What we talk about when we talk about species

Running title: Compressing genomes to understand their evolution

Apurva Narechania[1,2], Shyam Gopalakrishnan[2], M Thomas P Gilbert[2,3]

[1]Institute for Comparative Genomics, American Museum of Natural History, New York, NY, USA
[2]Center for Evolutionary Hologenomics, The Globe Institute, University of Copenhagen, Copenhagen, Denmark
[3]University Museum, NTNU, Trondheim, Norway
Corresponding authors: anarechania@amnh.org; tgilbert@sund.ku.dk

Keywords: Compression, information theory, pangenomes

Abstract

Genome annotation, alignment, and phylogenetics are at the center of most work in evolutionary genomics. These techniques function best when rooted in prior work. Genes are mined from new genomes using evidence from old gene models. These genomes are aligned to well-worn references to create matrices for tree reconstruction. And trees are often populated with well characterized genomes to add context to the newly sequenced. Genome inference traces a line back to model organisms, yoking the analysis of new genomes to layers of previous knowledge. We instead highlight methods that use unannotated and unaligned sequence to understand the information diversity of sequence ensembles. Any set of genomes can comprise our sequence ensemble. In a pandemic context, a sequence ensemble might be clinically isolated strains from one day. In a systematic context, a sequence ensemble could be the pangenome available for a clade. The normal bioinformatics playbook would have us align. But we instead compress. A sequence ensemble that compresses easily contains lower information diversity. For pandemics, we can use curves of information diversity to trace genomic novelty and monitor selective sweeps in new strains. For systematics, we can calculate compressibility quickly across all known bacterial taxa, leveling the criteria for species across clades. If we tolerate data loss, we can go one step further and capture structural evolution as we compress. Our approach sacrifices a lot. We skip many of the products of modern bioinformatics like variation anchored to known genes or genome alignment to prescribed references or pangenome graphs. But we gain speed, breadth, and the ability to respond to novelty.

Introduction (The problem)

Compression encodes information into reduced representations. Whether bits are eliminated through statistical redundancy (lossless compression), or shed entirely (lossy compression), compressed data always has a smaller footprint than the original. The act of compression – its difficulty or ease – communicates information about the original data source.

Highly redundant data with many common patterns will compress easily. In contrast, novelty or surprise with little repeated context is difficult to compress. Evolution creates ensembles of sequence. These ensembles can be represented as pangenomes. Pangenomes are compressible entities, but how compressible depends on evolutionary strategy.

Genomics is a retrospective field. Existing bioinformatic techniques often model new genomes on sequences annotated in the past[1]. Alignment to these reference genomes circumscribes our knowledge of diversity. Large swaths of the tree of life are presumably unknown[2]. For example, much of the sequence from environmental samples passes through annotation filters as undefined[3]. In a read streaming era[4], we need forward looking techniques that flag genomic novelty by dispensing with references, annotation or alignment. Standard methods are ill-equipped for these volumes. New species are not easily caught in the sparse web of the known.

As genomics has swept through biology, systematics has come to favor molecular character sets to help delimit species boundaries[5,6]. While morphology is still important, and holdouts have been more than vocal[7], phylogenomics has more recently carried the day. Phylogenomics extends the handful of marker genes that were the foundation of early molecular systematics to matrices that concatenate thousands of orthologous genes[8]. This character explosion has been a boon to systematics, but annotation is still anchored to the known.

The thousands of orthologous genes found in phylogenomic datasets are rarely evenly distributed among the genomes that describe a species[9]. The complete set of these genes is one definition of the pangenome, and its complexity was originally defined as the rate of gene accumulation with newly sequenced genomes[10]. Genes found universally comprise a genomic core and are considered indispensable for core species functions. Genes found sporadically may contribute to strain success in particular niches but may not be essential to their overall biology. The ratio of core genomes to accessory genomes informs genome fluidity[11]. Species whose genomes are mostly core have closed, less fluid pangenomes. Species with a large fraction of accessory genes are considered open and more fluid.

This gene-centric view of orthologs is blind to the diversity in the non-coding genome[12]. Whole genome alignment to annotated, chromosomal references[13,14] makes variation in non-coding genome accessible but again circumscribes its characterization. If all we know is a linear reference on a single coordinate system, our understanding of the non-coding genome will be limited to what will stick.

More recent pangenome methods attempt to enhance the reference by conveying it as a graph[15]. For example, a species graph through elements of the genomic core would collapse into a single consensus, punctuated by bubbles that code small scale variation like single nucleotide polymorphism and small insertion/deletion elements. In contrast, the accessory genome forks the pangenome graph along entirely disparate paths. Graph-based methods attempt to incorporate nuance and novelty into a more complex reference structure. But the game is still the same: new data is aligned to a set of old genomes bound together into a complex, branching network.

Is there another way? Can we measure some other property of whole genomes that isn’t contingent on their alignment? Can we de-center the gene so we aren’t limited to the protein coding genome? Can we dispense with phylogenomics so we aren’t spending CPU years deciphering a bifurcating set of species relationships that convey a mere shadow of a more reticulate truth[16]?

Here, we propose several new information theoretic techniques that reimagine genomes as ensembles of information, containers subject to compression. This view of genomic information does not require annotation. Because we aren’t concerned with genes or the contiguous arrangements of genomic elements, we also forgo alignment. We instead describe pangenomes with summary statistics of string-based intersections. In this article, we argue that compression can enhance existing comparative genomic strategies, highlight structural evolution through controlled information loss, and democratize the bacterial species question by applying a uniform mathematics across the Linnean taxonomy.

The toolkit (entropy)

Our approach is guided by two foundational concepts at the very root of information theory: entropy and relative entropy. Both ideas rely heavily on Claude Shannon's seminal ideas on information introduced in "A Mathematical Theory of Communication", the founding document of information theory[17]. Information is data that reduces uncertainty. Shannon's original formulation resembled the thermodynamic construction of entropy devised for statistical mechanics[18]. We measure information entropy as

$$H = -\sum_{i=1}^{N} p_i ln\, p_i$$

where N is the set of all possible states, i, and p_i is the probability of the ith state. This expression quantifies data into bits (base two logarithm) or nats (natural log). The bit is the most irreducible unit of information. A bit is gained when a binary variable is assigned either a 1 or a 0.

In genomics, our data comes in sequences. We can measure the entropy of sequences by digesting into substrings of specific size. In the bioinformatics literature, substrings of biological readouts (DNA, RNA, protein) are called k-mers. In a comparative setting, we're most interested in the entropy of a group of sequences, or sequence ensemble[19,20]. For genome sequence ensembles, alignment has been the tool of choice. But alignment is computationally arduous and breaks down with evolutionary distance. Fields as diverse as linguistics[21], neurobiology[22], and statistical mechanics[23] have successfully employed entropy to quantify ensemble complexity. In each of these fields, researchers code a linear string of observations and divide into subsequences, calculating the entropy of each set across the ensemble. In Figure 1, we show

how the entropy of genome sequence (e.g. DNA/RNA) typically increases with increasing subsequence size. This is a block entropy curve.

Block entropy curves contain information about the complexity of the ensemble[19]. Systems with more ensemble structure – repeated elements across sequences – will peak at lower entropy. More novelty across sequences yields higher entropy. In genomics, closed pangenomes, with a large core shared across all species genomes, have low entropy. Auxiliary genes unique to subsets of genomes add entropy to the ensemble system. The uneven distribution of these elements is the hallmark of an open pangenome. But to measure complexity we don't need the annotated and aligned genes. Signal is preserved in unaligned and unannotated k-mers.

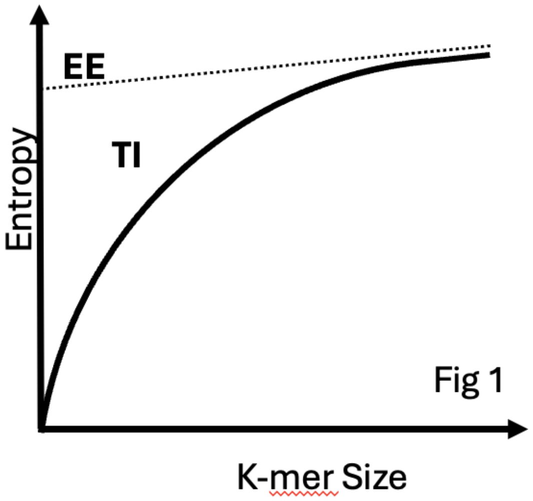

Figure 1. Block Entropy Curve. We show that entropy increases with k-mer size. We use this curve to calculate Excess Entropy (EE) and Transient Information (TI).

Block entropy curves asymptote at the minimum block size required to efficiently capture information across the sequences. We use three quantities calculated from these curves to describe the complexity of a pangenome: source entropy, excess entropy and transient information[19]. The source entropy (Hmu) is the irreducible randomness that remains even as larger block sizes capture most ensemble correlations. Hmu is a direct measure of randomness. Random distributions are hard to compress. A high source entropy is associated with the accumulation of unevenly distributed accessory genes, resulting in a more complex pangenome. The excess entropy (EE) is the non-random fraction of the total information in the system. It's the information we model from redundancies across the ensemble. Alignment is anchored to these same redundancies. In fact, alignment only works if enough of these redundancies are spread across the query genomes. Finally, the transient information (TI) measures how much information we must invest to learn Hmu and EE. In Figure 1 we show it as the area between the block entropy curve and the line defining Hmu. Species with closed pangenomes typically have a lower TI than those open to accumulating gene diversity. Closed pangenomes with a large core set of genes compress at lower k-mer sizes, approaching their Hmu quickly.

More tools in the toolkit (the information bottleneck)

Entropy is the workhorse of lossless compression. In fact, it defines lossless compression's limit. We cannot compress any further than the entropy of the source. In our context, the block entropy curve follows compression limits along a k-mer spectrum. Lossless compression preserves all data, but sacrifice can bring evolution into relief by isolating patterns from genomic noise. Using lossy compression, we can identify the core genome of any species without alignment or annotation. Along the way, we unlock the homologous and non-homologous recombination events that violate vertical signal.

To understand how we can detect structural evolution without annotation or alignment, we leverage Shannon's ideas on lossy compression. Shannon based his theory in communication. A sender passes a message to a receiver through a channel. The fundamental problem of communication is reconstructing that message. Communication channels suffer distortion. Data rarely reaches the receiver whole. Information entropy represents the limit on how efficiently a message can be compressed in the noise-less ideal.

No channel is noise-less. Still, the distortion introduced by noisy channels does not doom message passing. A sender can compensate for noise by encoding more information into a message, or a receiver can tolerate some level of distortion while ascertaining a sender's core meaning. Shannon formalized this concept as rate-distortion theory[17]. On the sender side, the rate is measured as bits of information per symbol. The sender's message is distorted as it passes through a channel. The sender’s rate and the receiver’s distortion are inversely related. The function describing the two variables for any given channel informs lossy compression. How much information loss can we tolerate in reconstructing a sender’s message?

This idea is central not only to information theory and lossy coding, but also to modern machine learning methods that use variational autoencoders to populate the compressive layers of a neural net[24,25]. The two key questions are 1. How well does a dataset compress, and 2. How much data can we afford to lose?

We use these concepts to further understand pangenome complexity. Imagine the compression regime in Figure 2. A set of genomes comprising a sequence ensemble are digested into k-mers and compressed into a set number of clusters. This compression is analogous to a communication channel. The more clusters we model, the higher the rate, and the lower the distortion. With fewer clusters, we force the k-mers through a narrower channel and suffer more distortion.

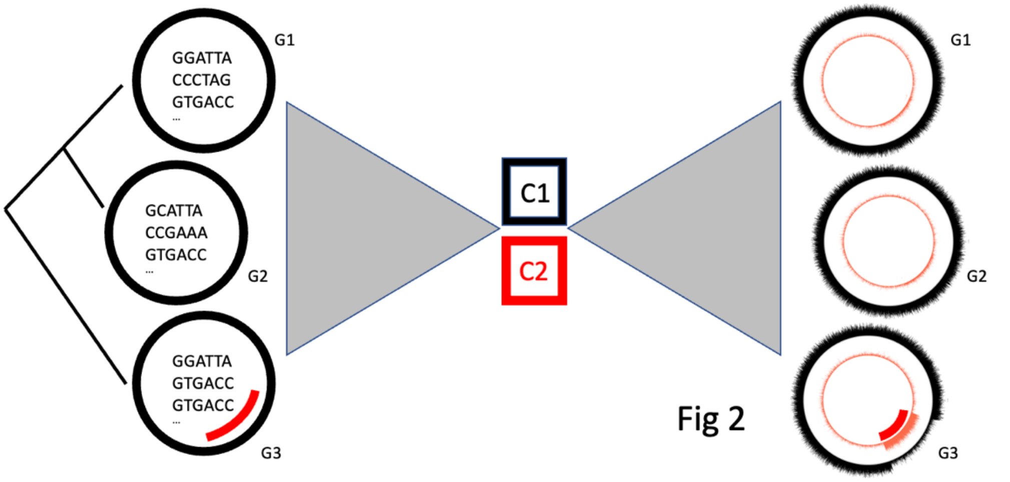

Figure 2. The information bottleneck. As k-mers from our input genomes are compressed into a narrow channel, patterns of structural evolution emerge from the resulting clusters.

If we hold the channel constant and model the same number of clusters across species ensembles, open pangenomes will suffer more data loss than their closed counterparts. Open pangenomes have more complex information to communicate. We employ the information bottleneck[26], an idea first proposed in the Natural Language Processing literature, to measure loss in our compression framework. Moreover, the clusters we glean from the information bottleneck comprise a model of structural evolution. The largest cluster usually represents the core. K-mers from recombination regions populate the others. The act of compression therefore deconstructs real biological events without the need to align genomes, build sets of orthologs, or calculate any trees.

Even more tools in the toolkit (relative entropy)

Block entropy curves measure the compressibility of any sequence ensemble. The bottleneck compresses information into clusters that communicate only the most salient bits. Compressibility is directly related to evolutionary strategy. Some species are open to genomic input, others have narrower, closed pangenomes. But as we've described it here, entropy treats the entire pangenome distribution as a single entity. This allows us to measure overall complexity, but doesn't account for each genome's departure from that distribution. Relative entropy, a measure of how one distribution (any given single genome) diverges from the overall distribution (our pangenome) adds nuance to our approach. Summed across all genomes, the relative entropy gives another, complementary angle on pangenome complexity and compression.

To formalize this concept, we turn to bedrock principals in ecology. Ecology has an extensive history of incorporating ideas from information theory and compression[27]. The Shannon Index has long been used to combine the effects of species richness, the absolute number of unique species in an environment, and species evenness, the relative abundances of those species[28]. But for ecologists the core equation is more general than Shannon entropy. Ecological datasets span many types of environments. Comparing diversity across those environments is crucial. Hill introduced the effective number of species as an intuitive solution[29]. The effective number of species of order q is given as

$$D_q = \left(\sum_{i=1}^{N} p_i^q\right)^{1/(1-q)}$$

where p_i is the frequency of a particular species i, and N, the total number of unique species. Sweeping through the parameter q controls the metric's responsiveness to rare ($q = 0$) or common species ($q = 2$ or more). At $q = 0$, the expression reduces to species richness, and at $q = 2$, the expression expands into the Simpson Index. But the sweet spot is at $q = 1$. The limit of this equation as q approaches 1 is Shannon's information entropy (or the Shannon Index if overheard in an ecology department). This transformation connects ideas from mathematical ecology to information theory. The exponent of Shannon entropy yields the Hill number at q = 1, or the effective number of species:

$$D_1 = exp\left(-\sum_{i=1}^{N} p_i ln\, p_i\right)$$

More diverse samples have higher Hill numbers. Hill numbers convey species diversity as an intuitive number. Because of its connection to information theory, Hill numbers are not the exclusive domain of ecologists. In Natural Language Processing, perplexity[30] is used to measure how well a language model can predict a string of text. Perplexity is the effective number of words in a library. Perplexity and Hill numbers draw from the same mathematical toolkit. This toolkit's simplicity allows for easy comparisons between entirely different experiments. But the expression collapses each experiment's observations into a single distribution.

We can enrich Hill numbers by extending beyond species measured as single variable distributions. To this point, we've defined what an ecologist would term alpha diversity[31], or the diversity of species in any one sample. One sample usually doesn't cut it. Ecologists sample multiple transects from their environment of interest. Sampling introduces several opportunities. First, the degree of sample overlap is a potential gauge of efficacy. Second, sample diversity yields insight into the overall, hypothetical, unapproachable diversity of the system, or the gamma diversity. If samples are highly diverse, ascertaining the diversity of the target environment may require more samples to be taken. If gamma diversity is too high, no sampling scheme may be enough. Beta diversity measures the degree of overlap between samples[32,33]. Grounding the concept in information theory, we extend the Hill number of species into a Hill number of samples. The following expression yields the effective number of samples:

$$D_{\beta} = exp\left(\sum_{S=1}^{M} w_s \sum_{i=1}^{N} p_{si} ln \frac{p_{si}}{p_i}\right)$$

This equation incorporates the Kullback-Leibler divergence or relative entropy, a formulation as frequently used as entropy in the information theory literature[34]. The relative entropy measures the divergence of any one genome's k-mer distribution against the k-mer distribution of the entire pangenome. Here, N is the number of unique species, M, the number of samples, p_{si} is the frequency of species i in sample s, p_i is the frequency of species i across all samples, and w_s weighs all observations in sample s relative to all individuals collected in the experiment.

The effective number of samples is another measure of compression. If species richness and evenness is the same across all samples, the effective number of samples reduces to 1. If the samples contain no species in common, or if species have wildly different occurrence counts, the effective number of samples approaches the number of samples taken. In the first case, we have perfect compression. In the latter, no compression at all.

We take this ecological concept and adapt it to genomics. Our goal is to calculate the information diversity embedded in sequence ensembles. This requires a complete reframe. Rather than species in a community (alpha diversity), we think k-mers in a genome. Rather than transects in an environment (beta diversity), we think genomes in a pangenome. The shift is in the container. Employed in this way, we recast Hill numbers as the effective number of genomes or genome equivalents. We coin KHILL, an intuitive metric that quantifies the information space of a pangenome, or the degree to which it will compress. We calculate KHILLs in a fraction of the time it takes to annotate genomes, run alignments, and build the orthologs required to compute pangenome fluidity.

The toolkit applied

Biological datasets are large and growing. Other fields also contend with large datasets, and some have been grappling with them for decades longer. For example, astronomers have big data, perhaps the biggest data in the sciences[35]. Processing and saving all astronomic data is impossible. Astronomers have known for years the importance of sensing data as it shines onto their mirrors. Compression normally happens at the point of collection. We are quickly reaching this point in biology.

Organized, collaborative genome sequencing projects began in earnest in the 1990s. Starting then and through the first two decades of this century, genomic datasets were sacrosanct. Groups held onto their data until every angle was exhausted. Though genomic data has always been big data, generating it back then was costly. This is no longer the case. The price of genome sequencing has seen steep decline. Storing this accumulating data has become nearly impossible. Perhaps it is time to let go. With the information bottleneck, we tolerate controlled data loss. New sequencing platforms emit data in nearly unending streams. Sensors are designed to glean information from data streams in real time. There are sensors that detect change in acceleration (engineers), in light (astronomers), in brain activity (doctors). Perhaps streams of biological sequence can also be processed and discarded. Can sequence become a sensor?

Take for example SARS-CoV-2. Fifteen million SARS-CoV-2 genomes are now available in various repositories around the world[36]. The state of the art in surveilling these genomes as they accumulate in time and space is phylodynamic[37,38]. But phylodynamics is retrospective. Investigators curate a fraction of the genomes available, compare them against an even more rarefied set of references, and embed the new alongside the old either in phylogenetic trees or networks. Alignment is the linchpin in this arrangement. Genome alignments feed tools like Nextstrain[39], which employ Bayesian and likelihood phylogenetic approaches – some of the most computationally costly algorithms in bioinformatics – to extend our view of SARS-CoV-2 biology slightly beyond the anointed references in a database.

We find this limiting. We can use KHILL to look forward, analyzing all the sequence available to us outright[40]. Whether it's 15 million clinical genomes or streams of wastewater, KHILL is capable of processing terabytes of streaming sequence and flagging the emergence of new variants without relying on the references that confine biological novelty. KHILL can also achieve rapid community analysis as exemplified in our study of the microbial shifts in the making of cheese (ref), and the microbiome perturbations caused by broad spectrum antibiotics (unpublished data). Whether it's a life threatening virus or the cheese you spread on crackers, we use *all* sequence, not just the bits that will stick to existing references.

For SARS-CoV-2, we calculate one KHILL number per day along a pandemic time course. Compiling these genome equivalents yields an information diversity curve through time. KHILL increases as variants of concern ascend in a population mixing with a prior background. KHILL decreases once these variants grow dominant and sweep away all other genomic heterogeneity. In this way, we detect the emergence of concerning strains well before annotation clearinghouses have blessed new database entries.

As a genomic measure of compression, KHILL also naturally lends itself to the analysis of pangenomes. In fact, with SARS-CoV-2, we used KHILL as a rolling measure of pangenome complexity. Because of their contracted timeline, pandemic genomes occupy a small information space. The KHILL of all the millions of sequenced SARS-CoV-2 compresses to about 1.15 effective genomes. But KHILL is not restricted to any one biological scale. We can measure the complexity of strains, species, genus, and collections at even higher taxonomic levels.

For example, we have used KHILL to calculate the pangenome complexity of all known bacterial species[41]. An analysis at this scale is impossible with current alignment-based bioinformatic techniques. But because KHILL is fast, we can compute genome equivalents for every species in the database. We couple this with metrics derived from block entropy curves (Hmu, EE and TI) to calculate the information space occupied by all known bacterial species. This information theoretic approach democratizes species classification, labeling each pangenome with a single number.

As we've defined it, KHILL species complexity mixes two separate phenomena. First, species definitions vary. The Linnaean taxonomy imposes a hierarchy on life, but this hierarchy is not uniformly applied. Species in one part of the taxonomic tree may not mean the same thing to its experts as species in another part of the tree. This is cultural. But it does influence the relative breadth of species buckets. We expect some variation in KHILL based just on these very human inconsistencies.

More interesting, however, is our second observation. Pangenome fluidity[11] has been shown to track with some gross aspects of bacteria phenotype[42]. For example, host-bound species accustomed to a uniform environment typically have less complex pangenomes. Cosmopolitan species occupying diverse niches tend towards more pangenome diversity. Obligate bacteria are less complex than their facultative counterparts. Non-motile organisms, less complex than those on the move. Complexity, in this case, was measured as pangenome fluidity. Pangenome fluidity is as near to measuring information-theoretic complexity as alignment-based techniques can get. We find that KHILL, a more direct, swifter measure of complexity, also corresponds to bacterial lifestyle. We see this borne out in KHILLs. For example, pathogens have significantly lower KHILL than mutualists. Challenging environments presumably encourage the accretion of pangenome complexity as species contend with instability. Our compression based techniques squeeze this information from genomes without the normal bioinformatics playbook.

Challenges? In sacrifice there is clarity!

Metrics based in compression can distort mechanism. KHILL increases with population heterogeneity, as in the case of our SARS-CoV-2 populations. But it also increases with genetic distance. This genetic diversity could be the result of environmental pressure, or it could simply be lazy, inconsistent categorization. Because block entropy curves and KHILL dispense with alignment, we also lose the ability to pinpoint change in genomic space. In criticism of this work, we've heard over and over how obscuring mechanism, sacrificing location, or conflating biological forces is a weakness. But in a field saturated with sequence data, our approach allows researchers to skim data streams without resorting to the heaviest, most cumbersome algorithms in bioinformatics.

The idea of conflating signal is a hallmark of information-based approaches. Shannon's communication problem is emblematic of this compromise. Distortion is inevitable as information is relayed from sender to receiver. This concept has been used in everything[43] from telecommunications, to thermodynamics, to data encoding in Natural Language Processing. More complex data requires a broader channel to communicate. But sometimes we must sacrifice nuance for meaning. In fact, compressing away the noise can sometimes distill signal. In other words, conflation sometimes yields clarity.

We take this concept to genomics[44]. Mutation, homologous recombination, and horizontal gene transfer all distort genomic signal. We can capture the degree of distortion by measuring how difficult it is to compress strings (k-mers) from a set of genomes into a set number of clusters. If the compression is easy, we need fewer clusters – a narrower channel – to achieve communication at an acceptable level of distortion. But if the genomes are labile, we need more clusters to communicate the added information diversity. The information bottleneck[26] quantifies complexity.

The clusters that comprise our information channel, are datasets that sort meaning. Where KHILL is a mark of compression, these clusters are actual compressed representations. We can measure the fidelity of the original ‘message’ carried by the genomes relative to these compressed representations. Clonal, tree-like, bifurcating species generally require fewer clusters to model modes of genomic change. Recombinogenic species require more clusters to achieve the same signal clarity.

Like KHILL, this approach conflates biological phenomena. Lossy compression through the bottleneck does not distinguish between mutation and recombination. But for both KHILL and the bottleneck, the compressibility of a set of genomes becomes a metric that can be used to compare sets of species.

We began this essay bemoaning genomics as a retrospective enterprise. We believe information theory allows us to shift our gaze forward. Eliminating references opens us to novelty. De-centering the gene offers a new view of pangenome complexity. And eliminating alignment boosts speed. Together these efficiencies recast sequencing as a sensor delimiting change. We can sense change along a pandemic trajectory. We can predict bacterial lifestyle from compression. And we can probe the unbalanced hierarchies of bacterial taxonomy.

Data Accessibility

All data discussed in this paper is freely available online at NCBI (https://www.ncbi.nlm.nih.gov).

Author Contributions

AN conceived the project and drafted the original and final versions; SG co-wrote the manuscript; MTPG co-wrote the manuscript.

References

1. Guigó, R. Genome annotation: From human genetics to biodiversity genomics. *Cell Genomics* **3**, 100375 (2023).
2. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056–1060 (2009).
3. Thomas, A. M. & Segata, N. Multiple levels of the unknown in microbiome research. *BMC Biol.* **17**, (2019).
4. Erlich, Y. A vision for ubiquitous sequencing. *Genome Res.* **25**, 1411–1416 (2015).
5. Baker, R. H. & DeSalle, R. Multiple Sources of Character Information and the Phylogeny of Hawaiian Drosophilids. *Syst. Biol.* **46**, 654–673 (1997).
6. Rokas, A., Williams, B. L., King, N. & Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798–804 (2003).
7. Neumann, J. S., Desalle, R., Narechania, A., Schierwater, B. & Tessler, M. Morphological Characters Can Strongly Influence Early Animal Relationships Inferred from Phylogenomic Data Sets. *Syst. Biol.* **70**, 360–375 (2021).
8. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, (2019).
9. Shi, T. & Falkowski, P. G. Genome evolution in cyanobacteria: The stable core and the variable shell. *Proc. Natl. Acad. Sci.* **105**, 2510–2515 (2008).
10. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome”. *Proc. Natl. Acad. Sci.* **102**, 13950–13955 (2005).

11. Kislyuk, A. O., Haegeman, B., Bergman, N. H. & Weitz, J. S. Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* **12**, 32 (2011).

12. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

13. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

14. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

15. Eizenga, J. M. *et al.* Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).

16. Doolittle, W. F. Phylogenetic Classification and the Universal Tree. *Science* **284**, 2124–2128 (1999).

17. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

18. Boltzmann, L. On the Relation between the Second Law of the Mechanical Theory of Heat and the Probability Calculus with respect to the Theorems of Thermal Equilibrium. *Wien. Berichte* **76**, 373–435 (1877).

19. Crutchfield, J. P. & Feldman, D. P. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos Interdiscip. J. Nonlinear Sci.* **13**, 25–54 (2003).

20. Bialek, W., Nemenman, I. & Tishby, N. Predictability, Complexity, and Learning. *Neural Comput.* **13**, 2409–2463 (2001).

21. Bentz, C. & Alikaniotis, D. The word entropy of natural languages. Preprint at https://doi.org/10.48550/ARXIV.1606.06996 (2016).

22. Bialek, W., Rieke, F., De Ruyter Van Steveninck, R. R. & Warland, D. Reading a Neural Code. *Science* **252**, 1854–1857 (1991).

23. Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.* **106**, 620–630 (1957).

24. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. Preprint at https://doi.org/10.48550/ARXIV.1312.6114 (2013).

25. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).

26. Tishby, N., Pereira, F. C. & Bialek, W. The information bottleneck method. (2000) doi:10.48550/ARXIV.PHYSICS/0004057.

27. Chao, A., Wang, Y. T. & Jost, L. Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods Ecol. Evol.* **4**, 1091–1100 (2013).

28. Jost, L. The Relation between Evenness and Diversity. *Diversity* **2**, 207–232 (2010).

29. Hill, M. O. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **54**, 427–432 (1973).

30. Brown, P., Pietra, S. D., Pietra, V. D., Lai, J. & Mercer, R. An Estimate of an Upper Bound for the Entropy of English. *CL* (1992).

31. Jost, L. PARTITIONING DIVERSITY INTO INDEPENDENT ALPHA AND BETA COMPONENTS. *Ecology* **88**, 2427–2439 (2007).

32. Marcon, E., Hérault, B., Baraloto, C. & Lang, G. The decomposition of Shannon’s entropy and a confidence interval for beta diversity. *Oikos* **121**, 516–522 (2012).

33. Marcon, E., Scotti, I., Hérault, B., Rossi, V. & Lang, G. Generalization of the Partitioning of Shannon Diversity. *PLoS ONE* **9**, e90289 (2014).

34. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).

35. Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLOS Biol.* **13**, e1002195 (2015).

36. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, (2017).

37. Grenfell, B. T. *et al.* Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science* **303**, 327–332 (2004).

38. Volz, E. M., Koelle, K. & Bedford, T. Viral Phylodynamics. *PLoS Comput. Biol.* **9**, e1002947 (2013).

39. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

40. Narechania, A. *et al.* Rapid SARS-CoV-2 surveillance using clinical, pooled, or wastewater sequence as a sensor for population change. *Genome Res.* **34**, 1651–1660 (2024).

41. Narechania, A., Bobo, D., Gilbert, M. T. P. & Gopalkrishnan, S. The information content of species: formal definitions of pangenome complexity track with bacterial lifestyle. Preprint at https://doi.org/10.1101/2025.03.28.645969 (2025).

42. Dewar, A. E., Hao, C., Belcher, L. J., Ghoul, M. & West, S. A. Bacterial lifestyle shapes pangenomes. *Proc. Natl. Acad. Sci.* **121**, e2320170121 (2024).

43. Noise and Distortion. in *Advanced Digital Signal Processing and Noise Reduction* 29–43 (Wiley, 2001). doi:10.1002/0470841621.ch2.

44. Narechania, A. *et al.* What Do We Gain When Tolerating Loss? The Information Bottleneck Wrings Out Recombination. *Mol. Biol. Evol.* **42**, msaf029 (2025).