The recent report by Liu and Zhang Claiming "No support for the adaptive hypothesis of lagging-strand encoding in bacterial genomes" is factually incorrect

Christopher Merrikh[1]* and Houra Merrikh[1]*

[1]Department of Biochemistry, Vanderbilt University

*Corresponding author
E-mail: christopher.merrikh@vanderbilt.edu, houra.merrikh@vanderbilt.edu

Abstract

In their manuscript[1] entitled "No support for the adaptive hypothesis of lagging-strand encoding in bacterial genomes", Liu and Zhang attempt to refute the findings of our 2018 *Nature Communications* paper[2]. Here, we demonstrate that multiple key claims made by Liu and Zhang are factually incorrect, and that their overall conclusions, including the title of their paper, are invalid. We show that multiple existing controls already preclude the null hypothesis proposed by Liu and Zhang. We also show that the authors falsely claim that we did not publish key data, when in fact, these data were presented in our manuscript. Similarly, we point out that the authors falsely claim that we made comparisons and logical arguments we never made. Furthermore, the authors claim that they reanalyzed our data when in fact they did not. Therefore, they could not, by definition, have identified errors in our data. Our reanalysis of the Liu and Zhang data reveals that the main problem is that the authors simply do not understand the meaning of the term "inversion". This caused them to misjudge the limitations of both their own inversion detection technique and ours. In fact, we find that the Liu and Zhang method cannot, specifically, identify inversions. Lastly, we provide new empirical evidence demonstrating further that our method works properly as described in our original manuscript. We conclude that Liu and Zhang's manuscript has no rational basis, and that our original hypothesis meets the criteria for acceptance based on their own standards.

Introduction

The influences driving the architecture of bacterial chromosomes are not fully understood. Specifically, the selective pressures driving genes to be encoded on the leading or lagging strand of the bacterial chromosome are still the subject of investigation. A gene's orientation is a critical concern because, when transcribed, it can lead to co-directional (leading strand genes) or head-on (lagging strand genes) collisions between the DNA replication and transcription machineries[5]. Compared to co-directional conflicts, head-on conflicts are much more detrimental to the cell[5,6]. Yet, a significant number of highly conserved genes are encoded in this orientation[2,5,6]. Multiple previous studies by our group strongly suggest that positive selection can drive the retention of certain head-on alleles[2,5,6]. In parallel, negative and neutral selection likely act on the head-on alleles of other genes. The recent manuscript by Liu and Zhang claims that positive selection does *not* drive the retention of head-on alleles, which we contest here.

Liu and Zhang did not reanalyze our data

In contrast to their statements, Liu and Zhang did not reanalyze our original GC skew data. Instead, the authors conducted an entirely new study in which they use a different method, involving phylogeny development, to identify genes that flipped between an arbitrarily chosen ancestor and a descendant ("focal") species. When they observe differences between the results of their method and the GC skew-based method, they assume that their own method must be correct, and that the GC skew method must be flawed. We demonstrate below that the differing results are not an indication that the GC skew method is flawed. Instead, this conclusion is the result of the authors' failure to recognize that their method cannot specifically identify inversions.

The Zhang method cannot, specifically, identify inversion events

Liu and Zhang attempted to identify gene inversions using: 1) a combination of phylogeny development, orthologue identification, and closed genome comparison (hereafter referred to as the "Zhang method"), and 2) although the data are not shown in their manuscript, a new analysis of the GC skew. The authors first identify inverted orthologues using the Zhang method, then apparently cross check the results with the GC skew method. They propose that if the GC skew method works, it should identify all of the flipped genes that the Zhang method identified. When they found only a partial overlap, they concluded that the GC skew method produces "false negative errors". This conclusion is completely invalid for the reasons described below.

The crux of the problem is that Liu and Zhang do not understand the meaning of the term "inversion". They incorrectly use the term in a relative sense to describe any gene that changed orientation. In fact, a core principle of the GC skew method is that a gene's status (inverted/not inverted) is *absolute*. If a gene has a negative GC skew, it has *recently* inverted. If it has a positive GC skew it has not. Therefore, there are two possible types of flipping events: gene inversion events and gene *reversion* events (Figure 1). The method used by Liu and Zhang is blind to the core concept of inversion versus reversion. In other words, their method only identifies flipping events, combining both inversion and reversion, rather than identifying inversions specifically. As such, by labelling many reversion events as "inversions", the Zhang method frequently produces false positive errors.

The Zhang method inaccurately identifies reversion events as inversions

To identify the source of the claimed errors in our data, we repeated three of the authors' six analyses (Tables 2, S1, S2, S3). We found that the GC skew method identifies most of the CD-to-HO flips identified by Liu and Zhang as inversions. However, our method did not identify most of the HO-to-CD flips as inversions. This disparity is the basis for Liu and Zhang's belief that the GC skew method does not work. However, as we demonstrate below, this result does not reflect an error in the GC skew method. Rather, this discrepancy is entirely expected.

First, Liu and Zhang made the unreasonable assumption that all genes in the ancestor species are in the non-inverted state. However, we show below that the majority of head-on genes are already in the inverted position in every ancestor species tested (Table 3). Therefore, most HO-to-CD flips identified by Zhang are likely reversion events (inversions/reversions are described in Figure 1). In accordance with these expectations, we found that most HO-to-CD gene flips identified by Liu and Zhang are indeed reversions (e.g. 88% in *M. gallisepticum*) (Table 2). This demonstrates that the majority of what Liu and Zhang call "false negative errors" are not errors at all. Instead, the Zhang method consistently produces false positive results by inaccurately identifying reversion events as inversions.

Liu and Zhang violated the limitations of the GC Skew method

Liu and Zhang also identified a second class of flipped genes that are not identified as inversions by the GC skew method but are also not reversion events. It is highly likely that these examples are the result of a combination of a gene flipping event and mutagenesis (see Figure 2 for a detailed explanation). As we stated in our manuscript, the GC skew method of inversion detection is designed to identify *recently* inverted genes. Therefore, by analyzing *ancient* gene flipping events using the GC skew, Liu and Zhang are breaking the limitations of the method. As such, the data points discussed here are not an indication that the GC skew method has flaws. Instead, Liu and Zhang's data confirm the limits of the GC skew method, showing that "recent" can be defined as speciation-level time scales.

The GC skew method works properly

The authors claim that the GC skew method for inverted gene detection does not work, asserting that our averaging of the GC skew across whole gene regions results in "erroneous inferences of gene inversion"[1]. Liu and Zhang provide no direct data to support their claim. However, they suggest that our method would be acceptable if we had simply used, what they call "the common practice" of calculating the GC skew using only nucleotides in the 3rd codon position[1]. Liu and Zhang do not acknowledge that, in direct opposition to their claim, it is standard practice to calculate the GC skew of whole chromosomes using steps of arbitrary lengths or a sliding window[34]. Therefore, Liu and Zhang's claim regarding the necessary use of the 3rd codon position is highly dubious. Nevertheless, we tested their assumption. We compared the GC skew of all genes in three species using our whole gene calculation, or only nucleotides in the 1st, 2nd, or 3rd codon positions (Table 1). Our analysis demonstrates that all methods yield consistent results for CD genes, in direct contrast to Liu and Zhang's assumption. Additionally, we find that the 3rd codon position analysis (what Liu and Zhang call the "standard method") is exceptionally unreliable when applied to head-on genes (Table 1). This is likely because mutations at this position have a high likelihood of producing synonymous changes and are therefore less subject to purifying selection. As a result, 3rd codon position nucleotides should mutate at an accelerated pace after a gene inverts, rapidly restoring the positive GC skew, and losing the hallmark signs of inversion that the GC skew method is designed to detect. Hence, Liu and Zhang's suggestion represents the

worst possible choice, both theoretically, and empirically. In fact, the GC skew calculation might even be improved by excluding these data points. Taken together, these data show that our whole-gene GC skew calculation method works properly and does not need to be modified.

Data the authors claim we did not show is actually in our original manuscript

The central claim of Liu and Zhang's manuscript is that there is "no support" for our conclusion that lagging strand encoding can be adaptive for some genes. They state that our primary evidence for the adaptive hypothesis stems from a comparison of the number of genes that flipped from HO-to-CD versus the number that flipped CD-to-HO. This is factually incorrect. At no point did we make this comparison. By extension, we could not have used this non-existent comparison to support our conclusions. Instead, the adaptive hypothesis was largely a product of our observation that HO genes have a higher frequency of dN/dS ratios significantly greater than 1 compared to CD genes. We showed these data in three instances: our original manuscript's Tables 1, 2, and 3[2]. Strangely, Liu and Zhang discuss these data in paragraph 5 of their manuscript, then immediately state that "Merrikh and Merrikh did not show any dN/dS ratio that significantly exceeds 1."[1] We direct the authors to our original manuscript's Table S3, which shows 147 examples of genes with a dN/dS ratio in excess of 1 (and as high as 3.18)[2].

Existing controls and empirical data have already overturned the null hypothesis

Liu and Zhang claim that the higher frequency of genes with a dN/dS ratio in excess of 1 that we observed among HO genes could be due to relaxed negative (i.e. neutral) selection, rather than positive selection. (Here we refer to this as the null hypothesis.) However, Liu and Zhang's interpretation of dN/dS ratios greater than 1 is in opposition to the standard interpretation in the field (our interpretation). A ratio greater than 1 is generally evidence of *positive* selection, rather than lack of purifying selection. In a clear internal contradiction, the authors themselves state that a dN/dS ratio greater than 1 is "the criterion for establishing positive selection."[1]

In support of the null hypothesis, the authors reference previous studies indicating that HO genes have a lower frequency of essential genes, implying that head-on genes may be more mutable. However, the authors missed the key point that we did not analyze all HO and CD genes, which might have skewed our results. Instead, we only analyzed genes that are highly conserved (i.e. core genes with 95% conservation of gene length and amino acid identity). Therefore, the influence of relaxed purifying selection should be equal for the HO and CD genes we analyzed, again making Liu and Zhang's null hypothesis highly unlikely.

Conclusions

In summary, despite the use of strong language, Liu and Zhang's manuscript utterly fails to undermine the validity of our work. First, the author's central claim that they identified errors in our analysis is factually incorrect. Second, they did not read our paper carefully, causing a wide variety of misunderstandings. For example, their paraphrasing of the logical framework of our manuscript is highly misleading, and at times completely wrong. In one instance, the authors assert that we addressed the topic of gene inversion *rates*, which we did not, thus relegating their many subsequent criticisms to the classification of logical fallacy. Liu and Zhang also overlook the existence of critical data (dN/dS ratios that exceed a value of 1), supporting our central conclusion. Furthermore, they missed the fact that we are studying

highly conserved core genes, rather than all genes, which negates their repeated claim that reduced purifying selection is the cause of accelerated evolution of lagging strand genes.

After reanalyzing the Liu and Zhang data in great detail, we have identified every discrepancy between the their data and our own. We have shown that the source of all claimed errors are the result of both technical and interpretation mistakes by Liu and Zhang. As such, the major criticisms raised by Liu and Zhang, including the title of their manuscript, are false.

As a last point, the authors have now overtly defined the support that they feel would justify the major conclusion we proposed in our manuscript – that the HO gene orientation can be adaptive, at least for some genes. Specifically, they correctly reference the accepted standard in the field, saying that dN/dS >1 is "the criterion for establishing positive selection."[1] In our manuscript, we showed that a higher fraction of HO genes have a dN/dS ratio significantly in excess of 1 -- precisely the data demanded by Liu and Zhang (our original manuscript's Tables, 2, 3, and Supplementary Table 3). Therefore, according to the authors' own standards, the adaptive hypothesis for lagging strand encoding of genes is both valid and well-supported by our work.

References

1) Liu H and Zhang J. No support for the adaptive hypothesis of lagging-strand encoding in bacterial genomes. bioRxiv. 2020. DOI: 10.1101/2020.01.14.906818

2) Merrikh CN and Merrikh H. Gene inversion potentiates bacterial evolvability and virulence. *Nature Communications*. 2018. DOI: 10.1038/s41467-018-07110-3

3) Arakawa K, and Tomita M. The GC Skew Index: A Measure of Genomic Compositional Asymmetry and the Degree of Replicational Selection. Evol Bioinform Online. 2007; 3: 159–168.

4) Luo H, and Gao F. DoriC 10.0: an updated database of replication origins in prokaryotic genomes including chromosomes and plasmids. Nucleic Acids Res. 2019 Jan 8; 47.

5) Lang KS, Hall AN, Merrikh CN, Ragheb M, Tabakh H, Pollock AJ, Woodward JJ, Dreifus JE, Merrikh H. Replication-Transcription Conflicts Generate R-loops That Orchestrate Bacterial Stress Survival and Pathogenesis. Cell 170, 787–799, 2017.

6) Paul S, Million-Weaver S, Chattopadhyay S, Sokurenko E, Merrikh H. Accelerated gene evolution through replication-transcription conflicts. Nature. 2013 Mar 28;495(7442):512-5.

Figures

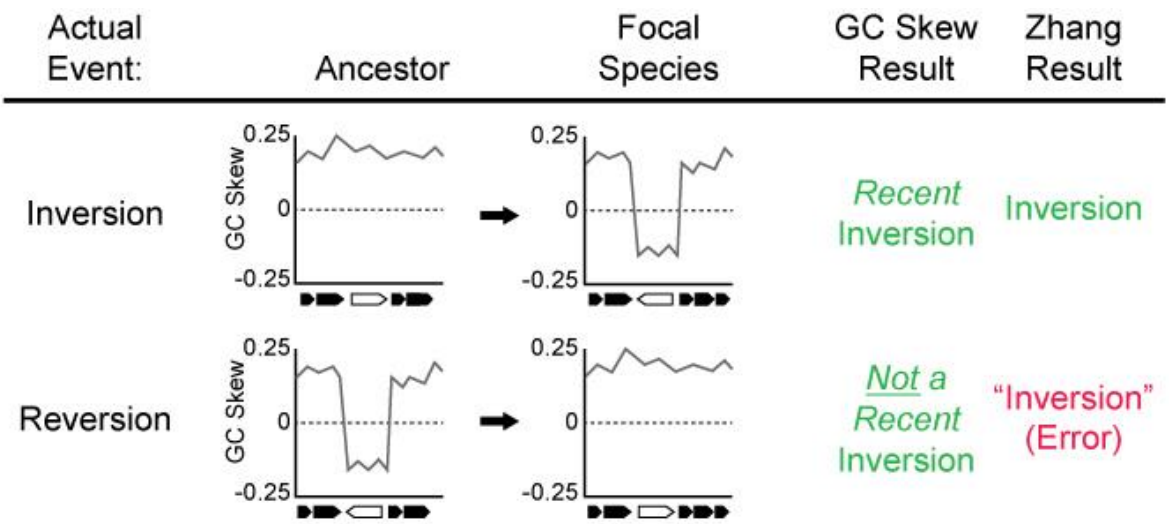

Figure 1. The Zhang method cannot distinguish between inversion and reversion. Hypothetical gene flipping data is shown for two situations: a true inversion (upper graphs) and a reversion event (lower graphs). The gene that changes orientation is annotated below (white), with surrounding genes (black). The white gene changes orientation at some point between the divergence of the ancestor and focal species (black arrows between graphs), resulting in an inversion of the GC skew within the gene region. Lower graphs: the white gene is already in the inverted state in the ancestor strain (as indicated by the negative GC skew). The flip reverts the gene to its typical orientation and restores the positive GC skew. The results of the two "inversion" detection methods are shown to the right. The GC skew identifies a gene as inverted if its GC skew is negative in a given focal species (the protocol we used in our original manuscript). The Zhang method identifies any change-of-orientation event between an ancestor and focal species as an "inversion" which is an inaccurate interpretation – these should have been labelled "flips", a term that is agnostic to directionality. In other words, the Zhang method cannot identify inversions because it cannot distinguish between inversion and reversion events. (Red text indicates that the lower Zhang method result is an inversion/reversion error.) This misinterpretation of their own data caused them to conclude (erroneously) that the GC skew method yields false negative results.

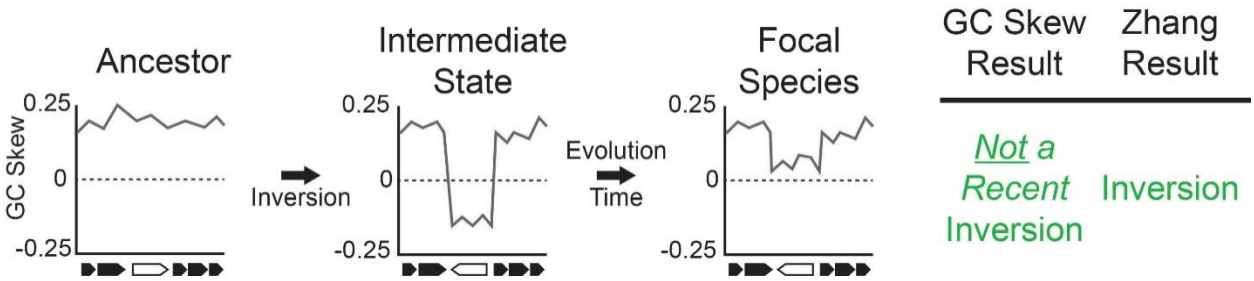

Figure 2. The GC skew method identifies only *recent* inversion events. The second cause of the so-called "errors" identified by Liu and Zhang is their inappropriate use of the GC skew method to identify ancient gene inversions. In this demonstration, a gene flips its orientation in the distant past (the left black arrow) leading to an intermediate ancestral state and a negative GC skew. Over a long time period, the gene mutates ("Evolution/Time", black arrow), increasing the GC skew. In the focal species, the GC skew method (*appropriately*) no longer detects the gene as recently inverted due to its long-term retention in the new orientation and its acclimatization (making the GC skew positive once again) through mutagenesis. Both results are accurate (green), though Liu and Zhang incorrectly interpret the GC skew result as being a false negative error.

Tables

GC Skew: Calculated for Full Gene versus:	Pearson Correlation Coefficients					
	M. gallisepticum		*B. subtilis*		*S. aureus*	
	CD	HO	CD	HO	CD	HO
Codon Pos. 1	0.54	0.32	0.51	0.55	0.58	0.65
Codon Pos. 2	0.51	0.59	0.43	0.47	0.54	0.58
Codon Pos. 3	0.44	0.07	0.53	0.45	0.42	0.35

Table 1. The GC skew calculation yields different results depending upon the method used. Pearson correlation coefficient values are shown for GC skew data sets (all HO or CD genes in the species indicated) calculated using the full gene sequence versus only the 1st, 2nd, or 3rd codon position, as indicated. The lowest correlation is almost always between the whole gene method and the codon position 3 method for head-on genes.

		# Inversions Detected		Data Agreement	Data do not match 100% Due to:	
Focal Species	*Inversion Direction*	*Zhang Method*	*GC Skew Method*	*GC Skew/ Zhang*	*Zhang Error Type 1*	*Zhang Error Type 2*
M. gallisepticum	CD > HO	25	22	88.0%	0%	12.0%
	HO > CD	40	1	2.5%	85%	12.5%
S. aureus	CD > HO	35	20	57.1%	2.9%	40.0%
	HO > CD	38	2	5.3%	52.6%	42.1%
B. subtilis	CD > HO	41	27	65.9%	0%	34.1%
	HO > CD	51	1	2.0%	70.6%	27.5%

Table 2. Two errors by Liu and Zhang are the source of all discrepancies between the two inversion detection methods. We reanalyzed the gene "inversions" (i.e. flipping events) reported by Liu and Zhang. Our analysis yielded a refined data set as not all flips could be verified (see Tables S1, S2, and S3 for additional information). For the agreed upon flip events, we calculated the GC skew of each gene in the ancestor and focal species. If a gene had a negative GC skew in the focal species (irrespective of the value in the ancestor), the gene was considered to be recently inverted. Agreement between the two methods was calculated by dividing the number of genes considered inverted by the GC skew analysis by the number identified by the Zhang method. According to Liu and Zhang, the flips identified as inverted only by their method indicate the presence of an error by the GC skew method. However, we show that this interpretation is incorrect: Two distinct errors by Liu and Zhang explain all discrepancies. Type 1 errors occurred when the authors counted a reversion event as an inversion (illustrated in Fig. 1). Type 2 errors occurred when the authors attempted to analyze an ancient gene flip using the GC skew method. This will frequently yield erroneous data because the GC skew method can only identify *recent* inversion events (illustrated in Fig. 2).

Ancestor Strain Information

Anecestor	*Focal Species*	*Orientation*	*# Genes In the Inverted State*	*# Not Inverted*	*Percent Inverted*
M. penetrans	*M. gallisepticum*	CD	71	797	8%
		HO	169	22	88%
S. saprophyticus	*S. aureus*	CD	0	1770	0%
		HO	381	234	62%
B. cereus	*B. subtilis*	CD	67	4072	2%
		HO	774	544	59%

Table 3. Many genes in the ancestor strains used by Liu and Zhang are in the inverted state. All bacterial strains tested possess a large number of inverted genes including the ancestor strains analyzed by Liu and Zhang. A gene's status as inverted or not inverted was determined using the GC skew method. Genes with a negative GC skew are considered inverted, whereas genes with a positive GC skew are considered to be in their typical orientation (i.e. not inverted).