Published in final edited form as:

Evolution. 2014 September ; 68(9): 2737–2744. doi:10.1111/evo.12460.

REMARKABLE SELECTIVE CONSTRAINTS ON EXONIC DINUCLEOTIDE REPEATS

Ryan J. Haas^{1,2,3} and Bret A. Payseur¹

¹Laboratory of Genetics, University of Wisconsin–Madison, Madison, Wisconsin 53706

Abstract

Long dinucleotide repeats found in exons present a substantial mutational hazard: mutations at these loci occur often and generate frame shifts. Here, we provide clear and compelling evidence that exonic dinucleotides experience strong selective constraint. In humans, only 18 exonic dinucleotides have repeat lengths greater than six, which contrasts sharply with the genome-wide distribution of dinucleotides. We genotyped each of these dinucleotides in 200 humans from eight 1000 Genomes Project populations and found a near-absence of polymorphism. More remarkably, divergence data demonstrate that repeat lengths have been conserved across the primate phylogeny in spite of what is likely considerable mutational pressure. Coalescent simulations show that even a very low mutation rate at these loci fails to explain the anomalous patterns of polymorphism and divergence. Our data support two related selective constraints on the evolution of exonic dinucleotides: a short-term intolerance for any change to repeat length and a long-term prevention of increases to repeat length. In general, our results implicate purifying selection as the force that eliminates new, deleterious mutants at exonic dinucleotides. We briefly discuss the evolution of the longest exonic dinucleotide in the human genome—a $10 \times CA$ repeat in *fibroblast growth factor receptor-like 1* (FGFRL1)—that should possess a considerably greater mutation rate than any other exonic dinucleotide and therefore generate a large number of deleterious variants.

Keywords

Mutations; natural selection; polymorphism; population genetics

Exonic dinucleotide microsatellites are hazardous genetic variants (Li et al. 2004). The large majority of mutations at dinucleotide repeats yields mutant alleles that differ from parent alleles by two or four base pairs (Sun et al. 2012). In exons, mutations of this size result in frameshifts that will frequently disrupt protein function. Moreover, dinucleotide mutation is not rare; the estimated rate of dinucleotide mutation in humans ($10^{-6} - 10^{-4}$; Sun et al. 2012; Baptiste et al. 2013) is orders of magnitude greater than that of point mutation (10^{-8} ;

³haaslr@uwplatt.edu.

²Current Address: Department of Biology, University of Wisconsin–Platteville, Platteville, Wisconsin 53818

The authors report no conflicts of interest.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Table S1: dN/dS ratios for the 18 genes harboring long exonic dinucleotides in human.

Table S2: Accession numbers for the cDNA and amino acid sequences used in PAML analysis.

Lynch 2010; Roach et al. 2010). Together, high potential for deleterious mutation and high mutation rate accord exonic dinucleotides with a dangerous volatility. Importantly, the mutational hazard presented by exonic dinucleotides should increase with repeat length, as microsatellite length is a primary, positive determinant of microsatellite mutation rate (Pemberton et al. 2009; Payseur et al. 2011; Sun et al. 2012; Ananda et al. 2013; Baptiste et al. 2013).

These facts suggest the existence of two related selective constraints at exonic dinucleotides. First, at any given point in the evolutionary history of a species, exonic dinucleotides should be constrained to maintain current repeat length due to the deleterious nature of most mutations at these loci. Second, over long periods of evolutionary history, there should be a tendency to maintain exonic dinucleotides at short repeat lengths, which keeps mutation rate low and thereby reduces genetic load in populations. Although several studies have examined the frequency and polymorphism of dinucleotides in untranslated regions, introns, and exons (e.g., Pemberton et al. 2009; Payseur et al. 2011; Baptiste et al. 2013), the nature of selective constraints on exonic dinucleotides remains poorly documented and poorly understood. Here, we demonstrate that patterns of polymorphism and divergence at exonic dinucleotides provide compelling support for both hypothesized selective constraints.

Dinucleotides appear to gain mutational instability at repeat lengths greater than or equal to five (Kelkar et al. 2010; Ananda et al. 2013). Scanning the human reference genome sequence (assembly GRCh37; RefSeq gene annotations) with the tandem repeat finder SciRoKo (Kofler et al. 2007), we found several hundred perfect exonic dinucleotides of repeat length greater than or equal to five (here, *perfect* refers to the uninterrupted nature of the repeat; e.g., CACACACA is perfect, whereas CACACTCA contains one imperfection). However, we only found 18 perfect dinucleotides of repeat length greater than or equal to seven that also fully overlapped exonic sequence (Table 1) despite the presence of at least 105,584 perfect dinucleotides of this size across the entire human genome (Payseur et al. 2011). Assuming 1% of the human genome is exonic (Ng et al. 2009), we expect 1056 of these dinucleotide microsatellites to fully overlap exons by chance. The observed number of 18 is therefore highly unlikely (χ^2 -test, $P < 10^{-15}$), supporting the idea that purifying selection prevents long alleles of exonic dinucleotides from accumulating in populations, and, ultimately, species over time. The size distribution of these 18 dinucleotides further supports the hypothesis that long dinucleotides are not tolerated in exons. Although Payseur et al. (2011) found 51,659 perfect dinucleotides of repeat length ≥ 11 across the genome, all 18 exonic dinucleotides in the human reference genome possess repeat lengths < 11 . Further evidence for a constraint against de novo emergence of long, exonic dinucleotides is provided by the chimpanzee genome (*Pan troglodytes*, assembly CHIMP2.1.4; RefSeq gene annotations). Despite the > 6.5 million years separating these two species (Chatterjee et al. 2009), an independent search for long exonic dinucleotides in chimpanzee identified the same set as that found in human with two exceptions: (1) the dinucleotide in *LUC7L2* was missing, as it is imperfect in chimpanzee, and (2) two copies of the *BRD7* dinucleotide due to an apparent duplication of this gene in chimpanzee. Thus, despite millions of years of independent evolution, neither human nor chimpanzee has produced a long, exonic dinucleotide not found in their common ancestor.

To test the hypothesis that exonic dinucleotides are constrained to current repeat length, we genotyped 18 long, exonic dinucleotides as well as several reference dinucleotides in 200 humans (400 chromosomes for each autosomal locus) from the 1000 Genomes project (1000 Genomes Project Consortium 2010). Eight worldwide populations were sampled: Yoruba (Nigeria, YRI; $n = 25$), Luhya (Kenya, LWK; $n = 25$), north-western Europe (CEU; $n = 25$), Toscani (Italy, TSI; $n = 25$), Finnish (FIN; $n = 31$), Gujarati Indian (GIH; $n = 26$), Han Chinese (CHB; $n = 27$), and Mexican (MXL, $n = 16$). Genotyping was conducted by Prevention Genetics (Marshfield, WI). Although high-coverage exome data are available from the 1000 Genomes Project and new methods to call microsatellite alleles from next-generation sequencing reads have been developed (e.g., Gymrek et al. 2012), it can be difficult to call microsatellite alleles with great certainty when using next-generation sequencing data. Therefore, we opted to obtain direct genotypes. Of the 18 exonic dinucleotides in our sample, 17 were monomorphic (observed heterozygosity, $H = 0$; Table 1). The only polymorphic microsatellite was an AG repeat in *CCAR1*. At this locus, two individuals—one each from the African populations of Yoruba (Nigeria) and Luhya (Kenya)—were heterozygous for the common AG₇ allele and a rare AG₆ allele. Subsequently, Sanger sequencing confirmed the heterozygosity of these two individuals.

The nearly complete absence of polymorphism at these loci suggests selection effectively eliminates size mutants from worldwide human populations. We also genotyped a number of reference dinucleotides of similar length in untranslated regions, intergenic sequence, and the exons of unannotated open reading frames. Loci in these nonexonic genomic contexts showed considerably greater polymorphism in the same human sample (Table 1). This result suggests that nonexonic dinucleotides are more free to experience size evolution within populations.

We used patterns of sequence divergence between humans and other primate species to test the hypothesis that there is a long-term selective constraint against the expansion of the 18 dinucleotides in question. These data are useful because they integrate a longer period of time and represent multiple independent instances of the evolution of the anomalous dinucleotides identified in humans. To assess divergence at the dinucleotides we genotyped in human populations, we used the Multiz Alignment and Conservation track of the UCSC Genome Browser (Meyer et al. 2013) to find the repeat number of each orthologous microsatellite in *Pan troglodytes* (chimpanzee), *Gorilla gorilla*, *Pongo pygmaeus* (orangutan), *Macaca mulatta* (rhesus macaque), *Papio hamadryas* (baboon), and *Callithrix jacchus* (marmoset). Remarkably, for most of the 18 long exonic dinucleotides identified in human, the most parsimonious interpretation of divergence data is that the human allele was already present in the common ancestor to all primates. All sampled primates bear the same repeat length as that observed in the human reference sequence at 15 of 18 loci. Of the remaining loci, the dinucleotides in *CCDC30* and *BRD7* show size conservation in the Catarrhini and Hominidae, respectively. The AG repeat in *HMGXB4* shows the least parsimonious pattern of divergence, with repeat lengths of seven and eight observed in the paraphyletic groups of human, chimpanzee, gorilla, and marmoset; and orangutan, macaque, and baboon, respectively. Because these 18 microsatellites have long repeat lengths, and, therefore, appreciable mutation rates, it is reasonable to assume that a very large number of

mutants for these loci have appeared in individuals across the primate phylogeny. Particularly when compared to levels of divergence at nonexonic dinucleotides (Table 1), the virtual absence of divergence suggests that strong and consistent purifying selection has eliminated all emergent mutations and maintained the repeat length found in the primate MRCA.

To obtain a more quantitative assessment of the support for a long-term selective constraint against the expansion of exonic dinucleotides, we next compared the pattern of divergence at the 18 loci identified in human to patterns of divergence at a genomic-scale set of dinucleotides. For each autosomal dinucleotide with perfect repeat length greater than or equal to seven in the human reference sequence, we recorded the repeat length of the orthologous locus in each of the non-human primate species listed above. When calculating empirical probabilities, we only included loci where alignment data were available for all seven primate species ($n = 27,360$).

The >40 million year depth of the primate phylogeny provides ample branch length for the size evolution of neutral microsatellites. In fact, microsatellites are frequently absent from a subset of species in the primate phylogeny due to microsatellite “death” that results from the loss of mutability due to reduction in size along one or more lineages (Kelkar et al. 2011; Sawaya et al. 2012). Thus, the temporal depth of the primate phylogeny suggests that the observed data—18 loci where human repeat length is greater than or equal to seven and all non-human primate species have repeat length greater than or equal to six—should be highly unusual. We find that 5579 (20.4%) of 27,360 autosomal dinucleotides with repeat length greater than or equal to seven in human also have repeat lengths greater than or equal to six in all sampled primates. Thus, the observed data in which 18 of 18 loci meet these conditions are highly unlikely under a null model of neutral evolution, which posits that these 18 loci are simply a random sample of independent dinucleotides ($P = 0.204^{18} = 3 \times 10^{-13}$).

We can also gauge support for the selective constraint against long-term expansion of exonic dinucleotides using variance in repeat length as a measure of divergence. Across the 400 human and six other primate chromosomes, 15 of 18 loci have a variance of 0, two have a variance of 0.286 (*BRD7* and *HMGXB4*), and one (*CCDC30*) has a variance of 0.571 (Table 1). We find that 241 (0.88%), 611 (2.2%), and 1124 (4.11%) of 27,360 genomic dinucleotides, respectively, possess these levels of divergence or less. The joint probability of the 18 observed values of divergence under the null, neutral model is then only $P = 0.0088^{15} \times 0.022^2 \times 0.411 = 3 \times 10^{-36}$.

An alternative explanation for the lack of variation at the 18 exonic dinucleotides is that due to an unknown molecular cause these loci possess unexpectedly low mutation rates relative to other dinucleotides of similar length. We used coalescent simulations of neutral microsatellite evolution (Hudson 2002; Haasl and Payseur 2010) on the primate phylogeny to determine how mutation rate (in the absence of natural selection) affects the probability of our observation: a dinucleotide locus that is monomorphic for repeat length greater than or equal to seven in a sample of 400 human chromosomes and one chromosome each for the primates listed above. This approach was particularly attractive because it allowed us to

simultaneously compare the observed divergence and polymorphism data to simulations that produced both types of data as output.

Because the positive relationship between repeat length and mutability is an important factor in our argument that the length of the 18 loci discussed here makes them anomalous, we used a model of mutation in which mutation rate depends on the current repeat length of the microsatellite (Fig. 1; Haasl and Payseur 2013). In addition, we allowed 5% of mutations to induce size changes of greater than one repeat (following a geometric distribution with parameter $p = 0.95$) and modelled contraction and expansion biases based on recent experimental results documenting both biases in dinucleotide microsatellites (Baptiste et al. 2013). Specifically, we modeled contraction bias as in Haasl and Payseur (2013) with contraction parameter $c = 0.005$ and incorporated recent empirical results regarding expansion bias in dinucleotides (Baptiste et al. 2013) by enforcing a 60% chance of expansion for repeat lengths >11 and <20 . Matching the most parsimonious interpretation of the observed data, we assumed that the most recent common ancestor of the primate phylogeny already possessed a repeat length of seven. We performed one million simulations each for three different mutation curves (Fig. 1; high, medium, and low mutation). The curve corresponding to high mutational pressure (navy) tracks a recent pedigree-based estimate of length-dependent dinucleotide mutation rate based on a sample of $>85,000$ humans (Sun et al. 2012). Despite labeling this the “high” mutation rate model, we note that this is the best empirical estimate of mutation rate available for arbitrary dinucleotides in human. However, because we were concerned with the possibility that very low mutation rates may explain the 18 dinucleotides in question, the other two mutation rate models assume considerably lower mutation rates than that of the best empirical estimate. In each of the one millions simulations, we assumed the phylogeny and point estimates of divergence times in Chatterjee et al. (2009).

0.009, 0.002, and 0% of simulated datasets for the high, medium, and low mutation models, respectively, were monomorphic. In all cases, however, the monomorphic allele had a repeat length <5 . In other words, for a variety of mutational profiles, simulations of neutral evolution categorically failed to recapitulate the observed characteristics of the observed data: monomorphic loci with repeat length greater than or equal to six on all human and primate chromosomes. Perhaps counterintuitively, our simulations also suggest that high mutation rates generate monomorphic datasets with slightly greater frequency. This is likely because over the large time period simulated, the greater total number of mutations generated at loci with high mutation rates increases the chance that repeat length will eventually wander to very short lengths where appreciable mutability is lost. In other words, while high mutation rate increases variance at existing microsatellite loci, it also decreases the duration of the birth-to-death period of a microsatellite on average. In general, then, abnormally low mutation rates are unlikely to account for the lack of variation at long exonic dinucleotides.

The posited purifying selection at the dinucleotides analyzed here is essentially a form of background selection (Charlesworth et al. 1993): recurrent removal of deleterious variants. Long-term background selection should cause local loss of linked genetic diversity near each dinucleotide. This may produce two detectable patterns of genetic variation. First, we

might expect elevated F_{ST} between two human populations near long exonic dinucleotides. This pattern is expected because purifying selection on frameshifting dinucleotide variants will randomly cause distinct linked alleles to rise in frequency in different populations. Second, we expect the ratio of polymorphism to divergence to decrease near the targeted dinucleotide; polymorphism decreases because neutral variants linked to deleterious dinucleotide alleles are lost, while the comparison to a measure of divergence normalizes the measure of polymorphism. To assess these patterns, we examined a 50,000-bp sequence (25,000 bp upstream and 25,000 bp downstream) flanking each dinucleotide. Using 1000 Genomes variant call data for the LWK and YRI populations, we calculated F_{ST} as

$\frac{\pi_{between} - \pi_{within}}{\pi_{between}}$, where π is the average number of pairwise differences between or within a sample of 25 chromosomes from each focal population. We calculated F_{ST} for windows of 2000 bp and a step size of 100 bp. For the same windows, we measured polymorphism as π in the YRI population and divergence as 1 minus fractional sequence identity between the human and chimpanzee reference sequences. At many of the 18 dinucleotides, these measures produced ambiguous results. However, several loci clearly show the expected patterns, including *HMGXB4* (Fig. 2 B–C). Additionally, *FGFRL1* is found very near the bottom of a dramatic decline in the polymorphism-to-divergence ratio (Fig. 2 D).

Of the 18 human loci in question, 16 have a repeat length of seven and one has a repeat length of eight. This suggests that the mutational load introduced by exonic dinucleotides becomes too extreme to support when repeat length exceeds seven or eight. Yet the remaining CA-repeat in fibroblast growth factor receptorlike 1 (*FGFRL1*) has a repeat length of 10 on all 400 sampled human chromosomes and in all primates. Thus, microsatellite *FGFRL1*_{CA} is an outlier among outliers. Interestingly, a CA₇ → CA₁₀ mutation (6 bp) would not result in a frameshift. This might partially explain the conspicuous jump between the relative abundance of 7× repeats and the repeat length of 10 observed at *FGFRL1*_{CA}. However, the CA₁₀ allele should have an appreciably higher mutation rate than a CA₇ allele. The lengths of *all* other exonic dinucleotides in the human genome certainly suggest that 10 should be an untenable repeat length for an exonic dinucleotide. Therefore, maintenance of this 10× repeat likely entails stronger purifying selection.

The *FGFRL1*_{CA} repeat is embedded within a sequence that consists of a 19× CA repeat with just three imperfections. Thus, it is plausible that point mutation has trimmed this repeat to a size that is less mutable. Of the other 18 repeats, none are such obvious candidates of progressive shortening. However, the AG dinucleotides in *LUC7L2* and *HMGXB4* are embedded in DNA strands that are unusually purine-rich. In particular, the *HMGXB4* is part of a 55-bp sequence that only contains one pyrimidine. Sequences such as this may be fertile nuclei for the evolution of new (or longer) microsatellites via point mutation.

At the protein level, the amino acid sequence coded for by *FGFRL1*_{CA} is conserved at least to mouse (Fig. 2 A). Unlike the four canonical fibroblast growth factor receptors, the intracellular domain of FGFRL1 is very short and characterized by a histidine-rich sequence, much of which is coded for by *FGFRL1*_{CA}. In general, histidine-rich sequences facilitate the protein–protein interactions of cytoskeletal proteins, and, in the specific case of FGFRL1, this sequence possesses high affinity for zinc and nickel ions (Zhuang et al. 2009). In

addition, it was recently shown that a tandem tyrosine-based motif immediately adjacent to the sequence coded for by *FGFRL1*_{CA} binds the enzyme SHP-1, which activates the ERK1/2 signaling pathway (Silva et al. 2013). The authors also showed that zinc binding of the histidine-rich region modulated binding of SHP-1, which suggests that the tyrosine-based and histidine-rich sequences act together to regulate FGFRL1 signaling. It is plausible that a threshold number of histidine residues is necessary for adaptive biochemical function and that this allowed an expansion allele of *FGFRL1*_{CA} to rise in frequency. However, the retention of *FGFRL1*_{CA} as a perfect repeat in all sampled primates except orangutan and marmoset remains puzzling. Incorporation of synonymous point mutations within *FGFRL1*_{CA} would decrease mutation rate while retaining the histidine-rich motif. Recently, Rieckmann et al. (2009) documented an individual heterozygous for a +2 CA allele at *FGFRL1*_{CA} who presented with craniosynostosis, radio-ulnar cynostosis, and genital abnormalities. This demonstration of the potentially existential hazard posed by *FGFRL1*_{CA} makes the maintenance of a perfect repeat at *FGFRL1*_{CA} even more puzzling.

Given their genomic context, the selective constraints we propose for exonic dinucleotides should not be controversial. However, our quantification of the extreme improbability of the observed data at the 18 exonic dinucleotides documented here begs the question, do the genes that harbor these loci share notable similarities? Using GO Term Finder (Boyle et al. 2004), we found an enrichment for the gene ontology term “mRNA processing” in four of 16 genes (25% versus the genomic frequency of 1.3%; corrected *P*-value = 0.006; *WDR64* and *CCDC30* not included in analysis because they are not annotated for GO terms). Although statistically significant, this result does not even include a majority of the 18 anomalous loci. Next, we calculated overall ω = dN/dS ratios for each of the 18 dinucleotide-containing genes in PAML using amino acid sequences (Yang 2007). ω quantifies the ratio of nonsynonymous to synonymous nucleotide substitutions between species. Values greater than one suggest positive natural selection, whereas values less than one suggest purifying natural selection. We assumed one value of ω for all branches and sites and used cDNA sequences for each of the genes from human, chimpanzee, gorilla, gibbon, rhesus macaque, baboon, marmoset, squirrel monkey (*Samiri boliviensis*), and galago (*Otolemur garnettii*); accession numbers are listed in Table S2. All estimated values of ω were <1, suggestive of purifying selection in all cases; specific values ranged from 0.005 (*IK*) through 0.718 (*MKI67*; Table S1). However, there is not a singular pattern among the loci (e.g., they are not uniformly close to zero) and most genes across the genome yield ω < 1.

One similarity that does exist among the loci is that 13 of the 18 dinucleotides code for arginine-glutamic acid (RE) dipeptides. The exceptions are three loci, including *FGFRL1*_{CA}, that code for HT dipeptides, and two microsatellites that code for a CV (*WDR64*) and AR (*NAGPA*) dipeptide. Although enrichment for RE dipeptides among the 18 loci is intriguing, it is difficult to ascribe biological significance to this enrichment. RE dipeptides have been implicated in protein–protein interactions (Yanagisawa et al. 2000), but the human gene *REER* named for its abundance of RE dipeptides does not use exonic dinucleotides to code for them. In other words, it is not necessary to use perfect CT or AG repeats to code for RE dipeptides.

As demonstrated here, long exonic dinucleotides are exceedingly rare in primate genomes due to mutation-based selective constraints. Those that do exist show remarkably low levels of polymorphism and divergence despite tens of millions of years of primate evolution. Moreover, lack of similarity among these loci in terms of biological function suggests that their long-standing existence is a product of chance more than anything else. It seems likely that rare three-step mutations—for example, from a 4× to a 7× repeat—were responsible for the initial production of expanded alleles, as these mutations would not incur frameshifts. Under specific circumstances, such as a population bottleneck, it is possible that the longer allele might then rise to fixation. The long allele might also offer some adaptive value that outweighs the increased mutational load associated with the increase in repeat length. In this way, a long exonic dinucleotide may become fixed at its current size for long stretches of evolutionary time. Although these conjectures are plausible, it remains unclear why a majority of the loci detailed in Table 1 have not fixed for synonymous point mutations in any of the sampled primates. Such mutations would reduce mutational load associated with a highly mutable microsatellite, yet have no effect on any selective advantage afforded by the current amino acid sequence. In what is otherwise a clear story of selective constraint, the conservation of these few hazardous perfect repeats remains a curious observation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This research was funded by the National Institutes of Health (grant number HG004498) awarded to BAP.

LITERATURE CITED

- 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467:1061–1073. [PubMed: 20981092]
- Ananda G, Walsh E, Jacob KD, Krasilnikova M, Eckert KA, Chiaromonte F, Makova KD. Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. *Genome Biol. Evol.* 2013; 5:606–620. Available at <http://dx.doi.org/10.1093/gbe/evs116>. [PubMed: 23241442]
- Baptiste BA, Ananda G, Strubczewski N, Lutzkanin A, Khoo SJ, Srikanth A, Kim N, Makova KD, Krasilnikova MM, Eckert KA. Mature microsatellites: mechanisms underlying dinucleotide microsatellite mutational biases in human cells. *G3 (Bethesda)*. 2013; 3:451–463. Available at <http://dx.doi.org/10.1534/g3.112.005173>. [PubMed: 23450065]
- Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. Go::termfinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*. 2004; 20:3710–3715. Available at <http://dx.doi.org/10.1093/bioinformatics/bth456>. [PubMed: 15297299]
- Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. *Genetics*. 1993; 134:1289–1303. [PubMed: 8375663]
- Chatterjee HJ, Ho SYW, Barnes I, Groves C. Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evol. Biol.* 2009; 9:259. Available at <http://dx.doi.org/10.1186/1471-2148-9-259>. [PubMed: 19860891]
- Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.* 2012; 22:1154–1162. [PubMed: 22522390]

- Haasl RJ, Payseur BA. The number of alleles at a microsatellite defines the allele frequency spectrum and facilitates fast, accurate estimation of θ . *Mol. Biol. Evol.* 2010; 27:2702–2715. [PubMed: 20605970]
- Haasl RJ, Payseur BA. Microsatellites as targets of natural selection. *Mol. Biol. Evol.* 2013; 30:285–298. [PubMed: 23104080]
- Hudson RR. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics.* 2002; 18:337–338. [PubMed: 11847089]
- Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, Makova KD. What is amicrosatellite: a computational and experimental definition based upon repeat mutational behavior at a/t and gt/ac repeats. *Genome Biol. Evol.* 2010; 2:620–635. Available at <http://dx.doi.org/10.1093/gbe/evq046>. [PubMed: 20668018]
- Kelkar YD, Eckert KA, Chiaromonte F, Makova KD. A matter of life or death: how microsatellites emerge in and vanish from the human genome. *Genome Res.* 2011; 21:2038–2048. Available at <http://dx.doi.org/10.1101/gr.122937.111>. [PubMed: 21994250]
- Kofler R, Schlötterer C, Lelley T. Sciroko: a new tool for whole genome microsatellite search and investigation. *Bioinformatics.* 2007; 23:1683–1685. Available at <http://dx.doi.org/10.1093/bioinformatics/btm157>. [PubMed: 17463017]
- Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.* 2004; 21:991–1007. [PubMed: 14963101]
- Lynch M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA.* 2010; 107:961–968. Available at <http://dx.doi.org/10.1073/pnas.0912629107>. [PubMed: 20080596]
- Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, et al. The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Res.* 2013; 41:D64–D69. Available at <http://dx.doi.org/10.1093/nar/gks1048>. [PubMed: 23155063]
- Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature.* 2009; 461:272–276. Available at <http://dx.doi.org/10.1038/nature08250>. [PubMed: 19684571]
- Payseur BA, Jing P, Haasl RJ. A genomic portrait of human microsatellite variation. *Mol. Biol. Evol.* 2011; 28:303–312. [PubMed: 20675409]
- Pemberton T, Sandefur C, Jakobsson M, Rosenberg N. Sequence determinants of human microsatellite variability. *BMC Genomics.* 2009; 10:612. [PubMed: 20015383]
- Rieckmann T, Zhuang L, Flück CE, Trueb B. Characterization of the first fgfr11 mutation identified in a craniosynostosis patient. *Biochim. Biophys. Acta.* 2009; 1792:112–121. Available at <http://dx.doi.org/10.1016/j.bbdis.2008.11.006>. [PubMed: 19056490]
- Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science.* 2010; 328:636–639. Available at <http://dx.doi.org/10.1126/science.1186802>. [PubMed: 20220176]
- Sawaya SM, Lennon D, Buschiazio E, Gemmell N, Minin VN. Measuring microsatellite conservation in mammalian evolution with a phylogenetic birth-death model. *Genome Biol. Evol.* 2012; 4:636–647. Available at <http://dx.doi.org/10.1093/gbe/evs050>. [PubMed: 22593552]
- Silva PN, Altamentova SM, Kilkenny DM, Rocheleau JV. Fibroblast growth factor receptor like-1 (fgfr11) interacts with shp-1 phosphatase at insulin secretory granules and induces beta-cell erk1/2 protein activation. *J. Biol. Chem.* 2013; 288:17859–17870. Available at <http://dx.doi.org/10.1074/jbc.M112.440677>. [PubMed: 23640895]
- Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, et al. A direct characterization of human mutation based on microsatellites. *Nat. Genet.* 2012; 44:1161–1165. Available at <http://dx.doi.org/10.1038/ng.2398>. [PubMed: 22922873]
- Yanagisawa H, Bundo M, Miyashita T, Okamura-Oho Y, Tadokoro K, Tokunaga K, Yamada M. Protein binding of a drpla family through arginine-glutamic acid dipeptide repeats is enhanced by extended polyglutamine. *Hum. Mol. Genet.* 2000; 9:1433–1442. [PubMed: 10814707]

- Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007; 24:1586–1591. Available at <http://dx.doi.org/10.1093/molbev/msm088>. [PubMed: 17483113]
- Zhuang L, Karotki AV, Bruecker P, Trueb B. Comparison of the receptor fgfr1 from sea urchins and humans illustrates evolution of a zinc binding motif in the intracellular domain. *BMC Biochem.* 2009; 10:33. Available at <http://dx.doi.org/10.1186/1471-2091-10-33>. [PubMed: 20021659]

Figure 1.

The three models of mutation rate used in simulations of microsatellite evolution on the primate phylogeny. In all cases, as empirical evidence suggests, mutation rate μ is a monotonically increasing function of repeat length. The blue, magenta, and orange lines correspond to models of high, medium, and low mutation rate, respectively.

Figure 2.

(A) Alignment of the C-terminus of *FGFRL1* orthologs in various vertebrate species. Bold-face residues are coded for by the CA₁₀ repeat. The final cysteine residue is the final amino acid of the protein in primates. Accession numbers and species: human, NP_001004356; chimpanzee (*Pan troglodytes*), JAA25770; gorilla, XP_004038363; orangutan (*Pongo abelii*), XP_002814545; baboon (*Papio anubis*), XP_003890969; macaque, XP_004847513; marmoset, XP_003732603; mouse, NP_00157731; chicken, NP_989787; frog (*Xenopus tropicalis*), NP_001011689; fish, (*Poecilia formosa*), XP_007576242. (B) F_{ST} between YRI and LWK populations at the *HMGXB4* dinucleotide. Vertical line indicates position of dinucleotide. (C) Ratio of polymorphism (π in YRI population) to divergence (from chimpanzee) at the *HMGXB4* dinucleotide. (D) Ratio of polymorphism (π in YRI population) to divergence (from chimpanzee) at the *FGFRL1* dinucleotide. The absence of

points immediately upstream of the dinucleotide is due to missing sequence in the chimpanzee genome.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Polymorphism and divergence data for human dinucleotide microsatellites.

Gene	Position	Motif	Intragenic	H	Refseq	Pt	Gg	Pp	Mm	Ph	Cj
<i>CCDC30</i>	1:43,002,201	AG	Exon	0	8	8	8	8	8	8	6
<i>CACNA1E</i>	1:181,680,102	AG	Exon	0	7	7	7	7	7	7	7
<i>WDR64</i>	1:241,846,828	GT	Exon	0	7	7	7	7	7	7	(7)
<i>ATP6V1C2</i>	2:10,917,819	AG	Exon	0	7	7	7	7	7	7	7
<i>MAGH</i>	3:65,372,850	CT	Exon	0	7	7	7	7	7	7	7
<i>FPIL1</i>	4:54,319,248	AG	Exon	0	7	7	7	7	7	7	7
<i>FGFRL1</i>	4:1,019,055	CA	Exon	0	10	10	10	(10)	10	10	(10)
<i>IK</i>	5:140,032,593	AG	Exon	0	7	7	7	7	7	7	7
<i>PAPD7</i>	5:6,755,013	CA	Exon	0	7	7	7	7	(7)	(7)	7
<i>LUC7L2</i>	7:139,094,366	AG	Exon	0	7	(7)	(7)	?	((7))	((7))	((7))
<i>MKI67</i>	10:129,905,113	CA	Exon	0	7	7	7	7	(7)	(7)	(7)
<i>CCAR1</i>	10:70,509,285	AG	Exon	0.01	7	7	7	7	7	7	(7)
<i>ACIN1</i>	14:23,528,502	CT	Exon	0	7	7	7	7	7	7	7
<i>SLTM</i>	15:59,182,564	CT	Exon	0	7	7	7	7	7	7	7
<i>NAGPA</i>	16:5,083,678	CG	Exon	0	7	7	7	7	7	7	(7)
<i>BRD7</i>	16:50,368,678	CT	Exon	0	7	7	7	7	6	6	6
<i>HMGXB4</i>	22:35,661,575	AG	Exon	0	7	7	7	8	8	8	7
<i>TXLNG</i>	X:16,850,849	AG	Exon	0	7	7	7	7	7	7	(7)
<i>C3orf27</i>	3:128,292,319	CT	Exon	0.005	9	9	9	6	(5)	(7)	(14)
<i>C10orf71</i>	10:50,534,970	AC	Exon	0.80	19	(22)	?	?	6/14*	8/6*	7
<i>RBM5</i>	3:50,155,888	AG	Exon/5'-UTR	0.04	11	10	(12)	15	10	10	3
<i>GPRASP1</i>	X:101,913,322	AT	3' UTR	0.03	7	6	?	6	(7)	(7)	((6))
Intergenic	13:104,538,371	CA	--	0.061	7	(10)	4	(14)	(10)	11	13
Intergenic	4:180,791,208	CA	--	0	7	8	11	9	(7)	(7)	(7)
Intergenic	11:23,584,479	GT	--	0	8	13	9	11	16	14	missing
Intergenic	18:65,819,377	CA	--	0.01	8	7	6	(6)	6	5	missing
Intergenic	6:23,512,730	CA	--	0.23	9	(13)	5	7	(11)	5	7

Gene	Position	Motif	Intragenic	H	Refseq	Pt	Gg	Pp	Mm	Ph	Cj
Intergenic	5:165,162,325	TG	--	0	9	6	5	4	7	9	5
Intergenic	2:123,914,689	CA	--	0.03	10	7	7	9	10	(10)	(10)
Intergenic	2:221,757,934	CA	--	0.43	10	12	6	(26)	((21))	missing	(7)

Notes:

¹ The upper section shows data for the only 18 coding dinucleotides in the human genome that are perfect repeats of length greater than or equal to seven. Reference sequence lengths are also reported for the following species: *Pt*: *Pan troglodytes*, chimpanzee; *Gg*: *Gorilla gorilla*; *Pp*: *Pongo pygmaeus*, orangutan; *Mm*: *Macaca mulatta*, rhesus macaque; *Ph*: *Papio hamadryas*, baboon; *Cj*: *Callithrix jacchus*, marmoset. Lengths reported within parentheses are imperfect repeats, where the number of parenthetical sets equals the number of internal imperfections. The lower section of the table shows comparative data from dinucleotide loci in open reading frames, UTRs, and intergenic regions.

² *H* = observed heterozygosity, ? = alignment unavailable, missing = no microsatellite.