Novel metrics for quantifying bacterial genome composition skews

Lena M. Joesch-Cohen[1,2], Max Robinson[1], Neda Jabbari[1], Christopher Lausted[1], Gustavo Glusman[1,*]

1 Institute for Systems Biology, 401 Terry Ave N, Seattle, WA 98109, USA
2 Brown University, Providence, RI 02912, USA

* Correspondence: Gustavo@SystemsBiology.org
Gustavo Glusman
Institute for Systems Biology
401 Terry Ave N
Seattle, WA 98109, USA
Tel: 206 732-1273

Abstract

Background. Bacterial genomes have characteristic compositional skews, which are differences in nucleotide frequency between the leading and lagging DNA strands across a segment of a genome. It is thought that these strand asymmetries arise as a result of mutational biases and selective constraints, particularly for energy efficiency. Analysis of compositional skews in a diverse set of bacteria provides a comparative context in which mutational and selective environmental constraints can be studied. These analyses typically require finished and well-annotated genomic sequences.

Results. We present three novel metrics for examining genome composition skews; all three metrics can be computed for unfinished or partially-annotated genomes. The first two metrics, (dot-skew and cross-skew) depend on sequence and gene annotation of a single genome, while the third metric (residual skew) highlights unusual genomes by subtracting a GC content-based model of a library of genome sequences. We applied these metrics to all 7738 available bacterial genomes, including partial drafts, and identified outlier species. A number of these outliers (i.e., Borrelia, Ehrlichia, Kinetoplastibacterium, and Phytoplasma) display similar skew patterns despite only distant phylogenetic relationship. While unrelated, some of the outlier bacterial species share lifestyle characteristics, in particular intracellularity and biosynthetic dependence on their hosts.

Conclusions. Our novel metrics appear to reflect the effects of biosynthetic constraints and adaptations to life within one or more hosts on genome composition. We provide results for each analyzed genome, software and interactive visualizations at http://db.systemsbiology.net/gestalt/skew_metrics.

Keywords

Nucleotide skew, leading strand, lagging strand, obligate intracellular, compositional bias, genome metrics, Lyme disease.

Background

Bacterial genomes display significant compositional biases, both in terms of G+C content and in compositional skews, i.e., strand asymmetries in ‘T’ vs. ‘A’ and ‘G’ vs. ‘C’ usage [1]. These biases are proposed to arise from the complex interplay of differential mutation rates and multiple selective constraints [2,3], particularly for energy efficiency [4], involving the replication, repair, and transcription enzymes. Bacterial chromosomes are replicated in both directions, from the origin of replication site to the terminator site; the "leading" strand is replicated continuously while the "lagging" strand is replicated in segments by different enzymes. Some genes are transcribed in the same direction as they are replicated (“leading strand genes”) while others are transcribed in the reverse direction (“lagging strand genes”). Each enzyme mediates both mutational and selective constraints, resulting in different compositional biases in different replicative and translational contexts. Analyses of skews in each context have the potential to expose multiple compositional constraints and their interactions, and ultimately inform about the DNA repair capacity, metabolism, and lifestyle of the species [5].

Compositional bias and strand asymmetry have been measured and analyzed in a variety of ways and contexts (for a review, see e.g. [5]). These methods include the original definitions (GC skew, (G-C)/(G+C); AT skew, (A-T)/(A+T)) [6], slight variants (e.g. G/(G+C)) [1], variants based on the three independent axes of Z Curves (x=R-Y, y=M-K, and z=S-W) [7,8], ANOVA [9], correspondence analysis of codon bias metrics [10,11], and competing mutational and selective parameters in an explicit evolutionary model [4], and have involved comparison of leading versus lagging contexts, transcribed versus intergenic regions, and restriction to each codon position.

Early examples of extreme compositional biases and asymmetries were found among species in the family Borreliaceae, tick-borne spirochetes including species causing Lyme disease (genus *Borreliella*, formerly *Borrelia*) as well as relapsing fever (genus *Borrelia*) [12]. Since its discovery in 1982 [13], the *Borreliella burgdorferi* spirochete has been of particular interest in the United States as the primary causative agent of Lyme disease. The sequencing of *B. burgdorferi* B31 in 1997 allowed an in depth exploration of the many intriguing features of the genome of this bacterium, from its small size and unusual structure (one large linear chromosome, several linear and circular plasmids) to its very low G+C content [14].

Significant skews in the third position of codons have been reported on both the leading (increased G and T) and lagging (increased A and C) strands in *B. burgdorferi* [10,15]. Among the first 43 genomes investigated [1], *B. burgdorferi* had the most extreme difference between leading and lagging strand nucleotide compositions. Both mutation and selection biases, variously induced by replication, transcription and translation constraints, have been suggested to play a role in *B. burgdorferi* [16,17] and

more generally, across all prokaryotes. The loss of some DNA repair genes may also contribute to the low G+C content and heightened skew seen in *B. burgdorferi* [18,19].

Thanks to the much expanded availability of complete genome sequences of bacterial species, it is now possible to perform large-scale comparative genomics studies [2,4,20]. A much larger number of bacterial genomes are in draft form, assembled to different levels of contiguity (contigs, scaffolds) and tentatively annotated using automated pipelines. Most of the existing methods for analyzing compositional biases and skews rely on fully or mostly contiguous genomic sequence and on the availability of detailed annotation of genes; such methods are much less applicable to the study of incomplete draft genomes.

We present here three novel metrics for quantitative analysis of genome skews. Our metrics are robust to assembly status and work well on incomplete genomes with draft annotation. Using these metrics, we analyzed a large collection of bacterial genomes—both complete and draft. We identified several groups of species and genera that present as outliers for one or more of the novel metrics. These outlier species are frequently pathogenic and tend to have unusual lifestyles, like *B. burgdorferi*.

Methods

Genomes studied

We obtained from the National Center for Biotechnology Information (NCBI) the genome sequence (in FASTA format) and current annotation (in General Feature Format, GFF) for 7948 bacterial species. We downloaded the “assembly_summary.txt” file from NCBI’s genome FTP site. This file provided various details on 86,822 genome assemblies including the organism name, RefSeq category (whether the genome considered “reference” for the species, “representative”, or otherwise) and assembly level (whether the genome is considered “completed”, or whether it is “incomplete” - assembled to chromosome, scaffold or contig level). Studying this file, we selected and downloaded:

1) 1581 “completed” genomes, (125 “reference”, 1456 “representative”),
2) 3303 “incomplete” genomes, (2 “reference”, 3301 “representative”), and
3) 3064 additional genomes, not repeating species names from the previous two sets, and prioritizing more advanced levels of completion where multiple assemblies are available for a given species.

For each genome, we included in the analysis all chromosomes, plasmids and sequence contigs at least 100 kb long. We removed from further analysis 210 genome assemblies for which the longest available sequence was shorter than 100 kb. The final set of genomes analyzed included 7738 assemblies.

Identification of origins of replication and terminator sites

For each sequence (chromosome, plasmid, scaffold and contig) in each genome assembly, we identified likely origins of replication and replication terminator sites using the GC disparity method [21], namely by identifying the minimum and maximum difference between the cumulative count of G and C along the genome. This method is independent of gene annotation and arbitrary window sizes; it can also efficiently

determine the likely direction of replication for sequence fragments (scaffolds and contigs), whether or not they include an origin of replication or a terminator site. When the resulting origin or terminator site lay within 1% of either end of the sequence, we corrected the location to coincide with the nearest sequence end.

Segmentation and analysis

We used available gene annotation (in the GFF files) to segment each sequence 100 kb or longer into a series of contiguous and disjoint segments (of variable lengths) which can be genes (including CDS, tRNA, and rRNA) or intergenic segments. We stratified intergenic segments by considering the relative orientations of the flanking genes: between two genes in the same orientation, or between two genes in opposite orientations ("head to head" or "tail to tail"). Infrequently, consecutive gene segments may be annotated as overlapping. We treated the overlapping segments as intergenic.

We computed for each segment (genic or intergenic) its length, G+C content, GC skew, and TA skew. We further determined for each oriented segment (namely genes and intergenic segments between genes transcribed in the same orientation) whether their orientation is the same or opposite to the direction of replication, i.e., whether they are on the leading or lagging strand, relative to origin and terminator sites predicted as described above.

Computation of characteristic skews

Given a set of comparable segments in a genome assembly (e.g., all genes on the leading strand), we computed skews (GC skew = (G-C)/(G+C) and TA skew = (T-A)/(T+A)) for the set as the average of the corresponding individual segment skews, weighted by segment length. We thus compute four characteristic skews for each species: $lead_{GC}$ and $lead_{TA}$ for leading strand genes, and lag_{GC} and lag_{TA} for lagging strand genes. We also evaluated weighted medians instead of weighted averages, which yielded very similar results (not shown).

Computation of the cross-skew and dot-skew

The four characteristic skews for a species can be interpreted as two characteristic skew vectors: one for the leading strand genes ($lead_{GC}$, $lead_{TA}$) and the other for the lagging strand genes (lag_{GC}, lag_{TA}). We computed the cross-skew as:

$$\text{cross-skew}(\text{lead}, \text{lag}) = |\text{lead}| \bullet |\text{lag}| \bullet \sin(\theta) \qquad (1)$$

where $|\text{lead}| = \text{sqrt}(\text{lead}_{GC}^2 + \text{lead}_{TA}^2)$, $|\text{lag}| = \text{sqrt}(\text{lag}_{GC}^2 + \text{lag}_{TA}^2)$, and θ is the angle between the two vectors. Similarly, we computed the dot-skew as:

$$\text{dot-skew}(\text{lead}, \text{lag}) = |\text{lead}| \bullet |\text{lag}| \bullet \cos(\theta) \qquad (2)$$

Computation of the residual skew

We modeled each of the four characteristic skews ($lead_{GC}$, $lead_{TA}$, lag_{GC} and lag_{TA}) as a function of the G+C content for 7738 bacterial genome assemblies. For each

characteristic skew we separated the genome assemblies with G+C content below or above 50% G+C (3635 and 4103 genomes, respectively), and fitted a robust regression line to each subset using the R function MASS::lqs() (least trimmed sum of squares, [22]. We then computed a single skew deviation magnitude metric (the residual skew) for each genome as the root mean square deviation (RMSD) from the regression line across the four characteristic skews.

Results

The genomic skews of *B. burgdorferi* are anti-correlated

A map of the GC and TA skews in 1 kb bins along the main chromosome of *B. burgdorferi* B31 shows that most bins have either a strong GC skew or a strong TA skew (Fig. 1A). Most genomic bins show a combination of both types of skew, such that the sum of the two skews (G+T vs. C+A) appears to be almost constant; this is particularly evident when plotting the TA skew vs. the GC skew (Fig. 1B). Indeed, the strength of the two skews is complementary and negatively correlated. When expressing the skews relative to the leading strand (i.e., changing the sign of skews where the leading strand corresponds to the bottom strand), we observe that the combined skew has a narrower distribution (0.270 +/- 0.099) than expected from those of the individual skews (GC skew: 0.179 +/- 0.132; TA skew: 0.091 +/- 0.112; expected deviation for their combination: 0.173). The correlation between GC and TA skews is -0.68. The plasmids of *B. burgdorferi* also have significant skews [23].

In *B. burgdorferi*, the majority of genes are transcribed in the same direction as they are replicated ('leading strand genes', blue in Fig. 1) while some are transcribed in the direction opposite to replication ('lagging strand genes', orange in Fig. 1). Leading strand genes tend to display stronger GC skew (Fig. 1C), while lagging strand genes have strong TA skews. In intergenic segments (red and green in Fig. 1C), the two skews tend to be positively correlated.

The characteristic skews of *B. burgdorferi* genes

Since the clear symmetry in the skew comparison plot for *B. burgdorferi* (Fig. 1) reflects the opposite characteristics of the two halves of the chromosome (and likewise for each plasmid), each leading from the origin of replication to a terminator (or telomere, for a linear chromosome or plasmid), it is appropriate and convenient to express the skews relative to the leading strand orientation. This transformation simplifies the representation, showing two main clusters of genes corresponding to genes transcribed on the leading strand vs. on the lagging strand (Fig. 2). To further simplify the representation (aiming to enable comparison of a large set of genomes), these two clusters can be summarized by their average TA and GC skews, weighted by gene length (see Methods). We thus computed the four characteristic skews for *B. burgdorferi*: $lead_{GC}$ = 0.259, $lead_{TA}$ = 0.022, lag_{GC} = 0.016 and lag_{TA} = 0.211.

We visualized these skews as two vectors (leading, lagging) and computed the angle between them θ = 80.89° (Fig. 2, inset). Then, based on the length and angles of

these two vectors (see Methods), we computed the *cross-skew* and *dot-skew* for *B. burgdorferi*: cross-skew(lead, lag) = 0.0584, dot-skew(lead, lag) = 0.0075. For other species within the Borrelia and Borreliella genera, these respectively ranged from 0.0562 to 0.0781 and from 0.0031 to 0.0293.

Learning from thousands of genomes

We similarly computed characteristic skews, angles and skew metrics for 7738 bacterial genome assemblies (see Methods). Visualization of these species-specific parameters demonstrates the wide diversity of bacterial genome composition; a few select examples are shown in Fig. 3. We observed genomes with strong skews and with negligible skews, at all possible angles between the skew vectors. We also created a web interface for generating species-specific skew plots and exploring their skew metrics, available at [24].

We compared the four characteristic skews of 7738 bacterial genome assemblies with their corresponding G+C content (Fig. 4). We observed that all skews are correlated with G+C content, and largely decrease in absolute value with increasing G+C content. These relationships are different for bacterial genomes with low vs. high G+C content. In fact, we observed a largely bimodal distribution of G+C content among sequenced bacterial genomes (Fig. 5, lower panel). We therefore fitted lines to the characteristic skews separately for bacterial genomes below and above 50% G+C content, and computed the deviations from the expected skews for each bacterial genome assembly. We used the R function MASS:lqs(), the leading robust linear regression method, to ensure that these lines accurately reflected the typical pattern, ignoring outliers.

The $lead_{GC}$ and lag_{TA} values of Borreliaceae genomes are large and are clear outliers relative to the entire data set of 7738 genomes (Fig. 4). On the other hand, while the Borreliaceae $lead_{TA}$ and lag_{GC} are close to zero and are not outliers relative to the entire data set, they are unusual for bacterial species with low G+C content, which tend to have negative values for these characteristic skews (Fig. 4). Thus, Borreliaceae genomes are unusual for all four characteristic skews. The deviations of characteristic skews for *B. burgdorferi* from the skews predicted by the fitted lines at the G+C content for *B. burgdorferi* are 0.091, 0.120, 0.106 and 0.124 for $lead_{GC}$, $lead_{TA}$, lag_{GC} and lag_{TA}, respectively. Borrelia species that cause relapsing fever have even larger deviations from the expected values.

Three novel metrics for analyzing genome skews

We described above several parameters for quantifying skews in individual bacterial genomes: the four characteristic skews and the magnitudes and angles of the vectors they define. Using these parameters, we defined two interrelated metrics for comparing and contrasting the skews of leading strand vs. lagging strand genes: the cross-skew and the dot-skew (see Methods). Furthermore, taking advantage of the availability of many thousand bacterial genome assemblies, we estimated expected values for each characteristic skew, as a function of the G+C content. We then used the observed deviations from these expected values to define a third metric: the *residual skew*.

We computed these three metrics for 7738 bacterial genome assemblies (available at [24]) and evaluated their relationship with G+C content (Fig. 5). For high G+C content bacteria, we observed that the cross-skew and the dot-skew are much more constrained than for lower G+C content species; these two metrics are most diverse for bacterial genomes under ~35% G+C. Compared to these two metrics, the residual skew is more diverse for all levels of G+C content. Borreliaceae genomes are clear outliers for all three metrics.

Finally, we combined all three metrics to generate a map of genome skews for all bacterial genomes (Fig. 6). In this map, most high G+C content bacteria are restricted to near the origin, while low G+C content bacteria show a more diverse spread. Borreliaceae genomes are seen as clear outliers, with the most extreme skew values corresponding to the group of Borrelia genomes that cause relapsing fever. Genomes in the genus Ehrlichia (see example in Fig. 3) are also outliers in all three metrics and show similar skew values as Borreliella genomes. Ehrlichia are intracellular vector-borne pathogens of vertebrates; like Borrelia, they have diminished biosynthetic abilities [25]. Ehrlichia are in the Rickettsiales order and are phylogenetically unrelated to Borreliaceae; the genome of *Ehrlichia canis* has a single circular chromosome and no plasmids [26]. Multiple other genera became evident as outliers of interest, discussed below.

Discussion

We have devised three novel metrics to study bacterial genome composition biases, integrating knowledge of the nucleotide skews in annotated genes, the direction of transcription relative to replication, and the G+C content of the genome.

The first two metrics (cross-skew and dot-skew) are computed based on knowledge of an individual genome's characteristic skew vectors, and they quantify the strength and relationship between the mutation and selection pressures on genes on the leading vs. lagging strands.

Positive dot-skew values (Fig. 6, top) indicate similar compositional constraints on all genes, relative to the direction of replication; an example of this pattern is observed in the obligate intracellular parasite *Chlamydia pneumoniae* [27] (Fig. 3). Conversely, negative dot-skew values (Fig. 6, bottom) reflect opposite compositional constraints on leading and lagging strand genes (i.e., transcribed in the same or opposite direction as they are replicated); extreme examples of this pattern are observed in fusobacteria including *Fusobacterium periodonticum* [28], *Leptotrichia buccalis* [29], and *Streptobacillus moniliformis* [30], the causal agent of rat bite fever. Positive dot-skew values can thus be interpreted as reflecting constraints driven mostly by the replication process, while negative dot-skew values largely reflect transcriptional and translational constraints [4].

The cross-skew quantifies the strength and orthogonality of the compositional skew vectors for leading and lagging strand genes. Genomes with high cross-skew values (Fig. 6, right) demonstrate skew patterns inconsistent with purely replicational or transcriptional constraints; Borreliaceae and Ehrlichia species are prime examples of this pattern.

Borreliaceae and Ehrlichia species lack amino acid and nucleotide synthesis pathways; the observed skew patterns in these pathogens may thus reflect a relaxation of the selection for energy efficiency that drives nucleotide usage and thus skews [4], possibly combined with more complex constraints imposed by the a life cycle that involves recurring transitions between mammalian and invertebrate (tick) hosts. We observed similar skew patterns in Kinetoplastibacteria (Fig. 6), which are endosymbionts of insect-infecting trypanosomatid flagellates [31] with multiple biosynthetic adaptations to life in the intracellular environment. Likewise, we observed distinct skew patterns among Blochmannia species (Fig. 6); these are also intracellular endosymbionts that lost multiple biosynthetic pathways and rely on the metabolic machinery of their carpenter ant hosts [32].

The third metric (residual skew) capitalizes on the current availability of thousands of complete or draft bacterial genomes to empirically assess how unusual a genome's skews are relative to the expected values as learned from other genomes. This analysis, which has not been possible until recent times, revealed that bacterial genomes with low G+C content typically have negative TA skews in leading strand genes and negative GC skews in lagging strand genes, and that these negative skews increase in magnitude as G+C content decreases (Fig. 4). On the background of these trends, the weakly positive skews observed in Borreliaceae species are highly unusual. This pattern is not evident relative to the global collection of genomes since the weakly positive Borreliaceae skews are comparable to those observed in high G+C content bacteria. Our regression analysis quantifies these deviations from expectation and integrates them into a unified metric that highlights the unusual skews in Borreliaceae species (Fig. 5) and also identifies other species as having skew patterns that are significantly unusual relative to the bulk of bacterial species. Of particular note are Phytoplasma species (Fig. 6); these are intracellular pathogens of multiple plant species that use insects as transmission vectors [33,34], in similarity to Borreliaceae and Ehrlichia for mammals.

Conclusions

We described here three novel metrics for quantifying bacterial genome composition skews and presented examples of their application to identify bacterial species with unusual skew patterns. Our metrics take advantage both of information about the genome of a single species and of patterns discernable from studying genomes of thousands of species - even those not yet finished and fully annotated. While some of the genera identified as skew outliers are phylogenetically close (e.g., Fusobacterium, Streptobacillus and Leptotrichia), our metrics identified similar skew patterns in genera of bacteria that are phylogenetically unrelated, like Borrelia, Ehrlichia and Kinetoplastibacterium, and (when considering the residual skew) Phytoplasma. These very disparate bacterial species share lifestyle characteristics (intracellularity and transmission via insect vectors), suggesting that our novel metrics successfully capture effects on genome composition of biosynthetic constraints and of interaction with the hosts.

List of abbreviations

GFF: General Feature Format; NCBI: National Center for Biotechnology Information; RMSD: root mean square deviation

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Software for computing skews for individual genomes, regression RMSD parameters, a table of skew characteristics for the 7738 bacterial genomes, and Vega-Lite [35] interactive visualizations are available at [24]. The table of characteristics for 7738 bacterial genomes is also included as supplementary material.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by generous donations from The Wilke Family Foundation, Jeff and MacKenzie Bezos, and The Steven & Alexandra Cohen Foundation. The donors were not involved in the design, implementation or reporting of this study.

Authors' contributions

GG conceived of the study. LMJC, MR and GG designed the methods and implemented the software. LMJC, NJ and GG performed analyses. LMJC, CL, MR and GG wrote the manuscript. All authors approved its final version.

Acknowledgements

We wish to thank Arian Smit and Jeff Boore for helpful discussions.

References

1. Lobry JR, Sueoka N. Asymmetric directional mutation pressures in bacteria. Genome Biol. 2002;3:RESEARCH0058.

2. Morton RA, Morton BR. Separating the effects of mutation and selection in producing DNA skew in bacterial chromosomes. BMC Genomics. 2007;8:369.

3. Vetsigian K, Goldenfeld N. Genome rhetoric and the emergence of compositional bias. Proc. Natl. Acad. Sci. U. S. A. 2009;106:215–20.

4. Chen W-H, Lu G, Bork P, Hu S, Lercher MJ. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. Nat. Commun. 2016;7:11334.

5. Dutta C, Paul S. Microbial lifestyle and genome signatures. Curr. Genomics. 2012;13:153–62.

6. Lobry JR. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 1996;13:660–5.

7. Guo F-B, Yu X-J. Re-prediction of protein-coding genes in the genome of Amsacta moorei entomopoxvirus. J. Virol. Methods. 2007;146:389–92.

8. Zhang R, Zhang CT. Z curves, an intutive tool for visualizing and analyzing the DNA sequences. J. Biomol. Struct. Dyn. 1994;11:767–82.

9. Tillier ER, Collins RA. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J. Mol. Evol. 2000;50:249–57.

10. McInerney JO. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proc. Natl. Acad. Sci. U. S. A. 1998;95:10698–703.

11. Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res. 1999;27:1642–9.

12. Cutler SJ, Ruzic-Sabljic E, Potkonjak A. Emerging borreliae - Expanding beyond Lyme borreliosis. Mol. Cell. Probes. 2017;31:22–7.

13. Burgdorfer W, Barbour AG, Hayes SF, Benach JL, Grunwaldt E, Davis JP. Lyme disease-a tick-borne spirochetosis? Science. 1982;216:1317–9.

14. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, et al. Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature. 1997;390:580–6.

15. Mackiewicza P, Gierlika A, Kowalczuka M, Szczepanika D, Dudek MR, and Cebrat S. Mechanisms generating long-range correlation in nucleotide composition of the Borrelia burgdorferi genome. Physica A: Statistical Mechanics and its Applications. 1999;273:103–15.

16. Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 2010;6:e1001107.

17. Wei W, Guo F-B. Strong Strand Composition Bias in the Genome of Ehrlichia canis Revealed by Multiple Methods. Open Microbiol. J. 2010;4:98–102.

18. Zhao H-L, Xia Z-K, Zhang F-Z, Ye Y-N, Guo F-B. Multiple Factors Drive Replicating Strand

Composition Bias in Bacterial Genomes. Int. J. Mol. Sci. 2015;16:23111–26.

19. Lind PA, Andersson DI. Whole-genome mutational biases in bacteria. Proc. Natl. Acad. Sci. U. S. A. 2008;105:17878–83.

20. Necşulea A, Lobry JR. A new method for assessing the effect of replication on DNA base composition asymmetry. Mol. Biol. Evol. 2007;24:2169–79.

21. Zhang R, Zhang C-T. Identification of replication origins in archaeal genomes based on the Z-curve method. Archaea. 2005;1:335–46.

22. Venables WN, Ripley BD. Modern Applied Statistics with S, 4th ed. Springer; 2002.

23. Picardeau M, Lobry JR, Hinnebusch BJ. Analyzing DNA strand compositional asymmetry to identify candidate replication origins of Borrelia burgdorferi linear and circular plasmids. Genome Res. 2000;10:1594–604.

24. Skew metrics of bacterial genomes [Internet]. [cited 2017 Nov 27]. Available from: http://db.systemsbiology.net/gestalt/skew_metrics/

25. Dunning Hotopp JC, Lin M, Madupu R, Crabtree J, Angiuoli SV, Eisen JA, et al. Comparative genomics of emerging human ehrlichiosis agents. PLoS Genet. 2006;2:e21.

26. Mavromatis K, Doyle CK, Lykidis A, Ivanova N, Francino MP, Chain P, et al. The genome of the obligately intracellular bacterium Ehrlichia canis reveals themes of complex membrane structure and immune evasion strategies. J. Bacteriol. 2006;188:4015–23.

27. Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, et al. Comparative genomes of Chlamydia pneumoniae and C. trachomatis. Nat. Genet. 1999;21:385–9.

28. Slots J, Potts TV, Mashimo PA. Fusobacterium periodonticum, a new species from the human oral cavity. J. Dent. Res. 1983;62:960–3.

29. Ivanova N, Gronow S, Lapidus A, Copeland A, Glavina Del Rio T, Nolan M, et al. Complete genome sequence of Leptotrichia buccalis type strain (C-1013-b). Stand. Genomic Sci. 2009;1:126–32.

30. Nolan M, Gronow S, Lapidus A, Ivanova N, Copeland A, Lucas S, et al. Complete genome sequence of Streptobacillus moniliformis type strain (9901). Stand. Genomic Sci. 2009;1:300–7.

31. Alves JMP, Serrano MG, Maia da Silva F, Voegtly LJ, Matveyev AV, Teixeira MMG, et al. Genome evolution and phylogenomic analysis of Candidatus Kinetoplastibacterium, the betaproteobacterial endosymbionts of Strigomonas and Angomonas. Genome Biol. Evol. 2013;5:338–50.

32. Gil R, Silva FJ, Zientz E, Delmotte F, González-Candelas F, Latorre A, et al. The genome sequence of Blochmannia floridanus: comparative analysis of reduced genomes. Proc. Natl. Acad. Sci. U. S. A. 2003;100:9388–93.

33. Tran-Nguyen LTT, Kube M, Schneider B, Reinhardt R, Gibb KS. Comparative genome analysis of “Candidatus Phytoplasma australiense” (subgroup tuf-Australia I; rp-A) and “Ca.

Phytoplasma asteris" Strains OY-M and AY-WB. J. Bacteriol. 2008;190:3979–91.

34. Hogenhout SA, Oshima K, Ammar E-D, Kakizawa S, Kingdom HN, Namba S. Phytoplasmas: bacteria that manipulate plants and insects. Mol. Plant Pathol. 2008;9:403–23.

35. Satyanarayan A, Moritz D, Wongsuphasawat K, Heer J. Vega-Lite: A Grammar of Interactive Graphics. IEEE Trans. Vis. Comput. Graph. 2017;23:341–50.

Figure legends

Figure 1. Bin-wise and gene-wise representations of genome skews in *B. burgdorferi*. A: TA (grey) and GC (black) skews in 1 kb bins along the (linear) main chromosome (top) and map of annotated genes and intergenic segments (bottom) showing strandedness and length of each segment: leading strand genes in blue, lagging strand genes in orange, intergenic segments flanked by genes in equal orientation in green, and intergenic segments flanked by genes in opposite orientations in red; circle area is proportional to segment length. The location of the origin of replication is evident from the sharp switch in skew sign from negative to positive. B: Comparison of skews per 1 kb bin. C: Comparison of skews per gene and intergenic segment; circle area is proportional to segment length.

Figure 2. TA vs. GC skews of gene and intergenic segments oriented relative to the leading strand. Graphic elements as in Fig. 1C. The vectors point from the origin (zero skews) to the weighted average of skews for genes on the leading strand (+) and genes on the lagging strand (-). Inset: definition of the angle θ between the two vectors.

Figure 3. Examples of TA vs. GC skew plots for several bacterial species. Graphic elements as in Figs. 1C and 2. Each plot displays skews in the range [-0.5, 0.5]. Lower-left inset for each plot: average genomic G+C content for that species.

Figure 4. Relationship between the four characteristic skew values and G+C content, for 7738 bacterial genomes, highlighting Borreliaceae species (red points). Red lines represent robust regression lines computed by least quantile of squares method.

Figure 5. A: Skew metrics vs. G+C content for 7738 bacterial genomes, highlighting Borreliaceae species (red points). From top to bottom: cross-skew, dot-skew, residual skew, and histogram of number of species studied.

Figure 6. Integration of skew metrics (cross-skew vs. dot-skew, point size represents residual skew) for 7738 bacterial genomes, highlighting some genera of interest. All genomes colored by G+C content, ranging from low (blue) to high (red).

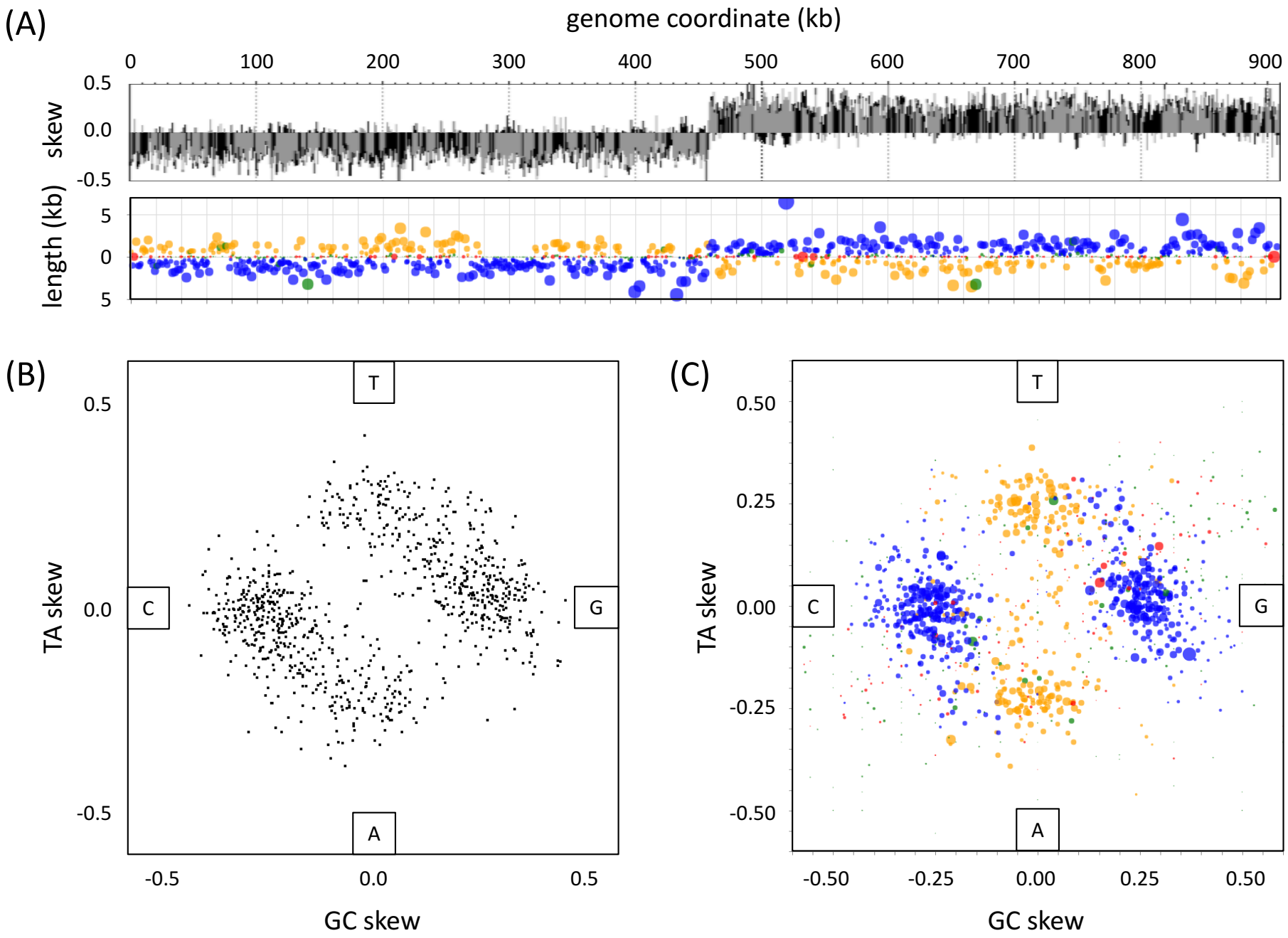
(A)
genome coordinate (kb)
0
100
200
300
400
500
600
700
800
900
skew
0.5
0.0
-0.5
length (kb)
5
0
5
(B)
T
C
G
A
TA skew
GC skew
0.5
0.0
-0.5
(C)
T
C
G
A
TA skew
GC skew
0.50
0.25
0.00
-0.25
-0.50

Figure 1

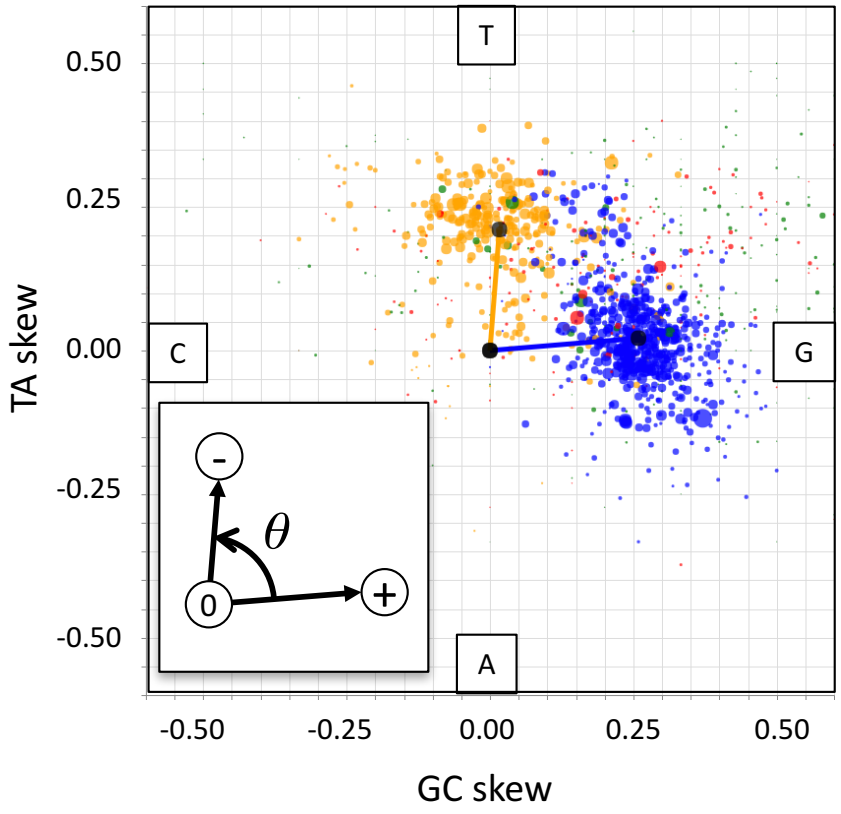

Figure 2

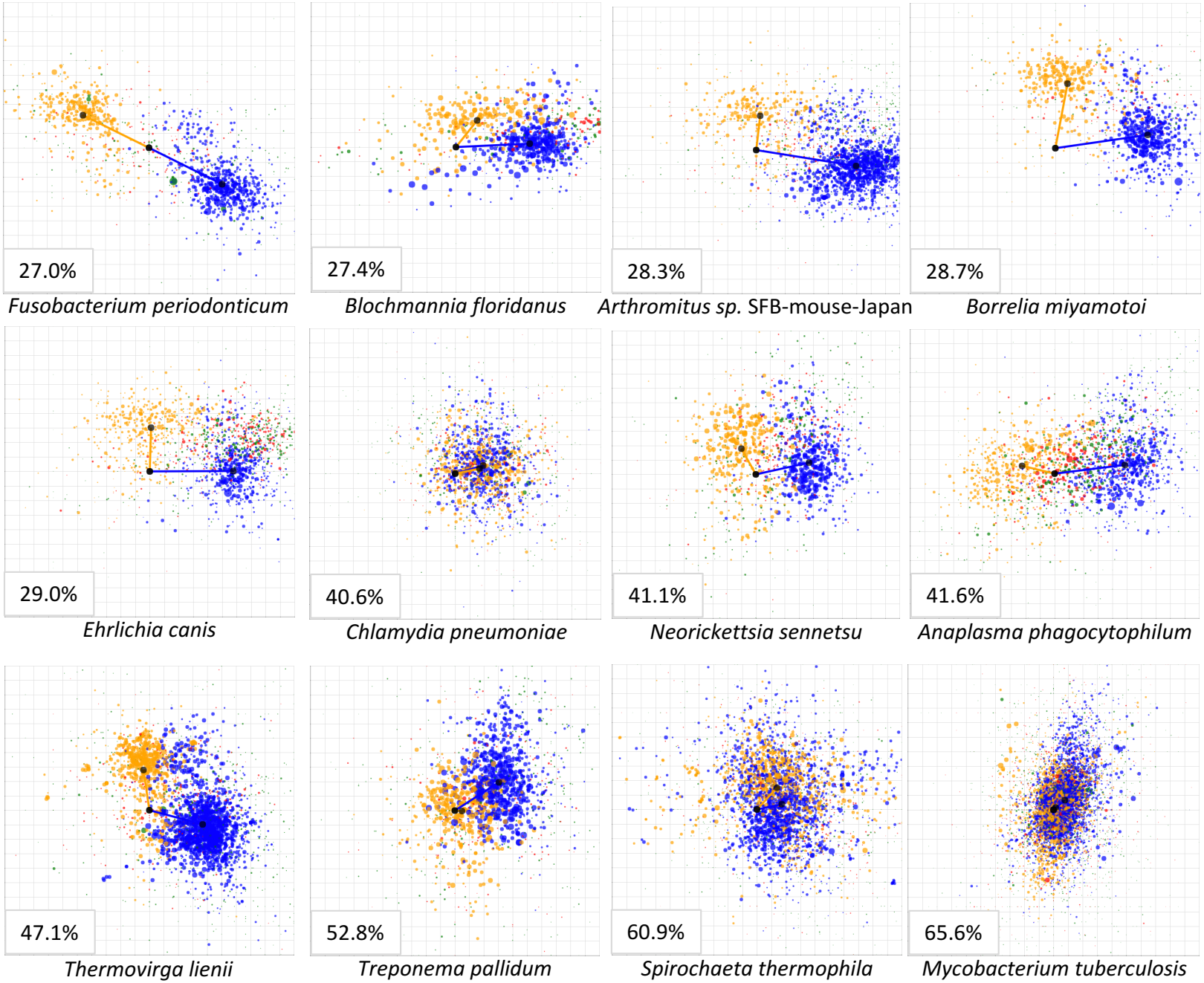
27.0%
Fusobacterium periodonticum
27.4%
Blochmannia floridanus
28.3%
Arthromitus sp. SFB-mouse-Japan
28.7%
Borrelia miyamotoi
29.0%
Ehrlichia canis
40.6%
Chlamydia pneumoniae
41.1%
Neorickettsia sennetsu
41.6%
Anaplasma phagocytophilum
47.1%
Thermovirga lienii
52.8%
Treponema pallidum
60.9%
Spirochaeta thermophila
65.6%
Mycobacterium tuberculosis

Figure 3

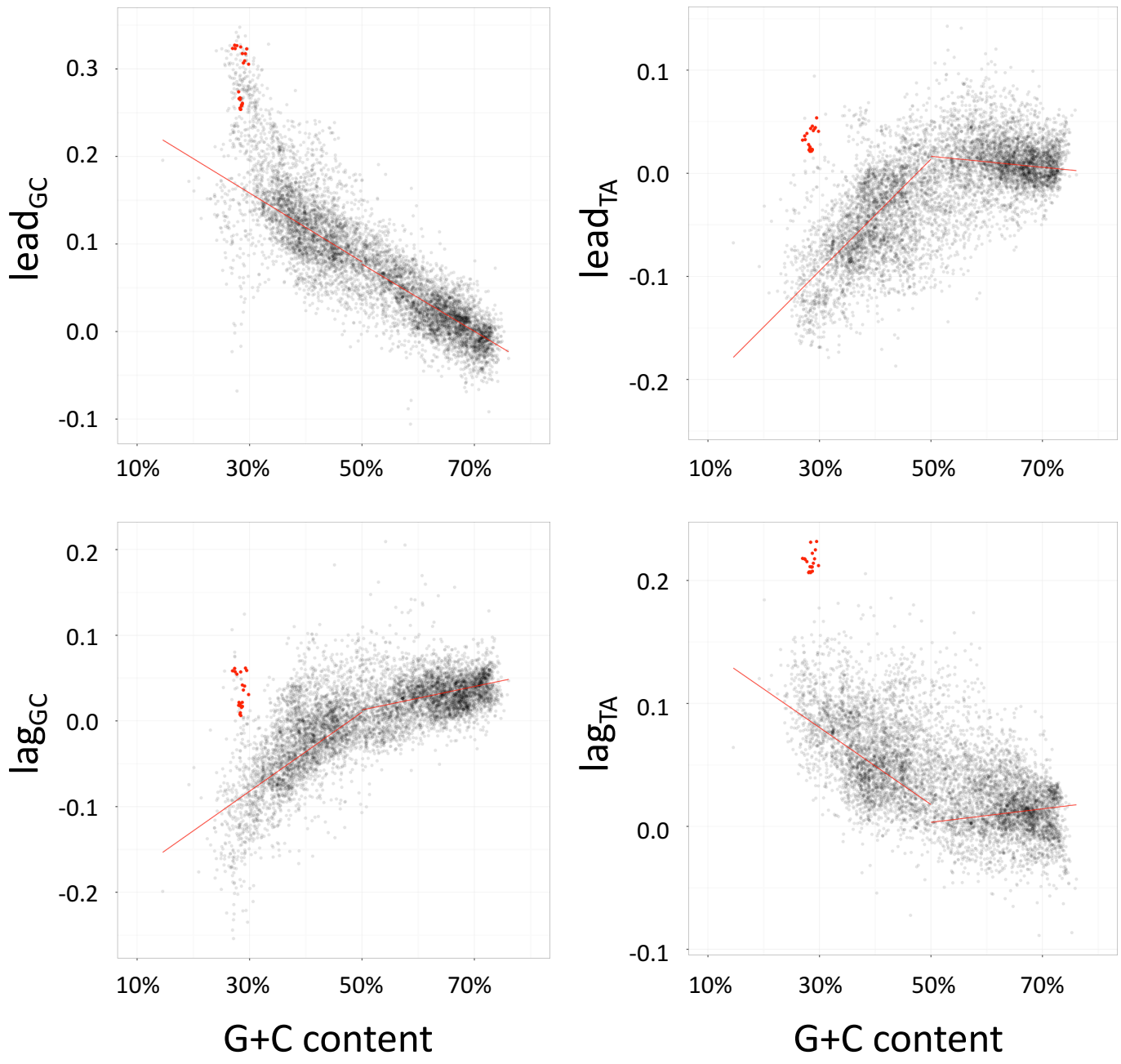
lead$_{GC}$
0.3
0.2
0.1
0.0
-0.1
10%
30%
50%
70%
lead$_{TA}$
0.1
0.0
-0.1
-0.2
10%
30%
50%
70%
lag$_{GC}$
0.2
0.1
0.0
-0.1
-0.2
10%
30%
50%
70%
G+C content
lag$_{TA}$
0.2
0.1
0.0
-0.1
10%
30%
50%
70%
G+C content

Figure 4

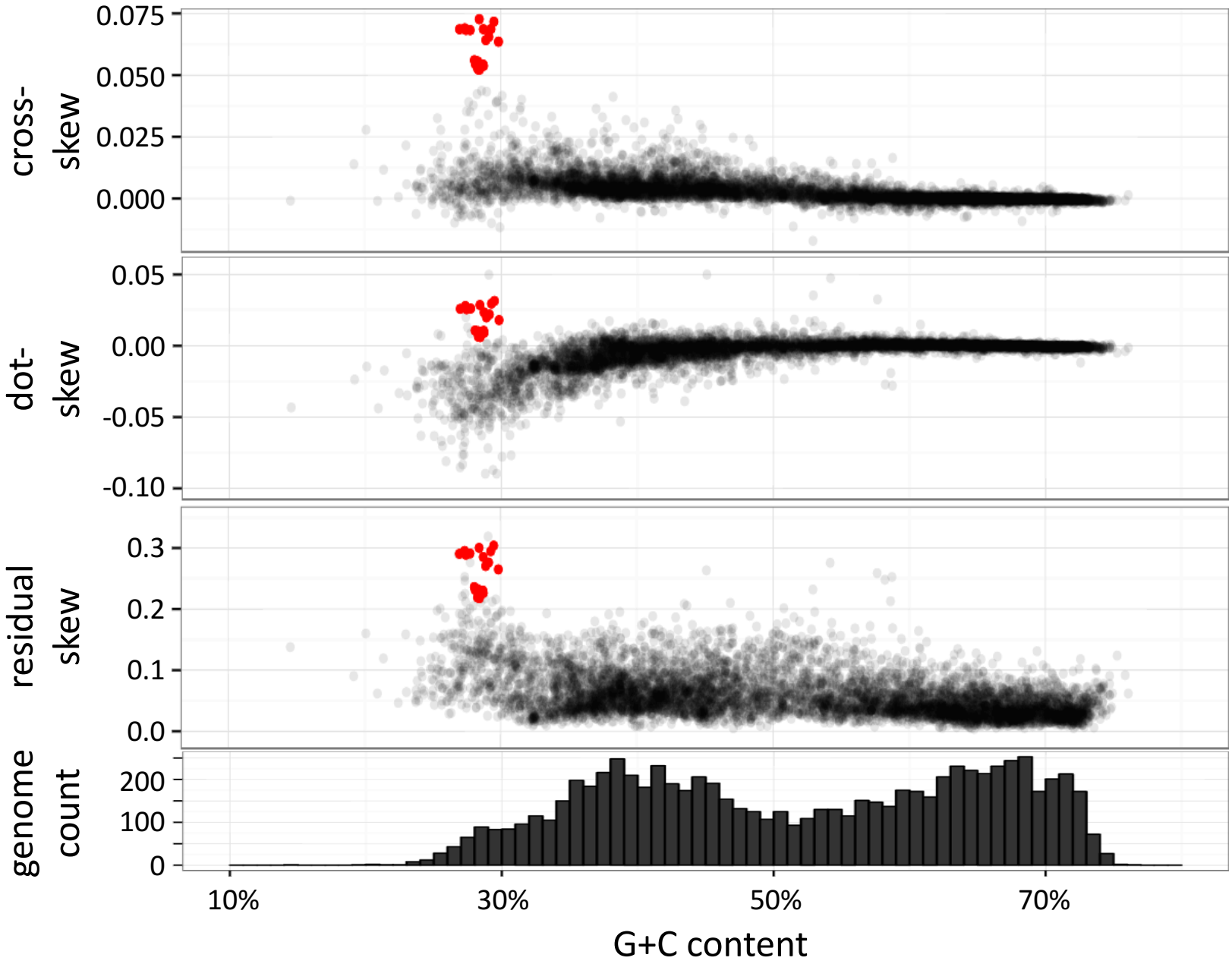
cross-skew
0.075
0.050
0.025
0.000
dot-skew
0.05
0.00
-0.05
-0.10
residual skew
0.3
0.2
0.1
0.0
genome count
200
100
0
10%
30%
50%
70%
G+C content

Figure 5

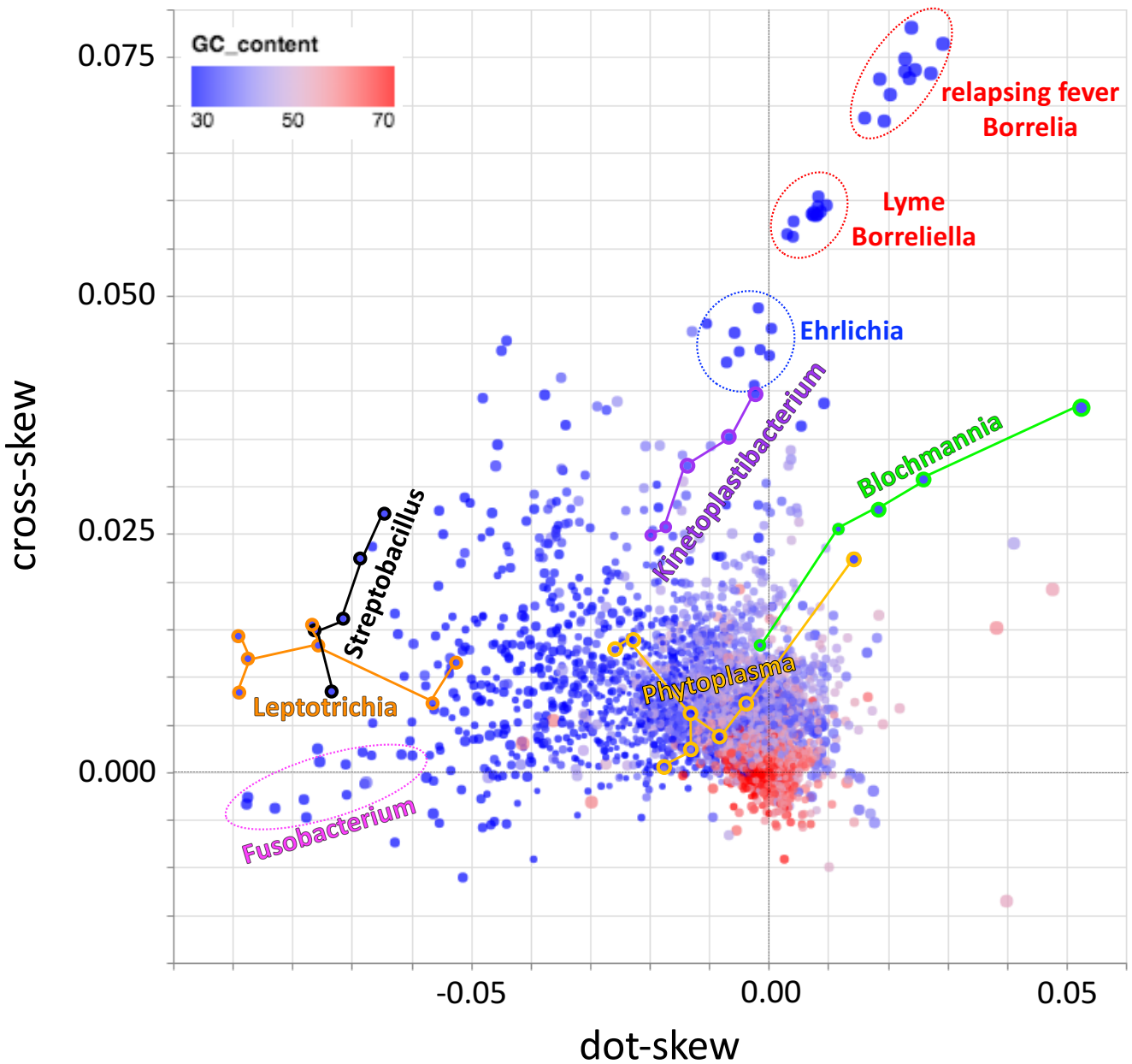
GC_content
30
50
70
0.075
0.050
0.025
0.000
cross-skew
-0.05
0.00
0.05
dot-skew
relapsing fever
Borrelia
Lyme
Borreliella
Ehrlichia
Kinetoplastibacterium
Blochmannia
Streptobacillus
Phytoplasma
Leptotrichia
Fusobacterium

Figure 6