# Triplet repeat length bias and variation in the human transcriptome

Michael Molla[a,1,2], Arthur Delcher[b,1], Shamil Sunyaev[c], Charles Cantor[a,d,2], and Simon Kasif[a,e]

[a]Department of Biomedical Engineering and [d]Center for Advanced Biotechnology, Boston University, Boston, MA 02215; [b]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742; [c]Department of Medicine, Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115; and [e]Center for Advanced Genomic Technology, Boston University, Boston, MA 02215

Contributed by Charles Cantor, July 6, 2009 (sent for review May 4, 2009)

**Length variation in short tandem repeats (STRs) is an important family of DNA polymorphisms with numerous applications in genetics, medicine, forensics, and evolutionary analysis. Several major diseases have been associated with length variation of trinucleotide (triplet) repeats including Huntington's disease, hereditary ataxias and spinobulbar muscular atrophy. Using the reference human genome, we have catalogued all triplet repeats in genic regions. This data revealed a bias in noncoding DNA repeat lengths. It also enabled a survey of repeat-length polymorphisms (RLPs) in human genomes and a comparison of the rate of polymorphism in humans versus divergence from chimpanzee. For short repeats, this analysis of three human genomes reveals a relatively low RLP rate in exons and, somewhat surprisingly, in introns. All short RLPs observed in multiple genomes are biallelic (at least in this small sample). In contrast, long repeats are highly polymorphic and some long RLPs are multiallelic. For long repeats, the chimpanzee sequence frequently differs from all observed human alleles. This suggests a high expansion/contraction rate in all long repeats. Expansions and contractions are not, however, affected by natural selection discernable from our comparison of human-chimpanzee divergence with human RLPs. Our catalog of human triplet repeats and their surrounding flanking regions can be used to produce a cost-effective whole-genome assay to test individuals. This repeat assay could someday complement SNP arrays for producing tests that assess the risk of an individual to develop a disease, or become part of personalized genomic strategy that provides therapeutic guidance with respect to drug response.**

computational biology | genomics | genome | polymorphisms | tandem repeats

Genomic variability is key to the adaptation, survival, and relative fitness of species. Nature uses mutations, duplications, recombinations, inversions, deletions, and many other genomic mechanisms to create diversity in populations. Evolution selects against most deleterious genomic changes. The remainder result in the spectacular range of traits observed in living organisms. Current research in human genetics has focused on documenting the full vocabulary of genomic polymorphisms and relating them to specific traits. In particular, medical genetics concentrates on identifying variations that are associated with disease. Most of this work focuses on identifying SNPs or other small genetic changes and on documenting both the common variations and disease-associated mutations (1, 2). This is due, in part, to the availability of relatively inexpensive technologies to probe the landscape of common mutations. Recently, copy number variation has become a prominent topic, in part due to the realization that a good fraction of observed genomic variation may result from copy number differences rather than SNPs (3–5).

Short tandem repeats (STRs) are an important family of genetic polymorphisms that have either documented or suspected significance in human genetics, personal medicine, and evolutionary analysis (6–8). Several diseases have been associated with tandem-repeat length variation (7–9). The most prominent among them is the polyglutamine (CAG) trinucleotide repeat, which has been implicated in a number of devastating neurodegenerative diseases including Huntington's disease (10) and hereditary ataxias (11, 12). All Huntington's patients exhibit an expanded number of copies in the CAG tandem repeat subsequence in the N terminus of the huntingtin gene. Moreover, an increase in the repeat length is anti-correlated to the onset age of the disease (13). Multiple other diseases have also been associated with copy number variation of tandem repeats (8, 14). Researchers have hypothesized that inappropriate repeat variation in coding regions could result in toxicity, incorrect folding, or aggregation of a protein. In noncoding regions, repeats are associated with increased instability or fragility of the DNA or in altering the binding properties of DNA binding proteins involved in DNA metabolism or biological regulation.

Tandem repeat variation in genomic DNA has already been shown to be important to morphological evolution (15). The work presented here is motivated by the hypothesis that an individual's repeatome—both common and private variation—plays a large role in the genetic basis of susceptibility to a wide range of diseases and other traits. The advent of whole-genome sequencing and analysis enables us to make progress toward supporting this hypothesis.

To investigate personal variation in repeat length polymorphisms (RLPs) in human populations, we undertook the first whole-genome triplet-repeat variation analysis for two of the available human genomes: those of J. Craig Venter and James Watson. Given that noncoding regions have extensive repeats for which there is less information about their level of impact on phenotype, we confined our analysis to triplet tandem repeats (3-STRs) in protein coding regions and surveyed the relative frequency of RLPs found in exons vs. introns.

Our analysis of this small sampling of personal genomes provides evidence for several thought-provoking phenomena. We observe a relatively low rate of short RLPs in both exons and introns, all of which are biallelic. We also see that, in the case of these short RLPs, one allele almost always agrees with the chimpanzee genome. Conversely, we observed a relatively high rate of long RLPs, some of which are multiallelic. Also unlike short RLPs, most long RLPs show deviation from the chimpanzee genome in all of the observed human alleles.

Additionally, we observe a surprising periodic pattern in the lengths of some repeat types in noncoding sequence. We also propose a design for a cost-effective, genome-wide assay of RLPs in humans that can be used for more extensive future studies.

## Results

In our survey of human 3-STRs, we have: (*i*) compared the frequency of these polymorphisms in exons versus introns and, in turn, compared this relationship to the analogous one for SNPs, to make informal estimates of the relative selective pressures on

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

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]M.M. and A.D. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: molla@bu.edu or ccantor@sequenom.com.

This article contains supporting information online at www.pnas.org/cgi/content/full/0907112106/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0907112106

**Table 1. Basic counts of 3-STRs and their combined length and the number of polymorphisms in human introns and exons and the number of exons and introns, their total length and the number of SNPs in each**

|  |  |  |  | Ratio |
|---|---|---|---|---|
|  |  | Exon | Intron | (exon/intron) |
| No. of Regions |  | 85,741 | 167,485 |  |
| Total Length |  | 31,918,308 | 901,736,782 |  |
| No. of 3-STRs |  | 201,850 | 3,629,645 |  |
| Total Length of 3-STRs |  | 1,576,309 | 27,681,474 |  |
| RLPs (Venter) | No. | 93 | 4,268 | 0.0218 |
|  | Rate | 1/3.43× $10^5$bp | 1/2.11× $10^5$ bp | 62% |
| SNPs (Venter) | No. | 17,765 | 899,970 | 0.0197 |
|  | Rate | 1/1.80× $10^3$bp | 1/1.00× $10^3$ bp | 56% |
| Nonsynonymous | No. | 8,452 |  | 0.0094 |
| SNPs (Venter) | Rate | 1/3.78× $10^3$bp |  | 26% |

3-STR RLPs; (*ii*) compared the genomes of Venter and Watson to the reference human genome sequence to estimate the 3-STR RLP rate in individuals and carefully documented the relative rate of RLPs in short and long repeats and both shared and different alleles in the human genomes and the chimp genome; and (*iii*) analyzed the length distribution of 3-STRs in the human genome with respect to their composition and uncovered a surprising periodic pattern possibly related to observed unusual DNA structures or other secondary structural properties (6). We also discuss methods for assaying the full set of such repeats in individuals to document personal RLPs that will lead to cost-effective approaches for genotyping using RLPs.

**The 3-STR RLPs in Venter and Watson Genomes.** We documented all 3-STRs in the gene regions (exons/introns) in the reference human genome and compared them to the polymorphisms observed in the genomes of J. Craig Venter (16) and James Watson (17). The full catalog of these 3-STRs and the Venter/Watson RLPs can be found on http://genomics10.bu.edu/RLPs. We counted 3,831,495 loci of these STRs in the transcribed regions of the human genome (Table 1). (Definition and identification methodology for 3-STRs are given in *Methods*.)

We identified 201,850 loci in exons and 3,629,645 in introns. The total length of these repeats in human transcripts—the size of the human 3-STR repeatome—is 29,257,783 bases: 1,576,309 in exons and 27,681,474 in introns. Based on the total length of all exons and introns (see *Methods*), this corresponds to ≈3.2% of the total amount of transcribed DNA in the human genome. The vast majority of these repeats are short, having an average length of 7.64 bases or 2.55 triplets: 7.81 bases in exons, 7.63 bases in introns. That the average length within exons is so similar to that of introns may seem surprising; however, these averages are heavily influenced by the abundance of short repeats that are likely to frequently occur by chance (18, 19).

We focused our analysis on length variation, i.e., insertions and deletions, rather than on point mutations within the repeats. By comparing the Venter genome to the reference (for the Watson genome, we have computed variations in only exon sequences), we counted 4,361 3-STR length differences (RLPs) between the two: 93 were in exons, 4,268 were in introns. This represents one such polymorphism per 3.43 × $10^5$ bases in the exons and one per 2.11 × $10^5$ bases in the introns. The level of repeat variation is lower than the level of SNP variation and similar to the frequency of small insertions and deletions in genomic DNA. These numbers, dominated by abundant short repeats, suggest that the mutation rate for expansions and contractions in short repeats is much greater than the overall rate of insertions and deletions. This observation is corroborated by the observed frequency of short repeats in the genome, which can be explained by a random sequence model with independent occurrence of nucleotides at frequencies matching their genome-wide prevalence.

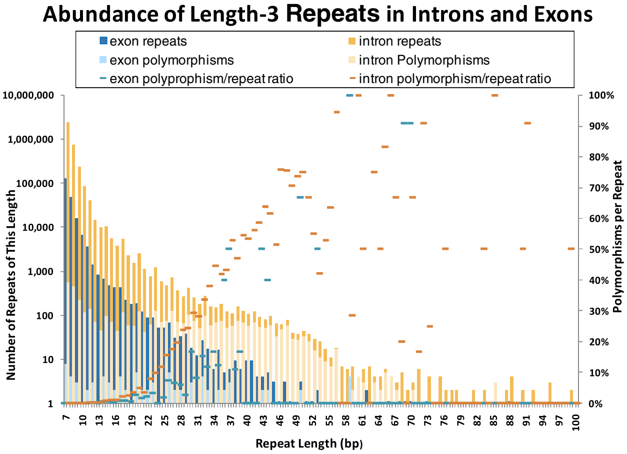

**Fig. 1.** Histogram of the frequency distribution of lengths of 3-STRs in the human genome. Superimposed on this histogram are data points showing the fraction of length-3 STRs that are length polymorphic for each length of STR. The absolute number of such polymorphisms for each repeat length is shown in the histogram as the lighter-colored bottom portion of each bar on the graph. For readability we have excluded from this figure the 17 intronic repeats that range in length from 101 bases to 362 bases.

Unlike shorter repeats, long repeats are highly polymorphic. The majority of repeats with length exceeding 40 bp appear polymorphic even in our small sample of individual genomes. Fig. 2 shows the fraction of 3-STRs that are length-polymorphic for each length 3-STR in the reference human genome. The absolute number of such polymorphisms for each repeat length is also shown.

The rate of occurrence of this type of polymorphism in exons is 62% that of introns. For comparison, we considered the analogous ratio for SNPs. Although most are synonymous (see Table 1), the SNP rate we observe in exons is ≈56% of that in introns, comparable to the 62% exon/intron rate that we observe for 3-STR RLPs.

As shown in Fig. 1, the fraction of polymorphic longer repeats for each length does not differ very much between introns and exons. In the shorter repeats, this difference is pronounced with a smaller proportion polymorphic in exons than in introns.

As shown in Fig. 2, based on our analysis, of the 201,850 3-STRs in human reference-sequence exons, the Watson genome contains RLPs at 60 genomic loci and the Venter genome contains such polymorphisms at 93 loci. These represent one polymorphism per 530 kb and one per 340 kb, respectively. This difference is not surprising given the small number of such polymorphisms we have observed in human exons.

**All Observed Short RLPs Appear Biallelic.** Only two allelic states occur in the combined set of both individual genomes and the reference. This is consistent with low variation and a low mutation rate for expansions and contractions of short repeats. Short repeats are similar in this respect to SNPs. The overwhelming majority of human SNPs are also biallelic (20). In contrast, 2 of 10 repeats with the length exceeding 25 were found multiallelic even in our small sample of two individual genomes and the reference. Theoretically, for events with the product of mutation rate and effective population size much less than one, the chance of observing a multiallelic event is very low. It is, therefore, possible that for expansions and contractions of long repeats, the mutation rate would not be substantially $<10^{-5}$ [approximately $(1/4)N_e$ for the human population]. This rate is much higher than the $2 \times 10^{-8}$ estimated for nucleotide substitutions.

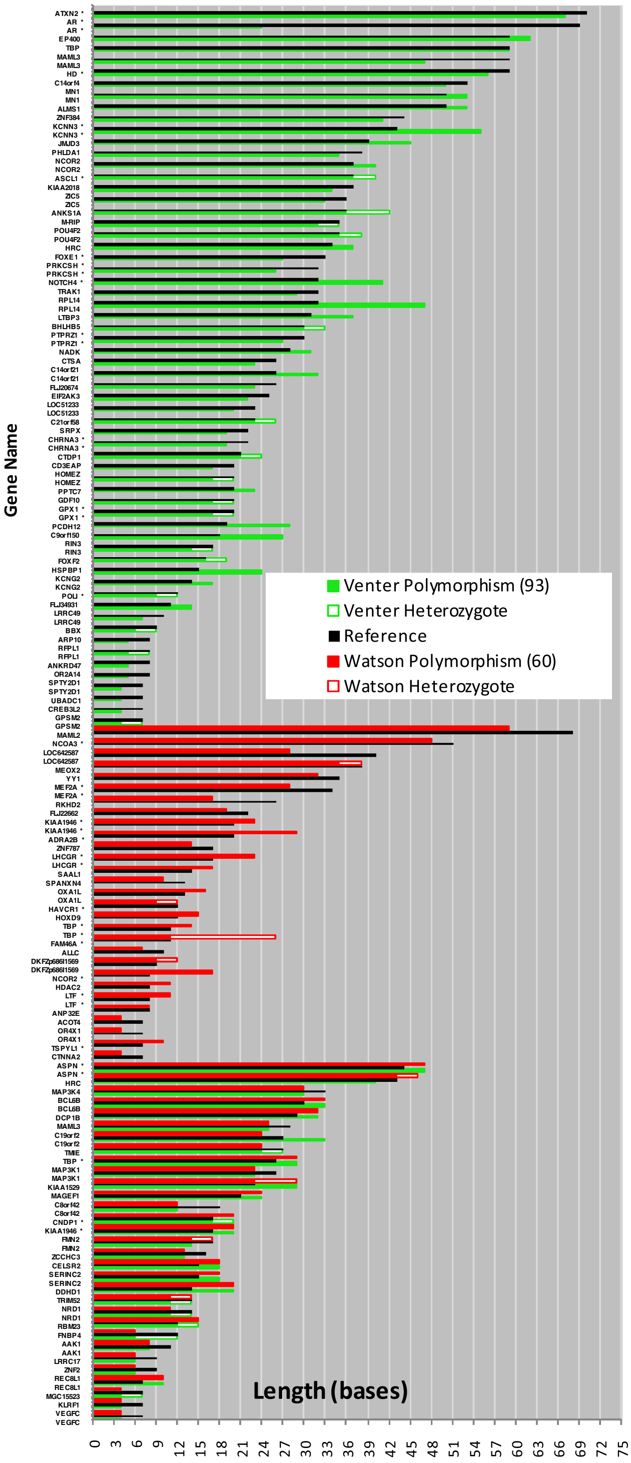

**Fig. 2.** Chart showing the length and containing gene of all exonic RLPs present in either the Watson or Venter genome. There are 60 exonic RLPs in the Watson genome, 93 in the Venter genome and 31 in both. In the case where the individual's genome is itself polymorphic with respect to a repeat length, this heterozygosity is shown as a white box indicating the extent of the difference. The 27 repeats whose labels are followed by asterisks overlap regions specifically associated with diseases in the Genetic Association Database (43).

**Length Bias in Human Intron Triplet Repeats.** Observing the lengths of 3-STRs in human transcripts, we were surprised by what seems to be a clear sawtooth pattern in the repeat lengths. Specifically, when the total length of a 3-STR is divided by 3, the remainder is more likely to be 2 than either 0 or 1. In other words, the nucleotide lengths of the 3-STRs are disproportionately = 2 (mod 3). When we plotted the number of repeats at each length, a zigzag pattern resembling the teeth of a saw emerged. As shown in Fig. 3, this sawtooth pattern is more pronounced in introns than exons.

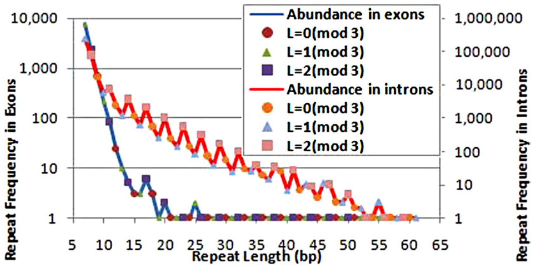

**Fig. 3.** Frequency of occurrence of 3-STRs of each length in 1 of the 10 repeat families. The red lines represent the abundance in introns and the blue lines represent the abundance in exons. More pronounced in the introns, the sawtooth pattern indicates a preference for repeats of length *L* where *L* = 2 (mod 3). All 10 are included in Fig. S2 and *SI Text*, *Sawtooth Pattern and Simulated Datasets*. For space, only the AAC repeat family is included here. AAC has the most prominent sawtooth pattern.

Following established methodology (21), we divided the 60 possible nonmonomer length-3 DNA sequences into 10 families. Each family consisted of 6 length-3 sequences that corresponded to a particular sequence in all 3 possible phases and their reverse complement. For example, the AGC family includes all genomic repeats composed of *AGC*, *GCA*, *CAG*, *GCT*, *TGC*, and *CTG*. The saw-tooth pattern that we have observed is more pronounced in some repeat families than in others. The AAC family of repeats, shown in Fig. 3, (AAC, GTT, ACA, TGT, CAA, and TTG), occurrences with lengths of 11, 14, 17, 20, 23, etc. are far more frequent than those with lengths 10, 13, 16, 19 and 22, respectively.

We considered the possibility that this pattern was related to the functional aspects of RNA transcripts (22). However, this is unlikely since we find the same pattern in intergenic DNA (as shown in Fig. S2 and *SI Text*, *Results in Simulated Datasets and Intergenic DNA*). We also investigated how the sawtooth pattern was related to short-range nucleotide correlations by training different fixed-length Markov models on the actual sequence, then generating random sequence using the model, and computing the lengths of 3-STRs in the generated sequence. This analysis (also detailed in Fig. S2 and *SI Text*, *Use of Simulated Datasets*) shows that the sawtooth pattern emerges as the length of the Markov model increases.

One observation of possible interest is that the amount of overrepresentation for these $3n + 2$ type lengths (as quantified by the number of values of $n$ where there are more $3n + 2$ length repeats than $3n + 1$ length repeats) corresponds inversely with the number of unusual DNA structures the specific sequence can exhibit. This is described in Table 2. Four types of unusual repeat-associated DNA structures have been described (6): (*i*) imperfect hairpins, (*ii*) G-quartets, (*iii*) slip-stranded DNA, and (*iv*) triplexes. These structures are associated with certain repeat families. Table 2 shows, for each repeat family, how many of these four types of structures it has been experimentally determined to exhibit and the preference for $3n + 2$ type lengths as described above. It appears that repeat families with lower totals tend to show more ($3n + 2$)-length bias. In particular, the AAC family of repeats, as shown in Fig. 3*A*, shows by far the most striking preference for these lengths. It is also the only repeat family not shown to take on any of the 4 unusual DNA structures. Although this relationship may prove tenuous, it may also provide a hint toward future study. A detailed analysis of the thermodynamic stability of the specific structures predicted and their likelihood of appearance may also shed more light on this pattern.

**Table 2. Unusual DNA structures and the sawtooth pattern**

| Triplet Sequence Family | Has the triplet sequence family been experimentally shown to exhibit the indicated structure? | | | | Total number of structures exhibited | Prominence of the sawtooth pattern |
|---|---|---|---|---|---|---|
| | Imperfect hairpins | G-quartets | Slip-Stranded DNA | DNA Triplexes | | |
| AAC | NO | NO | NO | NO | 0 | 12 |
| AAT | NO | NO | YES | NO | 1 | 11 |
| AAG | NO | NO | NO | YES | 1 | 10 |
| ACC | NO | YES | NO | NO | 1 | 9 |
| ACT | NO | NO | YES | NO | 1 | 9 |
| ACG | NO | NO | YES | NO | 1 | 7 |
| AGG | NO | YES | NO | NO | 1 | 6 |
| ATC | NO | NO | YES | NO | 1 | 6 |
| AGC | YES | NO | YES | NO | 2 | 6 |
| CCG | YES | YES | YES | NO | 3 | 5 |

This table shows, for each repeat family, the prominence of the sawtooth pattern, as quantified by the number of values of *n* where there are more $3n+1$ length repeats, for values of n from 2 through 19 (corresponding to repeat lengths from 8 through 59). Also shown are the number of known tandem-repeat-associated DNA secondary structures that the specific sequence has been experimentally determined to exhibit. There are 4 types of these DNA structures and each of these can only be made by certain repeat families (6). It appears that, in general, repeat families that can take on more of these structures tend to have less prominent sawtooth patterns. That is, they have a lower preference for lengths of the $3n + 2$ type.

**Comparison to Chimpanzee Variation.** Table 3 presents a comparison of 3-STR lengths in exons among Venter, Watson, and the chimpanzee genomes (23) at locations where both Venter and Watson exhibit variation from the reference. For short repeats, the chimpanzee sequence is always in agreement with one of the human alleles. This is another way in which short repeats behave similarly to human SNPs. This is also consistent with the hypothesis that the overall mutation rate for expansions and contractions of short repeats is low. However, for long RLPs, the chimpanzee sequence, in most cases, disagrees with all of the human alleles. Along with the multiallelic nature of these RLPs, this further suggests a high mutation rate among all long repeats, not just a subset of them.

To ascertain whether the repeat variation in the human exon sequences is under negative, neutral or positive selection (assuming RLPs in introns are evolving neutrally), we performed the generalized McDonald–Kreitman Test (24) to determine the human polymorphism rate versus the divergence rate from the chimpanzee genome. To apply the test to this context (as detailed in Fig. S1, Table S1, and *SI Text*, *McDonald-Kreitman Test*) we compared exonic RLPs to intronic RLPs. Our results indicate that the RLP rate in human exons is neutral since the difference in polymorphism and divergence rates were not significant (see *Methods*). This does not imply, however, that there is no selection pressure against specific 3-STR-length alleles. To measure this, more individual genomes would be required.

**Genes with RLPs are Highly Enriched in Transcriptional Regulators and Developmental Genes.** The set of 122 genes with exonic RLPs in the Venter/Watson genome (see Fig. 2) includes a number of well-documented disease genes including HD (huntingtin), AR (androgen receptor), ADRA2B, ALMS1 (Alström syndrome 1), ATXN2, TBP, and others. The full list of genes with their National Center for Biotechnology Information links is provided on http://genomics10.bu.edu/RLP and in Table S2 and *SI Text*, *Genes with Exonic Repeat-Length Polymorphisms (RLPs) in the Venter/Watson Genome*.

Many RLPs have been associated with neurological diseases (25), and studies have also identified RLPs in a number of transcriptional regulators and nuclear receptors such as HOXD9, HOMEZ, FOXF2, FOXE1, NCOR2, NCOR3, NCOA3, BBX, POU4F2, and MEF2A. Many of these transcriptional regulators are expressed in fetal brain tissue, leading to the hypothesis that these transcriptional factors and coactivators are involved in neural development (25).

We also measured GO enrichment (enrichment with respect to the Gene Ontology database) (26) of the set of exons containing RLPs among the full set of exons containing 3-STRs. The functional picture indicated by this analysis is consistent with the observation that these RLPs are associated with the regulation of transcription. The top 6 enriched GO terms all pertain to DNA transcription or transcription regulation: these are "transcription regulator activity," "transcription, DNA-dependent," "RNA biosynthetic process," "transcription," "regulation of gene expression," and "transcription from RNA polymerase II promoter" with *P* values of $5.9 \times 10^{-5}$, $3.9 \times 10^{-4}$, $4.0 \times 10^{-4}$, $5.0 \times 10^{-4}$, $8.1 \times 10^{-4}$, and $1.5 \times 10^{-3}$, respectively. Terms associated with development are also significantly enriched: "organ development" and "system development" are enriched with *P* values of $1.9 \times 10^{-3}$ and $3.5 \times 10^{-3}$, respectively.

Importantly, we do not assert a direct causal relationship between transcription factors and these RLPs. On the contrary, long 3-STRs are known to be overrepresented among transcription factors (27) and we show in this study that long 3-STRs are highly polymorphic. The overrepresentation of transcription factors among RLPs is probably secondary to these connections. We do claim, however, that, although long 3-STRs may be the causal mechanism by which these polymorphisms arise, the documented relationship between transcription factors, long repeats, and RLPs is of potential importance.

We also measured enrichment of other gene sets; notably, the enrichment of gene sets based on ESTs (28). We found strong enrichment for genes expressed in normal testis and colon. (See Table S3 and *SI Text*, *Complete List of GO-Terms and Other Gene-Set*.)

It is likely that the majority of the exonic 31 RLPs shared between the Watson and Venter genomes are common variations that will appear in a large subset of the appropriate population; however, a good number of RLPs we observed appear to be specific to their private repeatomes. This, we believe, is of both evolutionary and medical significance as follows: We identified a number of RLPs in relatively unexplored genes, such as C14ORF4 that also appears to be expressed and regulated in the brain (29). Other neural genes in the list include several olfactory receptors OR2A14 and OR4X1.

**Toward a Biological Assay of Genic RLPs: 2-Array Capture.** To assist future studies assaying RLPs, we compiled the complete set of 3-STRs in the human transcriptome. We then catalogued their flanking regions in the 3′ and 5′ directions. Each such locus is described by a sequence of the form $PR^kS$ where *P* and *S* are the prefix and suffix flanking regions of the repeat, respectively, and *R* is the 3-nucleotide repeat, *k* is the observed length in the reference genome. This information can be found at http://genomics10.bu.edu/RLPs.

To inexpensively assay the human repeatome, we have devised a process that relies on a two-step chip-based capture of the flanking

**Table 3. 3-STR Length variations among Venter, Watson, and Chimpanzee**

| | | Variation | | |
|---|---|---|---|---|
| Gene name | Reference 3-STR length | Venter | Watson | Chimpanzee |
| ASPN | 44 | +3 | +3 | −6* |
| HRC | 43 | −3 | +0/+3 | X |
| MAP3K4 | 33 | −3 | −3 | −12* |
| BCL6B | 30 | +3 | +0/+3 | +42* |
| DCP1B | 29 | +3 | +3 | +12* |
| MAML3 | 28 | −3 | −3 | −3 |
| TMIE | 27 | +0/−3 | −3 | −6* |
| C19orf2 | 27 | +6 | −3 | +0 (snp) |
| MAP3K1 | 26 | −3 | −3 | −9* |
| TBP | 26 | +3 | +3 | −3* |
| KIAA1529 | 23 | +6 | +0/+6 | +3*(snp) |
| MAGEF1 | 21 | +3 | +0/+3 | +9* |
| C8orf42 | 18 | −6 | −6 | +0 |
| FMN2 | 17 | −3 | +0/−3 | −3 |
| KIAA1946 | 17 | +3 | +3 | +0 |
| CNDP1 | 17 | +0/+3 | +0/+3 | +3 |
| ZCCHC3 | 15 | −3 | −3 | +6* |
| SERINC2 | 15 | +3 | +3 | +3 |
| CELSR2 | 15 | +3 | +3 | X |
| NRD1 | 14 | +0/−3 | −3 | +0 |
| TRIM52 | 14 | +0/−3 | +0/−3 | −3 |
| DDHD1 | 14 | +6 | +6 | +12* |
| FNBP4 | 12 | +0/−6 | −6 | +0 |
| RBM23 | 12 | +0/+3 | +3 | +0 |
| AAK1 | 11 | −3 | −3 | +0 |
| ZNF2 | 9 | −3 | −3 | −3 |
| LRRC17 | 9 | −3 | −3 | +0 |
| VEGFC | 7 | −3 | −3 | −3 |
| KLRF1 | 7 | −3 | −3 | −3 |
| REC8L1 | 7 | +3 | +3 | +3 |
| MGC15523 | 7 | +0/−3 | −3 | +0 |

Comparison of 3-STR length variations among Venter, Watson and chimpanzee genomes at exon locations where both Venter and Watson exhibit vartiation from the refenence genome. All values are in bases. Variations represent difference isn length from the reference value. Variation entries with two values separated by a slash represent heterozygous 3-STR lengths. An "x" in the chimpanzee column indicates that no match for that 3-STR was found in the chimpanzee genome. Entries with "(snp)" in the chimpanzee column have a single-base substitution difference from the reference repeat pattern in the chimpanzee pattern. The variations specific to the chimpanzee are marked with an "*".

sequences followed by a final sequencing step. We used our catalog of flanking sequences to produce designs two Nimblegen-style HD2 microarrays for this purpose. The first step in our dual-array process is to cleave the genome into ≈1,000-bp long fragments to hybridize to the microarray probes. These fragments can easily be sequenced using state-of-the-art next-generation sequencing technologies. A significant cost reduction is obtained by capturing the specific repeat loci as compared with currently available technologies for full exon sequencing.

As in standard chip-based sequence capture (30), we designed oligonucleotide probes to capture all fragments containing the flanking regions for subsequent sequencing. Instead of using a single chip, however, we use two to increase the proportion of sequences that include the entire repeat plus both flanking sequences. In brief, one chip contains all of the repeat prefix-flanking regions, *P*, and the other chip contains all of the repeat suffix-flanking regions, *S*. Following accepted protocols for exon capture arrays (30), we fragment the DNA and capture the sequences with the first array. However, instead of sequencing this set of fragments, we can run it on the second array: a second capture step. Thus, each of the sequences captured by this chip should have both a prefix and a suffix flank for one of the repeat regions. Sequencing this set will, therefore, provide the sequence of fragments fully spanning the repeats and their flanking regions, and little else. We have designed chips of this type and the designs can also be found at http://genomics10.bu.edu/chip_designs.

We hypothesize that there are only a small number of common 3-STR RLPs. If these STRs are identified, a low-throughput technique like mass spectrometry (26, 31) could be an inexpensive strategy for genotyping. However, a more comprehensive approach, as described above, is needed for de novo identification of novel RLPs.

## Discussion

We have conducted an initial analysis of the personal repeatome of two individuals. We focused on 3-nucleotide repeat variation in exonic regions. While we expected a strong bias against variation in these repeats, the fact that, out of 201,850 exonic 3-STRs, we found only 93 RLPs in the Venter exonic regions was still surprising. Even more unexpected is the relative ratio of polymorphisms in introns vs. exons. We expected a much higher ratio of polymorphisms to total DNA in introns, but in our study this ratio in exons is only 62% of the observed ratio in introns, which reinforces the apparent strong selection against length expansion or contractions in intronic regions (32), although other functional or evolutionary reasons may also play a role.

Selective pressure on specific alleles, however, seems surprisingly limited. Based on the relative polymorphism rates as measured by a variant of the McDonald–Kreitman test (24), there is no significant selective pressure between alleles in the context of RLPs. This is the case even when short and long repeat regions are analyzed separately.

However, in their amount of observed variation, long and short repeat regions seem to be following very different sets of rules. In terms of mutation rate, the number of observed alleles and divergence from the chimpanzee genome, short repeats behave like typical insertions and deletions. It seems plausible that the mechanisms behind short-repeat expansion and contraction are not dissimilar to those at work in the rest of the genome. Not so for longer repeats. Based on the number of multiallelic RLPs and the much greater divergence from the chimpanzee genome among long repeats, their mutation rate may be as high as $10^{-5}$. This difference could be an important clue to the role and behavior of repeats in the human genome.

We also discovered an unexpected pattern of length variation in intronic repeats. We confirmed this pattern in intergenic regions as well. We have suggested that this length bias may be related to structural constraints on DNA. However, many other explanations are possible and are currently being investigated. This pattern may shed light on the process of 3-STR expansion or contraction. We will consider this more thoroughly in future work.

Collecting RLP information from a much larger set of individuals will foster exploration of correlations among these polymorphisms. While it might be premature to speculate, there is a reasonable expectation of physiological consequences from interactions between RLPs. For example, Huntington's disease involves the formation of huntingtin aggregates. These aggregates have been shown to interact with heat shock proteins such as HSC70 (33, 34). Overexpression of HSC70 has been shown to reduce aggregation and decrease cell death. HSPBP1 (heat shock 70 kDa) is one of the RLP genes and has been shown to have an altered cellular distribution in the presence of aggregates. Thus, interaction between huntingtin and heat shock 70kDA may have a role in disease onset or progression.

It has been suggested that current databases under-represent the true number of STRs and their true length due to technological limitations of previous sequencing technologies. The availability of NextGen technologies, such as Roche 454 GS-FLX platform, that are not limited by bacterial clone libraries, may yield an increase in our dictionary of repeats in future studies.

This work revealed several preliminary but nevertheless thought provoking insights gained from mining the personal human repeatome in three genomes. However, much more remains to be done in this important area. Three percent of the sequenced human genome consists of tandem repeats (35). In genic regions, such repeats are disproportionately found in the 3′-UTR region of human genes (36). However, the personal repeatome analysis in these regions is likely to be technically difficult due to the problem of computing accurate alignments of these regions (35). Based on our initial findings and through increasingly efficient repeat-assay technologies such as the one proposed in this article, we anticipate major advances in this area in the near future.

## Methods

We used the National Center for Biotechnology Information human reference genome assembly build 36.2 to identify 3-STRs. We began by finding maximal, perfect 3-periodic regions of 7 or more bases, i.e., each base in the region matched the base 3 positions preceding it, and extended it as far as possible. Regions consisting of a single repeated base (monomer repeats) were excluded. Perfect repeats that were separated by a single-base variation from the 3-periodic pattern (substitution, insertion, or deletion), were joined together and counted as a single 3-STR. For example, the following shows, in capital letters, a 3-STR consisting of a perfect 8-bp region, a single-base insertion, and a perfect 7-bp region: . . .cagt**GTAGTAGTTAGTAGTA**agtc. . . .

The gene annotation used was all RefSeq genes downloaded from the UCSC Table Browser. The transcripts correspond to full-length premRNAs. For each gene, only the coding portion of the transcript was selected. To account for alternative transcription, regions contained in the exon of any isoform were classified as exons. All other regions were classified as introns. In addition to the coding transcripts, 100 bp of flanking sequence were also extracted to provide genomic context for each repeat, but these flanking regions were excluded from the search for STRs. We specifically chose to include all 3-STRs, without filtering based on a statistical approach (37).

Venter variations with respect to the reference genome were taken from files HuRef.InternalHuRef-NCBI.gff and HuRef.homozygous_indels.inversion.gff, downloaded from the J. Craig Venter Institute (ftp.jcvi.org/pub/data/huref) (16). Chimpanzee variations from the human reference were computed from the UCSC netted chained alignments at ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/vsPanTro2/axtNet, skipping files whose names contained the string ''random.''

To compute Watson exon variations, we downloaded all sequences for the James Watson sequencing project (17) from the National Center for Biotechnology Information Trace Archives (www.ncbi.nlm.nih.gov/Traces/traces.cgi, center_name = ''cshl'' and center_project = ''project jim''). We created a set of STR sequences by adding to each exon STR the 60 bp immediately preceding and following it in the reference genome. We used the sequence comparison program NUCmer (38) to identify traces with any type of match to an STR. A modified version of the Celera Assembler overlap program (39), which required a high-fidelity match in both flanking regions but allowed liberal indels in the repeat region, was then used to identify traces that matched each STR. Traces with matches to more than one position in the reference genome were excluded as being inconclusive. We required at least two traces to agree on the sequence variation to include it in our analysis.

We classified a reference 3-STR as having variation if the alternate sequence variation, in Venter, Watson or chimpanzee, occurred completely within the reference STR region and yielded a length difference that was a multiple of 3 bp. In the case of insertions, we also required that the inserted sequence extended the 3-STR pattern.

To gauge the functional contribution of these repeats, we measured GO enrichment (enrichment with respect to the Gene Ontology database) (26). We performed this analysis using the GORILLA (40) and DAVID (28, 41) web interfaces.

We measured the selective pressure on RLPs using a variant of the McDonald–Kreitman test (24). Typically, this test measures selective pressure by comparing the silent mutation rate ($K_s$) to the nonsilent mutation rate ($K_a$). We, instead, compared the rate of RLPs in the introns to the rate of RLPs in the exons. To estimate this rate, we measured the ratio of intronic RLPs to exonic RLPs among the chimpanzee variations from the human reference sequence. We then compared this ratio to the ratio of intronic RLPs to exonic RLPs among the Venter variations from the human reference using Fisher's Exact Test (42). No significant difference between these ratios was found. We repeated this test comparing long RLPs and short RLPs independently, with thresholds of 50, 55, and 60 bases (see Fig. S1, Table S1, and *SI Text*, *Methods for Generating Simulated Datasets* for further details of this analysis), with the same result.

**ACKNOWLEDGMENTS.** We thank Gary Benson, Alan Herbert, and Maxim Frank-Kamenetskii for informative exchanges on this topic; Laurie Goodman for her extensive editorial contribution; and the reviewers for their thoughtful and constructive comments. This research was supported by National Institutes of Health Grants R33 HG002850-02 and U54LM008748-02 and National Science Foundation Grants 0239435 and DMS-0602204.

1. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science* 322:881–888.
2. Andresen JM, et al. (2007) The relationship between CAG repeat length and age of onset differs for Huntington's disease patients with juvenile onset or adult onset. *Ann Hum Genet* 71(Pt 3):295–301.
3. Stranger BE, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315:848–853.
4. McCarroll SA (2008) Extending genome-wide association studies to copy-number variation. *Hum Mol Genet* 17(R2):R135–142.
5. Lee S, Kasif S, Weng Z, Cantor CR (2008) Quantitative analysis of single nucleotide polymorphisms within copy number variation. *PLoS ONE* 3:e3906.
6. Mirkin SM (2006) DNA structures, repeat expansions and human hereditary disorders. *Curr Opin Struct Biol* 16:351–358.
7. Mirkin SM (2007) Expandable DNA repeats and human disease. *Nature* 447:932–940.
8. Usdin K (2008) The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. *Genome Res* 18:1011–1019.
9. Haberman YAN, Rechavi G, Eisenberg E (2007) Trinucleotide repeats are prevalent among cancer-related genes. *Trends Genet* 24:14–18.
10. Penney JB, Jr, Vonsattel JP, MacDonald ME, Gusella JF, Myers RH (1997) CAG repeat number governs the development rate of pathology in Huntington's disease. *Ann Neurol* 41:689–692.
11. Sobczak K, Krzyzosiak WJ (2005) CAG repeats containing CAA interruptions form branched hairpin structures in spinocerebellar ataxia type 2 transcripts. *J Biol Chem* 280:3898–3910.
12. Ross CA (2002) Polyglutamine pathogenesis: Emergence of unifying mechanisms for Huntington's disease and related disorders. *Neuron* 35:819–822.
13. Li JL, et al. (2003) A genome scan for modifiers of age at onset in Huntington disease: The HD MAPS study. *Am J Hum Genet* 73:682–687.
14. Sharifi N, Figg WD (2007) Androgen receptor modulation: Lessons learned from beyond the prostate. *Cancer Biol Ther* 6:1358–1359.
15. Fondon JW, III, Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci USA* 101:18058–18063.
16. Levy S, et al. (2007) The diploid genome sequence of an individual human. *PLoS Biol* 5:e254.
17. Wheeler DA, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872–876.
18. Benson G, Waterman MS (1994) A method for fast database search for all k-nucleotide repeats. *Nucleic Acids Res* 22:4828–4836.
19. Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res* 27:573–580.
20. Ravi Sachidanandam, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933.
21. Bacolla A, et al. (2008) Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. *Genome Res* 18:1545–1553.
22. Jasinska A, et al. (2003) Structures of trinucleotide repeats in human transcripts and their functional implications. *Nucleic Acids Res* 31:5463–5468.
23. The Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.
24. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351:652–654.
25. Margolis RL, et al. (1997) cDNAs with long CAG trinucleotide repeats from human brain. *Hum Genet* 100:114–122.
26. Ashburner M, et al. (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25–29.
27. Nakamura Y, Koyama K, Matsushima M (1998) VNTR (variable number of tandem repeat) sequences as transcriptional, translational, or functional regulators. *J Hum Gen* 43:149–152.
28. Dennis G Jr SB, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Gen Biol* 4:3.
29. Heger S, et al. (2007) Enhanced at puberty 1 (EAP1) is a new transcriptional regulator of the female neuroendocrine reproductive axis. *J Clin Invest* 117:2145–2154.
30. Albert TJ, et al. (2007) Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 4:903–905.
31. Tang K, et al. (1999) Chip-based genotyping by mass spectrometry. *Proc Natl Acad Sci USA* 96:10016–10020.
32. Ogurtsov AY, Sunyaev S, Kondrashov AS (2004) Indel-based evolutionary distance and mouse-human divergence. *Genome Res* 14:1610–1616.
33. Jana NR, Tanaka M, Wang G, Nukina N (2000) Polyglutamine length-dependent interaction of Hsp40 and Hsp70 family chaperones with truncated N-terminal huntingtin: Their role in suppression of aggregation and cellular toxicity. *Hum Mol Genet* 9:2009–2018.
34. Swayne LA, Braun JE (2007) Aggregate-centered redistribution of proteins by mutant huntingtin. *Biochem Biophys Res Commun* 354:39–44.
35. Lander ES, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
36. Andken BB, et al. (2007) 3′-UTR SIRF: A database for identifying clusters of short interspersed repeats in 3′ untranslated regions. *BMC Bioinformatics* 8:274.
37. Gelfand Y, Rodriguez A, Benson G (2007) TRDB–the Tandem Repeats Database. *Nucleic Acids Res* 35:D80–D87.
38. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.
39. Miller JR, et al. (2008) Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* 24:2818–2824.
40. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48.
41. Huang DW SB, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocols* 4:44–57.
42. Agresti A (1992) A survey of exact inference for contegency tables. *Statitical Science* 7:131–153.
43. Becker KG BK, Bright TJ, Wang SA (2004) The Genetic Association Database. *Nat Genet* 36:431–432.