Evolution of Horizontal Gene Transfer

Paul Higgs

McMaster University,

Hamilton, Ontario.

Eric Collins – Part 1

Seyed Zamani Dahaj / Mohamed Okasha / Jakub Kosakowski – Part 2

Aaron Vogan – Part 3

Evolution of Horizontal Gene Transfer

Part 1 - The core genome and the pangenome

- Gene frequency distributions
- The infinitely many genes model

Part 2 – Phylogenetics with gene presence/absence patterns

- Models for gene gain and loss
- Estimating the frequency of horizontally transferred genes

Part 3 - Evolutionary theory / Cell biology

- What are the costs and benefits of HGT?
- Is there selection to increase or decrease HGT?

Background: Vertical and Horizontal Transmission

Vertical Transmission – genes passed from a parent
(usually whole chromosome at once)

Horizontal Transfer – genes gained from an unrelated individual
(usually single gene at a time or small number of linked genes)

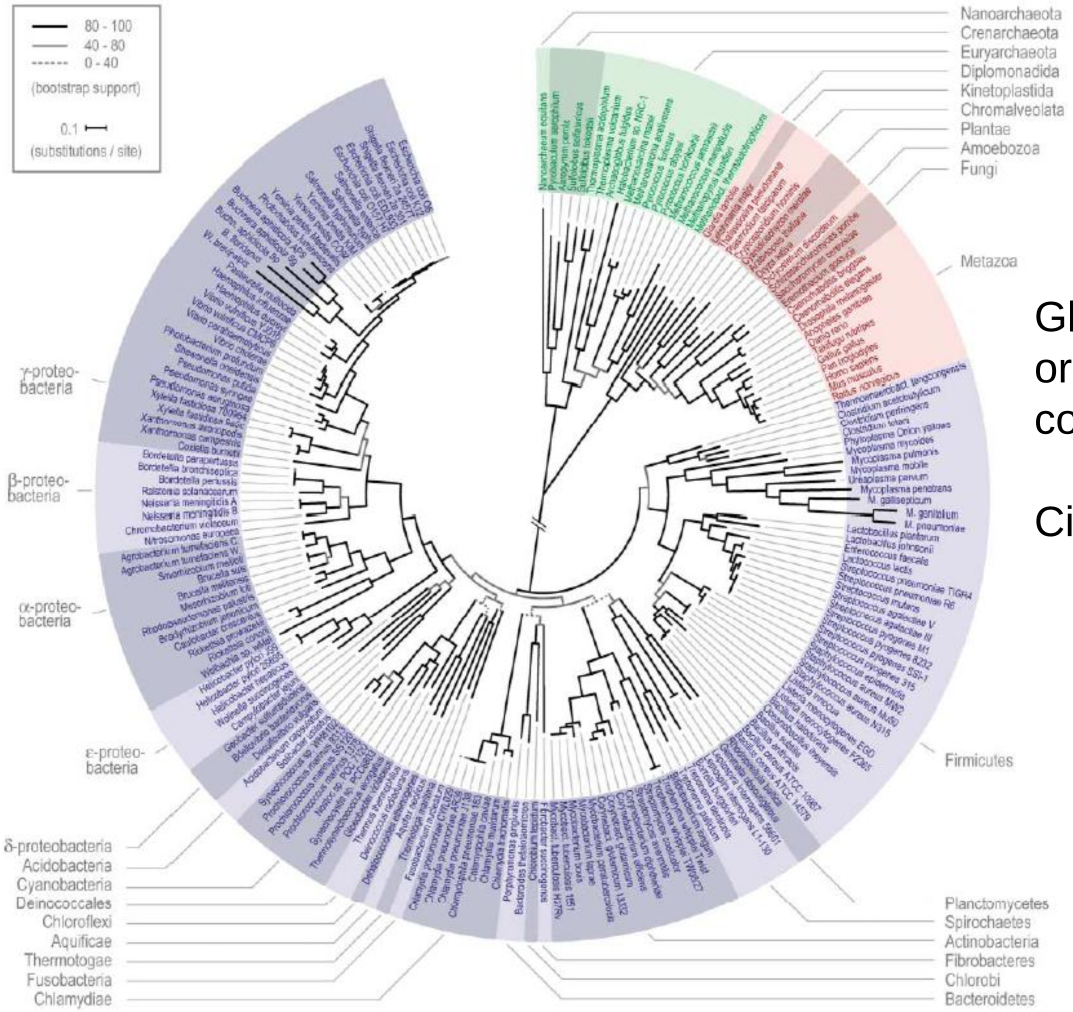
Mechanisms of transfer –

Transformation (import of DNA from environment)
Plasmids (conjugation can transfer plasmids and their genes)
Viruses (can tranfer host genes)

Once DNA gets inside a cell it still has to recombine with the genome in order to become a heritable part of the genome. Can be either

- homologous (replace another version of a gene)
- non-homologous (gain a new gene)

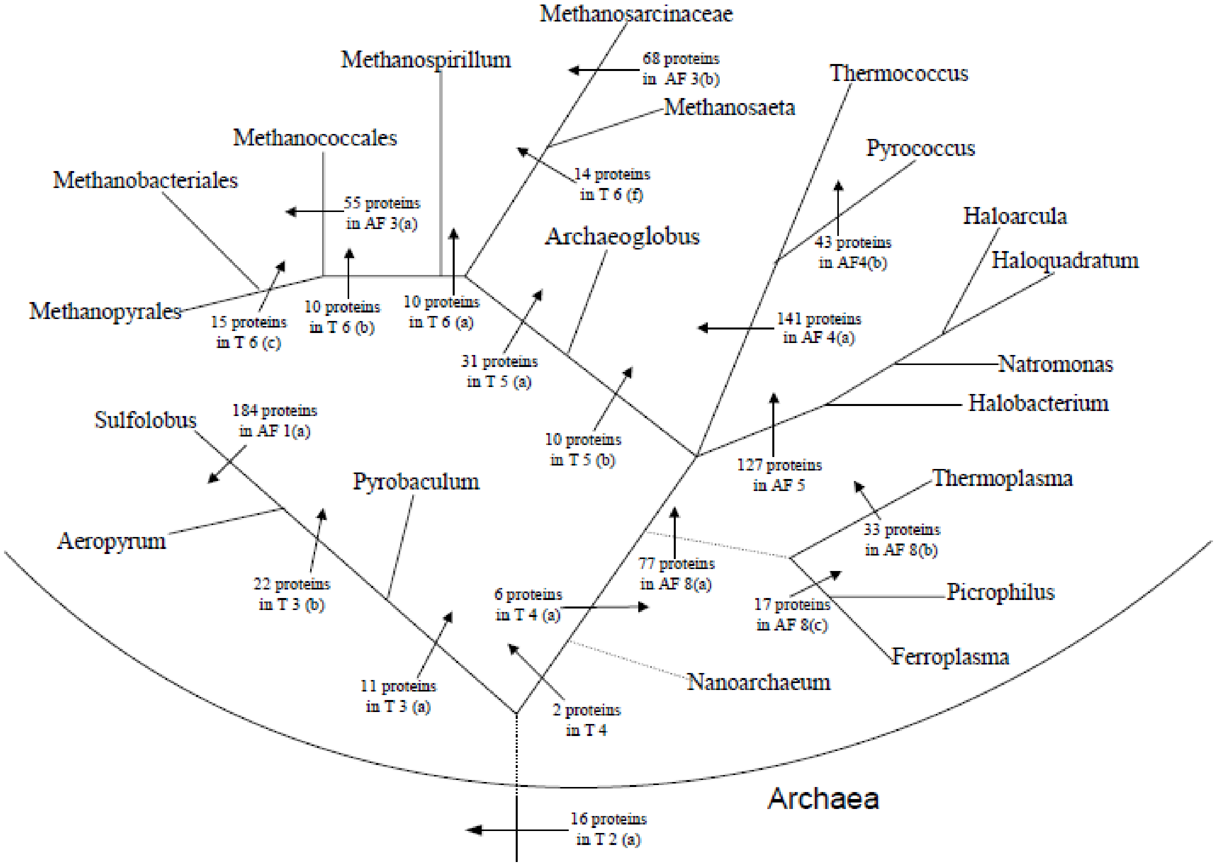
Background:
Large scale phylogenetic trees can be constructed **but** these use only a small number of genes



Global phylogeny of 191 organisms derived from 31 conserved protein genes.

Ciccarelli et al (2006) Science

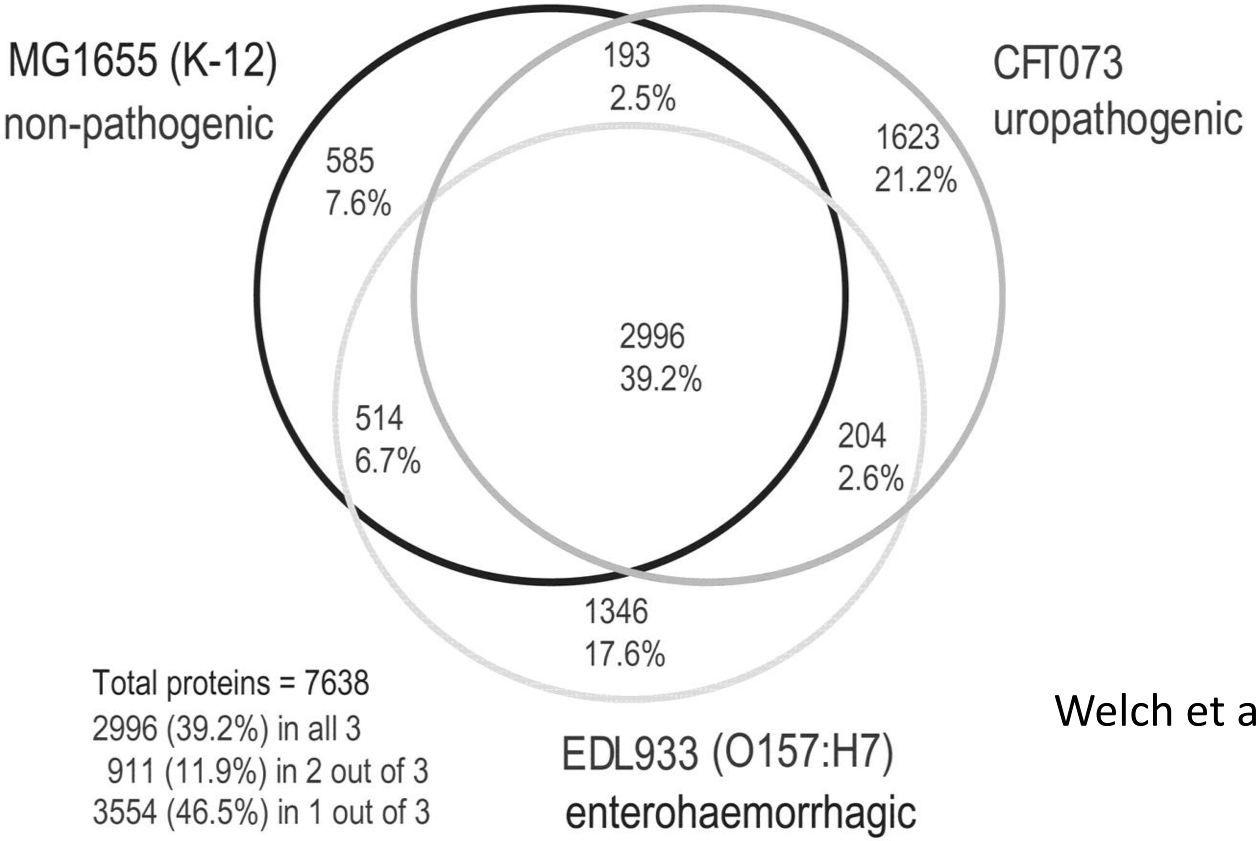
Signature genes are found in taxonomic groups of many different levels. This supports a tree-like picture of evolution.



Tree of Archaea based on signature genes

Gao and Gupta (2007) BMC Genomics

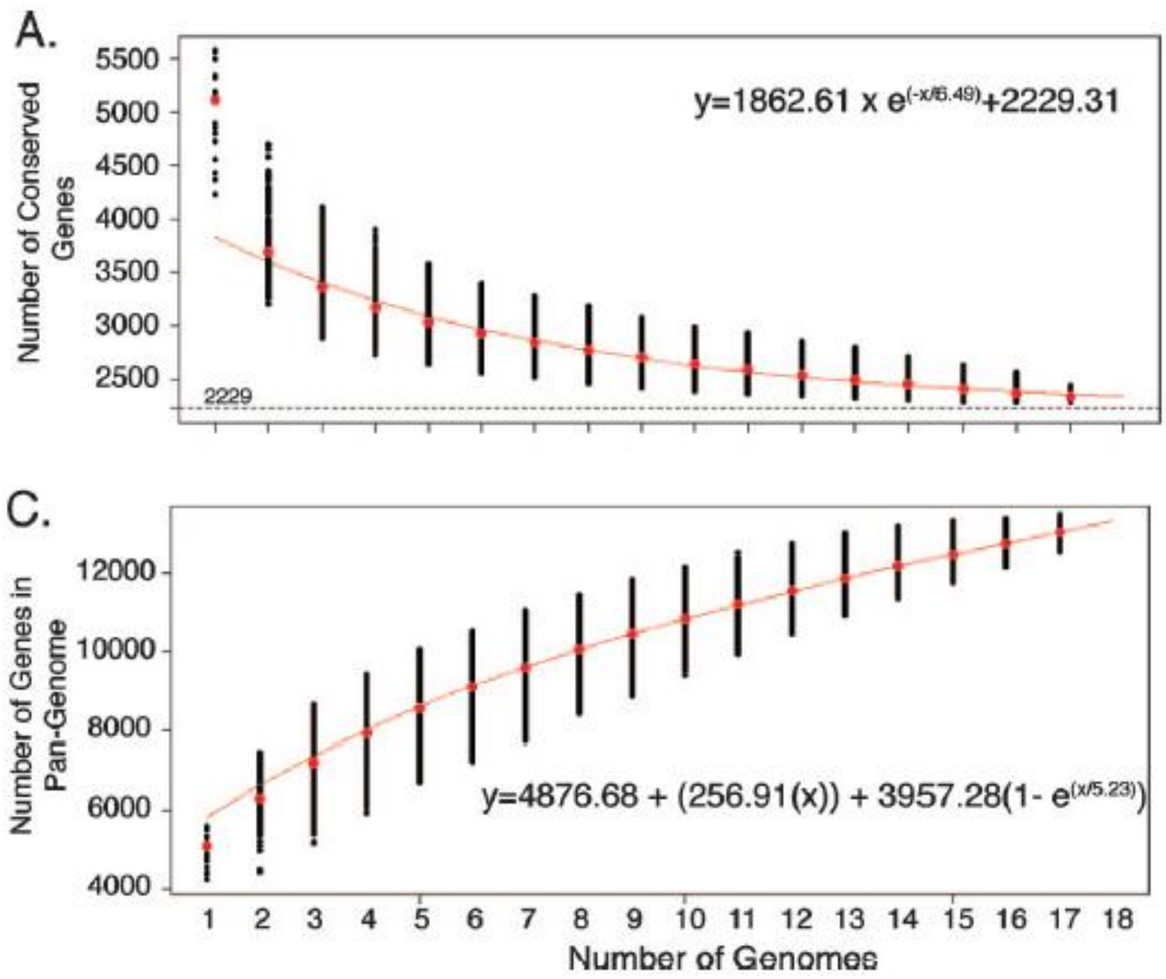
Background:
Gene Content Variation among E. coli genomes.
Evidence for horizontal transfer –



Welch et al (2002).

Core genome = intersection of sets
Pangenome = union of sets

Background:
Core and Pan-genomes in large data sets



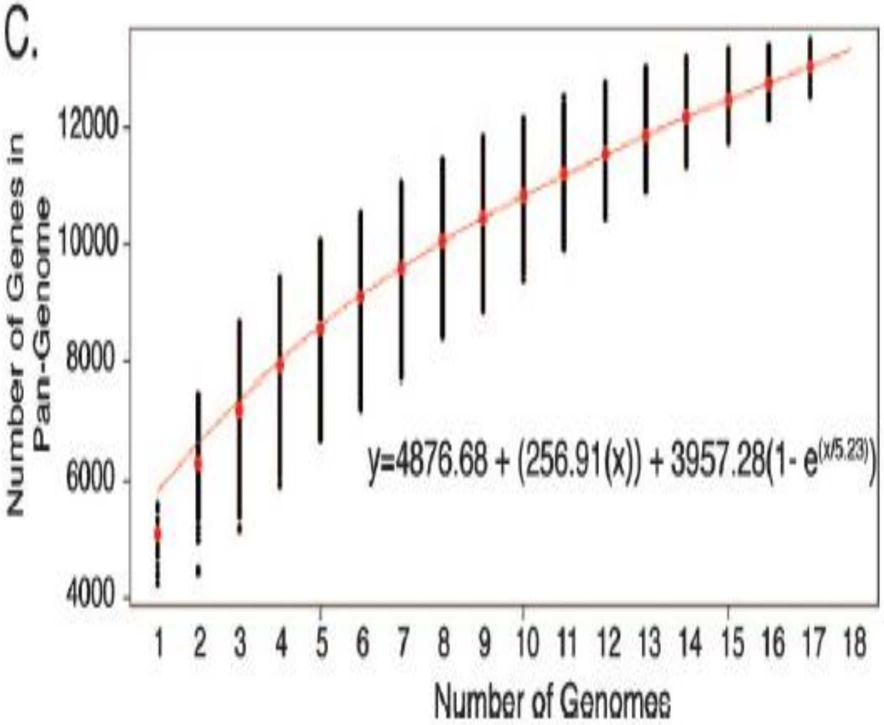
The Core genome is much smaller than the typical genome size

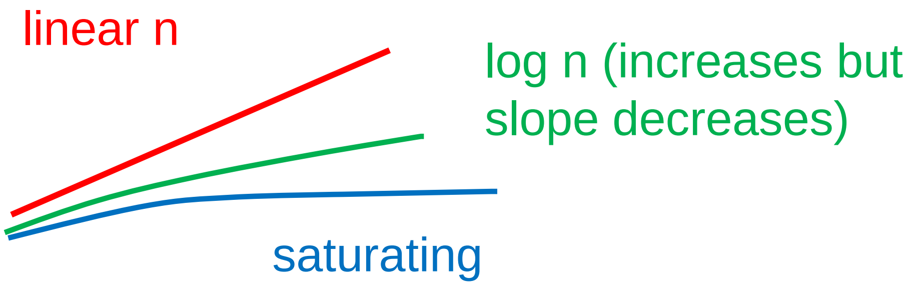
The Pan-genome is much larger than the typical genome size and keeps increasing

Core and Pan-genome of E. coli

Rasko et al (2008) J. Bacteriol.

Background:
Is the Pangenome Open or Closed?





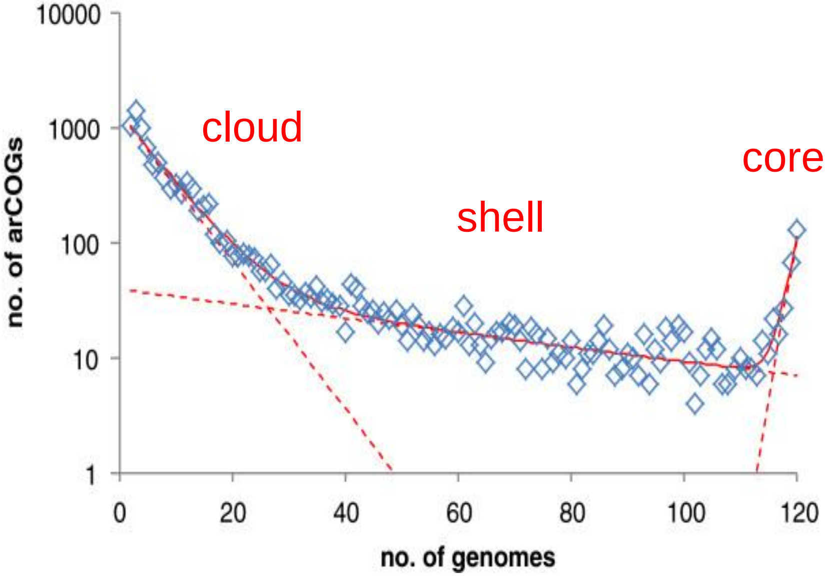
linear n

log n (increases but slope decreases)

saturating

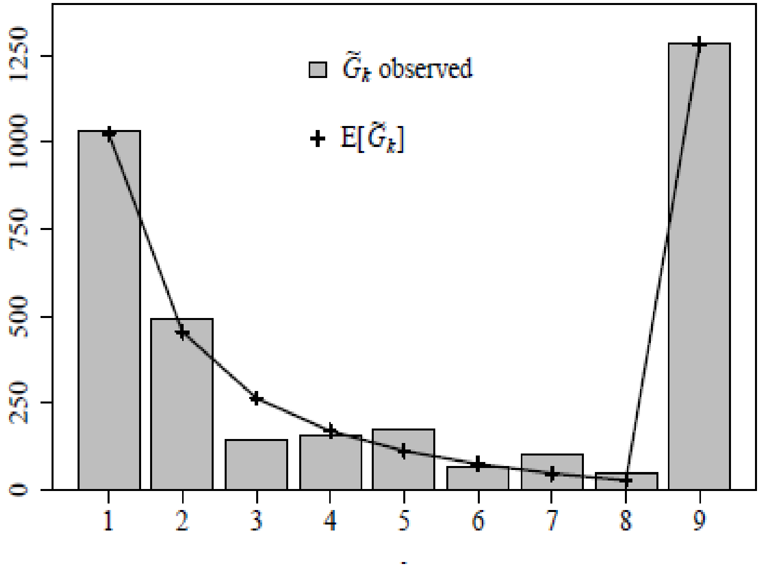
Gene Frequency Spectra
Characteristic U shape applies at large and small scales

120 Archaeal genomes



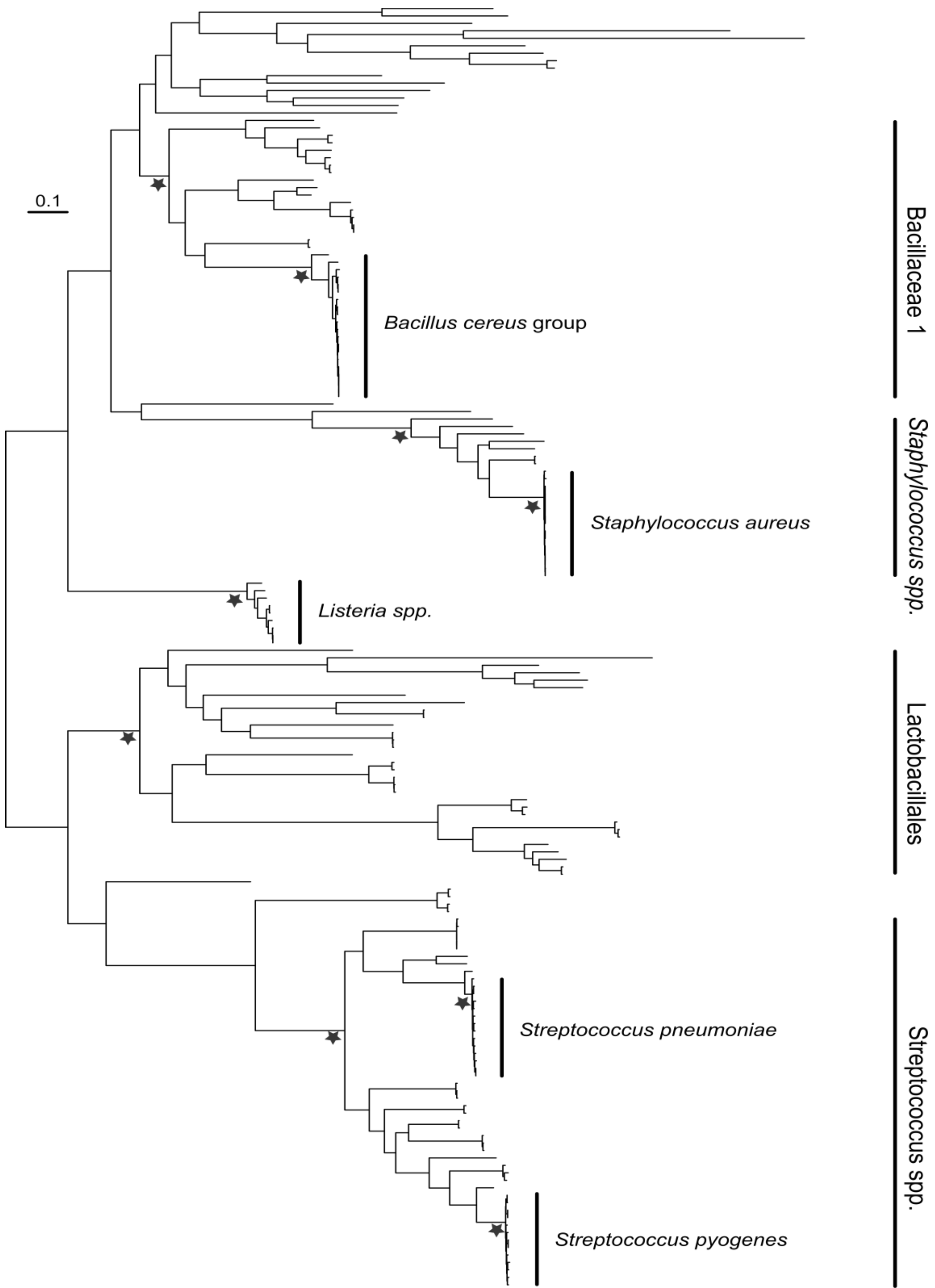
Wolf et al. (2012)

9 Prochlorococcus genomes



Baumdicker et al. (2009)

Introduced the Infinitely Many Genes model



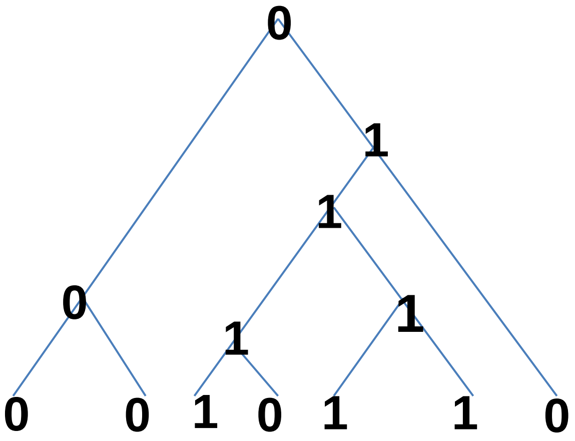
Part 1-
Collins and Higgs
Mol Biol Evol (2012)

Analyzed 172 complete genomes of Bacilli using the IMG

Looked and core and pangenomes and gene frequency spectra

Looked at full set and at clades of different sizes – indicated by *

The Infinitely Many Genes Model : IMG



Overall rate of gain of genes *u*
Loss rate per gene *v*

Each new gene is different from those already in the genome.

A gene may be deleted multiple times in different lineages

Gain could be **either** horizontal transfer from a diverse external pool **or** gene origin within the lineage (duplication or *de novo* ORF)

There are an infinite number of possible genes “out there”, but the number of genes in a genome remains finite

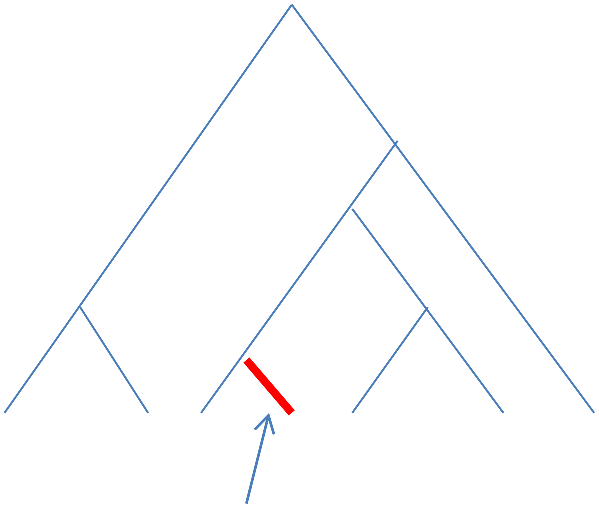
$$\begin{tikzcd} n\text{-}1 \arrow[r, shift left, "u"] & n \arrow[l, shift left, "nv"] \end{tikzcd}$$

Mean number of genes is *u/v*

Probability of having n genes

$$P(n) = \frac{(u/v)^n}{n!} \exp(-u/v)$$

IMG on a Coalescent Tree



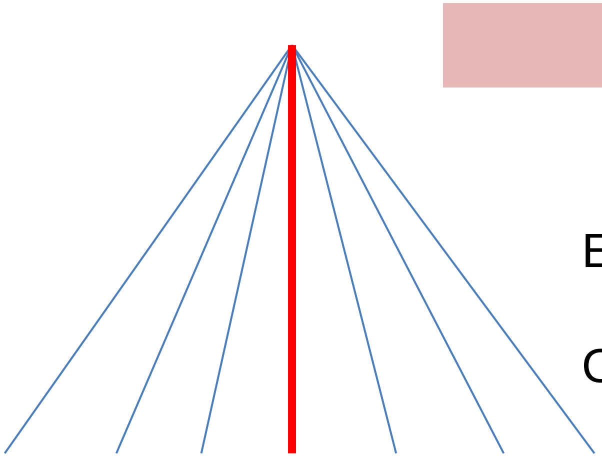
$$G_{\mathrm{pan}}(n) = \theta \sum_{k=0}^{n-1} \frac{1}{k+\rho}$$

$\theta = 2Nu$ and $\rho = 2Nv$.

Open pangenome with logarithmic increase

Extra branch added due to adding the nth genome gets shorter like 1/*n*

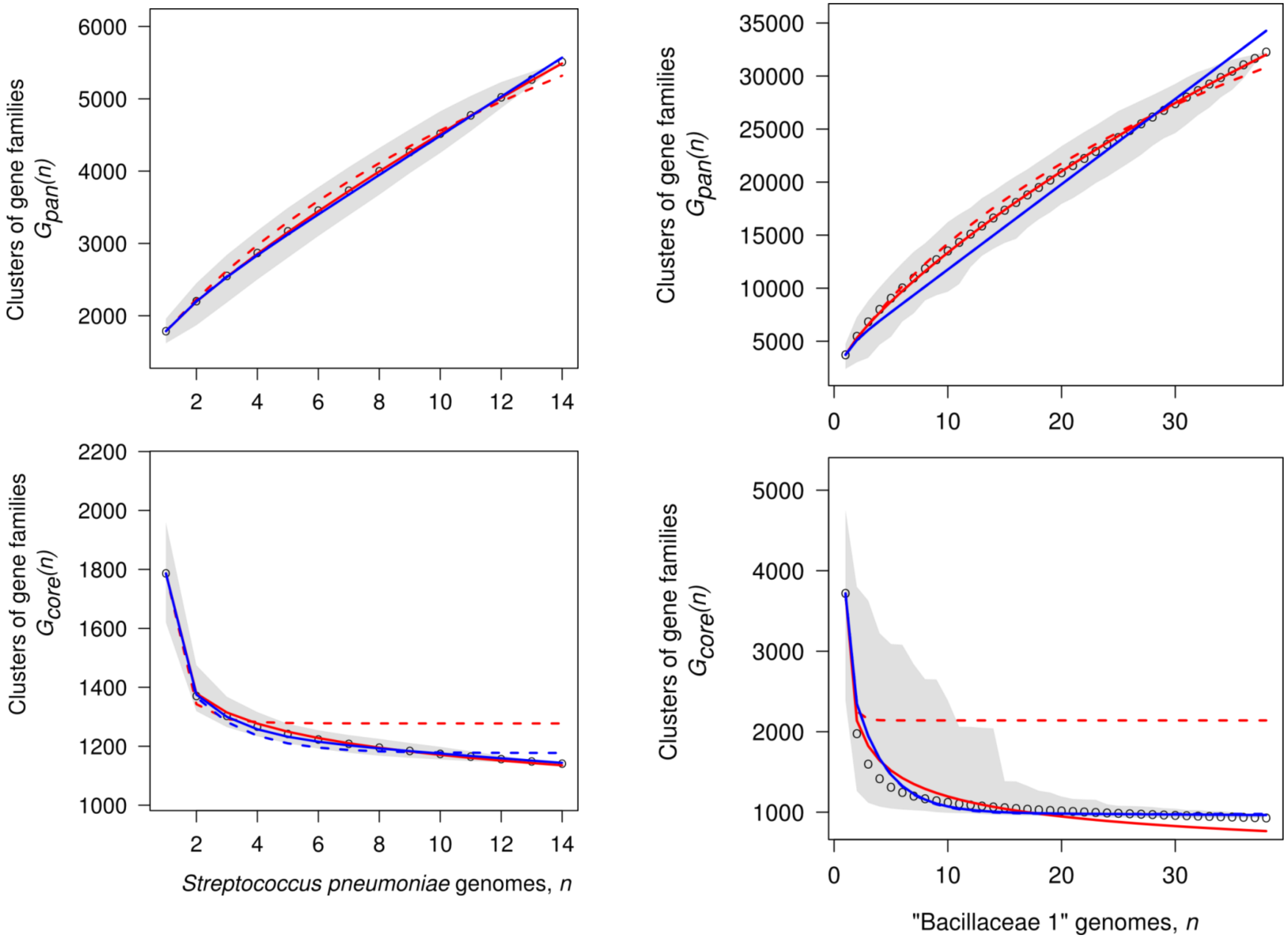
Star Tree



Extra branch added is constant length.

Open pangenome with linear increase

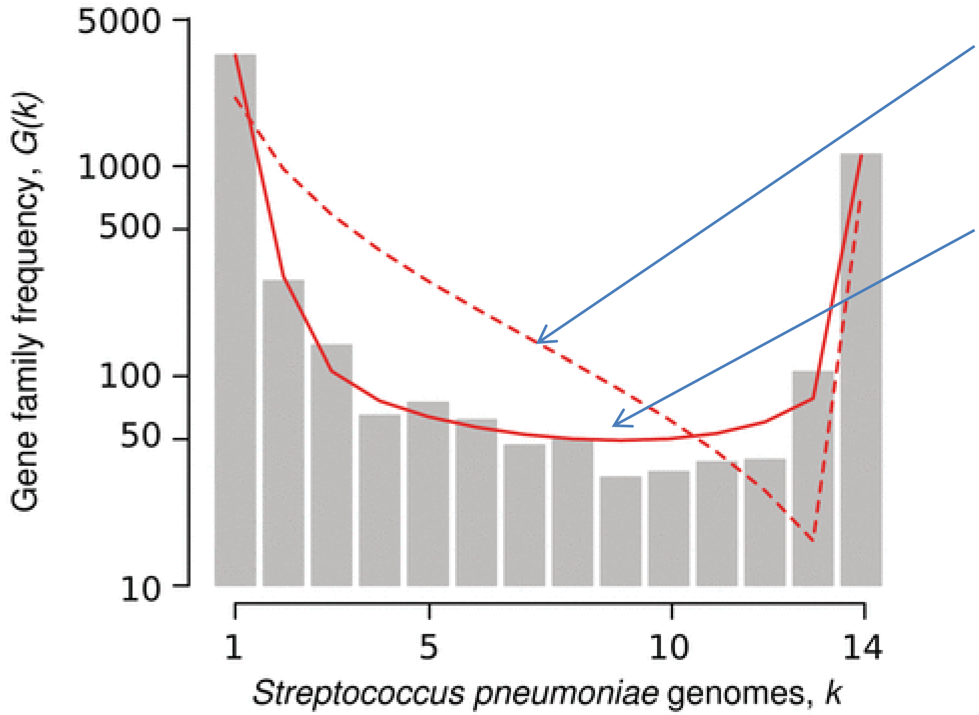
IMG Fits Pan and Core genome sizes with coalescent tree – but you need multiple rate classes



Collins and Higgs - Mol Biol Evol (2012)

Gene frequency spectrum can be calculated for IMG

$$G(k|n) = \frac{\theta}{k}\frac{n\ldots(n-k+1)}{(n-1+\rho)\ldots(n-k+\rho)}$$



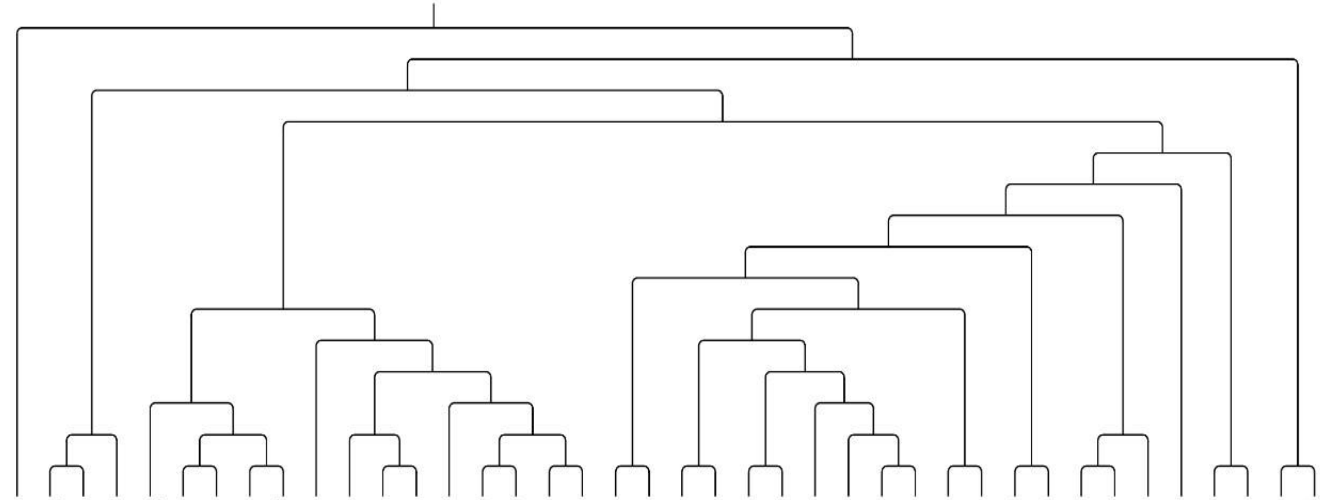
Essential + 1 Variable class

Essential + 2 Variable classes

Conclusion (Part 1) – IMG is useful description of Pangenome data and Gene frequency spectra

Part 2 – Phylogenetic pattern data

Current work of Seyed Zamani Dahaj



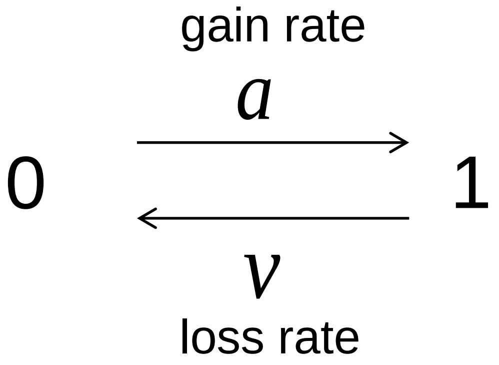
	GLVIO	THEEB	CYAP4	ACAM1	TRIEI	NOSA0	NOSP7	NOSS1	ANAVT	SYNP2	MICAN	CYAP7	CYAP2	SYNY3	CYAP8	CYAP0	UCYNA	CYAA5	PROMM	PROM3
P1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
P2	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0
P3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
P4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P5	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0
P6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
P7	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	PROMT	PROM1	PROMP	PROM5	PROM9	PROM2	PROMS	PROM0	PROM4	PRMAR	SYNS3	SYNPW	SYNS9	SYNSC	SYNPX	SYNR3	SYNP6	SYNE7	SYNJB	SYNJA	Scenario	Gamma Cat
P1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
P2	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	2
P3	1	1	1	0	1	0	1	1	1	1	1	1	0	1	0	1	1	1	0	0	1	4
P4	1	1	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	5
P5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	2	3
P6	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	2	1
P7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	2	4

Examples of patterns in cyanobacteria

Much more information than gene frequency spectra

Two-state Phylogenetic model for presence/absence
Finitely Many Genes model (FMG)



gain rate

a

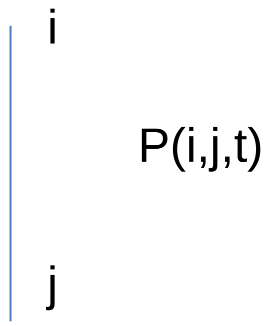
0

1

v

loss rate

What happens on one branch?



i

P(i,j,t)

j

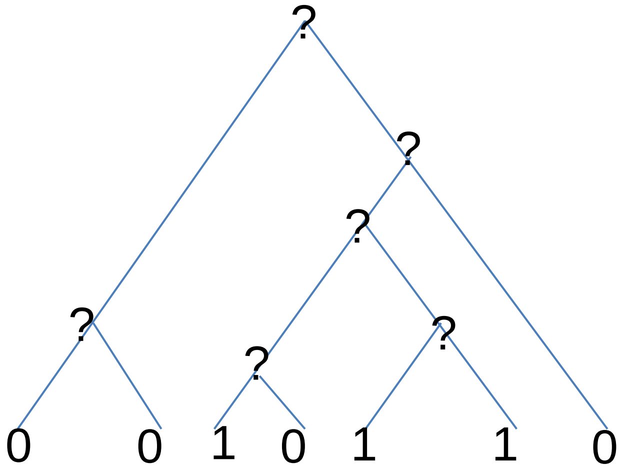
$$P(0,0,t)=\frac{v}{a+v}+\frac{a}{a+v}e^{-(a+v)t}$$

$$P(0,1,t)=\frac{a}{a+v}(1-e^{-(a+v)t})$$

$$P(1,0,t)=\frac{v}{a+v}(1-e^{-(a+v)t})$$

$$P(1,1,t)=\frac{a}{a+v}+\frac{v}{a+v}e^{-(a+v)t}$$

Likelihood of a data pattern on a tree L_{pat}



The importance of *M*

Number of possible kinds of genes M

The null pattern 000000000000 is invisible! We can't count the genes that are not there! We don't know what *M* is.

Probability that a gene is present $\dfrac{a}{a+v}$

Mean number of genes per genome $G = M\dfrac{a}{a+v}$

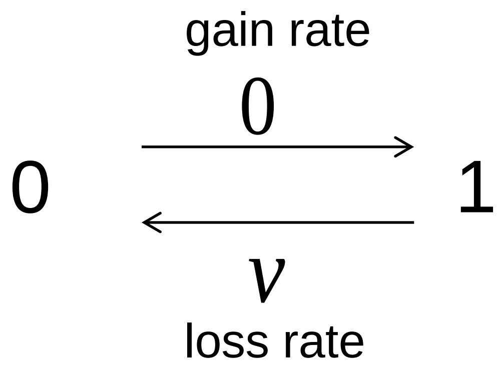
Expected number of occurrences of a pattern $N_{pat} = ML_{pat}$

Expected number of observable patterns $N_{obs} = M(1 - L_{null})$

Therefore number of occurrences of a pattern $N_{pat} = \dfrac{N_{obs}L_{pat}}{1 - L_{null}}$

From this we can calculate the likelihood of the full set of patterns and optimize the parameters to maximize this likelihood.

IMG as a limit of FMG



Take limit $a \to 0$ $M \to \infty$

With total gain rate fixed $u = Ma$

Expected number of genes in the genome is finite $G = u / v$

$$P(0,0,t) = 1$$

$$P(0,1,t) = 0$$

$$P(1,0,t) = 1 - e^{-vt}$$

$$P(1,1,t) = e^{-vt}$$

Problem:

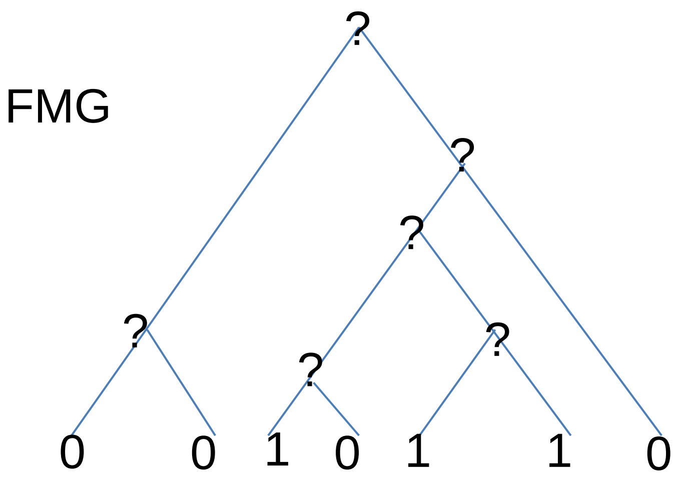
L_{null} = 1

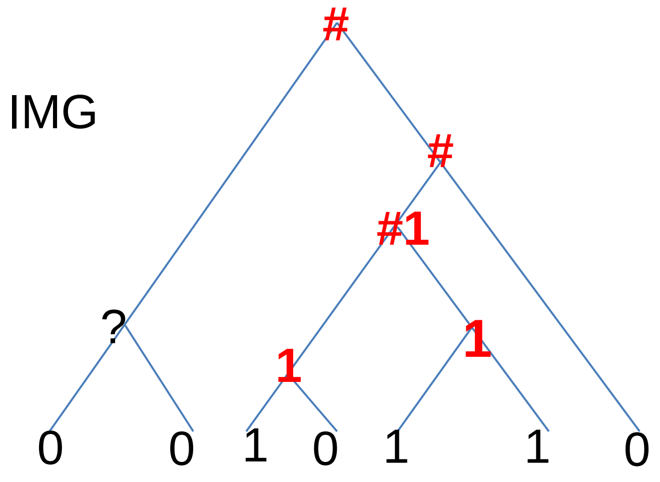
L_{pat} = 0 for all other patterns

$$N_{pat} = \frac{N_{obs} L_{pat}}{1 - L_{null}} = \frac{0}{1-1} = ?$$

Actually N_{pat} is finite, but we can't calculate it in the usual way

Calculating N_{pat} for the IMG model



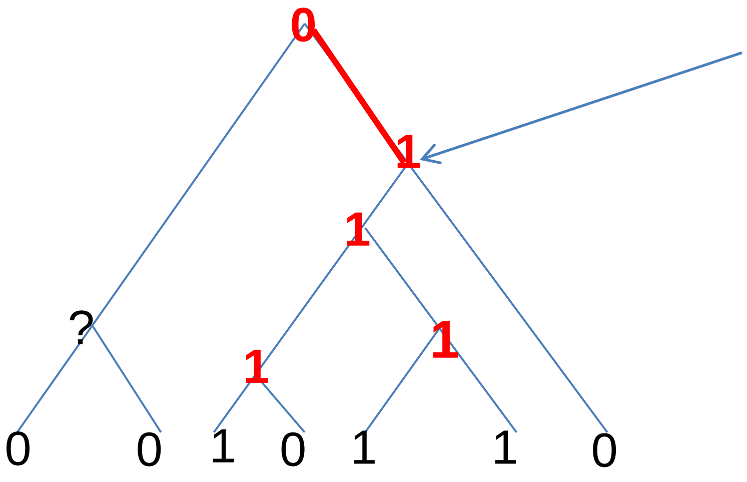


For FMG all the nodes labelled ? Can be 0 or 1

For IMG the nodes labelled **1** must be 1.
The **#** nodes are possible **origin nodes** for the gene.

Need to consider cases where genes originated on the branches leading to each of the **#** nodes

Calculating N_{pat} for the IMG model



N_{node} = number of new genes arising on the branch leading to that node

$$\frac{dN_{node}}{dt} = u - vN_{node}$$

$$N_{node} = \frac{u}{v}(1 - e^{-vt})$$

For the root node $N_{node} = u/v$

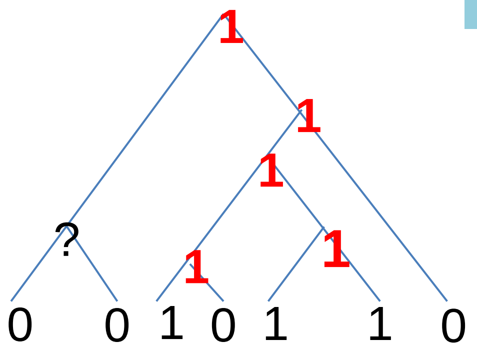
$P_{pat}(node)$ = probability that the pattern arises given that there was a 1 at the origin node. This is finite and can be calculated the usual way

$$N_{pat} = \sum_{\#-nodes} N_{node} P_{pat}(node)$$

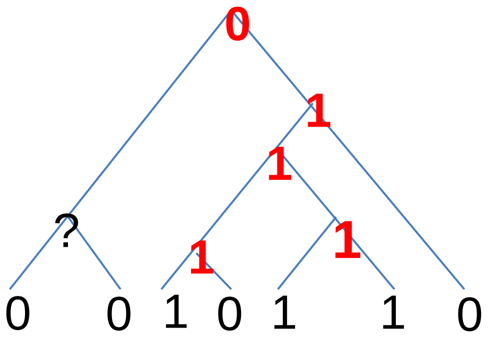
N_{pat} is finite for non-null patterns. We can normalize so that

$$N_{obs} = \sum_{pat} N_{pat}$$

Three Scenarios for Gene Histories



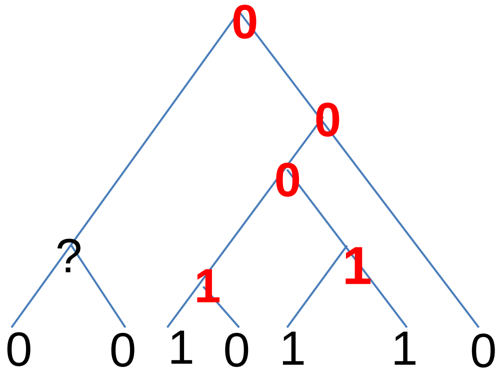
Scenario 0 means 0 gains
The gene was present at the root



Scenario 1 means 1 gain
The gene was not present at the root, and was gained once.

A single gain could be

- de novo (in the lineage)
- gene duplication (in the lineage)
- horizontal transfer (from outside the lineage)



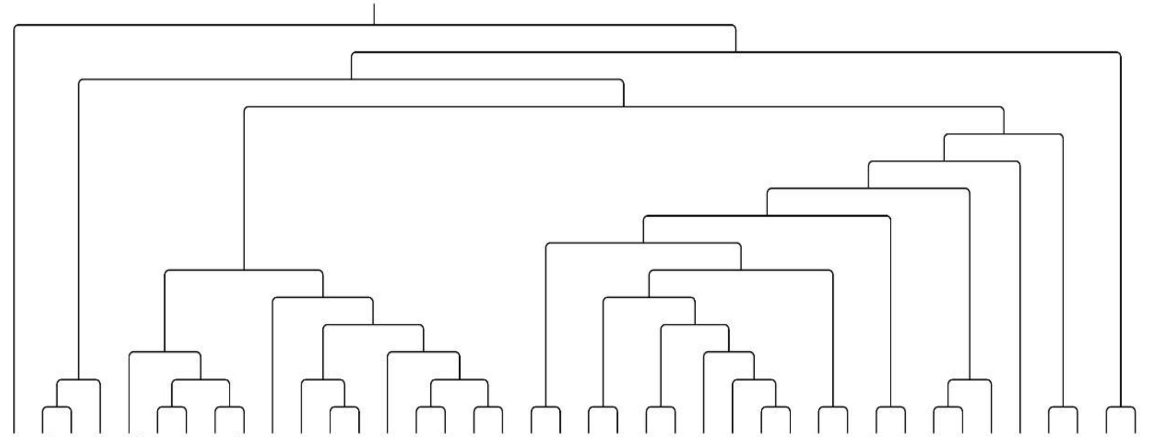
Scenario 2 means 2 or more gains

A second gain must be due to horizontal transfer.
This scenario is not possible for IMG

Example - 40 Cyanobacterial Genomes (Seyed Zamani Dahaj)

Gene clusters from HOGENOM database – (Penel et al BMC Bioinformatics (2009))
Contains 1470 complete genomes – release 06 (Dec 2011)

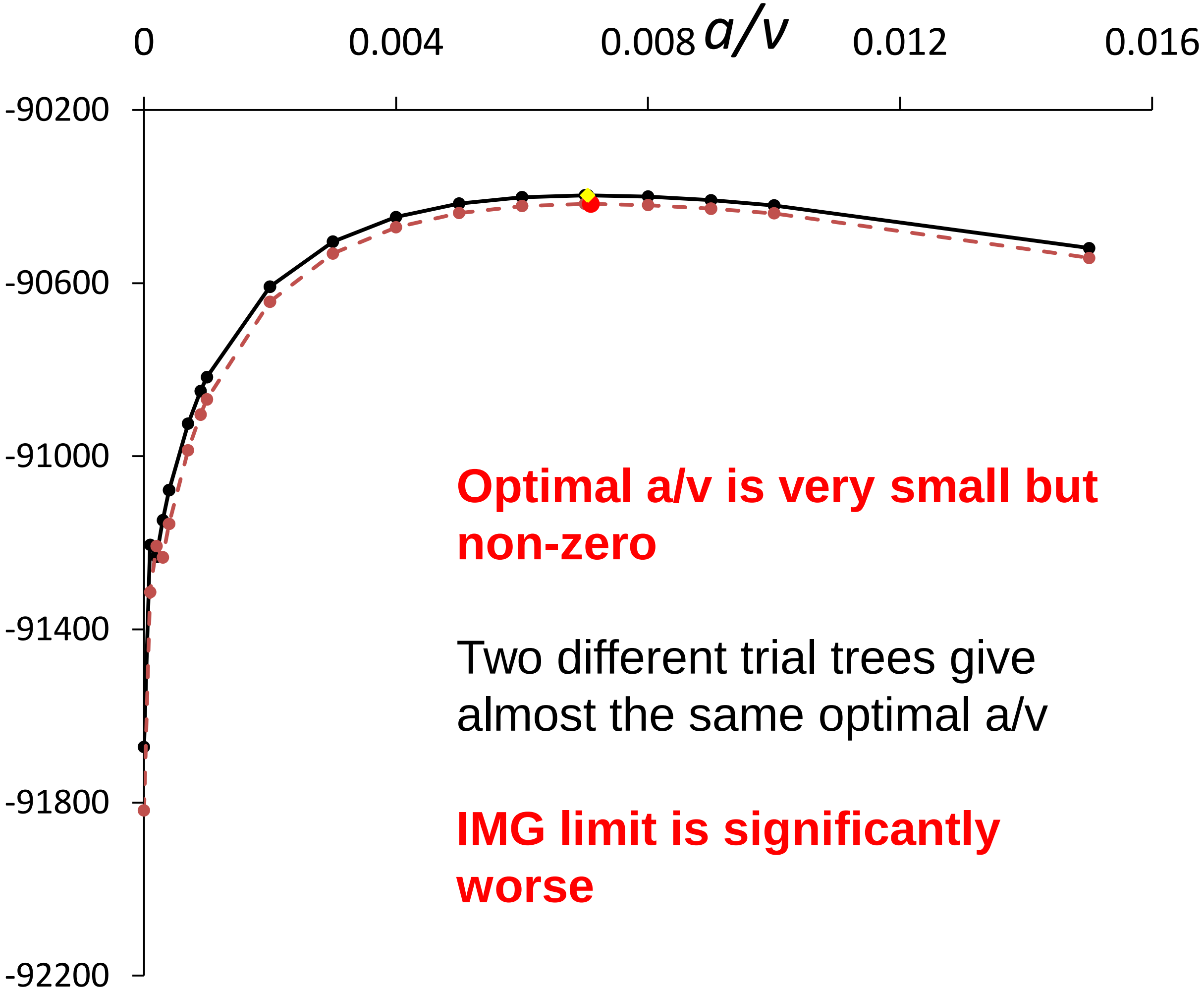
10304 gene clusters with at least two genes in Cyanobacteria
3510 distinct presence/absence patterns



	GLVIO	THEEB	CYAP4	ACAM1	TRIEI	NOSA0	NOSP7	NOSS1	ANAVT	SYNP2	MICAN	CYAP7	CYAP2	SYNY3	CYAP8	CYAP0	UCYNA	CYAA5	PROMM	PROM3	PROMT	PROM1	PROMP	PROM5	PROM9	PROM2	PROMS	PROM0	PROM4	PRMAR	SYNS3	SYNPW	SYNS9	SYNSC	SYNPX	SYNR3	SYNP6	SYNE7	SYNJB	SYNJA	Scenario	Gamma Cat
P1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
P2	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	2
P3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	1	0	1	1	1	1	1	1	0	1	0	1	1	1	0	0	1	4
P4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	5
P5	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	2	3
P6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	2	1
P7	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	2	4

Comparison of IMG and FMG gives statistical test for presence of HGT in presence/absence data

Log Likelihood as a function of a/v when all rate categories have same a/v



Optimal a/v is very small but non-zero

Two different trial trees give almost the same optimal a/v

IMG limit is significantly worse

Substantial variation in a/v ratio found among different rate categories

cat	G	v	a	a/v	M
1	749.14	0.0290	0.0102	0.3553	2857.58
2	583.20	0.189	$<10^{-7}$	$<10^{-7}$	5.70 ´ 10^{9}
3	228.27	0.4928	0.3749	0.7609	528.23
4	458.39	1.0471	0.0323	0.0309	15258.44
5	918.49	2.4093	0.0007	0.0003	2.94 ´ 10^{6}

These are almost IMG

If we use IMG in these two classes there is no significant difference (AIC prefers IMG)

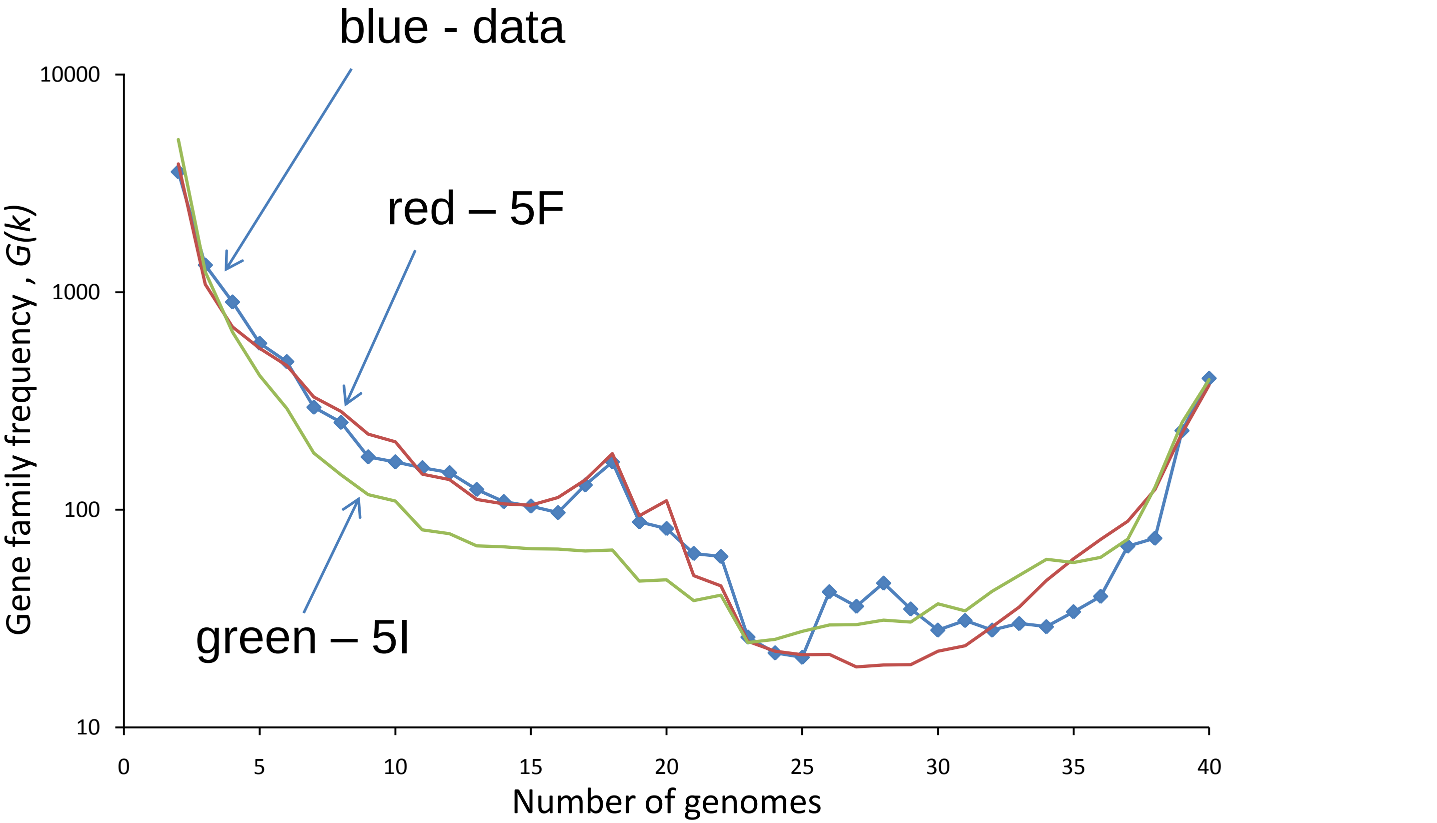
Remember that *M* diverges when *a/v* goes to zero.

$$M = G(1+\frac{1}{a/v})$$

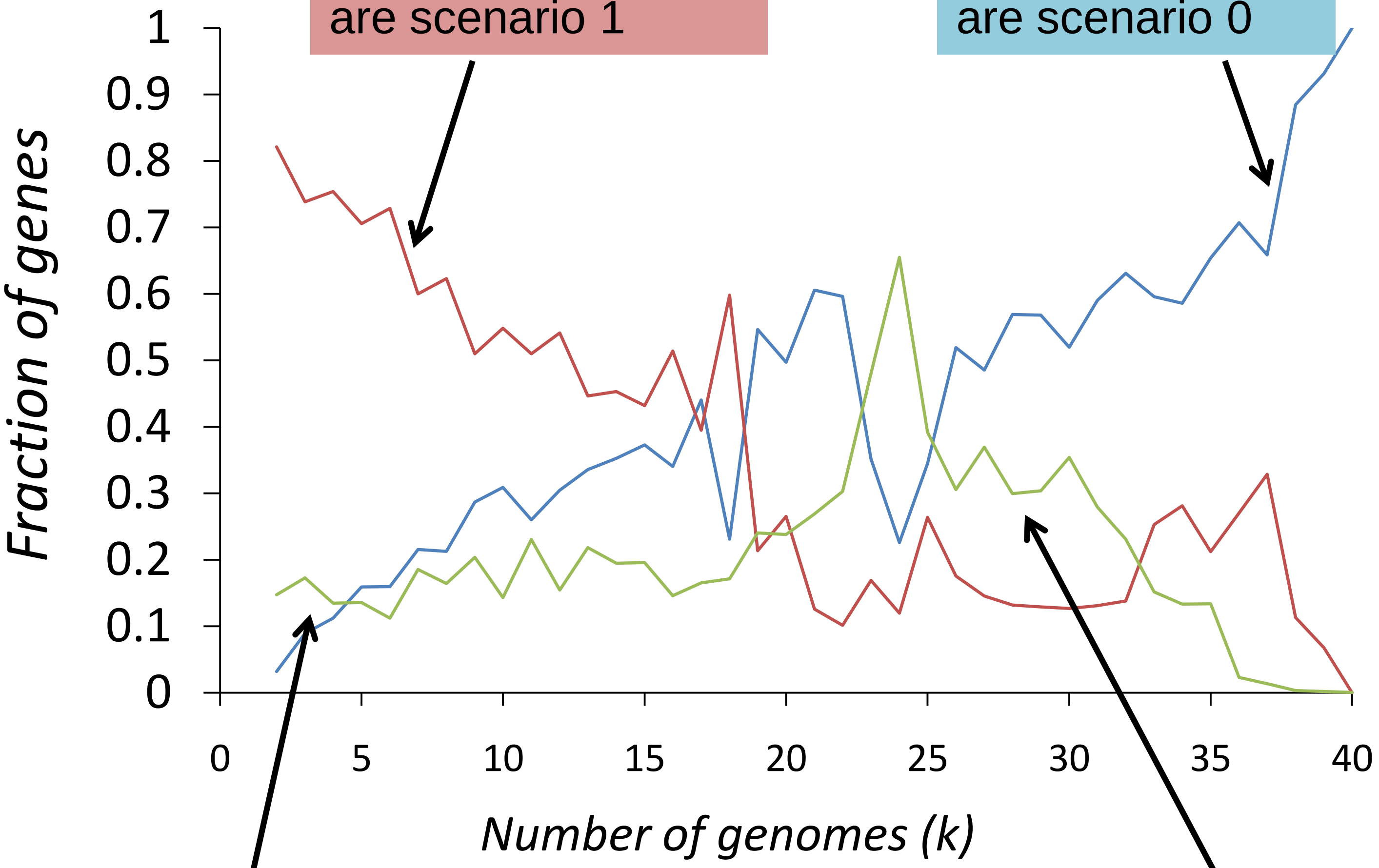
Expected number of observed patterns (gene clusters) in the three scenarios

	5F	FIFFI	5I
SCENARIO 0	2186.0 (21.2 %)	2141.6 (20.8 %)	2806.4 (27.2 %)
SCENARIO 1	6559.5 (63.7 %)	6613.1 (64.2 %)	7497.6 (72.8 %)
SCENARIO 2	1558.5 (15.1 %)	1549.2 (15.0 %)	0

Fitting the gene frequency spectrum



FMG gives noticeable improvement over IMG to predicting G(k)



Most Cloud genes are scenario 1

Most Core genes are scenario 0

Low probability here means rare genes are usually signatures of small related groups

Shell genes have the highest probability of being in scenario 2 BUT there are few of these genes

Conclusions – Part 2

- Comparison of IMG and FMG allows us to test for the presence of HGT
- Mean gain/loss ratio is small – a/v ~ 0.007
- Approx 15% of cyanobacterial gene clusters are best explained by scenario 2 (multiple insertions)
- Broad range of deletion rates among genes explains why there are signature genes even though there is rapid gain and loss
- Presence/absence patterns support the view that there is a strong signal of an underlying species tree

Where do new genes come from?
There must be a high rate of origin of genes within lineages.
Most of the genes seen in small related groups originated where we see them.

Part 3 - Does natural selection favour high or low rates of HGT?

Vogan and Higgs – Biology Direct (2011)

Advantages:

New beneficial genes arise rarely → Much quicker to acquire a new gene horizontally than to invent it for yourself.
Can gain new metabolic pathways (e.g. photosynthesis, antibiotic resistance).
Can replace lost or damaged genes

Disadvantages:

New gene may be a duplicate or non functional – cost of junk DNA
New gene may disrupt an existing gene
New gene may be a selfish replicator (transposable elements)
New gene may be a harmful parasite (virus)

Who Controls Gene Transfer?

The Recipient Cell: **YES**

Mechanisms of DNA uptake
Mechanisms recombination inside cell
Mechanisms of break-up of DNA fragments
Mechanisms of silencing inserted genes
Variation in cell wall thickness

Develop model to describe evolution of the recipient

The Donor Cell: **NO**

For transformation it is probably a fragment of DNA from a dead cell, so there isn't a donor.

When live cells transfer genes it is usually not controlled by the donor genome.

The Genes Themselves: **YES**

Transposable elements
Plasmids
Viruses

Model for evolution of HGT – simplest benefit v. simplest disadvantage

Population of *N* cells, each with a genome = list of genes, e.g. 3-4-7-1-9-6-1

Fitness $$w = \frac{1 + s n_{diff}}{1 + c n_{tot}}$$

n_{tot}= total number of genes, n_{diff} = number of different types

s = selective benefit for each new type of gene
c = cost per gene

$s > c$
Fitness increases with each new gene, but duplicate genes are penalized.

Moran model – birth and death model in population genetics.

Replication Process

Each gene in the parent is copied successfully to the offspring with probability 1-*v* or lost/deleted with probability *v*

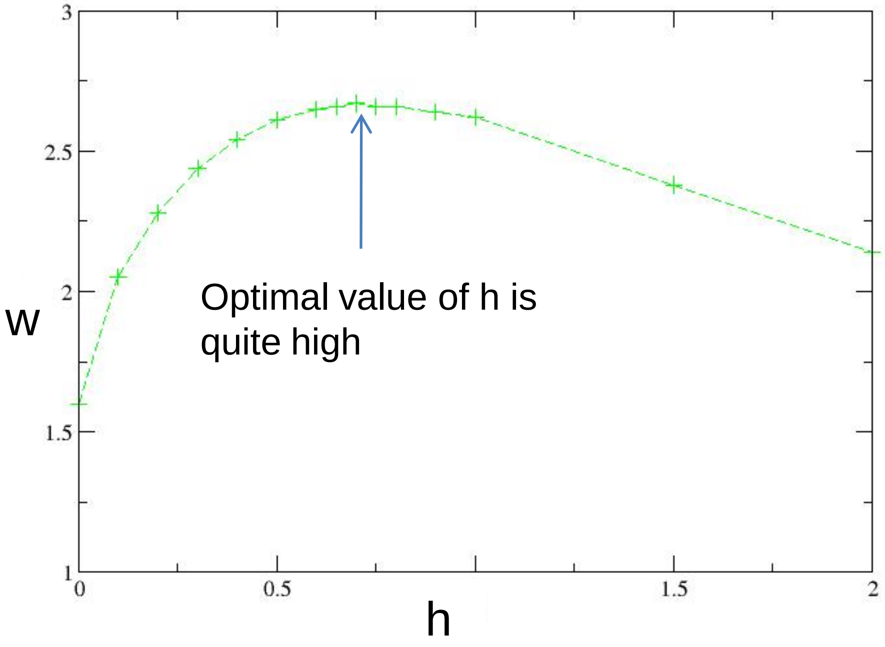
A new gene arises in the new cell with probability u_{new}

The new cell has the opportunity to acquire genes horizontally – mean number acquired is *h*.

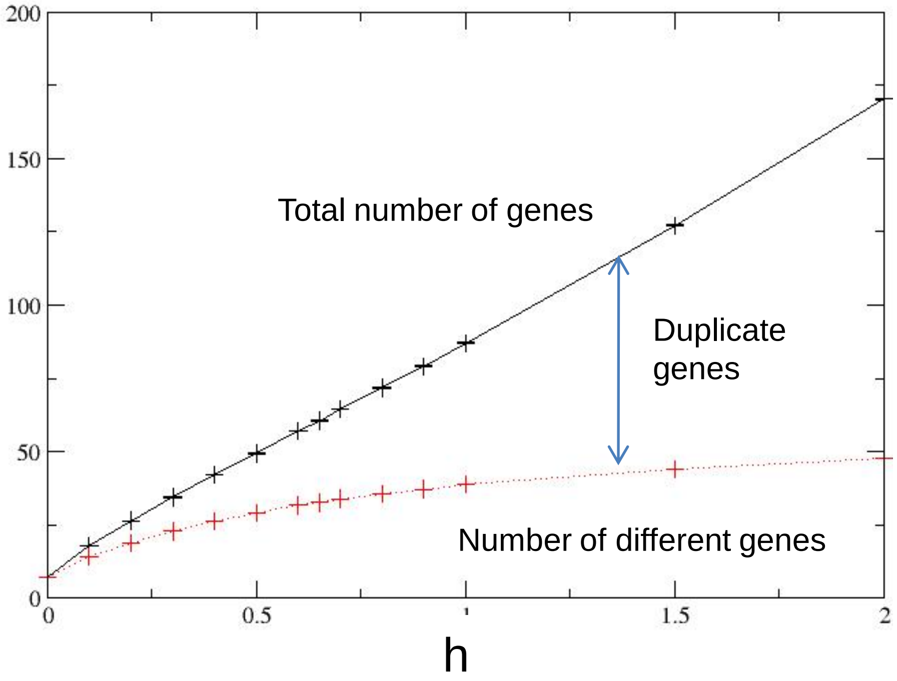
Probability of acquiring *k* genes is $$P(k) = \frac{h^k}{k!}\exp(-h)$$

Each acquired gene is a copy of a random gene from a random individual in the population (assumed to be representative of DNA fragments available).

Small rate of origin of genes u_{new} = 0.002 per genome
Large rate of deletion v = 0.01 per gene



Optimal value of h is quite high



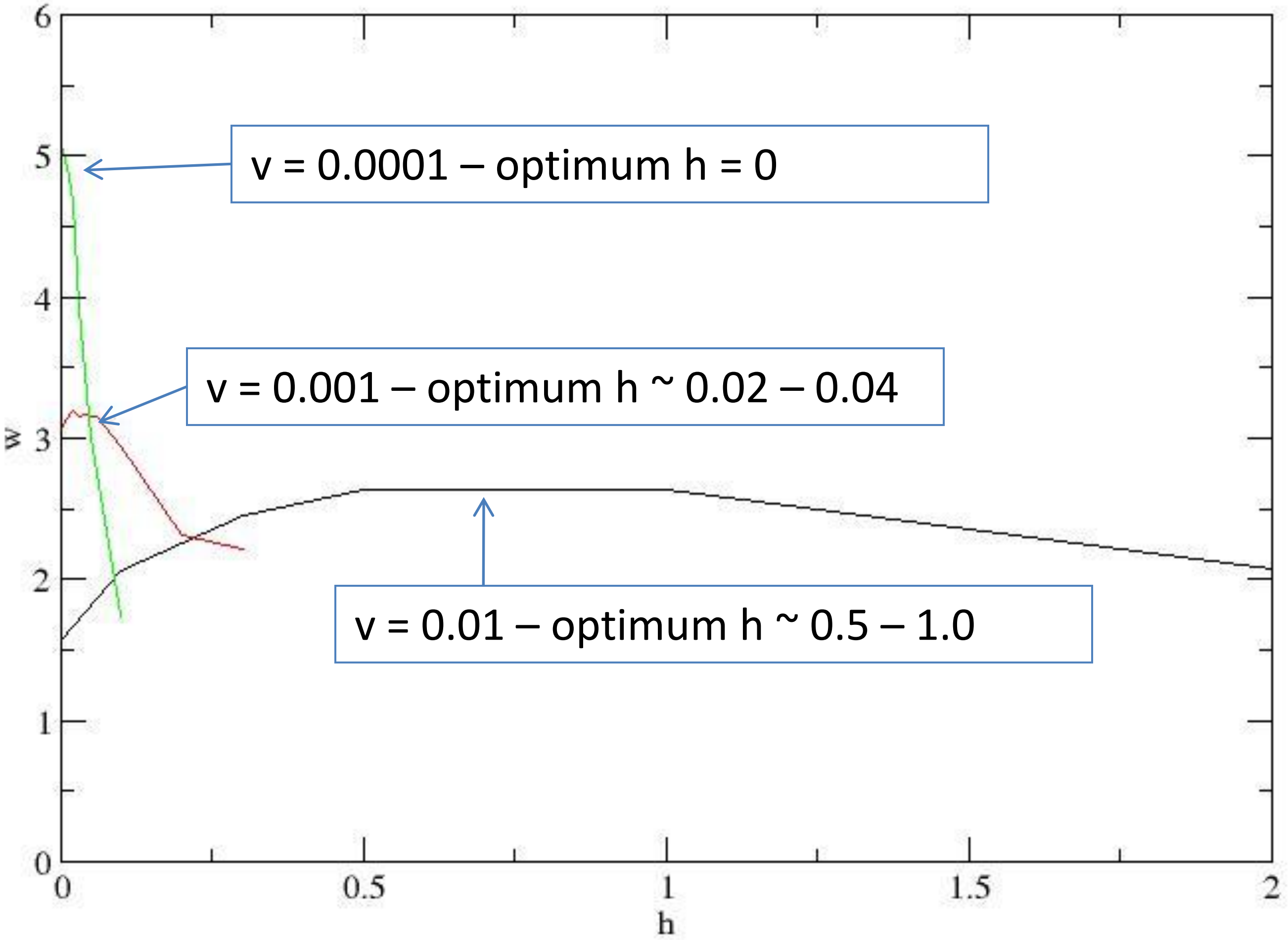
Total number of genes

Duplicate genes

Number of different genes

HGT allows large genomes to be built up and maintained in early phases of evolution when genome replication is very inaccurate.

The optimum h depends on the accuracy of replication

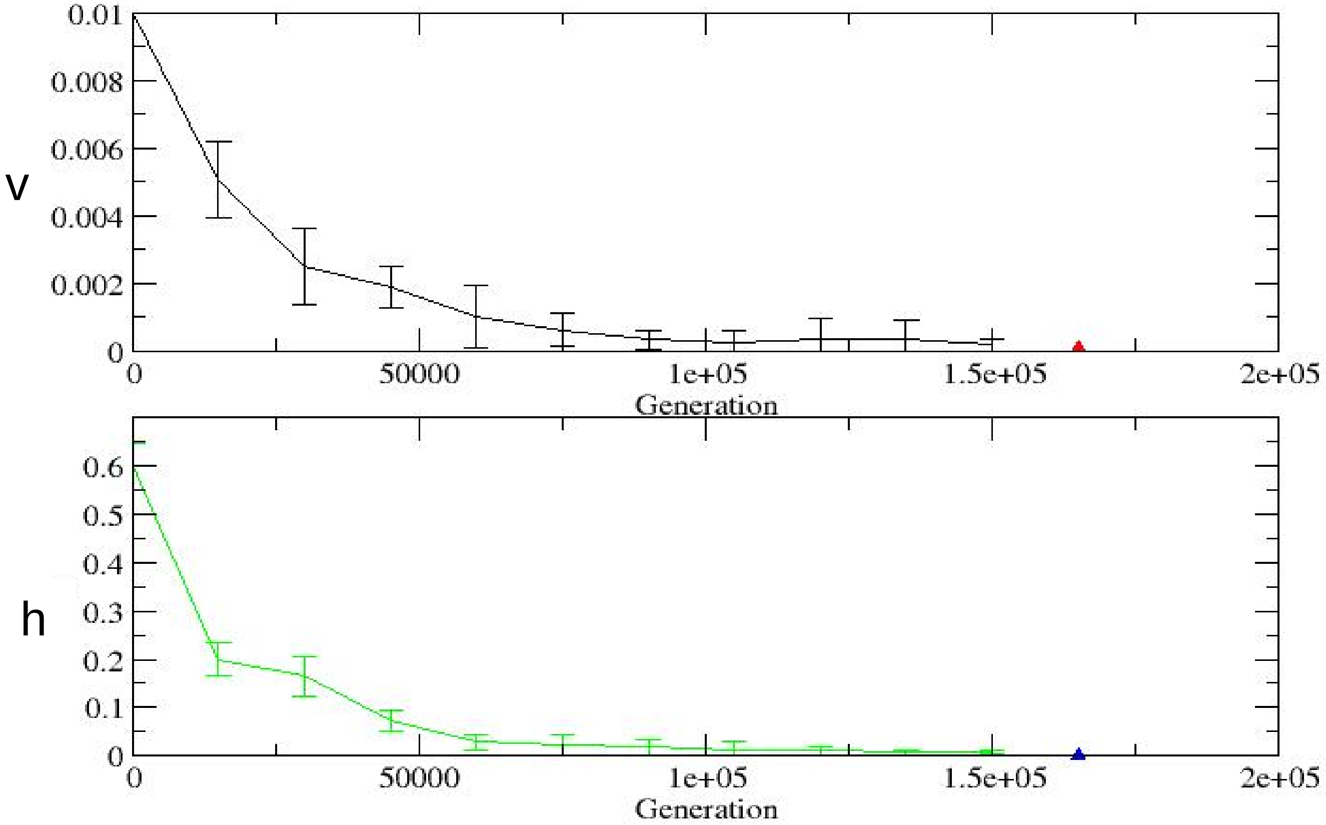


v = 0.0001 – optimum h = 0

v = 0.001 – optimum h ~ 0.02 – 0.04

v = 0.01 – optimum h ~ 0.5 – 1.0

Allow v and h to be heritable properties of cells

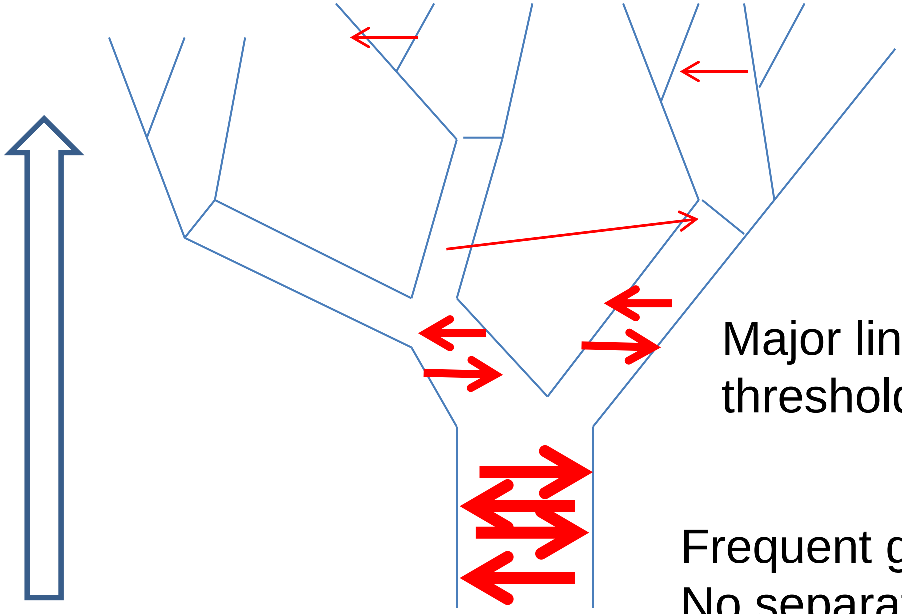


Accurate replication is advantageous → v decreases over time.
Low h is advantageous when v is low → h decreases over time.

Organisms evolve to a state with low HGT *if they can!*
Maybe HGT in modern organisms is a result of selfish transposable elements.

The Darwinian Threshold

Carl Woese argued that HGT was so frequent early on that there are no separate lineages. Lineages emerge later – "Darwinian Threshold"



Reduction in HGT over time

Many separate species with rare HGT

Major lineages emerge at the threshold

Frequent gene sharing. No separate lineages

The Vogan and Higgs model predicts why it was advantageous to have high HGT early on, why HGT should reduce over time; hence this supports the Darwinian threshold picture.

Interpretation

BEWARE – the name Darwinian threshold makes people think that Darwinian evolution begins at this point.

In my view – Darwinian evolution at the gene level was going on way before this.

What hapens at the threshold is the level of selection changes

Before the threshold – selection is on individual genes in an ever-changing mixture.

After the threshold – selection is on teams of linked genes that are inherited together vertically

Conclusions – Part 3

When replication accuracy is poor HGT is favourable.

Early organisms needed HGT to build up large genomes.

When replication accuracy is good HGT is unfavourable .
If other disadvantages were included, it would be even more unfavourable.

Modern organisms should avoid HGT if possible.

But it may not be possible – viruses and transposable elements are out for themselves.

Occasionally a cell may benefit by HGT even so (resistance genes etc).

There may be benefits of transformation that are not due to HGT (e.g. food source, or DNA repair) but HGT would be a side effect.

Evolution should move from a tangle to a tree as replication accuracy gets better.

Lineages should emerge as evolution proceeds.