Class 5

Symbiosis, Endosymbiosis, Symbiosis at the molecular level
Inteins

For Friday: Alphafold is hosted on a google server. To use it, you need to have a working Google account.

On Friday you will need your **Google username and password**.

Heinrich Anton de Bary

"Die Erscheinung der Symbiose" (Strausbourg, 1879)

- Parasitism/Antagonism
- Commensalism
- Mutualism
- Lichenism/Symbiogenesis

Beatrix Potter

Was not allowed to present her 1897 paper on lichen symbionts to the Royal Academy of Sciences due to her gender (an apology was issued on the 100 year anniversary of the event)

Symbiosis (from Wikipedia)

Symbiosis (from Greek συμβίωσις "living together", from σύν "together" and βίωσις "living") is any type of a close and long-term biological interaction between two different biological organisms, **be it mutualistic, commensalistic, or parasitic**. The organisms, each termed a **symbiont**, may be of the same or of different species. In 1879, Heinrich Anton de Bary defined it as "the living together of unlike organisms".

The term was subject to a century-long debate about whether it should specifically denote mutualism, as in lichens; biologists have now abandoned that restriction.

According to this definition,

all parasites that live closely with their host are symbionts!

Wikipedia is not infallible, but among most scientists this is accepted – but apparently not by all.

What might be the reason for the popularity of the more encompassing definition?

Examples of mutualistic symbiosis

- (Endo)symbiotic nitrogen-fixing bacteria in Alder (*Frankia*) and related tree, in Leguminosae (*Rhizobium*), in Gunera (*Nostoc)*
- Lichen: symbioses between fungi and eukaryotic algae or cyanobacteria
- Mycorrhiza in plants: fungi take over the role of root hairs
- Microbiota in the gut of animals (digestions of cellulose and other compounds, vitamin biosynthesis, production of group specific smell, synthesis of essential amino acids (especially in sap eating insects)
- Microorganisms in the rhizosphere of plants (provide nutrients and water).
  ?Communication between neighboring plants (the wood wide web)?

Aid from neighboring trees, reserves and effective wound healing, or two stems from the same tree?

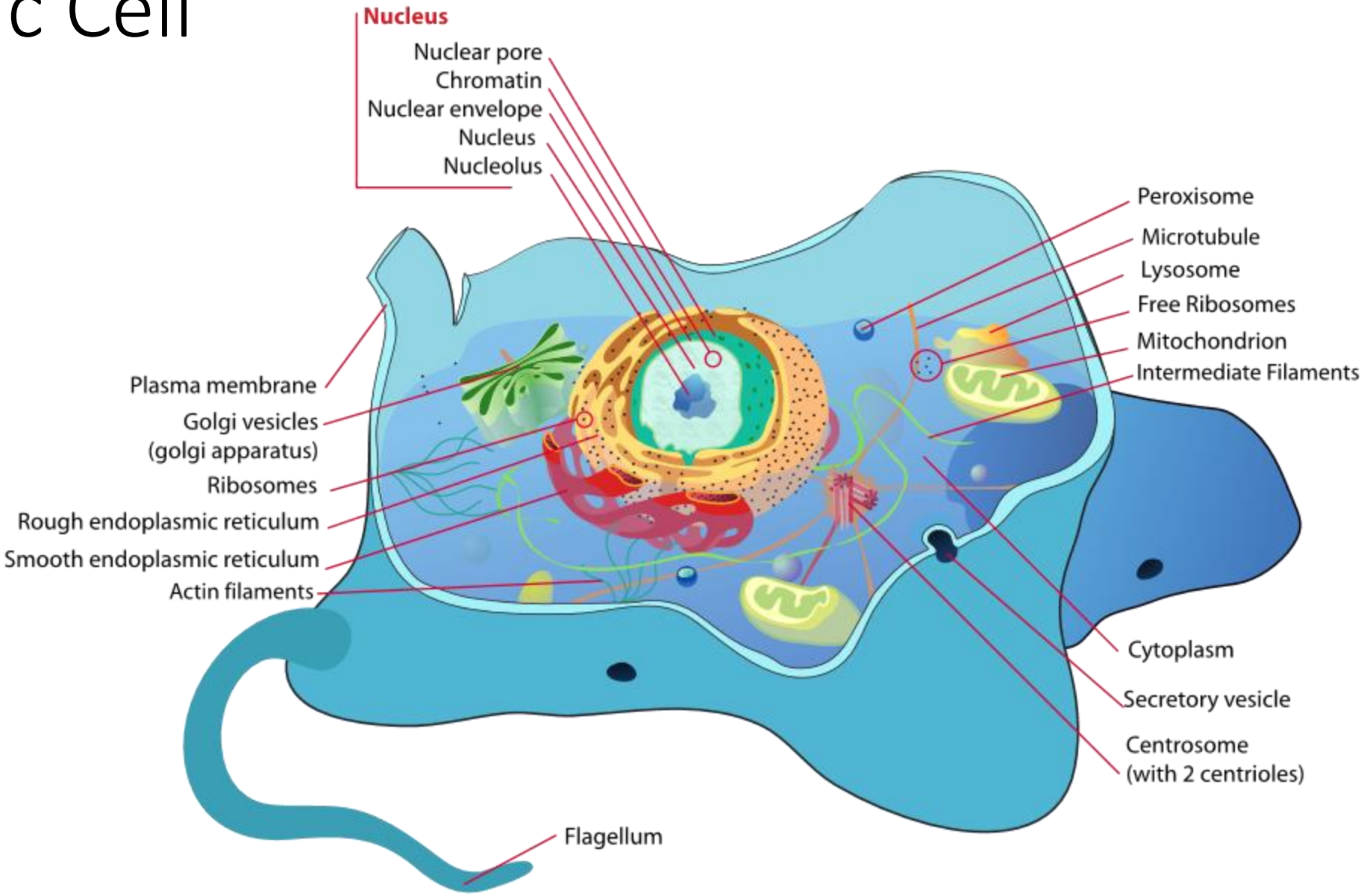
Tree stump on Vashon island (near Seattle, Washington). The stump's surface has been completely overgrown by new bark.

Example of the wood wide web?

See here for a critical assessment.

The Eukaryotic Cell

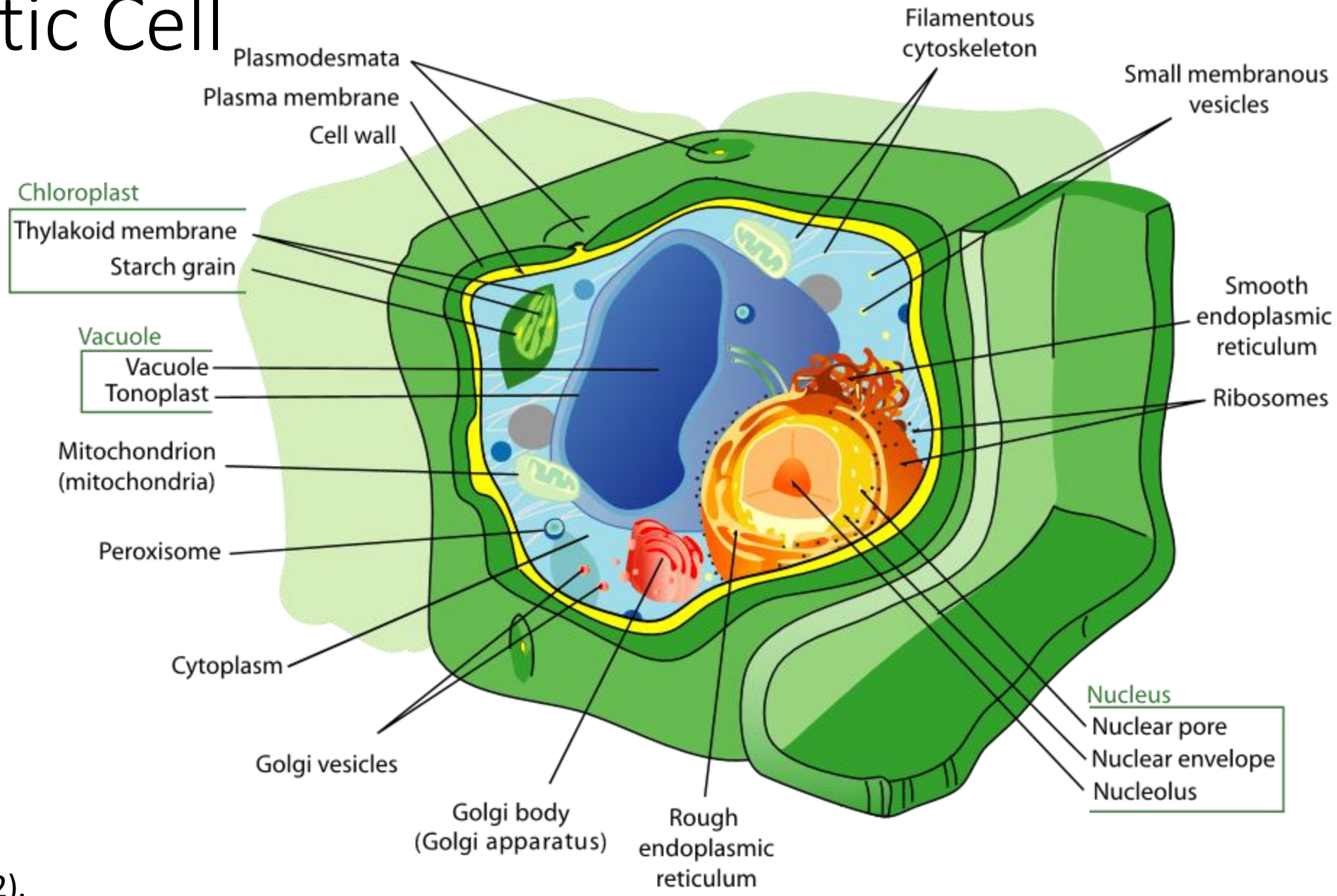
Animal Cell



Eukaryote. (2023, March 2).
In *Wikipedia*.https://en.wikipedia.org/wiki/Eukaryote

The Eukaryotic Cell

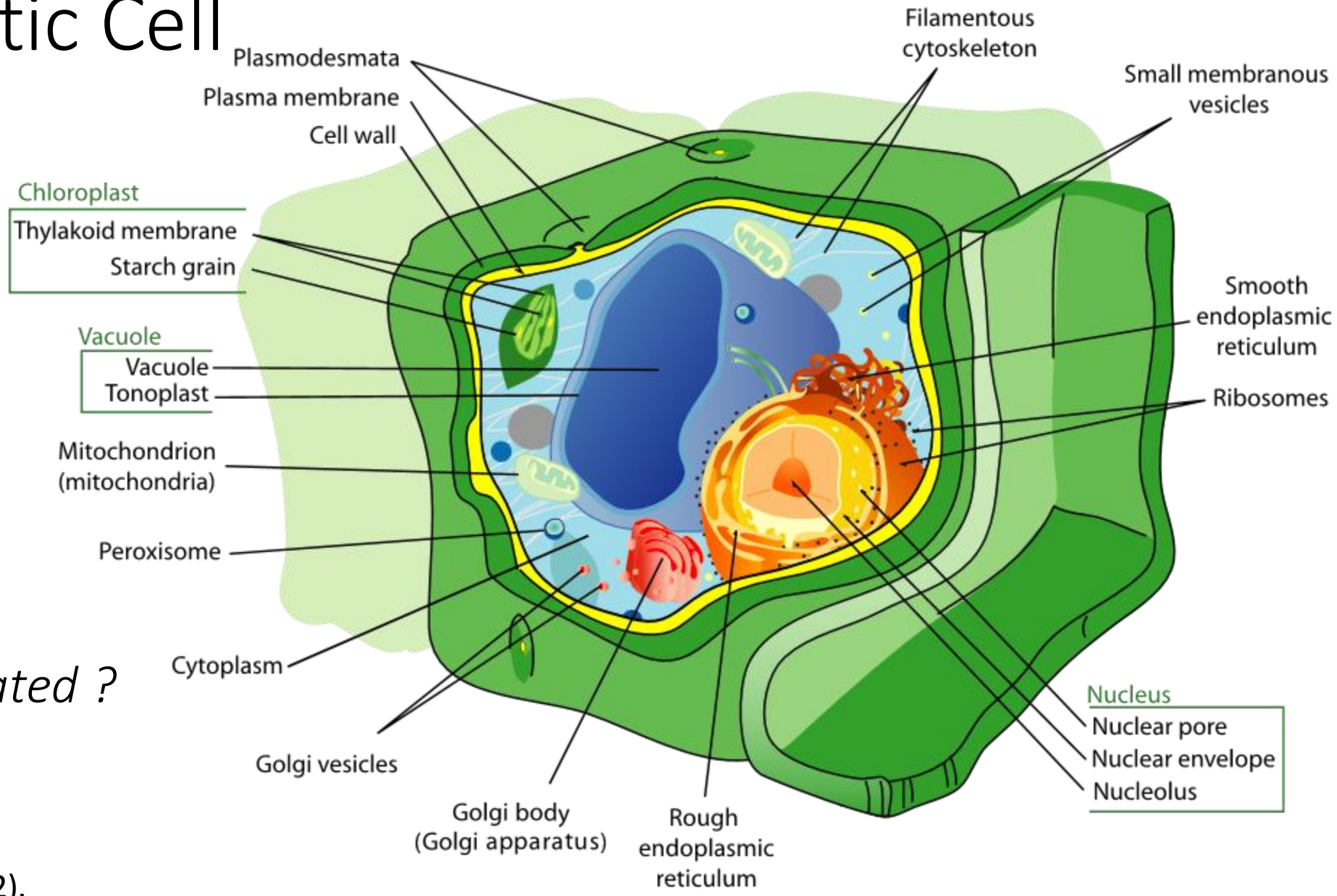
Plant cell



Eukaryote. (2023, March 2).
In *Wikipedia*. https://en.wikipedia.org/wiki/Eukaryote

The Eukaryotic Cell

Plant cell



Where is DNA located ?

Eukaryote. (2023, March 2).
In *Wikipedia*. https://en.wikipedia.org/wiki/Eukaryote

Mitochondria and Plastids

- Have their own circular chromosome.
- Surrounded by two membranes – the inner membrane is similar to bacterial membranes.
- Have their own ribosomes.
- Many of the enzymes functioning in these organelles are very similar to their bacterial homologs (ATP synthase, ribosomal proteins).

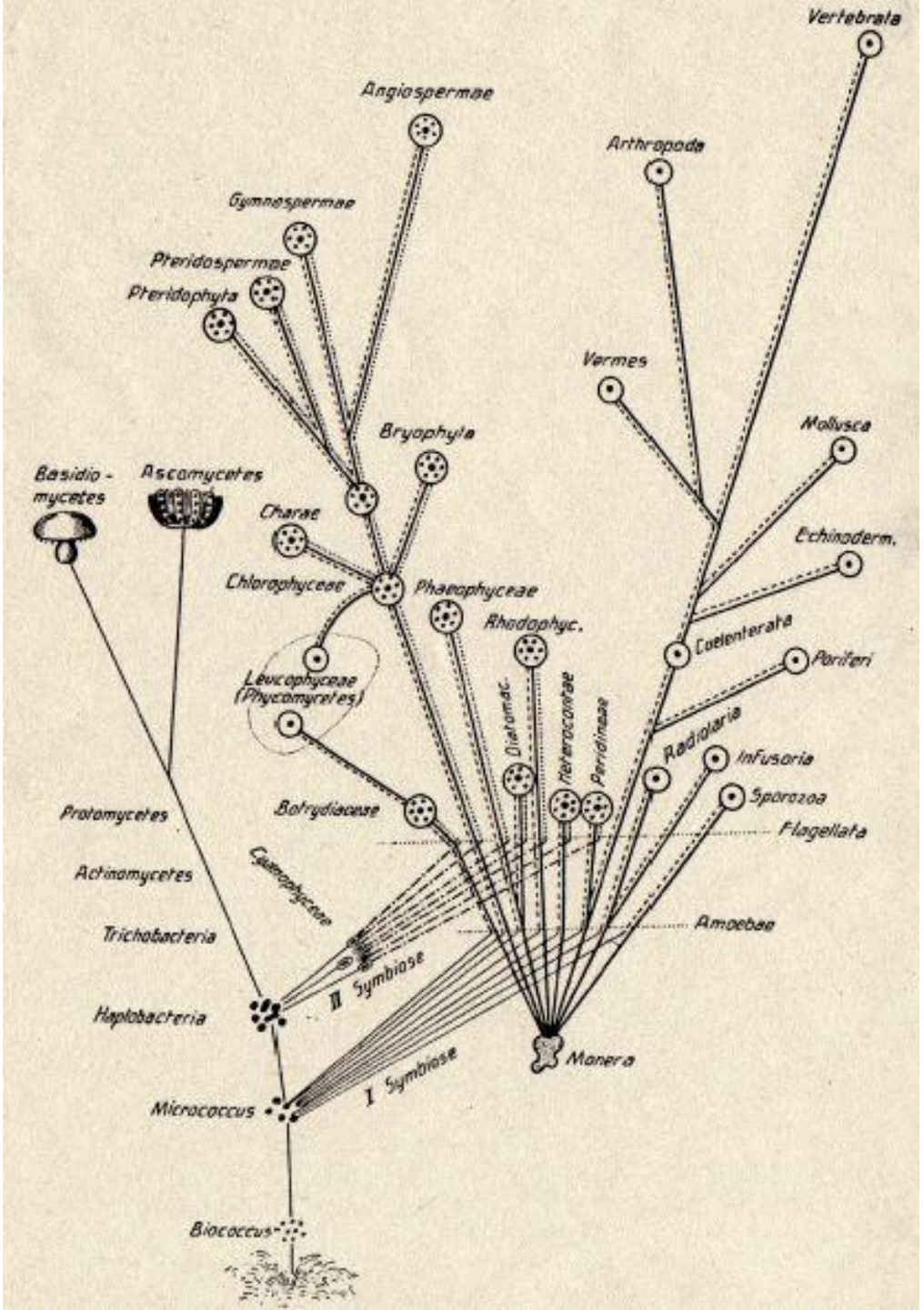
- **Today there is no doubt that plastids and mitochondria evolved from bacteria that lived as endosymbionts inside a host cell (related to Archaea).**

Symbiogenesis

The first proposal that endo-symbiosis played a role in the origin of Eukaryotes goes back to Konstantin Mereschkowski.

1905 tree-of-life diagram, showing the origin of complex life-forms by two episodes of symbiogenesis, the incorporation of symbiotic bacteria to form successively nuclei and chloroplasts.

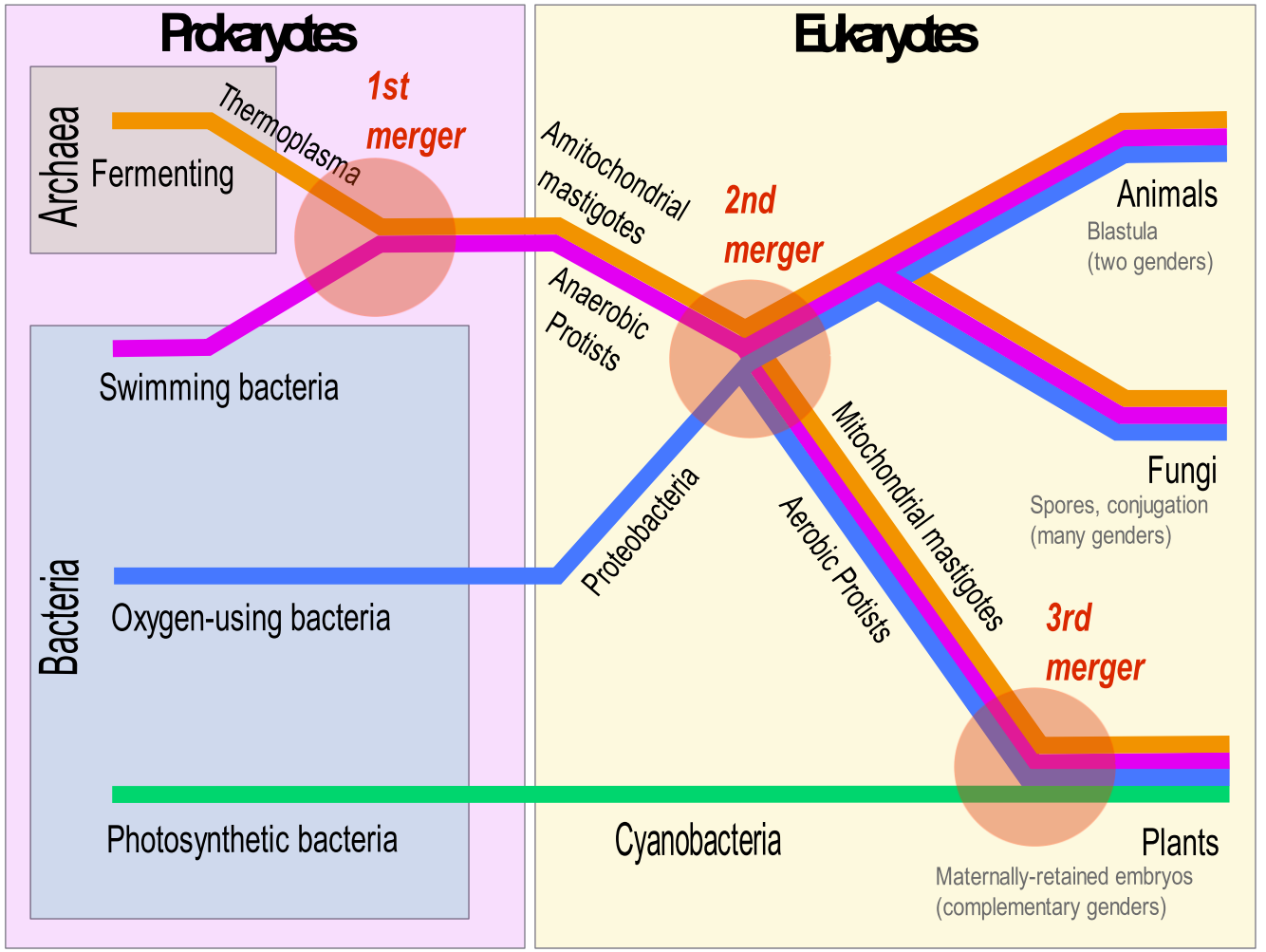
From Symbiogenesis. (2023, February 21). In *Wikipedia*. https://en.wikipedia.org/wiki/Symbiogenesis



Lynn Margulis and the Serial Endosymbiotic Theory

In 1967, Lynn Margulis revived the old but largely ignored theory of symbiogenesis. She proposed three mergers. The first merger is poorly supported and now not generally believed.

Molecular phylogenies provide overwhelming support for the bacterial origin of mitochondria and plastids.



Ian Alexander in Symbiogenesis. (2023, February 21). In *Wikipedia*. https://en.wikipedia.org/wiki/Symbiogenesis

Inteins

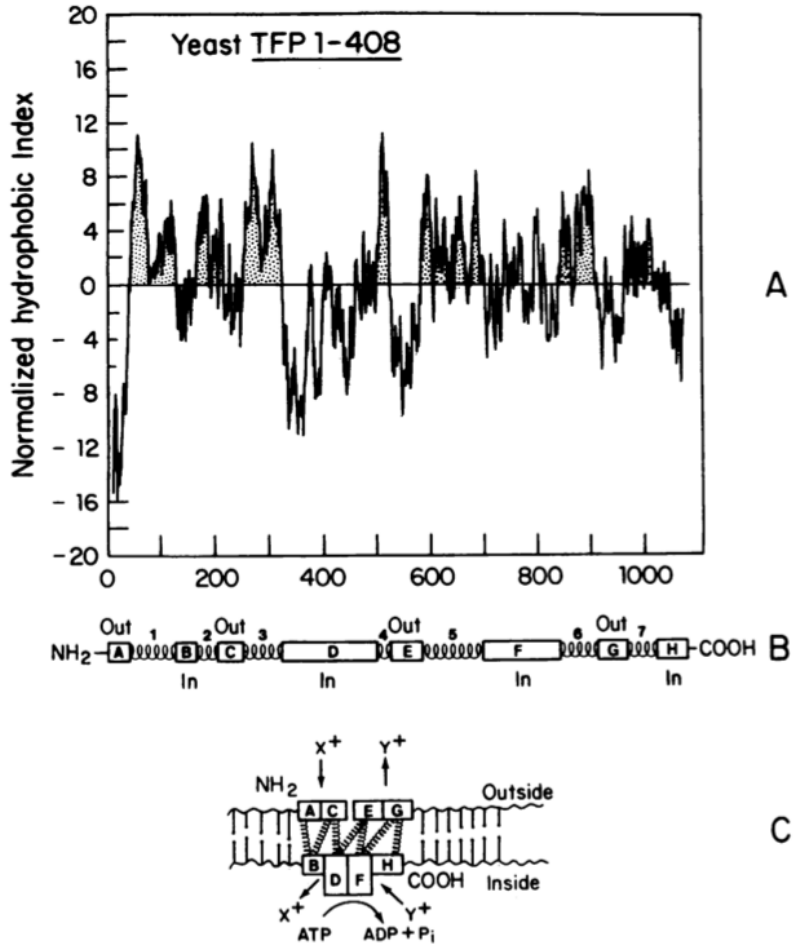
MOLECULAR AND CELLULAR BIOLOGY, Aug. 1988, p. 3094–3103
0270-7306/88/083094-10$02.00/0
Copyright © 1988, American Society for Microbiology
Vol. 8, No. 8

A Dominant Trifluoperazine Resistance Gene from *Saccharomyces cerevisiae* Has Homology with F_0F_1 ATP Synthase and Confers Calcium-Sensitive Growth

CHENG-KON SHIH, RONALD WAGNER, SABINE FEINSTEIN, CYNTHIA KANIK-ENNULAT, AND NORMA NEFF*

First published intein sequence, but not recognized as such.

(vacuolar ATPases catalytic subunit is not an integral membrane protein)



THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 265, No. 12, Issue of April 25, pp. 6726–6733, 1990
Printed in U.S.A.
© 1990 by The American Society for Biochemistry and Molecular Biology, Inc.

Molecular Structure of a Gene, *VMA1*, Encoding the Catalytic Subunit of H$^+$-Translocating Adenosine Triphosphatase from Vacuolar Membranes of *Saccharomyces cerevisiae**

(Received for publication, October 10, 1989)

Ryogo Hirata, Yoshinori Ohsumi‡, Akihiko Nakano, Hiroshi Kawasaki§, Koichi Suzuki§, and Yasuhiro Anraku¶

From the Department of Biology, Faculty of Science, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan and the §Department of Molecular Biology, The Tokyo Metropolitan Institute of Medical Science, Bunkyo-ku, Tokyo 113, Japan

```
S.cerevisiae (1071aa)   MAGAIENARKEIKRISLEDHAESEYGAIYSVSGPVVIAEN
                                                    * *********.**.
N.crassa     ( 607aa)              MAPQQNGAEVDGIHTGKIYSVSGPVVVAED

 41' MIGCAMYELVKVGHDNLVGEVIRIDGDKATIQVYEETAGLTVGDPVLRTGKPLSVELGPG
     *** ***********.********.**.***********. ********************
 31' MIGVAMYELVKVGHDQLVGEVIRINGDQATIQVYEETAGVMVGDPVLRTGKPLSVELGPG

101' LMETIYDGIQRPLKAIKEESQSIYIPRGIDTPALDRTIKWQFTPGKFQVGDHISGGDIYG
     *...*********. * *.*.********.******. **.*** ...*****.***..*
 91' LLNNIYDGIQRPLEKIAEASNSIYIPRGIATPALDRKKKWEFTP-TMKVGDHIAGGDVWG

161' SVFENSLISSHKILLPPRSRGTITWIAPAGEYTLDEKILEVEFDGKKSDFTLYHTWPVRV
     .*.***.** ********.*****.**  ****..************..... .******
150' TVYENSFISVHKILLPPRARGTITRIAEKGEYTVEEKILEVEFDGKKTEYPMMQTWPVRV

221' PRPVTEKLSADYPLLTGQRVLDALFPCVQGGTTCIPGAFGCGKTVISQSLSKYSNSDAII
     ***..** **. *.*.**********.*****. ****************.**.****.*.
210' PRPAAEKLSANQPLLTGQRVLDALFPSVQGGTVAIPGAFGCGKTVISQSVSKFSNSDVIV

281' YVGC--(454a.a.)--GERGNEMAEVLMEFPELYTEMSGTKEPIMKRTTLVANTSN
     ****               ***********..**** .*..* **********.*****
270' YVGC--           --GERGNEMAEVLKDFPELSIEVDGRKEPIMKRTTLIANTSN

779' MPVAAREASIYTGITLAEYFRDQGKNVSMIADSSSRWAEALREISGRLGEMPADQGFPAY
     ***************.********.**.*.*****************************
314' MPVAAREASIYTGITVAEYFRDQGMNVAMMADSSSRWAEALREISGRLGEMPADQGFPAY

839' LGAKLASFYERAGKAVALGSPDRTGSVSIVAAVSPAGGDFSDPVTTATLGITQVFWGLDK
     **************. ***** *.******.****.*********.*****.********
374' LGAKLASFYERAGKVQALGSPPREGSVSIVGAVSPPGGDFSDPVTSATLGIVQVFWGLDK

899' KLAQRKHFPSINTSVSYSKYTNVLNKFYDSNYPEFPVLRDRMKEILSNAEELEQVVQLVG
     ******************** ..*.*.*...**.** ****....**..***.*******
434' KLAQRKHFPSINTSVSYSKYLTILDKWYEREYPDFPRLRDRIRQLLSDSEELDQVVQLVG

959' KSALSDSDKITLDVATLIKEDFLQQNGYSTYDAFCPIWKTFDMMRAFISYHDEAQKAVAN
     ******.******.***************.**.******* **. ....*******.*.
494' KSALSDPDKITLDMATLIKEDFLQQNGYSDYDQFCPIWKTEWMMKLMMGFHDEAQKAIAQ

1019' GANWSKLADSTGDVKHAVSSSKFFEPSRGEKEVHGEFEKLLSTMQERFAESTD
      *.**.*. ..* *.. ...* **  ** *....  ..* .  * ..**. .*
554' GQNWNKVREATQDLQAQLKSLKFEVPSEGQEKICKKYEAIQQQMLDKFASVIDE
```

FIG. 6. **Amino acid sequence homology between the catalytic subunits of yeast and *Neurospora crassa* vacuolar membrane H$^+$-ATPases.** Identical residues and conserved amino acid replacements are indicated by *stars* and *dots*, respectively. *Dashes* represent gaps introduced to obtain maximum matching.

Protein Blot

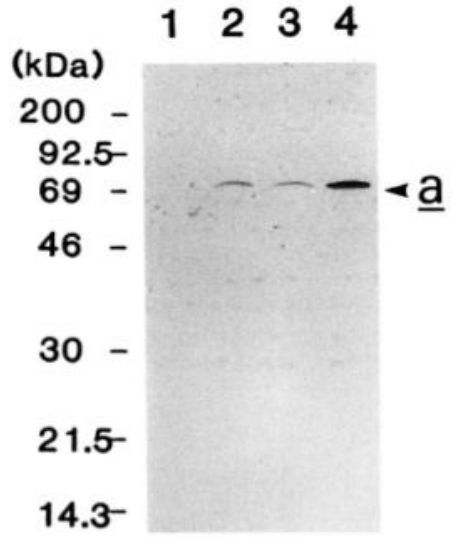


FIG. 5. **Lack of subunit *a* in total cell lysate from *vma1* null mutant cells.** Total cell lysate was prepared as described under "Materials and Methods" from RH101 (*vma1*, *lane 1*), ANY21 (*VMA1*, *lane 2*) and ANY21 harboring multicopy plasmid without or with *VMA1* gene (*lanes 3* and *4*). About 50 μg of proteins was separated on a 10% SDS-polyacrylamide gel and blotted onto a nitrocellulose membrane. Subunit *a* was detected using an anti-subunit *a* monoclonal antibody.

RNA Blot

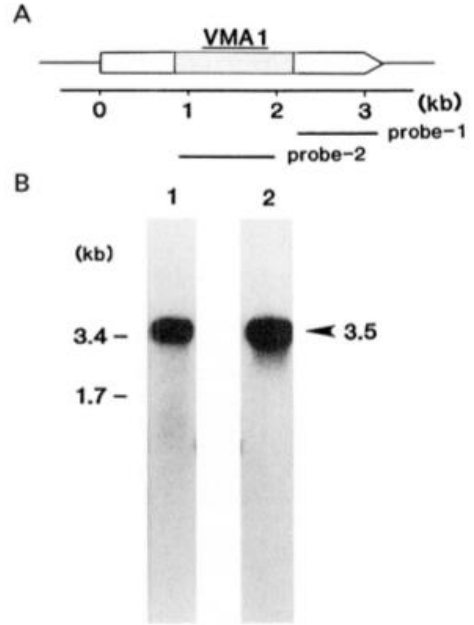
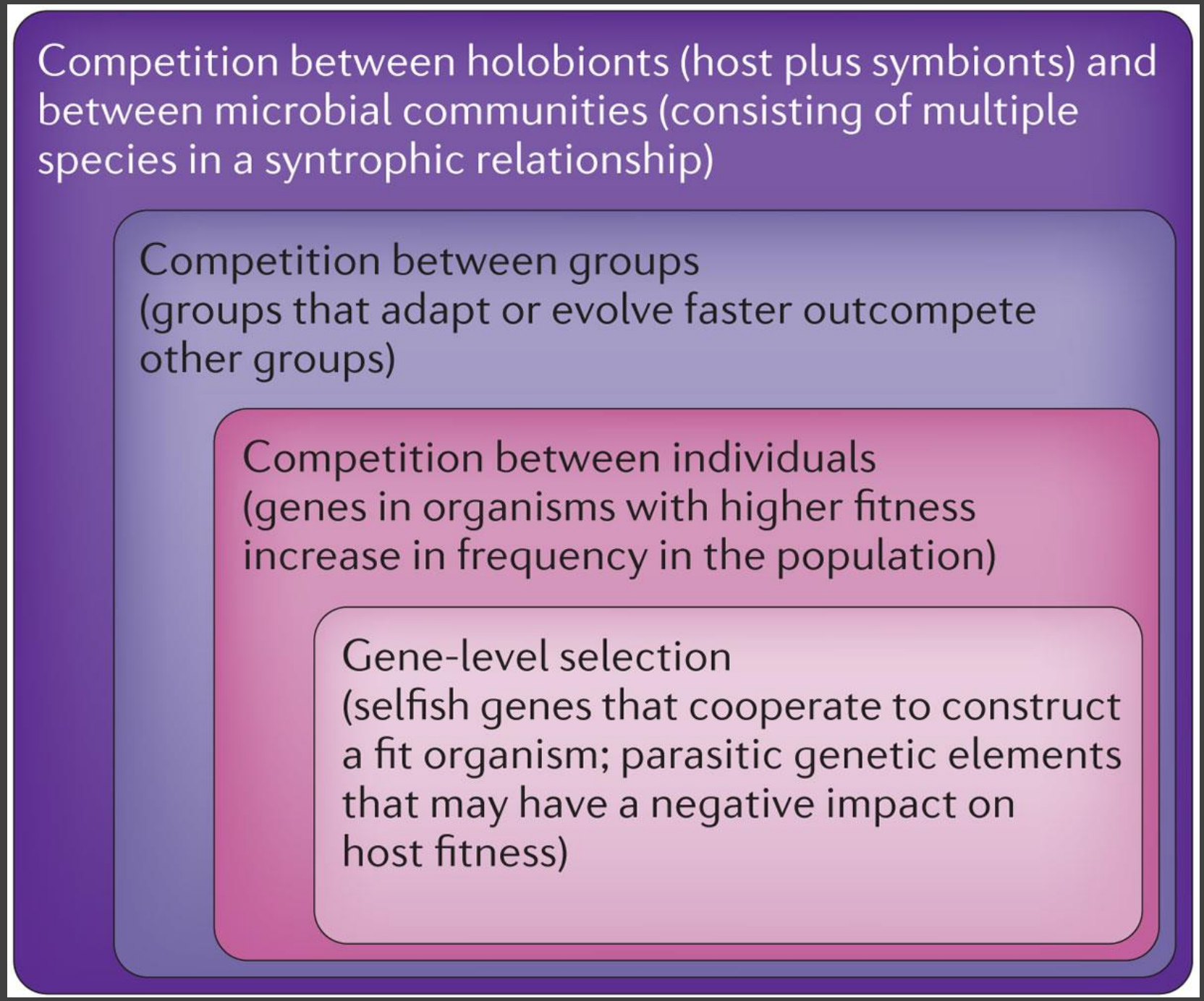


FIG. 8. **Northern blot analysis of the *VMA1* mRNA.** Poly(A)$^+$ RNA was isolated from ANY21, size fractionated on an agarose-formamide gel (1.2% gel), and transferred to a nylon membrane filter. Blots were hybridized with the DNA probes shown in *A*. *Hatched region* represents the nonhomologous insert (see "Results"). Hybridization patterns for probe-1 (*lane 1*) and probe-2 (*lane 2*) are shown in *B*. About 5 μg of RNA was loaded in each *lane*.

Levels of Selection

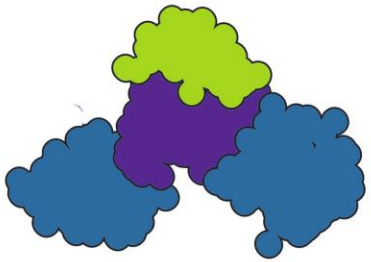


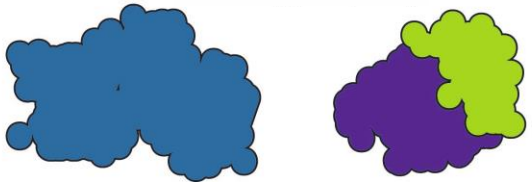
Inteins (molecular fleas): Self-Splicing Protein Mobile Elements

DNA

RNA

Protein





Intein Self-Splicing Domains

Auto catalytic splicing reaction removes the intein from the host protein

Intein Homing Endonuclease Domain

Recognizes unoccupied intein insertion sites (IIS) and disrupts the sequence with a double strand break

Extein

Host protein

Group I Intron

Exon: Intron: ORF on Intron

DNA

transcription

mRNA
or
rRNA

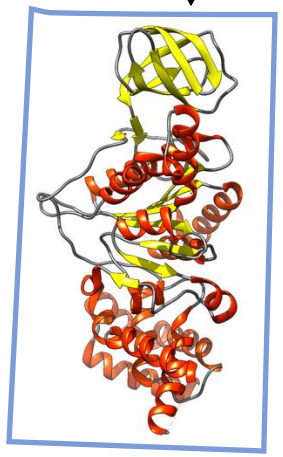
splicing

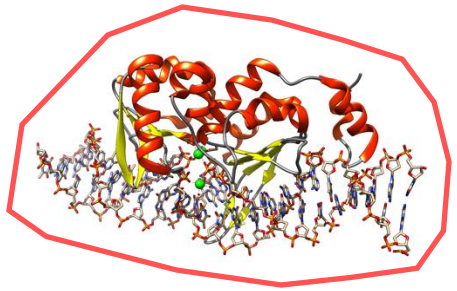
Exon

Intron with ORF

translation

Protein





Homing Endonuclease

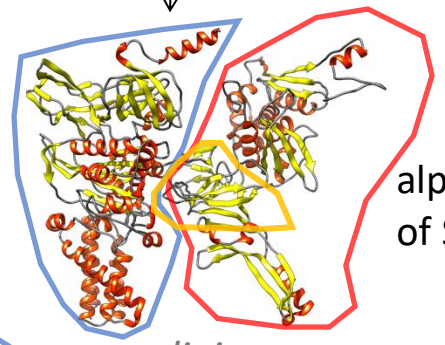
functional protein

Intein

Extein: Intein Splicing Domain: HE domain:

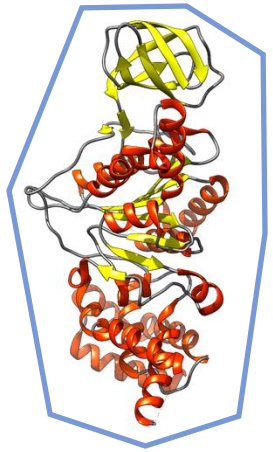
transcription

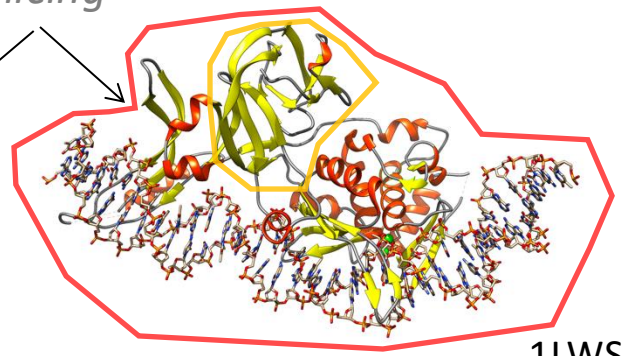
translation



alpha fold prediction
of Sce *vma*1 gene product

splicing





1LWS.pdb

functional protein

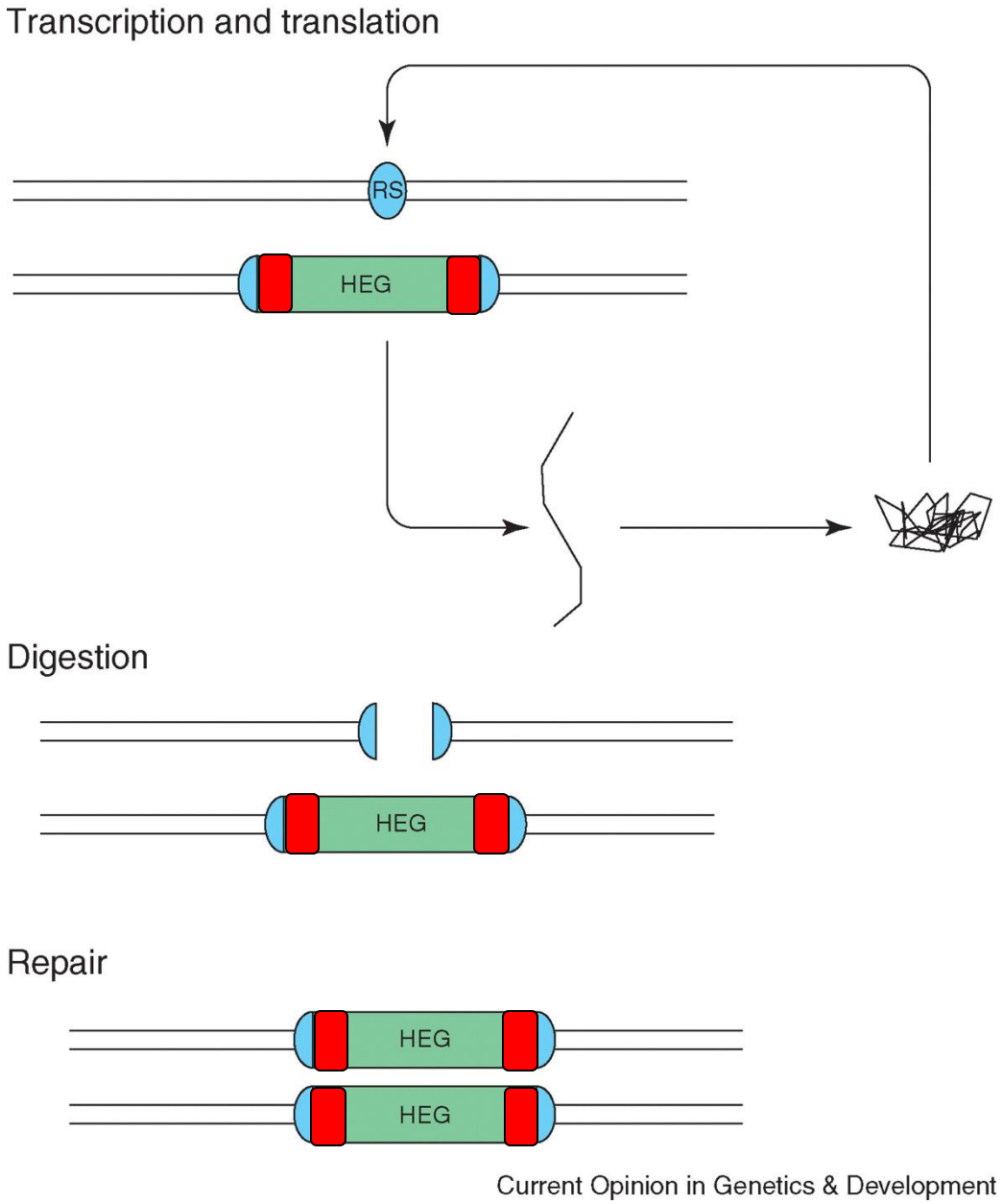
Intein with HE and
Splicing domain

Intein Homing

Intein = **Int**ervening Prot**ein** sequence
= “protein intron”

Self Splicing Domain

Homing Endonuclaease Domain



Modified from

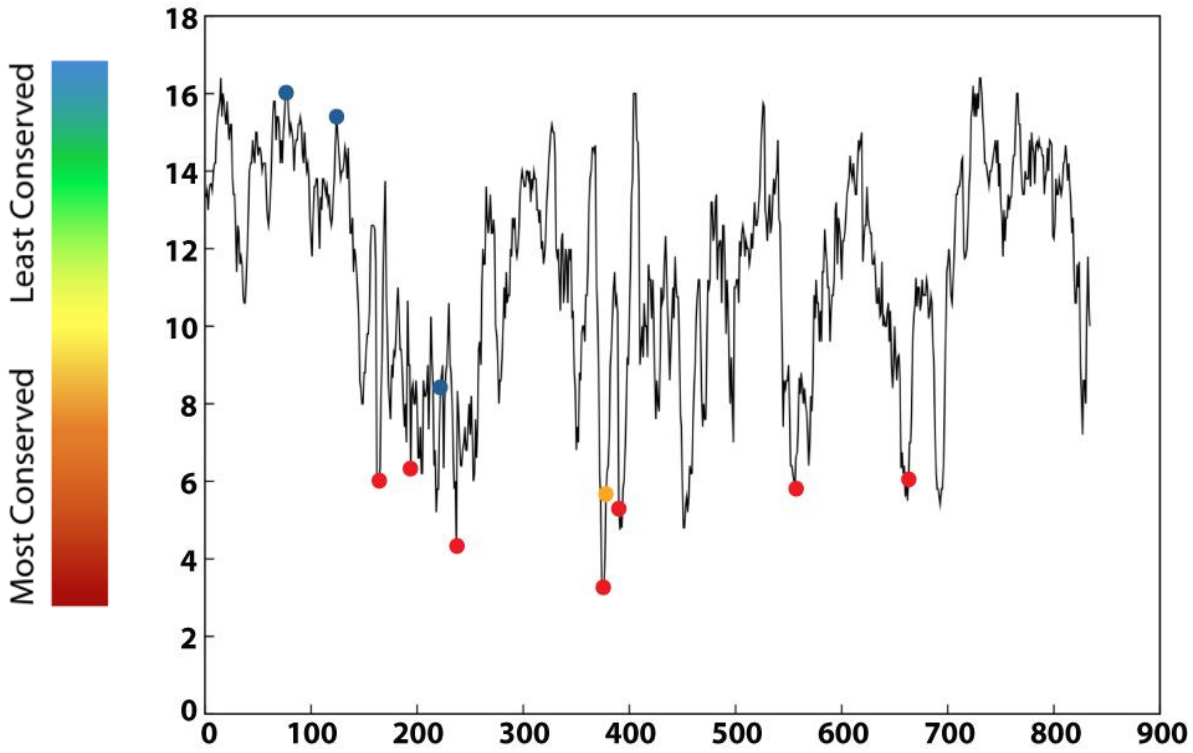
Homing endonuclease genes: the rise and fall and rise again of a selfish element

Austin Burt and Vassiliki Koufopanou

Current Opinion in Genetics & Development

Volume 14, Issue 6, December 2004, Pages 609-615

Ribonucleotide Reductase



- Intein
- Group I Introns
- Group II Introns

Swithers KS, Senejani AG, Fournier GP, Gogarten JP (2009)
Conservation of intron and intein insertion sites: Implications for life histories of parasitic genetic elements.
BMC Evolutionary Biology 2009, 9:303 doi:10.1186/1471-2148-9-303 (Highly Accessed)

In case of inteins the co-evolution of genes results in

- **Mutualism** between splicing and homing endonuclease domain

- **Commensalism** between host protein and intein without HE domain

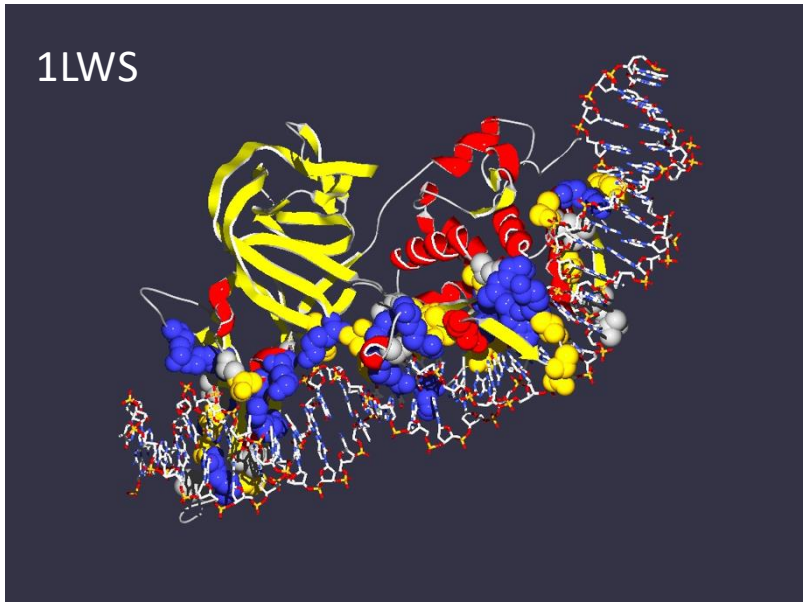
- **Parasitism** between host protein and intein with HE

Encoding

Extein = Host Protein

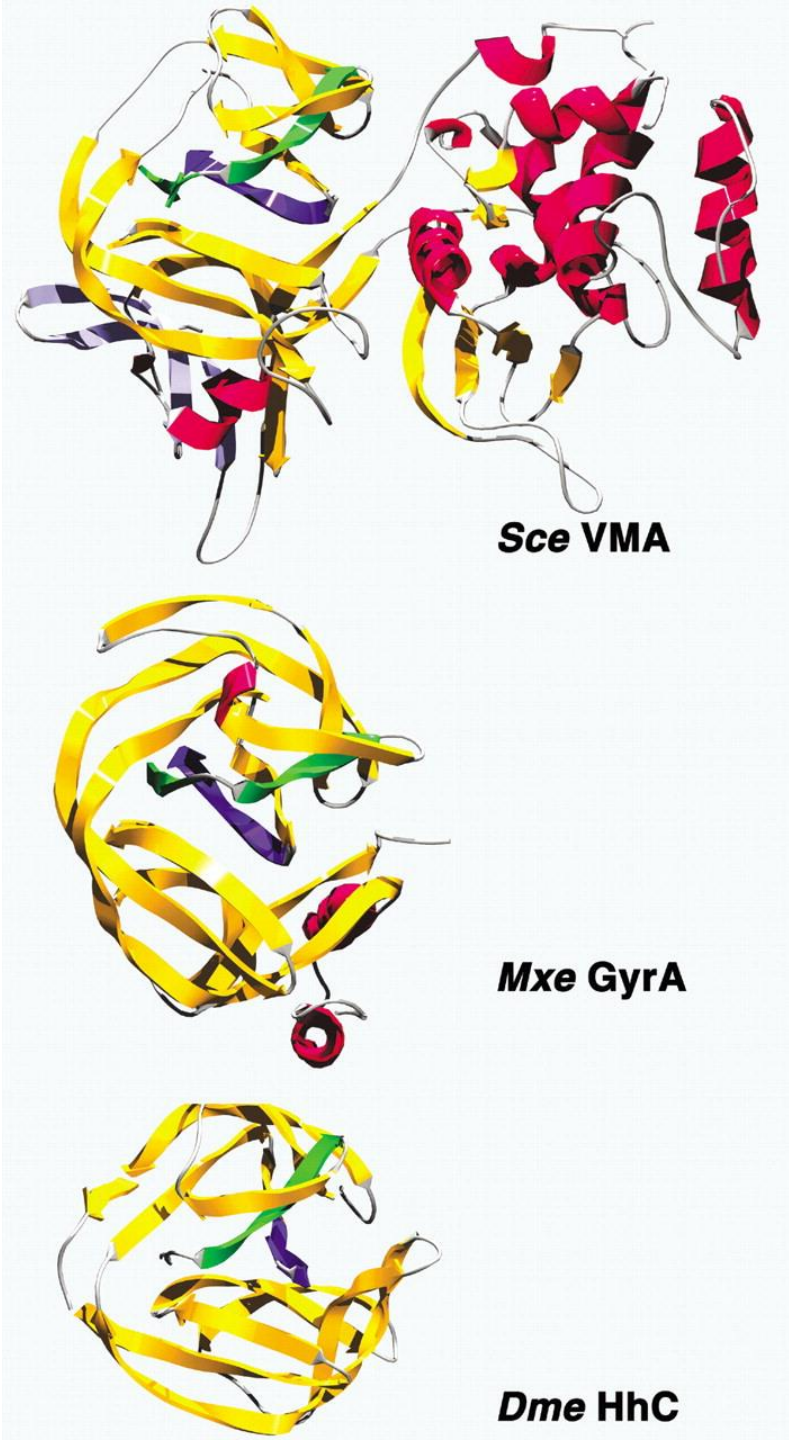
Intein Splicing Domain

Homing Endonuclease (HE) domain



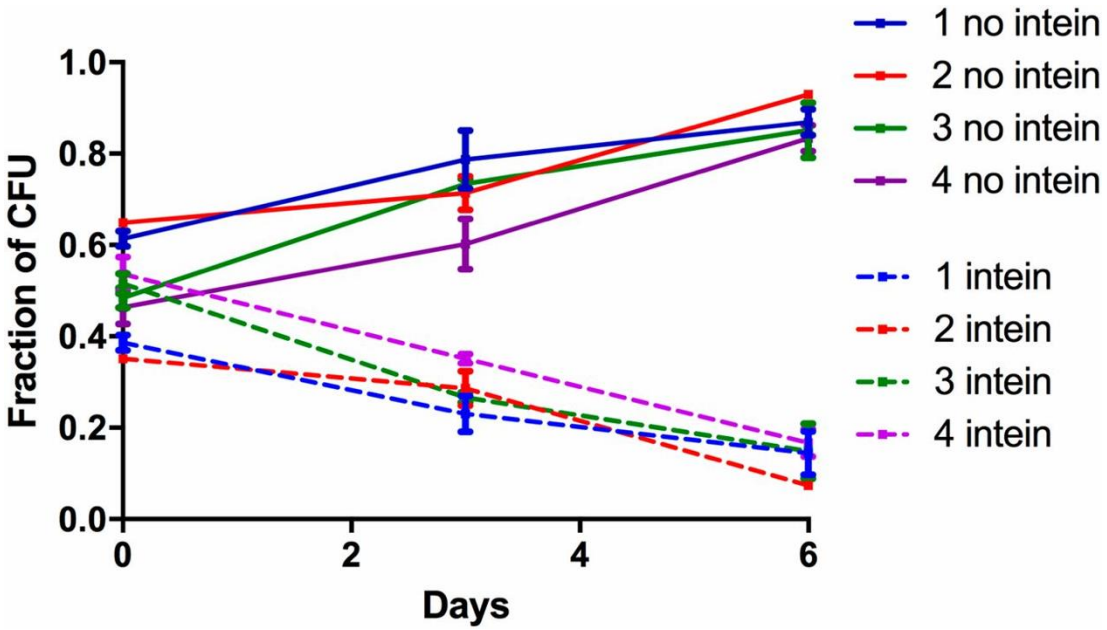
Structure from Moure et al. (2002) Nat.Struct.Biol. 9: 764-770

Comparison of intein and hedgehog protein structures. The structures of the *Sce* VMA, *Mxe* GyrA, and the autoprocessing domain of the hedgehog protein from *Drosophila melanogaster* have been determined by X-ray crystallography. The pdb-files were retrieved from the Protein Data Bank and processed using the Swiss-PdbViewer. The *orange arrows* and *red helices* indicate β-sheets and α-helices. The amino-terminal β-sheets are colored in *green*, and the carboxy-terminal β-sheets are shown as *blue arrows*. The endonuclease domain present in *Sce* VMA (*top panel*, *right*) forms a domain clearly distinct from the self-splicing domain (*top panel*, *left*). The part of the *Sce* VMA structure that is not part of the endonuclease domain, but partakes in DNA binding, is depicted in *light blue*.



Inteins have a high fitness cost for the host organism.

United States – Israel
Binational Science Foundation



Competition Experiment between Intein + and Intein – strains (otherwise isogenic)

From: Impact of a homing intein on recombination frequency and organismal fitness. Naor et al.
doi: 10.1073/pnas.1606416113

Dr. Adit Naor
Tel Aviv Univ / Stanford

Dr. Shannon Soucy
UConn / Dartmouth

Dr. Uri Gophna
Tel Aviv Univ

Dr. Thane Papke
UConn

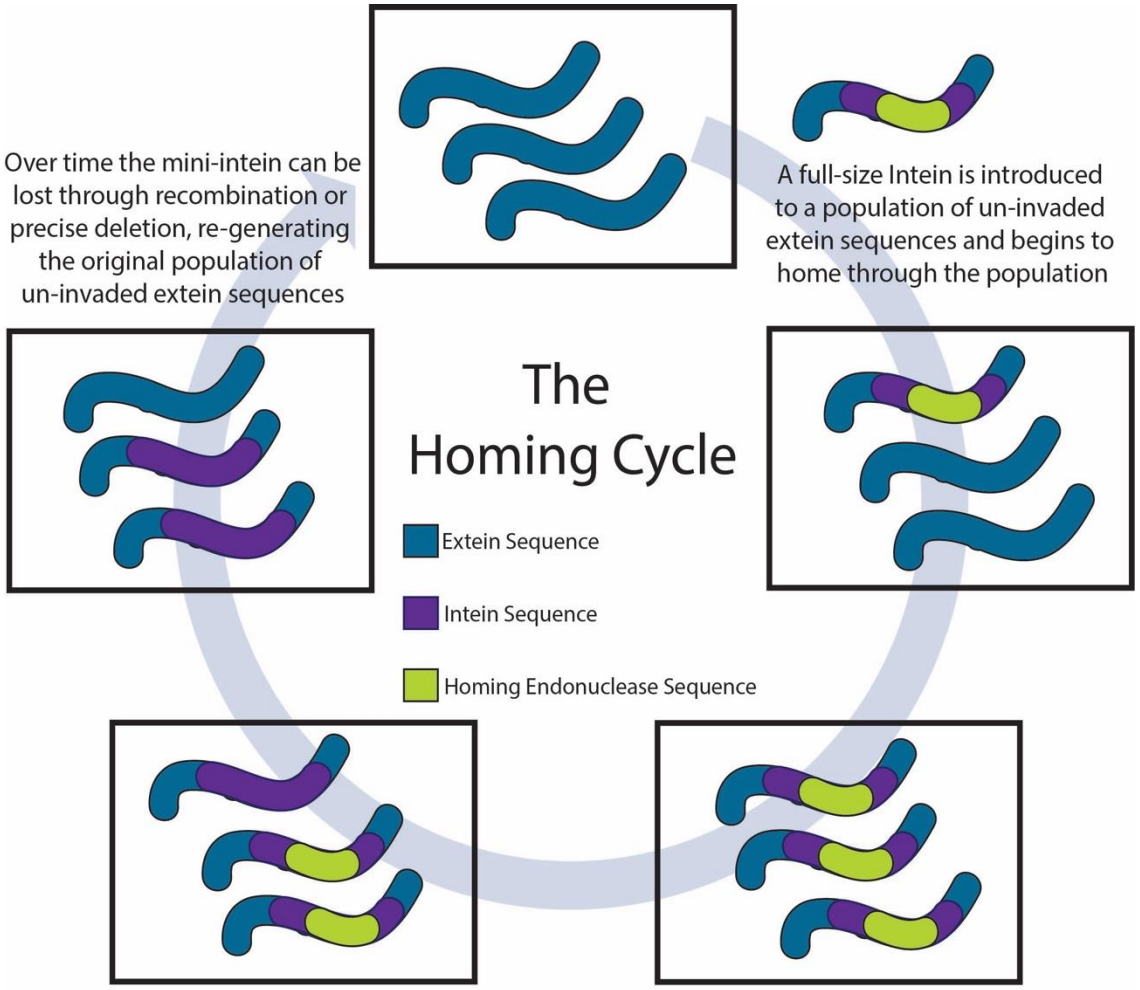
How can inteins with functional homing endonuclease survive on the long run?

- Homing cycle
- Coexistence of the three forms (preditor prey with three partners, in an intransitive fitness relationship)
- Diverse environment with regions that select for the intein, and those that don’t

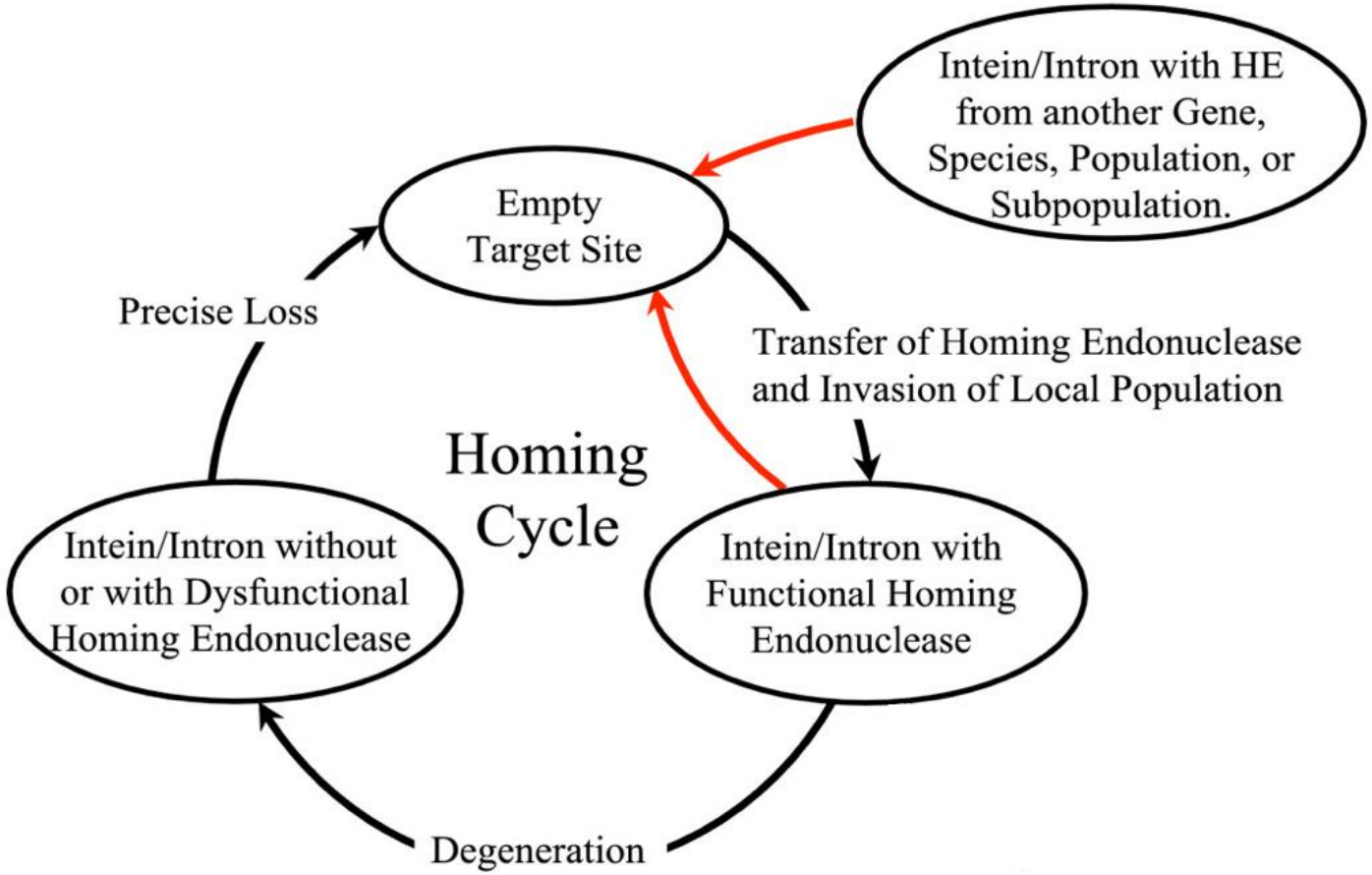
Three models for long term intein persistence

1) the homing cycle

The Homing Cycle



Homing Cycle of an Intein or Group 1 Intron



Local population with **only empty target sites**

Invasion of target sites

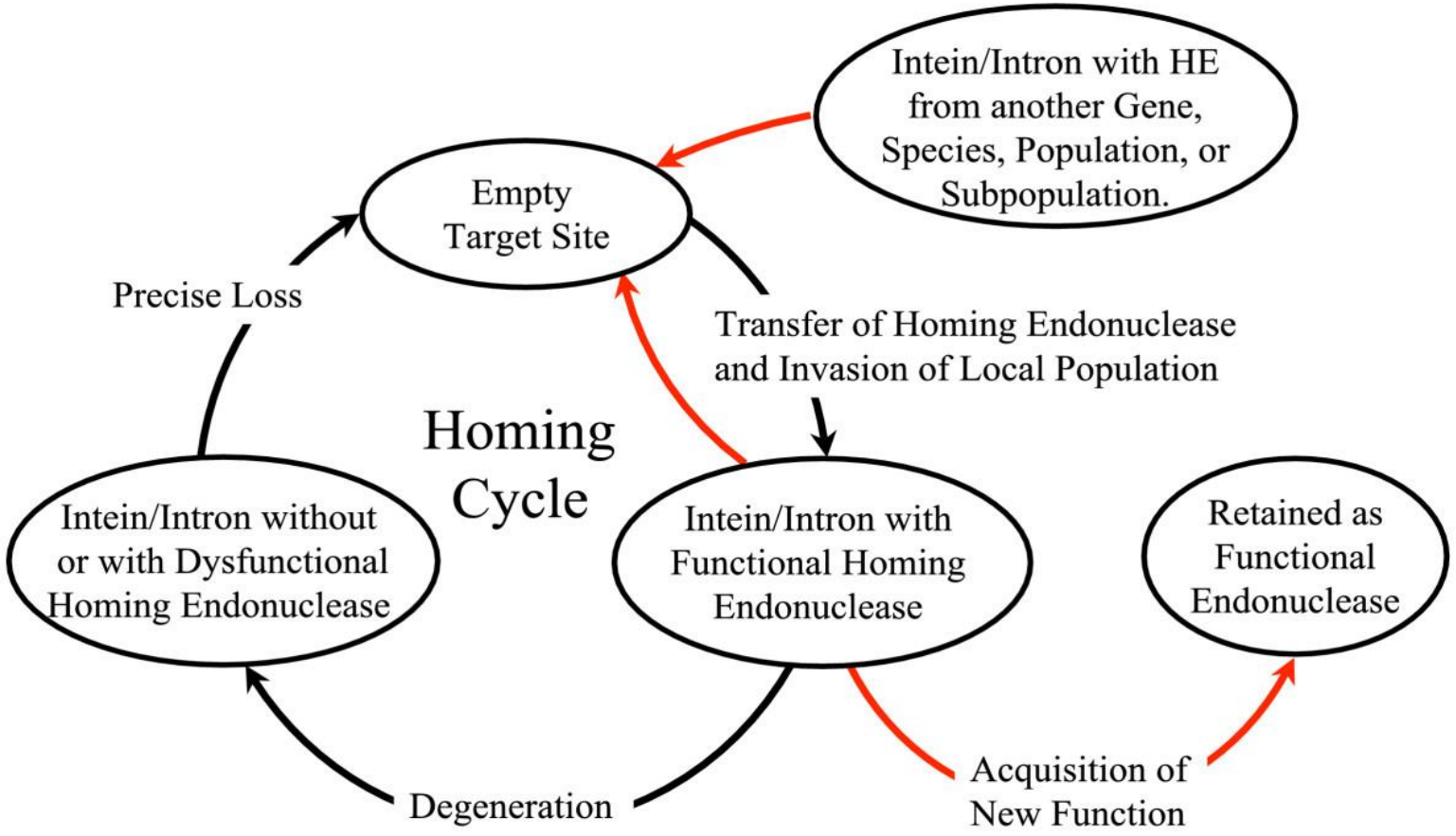
All sites are are occupied with **full intein**

Degeneration of Homing Endonuclease

Only Mini-Inteins are present

Loss – Difficult in case of inteins.
If invasion was complete,
no intein free template nucleotide
sequence is available as template

Each phase (invasion, degeneration, loss) goes to completion, before the next begins.



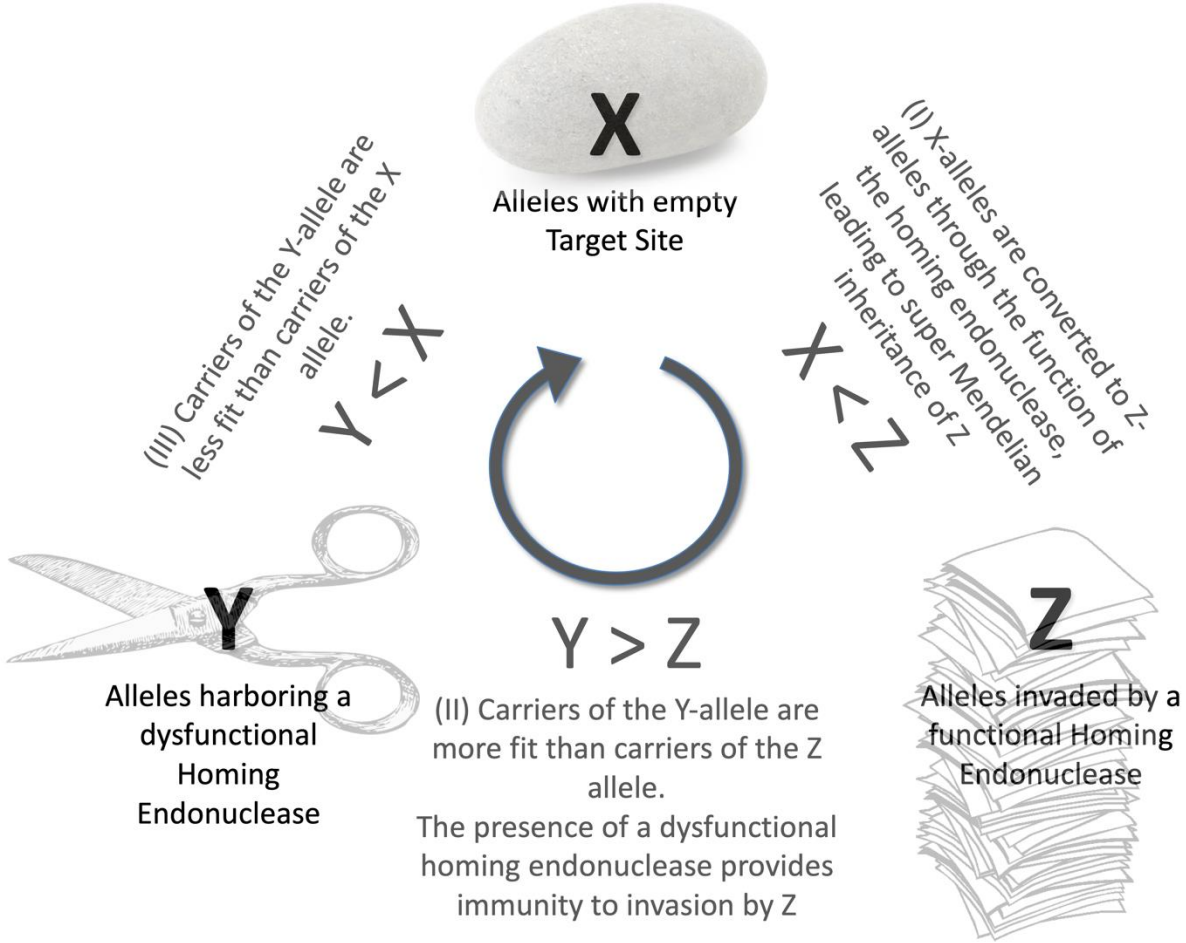
Homing cycle of a parasitic genetic element (modified from [3, 13]). Recent findings suggest that due to complex population structure the cycle might not operate in synchrony in different subpopulations. The red arrows indicate the trajectory of the functioning HE and the black arrows the fate of the host gene. The precise loss can occur through recombination with an intein or intron free allele, or, in case of introns, through recombination with a reverse transcript of the spliced mRNA [39, 40].

Gogarten, J. Peter and Elena Hilario (2006)
Inteins, introns, and homing endonucleases: Recent revelations about the life cycle of parasitic genetic elements.
BMC Evolutionary Biology 6:94 Open Access (Highly Accessed)

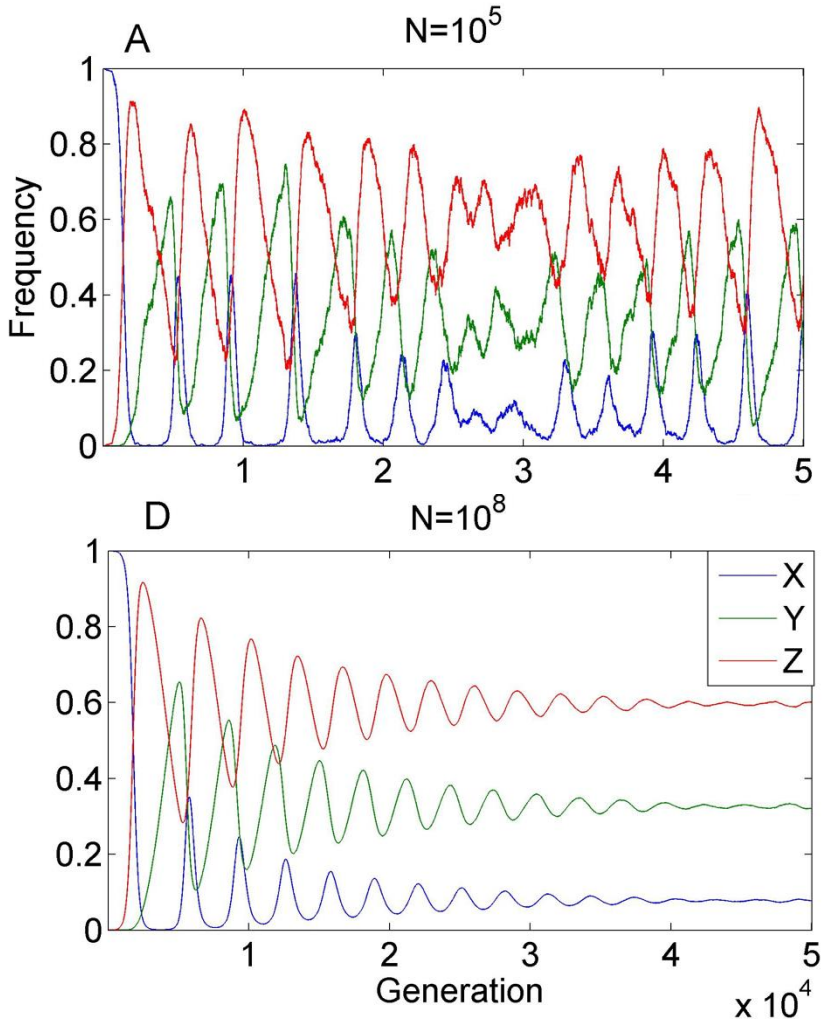
Three models for long term intein persistence

1) the homing cycle
2) Co-existence of the three allele types in a homogeneous population

Coexistence of the three forms



A Molecular Rock-Paper-Scissors game



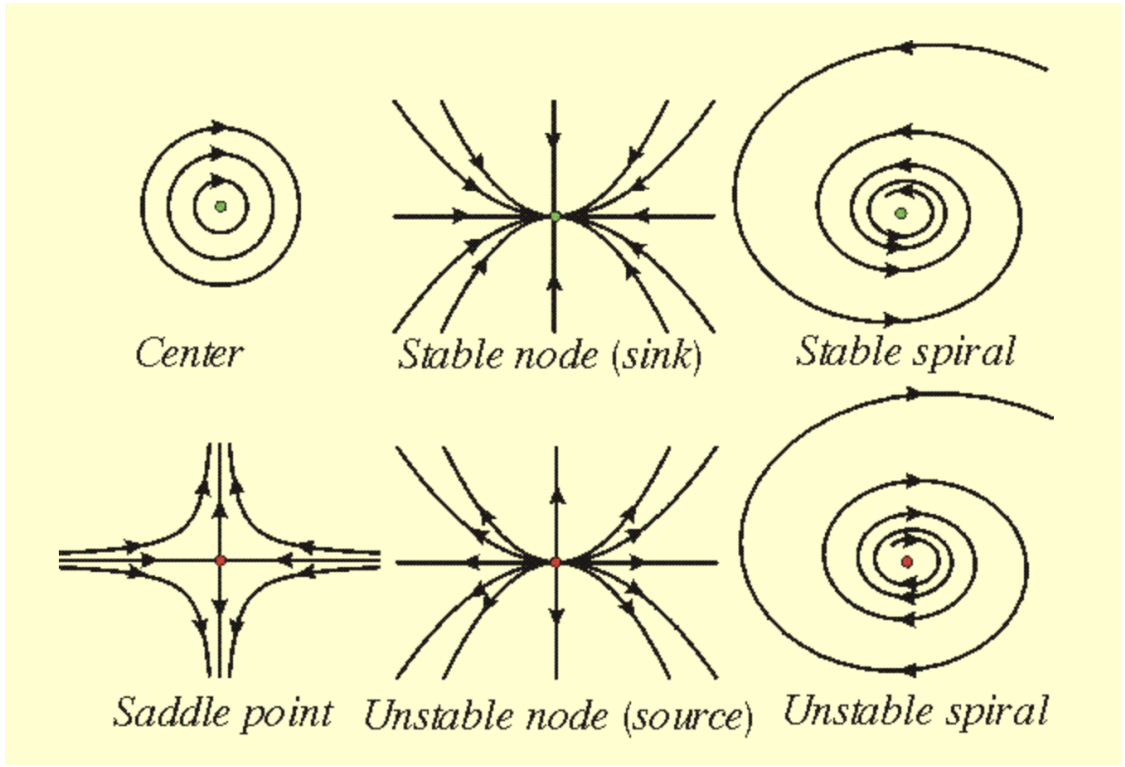
Simulations using difference-equations in populations of limited size

Barzel *et al. BMC Evolutionary Biology* 2011 **11**:324 doi:10.1186/1471-2148-11-324

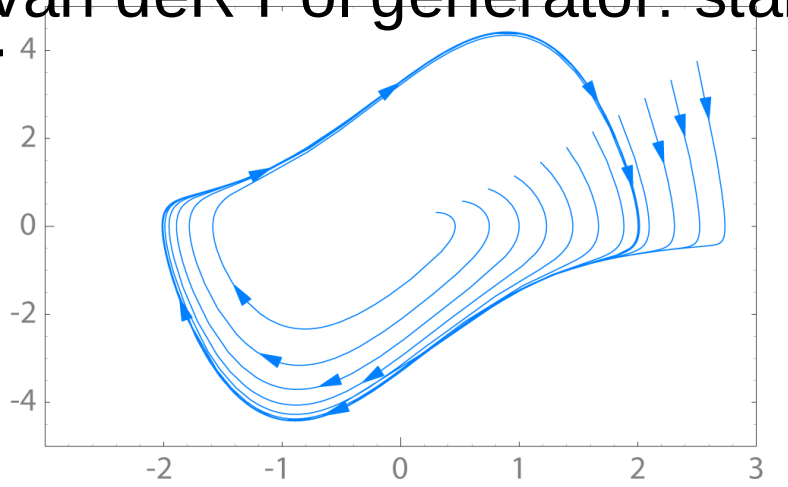
In a homogeneous well mixed population, this equilibrium can either be a limit cycle or a stable node.

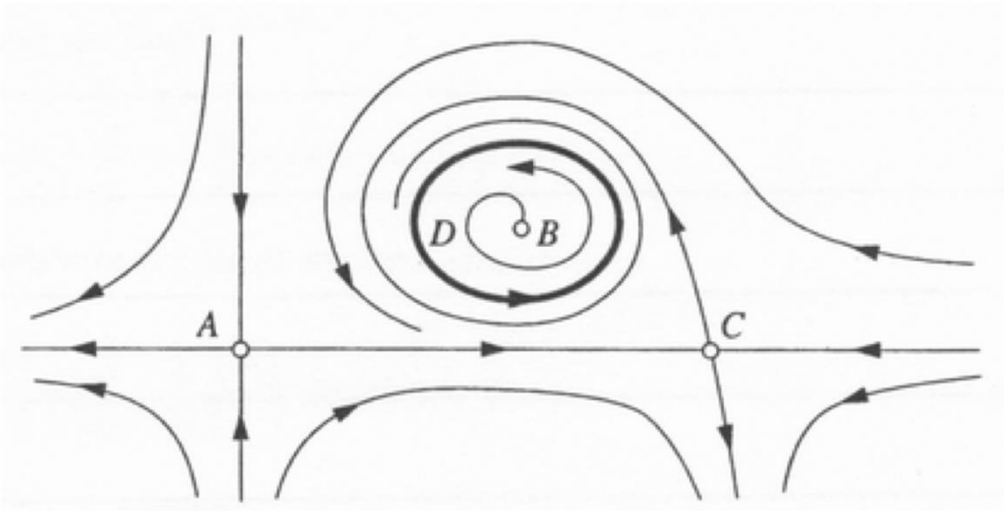
See: Yahara K, Fukuyo M, Sasaki A, Kobayashi I. Proc Natl Acad Sci U S A 106(44):18861-6
Barzel A, Obolski U, Gogarten JP, Kupiec M, Hadany L. BMC Evol Biol 11:324.

Aside: recall types of st



van deR Pol generator: stable limit c





From:

http://www.math.nthu.edu.tw/~sbhsu/1.3Linearization.pdf
http://staff.www.ltu.se/~larserik/applmath/chap9en/part7.html
https://en.wikipedia.org/wiki/Van_der_Pol_oscillator

Fixed points, like *A* , *B* , and *C,* satisfy $dfx_1/dt = dfx_2/dt$ 0, and correspond
to steady states or equilibria of the system. The ***closed orbits***, like *D* correspond to periodic solutions.

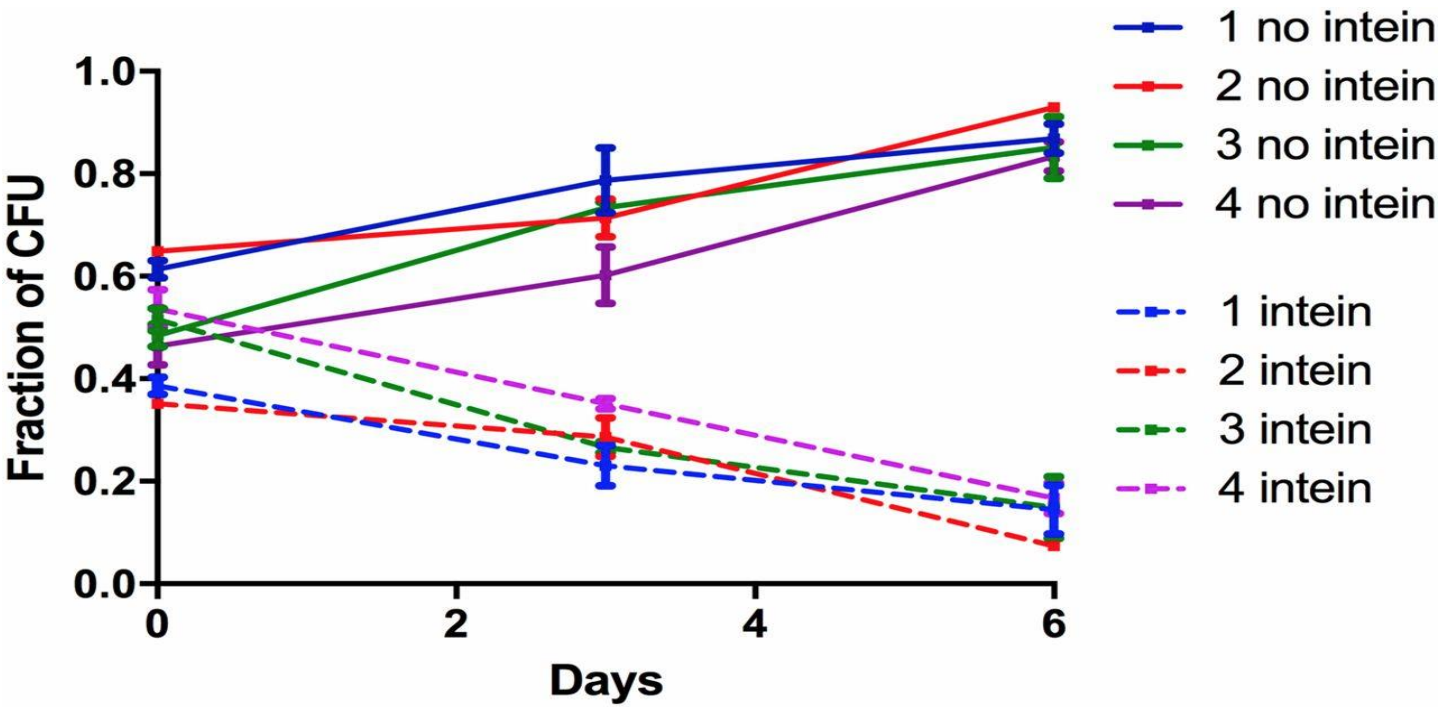
Three models for long term intein persistence

1) the homing cycle
2) Co-existence of the three allele types in a homogeneous population
3) Co-existence of the intein with homing endonuclease (P+) and the allele with an empty target site (P-) in a heterogenous environment

Co-existence in a **heterogenous** population

Arguments in favor for a co-existence in a heterogenous population

1) the intein can have a strong negative effect on organismal fitness



Competition Experiment between Intein + and Intein – strains (otherwise isogenic)
From: Impact of a homing intein on recombination frequency and organismal fitness. Naor et al. [doi: 10.1073/pnas.1606416113](https://doi.org/10.1073/pnas.1606416113)

Co-existence in a heterogenous population

Arguments in favor for a co-existence in a heterogenous population

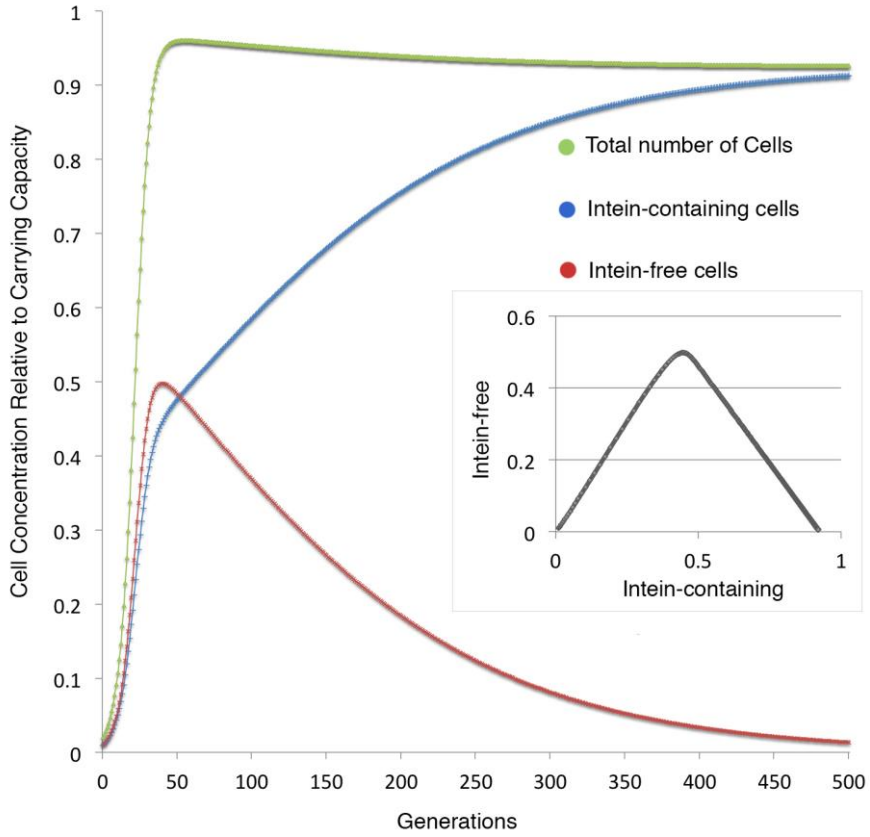
1) the intein can have a string negative effect on organismal fitness
2) Due to the large fitness cost the intein – alleles dominate during initial phase of the growth curve

On the right is a simple logistic growth curve
(with discrete generations) generated with an excel spreadsheet
The simulation starts with .01 intein + and .01 intein - cell

At the beginning of the growth curve
the intein - cells outcompete the intein + cells because
the intein - cells grow faster

At the end of the growth curve
the intein – are converted into intein + through HGT and
homing (close to carrying capacity, there is little growth,
the fitness cost does not play a big role

The carrying capacity (available resources) are the same
for both; with intein + cells requiring more resources



Intein invasion dynamics in a population with carrying capacity. Number of cells relative to the carrying capacity is plotted for the total number of cells (green line), the intein-containing cells (blue line), and the intein-free cells (red line) over 500 generations. The inset panel shows the concentration of intein-free alleles plotted against the concentration of intein-containing alleles. The simulation assumes that the fitness cost lowers the growth rate as well as the carrying capacity of the intein-containing cells. Parameters: homing efficiency = 0.01; fitness cost = 0.075; carrying capacity 1; growth rate 0.2 per generation; starting concentration 0.01 each for the intein-containing and intein-free cells.

Summary of NetLogo Procedures

Procedure	Description	Value	
Movement	Microbes move fixed Move_Distance in a random direction. Movement occurs at intervals of Move_Interval ticks.	Move_Distance = [0.1, 1] (patch-lengths) Move_Interval = [1, 10] (ticks)	
Resource Gathering	Microbes advance Replication_Timer in increments of 1, divided by the local cell density.	1 / (count turtles in radius 0.5)	
Replication	Microbes divide if their Replication_Timer exceeds Generation_Time. Intein expression results in an 8% fitness cost.	Generation_Time = [100 (Int-), 108 (Int+)] (ticks)	
Mating	Microbes co-localized to a set radius fuse according to Mating_Efficiency.	Mating_Efficiency = [1.00E-4 – 2.50E-4]	
Homing	Int(-) microbes are converted to Int(+) while mating with an Int(+) microbe based on Homing_Efficiency.	Homing_Efficiency = 0.68	
Death	Microbes die at random based on intrinsic Death_Rate, and die probabilistically based on zonal Lethality, at intervals of Dieoff_Frequency.	Death_Rate = 0.001 Dieoff_Frequency = 500 (ticks) Lethality = [0.20, 0.40, 0.60]	

Explains what occurs during every tick of each simulation

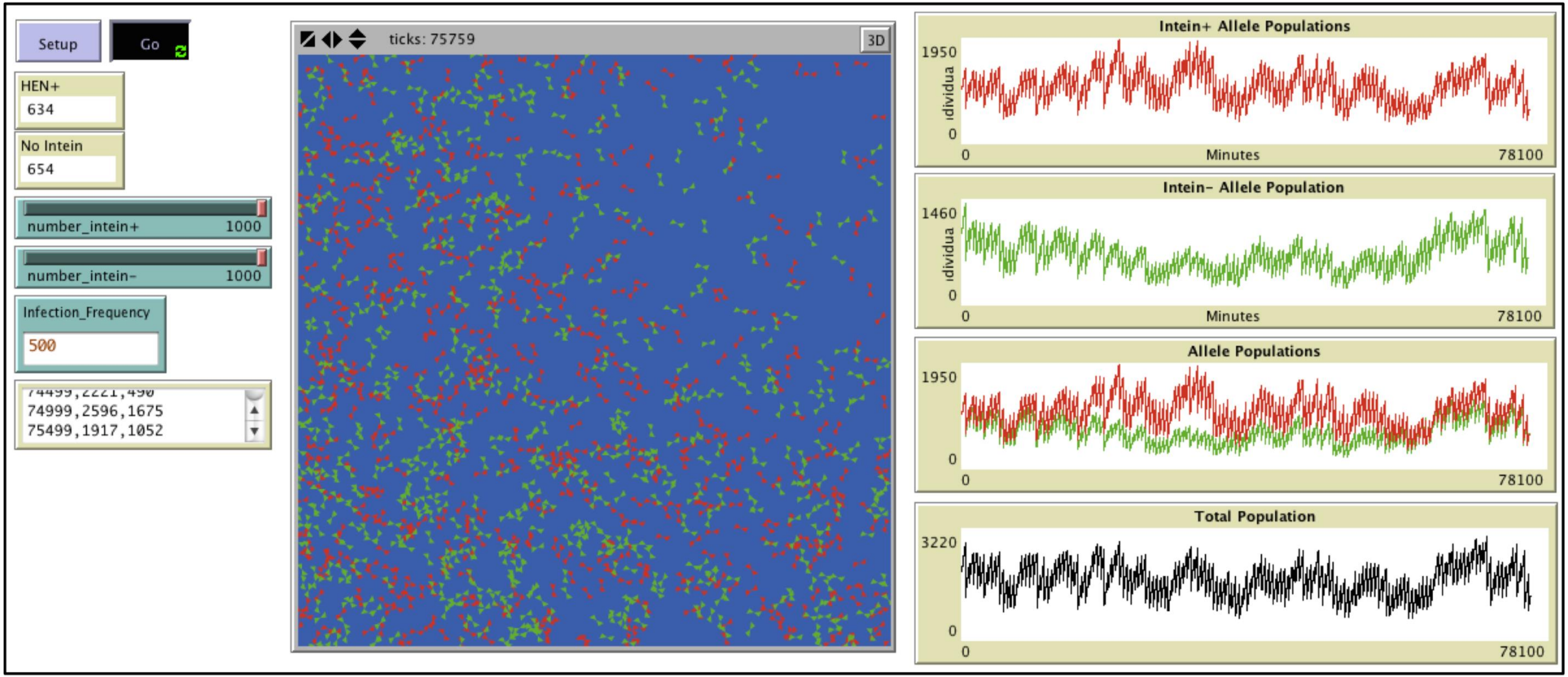
Somewhat wordy in its current form. May not be necessary with inclusion of a “Methods” section

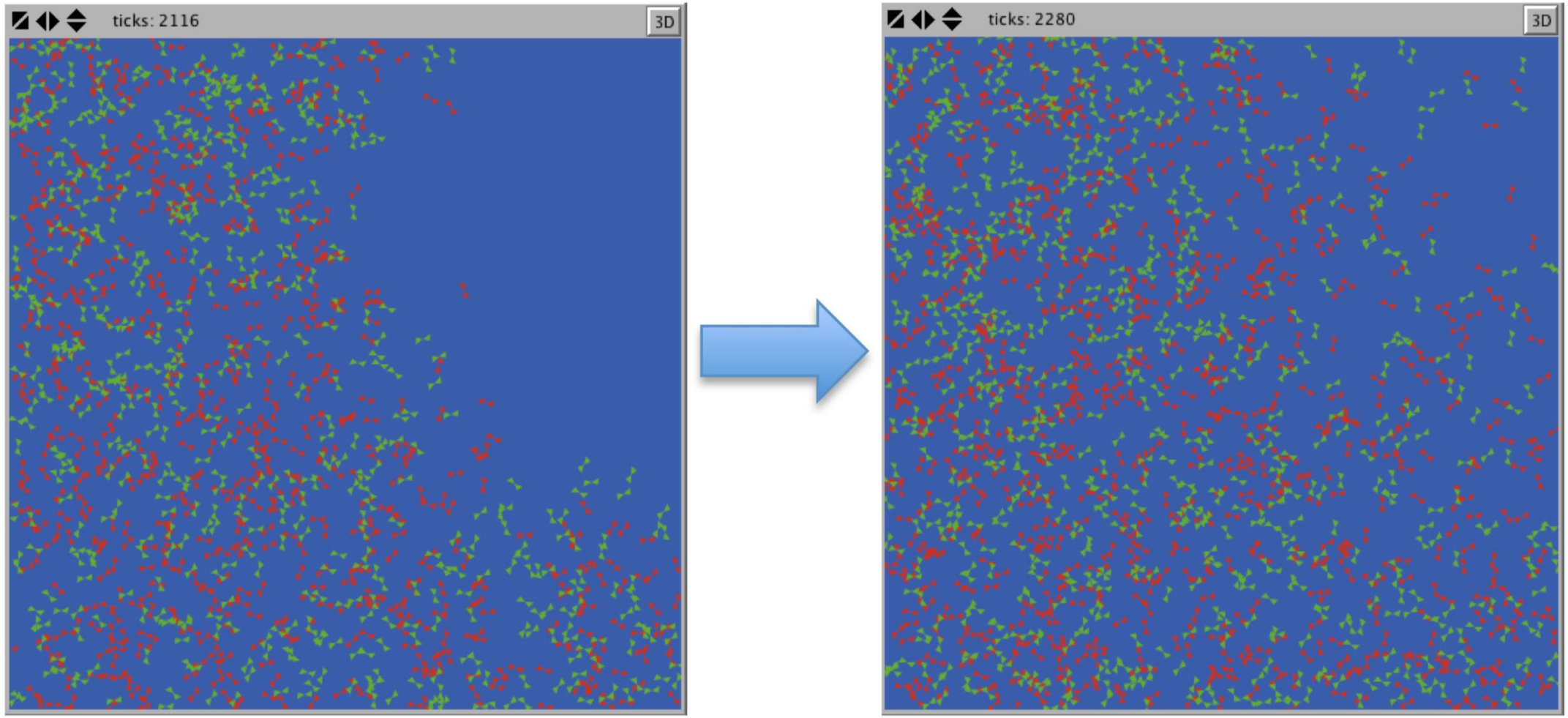
T.B.C. - In current form, doesn’t include fringe experiments (e.g. transfer rate between two environments, die-offs in random radii, etc.)

First Netlogo Experiments

Initially, we (i.e., Josh Skydel) explored a single environment into which local extinction events were placed at random

Agent Based Modeling of Intein-containing Populations



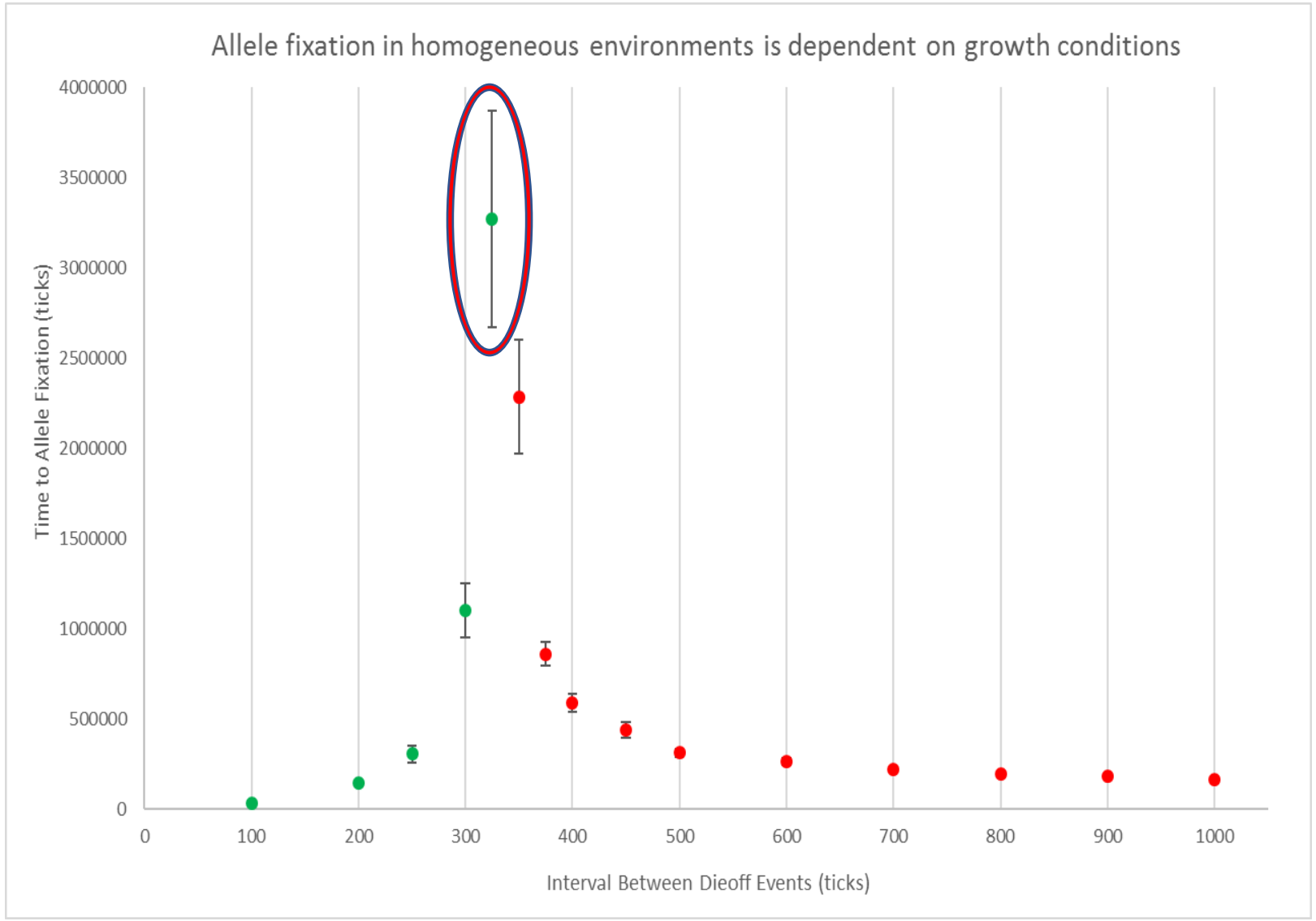


Viral infections create local environments of reduced cell density. Organisms in the modeled space are sensitive to the cell density of their surroundings in respect to their ability to replicate. Viral infections periodically kill all individuals in a randomly located radius. The cell density in that area is gradually restored as organisms migrate into the area; these individuals also gain a replicative “boost” from the lowered cell density in the region.

Sadly, this worked less well than expected
(the PI's wishful thinking failed).

While a set of parameters existed that gave rise to longer term coexistence; however, this "equilibrium" was unstable.

The asymptotically stable solutions had either the intein containing allele, or the intein free allele surviving.

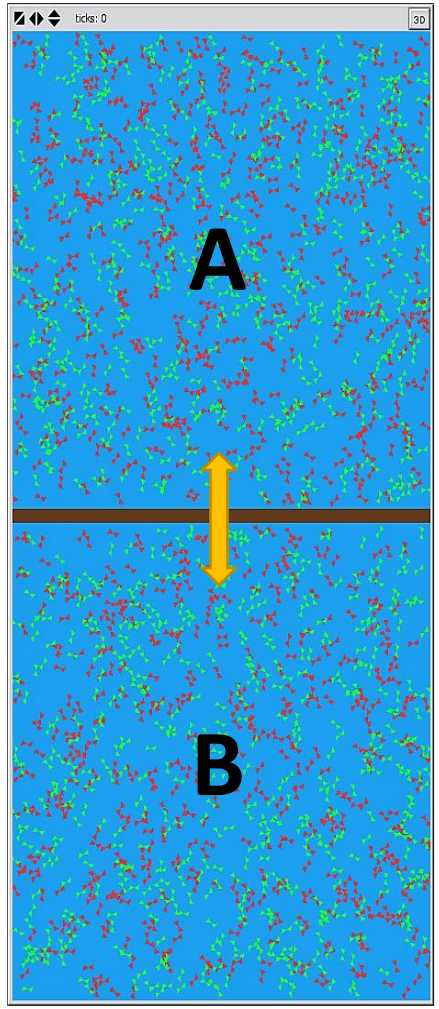


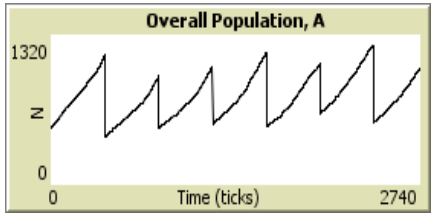
Environment without barrier and random, localized extinction events.

A solution was to have two distinct environments linked by migration between the two environments.

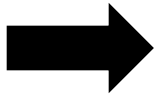
In one environment the intein plus alleles are favored, in the other the intein minus alleles

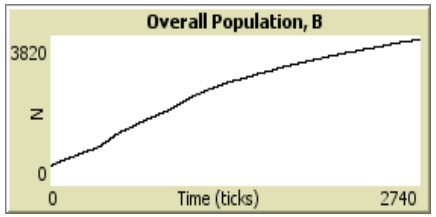
Agent-Based Modeling of Microbial Communities



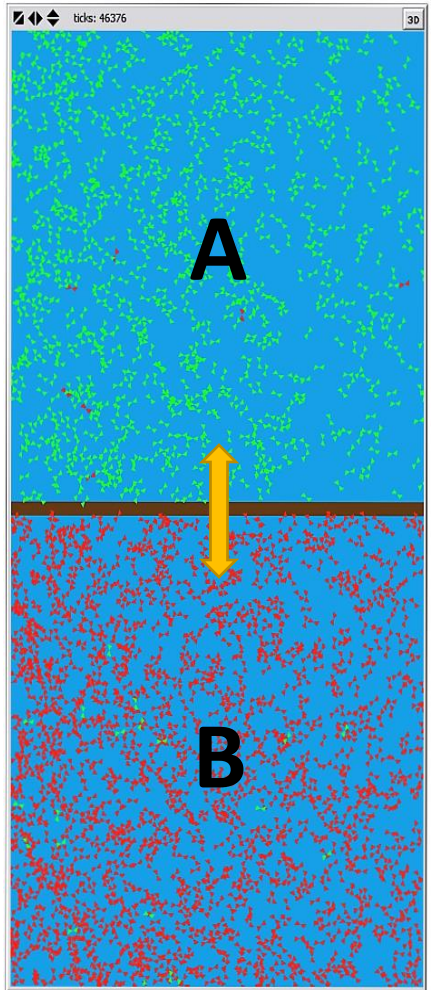


Population held in exponential phase





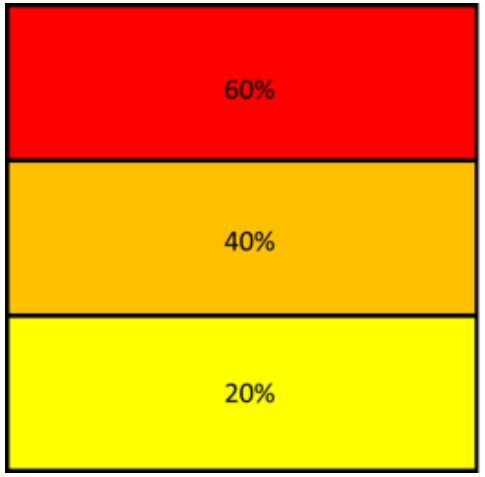
Population enters stationary phase



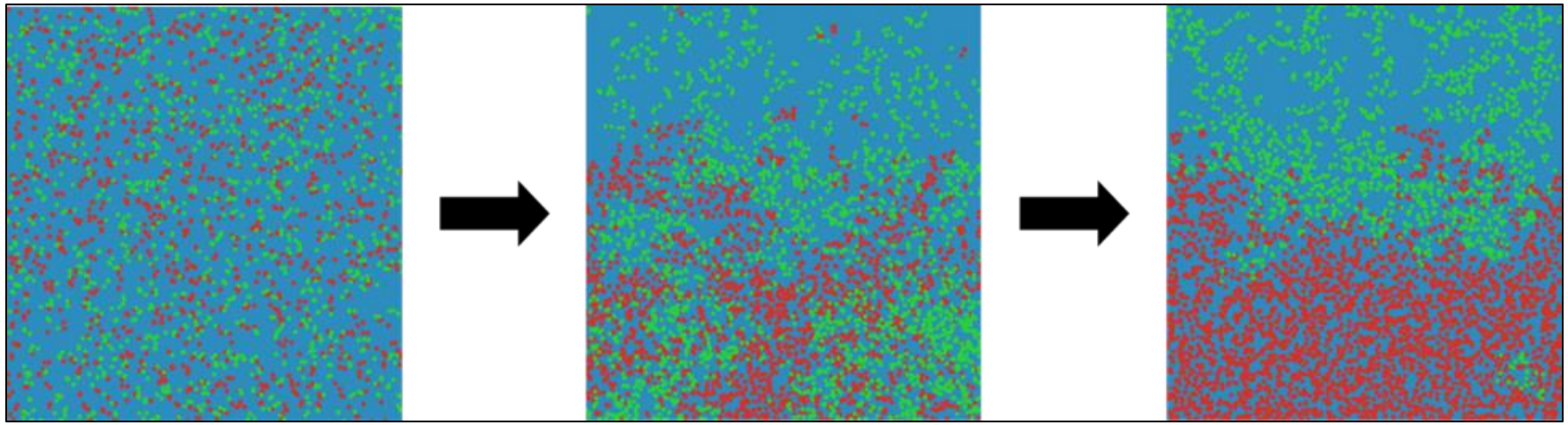
- The agent-based modeling platform NetLogo was used to assess the population dynamics of molecular parasites known as inteins
- Simulated microbial agents independently move, acquire resources, replicate, mate, and die
- User-defined parameters influence cellular movement, mating frequency, and community growth state
- The ultimate distribution of inteins was dependent on whether the community spent more time in exponential or stationary phase
- Limited transfer between sub-environments promoted limited allele coexistence

One environment with a gradient in efficiency of local extinction events

Extinction efficiency



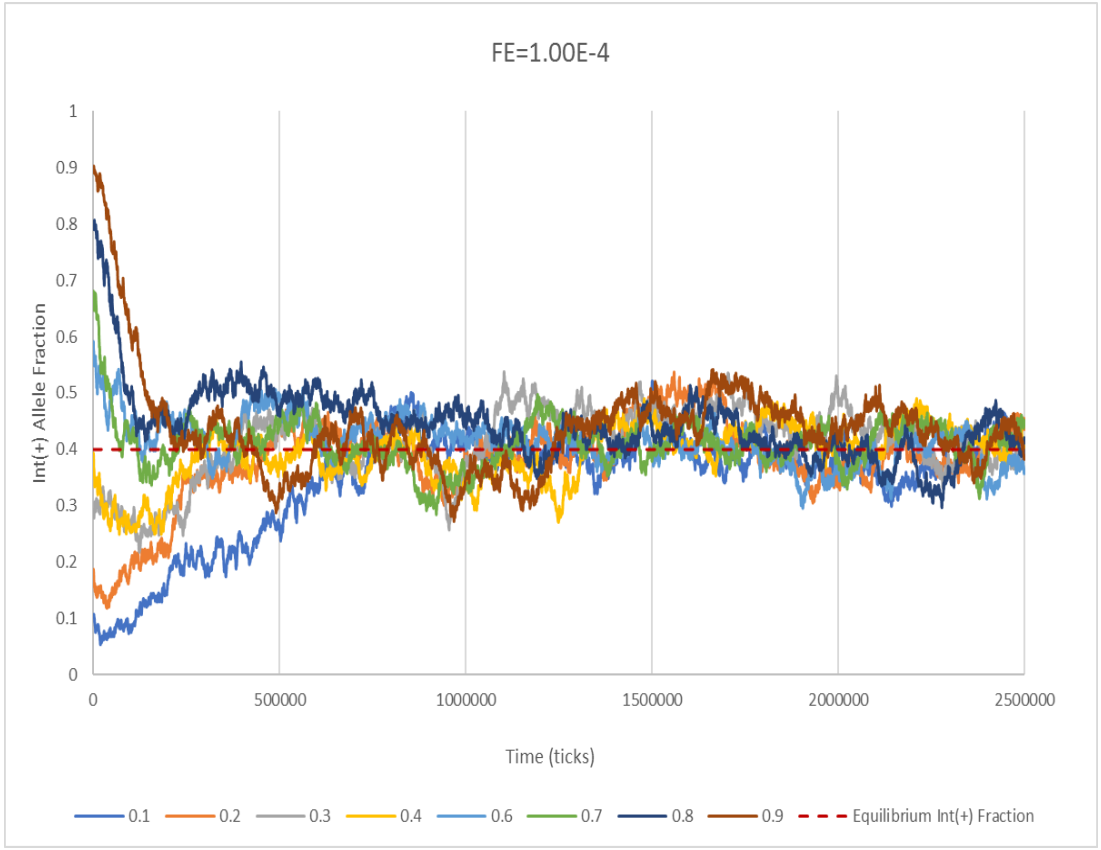
Development over time

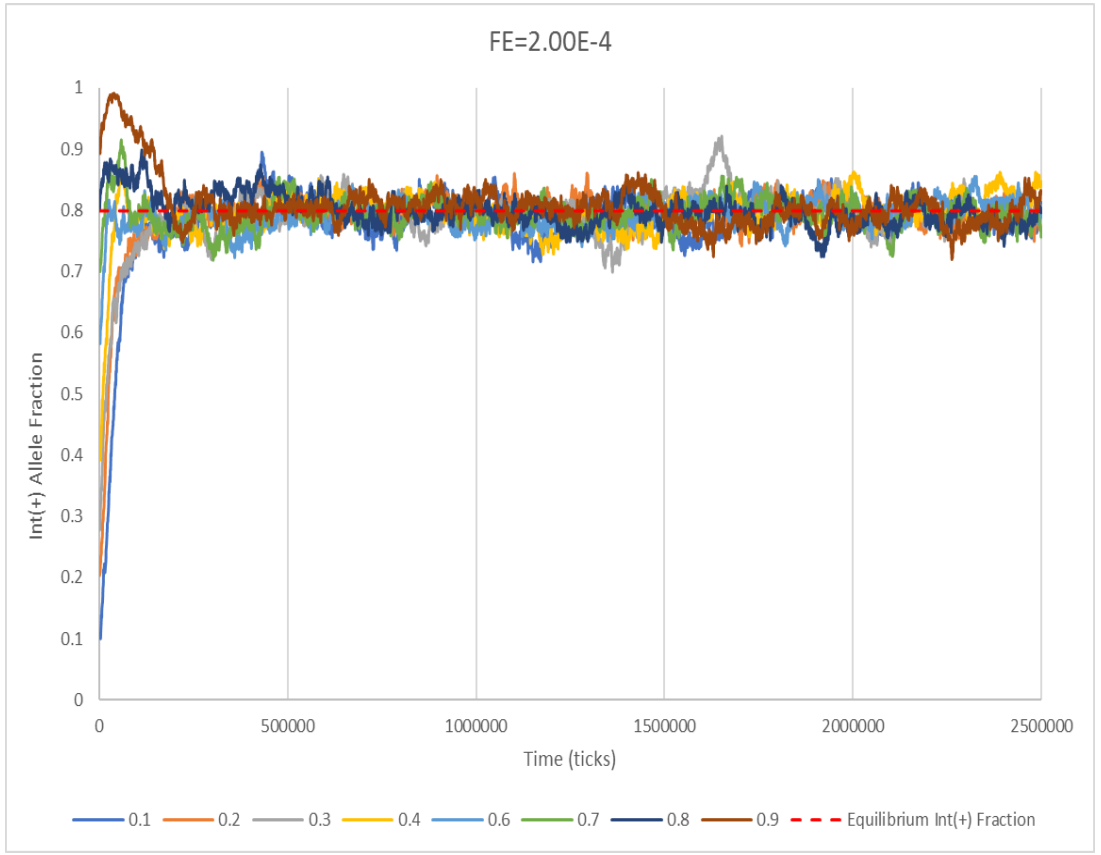


Close to stationary phase: Inteins transmission continues, increasing the frequency of the invaded allele

During active growth: Due to fitness cost of the intein, the frequency of the noninvaded allele increases

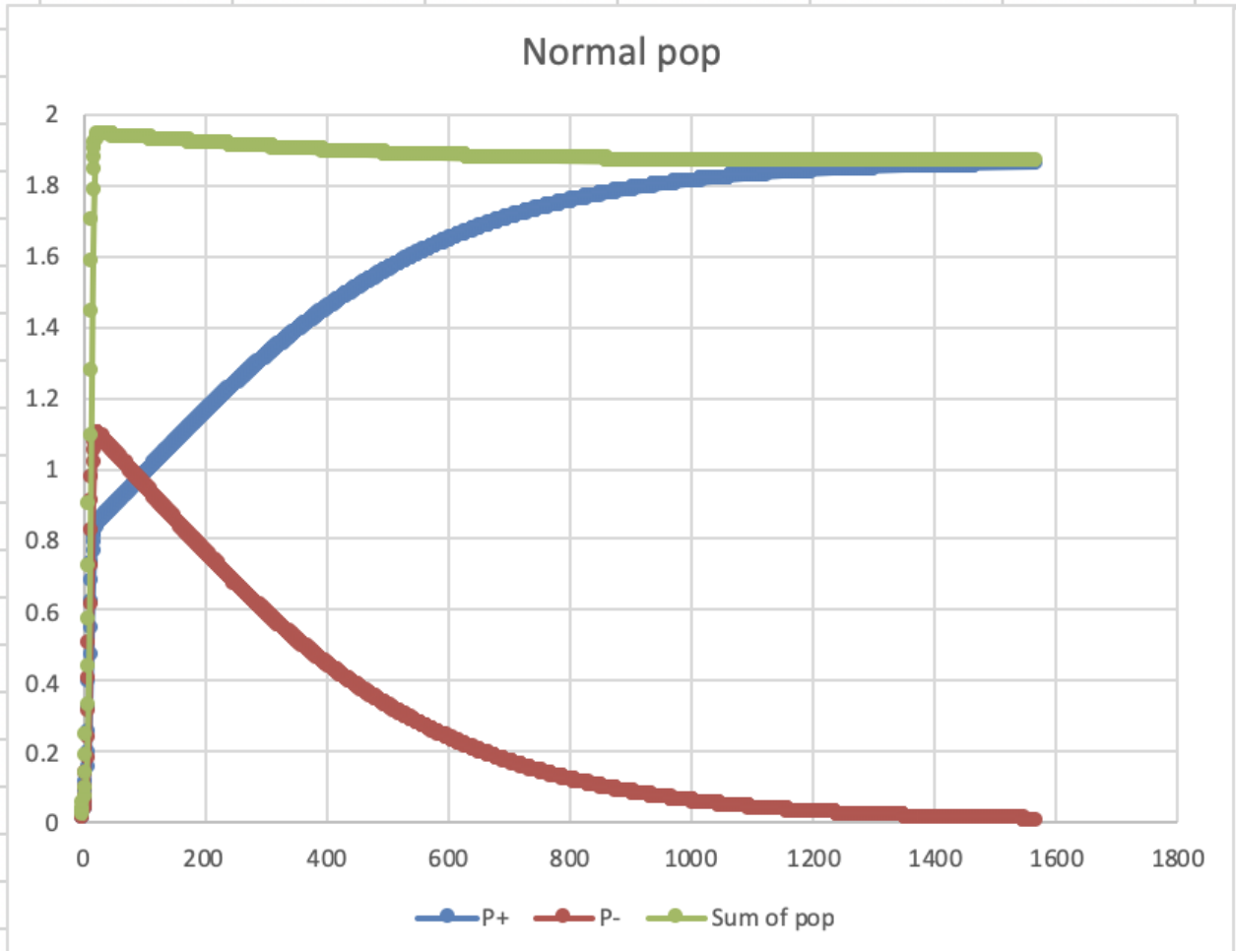
Convergence to a stable fraction of intein alleles

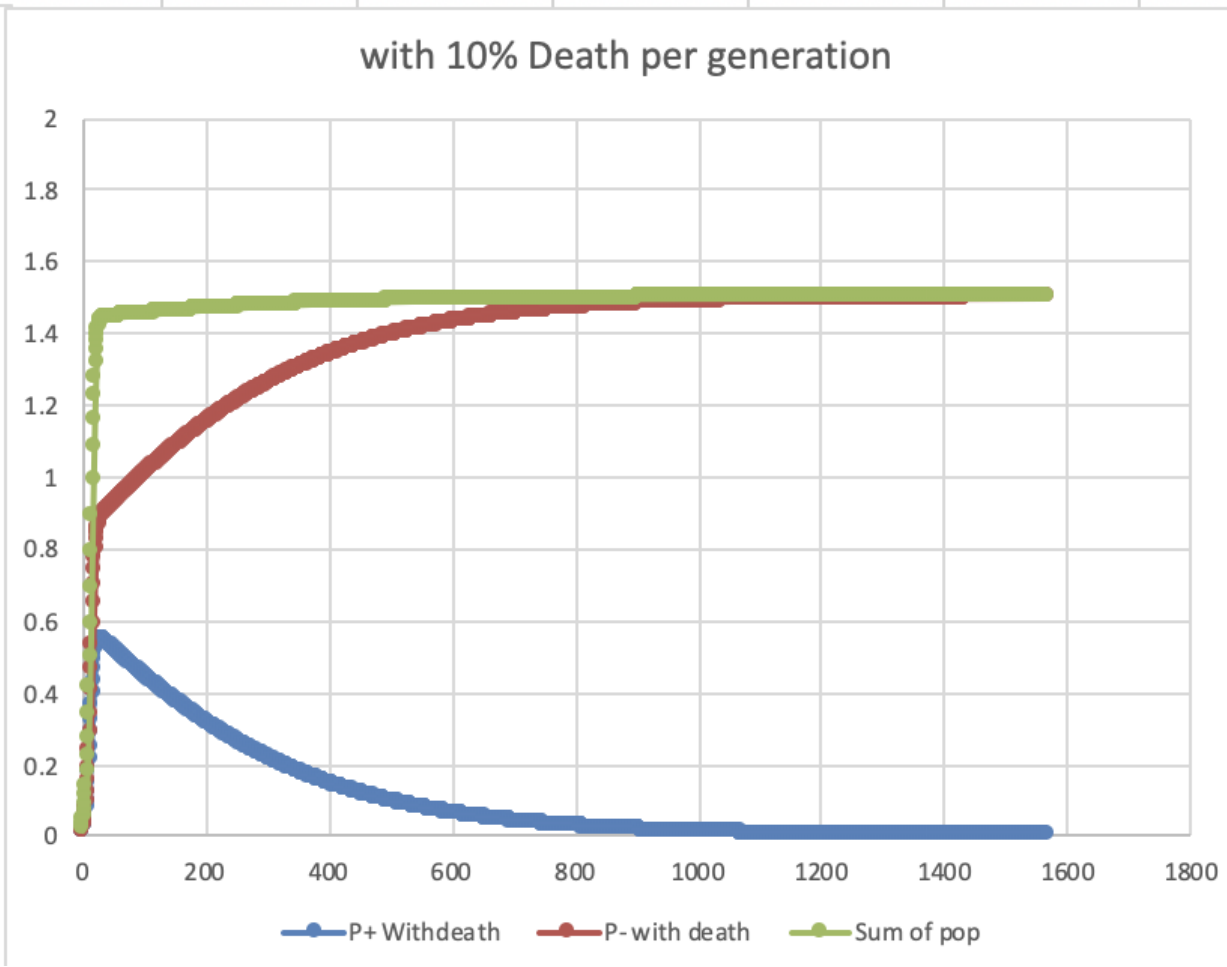




- Starting Int(+) allele fraction = 10% - 90%
- Equilibrium Int(+) fraction determined from final fraction for simulations beginning with equal allele representation
- Int(+) allele was fixed in two simulations when starting fraction was 90%

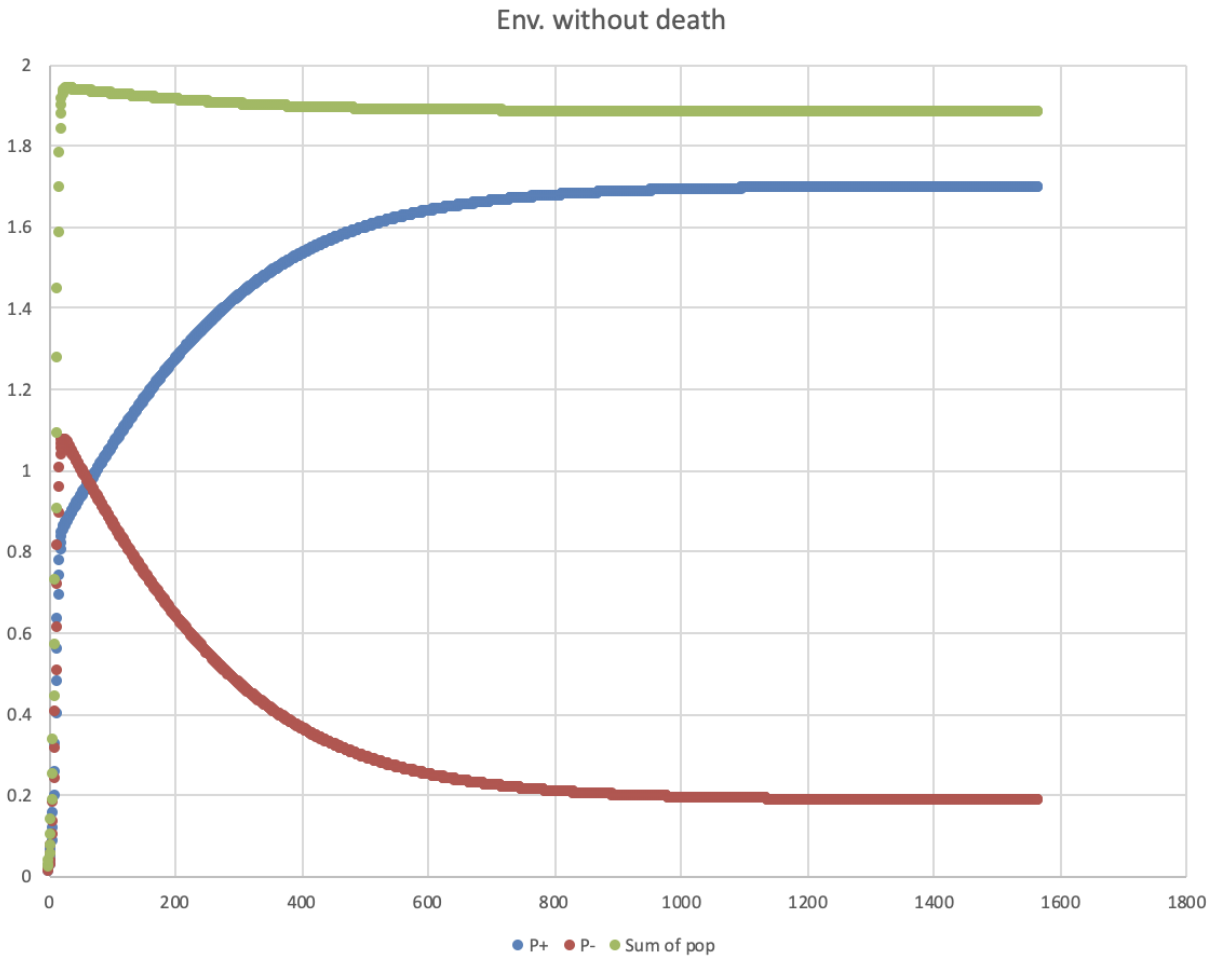
Using Excel and two populations following a discrete logistic growth
One population has a death rate of 10% per generation; no migration

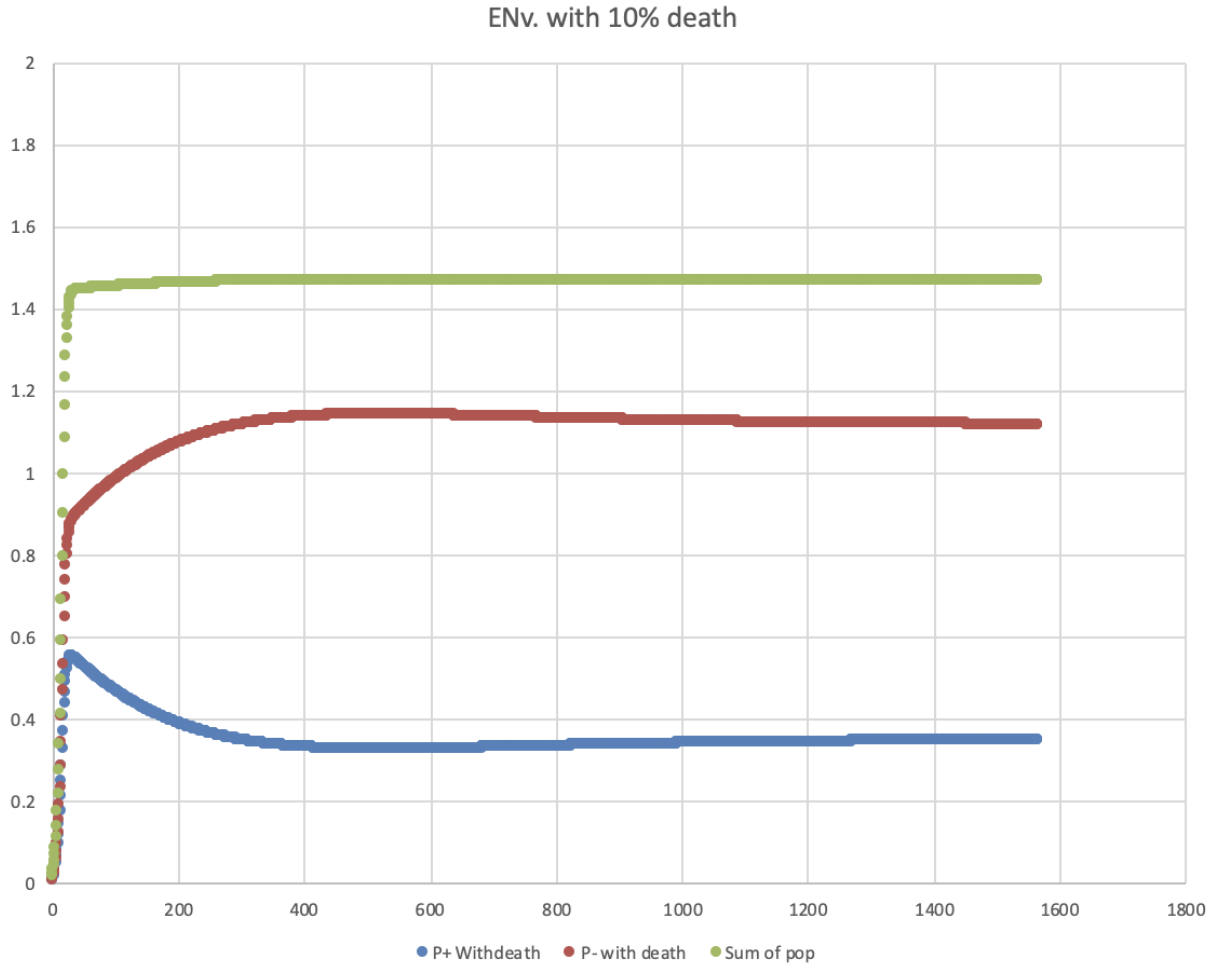




Rate for successful homing: 0.002; Fitness cost due to intein: 0.075; Growth rate: 0.2; Carrying capacity: 2

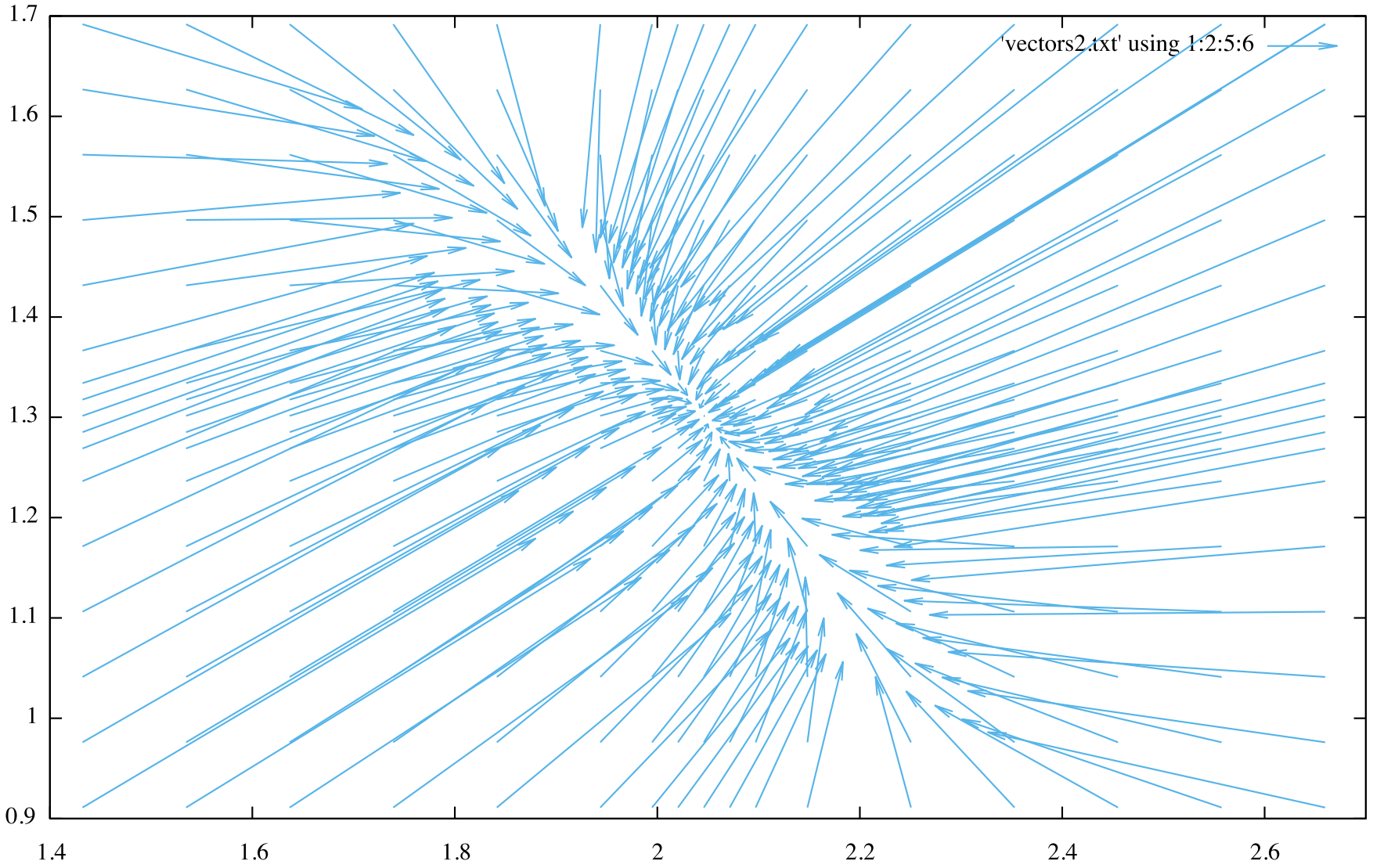
Using Excel and two populations following a discrete logistic growth
One population has a **death rate of 10% per generation**
Same as before, but every generation
0.1% moves from one environment to the other





(note each of the environments is homogenous, the P+ and P- are continuous variables)

The trajectories after disturbances plotted as vectors -- 0 (immediately after the disturbance) -> 4 generations later



Co-existence in a **heterogenous** population
Arguments in favor for a co-existence in a heterogenous population

1) the intein can have a strong negative effect on organismal fitness
2) Due to the large fitness cost, the intein + and intein – alleles dominate during different parts of the growth curve
3) Co-existence of the intein with homing endonuclease (P+) and the allele with an empty target site (P-) in a heterogenous environment.
   Intein containing and intein free alleles have been isolated from the same populations.
   (See Shannon Soucy and Yutian Feng)

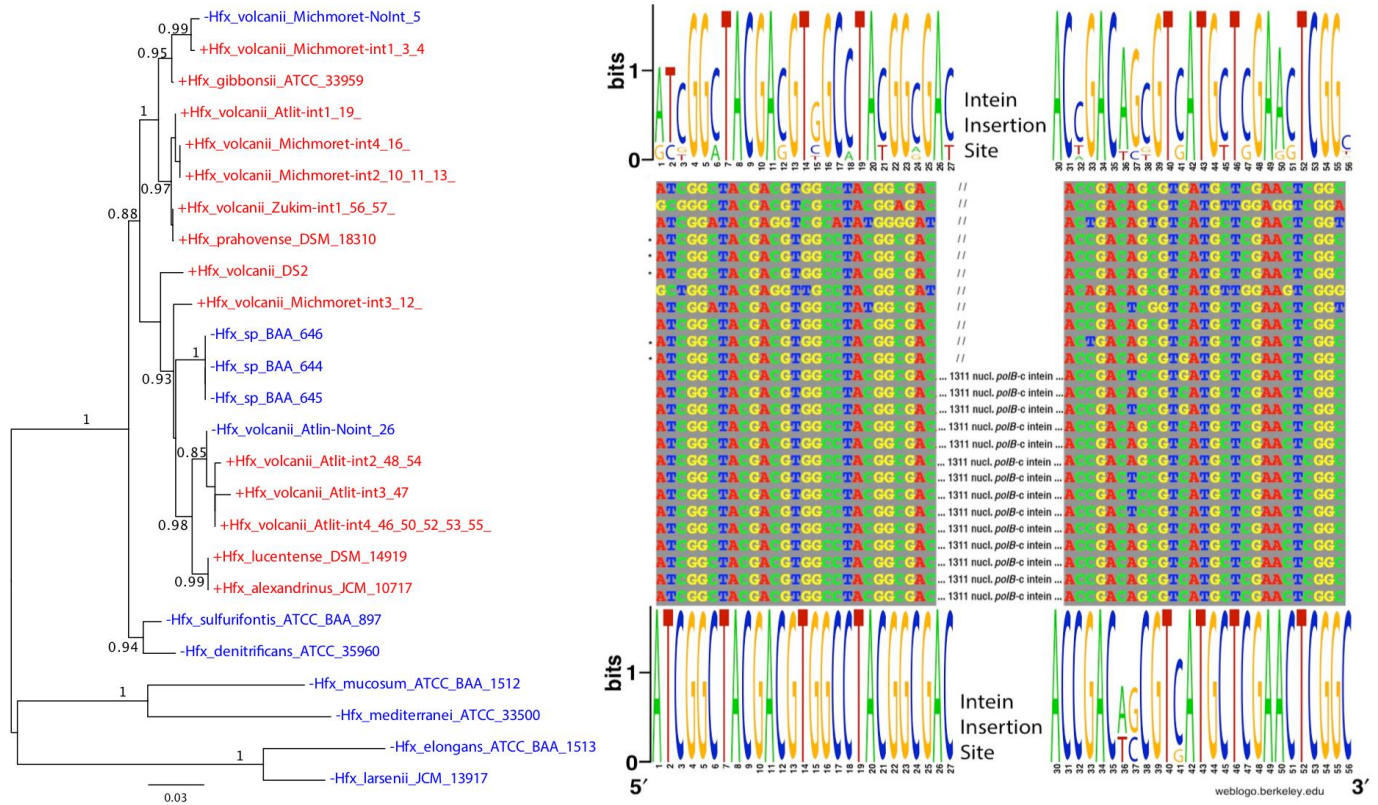
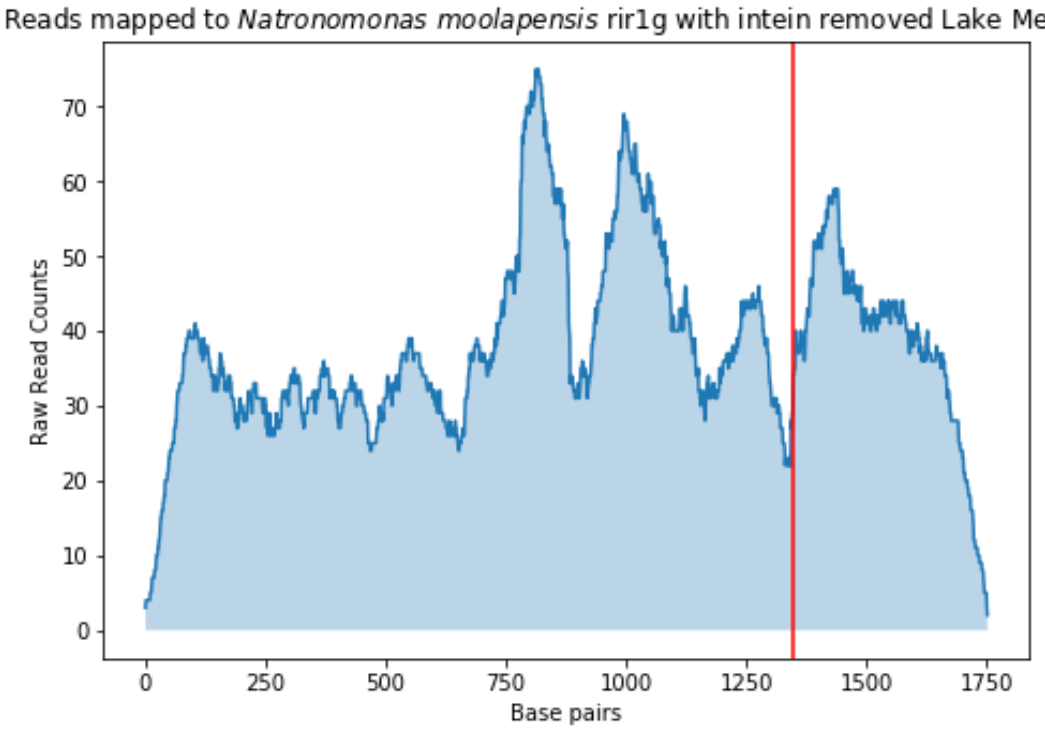
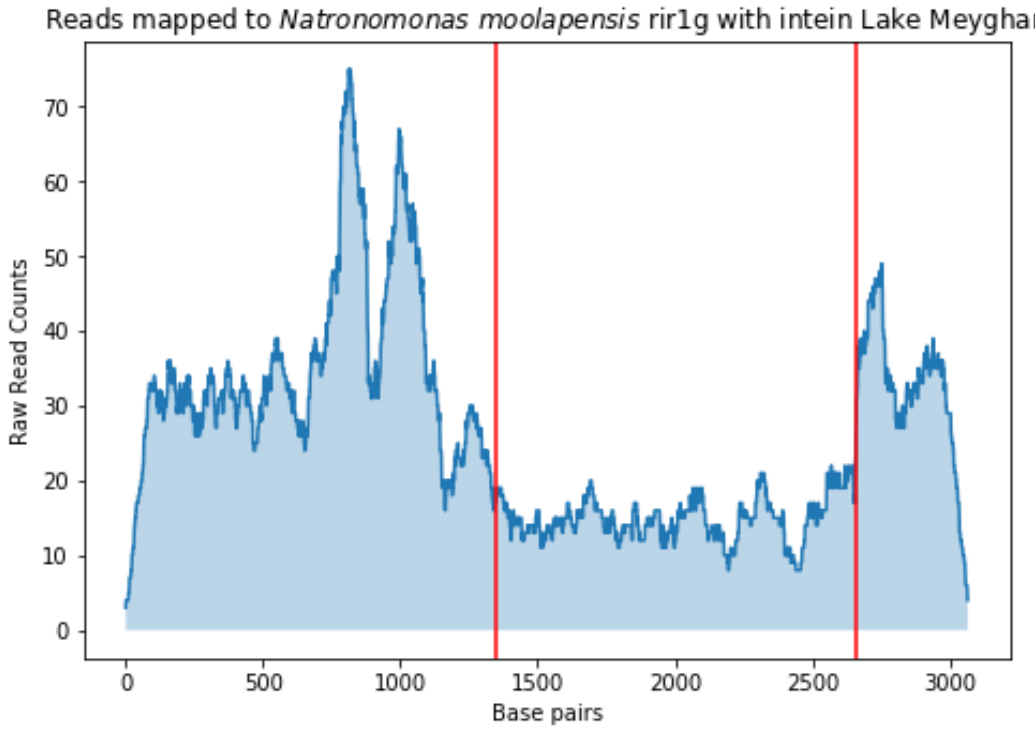


Fig 6. Maximum likelihood phylogeny for *polB* extein sequences (left) and conservation of *polB*-c intein insertion sites (right). Numbers give support values calculated using the approximate Likelihood Ratio Test as implemented in phyml 3.0 (32). Although drawn as a rooted, the tree should be considered unrooted. The finding that sequences without (blue) and with intein (red) do not always form distinct clans (34) reveals that invasion of the *Haloferax* genus with the *polB*-c intein is an ongoing process. The panel on the right shows a *polB* nucleotide sequence alignment around the intein insertion site c. Web logos (33) give the site conservation for intein minus (top) and intein plus sequences (bottom). The five intein minus sequences that group within the cluster of intein plus sequences are marked with an asterisk. The intein minus sequences show greater nucleotide diversity surrounding the intein insertion site, mainly in synonymous positions -- only two positions at the 5' and close to the 3' end of the alignment represent non-synonymous changes. Homing endonuclease site specificity was shown to tolerate substitutions that result in non-synonymous changes (35), suggesting that none of the depicted *Haloferax* sequences may be immune to intein invasion.

Frequency of reads mapped to *Natronomonas moolapensis* rir1g





- Metagenomic reads from Lake Meyghan* mapped back to reference sequences with and without the intein.
- Intein was artificially removed, to intact insertion site.
- Red lines indicate intein boundaries, sharp decrease in coverage when intein removed.

*Naghoni, A. *et al.* Microbial diversity in the hypersaline Lake Meyghan, Iran. *Sci. Rep.* **7,** (2017).

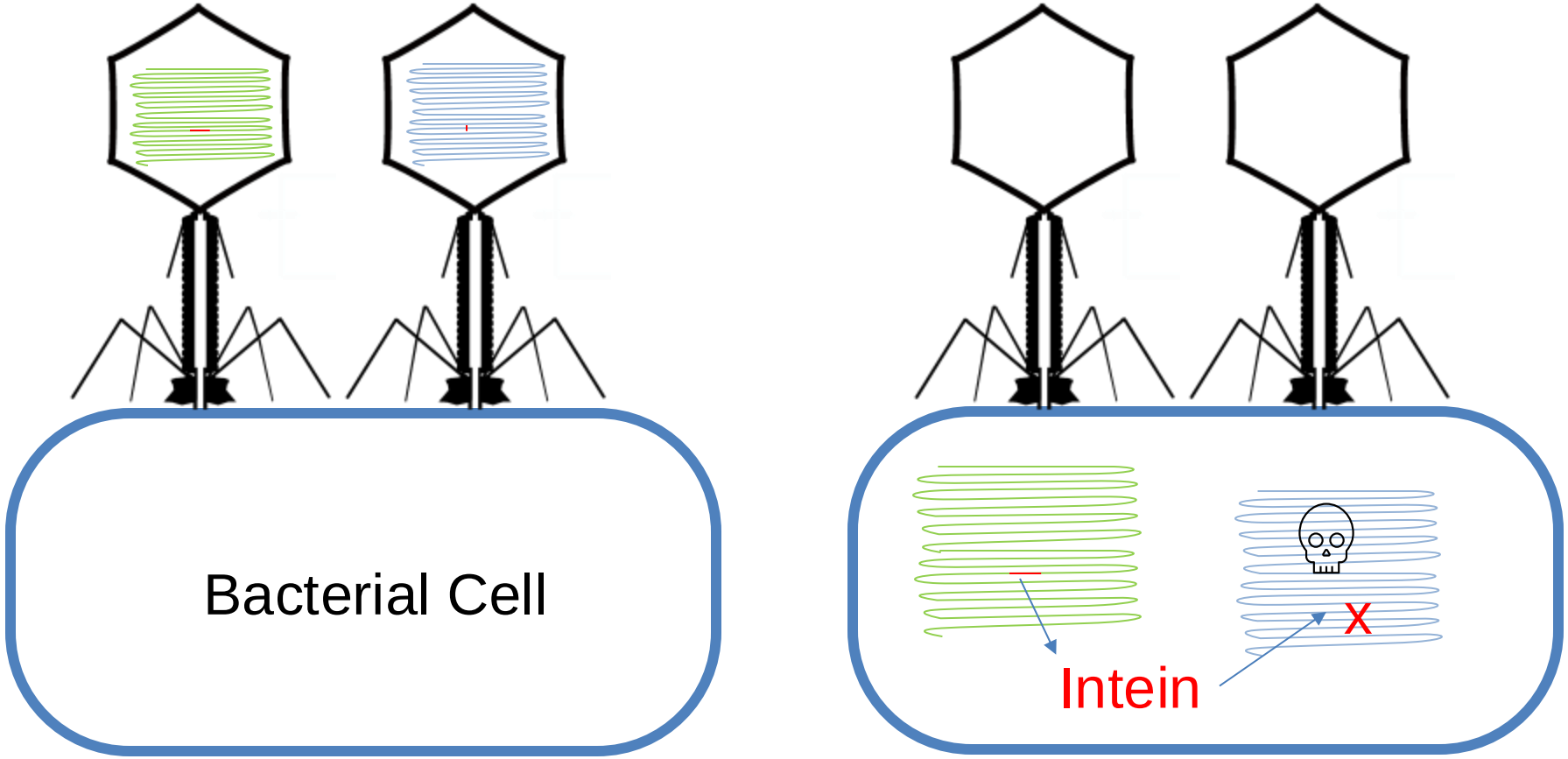
Do inteins have a function beneficial to the organisms?

Infinite Regress

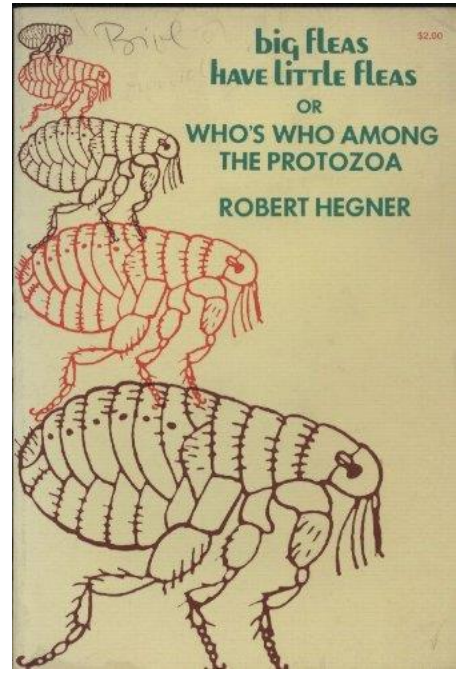
After a lecture on the structure of the solar system, the cosmologist William James was accosted by a little old lady.
"Your theory that the sun is the centre of the solar system, and the earth is a ball which rotates around it has a very convincing ring to it, Mr. James, but it's wrong. I've got a better theory," said the little old lady.
"And what is that, madam?" inquired James politely.
"**That we live on a crust of earth which is on the back of a giant turtle**."
Not wishing to demolish this absurd little theory by bringing to bear the masses of scientific evidence he had at his command, James decided to gently dissuade his opponent by making her see some of the inadequacies of her position.
"If your theory is correct, madam," he asked, "**what does this turtle stand on?**"
"You're a very clever man, Mr. James, and that's a very good question," replied the little old lady, "but I have an answer to it. And it's this: **The first turtle stands on the back of a second, far larger, turtle, who stands directly under him**."
"**But what does this second turtle stand on?**" persisted James patiently.
To this, the little old lady crowed triumphantly,
"**It's no use, Mr. James—it's turtles all the way down**."

From *Ross, John R. (1967). Constraints on variables in syntax (Doctoral dissertation). MIT. hdl:1721.1/15166.*
via https://en.wikipedia.org/wiki/Turtles_all_the_way_down

Phage with Intein Phage w/o Intein



Siphonaptera

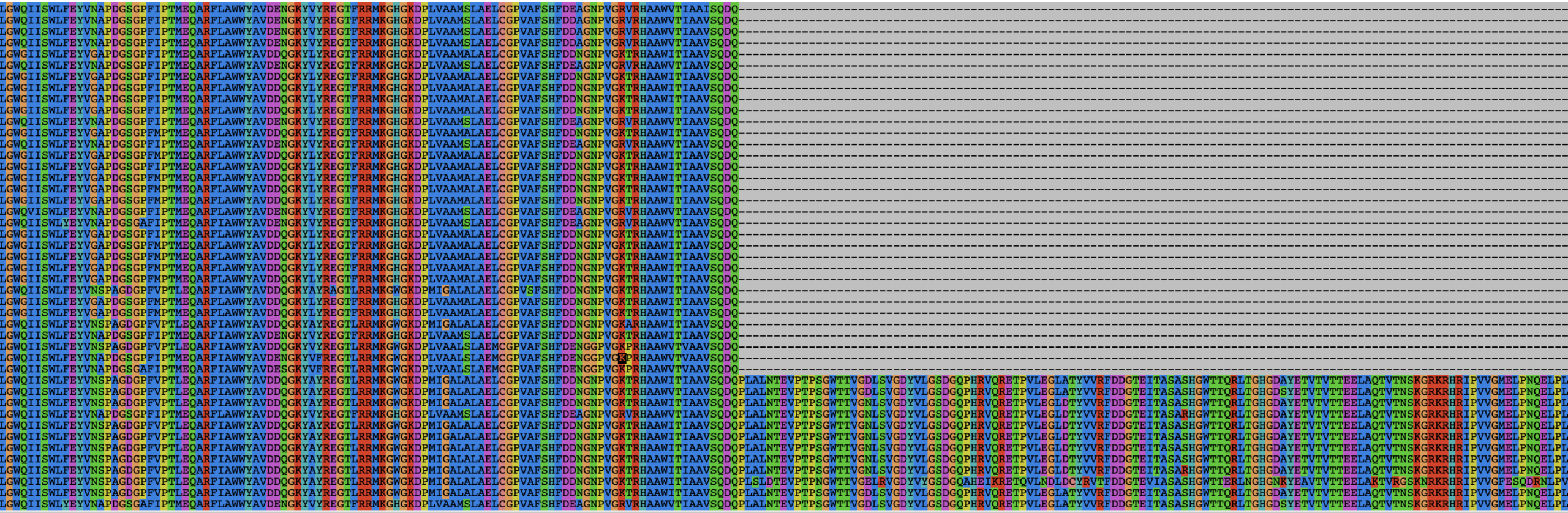


Great fleas have little fleas upon their backs to bite 'em,
And little fleas have lesser fleas, and so *ad infinitum*.
And the great fleas themselves, in turn, have greater fleas to go on ;
While these again have greater still, and greater still, and so on.

From: Augustus De Morgan's poem *Siphonaptera* (1872)
via https://en.wikipedia.org/wiki/Siphonaptera_(poem)

Terminase Subunits from Actinobacteriophages Cluster A1

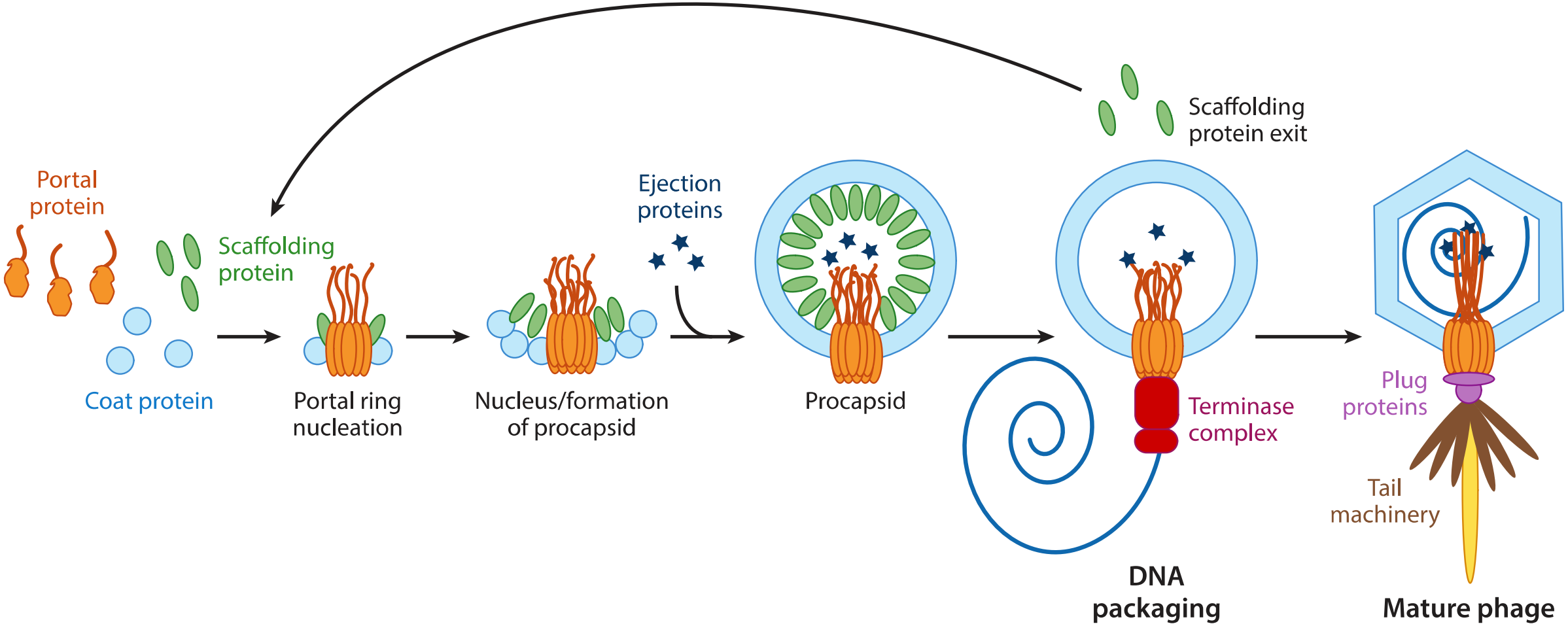
Alignment of protein sequences



N-Extein

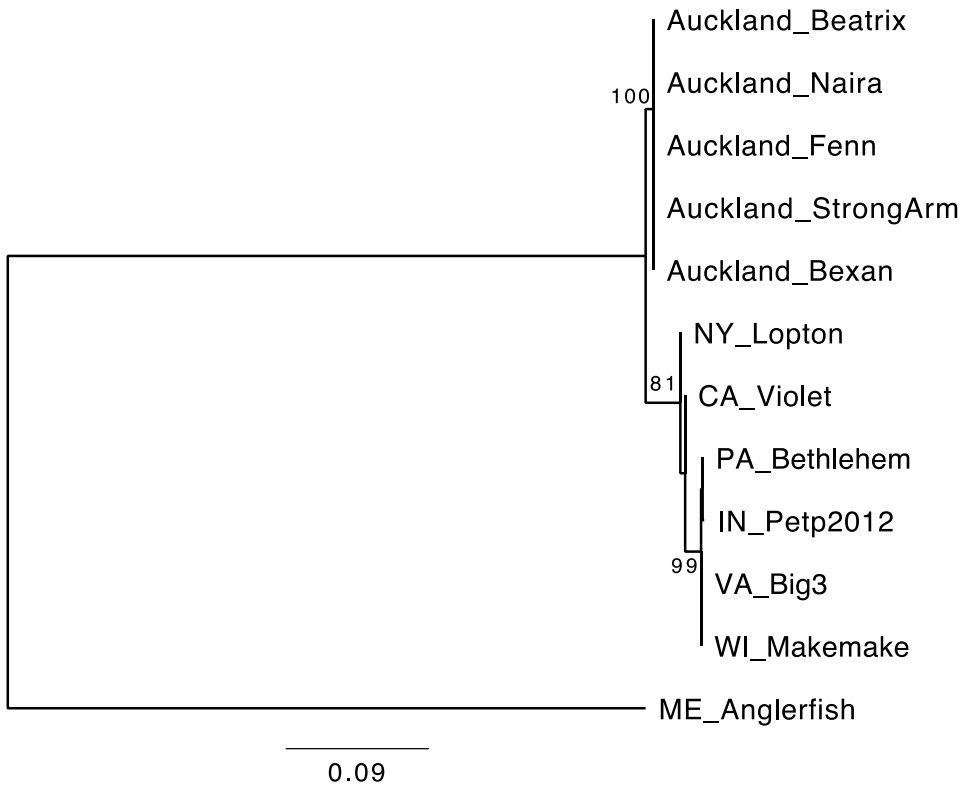
Intein

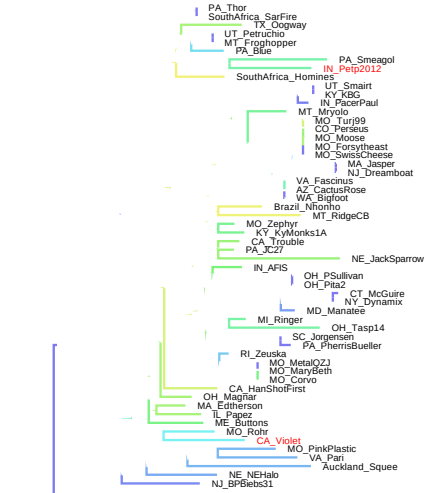
Phage Assembly

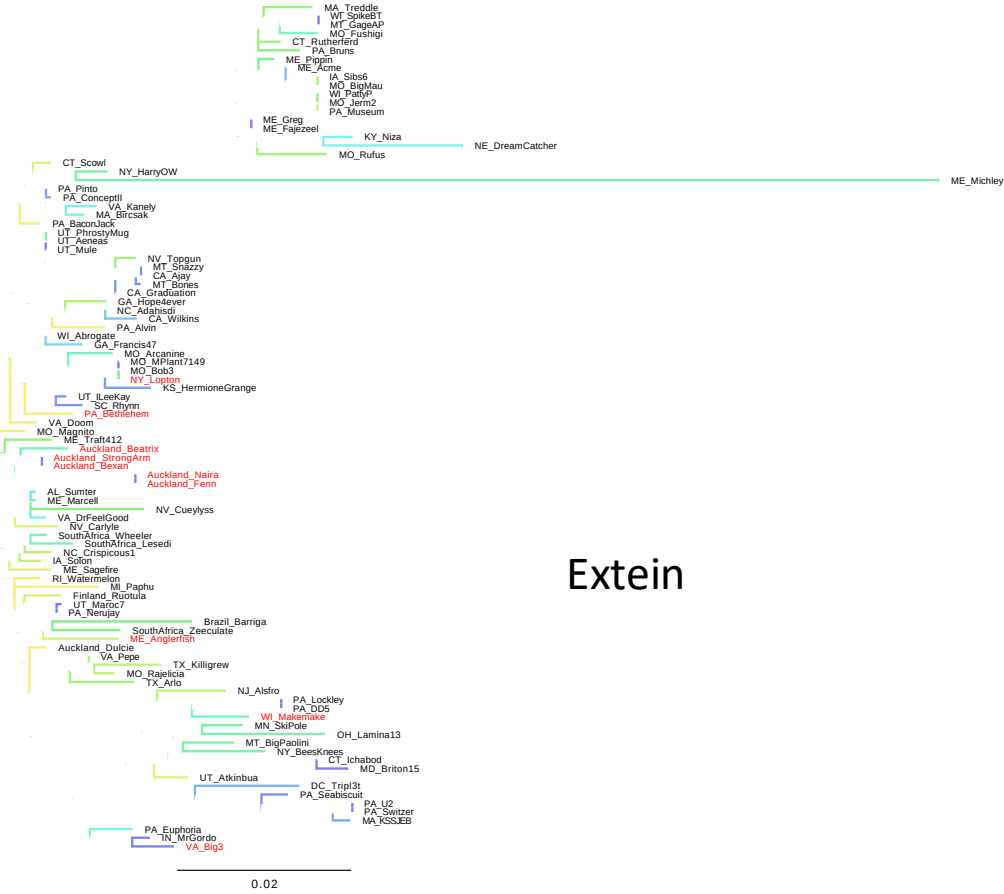


From: Corynne L. Dedeo, Gino Cingolani, and Carolyn M. Teschke:
Portal Protein: The Orchestrator of Capsid Assembly for the dsDNA Tailed Bacteriophages and Herpesviruses
Annual Review of Virology 2019

Exteins and their inteins from homologous terminase subunits in the A1 cluster

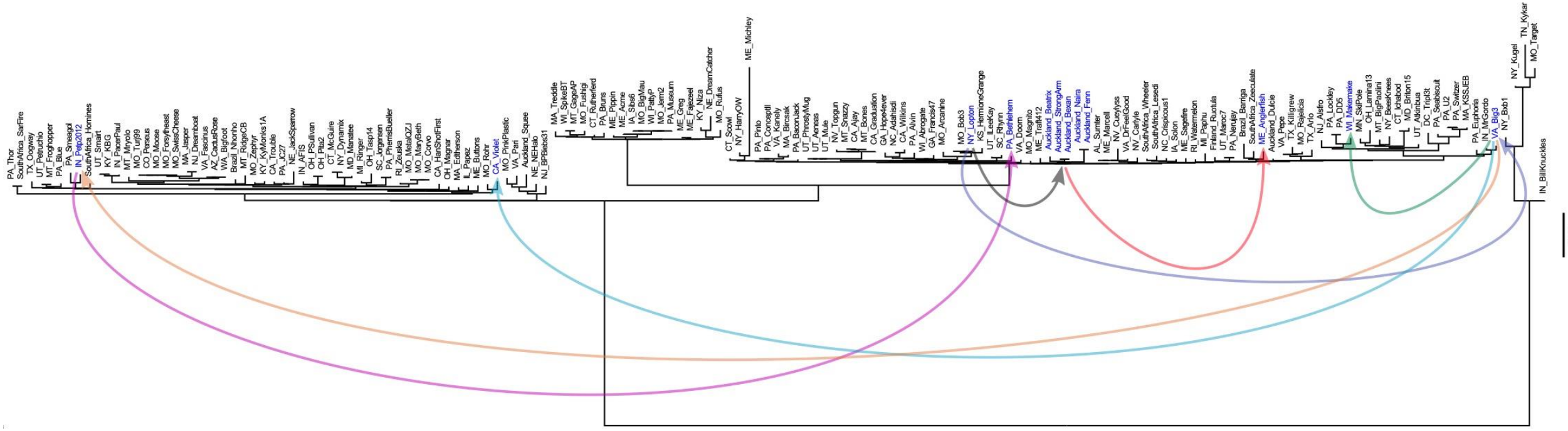






Extein

Reconciliation between extein and intein trees



This is one of many equally parsimonious reconciliations (allowing for transfer, duplication and loss).

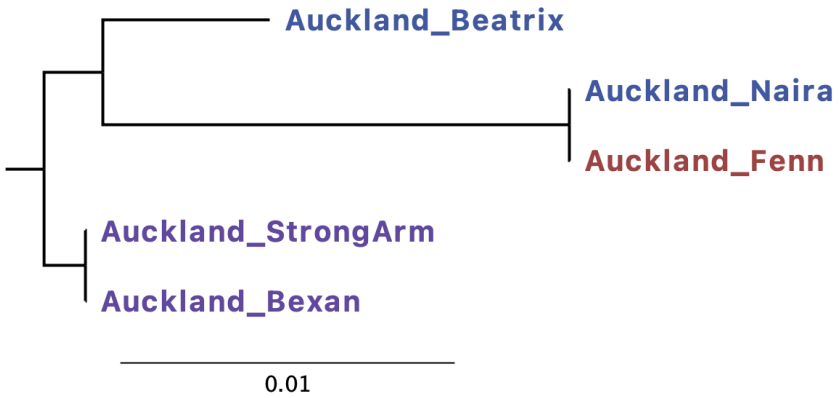
Note that the Auckland cluster requires only one intein acquisition at the base.

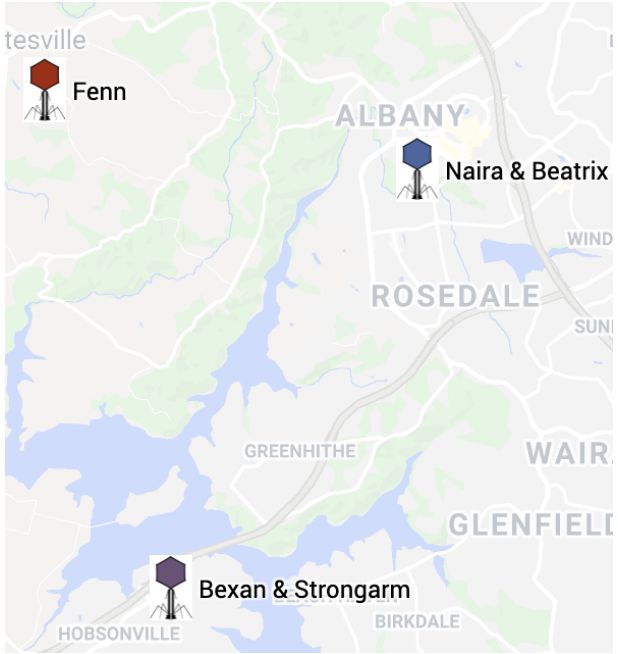
The A1 Phages from the Auckland area

Name	Year of Isolation	Location	Description of Location
Fenn	2017	36.724847 S, 174.635248 E	Soil on an old metal garage door in a field near Auckland
Naira	2017	36.734195 S, 174.700783 E	Massey University community garden compost bin
Beatrix	2017	36.734195 S, 174.700783 E	Massey University community garden compost bin
Bexan	2019	36.790857 S, 174.65929 E	Catalina community garden compost bin
Strongarm	2019	36.790857 S, 174.65929 E	Catalina community garden compost bin

	Number of sites	Number of sequences	Number of Polymorphic sites (only SNPs)	Watterson's Theta
Extein	1689	5	31	0.0088
Intein	1023	5	0	0

Phylogeny based on extein sequences (topology is the same if the whole genomes are used):





The Evolutionary History of a DNA Methylase Reveals Frequent Horizontal Transfer and Within-Gene Recombination

by Sophia P. Gosselin [1], Danielle R. Arsenault [1], Catherine A. Jennings [1] and Johann Peter Gogarten [1,2,*]

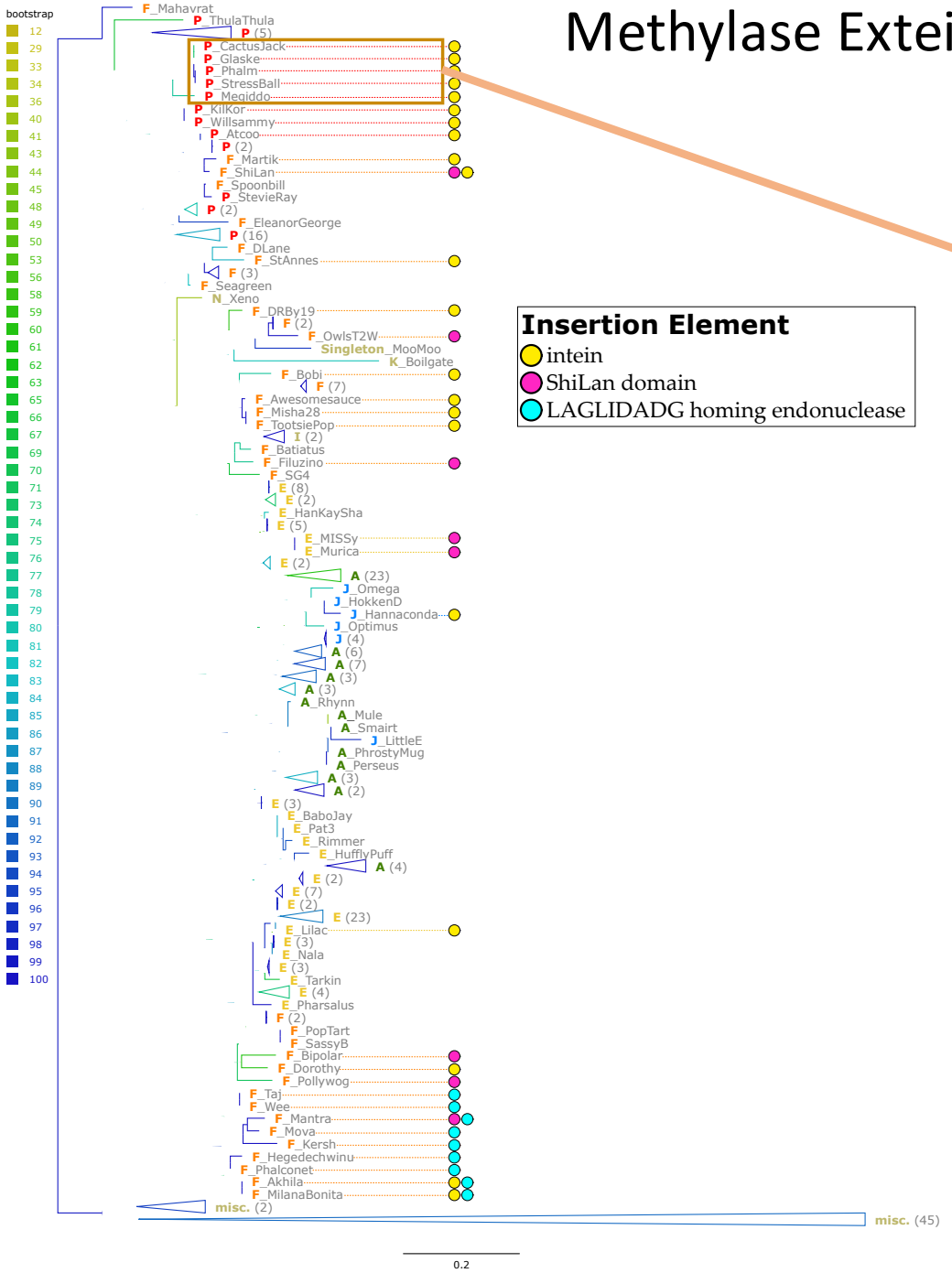
[1] Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06268-3125, USA

[2] Institute for Systems Genomics, University of Connecticut, Storrs, CT 06268-3125, USA

[*] Author to whom correspondence should be addressed.

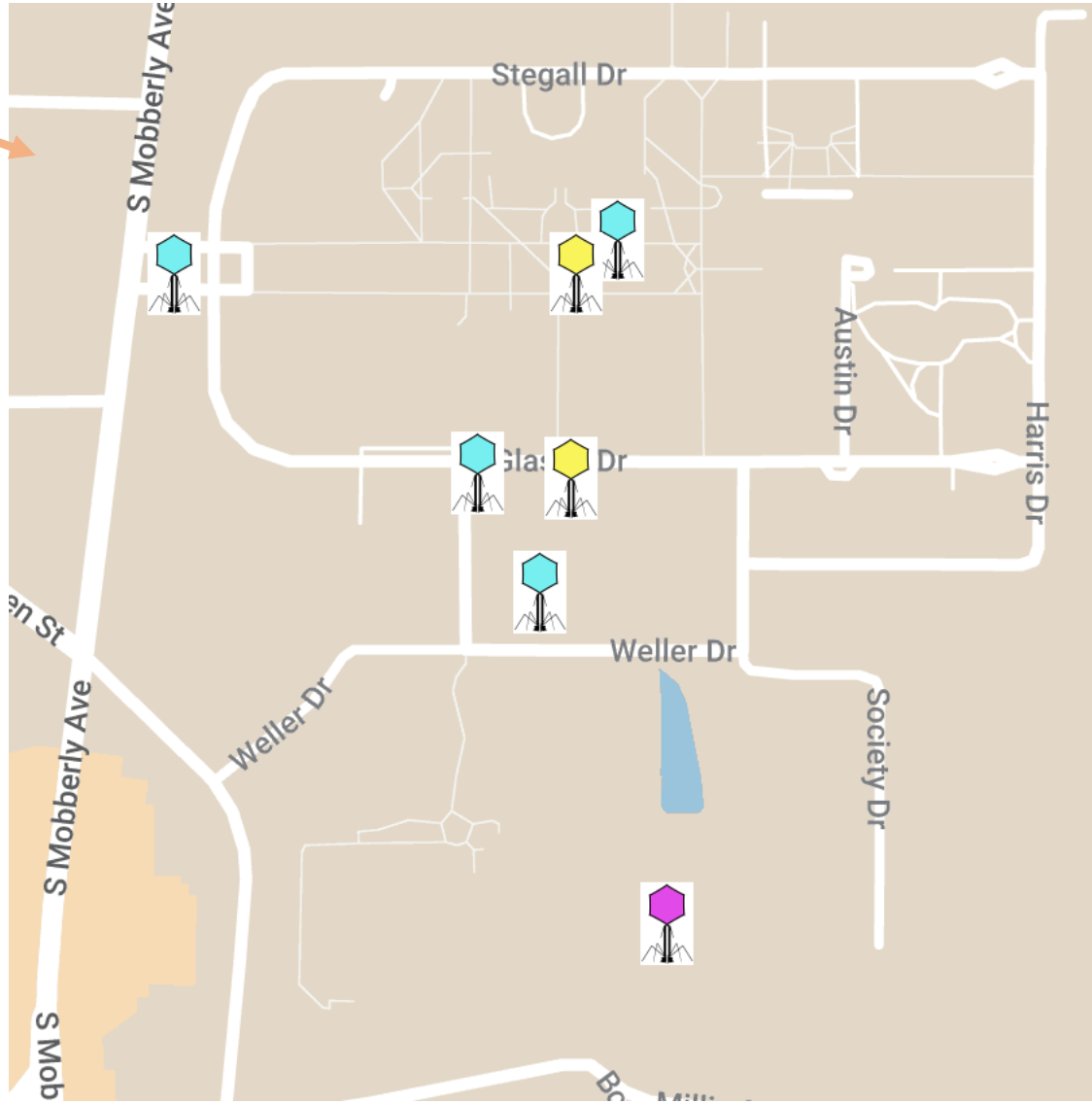
Genes **2023**, *14*(2), 288; https://doi.org/10.3390/genes14020288

Methylase Extein Phylogeny



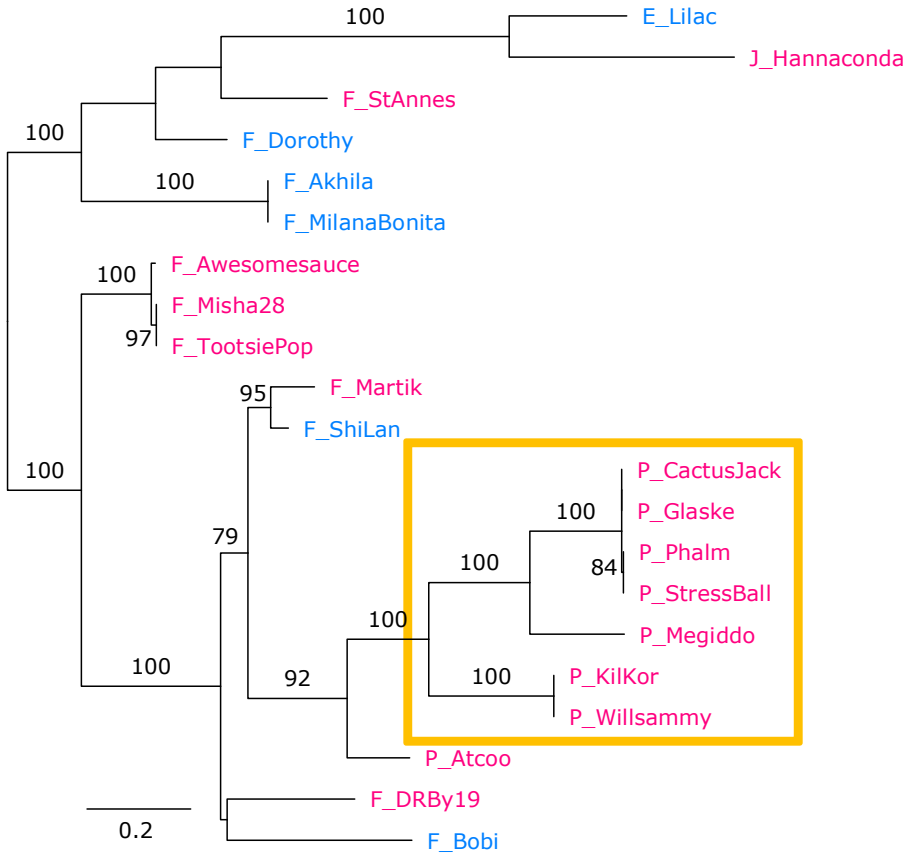
Insertion Element

- intein
- ShiLan domain
- LAGLIDADG homing endonuclease



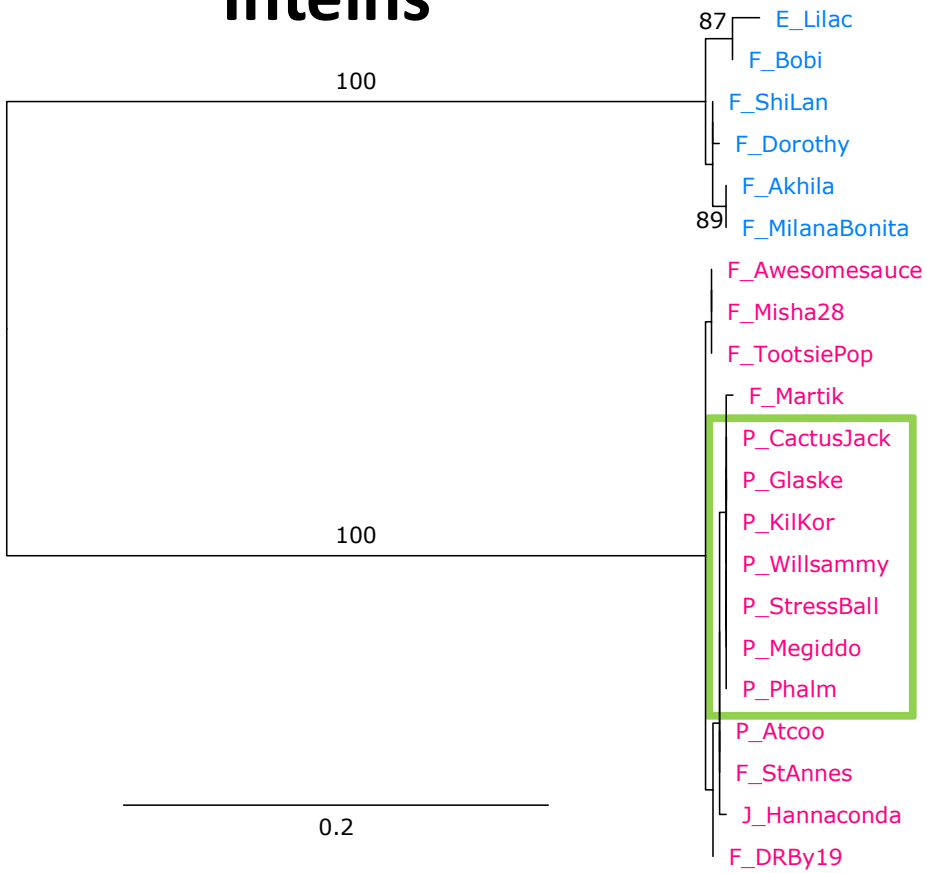
Phylogeny of an Actinophage Methylase Family

Exteins



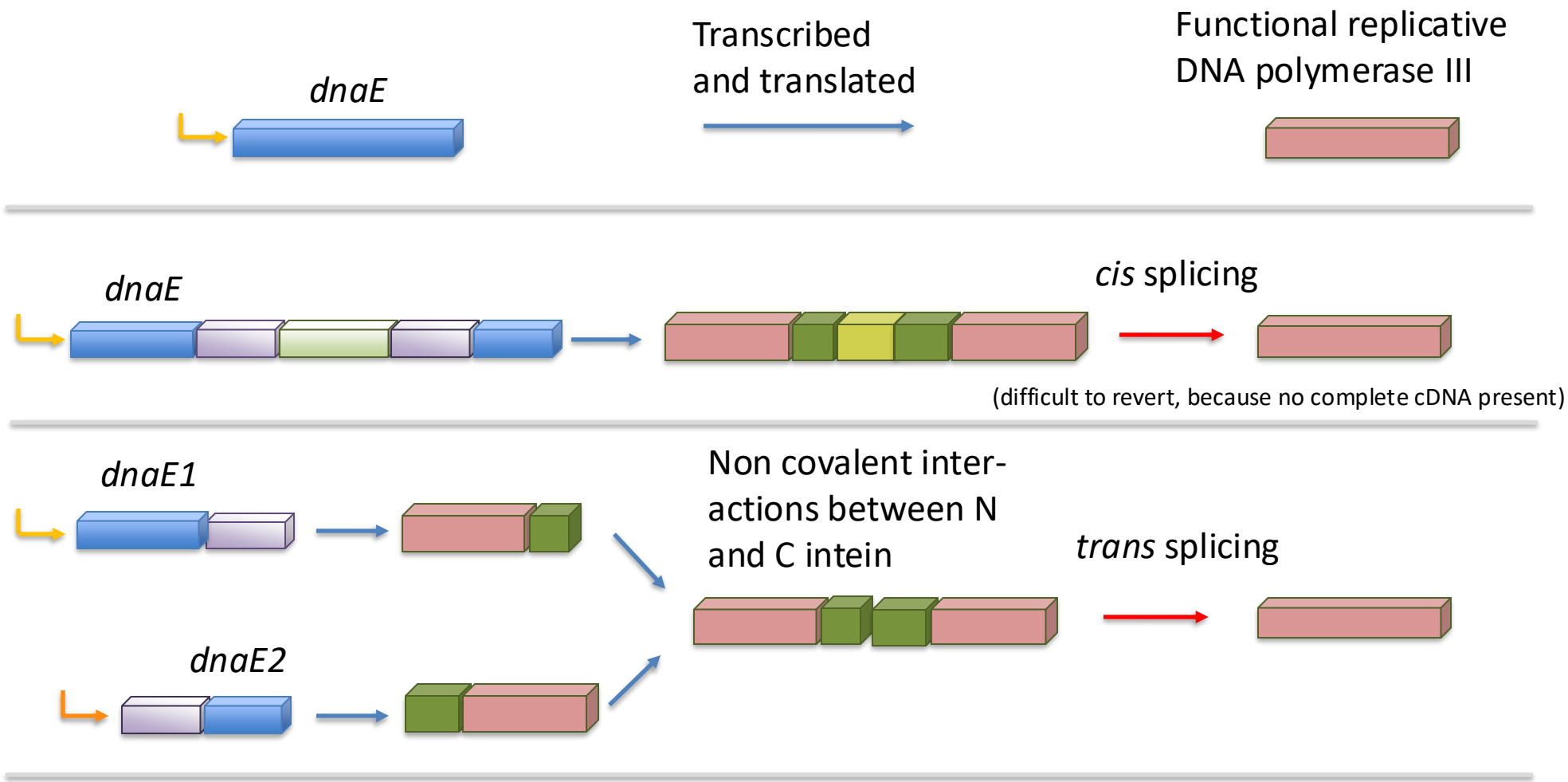
Max **13%** sequence divergence

Inteins



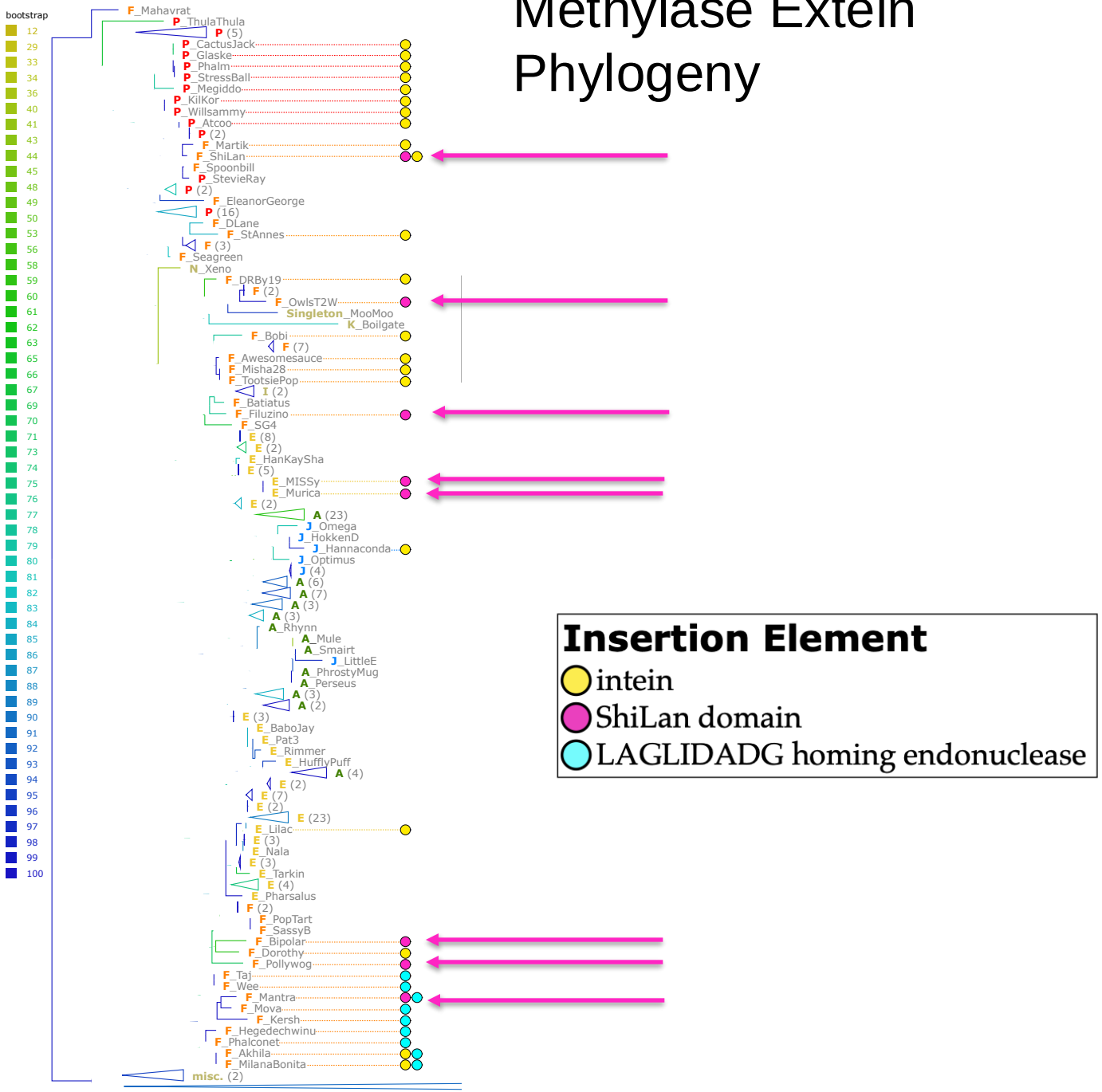
Max **0%** sequence divergence

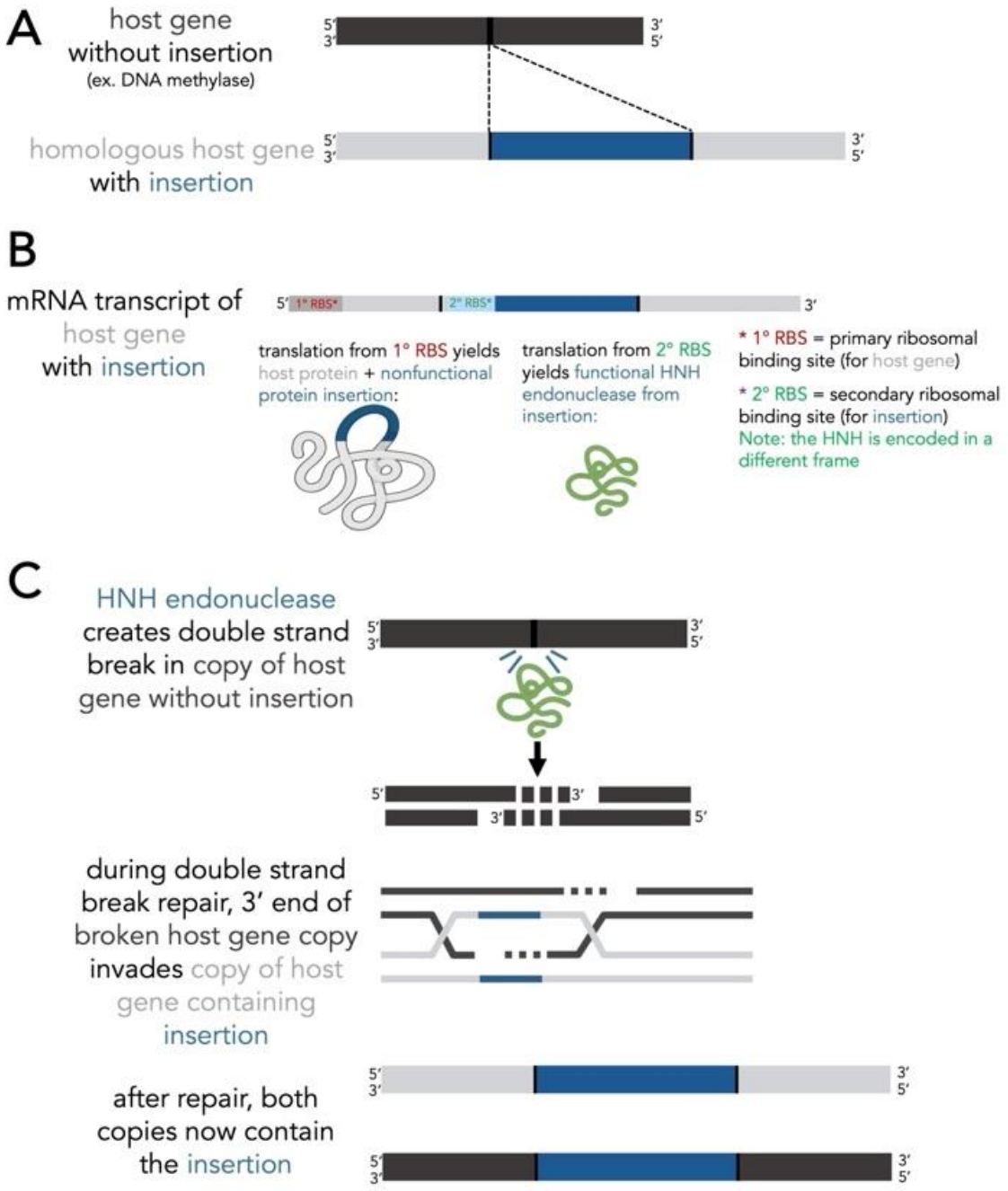
Split inteins as an example for CNE



The Shilan Domain

An new type of insertion sequence with disjunct distribution





(A) Host genes with and without insertion sequence.

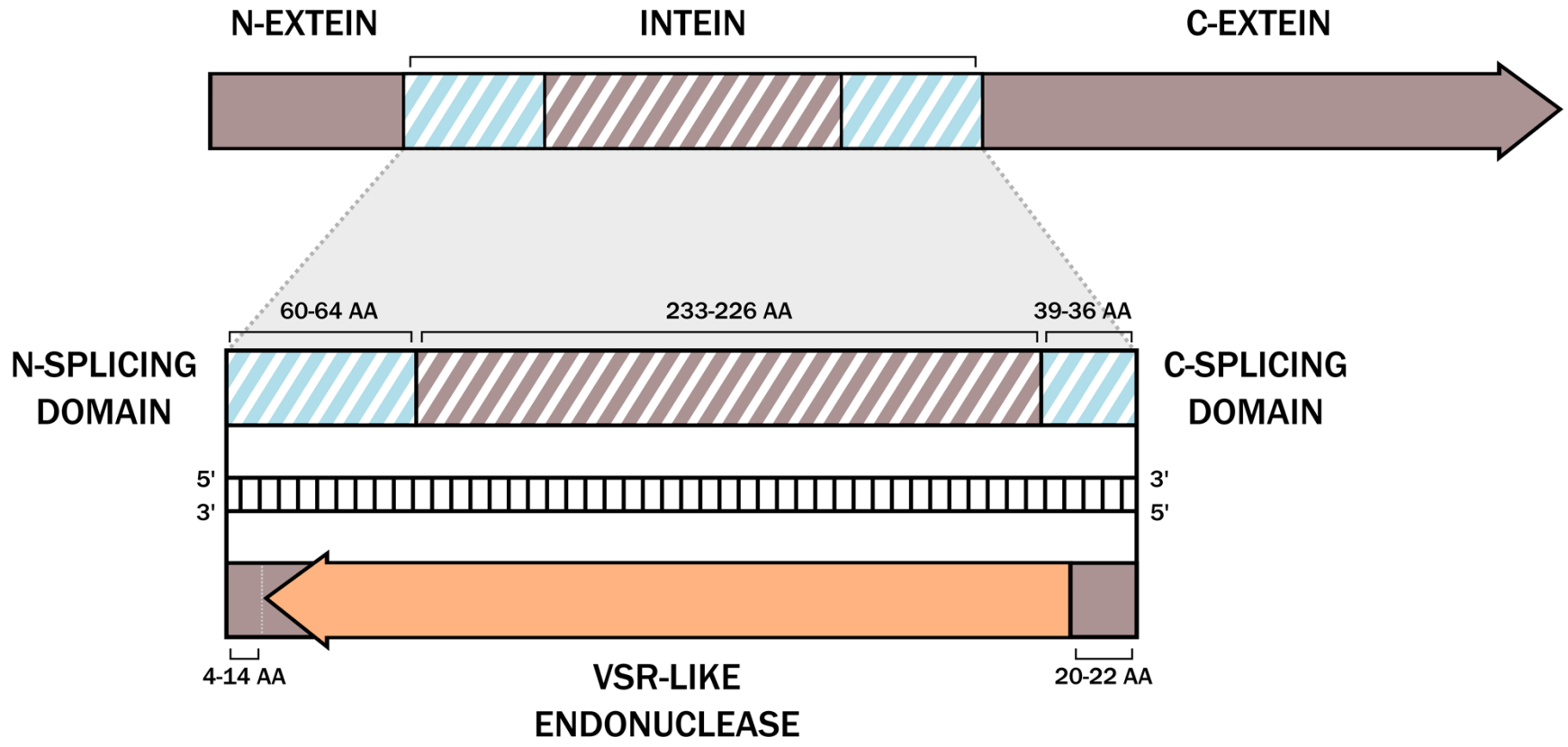
(B) (B) Depiction of second overlapping reading frame encoding an HNH endonuclease.

(C) Homing through double strand break repair.

Shilan domain

- No selfsplicing domain
- Insertion sequence has a second ORF that creates a transcript encoding an HNH endonuclease

Unexpected Variations in Inteins



The homing endonuclease is encoded on the other strand with its own transcription and translation start site.