4

Genomes in flux: the evolution of Archaeal and Proteobacterial gene content

Berend Snel, Peer Bork and Martijn A. Huynen

Genome Research **12** (2002) 17-25

Abstract

In the course of evolution, genomes are shaped by processes like gene loss, gene duplication, horizontal gene transfer, and gene genesis (the de novo origin of genes). Here we reconstruct the gene content of ancestral Archaea and Proteobacteria and quantify the processes connecting them to their present day representatives based on the distribution of genes in completely sequenced genomes. We estimate that the ancestor of the Proteobacteria contained around 2500 genes, and the ancestor of the Archaea around 2050 genes. Although it is necessary to invoke horizontal gene transfer to explain the content of present day genomes, gene loss, gene genesis, and simple vertical inheritance are quantitatively the most dominant processes in shaping the genome. Together they result in a turnover of gene content such that even the lineage leading from the ancestor of the Proteobacteria to the relatively large genome of *Escherichia coli* has lost at least 950 genes. Gene loss, unlike the other processes, correlates fairly well with time. This clock-like behavior suggests that gene loss is under negative selection, while the processes that add genes are under positive selection.

Introduction

How the gene content of a genome evolves is an important, complicated, and still largely open question. The evolution of the gene content has been studied with regard to both large-scale trends as well as specific processes. Many studies have focused on specific aspects of genome evolution or have tried to reconstruct a specific ancestral genome (Bruccoleri *et al.* 1998; de Rosa and Labedan 1998; Huynen and Bork 1998; Kyrpides *et al.* 1999; Makarova *et al.* 1999; Aravind *et al.* 2000; Ochman and Jones 2000; Jordan *et al.* 2001). Large-scale studies on the presence and absence of genes have shown that the number of shared genes between genomes depends on the size of genomes (Chapter 3.2), and their evolutionary distance (Gaasterland and Ragan 1998; Huynen and Bork 1998; Fitz-Gibbon and House 1999; Chapter 3.2; Tekaia *et al.* 1999). Correlation in the presence of genes has been used to predict functional interactions between genes (Pellegrini *et al.* 1999; Huynen and Snel 2000). These observations suggest that evolutionary history, genome size, and functional selection together determine gene content. The role of the specific processes involved in the evolution of gene content of specific genomes has also been emphasized. Massive gene duplication was postulated in the ancestor of *Vibrio cholerae* (Heidelberg *et al.* 2000), massive gene loss in the ancestor of *Buchnera* (Shigenobu *et al.* 2000), and massive horizontal gene transfer (HGT) to the ancestors of *E.coli* 0157:H7 and *E.coli* K12 (Perna *et al.* 2001). Such observations can however be rather species-specific, as indicated, for example, by the observation by Perna *et al.* 2001 that the amount of horizontal transfer into *E.coli* genomes appears to be much higher than that into *Helicobacter* or *Chlamydia* genomes. They therefore cannot be safely assumed to be representative for a large set of genomes.

Estimation of various aspects of gene content evolution such as the size of ancestral genomes and the amount of gene duplication are of course not independent. We therefore seek a general integrated approach to reconstruct explicitly which genes were present in the ancestral genomes and how the gene content of ancestral and present day genomes has been shaped by the processes of gene loss, gene duplication, HGT, gene fusion/fission,

and gene genesis. By gene genesis we mean the de novo origin of a gene. We define it as occurring in the lineage leading to the most recent common ancestor of the species in which the orthologous genes are present. For reasons regarding certainty of the phylogeny, doubts on the existence of a single last common ancestor (Doolittle 2000), and unreliable automated orthology determination at large evolutionary distances, we focus on two taxa for which multiple genomes are available at informative intermediate evolutionary distances: the Archaea and the Proteobacteria. Our reconstruction of the evolution of gene content is based on the presence and absence of genes in these two taxa and in the other complete genomes. The latter are used as outgroup to assess whether a gene potentially originated outside the taxon. The processes that shape gene content can also be studied by detailed sequence-based phylogenies. Such approaches do not scale up well among others because long branch attraction tends to draw fast-evolving sequences like the mycoplasmas (Teichmann and Mitchison 1999) or *Buchnera* (see below) towards the root of the tree. To correct for those effects and to create reliable sequence alignments, gene trees often require manual input. We therefore chose this complementary large-scale approach based on presence and absence of genes alone. The notion of a single common ancestor for a group of genomes might be a simplification; alternatives in the form of a community have been proposed (Woese 1998; Doolittle 2000). In such a scenario, our estimates for the gene content of early genomes represent rather that of a community of genomes.

Results

The processes that shape gene content

Horizontal Gene Transfer Versus Parallel Gene Loss

The central question is whether to explain patchy, nonphylogenetic gene distributions by multiple gene loss or by HGT (Fig. 4.1). We answer this by reconstructing the same gene distribution by the most parsimonious scenario without HGT (the non-HGT scenario, Fig. 4.1A), and with HGT (the HGT scenario, Fig. 4.1B). By comparing the two scenarios, we obtain the number of gene losses that become necessary when we explain the same distribution without HGT instead of with it. If this number of losses is lower than a variable "HGT penalty" we explain the distribution of these genes by including HGT; otherwise we explain it using only losses. By varying this HGT penalty we can differentiate between gene distributions that are to different degrees nonphylogenetic and that are thus more or less likely to be caused by horizontal transfer (Fig. 4.1B). In the final step, the presence pattern in the ancestral nodes from the most parsimonious scenario at each HGT penalty is used to determine the remaining processes: gene duplication (the number of genes within an orthologous group increases), gene fusion/fission (two orthologous groups fuse into one open reading frame (ORF), or the reverse one orthologous group splits into two ORFs), and gene genesis (a group of orthologous genes appears for the first time). Note that a patchy gene distribution does not necessarily imply HGT. Numerous cases can be retrieved in which such a distribution of genes is best explained by multiple gene losses based on independent evidence (see Fig. 4.2 for an example).

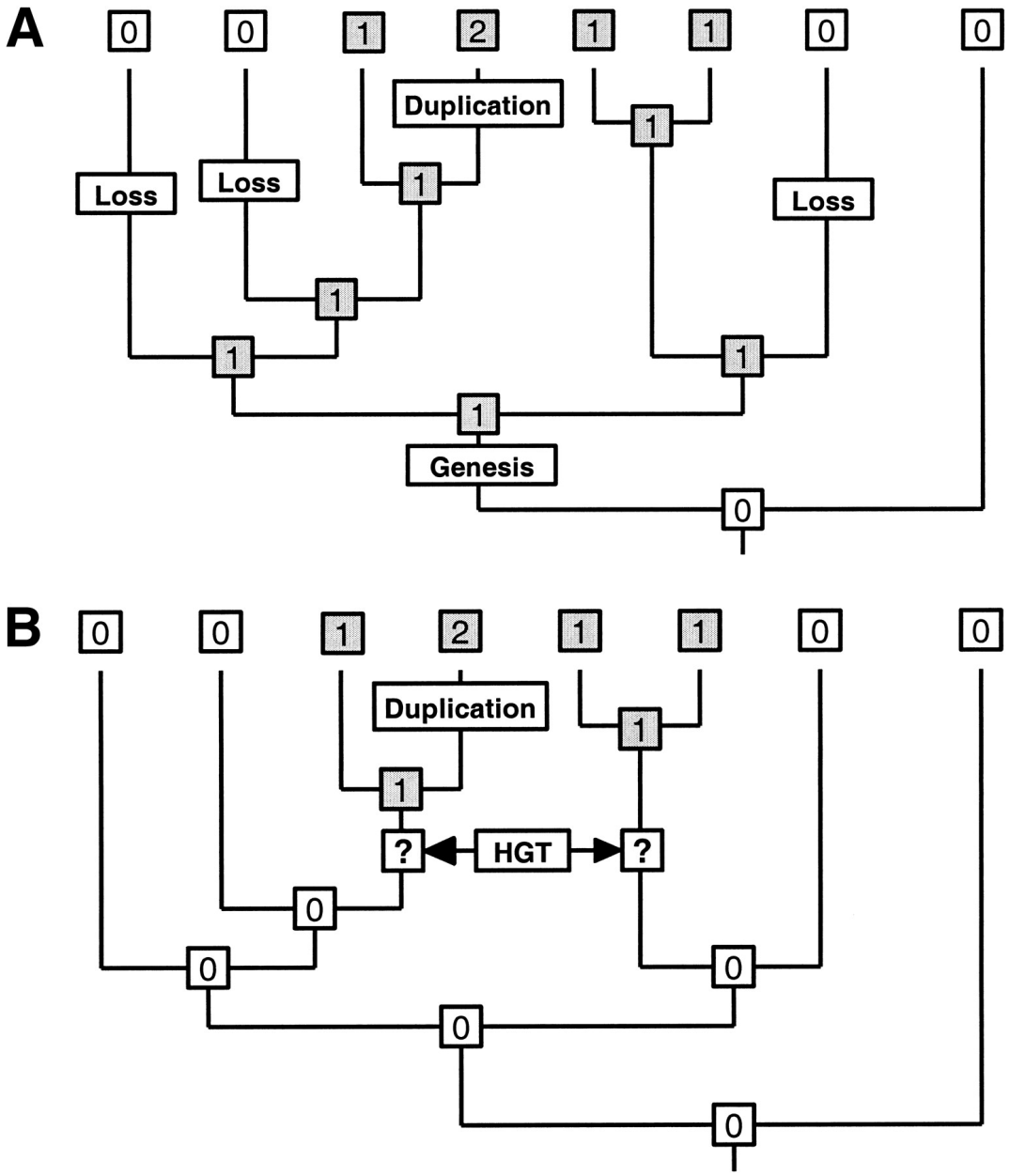


Figure 4.1. Schematic representation of the procedure used to explain presence patterns in terms of gene genesis, gene loss, gene duplication, and HGT. Panels A and B show the same species topology with the same present day presence pattern of a group of orthologs. The gray boxes with a "1" or "2" indicate that a gene from the group of orthologs is present one or two times, while the white boxes with the "0" indicate that the group is absent from that node. Panel A depicts, based on this distribution, what we infer about the presence of genes in the ancestral nodes assuming only vertical inheritance and using the minimum number of events necessary. It also shows where we determine gene genesis, gene duplication, and gene loss to have occurred based on this ancestral distribution pattern. Panel B shows how the same pattern can be explained by one duplication (the same as in A), one genesis, and one HGT. The boxes with question marks indicate that along one branch an HGT and along the other a gene genesis occurred, but we are unable to say which occurred where. Thus a question mark denotes either a gene genesis or the acceptance of a horizontally transferred gene. At an HGT penalty lower than 3, we explain the distribution of this orthologous group in terms of horizontal transfer, and at an HGT penalty higher or equal to 3 we explain the same distribution in terms of multiple losses.

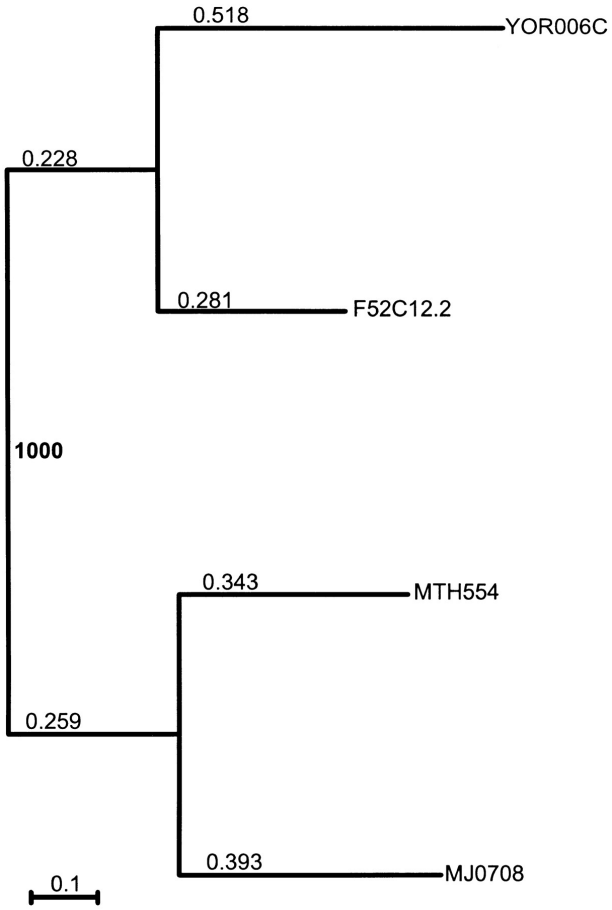


Figure 4.2. **Phylogeny of MTH554 and its orthologs**. The orthologous group is specific for the Archaea and the eukarya. Although the proteins are annotated as hypothetical, we find that it is homologous to a predicted rnase P component, and that it is conserved in an operon with rpl40 in three different species. It therefore probably has a function in translation/transcription. Despite its patchy species distribution, being in only the methanobacteria and the eukaryotes, the tree suggests simple vertical inheritance followed by gene loss in *A. fulgidus*, *A. pernix*, and the ancestor of the Pyrococci, rather than horizontal gene transfer. We propose these three losses because the gene phylogeny is consistent with the species phylogeny, and there is a long internal branch length separating the two groups, which is consistent with presence in the common ancestor of eukarya and Archaea. Moreover, any HGT explanation would contain unlikely events. When it would have taken place from a primitive eukaryote to an ancestor of methanobacteria, the receiving branch would be the very short branch separating the methanobacteria from the other Archaea. When alternatively it would have transferred from an ancestor of the methanobacteria to a primitive eukaryote, the donating branch would be the aforementioned (too) short branch.

Above a certain HGT penalty, horizontal transfer becomes absent from the results (Table 4.1). However, it also results in quite large ancestral genomes (Fig. 4.3), and extrapolation would suggest the last common ancestor of all species to have been a huge omnipotent organism (Doolittle 2000). We obtain a more realistic picture by allowing some HGT by decreasing the HGT penalty, because this allows genes from one organism to stem from "multiple" smaller ancestral genomes. Conversely, when HGT is considered as likely as gene loss (an HGT penalty of 1), ancestral genomes become unrealistically small, and extrapolation would suggest that a last common ancestor contained only a handful of genes. A reasonable window of truth can be obtained by discarding the most extreme scenarios (Fig. 4.4).

Table 4.1. Total number of events in the tree for different HGT penalties

Archaea HGT penalty	**Gene loss**	**Gene duplication**	**Genesis**	**Horizontal gene transfer**	**Vertically inherited genes**	**Gene fusion**
1	1894	1164	3120	1153	13285	221
2	2805	1164	3134	599	14486	221
3	3798	1164	3138	257	15501	221
4	4826	1164	3138	0	16529	221
Proteobacteria HGT penalty	**Gene loss**	**Gene duplication**	**Genesis**	**Horizontal gene transfer**	**Vertically inherited genes**	**Gene fusion**
1	9815	3684	5337	3181	24160	747
2	11201	3684	5337	2483	25546	747
3	11717	3677	5341	2289	26054	747
4	13976	3666	5341	1655	27689	747
5	18761	3663	5535	499	30576	747
6	18773	3663	5536	495	30586	747
7	18780	3663	5536	493	30591	747
8	22636	3663	5536	0	32541	747

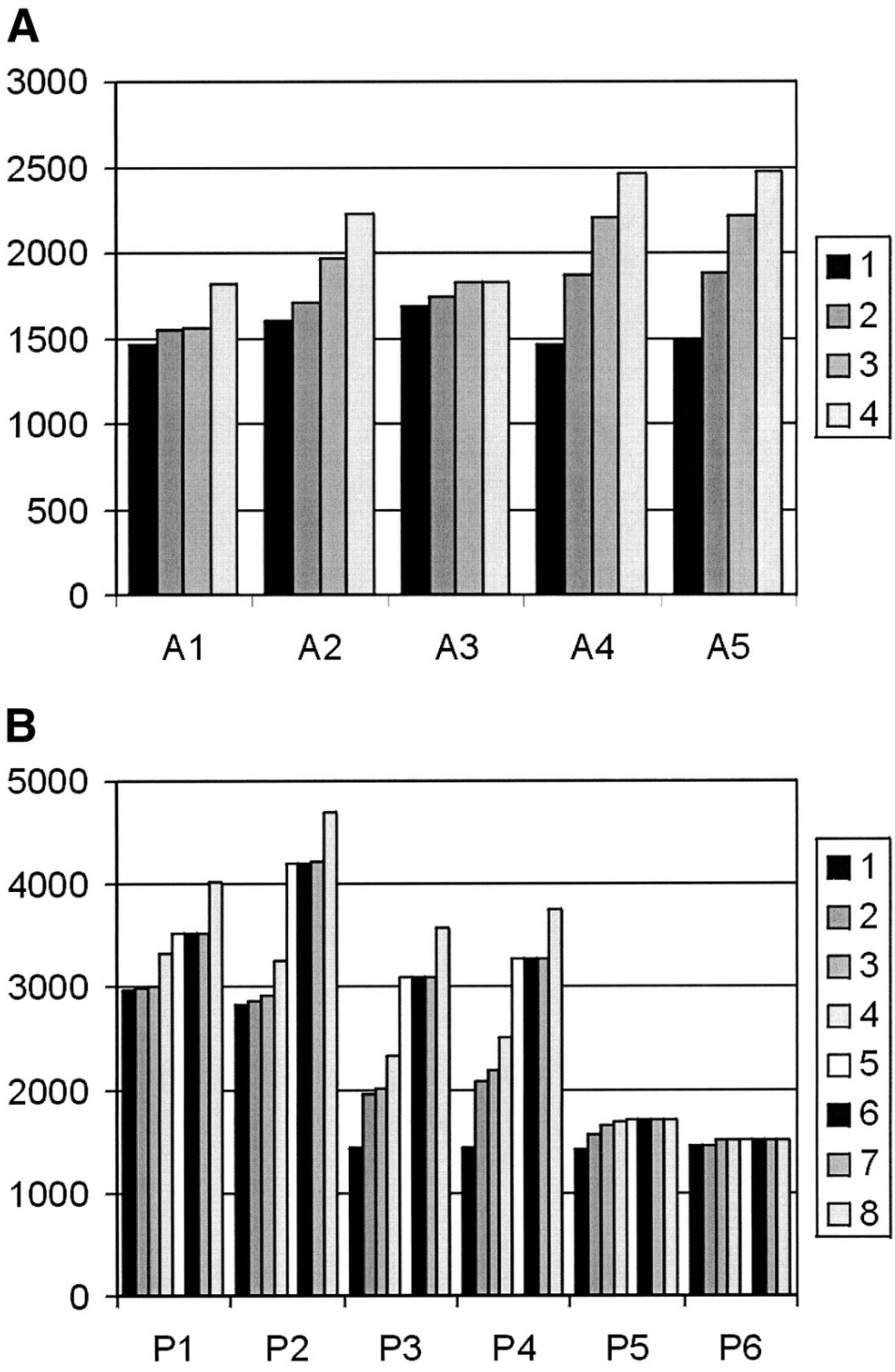
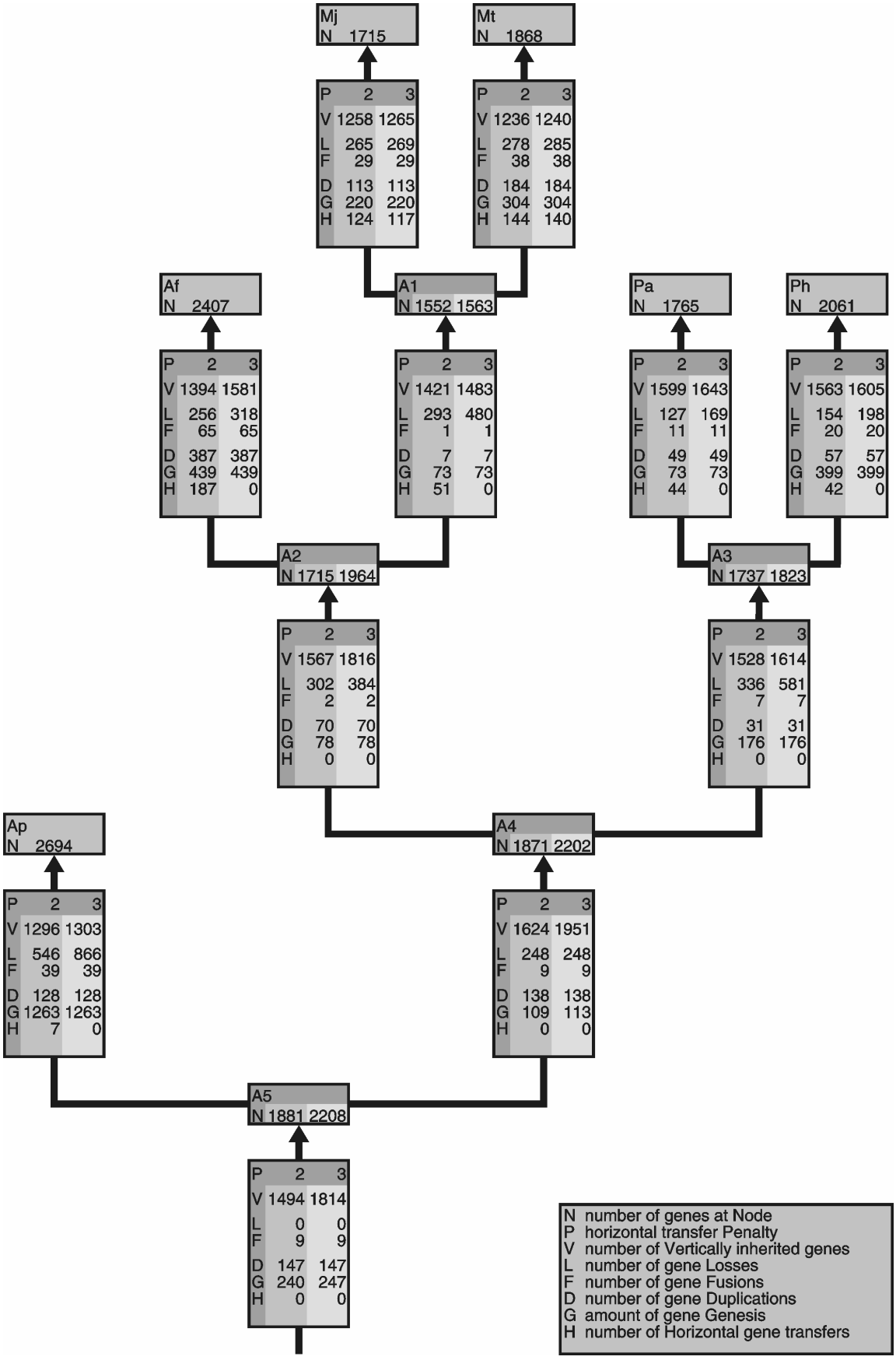


Figure 4.3 Histogram of the estimates for ancestral genome sizes for increasing HGT penalties. To see where in the tree the different ancestral nodes are present, see Figure 4.4. The different HGT penalties are given in the legend. The results for (A) Archaea and (B) Proteobacteria are shown.


Mj
N 1715
P 2 3
V 1258 1265
L 265 269
F 29 29
D 113 113
G 220 220
H 124 117
Mt
N 1868
P 2 3
V 1236 1240
L 278 285
F 38 38
D 184 184
G 304 304
H 144 140
Af
N 2407
P 2 3
V 1394 1581
L 256 318
F 65 65
D 387 387
G 439 439
H 187 0
A1
N 1552 1563
P 2 3
V 1421 1483
L 293 480
F 1 1
D 7 7
G 73 73
H 51 0
Pa
N 1765
P 2 3
V 1599 1643
L 127 169
F 11 11
D 49 49
G 73 73
H 44 0
Ph
N 2061
P 2 3
V 1563 1605
L 154 198
F 20 20
D 57 57
G 399 399
H 42 0
A2
N 1715 1964
P 2 3
V 1567 1816
L 302 384
F 2 2
D 70 70
G 78 78
H 0 0
A3
N 1737 1823
P 2 3
V 1528 1614
L 336 581
F 7 7
D 31 31
G 176 176
H 0 0
Ap
N 2694
P 2 3
V 1296 1303
L 546 866
F 39 39
D 128 128
G 1263 1263
H 7 0
A4
N 1871 2202
P 2 3
V 1624 1951
L 248 248
F 9 9
D 138 138
G 109 113
H 0 0
A5
N 1881 2208
P 2 3
V 1494 1814
L 0 0
F 9 9
D 147 147
G 240 247
H 0 0
N number of genes at Node
P horizontal transfer Penalty
V number of Vertically inherited genes
L number of gene Losses
F number of gene Fusions
D number of gene Duplications
G amount of gene Genesis
H number of Horizontal gene transfers

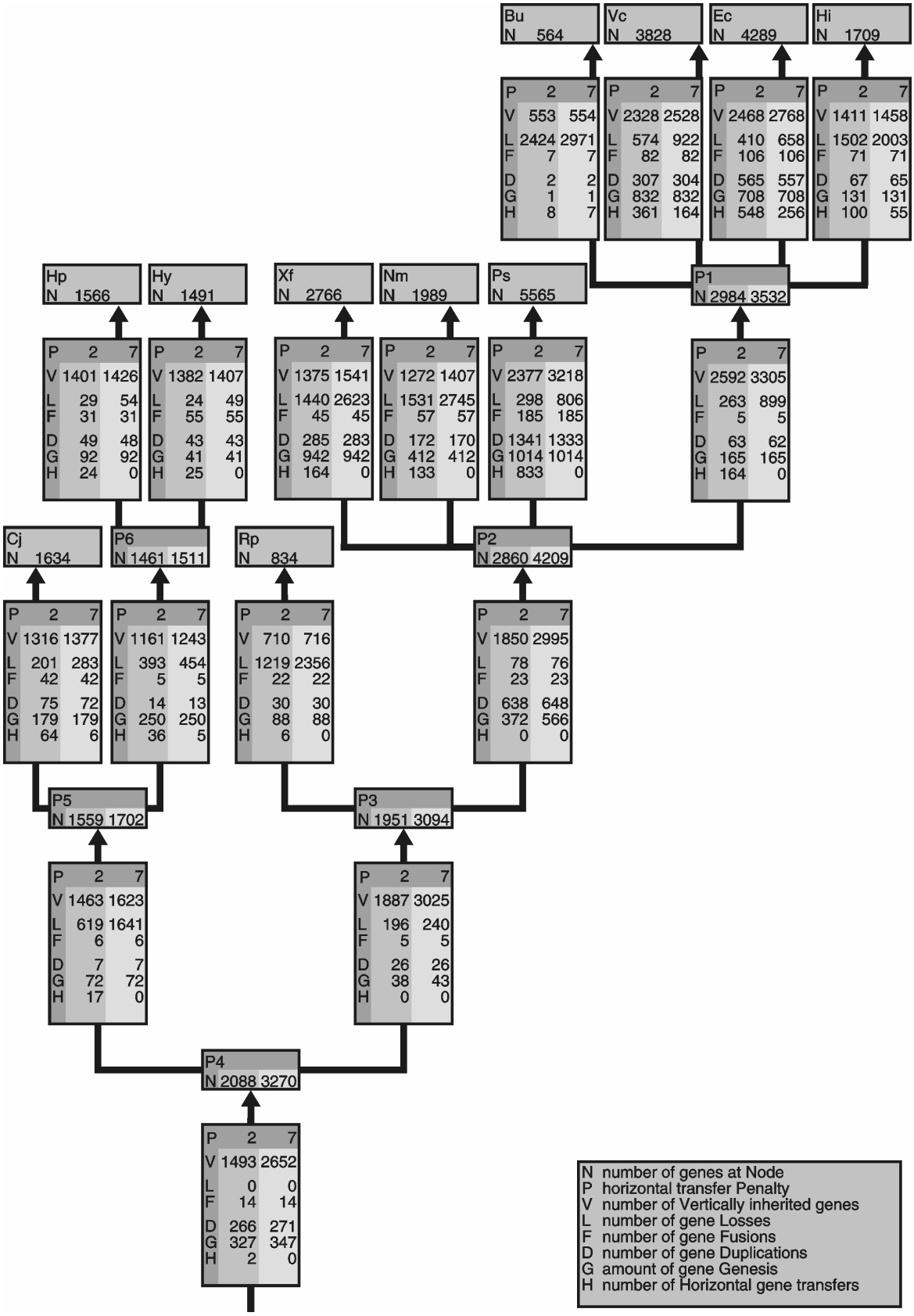


Figure 4.4. An integral reconstruction of genome evolution. The panels show a tree topology that reflects the assumed phylogeny of the species we analyzed, with our results for the evolution of gene content mapped onto them. We give the results for the reconstruction under two different transfer regimes. The branch lengths do not reflect evolutionary time. Two-letter abbreviations denote the species initials at

the leaves. The ancestral nodes have names that consist of one character to denote their taxon (`A' for Archaea and `P' for Proteobacteria) and a number to distinguish them from each other. At each node is indicated how many genes we propose to have been present in that ancestor under two different HGT penalties. On the branches, all the processes are enumerated by their character code followed by how often that event occurred under two different scenarios. The meaning of the character codes is shown in the insets. (A) The results for the Archaea. The two-letter codes for the Archaeal species are as follows: Af, *Archaeoglobus fulgidus*; Ap, *Aeropyrum pernix*; Mj, *Methanococcus jannaschii*; Mt, *Methanobacterium thermoautotrophicum*; Pa, *Pyrococcus abbysii*; and Ph *Pyrococcus horikoshi*. The first number for the processes and the ancestral genome sizes is at an HGT penalty of 2, and the second at an HGT penalty of 3. (B) shows the results for the *Proteobacteria*. Bu, *Buchnera* sp. APS; Cj, *Campylobacter jejuni* NCTC 11168; Ec, *Escherichia coli*; Hi, *Haemophilus influenzae*; Hp, *Helicobacter pylori* 26695; Hy, *Helicobacter pylori* J99; Nm, *Neisseria meningitidis* MC58; Ps, *Pseudomonas aeruginosa* PA01; Rp, *Rickettsia prowazekii*; Vc, *Vibrio cholerae*; and Xf, *Xylella fastidiosa*. The first number for the processes and the ancestral genome sizes is at an HGT penalty of 2, and the second at an HGT penalty of 7.

Duplication, fusion, and vertical inheritance

The occurrence of gene duplication and gene fusion is almost completely independent from the amount of horizontal gene transfer (Table 4.1). Note that our estimates for the number of recent duplications, which are the duplications along terminal branches in Figure 4.4, are similar, albeit slightly lower, to those found by Jordan *et al.* (2001). Since the estimate of the amount of losses is directly coupled to the amount of HGT, the total number of losses decreases with increasing frequencies of HGT (Table 4.1). This effect is most prominent in the primitive branches (Fig. 4.4). Most genes on a given branch are present in its starting node, and in the node to which it leads; that is, they are vertically inherited (Table 4.1, Fig. 4.4). On all but the most early branches, the number of vertically inherited genes is relatively independent of the HGT penalty (Fig. 4.4). There are less vertically inherited genes with more HGT, because this number depends strongly on how many genes are available in the node they start from as well as how many of these are lost, and both of these factors decrease with increasing levels of HGT. Paradoxically, the fraction of vertically inherited genes (the number of vertically inherited genes divided by the number of genes in the node from which a branch stems) increases with increasing levels of HGT. This is because the number of lost genes decreases faster than the number of genes in the ancestral node with increasing levels of HGT. Thus with more HGT the vertical component becomes more important in genome evolution.

Gene duplication versus HGT

We here compare the effects of gene loss and HGT using a penalty for transfer. However we cannot do that for duplication versus HGT, because one transfer origin of a gene in an organism with multiple copies of that gene is equivalent to one duplication event; that is, one duplication can be replaced by one HGT to obtain the same present day distribution. We therefore compiled a test set of orthologous groups, namely those groups for which only one of the species contains multiple copies of a gene. This is a suitable test set because otherwise we would need to explicitly reconstruct many different processes simultaneously. Phylogenetic analysis of these groups reveals that 65% of the duplicated genes clearly fall into one cluster within the trees. The origin of the rest of the genes is unclear. These can be explained by transfer, but as easily by problems in phylogenetic inference as well as nonparsimonious older duplication and independent loss scenarios (see page 24). Using relative sequence similarities to distinguish these cases, we find that an upper limit of 20% of the genes might actually be of xenelogous origin. A reclassification of 20% of the duplications as HGT would, except for the smallest HGT penalty scenario, not affect the relative order of importance of the various processes (Table 1).

Gene genesis
The total number of gene genesis events is almost independent of the HGT penalty (Table 1). Large genomes as well as genomes whose closest relatives are relatively distant have the most genesis events (Fig. 4.4). In addition there are branches leading to certain extant species that have a suspiciously high number of genesis events, most notably *Aeropyrum pernix* (Fig. 4.4A). The evaluation of a number of parameters (Table 4.2) suggests that *A. pernix*, *Pyrococcus horikoshi*, *Vibrio cholerae*, and *Xylella fastidiosa* contain ORFs that might mistakenly be annotated as genes, as has been noted before for some of these species (Cambillau and Claverie 2000; Huynen and Snel 2000). The number of genesis events in the branches leading to these species is thus probably an overestimate. The estimates for gene genesis also reveal that there are at least 240 genes that originated at the branch leading to the Archaea (Fig. 4.4A). For the Proteobacteria we estimate this number to be at least 320 (Fig. 4.4B). Such genes can be considered characteristic of a taxon, as they are unique to it and widespread within it. As implemented in the model, horizontal gene transfer is more abundant when the HGT penalty is lower, but the amount of HGT never dominates (Table 4.1). Notice that in estimating HGT we do not identify the recipient and the donor explicitly. Rather both branches are considered potential recipients. Thus the amount of HGT is a maximum estimate.

Table 4.2. Suspicious ORFs

Species	Genome size (bp)	No. of ORFs	No of gene genesis	No. ORFs without homolog
A. fulgidus	2178400	2407	349	275
A. pernix	1669695	2697	1212	1052
M. jannaschii	1664970	1715	186	143
M. thermoautothrophicum	1751377	1868	250	211
P. abyssi	1765118	1765	56	29
P. horikoshii	1738505	2064	368	310
C. jejuni	1641481	1634	162	125
E. coli	4639221	4289	673	497
H. influenzae	1830138	1709	108	95
H. pylori	1667867	1566	73	65
H. pylori J99	1643831	1490	21	18
N. meningitidis A Z2491	2272351	1989	189	167
R. prowazekii	1111523	834	84	71
V. cholerae	4033464	3828	823	674
X. fastidiosa	2679306	2766	907	760

Ancestral Genome Size
Our estimates of the ancestral genome sizes depend on the HGT penalty, albeit to a different extent for the different taxa (Fig. 4.3). Not unexpectedly, the general trend is that the number of genes in the older ancestral genomes decreases the more we interpret the

patchy presence patterns as horizontal gene transfer. In the following, we will use A(1-5) for denoting the ancestral Archaeal nodes and P(1-6) for denoting the ancestral Proteobacterial nodes (Figs. 4.3,4.4). The estimates for some ancestral genomes show almost no variation (e.g., the nodes A1, A3, P5, P6, and P1 in Fig. 4.3), which suggests that these are reasonable estimates for their number of genes. Other genomes show intermediate (e.g., A2, A4, and A5) through large (e.g., P2, P3, and P4) variation. In both clades, the primitive nodes are the most uncertain, in the Proteobacteria more so than in the Archaea. The reasonable amount of variation allows us to give, for the first time, explicit estimates for the genome size of ancestral genomes. Discarding the extremes we arrive at upper and lower boundaries for the ancestor of all Archaea (A5) between 1881 and 2208 genes, and for the Proteobacterial ancestor (P4) between 2088 and 3270 genes. Under the last common population model (Woese 1998; Doolittle 2000), the lower estimates represent the genes that were present in each organism in the ancestral population, while the higher estimates represent the genes that were present in at least one organism of that population.

Core genes

Under the model we use to interpret the presence patterns, the number of genes that are present in all nodes is independent from assumptions about horizontal gene transfer. For the Proteobacteria, that set consists of 252 genes, and for the Archaea of 480 genes. We find less genes in this Archaeal "stable core" than did Makarova *et al.* (1999). We therefore repeated our procedure with the same species they used, and we obtained 539 genes, closely approximating their number of 542. The difference between our stable core (480 genes) based on the species used here, and the core (540 genes) based on a limited set of genomes, is largely due to the addition of the crenarchaeum *A. pernix*. Obviously such a "core" group of genes is not independent of the number of genomes used to define it. The core, defined as those genes that are present in all organisms, has an opposite in the gene pool, the genes which are present in any of the organisms. Counting all orthologous groups, excluding single genes that do not have homologs (potentially dubious singletons), we estimate that the gene pool contains 6411 genes for the Proteobacteria, and 3496 genes for the Archaea.

Genome dynamics

The turnover of genes

The independence of certain processes and of the size of certain nodes to the amount of HGT allows a reconstruction of the dynamics of genome evolution (Figs. 4.4,4.5). Lineage-specific differences can be relatively safely inferred for branches that are invariant to the HGT penalty. For example, it can be concluded that *Haemophilus influenzae* has lost at least 1500 genes since its common ancestor with *V. cholerae* and *E. coli* (P1), while *E. coli* and *V. cholerae* each lost at least 400-500 genes (Fig. 4.4B). This means that gene loss is a major factor in explaining the difference in genome size between these organisms, as has been previously suggested for *H. influenzae* and *E. coli* (de Rosa and Labedan 1998). Figure 4.5B, which traces the history of a single genome in terms of how many ancestral genes from each ancestor survive, reveals that there was even a substantial increase in genome size leading to P1, followed by a substantial decrease leading to *H. influenzae*. Because *H. influenzae* and *E. coli* have the same genome history up to P1, Figure 5B also reveals that substantial loss occurred throughout

the history of *E. coli*. In total, the lineage leading to *E. coli* from P4, the common ancestor of the Proteobacteria, lost between 950 and 1500 genes (Figs. 4.4B, 4.5B). Furthermore, it also is not the case that the two large genomes, *E. coli* and *V. cholerae* simply inherited their size. Rather they independently underwent substantial amounts of gene genesis and gene duplications (Fig. 5.4B). Thus, the general trend is that there is gain and loss on each branch, including loss of genes that previously have been gained; that is, a turnover of the gene content (Fig. 5.5).

From numbers to rates

Further insights into genome evolution can be gained by evaluating the relationship of the number of events with the evolutionary time of the branches. As a measure of evolutionary time for the branches, we use the consensus from the consistent protein phylogenies of the core genes (see Methods). We can use this analysis to assess our results, because although we assumed a certain topology for our inferences, we did not assume specific branch lengths: that is, we did not require the processes to correlate with time. We normalize the events using fractions of genes for each process that we obtain by dividing the number of events on a branch by the genome size from which it stems. Although the fraction of lost genes on a branch correlates fairly well with the length of that branch (Table 4.3), there are lineage-specific differences such as the high number of losses in the branch leading to *H. influenzae*. HGT does not show significant correlation with time. Duplication and gene genesis only correlate with time in the Archaea. Whereas the relatively low correlation of gene genesis might be partly caused by wrongly annotated genes in certain species (Table 4.2), the amount of duplication has a low correlation with time (Table 4.3), and it shows a large variation among branches (Fig. 4.4). Specifically, large genomes and the branches leading to P4 and A4 contain relatively many duplications, suggesting an important role for duplication in genome size expansion and early genome evolution (Fig. 4.4). In general, the estimates of the correlation indicate to what extent a process is clock-like; that is, to what extent it has a constant rate in time. This in turn might reflect the type of selection a process is under. Processes that show a weak correlation with time could be under (strong) positive selection. On the other hand, the relatively clock-like behavior of gene loss likely reflects negative selection (Gillespie 1998).

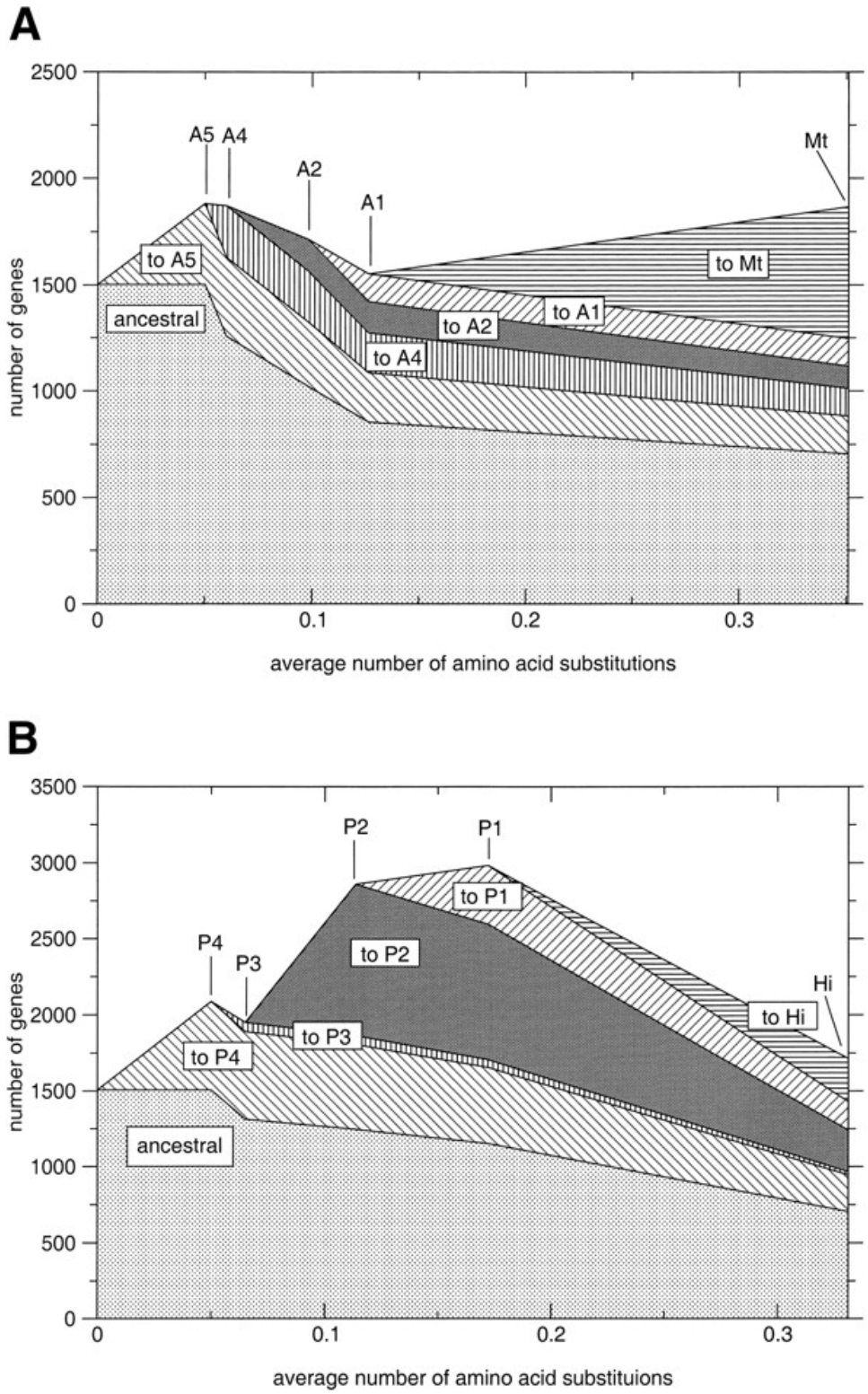


Figure 4.5. A genome history. The plot traces the lineage leading to a present day genome through time; that is, the events on the successive branches are plotted sequentially over the evolutionary time of the branches. Between all nodes the number of genes that is gained (i.e., gene genesis + gene duplication + horizontal gene transfer) leading to a node is plotted, and this set is marked. For each set stemming from a certain node/branch, the number of genes left in the succeeding nodes is traced, thereby denoting which genes are lost. The evolutionary time between the "root" and the common ancestor of the Archaea or Proteobacteria is unknown; we therefore used a fixed arbitrary distance for that branch lengths. (A) shows the lineage leading to *M. thermoautotrophicum* at an HGT penalty of 2. (B) shows the lineage leading to *H. influenzae* at an HGT penalty of 2.

Table 4.3 Correlation coefficient *r* of the fraction of events with evolutionary time

Archaea HGT penalty	Loss	Duplication	Genesis	HGT
1	0.25	0.57[a]	0.63[a,b]	0.05[a]
2	0.65[a,b]	0.57[a]	0.64[a,b]	0
3	0.57[a,b]	0.59[a]	0.65[a,b]	0
4	0.80[a,b]	0.57[a]	0.62[a,b]	0
Proteobacteria HGT penalty	**Loss**	**Duplication**	**Genesis**	**HGT**
1	0.74[a,b]	0.01	0.39	0.08
2	0.74[a,b]	0.06	0.38	0.21
3	0.75[a,b]	0.06	0.38	0.22
4	0.76[a,b]	0.05	0.38	0.22
5	0.77[a,b]	0.05	0.31	0.1
6	0.77[a,b]	0.05	0.31	0.1
7	0.77[a,b]	0.05	0.31	0.1
8	0.78[a,b]	0.05	0.32	0

[a] Significant at $P < 0.05$ when compared to protein evolution.

[b] Significant at $P < 0.05$ when compared to rRNA evolution.

Discussion

Relative importance of various processes

The complete set of results allows us to describe some general features of genome evolution. The branches and nodes early in the tree show the most variation. However, the estimates, excluding those from the extreme scenarios, do not differ too much, and are thus reasonable indications (Fig. 4). In all scenarios there is the same order of quantitative importance for the processes: gene loss, gene genesis, gene duplication, and, lastly, horizontal gene transfer. Although there is a significant number of HGT events, its contribution relative to the other processes is small. This result is logical to the extent that transferred genes behave phylogenetically normal before and after the transfer: they undergo gene loss or gene duplication, and along all branches except for the transfer branch, they are vertically inherited; that is, even if no single gene family would be without HGT, this would not necessarily imply its quantitative dominance. The quantitative dominance of the other processes was already suggested by the phylogenetic pattern in shared gene content (Gaasterland and Ragan 1998; Chapter 3.2; Tekaia *et al.* 1999).

The quantitatively important processes occur on all branches. For example, gene loss also operates along a branch where genome size increases. These processes thereby make

genome evolution very fluid, with a turnover of the gene content throughout the tree. Nevertheless on almost all branches these dynamic processes contribute less than the genes that are simply inherited from the ancestor. Surprisingly, the relative contribution of these vertically inherited genes increases with increased amounts of HGT, because the size of the ancestors decreases less drastically than gene loss with increased HGT. The estimates for evolutionary recent ancestral genome sizes are relatively invariant. For the early genomes we estimate the ancestor of the Archaea to have had between 1881 and 2208 genes, and for the ancestor of the Proteobacteria to have had between 2088 and 3270 genes.

Genome clock

Evaluation of the fraction of events over time per branch reveals their different modes of (genome) evolution: loss correlates fairly well with time, gene genesis correlates less well, and horizontal transfer as well as duplication hardly correlate at all. The clock-like behavior of gene loss suggests that it is under negative selection, while the processes resulting in the addition of a gene have a more adaptive character (Gillespie 1998). An explanation might be that there is a constant pressure to lose genes by gene deletion mutations, whereas the appearance of new genes only occurs as an adaptation to a new lifestyle. Note that we do not imply that gene loss is without functional interpretation as in the co-elimination of functionally interacting sets of proteins (Aravind *et al.* 2000), but only that it is under a different type of selection.

Nonparsimonious events

We reconstruct the evolution of genome content by explaining the present day species distribution of genes using the minimum number of events. However, evolution also proceeds nonparsimoniously. For example, we do not detect the transfer of genes to organisms where they replace an existing orthologous copy, that is, orthologous gene displacement (Huynen *et al.* 1999). Such displacement would lead us to miss one HGT event and one gene loss event. However, on average, only 24% of the trees in our core set of genes are inconsistent with the consensus species phylogeny, inconsistencies that to a large extent are due to unequal rates of evolution. Similarly, it would be impossible to detect a gene that originated early in evolution but that was subsequently lost from all following genomes. Thus, because of our parsimony methods our estimates are probably minimum estimates, except for gene genesis, which is probably a maximum estimate (see also Table 2).

Outlook

We provide here, based on an integrated approach and given explicit assumptions, estimates for the processes governing genome evolution and for the ancestral genomes. By including more species the result should converge, although it will probably be necessary to correct the HGT penalty for the numbers of species that are included in the analysis. Approaches such as the one presented here are required to move from distance-based genome phylogenies (Chapter 3.2) to genome trees that explicitly take ancestral nodes and the events connecting them into account. This avenue seems especially promising given the quantitative importance of processes that retain the phylogenetic signal such as vertical inheritance or gene genesis under all scenarios. In addition, approaches like this should improve the use of co-occurrence of genes for the prediction

of functional association (Huynen and Bork 1998; Pellegrini *et al.* 1999), because the information that genes were gained and lost together can be explicitly included.

Methods

Groups of orthologous genes

We constructed groups of orthologous genes starting from our set of pairwise orthologous genes (Huynen and Bork 1998), which are based on an all-against-all comparison of the complete set of proteins from each genome using smith-waterman searches (Smith and Waterman 1981, see http://www.tigr.org/tdb/mdb/mdbcomplete.html for an overview of currently available genomes). The Archaeal and Proteobacterial genomes we analyzed here are given in the caption of Figure 4. The other (outgroup) genomes we used are *Aquifex aeolicus*, *Bacillus subtilis*, *Borellia burgdorferi*, *Caenorhabditis elegans*, *Chlamydia pneumoniae*, *Chlamydia pneumoniae* AR39; *Chlamydia trachomatis* D/UW-3/CX, *Deinococcus radiodurans*, *Mycobacterium tuberculosis* Rv, *Mycoplasma genitalium* G37, *Mycoplasma pneumoniae* M129; *Saccharomyces cerevisiae*, *Synechocystis* PCC6803, *Thermotoga maritima*, *Treponema pallidum*, and *Ureaplasma urealyticum*. We mark the genes that have nonoverlapping orthologous hits with different genes as fused (Chapter 2). In order to find genes that have been duplicated since the first speciation event in the taxon, we first determine, for every gene, with which of its orthologs it has the lowest similarity to obtain a threshold. Subsequently we determine for each gene the homologs in its genome that are more similar than this threshold, and denominate these as "duplicates within the genome." Then we start from a seed gene, which is not allowed to be fused, and keep adding orthologs as well as duplicates, and, if they are not fused, use them as seeds, until no new genes are added. All genes hereby retrieved are considered an orthologous group of genes. This approach is conceptually similar to the COGs (Tatusov *et al.* 1997), GeneRAGE (Enright and Ouzounis 2000) or GEANFAMMER (Park and Teichmann 1998), where the latter two approaches, however, focus on homologs instead of groups of orthologous genes. Conceptually our approach assembles genes that have a single representative in the last common ancestor of the compared species into one orthologous group. Note that pairwise orthology, unlike homology, in principle is nontransitive; that is, when A is orthologous to B and B is orthologous to C, then A is not necessarily orthologous to C in the case of duplication events after the speciation event separating A, B, and C (Tatusov *et al.* 1996, 1997, Chapter 3.2). *Sensu stricto* our groups thus contains also paralogous relations. The group orthology concept as described here and as also implemented in COGs (Tatusov *et al.* 1997) is therefore the only approach that allows a quantification of the processes in which we are interested.

Phylogeny and divergence time

The phylogeny that we use here is based on the construction of 23S rRNA trees, the construction of gene order trees (Blanchette *et al.* 1999), and the construction of genome trees (Chapter 3.2). The tree partitions that consist of the same species in the trees from all three methods are implemented in the consensus phylogeny that we used for our analysis. To obtain evolutionary time for the branches, we used the orthologous groups that are present in all species. We constructed multiple sequence alignments using

CLUSTAL W (Thompson *et al.* 1994) and neighbor joining trees (Saitou and Nei 1987) based on these alignments with default parameters as implemented in CLUSTAL W (Thompson *et al.* 1994). Subsequently we took the trees that are consistent with the consensus phylogeny of these species, averaged their branch lengths, and used this as the measure of the evolutionary time for a branch. Although the individual phylogenies that we selected are decidedly not clock-like, the procedure gave a surprisingly clock-like average phylogeny for the species considered, to the extent that the distance of all end nodes to the root is very similar (available from http://www.bork.embl-heidelberg.de/~snel/flux/). The rRNA-based branch lengths that we used as an additional measure for computing the correlation of the different processes with evolutionary time was obtained from 23S RNA. After constructing an alignment of the 23S RNA sequences from the species analyzed in this study, we constructed a phylogeny that corresponds to the consensus using TREE-PUZZLE (Strimmer and von Haeseler 1996) and parsed the branch lengths from the tree for use in computing the correlation.

We found that 85% and 66% of the phylogenies of the core Archaeal and Proteobacterial genes, respectively, are consistent with the species phylogeny that we inferred. These inconsistencies could be the result of, among others, orthologous gene displacement or of gene duplication followed by differential loss. However, the higher fraction of inconsistent Proteobacterial trees relative to the Archaea is probably the result of another complication in constructing reliable phylogenies: unequal rates of sequence evolution, because more than half of the Proteobacterial inconsistent trees are classified as such due to *Buchnera* falling out of its grouping with *E. coli*, *H. influenzae*, and *V. cholerae*. Small genomes typically have higher rates of sequence evolution, which in combination with long branch attraction moves them towards the root of the tree.

Only vertical inheritance (the non-HGT-scenario)

Using perl scripts, we first determined the most parsimonious scenario without horizontal gene transfer: that is, we determined, given the presence/absence pattern of an orthologous group of genes, given the phylogeny of the species, and assuming only vertical inheritance, which ancestors of the genomes contained this gene (see Fig. 4.1A). The branch where a gene appears for the first time is the branch were the gene started (gene genesis). Because of our operational definition of gene genesis, we cannot explicitly determine whether they truly (1) represent genuine de novo gene origins (i.e., from noncoding DNA), (2) resulted from a gene duplication followed by such rapid sequence divergence that the original orthology/paralogy situation became unclear, or (3) resulted from an HGT followed by very rapid sequence divergence. Therefore, for the genes that resulted from a genesis, we performed an additional search to find homologs that are not members of their orthologous group, using PSI-BLAST (Altschul *et al.* 1997) to increase the sensitivity. These searches revealed that only 12% of the Archaeal and 14% of the Proteobacterial orthologous groups resulting from genesis have homologs that are not a member of the group. The exact origin of the remaining genes remains undetectable, and these percentages are thus a lower limit for the amount of an origin other than genuine de novo gene origins.

A branch where the number of members from an orthologous group increases is considered to have undergone gene duplication. A branch where the number of members from an orthologous group decreases is considered to have undergone gene loss.

Horizontal gene transfer

To include horizontal gene transfer we introduced a relative-to-gene-loss variable penalty for a transfer event. Transfer events are treated as independent gene genesis events, where each additional genesis event costs the penalty of horizontal gene transfer (see Fig. 4.1B). If then there is a scenario with independent genesis events that cost less than a scenario with only loss, that scenario is used to score events for the group. We take penalties for horizontal transfer of 1, 2, 3, 4, 5, 6, 7, and 8. Although mechanism and selection are tightly intertwined, we thus do not allow HGT to be easier than gene loss, because purely mechanistically, before selection, gene loss is "easier" than HGT (Brown 1999). One can interpret the HGT penalty as an "expected relative frequency" of HGT versus gene loss per group of orthologous genes.

Gene fusion and fission

When a gene is present in two or more groups of orthologous genes, it is thought to be a fusion gene. A specific fusion, that is, a group of orthologous genes consisting of the same set of domains (i.e., orthologous gene groups), is assumed to have occurred only once. Hence the fused genes are treated like their own group of orthologous genes. The groups of orthologous genes that gave rise to this fusion are then treated like a normal group, except that at the branch where the fusion occurred they both lose a member to the fused group. The result of this approach is that before the fusion event the components of the fused genes are treated like two or more separate genes, whereas after the fusion they are counted as one gene. We do not make a distinction between gene fusion and gene fission. In general, gene fusion is much more frequent than gene fission. A detailed gene tree-based analysis revealed that 85% of the nonoverlapping homology cases is caused by fusion rather than fission (Chapter 2). Our fusion category therefore contains a small fraction of fissions.

Acknowledgements

We thank J. Castresana, P. Hogeweg, and the members of the Bork group for discussion and comments.