Mobile Genetic Elements 4, e27829; January–December 2014; © 2014 Landes Bioscience

Insertion sequence distribution bias in Archaea

Morgan C Florek, Daniel P Gilbert, and Gordon R Plague*

Department of Biology; State University of New York at Potsdam; Potsdam, NY USA

Keywords: DNA transposon, transposable element, IS element, natural selection

Abbreviations: IS, insertion sequence; NGO, neighboring gene orientation

Insertion sequences (IS) are common transposable elements in Archaea. Intergenic IS elements are usually less harmful than intragenic ISs, simply because they are less likely to disrupt host gene function. However, because regulatory sequences are intergenic and upstream of genes, we hypothesized that not all intergenic regions are selectively equivalent for IS insertion. We tested this hypothesis by analyzing the distributions of intergenic IS elements within 155 fully sequenced archaeal genomes. Of the 22 genomes with enough IS elements for statistical analysis, five have significantly fewer ISs between divergently oriented neighboring genes than expected by chance, and seven have significantly more ISs between convergently oriented genes. Furthermore, of the 85 genomes with at least one expected IS within each of the three possible neighboring gene orientations (i.e., divergent, convergent, and tandem), 73 genomes have fewer ISs between divergently oriented genes than expected, and 60 have more ISs between convergently oriented genes than expected (both values deviate significantly from binomial probabilities of random distribution). We suspect that these non-random IS distributions are molded by natural selection resulting from differential disruption of neighboring gene regulation, and that this selective pressure has affected transposable element distributions in prokaryotes for billions of years.

Introduction

Insertion sequences (ISs) are widely distributed DNA transposons in prokaryotes.[1,2] Although IS elements occasionally generate beneficial mutations,[3,4] most are selectively neutral or deleterious for their hosts.[5,6] The fitness of an IS element is intimately tied to that of its host, so ISs that insert into deleterious chromosomal locations will be costly for their hosts and thus will usually be eliminated from most populations. On the other hand, ISs that insert into selectively neutral locations have a greater chance of long-term survival.[6,7]

Intergenic IS elements are usually longer-lived and thus more prevalent than intragenic ISs, simply because they are less likely to disrupt host genes.[8,9] However, not all intergenic regions in bacterial genomes are equivalent for IS occupancy; in many species, IS elements are much *less* common between divergently oriented genes (←→) (i.e., its neighbors are coded on the bottom [←] and top [→] DNA strands, respectively) than expected by chance, and much *more* common between convergently oriented genes (→←) than expected by chance.[10] These non-random distributions are likely due to differential selection imposed by the neighboring genes.[10] Specifically, because regulatory sequences (i.e., promoters, Shine-Dalgarno sequences, and transcription factor binding sites) are intergenic and upstream of genes, an IS inserting between: 1) ←→ oriented neighbors is potentially highly disruptive because the insertion site is upstream of both neighbors, possibly affecting the regulation of both genes, 2) tandemly oriented neighbors (→→ or ←←) is potentially moderately disruptive because the insertion site is upstream of only one neighbor, possibly affecting the regulation of only that gene, and 3) →← oriented neighbors will not disrupt either neighbor because the insertion site is downstream of both genes. Consequently, this differential selection pressure presumably leads to differential IS longevity and thus differential abundance in host genomes.[10]

These non-random intergenic IS distributions are pervasive across the domain of Bacteria.[10] Because Archaea generally have similar chromosomal architectures to Bacteria (e.g., comparable regulatory regions upstream of genes and many polycistronic genes), we hypothesized that they too may exhibit non-random intergenic IS element distributions, with IS elements being most common between →← neighbors and least common between ←→ neighbors.

Results

We tested this hypothesis by analyzing the neighboring gene orientations (NGOs) (i.e., →→/←←, →←, or ←→) for all intergenic IS elements in 155 fully sequenced archaeal genomes, then comparing these observed values to those expected under the null assumptions of random insertion and no natural selection (based on the premise that large and abundant NGO regions should receive more ISs than small and rare ones). Apart from five

*Correspondence to: Gordon Plague; Email: plaguegr@potsdam.edu
Submitted: 10/29/2013; Revised: 01/11/2014; Accepted: 01/13/2014
Citation: Florek MC, Gilbert DP, Plague GR. Insertion sequence distribution bias in Archaea. Mobile Genetic Elements 2014; 4:e27829; http://dx.doi.org/10.4161/mge.27829

genomes with non-standard annotations that precluded analysis, these 155 genomes constitute all completely sequenced and annotated RefSeq[11] archaeal genomes in the GenBank database (ncbi.nlm.nih.gov/bioproject/browse/) as of January 2014.

Of the 22 archaeal genomes with sufficient IS loads for χ^2 analysis,[12] nine (41%) have intergenic IS distributions that deviate significantly ($P \leq 0.05$) from expectations (**Table 1**). Significant deviations occur in all three NGOs, but most prominently between →← and ←→ oriented genes: seven genomes exhibit a significant excess of ISs between →← oriented genes, and five exhibit a significant deficit of ISs between ←→ oriented genes (**Table 1**). These two striking patterns are even prevalent among the 13 genomes that do not deviate significantly from χ^2 expectations (**Table 1**), as well as among the genomes with too few IS elements for χ^2 analysis (**Table S1**). Specifically, of the 85 archaeal genomes with ≥ 1 expected IS element in each NGO, 60 have more ISs between →← oriented genes than expected, and 73 have fewer ISs between ←→ oriented genes than expected. The binomial probabilities of having deviations this skewed just by chance are $P = 9.3 \times 10^{-5}$ and $P = 4.0 \times 10^{-12}$, respectively.

Discussion

Intergenic IS elements are not randomly distributed in archaeal genomes. Namely, fewer IS elements reside between ←→ neighbors and more between →← neighbors than expected by chance. These deviations, which span genomes with relatively large (**Table 1**) and small (**Table S1**) quantities of IS elements, could result from either insertion bias (i.e., ISs preferentially inserting between →← and away from ←→ oriented genes), or from natural selection molding IS distributions after insertion. Although some IS elements do exhibit target sequence specificity,[1] this is an unlikely explanation for the pervasive and consistent biases that we observed across Archaea (**Table 1**; **Table S1**) (see ref. 10). Instead, the most likely explanation is that natural selection molds IS distributions after they insert. In short, transposable elements that insert into locations that decrease host fitness will eventually be purged from most populations, while ISs that insert into relatively innocuous locations will have a greater chance of long-term survival.[7] Therefore, selection against deleterious IS genotypes will lead to a deficit of ISs in harmful locations and an excess in innocuous locations.[10] Since Archaea display a pervasive deficit of ISs between ←→ oriented genes and excess between →← genes (**Table 1**; **Table S1**), these are apparently relatively harmful and innocuous locations for IS elements to reside, respectively. We suspect this is because ISs that insert between ←→ oriented genes have the potential to disrupt regulatory sequences of both genes, causing the most harm, while ISs that insert between →← oriented genes have little chance of disruption and are therefore more likely to persist within populations.

Interestingly, Bacteria exhibit the same biased distributions of intergenic IS elements as Archaea, although the biases are even more pronounced in Bacteria.[10] Specifically, of the genomes with enough IS elements for χ^2 analysis, nearly twice as many Bacteria as Archaea have significantly more ISs between →← oriented genes than expected (59% vs. 32%, respectively), as well as significantly fewer ISs between ←→ oriented genes than expected (40% vs. 23%, respectively) (**Table 2**). Furthermore, of all analyzed genomes that have ≥ 1 expected IS element in each NGO (including those genomes that do not have enough ISs to meet χ^2 test assumptions),[12] 90% of Bacteria vs. 71% of Archaea have more ISs between →← oriented genes than expected, though 86% of both Bacteria and Archaea have fewer ISs between ←→ oriented genes than expected (**Table 2**). Therefore, if natural selection is indeed driving these biased IS distributions as we suggest, then selection must generally be more intense on intergenic IS elements in Bacteria than in Archaea. Because the efficacy of natural selection positively correlates to a population's effective size,[13] one possible explanation for this more intense selection in Bacteria may be that bacterial populations are generally larger than archaeal populations. Unfortunately, we have a poor understanding of the effective sizes of essentially every prokaryotic species,[14] so this is purely speculative.

Another possible explanation for this differential selection intensity may be that intergenic IS elements are generally less deleterious to Archaea than Bacteria. Interestingly, in many archaeal genomes, only internal genes within operons have Shine-Dalgarno sequences; single genes and operon-leading genes often lack Shine-Dalgarno sequences,[15,16] and instead apparently use the start codon as the primary signal for translation initiation.[17] These so-called leaderless genes also exist in some Bacteria, though much less commonly than in Archaea.[18] Therefore, at least as far as translation initiation is concerned, Archaea tend to have fewer regulatory sequences upstream of genes than do Bacteria, so the →→/←← and ←→ NGOs may be safer locations for IS occupancy, and thus more selectively equivalent (though not completely equivalent) to the →← NGO in Archaea.

Whatever the cause of this apparent differential selection between Archaea and Bacteria, these non-random distributions of intergenic IS elements (i.e., over- and under-abundance of ISs between →← and ←→ oriented neighbors, respectively) represent a remarkable genomic trend across all prokaryotes—and thus two of the three domains of life. Transposable elements are nearly universally distributed among all living organisms[19] and have probably existed for billions of years.[20] Therefore, these non-random IS distributions likely reflect an ancient selective pressure that has also persisted for billions of years, possibly even before Bacteria and Archaea diverged.

Materials and Methods

We first downloaded all protein coding sequences (CDS) for every RefSeq genome in GenBank. We then identified all chromosomal IS elements using the blastp program in the ISfinder database (https://www-is.biotoul.fr//),[21] considering all CDSs with a best blastp E value ≤ 10^{-10} to be an IS element.[10,22] Because we were only interested in intergenic IS elements that border two functional native genes, we took a relatively conservative approach when identifying intergenic IS elements (i.e., it is better to exclude some intergenic ISs than to include any intragenic ISs). Therefore, we eliminated all: 1) IS elements that neighbor a gene annotated as disrupted or partial, assuming that the gene was

Table 1. Observed (O) and expected (E) quantities of intergenic IS elements in fully sequenced archaeal chromosomes, and the χ^2 test statistic for each

	Neighboring gene orientation (NGO)[a]						
	→→,←←		→←		←→		
	O	E	O	E	O	E	χ^{2}[b]
Crenarchaeota							
Sulfolobus islandicus HVE10/4	27	26.2	12	7.5	13	18.4	4.3
S. islandicus L.D.8.5	37	36.5	**18**	**7.9**	*15*	*25.6*	17.1***
S. islandicus M.16.4	27	25.9	**17**	**6.9**	*5*	*16.2*	22.8***
S. islandicus REY15A	**37**	**29.3**	12	8.2	*8*	*19.4*	10.5**
S. islandicus Y.G.57.14	54	50.8	**22**	**13.4**	*22*	*33.8*	9.8**
Sulfolobus solfataricus 98/2	34	31.8	12	9.1	17	22.1	2.3
S. solfataricus P2	57	56.7	25	18.1	37	44.2	3.8
Sulfolobus tokodaii str. 7	34	32.0	8	11.4	17	15.6	1.3
Euryarchaeota							
Ferroplasma acidarmanus fer1	35	27.7	5	8.3	14	18.1	4.2
Haloquadratum walsbyi DSM 16790	13	15.2	**13**	**6.9**	4	7.9	7.7*
Methanococcoides burtonii DSM 6242	20	23.8	**12**	**5.9**	12	14.3	7.2*
Methanolobus psychrophilus R15	41	38.1	11	10.1	16	19.9	1.1
Methanosaeta concilii GP6	42	46.6	20	13.0	13	15.4	4.5
Methanosarcina acetivorans str. C2A	71	68.4	25	20.5	24	31.0	2.7
Methanosarcina barkeri str. Fusaro	46	45.8	**19**	**12.0**	13	20.1	6.5*
Methanosarcina mazei str. Goe1	**65**	**54.1**	19	13.4	*10*	*26.5*	14.8***
M. mazei Tuc01	20	20.8	**11**	**5.7**	5	9.5	7.2*
Methanospirillum hungatei JF-1	33	28.5	6	7.3	13	16.2	1.6
Natronomonas moolapensis 8.8.11	13	15.0	8	6.1	9	8.9	0.8
Pyrococcus furiosus COM1	31	26.2	2	5.2	12	13.6	3.0
Thaumarchaeota							
Candidatus Nitrososphaera gargensis	36	32.6	11	10.3	15	19.1	1.3
Unclassified Archaea							
halophilic archaeon DL31	9	11.6	7	5.1	9	8.3	1.4

[a] NGOs in bold contribute a significant excess of observed ISs to significant χ^2 deviations, and those in italics contribute a significant deficit of observed ISs. [b] Asterisks indicate significant *P* values: **P* ≤ 0.05, ***P* ≤ 0.01, ****P* ≤ 0.001.

Table 2. Comparison of intergenic IS element distributions in →← and ←→ neighboring gene orientations (NGOs) in Archaea and Bacteria

		Archaea	**Bacteria**[c]
Genomes with enough IS elements for χ^2 analysis	Percent with a significant IS excess in →← NGO	32% (7/22)[a]	59% (68/116)
	Percent with a significant IS deficit in ←→ NGO	23% (5/22)[a]	40% (46/116)
Genomes with ≥ 1 expected IS element in each NGO	Percent with more ISs in →← NGO than expected	71% (60/85)[b]	90% (222/247)
	Percent with fewer ISs in ←→ NGO than expected	86% (73/85)[b]	86% (212/247)

Raw values are in parentheses. [a] Data from Table 1; [b] Data from Table 1 and Table S1; [c] From reference 10.

fragmented due to IS insertion, 2) ISs that neighbor a pseudogene, conservatively assuming that the gene was non-functional when the IS inserted, and as such the IS never neighbored two functional native genes, 3) ISs bordered by non-consecutively numbered, and therefore presumably non-neighboring genes (e.g., some may be neighbored by non-annotated gene remnants that deteriorated following IS insertion), and 4) ISs that neighbor a phage gene (see **Table S2** for the number of IS elements eliminated for each reason in each genome). Also, we counted each intergenic region that harbors an IS element only once in the analysis, even if multiple IS elements reside in the same intergenic space.

For each genome, we calculated the observed number of intergenic IS elements residing within each of the three NGOs by analyzing the DNA strand on which each neighboring gene is coded (which is included in each GenBank CDS download). We then calculated the expected quantity of intergenic ISs within each NGO for each genome, based on the assumptions that IS insertion is random and that intergenic IS elements are not under selection. Since relatively large and abundant intergenic regions should receive more ISs than small, rare ones, we calculated the expected values individually for each genome using the product of 1) the mean intergenic distance between neighboring native archaeal genes in the three NGOs and 2) the global proportion of each native gene pair NGO. Finally, we assessed whether the observed quantities of intergenic IS elements within each NGO deviate from the expected quantities in each genome using a χ^2 goodness-of-fit test. We used an adjusted residual method to identify individual NGOs that contribute to each significant χ^2 deviation, and considered any adjusted residual with an absolute value > 2 to have made a significant contribution to the overall χ^2 deviation.[23] Because the assumptions of the χ^2 test are that no cell has an expected value < 1.0, and that ≤ 20% of cells have expected values < 5.0,[12] many archaeal genomes contain too few intergenic IS elements for χ^2 analysis. Despite this, we were able to combine all genomes that contain at least a minimal number of IS elements (we arbitrarily chose genomes with ≥ 1 expected IS element in each NGO)[10] in binomial analyses: if intergenic IS elements are indeed randomly distributed, then just by chance, 50% of analyzed genomes should have more ISs between →← oriented neighbors than expected (as calculated above), and 50% should have less (the same is true for ←→ oriented genes).

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Kevin Dougherty and an anonymous reviewer for helpful comments on this manuscript. This work was supported by the National Institutes of Health (7R15GM081862-02 to GRP) and the State University of New York at Potsdam.

References

1. Chandler M, Mahillon J. Insertion sequences revisited. In: Craig NL, Craigie R, Gellert M, Lambowitz A, eds. Mobile DNA II. Washington, D.C.: ASM Press, 2002:305-66
2. Filée J, Siguier P, Chandler M. Insertion sequence diversity in archaea. Microbiol Mol Biol Rev 2007; 71:121-57; PMID:17347521; http://dx.doi.org/10.1128/MMBR.00031-06
3. Héritier C, Poirel L, Nordmann P. Cephalosporinase over-expression resulting from insertion of IS*Aba1* in *Acinetobacter baumannii*. Clin Microbiol Infect 2006; 12:123-30; PMID:16441449; http://dx.doi.org/10.1111/j.1469-0691.2005.01320.x
4. Podglajen I, Breuil J, Rohaut A, Monsempes C, Collatz E. Multiple mobile promoter regions for the rare carbapenem resistance gene of *Bacteroides fragilis*. J Bacteriol 2001; 183:3531-5; PMID:11344163; http://dx.doi.org/10.1128/JB.183.11.3531-3535.2001
5. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. Nature 1980; 284:601-3; PMID:6245369; http://dx.doi.org/10.1038/284601a0
6. Kidwell MG, Lisch DR. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution 2001; 55:1-24; PMID:11263730
7. Lynch M. Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol 2006; 60:327-49; PMID:16824010; http://dx.doi.org/10.1146/annurev.micro.60.080805.142300
8. Campbell A. Eubacterial genomes. In: Craig NL, Craigie R, Gellert M, Lambowitz A, eds. Mobile DNA II. Washington, D.C.: ASM Press, 2002:1024-39
9. Zaghloul L, Tang C, Chin HY, Bek EJ, Lan R, Tanaka MM. The distribution of insertion sequences in the genome of *Shigella flexneri* strain 2457T. FEMS Microbiol Lett 2007; 277:197-204; PMID:18031340; http://dx.doi.org/10.1111/j.1574-6968.2007.00957.x
10. Plague GR. Intergenic transposable elements are not randomly distributed in bacteria. Genome Biol Evol 2010; 2:584-90; PMID:20697140; http://dx.doi.org/10.1093/gbe/evq040
11. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2005; 33:D501-4; PMID:15608248; http://dx.doi.org/10.1093/nar/gki025
12. Cochran WG. Some methods for strengthening the common χ^2 test. Biometrics 1954; 10:417-51; http://dx.doi.org/10.2307/3001616
13. Kimura M. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press, 1983
14. Fraser C, Hanage WP, Spratt BG. Recombination and the nature of bacterial speciation. Science 2007; 315:476-80; PMID:17255503; http://dx.doi.org/10.1126/science.1127573
15. Slupska MM, King AG, Fitz-Gibbon S, Besemer J, Borodovsky M, Miller JH. Leaderless transcripts of the crenarchaeal hyperthermophile *Pyrobaculum aerophilum*. J Mol Biol 2001; 309:347-60; PMID:11371158; http://dx.doi.org/10.1006/jmbi.2001.4669
16. Tolstrup N, Sensen CW, Garrett RA, Clausen IG. Two different and highly organized mechanisms of translation initiation in the archaeon *Sulfolobus solfataricus*. Extremophiles 2000; 4:175-9; PMID:10879562; http://dx.doi.org/10.1007/s007920070032
17. Soppa J. Initiation and regulation of translation in halophilic Archaea. In: Ventosa A, Oren A, Ma Y, eds. Halophiles and Hypersaline Environments. Berlin: Springer, 2011:191-205
18. Zheng X, Hu GQ, She ZS, Zhu H. Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. BMC Genomics 2011; 12:361; PMID:21749696; http://dx.doi.org/10.1186/1471-2164-12-361
19. Aziz RK, Breitbart M, Edwards RA. Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Res 2010; 38:4207-17; PMID:20215432; http://dx.doi.org/10.1093/nar/gkq140
20. Lynch M. The Origins of Genome Architecture. Sunderland, MA: Sinauer Associates, 2007
21. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 2006; 34:D32-6; PMID:16381877; http://dx.doi.org/10.1093/nar/gkj014
22. Touchon M, Rocha EPC. Causes of insertion sequences abundance in prokaryotic genomes. Mol Biol Evol 2007; 24:969-81; PMID:17251179; http://dx.doi.org/10.1093/molbev/msm014
23. Agresti A. An Introduction to Categorical Data Analysis. New York: John Wiley and Sons, 1996