2015

Analyzing Riboswitches as a Function of Genome Size and Genus Ancestry in Gammaproteobacteria

Robyn Reid
Portland State University

Analyzing Riboswitches as a Function of Genome Size and Genus Ancestry in Gammaproteobacteria

by

Robyn Reid

An undergraduate honors thesis submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science

in

University Honors

and

Mathematics and Chemistry

Thesis Adviser

Rahul Raghavan

Portland State University

2015

Analyzing Riboswitches as a Function of Genome Size and Genus Ancestry in Gammaproteobacteria

Robyn Reid
Advisor: Dr. Rahul Raghavan

Abstract:

Riboswitches are RNA molecules that regulate gene expression, without the need for protein factors, at the mRNA level in the bacterial kingdom. This paper focuses on nine riboswitches: TPP, FMN, SAM, Lysine, Cobalamine, Glycine, Molybdenum, Mg, and SAH, and whether or not there exists a relationship between genome size and riboswitch existence. The hypothesis of this paper is that there does exist a direct relationship because it is assumed that the more basepairs (bp) in a genome, the higher the chance that an old characteristic, such as a riboswitch, is conserved. Three hundred twenty members of the gammaproteobacteria class were selected using the Rfam and NCBI databases, and grouped by genome size. Each group was then analyzed via direct counting correlation and one way ANOVA for correlation and covariance. To check ANOVA assumptions, the gammaproteobacteria were grouped according to their respective genus ancestry, and statistics similarly ran. The Method of Most Likelihood was run for both sets via SPSS. The hypothesis of this paper was wrong. A direct, highly correlative relationship between genus ancestry and riboswitch existence was determined; whether or not each riboswitch was present or absent in the gammaproteobacteria analyzed was dependent upon whether the individual bacteria belonged to a genus that had that characteristic. A potential, stricter relationship between species and riboswitch existence was discovered, leaving room for further movement in this research.

Question:

Is there a relationship between genome size and riboswitch existence in gammaproteobacteria? If so, what is the relationship?

Introduction:

Riboswitches are RNA molecules that regulate gene expression, without the need for protein factors, at the mRNA level in the bacterial kingdom; they are usually found in the 5' untranslated region (UTR) of the mRNA. Riboswitches perform two main functions: molecular recognition and conformational switching. Each riboswitch is composed of a two aptamer domain: an aptamer that senses single ligands, as well as an expression platform that controls gene expression via a single mechanism. Riboswitches regulate the transcript in a cis fashion by *directly* binding to metabolites (McDaniel et al. 2003). The two distinct functional domains in riboswitches are highly conserved because there are only four monomer units used by RNA to form selective binding pockets, and therefore the function of the riboswitch is completely dependent on its form, which is in turn wholly dependent on sequence conservation. The aptamer domain, the effector molecule, exists in a compact three dimensional fold to scaffold the ligand binding pocket (Winkler and Breakler, 2003). The second domain, known as the expression platform, contains a secondary structural switch that interacts with the translational or transcriptional mRNA. Regulation of the transcript is achieved in the overlap region of these two domains, known as the switching sequence.

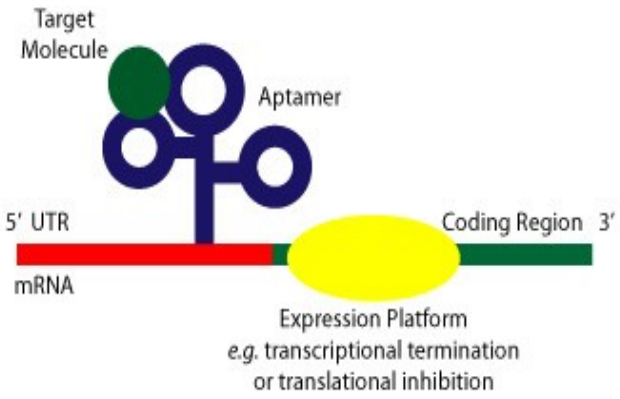
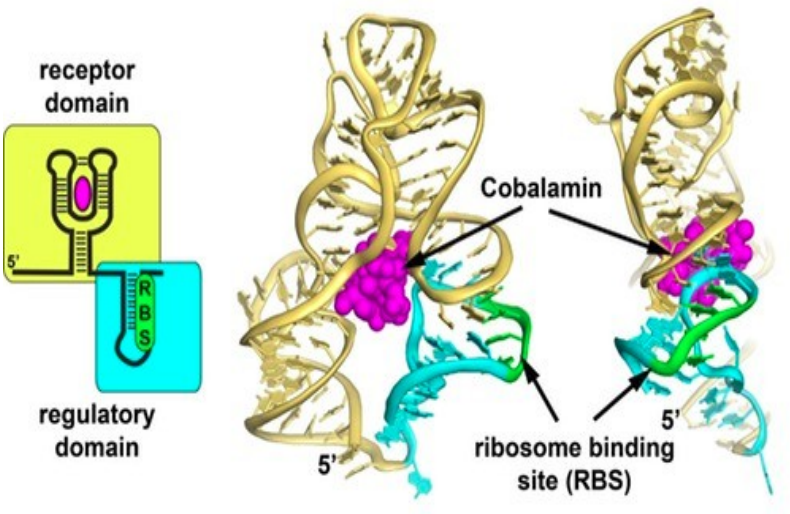


Illustration 1: Basic Structure of a Riboswitch

The pairing of the switching sequence directs RNA folding into one of two exclusive structures in the expression platform that represent the on and off states of the mRNA.

Illustration 1, taken from Batey, illustrates the basic structure of a riboswitch; the riboswitch exists in the 5' UTR of the mRNA and is characterized by its two aptamer domain. As seen in the effector molecule, the target metabolite directly interacts with the riboswitch, while the expression platform interfaces the translational and transcriptional mRNA.



Every year, about three new classes of riboswitches are being discovered, which have been shown to selectively bind metabolites such as coenzymes, amino acids, small ligands, nucleobases, and their derivatives. (Ames and Breaker, 2009). The characteristic binding mechanism of riboswitches indicates that riboswitches probably functioned as regulatory systems well before enzymes and genetic factors made of protein (Nahvi et al. 2002); if this is true, then riboswitches accurately reflect the capabilities of RNA sensors, switches, and regulation systems.

This paper focuses on nine riboswitches: TPP, FMN, SAM, Lysine, Cobalamine, Glycine, Molybdenum, Mg, and SAH, and whether or not there exists a relationship between genome size and riboswitch existence. The hypothesis of this paper is that there exists a direct relationship because it is assumed that the more basepairs (bp) in a genome, the higher the chance that an old characteristic such as a riboswitch, is conserved. If a genome is smaller, it is assumed that it is more probable that the genome would have decreased in size, and therefore less likely that any characteristic is conserved.

Literature Review:

The scope of research in riboswitches is extremely small because riboswitches are still a very new discovery, as of the early 2000s. Primarily, riboswitches have been researched to determine their function, form, and the scope of organisms that riboswitches exist in. Mehta Neel and Prashanna Balaji excellently summarized the research of Winkler, Breaker, Batey, and Nahvi, etc in *Riboswitches: Classification, Function, and Insilico Approach.* The basic function and classification of the ten

identified riboswitches are outlined in their paper, and referenced throughout this paper.

Further research has been conducted for how specific riboswitches function, primarily for the TPP, cobalamin, and glycine riboswitch. JD Stormo and Y Ji in *Do mRNAs Act as Direct Sensors of Small Molecules to Control Their Expression* were among the first to disclose that riboswitches are unique in their ability to directly bind the metabolite and the mRNA. This has been a profound discovery primarily because this characteristic does not exist otherwise, and consequently the question arose: Why aren't riboswitches preferred over proteins that perform the same function? They suggested that riboswitches are more efficient than their protein counterparts, which is discussed as a further expansion of the scope of this paper in the conclusion.

JE Barrick and RR Breaker are notable leader in identifying *how* riboswitches metabolize their ligands; the primary focus of their work has been comparing riboswitch mechanisms with protein mechanisms. As they discovered, the mechanisms are incredibly similar; as posed at the end of their paper *The Distributions, Mechanisms, and Structures of Metabolite Binding Riboswitches*, why do organisms that house riboswitches also house proteins and why would one not out-compete the other?

Methodolgy:

Three hundred twenty members of the gammaproteobacteria class were selected using the Rfam and NCBI databases. Selections were made based on the availability of the complete genome of the bacteria and if the existence or not of riboswitches in the genomes were mapped. Each bacteria was mapped for the existence of each of the nine riboswitches: TPP, FMN, SAM, Lysine, Cobalamine, Glycine, Molybdenum, Mg, and SAH. The presence of a riboswitch existed as a 1 while the absence existed as a 0 such that the total riboswitch count for each individual bacteria could be summed. Complete genome sizes of each bacteria were obtained from NCBI database, and one way ANOVA was ran on the entire set in order to obtain a standard deviation of the entire group. Each riboswitch was then categorized based on genome size, so that the standard deviation of each group was within one standard deviation of the standard deviation of the entire group; this was done so the groups could not vary too far from each other, to make the groups directly comparable to each other with a 95% confidence interval. Each group was then analyzed via direct counting correlation and one way ANOVA for correlation and covariance. In the one way ANOVA, it was assumed that the nominal

variable with two factors was riboswitch presence or absence taking on 1 or 0 values, respectively, and the independent variable was the genome size of each bacteria analyzed. It was also assumed that the nominal variable was dependent on the independent variable.

In order to check the last assumption about the nominal variable, the riboswitches were grouped according to their respective genera, and the same statistics were applied. The assumption with the one way ANOVA was that the riboswitch existence, the nominal variable with two inputs, was dependent on the genus ancestry, the independent variable. The Method of Most Likelihood was ran for both ANOVAs mentioned via SPSS.

Results and Discussion:

Confidence Matrix 1: Standard Deviations and Grouping Assignments

Genome Size (in million base pairs)	Standard Deviation of Group	Ratio of Group Stdev over Entire Stdev
Entire	1215505	1
[1, 2)	120992	0.1
[2, 3)	302213	0.24
[3, 4)	275207	0.23
[4, 4.651)	191686	0.16
(4.651, 5)	101233	0.1
[5, 6)	241475	0.2
[6, 7)	252818	0.21

The results and discussion are presented in the following sections:

1. The TPP Riboswitch
2. The FMN Riboswitch
3. The SAM Riboswitch
4. The Lysine Riboswitch
5. The Cobalamine Riboswitch
6. The Glycine Riboswitch
7. The Molybdenum Riboswitch
8. The Mg Riboswitch
9. The SAH Riboswitch

1. The TPP Riboswitch

Figure 1: Existence of the TPP Riboswitch as a Function of Size

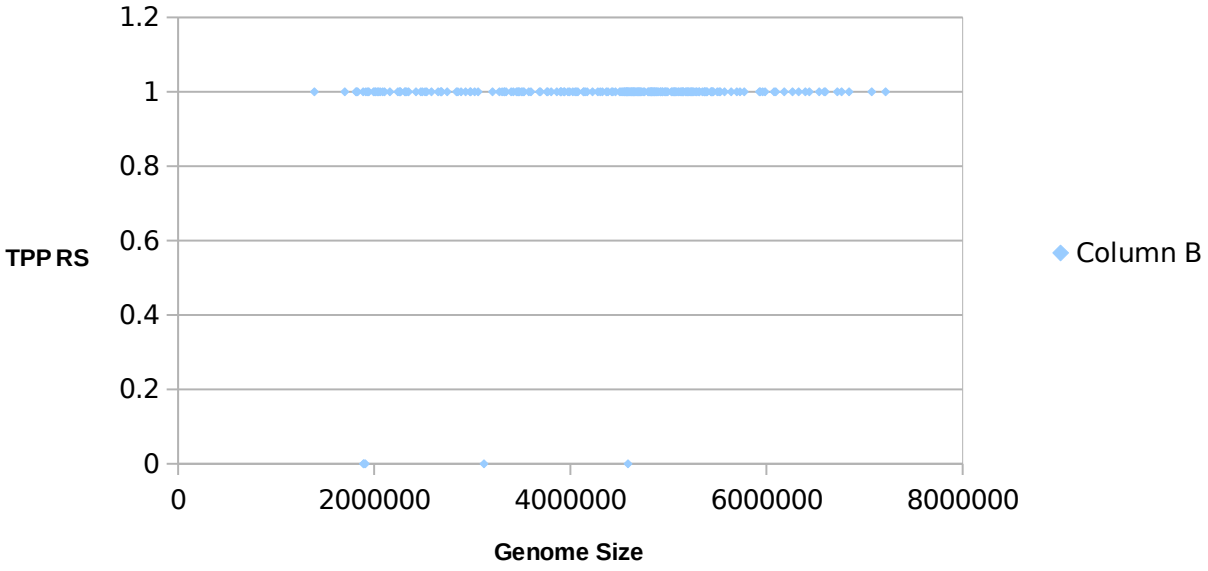


Table 1: TPP Riboswitch Presence based on Genome Size

Genome Size (in million base pairs)	Exact Ratio Present	Decimal Approximation
[1, 2)	11/22	0.5
[2, 3)	38/38	1
[3, 4)	34/35	0.9714
[4, 4.651)	66/67	0.9851
(4.651, 5)	65/65	1
[5, 6)	72/72	1
[6, 7)	13/13	1

Table 2: TPP Riboswitch Presence based on Genus

Genus	Exact Ratio Present	Decimal Approximation
Acidithiobacillus	4/4	1
Acinetobacter	12/12	1
Actinobacillus	4/4	1
Aeromonas	3/3	1
Coxiella	5/5	1
Dickeya	4/4	1
Enterobacter	5/5	1
Erwinia	4/4	1
Escherichia	48/48	1
Francisella	3/14	0.21
Haemophilus	11/11	1
Klibsiella	4/4	1
Legionella	6/6	1
Marinomonas	3/3	1
Nitrosococcus	3/3	1
Pantoea	5/5	1
Pseudomonas	25/25	1
Psychrobacter	3/3	1
Salmonella	21/21	1
Serratia	4/4	1
Shewanella	22/22	1
Shigella	8/8	1
Stenotrophomonas	3/3	1
Xanthomonas	12/12	1
Xylella	5/5	1
Yersinia	17/17	1

There are 301 TPP riboswitches present in the bacteria analyzed; the missing riboswitches belong to eleven of the fourteen bacteria of the Francisella genus, the one from the Alcanivorax genus, and the only gammaproteobacterium HdN1. Table 1 indicates how the probability of the riboswitch existence changes within each size class. When comparing Table 1 with Table 2, it is evident that the existence of the TPP riboswitch is mainly dependent on the genus relationship of the bacteria. This is because the ratio of riboswitch existence is either 0 (indicating the no species of that genus has the riboswitch), or 1 (indicating that all of the species of the genus contain the riboswitch) with the exception of Francisella to be discussed next.

Since bacteria of the same genus primarily have similar genome sizes, the direct correlation between genome size and riboswitch existence is invalid. Nevertheless, there are three species of the Francisella genus that do contain the riboswitch, denoted by F2: *novicid 3523* with 1,945,310 bp, *TX077308* with 2,016,427 bp, and *ATCC 25017* with 2,035,931 bp. This is significant because the F2 have larger genome sizes by 100 – 120 bp, when measured in the range of Francisella that do not contain the riboswitch, denoted by F1: 1,890,909 bp to 1,913,619 bp. Since the TPP riboswitch is only 100 – 120 bp, this implies that the F2 are only larger in genome size because of the addition of the riboswitch. Furthermore, two of the three F2 are in the second size group seen on Table 1. Consequently, the larger the genome size, the more likely the riboswitch will be present; however, this relationship has only been illustrated for the TPP riboswitch in the Francisella genus. Thus, among genera, species with larger genome sizes are more likely to contain a riboswitch.

2. The FMN Riboswitch

As illustrated by Figure 2, there are 261 total FMN riboswitches present in the analysis. The vast majority of the absent FMN are in the lower half of the size spectrum, around 2.5 to 3.5 million bp. As seen by Table 3, every size group, except the last, contains species without the riboswitch. There exists no general trends based solely on the data taken from Table 3, and thus for the FMN riboswitch, there is not a direct correlation with genome size.

Table 3: FMN Riboswitch Presence based on Genome Size

Genome Size (in million base pairs)	Exact Ratio Present	Decimal Approximation
[1, 2)	20/22	0.91
[2, 3)	22/38	0.5789
[3, 4)	18/35	0.5143
[4, 4.651)	57/67	0.8507
(4.651, 5)	63/65	0.969
[5, 6)	66/72	0.9167
[6, 7)	13/13	1

Table 4: FMN Riboswitch Presence based on Genus

Genus	Exact Ratio Present	Decimal Approximation
Acidithiobacillus	24	1
Acinetobacter	12/12	1
Actinobacillus	4/4	1
Aeromonas	3/3	1
Coxiella	5/5	1
Dickeya	4/4	1
Enterobacter	5/5	1
Erwinia	4/4	1
Escherichia	48/48	1
Francisella	14/14	1
Haemophilus	11/11	1
Klibsiella	4/4	1
Legionella	0/6	0
Marinomonas	3/3	1
Nitrosococcus	0/3	0
Pantoea	5/5	1
Pseudomonas	25/25	1
Psychrobacter	0/3	0
Salmonella	21/21	1
Serratia	4/4	1
Shewanella	21/22	0.95
Shigella	8/8	1
Stenotrophomonas	3/3	1
Xanthomonas	12/12	1
Xylella	0/5	0
Yersinia	17/17	1

Table 4 shows that the vast majority of riboswitches that are missing are individuals that did not belong to a genus having more than 2 species analyzed in this study. For the genera that are missing the FMN riboswitch, the Legionella, Nitrosococcus, and Psychrobacter all have genomes around 3.5 million, while the Xylella has a genome around 2.5 million. This is to be expected, as this range of genome size is where most of the FMN riboswitches are absent. 15 of the 33 FMN absent in the size group of 2 to 4 million bp are a result of these genera, however the remaining 18 are all individuals. For the remaining size groups, all of the absent FMN belong to individuals (45), with the exception of two, one for Shewanella and Pseudomonas. If the genera that were completely missing the FMN were removed, then the remaining bacteria missing the FMN would consist almost solely of individuals such that there is a skew to the right. This means that bacteria with larger genomes would be more likely to contain the FMN riboswitch; thus there exists a relationship between genome size and FMN riboswitch

presence that is both positive and direct. However, it is vital to be careful here, and examine the limitations of the previous statement. The genera of the individuals being discussed here were not able to be analyzed, and therefore it is wholly possible (and even probable) that there would be a strong correlation between the individuals analyzed here and the members of their respective genera. This is probable because the relationship in Table 4 indicates that the presence or absence of the FMN riboswitch was almost entirely genus based i.e. every genus ratio was either 0, 1, or one unit away from 1 (as in one member of the group did not contain the riboswitch while all others did).With this argument, one can only speculate that genera with larger genomes are more likely to contain the FMN riboswitch, as there is no evidence in this study that supports this. Based on the results presented here, it can be concluded that the existence of the FMN riboswitch is dependent on ancestry.

3. The SAM Riboswitch

Figure 3: The Existence of the SAM Riboswitch as a function of Genome Size

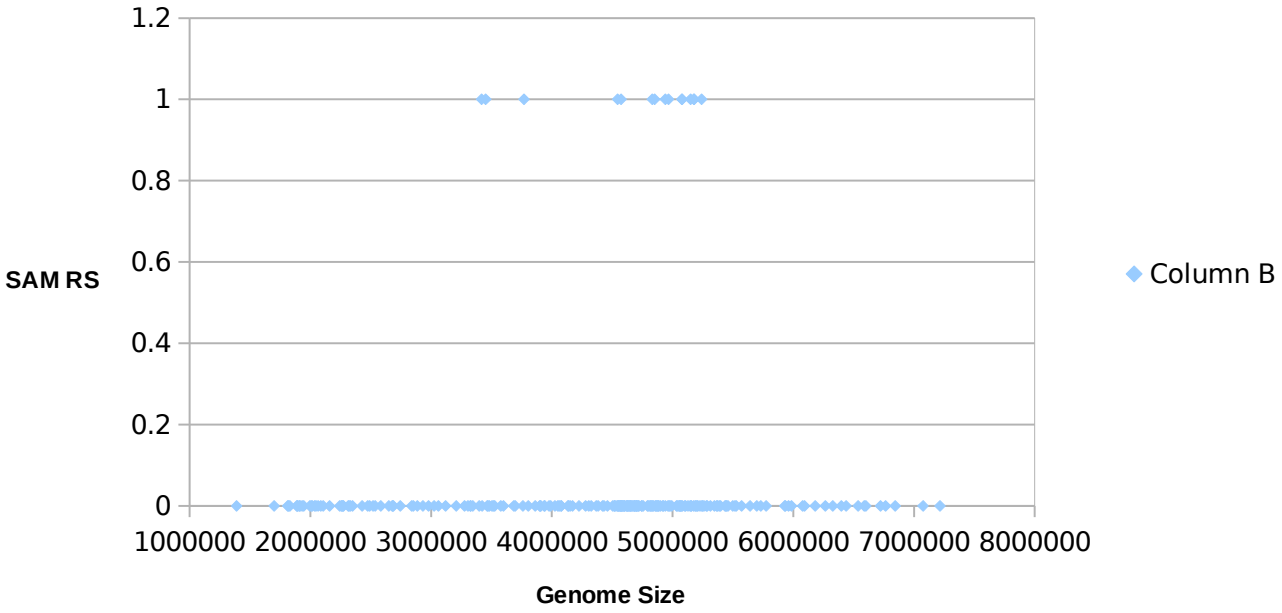


Table 5: SAM Riboswitch Presence based on Genome Size

Genome Size (in million base pairs)	Exact Ratio Present	Decimal Approximation
[1, 2)	0/22	0
[2, 3)	0/38	0
[3, 4)	3/35	0.0857
[4, 4.651)	2/67	0.02985
(4.651, 5)	6/65	0.09231
[5, 6)	6/72	0.08333
[6, 7)	0/13	0

There are 17 total SAM riboswitches. Table 5 and Figure 3 indicate that these riboswitches are fairly evenly disbursed between three to six million base pairs. Nevertheless, Table 6 shows that 15/17 of the SAM riboswitches are present between the Xanthomonas (12) and Stentrophomonas (3) genera, with the remaining two being the Pseudoxanthomonas genus that each have around 3.5 million bp. To further accredit the importance of ancestry, the Pseudoxanthomonas is more closely related to the Xanthomonas, in which all of the analyzed species contain the SAM riboswitch, than to the Pseudomonas, in which none of the species have the riboswitch; thus the Pseudoxanthomonas species contain the SAM due to close ancestry with the Xanthomonas. Furthermore, all of the ratios in Table 6 are either 0 or 1, which indicates that the genus relationship determines whether or not the SAM riboswitch exists in each species. To further illustrate that the relationship for the existence of the SAM riboswitch is dependent only on the genus relationship, it should be pointed out that the Xanthomonas genome size ranges from 3.5 million bp to 5.5 million bp. Thus the existence of the SAM cannot even be localized to a small range of genome sizes, and increasing genome size does not increase the likelihood of the SAM presence.

Table 6: SAM Riboswitch Presence based on Genus

Genus	Exact Ratio Present	Decimal Approximation
Acidithiobacillus	0/4	0
Acinetobacter	0/12	0
Actinobacillus	0/4	0
Aeromonas	0/3	0
Coxiella	0/5	0
Dickeya	0/4	0
Enterobacter	0/5	0
Erwinia	0/4	0
Escherichia	0/48	0
Francisella	0/14	0
Haemophilus	0/11	0
Klibsiella	0/4	0
Legionella	0/6	0
Marinomonas	0/3	0
Nitrosococcus	0/3	0
Pantoea	0/5	0
Pseudomonas	0/25	0
Psychrobacter	0/3	0
Salmonella	0/21	0
Serratia	0/4	0
Shewanella	0/22	0
Shigella	0/8	0
Stenotrophomonas	3/3	1
Xanthomonas	12/12	1
Xylella	0/5	0
Yersinia	0/17	0

4. The Lysine Riboswitch

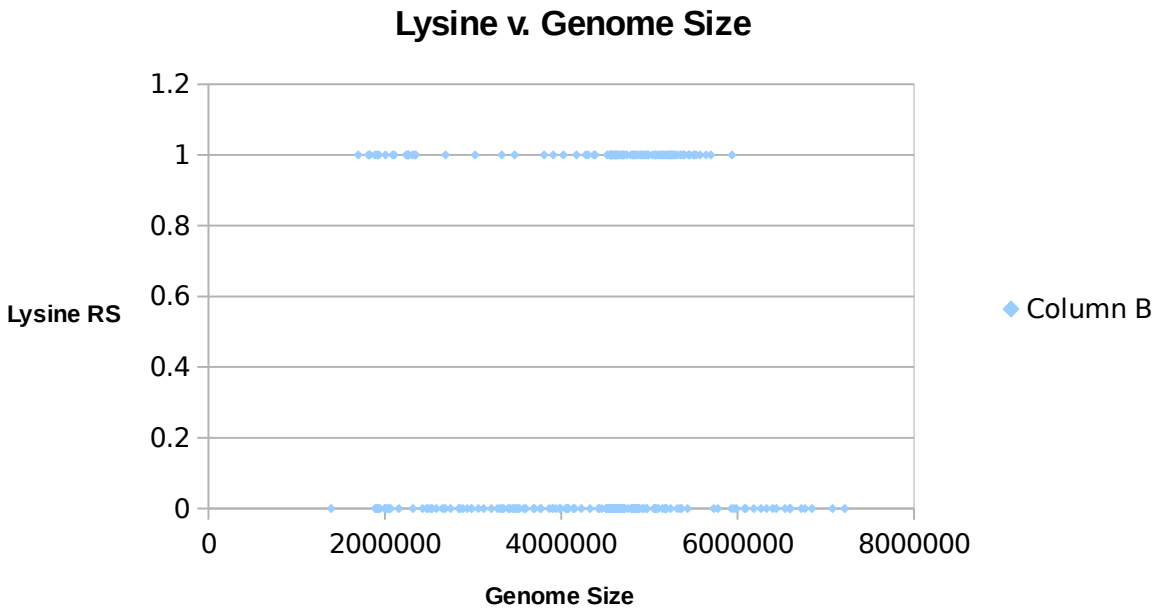


Table 7: Lysine Riboswitch Presence based on Genome Size

Genome Size (in million base pairs)	Exact Ratio Present	Decimal Approximation
[1, 2)	7/22	0.32
[2, 3)	13/38	0.34
[3, 4)	5/35	0.14
[4, 4.651)	29/67	0.43
(4.651, 5)	30/65	0.46
[5, 6)	49/72	0.68
[6, 7)	0/13	0

There are 133 Lysine riboswitches in this analysis. Table 4 and Figure 7 indicate that distribution has a slight skew to the right of genome size, and falls off completely at the end. But, Table 8 quickly refutes the implication that this skew may be due to a relationship between genome size and riboswitch existence because it magnifies that 115 of these 133 Lysine riboswitches belong to genera in which the all, or the vast majority, of the species belonging to the genus contain the riboswitch. For example, the higher decimal approximation in genome size [5, 6) million bp when compared to a group of lower genome size, say [3, 4) million bp is explained by the fact that the Escherichia genus was the largest

sample population in this analysis; 47 of the 49 riboswitches in [5, 6) belong to this genus, whereas there are only 5 in the sample size of Pantoea that belongs to [3, 4). All of the other genera that were under analysis contain no Lysine riboswitch. In summary, the species in the genera tend to behave similarly with the other species of their respective genera. It is only by coincidence that the sizes of these genera range the entire size scope. Whether or not the lysine riboswitch exists in a species is a function of what genus it belongs too, and not primarily a result of size as was hypothesized.

Table 8: Lysine Riboswitch Presence based on Genus

Genus	Exact Ratio Present	Decimal Approximation
Acidithiobacillus	0/4	0
Acinetobacter	0/12	0
Actinobacillus	3/4	0.75
Aeromonas	3/3	1
Coxiella	0/5	0
Dickeya	4/4	1
Enterobacter	1/5	0.2
Erwinia	4/4	1
Escherichia	47/48	0.98
Francisella	0/14	0
Haemophilus	11/11	1
Klibsiella	4/4	1
Legionella	0/6	0
Marinomonas	0/3	0
Nitrosococcus	0/3	0
Pantoea	5/5	1
Pseudomonas	0/25	0
Psychrobacter	0/3	0
Salmonella	0/21	0
Serratia	4/4	1
Shewanella	22/22	1
Shigella	8/8	1
Stenotrophomonas	0/3	0
Xanthomonas	0/12	0
Xylella	0/5	0
Yersinia	0/17	0

5. The Cobalamine Riboswitch

Figure 5: Cobalamine v. Genome Size

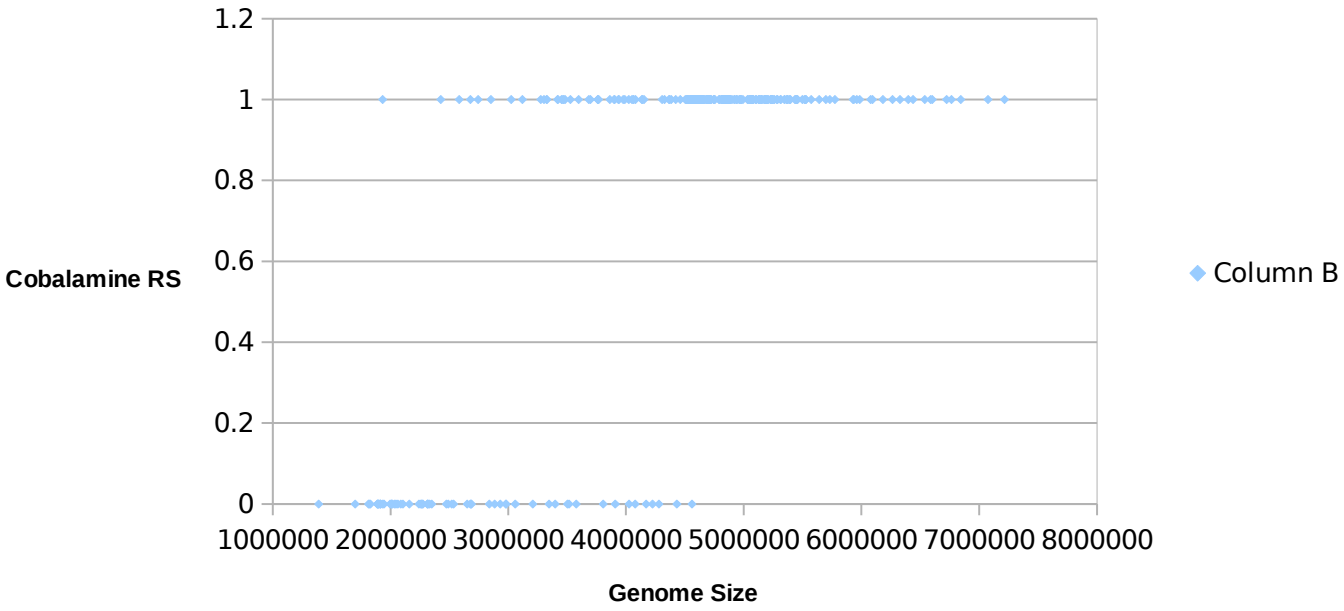


Figure 6: Cobalamine Riboswitch Decimal Approximation v. Genome Size

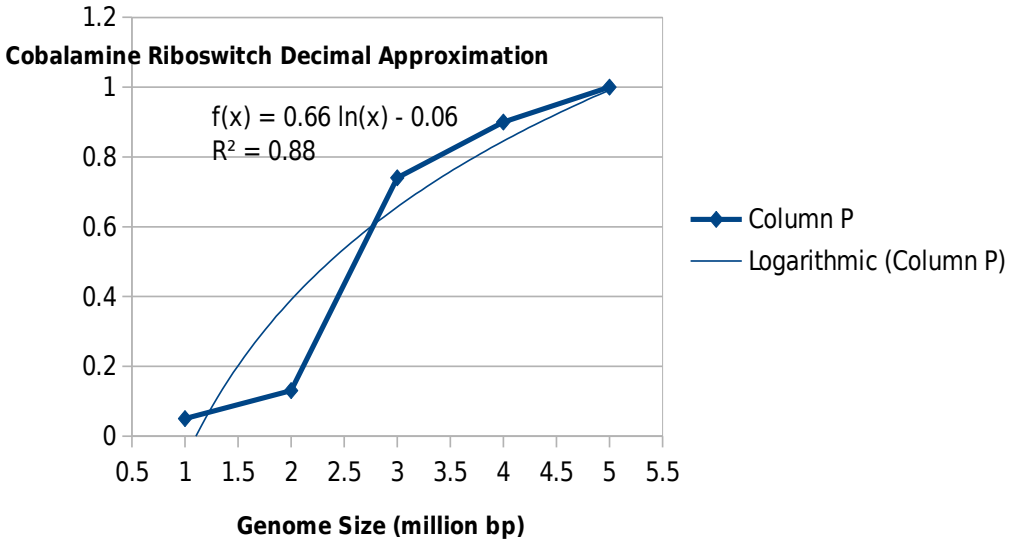


Table 9: Cobalamine Riboswitch Presence based on Genome Size

Genome Size (in million base pairs)	Exact Ratio Present	Decimal Approximation
[1, 2)	1/22	0.05
[2, 3)	5/38	0.13
[3, 4)	26/35	0.74
[4, 4.651)	60/67	0.9
(4.651, 5)	65/65	1
[5, 6)	72/72	1
[6, 7)	13/13	1

The Cobalamine Riboswitch is an excellent example of why the one way ANOVA in this analysis was misleading. As seen from Figure 6, there is a logarithmic correlation of degree two between the genome size and existence of the Cobalamine riboswitch. This relationship is correlative and the R^2 value of 0.88 is significant. However, the genus relationship with the riboswitch follows the same pattern as all of the previously discussed riboswitches. Nevertheless, the Cobalamine riboswitch is of particular interest thus far because the genera with larger genome sizes have a higher probability of containing the riboswitch. This can be exemplified visually by Figure 6 taken in context with Tables 9 and 10; Erwinia has a decimal approximation of only 0.25 in the genome size of [1, 2) while none of the other 3 genera in [1, 2) million bp have the riboswitch; in group [2, 3), only one genus shows the Cobalamine of a total of 5 analyzed. In the genera of genome size 4 million bp or higher, all of the genera tested contain the riboswitch, with 7 missing in [4, 4.651) million bp due to individuals in that group. Consequently, the R^2 value and logarithmic correlation of degree two between genome size and the existence of the Cobalamine riboswitch, coupled with the obvious genus relationship exemplified by all of the previous riboswitches discussed and exemplified in Table 10, it can be concluded that for the Cobalamine riboswitch, genera with larger genome sizes have a higher chance of containing the riboswitch. In summary, there exists a secondary, positive correlation between genome size and Cobalamine riboswitch existence.

Table 10: Cobalamine Riboswitch Presence based on Genus

Genus	Exact Ratio Present	Decimal Approximation
Acidithiobacillus	0/4	0
Acinetobacter	12/12	1
Actinobacillus	0/4	0
Aeromonas	3/3	1
Coxiella	0/5	0
Dickeya	4/4	1
Enterobacter	5/5	1
Erwinia	1/4	0.25
Escherichia	48/48	1
Francisella	0/13	0
Haemophilus	0/11	0
Klibsiella	4/4	1
Legionella	0/6	0
Marinomonas	3/3	1
Nitrosococcus	3/3	1
Pantoea	5/5	1
Pseudomonas	25/25	1
Psychrobacter	0/3	0
Salmonella	21/21	1
Serratia	4/4	1
Shewanella	22/22	1
Shigella	8/8	1
Stenotrophomonas	3/3	1
Xanthomonas	12/12	1
Xylella	0/5	0
Yersinia	17/17	1

Figure 7: Glycine v. Genome Size

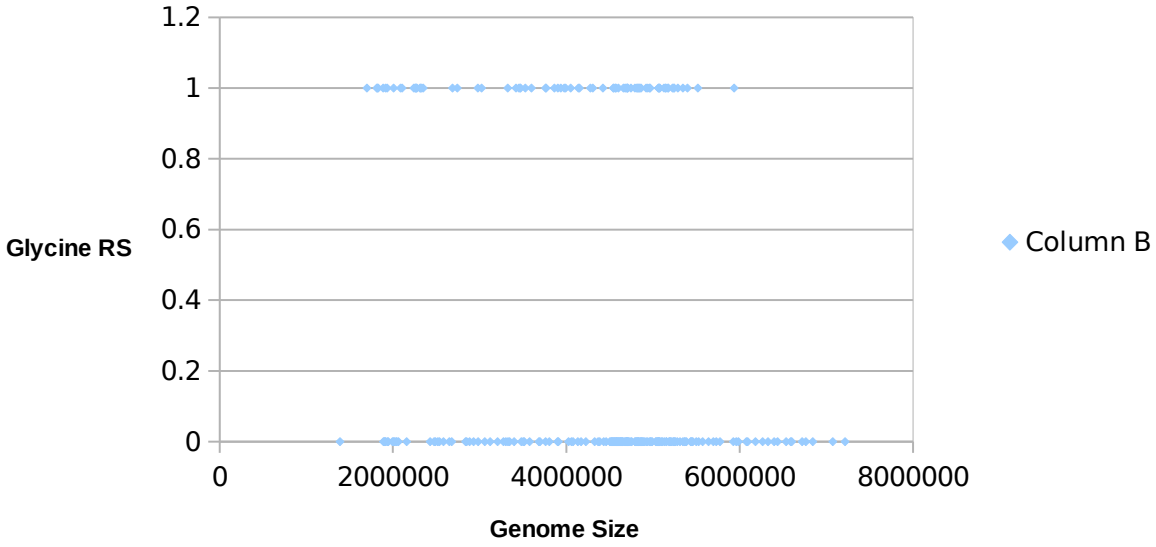


Table 11: Glycine Riboswitch Presence based on Genome Size

Genome Size (in million base pairs)	Exact Ratio Present	Decimal Approximation
[1, 2)	7/22	0.32
[2, 3)	16/38	0.42
[3, 4)	17/35	0.49
[4, 4.651)	12/67	0.18
(4.651, 5)	21/65	0.32
[5, 6)	20/72	0.28
[6, 7)	0/13	0

Across the genome sizes, there is a pretty consistent ratio for presence of the Glycine riboswitch, so no correlation can be determined. Automatically, the hypothesis of this paper is invalidated for the glycine riboswitch. As with all other previous riboswitches analyzed, the genus relationship is the determining factor of whether or not the riboswitch will be absent. A couple of points to note though, are the few missing riboswitches in the Shewanella genus and the few present riboswitches in the Pseudomonas, Erwinia, and Psychrobacter genera. For the latter, the Pseudomonas containing are the four stutzeri species in the genus, while the Psychrobacter, Erwinina, and Shewanella are different individual species within their respective genera. The difference among the stutzeri, however, does suggest a potential, more strict correlation between species and genome size. Since this minor trend cannot be observed elsewhere in this analysis, no conclusions can actually be made.

Table 12: Glycine Riboswitch Presence based on Genus

Genus	Exact Ratio Present	Decimal Approximation
Acidithiobacillus	0/4	0
Acinetobacter	12/12	1
Actinobacillus	4/4	1
Aeromonas	3/3	1
Coxiella	0/5	0
Dickeya	4/4	1
Enterobacter	0/5	0
Erwinia	1/4	0.25
Escherichia	0/48	0
Francisella	0/14	0
Haemophilus	11/11	1
Klibsiella	0/4	0
Legionella	0/6	0
Marinomonas	0/3	0
Nitrosococcus	0/3	0
Pantoea	0/5	0
Pseudomonas	4/25	0.16
Psychrobacter	1/3	0.333
Salmonella	0/21	0
Serratia	0/4	0
Shewanella	20/22	0.91
Shigella	0/8	0
Stenotrophomonas	3/3	1
Xanthomonas	12/12	1
Xylella	0/5	0
Yersinia	0/17	0

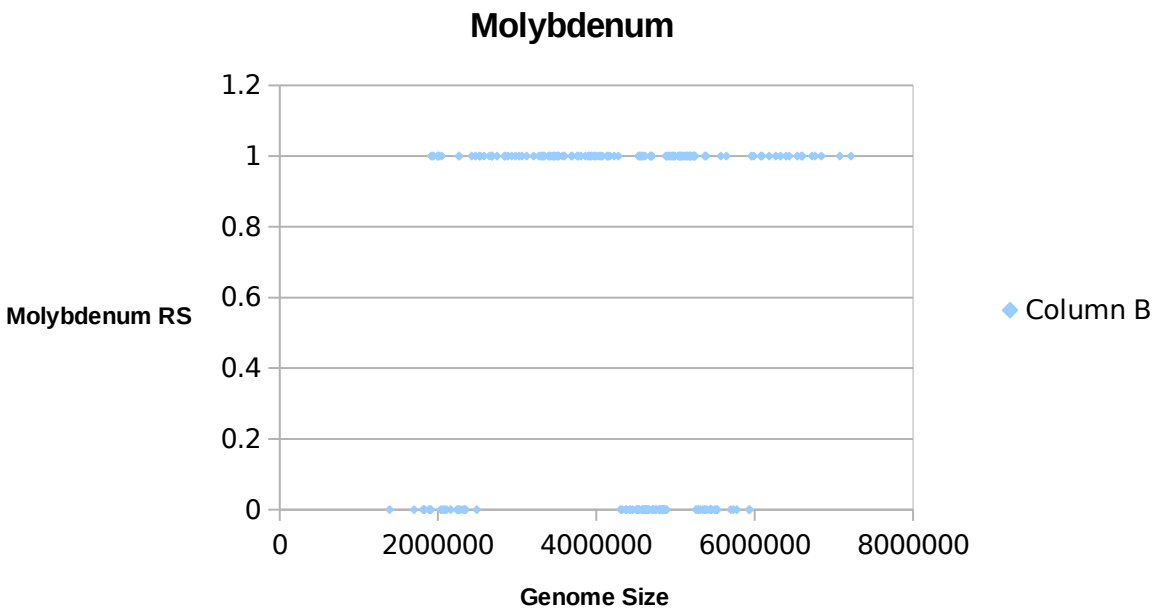


Table 13: Molybdenum Riboswitch Presence based on Genome Size

Genome Size (in million base pairs)	Exact Ratio Present	Decimal Approximation
[1, 2)	7/22	0.32
[2, 3)	14/38	0.37
[3, 4)	7/35	0.2
[4, 4.651)	52/67	0.78
(4.651, 5)	54/65	0.83
[5, 6)	53/72	0.74
[6, 7)	0/13	0

Unfortunately, the Molybdenum riboswitch is not an interesting or unique analysis. It has no genome size relationship and the existence of the riboswitch is fully correlated to genus ancestry within the scope of this analysis, and for all of the same reasons as already outlined.

Table 14: Molybdenum Riboswitch Presence based on Genus

Genus	Exact Ratio Present	Decimal Approximation
Acidithiobacillus	0/4	0
Acinetobacter	0/12	0
Actinobacillus	4/4	1
Aeromonas	3/3	1
Coxiella	0/5	0
Dickeya	4/4	1
Enterobacter	5/5	1
Erwinia	4/4	1
Escherichia	48/48	1
Francisella	0/14	0
Haemophilus	11/11	1
Klibsiella	4/4	1
Legionella	0/6	0
Marinomonas	0/3	0
Nitrosococcus	0/3	0
Pantoea	5/5	1
Pseudomonas	0/25	0
Psychrobacter	0/3	0
Salmonella	21/21	1
Serratia	4/4	1
Shewanella	22/22	1
Shigella	8/8	1

Stenotrophomonas	0/3	0
Xanthomonas	0/12	0
Xylella	0/5	0
Yersinia	17/17	1

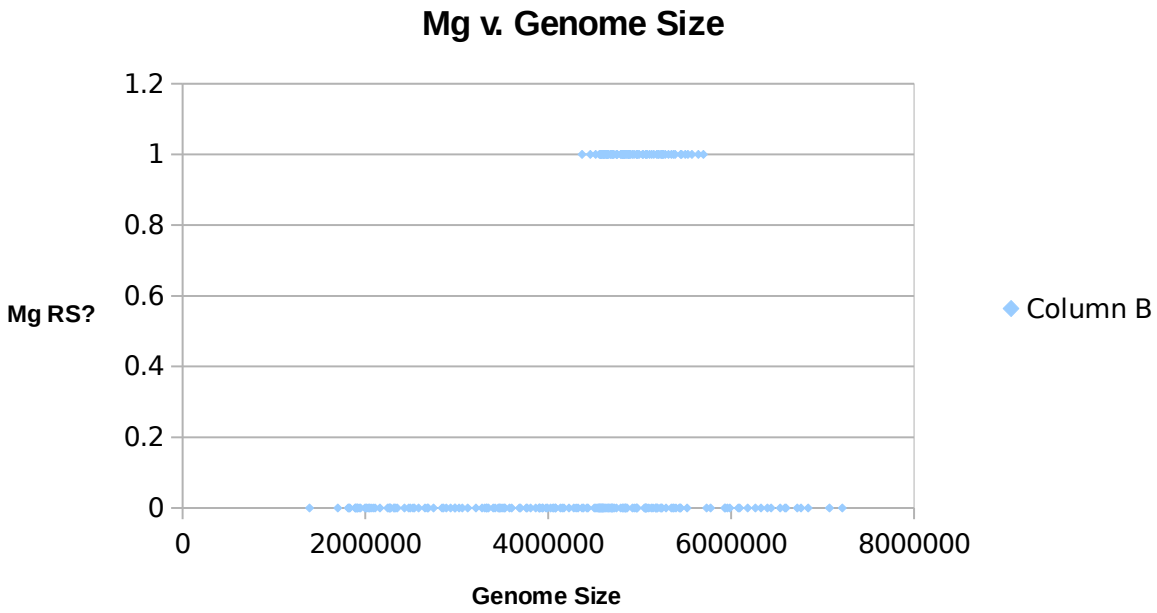


Table 15: Mg Riboswitch Presence based on Genome Size

Genome Size (in million base pairs)	Exact Ratio Present	Decimal Approximation
[1, 2)	0/22	0
[2, 3)	0/38	0
[3, 4)	0/35	0
[4, 4.651)	22/67	0.33
(4.651, 5)	34/65	0.52
[5, 6)	32/72	0.44
[6, 7)	0/13	0

As with the Molybdenum riboswitch, there is no genome size v. riboswitch existence relationship determined in the scope of this analysis, and the existence of the riboswitch is primarily dependent on the genus. There are two notable points: the only Escherichia without the Mg riboswitch is the only Escherichia that is not of the coli species (it is the fergusoni) implying again that there may be a more

informative, strict species relationship worth exploring. In direct opposition of this, however, is the Dickeya genus, where the only individual without the riboswitch belongs to the main species analyzed, dadantii. The zoea species of Dickeya follows the general Dickeya trend of having the Mg riboswitch, which further validates the conclusion that the genus relationship is the strongest indicator about whether or not a riboswitch will be absent in an individual.

Table 16: Mg Riboswitch Presence based on Genus

Genus	Exact Ratio Present	Decimal Approximation
Acidithiobacillus	0/4	0
Acinetobacter	0/12	0
Actinobacillus	0/4	0
Aeromonas	0/3	0
Coxiella	0/5	0
Dickeya	1/4	0.25
Enterobacter	5/5	1
Erwinia	0/4	0
Escherichia	47/48	0.98
Francisella	0/14	0
Haemophilus	0/11	0
Klibsiella	4/4	1
Legionella	0/6	0
Marinomonas	0/3	0
Nitrosococcus	0/3	0
Pantoea	0/5	0
Pseudomonas	0/25	0
Psychrobacter	0/3	0
Salmonella	21/21	1
Serratia	0/4	0
Shewanella	0/22	0
Shigella	8/8	1
Stenotrophomonas	0/3	0
Xanthomonas	0/12	0
Xylella	0/5	0
Yersinia	0/17	0

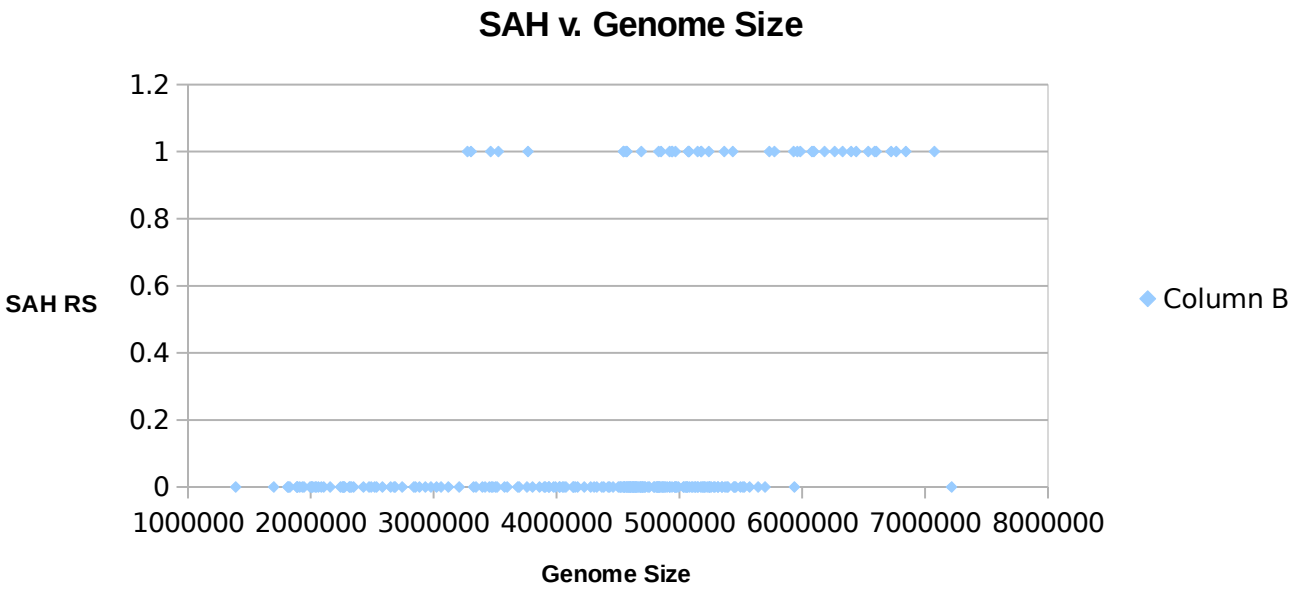


Table 17: SAH Riboswitch Presence based on Genome Size

Genome Size (in million base pairs)	Exact Ratio Present	Decimal Approximation
[1, 2)	0/22	0
[2, 3)	0/38	0
[3, 4)	5/35	0.14
[4, 4.651)	4/67	0.06
(4.651, 5)	8/65	0.12
[5, 6)	14/72	0.19
[6, 7)	13/13	1

At first glance, it appears as though there is a minor genome size to riboswitch existence relationship, as seen in Table 17. Glancing over Table 18 though, and this is void. As with the previous riboswitches under analysis, the relationship is genus based.

Table 18: SAH Riboswitch Presence based on Genus

Genus	Exact Ratio Present	Decimal Approximation
Acidithiobacillus	0/4	0
Acinetobacter	0/12	0
Actinobacillus	0/4	0
Aeromonas	0/3	0
Coxiella	0/5	0
Dickeya	0/4	0
Enterobacter	0/5	0
Erwinia	0/4	0
Escherichia	0/48	0
Francisella	0/14	0
Haemophilus	0/11	0
Klibsiella	0/4	0
Legionella	0/6	0
Marinomonas	0/3	0
Nitrosococcus	0/3	0
Pantoea	0/5	0
Pseudomonas	25/25	1
Psychrobacter	0/3	0
Salmonella	0/21	0
Serratia	0/4	0
Shewanella	0/22	0
Shigella	0/8	0
Stenotrophomonas	3/3	1
Xanthomonas	12/12	1
Xylella	0/5	0
Yersinia	0/17	0

Conclusion:

Retrospectively, the hypothesis of this paper was wrong. A direct, highly correlative relationship between genus ancestry and riboswitch existence was determined; whether or not each riboswitch was present or absent in the gammaproteobacteria analyzed was dependent upon whether the individual bacteria belonged to a genus that had that characteristic. There were few exceptions to this trend, and even the few exceptions were very minor and could be explained via differentiation of species within the genus, as seen with Pseudoxanthomonas, Escherichia, and Shewanella. That being said, there may be an even stricter, more informative relationship between species within a genus and riboswitch existence.

Also notably, the data collected in this paper was retrieved from Rfam and NCBI. While it was cross verified, it was not independently obtained, and therefore any conclusions made are potentially based off of false information. While Rfam and NCBI are generally reliable sources, there is a potential, however minor, for false information. Verifying the information obtained from these sources was beyond the scope of this analysis, but would make for an excellent project.

Furthermore, this analysis could be extended to comparing proteins with riboswitches, genome size, and genus ancestry in the same manner. Proteins perform exactly the same functions as their corresponding riboswitches, but ultimately thousands of basepairs are required for each and the metabolism process happens in steps, whereas riboswitches are smaller and more efficient. The size of each riboswitch explored in this paper and its metabolite that it acts on is summarized in Table 19. If riboswitches effectively regulate target metabolites, then why did protein and protein factors that equivalently control gene expression form? This fact, in and of itself, implies that riboswitches are not as effective as corresponding proteins; while there exists no evidence to support this implication, there exists no evidence to support the converse argument. It is wholly possible that riboswitches and proteins were developed via convergent evolution, and therefore there exists no direct, evolutionary relationship. Nevertheless, riboswitches *now* exist simultaneously with corresponding proteins, and therefore their relationship can be explored. If riboswitches are being actively selected over their corresponding proteins as genome size within a genus decreases in an organism, then it can be concluded that riboswitches are more efficient in some manner and biologically fit than the corresponding proteins.

Riboswitch	Metabolite	Length (bp)
Glycine	Glycine	100 -120
Lysine	Lysine	165 - 190
Adenosine-cobalamin (AdoCbl)	Vitamin B12	200 -220
S-adenosyl Methionine (SAM) *	cysteine	105 - 135
Thiamine Pyrophosphate (TPP)	pyrimidine and thiazolemoieties	100 - 120
Flavin Mononucleotide (FMN)	FMN biosynthesis (competes with NADP)	120 - 140
Molybdenum *	Molybdenum	60 - 80
Mg *	Magnesium	100 - 120
S-adenosyl homocysteine (SAH) *	Methionine adenosyltransferase	N/A

Special Thanks to:

Dr. Rahul Raghavan, Todd Smith, William Scott, and the rest of the Raghavan Research team for aiding me in this process and research, and for putting up with me.

Bibliography:

Barrick JE, Breaker RR. The distributions, mechanisms, and structures of metabolite-binding riboswitches. Genome Biol. 2007;8(11):R239.

Batey RT (2006). "Structures of regulatory elements in mRNAs". *Curr Opin Struct Biol* **16**: 299–306. doi:10.1016/j.sbi.2006.05.001. PMID 16707260.

Geer, LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH. The NCBI Biosystems Database. **Nucleic Acids Res**. 2010 Jan; 38 (Database Issue); D492-6 *(Epub 2009 Oct 23)* [PubMed PMID: 19854944].

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003). "Rfam: an RNA family database". *Nucleic Acids Res.* **31** (1): 439–41. doi:10.1093/nar/gkg006. PMC 165453. PMID 12520045.

Kwon, Miyun and Scott A. Strobel. *Chemical Basis of Glycine Riboswitch* Cooperativity. RNA. 2010 Nov; 16(11): 2291. PMCID: PMC2957066

Stormo GD, Ji Y (August 2001). "Do mRNAs act as direct sensors of small molecules to control their expression?". *Proc. Natl. Acad. Sci. U.S.A.* **98** (17): 9465–7. doi:10.1073/pnas.181334498. PMC 55472. PMID 11504932.

Tezuka, T and Ohnishi Y. *Two glycine riboswitches activate the glycine cleavage system essential for glycine detoxification in Streptomyces griseus.* J Bacteriol. 2014 Apr;1967:1369-76. doi: 10.1128/JB.01480-13. Epub 2014 Jan 17.

Tucker BJ, Breaker RR (2005). "Riboswitches as versatile gene control elements". *Curr Opin Struct Biol* **15**: 342–8. doi:10.1016/j.sbi.2005.05.003. PMID 1591915.

Wagner, Josef, Kirsty Short, Anthony G. Catto-Smith, Don J. S, Cameron, Ruth F. Bishop, Carl D. Kirkwood. *Identification and Characterisation of 16S Ribosomal DNA from Ileal Biopsies of Children with Crohn's Disease.* 2008 Oct 8. DOI: 10.1371/journal.pone.0003578

Weinberg Z; Wang JX; Bogue J et al. (March 2010). "Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes". *Genome Biol* **11**: R31. doi:10.1186/gb-2010-11-3-r31. PMC 2864571. PMID 20230605.