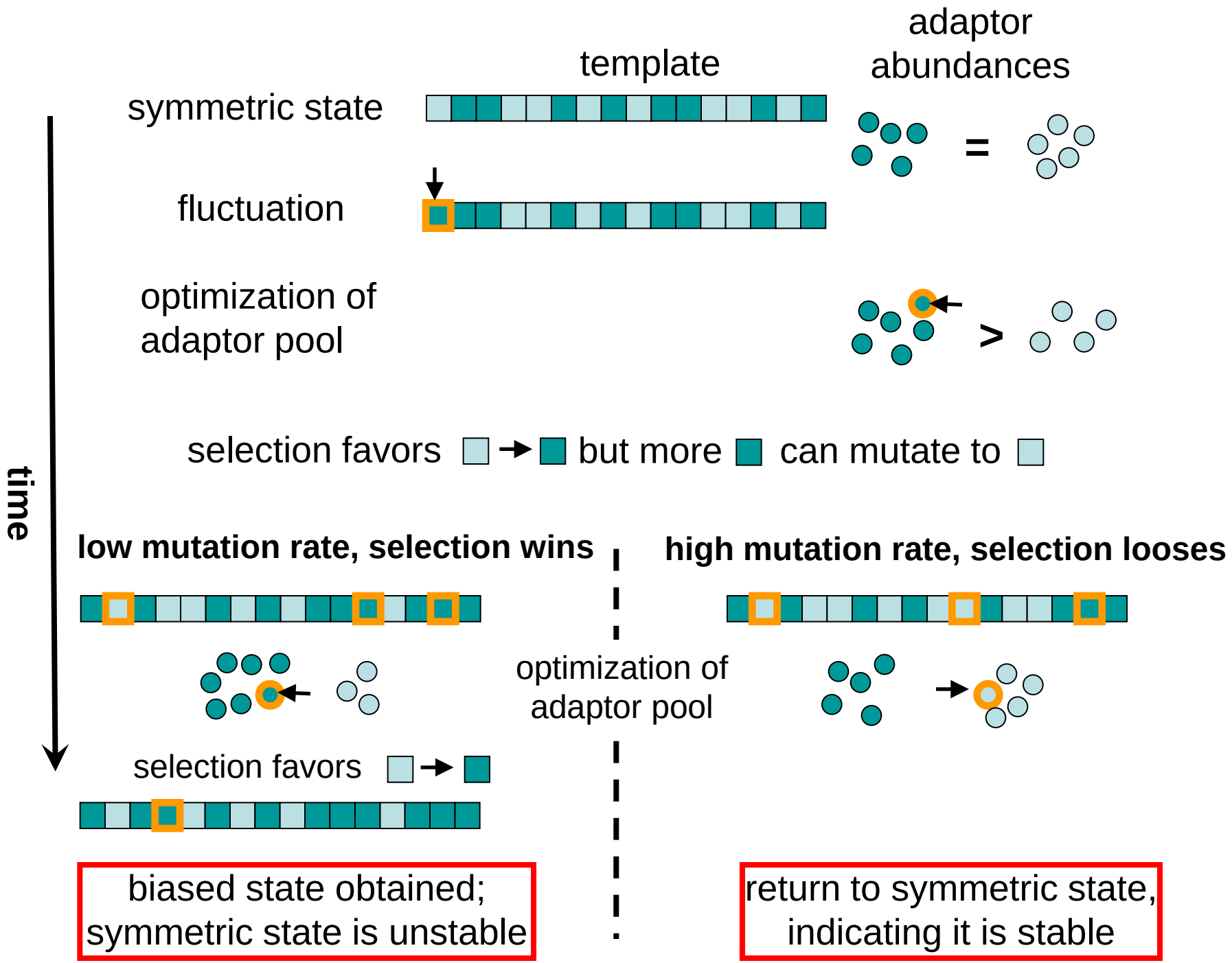
	at ■ :	at □ :
diffusion:	wait time ~ **1 / [●]**	wait time ~ **1 / [○]**
competition:	error rate ~ **[○] / [●]**	error rate ~ **[●] / [○]**

- Speed and accuracy of synthesis – want higher abundance of common adaptors
- Fluctuations in concentrations irrelevant at high mutation rates, but at low mutation rates, instabilities cause symmetry breaking

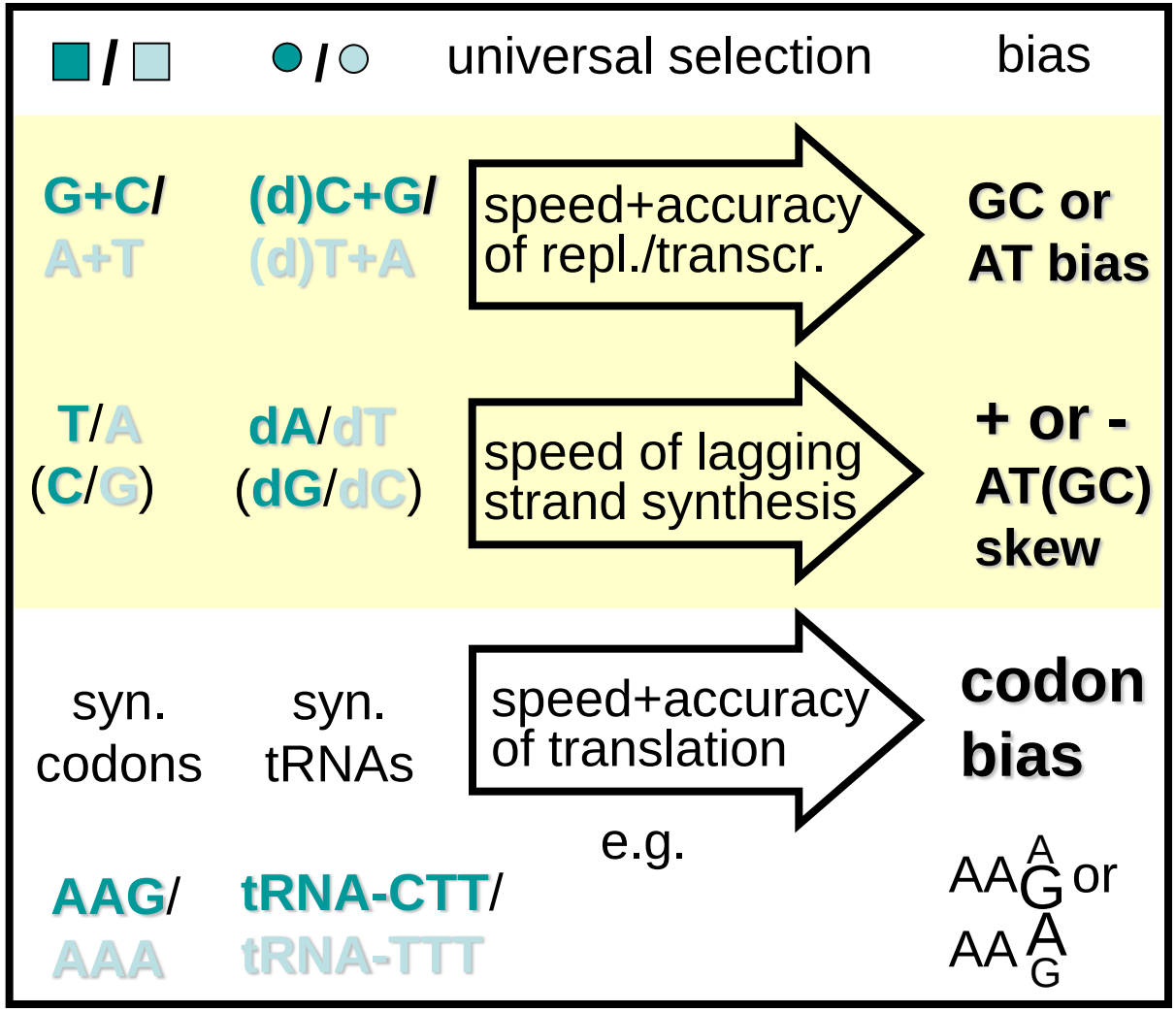
Combining mutation-selection framework with resource optimization



Feedback loops generate evolutionary instabilities


template
adaptor abundances
symmetric state
=
fluctuation
optimization of adaptor pool
>
selection favors
but more
can mutate to
time
low mutation rate, selection wins
high mutation rate, selection looses
optimization of adaptor pool
selection favors
biased state obtained; symmetric state is unstable
return to symmetric state, indicating it is stable

Unified theory of genome bias diversity



Quantify the coevolution

- Quasispecies fitness

Cost of adaptors

Adaptor concentrations

$$f(c,u) = \frac{G\left(\sum_a \alpha_a c_a\right)}{\Theta + \underbrace{\sum_{si} L_s u_{si}/c_i}_{\text{Synthesis time}}} \prod_{si} \left(R_{s,i}\right)^{L_s u_{si}}$$

- Change *c'* in adaptor pools invades if $f(c';u) > f(c;u)$
- Letter usage relaxes to a new mutation-selection-equilibrium

$$(\mathrm{MS})\mathbf{R_s u_s} = \lambda_s \mathbf{u_s}.$$

- Redundancy structure:

Fixed sequense of site types

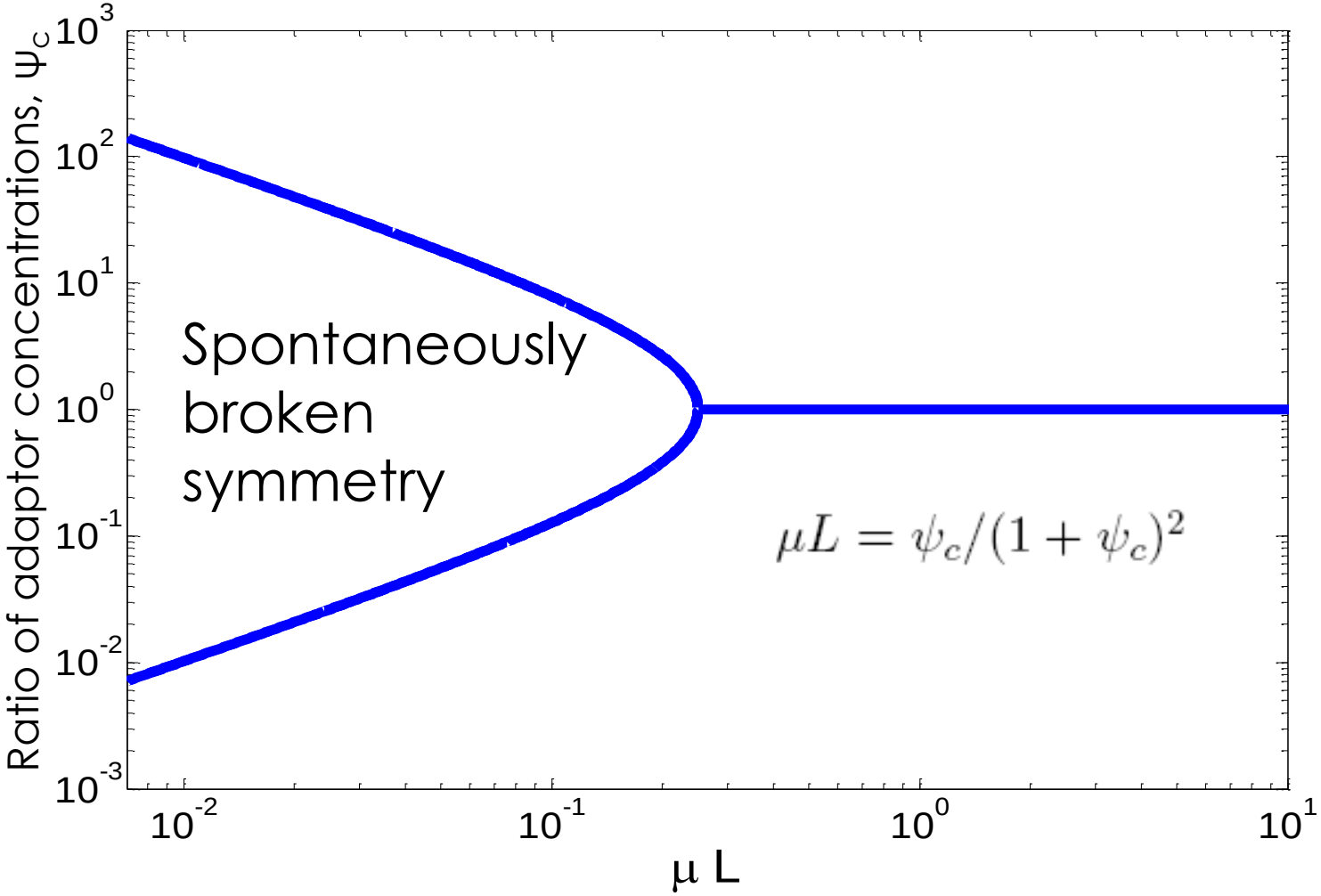
s_1	s_2	s_1	s_3	s_4	s_2	s_1	s_4	s_4	s_4	s_1	s_1	s_2	s_1	s_1	s_2

$R_{s,i} =$ Fitness of letter *i* at site type *s, e.g:* $R_{si} = \delta_{i,A} + \delta_{i,G}$

u_{si} Frequency of letter *i* at site type *s*

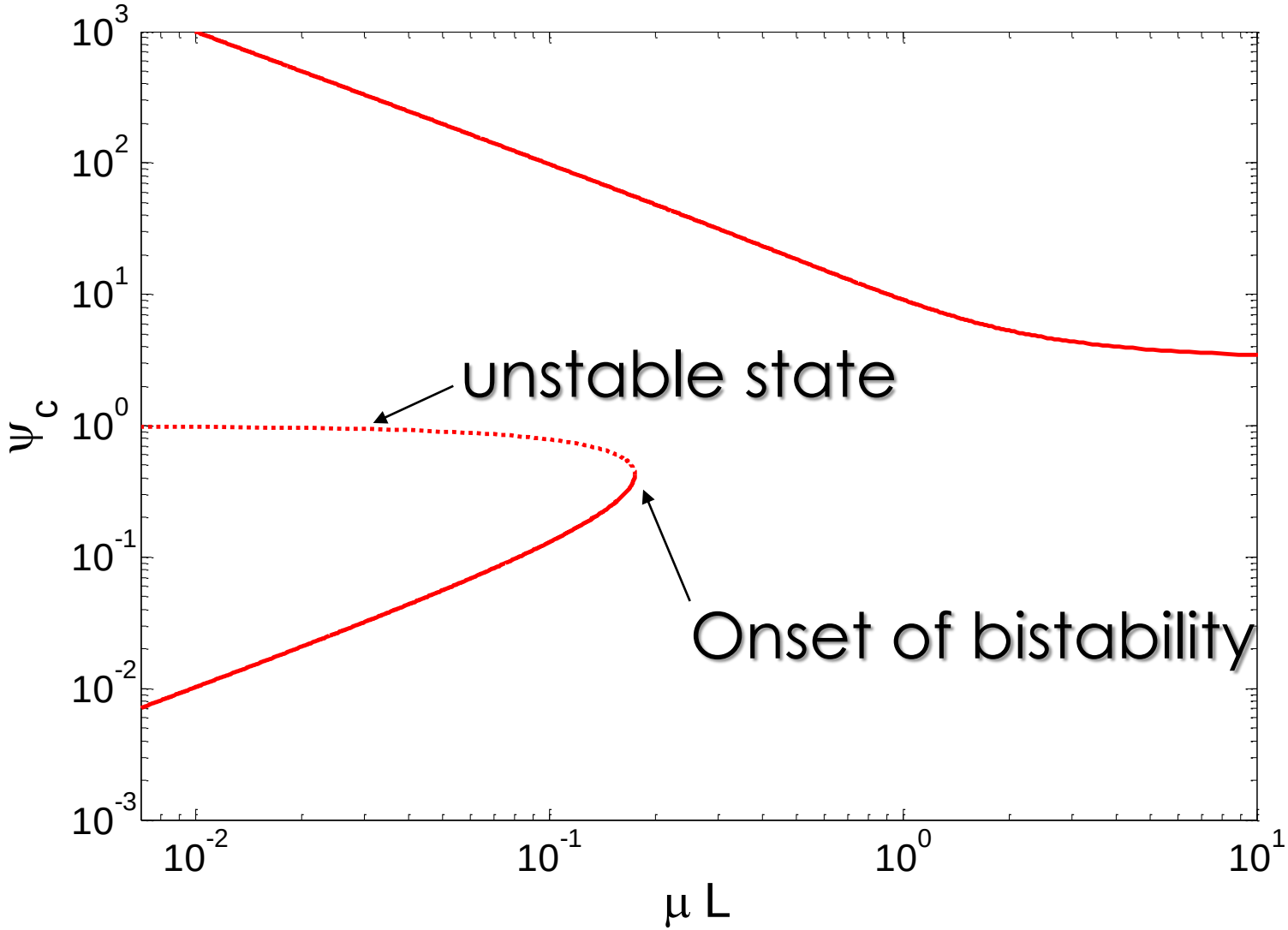
L_s – frequency of site type *s*

Neutral model – continuous phase transition



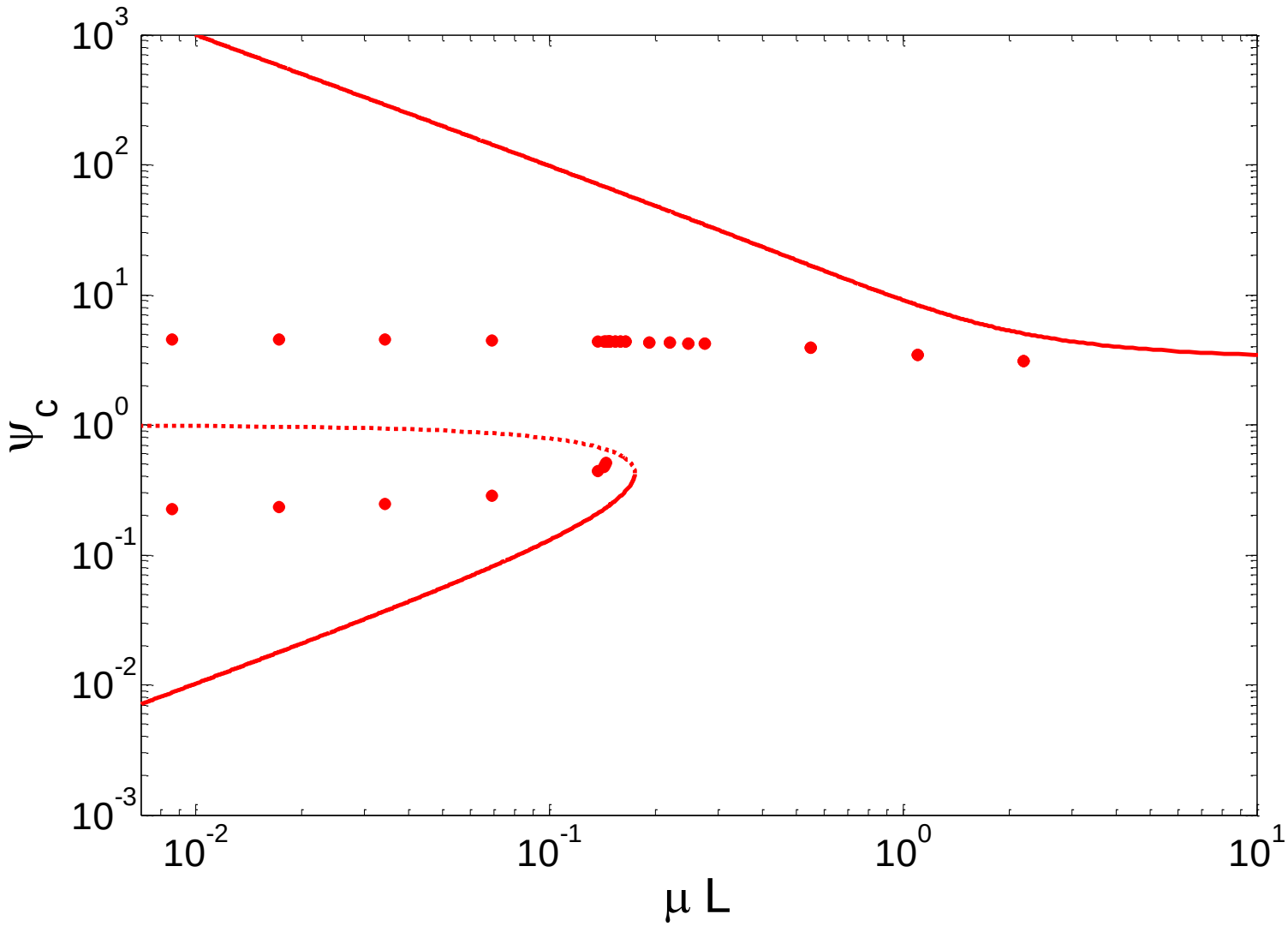
The controlling parameter is **number of mutations per genome per generation**

Neutral model with strong (10-fold) mutation bias



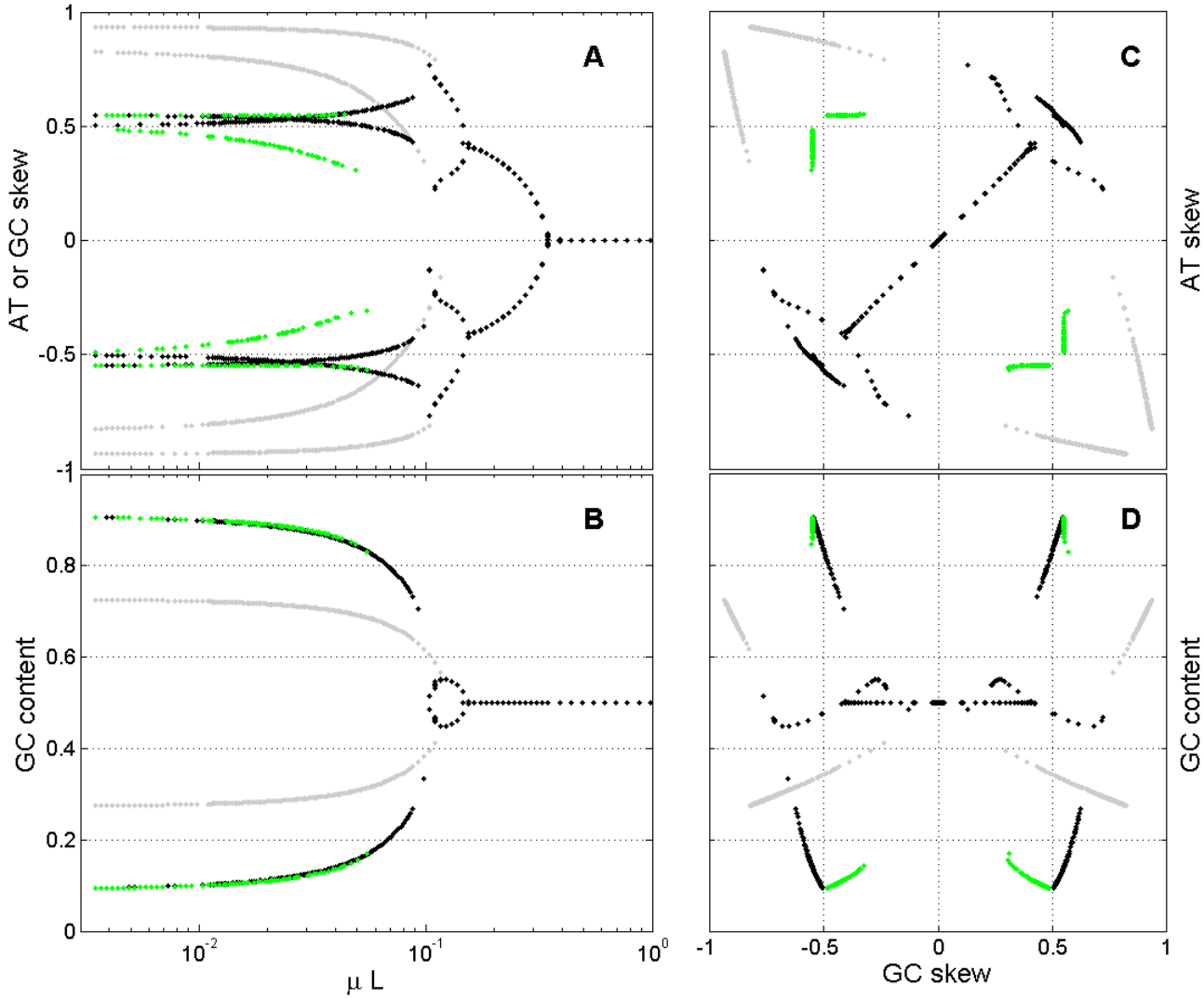
Asymmetry does not destroy the bistability

Non-neutral sites regularize the bias magnitude



(10% of the template letters are fixed)

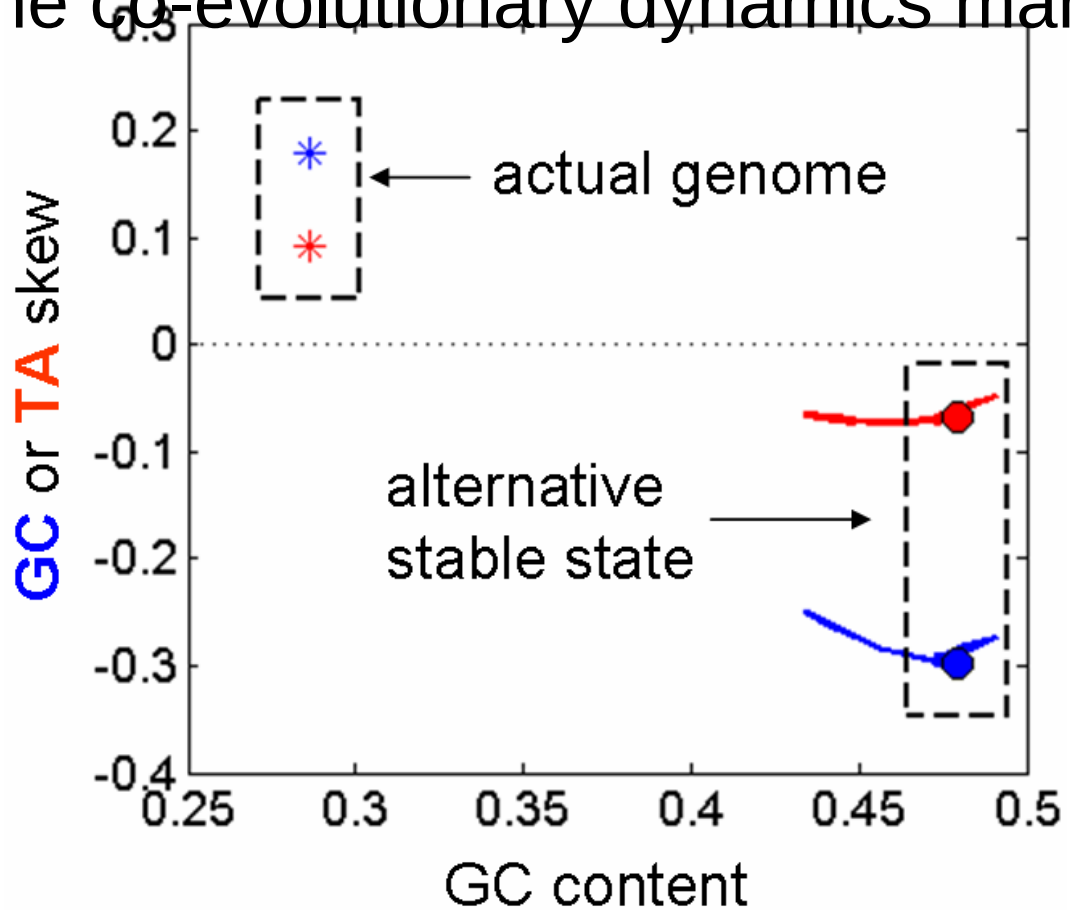
Cascade of transitions in a 4-letter model with transition-transversion bias



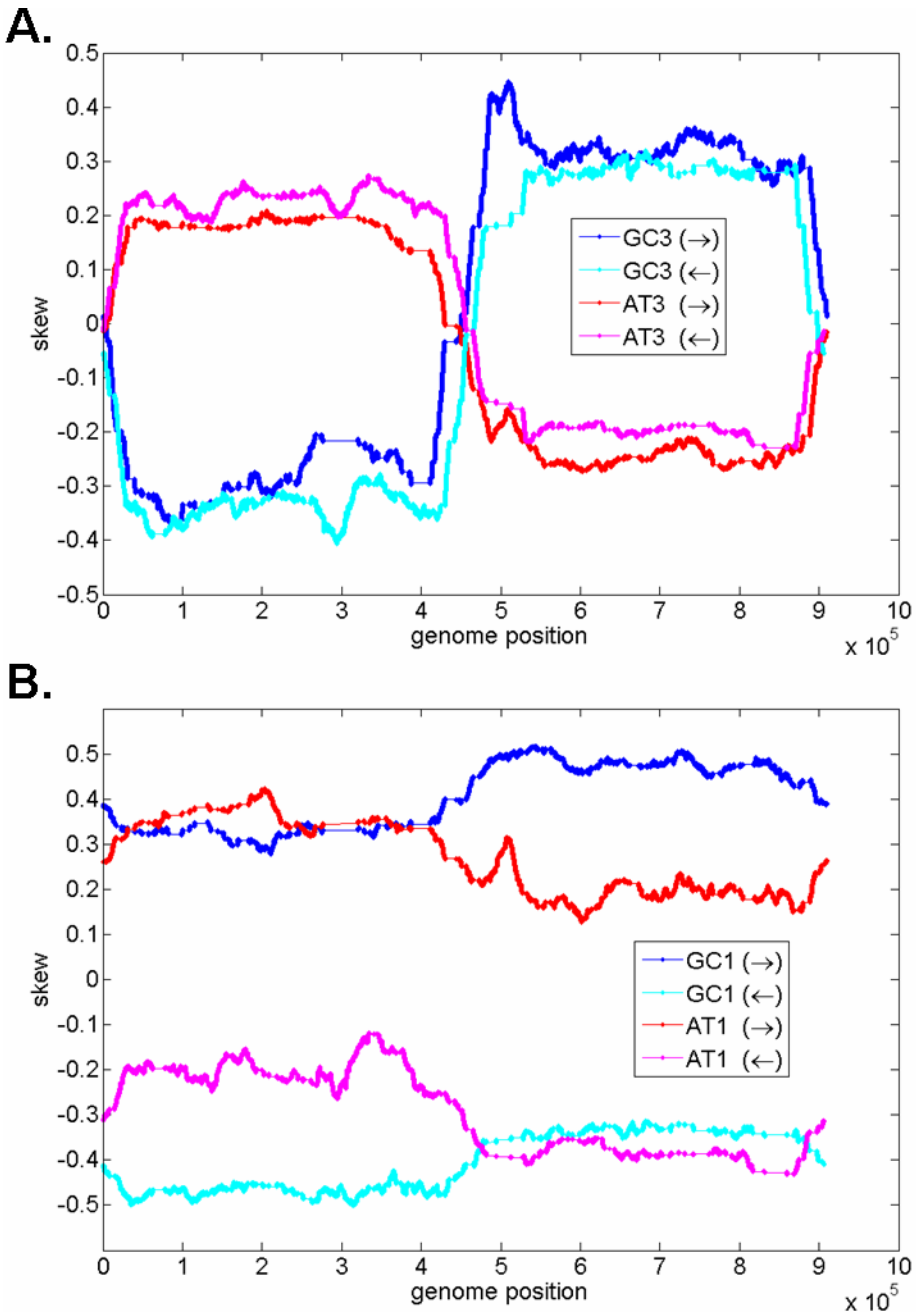
Both in-phase and **out-of-phase** GC and AT skews are possible

Model can be constrained with whole genome data

- Fit the parameters, assuming: 1) mutation-selection equilibrium at synonymous 3rd codon positions; 2) translational selection is same on both strands
- Run the co-evolutionary dynamics many times



Predicted bistability of nucleotide composition for *Borrelia burgdorferi*


A.
skew
genome position
x 10^5
GC3 (→)
GC3 (←)
AT3 (→)
AT3 (←)
B.
skew
genome position
x 10^5
GC1 (→)
GC1 (←)
AT1 (→)
AT1 (←)

Conclusion from this study

- Theory predicts that genome composition should not be uniform, due to nonlinear dynamics and multistability on evolutionary time-scales
- Universal selection towards bias but not its direction
- Classify patterns of diversity based on the patterns of stable solutions
- Implications for evolution of cell machinery

Conclusions

- Modern experimental techniques reveal a rich array of dynamical mechanisms at the genome level
  - Good problems for dynamical systems theory of evolution
- Predictions include speciation mechanisms, generic origin of compositional biases in microbial genomes