Novel metrics for quantifying bacterial genome composition skews

Lena M. Joesch-Cohen, Max Robinson, Neda Jabbari, Christopher Lausted, Gustavo Glusman

Institute for Systems Biology, 401 Terry Ave N, Seattle, WA 98109, USA

* Correspondence: Gustavo@SystemsBiology.org

Gustavo Glusman
Institute for Systems Biology
401 Terry Ave N
Seattle, WA 98109, USA
Tel: 206 732-1273

Abstract

We present three novel metrics for quantifying bacterial genome composition skews. Skews are asymmetries in nucleotide usage that arise as a result of mutational biases and selective constraints, particularly for energy efficiency. The first two metrics (dot product and cross product of average skew vectors) evaluate sequence and gene annotation of the genome of a single species, while the third metric (regression RMSD) discovers patterns only discernable from studying genomes of thousands of species. The three metrics can be computed for genomes not yet finished and fully annotated. We studied the genomes of 7738 bacterial species, including completed genomes and partial drafts, and identified multiple species with unusual skew parameters. A number of these outliers (i.e., Borrelia, Ehrlichia, Kinetoplastibacterium, and Phytoplasma) display similar skew patterns despite a lack of phylogenetic relation. These disparate bacterial species share lifestyle characteristics, suggesting that our novel metrics successfully capture effects on genome composition of biosynthetic constraints and of interaction with the hosts.

Introduction

Bacterial genomes display significant compositional biases, both in terms of G+C content and in skews (strand asymmetry in ‘T’ vs. ‘A’ and ‘G’ vs. ‘C’ usage). These biases arise from the complex interplay of differential mutation rates and multiple selective constraints (Morton and Morton 2007; Vetsigian and Goldenfeld 2009), particularly for energy efficiency (Chen et al. 2016). 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. 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”). These opposite combinations can lead to distinctive skew patterns; their study can reveal details of the multiple compositional constraints and their interactions, and ultimately inform about the DNA repair capacity, the metabolism and the lifestyle of the species (Dutta and Paul 2012).

Extreme examples of compositional biases are found among species in the family Borreliaceae, which comprises a variety of tick-borne spirochetes and includes species causing Lyme disease (genus *Borreliella*, originally *Borrelia*) as well as those causing relapsing fever (genus *Borrelia*) (Cutler, Ruzic-Sabljic, and Potkonjak 2017). Since its discovery in 1982 (Burgdorfer et al. 1982), the *Borreliella burgdorferi* spirochete has been of particular interest in the United States as the primary causative agent of Lyme disease. The subsequent sequencing of *B. burgdorferi* in 1997, has allowed for an in depth exploration of the many intriguing features of the genome of this bacterium, from is unusual structure (one large linear chromosome, several linear and circular plasmids) to its relatively low G+C content (Fraser et al. 1997). There is some evidence that smaller genomes tend to have lower G+C content than larger ones (Foerstner et al. 2005). Obligate intracellular organisms also tend to have low G+C content (Rocha 2004; Dutta and Paul 2012).

A number of studies have reported unusually high nucleotide skews in *B. burgdorferi*, with increased levels of ‘G’ and ‘T’ in third codon positions on the leading strand and increased ‘C’ and ‘A’ on the lagging strand (Mackiewicza et al. 1999; McInerney 1998). One study in particular (Lobry and Sueoka 2002) found *B. burgdorferi* to have the most extreme difference between leading and lagging strand skews among the 43 genomes investigated. Several hypotheses have been put forward for the high skew seen in *B. burgdorferi’s* high skews; both mutation and selection biases may play a role (Hildebrand, Meyer, and Eyre-Walker 2010; Wei and Guo 2010), variously induced by replication, transcription and translation constraints. Furthermore, the possible loss of some DNA repair genes may contribute to the low G+C content and heightened skew seen in *B. burgdorferi* (Zhao et al. 2015; Lind and Andersson 2008).

Multiple methods have been proposed for quantifying compositional biases and skews, ranging from simple computation in fixed-size windows along the genome, to very detailed stratification of nucleotides by direction of transcription and replication, codon position, codon adaptation index, and more. Some studies use the relative synonymous codon usage (RSCU) as a measure for skew, which focuses exclusively on the 3rd codon position. A recent study introduced several other metrics of qualifying and quantifying genome skew (Wei and Guo 2010); they computed Z curves (R. Zhang and Zhang 1994) using nucleotide counts and biases, and defined nine metrics on which to perform correspondence analysis (COA). COA was also performed on absolute codon counts and on the commonly used metric of RSCU. In combination, these metrics revealed significant differences in skew between genes transcribed on the leading and lagging strands.

Thanks to the much expanded availability of complete genome sequences of bacterial species, it is now possible to perform large-scale comparative genomics

studies (Morton and Morton 2007; Necşulea and Lobry 2007; Chen et al. 2016). A much larger number of bacterial genomes have been drafted, 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 drafted, incomplete genomes.

We present here three novel metrics for quantitative analysis of genome skews. Our metrics are resilient 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 drafted. 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*.

Materials and Methods

Genomes studied

We obtained from NCBI the genome sequence (in FASTA format) and current annotation (in GFF format) 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.

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 (Ren Zhang and Zhang 2005), 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 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 the average skews (GC and TA) for the set as the average of the corresponding individual segment skews, weighted by segment length. This yields four characteristic skews: $lead_{GC}$, $lead_{TA}$, lag_{GC} and lag_{TA}. We also evaluated using weighted medians; this yielded very similar results (not shown).

Computation of the skew cross product and dot product metrics

The set of 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 skew cross product metric as:

$$\text{cross_product}(\text{lead}, \text{lag}) = |\text{lead}| \bullet |\text{lag}| \bullet \sin(\theta) \quad (1)$$

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

$$\text{dot_product}(\text{lead}, \text{lag}) = |\text{lead}| \bullet |\text{lag}| \bullet \cos(\theta) \quad (2)$$

Computation of the skew deviation metric

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 using the least trimmed sum of squares as implemented in the R function MASS::lqs() (Venables and Ripley 2002). We then computed a single skew deviation

summary metric 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 genomic regions have either a strong GC skew or a strong TA skew (Fig. 1A). Some 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 anticorrelated. When normalizing the direction of the skews to be relative to the leading 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 (Picardeau, Lobry, and Hinnebusch 2000).

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 the figure), the two skews tend to be positively correlated.

The characteristic skews of *B. burgdorferi*

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). These two clusters can be represented by their average TA and GC skews, weighted by gene length (see Methods). We thus computed the four characteristic skews for *B. burgdorferi*: $lead_{GC}$ = 0.2590, $lead_{TA}$ = 0.0215, lag_{GC} = 0.0161 and lag_{TA} = 0.2110.

We visualized these skews as two vectors and computed the angle between them $\theta = 80.89°$ (Fig. 2, inset). Then, based on the length and angles of these two vectors, we computed the skew cross product and dot product metrics for *B. burgdorferi*: cross_product(lead, lag) = 0.0584, dot_product(lead, lag) = 0.0075. For other Borrelia and Borreliella species, 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: http://db.systemsbiology.net/gestalt/cgi-pub/skewSegmentPlot.pl.

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 thus fitted robust using the least quantile of squares method separately for bacterial genomes below and above 50% G+C content, and computed the deviations from the expected skews for each bacterial genome assembly.

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. 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). The deviations of characteristic skews for *B. burgdorferi* from the multi-genome fit 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. Thus, Borreliaceae genomes are unusual for all four characteristic skews.

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 skew dot product and the skew cross product (see Methods). Furthermore, the availability of many thousand bacterial genome assemblies allowed us to compute for each the expected values for each characteristic skew, as a function of the G+C content. We used the observed deviations from these expected values to define a third metric - the regression root mean squared deviation (RMSD).

We computed these three metrics for 7738 bacterial genome assemblies and evaluated their relationship with G+C content (Fig. 5). For high G+C content bacteria, we observed that the dot product and cross product metrics 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 regression RMSD metric 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 Figure 3) are also outliers in all three metrics and show similar skew values as Borreliella genomes. Ehrlichia are intracellular vector-borne pathogens of vertebrates (Dunning Hotopp et al. 2006); like Borrelia, they have diminished biosynthetic abilities. 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 (Mavromatis et al. 2006). Multiple other genera became evident as outliers of interest, discussed below. We provide a table of skew characteristics for the 7738 bacterial genomes Supplemental Table 1.

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 (dot product and cross product) 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 values of the dot product metric (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* (Kalman et al. 1999) (Fig. 3). Conversely, negative dot product values (Fig. 6, bottom) reflect opposite compositional constraints on leading and lagging strand genes; extreme examples of this pattern are observed in fusobacteria including *Fusobacterium periodonticum* (Slots, Potts, and Mashimo 1983), *Leptotrichia buccalis* (Ivanova et al. 2009), and *Streptobacillus moniliformis* (Nolan et al. 2009), the causal agent of rat bite fever. Positive dot product values can thus be interpreted as reflecting constraints driven mostly by the replication process, while negative dot product values largely reflect transcriptional and translational constraints. The cross product metric quantifies the strength and orthogonality of the compositional skew vectors for leading and lagging strand genes. Genomes with high values of the cross product metric (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 (Chen et al. 2016), 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 (Alves et al. 2013) 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 (Gil et al. 2003).

The third metric (regression RMDS) capitalizes on the current availability of thousands of complete or drafted 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 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 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 (Tran-Nguyen et al. 2008; Hogenhout et al. 2008), in similarity to Borreliaceae and Ehrlichia for mammals.

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 RMSD metric) Phytoplasma. These very disparate bacterial species share lifestyle characteristics, suggesting that our novel metrics successfully capture effects on genome composition of biosynthetic constraints and of interaction with the hosts.

Acknowledgements

We wish to thank Arian Smit and Jeff Boore for helpful discussions. This work was supported by generous donations from The Wilke Family Foundation, Jeff and MacKenzie Bezos, and The Steven & Alexandra Cohen Foundation.

References

Alves, João M. P., Myrna G. Serrano, Flávia Maia da Silva, Logan J. Voegtly, Andrey V. Matveyev, Marta M. G. Teixeira, Erney P. Camargo, and Gregory A. Buck. 2013. “Genome Evolution and Phylogenomic Analysis of Candidatus Kinetoplastibacterium, the Betaproteobacterial Endosymbionts of Strigomonas and Angomonas.” *Genome Biology and Evolution* 5 (2): 338–50.

Burgdorfer, W., A. G. Barbour, S. F. Hayes, J. L. Benach, E. Grunwaldt, and J. P. Davis. 1982. “Lyme Disease-a Tick-Borne Spirochetosis?” *Science* 216 (4552): 1317–19.

Chen, Wei-Hua, Guanting Lu, Peer Bork, Songnian Hu, and Martin J. Lercher. 2016. “Energy Efficiency Trade-Offs Drive Nucleotide Usage in Transcribed Regions.” *Nature Communications* 7 (April): 11334.

Cutler, Sally J., Eva Ruzic-Sabljic, and Aleksandar Potkonjak. 2017. “Emerging Borreliae - Expanding beyond Lyme Borreliosis.” *Molecular and Cellular Probes* 31 (February): 22–27.

Dunning Hotopp, Julie C., Mingqun Lin, Ramana Madupu, Jonathan Crabtree, Samuel V. Angiuoli, Jonathan A. Eisen, Jonathan Eisen, et al. 2006. “Comparative Genomics of Emerging Human Ehrlichiosis Agents.” *PLoS Genetics* 2 (2): e21.

Dutta, Chitra, and Sandip Paul. 2012. “Microbial Lifestyle and Genome Signatures.” *Current Genomics* 13 (2): 153–62.

Foerstner, Konrad U., Christian von Mering, Sean D. Hooper, and Peer Bork. 2005. “Environments Shape the Nucleotide Composition of Genomes.” *EMBO Reports* 6 (12): 1208–13.

Fraser, C. M., S. Casjens, W. M. Huang, G. G. Sutton, R. Clayton, R. Lathigra, O. White, et al. 1997. “Genomic Sequence of a Lyme Disease Spirochaete, Borrelia Burgdorferi.” *Nature* 390 (6660): 580–86.

Gil, Rosario, Francisco J. Silva, Evelyn Zientz, François Delmotte, Fernando González-Candelas, Amparo Latorre, Carolina Rausell, et al. 2003. “The Genome Sequence of Blochmannia Floridanus: Comparative Analysis of Reduced Genomes.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (16): 9388–93.

Hildebrand, Falk, Axel Meyer, and Adam Eyre-Walker. 2010. “Evidence of Selection upon Genomic GC-Content in Bacteria.” *PLoS Genetics* 6 (9): e1001107.

Hogenhout, Saskia A., Kenro Oshima, El-Desouky Ammar, Shigeyuki Kakizawa, Heather N. Kingdom, and Shigetou Namba. 2008. “Phytoplasmas: Bacteria That Manipulate Plants and Insects.” *Molecular Plant Pathology* 9 (4): 403–23.

Ivanova, Natalia, Sabine Gronow, Alla Lapidus, Alex Copeland, Tijana Glavina Del Rio, Matt Nolan, Susan Lucas, et al. 2009. “Complete Genome Sequence of Leptotrichia Buccalis Type Strain (C-1013-B).” *Standards in Genomic Sciences* 1 (2): 126–32.

Kalman, S., W. Mitchell, R. Marathe, C. Lammel, J. Fan, R. W. Hyman, L. Olinger, J. Grimwood, R. W. Davis, and R. S. Stephens. 1999. “Comparative Genomes of Chlamydia Pneumoniae and C. Trachomatis.” *Nature Genetics* 21 (4): 385–89.

Lind, Peter A., and Dan I. Andersson. 2008. “Whole-Genome Mutational Biases in Bacteria.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (46): 17878–83.

Lobry, Jean R., and Noboru Sueoka. 2002. “Asymmetric Directional Mutation Pressures in Bacteria.” *Genome Biology* 3 (10): RESEARCH0058.

Mackiewicza, P., A. Gierlika, M. Kowalczuka, D. Szczepanika, M. R. Dudek, and S. and Cebrat. 1999. “Mechanisms Generating Long-Range Correlation in Nucleotide Composition of the Borrelia Burgdorferi Genome.” *Physica A: Statistical Mechanics and Its Applications* 273 (1-2): 103–15.

Mavromatis, K., C. Kuyler Doyle, A. Lykidis, N. Ivanova, M. P. Francino, P. Chain, M. Shin, et al. 2006. “The Genome of the Obligately Intracellular Bacterium Ehrlichia Canis Reveals Themes of Complex Membrane Structure and Immune Evasion Strategies.” *Journal of Bacteriology* 188 (11): 4015–23.

McInerney, J. O. 1998. “Replicational and Transcriptional Selection on Codon Usage in Borrelia Burgdorferi.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (18): 10698–703.

Morton, Richard A., and Brian R. Morton. 2007. “Separating the Effects of Mutation and Selection in Producing DNA Skew in Bacterial Chromosomes.” *BMC Genomics* 8 (October): 369.

Necşulea, Anamaria, and Jean R. Lobry. 2007. “A New Method for Assessing the Effect of Replication on DNA Base Composition Asymmetry.” *Molecular Biology and Evolution* 24 (10): 2169–79.

Nolan, Matt, Sabine Gronow, Alla Lapidus, Natalia Ivanova, Alex Copeland, Susan Lucas, Tijana Glavina Del Rio, et al. 2009. “Complete Genome Sequence of Streptobacillus Moniliformis Type Strain (9901).” *Standards in Genomic Sciences* 1 (3): 300–307.

Picardeau, M., J. R. Lobry, and B. J. Hinnebusch. 2000. “Analyzing DNA Strand Compositional Asymmetry to Identify Candidate Replication Origins of Borrelia Burgdorferi Linear and Circular Plasmids.” *Genome Research* 10 (10): 1594–1604.

Rocha, Eduardo P. C. 2004. “The Replication-Related Organization of Bacterial Genomes.” *Microbiology* 150 (Pt 6): 1609–27.

Slots, J., T. V. Potts, and P. A. Mashimo. 1983. “Fusobacterium Periodonticum, a New Species from the Human Oral Cavity.” *Journal of Dental Research* 62 (9): 960–63.

Tran-Nguyen, L. T. T., M. Kube, B. Schneider, R. Reinhardt, and K. S. Gibb. 2008. “Comparative Genome Analysis of ‘Candidatus Phytoplasma Australiense’ (subgroup Tuf-Australia I; Rp-A) and ‘Ca. Phytoplasma Asteris’ Strains OY-M and AY-WB.” *Journal of Bacteriology* 190 (11): 3979–91.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S, 4th Ed*. Springer.

Vetsigian, Kalin, and Nigel Goldenfeld. 2009. “Genome Rhetoric and the Emergence of Compositional Bias.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (1): 215–20.

Wei, Wen, and Feng-Biao Guo. 2010. “Strong Strand Composition Bias in the Genome of Ehrlichia Canis Revealed by Multiple Methods.” *The Open Microbiology Journal* 4 (October): 98–102.

Zhang, Ren, and Chun-Ting Zhang. 2005. “Identification of Replication Origins in Archaeal Genomes Based on the Z-Curve Method.” *Archaea* 1 (5): 335–46.

Zhang, R., and C. T. Zhang. 1994. “Z Curves, an Intutive Tool for Visualizing and Analyzing the DNA Sequences.” *Journal of Biomolecular Structure & Dynamics* 11 (4): 767–82.

Zhao, Hai-Long, Zhong-Kui Xia, Fa-Zhan Zhang, Yuan-Nong Ye, and Feng-Biao Guo. 2015. “Multiple Factors Drive Replicating Strand Composition Bias in Bacterial Genomes.” *International Journal of Molecular Sciences* 16 (9): 23111–26.

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. 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. Skews with absolute values between 0.5 and 1 shown in compressed scale for clarity.

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. Lower-right inset for each plot: skew cross-product value 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 product, dot product, regression RMDS and histogram of number of species studied.

Figure 6. Integration of skew metrics (dot product vs. cross product, point size represents regression RMSD) for 7738 bacterial genomes, highlighting some genera of interest. All other genomes colored by G+C content: cyan for G+C<50%, pink for G+C>=50%.

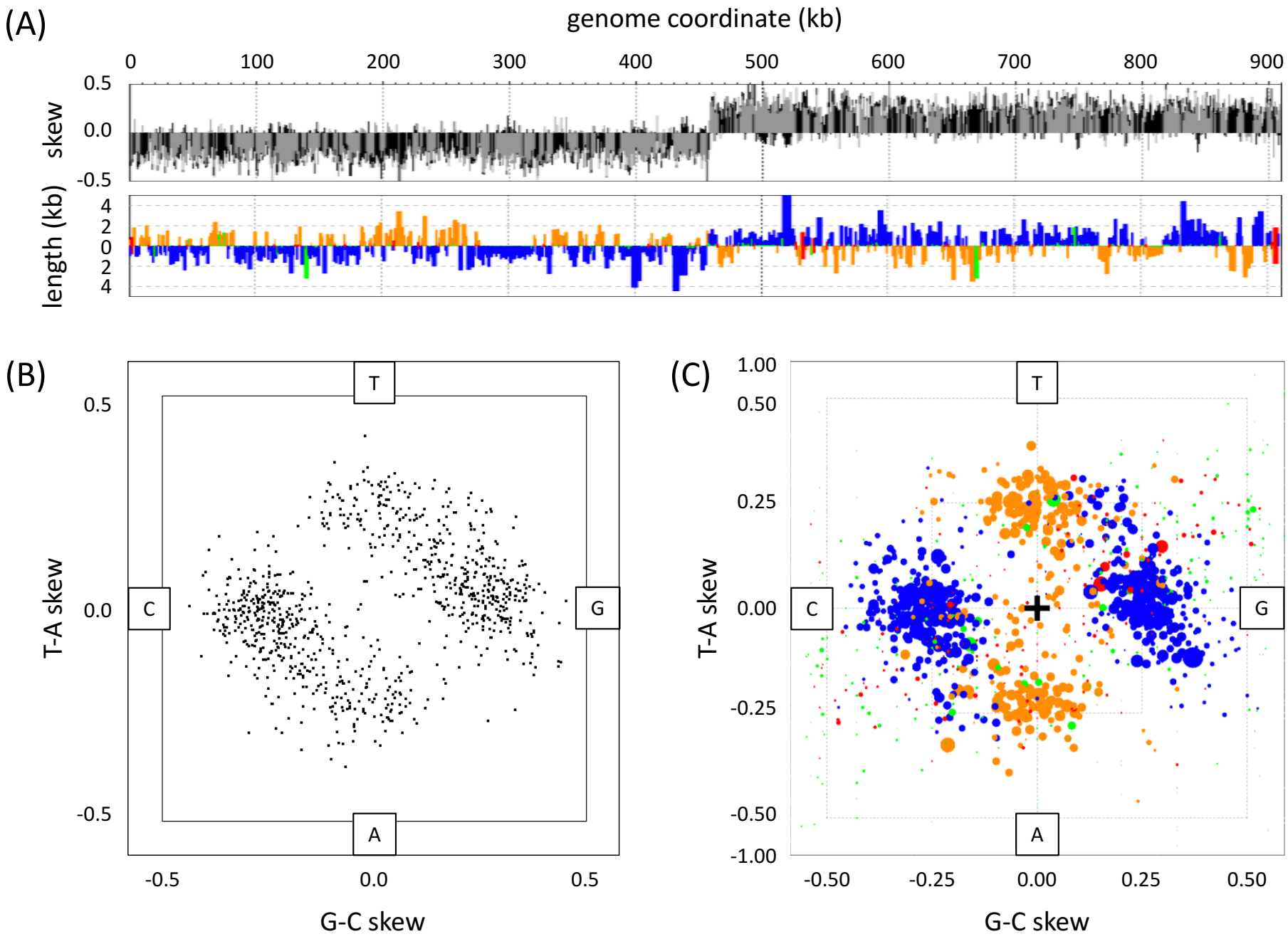
(A)
genome coordinate (kb)
0
100
200
300
400
500
600
700
800
900
skew
0.5
0.0
-0.5
length (kb)
4
2
0
2
4
(B)
T
C
G
A
0.5
0.0
-0.5
T-A skew
-0.5
0.0
0.5
G-C skew
(C)
T
C
G
A
1.00
0.50
0.25
0.00
-0.25
-0.50
-1.00
T-A skew
-0.50
-0.25
0.00
0.25
0.50
G-C skew

Figure 1

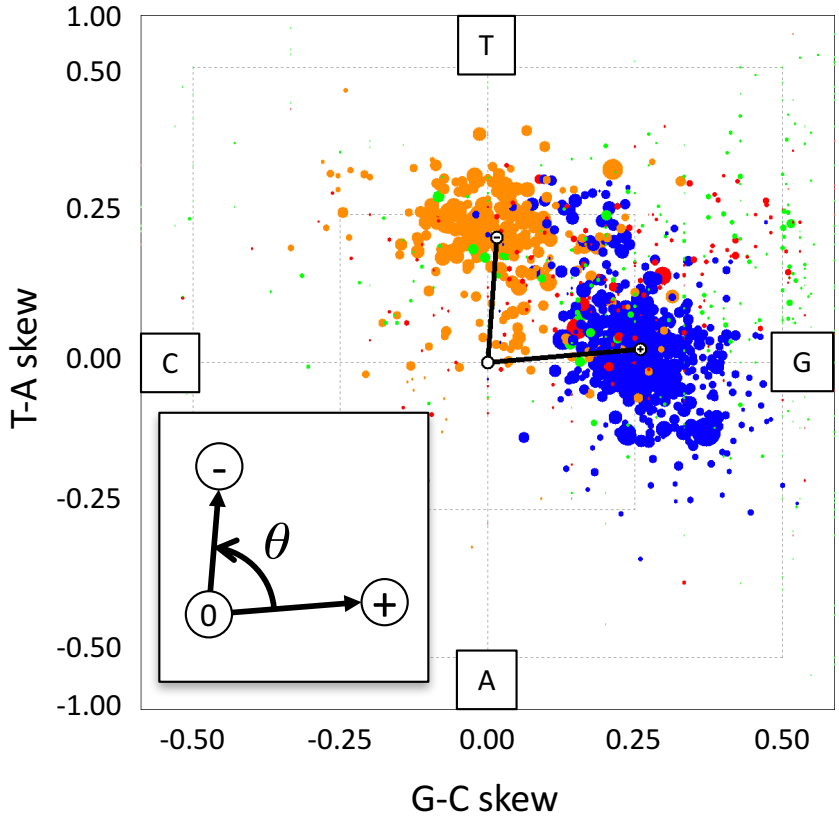
1.00
0.50
0.25
0.00
-0.25
-0.50
-1.00
T-A skew
-0.50
-0.25
0.00
0.25
0.50
G-C skew
T
C
G
A
-
0
+
θ

Figure 2

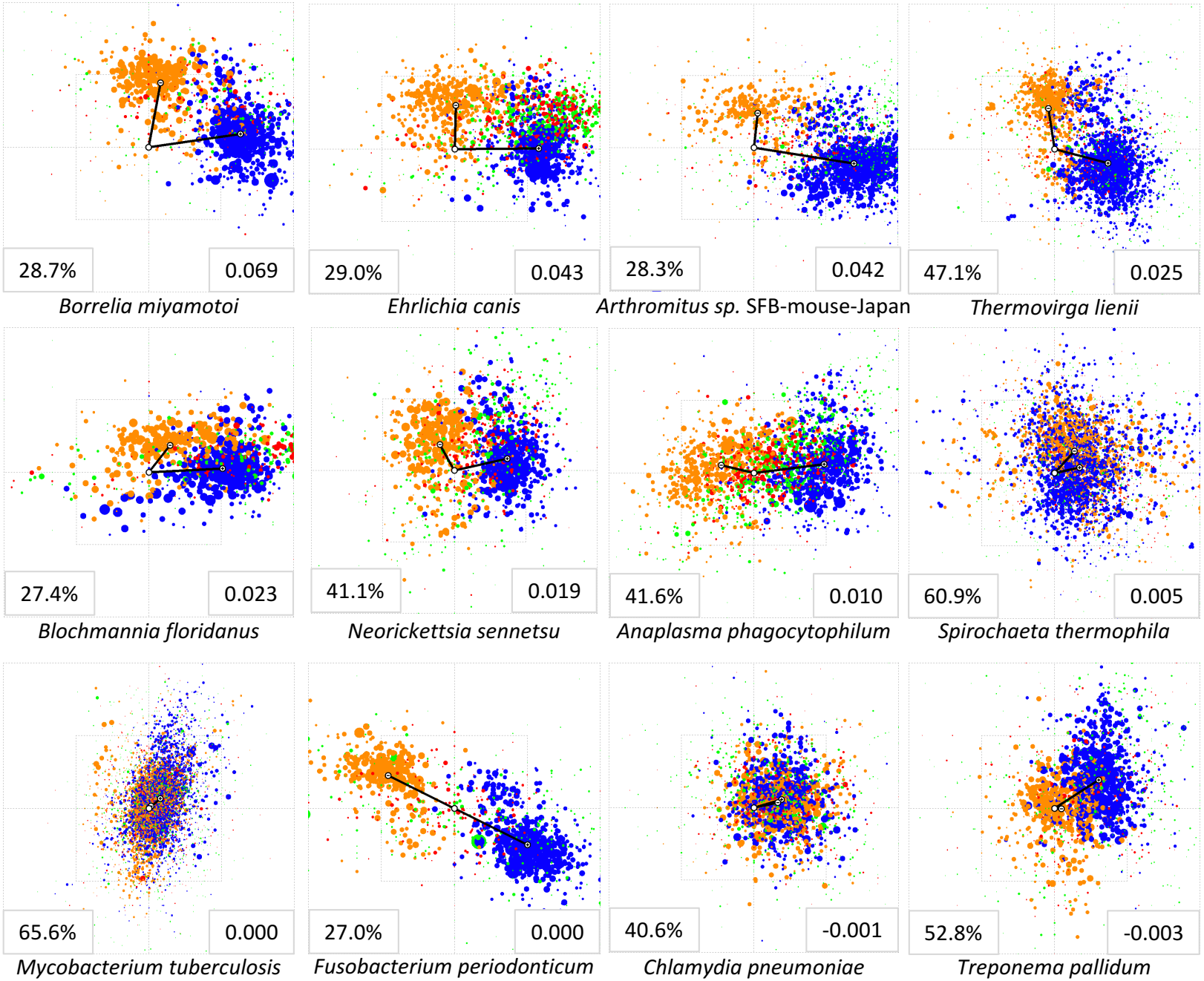
28.7%
0.069
Borrelia miyamotoi
29.0%
0.043
Ehrlichia canis
28.3%
0.042
Arthromitus sp. SFB-mouse-Japan
47.1%
0.025
Thermovirga lienii
27.4%
0.023
Blochmannia floridanus
41.1%
0.019
Neorickettsia sennetsu
41.6%
0.010
Anaplasma phagocytophilum
60.9%
0.005
Spirochaeta thermophila
65.6%
0.000
Mycobacterium tuberculosis
27.0%
0.000
Fusobacterium periodonticum
40.6%
-0.001
Chlamydia pneumoniae
52.8%
-0.003
Treponema pallidum

Figure 3

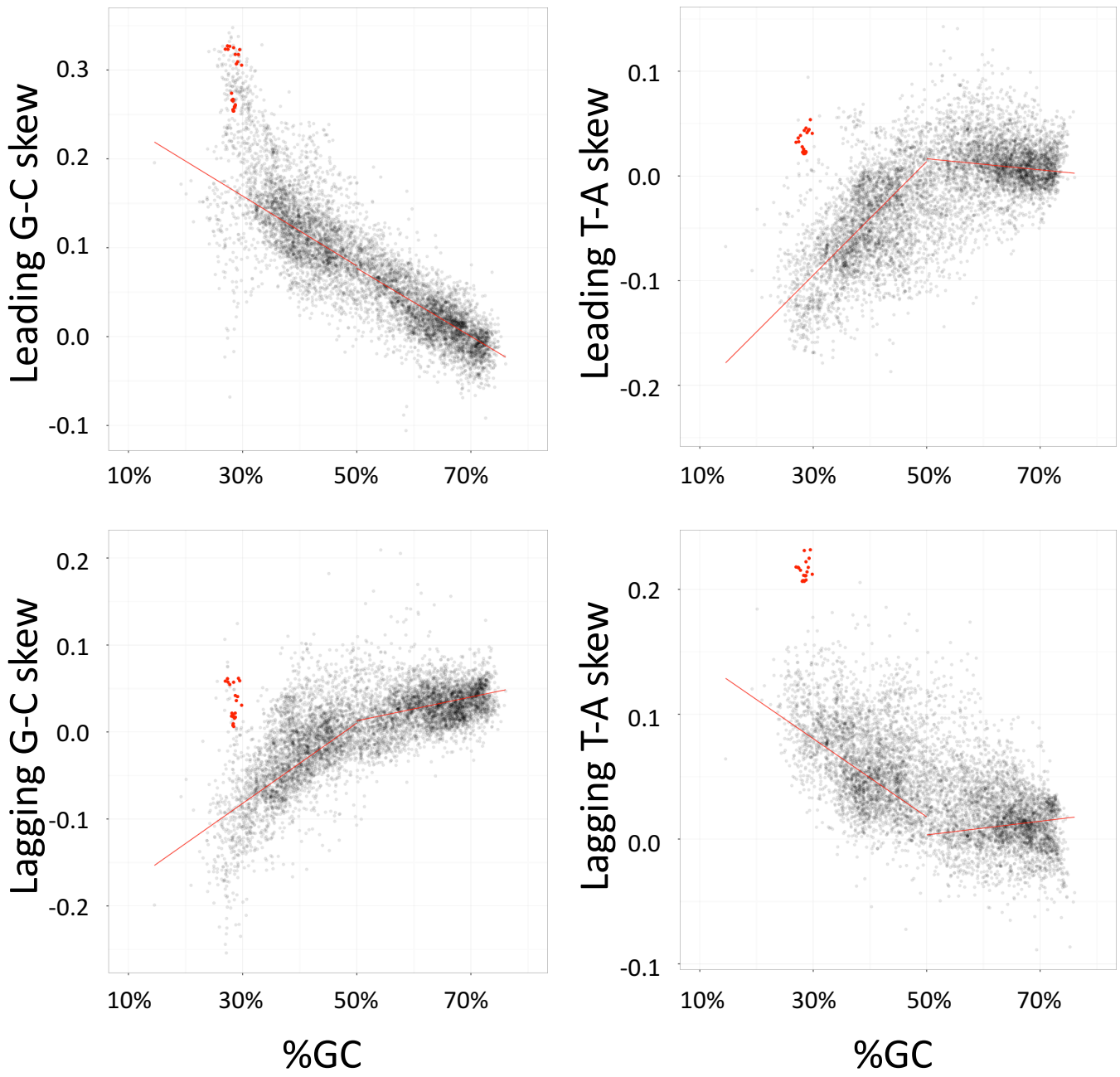
Leading G-C skew
0.3
0.2
0.1
0.0
-0.1
10%
30%
50%
70%
Leading T-A skew
0.1
0.0
-0.1
-0.2
10%
30%
50%
70%
Lagging G-C skew
0.2
0.1
0.0
-0.1
-0.2
10%
30%
50%
70%
%GC
Lagging T-A skew
0.2
0.1
0.0
-0.1
10%
30%
50%
70%
%GC

Figure 4

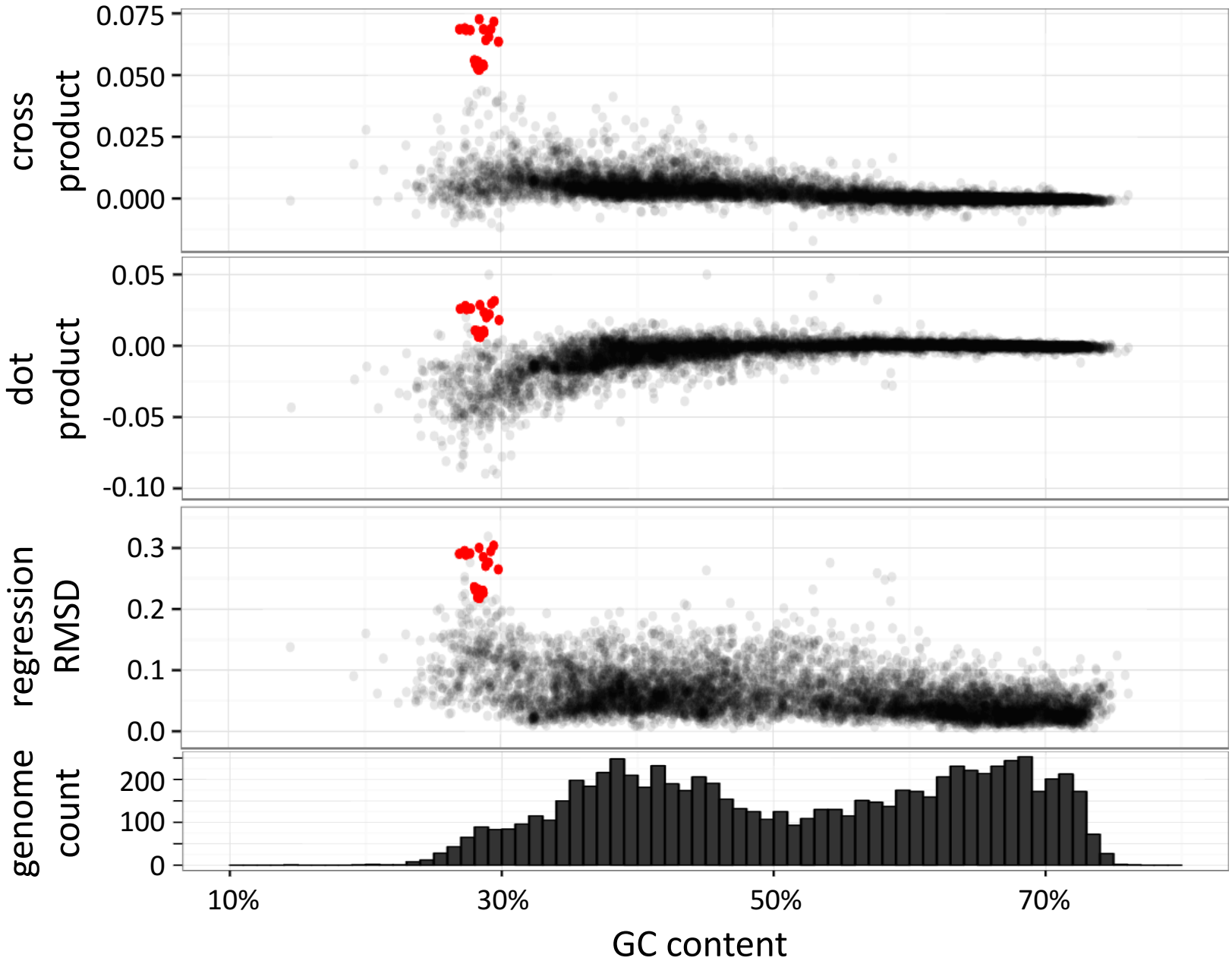
cross product
0.075
0.050
0.025
0.000
dot product
0.05
0.00
-0.05
-0.10
regression RMSD
0.3
0.2
0.1
0.0
genome count
200
100
0
10%
30%
50%
70%
GC content

Figure 5

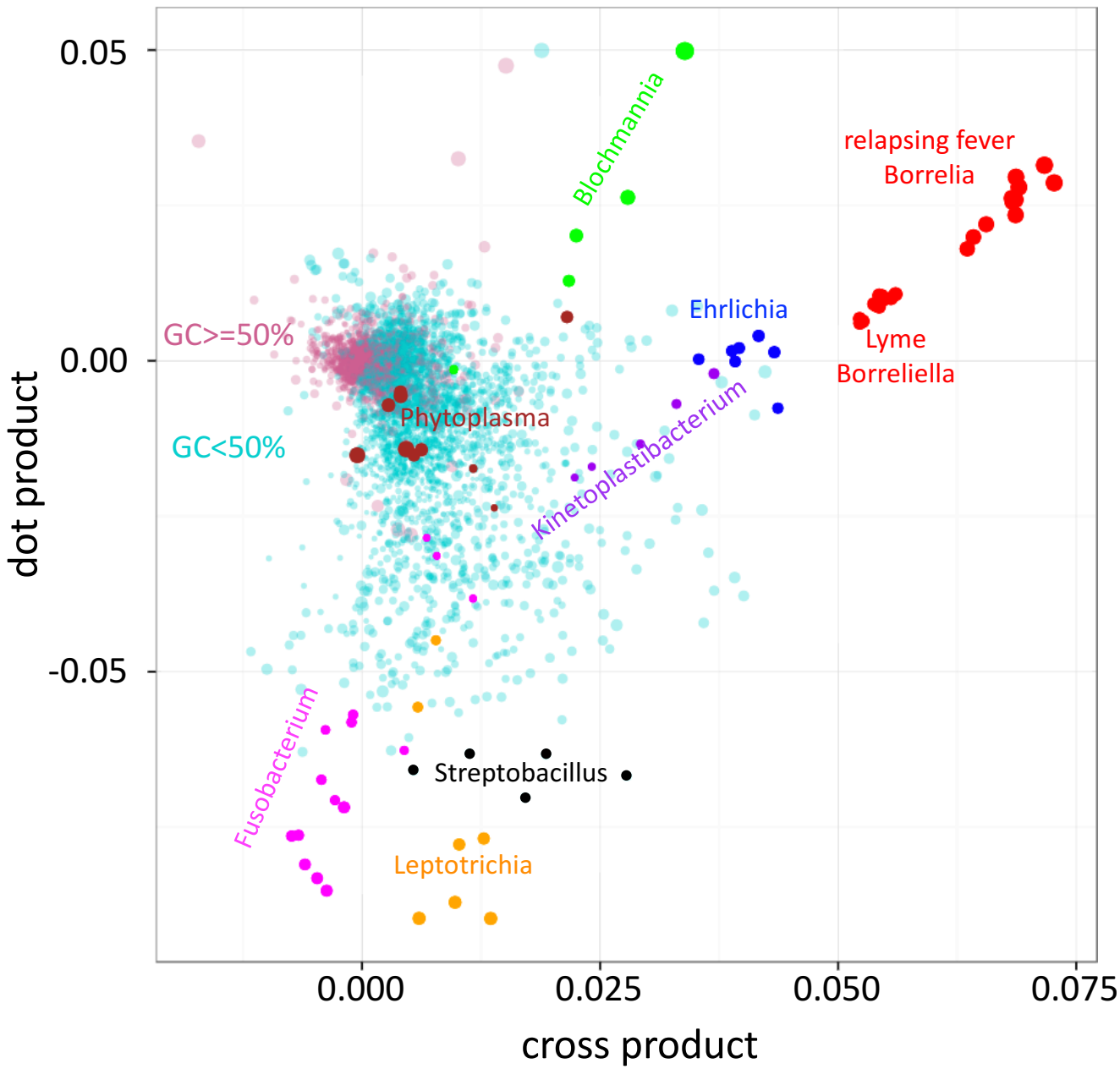
0.05
0.00
-0.05
dot product
0.000
0.025
0.050
0.075
cross product
Blochmannia
relapsing fever
Borrelia
Ehrlichia
Lyme
Borreliella
GC>=50%
GC<50%
Phytoplasma
Kinetoplastibacterium
Fusobacterium
Streptobacillus
Leptotrichia

Figure 6