Current Biology *21*, 1051–1054, June 21, 2011 ©2011 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2011.05.013

Report

Pervasive Multinucleotide Mutational Events in Eukaryotes

Daniel R. Schrider,[1,2] Jonathan N. Hourmozdi,[1]
and Matthew W. Hahn[1,2,*]
[1]Department of Biology
[2]School of Informatics and Computing
Indiana University Bloomington, Bloomington, IN 47405, USA

Summary

Many aspects of mutational processes are nonrandom, from the preponderance of transitions relative to transversions to the higher rate of mutation at CpG dinucleotides [1]. However, it is still often assumed that single-nucleotide mutations are independent of one another, each being caused by separate mutational events. The occurrence of multiple, closely spaced substitutions appears to violate assumptions of independence and is often interpreted as evidence for the action of adaptive natural selection [2, 3], balancing selection [4], or compensatory evolution [5, 6]. Here we provide evidence of a frequent, widespread multinucleotide mutational process active throughout eukaryotes. Genomic data from mutation-accumulation experiments, parent-offspring trios, and human polymorphisms all show that simultaneous nucleotide substitutions occur within short stretches of DNA. Regardless of species, such multinucleotide mutations (MNMs) consistently comprise ∼3% of the total number of nucleotide substitutions. These results imply that previous adaptive interpretations of multiple, closely spaced substitutions may have been unwarranted and that MNMs must be considered when interpreting sequence data.

Results and Discussion

A widely held assumption in the analysis of DNA sequences is that substitutions are independently Poisson distributed in time and space (but see [7]). This independence assumption is made despite the fact that adjacent nucleotide substitutions are found more often than expected as both polymorphisms [8–10] and fixed differences between species [5, 11]. An excess of such clustered mutations is often explained—even in polymorphism data—by the rapid emergence of separate mutations [9], the presence of mutational hot spots [8], or a series of independent substitutions that only appear simultaneous in phylogenetic studies [5, 7]. Indeed, a common interpretation of two nearby substitutions found on the same haplotype (such as those within a single codon) is that the initial, slightly deleterious substitution is compensated by the adaptive fixation of a second substitution [5, 6, 9]. Finding more than two closely spaced substitutions is often interpreted as evidence for the repeated fixation of adaptive alleles [2, 3] or balancing selection [4]. However, if there is a mutational process that can introduce multiple mutations to the same haplotype in a single generation (or a small number of generations), then natural selection need not be invoked. Although there is previous evidence for multinucleotide mutations (MNMs) in viruses, bacteria, yeast, and multicellular eukaryotic cell lines [12, 13], it is not clear whether a similar phenomenon occurs in the germline of multicellular eukaryotes. In order for these MNMs to play an important role in the evolution of multicellular eukaryotes, they must occur in the germline at an appreciable frequency. Below, we provide several lines of evidence that MNMs do occur at a high rate in the germline, accounting for ∼3% of de novo mutations across eukaryotes, and are therefore heritable and available as raw material for evolution.

We first examined the frequency of MNMs in previously published mutation-accumulation (MA) experiments from *Saccharomyces cerevisiae* [14], *Caenorhabditis elegans* [15, 16], *Arabidopsis thaliana* [17], and *Drosophila melanogaster* [18]. MA experiments reduce the efficacy of natural selection, thereby revealing the near-complete spectrum of mutation. Under a Poisson mutational model (see Supplemental Experimental Procedures), no closely spaced mutations are expected in any of these experiments (defined as at least two nucleotide substitutions within 20 bp of one another in the same MA line; results are also significant using windows of 10, 50, or 100 bp). Therefore, any closely spaced mutations are likely the result of MNMs. Examining nucleotide substitutions in sequenced MA line genomes that were validated by Sanger sequencing, we found at least one MNM in each organism, most often including only two substitutions but sometimes including three substitutions (Table 1). Across experiments, MNMs comprise between 2% and 16% of the total number of nucleotide substitutions (average across studies = 3.39%). Although multiple neighboring mutations in MA experiments do not necessarily have to arise within a single generation, they must have occurred only a few hundred (or in the case of *A. thaliana*, <30) generations apart.

To examine the frequency of MNMs in humans, we considered data from parent-offspring trios. A recent literature review of data from trios found many examples of multinucleotide mutational events but was not able to quantify their frequency [19]; the frequency of such events is important in understanding their relevance for evolutionary studies. Sequence data from trios consisting of unaffected parents and offspring affected by dominant disease mutations have been used previously to obtain a quantitative estimate of the per-nucleotide mutation rate [20, 21]. We used trio data on mutations resulting in premature stop codons in 44 autosomal genes (collected in [21]) to count the number of single-generation mutational events that involved multiple nucleotides. Although there are biases inherent in estimating the fraction of mutations in such studies that are due to MNMs, in general agreement with the data from MA lines, we found that multinucleotide events comprised 6.9% of base substitutions. There are several reasons why this number may deviate from the true rate of multinucleotide mutation (see Supplemental Experimental Procedures). Therefore, to get an unbiased view of human MNMs, we used whole-genome sequences of two trios from the 1000 Genomes Project Consortium [22] to find multinucleotide events. Using stringent criteria for base quality and coverage, 2.11% and 3.23% of all de novo nucleotide

*Correspondence: mwh@indiana.edu

Table 1. Multinucleotide Mutations in Mutation-Accumulation Lines

Species	Chromosome	Positions	Line	Reference
S. cerevisiae	14	688148[a], 688149[a], 688150[a]	C5	[14]
C. elegans	III	1933779[a], 1933788[a]	77	[15]
	V	18914852, 18914873	B526	[16]
	I	11042658, 11042669, 11042691	B529	[16]
	IV	1201160, 1201163	B538	[16]
	V	14433734, 14433737	B574	[16]
A. thaliana	4	13514562[a], 13514563[a]	119	[17]
D. melanogaster	3L	22741983, 22742032	M138	[18]
	3R	27545050, 27545069[a]	M138	[18]
	X	11668883[a], 11668884[a]	M126	[18]
	X	20669767, 20669802[a]	M158	[18]

[a] Validated by Sanger sequencing. None of the mutations listed were found to be false.

mutations were MNMs in the CEU (European) and YRI (Yoruban) trios, respectively (see Supplemental Experimental Procedures). No such clusters of substitutions would be expected if all mutations were independent ($p < 0.0005$ in each trio). Varying stringency thresholds always resulted in 1%–4% of de novo mutations being MNMs (Supplemental Experimental Procedures). The data from diseased and healthy trios are in quantitative agreement with those from the MA lines and published phylogenetic studies [11], suggesting that a similar mechanism is responsible and demonstrating that this mechanism can act within a single generation.

The results from MA lines and trios provide evidence that multinucleotide mutational events occur, but they do not tell us whether they represent a substantial proportion of variation within species. To determine whether MNMs are found within human populations, we first looked for pairs of nearby single-nucleotide polymorphisms (SNPs) within the Illumina-sequenced genome of a Han Chinese individual, referred to as YH01 [23]. This individual was used because of the high read depth and high quality of the sequence, both of which are necessary for accurate identification of polymorphisms. We independently called 1,665,824 high-confidence heterozygous SNPs and 975,211 homozygous SNPs that differ from the NCBI 36 reference genome (see Supplemental Experimental Procedures). After phasing different haplotypes using reads overlapping both polymorphic sites, we found 51,557 heterozygous pairs of SNPs that were within 20 bp of one another and that did not exhibit an intermediate (recombinant) haplotype in the NCBI reference genome (Figure 1A). We also found 33,266 homozygous pairs of SNPs where YH01 contains two sites that both differ from the reference genome (Figure 1B). We refer to such groups of polymorphisms with only two observed haplotypes as multinucleotide polymorphisms (MNPs; cf. [24]).

The proportion of MNPs due to multinucleotide mutational events can be inferred by polarizing these substitutions using the chimpanzee and orangutan genomes as outgroups. MNPs due to MNMs will have both substitutions on one haplotypic lineage (Figure 2A), whereas those due to separate single-nucleotide events can have substitutions on both lineages (Figure 2B). Under the assumption that mutations occur independently, the expected proportion of cases with both mutations on the same lineage is 50%. We were able to confidently infer the ancestral state for 71,019 MNPs (see Supplemental Experimental Procedures) and found that the majority (48,235, or 67.92%) represented mutations occurring on the same branch (Figure 2). Thus, we observe 25,451 more pairs of substitutions occurring within 20 bp of each other than expected; this is a highly significant excess ($\chi^2 = 9120.85$, $p < 2.2 \times 10^{-16}$), suggesting that these substitutions are due to MNMs. Notably, these 25,451 MNPs due to MNMs account for 1.93% of all nucleotide polymorphisms called in the YH01 genome using the same data and quality standards. The similarity in the proportion of MNMs found in the trio data and the polymorphism data strongly suggests a similar molecular mechanism—one that acts within a single generation. This fraction of all nucleotide polymorphisms due to MNMs is also highly similar to the excess of adjacent SNPs of the same frequency identified previously (~0.89%; [10]). Examining the genomic locations of all 48,235 possible MNMs in YH01 (i.e., both substitutions are on one lineage), we also found that they are at frequencies comparable to SNPs within exons, introns, and intergenic sequences (see Table S1 available online). This implies that MNMs are not on average subject to significantly stronger natural selection than SNPs. The fact that the majority of human MNMs occur in nonfunctional regions also excludes selective explanations, such as compensatory evolution, for their appearance.

In addition to describing the frequency of MNMs, our results also suggest that the mechanism (or mechanisms) responsible

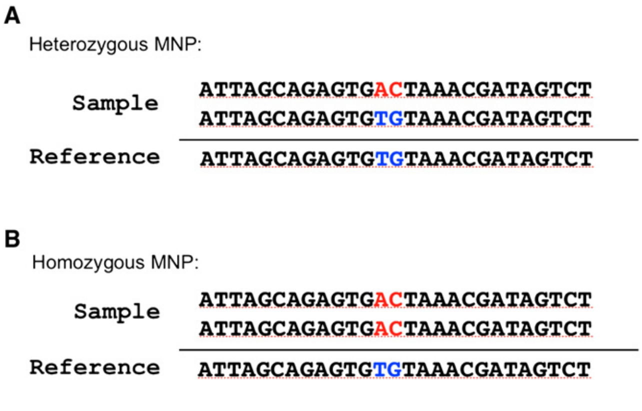
A
Heterozygous MNP:
Sample
ATTAGCAGAGTGACTAAACGATAGTCT
ATTAGCAGAGTGTGTAAACGATAGTCT
Reference
ATTAGCAGAGTGTGTAAACGATAGTCT
B
Homozygous MNP:
Sample
ATTAGCAGAGTGACTAAACGATAGTCT
ATTAGCAGAGTGACTAAACGATAGTCT
Reference
ATTAGCAGAGTGTGTAAACGATAGTCT

Figure 1. Definition of Multinucleotide Polymorphisms

(A) An example of a heterozygous multinucleotide polymorphism (MNP). In this case, a pair of single-nucleotide polymorphisms must be present (but do not have to be adjacent), with one haplotype exactly matching the reference genome and one differing at both positions.

(B) An example of a homozygous MNP, where the two haplotypes differ from the reference genome.

See also Figure S2.

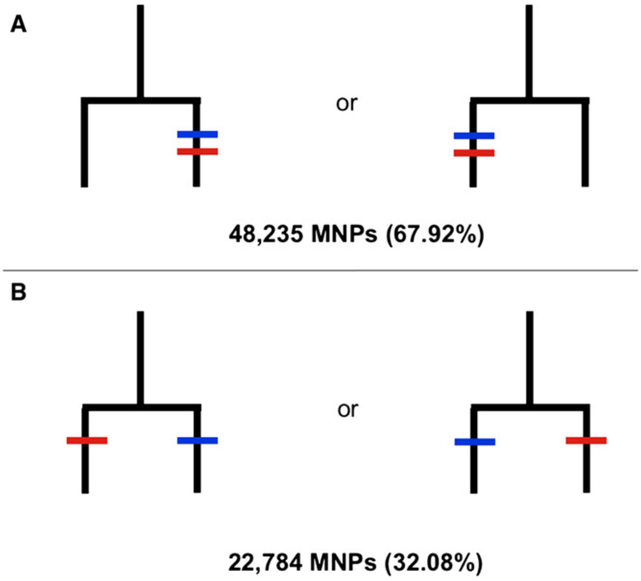

Figure 2. Multinucleotide Mutations Result in an Overrepresentation of Substitutions on the Same Lineage

(A) Two ways that two substitutions can occur on the same haplotypic lineage. The red and blue lines represent individual mutational events (e.g., A→T or C→G). The number of phased and polarized pairs of closely spaced substitutions in the Illumina-sequenced genome YH01 occurring in this manner is shown below the diagrams.

(B) Two ways that two substitutions can occur on different lineages, with the number of such pairs of substitutions in YH01 shown beneath the diagrams. If all substitutions are independent, the number of pairs occurring on the same lineage (A) should equal the number of pairs occurring on different lineages (B).

See also Tables S1 and S2.

for them acts in a highly local manner. An examination of the physical distance between the substitutions contained within the 48,235 possible MNMs reveals that the most common event comprises substitutions in two adjacent positions (Figure 3; Figure S1). In fact, 50% of all possible MNMs in this data set fall within four bases of one another, though there are significantly more pairs of substitutions on the same lineage at all distances up to 20 bp ($p < 6 \times 10^{-5}$ for each comparison at each distance). Of all pairs of single-nucleotide changes within 20 bp of each other, 16.8% are likely due to a multinucleotide mutational event.

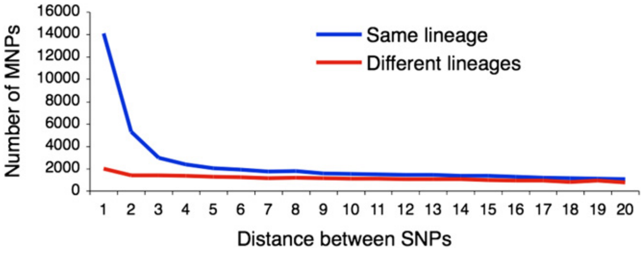

Figure 3. Multinucleotide Mutations Are More Likely to Be Close Neighbors

The distribution of nucleotide distances between phased and polarized pairs of substitutions in YH01 that occurred on the same haplotypic lineage (blue line) and on different lineages (red line). Multinucleotide mutations must have occurred on the same lineage. Adjacent substitutions are indicated by a distance of 1 bp, substitutions with one intervening base are indicated by a distance of 2 bp, etc. See also Figure S1.

MNMs may also involve distinct combinations of bases. We found that, of the 144 possible pairs of nucleotide substitutions, some were observed far more often than others (Table S2). CpG mutations may be in part responsible for the overrepresentation of certain MNMs, especially the CA→TG double substitution (where CG can represent an intermediate state). However, of the 48,235 possible MNMs in YH01, <10% (4,372) could possibly be explained by substitutions at CpGs. Pairs of substitutions that could not possibly have been due to CpGs were still significantly overrepresented on a single branch (62.56%; $\chi^2 = 6666.83$, $p < 2.2 \times 10^{-16}$; see Figure S1B).

Because the YH01 genome is based on Illumina short-read sequences, reads containing more than two differences from the reference genome were discarded during mapping [23]. Therefore, any MNMs that alter more than two positions will not have been detected in our analysis. In order to explore longer MNMs—and to ensure that our findings were not an artifact of the elevated error rate inherent to next-generation sequencing technologies—we considered 54,208 MNPs previously identified in the genome of J. Craig Venter [24]. As before, MNPs in this data set consist of clusters of SNPs exhibiting only two observed haplotypes (as illustrated in Figure 1). Because these data were obtained from longer Sanger sequencing reads, larger groups of mutations were detectable, including a few events that contain many neighboring substitutions (Figure S2). Consistent with results for YH01, most of the MNPs in the Venter genome consist of two substitutions within a few bases of one another. However, considering only substitutions at consecutive positions, a substantial number of MNPs involving three to nine bases were found (Figure S2C). These longer events do not include any MNPs that were possibly due to ectopic gene conversion from paralogous sequences or that were due to complementary deletions (Supplemental Experimental Procedures). In the same manner as with YH01, the chimpanzee and orangutan genomes were used to infer the ancestral state of these MNPs. Once again, the majority (67.59%) of polarized MNPs were found to consist solely of mutations occurring on the same branch (Supplemental Experimental Procedures), implying that many of these MNPs are the result of multinucleotide mutational events. We also found that 25.2% of potential MNMs with two substitutions in the Venter genome are present in YH01, strengthening the assertion that these represent true events (rather than sequencing errors) and the inference that they occur simultaneously or in rapid succession.

Far from being a peculiarity of the mutational process in a single organismal lineage, MNMs appear to occur across all domains of life [12, 13]. A number of different mechanisms may explain MNMs, including transient hypermutation due to incorrectly transcribed or translated DNA polymerases [25], or simply the normal activity of the more error-prone components of DNA repair pathways [26, 27]. It may also be the case that a single mutation at one site causes a second mutation at a nearby site (cf. [8, 28]), though this mechanism would act over a small number of generations rather than in a single generation. Because many of the competing hypotheses differ in the processivity of the polymerase invoked to explain the multiplicity of errors or in the specific base substitutions introduced, the genomic data presented here should provide a large number of events that can be used to distinguish among them.

Regardless of the precise molecular mechanisms, it is clear that the interpretation of patterns of molecular evolution—especially with regard to inferences of adaptive evolution—must take into account the pervasiveness of MNMs. For instance, the observation that 64% of substitutions in the same codon occurred along the same lineage led Bazykin et al. [5] to infer the action of positive selection; this proportion is almost exactly the same as the number of MNMs we observe across the genome, suggesting that no selective explanation is necessary. Although adaptive processes need not be invoked if MNMs are common, this does not exclude the possibility that MNMs can themselves be a target of selection. In fact, the activity of such a mutational mechanism also raises the possibility that organisms can "leap" across apparent fitness valleys [29] by simultaneously acquiring multiple substitutions required to reach higher fitness states [30, 31]. This result implies that we may have to reassess the probability of seemingly rare evolutionary events [32].

Supplemental Information

Supplemental Information includes two figures, two tables, and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.cub.2011.05.013.

Acknowledgments

We thank M. Lynch for encouragement and discussion and L. Moyle, M. Rockman, D. Begun, K. Montooth, A. Alverson, A. Tucker, A. Eyre-Walker, and an anonymous reviewer for comments on the manuscript. D.R.S. is supported by National Institutes of Health Genetics, Cellular & Molecular Sciences Training Grant GM007757. J.N.H. is supported by an Indiana University Cox Research Scholarship and a National Science Foundation Research Experience for Undergraduates supplement awarded to M.W.H. (DBI-0845494). M.W.H. is supported by a fellowship from the Alfred P. Sloan Foundation.

Received: March 10, 2011
Revised: April 12, 2011
Accepted: May 5, 2011
Published online: June 2, 2011

References

1. Arnheim, N., and Calabrese, P. (2009). Understanding what determines the frequency and pattern of human germline mutations. Nat. Rev. Genet. *10*, 478–488.
2. Gillespie, J.H., and Langley, C.H. (1979). Are evolutionary rates really variable? J. Mol. Evol. *13*, 27–34.
3. Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.
4. Hudson, R.R., Kreitman, M., and Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data. Genetics *116*, 153–159.
5. Bazykin, G.A., Kondrashov, F.A., Ogurtsov, A.Y., Sunyaev, S., and Kondrashov, A.S. (2004). Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. Nature *429*, 558–562.
6. Meer, M.V., Kondrashov, A.S., Artzy-Randrup, Y., and Kondrashov, F.A. (2010). Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. Nature *464*, 279–282.
7. Whelan, S., and Goldman, N. (2004). Estimating the frequency of events that cause multiple-nucleotide changes. Genetics *167*, 2027–2043.
8. Amos, W. (2010). Even small SNP clusters are non-randomly distributed: Is this evidence of mutational non-independence? Proc. Biol. Sci. *277*, 1443–1449.
9. Donmez, N., Bazykin, G.A., Brudno, M., and Kondrashov, A.S. (2009). Polymorphism due to multiple amino acid substitutions at a codon site within *Ciona savignyi*. Genetics *181*, 685–690.
10. Hodgkinson, A., and Eyre-Walker, A. (2010). Human triallelic sites: Evidence for a new mutational mechanism? Genetics *184*, 233–241.
11. Averof, M., Rokas, A., Wolfe, K.H., and Sharp, P.M. (2000). Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science *287*, 1283–1286.
12. Drake, J.W., Bebenek, A., Kissling, G.E., and Peddada, S. (2005). Clusters of mutations from transient hypermutability. Proc. Natl. Acad. Sci. USA *102*, 12849–12854.
13. Drake, J.W. (2007). Too many mutants with multiple mutations. Crit. Rev. Biochem. Mol. Biol. *42*, 247–258.
14. Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C.R., Dopman, E.B., Dickinson, W.J., Okamoto, K., Kulkarni, S., Hartl, D.L., et al. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc. Natl. Acad. Sci. USA *105*, 9272–9277.
15. Denver, D.R., Morris, K., Lynch, M., and Thomas, W.K. (2004). High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. Nature *430*, 679–682.
16. Denver, D.R., Dolan, P.C., Wilhelm, L.J., Sung, W., Lucas-Lledo, J.I., Howe, D.K., Lewis, S.C., Okamoto, K., Thomas, W.K., Lynch, M., et al. (2009). A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. Proc. Natl. Acad. Sci. USA *106*, 16310–16314.
17. Ossowski, S., Schneeberger, K., Lucas-Lledo, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science *327*, 92–94.
18. Keightley, P.D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M.L. (2009). Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res. *19*, 1195–1201.
19. Chen, J.M., Ferec, C., and Cooper, D.N. (2009). Closely spaced multiple mutations as potential signatures of transient hypermutability in human genes. Hum. Mutat. *30*, 1435–1448.
20. Kondrashov, A.S. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum. Mutat. *21*, 12–27.
21. Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. Proc. Natl. Acad. Sci. USA *107*, 961–968.
22. Altshuler, D.L., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., De la Vega, F.M., Donnelly, P., Egholm, M., et al. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
23. Wang, J., Wang, W., Li, R.Q., Li, Y.R., Tian, G., Goodman, L., Fan, W., Zhang, J.Q., Li, J., Zhang, J.B., et al. (2008). The diploid genome sequence of an Asian individual. Nature *456*, 60–65.
24. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The diploid genome sequence of an individual human. PLoS Biol. *5*, e254.
25. Ninio, J. (1991). Transient mutators: A semiquantitative analysis of the influence of translation and transcription errors on mutation rates. Genetics *129*, 957–962.
26. Loeb, L.A., and Monnat, R.J., Jr. (2008). DNA polymerases and human disease. Nat. Rev. Genet. *9*, 594–604.
27. Arana, M.E., and Kunkel, T.A. (2010). Mutator phenotypes due to DNA replication infidelity. Semin. Cancer Biol. *20*, 304–311.
28. Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., Hudson, R., Bergelson, J., and Chen, J.Q. (2008). Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature *455*, 105–108.
29. Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc. Sixth Int. Congr. Genet. *1*, 356–366.
30. Bridgham, J.T., Ortlund, E.A., and Thornton, J.W. (2009). An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature *461*, 515–519.
31. Weinreich, D.M., Delaney, N.F., DePristo, M.A., and Hartl, D.L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. Science *312*, 111–114.
32. Lynch, M., and Abegg, A. (2010). The rate of establishment of complex adaptations. Mol. Biol. Evol. *27*, 1404–1414.

Current Biology, Volume *21*

Supplemental Information

Pervasive Multinucleotide Mutational Events in Eukaryotes

Daniel R. Schrider, Jonathan N. Hourmozdi, and Matthew W. Hahn

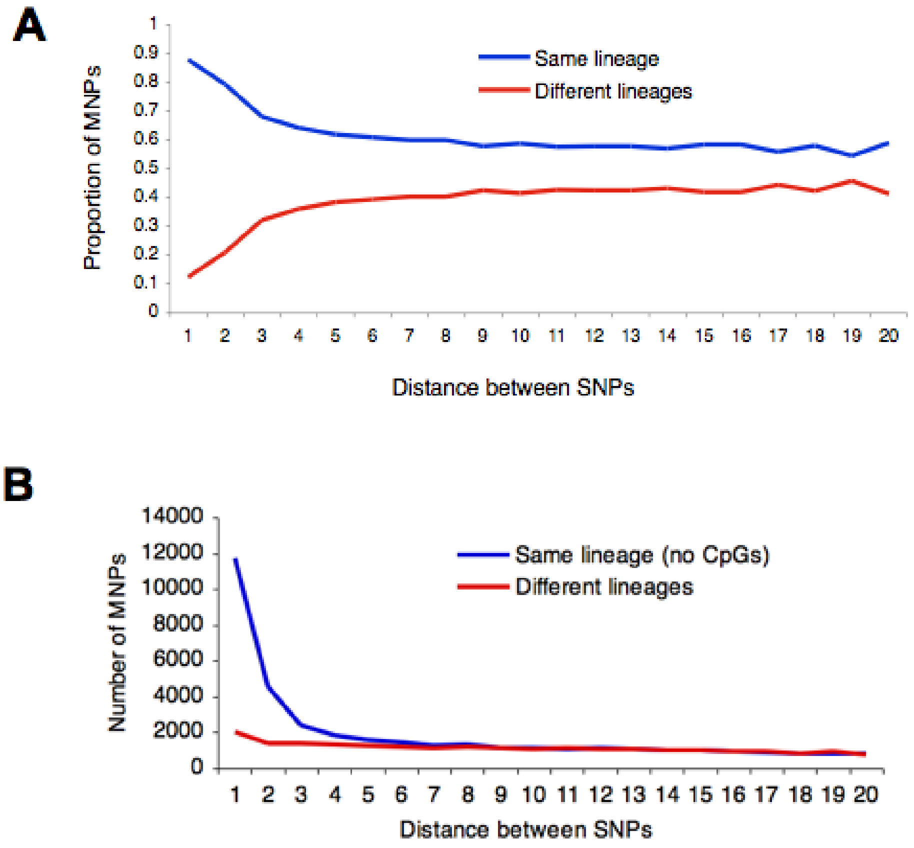

Figure S1, Related to Figure 3.

(A) The proportions of MNPs due to substitutions on the same or different lineages. The figure is based on the same data as in Figure 3 of the main text, but shows proportions rather than counts.

(B) The distribution of distances between phased and polarized pairs of substitutions in YH01 that occurred on the same lineage but could not be due to CpGs (blue line) and pairs that occurred on different lineages (red line). This figure is identical to Figure 3, except the numbers of pairs of substitutions occurring on the same lineage that could possibly be due at least in part to CpGs (based on flanking bases and both possible intermediate states) have been removed.

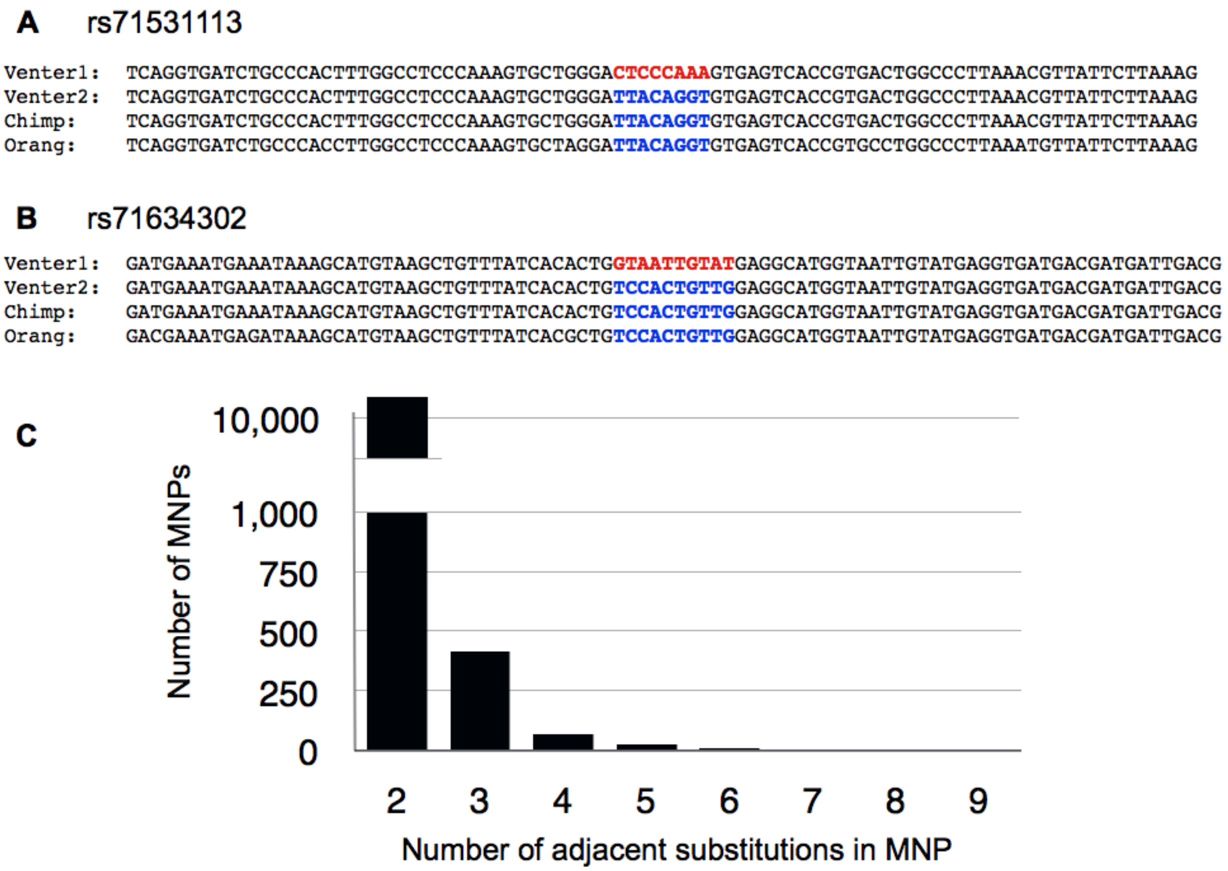

Figure S2, Related to Figure 1.

(A and B) Examples of heterozygous MNMs containing more than two substitutions from the diploid sequence of J. Craig Venter [24]. The MNM is highlighted in color (with the ancestral state in blue) and dbSNP IDs are given for each example.

(C) Here we plot only those MNPs with substitutions at consecutive positions in the Venter genome; for example, length 3 means that three consecutive positions differed between the two haplotypes (and also that the reference genome matched one haplotype exactly). There was one such polymorphism for the bins containing both 8 and 9 consecutive substitutions.

Table S1. Proportion of Single-Nucleotide and Multinucleotide Polymorphisms (with Both Substitutions on One Haplotypic Branch) across the Human Genome, Related to Figure 2

	SNPs	**MNPs**
Exon	0.40%	0.52%
Intron	34.50%	36.00%
Intergenic	65.10%	63.48%

Table S2. Frequency of All 144 Possible Two-Nucleotide Substitutions, Related to Figure 2

Mutation	Number of Events	Proportion of Events
TC->CT	1219	0.025272105
GA->AG	1217	0.025230642
CA->TG	1181	0.024484296
GG->AA	1169	0.024235514
CT->TC	1137	0.023572095
AG->GA	1133	0.023489168
CC->TT	1130	0.023426972
TG->CA	1101	0.022825749
GC->AT	1035	0.021457448
AA->GG	911	0.018886701
TT->CC	901	0.018679382
CG->TA	896	0.018575723
TA->CG	855	0.017725718
AT->GC	854	0.017704986
TA->AT	801	0.016606199
GT->AC	780	0.01617083
AC->GT	758	0.01571473
AT->TA	757	0.015693998
GC->TT	577	0.011962268
GC->AA	541	0.011215922
TC->AT	442	0.009163471
GA->AT	441	0.009142739
AG->GT	412	0.008541515
CT->AC	402	0.008334197
TA->CT	399	0.008272002
TA->CC	399	0.008272002
TA->GG	397	0.008230538
TA->AG	383	0.007940292
AT->TC	383	0.007940292
AT->GA	382	0.00791956
CA->TC	382	0.00791956
GA->TG	377	0.007815901
AG->CA	376	0.007795169
TC->AA	376	0.007795169
GA->TT	371	0.00769151
CC->TG	368	0.007629315
AA->TT	359	0.007442728
TT->AA	356	0.007380533
GC->AG	356	0.007380533
AC->CT	354	0.007339069
GT->AG	353	0.007318337
GG->AT	352	0.007297605
GG->CA	348	0.007214678
CT->TG	348	0.007214678
TC->CA	346	0.007173214
CC->AT	343	0.007111019
TG->CT	337	0.006986628

TG->GA	327	0.00677931
GT->AA	326	0.006758578
CA->AG	325	0.006737846
AG->TA	320	0.006634187
TG->AA	318	0.006592723
TT->AC	318	0.006592723
GC->CT	316	0.006551259
AC->TT	312	0.006468332
CG->TT	312	0.006468332
AA->GT	310	0.006426868
CT->TA	308	0.006385405
CG->AA	305	0.006323209
CC->TA	303	0.006281746
CA->TT	302	0.006261014
CT->GC	299	0.006198818
CC->GT	297	0.006157355
GT->TG	293	0.006074427
CA->AC	285	0.005908573
AC->GA	282	0.005846377
GG->AC	282	0.005846377
AA->TG	278	0.00576345
CA->GG	276	0.005721986
AG->GC	274	0.005680522
GA->AC	269	0.005576863
TT->CA	264	0.005473204
TG->GT	263	0.005452472
GG->TT	261	0.005411009
AC->CA	261	0.005411009
GG->TA	257	0.005328081
AA->GC	257	0.005328081
TG->CC	254	0.005265886
GT->TC	253	0.005245154
TC->GT	251	0.00520369
TG->AT	249	0.005162227
TT->GC	248	0.005141495
GA->CG	244	0.005058567
GT->CC	243	0.005037836
TT->CG	239	0.004954908
CC->AA	236	0.004892713
CG->GA	235	0.004871981
CG->TC	233	0.004830517
AT->GG	232	0.004809785
AT->CC	229	0.00474759
TC->CG	228	0.004726858
AC->GG	226	0.004685394
AA->CG	223	0.004623199
CA->AT	221	0.004581735
AT->CA	214	0.004436612
AT->TG	209	0.004332953

AG->TT	200	0.004146367
TA->GT	194	0.004021976
TA->AC	190	0.003939048
AC->TA	182	0.003773194
CT->AA	181	0.003752462
GT->TA	180	0.00373173
AA->CT	161	0.003337825
AA->CC	160	0.003317093
TT->GG	155	0.003213434
TT->AG	148	0.003068311
GC->TG	139	0.002881725
GC->CA	134	0.002778066
GA->CT	131	0.00271587
CT->AG	128	0.002653675
TC->AG	127	0.002632943
GC->TA	127	0.002632943
GA->CC	127	0.002632943
TC->GA	124	0.002570747
TC->GG	124	0.002570747
TG->AC	123	0.002550016
AA->TC	123	0.002550016
TT->GA	122	0.002529284
AG->CT	119	0.002467088
CC->AG	119	0.002467088
CC->GA	119	0.002467088
GG->CT	116	0.002404893
GA->TC	115	0.002384161
CT->GG	113	0.002342697
GC->CG	113	0.002342697
GG->CC	111	0.002301234
TG->GC	107	0.002218306
AG->TC	107	0.002218306
CC->GG	104	0.002156111
GT->CA	101	0.002093915
AG->CC	101	0.002093915
TA->GC	101	0.002093915
AC->TG	99	0.002052452
CT->GA	98	0.00203172
CA->GC	98	0.00203172
CG->GT	98	0.00203172
GG->TC	93	0.001928061
CG->GC	86	0.001782938
CG->AC	86	0.001782938
CG->AT	85	0.001762206
CA->GT	81	0.001679279
GT->CG	80	0.001658547
AC->CG	72	0.001492692
AT->CG	71	0.00147196

Supplemental Experimental Procedures

MA Lines and Trios

Data from mutation-accumulation lines was compiled from the main text and supplementary materials of the published literature [14-18]. To determine whether there were more multinucleotide mutation events than expected by chance, we calculated the probability of observing clusters of substitutions according to a Poisson process. In particular, for each study we calculated the per-base rate parameter by dividing the number of individual nucleotide substitutions observed by the number of bases in the genome at which substitutions could be detected, and then dividing this by the number of MA lines in the study (because MNMs must occur in the same MA line). To calculate the rate for each window size (10, 20, 50, or 100 bp), we multiplied the per-base rate by the corresponding length. The probability of seeing one or more windows with multiple substitutions was given by $1\text{-cdf}(1, \lambda)^n$, where cdf is the cumulative distribution function for a Poisson process with per-window rate λ, and there are n windows across the genome. The probability of seeing even one MNM is less than 0.05 in each study for all window sizes. The proportion of nucleotide substitutions due to MNMs was determined by summing the number of substituted bases within all MNMs and dividing by the total number of substituted bases including both single- and multinucleotide substitutions.

Data from published analyses of human disease trios was first obtained from the supplementary materials in ref. [21] and then by re-examining data for 12 genes individually. Close examination revealed several factors that might contribute to both the over- and under-counting of MNMs. First, this count could include so-called “complex” mutational events that involve the addition or subtraction of nucleotides in conjunction with base substitutions. This form of mutation does involve multiple nucleotide positions, but is not considered in our definition of MNMs. Second, there appears to be significant under-reporting of MNMs in the data. When only one site contributes to the creation of a stop codon, other nearby mutations may go unreported in the literature. In addition, adjacent nucleotide substitutions are often reported as “indels” or “delins,” implying a deletion followed by the insertion of the same number of bases within a single generation, rather than as multiple nucleotide substitutions. The number reported in the main text is therefore likely to be some average of true MNMs, complex mutations, and unreported MNMs, but they all must have occurred in a single generation.

Detecting and Phasing Pairs of Nearby Substitutions in Illumina Data

Mappings of Illumina reads from YH01 to the NCBI 36 human reference genome were downloaded from http://yh.genomics.org.cn. SOAPsnp [33] was then used to call single nucleotide polymorphisms in this individual using default parameters with the –t option. Note that the algorithm used to call SNPs in the publication describing YH01 has a strong bias against calling pairs of SNPs within 5 bp of one another [23]. SNPs with fewer than 5 reads supporting either base in the case of heterozygotes (or 10 reads supporting homozygous calls), or with more than 2 reads supporting a base differing from the consensus base calls were removed from the remainder of the analysis. Groups of SNPs within 20 bp of one another were then collected from the set of remaining SNP calls; 151,015 pairs of SNPs met these criteria. 19,143 of these pairs consisted of one heterozygous SNP and one homozygous SNP, and were removed from the remainder of the analysis as such pairs cannot be the result of multinucleotide mutation events. 98,606 of the pairs consisted of two heterozygous SNPs. These pairs were phased by examining the haplotype in each read spanning both SNPs in the pair. The two most common haplotypes supported by the reads were determined, and if the number of reads supporting each haplotype was at least 5, these two haplotypes were inferred to be the true haplotypes present in YH01. If more than one read supported a haplotype other than the two most common haplotypes, however, then this pair of SNPs was removed from further analysis. Of the 63,103 pairs of SNPs successfully phased, for 51,557 pairs no recombinant haplotype was observed in the NCBI reference genome (Figure 1A). We also found 33,266 pairs of homozygous SNPs in YH01 (Figure 1B).

For the 1000 genomes trios, mapping locations of Illumina reads to the NCBI 36 human reference genome were downloaded from http://www.1000genomes.org. SNPs were called in each individual in the two trios using the pileup command in SAMtools [34]. These calls were used to detect *de novo* single-nucleotide mutations and *de novo* MNMs (two mutations within 20 bp of one another). Single-nucleotide mutations were retained if both parents had at least 10 reads supporting the consensus base call, no reads supporting the mutant allele present in the offspring, and no more than 2 reads supporting any other allele. For *de novo* MNMs, the offspring was required to have at least 5 reads supporting each haplotype and no more than two reads supporting any other haplotype, and neither allele could have more than 1.5 times as many reads as the other. Each parent was required to have at least 10 reads matching the non-mutant haplotype, no reads supporting the mutant haplotype called in the offspring, and no more than two reads supporting any other haplotype. It is likely that many of the *de novo* mutations detected were somatic or occurred in the sequenced cell lines [22], but this does not change the conclusions with regard to the proportion of substitutions that are due to MNMs. We observed 2 mutations in the CEU trio (2.11% of all mutations) and 5 mutations in the YRI trio (3.23% of all mutations). Removing the 1.5-fold criterion resulted in 1.52% and 3.66% of *de novo* mutations being MNMs (in CEU and YRI, respectively), while increasing depth thresholds to 10 reads each for parental and *de novo* alleles resulted in 1.73% and 2.18% MNMs (no 1.5-fold threshold) and 1.40% and 1.62% MNMs (with 1.5-fold threshold). We used the same approach as with the MA data (see “MA lines and trios” above) to determine that the probability of observing clusters of two or more substitutions within 20 bp of one another is extremely low ($P<0.0005$) if mutations are identically Poisson distributed.

MNPs in the Venter Genome

MNPs were previously identified in the Sanger-sequenced genome of J. Craig Venter [24]. These MNPs were downloaded from NCBI dbSNP using the query: “multinucleotide polymorphism”[SNP_CLASS] AND “human”[ORGN] AND "HUMANGENOME_JCVI"[HANDLE]. The coordinates of these MNPs on the human reference genome are not available at dbSNP or the JCVI website. We therefore used BLAT [35] to attempt to unambiguously map the Venter MNPs to the GRCh37 human reference genome. For each MNP, the 500 bp of flanking sequence (250 bp on each side) was obtained from dbSNP for use with BLAT, and hits not having percent identity >95, bit-score >200, and spanning at least 50 bp of flanking sequence on either side of the MNP were removed from the analysis. Any MNP with multiple distinct hits (>90% identity) was removed from this set to prevent ambiguous mapping to the reference genome. This step also ensures that any MNPs found in the Venter genome are not the result of gene conversion events.

Polarizing MNPs

To determine whether pairs of SNPs exhibiting only two haplotypes (i.e. MNPs) were the result of mutation events happening on the same or different haplotypic lineages, we inferred the ancestral state at each pair of SNPs via parsimony using the chimpanzee and orangutan reference genomes. Coordinates of the SNPs in these genomes were determined using the liftOver tool at the UCSC Genome Browser [36]. In a small number of cases the location of the SNPs could not be determined or an indel was present; these cases were removed to prevent inaccurate mapping between genomes from affecting our results. These cases can also represent MNPs introduced by complementary deletions. For instance, if the ancestral state is the sequence ACGT, non-overlapping 2 bp deletions in each human haplotype (i.e. AC-- and --GT) could result in an apparent 2 bp substitution, but these were removed from our analysis. If both outgroup genomes agreed, or only one was found, the outgroup haplotype was inferred to be the ancestral state and substitutions were assigned to haplotypic lineages by parsimony. All other cases were removed to minimize the possibility of inaccurate inferences of the ancestral state. A comparison of the ancestral states inferred by parsimony and likelihood [using PAML; ref. 37] revealed 99.93% agreement.

Supplemental References

33. Li, R.Q., Li, Y.R., Fang, X.D., Yang, H.M., Wang, J., and Kristiansen, K. (2009). SNP detection for massively parallel whole-genome resequencing. Genome Research *19*, 1124-1132.

34. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, G.P.D.P. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078-2079.

35. Kent, W.J. (2002). BLAT--The BLAST-Like Alignment Tool. Genome Research *12*, 656-664.

36. Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., et al. (2011). The UCSC Genome Browser database: update 2011. Nucleic Acids Research, 10.1093/nar/gkq1963.

37. Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution *24*, 1586-1591.