Tandem repeat variation in human and great ape populations and its impact on gene expression divergence

Tugce Bilgin Sonay,[1,2,11] Tiago Carvalho,[3,11] Mark D. Robinson,[2,4] Maja P. Greminger,[5] Michael Krützen,[5] David Comas,[3] Gareth Highnam,[6] David Mittelman,[6] Andrew Sharp,[7] Tomàs Marques-Bonet,[3,8,9,11] and Andreas Wagner[1,2,10,11]

[1]*Institute of Evolutionary Biology and Environmental Studies, University of Zurich, CH-805 Zurich, Switzerland;* [2]*The Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland;* [3]*Institute of Evolutionary Biology (CSIC-UPF), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, 08003 Barcelona, Spain;* [4]*Institute of Molecular Life Sciences, University of Zurich, 8057 Zurich, Switzerland;* [5]*Evolutionary Genetics Group, Anthropological Institute and Museum, University of Zurich, CH-8057 Zurich, Switzerland;* [6]*Department of Biological Science and Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia 24061, USA;* [7]*Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai School, New York, New York 10029, USA;* [8]*Centro Nacional de Análisis Genómico (CNAG), PCB, Barcelona, 08028 Catalonia, Spain;* [9]*Catalan Institution for Research and Advanced Studies (ICREA), 08010 Barcelona, Spain;* [10]*The Santa Fe Institute, Santa Fe, New Mexico 87501, USA*

Tandem repeats (TRs) are stretches of DNA that are highly variable in length and mutate rapidly. They are thus an important source of genetic variation. This variation is highly informative for population and conservation genetics. It has also been associated with several pathological conditions and with gene expression regulation. However, genome-wide surveys of TR variation in humans and closely related species have been scarce due to technical difficulties derived from short-read technology. Here we explored the genome-wide diversity of TRs in a panel of 83 human and nonhuman great ape genomes, in a total of six different species, and studied their impact on gene expression evolution. We found that population diversity patterns can be efficiently captured with short TRs (repeat unit length, 1–5 bp). We examined the potential evolutionary role of TRs in gene expression differences between humans and primates by using 30,275 larger TRs (repeat unit length, 2–50 bp). Genes that contained TRs in the promoters, in their 3′ untranslated region, in introns, and in exons had higher expression divergence than genes without repeats in the regions. Polymorphic small repeats (1–5 bp) had also higher expression divergence compared with genes with fixed or no TRs in the gene promoters. Our findings highlight the potential contribution of TRs to human evolution through gene regulation.

[Supplemental material is available for this article.]

Tandem repeats (TRs) are DNA tracts in which a short base-pair motif, the repeat unit, is repeated several times in tandem. They are among the most variable loci, experiencing mutations in the number of repeat units that are 100 to 100,000 times more frequent than point mutations (Weber and Wong 1993; Brinkmann et al. 1998; Li et al. 2002; Legendre et al. 2007). Due to their unique properties, TRs have been extensively used as molecular markers in many population genetic studies (Ellegren 2004). Past technical constraints, however, limited the number of TRs that could be easily genotyped. For this reason, most TR-based studies of human diversity and intraspecies genetic divergence were restricted (Rosenberg et al. 2002; Molla et al. 2009; Pemberton et al. 2009, 2013; Tishkoff et al. 2009; Sun et al. 2012) or focused on comparing reference genomes (Webster et al. 2002; Kelkar et al. 2008, 2011; Payseur et al. 2011; Loire et al. 2013). However, recent advances in sequencing methodology (for review, see Mardis 2008; Metzker 2010) and the development of novel software that can systematically genotype repeats at a genome-wide scale (for review, see Lim et al. 2012) have permitted analysis of several thousand loci from multiple individuals in a cost-effective manner (McIver et al. 2011, 2013; Willems et al. 2014).

In eukaryotes, TRs located in coding regions and their promoters tend to occur in genes associated with transcriptional regulation, DNA binding, protein–protein binding, and developmental processes (Vinces et al. 2009; Gemayel et al. 2010), suggesting a regulatory role for TRs. In fact, TRs are emerging as good candidates for a type of genomic variation that can directly alter gene expression (Rockman and Wray 2002; Kashi and King 2006; Vinces et al. 2009; Gemayel et al. 2010). Because gene expression changes might contribute to the fundamental differences between humans and other species (King and Wilson 1975), it is imperative to study mechanisms that may permit rapid expression changes on short evolutionary time scales (Wray et al. 2003; Tirosh

[11]**These authors contributed equally to this work.**
Corresponding authors: andreas.wagner@ieu.uzh.ch, tomas.marques@upf.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.190868.115.

et al. 2006, 2009; Landry et al. 2007; Choi and Kim 2008). Promoter features such as TATA boxes, nucleosome density, and tracts of TRs can mediate such changes (Tirosh et al. 2009). Thus, the high incidence of TRs in regulatory regions in some species (Gemayel et al. 2010; Payseur et al. 2011) and their high mutability (Weber and Wong 1993; Brinkmann et al. 1998; Li et al. 2002; Legendre et al. 2007), suggests that it may be important to study TR variation to understand fundamental differences in gene expression across species and populations. In particular, since TRs constitute 3% of the human genome (Lander et al. 2001) and are dramatically enriched in promoter regions (Vinces et al. 2009; Sawaya et al. 2013), clarifying their functional role may provide important insights for the human biology field.

To characterize the extent of TR variation in humans and great apes, as well as its influence on patterns of gene expression, we examined both TR polymorphisms and divergence in different human populations and their closest primate relatives, chimpanzees and other great apes. We also investigated a potential expression divergence between human and chimpanzees based on nonrandomly associated TRs located at promoters.

Results

Divergence and population diversity of TRs

We first assessed the degree of polymorphism of TRs in human and nonhuman great ape species and, in order to do so, we analyzed genotype variation in a total of 6,965,726 TR loci in humans, 4,006,024 in chimpanzees and bonobos, 3,815,198 in gorillas, and 2,313,198 in orangutans. Specifically, these were genotyped in a panel of 27 humans (two from Europe, two from Asia, one from America, one from Oceania, and 21 sampled throughout Africa, of which nine are newly sequenced), 16 chimpanzees (four subspecies) and 12 bonobos, 18 gorillas (two species, three subspecies), and 10 orangutans (two species), respectively. To produce this repeat catalog, we used the human reference genome assembly (hg19) as a reference for all individuals, in order to facilitate genome coordinate comparison and to take advantage of the thorough human gene annotation. Additionally, to avoid potential biases resulting from using a single reference sequence, we also genotyped the corresponding TRs in the reference genome of each genus, and filtered out any repeats whose genotypes were not concordant when comparing the results using both references (for information about this and additional filtering criteria, see Supplemental Text S1, S2). For the genotyped repeat loci, we estimated average differences in absolute repeat number relative to those annotated on the hg19 human reference genome in different subspecies and species of nonhuman primate taxa, and in three human groups (non-Africans, African hunter-gatherers, and all other Africans). The total amount of repeat number differences accumulated by each group was consistent with what was expected based on their genetic distance to the human reference genome (Fig. 1A). We also found that the general patterns of heterozygosity between species and subspecies were highly concordant with those of SNP diversity in great apes on a similar data set (Fig. 1B; Prado-Martinez et al. 2013). Principal component analysis (PCA) performed with TR allele frequencies across the human samples replicated known diversity patterns among human populations (Supplemental Fig. S1A; Tishkoff et al. 2009). By using the PCA approach, we were also able to partition not only between

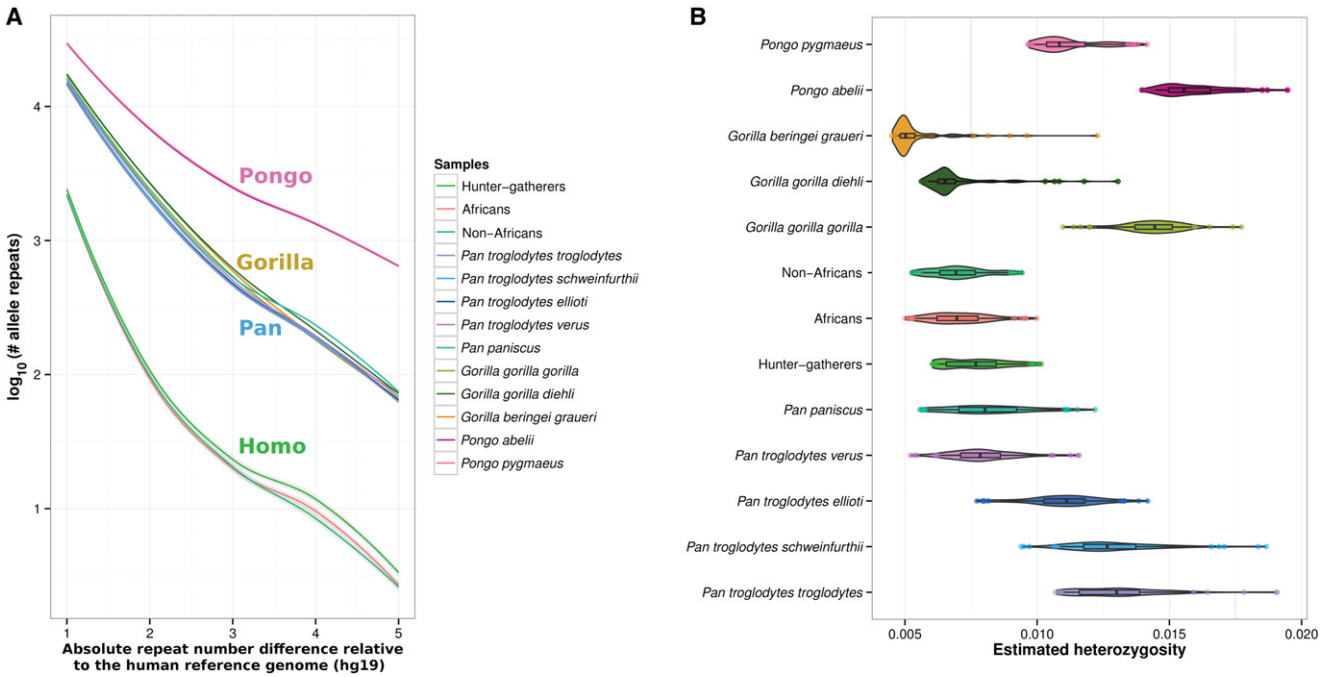

Figure 1. Repeat copy-number differences relative to human reference genome (hg19) and heterozygosity estimates for several nonhuman great apes species and human groups. (*A*) Absolute repeat number differences relative to the human reference genome estimated to occur for different human groups and subspecies/species of nonhuman primate taxa. The *x*-axis shows the number of repeat copy-number differences; the *y*-axis, the number of events for each repeat number difference in $\log_{10}$ scale. As expected, humans show the fewest differences relative to the human reference genome and are followed by chimpanzees, bonobos, gorillas, and orangutans. Within humans, African hunter-gatherers show the most variation, followed by Africans and non-Africans. (*B*) Heterozygosity estimates for different human groups and subspecies/species of nonhuman primate taxa. These results show great concordance with previous genetic surveys using millions of SNPs.

great ape subspecies but also between different great ape populations according to their geographical origin, with just up to four principal components. Even for the great apes, that display a very complex taxonomy (chimpanzees), each of the first three PCs separated the four chimpanzee subspecies currently described in the literature, while the fourth PC identified substructure in the central and eastern subspecies, which corresponds to the sample's geographical origin (Supplemental Fig. S1B). Likewise, for gorillas and orangutans, the PCA could clearly separate between species and subspecies (Supplemental Fig. S1C,E). Similar diversity patterns were observed when employing STRUCTURE (Pritchard et al. 2000) on the same set of markers used for the PCA for each taxon (see Supplemental Material Text S3), and when calculating for each taxon the standardized R_{ST} value (Slatkin 1995) between all pairs of individuals (see Supplemental Material Text S4). A list of ancestry-informative markers (AIMs) for the subspecies of chimpanzees, gorillas, and for the species of orangutans is provided in Supplemental Table S1 (see Supplemental Material Text S5).

The impact of TRs on gene expression

To identify all genes with any kind of TR in the 5 kbp upstream of the transcription start site (TSS), we used a set of 13,035 one-to-one orthologous genes in the reference genome assemblies of human (NCBI GRCh37.p10, replicated the results also on GRCh38) (Supplemental Fig. S4), chimpanzee (CHIMP2.1.4), and macaque (MMUL_1; see Methods). We found that on average 29%–31% of these genes harbored TRs (3820, 3910, and 4032 for human, chimpanzee, and macaque, respectively). To check the functionality of these repeats, we carried out a randomization test ($n = 500$) using data from ENCODE (The ENCODE Project Consortium 2012) on DNase hypersensitive site locations in human lymphoblastoid cells (Sabo et al. 2006), which are accessible regions of DNA associated with gene regulatory elements (Gross and Garrard 1988). We found that 60% of TRs overlap a DNase hypersensitive site. The significant enrichment of repeats in DNase hypersensitive sites (P-value = 10^{-350}) suggests that a substantial part of repeat sequences could potentially be involved in gene regulation.

Based on this premise, we used publicly available RNA-seq gene expression data (Brawand et al. 2011) to assess whether genes that contain TRs in their promoters have higher expression divergence than those that do not. To this end, we computed the mean of gene expression values belonging to different individuals for each gene and organ. In order to compute expression divergence between each pair of species, we calculated the difference between the mean expression values of the orthologous gene pairs, normalized by the sum of the mean expression values in a given organ. We then partitioned these pairwise expression differences into two subsets according to whether orthologous genes did or did not contain TRs in their promoters. We observed a significant increase in pairwise expression differences when genes have TRs in their promoters. More specifically, between human and chimpanzee orthologs with repeats within 5 kbp upstream of their TSS showed higher mean expression difference (0.264) compared with those without repeats (0.257, $P < 10^{-6}$, based on Wilcoxon rank-sum test [WRS]) (Mann and Whitney 1947). Similarly, human–macaque orthologs ($P < 0.01$, for all organs) and chimpanzee–macaque orthologs ($P < 10^{-5}$, for all organs) with TRs showed a higher mean expression difference (0.250 and 0.254, respectively) than orthologous genes without repeats (0.243 and 0.242, respectively). In order to avoid noise and bias for organ-specific gene expression variation differences, we next took a phylogenetic approach and performed a bootstrap-like resampling analysis, where gene expression values were sampled from different individuals of a species (see Methods). We computed two different expression distance matrices of (1000 replicates) × (three species pairs) for each organ and employed these matrices to construct neighbor-joining gene expression trees. Except for the macaque branch for liver- and heart-specific expression trees, all branches were significantly longer for repeat-containing genes in both species ($P < 10^{-10}$; based on a t-test with $n = 1000$, df = $n-1$, throughout unless otherwise mentioned) (Supplemental Fig. S5). The total tree length of genes with repeats was significantly greater in all organs ($P < 10^{-200}$ except for liver, where $P = 0.02$) (Fig. 2A). Repeats with total identity, higher Tandem Repeats Finder (TRF) (Benson 1999) score, and shorter repeat units yielded higher expression divergence differences between genes with TRs and other genes, consistent with positive correlation of such repeats to greater polymorphism (O'Dushlaine and Shields 2008). Our results were robust even when changing the tool to identify TRs in the genomes (http://www.ruhr-uni-bochum.de/ecoevo/cm/cm_phobos.htm) (Phobos 3.3.11) (Supplemental Fig. S6) or when adding changes in the parameters of TR identification (Supplemental Fig. S7).

When changing the distance of the upstream regions considered, we found that repeat-containing genes diverged more rapidly in their expression, and this difference was most pronounced for

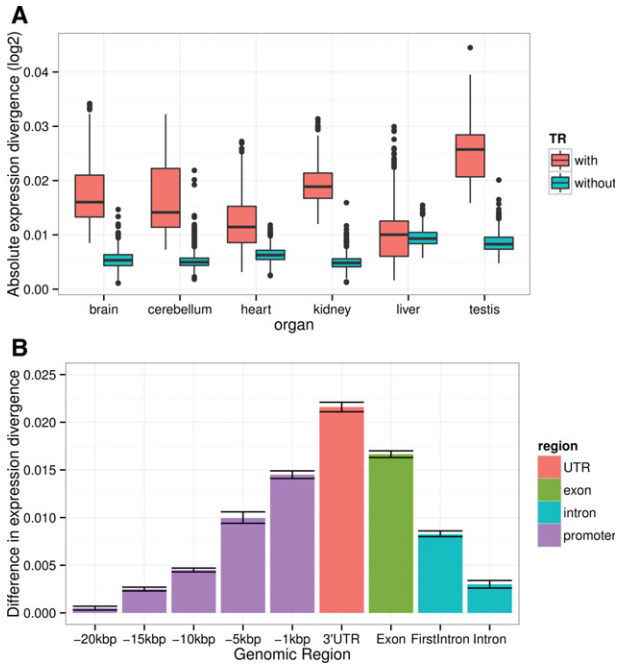

Figure 2. Relationship between expression divergence (normalized by mean tissue expression divergence) and the presence of repeats in gene promoters and other genic regions. (*A*) Boxplot of total tree lengths of genes with repeats (red) and genes without repeats (blue). Horizontal lines in the *middle* of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3% of the data points. (*B*) Presence of tandem repeats associate with higher expression divergence. Bars, mean differences in expression divergence, based on pairwise expression tree length differences between repeat-containing and non-repeat-containing genes. Repeats found in upstream regions of lengths of 20, 15, 10, 5, 1 kbp, as well as in 3′ UTRs, exons, first introns, and all introns were considered, as indicated on the horizontal axis. Note that all expression differences are positive, indicating that repeat-containing genes, regardless of category, diverged more rapidly. Whiskers represent 99.3% confidence intervals.

repeats within 1 kbp upstream of the TSS (95% CI: 0.0145, 0.0146). The difference got progressively smaller as we included repeats that are further away from the TSS (95% CI for windows of length 10 kbp: 0.003, 0.006; 15 kbp: 0.0021, 0.0024; 20 kbp: 0.0004, 0.0007) (Supplemental Fig. S8). Repeats in other potential regulatory regions also have an influence on the divergence of gene expression. Specifically, genes containing TRs within 1 kbp of the 3′ UTR also showed significantly greater expression divergence in all organs except the testis ($P < 10^{-47}$, for all organs) (Supplemental Fig. S9A). Moreover, those 2468 human genes with exon-containing repeats also showed greater expression divergence in all organs ($P < 10^{-269}$, for all organs) (Supplemental Fig. S9B). Finally, repeats in introns were also associated with greater expression divergence ($P < 10^{-198}$) for all organs except for the heart ($P < 10^{-85}$) and the liver ($P < 10^{-292}$), both of which show opposite patterns (Supplemental Fig. S9D). The mean difference of tree lengths for repeats found in any intron was smaller compared with the mean difference for the repeats found in first introns (95% CI: 0.0080, 0.0086). Overall, our gene expression trees indicate greater expression divergence between genes without repeats and genes that contain repeats, in the following order of decreasing divergence: repeats in 3′ UTR regions (mean difference in tree lengths: 0.022) > repeats in exons (0.017) > repeats in promoters (0.015) > repeats in 1st intron (0.008) > repeats in any intron (0.003) (Fig. 2B).

We wondered whether the observed association between TRs and expression divergence is simply due to relaxed selection. We therefore performed multiple analyses to compare the level of selection in genes and their promoters with and without TRs. Between these two sets of genes, we did not find any significant difference in the sequence divergence of the coding sequences (based on the d_N/d_S ratios for single-copy genes, 0.205 ± 0.35 and 0.212 ± 0.35, respectively, $P = 0.70$) (see Supplemental Text S6; Supplemental Fig. S10) or of the promoter sequences (based on the number of SNPs between human and chimpanzee genomes [identified by Prado-Martinez et al. 2013], WRS, $P = 0.36$) (see Supplemental Text S7; Supplemental Fig. S11). Moreover, we identified on average 5.09 recombination hotspots in repeat-containing promoters, whereas other promoters contained 5.27 hotspots on average (WRS, $P = 0.016$) (Supplemental Text S8; Supplemental Fig. S12), suggesting that genes with TRs do not experience more recombination events than genes without TRs. We also found that promoters with TRs are not in close proximity of particular chromosome locations, such as centromeres or telomeres. In fact, they were significantly away from both regions (WRS, $P = 0.001$ and $P = 10^{-14}$, respectively) (Supplemental Text S8). Finally, we wondered if repeat-containing genes are enriched for specific functions that associate with relaxed selective constraints. However, we found no enrichment for any function (see Supplemental Text S10). We then wondered if other regulatory factors were involved in the observed association between TRs and expression divergence. First, we repeated our analysis for highly and lowly expressed genes separately, as gene expression levels may play a role in expression divergence (Lehner 2008; Macneil and Walhout 2011; Pilpel 2011). We found that the association between TRs and the expression divergence holds for highly ($P < 10^{-117}$ in all organs) and lowly expressed genes ($P < 10^{-129}$ in all organs). Second, as CpG islands and promoter GC content have gene regulatory roles through epigenetic mechanisms (Fenouil et al. 2012), we asked whether promoters with and without TRs differ in their CpG content. Although we found a small increase in GC in promoters with TRs (WRS, $P = 0.02$) (Supplemental Fig. S13), the association between TRs and expression divergence was still significantly high ($P < 10^{-178}$, for all organs except for liver, where $P = 10^{-6}$) when TRs in CpG islands were excluded from the analysis (see Supplemental Text S11). Altogether, these different approaches suggest no cofactor or evidence for more relaxed selection in genes with TRs in their promoters compared to those without.

To investigate to what extent polymorphism within species might be affecting our results, we used our set of genotyped TRs and classified repeats either as polymorphic or fixed (no variation in copy number of repeat unit) within each taxon. To this end, we used our genotyped TR data in the panel of 27 humans and 16 chimpanzees. We found that for all tissues, genes with promoters containing repeats observed to be polymorphic in both taxa showed significantly more divergence than (1) genes with the same repeat genotype (repeat length) in both taxa and (2) genes without repeats (WRS, $P < 10^{-168}$ and $P < 10^{-163}$, respectively) (Fig. 3). We again did not find a difference between the three categories analyzed (Supplemental Fig. S14) in terms of selective constraints (Supplemental Fig. S15). By using the genotyping data for human and chimpanzee, as well as for the other three primate taxa (bonobos, gorillas, and orangutans), we finally checked for overrepresented biological process terms in genes that contained repeats in three categories: (1) genes whose repeat genotype length was different between human and chimpanzees and fixed within each taxon ($n = 2804$), (2) genes whose repeats were the same and fixed in all genotyped nonhuman primates but polymorphic in humans ($n = 1754$), and (3) genes whose repeats are fixed in humans but polymorphic in all genotyped nonhuman primates ($n = 2178$). We found that for the first two categories processes related to cell adhesion, neurogenesis, and neural development are enriched, while for category 3 processes related to detection and response to chemical and biotic stimuli, sensory perception of taste and smell, and skin development are enriched (Supplemental Table S2). Finally, and using a list of TRs known to be associated with human diseases (Supplemental Table S3), we observed that in half of the loci surveyed, the distribution of repeat allele length was significantly different between the two species and that humans showed on average less allele repeat length variation (Supplemental Fig. S16).

Discussion

TRs are abundant in human and nonhuman primate genomes, but their variation in natural populations and their potential regulatory role remain largely unexplored. Recent advances in sequencing technology, together with the development of computational tools, have finally made it affordable to accurately genotype up to millions of TRs at the genome-wide level and to circumvent much of the constraints that traditional repeat genotyping methods entail (Gymrek et al. 2012; Guilmatre et al. 2013; Highnam et al. 2013; Duitama et al. 2014; Willems et al. 2014; Carlson et al. 2015). Furthermore, quantitative deduction of gene expression divergence based on distance trees allowed reconstructing global evolutionary trends in more detail, as they are highly consistent with the known phylogeny. For example, total tree lengths correlate with organ-specific evolutionary rates, where the genes in testis show greater tree length compared with the genes in the other organs in mammals (Brawand et al. 2011). Likewise, genes located on the X Chromosome present higher expression divergence compared with genes on the autosomal chromosomes, consistent with the high evolutionary rates of sex chromosomes (Brawand et al. 2011).

Here, we surveyed the genome-wide diversity of TRs in a large collection of high-quality human and nonhuman great ape

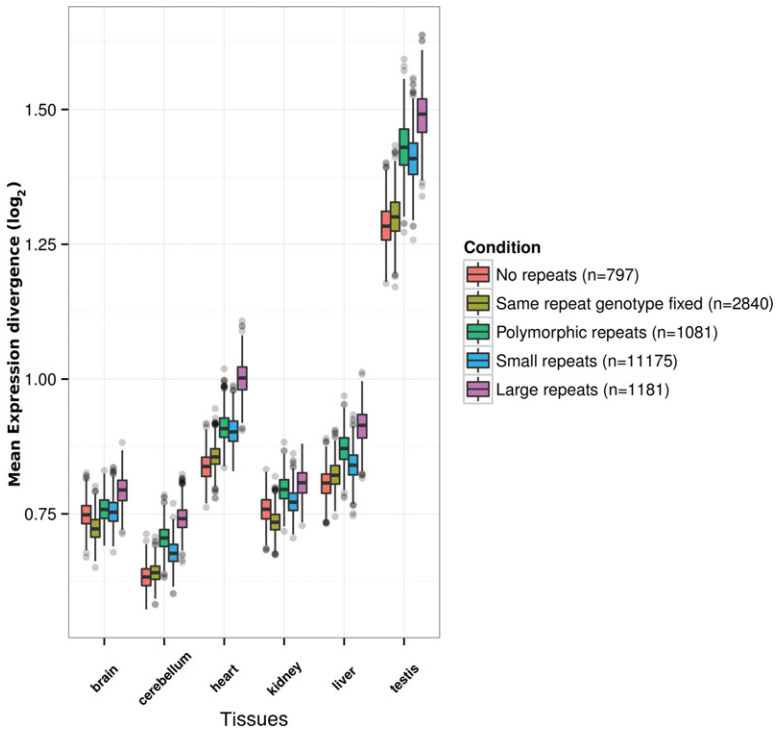

Figure 3. Relationship between expression divergence and within-species repeat genotype conservation in human and chimpanzees. Boxplots produced by resampling 1000 data points, each corresponding to the average log_2-transformed expression divergence value between human and chimpanzee for a particular tissue and for genes associated to a particular category. The tissues are shown on the *x*-axis, and the *y*-axis corresponds to the absolute mean expression divergence between humans and chimpanzees. The categories considered are as follows: genes with no repeats in promoters (red), genes containing exclusively repeats in promoters that have the same repeat length fixed across all human and chimpanzee samples (olive), genes containing exclusively repeats that are polymorphic in human and chimpanzees (green), genes containing small repeats (repeat unit length of 1–5 bp and <100 bp total repeat length) in their promoter (blue), and genes containing large repeats (repeat unit length of 2–50 bp) in the promoter (magenta). Genes lacking repeats in promoters or repeats for which the same repeat genotype length is found across human and chimpanzee samples show the least amount of expression divergence for all tissues.

genomes and showed for the first time their possible impact on gene expression divergence between human and other primates. We observed an association between polymorphic repeats in gene promoters and increased expression divergence, an observation that was robust to changes in the method used to identify TRs and to assess gene expression divergence. This association existed for most of all organ-specific expression data, except in some cases for testis and liver. Distinct expression patterns in these organs have also been observed by others (Hsieh et al. 2003; Somel et al. 2008; Brawand et al. 2011) in different contexts.

Repeats closer to the TSS were associated with greater expression divergence, an observation that might be explained through core promoter modules occurring preferentially close to this site and exerting a strong influence over transcriptional regulation (Wray et al. 2003; Spitz and Furlong 2012). Moreover, an association with expression divergence held also for repeats in other genic regions. The strongest of them was evident for 3′ UTRs, consistent with their known role in gene regulation (Yoon et al. 2012). In addition, repeats in first introns were associated with greater expression divergence than repeats in other introns. This observation is consistent with previous work showing that most intronic regulatory regions occur in the first intron (Rohrer and Conley 1998), that the first intron has the highest divergence between human and chimpanzees (Gazave et al. 2007), and that the first intron influences gene expression more than other introns (Jonsson et al. 1992; Charron et al. 2007).

Because the observed association between TRs and expression divergence could be a consequence of relaxed selection acting on sequence and gene expression, we performed multiple analyses to show that neither the coding nor the promoter sequences of the genes with promoter TRs experience relaxed selection compared with the genes without TRs. Also, we tested for other factors (functional category of genes, gene expression level, CpG islands, or genomic location) that could co-correlate with TR presence and changes in expression, to cause the observed correlation. Since none of the possible confounding factors could explain the expression divergence between the set of genes with and without TRs, our results support a potential causal relationship of the presence of TRs and changes in expression patterns in a substantial fraction of genes. This claim is also supported by polymorphic repeats, particularly compared with genes without repeats or with genes in which the same repeat genotype is fixed across human and chimpanzee samples, thus suggesting that repeat variation may elicit changes in gene expression levels across species.

Intriguingly, genes containing repeats whose genotype is conserved across nonhuman primates but polymorphic or fixed for a different genotype in humans seem to be enriched for functions related to neurogenesis and neural differentiation, as well as to development of the nervous system and cell adhesion. Furthermore, genes for which all repeats are fixed in humans but not fixed in other nonhuman primates seem to be enriched for processes related to stimulus detection, sensory perception, and skin development. Such biological processes may be associated with the evolution of human-specific cognitive traits and the response to new environments, respectively.

Our results provide further motivation for future studies to clarify the exact role of these genes in primate evolution and the extent to which repeats may have been involved in their regulation. Accumulating evidence from exhaustive genetic studies has already shown that TR variation has dramatic, often background-dependent phenotypic effects in model organisms (Verstrepen et al. 2005; Kashi and King 2006; Fondon et al. 2008; Borel et al. 2012; Egbert and Klavins 2012; Morrison et al. 2012; Raveh-Sadka et al. 2012). In yeasts, TR variation in promoters has been shown to alter gene expression (Vinces et al. 2009). An especially remarkable example in mammals is the features of a dog's snout, such as the degree of dorsoventral nose bend and midface length, which correlate with the ratio of the length of two TRs in a gene that regulates bone formation (Fondon and Garner 2004). Furthermore, a recent study (Hellen and Kern 2015) has found that TR insertions are significantly more frequent than expected

compared to other types of insertions and that they are mostly fixed in the human lineage. Taken together with all other studies, our observations further suggest a potential contribution of TRs in primate gene expression evolution.

We also showed how genome-wide short TRs genotyped from whole-genome sequencing data provide a valid means to uncover substructure and divergence patterns in human populations and great ape species by showing that they agreed with previous surveys in human and nonhuman great apes based on single-nucleotide polymorphisms (Li et al. 2008; Prado-Martinez et al. 2013). These markers could then potentially be used for conservation and breeding programs of great apes, since their higher mutation rate means that a low number of markers can be used, an advantage for conservation studies, which often use highly degraded noninvasive samples.

Several limitations are associated with this work. First, repeat content, repeat length, and repeat unit sizes may be affected by different mutation rates. While one would ideally want to take these differences into account, doing so would have limited our statistical power to detect structure patterns at several levels. Another limitation stems from whole-genome sequencing genotyping with short-read technologies and regards the maximum length of a repeat that can be genotyped. The reason is that genotyping a repeat requires it to be wholly encompassed by any given short read. Because we wanted to analyze only nonoverlapping repeats, the sequence reads of ~100 bp used in this study imply that we were limited to the analysis of only ~58% of the total fraction of TRs we identified in the human reference genome using Tandem Repeats Finder (TRF). Nonetheless, we were still able to genotype thousands of TRs and assess their conservation within and across different primate natural populations.

In the future, studies of the repeat landscape will be facilitated not only by the use of longer reads but also by a thorough subsequent genotyping of a subset of repeats, using one of the recently developed methods that specifically target repeats (Guilmatre et al. 2013; Duitama et al. 2014; Carlson et al. 2015). These methods rely on a presequencing enrichment step for repeats. They are currently able to target and genotype several thousand repeats in many individuals and do so in a much more accurate fashion than in silico methods. The combination of the two approaches will yield much reliable and important information regarding repeat variation in natural populations. Such TR genotyping from whole-genome sequencing data will have a profound impact on many fields, from conservation genetics to forensics, and on elucidating the role of TRs in complex trait heritability (Press et al. 2014).

In a seminal paper, King and Wilson (1975) observed about humans and chimpanzees that "their macromolecules are so alike that regulatory mutations may account for their biological differences." Since then, we have learned that such mutations, and in particular mutations that cause gene expression change, are indeed important in the evolution of primates and other organisms (Wren et al. 2000; Stranger et al. 2005, 2007; Fondon et al. 2008; Dimas et al. 2009; Vinces et al. 2009; Gemayel et al. 2010). Our work shows that TRs, a type of sequence with unusually high mutability, are a relevant class of regulatory mutations that might contribute to such species differences.

Methods

Genomic sequence data

We used a total of 83 samples sequenced with Illumina paired-end reads at high-coverage (more than 20×) and with read length ranging from 50–102 bp. Specifically, we used both publicly available data sets and newly sequenced human genomes. For the nonhuman primates, we used the complete collection from the Great Apes Genomic Project (Prado-Martinez et al. 2013). In addition, we selected a total of 27 human male samples from other genome sequencing studies (for sample information, see Supplemental Text S12; Supplemental Table S4). Raw sequence data from already published human and nonhuman great ape genomes are available through the Sequence Read Archive (SRA; SRP018689, SRP009145, SRP001139, and SRP001703) and at https://www.simonsfoundation.org/life-sciences/simons-genome-diversity-project-dataset/. For information on how to access data from nine newly sequenced African genomes, see the section "Data access."

TR identification

For the genomic diversity data analysis, we used the set of TRs that come with the RepeatSeq (Highnam et al. 2013) software. This is a set containing TRs with a repeat unit between 1 and 5 bp in length, identified in the human reference genome (hg19) using TRF (Benson 1999) v2.30, with parameters "2 5 5 80 10 14 5" (for further clarification regarding the parameters, see Supplemental Table S5), and further filtered so that any two repeats are at least 21 bp apart. By use of the same parameters, we also identified TRs in the panTro2.1.4, gorGor3, and ponAbe2 reference genomes, corresponding respectively to chimpanzee, gorilla, and orangutan.

Additionally, we also identified TRs occurring only in or near genes, using more stringent parameters and allowing for larger repeat motifs. Here, we considered TRs with repeat units from 2 nucleotides (nt) up to 50 nt in length (for details, see Supplemental Material Text S13).

Mapping and genotyping of TRs and their validation in nonhuman primates

In order to estimate the mean and standard deviation of the insert size between the reads in each set of paired-end reads, we first randomly sampled 50,000 reads from each sample and mapped them to the human reference genome (hg19) using Novoalign aligner version 2.08.02 (http://www.novocraft.com). We then mapped all the reads from each sample to the human genome reference using Novoalign with all parameters set to default and supplying the mean and standard deviation estimates previously computed. When mapping, we realigned reads located around indels using the Genome Analysis Toolkit (GATK) version 2.5 (McKenna et al. 2010) and used Picard Tools v1.7 (http://picard.sourceforge.net/) for removal of duplicate reads. Finally, using the set of TRs identified with TRF (Benson 1999) we genotyped all samples using RepeatSeq version 0.8.2 (Highnam et al. 2013), with default parameters and setting the "-emitconfidentsites" option. An approach similar to the one described above was repeated for one sample each from chimpanzee, gorilla, and orangutan sample sets, but using the corresponding species genome reference, in order to generate sets of TRs that were then compared with the ones obtained using the human genome reference. The set of TRs that were discordant between the two approaches was removed from further analyses (for details, see Supplemental Text S2).

Absolute repeat copy-number distance from human reference genome and heterozygosity estimates

We estimated absolute repeat copy-number distance from the human reference genome for groups corresponding to all chimpanzee and gorilla subspecies, bonobos, the two orangutan species, and three groups of humans, namely, African hunter-gatherers, all Africans that did not fit in the former category, and non-

Africans. For all repeat loci genotyped, we first subtracted the repeat copy number found by TRF in the human reference genome from the repeat copy number of each allele. By use of these values, we then implemented for each group a bootstrap-like resampling approach, which we repeated for 100 times, at each instance randomly selecting a new set of repeat loci and samples, to generate estimates of the amount of occurrences of the different repeat copy-number differences relative to the human genome reference (for details, see Supplemental Material Text S14).

We quantified the number of times each repeat copy-number difference was observed within each group in each of the 100 iterations and plotted these results with the geom_smooth function of ggplot (Wickham 2009), which performs a local regression fit to the data. Additionally, we estimated heterozygosity in each group using the same set same set of one hundred resamplings previously generated for estimating the absolute repeat copy-number distance from the hg19 reference (for details, see Supplemental Text S15).

TR genomic context

We characterized TRs according to whether they were located within promoters, splice sites, 3′ or 5′ UTRs, exons, or introns using the GenomicRanges package version 1.14.4 (Lawrence et al. 2013).

Repeat genotype conservation

We classified TRs in each taxon as being fixed or polymorphic across the genotyped samples. For a TR to be considered as fixed within a taxon, we required that >40% of the samples had been genotyped and that at least 95% of these share the same repeat genotype length. For a TR to be considered polymorphic, it needed to have at least four samples genotyped, with at least two different genotypes present. We classified each repeat locus as polymorphic or fixed in both human and nonhuman primates.

Gene ontology

We used the topGO R package (https://www.bioconductor.org/packages/release/bioc/html/topGO.html) to search for gene ontology terms related to biological processes that may be enriched both in human-specific repeats identified at the reference level and for particular TR genotype conservation categories within taxa (fixed or polymorphic). We used a Fisher's exact test to infer this overrepresentation, followed by an adjustment for multiple testing of the *P*-values produced using the Benjamini-Hochberg method (Kasen et al. 1990).

Gene expression and sequence data

The gene expression data we used were based on RNA sequencing of approximately 3.2 billion 76-bp Illumina Genome Analyzer IIx reads (Brawand et al. 2011). Expression levels are indicated as $\log_2$-transformed reads per kilobase pair of exon model per million mapped reads. This provided one-to-one gene expression measurements from multiple primates, where each gene's expression had been measured in six different organs (brain, cerebellum, heart, kidney, liver, testis) for between one and six individuals per species. From this data set, we used RNA-seq based expression values of all 13,035 one-to-one gene orthologs from humans, chimpanzees, and macaques. We obtained DNA sequences of the genes in our expression data set through the BioMart tool of Ensembl (Kinsella et al. 2011), using human annotation version GRCh37.p10, chimpanzee annotation CHIMP2.1.4, and macaque annotation MMUL_1.

Gene expression divergence

The gene expression data set we used (Brawand et al. 2011) contained gene expression measurements from several individuals of a species for each gene and organ. We took advantage of this fact to assess statistical differences in gene expression divergence with a bootstrap-like resampling procedure, where we sampled gene expression values from different individuals of a species to create 1000 replicate data sets ($n = 13{,}035$) for each organ and species.

We partitioned gene pairs in each such data set into two groups: gene pairs where genes of a given species contained TRs in a specific region of interest, such as a promoter, and gene pairs without such repeats. We then computed, separately for genes in the two groups, a pairwise matrix of Euclidean gene expression distance between all genes in a pair of species (for details, see Supplemental Material Text S16).

Overall, we created 12 separate expression distance matrices of size (1000×3) for two gene subsets based on repeat presence and for six organs. We used these matrices to construct gene expression trees using the neighbor-joining approach (implemented in the "ape" package [Paradis et al. 2004] in R [R Core Team 2015]). We used the branch lengths of the trees we constructed as a measure of gene expression divergence. To test the null hypothesis that the expression divergences (branch lengths) of the 1000 sampled trees were significantly different between the two gene subsets for each organ, we used paired *t*-tests ($N = 1000$, $df = n - 1$ unless otherwise mentioned). All *P* values are reported after Bonferroni correction (Dunn 1961) for multiple testing, and they were robust to number of bootstrap replicates. We performed all statistical analyses using MATLAB (7.10.0; The MathWorks, 2010).

Expression divergence versus repeat genotype conservation

We used again the expression data available for genes present both in human and chimpanzee (Brawand et al. 2011) and calculated the mean absolute expression divergence between human and chimpanzee for each tissue and TR genotype conservation category by randomly sampling 1000 genes with replacement. For each of these genes, we sampled the expression for both species from among the samples (up to six by tissue) with expression not equal to zero and calculated the $\log_2$ of their absolute expression levels. We then ranked these by their absolute expression divergence level and averaged their value after removing the top and bottom 25% of values. This step was repeated 100 times, and the values were used to generate the boxplot.

To test for significance across different categories' comparisons, we used the wilcox.test R function to implement a Wilcoxon rank-sum statistical test.

Data access

The raw sequencing data from nine newly sequenced African genomes have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession number SRP052818.

Acknowledgments

T.B.S. thanks Maria Anisimova for helpful discussions. T.C. thanks Javier Quilez for kindly providing the repeat disease-associated list, Marcos Fernandez for his technical assistance, and Irene Hernando-Herraez, Javier Prado-Martinez, and Rui Faria for their help and constructive comments. A.W. and T.B.S. acknowledge support through Swiss National Science Foundation grant

31003A_146137, as well as through the University Priority Research Program in Evolutionary Biology at the University of Zurich. T.M.B. was supported by a European Research Council (ERC) starting grant (260372), ICREA, EMBO Young Investigator Award, and the Ministerio de Ciencia e Innovación (MICINN) (BFU2014-55090-P, Spain). This work was also supported by the National Institutes of Health (NIH) grants HG006696, MH097018, and DA033660, and by the March of Dimes Foundation grant 6-FY13-92. Authors thank the reviewers for their helpful comments.

References

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573–580.

Borel C, Migliavacca E, Letourneau A, Gagnebin M, Béna F, Sailani MR, Dermitzakis ET, Sharp AJ, Antonarakis SE. 2012. Tandem repeat sequence variation as causative *Cis*-eQTLs for protein-coding gene expression variation: the case of *CSTB*. *Hum Mutat* **33:** 1302–1309.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478:** 343–348.

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B. 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet* **62:** 1408–1415.

Carlson KD, Sudmant PH, Press MO, Eichler EE, Shendure J, Queitsch C. 2015. MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals. *Genome Res* **25:** 750–761.

Charron M, Chern J-Y, Wright WW. 2007. The cathepsin L first intron stimulates gene expression in rat sertoli cells. *Biol Reprod* **76:** 813–824.

Choi JK, Kim Y-J. 2008. Epigenetic regulation and the variability of gene expression. *Nat Genet* **40:** 141–147.

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, et al. 2009. Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science* **325:** 1246–1250.

Duitama J, Zablotskaya A, Gemayel R, Jansen A, Belet S, Vermeesch JR, Verstrepen KJ, Froyen G. 2014. Large-scale analysis of tandem repeat variability in the human genome. *Nucleic Acids Res* **42:** 5728–5741.

Dunn OJ. 1961. Multiple comparisons among means. *J Am Stat Assoc* **56:** 52.

Egbert RG, Klavins E. 2012. Fine-tuning gene networks using simple sequence repeats. *Proc Natl Acad Sci* **109:** 16817–16822.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5:** 435–445.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I, et al. 2012. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res* **22:** 2399–2408.

Fondon JW, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci* **101:** 18058–18063.

Fondon JW 3rd, Hammock EA, Hannan AJ, King DG. 2008. Simple sequence repeats: genetic modulators of brain function and behavior. *Trends Neurosci* **31:** 328–334.

Gazave E, Marqués-Bonet T, Fernando O, Charlesworth B, Navarro A. 2007. Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biol* **8:** R21.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet* **44:** 445–477.

Gross DS, Garrard WT. 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57:** 159–197.

Guilmatre A, Highnam G, Borel C, Mittelman D, Sharp AJ. 2013. Rapid multiplexed genotyping of simple tandem repeats using capture and high-throughput sequencing. *Hum Mutat* **34:** 1304–1311.

Gymrek M, Golan D, Rosset S, Erlich Y. 2012. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res* **22:** 1154–1162.

Hellen EHB, Kern AD. 2015. The role of DNA insertions in phenotypic differentiation between humans and other primates. *Genome Biol Evol* **7:** 1168–1178.

Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. 2013. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* **41:** e32.

Hsieh W, Chu T, Wolfinger RD, Gibson G. 2003. Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* **165:** 747–757.

Jonsson JJ, Foresman MD, Wilson N, Mclvor RS. 1992. Intron requirement for expression of the human purine nucleoside phosphorylase gene. *Nucleic Acids Res* **20:** 3191–3198.

Kasen S, Ouellette R, Cohen P. 1990. Mainstreaming and postsecondary educational and employment status of a rubella cohort. *Am Ann Deaf* **135:** 22–26.

Kashi Y, King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet* **22:** 253–259.

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res* **18:** 30–38.

Kelkar YD, Eckert KA, Chiaromonte F, Makova KD. 2011. A matter of life or death: how microsatellites emerge in and vanish from the human genome. *Genome Res* **21:** 2038–2048.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188:** 107–116.

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* **2011:** bar030.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. 2007. Genetic properties influencing the evolvability of gene expression. *Science* **317:** 118–121.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9:** e1003118.

Legendre M, Pochet N, Pak T, Verstrepen KJ. 2007. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res* **17:** 1787–1796.

Lehner B. 2008. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol Syst Biol* **4:** 170.

Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol* **11:** 2453–2465.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319:** 1100–1104.

Lim KG, Kwoh CK, Hsu LY, Wirawan A. 2012. Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. *Brief Bioinform* **14:** 67–81.

Loire E, Higuet D, Netter P, Achaz G. 2013. Evolution of coding microsatellites in primate genomes. *Genome Biol Evol* **5:** 283–295.

Macneil LT, Walhout AJM. 2011. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res* **21:** 645–657.

Mann HB, Whitney DR. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* **18:** 50–60.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24:** 133–141.

McIver LJ, Fondon JW, Skinner MA, Garner HR. 2011. Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics* **97:** 193–199.

McIver LJ, McCormick JF, Martin A, Fondon JW, Garner HR. 2013. Population-scale analysis of human microsatellites reveals novel sources of exonic variation. *Gene* **516:** 328–334.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20:** 1297–1303.

Metzker ML. 2010. Sequencing technologies: the next generation. *Nat Rev Genet* **11:** 31–46.

Molla M, Delcher A, Sunyaev S, Cantor C, Kasif S. 2009. Triplet repeat length bias and variation in the human transcriptome. *Proc Natl Acad Sci* **106:** 17095–17100.

Morrison NA, Stephens AA, Osato M, Polly P, Tan TC, Yamashita N, Doecke JD, Pasco J, Fozzard N, Jones G, et al. 2012. Glutamine repeat variants in human RUNX2 associated with decreased femoral neck BMD, broadband ultrasound attenuation and target gene transactivation. *PLoS One* **7:** e42617.

O'Dushlaine CT, Shields DC. 2008. Marked variation in predicted and observed variability of tandem repeat loci across the human genome. *BMC Genomics* **9:** 175.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20:** 289–290.
Payseur BA, Jing P, Haasl RJ. 2011. A genomic portrait of human microsatellite variation. *Mol Biol Evol* **28:** 303–312.
Pemberton TJ, Sandefur CI, Jakobsson M, Rosenberg NA. 2009. Sequence determinants of human microsatellite variability. *BMC Genomics* **10:** 612.
Pemberton TJ, DeGiorgio M, Rosenberg NA. 2013. Population structure in a comprehensive genomic data set on human microsatellite variation. *G3 (Bethesda)* **3:** 891–907.
Pilpel Y. 2011. Noise in biological systems: pros, cons, and mechanisms of control. *Methods Mol Biol* **759:** 407–425.
Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* **499:** 471–475.
Press MO, Carlson KD, Queitsch C. 2014. The overdue promise of short tandem repeat variation for heritability. *Trends Genet* **30:** 504–512.
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959.
R Core Team. 2015. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, Zeevi D, Sharon E, Weinberger A, Segal E. 2012. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat Genet* **44:** 743–750.
Rockman MV, Wray GA. 2002. Abundant raw material for *cis*-regulatory evolution in humans. *Mol Biol Evol* **19:** 1991–2004.
Rohrer J, Conley ME. 1998. Transcriptional regulatory elements within the first intron of Bruton's tyrosine kinase. *Blood* **91:** 214–221.
Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. *Science* **298:** 2381–2385.
Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, et al. 2006. Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. *Nat Methods* **3:** 511–518.
Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N. 2013. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLoS One* **8:** e54710.
Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139:** 457–462.
Somel M, Creely H, Franz H, Mueller U, Lachmann M, Khaitovich P, Pääbo S. 2008. Human and chimpanzee gene expression differences replicated in mice fed different diets. *PLoS One* **3:** e1504.
Spitz F, Furlong EEM. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13:** 613–626.
Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavaré S, et al. 2005. Genome-wide associations of gene expression variation in humans. *PLoS Genet* **1:** e78.
Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315:** 848–853.
Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, et al. 2012. A direct characterization of human mutation based on microsatellites. *Nat Genet* **44:** 1161–1165.
Tirosh I, Weinberger A, Carmi M, Barkai N. 2006. A genetic signature of interspecies variations in gene expression. *Nat Genet* **38:** 830–834.
Tirosh I, Barkai N, Verstrepen KJ. 2009. Promoter architecture and the evolvability of gene expression. *J Biol* **8:** 95.
Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, et al. 2009. The genetic structure and history of Africans and African Americans. *Science* **324:** 1035–1044.
Verstrepen KJ, Jansen A, Lewitter F, Fink GR. 2005. Intragenic tandem repeats generate functional variability. *Nat Genet* **37:** 986–990.
Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. 2009. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324:** 1213–1216.
Weber JL, Wong C. 1993. Mutation of human short tandem repeats. *Hum Mol Genet* **2:** 1123–1128.
Webster MT, Smith NGC, Ellegren H. 2002. Microsatellite evolution inferred from human–chimpanzee genomic sequence alignments. *Proc Natl Acad Sci* **99:** 8748–8753.
Wickham H. 2009. *ggplot2*. Springer, New York.
Willems T, Gymrek M, Highnam G, Mittelman D, Erlich Y. 2014. The landscape of human STR variation. *Genome Res* **24:** 1894–1904.
Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* **20:** 1377–1419.
Wren JD, Forgacs E, Fondon JW, Pertsemlidis A, Cheng SY, Gallardo T, Williams RS, Shohet RV, Minna JD, Garner HR. 2000. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am J Hum Genet* **67:** 345–356.
Yoon OK, Hsu TY, Im JH, Brem RB. 2012. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS Genet* **8:** e1002882.

Received February 16, 2015; accepted in revised form August 14, 2015.