Natural selection driven by DNA binding proteins shapes genome-wide motif statistics

Long Qian[1] and Edo Kussell*[1,2]

[1]Department of Biology and Center for Genomics and Systems Biology, New York University,
12 Waverly Place, New York, New York, 10003, USA
[2]Department of Physics, New York University,
4 Washington Place, New York, 10003, USA
*To whom correspondence should be addressed; e-mail: edo.kussell@nyu.edu

(Dated: February 23, 2016)

Abstract

Ectopic DNA binding by transcription factors and other DNA binding proteins can be detrimental to cellular functions and ultimately to organismal fitness. The frequency of protein-DNA binding at non-functional sites depends on the global composition of a genome with respect to all possible short motifs, or k-mer words. To determine whether weak yet ubiquitous protein-DNA interactions could exert significant evolutionary pressures on genomes, we correlate *in vitro* measurements of binding strengths on all 8-mer words from a large collection of transcription factors, in several different species, against their relative genomic frequencies. Our analysis reveals a clear signal of purifying selection to reduce the large number of weak binding sites genome-wide. This evolutionary process, which we call *global selection*, has a detectable hallmark in that similar words experience similar evolutionary pressure, a consequence of the biophysics of protein-DNA binding. By analyzing a large collection of genomes, we show that global selection exists in all domains of life, and operates through tiny selective steps, maintaining genomic binding landscapes over long evolutionary timescales.

1 Introduction

DNA sequences encode information that is read and interpreted through molecular binding by proteins including transcription factors (TFs), nucleosomes, the RNA polymerase complex, and DNA replication machinery. A DNA binding factor must discriminate a small number of target sites from the set of all possible loci genome-wide. While some DNA binding proteins exhibit exquisite specificity, the majority display fuzziness in their binding preferences [1, 2], and DNA binding often relies on cooperative interactions and chromatin accessibility to increase specificity [3]. Since accessible genomic regions constitute a large amount of DNA (~450 Mb in the human genome) and a substantial proportion of transcribed regions [4], ectopic DNA binding can interfere with multiple processes, including transcription, replication, and nucleosome positioning. Moreover, it titrates copies of DNA binding proteins away from functional sites, reducing the efficiency of gene regulation. Since for any given TF there are exponentially more weak binding sites than strong ones [1, 2], weak ectopic binding to a large number of sites genome-wide is potentially more detrimental to cellular functions than strong ectopic binding which may be comparatively rare. It should therefore be beneficial for genome sequences to evolve to reduce the frequency of non-functional binding sites genome-wide.

Due to the large number of loci involved, such *global selection* on genomes is expected to involve tiny selective coefficients that may be difficult to detect by traditional methods. Indeed, previous studies have identified only a handful of binding motifs that appear to be globally selected against, mainly in bacteria, including promoter elements [5, 6], transcription/translation boundary signals [7, 8], and restriction sites [9, 10]. Other aspects of genome-wide composition have been extensively studied, including global G/C content and codon usage [11, 12, 13, 14, 15, 16]. Seminal work by Karlin and co-workers indicated that dinucleotide frequencies differ between species [17, 18], and more recently, further differences at the level of longer k-mer words have been detected [19, 20]. While these compositional differences could modulate genome-wide binding, there is little consensus on whether mutational biases, drift, or natural selection are their major driving force [11, 12, 21, 22, 23, 14, 24, 25]. It remains largely unknown whether genome sequences have been substantially shaped by DNA binding-related evolutionary pressures.

Here, we demonstrate that the distinct set of DNA binding proteins coded in each species' genome imposes a large set of global, evolutionary pressures that shape genome-wide motif composition. By correlating *in vitro* measurements of DNA binding with genome-wide word statistics, we show that genomes have evolved to reduce the occurrence of weak binding motifs. We introduce an evolutionary model of global selection, and use it to infer selective coefficients and to deduce the evolutionary timescales of global adaptation across all domains of life.

2 Results

2.1 Genomic binding landscapes of transcription factors.

To investigate the impact of DNA binding factors on genomic binding landscapes, we correlated *in vitro* datasets on protein-DNA binding specificities against genomic sequence composition. We used the UniPROBE datasets [26], which are based on a protein-binding microarray that measures the binding of a protein to every possible 8-mer sequence (total $32,896$). We initially studied the mouse dataset [1] in which binding of 109 TFs to each 8-mer on the microarray was determined. For each TF, the log binding intensity values were centered to the median, and normalized by their dispersion, yielding a binding score b_i for every possible 8-mer word i. Fig. 1A shows the distribution of binding scores for a single TF (Mafk) across all 8-mers (see File S1 for all TFs). Words in the positive or negative tails correspond to very strong or very weak binding, respectively. For most TFs, the majority of words lie along a continuum of binding levels without substantial gaps, consistent with previous observations that TFs typically exhibit degeneracy of their target preferences [27, 28].

We analyzed the intron regions of the mouse genome, which constitute $\sim$ 750Mb or nearly 30% of the total genomic DNA. Introns are ideal for detecting binding-related pressures because (i) they are largely devoid of locus-specific selective pressures (as found *e.g.* in exons) which confound detection of global effects, and (ii) they reside in genic regions and are thus generally accessible to binding factors. Simply correlating *in vitro* binding scores with genomic word frequencies, however, is highly misleading due to two effects. First, the nucleotide composition (G/C content) of the genome is a major predictor of word usage, and may be influenced by mutational biases, drift, and selection [22, 24]. TFs that bind words composed of more (less) frequent nucleotides are likely to exhibit positive (negative) correlations between binding scores and genomic k-mer frequencies regardless of evolutionary history. Thus, even if G/C content itself were evolutionarily shaped by binding-related global pressures, correlations between raw word frequencies and TF binding scores cannot be used as evidence. In our analysis, we therefore used *relative* word frequencies, i.e. normalized by expectation based on genome nucleotide composition. Second, the distribution of k-mer frequencies in mouse introns is bimodal due to differences in words' CpG content (Fig. 1B). In vertebrates and plants, the dinucleotide CpG is hypermutable (e.g. in mice its mutation rate is $\sim$10 times the average point mutation rate) causing genome-wide depletion of words that contain it [29]. To account for this large effect, which masks smaller differences among words, we correlated k-mer binding scores vs. relative word frequencies separately for words with different CpG content.

The example in Fig. 1C shows pronounced negative correlations in each CpG category, indicating that the stronger binding a word, the less frequently it is used in the genome relative to expectation. For most TFs, we observed that words with below-average binding ($b < 0$) exhibit highly significant negative correlations (Fig. 1D). For words with above-average binding ($b > 0$), both positive and negative correlations were found, depending on the TF (Fig. S1). Correlating binding scores of all words against their relative frequencies yielded negative correlations in each CpG category for a majority of TFs (Fig. S1). Similar results were obtained in worm (21 TFs), fly (14 TFs), human (8 TFs), and yeast (89 TFs) genomes (Fig. S2, Table S1, and File S1). We conclude that statistically significant correlations exist between binding scores and genomic relative word frequencies, and that in general, weak binding words are avoided compared to even weaker binding words.

2.2 A genomic hallmark of global selection.

We sought a more general method that could be applied in the absence of *in vitro* measurements to detect global selection due to DNA binding-related pressures. We noticed that, consistent with the biophysics of protein-DNA interaction [30], TF binding scores of words that differ at a single nucleotide are strongly correlated: for each word i we plot its binding score b_i vs. the average binding score $\tilde{b}_i$ of its 'mutational neighbors' – all words that differ from i at a single nucleotide (Fig. 1E & File S1). For all 241 TFs that we analyzed, we found a general statistical rule that *similar words have similar binding strengths*. Therefore, if genomes have adapted globally under DNA binding pressures, then we should detect a strong correlation between the frequencies of similar words, since *similar words would be under similar pressures*. In Fig. 1F we show the result for the mouse genome, where for each 8-mer word i, its frequency f_i is plotted vs. the average frequency $\tilde{f}_i$ of its mutational neighbors. Consistent with the hypothesis, we observed a strong correlation of f_i and $\tilde{f}_i$ ($\rho > 0.85$ within each CpG group).

We tested for word-neighbor correlations in a large collection of fully sequenced genomes spanning all domains of life. Word frequencies were measured separately in exon and intron regions, and normalized by using appropriate null models that account for context-dependent mutational biases and other compositional effects. For exons, we used synonymous codon and dicodon shuffling schemes to construct partially randomized DNA sequences that preserves amino acid sequences, genomic codon biases, and nucleotide base composition. The dicodon shuffling scheme additionally preserves the frequencies of all k-mers for $k \leq 4$. The randomized sequences were scanned to determine expected word frequencies for exons. For introns, we computed expected word frequencies based on genome-wide nucleotide (1-mer), dinucleotide (2-mer), or trinucleotide (3-mer) frequencies.

All tested genomes (947 bacterial/archaeal genomes, 1304 eukaryotic chromosomes from 75 species) exhibited striking correlation between f_i and $\tilde{f}_i$ in exons (Figs. 2A, S3, S4, S7, S23) and in introns (Figs. S5, S6, S6B) for each of the null models. Normalized word frequencies correlated strongly between exon and intron regions (Fig. S8). We separately analyzed all DNaseI hyper-sensitive regions of the human genome, which are verified binding-accessible regions, and these exhibited similar word-neighbor correlations (Fig. S22). Eukaryotic genomes were further analyzed on a chromosome-by-chromosome basis. In human chromosomes, for example, the overall shape of the f vs. $\tilde{f}$ plots from exons is qualitatively similar across chromosomes (Fig. 2B & Fig. S9). Comparing relative word frequencies f_i between different chromosomes, we found high correlation coefficients ($\rho > 0.9$) for most chromosome pairs within each genome (Table S4). Deviations were observed for short or Y chromosomes (Fig.

2B & Fig. S10A), due to insufficient word sampling as well as strong genetic linkage, discussed below.

2.3 Mathematical model of global selection.

We examined the consequences of a global selective process acting differentially on words across a genome. We represented a genome by its k-mer frequency vector f_i, where each word i experiences evolutionary pressure according to a selective coefficient s_i. A population of genomes evolves by successive rounds of mutation and selection, where mutations cause random changes to the word vectors with rate u (per bp), and genomes reproduce proportionally to the total fitness of their words. The model admits a unique equilibrium solution in the large population limit, which expresses the stable word frequencies in terms of the selective coefficients and mutation rate (*Supplementary Text*). Conveniently, it is possible to invert this relation, and for small u we obtain

$$s_i/u = k(1 - \tilde{f}_i/f_i) + \text{constant} . \tag{2.1}$$

Selective coefficients relative to u are determined up to an additive constant by the ratio of a word's frequency and the average frequency of its mutational neighbors. Moreover, this relation generalizes to include biased mutation rates (*Supplementary Text*). Fig. 3A-C shows examples of the model solution when selective coefficients are randomly sampled from a normal distribution with standard deviation σ. Words experiencing similar pressures fall on the lines in the $(\tilde{f}, f)$-plane defined by (1) (Fig. 3A). This result encapsulates a basic insight of our analysis: the genome-wide frequency of a word says little if anything about the global pressure it experiences. Words can be under- or over-represented simply because their mutational neighbors are under negative or positive pressures, respectively. Indeed, selective coefficients can only be inferred when a word is viewed relative to its mutational neighbors.

When selection is weak relative to mutation ($\sigma < u$, Fig. 3B) the word cloud collapses toward a point in which all words are effectively neutral, while under strong selection ($\sigma > u$, Fig. 3C) the word with the maximum selective coefficient dominates the distribution. Only in the intermediate regime ($\sigma \simeq u$, Fig. 3A) does the solution take the form of an extended word cloud, with frequencies varying from approximately twofold avoidance to twofold enrichment. A pronounced positive correlation between f_i and $\tilde{f}_i$ is seen, despite the fact that all words experience independent random pressures. This correlation results from selection, which modulates the frequency of neighbor words in order to alter mutational fluxes into words under selection (similar results are obtained with a skewed pressure distribution, Fig. S11). The correlation of frequencies can be increased further by introducing positive correlations in the pressures on similar words, resulting in an extended, rotated word cloud (Fig. 3D,E), while the equal-pressure lines remain unchanged.

2.4 Distribution of global selective coefficients in genomes.

We applied the mathematical model to infer the global selective coefficients in each genome. Word frequencies f_i were counted separately in introns and exons, and expected frequencies f'_i were obtained using the null models described above. We measured global selective coefficients using the *excess pressures* $\Delta s_i \equiv s_i - s'_i$, where s_i and s'_i were separately inferred from f_i and f'_i, respectively, using Eq. 1. By using excess pressures Δs_i rather than s_i, we measure the strength of selection needed to shift word frequencies to their observed values from their expected values based on the null models (*Supplementary Text*). For example, mutational biases such as CpG hypermutation and other less pronounced biases influence the distribution of dinucleotides observed in the genome. By measuring excess pressures relative to the 2-mer null model, we measure only the additional selective pressures that are not already accounted for by dinucleotide composition. In Fig. 4 the distribution of selective coefficients is shown for several genomes (see Figs. S3-S7 for additional genomes and null models). In all cases, the bulk of the distribution has a width comparable to u. The selective coefficients Δs_i measured on different chromosomes were strongly correlated, with overall very similar distributions (Fig. S9). Importantly, Eq. 1 allows us to determine the pressures on each word in spite of pressures acting on their mutational neighbors. Consistent with our primary hypothesis, we find that the selective pressures on similar words are indeed highly correlated (Fig. S13).

2.5 Neutral mechanisms are unable to account for observed word frequency distributions and word-neighbor correlations.

To determine whether non-selective mechanisms, such as mutational biases and repeat expansion, can account for the observations we ran a wide range of tests. Controls for mutational biases included performing an analysis of variance on k-mer frequencies using their dinucleotide and trinucleotide composition (Table S5), analysis of word-neighbor correlations using regression residuals (Fig. S16), a dicodon shuffling scheme that accounts for mutational biases in bacteria (Fig. S23), and explicit incorporation of mutational biases into evolutionary models (Fig. S12). In each of these tests, mutational biases were unable to account for the observed word frequency distributions and word-neighbor correlations (see *Supplementary Text* for a full discussion). Since large eukaryotic genomes have a substantial amount of repeat-derived sequence, we tested whether word-neighbor correlations might arise from a balance between amplification of specific classes of mobile and/or repeat-containing elements and mutational degradation. We analyzed the repeat-masked sequence of the human genome, a procedure that removes approximately 45% of the sequence (Fig. S21), and found that it exhibited strong word-neighbor correlations ($r \geq 0.92$ for all null models, Fig. S21C). Repeat expansion is therefore not responsible for word-neighbor correlations, and word frequencies were strongly correlated between repeat-masked and repeat-derived regions (Fig. S21B). A different possibility is that due to ubiquitous small insertions and deletions (indels) occurring in all genomic regions, word composition could be determined by slippage-based mutational mechanisms, a phenomenon known as 'cryptic simplicity' [31]. While repeat-masking cannot detect this finer-scale process, it is well-known from comparative genomics and mutation accumulation studies in different species that indels occur between 0.03 – 0.13 times as frequently as point mutations [32, 22, 33, 34, 35]. The mathematical model shows that processes that change word frequencies at much slower rates than the point mutation rate cannot yield the observed word-neighbor clouds (Fig. 3B). These results indicate that neither mutational biases nor neutral processes involving repetitive DNA can account for the ubiquitous word-neighbor correlations observed.

2.6 Ancient phylogenetic signal of global selection.

Evidence of a persistent global evolutionary process acting on k-mers can be found in a principal component analysis (PCA) we performed on intronic relative word frequencies (2-mer null) across eukaryotic species (Fig. 5). Chromosomes within a single genome form tight clusters that demarcate species. Although major groups of eukaryotes can be clustered to a certain extent using genome-wide di-, tri-, and tetra-nucleotide frequencies [17, 19] as well as codon biases [12], we achieved a significantly higher resolution using data exclusively from the intron regions of individual chromosomes. Closely related species were spatially proximate, with plants, invertebrates, and vertebrates following different directions in the PCA space (Fig. S17). As discussed below, these remarkable phylogenetic signals are not explained by neutral divergence processes (see Discussion).

Our analysis indicates that primates experienced a 'jump' in global pressures from the rest of the mammals (Fig. S18); while within this group, no species clusters were detected (Fig. 5), indicating that the global pressures have not significantly changed since the last common ancestor. In prokaryotes, phylogenetic signals are generally very weak or non-existent when assessed across all genomes (Fig. S19), but may persist at the genus and species level (Fig. S20), indicating that over the longest evolutionary timescales global pressures can change so extensively that the most ancient parts of the signal are extremely faint and difficult to detect.

2.7 Evolutionary dynamics of global selection.

Our mathematical model is only valid under mutation-selection balance, hence the selective coefficients that we infer (Fig. 4) correspond to the magnitude of purifying selection necessary to maintain genome-wide motif statistics over evolutionary timescales. Since the effect sizes on individual words are tiny, we asked whether the global fitness differences between individuals are sufficiently large to constitute non-negligible selective differences, i.e. larger than $\sim 1/N_{\mathit{eff}}$. If two individuals differ at l sites, or kl words, and each word contributes an average effect size $\pm\bar{s}$, their global fitness difference $\Delta S \sim \bar{s}\sqrt{kl}$. In human populations, pairs of individuals differ at 4.6×10^6 sites on average [36], and taking $k = 6$ and $\bar{s} \sim u$ (Fig. 4), with $u = 10^{-8}$ [37], we have $\Delta S \approx 5 \times 10^{-5}$. Effective global selection thus requires $N_{\mathit{eff}} \gtrsim 10^4$. In *E. coli*, isolates differ on average at 1 – 2% of sites [38], or approximately 7×10^4 sites. Taking

$\bar{s} \sim u$, and using $u = 2.2 \times 10^{-10}$ [33], we find $\Delta S \approx 1.4 \times 10^{-7}$, hence global selection requires $N_{\mathit{eff}} \gtrsim 10^{7}$. The per-word selective coefficients we measured in these genomes are thus sufficient to maintain genome-wide statistics given the present levels of diversity in these populations and previous estimates of their effective population sizes [39].

Since the number of loci affected by global selection is large, in chromosomal regions with low recombination rates Hill-Robertson effects will tend to oppose global adaptation [40] (*Supplementary Text*). Simulations of a non-recombining population initialized at the predicted mutation-selection balance evolved to lower fitness (Fig. S15), while recombination enabled genome composition to be efficiently maintained under global selection (Fig. S14). Similar behavior in a different simulation model has previously been shown [40]. Consistent with this prediction, we found that chromosomes whose word frequencies exhibit deviations from the genome-wide average tend to be non-recombining sex chromosomes or to have smaller genetic length than average (Fig. S10). In the human genome, the genome-wide average number of mutations per crossover per generation is 0.87 [41], while in bacteria homologous recombination replaces small fragments of a genome with homologous fragments from other cells, and occurs with rates per site that are comparable to or greater than u [42], allowing global selection to maintain genome motif composition.

3 Discussion

We presented several lines of evidence indicating that genome-wide word frequencies have been shaped by global selection over long evolutionary timescales. First, we showed that extensive *in vitro* binding data exhibit statistically significant correlations with genome-wide relative word frequencies. These correlations are usually negative, indicating that strength of selection against a word scales with binding strength. Second, we identified a general hallmark of global selection – the strongly correlated frequencies of similar words – which we observed in all genomes. We argued that this correlation results from the biophysics of protein-DNA binding: since similar words have similar binding strengths, the evolutionary pressures they experience should likewise be correlated, resulting in the observed word-neighbor correlations. Third, we analyzed a wide range of null models, which demonstrated that word-neighbor correlations cannot be attributed to mutational biases. Fourth, we introduced an evolutionary model that was used to infer the global selective coefficients of words. Fifth, we analyzed the phylogenetic signal of global selection, which we now discuss.

The persistence of a phylogenetic signal in relative word frequencies over the large evolutionary distances seen in Figs. 5, S17, & S18 cannot be explained by a neutral divergence process. For example, *D. melanogaster* and *M. musculus* diverged over 600 Mya, yet their relative word frequencies (2-mer null) exhibit a correlation of $\rho = 0.43$ (Table S4). Billions of generations separate these genomes, and considering their per bp per generation mutation rates are $\sim 10^{-8}$ [43] on average every neutral position will have mutated one or more times. Since our analysis was performed in introns, most of the ancient signal would have been destroyed leaving essentially no correlation. Furthermore, we showed that random drift does not play a major role in determining word frequencies by separately analyzing the chromosomes within each genome, which would drift independently of each other but instead cluster together in the PCA and exhibit strong correlations (Table S4). We propose that this phylogenetic signal persists because DNA binding motifs are conserved over long evolutionary timescales, and they continue to exert similar global pressures on genomes. While *cis*-regulatory architectures can evolve relatively rapidly, the motif specificities of homologous transcription factors from distant species have been shown *in vitro* to be highly conserved and evolve on a much slower timescale [44].

Our mathematical model, which represents global selection at mutation-selection balance, is appropriate for describing the maintenance of genome-wide word composition over evolutionary timescales. We find the selective coefficients of most words span within one order of magnitude of the point mutation rate u (Fig. 4), and since $u \sim 10^{-10}$ per generation (*E.coli* [33]) or $\sim 10^{-8}$ (humans [37]), we suspect these are among the smallest selective effects that have pervasively contributed to evolution. The model, however, cannot be used to infer the magnitude of selection that has acted periodically to change global composition. These potentially much larger effects – which we speculate would occur over timescales comparable to speciation events due to evolution of TFs with new binding motifs or substantial changes in expression levels of DNA-binding proteins – could be probed by a detailed comparative analysis of lineages that are at different stages of speciation. Given the key role of recombination in this process, further insights into global adaptation rates may result from detailed modeling of the interaction of recombination with global selection in combination with genomic analyses. Here, we showed that without recombination, genomes are only able to partially adapt globally, and cannot maintain genome compositions that have significantly higher global fitness (Figs. S14 & S15), a result which is consistent with analysis of chromosome sizes as well as Y chromosomes (Fig. S10). The presence of global selection should therefore result in a selective advantage for recombination [40].

Our extensive analysis of genomic data demonstrates that global selection is a universal evolutionary force that acts on genomes. We propose that this force arises from the functions of a diverse and distinct set of DNA binding processes within each genome that together generate a characteristic set of global pressures. As global selection maintains genome-wide motif compositions over long evolutionary timescales, its hallmark can be detected in the correlated frequencies of words and their mutational neighbors. Our findings introduce a new view on genomic evolution, in which molecular diversity that is effectively neutral over shorter timescales provides the raw material for global selection acting over much longer timescales.

Materials and Methods

All methods are provided in *Supplementary Methods*.

Supplementary Information

Includes: Methods (1), Text (1), Figures (S1-S23), Tables (S1-S5), and Data File (S1).

Acknowledgements

We wish to thank B. Greenbaum, A. Grosberg, O. Hallatschek, E. Koonin, S. Leibler, B. Levin, M. Lynch, D. Pollock, M. Purugannan, M. Rockman, D. Roos, I. Ruvinsky, E. Shakhnovich, B. Shraiman, S. Small, and J. Weitz for valuable discussions. This work was supported by a Human Frontiers Science Program Young Investigator's Grant.

References

[1] Badis G et al. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324(5935):1720–1723.

[2] Jolma A et al. (2013) DNA-binding specificities of human transcription factors. *Cell* 152(1-2):327–339.

[3] Lelli KM, Slattery M, Mann RS (2012) Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.* 46:43–68.

[4] Thurman RE et al. (2012) The accessible chromatin landscape of the human genome. *Nature* 489(7414):75–82.

[5] Hahn MW, Stajich JE, Wray GA (2003) The effects of selection against spurious transcription factor binding sites. *Mol. Biol. Evol.* 20:901–906.

[6] Froula JL, Pilar Francino M (2007) Selection against spurious promoter motifs correlates with translational efficiency across bacteria. *PLoS ONE* 2:e745.

[7] Itzkovit S, Hodis E, Segal E (2010) Overlapping codes within protein-coding sequences. *Genome Res.* 20(11):1582–1589.

[8] Whitaker WR, Lee H, Arkin AP, Dueber JE (2015) Avoidance of truncated proteins from unintended ribosome binding sites within heterologous protein coding sequences. *ACS Synth. Biol.* 4:249–257.

[9] Gelfand M, Koonin E (1997) Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Res.* 25:2430–2439.

[10] Rocha E, Danchin A, Viari A (2001) Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. *Genome Res.* 11:946–958.

[11] Hartl DL, Moriyama EN, Sawyer SA (1994) Selection intensity for codon bias. *Genetics* 138:227–234.

[12] Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH (2003) Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci USA* 101:3480–3485.

[13] Hershberg R, Petrov DA (2008) Selection on codon bias. *Annu. Rev. Genet.* 42:287–299.

[14] Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of Selection upon Genomic GC-Content in Bacteria. *PLoS Genetics* 6(9):e1001107.

[15] Tats A, Tenson T, Remm M (2008) Preferred and avoided codon pairs in three domains of life. *BMC Genomics* 9:463.

[16] Moura GR et al. (2011) Species-specific codon context rules unveil non-neutrality effects of synonymous mutations. *PLoS ONE* 6(10):e26817–e26817.

[17] Burge C, Campbell AM, Karlin S (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci USA* 89:1358–1362.

[18] Campbell A, Mrazek J, Karlin S (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc Natl Acad Sci USA* 96:9184–9189.

[19] Abe T, Sugawara H, Kanaya S, Knouchi M, Ikemura T (2006) Self-organizing map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes. *Gene* 365:27–34.

[20] Raymond J, Alsop EB (2013) Resolving prokaryotic taxonomy with rRNA: Longer oligonucleotide word lengths improve genome and metagenome taxonomic classification. *PLoS ONE* 8:e67337.

[21] Vetsigian K, Goldenfeld N (2009) Genome rhetoric and the emergence of compositional bias. *Proc Natl Acad Sci USA* 106(1):215–220.

[22] Lynch M (2010) Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA* 107:961–968.

[23] Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet* 6(9):e1001115.

[24] Rocha EPC, Feil EJ (2010) Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? *PLoS Genet* 6(9):e1001104.

[25] Greenbaum B, Cocco S, Levine AJ, Monasson R (2014) Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. *Proc Natl Acad Sci USA* 111:5054–5059.

[26] Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* 43(D1):D117–D122.

[27] Stormo GD (2000) DNA binding sites: representation and discovery. *Bioinformatics* 16:16.

[28] Stewart A, Plotkin JB (2013) The evolution of complex gene regulation by low-specificity binding sites. *Proc Royal Soc B* 280:20131313.

[29] Hodgkinson A, Eyre-Walker A (2011) Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* 12(11):756–766.

[30] Berg OG, von Hippel PH (1986) On the specificity of DNA-protein interactions. *Proc Natl Acad Sci USA* 83:1608–1612.

[31] Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in dna is a major source of genetic variation. *Nature* 322(6080):652–656.

[32] Keightley PD et al. (2009) Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res* 19(7):1195–1201.

[33] Lee H, Popodi E, Tang H, Foster PL (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc Natl Acad Sci USA* 109(41):E2774–83.

[34] Montgomery SB et al. (2013) The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Research* 23(5):749–761.

[35] Zhu Y, Siegal ML, Hall DW, Petrov DA (2014) Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci USA* 111:E2310–E2318.

[36] Consortium TGP (2015) A global reference for human genetic variation. *Nature* 526(7571):68–74.

[37] Roach JC, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328:636–639.

[38] Dixit PD, Pang TY, Studier FW, Maslov S (2015) Recombinant transfer in the basic genome of escherichia coli. *Proceedings of the National Academy of Sciences* 112(29):9070–9075.

[39] Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10:205.

[40] Charlesworth B, Betancourt A, Kaiser V, Gordo I (2009) Genetic recombination and molecular evolution. *Cold Spring Harbor Symposia on Quantitative Biology* 74:177–186.

[41] Jensen-Seaman MI, *et al.* (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* 14:528–538.

[42] Vos M, Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. *Isme J* 3(2):199–208.

[43] Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. *Genetics* 148(4):1667–1686.

[44] Nitta KR et al. (2015) Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife* 4:e04837.

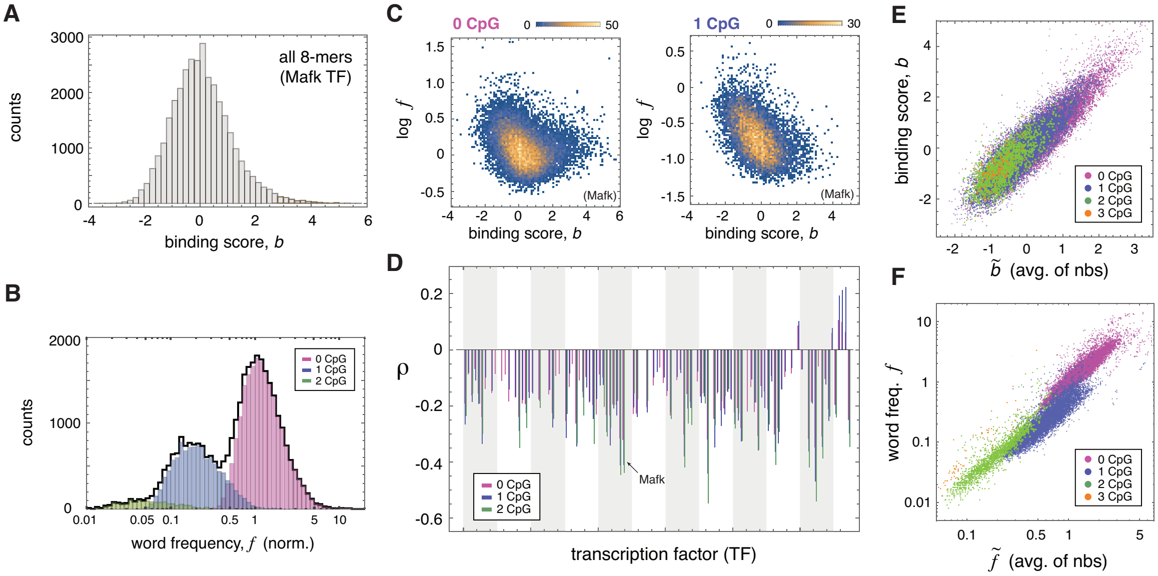


Figure 1: Mouse genomic binding landscape and *in vitro* DNA binding measurements. **A.** Distribution of binding scores b_i for the Mafk TF over all 8-mer words i. **B.** Distribution of 8-mer word frequencies f_i in mouse introns (black curve); f_i are shown normalized with respect to expectation based on genome-wide nucleotide composition. Words are separately histogramed according to their CpG counts (colored bars). **C.** Correlation of b_i and $\log f_i$ for the Mafk TF in separate CpG categories. Color indicates density of points. **D.** Correlation coefficients (Spearman's ρ) of b_i vs. f_i are shown for each mouse TF, using all weak-binding words ($b_i < 0$) separately computed conditioned on word CpG content. Bars are shown only for statistically significant correlations, with p-value $< 10^{-6}$. **E.** Binding scores of words (b_i) are correlated with the average binding score of their mutational neighborhoods ($\tilde{b}_i$); results shown for Mafk, $\rho(b, \tilde{b}) = 0.87$. **F.** Correlation of f_i and $\tilde{f}_i$ over all 8-mer words i for mouse introns; words are colored according to their CpG content, and frequencies are normalized as in **B** & **C**.

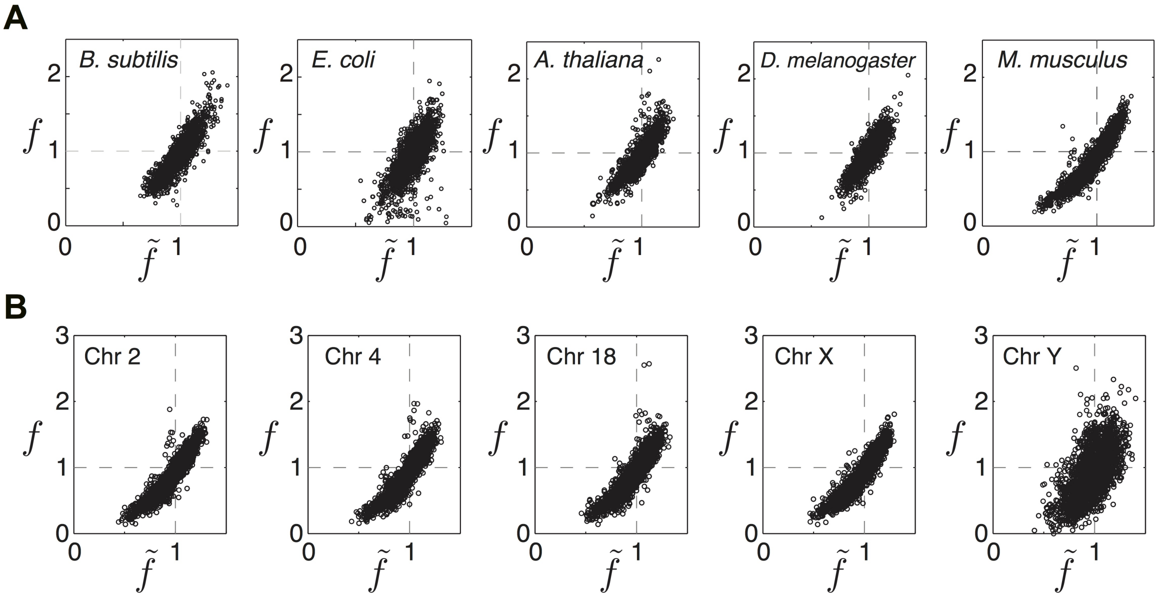


Figure 2: Word-neighbor correlations are a hallmark of global selection. **A**. Word frequency plots of f_i vs. $\tilde{f}_i$ in exons of representative bacterial and eukaryotic genomes. Points correspond to all possible 6 bp words. Frequencies are shown relative to the null expectation from synonymous codon-shuffling, where a value of 1 means the observed frequency is equal to the expected frequency. For eukaryotes, frequencies were computed across all chromosomes. **B**. Word frequency plots from exons of individual human chromosomes (see Fig. S9 for all data).

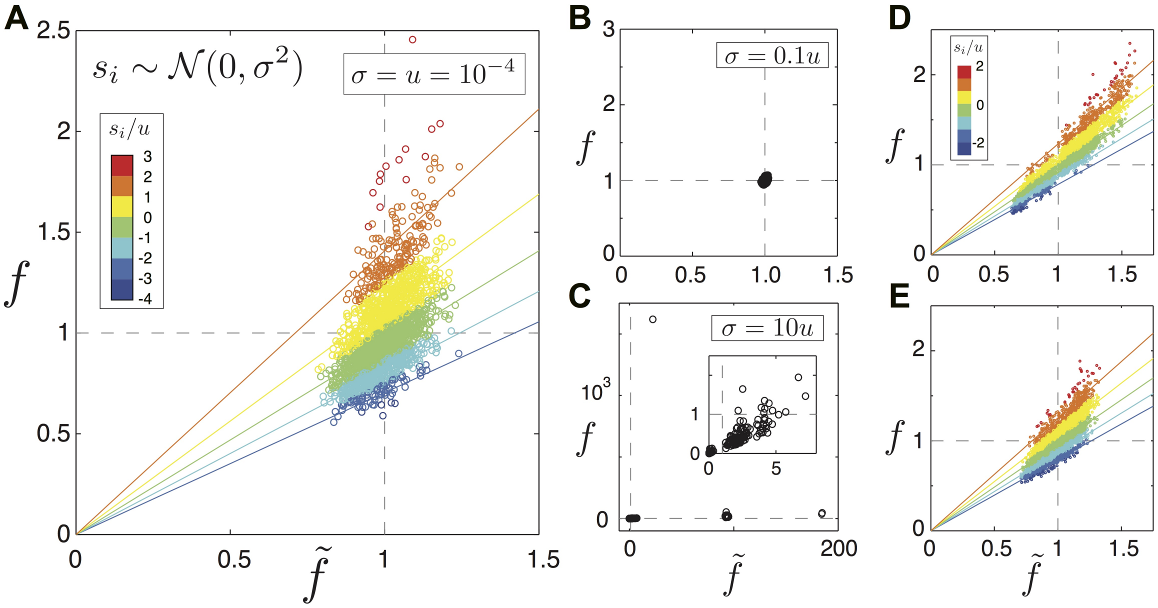


Figure 3: Mathematical model of global selection. **A-C**. Solution of the model at mutation-selection balance. Plots in the $(\tilde{f}, f)$-plane of the equilibrium word frequencies for independent, normally-distributed selective coefficients $s_i \sim \mathcal{N}(0, \sigma^2)$, using the indicated values of σ. We numerically solved the model's eigenvector equation (*Supplementary Text*, Eq. 2.2) to determine the equilibrium word frequencies. Color shows different bins of selective coefficients, and predicted equal-pressure lines are drawn using each interval's average pressure. Inset in C shows a zoomed view. Parameters were $k = 6$ and $u = 10^{-4}$, and frequencies are shown relative to the neutral expectation 4^{-k}. **D**. Solution when using strongly correlated pressures on neighboring words. **E**. Correlation is reduced by shuffling 50% of words' pressures from D.

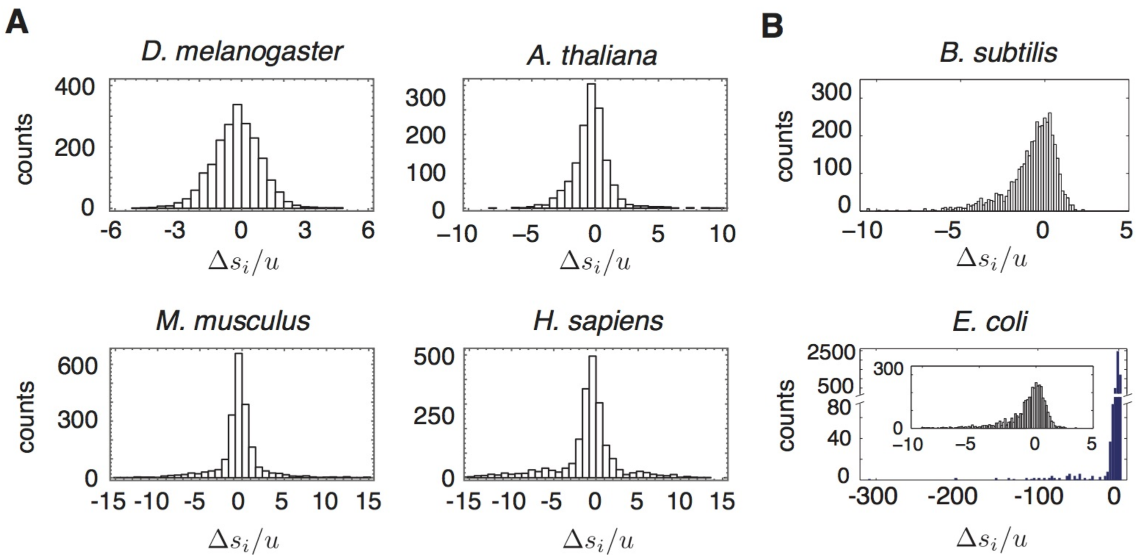


Figure 4: Distribution of global selective coefficients in different genomes. **A.** Selective coefficients of 6-mers in four eukaryotic species. Intron data was used and Δs_i values are given with respect to the 2-mer null model. **B.** Selective coefficients of 6-mers measured in two bacterial species. Exon data was used and Δs_i values are given with respect to the synonymous codon-shuffling model. In *E.coli* an inset shows the bulk of the distribution, since a small number of words including known restriction sites have large negative coefficients.

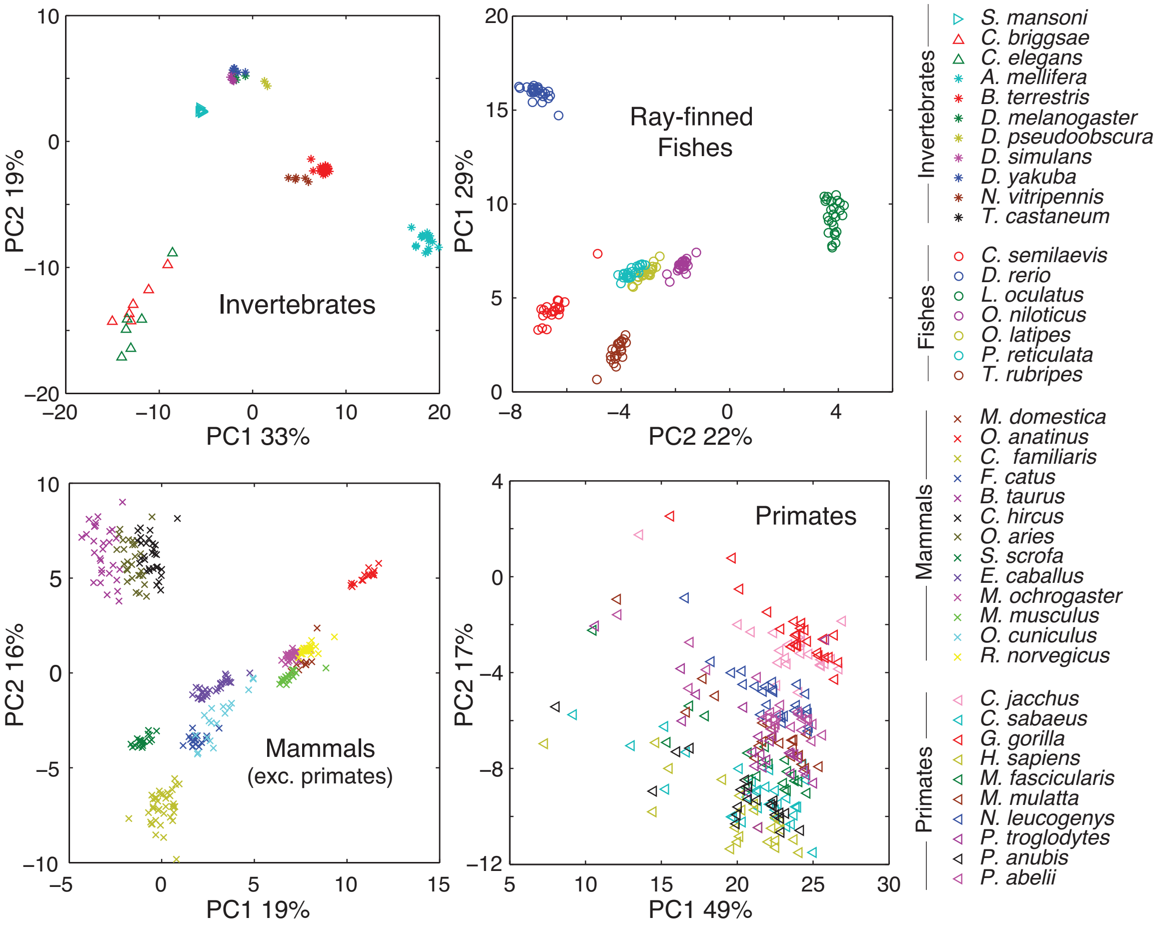


Figure 5: Cross-species comparison of intron word frequencies using principle components analysis (PCA). Each point in the PCA plots corresponds to a single chromosome projected on the two principal axes. Principal axes are computed using the 6-mer relative word frequency vectors (2-mer null model) of all chromosomes shown in each plot. See Table S3 for further species information.

Natural selection driven by DNA binding proteins shapes genome-wide motif statistics

Long Qian[1] and Edo Kussell*[1,2]

[1]Department of Biology and Center for Genomics and Systems Biology, New York University,
12 Waverly Place, New York, New York, 10003, USA
[2]Department of Physics, New York University,
4 Washington Place, New York, 10003, USA
*To whom correspondence should be addressed; E-mail: edo.kussell@nyu.edu.

Contents

1 Materials and Methods

1.1 Genomic sequence datasets

Sequences were downloaded from: ftp.ncbi.nlm.nih.gov/genomes/. The list of genomes for bacteria and archaea is found in the directory /GENOME_REPORTS/prokaryotes.txt, and for eukaryotes in the directory /GENOME_REPORTS/eukaryotes.txt. Bacterial plasmid sequences were not used. We manually curated a large, representative set of genomes across all bacterial and archaeal groups (947 genomes). For eukaryotes, all fully assembled and annotated genomes were included in the intron analysis (75 species, 1,304 chromosomes; mitochondria and plastid sequences were not included; see Table S3). Among eukaryotes, exon analysis was performed for 17 species (251 chromosomes; see Table S3). To extract the intron segments, CDS coordinates were obtained from the annotation files, and then mapped to the full length genome. For both exon and intron extraction, only one of the splicing isoforms was selected at random.

1.2 Measuring genomic word statistics

Word frequencies were collected by a sliding window of k-bp ($k = 2, \ldots, 8$) across the sequences. For the coding region analysis, frequencies were collected on each open reading frame (i.e. multiple exons were joined for eukaryotes) and then combined for all open reading frames on a chromosome. Start and stop codons were ignored. For the intron regions, each intron segment was read separately and then the frequencies were combined. Gaps (N tracts in the eukaryotic sequences), if encountered, were replaced randomly by four bases at equal frequencies. Word frequencies were counted on the plus strand of DNA, and the counts of reverse-complement words were combined.

1.3 Constructing the null models in exon and intron sequences

For the coding region analysis using codon shuffling as the null model, synonymous codons were shuffled across the entire chromosome sequence. Start and stop codons were ignored. Each shuffled sequence was then scanned as above. The expected word frequencies f'_i were calculated as the average of word counts from 1,000 such shuffled sequences. For the intron analysis using nucleotide base composition as the null model (1-mer null), base compositions were calculated on the chromosomal scale. For a word w with N_A, N_T, N_C, N_G counts of A, T, C, and G bases, respectively, the expected frequencies were calculated as $f'_w(\text{1-mer}) \equiv p_A^{N_A} p_T^{N_T} p_C^{N_C} p_G^{N_G}$, where p_A, p_T, p_C, p_G are base compositions measured from the full length sequences. For analysis using the dinucleotide (2-mer null) or trinucleotide (3-mer null) models, a Markov chain model was used to compute expectations. A k-mer word $w = (w_1 w_2 w_3 \ldots w_k)$ is composed of an overlapping set of dinucleotides $(w_1 w_2)$, $(w_2 w_3)$, $\ldots$, $(w_{k-1} w_k)$. The probability of observing each dinucleotide (ab) was measured across all introns conditional on a being at the first position, and denoted $p(b|a)$. Using these measured values, the expected frequency of the word w is given by

$$f'_w(\text{2-mer}) \equiv p_{w_1} p(w_2|w_1) p(w_3|w_2) \cdots p(w_k|w_{k-1}) \ ,$$

where p_{w_1} is the measured frequency of nucleotide w_1. Similarly, for the trinucleotide model, the word is composed of an overlapping set of trinucleotides, and

$$f'_w(\text{3-mer}) \equiv p_{w_1 w_2} p(w_3|w_1 w_2) p(w_4|w_2 w_3) \cdots p(w_k|w_{k-2} w_{k-1}) \ ,$$

where $p(c|ab)$ is the probability of observing trinucleotide (abc) conditional on the dinucleotide (ab) occupying the first two positions; and $p_{w_1 w_2}$ is the measured frequency of dinucleotide $(w_1 w_2)$.

1.4 Analysis of variance (ANOVA) of genomic word frequencies

To determine the extent to which the dinucleotide and trinucleotide composition of genomes can explain k-mer frequencies, we performed a multi-factor ANOVA of word frequencies vs. their dinucleotide and trinucleotide composition. The raw frequencies of 6-mer and 8-mer words were measured in introns, and their relative frequencies were computed under the 1-mer, 2-mer, and 3-mer null models. For the dinucleotide ANOVA, the composition of each word was recorded as a 10-dimensional vector, indicating the counts of each of the following dinucleotides, identifying reverse-complement pairs as a single factor: (AC/GT, CA/TG, AT, TA, CT/AG, TC/GA, CG, GC, AA/TT, CC/GG). For the trinucleotide ANOVA, the composition of each word was given by a 32-dimensional vector, indicating the counts of each possible trinucleotide, identifying reverse-complement pairs as a single factor.

Since dinucleotides and trinucleotides overlap within any k-mer word ($k \geq 4$), there exists a relation among the vector components which causes a singularity in performing regression analysis or ANOVA. A procedure to remove such degeneracies is known as differencing, in which two components are replaced by a single component that measures their difference. For dinucleotide analysis, we replaced the components corresponding to AT and TA by a single component that measures the difference in counts of these two dinucleotides. For trinucleotide analysis, we replaced the components corresponding to CGA and TAA by a single component that measures the difference in counts of these two trinucleotides. The resulting regression is then non-singular, and we verified that the results were not sensitive to the choice of differencing components.

Using the dinucleotide analysis, we also measured the fraction of the total variance explained by the CpG mutational neighborhood, which consists of four dinucleotides: CG, CT/AG, CA/TG, CC/GG. This was computed by performing the ANOVA using the four factors in the CpG mutational neighborhood.

1.5 Principle components analysis of genomic word frequencies

Principle component analyses were performed on the correlation matrix between relative word frequencies using the 2-mer null model. For eukaryotic intron regions (Figs. 5, S17, S18), relative word frequencies of 1,049 chromosomes from 61 species were used for the PCA (Fig. S17 legend and Table S3). Excluded were chromosomes that were shorter than 1 Mb or having < 0.8 correlation between the measured relative frequencies or excess pressures of reverse complement words on the plus strand, as well as the chromosomes of the bird *Ficedula albicollis* because its word frequencies take on a split distribution in the $(\tilde{f}, f)$ plane. The word frequencies on these chromosomes were so extreme that they dominated the PCA analysis. The plots for subgroups shown in each panel were produced by PCA on each subgroup alone. For bacterial coding regions (Fig. S19 & S20), relative word frequencies were computed using codon shuffling as the null model, and counts were combined on both strands of all analyzed exon sequences.

1.6 Simulation of global selection

In each simulation the population consisted of N word composition vectors $f^{(1)}, \ldots, f^{(N)}$, with $f^{(i)} = (f_1^{(i)}, f_2^{(i)}, \ldots, f_{4^k}^{(i)})$ giving the k-mer word counts in sequence i, and $\sum_j f_j^{(i)} = L$. The populations were initiated from random sequences by sampling N times from a multinomial distribution Mult(f^{rand}, L) with $f^{rand} = (1/4^k, \ldots, 1/4^k)$. Populations were propagated through rounds of Wright-Fisher reproduction, with each round consisting of (1) mutation, (2) recombination, and (3) selection. During the mutation step, for each sequence i, the total number of mutated words was sampled from a Poisson distribution with mean kuL. Each mutation caused a jump from word w to word w', where w was sampled based on the composition $f^{(i)}$, and w' was sampled uniformly from the mutational neighbors of w. With each jump, $f^{(i)}$ is updated, and the next mutation was sampled based on the updated $f^{(i)}$. During recombination random mating was implemented in the population, and for each mating pair i and j a fixed fraction of the sequence L corresponding to a total of l words was exchanged between $f^{(i)}$ and $f^{(j)}$: $f^{(i)} \to f^{(i)} + \Delta f^{(j \to i)} - \Delta f^{(i \to j)}$ and $f^{(j)} \to f^{(j)} + \Delta f^{(i \to j)} - \Delta f^{(j \to i)}$, where $\Delta f^{(j \to i)}$ and $\Delta f^{(i \to j)}$ were sampled from Mult$(f^{(j)}, l)$ and Mult$(f^{(i)}, l)$, respectively. During selection, the fitness of each sequence i was calculated as $e^{s \cdot f^{(i)}}$, and we define $p_i \equiv e^{s \cdot f^{(i)}} / \sum_j e^{s \cdot f^{(j)}}$ to be the probability distribution for selecting sequences to populate the next generation. The next generation of composition vectors is then obtained by a single sampling from Mult(p, N). To introduce bottlenecks, all steps were the same except that during population size transitions, from $N_{large} \to N_{small}$, or from $N_{small} \to N_{large}$, the sequences for the next generation were selected randomly from Mult(p, N_{small}) and Mult(p, N_{large}), respectively.

2 Supplementary Text

2.1 Mathematical Model of Global Selection

We use a generalization of the well-known quasi-species model [1], applied to the evolution of genome-wide word frequencies. We summarize the basic formulation of the model, which follows from our previous work [2], and then show how to invert the solution to obtain the exact mapping of observed frequencies to selective coefficients, which is used throughout the main text. Lastly, we discuss the simulation results on finite population sizes, Muller's ratchet, recombination, and bottlenecks.

2.1.1 Model derivation

A sequence of length L is represented by a k-mer composition vector $\mathbf{f}$ in the projected space $\mathcal{L} = \{\mathbf{f} \mid f_i \geq 0, \sum_{i=1}^{n} f_i = L\}$, where $n = 4^k$ is the total number of possible words, and f_i is the number of occurrences of word i in the sequence. We denote by $\mathbf{s}$ the selection vector whose i-th component is the selective coefficient s_i acting on each word of type i. We consider a large population of evolving sequences, each represented by a composition vector. At each generation, each sequence contributes offspring proportional to its total fitness, $\exp(\mathbf{s} \cdot \mathbf{f})$. The offspring mutate according to a mutational transfer operator, $G(\mathbf{f} \mid \mathbf{f}')$, which gives the probability of a mutational transition $\mathbf{f}' \to \mathbf{f}$. Under these dynamics, the expected frequency of $\mathbf{f}$ within the population at generation t, denoted by $P_t(\mathbf{f})$, evolves according to the equation

$$\lambda_t P_{t+1}(\mathbf{f}) = \sum_{\mathbf{f}' \in \mathcal{L}} e^{\mathbf{s} \cdot \mathbf{f}}\, G(\mathbf{f} \mid \mathbf{f}') P_t(\mathbf{f}') \tag{2.1}$$

where λ_t is a normalizing factor, which measures the average population fitness at generation t, given by summing both sides over all values of $\mathbf{f}$. The irreducibility of the transfer operator G guarantees (by the Perron-Frobenius theorem) that the linear mapping of $P_t \to P_{t+1}$ converges to a unique steady-state distribution, $P(\mathbf{f})$, which dictates the population structure at mutation-selection balance.

Since each locus along DNA mutates independently, the transfer operator can be decomposed as a product over all loci, which implies that $P(\mathbf{f})$ is a multinomial distribution [2]. A multinomial random variable $\mathbf{x}$, denoted $\mathbf{x} \sim \mathrm{Mult}(\mathbf{p}, L)$, results when sampling L independent events from a discrete probability distribution p_i, where x_i is the number of outcomes of type i. With this notation, we have $\mathbf{f} \sim \mathrm{Mult}(\mathbf{p}, L)$, where $\mathbf{p}$ is the single-locus distribution of words at mutation-selection balance. This distribution $\mathbf{p}$ is obtained as the unique positive solution of the eigenvalue problem

$$e^S A \mathbf{p} = c\, \mathbf{p}\,, \tag{2.2}$$

where $S = \mathrm{diag}(\mathbf{s})$, and A_{ij} is the probability that word j mutates into word i. Once more the Perron-Frobenius theorem guarantees a unique positive eigenvector $\mathbf{p}$ with associated eigenvalue c, which determines the equilibrium average fitness, $\lambda_t \to \lambda = c^L$.

The above eigenvalue problem can be solved numerically, e.g. to obtain the plots in Fig. 2. To invert the relationship, and obtain an expression for s_i given $\mathbf{p}$, we use the fact that the point mutation rate u (per bp per generation) is extremely small. We express A explicitly as $A = e^{uk(M-I)}$, where I is the identity matrix and M_{ij} is the probability that a single mutation converts word j into word i; M_{ij} is non-zero only for words i and j that are mutational neighbors. In the following analysis, we assume unbiased mutation, i.e. $M_{ij} = 1/(3k)$ for any pair of mutational neighbors i and j; generalization for biased mutations is straightforward, discussed below. We expand A in small u to first order, i.e. neglecting contributions from double (or higher order) mutants that occur within a single word locus. Subsituting $A \approx I + u\,k(M - I)$ in (2.2) yields,

$$c\,e^{-s_i}p_i \approx p_i + u\,k\left(-p_i + \frac{1}{3k}\sum_{j\in\mathcal{N}_i} p_j\right)\,, \tag{2.3}$$

where $\mathcal{N}_i$ is the set of mutational neighbors of word i. The second term within the parentheses is $\tilde{p}_i$, the average frequency of mutational neighbors. Dividing both sides by p_i and taking logarithms, we find

$$s_i \approx u\,k(1 - \tilde{p}_i/p_i) + \log c\,, \tag{2.4}$$

which after division by u yields equation (1) of the main text. This expression shows that the values of s_i are determined by $\tilde{p}_i/p_i$ up to an additive constant. In the main text, pressure distributions are shown using a value of $c = 1$, which corresponds to taking the average global fitness of a sequence to be zero.

Known mutational biases M_{ij} can easily be incorporated in the above, by defining $\tilde{p}_i$ to be a weighted average of the mutational neighbors:

$$\tilde{p}_i \equiv \sum_{j\in\mathcal{N}_i} M_{ij}p_j. \tag{2.5}$$

Since our goal was to survey a large number of species across all domains of life, most of which have not been characterized as far as mutational biases, we developed the excess pressure method, which correctly infers selective coefficients when mutational biases are not known (see next section).

We note that our mathematical model ignores the local, overlapping structure of words within a genome, effectively representing the genome as a ‘bag of words’. This approximation is fully justified because the operation of shifting a k-mer window by one position along a DNA sequence results typically in a jump to a very different word. The only exceptions occur within a very small subset of loci that contain short sequence repeats of size $\geq k$.

2.1.2 Finite Populations and Muller's Ratchet

The above analysis is strictly valid only in the infinitely large population limit. For finite populations and values of $L \gg 1$, deviations from this equilibrium can occur due to Muller's ratchet [3, 4]. This result has been discussed in [5], and we briefly review the same reasoning here. Due to the large genotype space, the maximally fit sequence – i.e. one in which the most favorable word is present at each locus – is rapidly lost in the population, because the rate of mutation away from this sequence, Lu, is much larger than the rate of back-mutation, u. Within a few generations, due to the finite population size, every sequence has accumulated at least one deleterious word, and the fittest sequence in the population is now one which carries at least one mutation. This process drives the population away from the fittest sequence, and reaches an equilibrium distribution some distance from $P(\mathbf{f})$, which depends in a complex manner on N, L, and u [5]. Similarly, in genomic regions with low recombination rates, interference effects of various kinds (known as Hill-Robertson effects; see [6]) will oppose global adaptation. Recombination, however, provides an accessible route to reverse the ratchet effect, by enabling in each generation the previously fittest sequence to be recovered. Our simulations demonstrate that the equilibrium $P(\mathbf{f})$ is indeed achieved in a finite size, recombining population (Figs. S14,S15).

2.2 Mutational Biases and Genomic Word Composition

In this section we analyze the role of mutational biases for the word frequency statistics of genomes. Direct measurements of mutational biases are performed by mutation accumulation experiments in which populations are propagated over many generations under conditions that minimize the strength of selection, typically by frequent passage through strong bottlenecks. Such studies have been performed in *E.coli* [7], *B. subtilis* [8], *S. cerevisiae* [9, 10], *C.elegans* [11], *D. melanogaster* [12], *A. thaliana* [13], and other species [14]. Additionally, measurements using known pedigrees have been made in humans [15, 16]. Mutational biases at the single nucleotide level are characterized by a mutational transition matrix among the basepairs (1-mers) AT, TA, GC, and CG, with six possible mutation types: AT → GC, GC → AT, AT → TA, GC → TA, AT → CG, and GC → CG. The rates of these mutation types vary within about one order of magnitude and differ between species. For example, in *H. sapiens* the rate of GC → AT is about 6 larger than AT → CG, while in *A. thaliana* it is about 14 times larger (see [15] Table 2). These rates are context independent, since they are measured at the single nucleotide level without considering neighboring nucleotides.

Context-dependent biases have also been measured in different species. The strongest known bias in vertebrates is due to CpG methylation, which increases the overall rate of mutation at CpG dinucleotides by about tenfold on average, and which varies across species [17]. In *E.coli*, different dinucleotide contexts were found to affect mutation rates, particularly for transitions [7]. In *B. subtilis*, trinucleotide contexts were assessed, and while dinucleotide contexts were found to account for most context dependencies, certain trinucleotide contexts were statistically significant [8]. In yeast, where mutation accumulation experiments have been the most extensive, measurements on trinucleotide contexts found that a small number of the 32 possible trinucleotide contexts explain most of the variance in mutation rates [9]. Among the 16 trinucleotide contexts for AT basepairs, no statistically significant differences were detected, while among the 16 contexts for GC basepairs two contexts (CCG/CGG and TCG/CGA) had elevated rates by a factor of two. Most of the rate variability could therefore be accounted for by two bits of data: the nucleotide identity of the central basepair (AT or CG), and whether or not the nucleotide is in either of CCG/CGG or TCG/CGA contexts.

2.2.1 Word frequency statistics

It is well-known that the equilibrium frequency of A/T nucleotides predicted based on context-independent mutational biases deviates significantly from the observed A/T genome composition at silent sites in many species (see e.g. [18, 19, 15]). To determine whether context-dependent mutational biases can explain the k-mer statistics of genomes, we performed the two separate analyses described next.

First, we constructed Markov chain null models of genomic sequences in which the frequency of a nucleotide was conditional on the previous nucleotide (2-mer null) or the previous dinucleotide (3-mer null) (see Methods, Sec. 1.3). Under the null hypothesis that dinucleotide or trinucleotide context-dependent mutational biases account for observed word frequencies, the frequency of any k-mer with $k > 3$ will be determined by the conditional probabilities of its constituent 2-mers or 3-mers. Under this null hypothesis, for each genome we parameterized the 2-mer and 3-mer null models using all intron sequence, and computed the expected frequency f_i' of each word i. For each analyzed genome, given total intron length L, we recorded the observed counts of each word i as $C_{i,obs}$, and computed the expected counts of each word, $C_{i,exp}$ = Lf_i', and its variance $\sigma_i^2 = Lf_i'(1 - f_i')$ under each null model. We then computed the z-score of each word as $z_i = (C_{i,obs} - C_{i,exp})/\sigma_i$. Given the large values of L for each genome, the distribution of z-scores is expected to be approximately normally distributed with mean zero and variance one. Instead, we observed that the distribution of z-scores spans 100's of standard deviations under both null models, indicating that the statistics of higher k-mers cannot be explained by context-dependent mutational biases (Figs. S6A & S6B).

Second, we performed a multi-factor analysis of variance (ANOVA) of word frequencies against their dinucleotide or trinucleotide composition. This analysis was designed to measure the fraction (r^2) of the total variance of word frequencies that could be explained by dinucleotide and trinucleotide composition, and hence by context-dependent mutational biases. Since the analysis is equivalent to regression of word frequencies on di- and tri-nucleotide factors, the r^2 value measures the best possible fit of the model to the data, and should thus be considered an upper bound on its explanatory power. For example, in the trinucleotide null model, the regression involves 32 factors, but in reality we know that far fewer trinucleotide contexts exhibit significant mutational biases (see above). The regression model will therefore over-fit the data, hence the calculated r^2 will be higher than its true explanatory capacity.

The ANOVA results are given in Table S5, performed using raw or relative word frequencies (either 6-mer or 8-mer) vs. the dinucleotide or trinucleotide factors. The ANOVA on 8-mer frequencies relative to the genome nucleotide composition (indicated in the f (1-mer) column), shows that for mouse and human genomes, respectively 34% and 29% of the variance is accounted for by dinucleotide composition, while these numbers increase only slightly to 39% and 33% for trinucleotide composition. This leaves more than 60% of the variance unexplained by these compositional models. Normalizing 8-mer frequencies relative to the 2-mer null model, and performing the ANOVA on trinucleotides (shown in the f (2-mer) column), we find the trinucleotide composition explains only 1% and 8% of relative word frequencies in mouse and human genomes, respectively. Thus, the increase of model complexity from dinucleotides to trinucleotides adds 22 degrees of freedom in regression analysis, but results in little additional explanatory power. In both of these genomes, the CpG mutational neighborhood (i.e. dinucleotides within 1 mutation of CpG) accounts for most of the dinucleotide signal, explaining 32% of the variance in mouse and 26% in human (see f (1-mer) column). Explained variance exhibits a wide range across genomes, where dinucleotides explain as little as 15% of the variance in *C.elegans* to as much as 55% in *G. gallus*, with similar trends observed for trinucleotides. A larger proportion of the variance can be explained for 6-mer frequencies, which is likely due to the much larger number of 8-mers (32,896) vs. 6-mers (2,080), since over-fitting by the regression becomes more difficult for the 8-mer data.

The ANOVA analysis also provides a simple consistency check on the Markov chain null models described above. The very small r^2 values given in the f (2-mer) column of the dinucleotide ANOVA indicate that normalization by expected frequencies computed according to the Markov chain dinucleotide model (Methods, Sec. 1.3) effectively accounts for most of the dinucleotide signal in the word frequency data. And, similarly, for the f (3-mer) column of the trinucleotide ANOVA.

We conclude that context-dependent mutational biases are unable to explain the wide distribution of word frequencies we observed. The fraction of the variance that is attributable to mutational biases is mainly due to dinucleotide biases, and largely dominated by the CpG mutational neighborhood in vertebrates and plants. For this reason, we used the 2-mer null model when presenting results on selection coefficients and PCA in the main text.

2.2.2 Word-neighbor correlations

The word-neighbor correlations we detected in all genomes were calculated using five different null models in several different genomic regions: in exons, synonymous codon shuffling (Figs. 2A, S3, S4, S7), synonymous dicodon shuffling (Fig. S23); in introns, 1-mer null (Fig. S5), 2-mer null (Fig. S6A), and 3-mer null (Fig. S6B); in repeat-masked intronic regions, 1-mer and 2-mer null (Fig. S21); and in DNAse I hypersensitive regions, 1-mer and 2-mer null (Fig. S22). The ANOVA results above show that our normalization by 2-mer and 3-mer expected frequencies effectively removes correlation with dinucleotide and trinucleotide composition. To separately verify that the word-neighbor correlations are not accounted for by 2-mer and 3-mer composition, we regressed out their compositional effects and examined word-neighbor correlations of the residuals (Fig. S16). The strongly correlated residuals we observed provide further evidence that word-neighbor correlations do not result from context-dependent mutational biases.

In bacteria, where genomes are dominated by exon regions, our most stringent test is the synonymous dicodon shuffling scheme, which swaps synonymous codons while preserving the coding sequence as well as the genome's dicodon composition. This means that the compositional biases on all k-mers for $k \leq 4$ are preserved. Any remaining signal therefore cannot be attributed to context-dependent mutational biases. In Fig. S23, we see that highly significant word-neighbor correlations persist in word frequencies normalized according to dicodon shuffling statistics. It is important to note that this shuffling scheme is overly conservative, and it constrains sequences much more than they would be in nature. Despite this fact, the excess pressures inferred based on the dicodon shuffling statistics (Fig. S23) span a range of width about $5u - 10u$ in the bacteria we examined indicating that lots of global pressures exist on word with $k \geq 5$ which cannot be adequately captured by the word statistics for $k \leq 4$.

2.2.3 Excess pressures, mutational biases, and global selective coefficients

Since mutational biases have been measured in only a small number of species, we developed a method that can accurately infer global selective coefficients without requiring measurements of mutational biases as input. We therefore defined the *excess pressure* to be the selective pressure required to shift the word frequencies away from their expectation based on a given null model. To do this we determine a set of effective selective coefficients s'_i that would maintain the genome's statistics at the expectation based on the given null model. We use an unbiased transition matrix and infer the values s'_i from the expected frequencies f'_i using Eq. 1. Next, we calculate the selective coefficients s_i needed to maintain the genome's actual statistics at the observed word frequencies f_i, again for an unbiased transition matrix, i.e. using Eq. 1. The excess pressures Δs_i are defined as the difference $s_i - s'_i$. In this way, unknown mutational biases are represented using effective selective coefficients s'_i, whose contribution is subtracted off from the measured values s_i.

Using genomes where mutational biases have been measured, we tested the excess pressure method. We used the direct measurements to yield the mutational transition matrix M_{ij} defined in Sec. 2.1.1. Using this matrix in Eq. 2.5, we then computed selective coefficients s^*_i using Eq. 2.4. Since the transition matrix explicitly accounts for mutational biases, s^*_i are the selective pressures required to shift the word frequencies away from the equilibrium established by mutation alone. In Fig. S12, we compared excess pressures Δs_i versus coefficients s^*_i, finding excellent agreement between the two methods. In human, plant, and fly genomes, the two methods exhibited correlations between 0.92 and 0.99. We conclude that excess pressures accurately assess global selective coefficients, and are thus useful when direct measurements of mutational biases are not available. Since most mutational biases are accounted for by dinucleotide effects, we expect that excess pressures computed in introns using the 2-mer null model (Figs. 4A, S6A) provide our most reliable estimate of global selective coefficients in eukaryotic genomes.

3 Legends

3.1 Supplemental Figures

Figure S1: Mouse TF binding score correlation with genomic word frequency data. Correlation coefficients (Spearman's ρ) computed between relative frequencies f_i and binding scores b_i for each mouse TF, where i range over all possible 8-mers. The binding scores used here corresponded to the z-scores reported in [20]. Frequencies f_i corresponded to the raw counts divided by expected counts based on nucleotide composition (i.e. the 1-mer null). Correlations are separately computed over subsets of words with the same CpG content. Bars are shown only for statistically significant correlations, with p-value $< 10^{-6}$. Vertical gray and white blocks span 10 TFs each. Panels (A) and (B) use weak binding words, $b_i < 0$ or $b_i < 2$, respectively. Panel (C) does not condition on binding strength, while panel (d) uses words with $b_i > 0$.

Figure S2: TF binding score correlation with genomic word frequency data for additional species. For each species and each TF, we computed the Spearman correlation (ρ) of binding scores b_i vs. genomic word frequencies f_i, where i indexes over all 8-mer words. We plot ρ values on three different subsets of k-mer words: (left panels) using all words i satisfying $b_i < 0$; (middle panels) using all words; and (right panels) using words satisfying $b_i > 0$. Correlation values are either plotted in color bars when statistically significant (p-val $< 10^{-6}$) or as white or gray bars otherwise. Negative correlations are observed for the majority of TFs in all cases, and occur most frequently over k-mers with $b_i < 0$ in mouse (Fig. S1), worm, fly, and human, and over k-mers with $b_i > 0$ in yeast. Frequencies f_i corresponded to the raw counts divided by expected counts based on nucleotide composition (i.e. the 1-mer null) for worm, fly, and human; counts were determined using all intron data in these species. In yeast, which has few introns, the raw counts were made across all exons, and the expected counts were obtained from the synonymous codon shuffling scheme. Binding scores b_i corresponded to z-scores for the fly dataset [21, 22, 23], and to e-values for worm [24] and yeast [25] datasets; the human dataset [26, 22] had 5 TFs with z-score data and 3 TFs with e-value data (see Table S1).

Figure S3: Word frequency plots and pressure distributions of exons for a range of eukaryotic species ($k = 6$). For each species, the top panel shows the $(\tilde{f}, f)$ plot, and the bottom panel gives the corresponding excess pressure distribution, with $\Delta s_i/u$ on the x-axis.

Figure S4: Word frequency plots and pressure distributions of exons for a range of bacterial species ($k = 6$). See Table S2 for species names, classification, and strain details. For each species, the top panel shows the $(\tilde{f}, f)$ plot, and the bottom panel gives the excess pressure distribution, with $\Delta s_i/u$ on the x-axis. Because certain bacteria exhibit a very wide pressure distribution, typically due to a small number of words with large negative pressures, we plot using red bars (3 on the left edge, 1 on the right edge) word counts with pressures in the ranges $(-1000, -500)$, $(-500, -100)$, $(-100, -20)$, and > 5, respectively.

Figure S5: Analysis of intron word frequency distributions using the 1-mer null model. Word frequencies ($k = 6$) are shown relative to expectation based on nucleotide composition of introns (1-mer null). **Row 1.** Relative word frequency histograms. Black curve indicates all words, purple bars (0 CpG), blue bars (1 CpG), green bars (2 CpG's). **Row 2.** Distribution of genomic word frequency z-score distributions. For each word i, its observed counts $C_{i,obs}$ were measured, and the expected counts $C_{i,exp}$ as well as expected variance $\sigma^2_{i,exp}$ were computed using the 1-mer null. The word z-scores were computed as $(C_{i,obs} - C_{i,exp})/\sigma_{i,exp}$. **Rows 3 & 4.** The $(\tilde{f}, f)$ plot for each species; colors in Row 3 indicate CpG content as in Row 1. **Row 5.** The distribution of excess pressures relative to the 1-mer null.

Figure S6A: Analysis of intron word frequency distributions using the 2-mer null model. Caption identical to S5, except all word frequencies are computed relative to the dinucleotide composition using the Markov chain model described in Methods.

Figure S6B: Analysis of intron word frequency distributions using the 3-mer null model. Caption identical to S5, except all word frequencies are computed relative to the trinucleotide composition using the Markov chain model described in Methods.

Figure S7: Analysis of exon word frequency distributions using the codon shuffling model. Word frequencies ($k = 6$) are shown relative to expectation based on the synonymous codon shuffling scheme in exons (exon null). **Row 1.** Relative word frequency histograms. Black curve indicates all words, purple bars (0 CpG), blue bars (1 CpG), green bars (2 CpG's). **Rows 2 & 3.** The $(\tilde{f}, f)$ plot for each species; colors in Row 2 indicate CpG content as in Row 1. **Row 4.** The distribution of excess pressures relative to the exon null.

Figure S8: Correlation of word frequencies f_i measured in exons vs. in introns for different species. Top row shows results for all 6-mers, colored by CpG content (see legend); bottom three rows plot separately within each CpG category. Spearman correlation ρ and p-value are indicated in each plot. Word frequencies f_i were computed relative to 1-mer null (introns) and relative to synonymous codon shuffling statistics (exons).

Figure S9: Word frequency plots of exons for all human chromosomes. For each chromosome, the top panel shows the $(\tilde{f}, f)$ plot, and the bottom panel gives the excess pressure distribution, with $\Delta s/u$ on the x-axis. Codon shuffling was performed separately for each chromosome.

Figure S10: Relation of physical size and recombination rates to chromosomal word frequency deviations. **A.** Using intron data of each chromosome in the eukaryotic dataset (1,304 chromosomes), we compute the correlation coefficient of its relative word-frequency vector f vs. the genome-wide average, and plot this value against the chromosome's total intron length. Relative word frequencies of each chromosome were calculated with respect to the 1-mer null, i.e.

expected frequencies based on chromosome-specific nucleotide composition; the genome-wide average was relative to the genome-wide nucleotide composition. The plot indicates that the relative word composition of smaller chromosomes tends to correlate less strongly with the rest of the genome than that of larger chromosomes. Y / W chromosomes are indicated as red points. **B.** Subplots show model species in which chromosomal recombination rates have been measured, showing the correlation coefficient of each chromosome's f vector vs. the genome-wide average. Relative word frequencies were computed as in panel (A). Y chromosomes are shown separately as squares. The x-axis was calculated as the recombination rate (cM / Mb) times the total length of introns analyzed (Mb). Recombination rates were obtained from [27] (human and mouse) and [28] (fly). In *C.elegans*, each chromosome crosses over exactly one time per meiosis, which corresponds to 50 cM [29]. We note that the further deviation of human chromosome 19 is potentially related to the prevalence of large, intrachromosomal duplications in its evolutionary history [30].

Figure S11: Model solution using Gamma-distributed random pressures, with given values of the shape parameter α and the standard deviation σ. The values of $-s_i$ were assigned from the corresponding Gamma distribution, and used in the model (Eq. 2.2) to obtain the equilibrium word frequency vector $\mathbf{f}$. Upper panels show the $(\tilde{f}, f)$ plots; lower panels show the histogram of selective coefficients in units of the mutation rate. All other details are as in Fig. 1.

Figure S12: Comparison of selective coefficients inferred using known mutational biases vs. using measured nucleotide or dinucleotide composition for *H. sapiens*, *A. thaliana*, and *D. melanogaster*. To compute s_i/u using known mutational biases, a mutation transition matrix M_{ij} was constructed using measured biases as follows. For the human genome, mutational biases measured in [15] were used, and the rate of dinucleotide CpG mutation was set to $11u$ as given in [17]. For plant and fly mutational biases, we used data from [13] and [12], respectively. Using the given mutational transition matrix in Eq. (2.5) and substituting in Eq. (2.4), we computed s_i^*/u using Eq. (2.4) from the raw (unnormalized) frequencies f_i measured in introns. To compute excess pressures $\Delta s_i/u$, we used an unbiased transition matrix and the unnormalized frequencies f_i to obtain s_i, and used the expected frequencies f_i' using the 2-mer null model (human) or 1-mer null model (plant and fly) to obtain s_i', which yielded $\Delta s_i = s_i - s_i'$. In all three cases, the plots show a very high correlation between the two different methods, which indicates that selective coefficients can be accurately inferred even in cases when mutational biases are not known, by using excess pressures relative to dinucleotide composition.

Figure S13: Correlation of excess pressures of words and their mutational neighbors in different species ($k = 6$). For each word, the excess pressure is plotted (x-axis) against the average excess pressure of its neighbors (y-axis). Excess pressures were calculated with respect to each of the null models (shown in separate rows). The correlation coefficient (r) from linear regression are given. All correlations are highly significant (p-val $\ll 10^{-10}$).

Figure S14: Dynamics of global per site fitness in simulated evolution with population bottlenecks (vertical gray bars) at regular intervals (see *Supplementary Methods*). **A.** Curves show the population average in five independent simulation runs with recombination (solid lines in color); the fittest sequence from one simulation with recombination (purple, top curve); and a run without recombination (dashed blue line). Dashed horizontal lines indicate the fitness for a uniform word composition (bottom line) and for the expected composition at mutation-selection balance (top line). Parameters: population size $N = 10^5$ (10^3 for bottlenecks); $L = 2000$bp; $u = 1 \times 10^{-4}$ per bp per generation; s_i are the measured values of s_i/u from human introns using $k = 4$, multiplied by 10^{-4}. Bottlenecks occur every 500 gen. and last 50 gen. Random pairs of sequences recombine with a cross-over length of 200bp. **B.** Snapshots of the evolution of the word frequency vector in the $(\tilde{f}, f)$ plane. Red dots correspond to the (4-mer) word frequencies in the human genome (i.e. the solution at mutation-selection balance) and the circles correspond to word frequencies in the simulated population at various time points.

Figure S15: *In silico* evolution using human intron 4-mer pressures, showing the Muller's ratchet effect in non-recombining populations. All parameters were the same as in Fig. S14, except that there was no recombination. One set of curves (lower) was initialized from random initial conditions; another set (upper) was initialized from the predicted equilibrium 4-mers frequencies. For each set, the top curve (purple or orange) corresponds to the maximally fit individual in the color-matched simulation run.

Figure S16: Word-neighbor plot of the residuals obtained from regression of word frequencies vs. word dinucleotide or trinucleotide composition (see Table S5 caption). Panels correspond to regression of 6-mer frequencies on dinucleotide composition (A) and trinucleotide composition (B); and 8-mer frequencies on dinucleotide composition (C) and trinucleotide composition (D). For each word, we plot its residual on the vertical axis vs. the average residuals of its mutational neighbors on the horizontal axis. Pearson correlation r is given for each plot. The figure demonstrates that even when dinucleotide or trinucleotide biases have been regressed out of the word frequency data, strong word-neighbor correlations persist.

Figure S17: Word frequency PCA plot of all eukaryotic chromosomes from all phylogenetic groups, indicated by shapes & colors according to the legend. See Table S3 for further species information. The analysis was based on intronic 6-mer frequencies. Frequencies were normalized using the 2-mer null model.

Figure S18: Word frequency PCA plot for all vertebrate chromosomes, showing the distinct separation of primates (left pointing triangles) away from the other mammals (x marks). See Table S3 for further species information. The analysis was based on intronic 6-mer frequencies. Frequencies were normalized using the 2-mer null model.

Figure S19: Word frequency PCA plot for all bacterial species (947 total). Color/symbol combinations denote different phyla (in the case of proteobacteria, further subdivisions are shown). Phyla with fewer than 20 genomes available were combined into a single group called 'other bacteria'. The analysis was based on exonic 6-mer frequencies. Frequencies were normalized by synonymous codon shufflings.

Figure S20: Word frequency PCA plot for γ-Proteobacteria (356 species). Each genus is denoted by a different color. In several genus (Acinetobacter, Buchnera, Salmonella, Escherichia, Legiononella) clusters appear extremely tight because they consist mostly of strains from a single species; however other clusters consisting of multiple species are likewise relatively tight. The analysis was based on exonic 6-mer frequencies. Frequencies were normalized by synonymous codon shufflings.

Figure S21: Word frequency distributions in the repeat and the non-repeat regions in introns of the human genome. Repeat sequences were extracted from intron regions from the repeat-masked sequence files (.mfa) in the RefSeq database, which uses the RepeatMasker program to identify repetitive or repeat-derived elements in the human genome. Word statistics were obtained separately from all repeat-derived or all repeat-masked segments of introns. The total length of repeat sequences extracted was 381.8Mb, leaving 454.2Mb of non-repeat intron sequences. (A, B) Correlations of word frequencies measured across all introns vs. across all repeat-masked (i.e. non-repeat-derived) intron regions (A), and across the repeat-derived intron regions vs. the repeat-masked regions (B). (C,D) Correlations between word and average neighbor frequencies measured in the repeat-masked regions (C) and in the repeat-derived regions (D). Plots from left to right are shown using raw word frequencies (normalized by the average word frequency), relative frequencies to the 1-mer null, and relative frequencies to the 2-mer null. To calculate relative frequencies, the 1-mer and 2-mer null models were constructed from nucleotide and di-nucleotide frequencies separately in the repeat-derived or repeat-masked portions accordingly. Pearson correlation coefficients (r) of log frequencies are indicated.

Figure S22: Word frequency distributions in DNase I hyper-sensitive regions in the human genome. DNase I hyper sensitive-regions (DNaseI peaks) were mapped by the UCSC ENCODE DNase analysis pipeline using the DNase-seq experimental data generated in [31]. DNaseI peaks were combined from 95 cell lines, and only the peaks in introns and intergenic regions were used (gene annotation was based on UCSC GENCODE v22). The total length of peak sequences analyzed was 431.4Mb. 6-mer frequencies and nucleotide and di-nucleotide frequencies were measured in each peak sequence and combined to calculate the genome-wide statistics. (A) Correlations of word frequencies obtained in DNaseI peaks vs. in the entire intronic regions. (B) Correlations between word and average neighbor frequencies in the DNaseI peaks. Pearson correlation coefficients (r) of log frequencies were indicated in the plots. (C, D) Same as panels (A, B) except that repeat-derived sequences (152.7Mb) were excluded from all the DNaseI peaks analyzed.

Figure S23: Word-neighbor plots for seven bacterial species relative to dicodon shuffling null model. For each species, the total exon sequence was used to construct 100 dicodon-shuffled randomizations, according to the shuffling scheme described in [32]. This scheme shuffles synonymous codons (i.e. maintaining amino-acid sequences) while preserving the dicodon pair statistics across the entire sequence. Since dicodons are 6 nucleotides long, this shuffling also preserves nucleotide, dinucleotide, trinucleotide, and tetranucleotide composition. Using the randomizations, we computed the expected word frequencies for all 7-mers. We used $k = 7$ instead of $k = 6$ to avoid any artifacts that could arise from the shuffling unit being of identical size to the k-mers we are measuring. Observed word frequencies in exon sequences are shown relative to the expected frequencies from the shuffled sequences (top row). Distributions of excess pressures relative to the dicodon shuffling null expectation are shown (bottom row). To show the extreme, red bins are used on the left and right of the distribution to indicate observations outside the range of $(-10, 5)$: on the left, red bins correspond to excess pressures in the ranges <-1000, $(-1000, -500)$, $(-500, -100)$, $(-100, -50)$, $(-50, -10)$ and on the right to $(5, 10)$, $(10, 50)$, $(50, 100)$, $(100, 500)$, and > 500. The following genomes were used in the above analyses: *Helicobacter pylori 26695*, *Bacillus subtilis 168*, *Neisseria meningitidis MC58*, *Synechococcus sp. WH8102*, *Escherichia coli K-12 W3110*, *Bifidobacterium animalis AD011*, *Streptococcus pyogenes M1 GAS*.

3.2 Supplemental Tables

Table S1: Transcription factor binding data correlated with genomic k-mer statistics in eukaryotic genomes. For each species and each TF, we computed the Spearman correlation (ρ) of binding scores b_i vs. genomic word frequencies f_i, where i indexes over all 8-mer words. Frequencies f_i corresponded to the raw counts divided by expected counts based on nucleotide composition (i.e. the 1-mer null) for mouse, fly, worm, and human; counts were determined using all intron data in these four species. In yeast, which has few introns, the raw counts were made across all exons, and the expected counts were obtained from the synonymous codon shuffling scheme. Binding scores corresponded to z-scores for mouse [20] and fly datasets [21, 22, 23], and to e-values for worm [24] and yeast [25] datasets; the human dataset [26, 22] had 5 TFs with z-score data and 3 TFs with e-value data as indicated. We report ρ on three different subsets of k-mer words: (1) Columns D-K report ρ using all words i satisfying $b_i < 0$; (2) columns L-S use all words; and (3) columns T-AA use words satisfying $b_i > 0$. The p-value associated with each measurement of ρ is reported in the column immediately to its right. For each of the subsets 1-3, we report four measurements of ρ using (a) all words in the subset, (b) conditional on having 0 CpG, (c) 1 CpG, and (d) 2 CpG's.

Table S2: List of bacterial genomes and species analyzed.

Table S3: List of eukaryotic genomes and species analyzed.

Table S4: Spearman correlation coefficients (ρ) of relative word frequency vectors between chromosome pairs either within genomes or between eukaryotic species. Relative frequencies of 6-mers were measured in exons (using the codon shuffling scheme) and in introns (using the 1-mer and 2-mer null models). Within-species ρ values were computed for all chromosome pairs of the same species; average and standard deviations of ρ across chromosome pairs are reported. Between-species ρ values are shown in three matrices (corresponding to the three null models). For each pair of species, the upper triangle of each matrix indicates the correlation between genome-wide relative word frequencies, and the lower triangle gives the mean and standard error of ρ across all pairs of chromosomes.

Table S5: Analysis of variance (ANOVA) of relative word frequencies regressed on dinucleotide and trinucleotide composition in different genomes (see Secs. 1.4 and 2.2.1 for details).

3.3 Additional Files

File S1: TF binding score and word frequency correlation plots for all analyzed data. For each transcription factor, we output a series of plots including (a) its reported primary binding motif, (b) its histogram of binding scores over all 8-mers, (c) the scatter plot of binding scores b_i vs. average mutational neighbor scores $\tilde{b}_i$, (d) a series of scatter plots showing b_i vs. f_i for words in different CpG categories (mouse, fly, human), or for all words (worm and yeast). Spearman correlation values ρ are shown, and significant values with $p < 10^{-6}$ are indicated as $***$. These plots were produced from the same data that was used to produce Table S1 (see caption for references).

4 Supplemental References

References

[1] Wilke C (2005) Quasispecies theory in the context of population genetics. *BMC Evolutionary Biology* 5:44.

[2] Qian L, Kussell E (2012) Evolutionary dynamics of restriction site avoidance. *Phys Rev Lett* 108:158105.

[3] Muller HJ (1964) The relation of recombination to mutational advance. *Mutat. Res.* 106:2–9.

[4] Felsenstein J (1974) The evolutionary advantage of recombination. *Genetics* 78:737–756.

[5] Peliti L (1997) Introduction to the statistical theory of Darwinian evolution. *arXiv:cond-mat/9712027* p. 14.

[6] Charlesworth B, Betancourt A, Kaiser V, Gordo I (2009) Genetic recombination and molecular evolution. *Cold Spring Harbor Symposia on Quantitative Biology* 74:177–186.

[7] Lee H, Popodi E, Tang H, Foster PL (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc Natl Acad Sci USA* 109(41):E2774–83.

[8] Sung W et al. (2015) Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. *Molecular Biology and Evolution*.

[9] Zhu Y, Siegal ML, Hall DW, Petrov DA (2014) Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci USA* 111:E2310–E2318.

[10] Lynch M et al. (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci USA* 105(27):9272–9277.

[11] Denver DR et al. (2009) A genome-wide view of Caenorhabditis elegans base-substitution mutation processes. *Proceedings of the National Academy of Sciences* 106(38):16310–16314.

[12] Keightley PD et al. (2009) Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res* 19(7):1195–1201.

[13] Ossowski S et al. (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327(5961):92–94.

[14] Halligan DL, Keightley PD (2009) Spontaneous mutation accumulation studies in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics* 40(1):151–172.

[15] Lynch M (2010) Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci USA* 107:961–968.

[16] Roach JC, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328:636–639.

[17] Hodgkinson A, Eyre-Walker A (2011) Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* 12(11):756–766.

[18] Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of Selection upon Genomic GC-Content in Bacteria. *PLoS Genetics* 6(9):e1001107.

[19] Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet* 6(9):e1001115.

[20] Badis G et al. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324(5935):1720–1723.

[21] Busser BW et al. (2012) Molecular mechanism underlying the regulatory specificity of a Drosophila homeodomain protein that specifies myoblast identity. *Development* 139(6):1164–1174.

[22] Nakagawa S, Gisselbrecht SS, Rogers JM, Hartl DL, Bulyk ML (2013) DNA-binding specificity changes in the evolution of forkhead transcription factors. *Proceedings of the National Academy of Sciences* 110(30):12349–12354.

[23] Soruco MML et al. (2013) The CLAMP protein links the MSL complex to the X chromosome during Drosophila dosage compensation. *Genes Dev* 27(14):1551–1556.

[24] Grove CA et al. (2009) A Multiparameter Network Reveals Extensive Divergence between C. elegans bHLH Transcription Factors. *Cell* 138(2):314–327.

[25] Zhu C et al. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* 19(4):556–566.

[26] Del Bianco C et al. (2010) Notch and MAML-1 complexation do not detectably alter the DNA binding specificity of the transcription factor CSL. *PLoS ONE* 5(11):e15034.

[27] Jensen-Seaman MI, *et al.* (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* 14:528–538.

[28] Comeron JM, Ratnappan R, Bailin S (2012) The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genet* 8(10):e1002905.

[29] Rockman MV, Kruglyak L (2009) Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genet* 5(3):e1000419.

[30] Grimwood J et al. (2004) The DNA sequence and biology of human chromosome 19. *Nature* 428(6982):529–535.

[31] Thurman RE et al. (2012) The accessible chromatin landscape of the human genome. *Nature* 489(7414):75–82.

[32] Itzkovit S, Hodis E, Segal E (2010) Overlapping codes within protein-coding sequences. *Genome Res.* 20(11):1582–1589.

Fig. S1

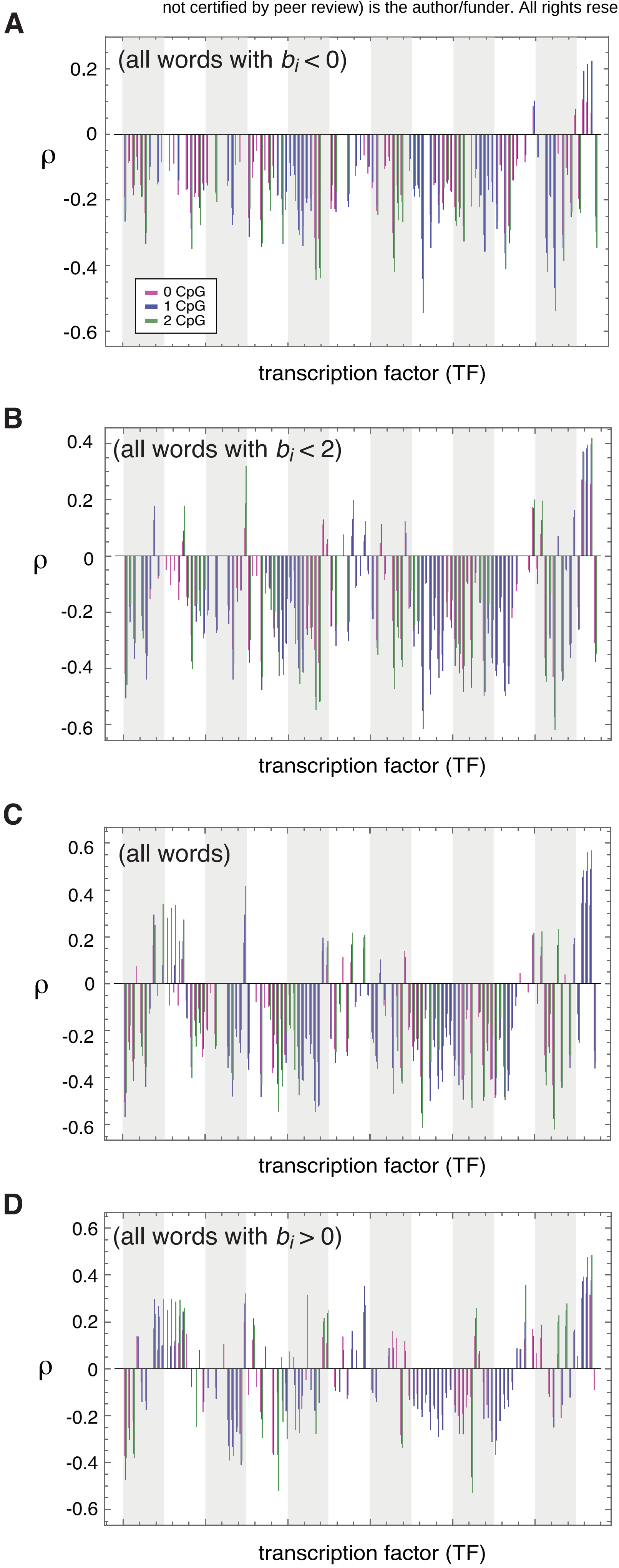

A
(all words with $b_i < 0$)
0.2
0
-0.2
-0.4
-0.6
ρ
0 CpG
1 CpG
2 CpG
transcription factor (TF)
B
(all words with $b_i < 2$)
0.4
0.2
0
-0.2
-0.4
-0.6
ρ
transcription factor (TF)
C
(all words)
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
ρ
transcription factor (TF)
D
(all words with $b_i > 0$)
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
ρ
transcription factor (TF)

Fig. S2

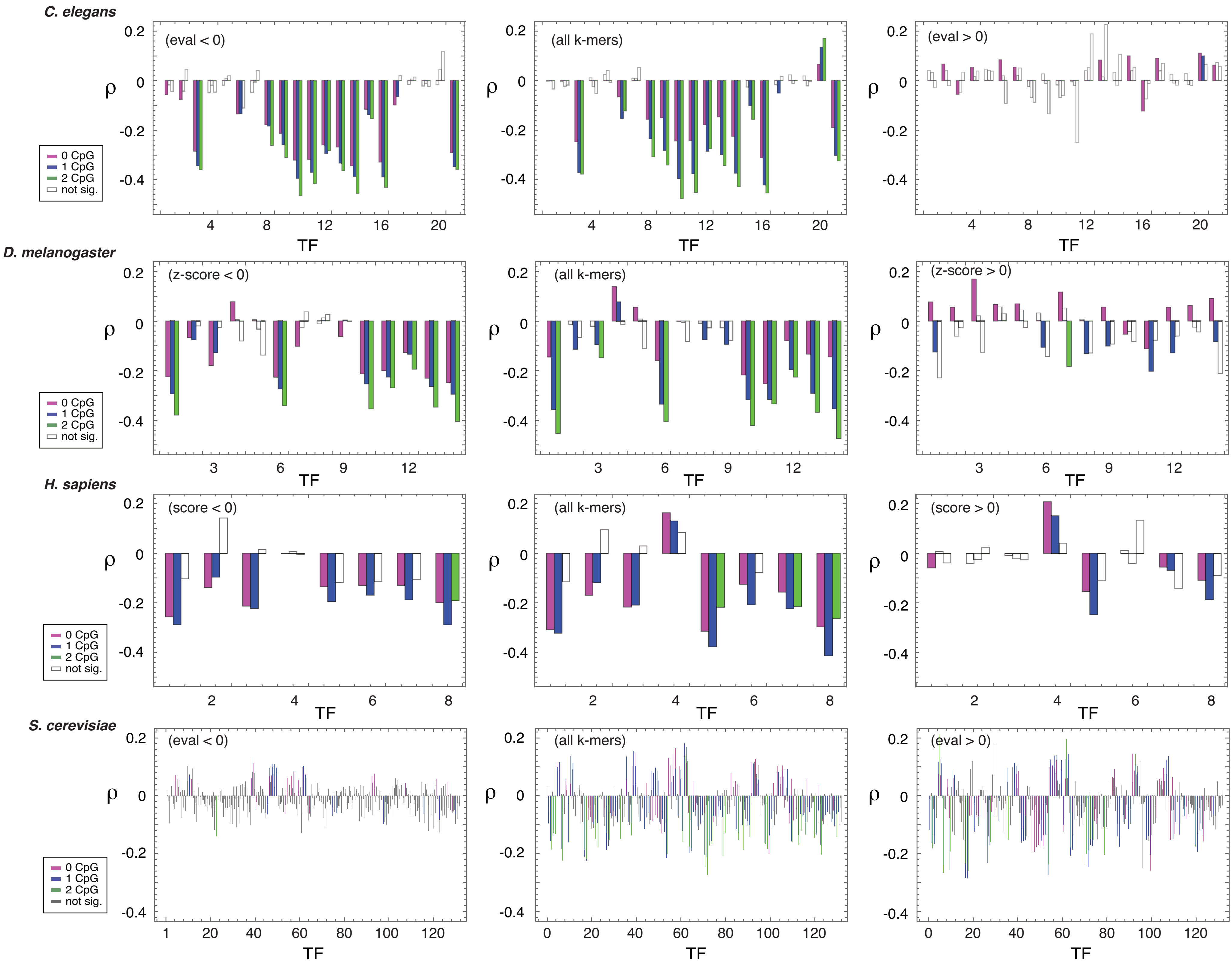

C. elegans
(eval < 0)
(all k-mers)
(eval > 0)
D. melanogaster
(z-score < 0)
(all k-mers)
(z-score > 0)
H. sapiens
(score < 0)
(all k-mers)
(score > 0)
S. cerevisiae
(eval < 0)
(all k-mers)
(eval > 0)
ρ
0.2
0
-0.2
-0.4
TF
4
8
12
16
20
3
6
9
12
2
4
6
8
1
0
20
40
60
80
100
120
0 CpG
1 CpG
2 CpG
not sig.

Fig. S3

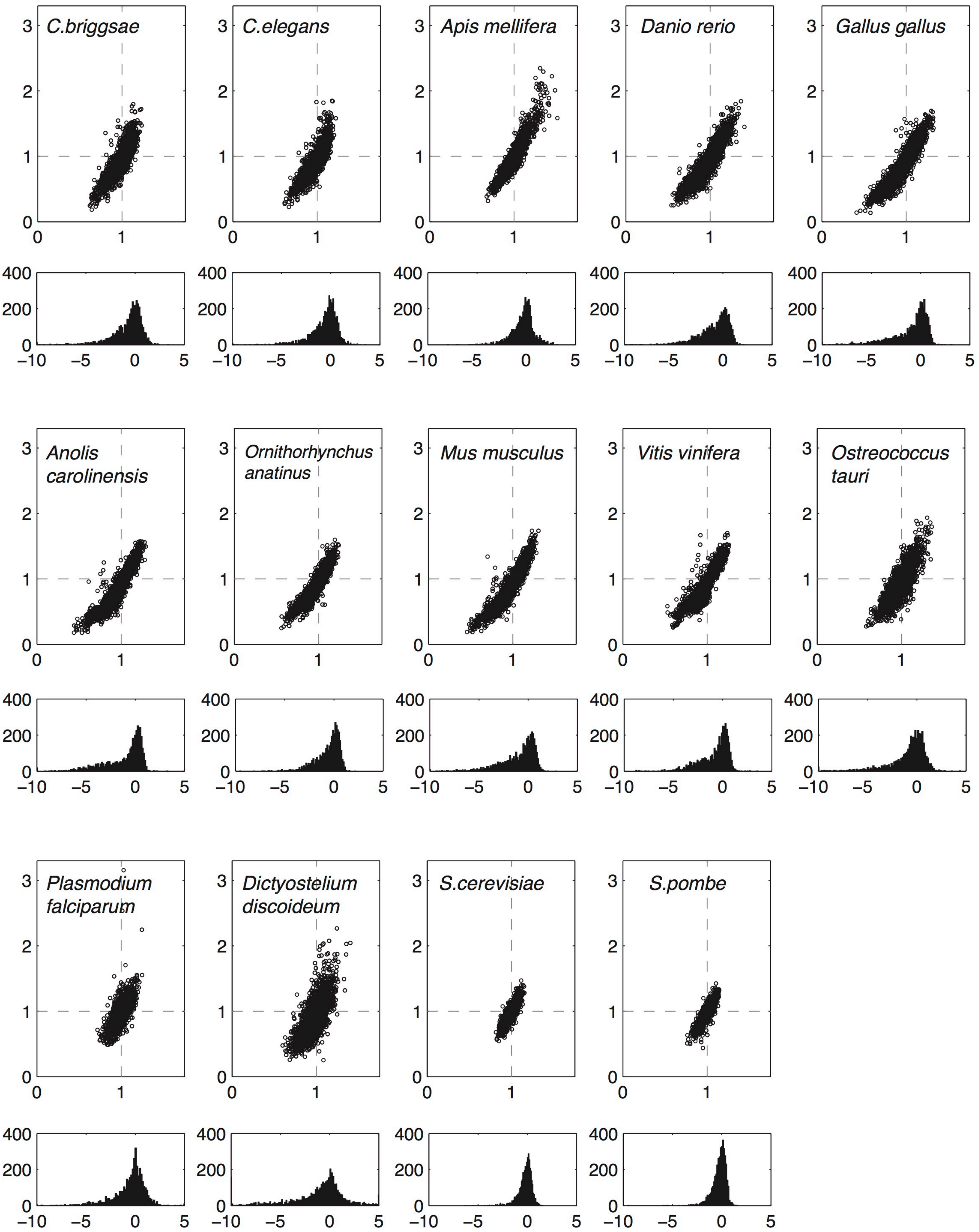

C.briggsae
C.elegans
Apis mellifera
Danio rerio
Gallus gallus
Anolis carolinensis
Ornithorhynchus anatinus
Mus musculus
Vitis vinifera
Ostreococcus tauri
Plasmodium falciparum
Dictyostelium discoideum
S.cerevisiae
S.pombe
0
1
2
3
400
200
–10
–5
5

Fig. S4

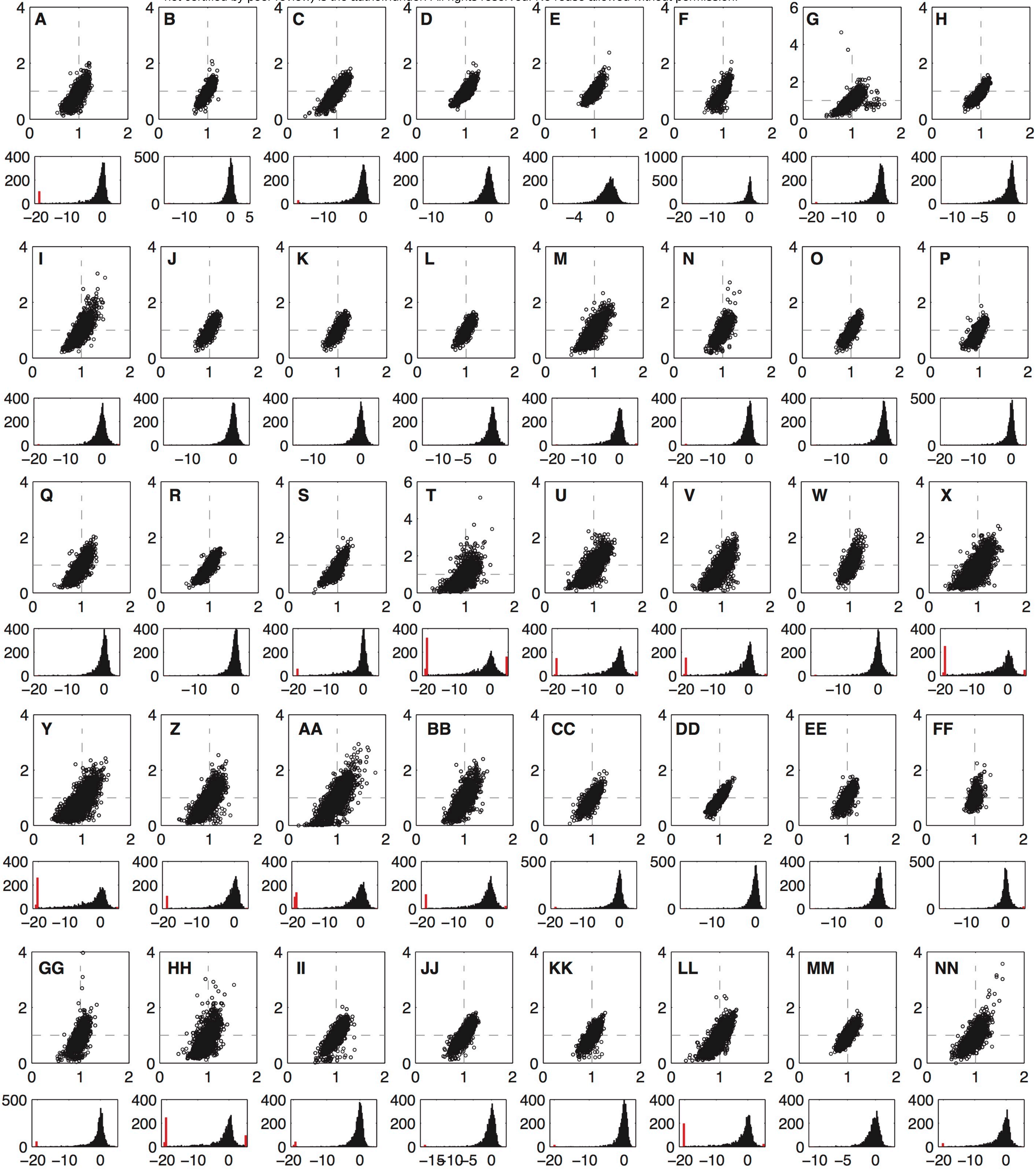
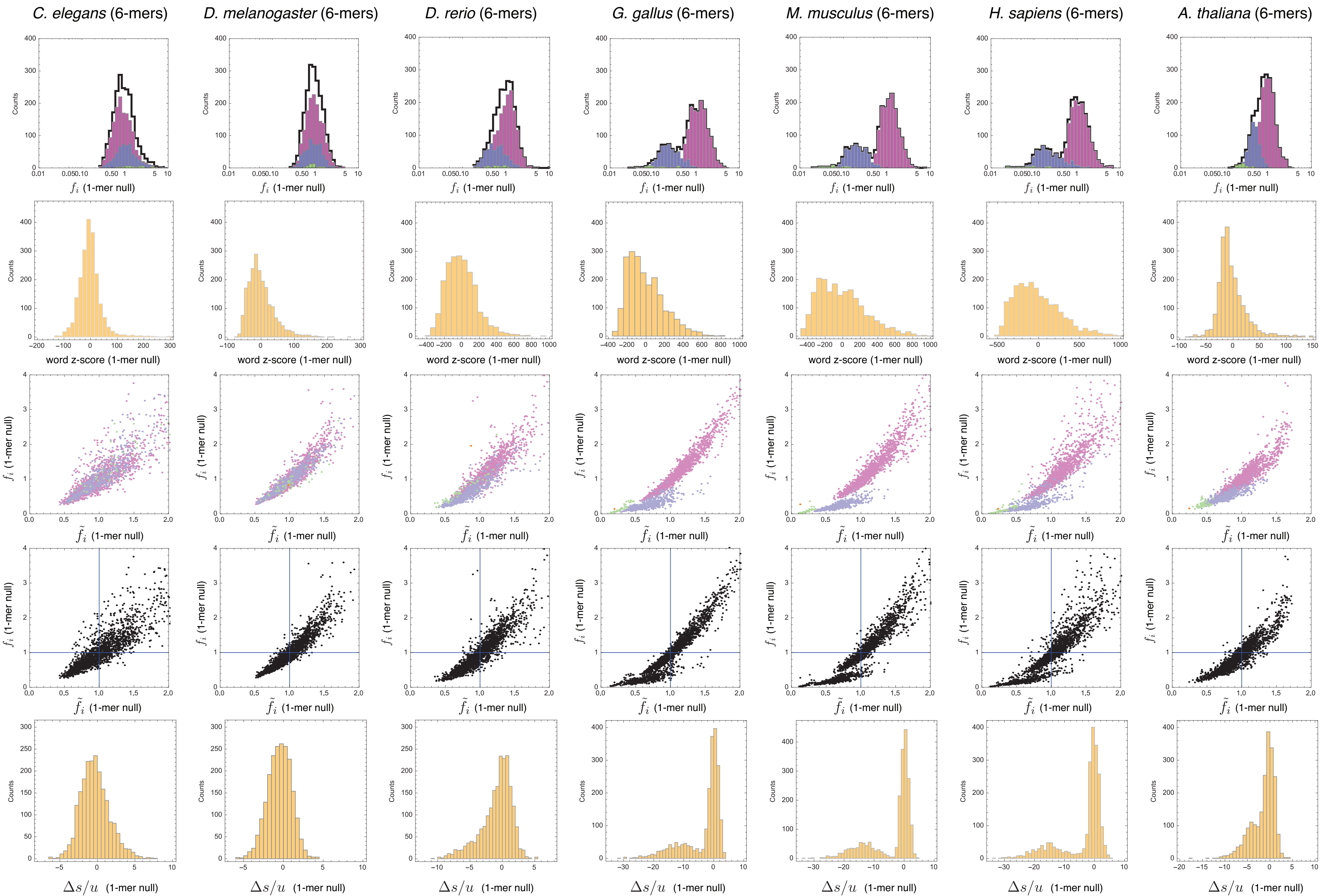
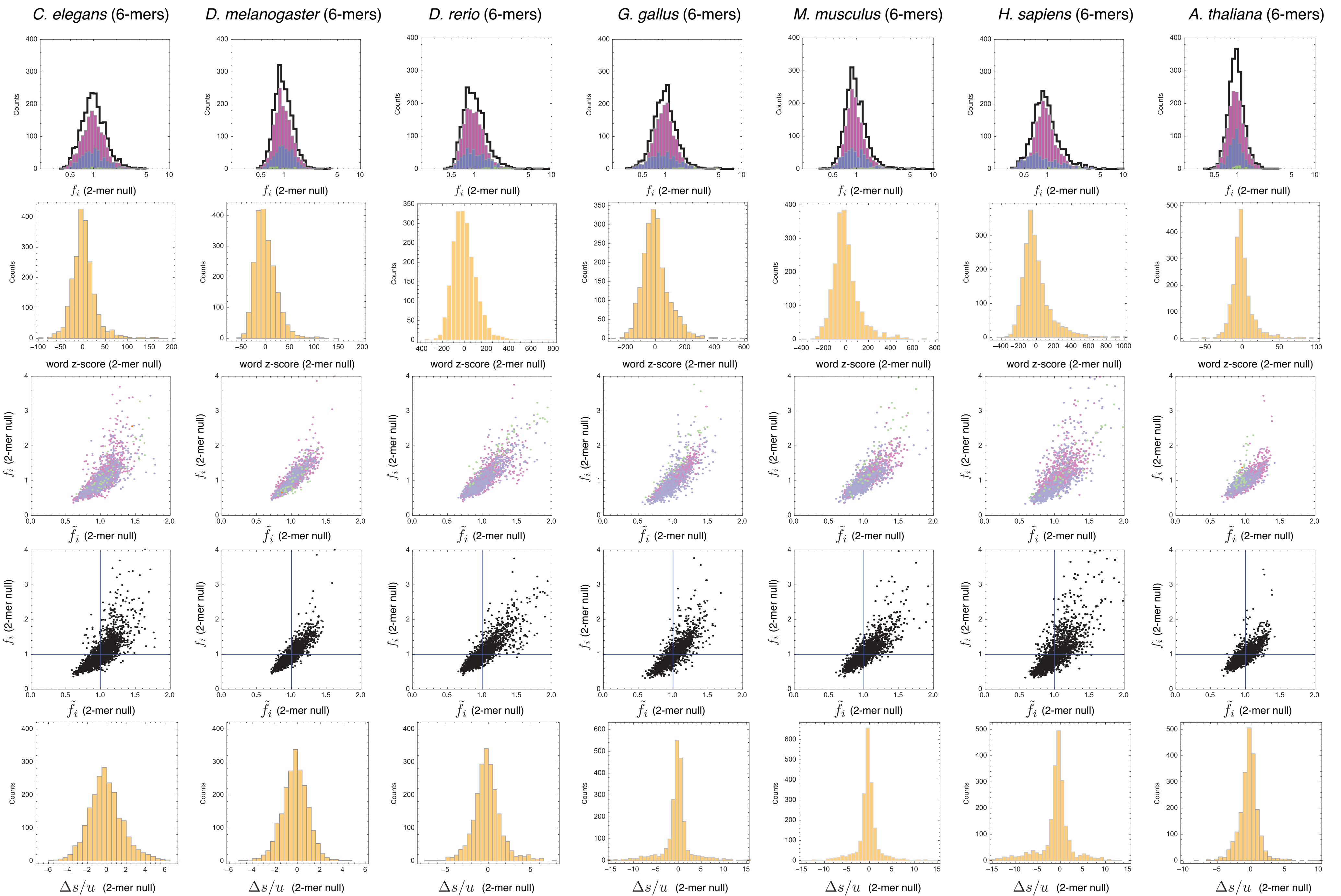

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z
AA
BB
CC
DD
EE
FF
GG
HH
II
JJ
KK
LL
MM
NN
0
1
2
4
6
200
400
500
1000
−20
−15
−10
−5
−4
0
5

Fig. S5



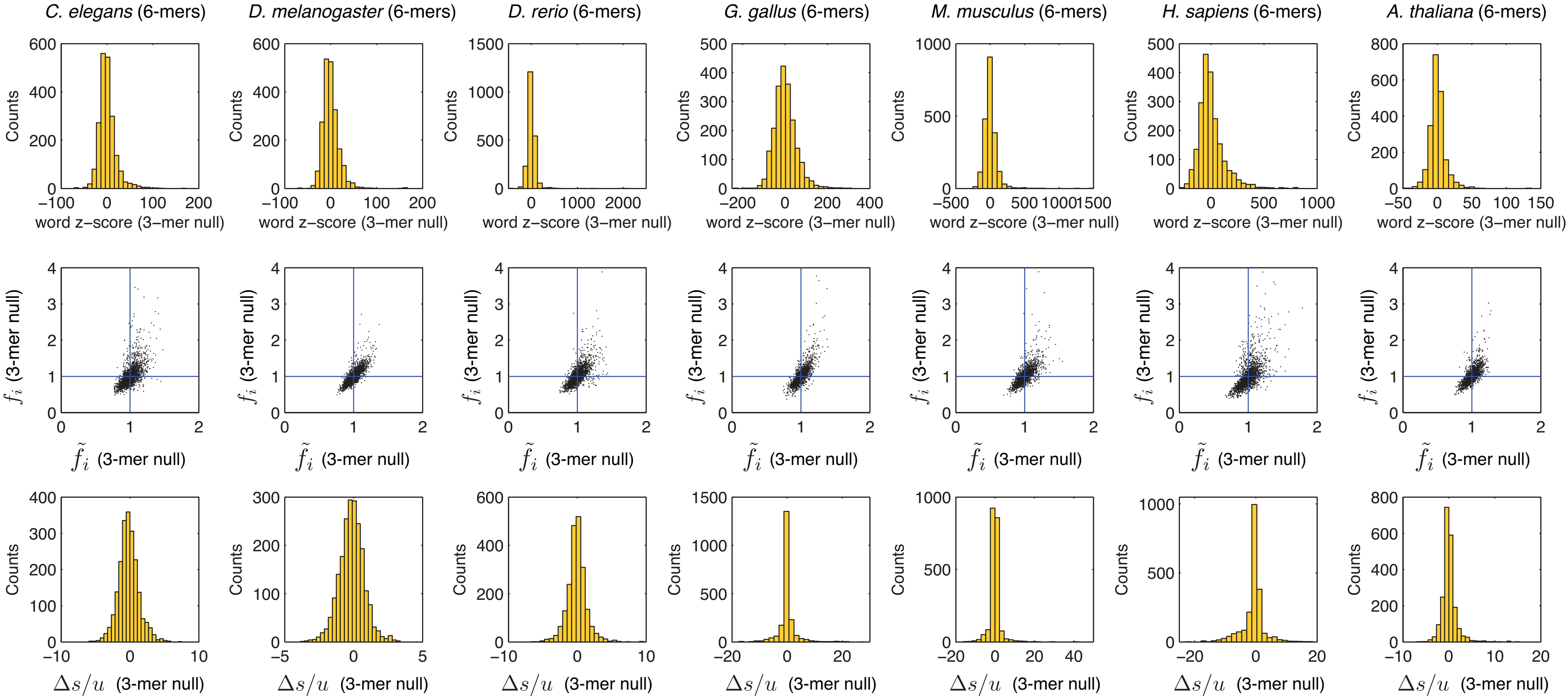
C. elegans (6-mers)
D. melanogaster (6-mers)
D. rerio (6-mers)
G. gallus (6-mers)
M. musculus (6-mers)
H. sapiens (6-mers)
A. thaliana (6-mers)
Counts
f_i (1-mer null)
word z-score (1-mer null)
f_i (1-mer null)
$\tilde{f}_i$ (1-mer null)
f_i (1-mer null)
$\tilde{f}_i$ (1-mer null)
Counts
$\Delta s/u$ (1-mer null)

Fig. S6A



C. elegans (6-mers)
D. melanogaster (6-mers)
D. rerio (6-mers)
G. gallus (6-mers)
M. musculus (6-mers)
H. sapiens (6-mers)
A. thaliana (6-mers)
Counts
0
100
200
300
400
0.5
1
5
10
f_i (2-mer null)
Counts
word z-score (2-mer null)
−100
−50
0
50
100
150
200
−400
−200
0
200
400
600
800
1000
f_i (2-mer null)
$\tilde{f}_i$ (2-mer null)
0.0
0.5
1.0
1.5
2.0
1
2
3
4
f_i (2-mer null)
$\tilde{f}_i$ (2-mer null)
Counts
$\Delta s/u$ (2-mer null)
−15
−10
−5
0
5
10
15
−6
−4
−2
2
4
6

Fig. S6B



C. elegans (6-mers)
D. melanogaster (6-mers)
D. rerio (6-mers)
G. gallus (6-mers)
M. musculus (6-mers)
H. sapiens (6-mers)
A. thaliana (6-mers)
Counts
600 400 200 0
−100 0 100 200
word z−score (3−mer null)
600 400 200 0
−100 0 100 200
1500 1000 500 0
0 1000 2000
500 400 300 200 100 0
−200 0 200 400
1000 500 0
−500 0 500 1000 1500
500 400 300 200 100 0
0 500 1000
800 600 400 200 0
−50 0 50 100 150
f_i (3−mer null)
$\tilde{f}_i$ (3−mer null)
4 3 2 1 0
0 1 2
Counts
400 300 200 100 0
−10 0 10
300 200 100 0
−5 0 5
600 400 200 0
−10 0 10
1500 1000 500 0
−20 0 20
1000 500 0
−20 0 20 40
1000 500 0
−20 0 20
800 600 400 200 0
−10 0 10 20
$\Delta s/u$ (3−mer null)

Fig. S7

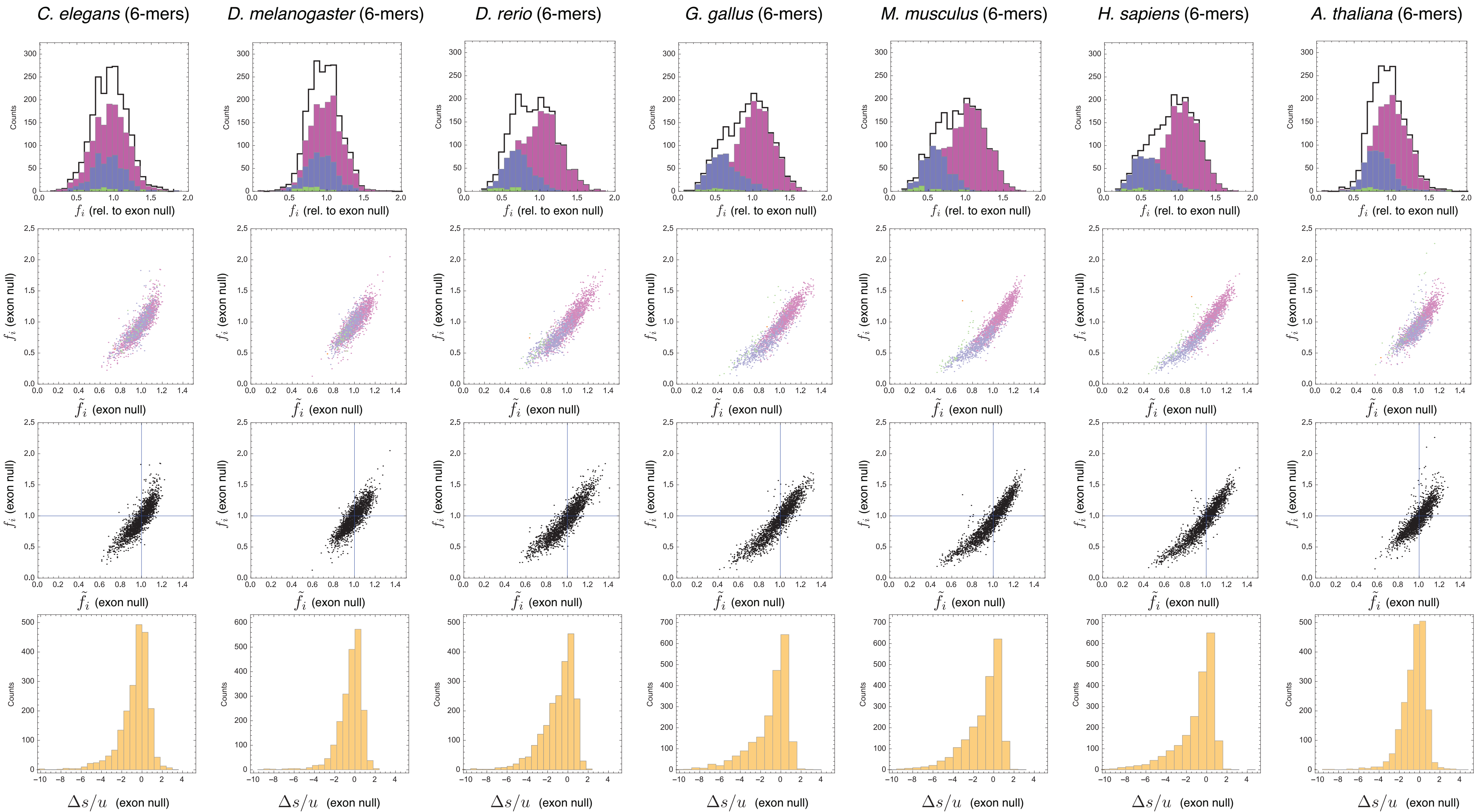
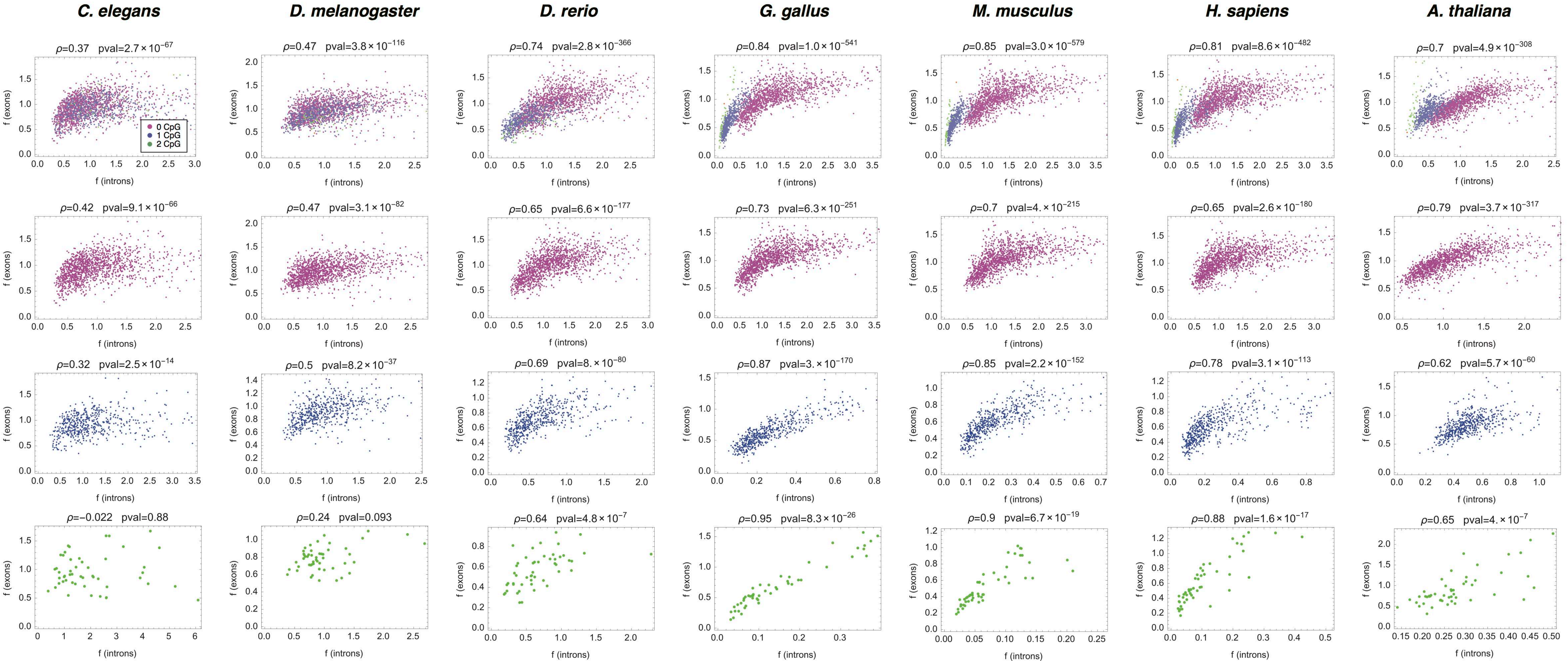

C. elegans (6-mers)
D. melanogaster (6-mers)
D. rerio (6-mers)
G. gallus (6-mers)
M. musculus (6-mers)
H. sapiens (6-mers)
A. thaliana (6-mers)
Counts
0
50
100
150
200
250
300
0.0
0.5
1.0
1.5
2.0
f_i (rel. to exon null)
f_i (exon null)
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
$\tilde{f}_i$ (exon null)
Counts
0
100
200
300
400
500
600
700
−10
−8
−6
−4
−2
0
2
4
$\Delta s/u$ (exon null)

Fig. S8



C. elegans
D. melanogaster
D. rerio
G. gallus
M. musculus
H. sapiens
A. thaliana
ρ=0.37 pval=2.7 × 10^{-67}
ρ=0.47 pval=3.8 × 10^{-116}
ρ=0.74 pval=2.8 × 10^{-366}
ρ=0.84 pval=1.0 × 10^{-541}
ρ=0.85 pval=3.0 × 10^{-579}
ρ=0.81 pval=8.6 × 10^{-482}
ρ=0.7 pval=4.9 × 10^{-308}
0 CpG
1 CpG
2 CpG
ρ=0.42 pval=9.1 × 10^{-66}
ρ=0.47 pval=3.1 × 10^{-82}
ρ=0.65 pval=6.6 × 10^{-177}
ρ=0.73 pval=6.3 × 10^{-251}
ρ=0.7 pval=4. × 10^{-215}
ρ=0.65 pval=2.6 × 10^{-180}
ρ=0.79 pval=3.7 × 10^{-317}
ρ=0.32 pval=2.5 × 10^{-14}
ρ=0.5 pval=8.2 × 10^{-37}
ρ=0.69 pval=8. × 10^{-80}
ρ=0.87 pval=3. × 10^{-170}
ρ=0.85 pval=2.2 × 10^{-152}
ρ=0.78 pval=3.1 × 10^{-113}
ρ=0.62 pval=5.7 × 10^{-60}
ρ=−0.022 pval=0.88
ρ=0.24 pval=0.093
ρ=0.64 pval=4.8 × 10^{-7}
ρ=0.95 pval=8.3 × 10^{-26}
ρ=0.9 pval=6.7 × 10^{-19}
ρ=0.88 pval=1.6 × 10^{-17}
ρ=0.65 pval=4. × 10^{-7}
f (exons)
f (introns)

Fig. S9

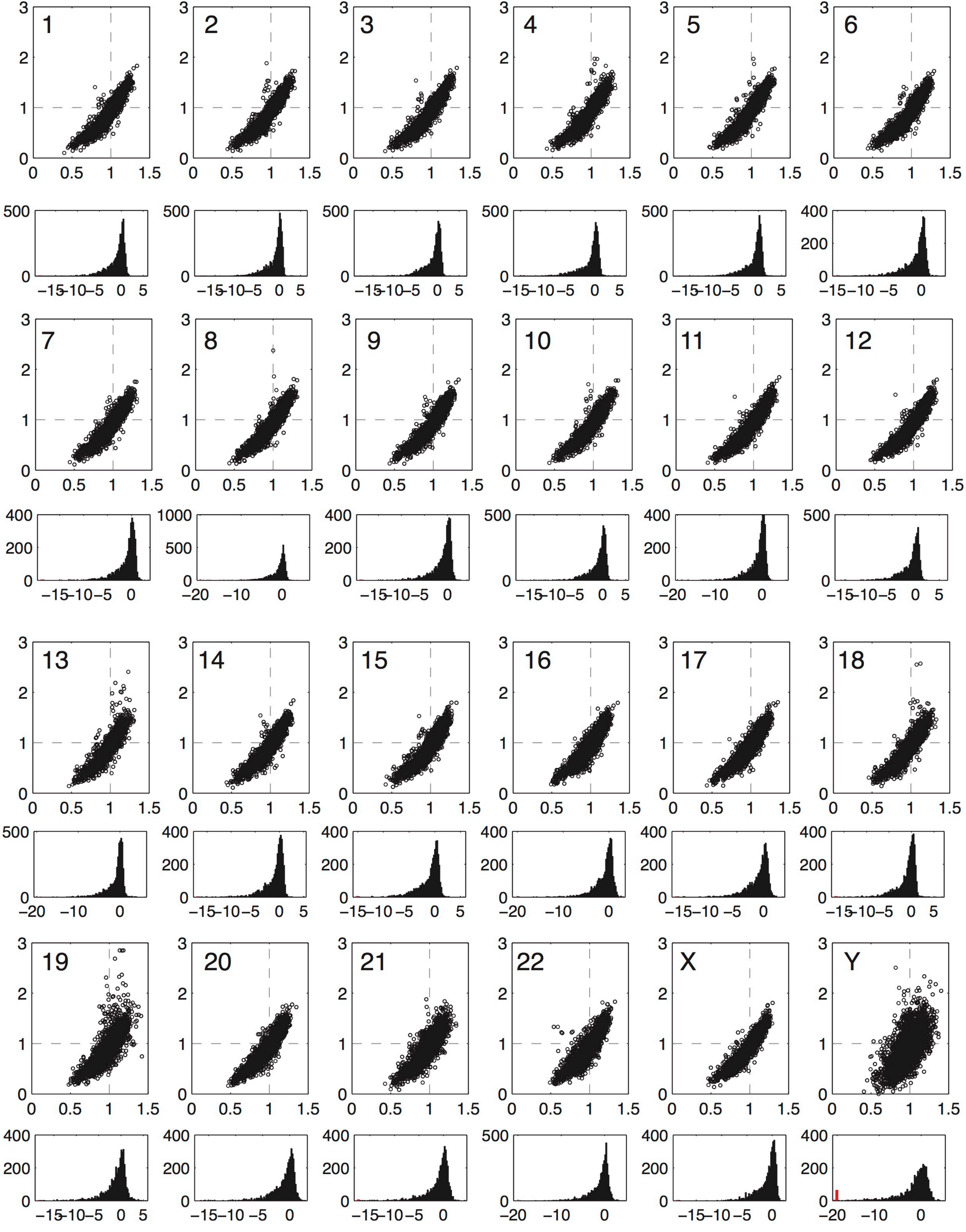

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
0
1
2
3
0
0.5
1
1.5
0
200
400
500
1000
−20
−15
−10
−5
0
5

Fig. S10

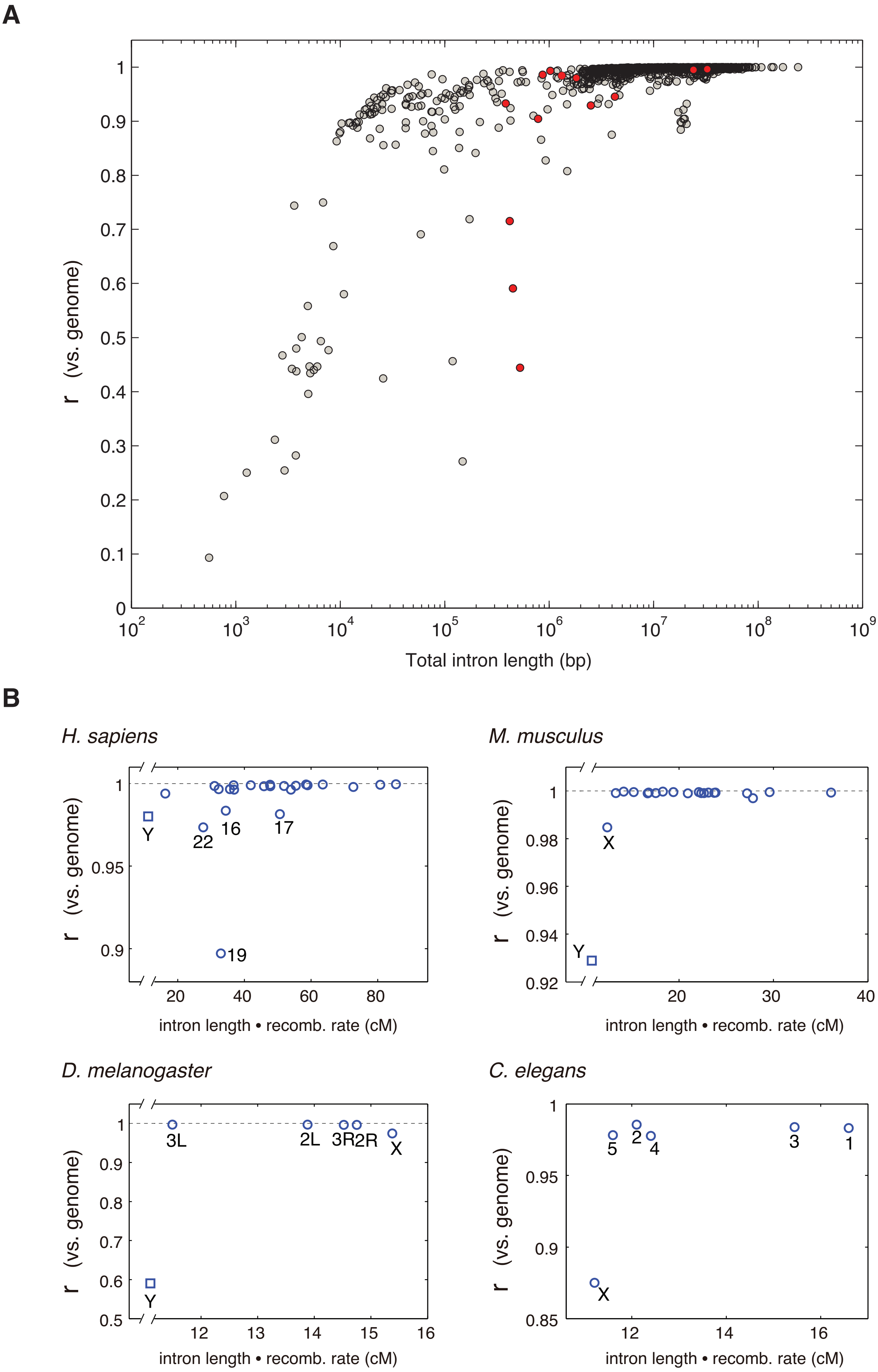

A
r (vs. genome)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10^2
10^3
10^4
10^5
10^6
10^7
10^8
10^9
Total intron length (bp)
B
H. sapiens
r (vs. genome)
1
0.95
0.9
Y
22
16
17
19
20
40
60
80
intron length • recomb. rate (cM)
M. musculus
r (vs. genome)
1
0.98
0.96
0.94
0.92
X
Y
20
30
40
intron length • recomb. rate (cM)
D. melanogaster
r (vs. genome)
1
0.9
0.8
0.7
0.6
0.5
3L
2L
3R
2R
X
Y
12
13
14
15
16
intron length • recomb. rate (cM)
C. elegans
r (vs. genome)
1
0.95
0.9
0.85
5
2
4
3
1
X
12
14
16
intron length • recomb. rate (cM)

Fig. S11

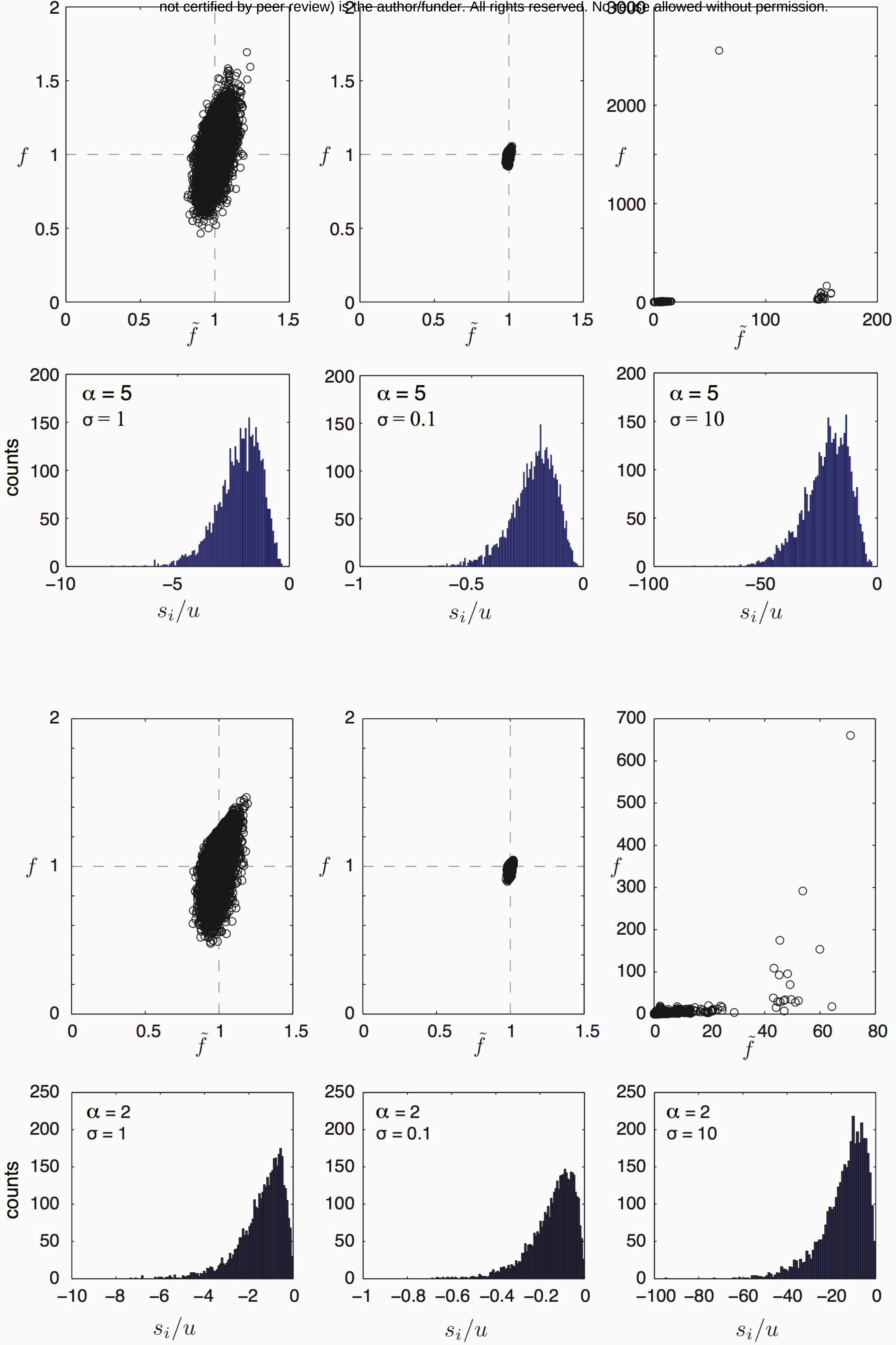
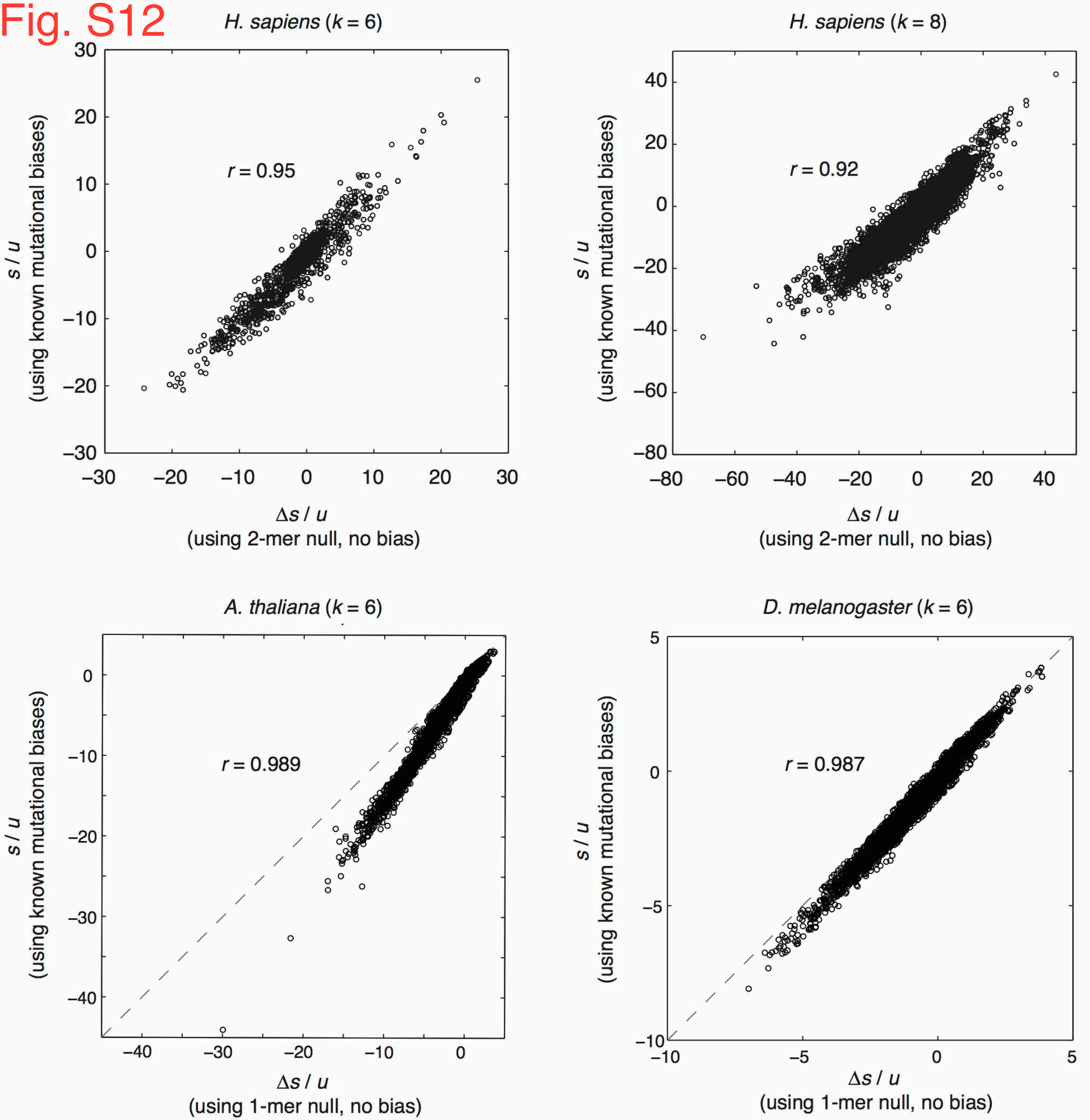

f
$\tilde{f}$
2
1.5
1
0.5
0
0
0.5
1
1.5
3000
2000
1000
0
100
200
counts
$\alpha = 5$
$\sigma = 1$
$\alpha = 5$
$\sigma = 0.1$
$\alpha = 5$
$\sigma = 10$
200
150
100
50
0
−10
−5
0
−1
−0.5
0
−100
−50
0
s_i/u
700
600
500
400
300
200
100
0
20
40
60
80
$\alpha = 2$
$\sigma = 1$
$\alpha = 2$
$\sigma = 0.1$
$\alpha = 2$
$\sigma = 10$
250
−10
−8
−6
−4
−2
0
−1
−0.8
−0.6
−0.4
−0.2
0
−100
−80
−60
−40
−20
0

Fig. S12



H. sapiens (k = 6)
r = 0.95
s / u
(using known mutational biases)
$\Delta s / u$
(using 2-mer null, no bias)
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30
H. sapiens (k = 8)
r = 0.92
s / u
(using known mutational biases)
$\Delta s / u$
(using 2-mer null, no bias)
40
20
0
−20
−40
−60
−80
−80
−60
−40
−20
0
20
40
A. thaliana (k = 6)
r = 0.989
s / u
(using known mutational biases)
$\Delta s / u$
(using 1-mer null, no bias)
0
−10
−20
−30
−40
−40
−30
−20
−10
0
D. melanogaster (k = 6)
r = 0.987
s / u
(using known mutational biases)
$\Delta s / u$
(using 1-mer null, no bias)
5
0
−5
−10
−10
−5
0
5

Fig. S13

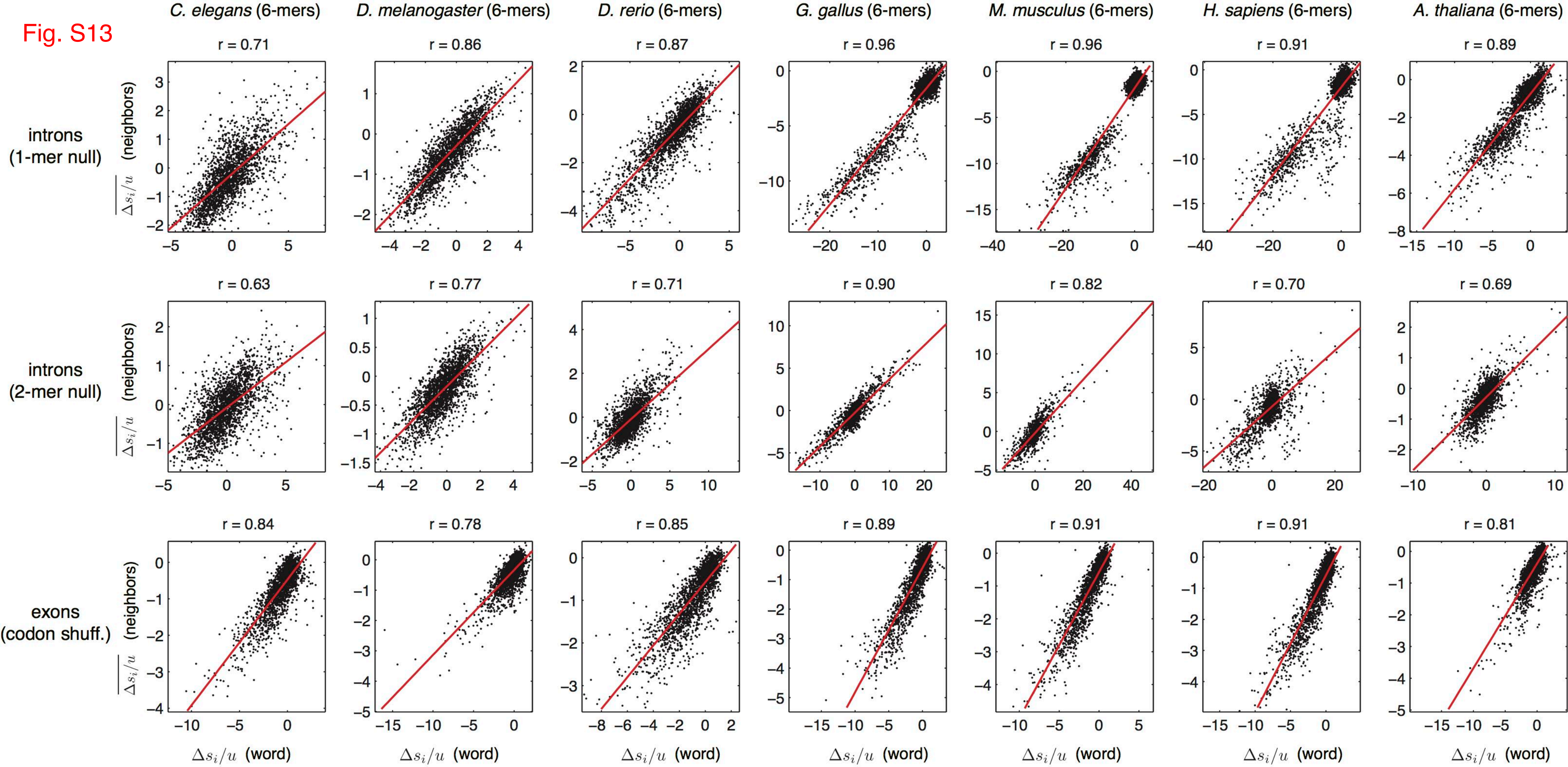
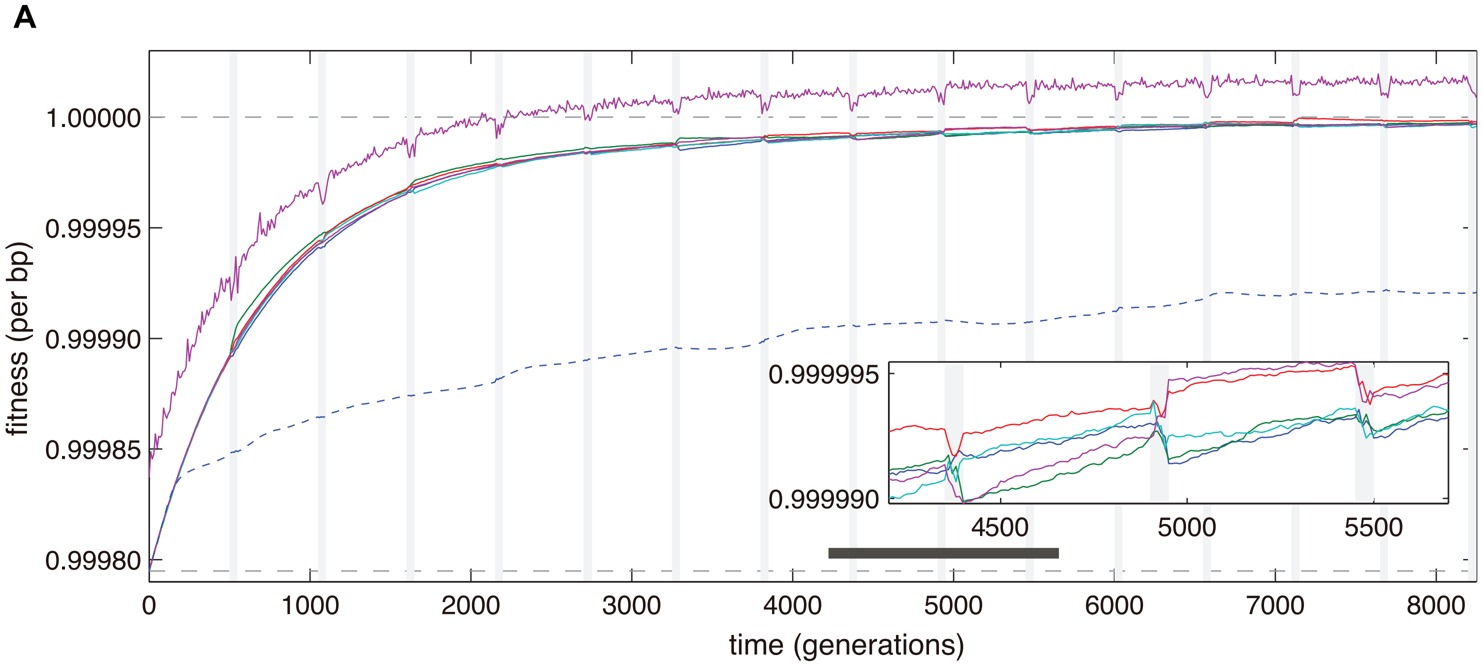

C. elegans (6-mers)
D. melanogaster (6-mers)
D. rerio (6-mers)
G. gallus (6-mers)
M. musculus (6-mers)
H. sapiens (6-mers)
A. thaliana (6-mers)
introns (1-mer null)
r = 0.71
r = 0.86
r = 0.87
r = 0.96
r = 0.96
r = 0.91
r = 0.89
introns (2-mer null)
r = 0.63
r = 0.77
r = 0.71
r = 0.90
r = 0.82
r = 0.70
r = 0.69
exons (codon shuff.)
r = 0.84
r = 0.78
r = 0.85
r = 0.89
r = 0.91
r = 0.91
r = 0.81
$\overline{\Delta s_i/u}$ (neighbors)
$\Delta s_i/u$ (word)

Fig. S14


A
fitness (per bp)
1.00000
0.99995
0.99990
0.99985
0.99980
0
1000
2000
3000
4000
5000
6000
7000
8000
time (generations)
0.999995
0.999990
4500
5000
5500

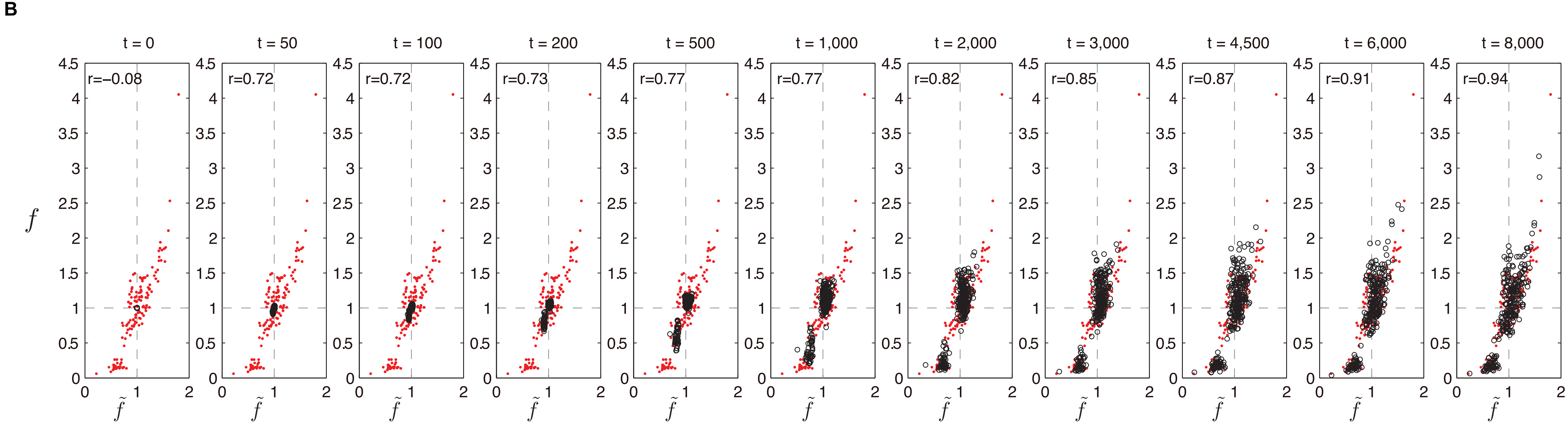

B
t = 0
t = 50
t = 100
t = 200
t = 500
t = 1,000
t = 2,000
t = 3,000
t = 4,500
t = 6,000
t = 8,000
r=−0.08
r=0.72
r=0.72
r=0.73
r=0.77
r=0.77
r=0.82
r=0.85
r=0.87
r=0.91
r=0.94
f
$\tilde{f}$
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

Fig. S15

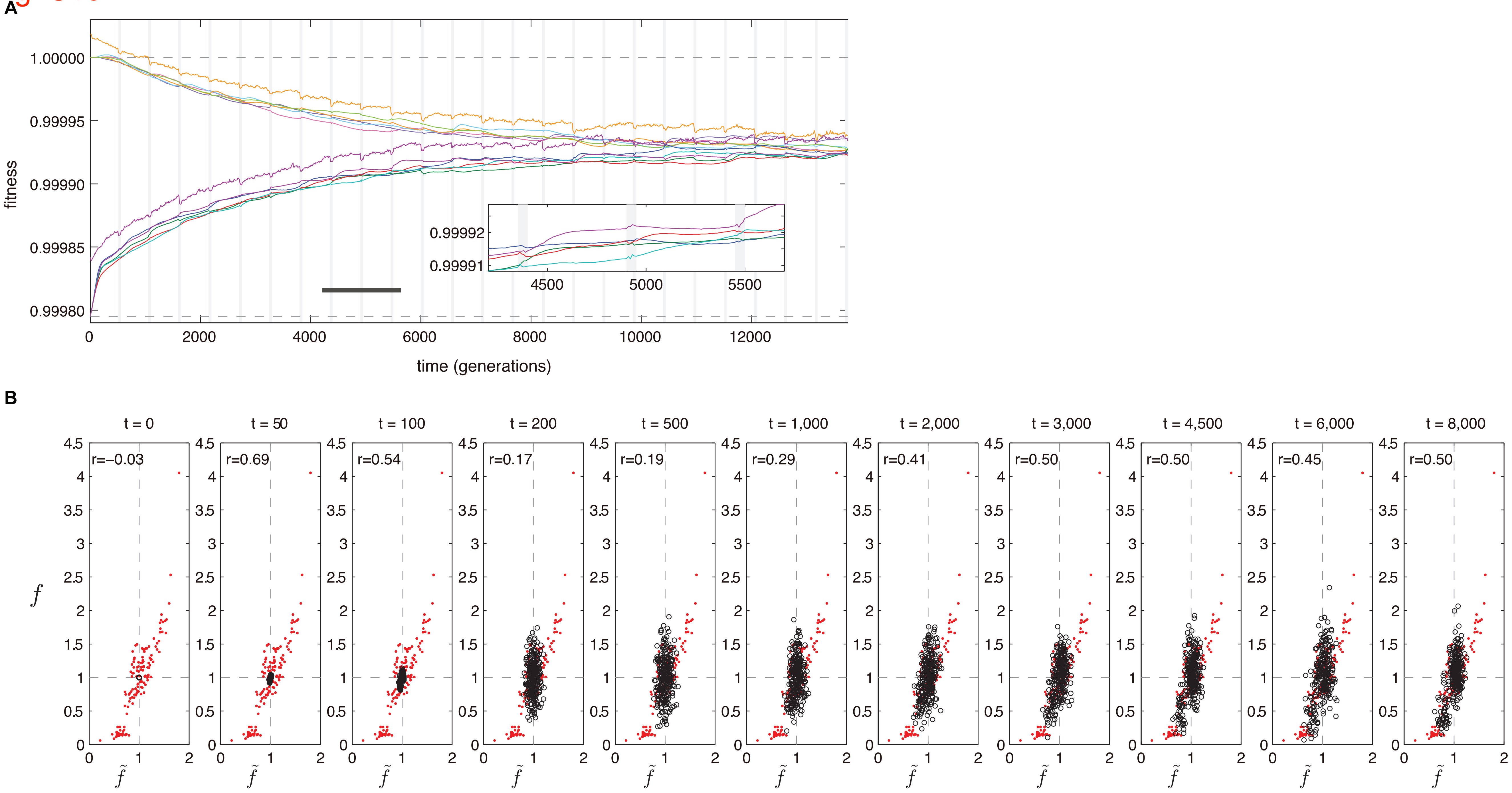

A
fitness
1.00000
0.99995
0.99990
0.99985
0.99980
0.99992
0.99991
4500
5000
5500
0
2000
4000
6000
8000
10000
12000
time (generations)
B
t = 0
t = 50
t = 100
t = 200
t = 500
t = 1,000
t = 2,000
t = 3,000
t = 4,500
t = 6,000
t = 8,000
r=−0.03
r=0.69
r=0.54
r=0.17
r=0.19
r=0.29
r=0.41
r=0.50
r=0.50
r=0.45
r=0.50
f
$\tilde{f}$
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
0
1
2

Fig. S16

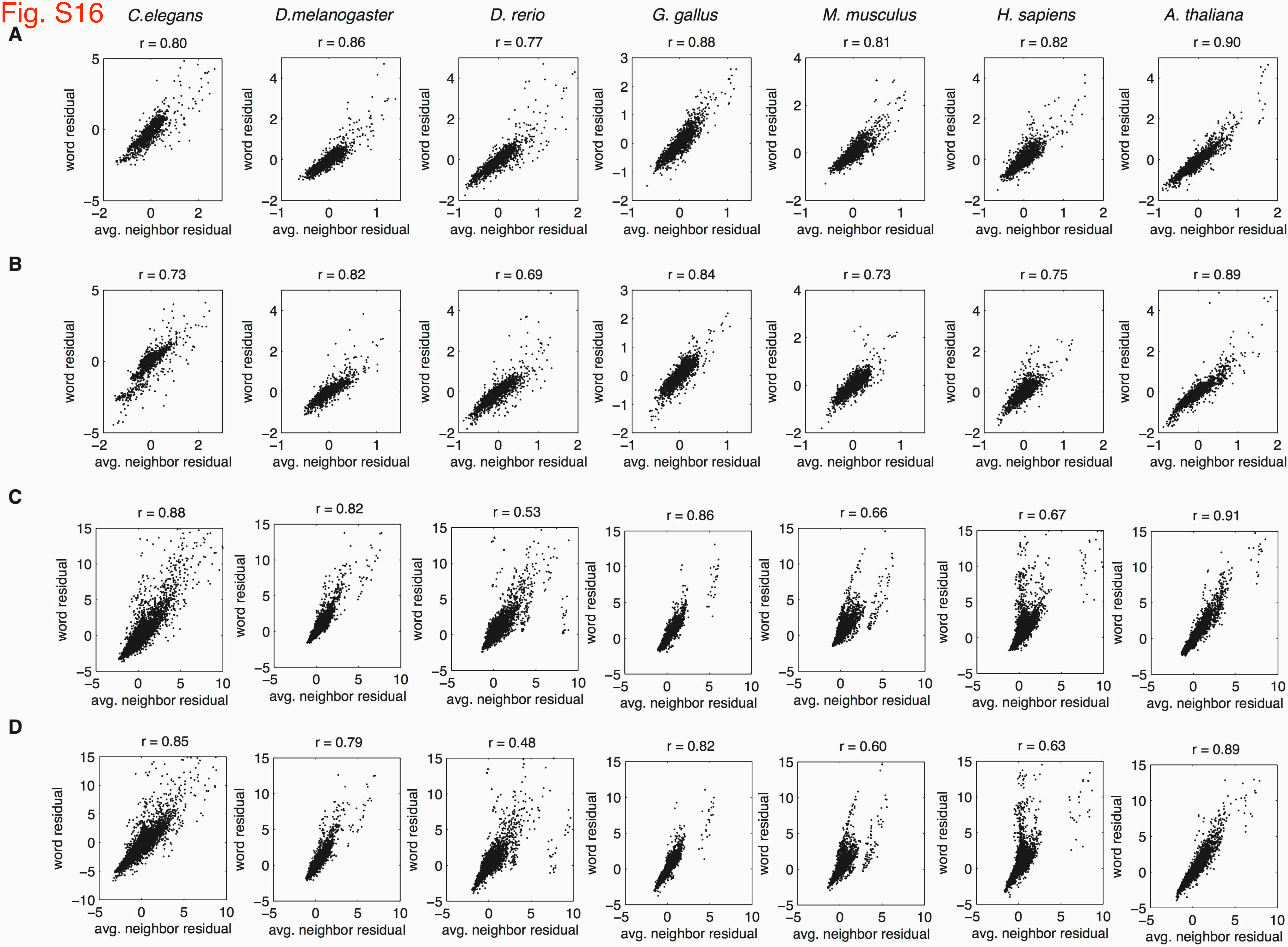

C.elegans
D.melanogaster
D. rerio
G. gallus
M. musculus
H. sapiens
A. thaliana
A
r = 0.80
r = 0.86
r = 0.77
r = 0.88
r = 0.81
r = 0.82
r = 0.90
B
r = 0.73
r = 0.82
r = 0.69
r = 0.84
r = 0.73
r = 0.75
r = 0.89
C
r = 0.88
r = 0.82
r = 0.53
r = 0.86
r = 0.66
r = 0.67
r = 0.91
D
r = 0.85
r = 0.79
r = 0.48
r = 0.82
r = 0.60
r = 0.63
r = 0.89
word residual
avg. neighbor residual

Fig. S17

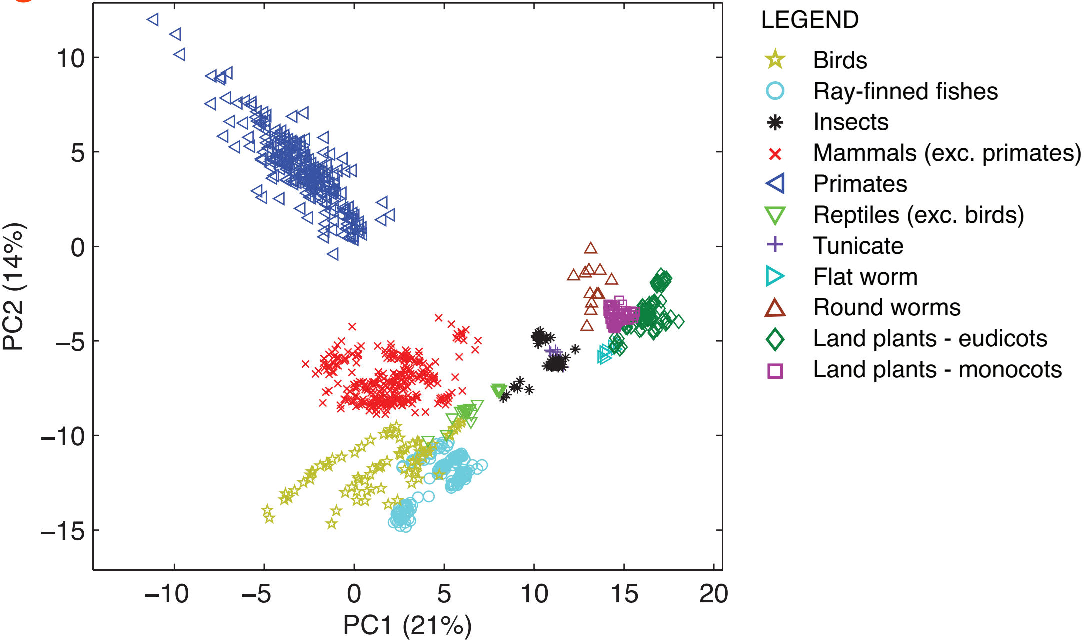
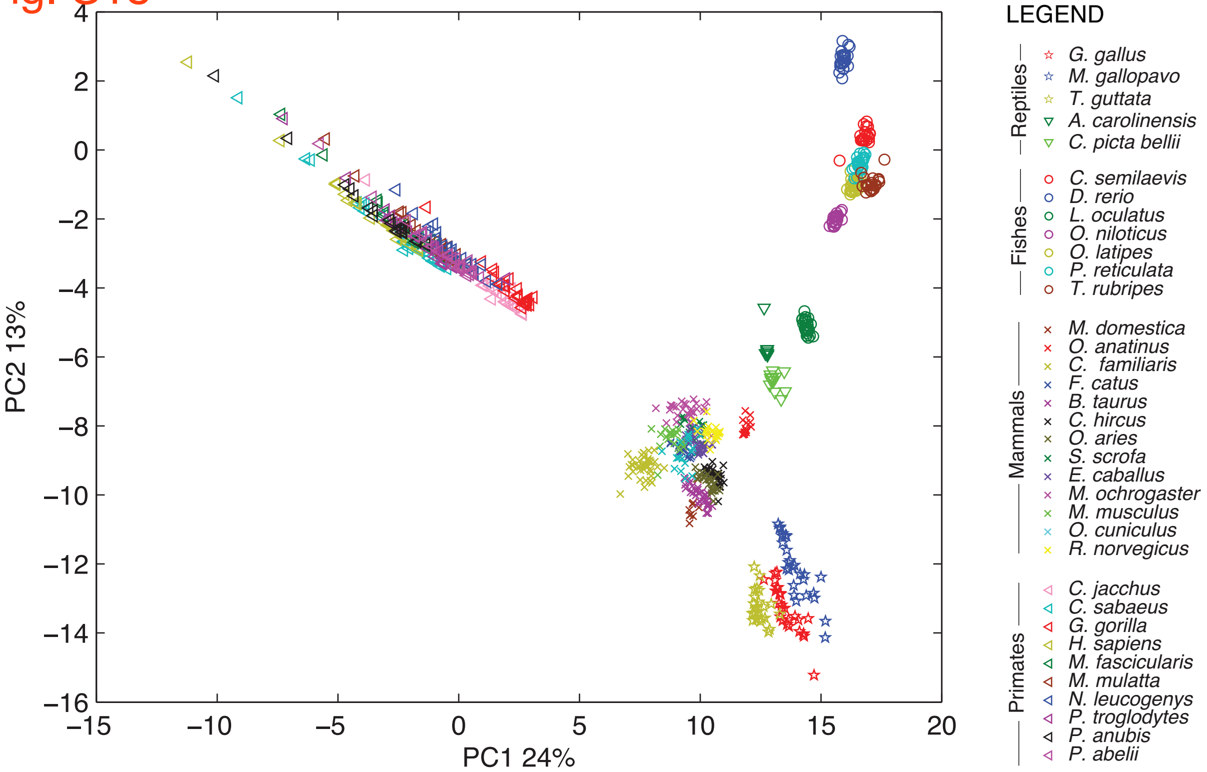
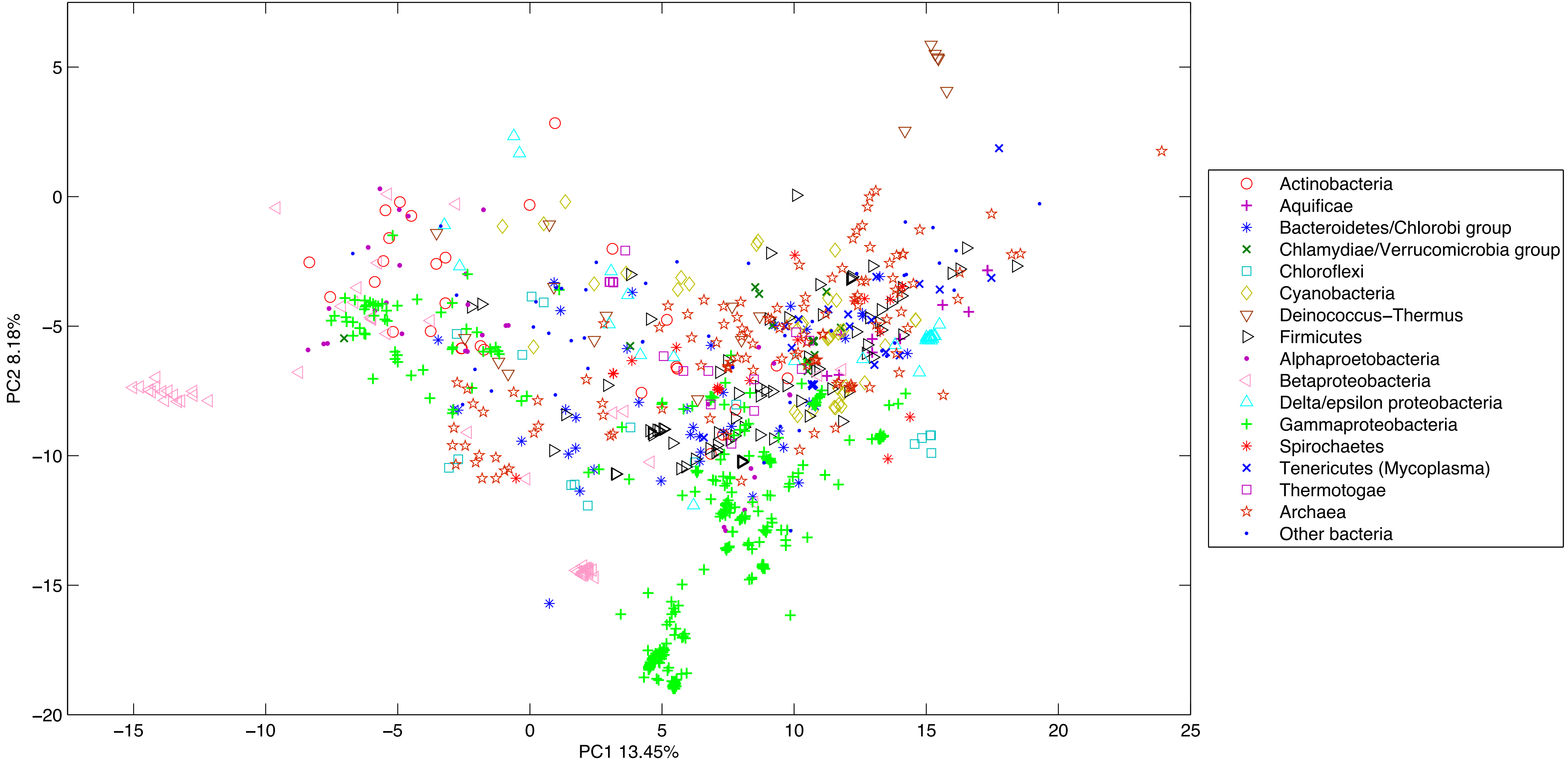

PC2 (14%)
10
5
0
−5
−10
−15
−10
−5
0
5
10
15
20
PC1 (21%)
LEGEND
Birds
Ray-finned fishes
Insects
Mammals (exc. primates)
Primates
Reptiles (exc. birds)
Tunicate
Flat worm
Round worms
Land plants - eudicots
Land plants - monocots

Fig. S18



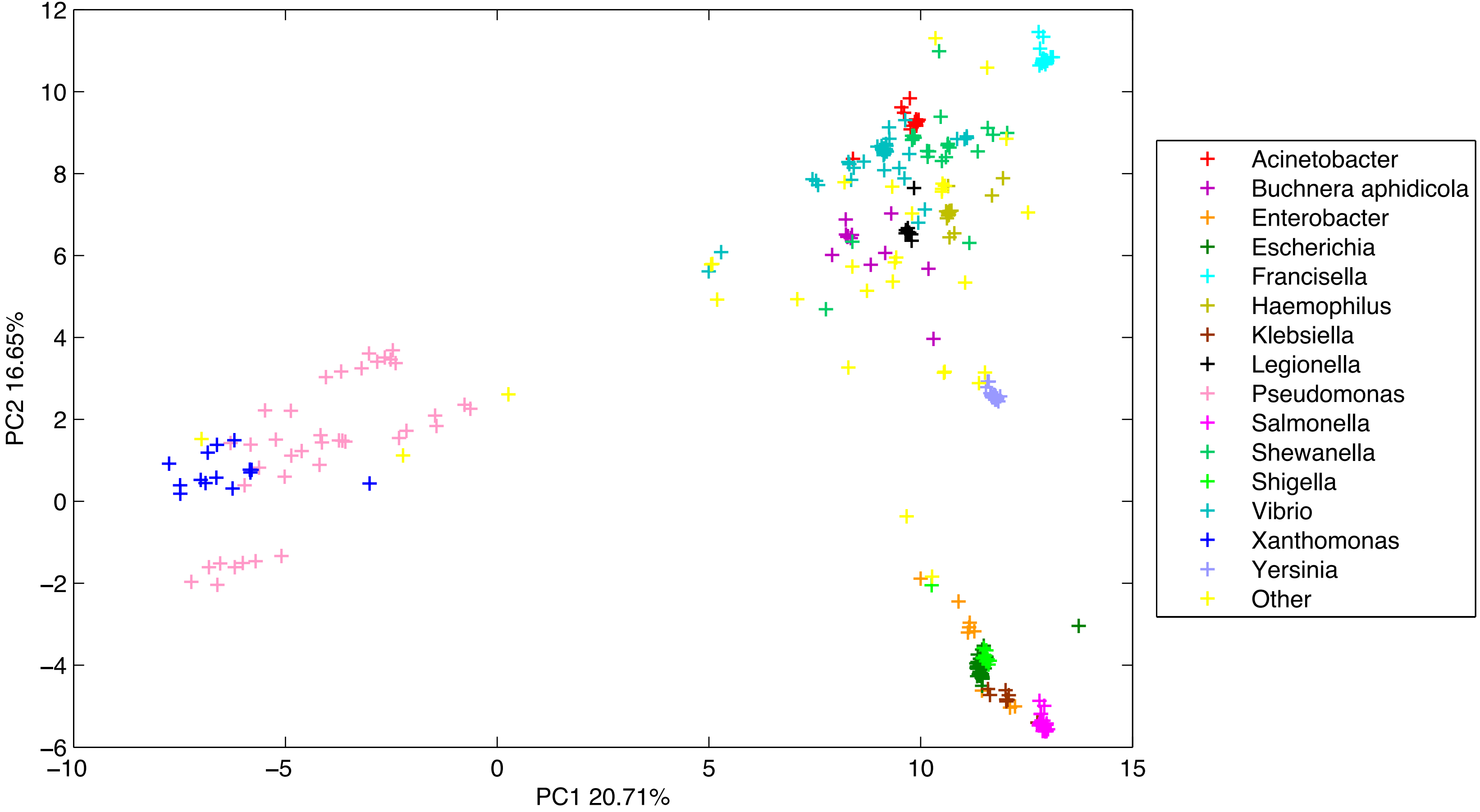
4
2
0
−2
−4
−6
−8
−10
−12
−14
−16
PC2 13%
−15
−10
−5
0
5
10
15
20
PC1 24%
LEGEND
Reptiles
G. gallus
M. gallopavo
T. guttata
A. carolinensis
C. picta bellii
Fishes
C. semilaevis
D. rerio
L. oculatus
O. niloticus
O. latipes
P. reticulata
T. rubripes
Mammals
M. domestica
O. anatinus
C. familiaris
F. catus
B. taurus
C. hircus
O. aries
S. scrofa
E. caballus
M. ochrogaster
M. musculus
O. cuniculus
R. norvegicus
Primates
C. jacchus
C. sabaeus
G. gorilla
H. sapiens
M. fascicularis
M. mulatta
N. leucogenys
P. troglodytes
P. anubis
P. abelii

Fig. S19



5
0
−5
−10
−15
−20
PC2 8.18%
−15
−10
−5
0
5
10
15
20
25
PC1 13.45%
Actinobacteria
Aquificae
Bacteroidetes/Chlorobi group
Chlamydiae/Verrucomicrobia group
Chloroflexi
Cyanobacteria
Deinococcus−Thermus
Firmicutes
Alphaproetobacteria
Betaproteobacteria
Delta/epsilon proteobacteria
Gammaproteobacteria
Spirochaetes
Tenericutes (Mycoplasma)
Thermotogae
Archaea
Other bacteria

Fig. S20



PC2 16.65%
12
10
8
6
4
2
0
−2
−4
−6
−10
−5
0
5
10
15
PC1 20.71%
Acinetobacter
Buchnera aphidicola
Enterobacter
Escherichia
Francisella
Haemophilus
Klebsiella
Legionella
Pseudomonas
Salmonella
Shewanella
Shigella
Vibrio
Xanthomonas
Yersinia
Other

Fig. S21

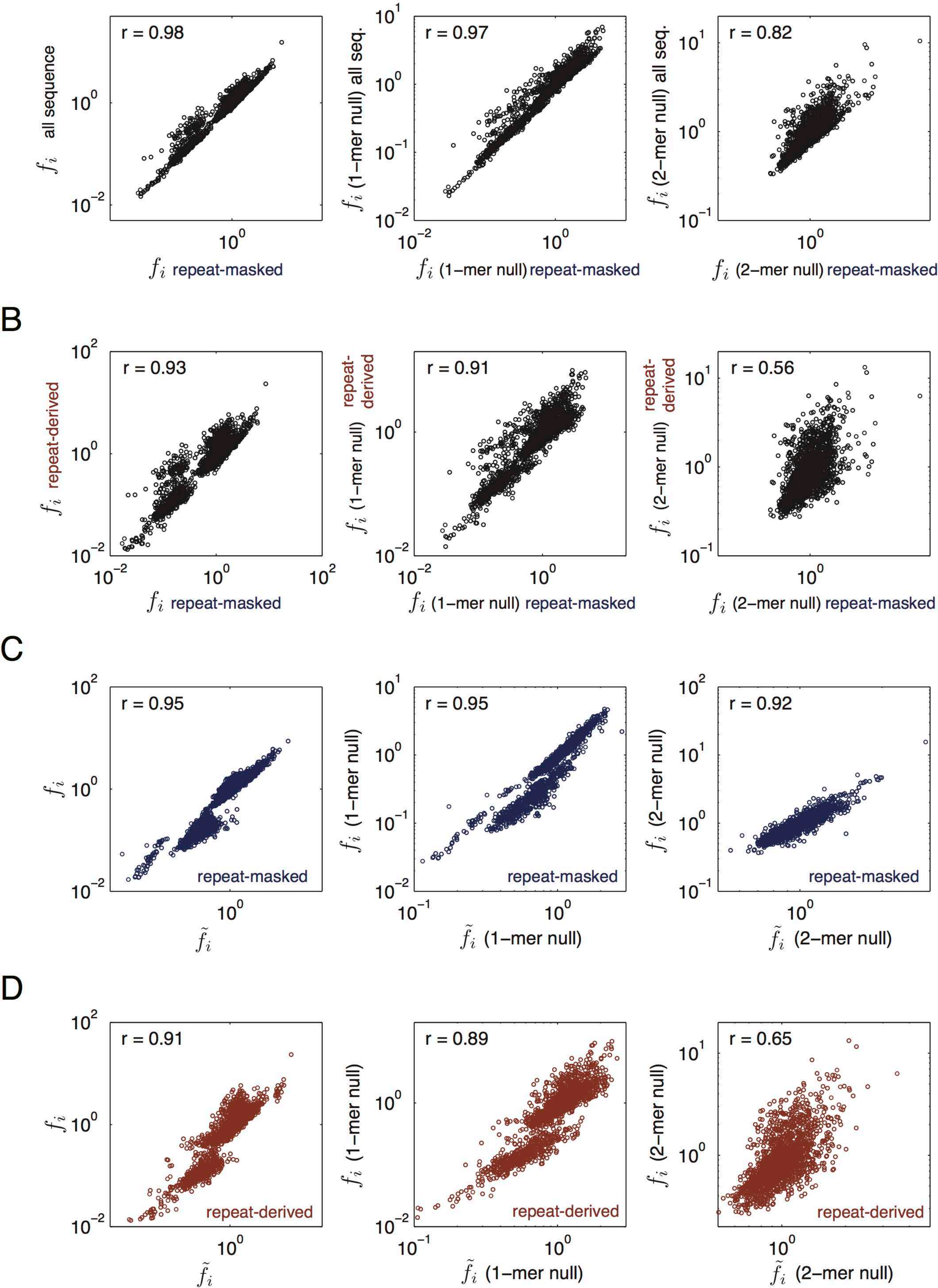

A
r = 0.98
f_i all sequence
f_i repeat-masked
10^{0}
10^{-2}
r = 0.97
f_i (1–mer null) all seq.
f_i (1–mer null) repeat-masked
10^{1}
10^{0}
10^{-1}
10^{-2}
r = 0.82
f_i (2–mer null) all seq.
f_i (2–mer null) repeat-masked
10^{1}
10^{0}
10^{-1}
B
r = 0.93
f_i repeat-derived
f_i repeat-masked
10^{2}
10^{0}
10^{-2}
r = 0.91
f_i (1–mer null) repeat-derived
f_i (1–mer null) repeat-masked
r = 0.56
f_i (2–mer null) repeat-derived
f_i (2–mer null) repeat-masked
C
r = 0.95
f_i
$\tilde{f}_i$
repeat-masked
r = 0.95
f_i (1–mer null)
$\tilde{f}_i$ (1–mer null)
repeat-masked
r = 0.92
f_i (2–mer null)
$\tilde{f}_i$ (2–mer null)
repeat-masked
D
r = 0.91
f_i
$\tilde{f}_i$
repeat-derived
r = 0.89
f_i (1–mer null)
$\tilde{f}_i$ (1–mer null)
repeat-derived
r = 0.65
f_i (2–mer null)
$\tilde{f}_i$ (2–mer null)
repeat-derived

Fig. S22

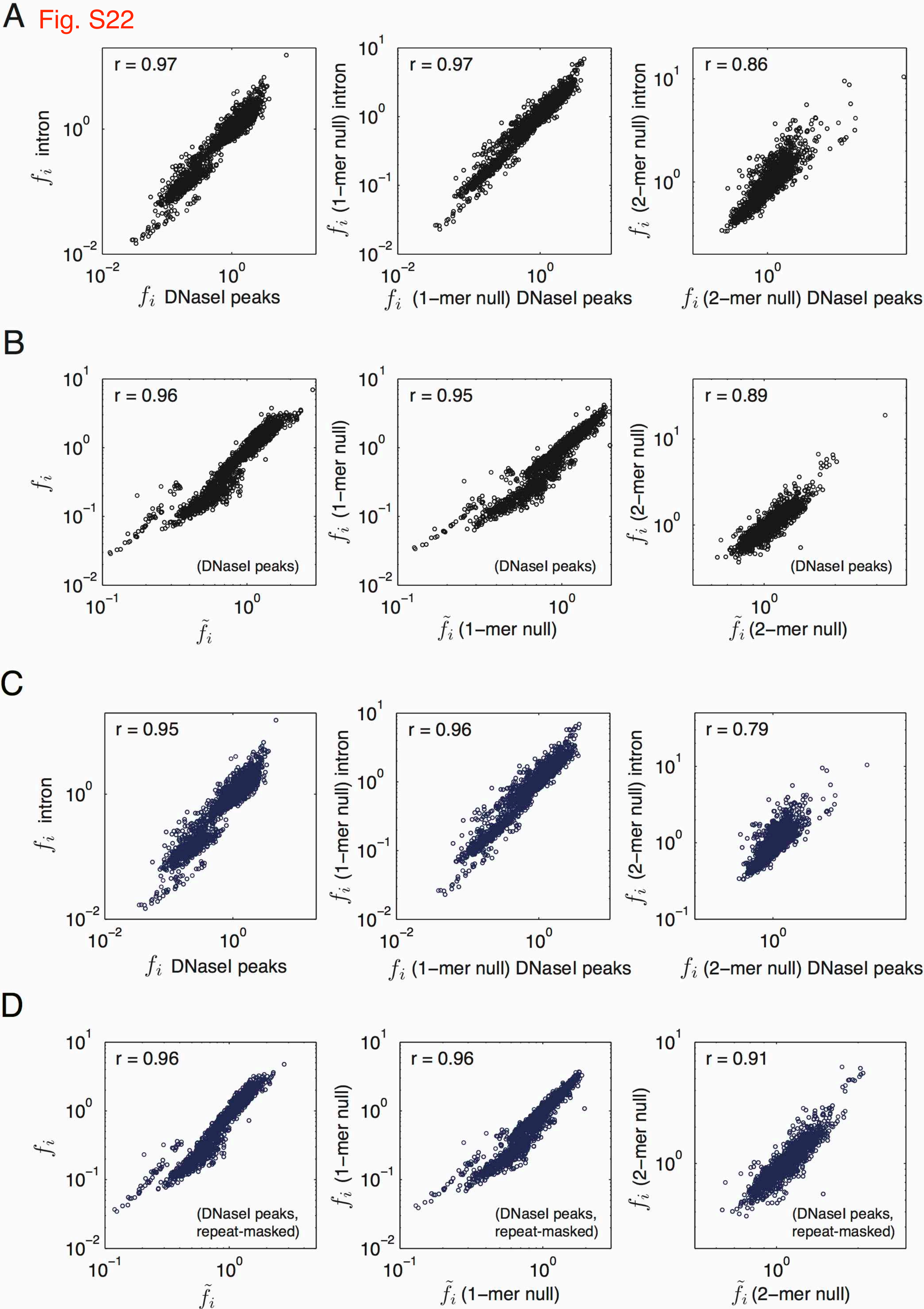

A
r = 0.97
f_i intron
f_i DNaseI peaks
r = 0.97
f_i (1–mer null) intron
f_i (1–mer null) DNaseI peaks
r = 0.86
f_i (2–mer null) intron
f_i (2–mer null) DNaseI peaks
10^{-2}
10^{-1}
10^{0}
10^{1}
B
r = 0.96
f_i
$\tilde{f}_i$
(DNaseI peaks)
r = 0.95
f_i (1–mer null)
$\tilde{f}_i$ (1–mer null)
(DNaseI peaks)
r = 0.89
f_i (2–mer null)
$\tilde{f}_i$ (2–mer null)
(DNaseI peaks)
C
r = 0.95
f_i intron
f_i DNaseI peaks
r = 0.96
f_i (1–mer null) intron
f_i (1–mer null) DNaseI peaks
r = 0.79
f_i (2–mer null) intron
f_i (2–mer null) DNaseI peaks
D
r = 0.96
f_i
$\tilde{f}_i$
(DNaseI peaks, repeat-masked)
r = 0.96
f_i (1–mer null)
$\tilde{f}_i$ (1–mer null)
(DNaseI peaks, repeat-masked)
r = 0.91
f_i (2–mer null)
$\tilde{f}_i$ (2–mer null)
(DNaseI peaks, repeat-masked)

Fig. S23

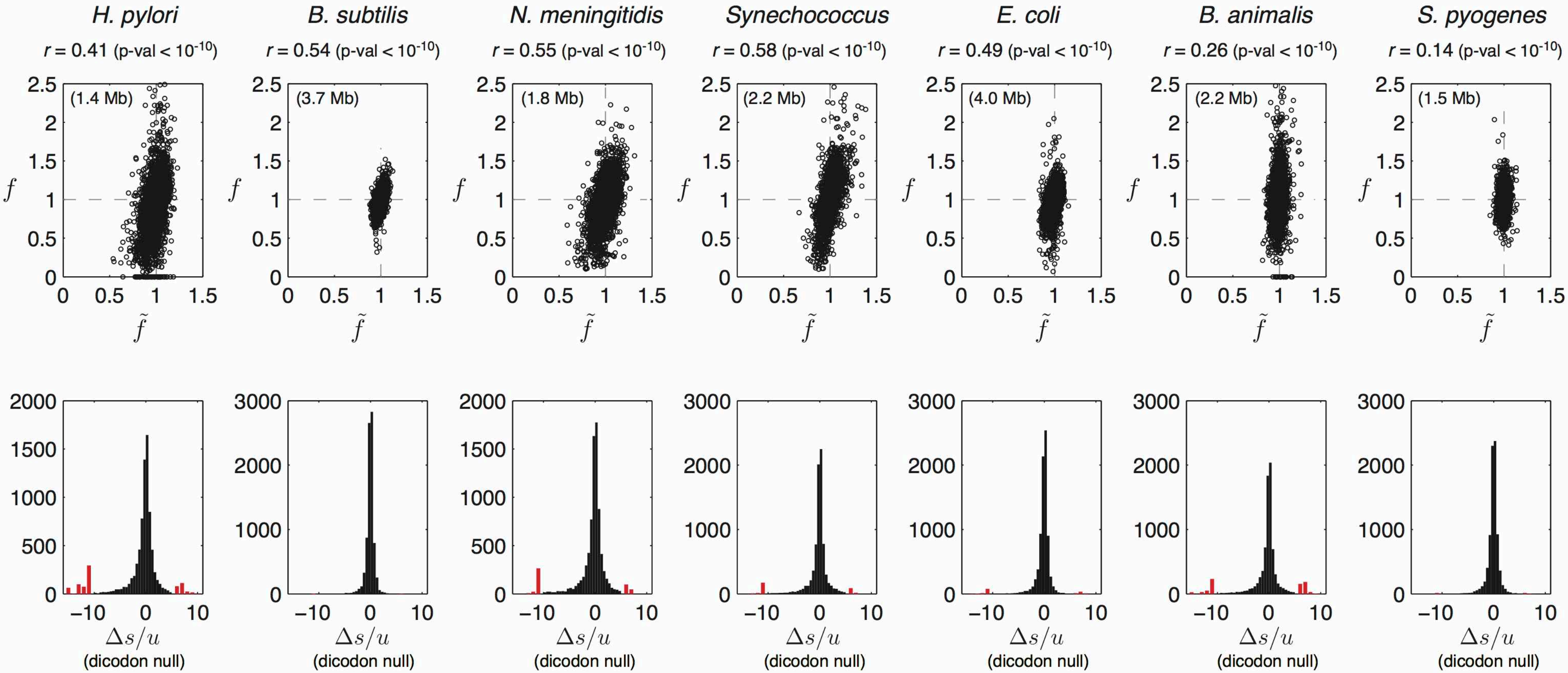
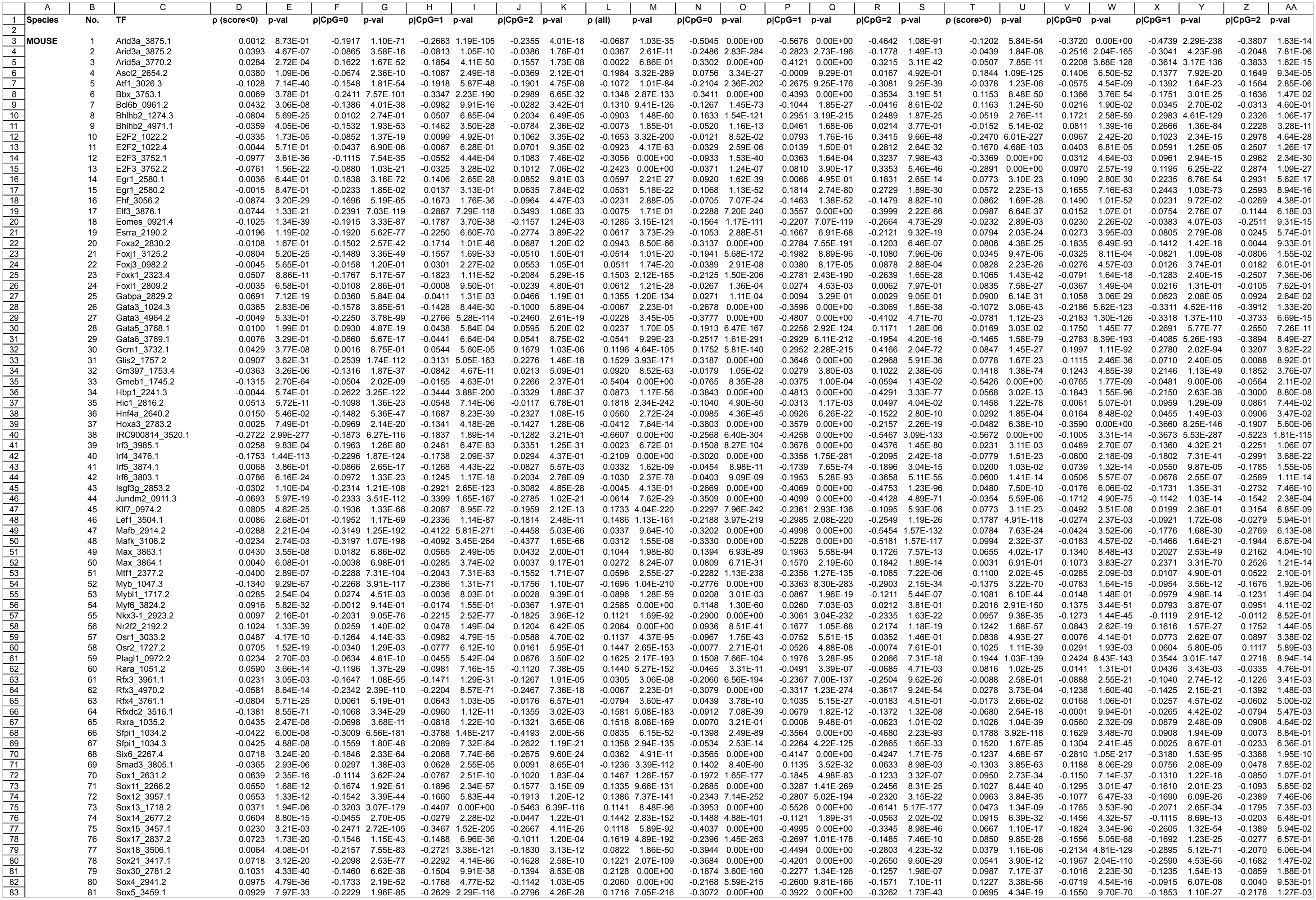

H. pylori
$r = 0.41$ (p-val < 10^{-10})
(1.4 Mb)
B. subtilis
$r = 0.54$ (p-val < 10^{-10})
(3.7 Mb)
N. meningitidis
$r = 0.55$ (p-val < 10^{-10})
(1.8 Mb)
Synechococcus
$r = 0.58$ (p-val < 10^{-10})
(2.2 Mb)
E. coli
$r = 0.49$ (p-val < 10^{-10})
(4.0 Mb)
B. animalis
$r = 0.26$ (p-val < 10^{-10})
(2.2 Mb)
S. pyogenes
$r = 0.14$ (p-val < 10^{-10})
(1.5 Mb)
f
0
0.5
1
1.5
2
2.5
$\tilde{f}$
0
0.5
1
1.5
0
500
1000
1500
2000
3000
−10
0
10
$\Delta s/u$
(dicodon null)

Table S1



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	Species	No.	TF	ρ (score<0)	p-val	ρ|CpG=0	p-val	ρ|CpG=1	p-val	ρ|CpG=2	p-val	ρ (all)	p-val	ρ|CpG=0	p-val	ρ|CpG=1	p-val	ρ|CpG=2	p-val	ρ (score>0)	p-val	ρ|CpG=0	p-val	ρ|CpG=1	p-val	ρ|CpG=2	p-val
2																											
3	MOUSE	1	Arid3a_3875.1	0.0012	8.73E-01	-0.1917	1.10E-71	-0.2663	1.19E-105	-0.2355	4.01E-18	-0.0687	1.03E-35	-0.5045	0.00E+00	-0.5676	0.00E+00	-0.4642	1.08E-91	-0.1202	5.84E-54	-0.3720	0.00E+00	-0.4739	2.29E-238	-0.3807	1.63E-14
4		2	Arid3a_3875.2	0.0393	4.67E-07	-0.0865	3.58E-16	-0.0813	1.05E-10	-0.0386	1.76E-01	0.0367	2.61E-11	-0.2486	2.83E-284	-0.2823	2.73E-196	-0.1778	1.49E-13	-0.0439	1.84E-08	-0.2516	2.04E-165	-0.3041	4.23E-96	-0.2048	7.81E-06
5		3	Arid5a_3770.2	0.0284	2.72E-04	-0.1622	1.67E-52	-0.1854	4.11E-50	-0.1557	1.73E-08	0.0022	6.86E-01	-0.3302	0.00E+00	-0.4121	0.00E+00	-0.3215	3.11E-42	-0.0507	7.85E-11	-0.2208	3.68E-128	-0.3614	3.17E-136	-0.3833	1.62E-15
6		4	Ascl2_2654.2	0.0380	1.09E-06	-0.0674	2.36E-10	-0.1087	2.49E-18	-0.0369	2.12E-01	0.1984	3.32E-289	0.0756	3.34E-27	-0.0009	9.29E-01	0.0167	4.92E-01	0.1844	1.09E-125	0.1406	6.50E-52	0.1377	7.92E-20	0.1649	9.34E-05
7		5	Atf1_3026.3	-0.1028	7.14E-40	-0.1548	1.81E-54	-0.1918	5.87E-48	-0.1901	4.75E-08	-0.1072	1.01E-84	-0.2104	2.36E-202	-0.2675	9.25E-176	-0.3081	9.25E-39	-0.0378	1.23E-06	-0.0575	4.54E-09	-0.1392	1.64E-23	-0.1564	2.85E-06
8		6	Bbx_3753.1	0.0069	3.78E-01	-0.2411	7.57E-101	-0.3347	2.23E-190	-0.2989	6.65E-32	0.1348	2.87E-133	-0.3411	0.00E+00	-0.4393	0.00E+00	-0.3534	3.19E-51	0.1153	8.48E-50	-0.1366	3.76E-54	-0.1751	3.01E-25	-0.1636	1.47E-02
9		7	Bcl6b_0961.2	0.0432	3.06E-08	-0.1386	4.01E-38	-0.0982	9.91E-16	-0.0282	3.42E-01	0.1310	9.41E-126	-0.1267	1.45E-73	-0.1044	1.85E-27	-0.0416	8.61E-02	0.1163	1.24E-50	0.0216	1.90E-02	0.0345	2.70E-02	-0.0313	4.60E-01
10		8	Bhlhb2_1274.3	-0.0804	5.69E-25	0.0102	2.74E-01	0.0507	6.85E-04	0.2034	6.49E-05	-0.0903	1.48E-60	0.1633	1.54E-121	0.2951	3.19E-215	0.2489	1.87E-25	-0.0519	2.76E-11	0.1721	2.58E-59	0.2983	4.61E-129	0.2326	1.06E-17
11		9	Bhlhb2_4971.1	-0.0359	4.05E-06	-0.1532	1.93E-53	-0.1462	3.50E-28	-0.0784	2.36E-02	-0.0073	1.85E-01	-0.0520	1.16E-13	0.0461	1.68E-06	0.0214	3.77E-01	-0.0152	5.14E-02	0.0811	1.39E-16	0.2666	1.36E-84	0.2228	3.28E-11
12		10	E2F2_1022.2	-0.0335	1.73E-05	-0.0852	1.37E-19	0.0099	4.92E-01	0.1062	3.35E-02	-0.1653	3.32E-200	-0.0121	8.52E-02	0.0793	1.76E-16	0.3415	9.66E-48	-0.2470	6.01E-227	0.0967	2.42E-20	0.1023	2.34E-15	0.2978	4.64E-28
13		11	E2F2_1022.4	-0.0044	5.71E-01	-0.0437	6.90E-06	-0.0067	6.28E-01	0.0701	9.35E-02	-0.0923	4.17E-63	-0.0329	2.59E-06	0.0139	1.50E-01	0.2812	2.64E-32	-0.1670	4.68E-103	0.0403	6.81E-05	0.0591	1.25E-05	0.2507	1.26E-17
14		12	E2F3_3752.1	-0.0977	3.61E-36	-0.1115	7.54E-35	-0.0552	4.44E-04	0.1083	7.46E-02	-0.3056	0.00E+00	-0.0933	1.53E-40	0.0363	1.64E-04	0.3237	7.98E-43	-0.3369	0.00E+00	0.0312	4.64E-03	0.0961	2.94E-15	0.2962	2.34E-30
15		13	E2F3_3752.2	-0.0761	1.56E-22	-0.0880	1.03E-21	-0.0325	3.28E-02	0.1012	7.06E-02	-0.2423	0.00E+00	-0.0371	1.24E-07	0.0810	3.90E-17	0.3353	5.46E-46	-0.2891	0.00E+00	0.0970	2.57E-19	0.1195	6.25E-22	0.2874	1.09E-27
16		14	Egr1_2580.1	0.0036	6.44E-01	-0.1838	3.16E-72	-0.1406	2.65E-28	-0.0852	9.81E-03	0.0597	2.21E-27	-0.0920	1.62E-39	0.0066	4.95E-01	0.1831	2.65E-14	0.0773	3.10E-23	0.1090	2.80E-30	0.2235	6.76E-54	0.2931	5.62E-17
17		15	Egr1_2580.2	-0.0015	8.47E-01	-0.0233	1.85E-02	0.0137	3.13E-01	0.0635	7.84E-02	0.0531	5.18E-22	0.1068	1.13E-52	0.1814	2.74E-80	0.2729	1.89E-30	0.0572	2.23E-13	0.1655	7.16E-63	0.2443	1.03E-73	0.2593	8.94E-16
18		16	Ehf_3056.2	-0.0874	3.20E-29	-0.1696	5.19E-65	-0.1673	1.76E-36	-0.0964	4.47E-03	-0.0231	2.88E-05	-0.0705	7.07E-24	-0.1463	1.38E-52	-0.1479	8.82E-10	0.0862	1.69E-28	0.1490	1.01E-52	0.0231	9.72E-02	-0.0269	4.38E-01
19		17	Elf3_3876.1	-0.0744	1.33E-21	-0.2391	7.03E-119	-0.2887	7.29E-118	-0.3493	1.06E-33	-0.0075	1.71E-01	-0.2288	7.20E-240	-0.3557	0.00E+00	-0.3999	2.22E-66	0.0987	6.64E-37	0.0152	1.07E-01	-0.0754	2.76E-07	-0.1144	6.18E-03
20		18	Eomes_0921.4	-0.1025	1.34E-39	-0.1915	3.33E-87	-0.1787	3.70E-38	-0.1157	1.24E-03	-0.1286	3.15E-121	-0.1564	1.17E-111	-0.2207	7.07E-119	-0.2664	4.73E-29	-0.0232	2.89E-03	0.0230	2.26E-02	-0.0383	4.07E-03	-0.2511	9.31E-15
21		19	Esrra_2190.2	-0.0196	1.19E-02	-0.1920	5.62E-77	-0.2250	6.60E-70	-0.2774	3.89E-22	0.0617	3.73E-29	-0.1053	2.88E-51	-0.1667	6.91E-68	-0.2121	9.32E-19	0.0794	2.03E-24	0.0273	3.95E-03	0.0805	2.79E-08	0.0245	5.74E-01
22		20	Foxa2_2830.2	-0.0108	1.67E-01	-0.1502	2.57E-42	-0.1714	1.01E-46	-0.0687	1.20E-02	0.0943	8.50E-66	-0.3137	0.00E+00	-0.2784	7.55E-191	-0.1203	6.46E-07	0.0806	4.38E-25	-0.1835	6.49E-93	-0.1412	1.42E-18	0.0044	9.33E-01
23		21	Foxj1_3125.2	-0.0804	5.20E-25	-0.1489	3.36E-49	-0.1557	1.69E-33	-0.0510	1.50E-01	-0.0514	1.01E-20	-0.1941	5.68E-172	-0.1982	8.89E-96	-0.1080	7.96E-06	0.0345	9.47E-06	-0.0325	8.11E-04	-0.0821	1.09E-08	-0.0806	1.55E-02
24		22	Foxj3_0982.2	-0.0045	5.65E-01	-0.0158	1.20E-01	0.0301	2.27E-02	0.0553	1.05E-01	0.0511	1.74E-20	-0.0389	2.91E-08	0.0380	8.17E-05	0.0878	2.88E-04	0.0828	2.23E-26	-0.0276	4.57E-03	0.0126	3.74E-01	0.0182	6.01E-01
25		23	Foxk1_2323.4	0.0507	8.86E-11	-0.1767	5.17E-57	-0.1823	1.11E-52	-0.2084	5.29E-15	0.1503	2.12E-165	-0.2125	1.50E-206	-0.2781	2.43E-190	-0.2639	1.65E-28	0.1065	1.43E-42	-0.0791	1.64E-18	-0.1283	2.40E-15	-0.2507	7.36E-06
26		24	Foxl1_2809.2	-0.0035	6.58E-01	-0.0108	2.86E-01	-0.0008	9.50E-01	-0.0239	4.80E-01	0.0612	1.21E-28	-0.0267	1.36E-04	0.0274	4.53E-03	0.0062	7.97E-01	0.0835	7.58E-27	-0.0367	1.49E-04	0.0216	1.31E-01	-0.0105	7.62E-01
27		25	Gabpa_2829.2	0.0691	7.12E-19	-0.0360	5.84E-04	-0.0411	1.31E-03	-0.0466	1.19E-01	0.1355	1.20E-134	0.0271	1.11E-04	-0.0094	3.29E-01	0.0029	9.05E-01	0.0900	6.14E-31	0.1058	3.06E-29	0.0623	2.08E-05	0.0924	2.64E-02
28		26	Gata3_1024.3	0.0365	2.83E-06	-0.1578	3.85E-51	-0.1428	8.44E-30	-0.1000	5.89E-04	-0.0067	2.23E-01	-0.2678	0.00E+00	-0.3596	0.00E+00	-0.3069	1.85E-38	-0.1072	3.06E-43	-0.2186	5.62E-123	-0.3311	4.52E-116	-0.3912	1.33E-20
29		27	Gata3_4964.2	-0.0049	5.33E-01	-0.2250	3.78E-99	-0.2766	5.28E-114	-0.2460	2.61E-19	-0.0228	3.45E-05	-0.3777	0.00E+00	-0.4807	0.00E+00	-0.4102	4.71E-70	-0.0781	1.12E-23	-0.2183	1.30E-126	-0.3318	1.37E-110	-0.3733	6.69E-15
30		28	Gata5_3768.1	0.0100	1.99E-01	-0.0930	4.87E-19	-0.0438	5.84E-04	0.0595	5.20E-02	0.0237	1.70E-05	-0.1913	6.47E-167	-0.2256	2.92E-124	-0.1171	1.28E-06	-0.0169	3.03E-02	-0.1750	1.45E-77	-0.2691	5.77E-77	-0.2550	7.26E-11
31		29	Gata6_3769.1	0.0076	3.29E-01	-0.0860	5.67E-17	-0.0441	6.64E-04	0.0541	8.75E-02	-0.0541	9.29E-23	-0.2517	1.61E-291	-0.2929	6.11E-212	-0.1954	4.20E-16	-0.1465	1.58E-79	-0.2783	8.39E-193	-0.4085	5.26E-193	-0.3894	8.49E-27
32		30	Gcm1_3732.1	0.0429	3.77E-08	0.0016	8.75E-01	0.0544	5.60E-05	0.1679	1.03E-06	0.1196	4.64E-105	0.1752	5.81E-140	0.2952	2.28E-215	0.4166	2.04E-72	0.0847	1.45E-27	0.1997	1.11E-92	0.2780	2.02E-94	0.3207	3.82E-22
33		31	Glis2_1757.2	0.0907	3.62E-31	-0.2539	1.74E-112	-0.3131	5.05E-163	-0.2276	1.46E-18	0.1529	3.93E-171	-0.3187	0.00E+00	-0.3646	0.00E+00	-0.2968	5.91E-36	0.0778	1.67E-23	-0.1115	2.46E-36	-0.0710	2.40E-05	0.0088	8.92E-01
34		32	Gm397_1753.4	-0.0363	3.26E-06	-0.1316	1.87E-37	-0.0842	4.67E-11	0.0213	5.09E-01	0.0920	8.52E-63	-0.0179	1.05E-02	0.0279	3.80E-03	0.1022	2.38E-05	0.1418	1.38E-74	0.1243	4.85E-39	0.2146	1.13E-49	0.1852	3.76E-07
35		33	Gmeb1_1745.2	-0.1315	2.70E-64	-0.0504	2.02E-09	-0.0155	4.63E-01	0.2266	2.37E-01	-0.5404	0.00E+00	-0.0765	8.35E-28	-0.0375	1.00E-04	-0.0594	1.43E-02	-0.5426	0.00E+00	-0.0765	1.77E-09	-0.0481	9.00E-06	-0.0564	2.11E-02
36		34	Hbp1_2241.3	-0.0044	5.74E-01	-0.2622	3.25E-122	-0.3444	3.88E-200	-0.3329	1.88E-37	0.0873	1.17E-56	-0.3843	0.00E+00	-0.4813	0.00E+00	-0.4291	3.33E-77	0.0568	3.02E-13	-0.1843	1.55E-96	-0.2150	2.63E-38	-0.3000	8.80E-08
37		35	Hic1_2816.2	0.0513	5.72E-11	-0.1098	1.36E-23	-0.0548	7.14E-06	-0.0117	6.78E-01	0.1818	2.34E-242	-0.1040	4.90E-50	-0.0313	1.17E-03	0.0497	4.04E-02	0.1458	1.22E-78	0.0061	5.07E-01	0.0959	1.29E-09	0.0861	7.44E-02
38		36	Hnf4a_2640.2	0.0150	5.46E-02	-0.1482	5.36E-47	-0.1687	8.23E-39	-0.2327	1.08E-15	0.0560	2.72E-24	-0.0985	4.36E-45	-0.0926	6.26E-22	-0.1522	2.80E-10	0.0292	1.85E-04	0.0164	8.48E-02	0.0455	1.49E-03	0.0906	3.47E-02
39		37	Hoxa3_2783.2	0.0025	7.49E-01	-0.0969	2.14E-20	-0.1341	4.18E-26	-0.1427	1.28E-06	-0.0412	7.64E-14	-0.3803	0.00E+00	-0.3579	0.00E+00	-0.2157	2.26E-19	-0.0482	6.38E-10	-0.3590	0.00E+00	-0.3660	8.25E-146	-0.1907	5.60E-06
40		38	IRC900814_3520.1	-0.2722	2.99E-277	-0.1873	6.27E-116	-0.1837	1.89E-14	-0.1282	3.21E-01	-0.6607	0.00E+00	-0.2568	6.40E-304	-0.4258	0.00E+00	-0.5467	3.09E-133	-0.5672	0.00E+00	-0.1005	3.31E-14	-0.3673	5.53E-287	-0.5223	1.81E-115
41		39	Irf3_3985.1	-0.0258	9.83E-04	-0.1963	1.26E-80	-0.2461	6.47E-83	-0.3351	1.25E-31	-0.0023	6.72E-01	-0.1508	8.27E-104	-0.3678	0.00E+00	-0.4376	1.45E-80	0.0231	3.11E-03	0.0489	2.70E-07	-0.1360	4.32E-21	-0.2251	1.06E-07
42		40	Irf4_3476.1	-0.1753	1.44E-113	-0.2296	1.87E-124	-0.1738	2.09E-37	0.0294	4.37E-01	-0.2109	0.00E+00	-0.3020	0.00E+00	-0.3356	1.75E-281	-0.2095	2.42E-18	-0.0779	1.51E-23	-0.0600	2.18E-09	-0.1802	7.31E-41	-0.2991	3.68E-22
43		41	Irf5_3874.1	0.0068	3.86E-01	-0.0866	2.65E-17	-0.1268	4.43E-22	-0.0827	5.57E-03	0.0332	1.62E-09	-0.0454	8.98E-11	-0.1739	7.65E-74	-0.1896	3.04E-15	0.0200	1.03E-02	0.0739	1.32E-14	-0.0550	9.87E-05	-0.1785	1.55E-05
44		42	Irf6_3803.1	-0.0786	6.16E-24	-0.0972	1.33E-23	-0.1245	1.17E-18	-0.2034	2.78E-09	-0.1030	2.37E-78	-0.0403	9.09E-09	-0.1953	5.28E-93	-0.3658	5.11E-55	-0.0600	1.41E-14	0.0506	5.57E-07	-0.0678	2.55E-07	-0.2589	1.11E-14
45		43	Isgf3g_2853.2	-0.0302	1.10E-04	-0.2314	1.21E-108	-0.2921	2.65E-123	-0.3082	4.85E-28	-0.0045	4.13E-01	-0.2669	0.00E+00	-0.4069	0.00E+00	-0.4753	1.23E-96	0.0480	7.50E-10	-0.0176	6.06E-02	-0.1731	1.35E-31	-0.2732	7.46E-10
46		44	Jundm2_0911.3	-0.0693	5.97E-19	-0.2333	3.51E-112	-0.3399	1.65E-167	-0.2785	1.02E-21	-0.0614	7.62E-29	-0.3509	0.00E+00	-0.4099	0.00E+00	-0.4128	4.89E-71	-0.0354	5.59E-06	-0.1712	4.90E-75	-0.1142	1.03E-14	-0.1542	2.38E-04
47		45	Klf7_0974.2	0.0805	4.62E-25	-0.1936	1.33E-66	-0.2087	8.95E-72	-0.1959	2.12E-13	0.1733	4.04E-220	-0.2297	7.96E-242	-0.2361	2.93E-136	-0.1095	5.93E-06	0.0773	3.11E-23	-0.0492	3.51E-08	0.0199	2.36E-01	0.3154	6.85E-09
48		46	Lef1_3504.1	0.0086	2.68E-01	-0.1952	1.17E-69	-0.2336	1.14E-87	-0.1814	2.48E-11	0.1486	1.13E-161	-0.2188	3.97E-219	-0.2985	2.08E-220	-0.2549	1.19E-26	0.1787	4.91E-118	-0.0274	2.37E-03	-0.0921	1.72E-08	-0.0279	5.94E-01
49		47	Mafb_2914.2	-0.0288	2.21E-04	-0.3149	1.25E-192	-0.4122	5.81E-271	-0.4458	5.03E-66	0.0337	9.64E-10	-0.3202	0.00E+00	-0.4998	0.00E+00	-0.5454	1.57E-132	0.0784	7.63E-24	-0.0424	3.52E-06	-0.1776	1.68E-30	-0.2769	6.13E-08
50		48	Mafk_3106.2	-0.0234	2.74E-03	-0.3197	1.07E-198	-0.4092	3.45E-264	-0.4377	1.65E-66	0.0312	1.55E-08	-0.3330	0.00E+00	-0.5228	0.00E+00	-0.5181	1.57E-117	0.0994	2.32E-37	-0.0183	4.57E-02	-0.1466	1.64E-21	-0.1944	6.67E-04
51		49	Max_3863.1	0.0430	3.55E-08	0.0182	6.86E-02	0.0565	2.49E-05	0.0432	2.00E-01	0.1044	1.98E-80	0.1394	6.93E-89	0.1963	5.58E-94	0.1726	7.57E-13	0.0655	4.02E-17	0.1340	8.48E-43	0.2027	2.53E-49	0.2162	4.04E-10
52		50	Max_3864.1	0.0040	6.08E-01	-0.0038	6.98E-01	-0.0285	3.74E-02	0.0037	9.17E-01	0.0272	8.24E-07	0.0809	6.71E-31	0.1570	2.19E-60	0.1842	1.89E-14	0.0031	6.91E-01	0.1073	3.83E-27	0.2371	3.31E-70	0.2526	1.21E-14
53		51	Mtf1_2377.2	-0.0400	2.89E-07	-0.2288	7.31E-104	-0.2043	7.31E-63	-0.1552	1.71E-07	0.0596	2.55E-27	-0.2282	1.13E-238	-0.2356	1.27E-135	-0.1085	7.22E-06	0.1100	2.02E-45	-0.0285	2.09E-03	0.0107	4.90E-01	0.0522	2.10E-01
54		52	Myb_1047.3	-0.1340	9.29E-67	-0.2268	3.91E-117	-0.2386	1.31E-71	-0.1756	1.10E-07	-0.1696	1.04E-210	-0.2776	0.00E+00	-0.3363	8.30E-283	-0.2903	2.15E-34	-0.1375	3.22E-70	-0.0783	1.64E-15	-0.0954	3.56E-12	-0.1676	1.92E-06
55		53	Mybl1_1717.2	-0.0285	2.54E-04	0.0274	4.51E-03	-0.0036	8.03E-01	-0.0028	9.39E-01	-0.0896	1.28E-59	0.0208	3.01E-03	-0.0867	1.96E-19	-0.1211	5.44E-07	-0.1081	6.10E-44	-0.0148	1.48E-01	-0.0979	4.98E-14	-0.1231	1.49E-04
56		54	Myf6_3824.2	0.0916	5.82E-32	-0.0012	9.14E-01	-0.0174	1.55E-01	-0.0367	1.97E-01	0.2585	0.00E+00	0.1148	1.30E-60	0.0260	7.03E-03	0.0212	3.81E-01	0.2016	2.91E-150	0.1375	3.44E-51	0.0793	3.87E-07	0.0951	4.11E-02
57		55	Nkx3-1_2923.2	0.0097	2.16E-01	-0.2031	9.05E-76	-0.2215	2.52E-77	-0.1825	3.96E-12	0.1121	1.69E-92	-0.2900	0.00E+00	-0.3061	3.04E-232	-0.2335	1.63E-22	0.0957	9.38E-35	-0.1273	1.44E-45	-0.1119	2.91E-12	-0.0112	8.52E-01
58		56	Nr2f2_2192.2	0.1024	1.33E-39	0.0259	1.40E-02	0.0478	1.49E-04	0.1204	6.42E-05	0.2064	0.00E+00	0.0936	8.51E-41	0.1677	1.05E-68	0.2174	1.18E-19	0.1242	1.68E-57	0.0843	2.62E-19	0.1616	1.57E-27	0.1752	1.44E-05
59		57	Osr1_3033.2	0.0487	4.17E-10	-0.1264	4.14E-33	-0.0982	4.79E-15	-0.0588	4.70E-02	0.1137	4.37E-95	-0.0967	1.75E-43	-0.0752	5.51E-15	0.0352	1.46E-01	0.0838	4.93E-27	0.0076	4.14E-01	0.0773	2.62E-07	0.0897	3.38E-02
60		58	Osr2_1727.2	0.0705	1.52E-19	-0.0340	1.29E-03	-0.0777	6.12E-10	0.0161	5.95E-01	0.1447	2.65E-153	-0.0077	2.71E-01	-0.0526	4.88E-08	-0.0074	7.61E-01	0.1025	1.11E-39	0.0291	1.93E-03	0.0604	5.80E-05	0.1117	5.89E-03
61		59	Plagl1_0972.2	0.0234	2.70E-03	-0.0634	4.61E-10	-0.0455	5.42E-04	0.0676	3.50E-02	0.1625	2.17E-193	0.1508	7.66E-104	0.1976	3.28E-95	0.2066	7.31E-18	0.1944	1.03E-139	0.2424	8.43E-143	0.3544	3.01E-147	0.2718	8.94E-14
62		60	Rara_1051.2	0.0590	3.66E-14	-0.1196	1.37E-29	-0.0981	7.16E-15	-0.1120	7.38E-05	0.1440	5.27E-152	-0.0465	3.31E-11	-0.0491	3.39E-07	-0.0685	4.71E-03	0.0816	1.02E-25	0.0141	1.31E-01	0.0436	3.43E-03	-0.0335	4.76E-01
63		61	Rfx3_3961.1	0.0231	3.05E-03	-0.1647	1.08E-55	-0.1471	1.29E-31	-0.1267	1.91E-05	0.0305	3.06E-08	-0.2060	6.56E-194	-0.2367	7.00E-137	-0.2504	9.62E-26	-0.0088	2.58E-01	-0.0888	2.55E-21	-0.1040	2.74E-12	-0.1226	3.41E-03
64		62	Rfx3_4970.2	-0.0581	8.64E-14	-0.2342	2.39E-110	-0.2204	8.57E-71	-0.2467	7.36E-18	-0.0067	2.23E-01	-0.3079	0.00E+00	-0.3317	1.23E-274	-0.3617	9.24E-54	0.0278	3.73E-04	-0.1238	1.60E-40	-0.1425	2.15E-21	-0.1392	1.48E-03
65		63	Rfx4_3761.1	-0.0804	5.71E-25	0.0061	5.19E-01	0.0643	1.03E-05	-0.0176	6.57E-01	-0.0794	3.60E-47	0.0439	3.78E-10	0.1035	5.15E-27	-0.0183	4.51E-01	-0.0173	2.66E-02	0.0168	1.06E-01	0.0257	4.57E-02	-0.0602	5.00E-02
66		64	Rfxdc2_3516.1	-0.1381	8.55E-71	-0.1068	3.34E-29	-0.0960	1.12E-11	-0.1355	3.02E-03	-0.1581	5.08E-183	-0.0912	7.08E-39	-0.0679	1.82E-12	-0.1372	1.32E-08	-0.0680	2.54E-18	-0.0001	9.94E-01	-0.0265	4.42E-02	-0.0794	5.47E-03
67		65	Rxra_1035.2	0.0435	2.47E-08	-0.0698	3.68E-11	-0.0818	1.22E-10	-0.1321	3.65E-06	0.1518	8.06E-169	0.0070	3.21E-01	0.0006	9.48E-01	-0.0623	1.01E-02	0.1026	1.04E-39	0.0560	2.32E-09	0.0879	2.48E-09	0.0908	4.64E-02
68		66	Sfpi1_1034.2	-0.0422	6.00E-08	-0.3009	6.56E-181	-0.3788	1.48E-217	-0.4193	2.00E-56	0.0835	6.15E-52	-0.1398	2.49E-89	-0.3564	0.00E+00	-0.4680	2.23E-93	0.1788	3.92E-118	0.1629	3.48E-70	0.0908	1.94E-09	0.0073	8.84E-01
69		67	Sfpi1_1034.3	0.0425	4.88E-08	-0.1559	1.80E-48	-0.2089	7.32E-64	-0.2622	1.19E-21	0.1358	2.94E-135	-0.0534	2.53E-14	-0.2264	4.22E-125	-0.2865	1.65E-33	0.1520	1.67E-85	0.1304	2.41E-45	0.0025	8.67E-01	-0.0233	6.36E-01
70		68	Six6_2267.4	0.0718	3.24E-20	-0.1846	2.33E-64	-0.2068	7.74E-66	-0.2675	9.60E-24	0.0362	4.91E-11	-0.3565	0.00E+00	-0.4147	0.00E+00	-0.4247	1.71E-75	-0.1237	4.68E-57	-0.2810	1.05E-217	-0.3180	1.53E-95	-0.3368	1.95E-10
71		69	Smad3_3805.1	-0.0365	2.93E-06	0.0297	1.38E-03	0.0628	2.55E-05	0.0091	8.65E-01	-0.1236	3.39E-112	0.1402	8.40E-90	0.1135	3.52E-32	0.0633	8.98E-03	-0.1303	3.85E-63	0.1188	8.06E-29	0.0756	2.08E-09	0.0478	7.85E-02
72		70	Sox1_2631.2	0.0639	2.35E-16	-0.1114	3.62E-24	-0.0767	2.51E-10	-0.1020	1.83E-04	0.1467	1.26E-157	-0.1972	1.65E-177	-0.1845	4.98E-83	-0.1233	3.32E-07	0.0950	2.73E-34	-0.1150	7.14E-37	-0.1310	1.22E-16	-0.0850	1.07E-01
73		71	Sox11_2266.2	0.0550	1.68E-12	-0.1674	1.92E-51	-0.1896	2.34E-57	-0.1577	3.15E-09	0.1335	9.66E-131	-0.2685	0.00E+00	-0.3287	1.41E-269	-0.2456	8.31E-25	0.1027	8.44E-40	-0.1295	3.01E-47	-0.1610	2.01E-23	-0.1093	5.65E-02
74		72	Sox12_3957.1	0.0553	1.33E-12	-0.1542	3.39E-44	-0.1660	5.83E-44	-0.1913	1.20E-12	0.1386	7.37E-141	-0.2343	7.14E-252	-0.2807	5.02E-194	-0.2320	3.15E-22	0.0963	3.84E-35	-0.1077	6.47E-33	-0.1690	6.09E-26	-0.2389	7.46E-06
75		73	Sox13_1718.2	0.0371	1.94E-06	-0.3203	3.07E-179	-0.4407	0.00E+00	-0.5463	6.39E-116	0.1141	8.48E-96	-0.3953	0.00E+00	-0.5526	0.00E+00	-0.6141	5.17E-177	0.0473	1.34E-09	-0.1765	3.53E-90	-0.2071	2.65E-34	-0.1795	7.35E-03
76		74	Sox14_2677.2	0.0604	8.80E-15	-0.0455	2.70E-05	-0.0279	2.28E-02	-0.0447	1.22E-01	0.1442	2.83E-152	-0.1488	4.88E-101	-0.1121	1.89E-31	-0.0563	2.02E-02	0.0915	6.39E-32	-0.1456	4.32E-57	-0.1115	8.69E-13	-0.0203	6.48E-01
77		75	Sox15_3457.1	0.0230	3.21E-03	-0.2471	2.72E-105	-0.3467	1.52E-205	-0.2667	4.11E-26	0.1118	5.89E-92	-0.4037	0.00E+00	-0.4995	0.00E+00	-0.3345	8.98E-46	0.0667	1.10E-17	-0.1824	3.34E-96	-0.2605	1.32E-54	-0.1389	5.94E-02
78		76	Sox17_2837.2	0.0723	1.73E-20	-0.1546	1.15E-43	-0.1488	6.96E-36	-0.1011	1.20E-04	0.1619	4.89E-192	-0.2396	1.45E-263	-0.2697	1.01E-178	-0.1485	7.46E-10	0.0850	9.85E-28	-0.1556	5.05E-68	-0.1692	1.23E-25	-0.0277	6.57E-01
79		77	Sox18_3506.1	0.0064	4.08E-01	-0.2157	7.55E-83	-0.2721	3.38E-121	-0.1830	3.13E-12	0.0822	1.86E-50	-0.3944	0.00E+00	-0.4494	0.00E+00	-0.2803	4.23E-32	0.0379	1.16E-06	-0.2134	4.81E-129	-0.2895	5.12E-71	-0.2070	6.06E-04
80		78	Sox21_3417.1	0.0718	3.12E-20	-0.2098	2.53E-77	-0.2292	4.14E-86	-0.1628	2.58E-10	0.1221	2.07E-109	-0.3684	0.00E+00	-0.4201	0.00E+00	-0.2650	9.60E-29	0.0541	3.90E-12	-0.1967	2.04E-110	-0.2590	4.53E-56	-0.1682	1.47E-02
81		79	Sox30_2781.2	0.1031	4.33E-40	-0.1460	6.62E-38	-0.1504	9.91E-38	-0.1394	8.53E-08	0.2128	0.00E+00	-0.1874	3.60E-160	-0.2277	1.34E-126	-0.1257	1.98E-07	0.0987	7.17E-37	-0.1016	2.23E-30	-0.1235	1.54E-13	-0.0859	1.88E-01
82		80	Sox4_2941.2	0.0975	4.79E-36	-0.1733	2.19E-52	-0.1768	4.77E-52	-0.1142	1.03E-05	0.2060	0.00E+00	-0.2168	5.59E-215	-0.2600	9.81E-166	-0.1571	7.10E-11	0.1227	3.38E-56	-0.0719	4.54E-16	-0.0915	6.07E-08	0.0040	9.53E-01
83		81	Sox5_3459.1	0.0929	7.97E-33	-0.2229	1.96E-85	-0.2629	2.29E-116	-0.2796	4.26E-28	0.1716	7.05E-216	-0.3072	0.00E+00	-0.3922	0.00E+00	-0.3262	1.73E-43	0.0695	4.34E-19	-0.1550	9.70E-70	-0.1853	1.10E-27	-0.2178	1.27E-03

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
84		82	Sox7_3460.1	0.0514	4.18E-11	-0.2069	3.27E-76	-0.2522	9.12E-104	-0.2268	3.53E-18	0.1113	3.05E-91	-0.3444	0.00E+00	-0.4319	0.00E+00	-0.3488	7.26E-50	0.0341	1.24E-05	-0.2018	2.45E-115	-0.2775	2.67E-65	-0.2040	7.82E-04
85		83	Sox7_4972.2	-0.0062	4.27E-01	-0.2783	5.40E-139	-0.3267	9.98E-175	-0.3247	2.47E-38	0.0762	1.40E-43	-0.4078	0.00E+00	-0.4936	0.00E+00	-0.3907	3.46E-63	0.0594	2.55E-14	-0.1914	1.24E-103	-0.2784	3.23E-67	-0.2223	1.79E-03
86		84	Sox8_1733.2	0.0366	2.61E-06	-0.0455	2.11E-05	-0.0455	2.50E-04	-0.0322	2.70E-01	0.1157	2.31E-98	-0.1507	1.11E-103	-0.1123	1.49E-31	-0.0863	3.66E-04	0.0724	1.47E-20	-0.1640	9.00E-71	-0.1113	2.88E-13	-0.0113	7.96E-01
87		85	Sp100_2947.2	-0.2820	2.87E-298	-0.2197	1.48E-170	-0.1276	3.81E-04	0.0000	0.00E+00	-0.7463	0.00E+00	-0.2961	0.00E+00	-0.4970	0.00E+00	-0.5283	5.64E-123	-0.6682	0.00E+00	-0.1195	2.61E-16	-0.4624	0.00E+00	-0.5283	5.64E-123
88		86	Sp4_1011.2	0.0270	5.71E-04	-0.1318	3.35E-35	-0.1052	3.20E-17	-0.0238	4.24E-01	0.1543	2.04E-174	-0.0227	1.23E-03	0.0153	1.12E-01	0.0969	6.19E-05	0.1451	7.25E-78	0.1388	1.56E-50	0.2172	4.67E-47	0.2615	3.20E-10
89		87	Spdef_0905.2	-0.0723	1.73E-20	-0.2222	1.73E-112	-0.1866	2.45E-43	-0.1866	5.35E-09	-0.0749	4.31E-42	-0.1389	3.31E-88	-0.1249	1.12E-38	-0.1205	6.23E-07	-0.0754	3.69E-22	0.0635	1.16E-10	0.0752	3.32E-08	0.0689	6.14E-02
90		88	Srf_3509.1	0.0078	3.18E-01	-0.3070	1.71E-174	-0.3582	4.22E-208	-0.3572	8.46E-45	0.0785	4.02E-46	-0.3434	0.00E+00	-0.4980	0.00E+00	-0.4829	3.77E-100	0.1085	2.97E-44	-0.0583	9.39E-11	-0.2141	3.42E-41	-0.1774	4.58E-03
91		89	Sry_2833.2	0.0715	4.35E-20	-0.1287	4.99E-30	-0.1696	2.17E-47	-0.1232	2.86E-06	0.1681	5.85E-207	-0.2511	3.88E-290	-0.2850	3.25E-200	-0.1693	2.05E-12	0.0755	3.07E-22	-0.1555	6.28E-69	-0.2109	2.04E-37	-0.0719	2.41E-01
92		90	Tbp_pr781.1	0.0555	1.09E-12	-0.1466	1.05E-38	-0.2069	7.85E-70	-0.1252	1.81E-06	0.0903	1.63E-60	-0.4098	0.00E+00	-0.4070	0.00E+00	-0.2258	4.15E-21	0.0133	8.90E-02	-0.2618	8.35E-196	-0.3107	7.05E-82	-0.1301	3.68E-02
93		91	Tcf1_2666.2	-0.0111	1.55E-01	-0.2129	2.04E-89	-0.2884	3.35E-122	-0.2665	4.38E-23	-0.0596	2.87E-27	-0.4876	0.00E+00	-0.4744	0.00E+00	-0.3998	2.44E-66	-0.1061	2.26E-42	-0.3683	0.00E+00	-0.3066	3.58E-96	-0.1213	1.96E-02
94		92	Tcf1_2666.3	0.0246	1.61E-03	-0.0920	2.98E-18	-0.1135	3.37E-19	-0.1113	8.29E-05	0.0377	7.57E-12	-0.2383	1.22E-260	-0.2309	3.38E-130	-0.1467	1.20E-09	-0.0193	1.32E-02	-0.2243	5.59E-130	-0.2225	2.25E-52	-0.0903	5.40E-02
95		93	Tcf3_3787.1	0.0278	3.59E-04	-0.3021	1.21E-164	-0.3624	2.44E-219	-0.4103	4.59E-60	0.1043	2.77E-80	-0.3528	0.00E+00	-0.4816	0.00E+00	-0.4956	3.31E-106	0.0939	1.54E-33	-0.1096	9.19E-35	-0.1720	1.17E-25	-0.1043	1.00E-01
96		94	Tcf7_0950.2	0.0018	8.13E-01	-0.2796	2.64E-137	-0.3321	1.44E-185	-0.2930	6.64E-31	0.1146	1.49E-96	-0.3658	0.00E+00	-0.4552	0.00E+00	-0.3889	1.44E-62	0.1034	2.72E-40	-0.1320	2.07E-50	-0.1633	1.51E-22	-0.1492	3.07E-02
97		95	Tcf7l2_3461.1	-0.0051	5.14E-01	-0.1402	8.37E-38	-0.1412	3.48E-32	-0.0312	2.91E-01	0.1283	1.12E-120	-0.1997	3.54E-182	-0.1865	7.82E-85	-0.0204	3.99E-01	0.1535	2.68E-87	-0.0538	3.60E-09	-0.0903	2.15E-08	0.0139	7.45E-01
98		96	Tcfap2a_2337.3	0.0185	1.80E-02	-0.0971	6.22E-21	-0.0763	3.26E-09	-0.0669	2.85E-02	0.0658	6.82E-33	-0.0965	2.44E-43	-0.0557	7.34E-09	-0.0050	8.38E-01	0.0658	2.88E-17	0.0075	4.31E-01	0.0870	2.00E-09	0.1742	1.11E-05
99		97	Tcfap2b_3988.1	0.0604	8.60E-15	-0.0024	8.19E-01	-0.0243	5.91E-02	-0.0127	6.81E-01	0.1332	3.89E-130	0.0469	2.27E-11	0.0241	1.23E-02	0.0482	4.69E-02	0.1040	8.92E-41	0.0837	1.24E-18	0.0856	3.60E-09	0.0817	3.77E-02
100		98	Tcfap2c_2912.2	0.0518	3.78E-11	-0.0855	1.04E-15	-0.0632	5.57E-07	-0.0317	2.71E-01	0.1789	1.10E-234	0.0252	3.18E-04	0.0293	2.36E-03	0.1076	8.67E-06	0.1557	1.01E-89	0.1268	2.24E-42	0.1998	4.02E-41	0.3593	2.08E-16
101		99	Tcfap2e_3713.1	0.0471	1.50E-09	-0.0498	2.34E-06	-0.0534	2.34E-05	-0.0178	5.52E-01	0.1292	1.96E-122	-0.0375	8.48E-08	-0.0470	1.10E-06	0.0835	5.64E-04	0.0959	7.16E-35	0.0147	1.18E-01	0.0283	5.79E-02	0.1290	1.80E-03
102		100	Tcfe2a_3865.1	0.0898	8.12E-31	0.0858	1.31E-16	0.1023	1.88E-15	0.1188	7.24E-05	0.2267	0.00E+00	0.2083	2.50E-198	0.2047	3.39E-102	0.2160	2.01E-19	0.1719	2.99E-109	0.1692	7.52E-72	0.1388	7.19E-22	0.1537	1.78E-04
103		101	Zbtb12_2932.2	0.0153	4.91E-02	-0.0701	1.29E-11	-0.0700	5.92E-08	-0.0553	6.65E-02	0.0940	1.80E-65	-0.0213	2.42E-03	-0.0831	6.08E-18	-0.1109	4.55E-06	0.1074	2.09E-43	0.0666	2.53E-12	-0.0111	4.45E-01	-0.0660	1.06E-01
104		102	Zbtb3_1048.2	0.1121	3.79E-47	0.0145	1.79E-01	0.0418	6.55E-04	0.1116	1.74E-04	0.2619	0.00E+00	0.1198	6.13E-66	0.1581	3.52E-61	0.2227	1.45E-20	0.1965	9.63E-143	0.1330	2.29E-47	0.1876	6.97E-34	0.1331	1.38E-03
105		103	Zbtb7b_1054.2	0.0960	5.27E-35	-0.3177	1.14E-176	-0.3616	7.04E-226	-0.4190	2.47E-63	0.1661	4.44E-202	-0.3042	0.00E+00	-0.3759	0.00E+00	-0.4320	2.42E-78	0.0902	4.50E-31	-0.0361	4.55E-05	0.0570	8.62E-04	-0.0167	7.98E-01
106		104	Zfp105_2634.2	0.0414	1.10E-07	-0.1862	2.80E-62	-0.1909	1.92E-59	-0.1072	6.14E-05	0.1460	3.86E-156	-0.2704	0.00E+00	-0.2948	8.91E-215	-0.1304	6.71E-08	0.1089	1.51E-44	-0.1052	4.97E-32	-0.1163	1.99E-12	-0.0332	5.60E-01
107		105	Zfp128_2806.2	-0.0096	2.18E-01	-0.3471	6.29E-235	-0.4691	0.00E+00	-0.5395	3.41E-100	-0.0143	9.40E-03	-0.4319	0.00E+00	-0.5757	0.00E+00	-0.6209	5.12E-182	-0.0577	1.32E-13	-0.2060	4.02E-115	-0.2485	5.34E-58	-0.1903	1.69E-04
108		106	Zfp161_2858.2	-0.1586	4.57E-93	-0.0598	5.35E-12	-0.0114	5.29E-01	0.0186	8.51E-01	-0.3961	0.00E+00	-0.0101	1.51E-01	0.1635	2.17E-65	0.2335	1.66E-22	-0.3382	0.00E+00	0.0630	1.19E-07	0.2012	2.80E-71	0.2279	2.91E-20
109		107	Zfp187_2626.2	0.0453	6.44E-09	-0.3078	1.06E-168	-0.3463	2.69E-202	-0.3867	5.99E-53	0.0968	2.59E-69	-0.4159	0.00E+00	-0.4448	0.00E+00	-0.4379	1.17E-80	0.0173	2.65E-02	-0.2095	2.87E-125	-0.1538	3.27E-20	-0.1725	6.16E-03
110		108	Zfp281_0973.2	0.0407	1.80E-07	-0.1025	2.58E-22	-0.1234	1.28E-22	-0.1257	1.41E-05	0.1819	8.94E-243	0.0399	1.29E-08	0.0404	2.71E-05	0.0080	7.40E-01	0.2018	1.40E-150	0.1835	7.87E-87	0.2495	5.20E-65	0.2781	1.40E-10
111		109	Zfp410_3034.2	0.0844	2.14E-27	-0.2093	3.91E-76	-0.2525	1.67E-106	-0.2377	2.26E-20	0.1621	2.21E-192	-0.3034	0.00E+00	-0.3578	0.00E+00	-0.3074	1.45E-38	0.0808	3.17E-25	-0.1228	5.80E-44	-0.1226	4.48E-13	-0.2036	2.01E-03
112		110	Zfp691_0895.2	0.0478	8.48E-10	0.0585	3.43E-09	0.0772	1.30E-08	0.0711	4.44E-02	0.0955	1.64E-67	0.1694	7.13E-131	0.1948	1.61E-92	0.0709	3.45E-03	0.0974	6.17E-36	0.1587	3.65E-58	0.1675	5.79E-35	0.0524	1.16E-01
113		111	Zfp740_0925.3	0.1169	3.65E-51	-0.1963	2.54E-65	-0.2280	3.60E-88	-0.2407	1.76E-21	0.2637	0.00E+00	-0.1299	2.89E-77	-0.2393	5.75E-140	-0.2518	5.06E-26	0.1929	1.10E-137	0.0532	1.44E-09	0.0033	8.47E-01	-0.1195	1.09E-01
114		112	Zic1_0991.2	0.0271	5.13E-04	0.1063	2.43E-29	0.1924	1.60E-41	0.1721	1.71E-04	0.0832	1.42E-51	0.3433	0.00E+00	0.4531	0.00E+00	0.4829	3.99E-100	0.1059	3.02E-42	0.3025	3.97E-194	0.3748	6.72E-197	0.3933	9.80E-47
115		113	Zic2_2895.2	0.0363	3.13E-06	0.0982	5.72E-25	0.2131	3.86E-52	0.1532	7.10E-04	0.0987	4.58E-72	0.3443	0.00E+00	0.4808	0.00E+00	0.5595	8.71E-141	0.0965	2.47E-35	0.3199	1.42E-221	0.3900	1.49E-209	0.4771	3.34E-70
116		114	Zic3_3119.2	-0.0313	5.94E-05	0.0641	1.37E-11	0.2238	3.70E-56	0.1923	2.35E-05	0.0753	1.47E-42	0.3348	0.00E+00	0.4890	0.00E+00	0.5689	1.49E-146	0.1007	2.61E-38	0.3155	1.62E-212	0.3770	6.95E-199	0.4869	6.64E-74
117		115	Zscan4_2667.2	0.0023	7.71E-01	-0.2504	3.48E-115	-0.2985	3.13E-143	-0.3467	4.96E-40	0.1036	3.64E-79	-0.2866	0.00E+00	-0.3598	0.00E+00	-0.3338	1.43E-45	0.0794	2.17E-24	-0.0908	5.91E-24	-0.0694	1.87E-05	-0.0947	8.59E-02
118																											
119																											
120	**WORM**	1	CND-1	-0.0608	5.25E-15	-0.0564	8.72E-08	-0.0181	1.54E-01	-0.0435	1.28E-01	-0.0387	2.20E-12	-0.0036	6.06E-01	-0.0025	7.99E-01	-0.0346	1.53E-01	0.0174	2.65E-02	0.0416	9.78E-06	0.0330	2.69E-02	-0.0278	5.46E-01
121		2	HLH-1	-0.0706	3.24E-20	-0.0753	1.49E-12	-0.0433	3.21E-04	0.0453	1.08E-01	-0.0547	3.37E-23	-0.0057	4.15E-01	-0.0246	1.06E-02	-0.0195	4.20E-01	0.0342	1.76E-05	0.0671	6.67E-13	0.0414	1.05E-02	-0.0207	6.65E-01
122		3	HLH-2	-0.3479	0.00E+00	-0.2843	2.77E-141	-0.3431	3.94E-218	-0.3594	5.23E-51	-0.3346	0.00E+00	-0.2458	1.03E-277	-0.3707	0.00E+00	-0.3768	1.58E-58	-0.0485	1.43E-09	-0.0552	4.72E-10	-0.0465	1.39E-02	0.0344	7.84E-01
123		4	HLH-3	-0.0549	1.46E-12	-0.0496	2.37E-06	-0.0192	1.29E-01	-0.0469	1.01E-01	-0.0384	3.10E-12	0.0111	1.13E-01	-0.0254	8.29E-03	-0.0529	2.92E-02	0.0246	1.76E-03	0.0531	1.87E-08	0.0173	2.47E-01	0.0389	3.97E-01
124		5	HLH-4	-0.0177	2.40E-02	-0.0198	5.70E-02	0.0078	5.50E-01	0.0194	5.22E-01	0.0060	2.79E-01	0.0250	3.54E-04	0.0405	2.64E-05	-0.0078	7.48E-01	0.0303	1.03E-04	0.0466	1.03E-06	0.0417	3.82E-03	0.0385	3.45E-01
125		6	HLH-8	-0.1575	6.51E-96	-0.1344	2.18E-35	-0.1324	1.46E-29	-0.1110	2.35E-05	-0.1499	1.76E-164	-0.0663	2.68E-21	-0.1520	1.30E-56	-0.1222	4.24E-07	0.0470	4.35E-09	0.0839	6.20E-20	0.0193	2.54E-01	-0.0921	1.41E-01
126		7	HLH-10	-0.0377	1.23E-06	-0.0487	4.35E-06	-0.0061	6.26E-01	0.0405	1.71E-01	-0.0212	1.19E-04	0.0090	1.98E-01	0.0091	3.44E-01	0.0523	3.11E-02	0.0255	1.16E-03	0.0537	1.04E-08	0.0217	1.53E-01	0.0512	2.31E-01
127		8	HLH-11	-0.2173	5.53E-180	-0.1781	5.05E-61	-0.1830	1.06E-53	-0.2606	1.36E-23	-0.2326	0.00E+00	-0.1557	1.42E-110	-0.2338	1.67E-133	-0.3072	1.58E-38	-0.0440	2.98E-08	-0.0246	7.38E-03	-0.0685	2.84E-05	-0.0872	1.54E-01
128		9	HLH-14	-0.2749	5.93E-291	-0.2122	1.03E-84	-0.2583	5.19E-108	-0.3092	8.40E-35	-0.2493	0.00E+00	-0.1502	5.42E-103	-0.2813	7.09E-195	-0.3400	2.46E-47	0.0095	2.32E-01	0.0274	2.70E-03	-0.0112	4.97E-01	-0.1334	6.79E-02
129		10	HLH-15	-0.3920	0.00E+00	-0.3205	4.21E-192	-0.3943	7.91E-266	-0.4641	5.38E-83	-0.3426	0.00E+00	-0.2432	8.82E-272	-0.3950	0.00E+00	-0.4756	8.38E-97	-0.0309	9.28E-05	-0.0184	4.18E-02	-0.0685	4.23E-05	-0.0591	4.55E-01
130		11	HLH-19	-0.3754	0.00E+00	-0.3172	1.17E-185	-0.3700	5.05E-238	-0.4156	8.98E-66	-0.3362	0.00E+00	-0.2409	1.74E-266	-0.3753	0.00E+00	-0.4508	5.88E-86	-0.0125	1.15E-01	-0.0053	5.52E-01	-0.0210	2.24E-01	-0.2482	2.12E-03
131		12	HLH-25	-0.2985	0.00E+00	-0.2595	6.38E-145	-0.2932	8.60E-138	-0.2824	1.43E-24	-0.2446	0.00E+00	-0.1781	1.26E-144	-0.2853	1.17E-200	-0.2746	7.87E-31	0.0532	5.87E-11	0.0399	3.21E-05	0.0540	8.78E-04	0.1886	7.15E-05
132		13	HLH-26	-0.3316	0.00E+00	-0.2676	4.86E-134	-0.3324	3.12E-182	-0.3627	2.49E-47	-0.2498	0.00E+00	-0.1461	1.77E-97	-0.2980	1.18E-219	-0.3418	7.97E-48	0.0606	1.74E-14	0.0831	5.15E-20	0.0152	3.57E-01	0.2258	8.09E-04
133		14	HLH-27	-0.3940	0.00E+00	-0.3435	2.48E-225	-0.3862	6.22E-252	-0.4551	3.43E-75	-0.3197	0.00E+00	-0.2240	8.33E-230	-0.3730	0.00E+00	-0.4277	1.21E-76	0.0217	6.02E-03	0.0326	3.36E-04	-0.0179	2.82E-01	0.1059	9.83E-02
134		15	HLH-29	-0.1487	9.08E-85	-0.1155	4.60E-28	-0.1379	6.31E-30	-0.1532	5.64E-08	-0.0942	9.46E-66	-0.0274	9.06E-05	-0.1005	1.41E-25	-0.1559	9.94E-11	0.0672	3.18E-17	0.0999	1.80E-26	0.0396	1.24E-02	0.0543	2.50E-01
135		16	HLH-30	-0.3862	0.00E+00	-0.3283	2.59E-206	-0.3881	6.88E-266	-0.4308	2.23E-70	-0.3846	0.00E+00	-0.3112	0.00E+00	-0.4207	0.00E+00	-0.4531	6.40E-87	-0.1024	1.83E-37	-0.1225	1.64E-41	-0.0737	2.29E-05	-0.0112	8.86E-01
136		17	LIN-32	-0.1059	3.36E-44	-0.0985	1.09E-19	-0.0649	3.35E-08	0.0201	4.40E-01	-0.0882	7.57E-58	-0.0184	8.68E-03	-0.0510	1.21E-07	0.0161	5.07E-01	0.0501	4.17E-10	0.0899	1.14E-22	0.0393	2.00E-02	0.0703	3.01E-01
137		18	MDL-1	-0.0168	2.76E-02	-0.0156	1.20E-01	0.0066	6.05E-01	0.0146	6.34E-01	-0.0103	6.18E-02	0.0231	1.00E-03	-0.0134	1.66E-01	0.0004	9.87E-01	0.0002	9.80E-01	0.0259	8.47E-03	-0.0095	5.23E-01	-0.0178	6.57E-01
138		19	MXL-1	-0.0330	1.94E-05	-0.0228	2.65E-02	-0.0121	3.42E-01	-0.0237	4.37E-01	-0.0208	1.56E-04	0.0190	6.74E-03	-0.0221	2.21E-02	-0.0051	8.33E-01	0.0019	8.11E-01	0.0292	2.35E-03	-0.0150	3.08E-01	-0.0198	6.22E-01
139		20	MXL-3	0.0072	3.42E-01	-0.0145	1.60E-01	0.0450	1.99E-04	0.1175	3.42E-05	0.0547	3.42E-23	0.0654	9.38E-21	0.1338	3.60E-44	0.1702	1.56E-12	0.0712	1.37E-18	0.1111	4.04E-31	0.0998	4.29E-10	0.0644	1.69E-01
140		21	REF-1	-0.3438	0.00E+00	-0.2903	2.59E-167	-0.3468	1.07E-208	-0.3582	1.75E-46	-0.2728	0.00E+00	-0.1890	5.50E-163	-0.3015	6.01E-225	-0.3237	8.45E-43	0.0550	1.22E-11	0.0627	1.23E-11	0.0734	2.33E-05	0.0568	4.21E-01
141																											
142																											
143	**FLY**	1	Abd-B	-0.2567	1.06E-245	-0.2249	8.02E-96	-0.2937	8.41E-133	-0.3790	2.18E-48	-0.2170	0.00E+00	-0.1453	2.33E-96	-0.3571	0.00E+00	-0.4530	7.21E-87	0.0445	1.15E-08	0.0767	3.70E-17	-0.1247	1.24E-15	-0.2295	3.31E-05
144		2	Bap	-0.0662	1.95E-17	-0.0671	2.80E-09	-0.0758	1.57E-10	-0.0196	4.61E-01	-0.0405	2.04E-13	-0.0150	3.29E-02	-0.1141	1.61E-32	-0.0665	6.05E-03	0.0490	3.20E-10	0.0556	4.98E-10	-0.0608	2.44E-04	-0.0258	6.62E-01
145		3	CHES-1-like	-0.1528	2.05E-86	-0.1789	1.93E-71	-0.1274	7.76E-23	-0.0275	4.73E-01	-0.0484	1.57E-18	-0.0219	1.76E-03	-0.0948	6.37E-23	-0.1476	9.51E-10	0.1079	1.80E-43	0.1693	5.46E-68	0.0208	1.49E-01	-0.1261	5.83E-05
146		4	CLAMP_Cterm4Zn	0.0483	5.72E-10	0.0774	8.55E-16	0.0066	6.41E-01	-0.0803	3.92E-02	0.1065	1.41E-83	0.1385	1.11E-87	0.0771	1.16E-15	-0.0133	5.83E-01	0.0576	1.43E-13	0.0663	8.75E-11	0.0566	1.71E-05	0.0290	3.49E-01
147		5	CLAMP_Cterm6Zn	-0.0146	6.14E-02	0.0059	5.51E-01	-0.0334	1.47E-02	-0.1371	1.07E-04	0.0302	4.42E-08	0.0557	1.75E-15	0.0082	3.97E-01	-0.1113	4.17E-06	0.0544	2.87E-12	0.0690	4.33E-12	0.0441	1.14E-03	-0.0267	4.22E-01
148		6	Eve	-0.2490	1.14E-230	-0.2264	5.09E-95	-0.2735	2.53E-117	-0.3409	1.40E-39	-0.2140	0.00E+00	-0.1593	9.61E-116	-0.3349	2.96E-280	-0.4051	3.10E-68	0.0204	8.88E-03	0.0325	3.29E-04	-0.1056	3.03E-11	-0.1437	1.33E-02
149		7	Jumeau	-0.0650	7.29E-17	-0.1013	1.56E-23	-0.0249	5.41E-02	0.0367	3.21E-01	-0.0044	4.23E-01	-0.0021	7.68E-01	-0.0073	4.50E-01	-0.0819	7.16E-04	0.0784	7.63E-24	0.1170	9.52E-34	0.0519	3.42E-04	-0.1823	1.13E-08
150		8	Lbl	0.0015	8.46E-01	-0.0117	2.88E-01	0.0123	3.13E-01	0.0262	3.41E-01	-0.0274	6.93E-07	-0.0103	1.43E-01	-0.0751	6.13E-15	-0.0281	2.47E-01	-0.0161	3.93E-02	0.0069	4.51E-01	-0.1308	1.10E-16	-0.1290	1.16E-02
151		9	Msh	-0.0298	1.33E-04	-0.0618	1.04E-08	0.0041	7.37E-01	-0.0020	9.43E-01	-0.0483	1.83E-18	-0.0279	6.95E-05	-0.0937	2.11E-22	-0.0778	1.31E-03	0.0245	1.70E-03	0.0561	1.08E-09	-0.1008	8.80E-11	-0.0928	3.94E-02
152		10	Ptx1	-0.2297	9.56E-196	-0.2128	3.07E-84	-0.2534	6.28E-100	-0.3545	2.98E-43	-0.2408	0.00E+00	-0.2175	2.02E-216	-0.3175	1.14E-250	-0.4214	2.99E-74	-0.0308	7.68E-05	-0.0527	6.10E-09	-0.0428	7.04E-03	-0.0823	1.68E-01
153		11	Six4	-0.2084	9.50E-161	-0.1998	7.24E-79	-0.2256	2.14E-75	-0.2696	6.01E-22	-0.2656	0.00E+00	-0.2527	4.72E-294	-0.3159	5.28E-248	-0.3340	1.26E-45	-0.1194	2.53E-53	-0.1121	6.00E-34	-0.2023	6.97E-41	-0.0774	9.36E-02
154		12	Slou	-0.1300	7.53E-63	-0.1266	1.51E-30	-0.1336	1.75E-28	-0.1939	3.41E-13	-0.1163	1.98E-99	-0.0793	9.19E-30	-0.1962	7.51E-94	-0.2258	4.11E-21	0.0231	3.09E-03	0.0554	9.50E-10	-0.1283	5.81E-16	-0.0606	2.83E-01
155		13	Tin	-0.2356	3.14E-206	-0.2303	2.66E-91	-0.2636	7.53E-116	-0.3467	4.85E-45	-0.1713	3.59E-215	-0.1333	2.64E-81	-0.2912	2.70E-209	-0.3675	1.47E-55	0.0738	2.59E-21	0.0623	1.67E-12	-0.0245	1.49E-01	-0.0439	5.96E-01
156		14	Ubx	-0.2712	5.72E-275	-0.2485	8.82E-116	-0.2947	1.89E-135	-0.4039	5.79E-56	-0.2163	0.00E+00	-0.1444	3.22E-95	-0.3544	0.00E+00	-0.4730	1.35E-95	0.0651	6.88E-17	0.0907	1.38E-23	-0.0829	1.52E-07	-0.2122	1.94E-04
157																											
158																											
159	**HUMAN**	1	FOXN2	-0.1215	3.72E-55	-0.2565	9.51E-143	-0.2876	1.54E-116	-0.1034	3.11E-03	-0.1257	4.94E-116	-0.3083	0.00E+00	-0.3224	1.00E-258	-0.1156	1.75E-06	-0.0349	7.64E-06	-0.0589	8.33E-10	0.0080	5.86E-01	-0.0393	2.43E-01
160		2	FOXN4	-0.1247	5.66E-58	-0.1379	5.13E-48	-0.0959	9.67E-12	0.1422	5.31E-03	-0.2133	0.00E+00	-0.1694	6.28E-131	-0.1184	6.75E-35	0.0950	8.74E-05	-0.1684	8.71E-105	-0.0423	4.42E-05	-0.0253	5.51E-02	0.0225	4.15E-01
161		3	FOXR1	-0.1782	2.40E-117	-0.2131	2.42E-111	-0.2228	2.37E-59	0.0149	7.59E-01	-0.2172	0.00E+00	-0.2168	4.29E-215	-0.2091	1.21E-106	0.0295	2.24E-01	-0.0852	6.95E-28	-0.0095	3.56E-01	-0.0223	9.59E-02	-0.0265	3.44E-01
162		4	CSL	0.0013	8.66E-01	-0.0015	8.81E-01	0.0062	6.42E-01	-0.0059	8.56E-01	0.1348	3.05E-133	0.1630	3.58E-121	0.1300	9.45E-42	0.0841	5.18E-04	0.1654	3.76E-101	0.2077	1.79E-103	0.1509	3.52E-27	0.0410	2.64E-01
163		5	TBP	0.0412	1.29E-07	-0.1351	4.86E-33	-0.1946	8.05E-62	-0.1186	6.27E-06	0.0957	8.49E-68	-0.3143	0.00E+00	-0.3775	0.00E+00	-0.2178	1.01E-19	0.0684	1.59E-18	-0.1533	8.03E-67	-0.2475	1.10E-51	-0.1110	7.50E-02
164		6	Sox4(Eval)	0.0583	1.11E-15	-0.1298	2.59E-37	-0.1688	3.18E-50	-0.1141	9.65E-06	0.1703	1.98E-212	-0.1249	1.60E-71	-0.2077	3.51E-105	-0.0772	1.44E-03	0.1550	6.51E-76	0.0115	2.36E-01	-0.0423	1.94E-02	0.1334	5.78E-02
165		7	Jun-fos(Eval)	-0.0918	2.06E-32	-0.1293	8.36E-39	-0.1881	1.16E-45	-0.1060	1.42E-03	-0.0864	1.54E-55	-0.1564	1.34E-111	-0.2233	9.93E-122	-0.2146	3.57E-19	-0.0230	3.50E-03	-0.0556	1.98E-08	-0.0684	9.47E-07	-0.1419	6.16E-05
166		8	Oct1(Eval)	-0.0004	9.52E-01	-0.1990	5.57E-99	-0.2888	1.29E-155	-0.1914	2.58E-14	0.0393	1.04E-12	-0.2974	0.00E+00	-0.4136	0.00E+00	-0.2631	2.42E-28	0.0416	4.98E-06	-0.1088	8.67E-26	-0.1876	2.76E-22	-0.0893	2.89E-01

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
167																											
168																											
169	**YEAST**	1	Aft1	0.0081	2.93E-01	-0.0008	9.35E-01	0.0104	4.46E-01	0.0025	9.39E-01	-0.0149	7.01E-03	0.0323	4.15E-06	-0.0954	3.53E-23	-0.0575	1.77E-02	-0.0585	1.28E-13	-0.0117	2.48E-01	-0.1177	4.55E-18	-0.0151	6.74E-01
170		2	Aro80	-0.0250	2.90E-04	-0.0162	4.42E-02	-0.0002	9.88E-01	-0.0957	2.54E-02	-0.1586	3.32E-184	-0.0354	4.45E-07	-0.1557	2.18E-59	-0.1852	1.35E-14	-0.2136	1.23E-122	0.0088	5.33E-01	-0.1615	1.81E-34	-0.1813	5.22E-10
171		3	Asg1	0.0163	1.64E-02	0.0342	1.71E-05	0.0334	1.60E-02	-0.0054	8.97E-01	-0.1263	5.02E-117	0.0140	4.53E-02	-0.1380	6.71E-47	-0.1283	1.08E-07	-0.2245	2.63E-129	-0.0450	2.37E-03	-0.1662	1.10E-35	-0.1371	3.40E-06
172		4	Bas1	-0.0151	4.92E-02	-0.0276	8.29E-03	-0.0384	1.76E-03	-0.0450	1.39E-01	0.0063	2.56E-01	-0.0018	7.94E-01	-0.1185	6.00E-35	-0.1304	6.64E-08	0.0222	5.02E-03	-0.0026	7.81E-01	-0.0756	1.11E-06	-0.0699	8.11E-02
173		5	Cbf1	0.0455	2.59E-10	0.0714	5.29E-15	0.0201	1.11E-01	-0.0216	4.98E-01	0.0978	9.32E-71	0.1147	1.41E-60	0.0996	4.00E-25	0.0863	3.65E-04	0.0693	6.14E-16	0.0760	3.31E-12	0.1174	3.99E-15	0.2135	7.25E-09
174		6	Cbf1	0.0478	1.80E-11	0.0564	4.57E-10	0.0304	1.36E-02	0.0282	3.80E-01	0.1051	2.17E-81	0.1118	1.40E-57	0.1163	1.06E-33	0.1151	1.93E-06	0.0591	1.18E-11	0.0705	1.78E-10	0.1270	1.64E-16	0.1117	2.51E-03
175		7	Cep3	0.0113	1.31E-01	0.0106	2.59E-01	0.0115	3.94E-01	-0.0764	3.46E-02	-0.0884	4.25E-58	0.0054	4.44E-01	-0.1823	4.33E-81	-0.2243	7.44E-21	-0.1582	1.51E-85	-0.0073	4.93E-01	-0.2390	4.19E-69	-0.2664	1.10E-16
176		8	Cha4	0.0077	2.70E-01	0.0196	1.66E-02	0.0165	2.40E-01	-0.0406	4.00E-01	-0.1023	3.16E-77	0.0054	4.39E-01	-0.0416	1.55E-05	-0.0623	1.02E-02	-0.1630	4.99E-75	-0.0408	2.60E-03	-0.0466	4.36E-04	-0.0345	2.20E-01
177		9	Cup9	0.0284	2.17E-04	0.0180	7.69E-02	-0.0059	6.46E-01	-0.0522	7.56E-02	0.0812	3.40E-49	0.0578	1.59E-16	0.0197	4.14E-02	0.0376	1.21E-01	0.0693	1.81E-18	0.0573	2.91E-09	0.0285	5.15E-02	0.0052	9.04E-01
178		10	Ecm22	0.0160	2.41E-02	0.0041	6.32E-01	0.0516	1.78E-04	0.1027	1.63E-02	-0.1369	2.08E-137	-0.0274	9.13E-05	-0.1608	3.08E-63	-0.1589	4.31E-11	-0.2576	6.92E-198	-0.0767	7.93E-10	-0.2548	4.02E-82	-0.1930	3.76E-11
179		11	Fhl1	0.0312	1.45E-05	0.0027	7.60E-01	0.0727	1.99E-08	0.0526	1.94E-01	0.0246	8.44E-06	0.0406	7.01E-09	0.1380	6.88E-47	-0.0118	6.27E-01	-0.0544	2.07E-10	0.0306	7.32E-03	0.0900	3.80E-10	-0.1234	4.40E-05
180		12	Fhl1	0.0612	2.70E-17	0.0578	7.58E-11	0.0516	7.45E-05	0.0936	2.81E-02	0.0383	3.63E-12	0.0905	2.82E-38	0.1137	2.63E-32	-0.0445	6.65E-02	-0.0566	2.77E-11	0.0602	1.27E-07	0.0725	4.15E-07	-0.1347	4.54E-06
181		13	Fkh1	0.0241	1.39E-03	-0.0060	5.65E-01	0.0154	2.01E-01	-0.0635	2.19E-02	0.0693	2.41E-36	0.0127	6.99E-02	0.0038	6.92E-01	-0.0701	3.80E-03	0.0583	4.90E-13	0.0088	3.52E-01	0.0189	2.40E-01	0.0005	9.92E-01
182		14	Fkh1	0.0041	5.84E-01	-0.0243	1.99E-02	-0.0072	5.45E-01	-0.0889	1.14E-03	0.0617	3.76E-29	-0.0069	3.23E-01	-0.0012	9.05E-01	-0.1122	3.52E-06	0.0697	8.27E-18	0.0059	5.30E-01	0.0516	1.92E-03	-0.0228	6.64E-01
183		15	Fkh2	0.0358	1.25E-06	0.0391	7.53E-05	0.0283	1.81E-02	-0.0321	2.91E-01	0.0597	2.43E-27	0.0211	2.63E-03	0.0367	1.40E-04	-0.0955	7.97E-05	0.0420	4.00E-07	-0.0259	9.34E-03	0.0433	7.67E-03	-0.0886	2.72E-02
184		16	Fkh2	0.0226	2.21E-03	-0.0324	1.92E-03	-0.0038	7.41E-01	-0.0312	2.42E-01	0.0698	8.23E-37	-0.0265	1.57E-04	0.0171	7.56E-02	-0.0896	2.13E-04	0.0343	3.41E-05	-0.0188	4.66E-02	0.0241	1.79E-01	-0.1518	9.12E-03
185		17	Gal4	0.0037	6.02E-01	0.0132	1.22E-01	-0.0144	2.97E-01	-0.0148	7.10E-01	-0.1288	1.35E-121	0.0097	1.68E-01	-0.2103	8.19E-108	-0.1876	6.11E-15	-0.2600	8.65E-204	-0.0161	1.90E-01	-0.2843	9.21E-103	-0.2337	1.03E-14
186		18	Gal4	-0.0223	2.08E-03	-0.0290	1.10E-03	-0.0337	1.22E-02	-0.0052	8.87E-01	-0.1260	1.95E-116	-0.0091	1.93E-01	-0.2242	1.15E-122	-0.2174	1.19E-19	-0.2225	1.75E-155	0.0084	4.60E-01	-0.2854	1.07E-98	-0.2591	4.07E-16
187		19	Gat1	-0.0173	1.11E-02	-0.0402	5.55E-06	-0.0105	3.67E-01	-0.0533	6.14E-02	-0.0441	1.31E-15	-0.0826	3.81E-32	-0.0401	3.18E-05	-0.0487	4.45E-02	-0.0358	1.33E-04	-0.0681	3.04E-09	-0.0028	8.70E-01	0.0931	4.38E-02
188		20	Gat1	-0.0013	8.54E-01	-0.0292	1.23E-03	-0.0321	5.70E-03	-0.0882	1.51E-03	-0.0226	4.28E-05	-0.0721	7.35E-25	-0.0363	1.62E-04	-0.1128	3.09E-06	-0.0406	9.28E-06	-0.0664	2.13E-09	0.0017	9.23E-01	0.1189	1.59E-02
189		21	Gat3	-0.0015	8.26E-01	-0.0225	1.61E-02	-0.0438	1.68E-04	-0.0723	7.62E-03	-0.0366	3.17E-11	-0.0899	8.49E-38	-0.0970	6.43E-24	-0.1105	4.92E-06	-0.0766	6.73E-18	-0.1123	2.08E-26	-0.1240	5.08E-13	-0.0257	6.37E-01
190		22	Gat4	0.0010	8.96E-01	0.0115	2.52E-01	-0.0352	3.61E-03	-0.1180	4.62E-05	0.0029	5.93E-01	-0.0303	1.57E-05	-0.0810	4.03E-17	-0.1489	6.75E-10	-0.0169	3.83E-02	-0.0734	5.00E-14	-0.0661	3.25E-05	-0.0589	1.82E-01
191		23	Gat4	-0.0103	1.67E-01	-0.0156	1.28E-01	-0.0550	3.34E-06	-0.1397	3.86E-07	0.0027	6.24E-01	-0.0495	1.66E-12	-0.0933	3.01E-22	-0.1787	1.11E-13	-0.0210	1.04E-02	-0.0698	3.31E-13	-0.1035	4.23E-10	-0.0019	9.69E-01
192		24	Gcn4	-0.0130	6.97E-02	-0.0133	1.54E-01	-0.0016	8.91E-01	-0.0372	2.11E-01	0.0182	9.69E-04	-0.0049	4.81E-01	0.0014	8.83E-01	0.0038	8.74E-01	0.0404	3.24E-06	0.0200	5.94E-02	0.0128	4.32E-01	0.0207	6.22E-01
193		25	Gcn4	0.0075	2.99E-01	-0.0195	4.16E-02	0.0161	1.79E-01	0.0213	4.56E-01	0.0345	3.66E-10	-0.0113	1.08E-01	0.0186	5.37E-02	0.0133	5.83E-01	0.0414	1.27E-06	0.0283	6.14E-03	0.0179	2.72E-01	-0.0135	7.70E-01
194		26	Gln3	-0.0254	1.65E-04	-0.0259	2.65E-03	-0.0257	2.84E-02	-0.0723	1.20E-02	-0.0536	2.41E-22	-0.0712	2.85E-24	-0.0431	7.68E-06	-0.1046	1.53E-05	-0.0429	7.57E-06	-0.0765	2.07E-10	-0.0255	1.33E-01	0.0125	7.83E-01
195		27	Gsm1	-0.0300	1.19E-05	-0.0104	1.96E-01	-0.0390	4.58E-03	-0.0410	3.25E-01	-0.1614	6.97E-191	-0.0339	1.31E-06	-0.1863	1.23E-84	-0.1289	9.59E-08	-0.2505	1.83E-165	-0.0570	6.13E-05	-0.2325	3.77E-68	-0.1396	2.64E-06
196		28	Gsm1	-0.0402	2.88E-08	-0.0380	1.80E-05	-0.0409	2.24E-03	-0.0476	2.13E-01	-0.1228	9.61E-111	-0.0317	6.08E-06	-0.1722	2.12E-72	-0.1529	2.30E-10	-0.1644	1.56E-84	-0.0143	2.12E-01	-0.1703	4.63E-35	-0.1598	3.09E-07
197		29	Gzf3	0.0025	7.17E-01	-0.0108	2.45E-01	-0.0395	6.46E-04	-0.0897	1.21E-03	-0.0163	3.04E-03	-0.0587	5.23E-17	-0.0683	1.30E-12	-0.0773	1.42E-03	-0.0292	1.06E-03	-0.0551	2.31E-07	-0.0217	2.12E-01	0.0733	1.41E-01
198		30	Gzf3	0.0136	5.68E-02	-0.0306	1.72E-03	-0.0415	3.07E-04	-0.0966	2.73E-04	0.0100	7.06E-02	-0.0690	6.30E-23	-0.0514	9.54E-08	-0.1053	1.33E-05	-0.0336	1.01E-04	-0.0671	2.42E-11	-0.0056	7.50E-01	0.1831	1.88E-03
199		31	Hal9	0.0103	1.30E-01	0.0314	6.77E-05	0.0342	1.65E-02	-0.0459	3.59E-01	-0.1146	1.42E-96	0.0317	6.13E-06	-0.0701	3.22E-13	-0.0749	1.98E-03	-0.1704	2.02E-75	-0.0033	8.27E-01	-0.0403	2.05E-03	-0.0684	1.36E-02
200		32	Leu3	-0.0200	3.85E-03	-0.0114	1.53E-01	0.0321	2.61E-02	-0.1087	1.86E-02	-0.1384	2.32E-140	-0.0167	1.70E-02	-0.0915	1.87E-21	-0.0420	8.35E-02	-0.1756	1.52E-83	-0.0069	6.37E-01	-0.0875	1.26E-11	-0.0185	5.16E-01
201		33	Lys14	-0.0307	3.27E-05	-0.0298	1.76E-03	-0.0379	2.79E-03	-0.0954	2.33E-03	-0.0109	4.71E-02	0.0032	6.53E-01	-0.0467	1.24E-06	-0.1494	5.90E-10	0.0078	3.45E-01	0.0324	1.76E-03	-0.0140	3.46E-01	-0.0493	1.97E-01
202		34	Matalpha2	0.0170	2.39E-02	0.0362	1.93E-04	-0.0123	3.36E-01	-0.0002	9.95E-01	0.0133	1.61E-02	0.0245	4.77E-04	-0.0430	8.04E-06	-0.0895	2.18E-04	-0.0058	4.72E-01	-0.0246	1.53E-02	-0.0438	2.96E-03	-0.1242	6.42E-04
203		35	Matalpha2	0.0203	5.95E-03	0.0412	1.96E-05	-0.0083	5.02E-01	-0.0509	1.03E-01	0.0231	2.78E-05	0.0168	1.64E-02	-0.0279	3.74E-03	-0.0739	2.30E-03	0.0208	1.23E-02	-0.0291	4.28E-03	-0.0176	2.52E-01	-0.0490	2.03E-01
204		36	Mbp1	-0.0049	4.77E-01	0.0305	1.31E-04	0.0496	6.09E-04	-0.0029	9.63E-01	-0.0534	3.31E-22	0.0566	6.72E-16	0.1355	2.74E-45	-0.0490	4.33E-02	-0.0995	4.26E-28	0.0481	1.04E-03	0.0966	6.82E-14	-0.1519	7.63E-09
205		37	Mcm1	0.0229	2.60E-03	0.0276	6.11E-03	-0.0261	4.12E-02	-0.0743	9.75E-03	0.0141	1.04E-02	0.0117	9.47E-02	-0.0890	2.19E-20	-0.1001	3.49E-05	0.0096	2.29E-01	-0.0019	8.42E-01	-0.0440	2.65E-03	-0.1011	2.48E-02
206		38	Mcm1	-0.0057	4.49E-01	-0.0457	8.26E-06	-0.0567	3.95E-06	-0.1017	1.42E-04	0.0087	1.16E-01	-0.0296	2.41E-05	-0.0950	5.48E-23	-0.1370	1.40E-08	0.0059	4.62E-01	0.0003	9.75E-01	-0.0351	2.39E-02	0.0979	8.75E-02
207		39	Met32	0.1042	1.29E-42	0.0597	4.19E-09	0.1313	6.40E-26	0.0829	6.34E-03	0.1446	3.39E-153	0.1017	6.10E-48	0.1554	4.06E-59	0.0378	1.19E-01	0.0729	5.38E-20	0.0443	4.57E-06	0.1138	4.06E-14	0.0262	5.16E-01
208		40	Met32	0.0799	1.05E-26	0.1140	9.05E-35	0.0234	7.61E-02	0.0777	6.58E-02	0.0880	1.47E-57	0.1461	2.06E-97	0.1019	3.07E-26	0.0407	9.33E-02	0.0376	4.08E-06	0.0761	9.05E-13	0.1025	3.01E-13	0.0251	3.97E-01
209		41	Mga1	-0.0265	1.49E-04	-0.0455	3.26E-07	-0.0208	8.81E-02	0.0133	6.58E-01	0.0083	1.31E-01	0.0285	4.74E-05	-0.0282	3.48E-03	-0.0724	2.81E-03	0.0379	2.30E-05	0.0859	3.41E-14	-0.0095	5.44E-01	-0.0601	1.46E-01
210		42	Mig1	0.0364	5.77E-07	0.0131	1.79E-01	-0.0320	7.67E-03	-0.0332	2.25E-01	0.0153	5.58E-03	-0.0088	2.09E-01	-0.0888	2.76E-20	-0.0874	3.05E-04	-0.0814	5.07E-22	-0.0595	3.71E-09	-0.1575	1.20E-22	-0.1629	1.77E-03
211		43	Mig2	0.0240	5.94E-04	0.0348	6.25E-05	0.0194	1.20E-01	-0.0581	9.69E-02	-0.0341	6.44E-10	0.0097	1.64E-01	-0.0435	6.39E-06	-0.0382	1.15E-01	-0.1513	2.37E-64	-0.1134	8.14E-22	-0.1550	1.29E-24	-0.0714	3.37E-02
212		44	Mig3	0.0189	9.93E-03	0.0224	1.40E-02	0.0121	3.54E-01	-0.0467	2.01E-01	-0.0604	5.51E-28	0.0061	3.85E-01	-0.1149	6.04E-33	-0.1129	3.00E-06	-0.1362	5.67E-60	-0.0668	1.02E-09	-0.1530	4.84E-27	-0.1402	1.42E-05
213		45	Ndt80	0.0109	1.52E-01	0.0261	1.26E-02	-0.0576	2.28E-06	-0.0780	5.01E-03	0.0660	4.83E-33	0.0444	2.25E-10	-0.0748	7.83E-15	-0.0739	2.29E-03	0.0707	7.26E-19	0.0352	1.92E-04	-0.0161	3.05E-01	-0.0125	8.01E-01
214		46	Ndt80	0.0000	9.97E-01	-0.0285	8.48E-03	-0.0638	3.52E-08	-0.0800	2.05E-03	0.0749	3.51E-42	0.0025	7.18E-01	-0.0588	1.03E-09	-0.0982	4.96E-05	0.0767	1.78E-21	0.0301	1.07E-03	-0.0099	5.68E-01	-0.0684	3.16E-01
215		47	Nhp6a	0.1011	3.02E-45	0.0665	2.30E-11	0.0947	5.32E-17	0.0471	7.37E-02	0.0570	4.71E-25	-0.0826	4.05E-32	0.0858	4.73E-19	0.0205	3.98E-01	-0.1028	4.46E-33	-0.1921	5.48E-86	-0.0479	9.16E-03	-0.0090	8.85E-01
216		48	Nhp6a	0.1178	2.30E-61	0.0701	1.26E-12	0.1106	1.25E-22	0.0991	1.47E-04	0.0534	3.43E-22	-0.0882	1.81E-36	0.0793	1.71E-16	0.0552	2.28E-02	-0.1097	5.30E-37	-0.1936	3.52E-86	-0.0518	4.70E-03	-0.0317	6.26E-01
217		49	Nhp6b	0.0941	2.00E-39	0.0743	2.32E-14	0.1083	7.15E-21	0.0296	2.77E-01	0.0495	2.60E-19	-0.0656	7.29E-21	0.0968	8.29E-24	0.0100	6.80E-01	-0.0906	5.15E-26	-0.1769	6.99E-70	-0.0537	2.02E-03	-0.0008	9.88E-01
218		50	Nhp6b	0.0889	5.60E-35	0.0676	3.15E-12	0.0966	1.71E-16	0.0775	5.51E-03	0.0350	2.28E-10	-0.0759	2.27E-27	0.0869	1.76E-19	0.0520	3.20E-02	-0.0897	7.14E-26	-0.1698	6.00E-64	-0.0483	4.21E-03	-0.0982	4.40E-02
219		51	Nrg1	0.0254	3.27E-04	0.0008	9.34E-01	-0.0353	2.62E-03	-0.0467	8.33E-02	-0.0408	1.26E-13	-0.0911	8.72E-39	-0.1176	1.93E-34	-0.0980	5.15E-05	-0.1403	1.70E-57	-0.1801	9.41E-67	-0.1549	3.99E-20	-0.0837	1.31E-01
220		52	Nrg1	0.0222	1.31E-03	-0.0279	2.24E-03	-0.0120	2.99E-01	-0.0230	3.96E-01	-0.0536	2.49E-22	-0.1295	8.99E-77	-0.0804	6.85E-17	-0.0486	4.49E-02	-0.1432	8.75E-56	-0.1732	2.51E-57	-0.1358	8.97E-15	-0.1840	6.49E-04
221		53	Oaf1	0.0062	4.19E-01	0.0262	6.28E-03	-0.0190	1.81E-01	-0.0563	1.02E-01	-0.0556	6.55E-24	0.0126	7.20E-02	-0.1250	9.38E-39	-0.0920	1.44E-04	-0.0916	1.99E-31	-0.0243	1.83E-02	-0.1339	1.23E-24	-0.0308	3.69E-01
222		54	Oaf1	-0.0456	1.52E-10	-0.0617	8.37E-13	-0.0312	2.06E-02	0.0189	6.08E-01	-0.1696	6.66E-211	-0.0849	6.54E-34	-0.2133	5.59E-111	-0.2008	6.10E-17	-0.2445	6.99E-179	-0.0601	5.63E-07	-0.2741	1.82E-91	-0.2371	7.75E-14
223		55	Pbf1	0.0269	1.53E-04	0.0606	1.24E-11	-0.0264	3.35E-02	0.0130	7.05E-01	0.0737	8.20E-41	0.1293	1.27E-76	0.0330	6.29E-04	0.0036	8.80E-01	0.0854	1.57E-22	0.0999	6.66E-19	0.1247	3.19E-16	0.1151	8.01E-04
224		56	Pbf1	0.0264	2.14E-04	0.0376	3.88E-05	-0.0071	5.60E-01	-0.0222	4.95E-01	0.0777	3.54E-45	0.1041	3.85E-50	0.0482	5.73E-07	-0.0452	6.25E-02	0.0898	3.33E-25	0.1052	4.87E-22	0.1039	2.71E-11	0.0240	5.10E-01
225		57	Pbf2	0.0089	1.99E-01	0.0691	2.06E-16	-0.0290	2.49E-02	0.0251	5.15E-01	0.0499	1.36E-19	0.1426	7.02E-93	0.0692	6.82E-13	0.0075	7.57E-01	0.1002	1.87E-28	0.0941	9.00E-14	0.1391	3.86E-22	0.0771	1.34E-02
226		58	Pbf2	0.0316	4.79E-06	0.0833	1.53E-22	0.0090	4.71E-01	-0.0273	4.29E-01	0.0914	5.35E-62	0.1642	5.28E-123	0.0938	1.76E-22	-0.0128	5.98E-01	0.1178	3.99E-38	0.1242	2.89E-24	0.1392	2.18E-20	0.0520	1.26E-01
227		59	Pdr1	0.0030	6.99E-01	-0.0142	1.70E-01	-0.0444	4.52E-04	-0.0653	1.98E-02	0.0278	4.63E-07	0.0060	3.92E-01	-0.0687	9.77E-13	-0.0761	1.69E-03	0.0227	4.13E-03	0.0185	5.29E-02	-0.0095	5.21E-01	0.0400	4.09E-01
228		60	Phd1	0.0370	5.76E-07	0.0539	5.32E-09	0.0163	2.13E-01	-0.0292	4.44E-01	0.0684	1.90E-35	0.1011	2.23E-47	0.0900	8.68E-21	0.0587	1.54E-02	0.0763	2.13E-20	0.0909	2.53E-17	0.1231	3.39E-18	0.0932	2.97E-03
229		61	Pho2	-0.0094	1.79E-01	0.0282	2.22E-03	-0.0564	1.19E-06	-0.1288	3.09E-06	-0.0414	5.77E-14	-0.0660	4.33E-21	-0.1208	2.77E-36	-0.1583	5.17E-11	-0.0395	1.13E-05	-0.1233	1.70E-30	-0.1015	3.82E-09	-0.0764	1.28E-01
230		62	Pho4	0.0964	1.68E-40	0.0992	2.19E-27	0.1033	2.58E-16	0.0448	1.96E-01	0.1281	2.58E-120	0.1360	1.36E-84	0.1810	5.79E-80	0.1375	1.23E-08	0.0382	6.62E-06	0.0590	5.67E-08	0.1273	9.87E-18	0.1962	5.51E-09
231		63	Pho4	0.0796	5.53E-28	0.0765	7.93E-17	0.0724	1.03E-08	0.0621	6.95E-02	0.1135	7.99E-95	0.1134	3.42E-59	0.1677	1.08E-68	0.1196	7.49E-07	0.0424	5.42E-07	0.0606	2.13E-08	0.1438	2.38E-22	0.1233	3.24E-04
232		64	Put3	-0.0440	2.62E-10	-0.0327	7.76E-05	-0.0582	2.17E-05	-0.0853	2.63E-02	-0.1673	3.44E-205	-0.0581	1.07E-16	-0.1967	2.64E-94	-0.1755	3.04E-13	-0.2335	1.57E-151	-0.0596	6.16E-06	-0.2064	1.82E-53	-0.1783	9.34E-09
233		65	Put3	-0.0471	1.06E-10	-0.0612	8.72E-12	-0.0407	2.44E-03	-0.0716	4.20E-02	-0.1499	1.56E-164	-0.0843	1.92E-33	-0.1901	4.09E-88	-0.1803	6.68E-14	-0.1899	1.43E-114	-0.0560	6.05E-07	-0.1855	1.33E-41	-0.1333	6.37E-05
234		66	Rap1	0.0379	9.44E-07	0.0394	8.55E-05	0.0066	6.19E-01	-0.0077	7.96E-01	0.0362	4.85E-11	0.0498	1.15E-12	-0.0561	5.87E-09	-0.0255	2.93E-01	-0.0148	6.00E-02	-0.0073	4.53E-01	-0.0815	5.14E-09	0.0496	2.40E-01
235		67	Rap1	0.0144	6.13E-02	-0.0413	1.24E-04	-0.0330	6.35E-03	-0.0070	7.98E-01	0.0277	5.09E-07	-0.0492	2.12E-12	-0.0685	1.09E-12	-0.0261	2.81E-01	0.0081	3.03E-01	-0.0332	3.27E-04	-0.0462	3.78E-03	-0.0692	1.93E-01
236		68	Rdr1	-0.0111	1.06E-01	-0.0032	6.92E-01	-0.0080	5.64E-01	0.0446	2.82E-01	-0.1375	1.54E-138	-0.0313	7.89E-06	-0.1052	7.23E-28	-0.1180	1.05E-06	-0.1689	4.03E-75	-0.0197	1.66E-01	-0.0647	1.59E-06	-0.1055	4.10E-04
237		69	Rdr1	0.0008	9.02E-01	0.0118	1.41E-01	0.0227	1.02E-01	-0.0159	7.19E-01	-0.1243	2.43E-113	-0.0045	5.19E-01	-0.0933	2.97E-22	-0.0981	5.08E-05	-0.1814	6.36E-86	-0.0338	2.00E-02	-0.0758	1.52E-08	-0.0982	7.10E-04
238		70	Rds1	-0.0227	9.42E-04	-0.0007	9.26E-01	-0.0066	6.49E-01	-0.0210	7.07E-01	-0.1383	4.23E-140	0.0005	9.45E-01	-0.0649	1.57E-11	-0.0865	3.52E-04	-0.1819	4.11E-87	-0.0260	9.47E-02	-0.0599	3.14E-06	-0.0559	3.78E-02
239		71	Rds2	-0.0457	1.18E-10	-0.0263	1.40E-03	-0.0277	5.81E-02	-0.0766	1.22E-01	-0.1969	8.13E-285	-0.0581	1.11E-16	-0.2028	3.13E-100	-0.2462	6.51E-25	-0.2386	1.25E-168	-0.0389	3.51E-03	-0.2366	6.17E-78	-0.2006	3.25E-13
240		72	Rds2	-0.0414	6.59E-09	-0.0226	7.14E-03	-0.0432	2.56E-03	-0.1094	1.57E-02	-0.1762	1.28E-227	-0.0322	4.25E-06	-0.2131	8.78E-111	-0.2740	1.10E-30	-0.2321	3.55E-162	-0.0071	5.80E-01	-0.2458	1.03E-81	-0.1875	4.48E-11
241		73	Rgt1	-0.0286	4.45E-05	-0.0210	1.16E-02	-0.0118	3.92E-01	0.0245	5.52E-01	-0.1324	1.63E-128	-0.0187	7.77E-03	-0.1550	6.99E-59	-0.1716	1.03E-12	-0.1671	1.58E-79	0.0114	3.81E-01	-0.1696	9.70E-37	-0.1496	5.24E-07
242		74	Rgt1	-0.0232	9.40E-04	-0.0132	1.14E-01	-0.0084	5.39E-01	-0.0297	4.49E-01	-0.1224	6.42E-110	-0.0111	1.13E-01	-0.1554	3.87E-59	-0.1654	6.53E-12	-0.1706	3.01E-82	-0.0009	9.42E-01	-0.1729	1.44E-37	-0.1517	8.04E-07
243		75	Rph1	0.0258	6.27E-04	-0.0442	5.40E-05	-0.0694	1.39E-09	-0.0377	1.34E-01	0.0344	4.40E-10	-0.0911	8.57E-39	-0.0704	2.63E-13	-0.0555	2.21E-02	-0.0299	2.20E-04	-0.0863	2.85E-21	-0.0405	2.30E-02	-0.0463	6.16E-01
244		76	Rph1	0.0298	6.21E-05	0.0114	2.82E-01	-0.0555	1.29E-06	-0.0974	1.73E-04	0.0312	1.57E-08	-0.0680	2.85E-22	-0.0736	2.15E-14	-0.1045	1.57E-05	-0.0381	3.39E-06	-0.0961	6.41E-25	-0.0764	1.70E-05	-0.1504	2.54E-02
245		77	Rpn4	0.0144	5.92E-02	0.0161	1.29E-01	-0.0596	9.14E-07	-0.0401	1.44E-01	0.0578	9.57E-26	0.0177	1.14E-02	-0.0690	7.93E-13	-0.0364	1.33E-01	0.0427	8.12E-08	0.0110	2.38E-01	-0.0619	8.92E-05	0.0028	9.57E-01
246		78	Rpn4	0.0121	1.10E-01	-0.0351	1.10E-03	-0.0338	4.04E-03	-0.0095	7.19E-01	0.0605	4.38E-28	-0.0189	6.92E-03	-0.0480	6.39E-07	-0.0374	1.23E-01	0.0566	1.85E-12	0.0024	7.94E-01	0.0049	7.73E-01	-0.1130	6.46E-02
247		79	Rsc3	-0.0137	4.44E-02	0.0134	7.36E-02	0.0170	3.16E-01	-0.0592	5.62E-01	-0.1781	1.28E-232	0.0087	2.13E-01	-0.0730	3.32E-14	-0.2086	3.51E-18	-0.2556	1.04E-170	-0.0035	8.61E-01	-0.1076	3.36E-20	-0.2043	1.43E-16
248		80	Rsc30	-0.0055	4.25E-01	0.0196	9.09E-03	0.0307	6.92E-02	-0.0065	9.47E-01	-0.1628	3.46E-194	0.0197	5.01E-03	-0.0337	4.65E-04	-0.1277	1.26E-07	-0.2174	3.03E-124	0.0006	9.74E-01	-0.0455	1.06E-04	-0.1066	1.96E-05
249		81	Rtg3	0.0188	1.46E-02	0.0229	2.60E-02	0.0077	5.38E-01	0.0015	9.63E-01	0.0376	9.02E-12	0.0204	3.53E-03	-0.0184	5.58E-02	-0.0630	9.27E-03	0.0138	8.20E-02	-0.0170	7.67E-02	-0.0404	7.73E-03	0.0011	9.77E-01

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
250		82	Sfl1	-0.0116	1.18E-01	-0.0454	5.68E-06	-0.0404	8.58E-04	-0.0244	3.85E-01	0.0245	8.57E-06	-0.0154	2.84E-02	-0.0372	1.14E-04	-0.0610	1.18E-02	0.0457	2.85E-08	0.0438	8.24E-06	-0.0106	5.04E-01	-0.0278	5.64E-01
251		83	Sfp1	0.0412	5.82E-08	0.0342	6.87E-04	0.0156	2.20E-01	-0.0030	9.17E-01	0.0825	9.16E-51	0.0774	1.99E-28	0.0013	8.93E-01	-0.0135	5.77E-01	0.0726	1.10E-19	0.0654	2.16E-11	0.0157	2.88E-01	0.0604	1.70E-01
252		84	Sfp1	0.0335	3.95E-06	-0.0412	6.03E-05	-0.0147	1.93E-01	-0.0248	3.28E-01	0.1080	7.33E-86	0.0310	9.80E-06	0.0004	9.63E-01	-0.0355	1.43E-01	0.1123	2.54E-40	0.0934	1.61E-22	0.0337	7.03E-02	-0.0521	5.28E-01
253		85	Sip4	0.0214	3.95E-03	0.0297	9.88E-04	0.0063	6.54E-01	-0.0133	7.18E-01	-0.0533	3.97E-22	0.0502	7.75E-13	-0.1184	6.83E-35	-0.1149	2.00E-06	-0.1134	2.36E-43	0.0289	9.28E-03	-0.1400	5.08E-26	-0.0968	2.73E-03
254		86	Sip4	0.0197	7.43E-03	0.0222	1.34E-02	0.0159	2.47E-01	-0.0044	9.13E-01	-0.0695	1.52E-36	0.0386	3.73E-08	-0.1271	5.17E-40	-0.1526	2.51E-10	-0.1570	1.23E-80	0.0055	6.23E-01	-0.1761	4.70E-39	-0.1397	4.75E-06
255		87	Skn7	0.0033	6.64E-01	0.0096	3.14E-01	0.0005	9.72E-01	-0.0884	1.28E-02	-0.0014	7.99E-01	0.0331	2.33E-06	-0.0193	4.50E-02	-0.0097	6.89E-01	-0.0029	7.15E-01	0.0375	2.82E-04	-0.0075	5.84E-01	0.0056	8.67E-01
256		88	Smp1	0.0097	2.04E-01	0.0372	1.33E-04	-0.0099	4.48E-01	-0.0713	4.82E-02	-0.0434	3.32E-15	-0.0011	8.71E-01	-0.1018	3.56E-26	-0.1862	9.62E-15	-0.0649	4.13E-16	-0.0661	5.86E-11	-0.0904	2.15E-10	-0.0974	2.88E-03
257		89	Spt15	0.0256	6.72E-04	0.0327	1.23E-03	-0.0110	3.69E-01	-0.0759	8.68E-03	0.0212	1.18E-04	-0.0216	2.10E-03	-0.0490	3.57E-07	-0.0787	1.15E-03	-0.0148	6.82E-02	-0.0804	1.13E-16	-0.0295	5.94E-02	0.0187	6.75E-01
258		90	Srd1	0.0051	5.01E-01	0.0019	8.49E-01	-0.0402	1.40E-03	-0.0418	1.42E-01	0.0047	3.95E-01	-0.0061	3.86E-01	-0.1022	2.24E-26	-0.0872	3.18E-04	-0.0300	2.04E-04	-0.0419	1.84E-05	-0.1144	1.96E-14	-0.0398	3.92E-01
259		91	Srd1	-0.0054	4.75E-01	0.0061	5.42E-01	-0.0411	8.57E-04	-0.0560	6.47E-02	-0.0157	4.32E-03	-0.0274	9.10E-05	-0.1019	2.97E-26	-0.1414	4.72E-09	-0.0220	6.65E-03	-0.0464	2.21E-06	-0.0960	4.84E-10	0.0058	8.85E-01
260		92	Stb3	0.0279	1.67E-04	0.0142	1.57E-01	-0.0081	5.04E-01	-0.0492	7.55E-02	0.1257	5.17E-116	0.1162	4.61E-62	0.0096	3.17E-01	-0.0278	2.52E-01	0.1482	2.93E-73	0.1452	1.72E-50	0.0471	3.37E-03	-0.0271	5.90E-01
261		93	Stp2	0.0340	1.30E-06	0.0673	1.24E-15	0.0453	7.01E-04	-0.0676	1.10E-01	0.0310	1.85E-08	0.1103	4.34E-56	0.1115	4.26E-31	0.1181	1.04E-06	0.0069	4.41E-01	0.0867	6.43E-12	0.1052	3.48E-14	0.1444	9.64E-07
262		94	Stp2	0.0490	1.29E-11	0.0768	3.92E-19	0.0239	8.66E-02	-0.0216	6.57E-01	0.0167	2.47E-03	0.1271	5.23E-74	0.0614	1.80E-10	0.0751	1.94E-03	-0.0377	9.14E-06	0.0735	1.16E-09	0.0805	1.42E-09	0.1020	2.62E-04
263		95	Stp4	0.0211	5.94E-03	0.0319	8.41E-04	0.0202	1.43E-01	-0.0002	9.96E-01	0.0259	2.72E-06	0.0424	1.44E-09	0.0560	6.22E-09	0.1068	1.01E-05	-0.0102	2.00E-01	-0.0027	7.91E-01	0.0325	1.60E-02	0.1246	1.31E-04
264		96	Sum1	0.0494	6.32E-13	-0.0056	5.53E-01	0.0163	1.40E-01	-0.0489	5.22E-02	0.0596	2.93E-27	-0.0296	2.37E-05	-0.0023	8.09E-01	-0.0656	6.77E-03	0.0146	1.15E-01	-0.0277	8.32E-03	-0.0238	2.30E-01	-0.2501	4.91E-03
265		97	Sum1	0.0368	1.50E-07	0.0149	1.19E-01	-0.0144	2.01E-01	-0.0013	9.60E-01	0.0507	3.37E-20	-0.0120	8.72E-02	-0.0333	5.60E-04	-0.0257	2.90E-01	0.0208	1.99E-02	-0.0271	8.74E-03	-0.0197	2.89E-01	-0.0213	7.46E-01
266		98	Sut2	-0.0449	1.25E-09	-0.0425	7.36E-06	-0.0972	3.89E-14	-0.0913	2.57E-03	-0.0686	1.27E-35	-0.0299	2.05E-05	-0.1630	5.70E-65	-0.1655	6.35E-12	-0.0918	1.17E-28	-0.0060	5.66E-01	-0.1511	1.37E-25	-0.1495	2.00E-04
267		99	Sut2	-0.0281	1.90E-04	-0.0209	3.19E-02	-0.0741	1.22E-08	-0.0997	1.05E-03	-0.0360	6.62E-11	0.0065	3.57E-01	-0.1353	3.89E-45	-0.1838	2.17E-14	-0.0644	1.64E-15	0.0062	5.39E-01	-0.1388	2.17E-22	-0.1548	1.04E-04
268		100	Tbf1	0.0272	9.94E-05	-0.0289	2.20E-03	-0.0387	6.79E-04	-0.0608	1.98E-02	-0.0898	8.09E-60	-0.2001	8.86E-183	-0.1369	3.67E-46	-0.0898	2.06E-04	-0.2002	1.41E-112	-0.2578	8.00E-139	-0.2036	8.03E-30	-0.1362	3.73E-02
269		101	Tbs1	-0.0099	1.50E-01	0.0091	2.58E-01	-0.0337	1.57E-02	-0.0468	2.81E-01	-0.1163	2.24E-99	0.0084	2.31E-01	-0.0811	3.70E-17	-0.1128	3.09E-06	-0.1586	1.06E-66	0.0024	8.67E-01	-0.0123	3.58E-01	-0.0969	8.97E-04
270		102	Tbs1	0.0121	8.18E-02	0.0320	7.35E-05	-0.0120	3.99E-01	-0.0727	1.16E-01	-0.0927	1.18E-63	0.0432	6.78E-10	-0.0512	1.06E-07	-0.1014	2.79E-05	-0.1785	2.94E-87	0.0024	8.65E-01	-0.0319	1.49E-02	-0.0940	9.50E-04
271		103	Tea1	-0.0141	5.45E-02	0.0266	1.84E-03	-0.0302	3.80E-02	0.0602	3.01E-01	-0.0962	1.62E-68	-0.0032	6.46E-01	0.0247	1.04E-02	-0.0645	7.79E-03	-0.1224	1.63E-48	-0.0769	3.15E-10	0.0520	5.29E-05	-0.0008	9.75E-01
272		104	Tec1	0.0145	4.02E-02	0.0382	1.53E-05	0.0548	1.30E-05	-0.0428	2.13E-01	0.0644	1.28E-31	0.1019	3.91E-48	0.0823	1.17E-17	0.0325	1.81E-01	0.0732	8.85E-17	0.0581	4.61E-07	0.0551	2.38E-04	0.0375	2.74E-01
273		105	Tye7	0.0214	4.87E-03	0.0330	5.41E-04	0.0223	9.86E-02	-0.0103	7.72E-01	0.0420	2.66E-14	0.0660	4.49E-21	0.0410	2.10E-05	0.0679	5.09E-03	0.0395	8.05E-07	0.0499	1.35E-06	0.0656	1.88E-06	0.0226	4.95E-01
274		106	Tye7	0.0506	7.01E-12	0.0313	1.21E-03	0.0360	3.53E-03	0.0293	3.35E-01	0.1171	9.44E-101	0.1042	3.30E-50	0.0871	1.36E-19	0.0664	6.17E-03	0.0780	4.81E-21	0.0947	1.05E-20	0.0747	1.29E-06	0.0508	2.08E-01
275		107	Ume6	-0.0378	5.00E-08	-0.0237	3.75E-03	-0.0059	6.62E-01	-0.0262	6.10E-01	-0.0926	1.37E-63	-0.0407	6.55E-09	0.0908	3.87E-21	-0.0135	5.77E-01	-0.0995	5.03E-28	-0.0490	3.01E-04	0.1245	1.03E-19	0.0747	6.65E-03
276		108	Ume6	0.0216	2.75E-03	0.0439	5.06E-07	-0.0036	7.89E-01	0.0273	5.59E-01	-0.0272	7.71E-07	0.0275	8.56E-05	0.0914	2.05E-21	-0.0192	4.28E-01	-0.0857	8.78E-24	-0.0213	6.93E-02	0.1368	5.92E-23	0.0120	6.73E-01
277		109	Usv1	0.0277	9.72E-05	0.0024	8.04E-01	0.0054	6.42E-01	-0.0410	1.46E-01	-0.0278	4.80E-07	-0.0957	1.36E-42	-0.0594	6.82E-10	-0.0559	2.10E-02	-0.0830	1.88E-21	-0.1522	3.14E-49	-0.0594	6.19E-04	-0.0504	2.89E-01
278		110	Xbp1	-0.0015	8.36E-01	0.0122	1.63E-01	-0.0250	6.48E-02	-0.0675	4.69E-02	-0.0671	4.17E-34	0.0587	5.18E-17	-0.1574	1.11E-60	-0.1582	5.33E-11	-0.1401	3.41E-60	0.0639	4.32E-08	-0.1490	1.04E-27	-0.0472	1.73E-01
279		111	Yap1	0.0248	7.56E-04	0.0039	6.97E-01	0.0040	7.36E-01	-0.0151	5.82E-01	0.0049	3.72E-01	-0.0851	4.83E-34	-0.0137	1.54E-01	0.0239	3.25E-01	-0.0881	2.96E-26	-0.1316	1.14E-41	-0.0897	8.47E-08	0.0185	7.23E-01
280		112	Yap6	0.0217	4.69E-03	-0.0140	1.94E-01	-0.0282	1.88E-02	-0.0706	8.05E-03	0.0683	2.80E-35	0.0068	3.33E-01	-0.0439	5.15E-06	-0.0808	8.54E-04	0.0562	1.36E-12	0.0245	7.94E-03	-0.0209	1.97E-01	-0.0977	9.44E-02
281		113	Ybr239c	-0.0076	2.92E-01	0.0004	9.63E-01	-0.0258	4.98E-02	-0.0516	1.36E-01	-0.0650	4.13E-32	0.0209	2.89E-03	-0.1343	1.78E-44	-0.1693	2.05E-12	-0.1003	2.89E-32	0.0463	3.79E-05	-0.1336	2.86E-21	-0.0952	5.04E-03
282		114	Ydr520c	-0.0278	3.10E-04	-0.0411	1.05E-04	-0.0543	9.08E-06	-0.0888	1.56E-03	0.0324	4.01E-09	-0.0052	4.62E-01	-0.0617	1.51E-10	-0.1204	6.24E-07	0.0364	4.05E-06	0.0095	3.08E-01	-0.0157	3.15E-01	-0.0033	9.45E-01
283		115	Yer130c	0.0351	4.54E-06	0.0480	1.25E-06	-0.0007	9.55E-01	-0.0382	2.19E-01	0.0045	4.17E-01	0.0211	2.67E-03	-0.0705	2.50E-13	-0.0675	5.35E-03	-0.0394	7.08E-07	-0.0461	3.46E-06	-0.0701	6.48E-07	-0.0335	3.89E-01
284		116	Yer130c	-0.0063	4.08E-01	0.0092	3.64E-01	-0.0631	4.12E-07	-0.0936	1.40E-03	-0.0240	1.34E-05	-0.0575	2.28E-16	-0.0912	2.60E-21	-0.1336	3.16E-08	-0.0335	2.84E-05	-0.0796	1.78E-16	-0.0372	1.43E-02	-0.0364	3.99E-01
285		117	Ygr067c	0.0291	1.39E-04	0.0409	2.72E-05	0.0023	8.64E-01	0.0210	5.20E-01	0.0221	6.28E-05	0.0574	2.53E-16	-0.0379	8.34E-05	-0.0896	2.16E-04	-0.0362	5.56E-06	-0.0179	7.65E-02	-0.0635	5.75E-06	-0.0714	4.82E-02
286		118	Ykl222c	-0.0187	7.48E-03	0.0038	6.50E-01	-0.0304	2.57E-02	-0.1214	3.24E-03	-0.1006	8.15E-75	-0.0094	1.78E-01	-0.0555	8.22E-09	-0.0595	1.41E-02	-0.1290	5.14E-47	-0.0280	3.28E-02	-0.0172	2.06E-01	-0.0408	1.74E-01
287		119	Ykl222c	-0.0068	3.29E-01	0.0055	5.08E-01	0.0039	7.74E-01	-0.0648	1.35E-01	-0.0851	6.84E-54	0.0082	2.41E-01	-0.0425	1.03E-05	-0.0223	3.57E-01	-0.1549	2.84E-67	-0.0306	1.97E-02	-0.0607	9.59E-06	-0.0575	4.95E-02
288		120	Yll054c	-0.0208	2.41E-03	0.0019	8.06E-01	0.0094	5.29E-01	-0.1130	3.46E-02	-0.1740	5.64E-222	-0.0203	3.81E-03	-0.1418	1.92E-49	-0.1443	2.26E-09	-0.2461	2.05E-158	-0.0472	3.45E-03	-0.1686	4.80E-41	-0.1754	8.40E-11
289		121	Yll054c	0.0019	7.78E-01	0.0211	6.63E-03	0.0367	1.33E-02	-0.0596	3.17E-01	-0.1572	5.09E-181	0.0087	2.16E-01	-0.1333	7.26E-44	-0.1287	9.98E-08	-0.2665	9.29E-186	-0.0484	3.02E-03	-0.1928	3.88E-53	-0.1240	2.80E-06
290		122	Yml081w	0.0590	3.60E-16	0.0587	7.90E-10	-0.0032	7.92E-01	-0.0112	6.95E-01	0.0429	7.45E-15	0.0406	6.93E-09	-0.0503	1.82E-07	-0.0932	1.18E-04	-0.0487	9.73E-09	-0.0413	6.08E-05	-0.1190	7.18E-14	-0.0952	3.80E-02
291		123	Ynr063w	-0.0513	9.56E-12	-0.0177	6.50E-02	-0.0832	2.11E-10	-0.1269	1.43E-04	-0.0845	3.54E-53	-0.0267	1.43E-04	-0.1695	3.66E-70	-0.2033	2.46E-17	-0.0731	1.54E-19	-0.0142	1.68E-01	-0.1370	3.47E-22	-0.1151	1.05E-03
292		124	Yox1	0.0079	2.54E-01	0.0250	7.57E-03	-0.0504	5.29E-06	-0.0706	6.96E-03	-0.0127	2.15E-02	-0.0860	9.89E-35	-0.0803	7.47E-17	-0.0982	4.99E-05	-0.0685	1.02E-13	-0.1571	2.27E-50	-0.0524	7.58E-03	0.0264	6.83E-01
293		125	Ypr013c	0.0023	7.57E-01	0.0331	5.13E-04	-0.0433	8.16E-04	-0.0672	3.11E-02	-0.0155	4.87E-03	0.0056	4.23E-01	-0.0664	5.21E-12	-0.0662	6.28E-03	-0.0352	1.79E-05	-0.0475	4.51E-06	-0.0140	3.31E-01	-0.0282	4.67E-01
294		126	Ypr013c	0.0003	9.73E-01	0.0402	1.95E-05	-0.0372	4.07E-03	-0.0755	2.89E-02	-0.0280	3.77E-07	0.0088	2.09E-01	-0.0546	1.48E-08	-0.1049	1.45E-05	-0.0391	1.90E-06	-0.0386	2.35E-04	-0.0222	1.24E-01	-0.0829	1.49E-02
295		127	Ypr015c	0.0262	5.51E-04	0.0476	6.44E-07	0.0019	8.87E-01	-0.0416	2.88E-01	-0.0350	2.28E-10	0.0040	5.66E-01	-0.0624	8.93E-11	-0.0671	5.65E-03	-0.0659	1.80E-16	-0.0523	3.89E-07	-0.0681	1.25E-06	-0.0739	1.69E-02
296		128	Ypr196w	-0.0030	6.97E-01	0.0069	4.89E-01	-0.0488	2.19E-04	-0.0510	9.74E-02	-0.0059	2.87E-01	0.0075	2.86E-01	-0.0893	1.71E-20	-0.0784	1.20E-03	-0.0139	7.80E-02	-0.0038	7.00E-01	-0.0691	9.28E-07	-0.1015	9.93E-03
297		129	Ypr196w	-0.0313	4.00E-05	-0.0522	4.51E-07	-0.0579	2.57E-06	-0.0536	5.93E-02	-0.0123	2.59E-02	-0.0507	4.70E-13	-0.1024	1.83E-26	-0.1053	1.35E-05	0.0158	4.85E-02	-0.0058	5.45E-01	-0.0707	4.89E-06	-0.1157	1.25E-02
298		130	Yrm1	-0.0344	6.56E-06	-0.0494	2.87E-06	-0.0641	1.11E-07	-0.0968	5.37E-04	0.0293	1.10E-07	-0.0177	1.18E-02	-0.0636	3.99E-11	-0.1424	3.58E-09	0.0714	2.72E-19	0.0215	2.16E-02	0.0495	1.96E-03	-0.0563	2.46E-01
299		131	Yrr1	-0.0332	3.19E-06	-0.0288	7.91E-04	-0.0255	5.75E-02	-0.0782	5.71E-02	-0.1103	1.28E-89	-0.0473	1.53E-11	-0.0571	3.04E-09	-0.0992	4.14E-05	-0.1073	4.25E-35	-0.0348	4.12E-03	0.0022	8.75E-01	-0.0412	1.70E-01
300		132	Yrr1	-0.0243	1.27E-03	-0.0237	1.62E-02	-0.0358	4.40E-03	-0.0377	2.46E-01	-0.0042	4.45E-01	-0.0063	3.72E-01	-0.0399	3.49E-05	-0.0904	1.87E-04	0.0041	6.11E-01	0.0028	7.75E-01	0.0182	2.25E-01	-0.0333	3.61E-01

Table S2

Symbol	**RefSeq**	**Phylum**	**Class**	**Genus**	**Species**	**Strain**
A	NC_014169	Actinobacteria	Actinobacteria	Bifidobacterium	longum	subsp. longum JDM301
B	NC_017031	Actinobacteria	Actinobacteria	Corynebacterium	pseudotuberculosis	P54B96
C	NC_009565	Actinobacteria	Actinobacteria	Mycobacterium	tuberculosis	F11
D	NC_007429	Chlamydiae/Verrucomicrobia group	Chlamydiae	Chlamydia	trachomatis	A/HAR-13
E	NC_017290	Chlamydiae/Verrucomicrobia group	Chlamydiae	Chlamydophila	psittaci	08DC60
F	NC_012588	Crenarchaeota	Thermoprotei	Sulfolobus	islandicus	M.14.25
G	NC_006576	Cyanobacteria	Oscillatoriophycideae	Synechococcus	elongatus	PCC 6301
H	NC_005042	Cyanobacteria	Prochlorales	Prochlorococcus	marinus	subsp. marinus str. CCMP1375
I	NC_005791	Euryarchaeota	Methanococci	Methanococcus	maripaludis	S2
J	NC_011658	Firmicutes	Bacilli	Bacillus	cereus	AH187
K	NC_012659	Firmicutes	Bacilli	Bacillus	anthracis	str. A0248
L	NC_018500	Firmicutes	Bacilli	Bacillus	thuringiensis	HD-771
M	NC_017469	Firmicutes	Bacilli	Lactobacillus	delbrueckii	subsp. bulgaricus 2038
N	NC_008527	Firmicutes	Bacilli	Lactococcus	lactis	subsp. cremoris SK11
O	NC_003210	Firmicutes	Bacilli	Listeria	monocytogenes	EGD-e
P	NC_016912	Firmicutes	Bacilli	Staphylococcus	aureus	subsp. aureus VC40
Q	NC_017591	Firmicutes	Bacilli	Streptococcus	pneumoniae	INV104
R	NC_008021	Firmicutes	Bacilli	Streptococcus	pyogenes	MGAS9429
S	NC_009698	Firmicutes	Clostridia	Clostridium	botulinum	A str. Hall
T	NC_007493	Proteobacteria	Alphaproteobacteria	Rhodobacter	sphaeroides	2.4.1
U	NC_005296	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas	palustris	CGA009
V	NC_017246	Proteobacteria	Alphaproteobacteria	Brucella	melitensis	M5-90
W	NC_017056	Proteobacteria	Alphaproteobacteria	Rickettsia	prowazekii	str. BuV67-CWPP
X	NC_008060	Proteobacteria	Betaproteobacteria	Burkholderia	cenocepacia	AU 1054
Y	NC_007434	Proteobacteria	Betaproteobacteria	Burkholderia	pseudomallei	1710b
Z	NC_003112	Proteobacteria	Betaproteobacteria	Neisseria	meningitidis	MC58
AA	NC_000915	Proteobacteria	delta/epsilon subdivisions	Helicobacter	pylori	26695
BB	NC_008787	Proteobacteria	delta/epsilon subdivisions	Campylobacter	jejuni	subsp. jejuni 81-176
CC	NC_006570	Proteobacteria	Gammaproteobacteria	Francisella	tularensis	subsp. tularensis SCHU S4
DD	NC_002942	Proteobacteria	Gammaproteobacteria	Legionella	pneumophila	subsp. pneumophila str. Philadelphia 1
EE	NC_009085	Proteobacteria	Gammaproteobacteria	Acinetobacter	baumannii	ATCC 17978
FF	NC_002528	Proteobacteria	Gammaproteobacteria	Buchnera	aphidicola	str. APS (Acyrthosiphon pisum)
GG	NC_000907	Proteobacteria	Gammaproteobacteria	Haemophilus	influenzae	Rd KW20
HH	NC_002516	Proteobacteria	Gammaproteobacteria	Pseudomonas	aeruginosa	PAO1
II	NC_003197	Proteobacteria	Gammaproteobacteria	Salmonella	enterica	subsp. enterica serovar Typhimurium str. LT2
JJ	NC_002505	Proteobacteria	Gammaproteobacteria	Vibrio	cholerae	O1 biovar El Tor str. N16961
KK	NC_003143	Proteobacteria	Gammaproteobacteria	Yersinia	pestis	CO92
LL	NC_003902	Proteobacteria	Gammaproteobacteria	Xanthomonas	campestris	pv. campestris str. ATCC 33913
MM	NC_010741	Spirochaetes	Spirochaetia	Treponema	pallidum	subsp. pallidum SS14
NN	NC_017502	Tenericutes	Mollicutes	Mycoplasma	gallisepticum	str. R(high)

Table S3

Organism	Group	SubGroup	N_chromosomes	N_chromosomes in PCA (1-mer null)	N_chromosomes in PCA (2-mer null)	Full length intr	Exon analysis	Full length exon
Ciona intestinalis	Animal	Primitive	14	14	12	3.63E+07		
Schistosoma mansoni strain Puerto Rico	Animal	Flatworm	8	8	8	1.05E+08		
Caenorhabditis briggsae AF16	Animal	Roundworm	6	6	6	2.54E+07	Y	2.03E+07
Caenorhabditis elegans Bristol N2	Animal	Roundworm	6	6	6	2.61E+07	Y	2.34E+07
Anopheles gambiae str. PEST	Animal	Insects	5	0	0	1.66E+06		
Apis mellifera	Animal	Insects	16	16	16	8.22E+07	Y	1.46E+07
Bombus terrestris	Animal	Insects	18	18	17	9.05E+07		
Drosophila melanogaster	Animal	Insects	6	5	5	3.08E+07	Y	2.07E+07
Drosophila pseudoobscura pseudoobscura	Animal	Insects	2	2	2	1.27E+07		
Drosophila simulans	Animal	Insects	6	5	4	3.16E+07		
Drosophila yakuba	Animal	Insects	6	5	5	3.19E+07		
Nasonia vitripennis	Animal	Insects	5	5	5	6.30E+07		
Tribolium castaneum	Animal	Insects	10	7	0	4.83E+07		
Cynoglossus semilaevis	Animal	Fishes	22	22	21	1.86E+08		
Danio rerio	Animal	Fishes	25	25	25	5.85E+08	Y	4.18E+07
Lepisosteus oculatus	Animal	Fishes	29	27	26	3.69E+08		
Oreochromis niloticus	Animal	Fishes	22	22	22	2.73E+08		
Oryzias latipes	Animal	Fishes	24	24	24	2.61E+08		
Poecilia reticulata	Animal	Fishes	23	23	23	3.10E+08		
Takifugu rubripes	Animal	Fishes	22	22	22	1.04E+08		
Ficedula albicollis	Animal	Birds	33	0	0	3.89E+08		
Gallus gallus	Animal	Birds	32	29	27	3.80E+08	Y	2.52E+07
Meleagris gallopavo	Animal	Birds	31	29	27	3.39E+08		
Taeniopygia guttata	Animal	Birds	33	29	25	3.79E+08		
Anolis carolinensis	Animal	Reptiles	13	10	7	3.92E+08	Y	1.41E+07
Chrysemys picta bellii	Animal	Reptiles	18	15	13	1.55E+08		
Monodelphis domestica	Animal	primitive mam	9	9	9	1.06E+09		
Ornithorhynchus anatinus	Animal	primitive mam	19	19	15	1.32E+08	Y	4.25E+06
Canis lupus familiaris	Animal	Carnivores	39	39	39	7.84E+08		
Felis catus	Animal	Carnivores	19	19	19	7.50E+08		
Bos taurus	Animal	hoofed mamm	30	30	29	8.20E+08		
Capra hircus	Animal	hoofed mamm	30	30	30	7.77E+08		
Ovis aries	Animal	hoofed mamm	27	27	27	8.16E+08		
Sus scrofa	Animal	hoofed mamm	20	19	19	6.57E+08		
Equus caballus	Animal	hoofed mamm	32	32	32	7.51E+08		
Microtus ochrogaster	Animal	glires	28	28	28	5.40E+08		
Mus musculus	Animal	glires	21	20	20	7.52E+08	Y	3.43E+07
Oryctolagus cuniculus	Animal	glires	22	22	18	5.75E+08		
Rattus norvegicus	Animal	glires	21	21	21	7.28E+08		
Callithrix jacchus	Animal	primates	24	24	23	8.99E+08		
Chlorocebus sabaeus	Animal	primates	31	30	30	9.04E+08		
Gorilla gorilla gorilla	Animal	primates	24	24	24	8.44E+08		
Homo sapiens	Animal	primates	24	24	23	8.71E+08	Y	3.14E+07
Macaca fascicularis	Animal	primates	21	21	21	9.01E+08		
Macaca mulatta	Animal	primates	21	21	21	8.62E+08		
Nomascus leucogenys	Animal	primates	26	26	26	9.13E+08		
Pan troglodytes	Animal	primates	25	25	24	8.98E+08		
Papio anubis	Animal	primates	21	21	21	7.82E+08		
Pongo abelii	Animal	primates	24	24	24	7.80E+08		
Solanum lycopersicum	Land Plant	eudicots	12	12	11	7.31E+07		
Cicer arietinum	Land Plant	eudicots	8	8	8	4.84E+07		
Fragaria vesca subsp. vesca	Land Plant	eudicots	7	7	6	3.30E+07		
Glycine max	Land Plant	eudicots	20	20	20	1.11E+08		
Malus domestica	Land Plant	eudicots	17	17	0	5.06E+07		
Medicago truncatula	Land Plant	eudicots	8	8	8	5.12E+07		

Phaseolus vulgaris	Land Plant eudicots	11	11	0	5.14E+07		
Populus trichocarpa	Land Plant eudicots	19	0	0	2.67E+06		
Prunus mume	Land Plant eudicots	8	8	8	3.37E+07		
Arabidopsis thaliana	Land Plant eudicots	5	5	0	1.77E+07	Y	3.34E+07
Citrus sinensis	Land Plant eudicots	9	9	7	3.54E+07		
Theobroma cacao	Land Plant eudicots	10	10	10	5.72E+07		
Vitis vinifera	Land Plant eudicots	19	19	19	1.00E+08	Y	2.81E+07
Brachypodium distachyon	Land Plant monocots	5	5	5	4.54E+07		
Oryza brachyantha	Land Plant monocots	12	12	12	4.67E+07		
Oryza sativa Indica Group	Land Plant monocots	12	12	12	5.17E+07		
Oryza sativa Japonica Group	Land Plant monocots	12	12	12	3.83E+07		
Sorghum bicolor	Land Plant monocots	10	10	10	4.54E+07		
Zea mays	Land Plant monocots	10	10	10	7.75E+07		
Micromonas sp. RCC299	Green Algae	17	0	0	8.10E+05		
Ostreococcus lucimarinus CCE9901	Green Algae	21	0	0	3.51E+05		
Ostreococcus tauri	Green Algae	20	0	0	4.96E+05	Y	1.00E+07
Plasmodium falciparum 3D7	Protist	14	0	0	1.30E+06	Y	1.17E+07
Dictyostelium discoideum AX4	Protist	6	0	0	2.23E+06	Y	2.07E+07
Saccharomyces cerevisiae S288c	Fungi	16	0	0	6.44E+04	Y	8.68E+06
Schizosaccharomyces pombe 972h	Fungi	3	0	0	3.89E+05	Y	7.09E+06

Table S4

Spearman correlation (ρ) of relative frequencies between chromosome pairs (both within and between genomes)

Word frequencies using k=6
Lower triangle: avg. (std. err.) of ρ for all inter-species chromosome pairs
Upper triangle: ρ of genomic pooled data

non-significant entries flagged in gray:
flagged in gray if >30% of chromosome pairs with p-value > 1e-6
flagged in gray if p-value > 1e-6

DATA: Exon/codon shuffling

	C. elegans	*D. melanogaster*	*D. rerio*	*G. gallus*	*M. musculus*	*H. sapiens*	*A. thaliana*	*D. discoideum*	*S. cerevisiae*
within-genome mean ρ (std. dev.)	0.9599 (0.0071)	0.8148 (0.0574)	0.9603 (0.0013)	0.8675 (0.0060)	0.9250 (0.0090)	0.9206 (0.0048)	0.9901 (0.0004)	0.9264 (0.0041)	0.7234 (0.0068)
C. elegans		0.4735	0.4073	0.4094	0.3491	0.349	0.6279	0.3248	0.4891
D. melanogaster	0.4510 (0.0064)		0.5091	0.4734	0.4665	0.4597	0.4226	0.2835	0.5008
D. rerio	0.3981 (0.0036)	0.4616 (0.0066)		0.8382	0.844	0.8139	0.6003	0.3098	0.5828
G. gallus	0.3780 (0.0040)	0.4296 (0.0055)	0.7745 (0.0026)		0.942	0.9591	0.5265	0.3047	0.6069
M. musculus	0.3362 (0.0049)	0.4184 (0.0075)	0.7980 (0.0038)	0.8478 (0.0048)		0.974	0.5122	0.272	0.5727
H. sapiens	0.3400 (0.0038)	0.4149 (0.0059)	0.7673 (0.0022)	0.8611 (0.0039)	0.8995 (0.0048)		0.4588	0.2885	0.5655
A. thaliana	0.6184 (0.0059)	0.3899 (0.0109)	0.5851 (0.0014)	0.4813 (0.0039)	0.4923 (0.0050)	0.4408 (0.0029)		0.3039	0.4749
D. discoideum	0.3154 (0.0027)	0.2670 (0.0024)	0.2981 (0.0007)	0.2840 (0.0016)	0.2610 (0.0024)	0.2756 (0.0014)	0.2980 (0.0015)		0.4604
S. cerevisiae	0.4274 (0.0046)	0.4158 (0.0042)	0.4969 (0.0020)	0.4886 (0.0028)	0.4784 (0.0033)	0.4738 (0.0023)	0.4148 (0.0030)	0.3891 (0.0032)	

DATA: Intron/1-mer

	C. elegans	*D. melanogaster*	*D. rerio*	*G. gallus*	*M. musculus*	*H. sapiens*	*A. thaliana*	*D. discoideum*	*S. cerevisiae*
within-genome mean ρ (std. dev.)	0.9485 (0.0107)	0.8971 (0.0347)	0.9964 (0.0002)	0.9714 (0.0015)	0.9927 (0.0012)	0.9924 (0.0005)	0.9913 (0.0004)	0.7179 (0.0125)	0.1306 (0.0037) (7% of pairs not sig.)
C. elegans		0.6671	0.4432	0.2233	0.2041	0.2826	0.0929	0.5871	0.3462
D. melanogaster	0.6498 (0.0071)		0.6387	0.4214	0.3748	0.4359	0.2934	0.464	0.4473
D. rerio	0.4443 (0.0054)	0.6011 (0.0063)		0.8222	0.7878	0.7987	0.5648	0.4114	0.4072
G. gallus	0.2458 (0.0045)	0.4051 (0.0069)	0.8177 (0.0009)		0.9574	0.932	0.7586	0.1837	0.4191
M. musculus	0.2074 (0.0044)	0.3381 (0.0068)	0.7812 (0.0010)	0.9342 (0.0010)		0.9548	0.7448	0.2304	0.3693
H. sapiens	0.2866 (0.0039)	0.3987 (0.0070)	0.7955 (0.0007)	0.9137 (0.0009)	0.9463 (0.0009)		0.7009	0.2706	0.3898
A. thaliana	0.1032 (0.0123)	0.2534 (0.0173)	0.5613 (0.0012)	0.7394 (0.0013)	0.7404 (0.0006)	0.6932 (0.0013)		-0.0463	0.3975
D. discoideum	0.4470 (0.0074)	0.3827 (0.0061)	0.3792 (0.0017)	0.2566 (0.0023)	0.2908 (0.0022)	0.3064 (0.0018)	0.0768 (0.0082)		0.2502
S. cerevisiae	0.0295 (0.0076)	0.0954 (0.0073)	0.1120 (0.0036)	0.1758 (0.0030)	0.1504 (0.0037)	0.1461 (0.0035)	0.2314 (0.0063)	-0.0175 (0.0087)	

DATA: Intron/2-mer

	C. elegans	*D. melanogaster*	*D. rerio*	*G. gallus*	*M. musculus*	*H. sapiens*	*A. thaliana*	*D. discoideum*	*S. cerevisiae*
within-genome mean ρ (std. dev.)	0.9139 (0.0145)	0.8334 (0.0536)	0.9919 (0.0003)	0.8937 (0.0053)	0.9457 (0.0088)	0.9793 (0.0012)	0.9405 (0.0023)	0.6821 (0.0130)	0.1033 (0.0035) (>25% of pairs not sig.)
C. elegans		0.3414	0.3423	0.2671	0.3347	0.3688	-0.0013	0.3578	0.0925
D. melanogaster	0.3264 (0.0066)		0.4684	0.3199	0.4334	0.4064	0.1557	0.4295	0.1734
D. rerio	0.3388 (0.0045)	0.4475 (0.0030)		0.5655	0.6685	0.6094	-0.0272	0.4938	0.1119
G. gallus	0.2175 (0.0045)	0.2863 (0.0034)	0.5148 (0.0022)		0.8034	0.6937	0.108	0.2407	0.2958
M. musculus	0.3245 (0.0039)	0.4095 (0.0036)	0.6474 (0.0027)	0.7243 (0.0041)		0.8015	0.1273	0.3119	0.2447
H. sapiens	0.3569 (0.0030)	0.3790 (0.0039)	0.6011 (0.0012)	0.6365 (0.0028)	0.7732 (0.0036)		0.0498	0.3354	0.1991
A. thaliana	0.0039 (0.0032)	0.1425 (0.0064)	-0.0244 (0.0012)	0.1289 (0.0029)	0.1289 (0.0014)	0.0497 (0.0013)		0.0227	0.2221
D. discoideum	0.2739 (0.0061)	0.3553 (0.0069)	0.3610 (0.0036)	0.1357 (0.0027)	0.2113 (0.0034)	0.2440 (0.0030)	0.0424 (0.0054)		0.1662
S. cerevisiae	-0.0196 (0.0041)	0.0331 (0.0037)	-0.0408 (0.0021)	0.0740 (0.0014)	0.0321 (0.0021)	0.0365 (0.0020)	0.1195 (0.0037)	0.0001 (0.0058)	

Table S5

6-mer data regressed on 2-mer counts

Species	f		f (1-mer)		f (2-mer)	
	r^2	r^2(CpG)	r^2	r^2(CpG)	r^2	r^2(CpG)
C.elegans	0.43	0.20	0.43	0.18	0.13	0.043
D.melanogaster	0.63	0.33	0.54	0.16	0.030	0.014
D.rerio	0.49	0.35	0.47	0.38	0.13	0.11
G.gallus	0.64	0.47	0.67	0.57	0.11	0.083
M.musculus	0.58	0.48	0.63	0.58	0.047	0.036
H.sapiens	0.54	0.39	0.58	0.52	0.074	0.057
A.thaliana	0.67	0.47	0.64	0.43	0.075	0.049

6-mer data regressed on 3-mer counts

Species	r^2			
	f	f (1-mer)	f (2-mer)	f (3-mer)
C.elegans	0.60	0.56	0.41	0.079
D.melanogaster	0.73	0.66	0.42	0.014
D.rerio	0.62	0.62	0.45	0.066
G.gallus	0.75	0.77	0.42	0.17
M.musculus	0.70	0.70	0.16	0.087
H.sapiens	0.67	0.66	0.32	0.067
A.thaliana	0.74	0.72	0.39	0.16

8-mer data regressed on 2-mer counts

Species	f		f (1-mer)		f (2-mer)	
	r^2	r^2(CpG)	r^2	r^2(CpG)	r^2	r^2(CpG)
C.elegans	0.30	0.14	0.15	0.075	0.054	0.024
D.melanogaster	0.43	0.23	0.34	0.11	0.030	0.017
D.rerio	0.19	0.15	0.18	0.15	0.077	0.067
G.gallus	0.49	0.36	0.55	0.47	0.095	0.082
M.musculus	0.32	0.26	0.34	0.32	0.0044	0.0034
H.sapiens	0.29	0.21	0.29	0.26	0.030	0.025
A.thaliana	0.52	0.38	0.50	0.32	0.046	0.032

8-mer data regressed on 3-mer counts

Species	r^2			
	f	f (1-mer)	f (2-mer)	f (3-mer)
C.elegans	0.43	0.20	0.14	0.033
D.melanogaster	0.50	0.42	0.21	0.016
D.rerio	0.25	0.25	0.17	0.050
G.gallus	0.58	0.65	0.19	0.12
M.musculus	0.39	0.39	0.012	0.011
H.sapiens	0.36	0.33	0.081	0.028
A.thaliana	0.57	0.58	0.20	0.11