The information signature of diverging lineages

Douglas G. Moore[1,2]*, Matheo Morales[3,4], Sara I. Walker[1,5], Greer A. Dolby[3,6]*

[1] The Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe, AZ

[2] 39Alpha Research, Tempe, AZ

[3] School of Life Sciences, Arizona State University, Tempe, AZ

[4] Department of Genetics, Yale University, New Haven, CT

[5] School of Earth and Space Exploration, Arizona State University, Tempe, AZ

[6] Center for Mechanisms of Evolution, Biodesign Institute, Arizona State University, Tempe, AZ

Corresponding author(s):

DGM - douglas.g.moore@asu.edu

GAD - gadolby@asu.edu

Running title: Genomic information decomposition

Keywords: speciation, genomics, partial information decomposition, theory

INTRODUCTION

The process of forming new species is the driving force behind the diversity of life on Earth. Yet, we have not answered the basic question: why are species unevenly distributed across taxonomic groups and geographic settings? This is because we lack the means to directly compare aspects of population (lineage) divergence across unrelated species because taxon-specific effects make comparisons difficult or impossible. Here, we present a new solution to this challenge by identifying the information signature of diverging lineages, calculated using partial information decomposition (PID), under different evolutionary scenarios. We show **in silico** ***how the informational decomposition of genetic metrics varies over time since divergence. Calculating PID over 97,200 lattices reveals that the decomposed nodes of Tajima's D, θ_W, and π have strong information signatures, while F_{ST} was least useful for discriminating among divergence scenarios. The presence or absence of gene flow during divergence was the most detectable signature; mutation rate and effective population size (N_e) were also detectable whereas differences in recombination rate were not. This work demonstrates that PID can reveal evolutionary patterns that are minimally detectable using the raw metrics themselves; it does so by leveraging the architecture of the genome and the partial redundancy of information contained in genetic metrics. Our results demonstrate for the first time how to directly compare characteristics of diverging populations even among distantly related species, providing a foundational tool for understanding the diversity of life across Earth.***

Many central themes of evolutionary biology deal with how new species differentiate from related ones, but many core questions about speciation remain unsolved. Among these are how often lineages diverge in isolation or with gene flow, how often certain geologic or geographic settings promote speciation, and which taxonomic groups are most affected (1). Answering these questions depends upon comparing aspects of lineage (population) divergence, such as the amount of genetic divergence and gene flow between populations, in different taxa. Yet, the organismal traits that cause population pairs of species to diverge differently under a common set of conditions also influence the measures of population diversity and divergence (e.g., π, F_{ST}) that we use to characterize those differences, making it difficult to compare characteristics of diverging populations across distantly related species (Figure 1). Some such aspects can be tackled using the site frequency spectrum (SFS; 2), but the SFS has limitations because different histories can produce similar geometric distributions (3).

In this study we take an altogether new approach by assessing the information signal of genetic metrics under different evolutionary scenarios. Multivariate analysis is helpful when individual variables convey partial knowledge about a system of study. Taking an example from the kinetic theory of ideal gasses, macroscopic observables such as temperature and pressure provide information about the microscopic state of the system; they each provide some information, but it is only when taken together that the density of the gas can be determined solely through observation of those properties. We can characterize and identify observables which have utility for answering specific questions about the system under study by decomposing the information the

observables provide over combinations of those observables, identifying the redundant, unique and synergistic contributions of each observable. Along this vein, we chose to decompose the multivariate information signal of different combinations of genetic metrics calculated on datasets simulated under a suite of evolutionary scenarios. Using the architecture of the genome, this information signal is effectively standardized within a dataset making comparison across species possible.

DNA is an information storage molecule because it encodes directions for the formation of proteins and those directions are heritable across generations. Information processing across levels of biological organization is also thought to be a fundamental, measurable characteristic that sets living systems apart from nonliving systems (4). It may therefore be appropriate that information theoretic measurements reveal higher-order patterns about biological evolution, which fundamentally based on the context-dependency and patterned cross-generational transmission of genetic information (DNA) (5).

Claude Shannon introduced mutual information in 1948 to quantify the degree to which observations of some random variable reduces the observer's uncertainty about another variable (6). Mutual information and its related information-theoretic quantities have been applied in biology (7-10), machine learning and data science (11-13), complex systems theory (14-16), and ecohydrology (17). As with linear measures of correlation, mutual information can be generalized to a multivariate setting where total correlation (18), dual total correlation (19) and interaction information (20) are exemplars. We calculated the PID formalized by Williams and Beer (21) over eight common genetic metrics applied to 'genic' and 'intergenic' regions of genetic data

93 simulated for diverging populations. PID decomposes the total correlation between a
94 target variable and a collection of source variables over a lattice (Figure 1E). The nodes
95 of the lattice represent redundant, unique and synergistic contributions of combinations
96 of the source variables, and each combination yields a non-negative information
97 contribution.

98 Application of PID to genomic data here (Figure 1) is based on the premise that
99 how information about evolutionary processes is encoded across genomic elements
100 (e.g., genes), and how that information decomposes over genetic metrics (e.g., F_{ST}),
101 provides information about both the evolutionary and organismal context of that
102 sequence (Figure 1B). Consider populations diverging without gene flow (i.e.
103 reproduction). We expect these populations to accrue divergence in intergenic
104 (noncoding) elements of the genome faster than coding elements, and this discrepancy
105 would be higher in genetically isolated populations than if the populations had some
106 level of gene flow (22). To understand the information signature of genetic metrics
107 under different evolutionary settings, we simulated 60 kbp of sequence data for
108 individuals in each of two populations under 24 evolutionary scenarios. Evolutionary
109 scenarios varied mutation rate (μ), recombination rate (r), effective population size (N_e),
110 and gene flow (M). We sampled genomes of 10 individuals per population at 10
111 timepoints per simulation to obtain a timeseries signature during divergence. Simulated
112 genomes contained 10 'intergenic' regions that evolved unconstrained and 10 'genic'
113 regions that evolved under purifying selection where mutations that changed the amino
114 acid from the ancestral state penalized the individual's fitness (see methods). We
115 simulated 100 replicates per scenario which were concatenated and used to calculate

116 eight genetic metrics (F_{ST}, D_a, Tajima's D, θ_W, π, S, Fu's FS, and D_{XY}) per genomic
117 region. For each set of genetic metrics, we calculated PID over lattices constructed from
118 combinations of up to four variables using genic/intergenic as the target variable.
119 Results yielded 97,200 PID lattices that contained 7,648,800 nodes in total. The results
120 were vectorized over timepoints to determine how the information of nodes within
121 lattices varied with time since divergence and we grouped these into temporal patterns
122 using k-means clustering. Results for the long time series (10^5 generations) showed that
123 nodes within decomposed lattices exhibited partial information patterns that fell into
124 seven timeseries classes (Figure 2B). Most nodes were invariable (N=302,207), while
125 some reliably decreased (Class 2, N=196) or increased (Class 3, N=128). Complex time
126 signatures are in SI (Figure S1).

127 We find that information signatures depended almost entirely on the
128 informativeness of a few nodes representing key genetic metrics and not on the set of
129 variables being decomposed (Figures 2a, 3, S2, S3). The only exception to this pattern
130 was Tajima's D, which increases in informativeness over time in the absence of gene
131 flow (Figure 3c). When included among the variables, Tajima's D redistributed
132 information over the lattice, making other nodes more informative even though Tajima's
133 D itself was rarely (if ever) part of the node with an informative time signature (class 2 or
134 3) or an ability to discriminate among scenarios (e.g., Figure 3d). The variables lending
135 greatest informativeness (i.e. ability to discriminate among evolutionary scenarios) were
136 nucleotide diversity (π) and Watterson's estimator (θ_W) (Tajima's D was influential for
137 the reason described). The types of nodes with greatest discrimination potential were
138 either unique or redundant; synergistic nodes were never discriminatory among

evolutionary scenarios (Figure 3). We interpret this to be due to the redundancy or nonindependence of the genetic metrics and not a behavior of PID itself.

Our exploration of evolutionary parameter space (N_e, M, r, and μ) reveals the presence of gene flow to be easiest to detect (Figure 3), which is consistent with results based on analysis of the Site Frequency Spectrum (SFS, 23). The discrimination is binary rather than continuous, identifying the presence not the degree of gene flow in our simulations. However, sampling more rates of gene flow could further resolve this pattern. The variable most able to recover the gene flow signal is π while Tajima's D and F_{ST} reflect the presence/absence of gene flow only at higher mutation rates (Figure S4). Mutation rate is recovered to a lesser extent than the presence of gene flow and is reflected best by Fu's FS. Segregating sites (S) and Watterson's estimator (θ_W) reflect differences in mutation rate but are also influenced by presence/absence of gene flow (Figure S5). S and θ_W yielded identical results (Figure S5); this is expected because θ_W is calculated by correcting S for the number of samples observed which was consistent across our simulations. This nonetheless demonstrates PID's consistency to record equivalent patterns despite calculation over raw values of different scales. If we varied sample size, these metrics might no longer produce identical decompositions.

No nodes effectively discriminated recombination rate, which could be because our metrics did not reflect patterns of genomic linkage. We speculate that decomposing alternative genetic metrics could recover a difference in recombination rates. For differences in N_e (Figure S6), the magnitude of partial information was slightly inflated under smaller effective population sizes (Figure 3). It is possible that additional targeted exploration of 'node-space' with a focus on discriminating among N_e could uncover a

more useful decomposition.

We assessed the reproducibility of PID results by selecting four scenarios ($M = 0$, $\mu = 2.5 \times 10^{-7}$, $N_e \in \{500,1000\}$, $r \in \{0, 10^{-7}\}$) and simulating 300 additional genetic datasets in batches of 100 per scenario (Table S2). As done previously, the eight metrics were calculated, their results concatenated per batch, and PID computed independently on the three new batches of simulations. To assess technical variance, lattices across the four batches (the one original batch and three additional replicate batches) were compared using mean-absolute-difference (MAD). Results showed the MAD between any pair of lattices never exceeded $1e^{-4}$ bits, which is orders of magnitude below the differences we observe between evolutionary scenarios. In other words, the technical variance among genetic simulations and genomic PIDs was exceedingly small.

After as few as 10^4 generations, many of the PID nodes exhibited partial information near the theoretical maximum of 1 bit. However, it is expected that before divergence begins that all nodes have negligible partial information because the populations start out from a single sequence. To confirm this, we re-simulated the first 10^4 generations and sampled at a higher frequency (every 10^3 generations) with the expectation that the nodes exhibiting non-zero partial information would rapidly depart from 0 bit. Results are consistent with the assumption that the information content begins at 0 bit; results show it becomes nonzero following the first generation. There is slightly less stability in the early generations evidenced by the greater number of time series clusters compared to the 10^5-generation dataset (N=20 vs N=7 groups, respectively; Figure S1 vs. S7). With additional work, it is possible that this latency

185 period could be leveraged to measure aspects of very early divergence ($\lesssim$ 1000
186 generations, Figure S8). Aside from slightly reduced power in early generations to
187 discriminate among evolutionary scenarios, the information patterns stabilize once a
188 critical threshold is met (Figure 3A). This reliability-of-pattern is favorable for the
189 applicability of PID to genomic datasets of wild populations in which age-of-divergence
190 dates can be reasonably presumed to be $\gtrsim$3000 generations.

191 Several broadscale observations can be made from these results about both the
192 behavior of information decomposition and its application to genetic metrics and
193 evolution. First, it is surprising that F_{ST}, which is one of the most broadly used and
194 interpretable statistics in evolution, was the least informative when decomposed, while
195 more fundamental measures of diversity (π, S or θ_W) drove information patterns. Also,
196 key variables determined information patterns rather than the combination of variables
197 in the decomposition (the variable set). This finding is important because computing PID
198 on a large number of variables is costly and difficult to interpret; it suggests that small
199 lattices with a few key variables are sufficient for many purposes, and new applications
200 of PID should assess patterns within small decompositions first. Likewise, we were
201 surprised to find that the unique and redundant nodes far outweighed information in
202 synergistic nodes. We interpret this to be a property of the genetic metrics analyzed,
203 particularly their non-independence under different evolutionary settings. If, unlike other
204 multivariate statistics, PID can leverage interpretational power from the
205 nonindependence of variables, that may be its biggest strength. It is possible that
206 synergistic terms could gain power under more evolutionarily complex scenarios, such
207 as reduced gene flow from a physical barrier in combination with differential adaptation,

which remains a stumbling block for analyses using the SFS (24).

A final, related discovery was that combinations of variables exhibited “additive” discriminatory behavior. As an example, the single variable decomposition [S] displayed sensitivity to both M (presence/absence) and μ (Figure 4a). The single variable decomposition of [D] reflected primarily presence or absence of gene flow (M, Figure 3c). The two-variable decomposition [S,D] presents a simple combination of [S] and [D], with a primary signal of gene flow/no gene flow (present in both [S] and [D]) and a secondary signal of μ (present in [S]). The discovery of additive behavior among variables should help biologists perform a targeted search of ‘decomposition-space’ for lattices that will discriminate between biological parameters of interest. In theory, choice of metrics to decompose can be optimized based on which properties of a diverging population are of interest.

In summary, answering many outstanding questions in speciation and macroevolution depends on our ability to compare characteristics of diverging lineages across diverse organisms. We outline a new information theoretic approach to do this which leverages the architecture of the genome and redundancy of well-known genetic metrics. Application of PID here shows the presence/absence of gene flow is easily detectable by several decompositions. But more importantly, PID is a new class of tool that can be applied to other variables and the large decomposition-space can be mined for nodes targeted to specific evolutionary characteristics and timepoints.

MATERIALS & METHODS

<u>Genetic simulations & evolutionary analysis</u>

231 We simulated population divergence scenarios in SLiM v3.3.2 (25) using house scripts
232 that automated the parameter generation, simulations, and molecular statistic
233 calculations with EggLib (Evolutionary Genetics and Genomics Library) v3.0.0b21 (26).
234 The automated pipeline was written in python, Jinja, linux, and Eidos and allows for
235 parallelizing SLiM simulations to be run in batch on high performance computing
236 clusters (available at https://github.com/mmoral31/SLiM_pipeline); it processes
237 intermediate files, and calculates diversity statistics. All simulations started with a
238 random seed and initial genome sequence of 60,000 nucleotides divided into ten
239 "coding" and ten "noncoding" regions of 3,000 nucleotides each to simulate a small
240 eukaryotic genome. The "coding" and "noncoding" regions were composed of a random
241 assortment of 1,000 codons (excluding stop codons) and 3,000 nucleotides,
242 respectively. We simulated data under all combinations of the following four parameters:
243 recombination rate $r \in \{0, 10^{-7}\}$, mutation rate $\mu \in \{2.5 \times 10^{-7}, 2.5 \times 10^{-6}\}$, symmetrical
244 gene flow $M \in \{0.0, 0.01, 0.10\}$, and effective population size $N_e \in \{500, 1000\}$ for a total of
245 24 scenarios (Table S1). To simulate purifying selection on coding regions and calculate
246 fitness, in SLiM we translated and compared the coding regions for each individual per
247 generation to its initial (ancestral) amino acid sequence. The individual was assigned a
248 reduced fitness of 0.9 based on one amino acid mismatch and 0.5 for two or more
249 mismatches. Simulations were run for 100,000 generations and VCF (Variant Call
250 Format) files were generated every 10,000 generations; each parameter combination
251 was replicated 100 times, yielding 2,400 VCF files that were then concatenated for
252 genetic calculations.

253 To calculate genetic metrics, we used EggLib (Evolutionary Genetics and

Genomics Library) on the concatenated VCFs generated by SLiM. The house script calculated eight statistics: F_{ST}, D_a, Tajima's D, θ_W, π, S, Fu's FS, and D_{XY} in each 3-kb coding and noncoding region individually (i.e. window size and step size of 3 kb; Table S2). Whether the region was 'coding' or 'noncoding' was recorded in the output for each window. These statistics were then passed into Imogen.jl for PID calculations.

Technical variation & early divergence simulations

To assess technical variance of simulations and its effect on PID calculations, we ran an additional 300 replicates (for a total of N=400 per parameter combination) for a subset of parameters: no gene flow, mutation rate = 2.5×10^{-7}, with both population sizes of 500 and 100, and with and without recombination (Table S2). To confirm the theoretical assumption that the genetic metrics hold no information at the immediate onset of divergence, and to test the reliability of decompositions during early-divergence, we ran another set of simulations focusing on early time-series sampling. For this we re-ran all 24 scenarios for 10,000 generations and sampled every 1,000 generations, with another 100 replicates per scenario (Table S2). The VCFs were generated at generation two instead of generation one due to software limitations. This set of data also consisted of 2,400 VCF files that were concatenated for genetic analysis.

Partial information decomposition

Implementation & running of analyses

The preceding analysis provided a sequence of values for each genetic metric along

with whether or not the associated 3-kb region was a coding or noncoding region. In other words, over the length of the genome, we had a binary vector $\vec{G}$ representing whether or not a given region is coding, and eight vectors $\overrightarrow{d_i}$ – one for each evolutionary statistic estimated in the region. We treated these as sequences of observations of random variables G and d_i which allowed us to ask how the information about whether a given region is coding decomposes over combinations of the evolutionary statistics using the PID formalism. However, in general the $\overrightarrow{d_i}$ are real-valued, and since the Williams and Beer PID formalism is limited to discrete-valued data, we performed a discretization process on each $\overrightarrow{d_i}$. For the sake of simplicity and to reduce systematic errors due to the short genome lengths (60,000 bp), we opted to use a "mean-threshold" binning scheme: the jth value of $\overrightarrow{d_i}$, written here as d_{ji}, was replaced with a 1 if it was greater than the average of all values in $\overrightarrow{d_i}$, and 0 otherwise. This yielded a set of eight binary vectors $\overrightarrow{D_i}$. PID could then be applied to the resulting vectors $\vec{G}$ and $\overrightarrow{D_i}$. In this process, the multivariate mutual information between $\vec{G}$ and $\overrightarrow{D_i}$ was estimated and subsequently decomposed into a sum of non-negative terms, roughly describing the degree to which combinations of the variables $\overrightarrow{D_i}$ provided unique, redundant, and synergistic information about whether or not a given region is coding. Due to computational limitations of PID, the process can only be applied to at most five of the evolutionary statistics at time. In this work, we limited calculations to at most combinations of four statistics considered simultaneously to limit computation time and keep the amount of data generated manageable. We carried out the computation for all lattices with between one and four of the genetic metrics as source variables and the region classification $\vec{G}$ as the target to yield 162 PID lattices per simulation per sampled

generation for a total of 97,200 lattices containing 7,648,800 nodes in total.

Information clusters & informative metrics

For each node in each lattice per simulation, we created a vector of the partial information (Π) for that node over time. Many of the resulting curves had similar temporal dynamics but started out with a different initial value. To account for this, we shifted each vector by subtracting the zeroth element, element-wise, to make all vectors start at the same zero value. To reduce the 764,880 vectors into classes of information signatures over time, we performed k-means clustering based on Euclidean norm distance in the 10-dimensional space formed by the 10 sampled timepoints. We used the “elbow method” to identify a reasonable number of clusters for the k-means analysis. That is, we plotted the explained variance against the number of clusters for a range of values of k, and selected k at or near a prominent elbow (bend) in the curve. Based on this method, we chose a 7-means clustering to reduce the number of patterns on which to conduct further analyses without oversimplifying the diversity of results.

Limitations

We note several aspects of this analysis that could affect the results obtained. First, the genomes simulated were small (60 kbp) relative to the gigabasepair-size genome of vertebrates, for example. For simplicity, simulations did not incorporate standing variation in the population, which is a simplification of reality. Further, certain technical aspects of the Williams and Beer formalization of PID, such as the lack of localizability, continuity and differentiability, leave room for further exploration. More recent work by

(27,28) made progress in addressing these points, and have arguably developed notions of information decomposition that are easier to interpret (which may be important for the future of PID for biological applications). The choice to use Williams and Beer's formalism necessitated a binning procedure on the computed evolutionary quantities before analysis, and the particular method of binning can have effects on the resulting decompositions. It would be worth considering different discretization methods (i.e. 29-31) to assess the sensitivity of the results presented in this work.

ACKNOWLEDGEMENTS

We thank the Baja GeoGenomics consortium for useful feedback and discussions and Kenro Kusumi for support.

DATA AVAILABILITY

Code is available on Github: https://github.com/elife-asu/gPID and https://github.com/mmoral31/SLiM_pipeline

All data files will be made available upon publication.

AUTHOR CONTRIBUTIONS

DGM and GAD conceived of this project; DGM, SIW, and GAD supervised analyses. MM carried out genetic simulations and pipeline automation, DGM performed information theoretic analyses. GAD and DGM drafted and all authors critically revised the manuscript.

FUNDING

This work was funded by NSF-EAR award #1925535 to GD, NSF GRFP to MM, award #61184 from the John Templeton Foundation to SIW.

REFERENCES

1. Antonelli, A. *et al.* Geological and climatic influences on mountain biodiversity. *Nature Geoscience* 1–10 (2018). doi:10.1038/s41561-018-0236-z
2. Xue, A. T. & Hickerson, M. J. The aggregate site frequency spectrum for comparative population genomic inference. *Mol Ecol* **24,** 6223–6240 (2015).
3. Myers, S., Fefferman, C. & Patterson, N. Can one learn history from the allelic spectrum? *Theoretical Population Biology* **73,** 342–348 (2008).
4. Kim, H., Valentini, G., Hanson, J. & Walker, S. I. Informational architecture across non-living and living collectives. *Theory in Biosciences* 1–18 (2021).
5. Donaldson-Matasci, M. C., Bergstrom, C. T. & Lachmann, M. The fitness value of information. *Oikos* **119,** 219–230 (2010).
6. Shannon, C. E. The mathematical theory of communication. *The Bell Systems Technical Journal* **27,** 379–423 (1948).
7. Farber, R., Lapedes, A. & Sirotkin, K. Determination of eukaryotic protein coding regions using neural networks and information theory. *J Mol Biol* **226,** 471–479 (1992).
8. Fath, B. D., Cabezas, H. & Pawlowski, C. W. Regime changes in ecological systems: an information theory approach. *J Theor Biol* **222,** 517–530 (2003).
9. Hegyi, G. & Garamszegi, L. Z. Using information theory as a substitute for stepwise regression in ecology and behavior. *Behav Ecol Sociobiol* **65,** 69–76 (2011).
10. Vicente, R., Wibral, M., Lindner, M. & Pipa, G. Transfer entropy--a model-free measure of effective connectivity for the neurosciences. *J Comput Neurosci* **30,** 45–67 (2011).
11. Huggins-Daines, D. & Rudnicky, A. I. A constrained Baum-Welch algorithm for improved phoneme segmentation and efficient training. *Ninth International Conference on Spoken Language Processing* (2006).
12. Zhang, X., Wang, Y. & Zhao, Z. A hybrid speech recognition training method for hmm based on genetic algorithm and baum welch algorithm. *IEEE* **IPICIC 2007,** 572–572 (2007).
13. Shlezinger, N., Farsad, N., Eldar, Y. C. & Goldsmith, A. J. ViterbiNet: A Deep Learning Based Viterbi Algorithm for Symbol Detection. *IEEE Trans. Wireless Commun.* 1–1 (2020).
14. Lizier, J. T., Prokopenko, M. & Zomaya, A. Y. Information modification and particle collisions in distributed computation. *Chaos* **20,** 037109 (2010).
15. Tehrani-Saleh, A. & Adami, C. Can Transfer Entropy Infer Information Flow in Neuronal Circuits for Cognitive Processing? *Entropy* **22,** (2020).
16. Flecker, B., Alford, W., Beggs, J. M., Williams, P. L. & Beer, R. D. Partial information decomposition as a spatiotemporal filter. *Chaos* **21,** 037104 (2011).
17. Goodwell, A. E., Jiang, P., Ruddell, B. L. & Kumar, P. Debates—Does Information Theory Provide a New Paradigm for Earth Science? Causality, Interaction, and Feedback. *Water Resour. Res.* **56,** e2019WR024940 (2020).
18. Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM Journal* **January,** (1960).
19. Han, T. S. Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control* **36,** 133–156 (1978).
20. McGill, W. J. Multivariate information transmission. *Psychometrika* **19,** 97–116 (1954).
21. Williams, P. L. & Beer, R. D. Nonnegative Decomposition of Multivariate Information. *bioRxiv* (2010).
22. Samuk, K. *et al.* Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Mol Ecol* **26,** 4378–4390 (2017).

23. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet* **5,** e1000695–11 (2009).
24. Mathew, L. A. & Jensen, J. D. Evaluating the ability of the pairwise joint site frequency spectrum to co-estimate selection and demography. *Front. Genet.* **6,** 268 (2015).
25. Haller, B. C. & Messer, P. W. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution* **36,** 632–637 (2019).
26. De Mita, S. & Siol, M. EggLib: Processing, analysis and simulation tools for population genetics and genomics. *BMC Genetics* **13,** 27–12 (2012).
27. Finn, C. & Lizier, J. T. Pointwise Partial Information Decomposition Using the Specificity and Ambiguity Lattices. *Entropy* **20,** (2018).
28. Makkeh, A., Gutknecht, A. J. & Wibral, M. A differentiable measure of pointwise shared information. *arXiv [cs.IT]* (2020).
29. Doane, D. P. Aesthetic Frequency Classifications. *Am. Stat.* **30,** 181–183 (1976).
30. Scargle, J. D., Norris, J. P., Jackson, B. & Chiang, J. Studies in astronomical time series analysis. Vi. Bayesian block representations. *ApJ* **764,** 167 (2013).
31. Scott, D. W. Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions. *Ann. Stat.* **13,** 1024–1040 (1985).

FIGURES

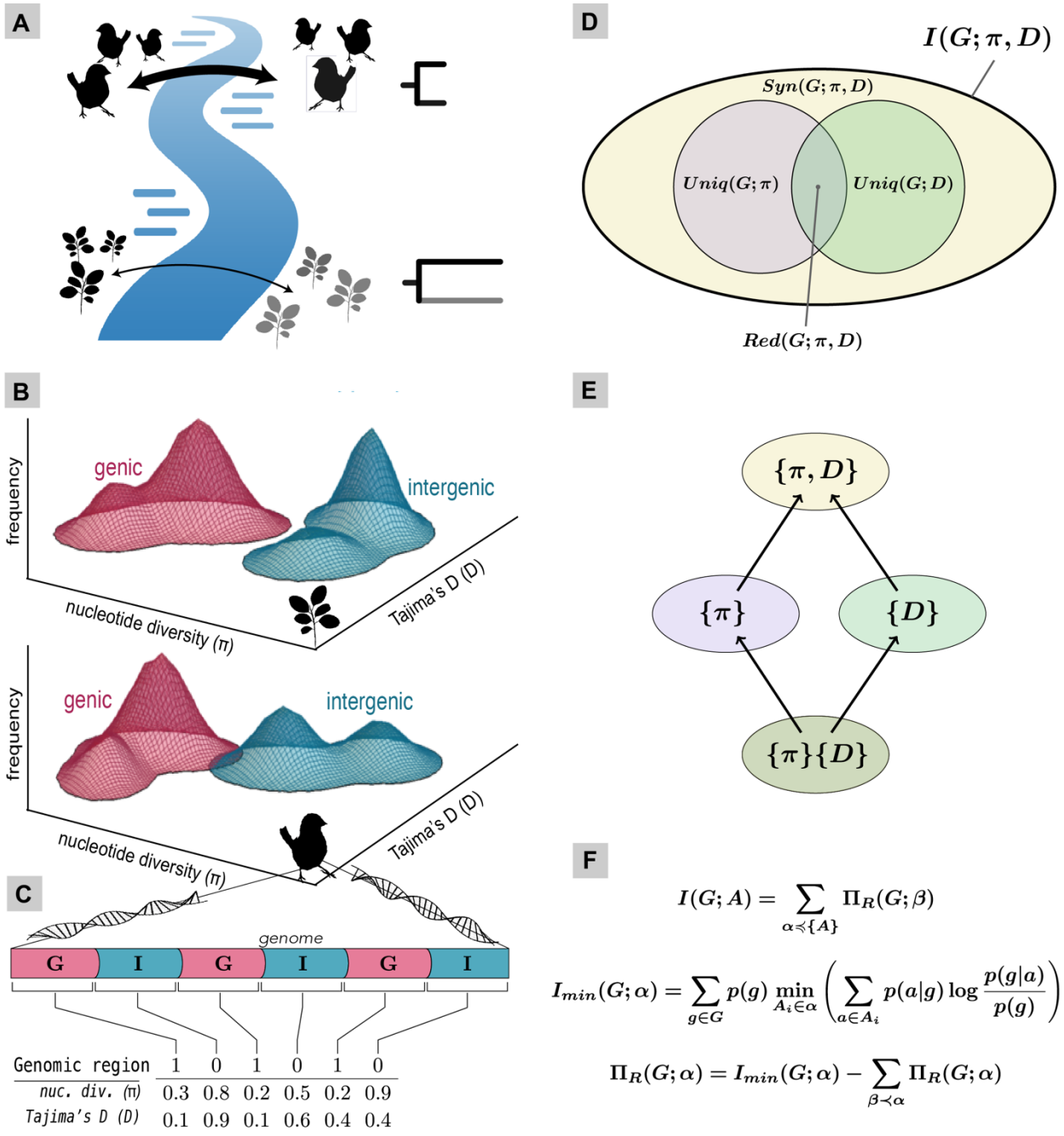


Figure 1 Conceptualization of partial information decomposition (PID) applied to genomic data (gPID). (A) Organisms may be affected differently by the same extrinsic events (e.g., formation of a river). **(B)** Genic and intergenic parts of the genome are differentially shaped by extrinsic (e.g., barrier to gene flow) and intrinsic (e.g., selection) forces that can be quantified with genetic metrics (e.g., F_{ST}, Tajima's D), the values of which will vary by taxon. **(C)** We calculated common metrics (π-nucleotide diversity, D-

Tajima's D) in genic and intergenic regions of the genome on simulated data and **(D,E)** calculated mutual information on these genetic statistics which is decomposed over a lattice into unique, redundant, synergistic information terms given by **(F)**.

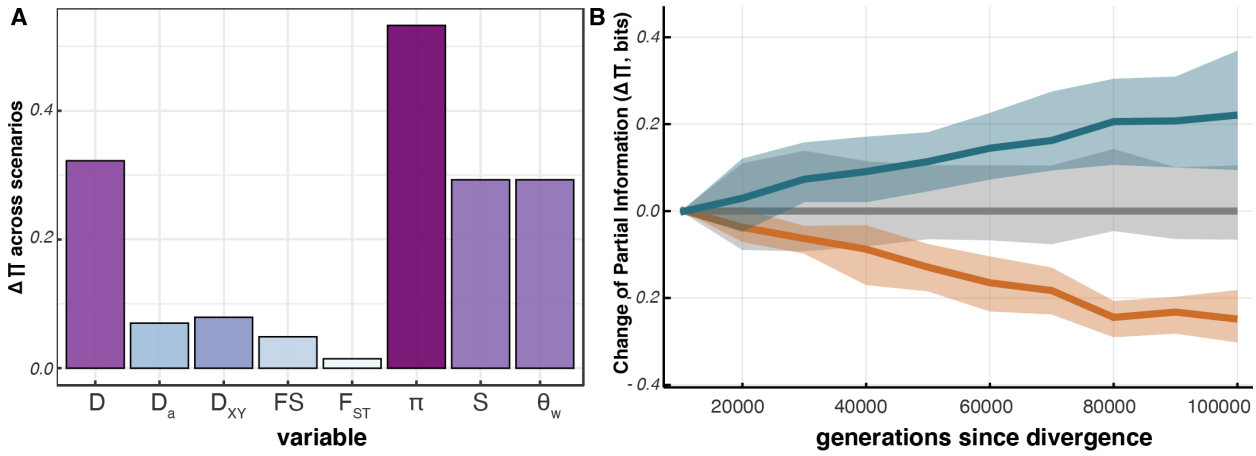


Figure 2. Summary of information content in genomic metrics. (A) Change in information content ($\Delta\prod$) in single-variable decompositions across evolutionary scenarios. A higher value reflects a greater ability for that variable to discriminate among evolutionary scenarios. **(B)** Time series summary of how nodes from the decomposed lattices change over time. Patterns were summarized into seven time-signature clusters using k-means. Three of those are shown here: flat (grey, N=303151), increasing (turquoise, N=368), and decreasing (orange, N=196).

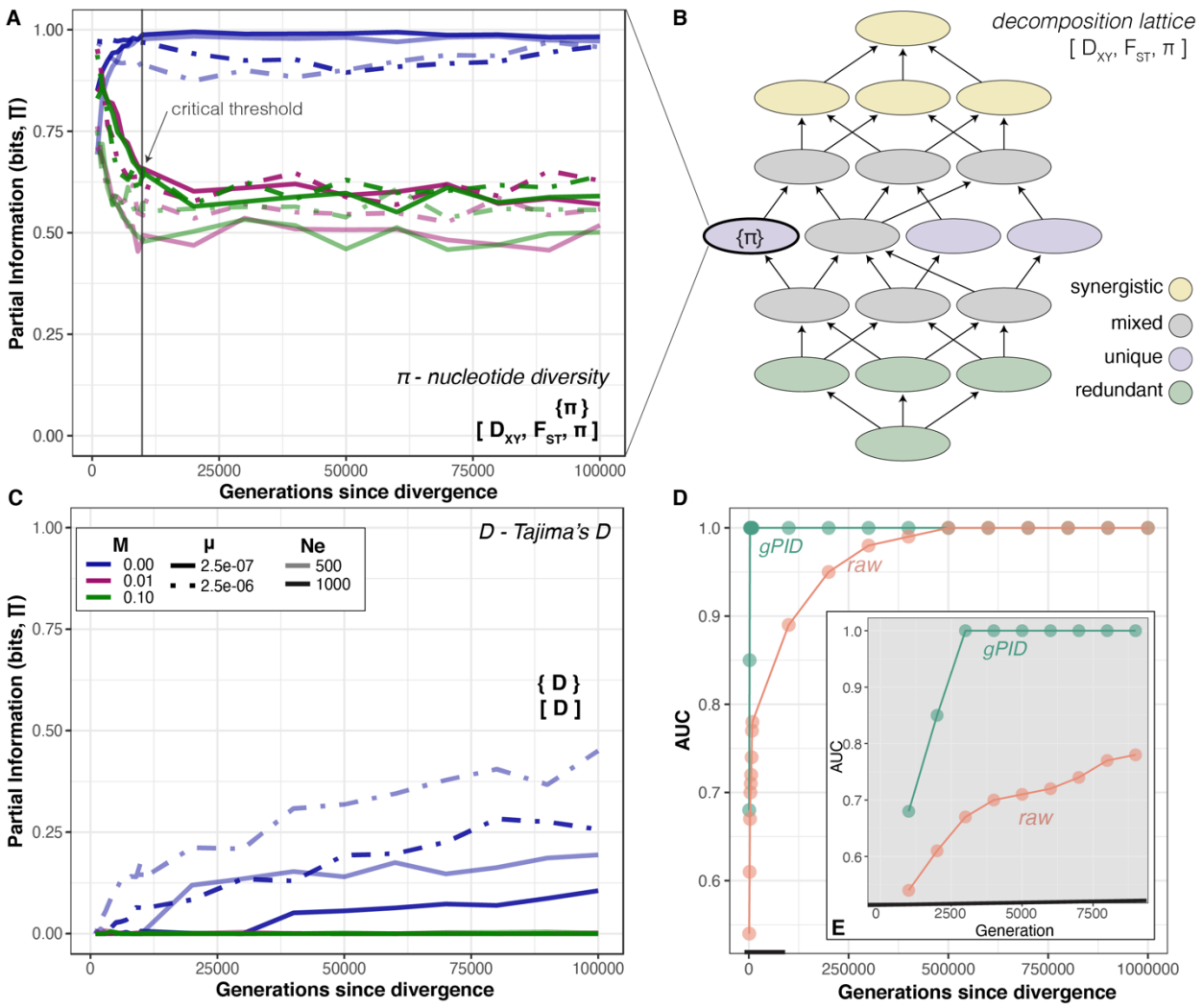
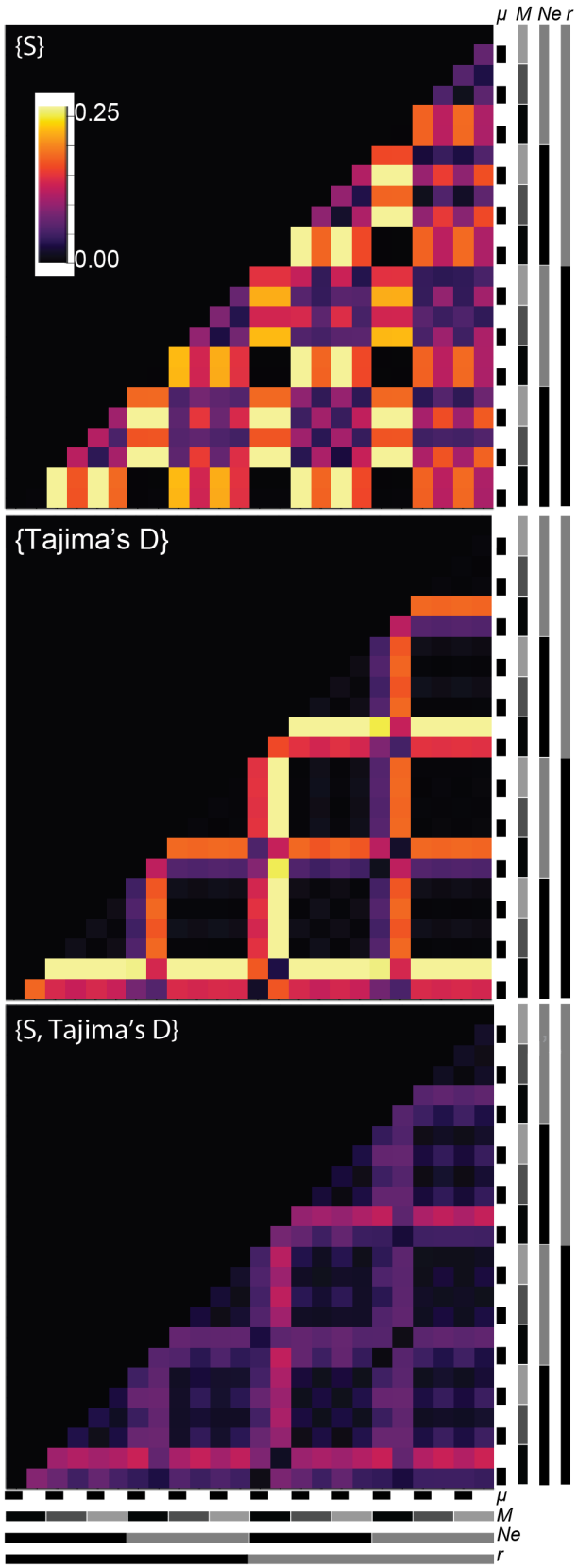


Figure 3. Discrimination of time-series signatures - (A) Partial information for Unique{Pi} from [D_{XY}, F_{ST}, π] over 100,000 generations, which shows the strongest discrimination capability between gene flow (pink, green) and no gene flow (blue) following a critical threshold when the pattern stabalizes. **(B) The 18-node decomposition lattice for** [D_{XY}, F_{ST}, π], from which a single node {π} is plotted in (a). (C) Single variable decomposition of Tajima's D showing increasing partial information for {D} with time since divergence in the absence of gene flow, causing redistribution of information across lattices. **(D)** Comparison of AUC values for a receiver-operator curve (ROC) for the ability of π to discriminate between presence/absence of gene flow using gPID (green) versus raw π values (orange).



454

Figure 4. Additivity behavior of nodes in decomposed lattices. (A) Information variability across evolutionary simulations for single decomposition of segregating sites (S) and **(B)** Tajima's D. Higher values (yellow) indicate stronger discrimination ability. Grey bars at bottom and right are guides for the heatmap pattern expected for each variable simulated (μ - mutation rate, M - gene flow, N_e - effective population size, r - recombination rate). For example, Tajima's D is discriminatory among gene flow whereas S detects both gene flow and mutation rate. **(C)** The information signatures are additive, wherein {S, Tajima's D} displays the additive effects from panels A and B and the signal in heatmap **(C)** is exactly 1/4 the sum of the signals in **(A)** and **(B)**, demonstrating true additivity.