Gene 421 (2008) 27–31

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

High rates of lateral gene transfer are not due to false diagnosis of gene absence

Weilong Hao, G. Brian Golding *

Department of Biology, McMaster University, Hamilton, Ontario, Canada L8S 4K1

ARTICLE INFO

Article history:
Received 30 October 2007
Received in revised form 20 May 2008
Accepted 3 June 2008
Available online 14 June 2008

Keywords:
Lateral gene transfer
Truncated gene
Gene gain
Genome evolution

ABSTRACT

Methods for assessing gene presence and absence have been widely used to study bacterial genome evolution. A recent report by Zhaxybayeva et al. [Zhaxybayeva, O., Nesbo, C. L., and Doolittle, W. F., 2007. Systematic overestimation of gene gain through false diagnosis of gene absence. Genome. Biol. 8, 402] suggests that false diagnosis of gene absence or the presence of undetected truncated genes leads to a systematic overestimation of gene gain. Here (1) we argue that these annotation errors can cause more complicated effects and are not necessarily systematic, (2) we argue that current annotations (supplemented with BLAST searches) are the best way to consistently score gene presence/absence and (3) that genome wide estimates of gene gain/loss are not strongly affected by small differences in gene annotations but that the number of related gene families is strongly affected. We have estimated the rates of gene insertions/deletions using a variety of cutoff thresholds and match lengths as a way in which to alter the recognition of genes and gene fragments. The results reveal that different cutoffs for match length only cause a small variation of the estimated insertion/deletion rates. The rates of gene insertions/deletions on recent branches remain relatively high regardless of the thresholds for match length. Lastly (4), the dynamic process of gene truncation needs to be further considered in genome comparison studies. The data presented suggest that gene truncation tends to take place preferentially in recently transferred genes, which supports a fast turnover of recent laterally transferred genes. The presence of truncated genes or false diagnosis of gene absence therefore does not significantly affect the estimation of gene insertions/deletions rates, but there are several other factors that bias the results toward an under-estimation of the rate of gene insertion/deletion. All of these factors need to be considered.

Gene insertions and deletions have been widely acknowledged to play an essential role in shaping bacterial genomes during evolution (Garcia-Vallve' et al., 2000; Gogarten et al., 2002; Fraser-Liggett, 2005; Gogarten and Townsend, 2005). A general consequence of gene insertions and deletions is a patchy distribution of genes. For instance, genes that are present in some genomes might be absent from other closely related genomes (Welch et al., 2002; Tettelin et al., 2005). Various methods have been developed to understand the process of gene insertions and deletions (Daubin et al., 2003; Kunin and Ouzounis, 2003; Mirkin et al., 2003; Boussau et al., 2004; Hao and Golding, 2004; Hao and Golding, 2006; Dagan and Martin, 2007). Previous studies have shown that most genes, if not all, are subject to the possibility of gene transfer (Dagan and Martin, 2007). By examining gene presence/absence in closely related species, the rates of gene insertion/deletion can be inferred (Hao and Golding, 2006). Recently Zhaxybayeva et al. (2007) reported that genomes with truncated homologues might erroneously lead to false inferences of "gene gain" rather than multiple instances of "gene loss". This raises the question of how a false diagnosis of gene absence affects the estimation of insertion/deletion rates, if at all.

To examine how gene diagnosis will affect inferred rates of gene insertion/deletion we have recalculated maximum likelihood gene indel rate estimates for the *Bacillaceae* genomes under various conditions. In Hao and Golding (2006) annotated genes were used to indicate presence/absence. TBLASTN (Altschul et al., 1997) was used to check that a gene annotated in one genome was indeed missing in another and was considered an annotation error if the sequence matched over a given length criteria. Finally, since we were concerned about gene transfer rather than intraspecific gene family duplication, BLAST was used with a match length criteria to delineate gene family members. The end result considers only presence/absence of the family and used a method that assumes gene insertions and gene deletions occur at an equal rate, as must be true if the genome sizes are evolutionarily stable given the compact nature of bacterial genomes.

It has long been recognized that the presence of a gene in one taxon but not in others could be due to either gene gain or to gene loss. Parsimony methods have been used to explore the effects of different weightings between these two events (Kunin and Ouzounis, 2003; Mirkin et al., 2003). Different choices are influenced by different ancestral patterns (Fig. 1). Knowledge of the presence of truncated gene fragments also influences this choice. Trying to create an estimate of the rate of gene gain/loss adds another layer of complexity. A method that is based on a maximum likelihood model can be used to avoid arbitrary weightings of loss versus gain. A model with equal

Abbreviations: BLAST, Basic Local Alignment Search Tool; LGT, Lateral Gene Transfer; MLE, Maximum Likelihood Estimation; ORFans, Orphan Genes/ORFs.

* Corresponding author. Tel.: +1 905 525 9140; fax: +1 905 522 6066.
E-mail address: Golding@McMaster.CA (G.B. Golding).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.06.015

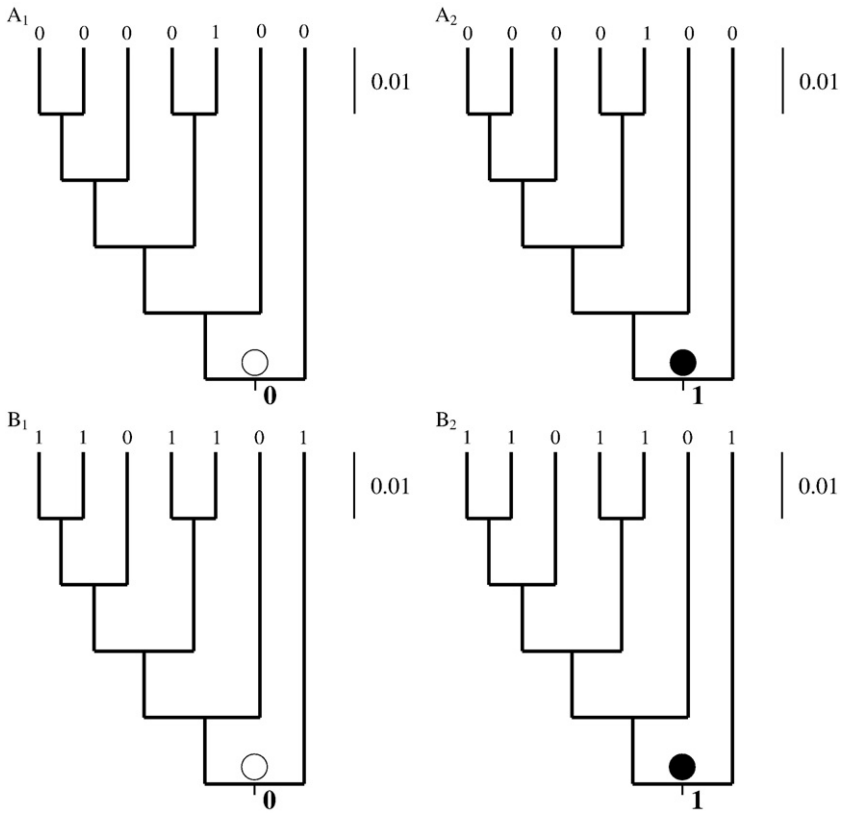

Fig. 1. Inferences for the gene presence/absence scenario from Zhaxybayeva et al., 2007 (see also Figure S.1). Truncated genes are inferred as either absent (A) or present (B). In each case, the ancestral pattern can be absent (A_1, B_1) or present (A_2, B_2).

rates of gain/loss prefers a situation where the number of gain and loss events are more nearly equal and will therefore give some proportion of the likelihood to intermediate scenarios and the exact proportion will depend on the number and pattern of events at each locus within the genome.

Zhaxybayeva et al. (2007) compare a situation where one taxon out of seven has a gene present to a situation where one taxon has a full length sequence, four taxa have truncated copies and two taxa have the gene genuinely absent (reproduced in Figure S.1). They note that in the first situation, that a single gene gain would be inferred but if the second situation was seen to be correct, the inference would be multiple gene loss rather than gene gain. They therefore suggest that failure to recognize truncated genes leads to a systematic over-estimation of gene gain.

There are many uncertainties that enter into models of genome evolution. There are general and pervasive problems with genome annotations, in our specific case unique genes have been ignored, gene duplications have been ignored, orthologous replacement via LGT has been ignored, and there are complications imposed by necessarily simplified models that make inferences of gene indel rates. If fragments of a gene are in fact present and do constitute a reasonable length of the gene, annotators are most likely to include the truncated fragments as a true gene. For this reason (and possibly others) there is a large number of short proteins in bacterial genomes. This bias will cause more genes to be listed as "present" when in fact they should be listed as pseudogenes. Given all these problems, the inference that a systematic bias to favour higher rates of gene gain exists is not immediately clear.

The likelihood values for a particular locus can be calculated assuming specific ancestral states. The situation of Zhaxybayeva et al. (2007), ignoring gene fragments, is shown in Fig. 1A_1. If gene fragments are considered, their scenario would add four gene fragments in the taxa as shown in Fig. 1B_2. A major difference in the ancestral state is inferred between these two cases; one suggesting the ancestral state was absent while the second suggests the ancestral state was present.

In lieu of a model that directly incorporates fragment information, we have considered the two extremes; only one full length gene present, versus gene fragments that are considered fully present. The rate estimates and likelihood estimates for these scenarios are shown in Fig. 2. A model considering only one full length gene and an ancestral state absent, has the highest likelihood.

However, it also has the lowest rate estimate (the corresponding likelihood with an ancestral state present (A_2) asymptotically increases with no maximum rate). Estimates with each gene fragment considered as fully present, also yield estimates of higher rates (and lower likelihoods). The reason why a lower rate estimate is made contrary to the rational reasoning of Zhaxybayeva et al. (2007) is that the model being used requires equal rates of gene insertion and gene deletion. Adding in gene fragment information has lowered the number of gene insertions that should be inferred but increased the number of gene deletions. Hence, the nature and the limitations of genome evolution models also influence the rate estimates. But minimally, the presence of truncated genes need not always lead to an overestimation of gene indel rate.

To determine how these factors interact and how different considerations of gene fragments might affect estimates of genome wide rates of gene gain/loss we have re-examined seven of the most closely related *Bacillus* genomes from Hao and Golding (2006). Four different expect-values (10^{-5}, 10^{-10}, 10^{-15}, or 10^{-20}) were used to identify homologues via BLASTP for annotated protein sequences and TBLASTN to uncover non-annotated genes (Altschul et al., 1997). For each expect-value, various thresholds for match length (25%, 50%, 70%, or 85%) were used. For many proteins a match length of just 25% will identify even some random similarities as part of the match in the absence of any homology. If a predicted ORF is present in only one

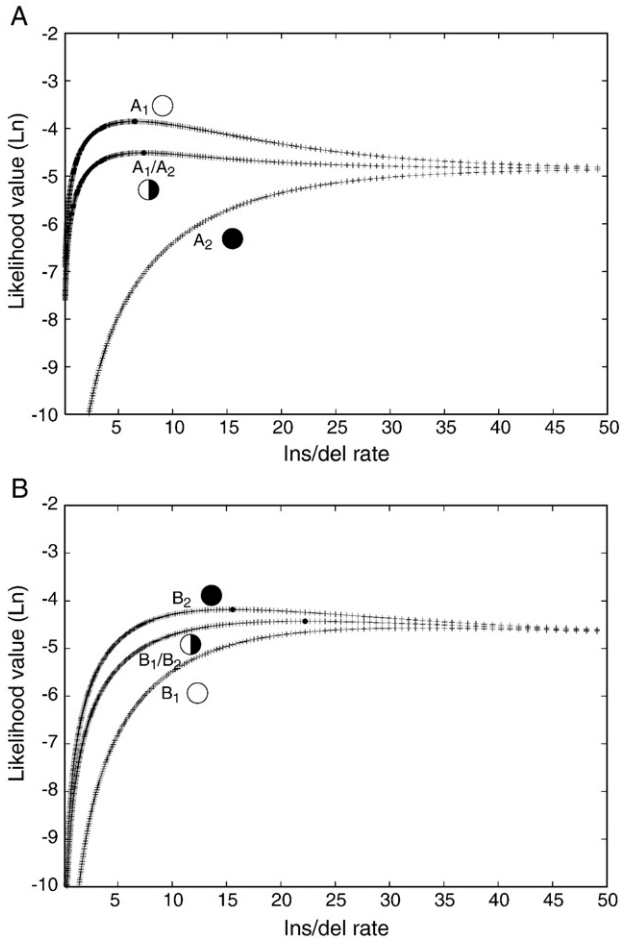

Fig. 2. Likelihood values for different indel rates. Different ancestral patterns were assumed; gene was present, absent, or uncertain (each with a probability of 50%). A, truncated genes were assumed absent (as shown in Fig. 1A); B, truncated genes were assumed present (as shown in Fig. 1B). The small black dots indicate the maximum likelihood values.

genome and does not have any known homologues within the NCBI nr database, it was removed from the analysis (such genes are known as ORFans; Fischer and Eisenberg, 1999).

For a given gene, it is expected that more "homologous" sequences might be detected when more relaxed criteria are used. Hence, with relaxed criteria, homologues clustered together as gene families might become broader. To quantify the changes that might occur in phyletic pattern, annotated genes from the *Bacillus anthracis* (Ames strain; abbreviated here Ba_1) genome were re-examined relative to genes in the other sequenced genomes of two *B. anthracis* strains, three *B. cereus* strains, and a *B. thuringiensis* strain (abbreviated collectively here as Bc, see Supplementary for all abbreviations).

It is note worthy that the most striking change with relaxed criteria is in the number of total gene families rather than in gene presence/absence (observe the large changes in overall totals in Table 1 particularly as the match length criteria is decreased; reading horizontally across the totals). Relaxed criteria for match length result in more "homologues" detected within the same genome and hence inferred as gene duplicates. As a consequence, the number of gene families declines when more genes are identified as homologues and are clustered into the same gene family. The distribution of pattern categories, however, remains remarkably similar (Table 1). Hence, many gene families still have a patchy distribution, even when the criteria for match length are relaxed. How some of the gene families moved between patterns is illustrated in Table S.1.

A maximum likelihood method in Hao and Golding (2006) was employed to estimate insertion/deletion rates based on the phyletic pattern of all gene families. It is clear that different cutoff thresholds for match length do have an effect on the estimated insertion/deletion rates (Table 2) but this effect is numerically small. Even though the effect of inference on gene presence/absence is complex, the indel rate estimations are still comparable. The rate on recent branches (rate α_1) is still consistently higher than the rate on remaining branches (rate α_2). Assuming that rates have not increased recently, this can be explained by the rapid loss of recent gene insertions. It also suggests that the observed higher insertion/deletion rates on recent branches are not an artifact of false diagnosis of gene absence. Moreover, ORFans were not included in this analysis and the addition of ORFans would further inflate the rate at the tips of phylogeny (Table S.2).

The designation of gene presence/truncation/absence relies heavily on genome annotation. Gene truncation or gene absence can be detected only when the gene is annotated in at least one other genome in the analysis. The presence of non-annotated genes can also affect the rate estimation. For instance, significant matches to the *Salmonella* ORF16-*lacZ* fusion protein (protein ID: AAX64154; expect-value $<10^{-20}$; match length >85%) have been detected in all *Bacillus* genomes, but no gene was annotated in any *Bacillus* genomes. The effect of inference on gene presence/absence, therefore, is clearly not systematic but instead rather complex.

To gain a greater understanding of the dynamics of truncated genes, annotated protein sequences in Ba_1 genome were used as query sequences to search against the Bc_3 complete genome sequence via TBLASTN (Altschul et al., 1997). Significant matches were required to have an expect-value less than 10^{-5} with no restriction on match length but sequences had to be syntenic between Ba_1 and Bc_3 (Figure S.2) to be considered homologous. Protein sequences that have paralogues within the Ba_1 genome or have more than one significant hit in the Bc_3 genome were removed from further analysis. These sequences were then subdivided into two categories, those that did

Table 1

E-value	Pattern	Match length			
		85%	70%	50%	25%
10^{-20}	1100000	3	2	2	2
	1110000	252	221	192	179
	1111000	68	56	53	52
	1111100	149	129	114	100
	1111110	247	198	162	142
	1111111	3631	3411	3215	2971
	Others	1525	1318	1168	1020
	Total	5875	5335	4906	4466
10^{-15}	1100000	4	3	3	3
	1110000	257	222	189	179
	1111000	72	60	52	50
	1111100	159	132	120	107
	1111110	241	191	146	124
	1111111	3657	3425	3184	2898
	Others	1566	1331	1173	995
	Total	5956	5364	4867	4356
10^{-10}	1100000	1	0	0	0
	1110000	237	197	162	150
	1111000	77	67	58	54
	1111100	156	122	111	94
	1111110	266	210	163	145
	1111111	3678	3405	3098	2705
	Others	1604	1344	1180	988
	Total	6019	5345	4772	4136
10^{-05}	1100000	1	0	0	0
	1110000	223	180	132	117
	1111000	73	55	40	30
	1111100	144	118	93	76
	1111110	245	181	122	105
	1111111	3607	3269	2833	2318
	Others	1612	1317	1139	903
	Total	5905	5120	4359	3549

Table 2
Maximum likelihood estimation (MLE) inferred on external branches α_1 and internal branches α_2 in the Bc group using various cutoff thresholds

E-value	Rate	Match length			
		85%	70%	50%	25%
10^{-20}	$\alpha_1=\alpha_2$	4.52	4.10	3.72	3.54
10^{-15}		4.52	4.10	3.72	3.54
10^{-10}		4.52	4.10	3.71	3.72
10^{-5}		4.52	4.10	3.72	3.72
10^{-20}	α_1	7.55	7.19	6.52	5.92
	α_2	1.22	1.11	1.22	1.22
10^{-15}	α_1	7.93	6.85	6.52	5.92
	α_2	1.22	1.22	1.11	1.11
10^{-10}	α_1	7.55	6.85	6.52	6.21
	α_2	1.34	1.22	1.16	1.22
10^{-5}	α_1	7.55	6.85	6.52	6.52
	α_2	1.34	1.22	1.11	1.11

The branch leading to Ba_1, Ba_2, and Ba_3 was treated as an external branch as done in Hao and Golding, 2006.

have similar sequences in the other *Bacillaceae* genomes (Gk, Bl, Bs, Bk, Bh, Oi) and those that do not have significant match (with an expect-value less than 10^{-5} in a TBLASTN search) in any other *Bacillaceae* genomes. Thus one set is identified as Bc (group) specific and the other set as non-Bc specific. The length of the match in the TBLASTN search is shown in Fig. 3 as a proportion of the query sequence. Strikingly, the Bc specific genes have a much higher proportion of truncated genes than do non-Bc specific genes. These results are robust regardless of the expect-values used in homologue identification (*E*-values 10^{-5} to 10^{-20}).

The results were further evaluated by a BLASTN search in the Bc_3 genome using annotated protein coding nucleotide sequences in Ba_1 as query sequences, since frameshift mutations that disrupt the translational reading frame could be detected as truncated genes by TBLASTN. Here, we employed a BLASTN search with low stringency by modifying match/mismatch scoring parameters (Figure S.6) to eliminate the effect of possible frameshift mutations. Truncated genes detected in a BLASTN search are fewer than those detected in a TBLASTN search, suggesting that some genes might have undergone frameshift. The BLASTN results, however, are still consistent with the TBLASTN results that the Bc specific genes have high proportion of truncated genes. Please note that the BLASTN search is using low stringent parameters instead of default parameters, and the identification of truncated genes should be considered conservative. In fact, the low stringency of BLASTN parameters is known to result in elongation of match length (Korf et al., 2003). Together, the Bc specific genes have high proportion of truncated genes/pseudogenes. Since the Bc specific genes are more likely to include a greater proportion of recently transferred genes, this suggests that such recently transferred genes are more likely to be truncated or frameshifted. This is also consistent with the findings that many recently transferred genes are transient (Hao and Golding, 2004; Hao and Golding, 2006).

Gene truncation is useful for detecting pseudogenes. However even here, the criteria that have been used to classify pseudogenes are far from consistent (Chain et al., 2004; Lerat and Ochman, 2004) and in any case, some shortened homologues might still carry out some function (Ogata et al., 2001). Detection of pseudogenes is difficult and requires the knowledge of each gene's transcription and its protein's function. The boundary between gene and pseudogene might, indeed, be ambiguous (Zheng and Gerstein, 2007). Even when a gene is present and annotated within a genome, it is usually not clear that this gene is actually functional. In this study, Ka/Ks ratios (Fig. 4) were calculated for intact genes and truncated genes inferred by TBLASTN using the PAML package (Yang, 2007). It is striking that truncated genes tend to have higher Ka/Ks ratios than intact and nearly intact genes, suggesting that truncated genes are under relaxed functional constraints and likely many of them are pseudogenes. Most of truncated genes have Ka/Ks ratios less than 1, this suggests that pseudogenization in many genes occurred after the two strains diverged. On the other hand, the possibility that some genes might still have function could not be completely ruled out. Evidence for the functionality of each gene within a bacterial genome is generally beyond most current studies. Given these limitations, it is probably best to continue to use the annotations provided by and updated by the experts for each bacterial genome.

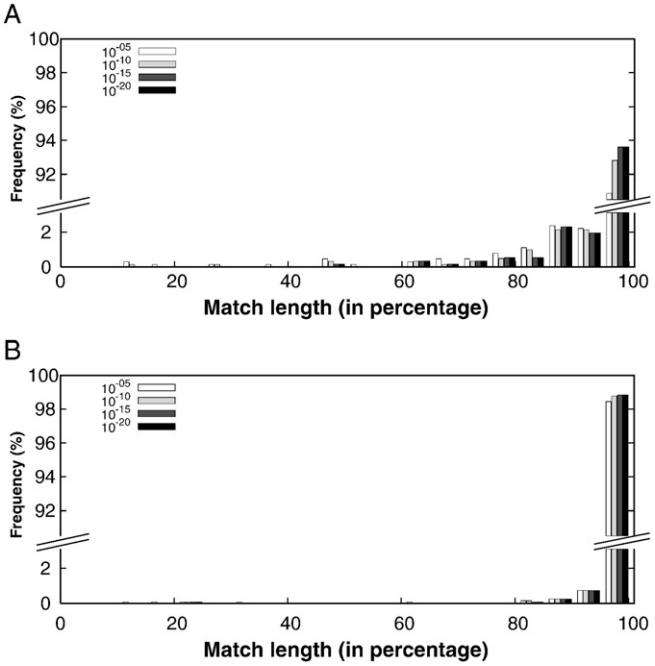

Fig. 3. Fraction of syntenic sequences with different lengths (expressed as a percentage of the maximum total length) in a TBLASTN search. Only syntenic sequences between Ba_1 and Bc_3 are considered. Results using different *E*-values are shown. A: genes specific to the Bc group (more likely recently transferred); B: genes that are not Bc group specific genes (less likely recently transferred).

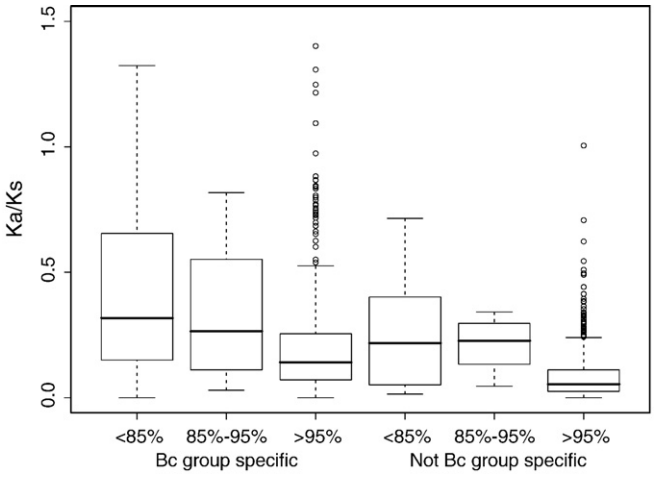

Fig. 4. Boxplot of Ka/Ks values for different groups of truncated genes. Ka/Ks values are higher in truncated genes (<85% in length and 85%–95% in length) than those in intact or nearly intact genes (>95% in length). In Bc group specific genes, $P=0.0004$ ("<85%" versus ">95%") and $P=0.0017$ ("85%–95%" versus ">95%") in a Wilcoxon signed-rank test. In not Bc group specific genes, $P=0.064$ ("<85%" versus ">95%") and $P=5.0\times10^{-5}$ ("85%–95%" versus ">95%") in a Wilcoxon signed-rank test. Truncated genes were identified by a TBLASTN search using cutoff threshold *E*-value $<10^{-5}$. The results based on a BLASTN search using the same threshold reveal an essentially same pattern (data not shown).

Current studies have not yet incorporated gene fragment knowledge into rate analyses. More complicated models will be required to incorporate this knowledge but from the arguments presented here, other factors of possibly greater importance also need to be added to genome evolution models. With the additional consideration of gene fragments, the rate of gene insertion/deletion might either increase or decrease. Their addition to models attempting to describe gene evolution will be complicated as there is evidence that gene fragments themselves can be internally duplicated and might themselves be possibly laterally transferred (work in progress). The use of more complicated models is warranted but never-the-less, the estimation of already very high gene indel rates is most likely to increase and not to decrease.

Acknowledgements

This work was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) grant to GBG.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2008.06.015.

References

Altschul, S.F., Madden, T.L., Schffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Boussau, B., Karlberg, E.O., Frank, A.C., Legault, B.A., Andersson, S.G., 2004. Computational inference of scenarios for alpha-proteobacterial genome evolution. Proc. Natl. Acad. Sci. U. S. A. 101, 9722–9727.

Chain, P.S., et al., 2004. Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. Proc. Natl. Acad. Sci. U. S. A. 101, 13826–13831.

Dagan, T., Martin, W., 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proc. Natl. Acad. Sci. U. S. A. 104, 870–875.

Daubin, V., Lerat, E., Perriere, G., 2003. The source of laterally transferred genes in bacterial genomes. Genome. Biol. 4, R57.

Fischer, D., Eisenberg, D., 1999. Finding families for genomic ORFans. Bioinformatics 15, 759–762.

Fraser-Liggett, C.M., 2005. Insights on biology and evolution from microbial genome sequencing. Genome. Res. 15, 1603–1610.

Garcia-Vallve', S., Romeu, A., Palau, J., 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome. Res. 10, 1719–1725.

Gogarten, J.P., Doolittle, W.F., Lawrence, J.G., 2002. Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. 19, 2226–2238.

Gogarten, J.P., Townsend, J.P., 2005. Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Microbiol. 3, 679–687.

Hao, W., Golding, G.B., 2004. Patterns of bacterial gene movement. Mol. Biol. Evol. 21, 1294–1307.

Hao, W., Golding, G.B., 2006. The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome. Res. 16, 636–643.

Korf, I., Yandell, M., Bedell, J., 2003. BLAST. O'Reilly.

Kunin, V., Ouzounis, C.A., 2003. The balance of driving forces during genome evolution in prokaryotes. Genome. Res. 13, 1589–1594.

Lerat, E., Ochman, H., 2004. Ψ–Φ: exploring the outer limits of bacterial pseudogenes. Genome. Res. 14, 2273–2278.

Mirkin, B.G., Fenner, T.I., Galperin, M.Y., Koonin, E.V., 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC. Evol. Biol. 3, 2.

Ogata, H., Audic, S., Renesto-Audiffren, P., Fournier, P.E., Barbe, V., Samson, D., Roux, V., Cossart, P., Weissenbach, J., Claverie, J.M., Raoult, D., 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 293, 2093–2098.

Tettelin, H., et al., 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. U. S. A. 102, 13950–13955.

Welch, R.A., et al., 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. 99, 17020–17024.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.

Zhaxybayeva, O., Nesbo, C.L., Doolittle, W.F., 2007. Systematic overestimation of gene gain through false diagnosis of gene absence. Genome. Biol. 8, 402.

Zheng, D., Gerstein, M.B., 2007. The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends. Genet. 23, 219–224.