Segmenting the human genome based on states of neutral genetic divergence

Prabhani Kuruppumullage Don^{a,1}, Guruprasad Ananda^{b,1,2}, Francesca Chiaromonte^{a,3}, and Kateryna D. Makova^{c,3}

Departments of ^aStatistics and ^cBiology, and ^bHuck Institute of the Life Sciences, The Pennsylvania State University, University Park, PA 16802

Edited by Wen-Hsiung Li, The University of Chicago, Chicago, IL, and approved July 23, 2013 (received for review December 13, 2012)

Many studies have demonstrated that divergence levels generated by different mutation types vary and covary across the human genome. To improve our still-incomplete understanding of the mechanistic basis of this phenomenon, we analyze several mutation types simultaneously, anchoring their variation to specific regions of the genome. Using hidden Markov models on insertion, deletion, nucleotide substitution, and microsatellite divergence estimates inferred from human–orangutan alignments of neutrally evolving genomic sequences, we segment the human genome into regions corresponding to different divergence states—each uniquely characterized by specific combinations of divergence levels. We then parsed the mutagenic contributions of various biochemical processes associating divergence states with a broad range of genomic landscape features. We find that high divergence states inhabit guanine- and cytosine (GC)-rich, highly recombining subtelomeric regions; low divergence states cover inner parts of autosomes; chromosome X forms its own state with lowest divergence; and a state of elevated microsatellite mutability is interspersed across the genome. These general trends are mirrored in human diversity data from the 1000 Genomes Project, and departures from them highlight the evolutionary history of primate chromosomes. We also find that genes and noncoding functional marks [annotations from the Encyclopedia of DNA Elements (ENCODE)] are concentrated in high divergence states. Our results provide a powerful tool for biomedical data analysis: segmentations can be used to screen personal genome variants—including those associated with cancer and other diseases—and to improve computational predictions of noncoding functional elements.

Whole-genome sequencing studies have demonstrated that divergence estimates for several mutation types (e.g., nucleotide substitutions, insertions, and deletions) vary substantially across the human genome. This phenomenon has been studied at various genomic scales and evolutionary distances (reviewed in ref. 1), and—whereas initially of interest solely to evolutionary biologists—is now entering the purview of main biomedical research. Specifically, human population (e.g., ref. 2) and cancer (3, 4) genome resequencing projects have revealed that incidences of single nucleotide polymorphisms (SNPs), insertions and deletions (indels), and copy number variants (CNVs) vary across the genome. Divergence estimates for different mutation types also covary across the genome (5, 6)—e.g., substitution rates increase in regions with high indel rates (7)—suggesting that regional variation is an important and general characteristic of mutations.

Variation in divergence is often linked to genomic landscape features such as base composition, replication timing, and recombination rates (1). For instance, nucleotide substitution rates are elevated in late-replicating regions because of an accumulation of single-stranded DNA susceptible to endogenous damage (8) and are affected by chromatin structure (9) and recombination as a result of either biased gene conversion (BGC) (10) or the mutagenicity of recombination (2). Moreover, nucleotide substitution rates depend nonlinearly on guanine and cytosine (GC) content (11) and are affected by methylation levels and GC content at cytosine–phosphate–guanine (CpG) sites (12) and by replication timing and distance to telomeres at non-CpG sites (13). Covariation in divergence among rates of different mutation types can also be at least partly attributed to the influence of

a common genomic landscape (5). Importantly, we note that, whereas selection may indeed operate in noncoding regions, it is unlikely to explain the large degree of variation and covariation in divergence estimates computed from putatively neutral sequences (1, and references therein). Divergence computed for neutral DNA ought to reflect mutation, BGC, and—for relatively distant species—only a minimal amount of diversity.

Anchoring variation and covariation of divergence estimates for different mutation types to specific regions of the genome is crucial for elucidating how biochemical processes—e.g., replication and recombination—drive mutagenesis (8, 10) and for understanding genome evolution. Such a “geographic” characterization may correlate with the spatial distribution of genes; for instance, cellular receptor and housekeeping genes tend to locate, respectively, in high and low nucleotide substitution rate regions (14). It may also aid prediction of noncoding functional elements (15, 16) and studies of the genetic basis of disease. For instance, it could assist in (i) discerning whether a locus exhibits an excess of mutations because it resides in a hotspot, thus preventing false positive associations with a disease; and (ii) identifying loci with mutational signatures typical of a disease, e.g., explaining frequent coincidence in tumors of sites prone to DNA damage and chromosomal instability (17).

With this motivation, we used hidden Markov models (HMMs) (18), a well-established statistical tool, to analyze human divergence for different mutation types. An HMM models a sequence of observations as governed by underlying states that are not directly observable (hidden) but can be inferred based on the data. These states alternate along the sequence following a Markovian structure, i.e., the state governing a given observation may depend on the state governing the preceding observation.

Significance

In addition to a significant contribution to our understanding of the intricacies of mutagenesis, this study provides a powerful platform for mining biomedical data—which we make publicly available through the University of California Santa Cruz Genome Browser and the Galaxy portal. The divergence states we characterize serve as local background to benchmark signals used in computational algorithms for prediction of noncoding functional elements and in screening variants from cancer and other disease-affected genomes.

Author contributions: P.K.D., G.A., F.C., and K.D.M. designed research; P.K.D. and G.A. performed research; P.K.D. and G.A. analyzed data; and P.K.D., G.A., F.C., and K.D.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The segmentations have been deposited into the University of California Santa Cruz (UCSC) Genome Browser (<http://tinyurl.com/AR6states>) and Galaxy (<https://main.g2.bx.psu.edu/u/guru/p/ar-divergence-states>).

¹P.K.D. and G.A. contributed equally to this work.

²Present address: Computational Sciences, The Jackson Laboratory, Bar Harbor, ME 04609.

³To whom correspondence may be addressed. E-mail: kdm16@psu.edu or chiaro@stat.psu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1221792110/-DCSupplemental.

Maximum likelihood techniques are used to select an appropriate number of states, characterize them, and partition the sequence into contiguous segments governed by each state.

In genomics, HMMs have been used to model stretches of DNA—the sequences of observations—in a variety of applications ranging from gene finders (19) to epigenomic segmentations (20, 21). In our study, we compute divergence estimates for four mutation types—substitutions, insertions, deletions, and microsatellite repeat number alterations—in nonoverlapping windows of neutrally evolving sequences present in human–orangutan genomic alignments. Modeling the resulting observations with HMMs, we identify distinct divergence states characterized by biologically meaningful combinations of elevated, average, or depressed divergence levels for the four mutation types, e.g., a state where only microsatellite mutations are elevated, whereas the other divergence types are average; one where substitutions, insertions, and deletions are all elevated, whereas microsatellite divergence is average, etc. Correspondingly, we partition the genome into chromosomal segments governed by these states. Our analysis departs from previous applications in considering several mutation types simultaneously—thus accounting for their interdependencies—and in generating segmentations not on a small 100-bp scale (22), but on a larger 1-Mb scale that robustly captures variation in divergence for mammalian genomes (11, 23). Additionally, we investigate whether divergence states differ in genomic landscape features that proxy underlying mutagenic processes, correlate with the spatial organization of functional elements, and persist when assessed from human diversity estimates or for varying genomic scales.

Results

We estimated divergence for nucleotide substitutions, small (≤ 30 bp) insertions, and deletions, and mononucleotide microsatellite repeat number alterations in nonoverlapping 1-Mb windows of human–orangutan alignments. This was done separately for two mutually exclusive subgenomes frequently used as proxies for neutral evolution (5, 6, 23): ancestral repeats (ARs) and non-coding nonrepetitive (NCNR) sequences, covering, respectively, 817 and 822 Mb of alignments—i.e., 28.4% and 28.5% of the human genome (*Materials and Methods* and Fig. 1 for data preprocessing). To minimize the contribution of diversity to divergence estimates, a known issue for human–chimpanzee comparisons (24), we used human–orangutan alignments. We inferred divergence estimates for the four mutation types in the AR and NCNR subgenomes and used these estimates to fit separate multivariate HMMs for each subgenome (*Materials and Methods* and *SI Appendix*, Fig. S1). Models with six and five divergence states, respectively, for AR and NCNR had optimal Bayes

information criterion (BIC) values (25) (*SI Appendix*, Fig. S24), allowed us to isolate signals specific to microsatellite mutability and the X chromosome observed in prior studies (26), and were qualitatively consistent with one another (*SI Appendix*, Table S14). Results were robust with respect to key parameters of the data preprocessing pipeline (*Materials and Methods*). Below, we focus on the six-state AR model (five- and seven-state AR models are described in *SI Appendix*, Table S1).

Divergence States. The six divergence states have distinct profiles (Fig. 2A and Table 1). Five are characterized by substitutions, insertions, and deletions ranging from “hot” to “cold,” but background levels of microsatellite mutability (*SI Appendix*, Table S2 and Fig. S1H). At one extreme, we have a state—which covers $\sim 8\%$ of the windows under analysis—with indels and substitutions enhanced by 1/5–1/3 compared with genome medians (hot state). Next comes a state—covering $\sim 18\%$ of windows—with deletions and substitutions enhanced by $\sim 1/10$ (deletion/substitution-warm state), followed by a state—covering as much as $\sim 28\%$ of windows—with insertions enhanced by $\sim 1/10$ (insertion-warm state). As a rule, hot and deletion/substitution-warm segments locate in autosomal subtelomeric regions and are of short or medium length, whereas insertion-warm segments locate in more-inward regions of autosomes and are of medium or long length (Fig. 3 and Table 1). At the other end of the spectrum, we have a state with indels and substitutions depressed by $\sim 1/5$, which covers $\sim 4\%$ of windows exclusively on chromosome X (cold X-chromosomal state) and a state with indels and substitutions depressed by $\sim 1/10$, which occurs on autosomes and covers $\sim 36\%$ of windows (cold autosomal state). Cold autosomal segments are long, removed from telomeres by ≥ 60 Mb on most large autosomes, and absent from the short arms of many small autosomes—suggesting that a certain distance from the telomeres may be needed for the DNA to “cool off” (Fig. 3 and Table 1). Finally, a sixth state is characterized by strongly enhanced microsatellite mutability, but background levels for indels and substitutions (microsatellite state); this state covers $\sim 5\%$ of windows in very short interspersed segments.

Some exceptions to these general geographic trends highlight the evolutionary history of primate chromosomes. For instance, chromosome 7 was formed by a Robertsonian fusion of two chromosomes in the common ancestor of Old World monkeys and apes (27); thus, hot and deletion/substitution-warm segments near the centromere of chromosome 7 (Fig. 3) might be relics of the ancestral telomeres. Additionally, these exceptions point toward a potential role of heterochromatin repeats located near centromeres in driving divergence (28).

Genomic Landscapes. We considered a set of 35 genomic features—measured over the entire lengths of our 1-Mb windows (*SI Appendix*, Table S4 and Fig. S2C)—as proxies for different biochemical processes and contexts affecting DNA mutations, and we evaluated their enrichment in each divergence state (*Materials and Methods*). Below we focus on the 20 features showing the most prominent associations with divergence (Fig. 2B and Table 2).

The hot state occurs in subtelomeric autosomal regions (Fig. 3) that are known to have aberrant repair—e.g., high levels of non-homologous end-joining repair (29)—and, as expected (30), exhibits enhanced male recombination. Whereas telomere-specific aberrant repair might elevate substitutions directly (29), the mutagenicity of recombination has been debated (2, 31). GC richness—measured at 45% vs. 39.8% genome median—agrees well with abundance of genes and functional elements in the hot state, as do abundance of open chromatin (32), enrichment for 5-hydroxy methylcytosines (33), and early replication timing (34) (Fig. 2B and Table 2; see also below). As this state has high transcriptional potential, it might also host transcription-associated mutagenesis (35). Subtelomeric effects and functional activity here appear to override previously observed associations between substitution rates and some genomic features; contrary to previous reports (8, 36, 37), high substitution rates are associated with early replication timing, open chromatin, and depleted fraction of methylated CpGs.

Fig. 1. Schematic overview of the data generation and filtering steps for AR divergence estimation in 1-Mb windows. Dashed arrows point to the number of windows discarded at each filtering step.

Fig. 3. Genomic locations of segments belonging to the divergence states of the six-state AR model. Bars represent human chromosomes, reported in scale and with positions indicated on the vertical axis. Gray regions correspond to windows excluded from the analysis due to assembly gaps or data preprocessing filters.

the proportion of all genes located in this state (14.2%). Thus, resonating with a previous study (14), we find that coding sequences—including those that need to evolve rapidly—favor regions where germ-line mutation rates are highest, likely because this can confer an adaptive advantage. It should be noted that this observation might be confounded by the higher power of selection tests in regions with a larger number of substitutions. Cancer-related genes are significantly enriched in the insertion-warm state (Fig. 3)—suggesting that its open chromatin landscape offers opportunities for genes to become tumorigenic, e.g., via epigenetic modifications.

CpG and Non-CpG Substitutions. Unlike those at non-CpG sites, most nucleotide substitutions at CpG sites result from spontaneous deamination of methylated cytosines—a process with high

rates but unrelated to replication errors (45, 46). To explore these biological differences, we used HMMs to model rates of CpG substitutions, non-CpG substitutions, small insertions, small deletions, and microsatellite alterations (*SI Appendix, Fig. S3 A–E*). The optimal model has seven states (*SI Appendix, Fig. S3A*) and captures details of the differentiation between CpG and non-CpG substitutions, whereas confirming the patterns of the AR six-state model described above (Figs. 24 and S3C). Both substitution types are elevated in a deletion/substitution-warm state and depressed in a cold X-chromosomal state. However, we now identify a hot state with elevated non-CpG substitutions and depleted CpG substitutions, and an insertion-warm state with depleted CpG substitutions. Moreover, the cold autosomal state of the AR six-state model splits into a state with depressed indels and elevated CpG substitutions and a state with depressed deletions and both CpG and non-CpG substitutions (*SI Appendix, Table S1C and Fig. S3C*). Interestingly, and in line with previous observations (12, 13), CpG substitutions are always enhanced in states with high methylation and depleted in states with low methylation (*SI Appendix, Fig. S3D*).

Genomic Scales. HMMs ran on divergence estimates from smaller (0.1 Mb and 0.5 Mb) and larger (2 Mb and 3 Mb) windows (*SI Appendix, Figs. S4–S7*) also confirmed the patterns of the AR six-state model at 1 Mb. However, at the coarser 2-Mb and 3-Mb scales, the microsatellite state is redistributed among other states (*SI Appendix, Figs. S6 and S7 and Table S1 F and G*). In contrast, reiterating previous findings (5), models at the finer 0.1- and 0.5-Mb scales comprise additional states that reflect associations between microsatellite alterations and other mutation types (*SI Appendix, Figs. S4 and S5*). Using smaller windows also allowed us to retain portions of chromosome 19, which—due to its gene richness—was entirely excluded from the 1-Mb analysis (*Materials and Methods*). Substantiating our observations on the enrichment of genes, the recovered portions of chromosome 19 belong to hot states (*SI Appendix, Figs. S4D and S5D*).

Human Diversity. To verify whether the patterns we observed based on a primate comparison could be retraced in present-day human diversity, we used HMMs to model SNPs and small insertions and deletions derived from the 1000 Genomes Project (2). Microsatellite alterations and all of chromosome X were excluded from this analysis due to data limitations (*Materials and Methods*). Reassuringly, the most satisfactory model (*SI Appendix, Fig. S8A*) is qualitatively consistent with the AR six-state model described above. It comprises a hot state occurring near telomeres (*SI Appendix, Fig. S8D*), two cold states located away from telomeres, and a widespread “average” state where all three rates are at genome-wide baselines (*SI Appendix, Table S1*).

Table 2. Genomic landscapes and biochemical processes associated with the divergence states of the six-state AR model

Landscapes/states	Hot	Del/sub warm	Ins warm	Cold auto	Cold chr X	Microsat
GC content	High 44.9%	Low 38.3%	High 41.5%	Ave 38.4%	Ave 38.9%	Ave 39.5%
Replication: enhanced replication timing ratio (34) and GC content (60) signal early replication	Early	Late	Early	Mid	Mid	Late
Recombination rates	High Male >> Female	High Male > Female	High Male < Female	Low	Low Female Only	Ave
Chromatin: high H3K4Me1 (61) and DNase 1 (62), and low nuclear lamina (5) and methylated CpGs (63) signal open chromatin	Open	Closed	Open	Closed	Closed	Mid
Transcription: enhanced coding exons and CpG island density (64) signal high transcription	High	Low	High	Ave	Low	Ave
CpG methylation: enhanced fraction of methylated CpG (63) and depressed CpG island count signal high methylation	Low	High	Low	High	Ave	Ave
Telomeric mechanisms: active at low distance to telomeres (29)	Active	Active	Active	Inactive	Ave	Ave

Ave, average; >, higher; >>, much higher; <, lower. See Table 1 for other abbreviations.

Feature/State	Hot	Ins-warm	Del/sub-warm	Cold auto	Micro sat
All genes (41)	0.001	0.002	0.009	n.s.	0.018
Positively selected genes (42)	0.000	n.s.	n.s.	n.s.	n.s.
Cancer census genes (43)	n.s.	0.032	0.080	n.s.	n.s.
All functional marks (21)	0.006	0.002	0.055	n.s.	0.025
Predicted CTCF (21)	0.005	0.012	n.s.	n.s.	0.029
Predicted enhancers (21)	0.001	0.001	0.032	n.s.	0.056
Predicted promoter regions (21)	0.001	0.001	0.016	n.s.	0.023
ChromHMM q/dead state* (20)	0.089	n.s.	0.036	0.059	n.s.
Segway q/dead state* (65)	n.s.	n.s.	0.056	n.s.	n.s.

Fig. 4. Genes and functional marks in the divergence states of the six-state AR model. Features are measured as counts or base coverage (the latter denoted by an asterisk). Only statistically significant P values (≤ 0.1) are shown with yellow and blue denoting enrichment and depletion, respectively; NS, nonsignificant. *SI Appendix, Table S5 A and B* contain P values and observed and expected counts for a broader set of features. The cold X-chromosomal state is excluded from this analysis due to limited functional annotation on X.

and Fig. S8 A–D). The latter state is attributable to the short evolutionary time in which polymorphisms segregated in human populations and to variation in coalescent time along the genome (47). Hot diversity windows are mostly hot or warm for divergence (84% belong to the hot, deletion/substitution-warm, or insertion-warm states of the AR six-state model); cold diversity windows are mostly cold autosomal (45%) or insertion-warm (36%) for divergence; and, similarly, average diversity windows map mostly to cold autosomal (47%) and insertion-warm (28%) states (*SI Appendix, Table S1*). This all suggests substantial conservation in the patterns unveiled by our study.

Discussion

We provide a comprehensive geographic characterization of variation of several divergence estimates across the human genome. Our results confirm a strong covariation of indels and substitutions (6) and a separate variation pattern for microsatellites (5) while also drawing a rich and multifaceted picture of mutationally hot, warm, and cold DNA regions.

The six divergence states we identify present distinct genomic landscapes, suggesting associations between mutagenesis and genomic features. Some associations recur across states and largely agree with previous observations. Others are linked to one or a few specific states and provide unique insights. For instance, compact chromatin and late replication timing have been shown to associate with divergence (8, 9); we confirm this for the abundant deletion/substitution-warm state, but not for the rarer hot state, which carries functional marks and is strongly affected by subtelomeric effects and transcription-associated mutagenesis.

Finally, our divergence states correlate with the spatial distribution of functional sequences. The hot and insertion-warm states are enriched in genes and noncoding functional marks, suggesting that their open chromatin architecture favors both mutagenesis and functional activity. The deletion/substitution-warm state has a compact chromatin structure and does not show a similar enrichment of functional elements. The insertion-warm state is also enriched in cancer-related genes, perhaps again because of its open chromatin structure.

Whole-genome sequencing of human populations and pedigrees is providing fine-resolution data to pinpoint de novo mutations and estimate mutation rates (e.g., refs. 2, 48). Information on divergence states can be critical in analyzing these data. For instance, given the heterogeneity of divergence levels in the genome, it is likely that evolutionarily recent mutations may emerge from hot regions of the genome and thus lead to

overestimation of mutation rates in pedigree studies. Indeed, we observe a significant enrichment of SNPs from the 1000 Genomes Project in the hot state (P value = 0.048).

More generally, information on divergence states can improve prediction of noncoding functional elements as well as screening for disease variants. Conservation-based algorithms predicting functional elements flag sequences whose divergence is lower than a genome-wide or local background defined in terms of specific rates—e.g., substitutions (15) or indels (16). The accuracy of these algorithms can substantially increase accounting for covariation among different rates—e.g., benchmarking on substitutions alone, we may have false positive findings in a region that is cold for substitutions but hot for other mutation types, and we may have false negatives in a region that is hot for substitutions but cold for other mutation types. Similar considerations apply when evaluating genetic risk factors in personal genomes and tailoring clinical counseling and intervention. Although baseline somatic rates are used to evaluate somatic variants (49), screens for relevant germ-line mutations—including those predisposing carriers to cancer—could be improved by accounting for covariation among different divergence rates.

Materials and Methods

Neutral Subgenomes and Rates. AR and NCNR subgenomes were constructed based on the build 36.1 of the UCSC human genome assembly (hg18). AR included transposable elements inserted before the human–macaque divergence—excluding lineage-specific elements L1PA1-7 (primate-specific L1s), L1HS (human-specific L1s), and *AluY* (active human *Alus*). The NCNR sequence was obtained by removing genes (41) and 5-kb flanking regions up- and downstream, as well as known functional elements (50–53) and interspersed repeats (54). These subgenomes were delineated in human–orangutan–macaque BLASTZ alignments broken into nonoverlapping 1-Mb windows based on human coordinates. For nucleotides belonging to each subgenome in each window, we computed human-specific insertion and deletion rates, human-specific nucleotide substitution rates (using the Jukes–Cantor model) (55), and human–orangutan pairwise microsatellite mutability (56). The latter was obtained for mononucleotide microsatellites with ≥ 9 repeats that were sufficiently abundant per 1 Mb and for repeat number bins of size four—e.g., microsatellites with 9–12 repeats formed one bin, etc.; mutability was regressed on mean repeat numbers of the bins, and the resulting residuals were used below.

Data Preprocessing Pipeline. See Fig. 1 for a graphical illustration. First, we removed 1-Mb windows with low neutral subgenome coverage, which was defined as <200 kb, based on the distribution of coverage (*SI Appendix, Fig. S1A*) and on the balance of coverage between the two subgenomes (*SI Appendix, Fig. S1B*). Second, divergence rates underwent log transformation, removal of grossly outlying windows (*SI Appendix, Fig. S1B and C*), and standardization (*SI Appendix, Fig. S1D*). Third, we only retained windows forming contiguous strings of ≥ 10 , because HMMs cannot use isolated windows. The histogram for the length of contiguous strings (*SI Appendix, Fig. S1E*) and the number of windows in a continuous string plotted as a function of their lengths (*SI Appendix, Fig. S1F*) suggested a length of 10 as a natural breaking point. These steps led to the retention of 2,558 windows out of 2,798, covering 86% of the genome. Chromosome 19 was not captured by the analysis as it is the most gene-rich chromosome and thus does not provide reliable estimates of neutral divergence; only 35 of its windows have neutral coverage ≥ 200 kb, and all of these windows are in strings of <10 . We also missed some subcentromeric and/or subtelomeric locations of chromosomes 1, 9, 11, 12, 16, 17, and X due to assembly gaps. To evaluate robustness of our results with respect to filtering steps, we used HMMs to model windows with ≥ 100 kb of neutral sequence and requiring 5 contiguous windows to retain a string (*SI Appendix, Fig. S1G*). In both cases results are similar to those obtained in the original analysis (*SI Appendix, Figs. S9 and S10*). In addition, based on the HMM parameter values estimated on the 2,558 retained windows, we used the Viterbi algorithm (57) to assign states to the 222 windows excluded from our original analysis due to coverage or contiguous length restrictions. The divergence profiles of these windows are consistent with our results (*SI Appendix, Figs. S11 and S12*).

Multivariate Hidden Markov Models and Simulation-Based P Values. Details are provided in *SI Appendix*.

1000 Genomes Project Analysis. We excluded microsatellite mutability and the X chromosome from this analysis because of the limitations of the 1000 Genomes Project data (58). Thus, we worked with three mutation types—insertions, deletions, and SNPs—and 2,381 autosomal windows, restricting diversity computation to the AR subgenome and focusing on variation within the European population (CEU).

UCSC Tracks and Galaxy Tools. The segmentation corresponding to the AR six-state model is available as a track at the UCSC Genome Browser (<http://tinyurl.com/AR6states>) and at Galaxy (<https://main.g2.bx.psu.edu/u/guru/p/ar-divergence-states>). We also integrated the following tools into the open-source web-based genomics portal Galaxy (59) (<http://test.g2.bx.psu.edu>): “Fit HMM”, “Heatmap”, and “Karyotype plot”.

ACKNOWLEDGMENTS. We thank two anonymous reviewers whose comments greatly improved the manuscript; S. Palmer for his valuable help with the final text; R. Hardison, K. Eckert, L. Carrel, L. Duret, and E. Kvikstad for suggestions on earlier drafts; A. Fungtammasan and R. Campos Sanchez for their help with collection of genomic features; and the Wellcome Trust Sanger Institute Cancer Genome Project for making data on cancer genes available. This study was supported by National Science Foundation Grant DBI-0965596 (to K.D.M. and F.C.), National Institutes of Health Grant GM087472 (to K.D.M. and K. Eckert), a Marie Curie Fellowship (to K.D.M.), the Penn State Clinical and Translational Science Institute, and Pennsylvania Tobacco Settlement Funds from the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analysis, interpretations or conclusions.

- Hodgkinson A, Eyre-Walker A (2011) Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* 12(11):756–766.
- 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.
- Hodgkinson A, Chen Y, Eyre-Walker A (2012) The large-scale distribution of somatic mutations in cancer genomes. *Hum Mutat* 33(1):136–143.
- Schuster-Böckler B, Lehner B (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* 488(7412):504–507.
- Ananda G, Chiaromonte F, Makova KD (2011) A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome Biol* 12(3):R27.
- Hardison RC, et al. (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* 13(1):13–26.
- Tian D, et al. (2008) Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* 455(7209):105–108.
- Stamatoyannopoulos JA, et al. (2009) Human mutation rate associated with DNA replication timing. *Nat Genet* 41(4):393–395.
- Chen X, et al. (2012) Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science* 335(6073):1235–1238.
- Duret L, Arnott PF (2008) The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet* 4(5):e1000071.
- Hellmann I, et al. (2005) Why do human diversity levels vary at a megabase scale? *Genome Res* 15(9):1222–1231.
- Mugal CF, Ellegren H (2011) Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. *Genome Biol* 12(6):R58.
- Chen CL, et al. (2010) Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* 20(4):447–457.
- Chuang JH, Li H (2004) Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol* 2(2):E29.
- Taylor J, et al. (2006) ESPERR: Learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Res* 16(12):1596–1604.
- Lunter G, Ponting CP, Hein J (2006) Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* 2(1):e5.
- Degtyareva NP, Chen L, Mieczkowski P, Petes TD, Doetsch PW (2008) Chronic oxidative DNA damage due to DNA repair defects causes chromosomal instability in *Saccharomyces cerevisiae*. *Mol Cell Biol* 28(17):5432–5445.
- Eddy SR (2004) What is a hidden Markov model? *Nat Biotechnol* 22(10):1315–1316.
- Majoros WH, Pertea M, Antonescu C, Salzberg SL (2003) GlimmerM, Economy and Unveil: Three ab initio eukaryotic genefinders. *Nucleic Acids Res* 31(13):3601–3604.
- Ernst J, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473(7345):43–49.
- Dunham I, et al.; ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
- Oldmeadow C, Mengersen K, Mattick JS, Keith JM (2010) Multiple evolutionary rate classes in animal genome evolution. *Mol Biol Evol* 27(4):942–953.
- Gaffney DJ, Keightley PD (2005) The scale of mutational variation in the murid genome. *Genome Res* 15(8):1086–1094.
- Makova KD, Li WH (2002) Strong male-driven evolution of DNA sequences in humans and apes. *Nature* 416(6881):624–626.
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461–464.
- Ellegren H (2011) Sex-chromosome evolution: Recent progress and the influence of male and female heterogamety. *Nat Rev Genet* 12(3):157–166.
- Richard F, Lombard M, Dutrillaux B (2000) Phylogenetic origin of human chromosomes 7, 16, and 19 and their homologs in placental mammals. *Genome Res* 10(5):644–651.
- Ellegren H, et al. (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* 491(7426):756–760.
- Linaropoulou EV, et al. (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* 437(7055):94–100.
- Kong A, et al. (2002) A high-resolution recombination map of the human genome. *Nat Genet* 31(3):241–247.
- Webster MT, Smith NG, Hultin-Rosenberg L, Arnott PF, Ellegren H (2005) Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. *Mol Biol Evol* 22(6):1468–1474.
- Lander ES, et al.; International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.
- Szulwach KE, et al. (2011) Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS Genet* 7(6):e1002154.
- Ryba T, et al. (2010) Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res* 20(6):761–770.
- Kim N, Jinks-Robertson S (2012) Transcription as a source of genome instability. *Nat Rev Genet* 13(3):204–214.
- Ying H, Epps J, Williams R, Huttley G (2010) Evidence that localized variation in primate sequence divergence arises from an influence of nucleosome placement on DNA repair. *Mol Biol Evol* 27(3):637–649.
- Walser JC, Ponger L, Furano AV (2008) CpG dinucleotides and the mutation rate of non-CpG DNA. *Genome Res* 18(9):1403–1414.
- Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD (2007) A macaque’s-eye view of human insertions and deletions: Differences in mechanisms. *PLoS Comput Biol* 3(9):1772–1782.
- Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T (1987) Male-driven molecular evolution: A model and nucleotide sequence analysis. *Cold Spring Harb Symp Quant Biol* 52:863–867.
- Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res* 18(1):30–38.
- Hsu F, et al. (2006) The UCSC known genes. *Bioinformatics* 22(9):1036–1046.
- Kosiol C, et al. (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genet* 4(8):e1000144.
- Futrel PA, et al. (2004) A census of human cancer genes. *Nat Rev Cancer* 4(3):177–183.
- Pique-Regi R, et al. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* 21(3):447–455.
- Ehrlich M, Wang RY (1981) 5-Methylcytosine in eukaryotic DNA. *Science* 212(4501):1350–1357.
- Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* 156(1):297–304.
- Hobolth A, Christensen OF, Mailund T, Schierup MH (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet* 3(2):e7.
- Kong A, et al. (2012) Rate of de novo mutations and the importance of father’s age to disease risk. *Nature* 488(7412):471–475.
- Hodis E, et al. (2012) A landscape of driver mutations in melanoma. *Cell* 150(2):251–263.
- Siepel A, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15(8):1034–1050.
- Kim TH, et al. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 128(6):1231–1245.
- Carroll JS, et al. (2006) Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* 38(11):1289–1297.
- Karolchik D, et al. (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* 36(Database issue):D773–D779.
- Smit AFA, Hubley R, Green P (1996–2004) RepeatMasker, open-3.1.2.
- Jukes TH, Cantor CR (1969) *Evolution of Protein Molecules* (Academic, New York).
- Webster MT, Smith NG, Ellegren H (2002) Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proc Natl Acad Sci USA* 99(13):8748–8753.
- Taramasco O, Bauer S (2010) RHMm: Hidden Markov models simulations and estimations. R package version 1.4.4. Available at <http://CRAN.R-project.org/package=RHMm>.
- Montgomery SB, et al.; 1000 Genomes Project Consortium (2013) The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* 23(5):749–761.
- Blankenberg D, et al. (2010) Galaxy: A web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* Chap19:Unit 19.10.1–10.21.
- Woodfine K, et al. (2004) Replication timing of the human genome. *Hum Mol Genet* 13(2):191–202.
- Barski A, et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129(4):823–837.
- Crawford GE, et al. (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* 16(1):123–131.
- Cooper DN, Krawczak M (1989) Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum Genet* 83(2):181–188.
- Cross SH, Bird AP (1995) CpG islands and genes. *Curr Opin Genet Dev* 5(3):309–314.
- Hoffman MM, et al. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* 9(5):473–476. 10.1038/nmeth.1937.