"Do all mammals exhibit heterogeneity in their mutation rates? Do all yeasts exhibit uniformity?"

VARYING PATTERNS OF MUTATION

Measuring the Universality of Regional Mutation Rates

ALEAH FOX

MUTATIONS ARE THE ULTIMATE SOURCE OF GENETIC VARIATION IN DNA. THE PATTERNS OF MUTATION, HOWEVER, CAN VARY BOTH WITHIN AND ACROSS GENOMES. IT HAS PREVIOUSLY BEEN SHOWN THAT SEVERAL MAMMALS HAVE HETEROGENEOUS MUTATION RATES, WHILE FOUR YEASTS HAVE BEEN OBSERVED TO HAVE UNIFORM RATES. THE GENERALITY OF THESE OBSERVATIONS HAS NOT BEEN KNOWN. HERE WE EXAMINE SILENT SITE SUBSTITUTIONS IN CODING REGIONS OF 20 MAMMALS, 27 YEAST, AND 4 INSECTS, TO DETERMINE WHICH GENOMES DEMONSTRATE THIS MOSAIC RATE DISTRIBUTION AND WHICH ARE UNIFORM. OUR FINDINGS SHOW THAT MUTATIONAL HETEROGENEITY OCCURS IN ALL BRANCHES OF THE MAMMALIAN PHYLOGENY, AS WELL AS IN FLIES AND MOSQUITOES. ALL YEASTS HAVE A UNIFORM RATE ACROSS THEIR GENOMES WITH THE EXCEPTION OF THREE *CANDIDA* SPECIES: *C. ALBICANS*, *C. DUBLINIENSIS*, AND *C. TROPICALIS*. WE HYPOTHESIZE THAT THIS IS DUE TO THE LACK OF SEXUAL RECOMBINATION IN THESE SPECIES, LEADING TO THE REGIONAL ACCUMULATION OF MUTATIONS.

BACKGROUND

Popular understanding of gene mutations often focuses on those that have obvious effects, such as insects that develop resistance to pesticides. However, the majority of mutations have neither a positive nor a negative impact because they do not affect functional regulatory elements, which generally make up a small percentage of the genome. These mutations are known as "neutral mutations." The presence and rate at which these mutations occur can be used to illuminate evolutionary relationships between different species, as well as to distinguish areas of functional regulatory elements from non-functional DNA, a process called "phylogenetic footprinting." Though neutral mutation rates were once considered to be uniform, it has been discovered that they can vary dramatically, not only from one species to the next but also within a single genome.[i] Within-genome heterogeneity has been demonstrated in several investigations of mammalian species.[ii-vii] In these species, the rate of neutral mutations is different in different regions of the genome, such as that of Chromosome 1 compared to that of Chromosome 4. In contrast, uniform mutation rates have been observed only in the phylogeny of the *sensu stricto* yeasts *S. cerevisiae*, *S. paradoxus*, *S. bayanus*, and *S. mikatae*.[viii] In these species, neutral mutations occur at the same rate in the entire genome. The reason for this differing mutational behavior is not known, and little is known about regional biases in other species. A better characterization of regional biases would improve fundamental understanding of DNA mutation. It would also aid in the calibration of phylogenetic footprinting methods, which are used to detect sequences under purifying selection. For example, uncertainties in regional mutation biases have hindered estimates of the amount of functional human DNA.[ix]

Most previous studies of mutation rate variation have focused on mammals, including human, chimp, mouse, rat, dog, and cow. Such studies have identified regional effects by studying the correlation of independent neutral measures, such as single nucleotide polymorphism (SNP) density, insertion-deletion (indel) density, substitution rate in ancestral repeats, or substitution rates in silent sites.[x] Regional effects have also been characterized via the length scales at which nearby neutral sequences are correlated, as well as via the strength of correlations within single genes.[xi] Proposed causes of mutation rate variation have included regional variations in base composition, recombination, gene density, or pattern of gene expression.[xii] Unfortunately, the causes of variation in these other features are not clearly understood either. Regional effects must be different in yeast since yeast chromosomes are typically only one-hundredth as long as human chromosomes. The length scales of these various types of mammalian regional variation are often as large as a yeast chromosome.

An important step toward understanding the causes of mutational heterogeneity is to measure which species have heterogeneous mutation rates and which species have homogeneous mutation rates. In this work, we analyze regional mutational biases in 20 mammalian, 27 yeast, and 4 insect genomes, determining which clades (taxonomic groups) have uniform and which have heterogeneous mutation rates. We answer the questions: Do all mammals exhibit heterogeneity in their mutation rates? Do all yeasts exhibit uniformity? Studying these phylogenies together provides a valuable contrast.

RESULTS

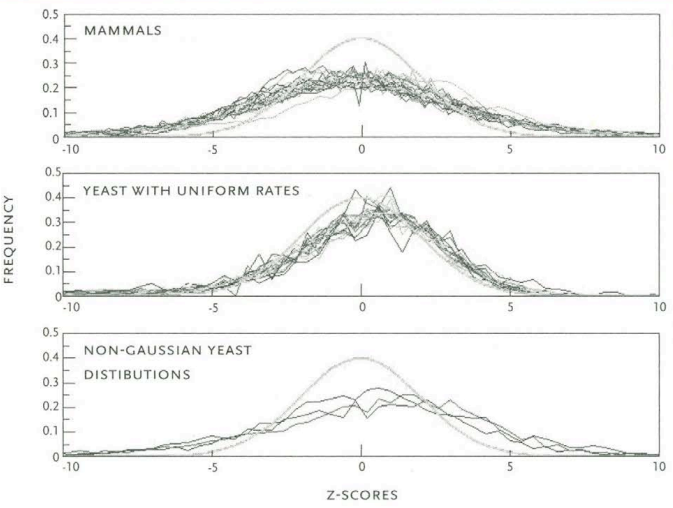
There are many sequences in the genome that can be used to determine the neutral mutation rate, such as pseudogenes, ancestral repeats, intergenic regions and synonymous sites. In this work, we focus on synonymous sites, and in particular, four-fold degenerate sites, from coding regions in 51 species. These sites can be altered without affecting the encoded amino acid sequence. While recent studies indicate that some synonymous sites are under selection, the majority are still likely to be neutral.[xiii] To isolate lineage-specific effects, species were analyzed in pairs with species close to them in the phylogenetic tree. Yeast species pairs were determined from the tree of 42 fungal species of Fitzpatrick et al.[xiv] For the mammals, species pairs were chosen based on recent ENCODE (the

Encyclopedia of DNA Elements Project) and whole-genome-based phylogenies.[xv-xvi] Raw substitution rates were calculated for each orthologous gene pair (pairs of similar genes in different species) by counting the fraction of observed substitutions at four-fold sites and then normalizing to z-score values (see Methods).[xvii] This z-score normalization corrects for the stochastic finite-size effects that result from genes having different numbers of four-fold sites.

YEAST

The distribution of normalized substitution rates provides a test of whether substitution rates are uniform throughout a genome. Our yeast phylogeny is comprised of 27 species, and we have calculated substitution rates between 26 species pairs, based on choosing those which are closely related in the phylogeny. In genomes with uniform rates, such as *S. cerevisiae*, the distribution of z-scores is very close to a normal distribution with a standard deviation of one.[xviii] However, in heterogeneous genomes such as mouse and human, the width of the z-score distribution is considerably larger, due to the tendency of sites in the same gene to be subject to similar mutational pressures (Figure 1a). We find that the distributions of normalized rates for 22 of the 27 yeast species (23 species pairs) fit the normal distribution with unit standard deviation (Average s = 1.32, range = [1.02 – 1.40]). This indicates that there are generally not regions of high or low neutral mutation rates within these genomes (Figure 1b). As in the *sensu stricto* yeasts, these 22 yeast species tend to have a slight excess of genes with low silent substitution rates. This can be explained by codon usage selection. The tail at negative z-scores is mainly comprised of ribosomal and carbohydrate catabolism genes that are under selection for codon usage bias.[xix] Of the 432 genes with z-scores ≤ -2 in the *S. cerevisiae* - *S. bayanus* comparison, 201 (46.5%) have a ribosomal or metabolism GO annotation. Similarly, in the distinct lineage *D. hansenii* - *C. guilliermondii*, 138 of 253 genes (54.5%) with z-score ≤ -2 map to ribosomal or metabolism GO categories. One caveat is that this z-score approach is less applicable for comparisons of species with saturated divergence. Of the full set of 26 species pairs, we observe 8 with less than 90% of the divergence that would be expected at saturation, given the base

FIGURE 1: NORMALIZED DISTRIBUTION OF SUBSTITUTION RATES



THIS IS THE DISTRIBUTION OF NORMALIZED 4-FOLD SUBSTITUTION RATES FOR MAMMALS AND YEASTS. THE NORMAL GAUSSIAN DISTRIBUTION (SMOOTH GRAY LINE) IS WHAT WOULD BE EXPECTED IF ALL 4-FOLD SITES IN EACH GENE HAVE AN EQUAL AND INDEPENDENT PROBABILITY OF BEING SUBSTITUTED.

TOP: *ALL MAMMALIAN DISTRIBUTIONS HAVE A BIAS TOWARD HIGH AND LOW SUBSTITUTION RATES, CONSISTENT WITH REGIONAL MUTATION BIASES IN MAMMALIAN GENOMES.*

MIDDLE: *MOST YEAST DISTRIBUTIONS FIT MORE CLOSELY TO THE GAUSSIAN, EXCEPT FOR A TAIL OF GENES WITH LOW SUBSTITUTION RATES STEMMING FOR CODON USAGE SELECTION.*

BOTTOM: *THIS IS THE DISTRIBUTION OF C. ALBICANS, C. DUBLIENSIS, C. TROPICALIS, N. CRASSA, AND C. GLOBOSUM. THESE SPECIES HAVE A WIDER DISTRIBUTION SIMILAR TO THE MAMMALIAN DISTRIBUTION.*

TABLE 1: PEARSON CORRELATION FOR THE SUBSTITUTION RATE OF NEIGHBORING GENES

SPECIES 1	SPECIES 2	AVERAGE DIVERGENCE	NUMBER OF ORTHOLOGS	PEARSON CORRELATION	P-VAL
S_CEREVISIAE	S_PARADOXUS	0.2592	4434	0.0234	0.1188
S_CEREVISIAE	S_BAYANYUS	0.4709	3781	0.0305	0.0606
S_CASTELLII	S_MIKATAE	0.6585	3247	0.0029	0.8649
K_LACTIS	C_GLABRATA	0.6545	4077	-0.0135	0.3887
D_HANSENII	C_LUSITANIAE	0.6926	3425	0.0417	0.0144
D_HANSENII	C_GUILLIERMONDII	0.6842	3538	0.0221	0.1883
D_HANSENII	L_ELONGISPORUS	0.6628	3345	0.0083	0.6275
D_HANSENII	C_PARAPSILOSIS	0.6561	3324	0.0257	0.1378
D_HANSENII	C_TROPICALIS	0.6104	3358	0.0073	0.6685
E_GOSSYPII	S_POMBE	0.7616	1536	0.0837	0.0010
E_GOSSYPII	Y_LIPOLYTICA	0.7114	2441	0.0297	0.1414
Y_LIPOLYTICA	M_GRISEA	0.6751	2464	0.0410	0.0418
Y_LIPOLYTICA	A_NIDULANS	0.7072	2515	0.0162	0.4156
Y_LIPOLYTICA	C_IMMITIS	0.7109	2584	0.0021	0.9117
Y_LIPOLYTICA	S_POMBE	0.7271	1270	0.0404	0.1495
S_POMBE	S_JAPONICUS	0.6551	1461	0.0549	0.0358
C_DUBLINIENSIS	C_ALBICANS	0.2893	3796	0.2176	6.3 E -42
C_DUBLINIENSIS	C_TROPICALIS	0.5522	3514	0.1029	9.8 E -10
C_TROPICALIS	C_PARAPSILOSIS	0.6059	3390	0.0426	0.0130
C_TROPICALIS	L_ELONGISPORUS	0.6257	3442	-0.0227	0.1818
N_CRASSA	C_GLOBOSUM	0.5842	2369	-0.0085	0.6757
U_REESII	C_IMMITIS	0.5293	917	-0.0513	0.1201
A_NIDULANS	A_TERREUS	0.6364	1121	0.0370	0.2147
H_CAPSULATUM	C_IMMITIS	0.6839	847	0.0356	0.2999
H_CAPSULATUM	U_REESII	0.6655	393	0.0034	0.9459
L_ELONGISPORUS	C_PARAPSILOSIS	0.6363	3152	0.0265	0.1366

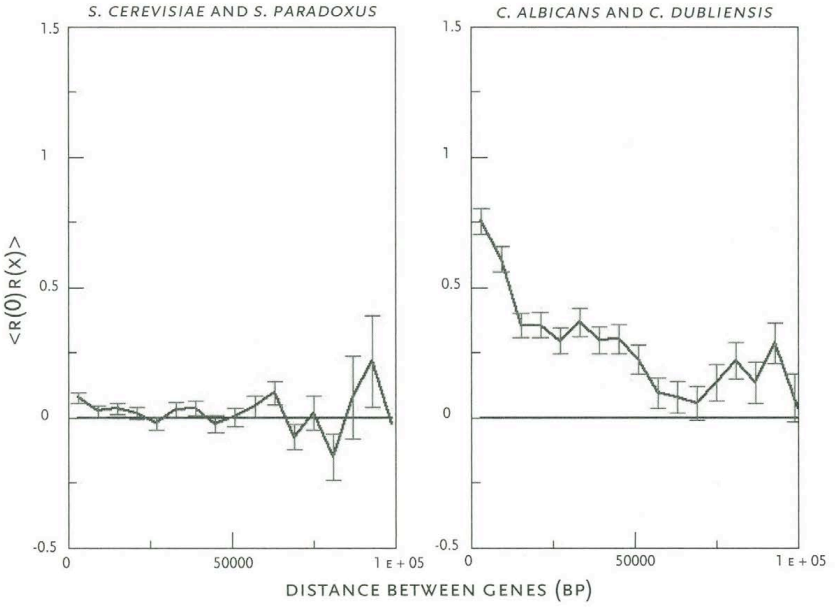
THE DATA COLLECTED WAS FROM 26 SPECIES PAIRS AMONG 27 YEASTS. AVERAGE DIVERGENCE IS THE FRACTION OF ALL ALIGNED 4-FOLD SITES WHICH DIFFER BETWEEN THE TWO SPECIES. PEARSON CORRELATIONS FOR THE NORMALIZED SUBSTITUTION RATES OF NEIGHBORING GENES ARE SHOWN IN COLUMN 5. THE ONLY SPECIES WITH PEARSON CORRELATION WITH SIGNIFICANCE < 0.001 ARE THOSE AMONG *C. DUBLINIENSIS*, *C. ALBICAN*, AND *C. TROPICALIS*.

TABLE 2: MAMMALIAN PAIRWISE COMPARISONS

PAIR NUMBER	SPECIES 1	SPECIES 2	AVERAGE DIVERGENCE	NUMBER ORTHOLOGS	PEARSON CORRELATION	P-VAL
1	HUMAN	MOUSE	0.3317	17112	0.2306	2.0E -205
2	HUMAN	DOG	0.2451	16532	0.3017	9.8E -247
3	HUMAN	CAT	0.2417	7529	0.3263	2.5E -186
4	HUMAN	COW	0.5761	6972	0.0540	6.2E -06
5	DOG	CAT	0.1831	12931	0.2438	1.9E -174
6	CAT	MOUSE	0.3581	6726	0.2292	6.0E -81
7	RAT	MOUSE	0.1586	20054	0.1576	1.0E -111
8	MOUSE	RABBIT	0.3606	12442	0.1851	2.2E -96
9	EURO. HEDGEHOG	TENREC	0.3679	2360	0.1968	4.6E -22
10	EURO. HEDGEHOG	TREE SHREW	0.3322	1650	0.3132	6.7E -39
11	BUSHBABY	OPOSSUM	0.4599	12922	0.2097	2.0E -128
12	COW	DOG	0.2532	16941	0.2606	4.4E -261
13	COW	CAT	0.2530	13355	0.2626	1.5E -209
14	DOG	MOUSE	0.3550	16522	0.2152	2.4E -172
15	COW	MOUSE	0.3646	17340	0.2152	6.5E -181
16	HUMAN	MACAQUE	0.0673	16000	0.2754	2.0E -276
17	BAT	RAT	0.3627	5938	0.1995	2.2E -54
18	MACAQUE	CHIMP	0.0710	17621	0.2231	1.1E -197
19	ARMADILLO	ELEPHANT	0.2527	2045	0.1948	5.9E -19
20	MOUSE	SQUIRREL	0.3304	12794	0.1789	1.4E -92
21	TENREC	ELEPHANT	0.2741	2244	0.1667	1.8E -15
22	COMMON SHREW	TREE SHREW	0.3442	2124	0.2503	1.0E -31
23	DOG	BAT	0.2459	13669	0.3527	< E -276
24	COW	BAT	0.2579	14174	0.3398	< E -276
25	OPOSSUM	PLATYPUS	0.4992	9786	0.2701	3.3E -163

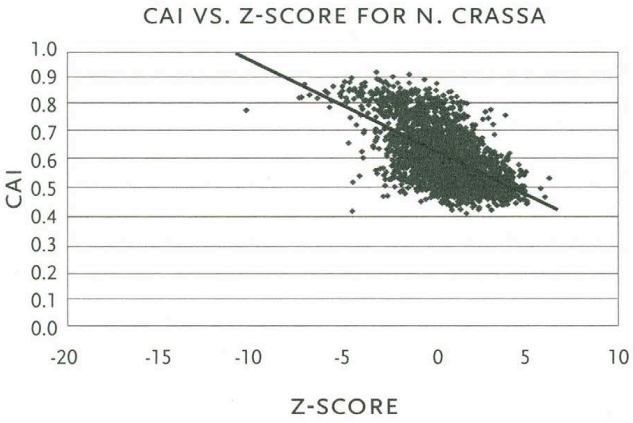
MAMMALIAN PAIRWISE COMPARISONS. 4-FOLD SITE DIVERGENCE BETWEEN THE TWO SPECIES IS SHOWN IN COLUMN 3. PEARSON CORRELATIONS FOR THE NORMALIZED SUBSTITUTION RATES OF NEIGHBORING GENES ARE SHOWN IN COLUMN 5. EVERY MAMMALIAN COMPARISON SHOWS A SIGNIFICANT CORRELATION BETWEEN RATES OF NEIGHBORING GENES (COLUMN 6).

FIGURE 2: AUTOCORRELATION OF NORMALIZED SUBSTITUTION RATES FOR PAIRWISE ANALYSES



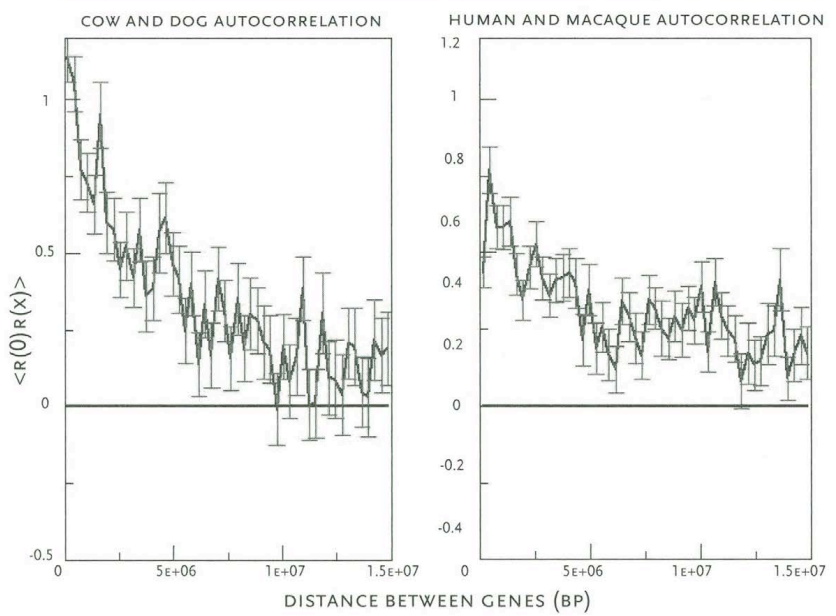
RATES ARE NOT CORRELATED ALONG THE GENOMES IN FIGURE 2A. ANALAGOUS GRAPHS WERE PRODUCED FOR ALL YEAST SPECIES PAIRS, EACH INDICATING NO AUTOCORRELATION ALONG THE GENOME, EXCEPT FOR THE C. DUBLIENSIS/C. ALBICANS *AND* C. TROPICALIS/C. DUBLIENSIS *COMPARISONS. THOSE PAIRS SHOW THAT RATES OF NEIGHBORING GENES ARE CORRELATED UP TO 50 KB APART.*

FIGURE 3: CAI VS. NORMALIZED SUBSTITUTION RATES FOR *SENSU STRICTO* YEASTS AND *N. CRASSA*



THESE ARE GRAPHS OF CAI VS. NORMALIZED SUBSTITUTION RATES FOR SENSU STRICTO YEASTS (S. CEREVISIAE, S. BAYANUS, S. MIKATAE, S. PARADOXUS) *(BOTTOM) AND* N. CRASSA *(TOP). IN EACH GRAPH, THERE IS A LARGE CLUSTER OF GENES WITH LOWER CAI VALUES AND SUBSTITUTION RATES DISTRIBUTED AROUND ZERO. EACH GRAPH ALSO CONTAINS A SECOND GROUP OF GENES WITH HIGHER CAI VALUES AND LOWER SUBSTITUTION RATES, AND SUCH GENES ARE LIKELY TO BE UNDER CODON USAGE SELECTION. THE* SENSU STRICTO *YEASTS HAVE 121 GENES THAT HAVE CAI VALUES GREATER THAN 0.40 WHILE* N. CRASSA *HAS 949 WITH CAI VALUES GREATER THAN 0.70, SUGGESTING THAT MORE GENES ARE UNDER CODON USAGE SELECTION IN* N. CRASSA. *THE* N. CRASSA *CODON USAGE TABLE WAS USED TO CALCULATE CAI VALUES IN THE TOP GRAPH AND THE* S. CEREVISIAE *CODON USAGE TABLE WAS USED TO COMPUTE THE CAI VALUES IN THE BOTTOM GRAPH.*

FIGURE 4: AUTOCORRELATION OF NORMALIZED SUBSTITUTION RATES AT 4-FOLD DEGENERATE SITES



AUTOCORRELATION FOR NORMALIZED SUBSTITUTION RATES AT 4-FOLD DEGENERATE SITES BETWEEN COW AND DOG (LEFT) AND HUMAN AND MACAQUE (RIGHT). BOTH GRAPHS SHOW CORRELATION OF RATES FOR GENES WITHIN 10Mb OF EACH OTHER.

CAI VS. Z-SCORE FOR SENSU STRICTO YEAST

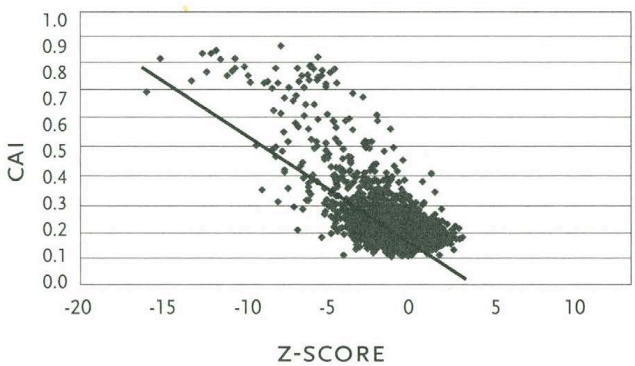
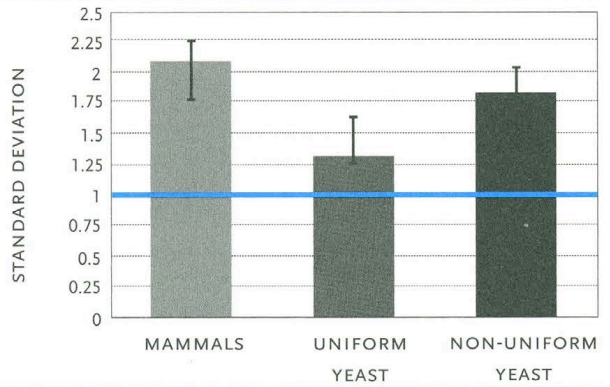


FIGURE 5: COMPARISON OF THE WIDTHS OF THE RATE DISTRIBUTIONS



THE BLUE LINE AT s = 1 SHOWS THE EXPECTED STANDARD DEVIATION FOR A NORMAL DISTRIBUTION.

THE MAMMALS HAVE AN AVERAGE STANDARD DEVIATION OF 2.056, INDICATING THEIR BIAS TOWARD BOTH HIGH AND LOW RATES.

THERE ARE THREE YEAST PAIRWISE COMPARISONS (C. DUBLINIENSIS/C. ALBICANS, C. TROPICALIS/C. DUBLINIENSIS, AND N. CRASSA/C. GLOBOSUM) WHICH HAVE UNUSUALLY WIDE RATE DISTRIBUTIONS (SEE FIGURE 1C). FOR THESE COMPARISONS, THE AVERAGE s IS 1.825. THE REMAINING 22 YEASTS HAVE A NOTICEABLY LOWER VALUE OF s (AVERAGE s=1.32, RANGE = [1.02 – 1.40], CONSISTENT WITH THEIR HAVING MORE UNIFORMLY MUTATING GENOMES.

compositions of these species (see Methods).

The inference that there is no neutral mutational variation within these 22 yeast genomes is further supported by a neighboring gene analysis. For each pair of species, the Pearson correlation was calculated for the rates of substitution of neighboring genes (see Methods). These results are shown in Table 1, with the 22 homogeneous genomes shown in black. There was no significant correlation for any of the species pairs made up of two members from these 22 genomes. The same results were obtained irrespective of which species in the pair was used to specify gene locations.

More generally, an autocorrelation function $\langle r^{16} r(0) \rangle$ was computed, where r(0) is the normalized substitution rate of a gene and r^{16} is the normalized substitution rate of a gene that is x base pairs downstream of the first gene.[xx] The rates are normalized around r = 0 so we would expect $\langle r(0) r^{16} \rangle \sim 0$ if there are no regional rate biases (see Methods). The autocorrelation reveals no significant correlation between neighboring genes at any finite separation (Figure 2a). This shows that these yeast species do not demonstrate the heterogeneous mutation rates that have been observed in mammals.

Species pairs from two subclades [(*C. albicans, C. dubliniensis, C. tropicalis*) and (*N. crassa, C. globosum*)] were found to have distributions that did not fit the normal curve. These species have substitution rate distributions biased toward high and low rates ($2.07 \leq s \leq 2.20$) (Figure 1c). This phenomenon mirrors that which has been observed for human-mouse substitution rates (see also Figure 1a).[xxi]

For the three *Candida* species, the autocorrelation analysis shows that substitution rates are significantly correlated for genes within 50,000 base pairs of each other (Figure 2b). This result implies that regional biases extend over scales encompassing over 20 genes, since in *C. albicans* the spacing between genes is 2300 bp.[xxii] The neighboring gene Pearson correlations were also significant (*C. dubliniensis/C. albicans* Pearson correlation = 0.218, $p = 6.4 \times 10^{-42}$; C. tropicalis/C. dubliniensis Pearson correlation = 0.103, $p = 9.8 \times 10\text{-}10$). SNPs within the *C. albicans* genome have been shown to be unevenly distributed along contigs,[xxiii] in agreement with regional mutational biases. These regional effects are not due to a CpG dinucleotide effect. When we excluded CpG sites from the analysis, neighboring genes were still observed to have correlated substitution rates (Pearson-correlation = 0.1096, $p < 10^{-8}$). These *Candida* species translate CUG as serine instead of the usual leucine, and it might be hypothesized that this is relevant to the rate inferences.[xxiv] However, ignoring these codons does not significantly diminish the correlation (*C.*

dubliniensis/*C. albicans* Pearson correlation = 0.208, $p = 10^{-27}$).

Like the *Candida* species, the species *N. crassa* and *C. globosum* have a rate distribution that is wider than the normal Gaussian. However, there is no significant rate correlation (Pearson correlation = -0.0085, p = 0.6757) between neighboring genes, and the autocorrelation plot appears nearly identical to those of the yeast with uniform rates. These seemingly contradictory results suggest that the wide distribution of rates ($s = 2.07$) is due to pressures on individual genes, rather than regional effects.

We hypothesized that the large *s* without apparent regional correlation could be related to increased selection on the silent sites of *N. crassa* and *C. globosum* genes. Such selection could cause some genes to have more extreme conservation, broadening the rate distribution. To test this, we considered the codon usage bias in these species, which is the best understood type of silent site selection in yeasts. Our results indicate that codon usage selection is a stronger effect in *N. crassa* and *C. globosum* genomes than in *sensu stricto* yeasts. CAI values were calculated for each of *N. crassa* and *S. cerevisiae* based on their respective codon biases (Figure 3). *N. crassa* genes generally have higher CAI values (median 0.63) than *S. cerevisiae* genes (median 0.14). In each of the genomes, there is one main cluster of genes having z-scores distributed around zero and low CAI values. Then there is another group having negative z-scores and high CAI, and this group is presumably under codon usage selection. *N. crassa* appears to have more genes in the group under codon usage selection (949 genes above CAI = 0.7) than *S. cerevisiae* (121 genes above CAI = 0.4). This suggests that codon usage selection affects more genes in *N. crassa*/*C. globosum*, and could be responsible for the large *s* in these genomes.

MAMMALS

All 20 mammalian species appear to have heterogeneous mutation patterns, as evidenced by their wide rate distributions ($1.80 \leq s \leq 2.23$) (Figures 1a and 5). The Pearson correlations (Table 2) and autocorrelations of nearby genes (Figure 4) for each pair-wise mammalian comparison are all significant. For example, in the cow/human comparison, the autocorrelation graph suggests mutational blocks along each chromosome as large as 10Mb, similar to the length scale that has previously been observed in mouse and human.

The regional variations are also apparent from correlation analysis of ancestral repeats, another largely neutral scattered throughout the genome.[xxv] We analyzed the 18-way vertebrate multi-species alignments from UCSC to obtain

TABLE 3: CORRELATION OF SUBSTITUTIONAL RATES IN NEIGHBORING ANCIENT REPEATS

SPECIES 1	SPECIES 2	NUMBER OF REPEATS	PEARSON CORRELATION	P-VAL
HUMAN	MACAQUE	5081426	0.32192	< E -276
DOG	COW	1378699	0.18829	< E -276
ELEPHANT	TENREC	248577	0.12158	< E -276
RAT	MOUSE	231283	0.1292	< E -276

IN ALL FOUR MAMMALIAN PAIRS CONSIDERED (HUMAN-MACAQUE), (DOG-COW), (ELEPHANT-TENREC), AND (RAT-MOUSE), THE CORRELATIONS WERE EXTREMELY SIGNIFICANT ($<10^{-276}$). THIS SUPPORTS THE WIDESPREAD HETEROGENEITY OF MUTATION RATES IN MAMMALIAN GENOMES.

a stringent set of ancestral repeats aligned orthologously across the species pairs (human, macaque), (mouse, rat), (elephant, tenrec), and (dog, cow). We observed that an ancestral repeat's substitution rate is significantly correlated with that of the neighboring ancestral repeat along the genome, for each of the species pairs (Table 3). In each case the significance of the correlation was at the limit of computational precision ($p < 10^{-276}$). The wide distribution of ancestral repeat normalized rates mimics those in Figures 1a and 1c, with standard deviations ranging from 1.41 (elephant-tenrec) to 1.96 (human-macaque).

INSECTS

The yeast species we have studied are at phylogenetic distances that are generally larger than those for the mammalian species. In principle, it is more difficult to measure regional variations when inferring rates from more distantly related species. This is because as species approach saturated divergence, all mutation rate inferences become increasingly uncertain. Therefore one might be concerned that the greater divergence among the yeast species obscures regional effects.

This hypothesis can be tested by using insect genomes, several of which are at phylogenetic divergences as large as that of the yeasts. In particular, *D. melanogaster* and *D. pseudoobscura* are at a separation generally larger than that of the mammalian species (four-fold divergence = 0.514) and comparable to that of most of the yeast pairs, as are the two mosquitoes *A. aegypti* and *A. gambiae* (4-fold divergence = 0.632).

In contrast to the yeasts, flies and mosquitoes both show clear evidence of regional effects. *D. melanogaster* and *D. pseudoobscura* have neighboring genes with significant rate correlations at distances up to ~1Mb apart (Pearson correlation = 0.1642, p = 9.0e-59) and a wide rate distribution (s = 2.152). This effect is also observed in *A. aegypti* and *A. gambiae* (Pearson correlation = 0.2099, p = 2.1e-89). The width of the distribution of normalized rates is s = 1.804. Each of these insect lineages demonstrate correlations which are more statistically significant than the most significant yeast comparison, *C. albicans* and *C. dubliniensis*, despite the fact that the two *Candida* species have a lower divergence (0.289) than the insects. Therefore, we can conclude that the predominant lack of correlations for the yeasts is not an artifact of their larger phylogenetic separations.

CONCLUSIONS AND DISCUSSION

Based on our examination of 20 mammalian species, we conclude that all mammals have regional biases of neutral mutation rates. While the factors controlling these regional biases are still not well understood (e.g. base composition, local recombination rate, pattern of gene expression, gene density and DNA repair domain),[xxvi] our findings indicate that any valid explanations must occur throughout the mammalian phylogeny. In contrast, we find that almost all yeasts have a uniform neutral mutation rate. This conclusion is supported by a recent report that *S. cerevisiae* polymorphism rates are not correlated along the genome.[xxvii]

The monophyletic group of the three *Candida* species (*C. albicans*, *C. dubliniensis*, and *C. tropicalis*) is an exception to the other yeasts. Both the distribution of rates and gene-to-gene correlations indicate that these species have heterogeneous mutation rates, a trait which had not previously been observed outside the mammals. Previous studies of SNP data also indicate hotspots of polymorphism, supporting the concept of regional biases.[xxviii]

What characteristic sets these three yeasts apart from the others? One intriguing trait is meiosis. Unlike other yeasts, sexual reproduction appears to be rare in *C. albicans*, *C. dubliniensis*, and *C. tropicalis*. Sexual reproduction in *C. albicans* and *C. dubliniensis* was recently discovered under specialized laboratory conditions; however, no evidence for meiosis has been found.[xxix] The sexual cycle of *C. tropicalis* has not been extensively explored, but meiosis has not been observed for it either. On the other hand, comparative genomic studies of the uniformly mutating *C. glabrata* support a complete sexual cycle,[xxx] and so-called "defects" in

the mating type locus of *C. parapsiolosis* suggest that it is unlikely to have mating similar to *C. albicans* and *C. dubliniensis*.[xxxi] Thus, the *Candida* species without evidence for meiosis are the ones with heterogeneous neutral mutation rates. How could a lack of meiosis influence regional mutation rates? One possible explanation is that the lack of meiosis prevents recombination between homologous chromosomes. If recombination were to cause random fluctuations in chromosome length, this could have the effect of smoothing out regional mutational biases.

A comparison of *N. crassa* and *C. globosum* also fails to give a normal distribution of substitution rate z-scores, but this is due to stronger codon usage selection, rather than regional mutation effects. This stronger codon usage selection is more apparent in the context of a mutational defense mechanism that *N. crassa* employs against selfish DNA. The mechanism, known as repeat-induced point mutation (RIP), protects against selfish DNA by inducing G:C to A:T mutations during sexual reproduction and by methylating cytosine residues in duplicated sequences.[xxxii] This process affects all duplicated regions except for ribosomal RNA genes, even though these occur in large copy number.[xxxiii] This observation is consistent with ribosomal genes being under strong codon usage selection and hence being more conserved at synonymous sites. RIP drives other duplicated genes in *N. crassa* to unusually high substitution rates, which would explain the genes with high scores in the rate distribution. The wide substitution rate distribution is likely due to RIP in just *N. crassa*, as the phenomenon has not been observed in *C. globosum*.

METHODS

ORTHOLOG GENERATION

For the yeast analysis, FASTA files of coding regions and an ortholog tree listing predicted gene relationships between all 32 yeast species were obtained as described in Tsong et al.[xxxiv] Mammalian and mosquito genes were obtained using the ENSEMBL BioMart (release 45) database. Fly CDS and peptide files were downloaded from FlyBase (version FB2006_01). For the fly species, the amino acid sequences were run through BLAST and tagged as true orthologs if they were each other's mutual best hit when applying BLASTALL with a worst-case E-value cutoff of 10^{-10}.

CALCULATION OF SUBSTITUTION RATES

Our substitution rate calculations parallel those in previous works.[xxxv-xxxvi] Nucleotide sequences of orthologous coding regions were translated into amino acid residues, aligned using CLUSTALW, and then back-translated to determine the aligned DNA sequence. The four-fold synonymous sites—the third base in a codon for which the amino acid is determined by the first two positions—were analyzed. If a sequence contained fewer than 20 four-fold sites before occurrence of a stop codon, the entire sequence was discarded. To ensure equivalent sequence context, only four-fold sites for which both the preceding and succeeding base matched for the two species were considered.

The raw neutral substitution rate was calculated based on the fraction of observed differences at silent sites within a gene. Individual gene rates were then normalized in order to correct for the finite-size of each gene (Table 1 & 2), and this new rate was defined to be

$$r = (p - \langle p \rangle)/s(N),$$

where p is the observed four-fold substitution rate for the gene in question and <p> is the average substitution rate for all ortholog pairs for the two species in question. $s(N)$ was defined to be the expected standard deviation for a gene with N independent four-fold sites, i.e. $s(N) = (\langle p \rangle(1-\langle p \rangle)/N)^{1/2}$. The distribution of the normalized substitution rates would be expected to follow a normal Gaussian ($f(x) = e^{x*x/2} / \sqrt{(2\pi)}$) if each four-fold site was mutating at the same rate and independently of each other.

CORRELATION CALCULATIONS

We tested whether or not neighboring genes had similar substitution rates by calculating a Pearson correlation between the rate of gene $r(0)$ and the rate of gene r^{16} which is located x base pairs downstream. Gene pairs in orthologous blocks up to the 35th gene downstream from the starting gene were considered. Blocks were determined by

genes located on the same chromosome (scaffolds were used when chromosomal data was not available). Correlations were measured twice, in each case using location data from one of the species, except in cases where location data was available in only one species. For yeast, the data for each pairwise calculation was binned into 50 uniformly spaced groups covering x = {0, 300000} and averaged over each bin to determine the autocorrelation function $\langle r(0)r^{16} \rangle$. Error bars were assigned based on the standard deviation of the values in each bin. For the larger genomes of mammals and insects, data was binned into 200 groups where x ranged from 0 to 15Mb.

CAI

CodonW (Peden 1999) was downloaded and used to calculate the CAI values for the yeast species. The input file for each species was a CDS FASTA file of all genes (predicted and known) and the background CAI was set to *Saccharomyces cerevisiae* for all *sensu stricto*. The EMBOSS package was also downloaded locally. This includes codon usage tables for a number of species including *N. crassa*.[xxxvii] This table was used to calculate the CAI for the genes in *N. crassa*.

ENDNOTES

i. Baer 2007
ii. Chuang 2004
iii. Lercher 2001
vi. Liu 2006
v. Malcom 2003
vi. Matassi 1999
vii. Wolfe 1989
viii. Chin 2005
ix. Pheasant 2007
x. Hardison 2003
xi. Chuang 2004
xii. Baer 2007
xiii. Chamary 2006
xiv. Fitzpatrick 2006
xv. Murphy 2007
xvi. Nikolaev 2007
xvii. Chuang 2004
xviii. Chin 2005
xix. Ibid.
xx. Ibid.
xxi. Chuang 2004
xxii. MI 2007
xxiii. Lott 2005
xxiv. Yokogawa 1992
xxv. Lunter 2006
xxvi. Baer 2007
xxvii. Ruderfe 2006
xxviii. Jones 2004
xxix. Pujol 2004
xxx. Wong 2003
xxxi. Logue 2005
xxxii. Selke 1990
xxxiii. Vyas 2005
xxxiv. Tsong 2006
xxxv. Chuang 2004
xxxvi. Chin 2005
xxxvii. Ikemura 1985

REFERENCES

Baer, C. F., Miyamoto, M. M. & Denver, D. R. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat Rev Genet 8, 619-31 (2007).

Chamary, J. V., Parmley, J. L. & Hurst, L. D. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7, 98-108 (2006).

Chin, C. S., Chuang, J. H. & Li, H. Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. Genome Res 15, 205-13 (2005).

Chuang, J. H. & Li, H. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. PLoS Biol 2, E29 (2004).

Fitzpatrick, D., Logue, M., Stajich, J. & Butler, G. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evolutionary Biology 6, 99 (2006).

Hardison, R. C. et al. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res 13, 13-26 (2003).

Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2, 13-34 (1985).

Jones, T. et al. The diploid genome sequence of Candida albicans. Proceedings of the National Academy of Sciences 101, 7329-7334 (2004).

Lercher, M. J., Williams, E. J. & Hurst, L. D. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for

understanding the mechanistic basis of the male mutation bias. Mol Biol Evol 18, 2032-9 (2001).

Liu, G. E., Matukumalli, L. K., Sonstegard, T. S., Shade, L. L. & Van Tassell, C. P. Genomic divergences among cattle, dog and human estimated from large-scale alignments of genomic sequences. BMC Genomics 7, 140 (2006).

Logue, M. E., Wong, S., Wolfe, K. H. & Butler, G. A genome sequence survey shows that the pathogenic yeast Candida parapsilosis has a defective MTLa1 allele at its mating type locus. Eukaryot Cell 4, 1009-17 (2005).

Lott, T. J., Fundyga, R. E., Kuykendall, R. J. & Arnold, J. The human commensal yeast, Candida albicans, has an ancient origin. Fungal Genet Biol 42, 444-51 (2005).

Lunter, G., Ponting, C. P. & Hein, J. Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput Biol 2, e5 (2006).

Malcom, C. M., Wyckoff, G. J. & Lahn, B. T. Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. Mol Biol Evol 20, 1633-41 (2003).

Matassi, G., Sharp, P. M. & Gautier, C. Chromosomal location effects on gene sequence evolution in mammals. Curr Biol 9, 786-91 (1999).

MIT, B. I. o. H. a. (2007).

Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S. & Miller, W. Using genomic data to unravel the root of the placental mammal phylogeny. Genome Res 17, 413-21 (2007).

Nikolaev, S. et al. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. PLoS Genet 3, e2 (2007).

Pheasant, M. & Mattick, J. S. Raising the estimate of functional human sequences. Genome Res. 17, 1245-1253 (2007).

Pujol, C. et al. The closely related species Candida albicans and Candida dubliniensis can mate. Eukaryot Cell 3, 1015-27 (2004).

Ruderfer, D. M., Pratt, S. C., Seidel, H. S. & Kruglyak, L. Population genomic analysis of outcrossing and recombination in yeast. Nat Genet 38, 1077-81 (2006).

Selker, E. U. Premeiotic instability of repeated sequences in Neurospora crassa. Annu Rev Genet 24, 579-613 (1990).

Tsong, A. E., Tuch, B. B., Li, H. & Johnson, A. D. Evolution of alternative transcriptional circuits with identical logic. Nature 443, 415 (2006).

Vyas, M. & Kasbekar, D. P. Collateral damage: spread of repeat-induced point mutation from a duplicated DNA sequence into an adjoining single-copy gene in Neurospora crassa. J Biosci 30, 15-20 (2005).

Wolfe, K. H., Sharp, P. M. & Li, W. H. Mutation rates differ among regions of the mammalian genome. Nature 337, 283-5 (1989).

Wong, S., Fares, M. A., Zimmerman, W., Butler, G. & Wolfe, K.H. Evidence from comparative genomics for a complete sexual cycle in the 'asexual' pathogenic yeast Candida glabrata. Genome Biol 4, R10 (2003).

Yokogawa, T. et al. Serine tRNA complementary to the nonuniversal serine codon CUG in Candida cylindracea: evolutionary implications. Proc Natl Acad Sci U S A 89, 7408-11 (1992).