SNP Formation Bias in the Murine Genome Provides Evidence for Parallel Evolution

Zackery E. Plyler[1], Aubrey E. Hill[2], Christopher W. McAtee[3], Xiangqin Cui[4], Leah A. Moseley[3], and Eric J. Sorscher[5,*]

[1]Department of Biology, University of Alabama at Birmingham

[2]Department of Computer and Information Sciences, University of Alabama at Birmingham

[3]Gregory Fleming James Cystic Fibrosis Research Center, University of Alabama at Birmingham

[4]Department of Biostatistics, University of Alabama at Birmingham

[5]Department of Pediatrics, Emory University School of Medicine

*Corresponding author: E-mail: esorscher@emory.edu.

Associate editor: Maria Costantini

Accepted: August 2, 2015

Abstract

In this study, we show novel DNA motifs that promote single nucleotide polymorphism (SNP) formation and are conserved among exons, introns, and intergenic DNA from mice (Sanger Mouse Genomes Project), human genes (1000 Genomes), and tumor-specific somatic mutations (data from TCGA). We further characterize SNPs likely to be very recent in origin (i.e., formed in otherwise congenic mice) and show enrichment for both synonymous and parallel DNA variants occurring under circumstances not attributable to purifying selection. The findings provide insight regarding SNP contextual bias and eukaryotic codon usage as strategies that favor long-term exonic stability. The study also furnishes new information concerning rates of murine genomic evolution and features of DNA mutagenesis (at the time of SNP formation) that should be viewed as "adaptive."

Key words: parallel evolution, mutation, SNP formation bias.

Introduction

Mutations and specific constraints that govern sites of DNA polymorphism are not well understood. Although eukaryotic single nucleotide polymorphism (SNP) formation is viewed as stochastic, preferences in the nature and location of certain single base replacements (including C⇔T and A⇔G [transition] mutations) are described among metazoans (Lederberg JA and Lederberg EM 1952; Collins and Jukes 1994; Wakeley 1994; Drake et al. 1998; Freeland and Hurst 1998; Sung et al. 2012; Hill et al. 2014). Isolated "hot spots" for single nucleotide variants exist, but there is little or no information regarding the overall contribution of such regions to eukaryotic SNP formation. A bias underlying hominid or murine single base polymorphism as determined by neighboring sequence context has been suggested, including effects from nucleotides upwards of 200 bp away from a polymorphic site (Koch 1971; Krawczak et al. 1998; Zhao and Boerwinkle 2002; Zhao and Zhan 2004; Zhang and Zhao 2005). Others have concluded that although human and nonhuman primate SNPs exhibit striking homology to each other, the surrounding DNA context is not responsible (Duret 2009; Hodgkinson et al. 2009; Hodgkinson and Eyre-Walker 2010). These earlier studies point to gaps in knowledge regarding formation and ongoing distribution of SNPs among higher organisms.

Massive sequence compendia from inbred murine strains furnish a powerful tool for investigating the "randomness" of SNP accumulation. In the present analysis, full genomic files from the Sanger Institute Mouse Genomes Project (http://www.sanger.ac.uk/resources/mouse/genomes/, last accessed August, 2015), including deep sequence and polymorphism data, were examined for 17 strains to investigate SNP formation bias. The murine strains (DBA, CBA, Balb/c, etc.) have each been backcrossed and/or inbred for well over 50 filial generations to reach allelic fixation with respect to ancestral polymorphism (Bailey 1978; Green 1981). Stringent parameters (depth of coverage, Phred quality score) were established to ensure that polymorphisms culled from the database were

correctly delineated, and representative SNP cohorts were manually inspected to confirm authenticity. Our findings revealed a pronounced bias underlying SNP location in mice and we verified the same observations prospectively in human germ line DNA based on 1000 Genomes data (http://www.1000genomes.org, last accessed August, 2015) and acquired somatic mutations in human breast cancer (https://tcga-data.nci.nih.gov/tcga/dataAccess Matrix.htm, last accessed August, 2015). The results also establish a remarkable level of recent parallel evolution within the murine genome. Here we show regulatory patterns that underlie SNP formation, and provide a framework for investigating novel aspects of genomic diversification.

Materials and Methods

Acquisition of Data from Sanger Institute Compendium

SNP data from discrete regions (intronic, exonic, and intergenic) were queried and downloaded from the Sanger Mouse Genomes Project database (http://www.sanger.ac.uk/re sources/mouse/genomes/, last accessed August, 2015) for 17 highly inbred *Mus musculus* strains (Keane et al. 2011; Yalcin et al. 2011). Intronic and intergenic SNPs were obtained from chromosomes 1–3 after being shown to meet screening thresholds that included: 1) Sequencing depth of $\geq$30, and 2) Phred score of at least 60 for three categories (SNP quality, mapping quality, and consensus quality). Exonic SNPs were obtained from murine chromosomes 1–8 at the same level of sequencing depth and quality. In this manner, a very large region of chromosomal DNA (>500 million bp in each of 17 strains) was evaluated. Because 1) substantial variability in SNP patterns has not been suggested among different chromosomes; 2) previous studies in mouse, human, and other species have drawn valuable conclusions using comparable (or smaller) genomic samples (Zhang et al. 2005; Xue et al. 2009; Billings et al. 2010; Miller et al. 2012); and 3) practical aspects were found to limit a more extensive computational analysis, the chromosomal regions tested here were viewed as sufficient for purposes of this report. Polymorphism data were examined for every homozygous site for all 17 murine lines and with at least 4 strains (for noncoding DNA regions) or 2 strains (for exonic SNPs) bearing a minor allele in homozygous form. Positions of heterozygosity were also investigated. Because very few heterozygous alleles are anticipated among highly inbred (congenic) lines, well-validated heterozygous sites were provisionally interpreted as recent mutations (acquired during laboratory inbreeding [Bailey 1978; Green 1981; Silver 1995; Peters 2007; see also Considerations Regarding Murine Heterozygosity]). These data sets (termed "homosites" and "heterosites," respectively) were distributed among six groups, according to type of SNP (A⇔G, C⇔T, A⇔C, etc.).

Manual Inspection of Murine SNPs

As a further test of authenticity, a representative sampling of SNPs from each category (heterosite, homosite, exonic, intronic, and intergenic) and a random cohort of nucleotide positions without known polymorphism were evaluated using Interactive Genomics Viewer (IGV) software (http://www.broadinstitute.org/igv/, last accessed August, 2015). Primary data from the Sanger Mouse Genomes Project (http://www.sanger.ac.uk/re sources/mouse/genomes/, last accessed August, 2015) were downloaded and a 160–200 bp interval surrounding each SNP or random (non-SNP) position (selected by computer algorithm), and formally inspected for parameters associated with next generation sequencing (NGS) artifact including 1) diminished (local) Phred score, 2) low regional sequencing depth or map quality, 3) nearby short repeats (which suggest incorrect SNP alignment), 4) DNA motifs linked previously to sequence error (Dohm et al. 2008; Harismendy et al. 2009), 5) indels in the immediate vicinity, 6) obvious misalignment among multiple reads, 7) unexplained increase in reported coverage (i.e., "pile-up"; in which the number of sequences obtained over a particular DNA segment is markedly increased, indicating gene duplication and/or aberrant sequence alignment), and 8) evidence of greater than two haplotype blocks from the same region (suggestive of possible template contamination, as any gene in a particular murine strain should be represented by only two haplotypes). A screening algorithm for viewing the data was depicted in tabular form.

Acquisition of Neighboring Nucleotide Context and Genomic Representation

Neighboring nucleotides that flank SNP positions were retrieved from the Ensembl genome browser (http://uswest.ensembl.org/Mus_musculus/Info/Index, last accessed August, 2015). DNA sequences 50 bp 5′ or 3′ to a position of interest were extracted from Ensembl so that reads surrounding each position could be compiled as output text files. The sequences were aligned at the time of acquisition and loaded into an Excel spreadsheet for analysis (see also Computer Simulation) with reverse complement SNPs combined (A⇔G with C⇔T, A⇔C with G⇔T). Base frequency at nucleotide positions relative to an SNP site (±50 bp) was monitored to obtain "bias (%)" (overall base representation across each respective region of the murine genome was subtracted from base representation observed experimentally for every relative position surrounding each SNP on murine chromosomes 1–8 [exonic] or 1–3 [intronic, intergenic]—larger coding DNA samples being necessary to record sufficient exonic variants). Phred score, depth cutoff, and so forth were held constant among all SNPs to maintain stringency. The analysis also was conducted for a comparable number of randomly chosen (nonpolymorphic) positions within murine chromosomes 1–4 (exonic) or 1–3 (intronic, intergenic), as further control to

assess context-dependent SNP location. Standard deviation for base representation was measured by bootstrapping individual samples from the data set over a series of 2,000 repeats.

Dinucleotide Quartet Analysis

For this study, dinucleotide quartets were defined as two base pairs upstream and two downstream of a polymorphic site or other nucleotide position being evaluated. The dinucleotide quartet frequency surrounding each SNP type was collected from 17 murine strains (DBA/2J, CBA/J, BALB/cJ, 129P2/OlaHsd, 129S1/SvImJ, 129S5SvEvBr, A/J, AKR/J, C3H/HeJ, C57BL/6NJ, CAST/EiJ, FVB/NJ, LP/J, NOD/ShiLtJ, NZO/HlLtJ, PWK/PhJ, WSB/EiJ) and normalized for expected values among all 256 possible quartets (5′-XX|XX-3′, where vertical line denotes SNP location) in areas of interest (exonic, intronic, and intergenic) as determined by a computer program that directly tallies occurrence of all possible quartets on murine chromosomes 1–3 using sequence data (.txt files) downloaded from Ensembl. The observed incidence for each quartet surrounding a particular SNP type was compared with a stochastic representation of SNP patterns (see Statistics). Quartets were considered "permissive" or "shielded" to polymorphism only if SNP representation was significantly different from expected; for example, for quartet context *E* surrounding an A⇔G SNP, *E* would be termed permissive for A⇔G variants only if observed association with single base replacement was statistically greater than the incidence at which both adenine and guanine are expected stochastically (only if *P* values for both nucleotides were significant).

Acquisition and Analysis of Human SNP Data from 1000 Genomes

SNP data were queried and downloaded from 1000 Genomes (http://www.1000genomes.org, last accessed August, 2015) for 19 human genes of interest taken from a list extensively characterized by our laboratory for features, such as transition bias, genetic founder alleles, well-defined minor allelic frequency, intronic versus exonic SNP prevalence, haplotype block formation, synonymous versus nonsynonymous polymorphism, and conservation among multiple species (including horse, frog, zebrafish, opossum, shark, and chicken) (Fortini et al. 2000; Hill et al. 2014). Intronic DNA was studied to minimize evolutionary selection bias, and data were collated according to SNP type (A⇔G, C⇔T, A⇔C, etc.). Computer-based summation was used (as above) to measure incidence of each nucleotide context within human intronic DNA. Site-specific base frequencies (as well as "Bias (%)") and dinucleotide quartet representation were calculated to determine over- or underrepresented contexts versus the incidence measured directly for each of 256 possible contexts across a greater than 300-million-bp region of the human introme.

Analysis of Human Breast Cancer

SNP data from intronic regions were queried and downloaded from the TCGA database ([https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm, last accessed August, 2015] for somatic mutations specific to human breast carcinoma [i.e., differing from germ line]). Data were divided as above (A⇔G, C⇔T, A⇔C, etc.) and computer summation used to determine surrounding nucleotides for each somatic SNP. Site-specific base frequencies and dinucleotide quartet representation were calculated to determine over- and underrepresented contexts versus the incidence measured directly across greater than 300 million bases.

Statistics

Because statistical analysis in this study involved comparisons between observed and expected SNP frequencies, as well as incidence of specific nucleotide contexts, *P* values were calculated by chi-square. We considered using Fisher's exact test for analyses of the "hit" (SNP) and "no-hit" (no SNP) findings within 2×2 tables as magnitudes of these frequencies varied. However, chi-square using observed counts indicated that all assumptions for the test were satisfied. In particular, none of the four expected values in 2×2 contingency tables was less than 5 (Rosner 2011). We therefore concluded that chi-square was the appropriate test. Observed and expected SNP and context tallies were obtained (with frequencies calculated) and compared by 2×2 contingency tables with Yates's correction for continuity (Yates 1934) to minimize Type I error. For tables of sequential tests (e.g., dinucleotide pairings in quartet representation; i.e., 256 sequential comparisons), the Bonferroni (Rice 1988) technique was used to further diminish Type I error. In other analyses, as a baseline for "CG" representation, triplet frequencies derived from murine codon usage (and equivalent counts of CG recognition by anticodons) were calculated and compared with a random distribution of dinucleotide frequency rates (based on computer summation of A, T, C, and G incidence from currently identified murine exonic DNA).

Computer Simulation

Computer programs (written in "java") were designed to perform the following tasks. Specific sequences ±50 bp for a given SNP site were retrieved from the Ensembl genome browser (http://uswest.ensembl.org/Mus_musculus/Info/Index, last accessed August, 2015). Randomly selected (nonpolymorphic) bases corresponding to a given SNP type were obtained as a control and aligned to the ±50-bp region surrounding these sites (i.e., for A⇔G SNPs, adenine or guanine sites were randomly selected). The program was run with Eclipse Integrated Development Environment (IDE) software (https://eclipse.org/downloads/packages/eclipse-ide-java-developers/marsr, last accessed August, 2015). Results were used as a further control for comparison to relative

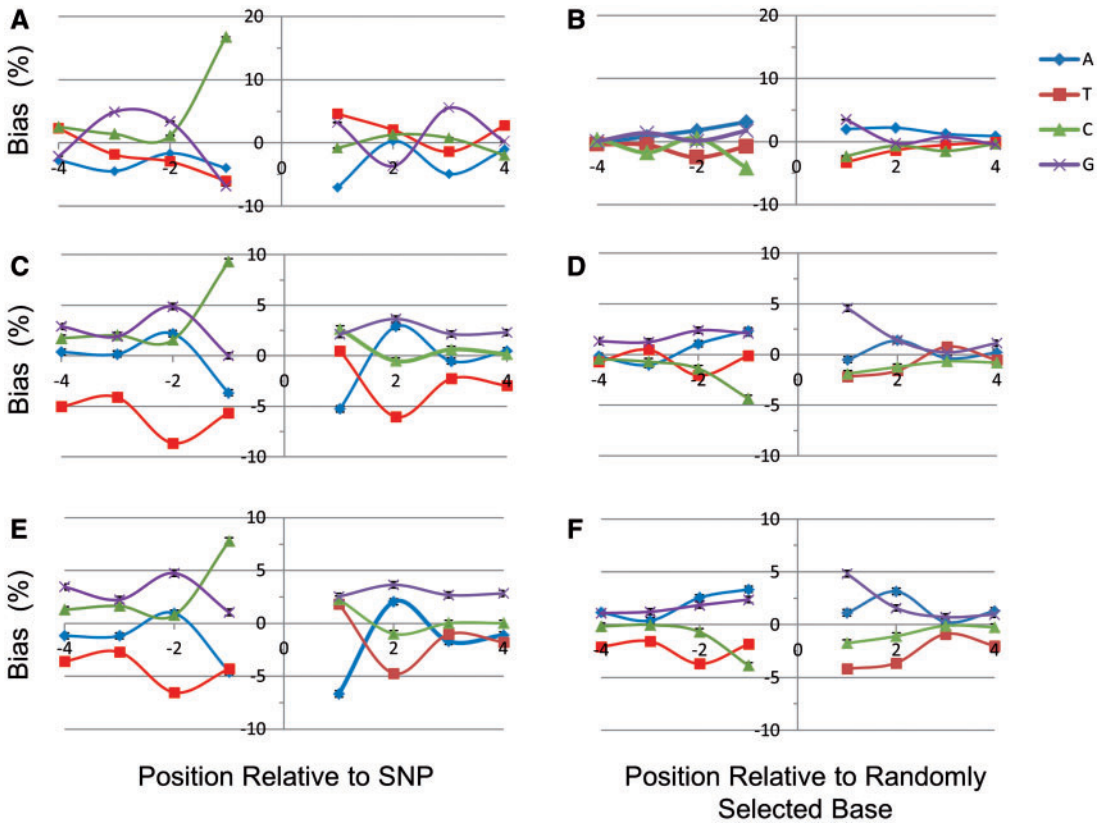

Fig. 1.—Base frequency bias of murine A⇔G SNPs. (*A*) Base location frequency bias (Bias %) immediately surrounding (±4 bp) 20,603 exonic A⇔G homozygous SNPs, compiled from coding regions of murine chromosomes 1–8. Bias % was calculated as described in Materials and Methods. (*B*) Genomic base frequency bias (±4 bp) relative to 20,000 randomly chosen (not SNP-associated) adenine (A) and guanine (G) nucleotides within murine exons (chromosomes 1–4) studied in a fashion otherwise identical to Panel (*A*). (*C*) Base location frequency bias immediately surrounding (±4 bp) 50,244 A⇔G SNPs compiled from intronic regions of murine chromosomes 1–3. (*D*) Genomic base frequency bias relative to 50,000 randomly chosen "A" or "G" sites within murine introns. (*E*) Base location frequency bias relative to 67,663 A⇔G SNPs compiled from intergenic regions on murine chromosomes 1–3. (*F*) Genomic base frequency bias relative to 50,000 randomly chosen "A" or "G" sites from intergenic regions of murine chromosome 2. In all cases, standard deviation (as judged by bootstrap analysis) was very low (on the order of ~0.1–0.3%).

nucleotide frequencies associated with random SNP locations (see fig. 1). In addition, counts representing all dinucleotide quartets surrounding all base positions (A, C, G, or T) for downloaded genomic sequence (.txt files; e.g., exons, introns, and intergenic regions of murine chromosomes 1–3; >300 million bp of human intronic sequence) were established. These results served as an additional, independent control for evaluating quartet bias and SNP distribution.

Results

Context Biases Associated with Homozygous Murine and Hominid SNPs

To investigate positional bias contributing to SNP frequency and/or location, we studied genomic data from the Sanger compendium (Keane et al. 2011; Yalcin et al. 2011) stratified according to intergenic, intronic, or exonic regions of murine DNA. SNP data were filtered in our studies to include polymorphic sites for which multiple distinct murine strains in the repository encoded the minor allele in homozygous form. Sequence depth (≥30-fold coverage) and quality (Phred score) utilized by Sanger (Keane et al. 2011; Yalcin et al. 2011) were robust, and the convention of focusing on positions of redundant homozygous polymorphism allowed an additional level of stringency. For example, the probability of eight minor alleles being called incorrectly (e.g., due to sequence artifact) for a certain SNP identified as homozygous in 4 of 17 inbred murine strains by these criteria based on minimum sequence quality (i.e., Phred score) is less than 1×10^{-22}. For any minor allele (i.e., the less common nucleotide in the population at a specific position), therefore, the likelihood of sequence artifact or miscall was remote.

Data were divided into six categories based on the type of SNP identified (A⇔G, A⇔C, A⇔T, etc.), and a computer algorithm established to retrieve and align surrounding nucleotides for each polymorphic site. In figure 1, exonic SNPs identified across chromosomes 1–8 in 17 murine strains were

analyzed for position-specific bias. Because A⇔G transition SNPs (and reverse compliment C⇔T transitions on the opposite strand) are 1) present in higher numbers across the murine genome than other SNP categories, 2) of mechanistic interest, that is, enhanced in both pro- and eukaryotes (Collins and Jukes 1994; Wakeley 1994), and 3) were found to exhibit significantly conserved patterns in exonic, intronic, and intergenic regions, we focused the analysis on polymorphisms of this type. As described below, the same considerations also apply to other SNP categories.

A marked overrepresentation of cytosine (immediately 5′ to A⇔G SNP location) and underrepresentation of adenine (immediately 3′) were noted for exonic single base replacements (fig. 1*A*). This suggested a sequence bias in the immediate vicinity of A⇔G polymorphism. The same was observed prospectively for both intronic (fig. 1*C*) and intergenic (fig. 1*E*) murine SNPs (compared with randomly chosen controls; fig. 1*B*, *D*, and *F*). Bootstrapping indicated that standard deviations in all cases were small (0.1–0.3%; fig. 1). When data from 1000 Genomes for intronic SNPs among 19 randomly selected human genes were investigated, a similar pattern was observed (supplementary fig. S1*A*, Supplementary Material online). The same was true for somatic SNPs differing from germ line in human breast carcinoma (supplementary fig. S1*B*, Supplementary Material online).

We next investigated sets of "dinucleotide quartets," or groups of four bases immediately 5′ or 3′ to each SNP site ($X_1X_2|X_3X_4$; where vertical line represents SNP location). Data are shown for A⇔G polymorphisms (table 1), but similar patterns were observed for other SNP categories (examples in supplementary tables S1–S3, Supplementary Material online). Each of 256 possible quartets was assembled and frequencies collated among homozygous SNPs for 17 murine strains in which 1) complete genomic sequence data were available and 2) multiple strains were homozygous for the minor allele at a specific position (see Materials and Methods). Statistical analysis was performed by comparing incidence of each quartet surrounding a particular SNP versus the observed (non-SNP associated) occurrence (as measured by computer summation of all possible quartets present on murine chromosomes 1–3 collated by exonic, intronic, and intergenic location). The results define a pronounced bias for SNP prevalence; ranking of quartets from among 256 possibilities for murine exonic A⇔G SNPs is shown in the far left column (table 1 and supplementary fig. S2, Supplementary Material online). Across murine exonic DNA (chromosomes 1–8; >20,000 SNPs), the 20 most frequent A⇔G SNP-associated quartets describe approximately 2,350 single nucleotide variants at preferential sites that would not have occurred on a random basis. For the 20 most shielding A⇔G contexts, approximately 1,840 SNPs expected at random were instead directed away from these specific motifs. Note that the quartets described here are not characteristic of stereotypic repetitive elements in human or murine DNA (LINE, SINE, ALU, B1 sequences, etc.). Moreover, the same SNP promoting or shielding contexts were observed in both exonic DNA (where ancient transposable or repetitive elements rarely occur) and the noncoding compartment, indicating that purifying selection for improved protein folding or function does not account for the contextual bias described here.

The frequent observation of cytosine immediately 5′ to A⇔G polymorphism is likely attributable (at least in part) to DNA methylation on the complementary strand, followed by deamination (resulting in G:T mispairing). However, this cannot explain SNP-associated motifs in nontransition categories (e.g., A⇔C, C⇔G, etc.; supplementary tables S1 and S3, Supplementary Material online), nor does methylation account for overrepresentation of thymine (or dramatic underrepresentation of adenine) immediately 3′ to A⇔G base replacement (table 1). In addition, CG motifs were predictive not only of DNA transition but also A⇔C polymorphism (supplementary table S1, Supplementary Material online; note CG prominence immediately 5′ to A⇔C substitution [5′-CG|XX-3′]; a finding that cannot be attributed to cytosine methylation/deamination). Moreover, many of the 64 possible quartets with "C" preceding an A⇔G transition (5′-XC|XX-3′) showed no evidence of predisposition to SNP formation (table 2). Therefore, although a subset of transition mutations with 5′ cytosine are likely attributable to DNA methylation, this cannot account for either specificity or context of A⇔G SNPs identified here (preponderance of 5′-XC|TX-3′ and underrepresentation of 5′-XC|AX-3′), or the strong contextual patterns observed for other SNP categories.

Considerations Regarding Murine Heterozygosity

Heterozygosity among inbred murine strains was reported by Sanger when thresholds for sequencing accuracy were set at high stringency. Unlike the analysis described above which delineates homozygous SNP locations (the majority of which are typically attributed to alleles from genetic founders of murine lines and therefore comparatively ancient), authentic heterozygous positions are likely to reflect recent mutations on an otherwise highly inbred background (up to 290 filial inbreedings for certain strains tested here). For example, in a murine line that has undergone 50 filial crossings, 99.998% of ancestral heterozygous loci should be fixed; that is, only approximately 2 in 100,000 of the originally heterozygous positions should remain heterozygous (Bailey 1978; Green 1981; Silver 1995; Peters 2007). This calculation is conservative, as it does not account for finite genome size or the fact that physically linked SNPs are nonrandomly assorted (correction for either factor would substantially decrease likelihood of observing heterozygosity). Furthermore, because all classically inbred murine lines exceed filial generation 50 (F_{50}), these mice are widely viewed as homozygous at every position in the genome (Peters 2007), barring a recent single nucleotide replacement.

Table 1

Overrepresented (Permissive) 5′-XC|XX-3′ Quartets Surrounding A⇔G Transition SNPs

A	Rank	Dinucleotide Quartet	*P* Value	Frequency (%)		Fold Increase in Frequency	Rank		
				Occurrence Observed	Occurrence Expected		Human Introns	Murine Introns	Murine Intergenic
	1	TC|TA	3.1E-134	0.699	0.114	6.153	—	—	—
	2	TC|TC	7.4E-125	1.379	0.368	3.742	—	—	15
	3	GC|TT	1.21E-98	0.820	0.184	4.448	—	16	6
	4	CC|TC	1.24E-84	1.243	0.390	3.184	8	—	11
	5	GC|TC	8.82E-68	0.951	0.292	3.256	15	6	16
	6	CC|TT	2.17E-65	0.815	0.235	3.473	11	8	12
	7	TC|TT	4.14E-63	0.806	0.235	3.427	—	—	8
	8	GC|TA	3.4E-61	0.466	0.099	4.692	16	9	13
	9	TC|GG	1.64E-58	0.985	0.333	2.958	—	—	18
	10	AC|TA	1.03E-56	0.539	0.132	4.070	—	—	14
	11	CC|TA	1.31E-54	0.544	0.138	3.953	9	7	10
	12	AC|TT	1.94E-51	0.748	0.235	3.186	14	—	5
	13	AC|TC	8.94E-47	0.898	0.325	2.767	—	13	7
	14	AC|GT	2.06E-41	0.849	0.317	2.678	—	12	—
	15	TC|GA	4.51E-41	1.087	0.454	2.394	—	—	—
	16	AC|CT	1.19E-34	0.723	0.273	2.649	—	11	17
	17	GC|CT	3.84E-30	0.718	0.288	2.490	—	14	22
	18	TC|TG	5.1E-30	0.748	0.306	2.444	17	5	4
	19	CC|TG	2.19E-29	0.961	0.439	2.190	2	3	3
	20	CC|CT	4.42E-29	0.825	0.357	2.314	—	—	—

B	Rank	Dinucleotide Quartet	*P* Value	Frequency (%)		Fold Increase in Frequency	Rank		
				Occurrence Observed	Occurrence Expected		Human Introns	Murine Exons	Murine Intergenic
	1	AC|TG	2.010E-108	1.150	0.340	3.383	12	12	1
	2	GC|TG	5.949E-96	0.953	0.272	3.506	1	8	2
	3	CC|TG	4.733E-79	0.957	0.304	3.145	2	6	3
	4	GC|CA	8.703E-53	0.728	0.249	2.930	—	—	9
	5	TC|TG	2.649E-41	0.756	0.295	2.558	17	1	4
	6	GC|TC	1.079E-37	0.524	0.180	2.907	15	14	16
	7	CC|TA	4.196E-32	0.520	0.193	2.694	9	2	10
	8	CC|TT	5.829E-31	0.705	0.303	2.323	11	-	12
	9	GC|TA	4.492E-30	0.488	0.182	2.687	16	4	13
	10	AC|GG	1.757E-27	0.650	0.285	2.283	—	—	20
	11	AC|CT	1.866E-26	0.618	0.270	2.292	—	—	17
	12	AC|GT	5.827E-26	0.669	0.304	2.205	—	7	—
	13	AC|TC	2.766E-24	0.500	0.208	2.408	—	5	7
	14	GC|CT	2.472E-22	0.472	0.199	2.377	—	—	22
	15	AC|CC	7.458E-22	0.453	0.189	2.396	5	—	—
	16	GC|TT	5.879E-19	0.532	0.251	2.122	—	—	6
	17	GC|CC	2.683E-16	0.327	0.136	2.408	3	—	—
	18	GT|CA	1.215E-12	0.638	0.367	1.740	—	—	19
	19	GC|CG	1.533E-12	0.150	0.049	3.040	6	—	21
	20	GG|CA	4.365E-10	0.622	0.379	1.640	—	—	28

C	Rank	Dinucleotide Quartet	*P* Value	Frequency (%)		Fold Increase in Frequency	Rank		
				Occurrence Observed	Occurrence Expected		Human Introns	Murine Exons	Murine Introns
	1	AC|TG	6.089E-289	1.126	0.327	3.440	12	12	1
	2	GC|TG	3.910E-268	0.888	0.235	3.773	1	8	2
	3	CC|TG	1.322E-232	0.953	0.285	3.346	2	6	3
	4	TC|TG	1.323E-133	0.780	0.280	2.788	17	1	5
	5	AC|TT	5.371E-84	0.800	0.355	2.254	14	3	—

(continued)

Table 1 Continued

C	Rank	Dinucleotide Quartet	*P* Value	Frequency (%)		Fold Increase in Frequency	Rank		
				Occurrence Observed	Occurrence Expected		Human Introns	Murine Exons	Murine Introns
	6	GC|TT	6.665E-65	0.539	0.227	2.374	—	—	16
	7	AC|TC	1.210E-64	0.529	0.221	2.389	—	5	13
	8	TC|TT	3.628E-62	0.755	0.367	2.056	—	—	—
	9	GC|CA	1.001E-61	0.563	0.246	2.286	—	—	4
	10	CC|TA	1.002E-60	0.482	0.199	2.419	9	—	7
	11	CC|TC	7.293E-58	0.560	0.251	2.232	8	10	—
	12	CC|TT	8.402E-58	0.631	0.296	2.134	11	—	8
	13	GC|TA	1.002E-57	0.439	0.178	2.461	16	4	9
	14	AC|TA	7.519E-52	0.690	0.347	1.990	—	11	—
	15	TC|TC	2.945E-50	0.508	0.232	2.192	—	15	—
	16	GC|TC	3.995E-48	0.403	0.171	2.358	15	14	6
	17	AC|CT	2.395E-47	0.570	0.277	2.056	—	—	11
	18	TC|GG	1.429E-45	0.581	0.288	2.016	—	13	—
	19	GT|CA	1.993E-36	0.618	0.336	1.836	—	—	18
	20	AC|GG	1.272E-35	0.554	0.294	1.885	—	—	10

D	Rank	Dinucleotide Quartet	*P* Value	Frequency (%)		Fold Increase in Frequency	Rank		
				Occurrence Observed	Occurrence Expected		Murine Exons	Murine Introns	Murine Intergenic
	1	GC|TG	3.682E-125	1.683	0.235	7.148	8	1	2
	2	CC|TG	3.891E-33	1.091	0.285	3.828	6	3	3
	3	GC|CC	6.589E-30	0.654	0.132	4.961	—	17	—
	4	CC|CC	2.250E-29	0.950	0.246	3.858	—	—	—
	5	AC|CC	1.447E-26	0.810	0.203	3.991	—	15	—
	6	GC|CG	2.807E-26	0.327	0.043	7.529	—	19	21
	7	CC|GG	7.195E-24	1.044	0.323	3.234	—	—	—
	8	CC|TC	8.408E-23	0.873	0.251	3.478	10	—	11
	9	CC|TA	2.512E-22	0.748	0.199	3.755	—	7	10
	10	CT|TA	1.750E-21	1.262	0.455	2.775	—	—	—
	11	CC|TT	8.346E-19	0.904	0.296	3.056	—	8	12
	12	AC|TG	8.919E-19	0.966	0.327	2.951	12	1	1
	13	TC|CT	2.110E-17	0.857	0.285	3.010	—	—	—
	14	AC|TT	2.371E-15	0.950	0.355	2.679	3	—	5
	15	GC|TC	1.280E-14	0.577	0.171	3.370	5	6	16
	16	GC|TA	1.283E-14	0.592	0.178	3.320	8	9	13
	17	TC|TG	1.492E-14	0.795	0.280	2.839	18	5	4
	18	GC|GT	6.518E-12	0.577	0.193	2.991	24	—	—
	19	CC|AG	3.315E-11	0.717	0.275	2.604	—	—	—
	20	GC|GG	4.116E-11	0.654	0.242	2.705	—	—	—

NOTE.—(A) Statistically overrepresented dinucleotide quartets surrounding 20,603 A⇔G coding (exonic) SNPs from murine chromosomes 1–8 (vertical line in each quartet indicates position of polymorphic base). Dinucleotide quartets were defined as two base pairs upstream and two downstream of a polymorphic site or other nucleotide position being evaluated. Rankings that strongly overlap between exonic murine SNPs and other murine and human SNP categories (quartets from among 256 possibilities significantly overrepresented in three of four murine and human DNA compartments) are indicated by yellow highlight; (B) same analysis for 50,244 intronic A⇔G SNPs from murine chromosomes 1–3; (C) findings for 67,663 intergenic A⇔G SNPs for murine chromosomes 1–3; (D) findings for 6,419 A⇔G SNPs from introns of 22 human genes. "—," nonoverlapping in range shown. Note increased incidence of 5′ cytosine and 3′ thymine in quartet motifs predictive of SNP location.

If one applies a conservative estimate of heterozygosity for ancestral (founder) alleles (e.g., 1 in 1,500 genomic positions), no more than 1 in 70 million locations among modern lines should remain heterozygous after 50 filial generations of inbreeding. Authentic heterozygosity is therefore likely to be quite recent.

A conceptual framework for estimating steady-state levels of population-based heterozygosity has been described previously (Charlesworth 2009; Lynch 2010). For example, the sex-averaged germline substitution rate among murine strains (μ) is approximately 30×10^{-9} SNPs per base pair per generation (Lynch 2010). An equilibrium level of heterozygosity (π_s) can be calculated as $4N_e\mu$ (Charlesworth 2009; Lynch 2010), where N_e is the effective population size (a value of 2 in the setting of filial inbreeding). For a haploid genomic region of 500 Mb, therefore, approximately 120 (i.e., $4 \times 2 \times \mu \times 500{,}000{,}000$ bp) steady-state heterozygous positions would be expected per generation on murine chromosomes 1–3

Table 2

Nonpermissive 5′-XC|XX-3′ Quartets Surrounding A⇔G Transition SNPs

A	Dinucleotide Quartet	Frequency (%)	
		Occurrence Observed	Occurrence Expected
	TC|AC	0.277	0.275
	CC|AG	0.597	0.630
	GC|AG	0.388	0.400
	AC|AG	0.456	0.444
	GC|AC	0.267	0.256
	TA|CG	0.116	0.100
	TC|AG	0.393	0.427
	AC|AC	0.340	0.306
	CC|AA	0.461	0.390
B	**Dinucleotide Quartet**	**Frequency (%)**	
		Occurrence Observed	Occurrence Expected
	CC|GC	0.362	0.286
	GC|AA	0.283	0.235
	TC|AC	0.185	0.148
	TC|CG	0.060	0.042
	TC|AA	0.366	0.340
	CC|CG	0.062	0.052
C	**Dinucleotide Quartet**	**Frequency (%)**	
		Occurrence Observed	Occurrence Expected
	AC|AA	0.473	0.514
	CC|AA	0.386	0.324
	TC|AA	0.372	0.385
	CC|CC	0.324	0.246
	CC|GC	0.300	0.268
	AC|AC	0.288	0.222
	TC|AT	0.279	0.254
	GC|AA	0.269	0.266
	GC|AT	0.247	0.176
	GC|GC	0.222	0.197
	CC|AC	0.207	0.188
	TC|AC	0.188	0.183
	GC|AC	0.185	0.129
	CC|CG	0.081	0.052
	TC|CG	0.069	0.040

NOTE.—Nonpermissive 5′-XC|XX-3′ quartets in the setting of (A) 20,603 exonic (chromosomes 1–8), (B) 50,244 intronic (chromosomes 1–4), and (C) 67,663 intergenic (chromosomes 1–4) A⇔G homozygous SNPs. Vertical line in each quartet indicates position of polymorphic base. Yellow highlight indicates nonpermissiveness for SNP formation in at least two of the three regions shown in (A)–(C). Numerous contexts with cytosine immediately 5′ do not predict the location of an A⇔G SNP. All quartet *P* values are greater than 0.05 (Compare with table 1).

for each inbred strain, or 1,900 heterozygous positions among 16 murine lines analyzed here. (Murine strain AKR was omitted from the analysis based on an aberrantly high spontaneous mutation rate [Schlager and Dickie 1967].) This value represents a lower limit for the equilibrium level of SNPs, as the estimate does not account for somatic mutations early in development.

When we tested heterozygosity within murine chromosomes 1–3 (depth > 30; confidence [Phred] = 60) for 16 inbred strains, 18,558 heterozygous positions were observed. Note that Phred of 60 and depth of 30 represent very stringent benchmarks for SNP identification—the threshold is set to permit $<<1$ in a million miscalls from among approximately 18,500 heterosites. However, because approximately 120 heterosites should have been expected per murine line (i.e., ~1,900 heterosites for 16 strains across ~500 Mb of chromosomes 1–3), one must also consider the possibility that NGS artifact has led to a significant burden of erroneous SNPs.

Studies to Minimize NGS Misalignment and Other Sequence Artifact

High-volume genomic sequencing is subject to miscalls, even when utilizing robust map quality, sequence depth, and Phred. With regard to the threshold for identifying authentic SNPs, we tested exonic, intronic, and intergenic DNA compartments in mice and human in a rigorous fashion to exclude sequencing error. The use of phred greater than 60 was incorporated to help assure absence of sequence artifact, which we confirmed by detailed inspection of representative SNPs and by utilizing regions with depth coverage greater than 30. Such criteria are exacting, but provide high levels of confidence in the data being evaluated. Supplementary tables S5–S7, Supplementary Material online, describe assessment of erroneous SNP assignment by manual inspection. Polymorphisms with low-quality score, surrounded by inconsistent consensus sequence data, artifactually high "coverage" (pile-up) due to homologous sequences elsewhere in the genome, or clearly duplicated reads, for example, can often be dispatched by direct visualization of a specific genomic interval. Other features of SNP environment—such as location within a short dinucleotide repeat or nearby indel—are sometimes more difficult to evaluate, as these regions are known to be genomically unstable, and represent common sites of true SNP formation (Pearson et al. 2005; Tian et al. 2008; Lopez Castel et al. 2010).

Manual Inspection of Representative Homosites

Supplementary table S5, Supplementary Material online, describes 150 intronic, exonic, and intergenic SNPs selected randomly from among homosites identified by this study. Because all homosites were 1) based on robust Phred score and sequencing depth, 2) required to exhibit the minor allele in multiple distinct strains, and 3) found to occur at roughly the expected incidence of genomic variation among murine lines (i.e., one SNP per every few thousand nucleotide positions), the prior likelihood of error was very low (estimated at $<10^{-22}$ per SNP). This assertion is borne out by the absence

of misalignment, duplicate reads, indels, short local repeats, and so forth in the majority of homozygous SNPs (supplementary table S5, Supplementary Material online). The pattern is similar to a randomly selected region of high-quality DNA sequence data (supplementary table S7, Supplementary Material online) and indicates that the substantial majority of homosites reported here is authentic.

Manual Inspection of Representative Heterosites

A significant number of heterosite positions are clearly artifactual and exhibit surrounding sequence misalignment, duplicate reads from multiple genomic regions, low quality, and so forth (supplementary table S6, Supplementary Material online). Nonetheless, a meaningful subset of representative heterosites (~16%) fail to exhibit any evidence whatsoever of NGS artifact. These heterozygous SNPs exhibit high local mapping scores, depth of coverage, Phred, consistent surrounding sequence, and are without evidence of "pile-up," local read duplication, homologues elsewhere in the genome, and so forth. The sampling analysis therefore suggests that from among 18,558 putative heterozygous positions, much smaller numbers (e.g., ~16% or 3,000) are likely to represent the authentic sites of recent mutation. This agrees well with expected heterozygosity calculated above based on population accumulation and the known murine mutation rate (an estimated 1,900 heterozygous positions at steady state), particularly when one considers that de novo SNPs formed during early embryogenesis would further increase the total number of expected heterosites (i.e., by ~2-fold) above the value shown here (Lynch 2010).

Permissive and Nonpermissive SNP Contexts in the cDNA of Human Genes

We and others have suggested that random mutation accrual over billions of years could otherwise degrade the integrity of core metabolic genes, and that regulatory mechanisms may therefore exist to influence where (and possibly when) SNPs are most likely to occur in genomic DNA (Charlesworth B and Charlesworth D 1997; Loewe and Lamatsch 2008; Hill et al. 2014). One such mechanism is a transition bias that favors both synonymous and conservative exonic SNPs (Collins and Jukes 1994; Wakeley 1994; Freeland and Hurst 1998; Hill et al. 2014). To further investigate relevance of the present findings to exonic patterns of evolutionary SNP accumulation, we located the four greatest and four least permissive quartets for A⇔G polymorphism within cDNAs of *CFTR* and dystrophin, two genes of ancient vertebrate origin (fig. 2). We observed elevated representation of quartets that minimize SNP formation, and underrepresentation of motifs predisposed to augment the accrual of new SNPs. Either of these genes can be lethal when deleted from the mammalian genome, and therefore cannot be taken as representative of exonic DNA as a whole. However, because gene products such as these are likely to incur strong selective pressure, they provide a stringent (and nonneutral) test for SNP distribution bias. From this perspective, because the same contextual preferences shown here also apply across exonic, intronic, and intergenic DNA (fig. 1 and tables 1 and 2), the distributions cannot be ascribed to ongoing natural selection for optimizing or conserving amino acid sequence. Instead, we believe that constraints of this type offer a hint as to mechanisms that underlie SNP production in the murine genome (i.e., at the time of SNP formation, see below).

Codon Usage, Exonic Dinucleotide Representation, and Relevance to SNP Formation Bias

The above analysis indicates that SNPs occur with greater frequency within the immediate vicinity of a CpG (i.e., 5′-CG|-3′), or within a CpG dinucleotide itself (5′-C|-3′). It is of interest that codon usage in mouse and human is underrepresented by CpG dinucleotides. For example, of the multiple nucleotide triplets available for serine, proline, threonine, and alanine (six for serine; four each for proline, threonine, and alanine), CG-containing codons are markedly underutilized. In mice, among six codons that designate serine, the triplet containing CG is preferred only 5.1% ($P < 1.0 \times 10^{-30}$; Materials and Methods). For proline, threonine, and alanine, the triplet containing CG is preferred at 10.3, 10.4, and 9.4%, respectively ($P < 1.0 \times 10^{-30}$ in all cases). Underrepresentation of CG dinucleotides within murine exons (supplementary fig. S3, Supplementary Material online) has been reported previously (International Human Genome Sequencing Consortium et al. 2001), connotes significance of the results shown in figure 2, and, based on findings presented here, would diminish exonic SNP formation and help preserve protein-coding DNA over the evolutionary timescale.

Discussion

Our results demonstrate that nucleotide positions within murine and human DNA have distinct likelihoods of single base replacement that can be predicted in part from local sequence environment (e.g., see supplementary fig. S2, Supplementary Material online). A preference for transition substitutions (A⇔G and C⇔T), as well as transversion SNPs, was observed in the immediate vicinity of well-defined DNA quartets (fig. 1, supplementary figs. S1 and S2, Supplementary Material online, tables 1 and 2, and supplementary tables S1 and S3, Supplementary Material online). The findings are not compatible with a "neutral"-type DNA evolutionary model, as SNP accumulation genome wide is strongly nonrandom, as evidenced by local context, predisposition toward synonymous alterations (table 3), as well as a tendency to preserve exonic sequence (fig. 2 and supplementary fig. S3, Supplementary Material online). Frame of reference for these studies was based on expected incidence for random (i.e., stochastic) SNP accumulation.

The quartets described here are not attributable (at least in strict sense) to conventional evolutionary models in which SNP distribution is due to selective pressures that govern (along with drift, shift, hitchhiking, etc.) the tendency for allelic fixation. The same quartet preferences observed in coding DNA were also noted in noncoding (regulatory) DNA in mouse and human (yellow highlight, table 1), providing evidence that SNP-associated motifs are not a consequence of purifying selection, but are the result of a mechanism that governs their formation. This is particularly true with regard to the highly conserved sequences within nonprotein coding or regulatory DNA (ENCODE Project Consortium 2012; Gerstein et al. 2012), as there would be no reason to expect that the same quartets that provide an exonic (protein based) fitness advantage would also be advantageous for the intronic or intergenic DNA compartments. Our findings therefore support the existence of previously undescribed mechanism(s) by which single base replacements exhibit the same biased formation patterns in coding and noncoding, germline and somatic, and both murine and human DNA.

This study also investigated ostensibly more recent genetic polymorphism among 16 murine lines. Because these strains are heavily inbred (to over 50 filial matings), homozygosity at virtually every position would be expected, and the likelihood of a heterozygous site due to anything other than recent mutation is remote (Bailey 1978; Green 1981). Such strains are removed from many forms of selection that act during the evolutionary time scale; that is, most new single base replacements described above were generated in a "minimally selective" laboratory environment. We identified approximately 3,000 very recent SNPs across a 500-Mb region of DNA in multiple strains. The true extent of recent polymorphism must be significantly greater if one considers the high rate of fixation (drift; $1/2N_e$) for de novo heterosites (e.g., >90% of recent heterozygous positions are expected to become fixed after just 40 inbred generations [Bailey 1978; Green 1981]). This means a sizable majority of heterosites during 50–100 years of inbreeding should become fixed as homosites (typically retaining the original allele). Interestingly, when precise location for heterozygous SNPs was compared with sites with at least one (of 16) strain homozygous for the same minor allele, identity was substantial, in that 25–30% of all heterosites in one strain were also represented by a homosite in at least one other strain at exactly the same position. Moreover, although heterosites were rare, approximately 20% were represented in more than one strain as heterozygous. As the likelihood of a homosite miscall is negligible (<1 in 10^{22}), such findings provide further support for the notion that significant numbers of heterosites have formed independently and in multiple strains (parallel evolution) and subsequently become "fixed"; that is, detected as homozygous SNPs of very high quality judged by Phred score, depth, mapping quality, and other criteria described above. For the frequent parallel variants described here (at least 20–25% of all recent point mutations), SNP formation is approximately 9 orders of magnitude greater than expected on a stochastic basis. Quantitative estimates of murine evolutionary rate have not previously considered parallel SNP formation or the

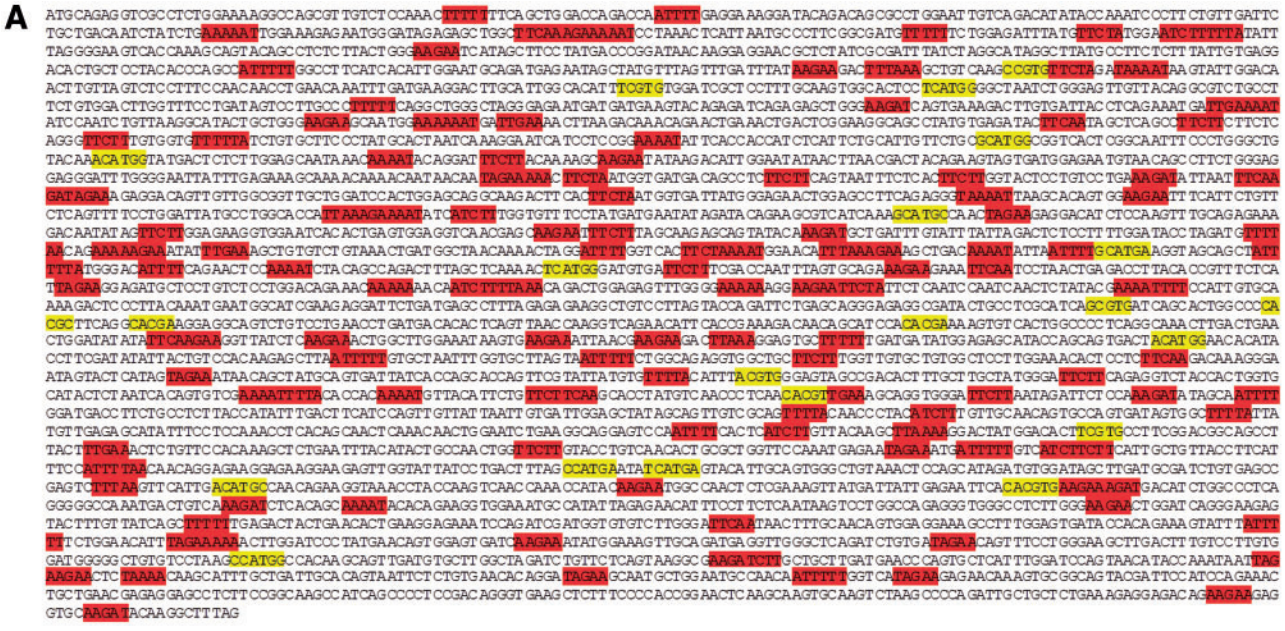

FIG. 2.—SNP-permissive and shielding quartets in human ORFs. (*A*) Most frequent SNP-permissive (yellow) and SNP-shielding (red) A⇔G quartet contexts within the cDNA of *CFTR*. (*B*) Same analysis for cDNA of dystrophin. A statistical (chi square 2 × 2 contingency table) analysis indicated overrepresentation of shielding dinucleotide quartets in both genes ($P < 0.00001$), and underrepresentation of quartets permissive for SNP formation ($P < 0.00001$).

(continued)

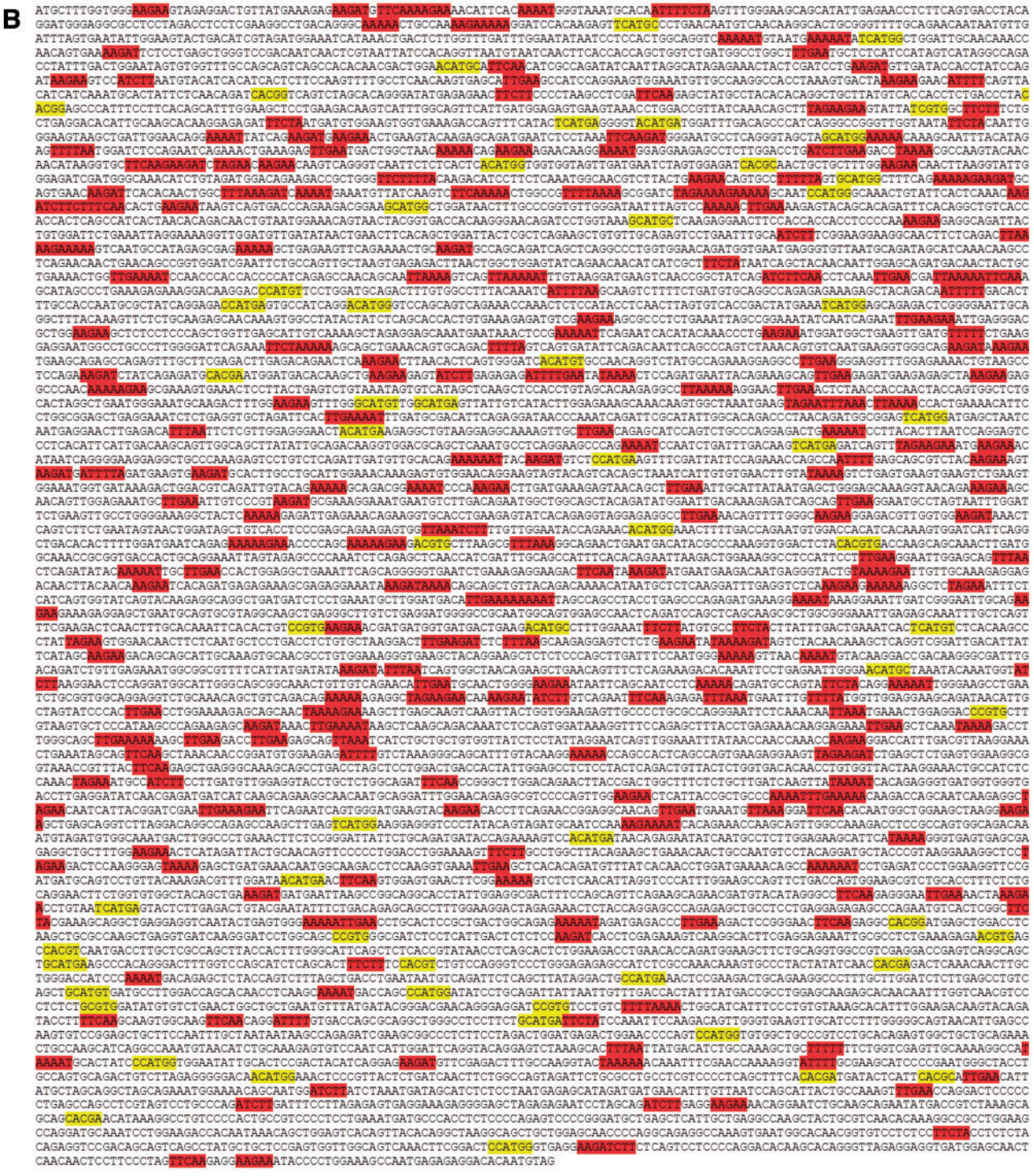

FIG. 2.—Continued.

positional constraints that appear to underlie a very large subset of newly accumulating SNPs.

Although it is expedient (and appealing) to simply attribute an unexpectedly large number of parallel SNPs to NGS artifact, we stringently tested this possibility and were unable to arrive at such a conclusion. Care was taken to exclude sequence artifact as a confounding variable. As noted above, the likelihood that a homozygous SNP in 4 of 17 murine strains represents sequence error is minimal (<1 in 10^{22} based on sequencing depth and high Phred quality) and manual SNP

Table 3

Comparison of Transition: Transversion and Nonsynonymous:Synonymous SNP Frequencies in Homozygous (Homosite) versus Heterozygous (Heterosite) Murine SNPs

SNP Category	Percent Transition	Percent Transversion	Nonsynonymous (NS) Coding SNPs	Synonymous (S) Coding SNPs	NS:S (*P* Value[a])
Homozygous	77.7	22.3	708	1,245	1:1.76 (1.65E-14)
Heterozygous	66.3	33.7	272	425	1:1.56 (2.64E-11)

NOTE.—Murine homosites from chromosomes 1–3 with at least four lines exhibiting the minor allele are shown.

[a]*P* values represent observed SNP frequencies versus those that would be expected if SNPs formed stochastically (Materials and Methods).

assessment indicates authenticity of the homosites identified here. In addition, DNA contexts found by our studies to predict SNP location have not been associated with sequencing error in the past, and motifs shown previously to increase rates of sequence artifact (e.g., SNPs preceded by "G" immediately 5′, poly-A tracts, etc. [Dohm et al. 2008; Harismendy et al. 2009]) were not identified as "permissive" (table 1 and supplementary tables S1–S3, Supplementary Material online). Manual analysis of local DNA environment did implicate NGS artifact as a primary source for most heterozygous calls, even at 30-fold coverage. Nonetheless, a meaningful subset of heterosites identified by Sanger appear authentic (supplementary table S6, Supplementary Material online). Also, as shown in table 3, enhancement of synonymous (vs. nonsynonymous) heterosite SNPs and a very strong transition bias (neither of which are associated with sequencing error) was observed in the heterosite population, further indicating authenticity. Heterozygous SNPs exhibited a synonymous:nonsynonymous ratio of approximately 1.6:1, which is very similar to the homosite value (1.76:1). Moreover, when we used established methods to estimate the equilibrium level of heterozygosity (π_S) expected from germline mutation among 16 strains, we obtained a value of approximately 1,900 heterosites across 500 Mb (chromosomes 1–3), which is in reasonable agreement with a conservative estimate of approximately 3,000 authentic variants. Finally, even the most restrictive estimates indicate that meaningful numbers of heterozygous SNPs described by Sanger must exist. If all heterozygous positions are artifact, the present findings contradict known mutation rates in murine DNA, and debase a large number of past and ongoing genome-scale sequencing projects in multiple species, including human, that employ leading-edge data acquisition and analysis methods comparable to those used here. For the present interpretation of our findings to be discounted, therefore, one must assert that 1) parallel SNPs observed in our experiments are largely the result of NGS artifact (despite manual evaluation and other evidence to the contrary); 2) sequence artifact exhibits an inexplicable transition and synonymous bias, very similar to what occurs in living cells; 3) both heterozygous and homozygous calls are grossly in error (i.e., ~25% of homozygosity is seriously tainted by NGS artifact); and 4) not only are the identified heterozygous SNPs incorrect but also the true heterozygous positions (of which 2,000–3,000 would be expected) are missing—and not detectable by the best available DNA sequencing technology.

As expected, well-validated heterozygous SNPs were rare, yet a high degree of parallel occurrence with homosites (25–30%) and other heterosites (20%) was observed after just a few decades of laboratory breeding. We believe that still higher levels of concordance would be obtained were additional time allowed for SNPs to accumulate, or if it were possible to directly measure the number of recent SNPs that have subsequently undergone fixation. In either case, the extent of strain propagation in the present studies essentially precludes ancestral haplotype as an explanation for observed patterns of heterozygosity, and instead points to a robust positional bias for recent mutation. The finding of context dependent, parallel, and recent SNP formation (many orders of magnitude beyond that predicted on stochastic basis) has not been considered in previous studies of murine genomic mutation rate (Ellison et al. 1996; Ananda et al. 2011), evolutionary "clocks" based on SNP genesis (Easteal et al. 1995; Hedges and Kumar 2003), ultravariable versus ultraconserved DNA otherwise ascribed to purifying selection (Ellison et al. 1996; Ahituv et al. 2007), or somatic mutational patterns in neoplasia (Song et al. 2013), but should be considered as part of future analyses in these areas.

We and others have characterized longevity of core metabolic genes in the face of a mutational "ratchet" (that over hundreds of millions of years would be capable of decimating eukaryotic exons), and suggested existence of adaptive mechanisms that regulate DNA mutation and serve to promote long-term genomic survival (Gabriel et al. 1993; Lynch 2010; Koonin 2012; Hill et al. 2014). Findings from the present study furnish new evidence in support of this hypothesis. We show that SNP distribution in congenic mice exhibits contextual bias that may divert single nucleotide variants away from protein-coding DNA (fig. 2 and supplementary fig. S3, Supplementary Material online). We also provide evidence that recently acquired (heterozygous) SNPs (produced in laboratory mice under minimal selective pressure) nonetheless exhibit a strong synonymous predisposition (table 3). In addition, our data point to modes of rapid DNA evolution restricted by specific sequences (e.g., CG dinucleotides) repletes in noncoding DNA (table 1 and supplementary table S1, Supplementary Material online). Contextual and other SNP preferences must,

in part, reflect rates of nucleotide misincorporation, errors during proofreading, biased gene conversion, and/or failure to conduct mismatch-mediated repair with regard to certain DNA motifs. Findings presented here provide evidence that mutational fault tolerance has been adapted to spare eukaryotic reading frames. We also note that highly specialized sequence motifs favoring SNPs within promoters and other crucial regions of noncoding DNA (e.g., CpG islands) could serve to preferentially facilitate polymorphism and diversity in a manner that directs single base mutations to the regulatory compartment (ENCODE Project Consortium 2012; Gerstein et al. 2012), while shielding against uncontrolled mutation accrual within essential protein-coding elements (table 1 and supplementary table S1 and fig. S3, Supplementary Material online) (Hill et al. 2014). In either case, future studies of mouse genomic evolution should consider the role of contextual preference and parallel SNP formation shown here for highly inbred murine strains.

Supplementary Material

Supplementary figures S1–S3 and tables S1–S7 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Dr J. Hartman and Dr R. Oster for review of the article and valuable suggestions. They also thank Jenny Mott, Jan Tindall, and Cheryl Owens for help preparing the manuscript.

Literature Cited

Ahituv N, et al. 2007. Deletion of ultraconserved elements yields viable mice. PLoS Biol. 5(9):e234.
Ananda G, Chiaromonte F, Makova KD. 2011. A genome-wide view of mutation rate co-variation using multivariate analyses. Genome Biol. 12:R27.
Bailey DW. 1978. Sources of subline divergence and their relative importance for sublines of six major inbred strains of mice. In: Morse HC, editor. Origins of inbred mice. New York: Academic Press.
Billings T, et al. 2010. Patterns of recombination activity on mouse chromosome 11 revealed by high resolution mapping. PLoS One 5(12):e15340.
Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 10:195–205.
Charlesworth B, Charlesworth D. 1997. Rapid fixation of deleterious alleles can be caused by Muller's ratchet. Genet Res. 70(1):63–73.
Collins DW, Jukes TH. 1994. Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics 20(3):386–396.
Dohm J, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36(16):e105.
Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics 148(4):1667–1686.
Duret L. 2009. Mutation patterns in human genome: more variable than expected. PLoS Biol. 7(2):e1000028.
Easteal S, Collet C, Betty D. 1995. The mammalian molecular clock. Austin (TX): RG Landes.
Ellison JW, Li X, Francke U, Shapiro LJ. 1996. Rapid evolution of human pseudoautosomal genes and their mouse homologs. Mamm Genome. 7(1):25–30.
ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.
Fortini ME, Skupski MP, Boguski MS, Hariharan IK. 2000. A survey of human disease gene counterparts in the *Drosophila* genome. J Cell Biol. 150(2):F23–F30.
Freeland SJ, Hurst LD. 1998. The genetic code is one in a million. J Mol Evol. 47(3):238–248.
Gabriel W, Lynch M, Bürger R. 1993. Muller's Ratchet and mutational meltdowns. Evolution 47:1744–1757.
Gerstein MB, et al. 2012. Architecture of the human regulatory network derived from ENCODE data. Nature 489(7414):91–100.
Green EL. 1981. Genetics and probability in animal breeding experiments. New York: Oxford University Press.
Harismendy O, et al. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 10(3):R32.
Hedges SB, Kumar S. 2003. Genomic clocks and evolutionary timescales. Trends Genet. 19(4):200–206.
Hill AE, et al. 2014. Longevity and plasticity of *CFTR* provide an argument for noncanonical SNP organization in hominid DNA. PLoS One 9:e109186.
Hodgkinson A, Eyre-Walker A. 2010. The genomic distribution and local context of coincident SNPs in human and chimpanzee. Genome Biol Evol. 2:547–557.
Hodgkinson A, Ladoukakis E, Eyre-Walker A. 2009. Cryptic variation in the human mutation rate. PLoS Biol. 7(2):e1000027.
International Human Genome Sequencing Consortium, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409 (6822):860–921.
Keane TM, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477(7364):289–294.
Koch RE. 1971. The influence of neighboring base pairs upon base-pair substitution rates. Proc Natl Acad Sci U S A. 68(4):773–776.
Koonin EV. 2012. The Logic of Chance—the natures and origin of biological evolution. 1st ed. Upper Saddle River (NJ): Pearson.
Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet. 63:474–488.
Lederberg JA, Lederberg EM. 1952. Replica plating and indirect selection of bacterial mutants. J Bacteriol. 63(3):399–406.
Loewe L, Lamatsch DK. 2008. Quantifying the threat of extinction from Muller's ratchet in the diploid Amazon molly (*Poecilia formosa*). BMC Evol Biol. 8:88.
Lopez Castel A, Cleary JD, Pearson CE. 2010. Repeat instability as the basis for human diseases and as a potential target for therapy. Nat Rev Mol Cell Biol. 11(3):165–170.
Lynch M. 2010. Evolution of the mutation rate. Trends Genet. 26(8):345–352.
Miller DE, et al. 2012. A whole-chromosome analysis of meiotic recombination in *Drosophila melanogaster*. G3 (Bethesda) 2(2):249–260.
Pearson CE, Nichol Edamura K, Cleary JD. 2005. Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet. 6(10):729–742.
Peters LL. 2007. The mouse as a model for human biology: a resource guide for complex trait analysis. Nat Rev Genet. 8(1):58–69.
Rice WR. 1988. Analyzing tables of statistical tests. Evolution 43(1):223–225.
Rosner B. 2011. Fundamentals of biostatistics. 7th ed. Boston: Brooks/Cole. Cengage Learning.

Schlager G, Dickie MM. 1967. Spontaneous mutations and mutation rates in the house mouse. Genetics 57:319–330.
Silver LM. 1995. Mouse genetics: concepts and applications. New York: Oxford University Press.
Song Y, et al. 2013. Evolutionary etiology of high-grade astrocytomas. Proc Natl Acad Sci U S A. 110(44):17933–17938.
Sung W, Ackerman M, Miller S, Doak T, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci U S A. 109(45):18488–18492.
Tian D, et al. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature 455(7209):105–109.
Wakeley J. 1994. Substitution-rate variation among sites and the estimation of transition bias. Mol Biol Evol. 11(3):436–442.
Xue Y, et al. 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol. 19(17):1453–1457.
Yalcin B, et al. 2011. Sequence-based characterization of structural variation in the mouse genome. Nature 477(7364):326–329.
Yates F. 1934. Contingency table involving small numbers and the $\chi 2$ test. Suppl J R Stat Soc. 1(2):217–235.
Zhang F, Zhao Z. 2005. SNPNB: analyzing neighboring-nucleotide biases on single nucleotide polymorphisms (SNPs). Bioinformatics 21(10):158–168.
Zhang J, et al. 2005. A high-resolution multistrain haplotype analysis of laboratory mouse genome reveals three distinctive genetic variation patterns. Genome Res. 15(2):241–249.
Zhao Z, Boerwinkle E. 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. Genome Res. 12(11):1679–1686.
Zhao Z, Zhan F. 2004. The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. Genomics 84:785–795.