Finding and extending ancient simple sequence repeat-derived regions in the human genome

Jonathan A. Shortt[1], Robert P. Ruggiero[2], Corey Cox[1], Aaron C. Wacholder[3], and David D. Pollock[4, *]

[1]Colorado Center for Personalized Medicine, University of Colorado School of Medicine, Aurora, CO 80045, USA

[2]Southeast Missouri State University, Department of Biology, Cape Girardeau, MO 63701, USA

[3]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, United States.

[4]Department of Biochemistry & Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA

*Author for Correspondence: David D. Pollock, Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO, 80045 USA, 303-724-3234, 303-724-3215, David.Pollock@ucdenver.edu

Abstract

Background

Previously, 3% of the human genome has been annotated as simple sequence repeats (SSRs), similar to the proportion annotated as protein coding. The origin of much of the genome is not well annotated, however, and some of the unidentified regions are likely to be ancient SSR-derived regions not identified by current methods. The identification of these regions is complicated because SSRs appear to evolve through complex cycles of expansion and contraction, often interrupted by mutations that alter both the repeated motif and mutation rate. We applied an empirical, kmer-based, approach to identify genome regions that are likely derived from SSRs.

Results

The sequences flanking annotated SSRs are enriched for similar sequences and for SSRs with similar motifs, suggesting that the evolutionary remains of SSR activity abound in regions near obvious SSRs. Using our previously described P-clouds approach, we identified 'SSR-clouds', groups of similar kmers (or 'oligos') that are enriched near a training set of unbroken SSR loci, and then used the SSR-clouds to detect likely SSR-derived regions throughout the genome.

Conclusions

Our analysis indicates that the amount of likely SSR-derived sequence in the human genome is 6.77%, over twice as much as previous estimates, including millions of newly identified ancient SSR-derived loci. SSR-clouds identified poly-A sequences adjacent to transposable element

termini in over 74% of the oldest class of *Alu* (roughly, *AluJ*), validating the sensitivity of the approach. Poly-A's annotated by SSR-clouds also had a length distribution that was more consistent with their poly-A origins, with mean about 35 bp even in older *Alus*. This work demonstrate that the high sensitivity provided by SSR-Clouds improves the detection of SSR-derived regions and will enable deeper analysis of how decaying repeats contribute to genome structure.

Keywords

SSR, genome structure, repeats, microsatellites, tandem repeats, genome evolution

Background

Simple sequence repeats (SSRs) are 1-6 bp tandem repeats that have been estimated to comprise 3% of the human genome (1, 2). SSRs are notable for their unusual mutation process; after they reach a threshold length (3-5 tandem motif repeats), the rate of slippage during DNA replication dramatically increases, resulting in rapid expansion or contraction of SSR loci. These events may occur at a rate of 1×10^{-3} per locus per generation (3, 4), many orders of magnitude faster than point mutation rates, and can modify structural and regulatory functions, contributing to disease (5). In addition, because they are enriched in promoters, highly mutable, and provide a rich source of heritable variation, SSRs were proposed to be evolutionary "tuning knobs" (6-10). Numerous recent studies have highlighted the potential functional role of SSRs in gene regulation (11-14) and a better understanding of SSR evolution may therefore allow insights into how function can arise from constantly changing genomic structure.

A proposed life cycle for SSRs includes intertwined stages of birth, adulthood, and death (15-18). *De novo* birth of an SSR at a location occurs when a short series of repeats arises by chance mutations, and aided and extended by the tendency of duplications to occur via normal (non-SSR) slippage events that result in tandem duplication of short motifs (15, 18). If the number of simple sequence repeats exceeds some threshold length, which can depend on the composition and purity of the repeated motif (19), then the probability of slippage will increase with a slight bias towards increasing numbers of repeats (4, 20-22). Additionally, although there is a clear lower bound on repeat lengths (zero, obviously) and the slippage rates for small numbers of repeats is low, there is no upper bound on repeat lengths unless it is biologically imposed. These factors together are thought to result in rapid expansion in the number of motifs

at SSR loci and suggests that accurately describing the length and distribution of SSRs may provide a new source of insights into genome biology.

It is thought that during SSR "adulthood", slippage-induced expansions and contractions (usually one repeat at a time) can rapidly alter the length of SSR loci, but mutations that disrupt the composition of tandem repeats also accumulate and slow or stop the slippage process (23, 24). The SSR life cycle is potentially complicated by rare multiple-motif copy number mutations that are thought to be biased towards large deletions, and by selection against long repeat lengths that may lead to upper size limits (20, 21, 25). Transposable elements (TEs) also contribute to SSR generation by introducing pre-existing repeats at the time of TE replication, by introducing poly-A tails (in the case of retroelements), or by repeatedly introducing sequences that are likely to give birth to new SSRs (16, 26, 27).

SSR death presumably occurs after either sufficiently large deletions at a locus have occurred or after enough mutations have accumulated so that there are no longer uninterrupted tandem motif stretches above the threshold length (17). After the death of an SSR, remnants of the formerly active SSR locus may remain in the genome, sometimes spawning an active SSR locus (with the same or similar motif) capable of expansion by slippage; this phenomenon has been observed but not characterized in great depth (15).

The abundance of active SSRs in the genome and their finite lifetime suggest that dead SSRs may also be abundant, although their high slippage mutation rate and complex, motif-dependent evolution makes modeling their evolutionary outcomes difficult. The identification of dead SSRs remains important if for no other reason than because their presence in the genome

can confound the detection and annotation of other genomic elements (28). Several reports have noted that the sequence composition near SSRs is biased towards the adjacent SSR motif, and it has been proposed that such sequences are SSR-derived (29, 30); however, the origin of this biased sequence has not been explored in detail. Part of the problem is that Tandem Repeats Finder (TRF) (31), the current predominant method for finding genomic repeats, although mathematically elegant and computationally efficient, is designed to detect perfect and near-perfect repeats, and provides little information about more degenerate SSR-derived loci. The ability to better identify degraded SSRs at various ages and stages of their life cycle would thus aid in annotation of the genome and inform on the origins and history of regions in the genome where they reside.

Here, we report a new method to detect SSR-derived sequence using a probability-clouds (*P-clouds*) (32, 33) based approach. This approach uses empirical counts of oligonucleotides (oligos) to find clusters (or clouds) of highly enriched and related oligos that, as a group, occur more often than predicted by chance. The *P-clouds* method has been applied to identify various repetitive structures in the human genome (32, 33), including transposable elements, but has not yet been applied to identify SSRs (which were specifically excluded from the original method). The use of empirical oligo enrichment, coupled with alignment-free and library-free detection, makes *P-clouds* both fast and particularly well-suited to annotate regions resulting from the complex mutational processes associated with SSR loci. We obtained sets of p-clouds in regions flanking perfect live SSRs under the hypothesis that such regions will be enriched in the mutated detritus of the SSRs (34). These SSR p-clouds, called SSR-clouds, were then used to re-define the spans of active SSR regions and locate dead SSR loci that were not previously identified. We

also provide further evidence that SSRs frequently spawn new SSR loci with similar motifs, presumably because the low sequence degeneracy of SSR detritus regions makes them fertile spawning grounds.

Results

Characterization of perfect SSR loci in the human genome

Uninterrupted perfect SSR loci abound in the genome. SSR sequence motifs of 1-6 bp were grouped into motif families comprised of a motif, its reverse complement, and any possible alternate phase of the motif or its reverse complement (e.g., AAC, ACA, CAA, GTT, TGT, and TTG all belong to the same motif family) to create a total of 501 separate SSR motif families. If a longer motif was a repeated multiple of a shorter motif (e.g., ATAT versus AT), that motif was assigned to the shorter motif. The unmasked human genome (hg38) was annotated (Table S2) with these motif families to locate every *perfectly* repeated contiguous SSR locus (one that contains no point mutation, insertion, deletion, or motif phase shift; loci separated by 1 or more bp were assigned different loci in this analysis) at least 12 bp in length. A total of 4,551,080 perfect (uninterrupted) SSR annotations were found, covering 68.8 Mb (~2.2% of the genome). These perfect repeats constitute over three-quarters (77.8%) of the 88.4 Mb SSR sequence (2.85% of the human genome) annotated using standard TRF settings.

The 12 bp minimum length for SSR loci is consistent with reports that established an SSR expansion threshold cutoff at around 10bp for motifs ≤ 4 bp (15, 38, 39), and is consistent with our own analyses of when perfect SSR frequencies significantly exceed expectations based on genomic dinucleotide frequencies (see Figure S1). The most highly-represented SSR is the

mononucleotide repeat poly-A/poly-T (henceforth referred to as just poly-A) with 703,012 separate loci. Consistent with previous reports, many (467,092, or 66.44%) of these poly-A's overlap with an annotated *Alu* (40), and 536,938 (76.38%) overlap with any annotated transposable element. Some caution is warranted in interpreting this result, both because the poly-A tail and the A-rich region in the center of many *Alus* may or may not contain a perfect repeat, and because RepeatMasker is inconsistent about whether it includes a poly-A tail in a repeat annotation. Nevertheless, this result indicates the minimum extent to which transposable elements contribute to the frequency of poly-A loci in the genome. Other than poly-A, the next most represented motif is CA/TG with 170,729 separate annotations, only 3,206 (1.88%) of which are found in an *Alu* element. Although all possible SSR motifs families have at least one locus in the genome, the most common motif families tend to have much simpler motifs than the least common (64% of the 50 most common motifs contain only 1 or 2 nucleotides, and only three of the most common motifs contain all 4 nucleotides, while 82% of the least common motifs contain all four bases (see Table S2), suggesting more frequent rates of origination for these simpler motifs. There is also an enrichment of shorter motifs amongst the most common SSRs, a trend that is consistent with previous observations (4, 41).

Characterization of sequence bias in the regions flanking perfect SSRs

Sequence biases in the regions flanking SSRs are a rich resource for understanding the evolutionary remains of SSR activity. Perfect SSR loci are often closer to each other than expected by chance, with an extremely high peak under 10 bp separation, and leveling off before 100bp (Figure S2). Reasonable explanations for close repeats include that they were previously a

single locus that was divided by imperfections, or that new repeats were spawned from a single repeat's detritus. Indeed, the repeated motifs of adjacent SSR loci often share high sequence similarity. The most represented repeated motif near a perfect SSR locus is often the repeated reference motif itself, and other similar motifs are also highly over-represented (Figure 1). As an example of more complex families, we considered $(ATGC)_n$ loci, and adjacent SSRs that had 1, 2, or 3 different nucleotides. As with the simpler motifs in Figure 1, similar motifs are highly enriched at short distances from $(ATGC)_n$ repeats (Figure 2), while dissimilar motifs are far less enriched. These observations suggest that SSRs can originate from the periphery of existing SSR loci where sequence is already biased towards simple sequences (30). Under this hypothesis, dissimilar families that require multiple mutations to reach a threshold slippage length are found at lower frequencies because they are more difficult to seed.

To better describe the extent of the periphery around SSRs, which is known to deviate from random sequence (29, 30) and may represent a detritus field of mutated repeats (34), we measured similarity to each repeated perfect motif within 200 bp on either side of the repeat. There are differences depending on the size and repeat motif, but in general similarity extends at least 50-100 bp on either side of motifs (Figure 3). This size of detritus field is consistent with the idea that regular SSR seeding occurs from this detritus. As a side note, poly-A sequences had detritus fields on their 3' side, but not their 5' side, because they commonly originate from transposable elements (Figure S3) whose uniform sequence obscured the presence of detritus fields.

Construction and evaluation of SSR-clouds for detection of SSRs

To characterize and detect oligos in SSR detritus fields, we used the probability clouds (*P-clouds*) method (32, 33), which annotates empirically identified clusters (or clouds) of related oligos that are over-represented in a sequence. This approach has the potential to identify ancient repeats that have diverged considerably from their original sequence. By using increasingly relaxed threshold enrichment parameters, we built nested oligo clouds for each SSR motif family. There are relatively few highly enriched oligos with high similarity to the parent motif, and larger sets of more diverse but less-enriched oligos (Figure 4). High count, high similarity oligos are included in high stringency clouds, and low count, low similarity oligos are built into lower stringency clouds. We note here that although the largest motif families identified over 50,000 16-mer oligos in their low-stringency clouds, this represents only a very small fraction (0.0000116) of all possible 16-mer oligos. We conclude that finding *extended* regions in the genome made up of such oligos by chance alone is improbable. For example, if 50,000 oligos were distributed evenly across the genome, one might expect to find only about one oligo every 100,000 bp.

SSR-cloud loci were ranked according to the highest-stringency oligo contained in the locus, but annotations of high-stringency oligos can be extended using oligos contained in lower stringency clouds. The extension of locus annotations with lower-stringency oligo clouds has a striking impact on the length distributions of SSR loci (Figure 5). For example, poly-A SSR loci go from a highly skewed, almost exponential length distribution with a mean at 17.2 bp when only perfect repeats are considered, to something much closer to a normal distribution (although still right skewed) with a mean near 36 bp when extended using lower-stringency SSR-cloud sets

(Figure 5A). The latter distribution is consistent with the biology of poly-A origins through retrotransposition and reports that *Alu* transposition efficacy increases with poly-A tail length up to 50 bp (42, 43). Thus, the lower-stringency oligos enable detection of a region that is consistent with the *entire* ancient sequence derived from the poly-A tail at the time of insertion. However, it should be recognized that some of the detected length could be due to slippage in either direction post-insertion and prior to degradation. The length distributions of other SSR loci are similarly expanded, but with tails often extending to much larger regions (Figure 5B). Annotation and locus extension may occur infrequently by chance and can be accounted for with false discovery rates. Nevertheless, to ensure that the SSR locus length distributions we observe are not biased towards the loci used in cloud building, we tested the length distributions of the 10% of SSR loci that were not used in cloud building (see Methods). Figure S4 shows that the length distributions of these sets of loci do not substantially change, even at low cloud stringency.

SSR-clouds annotation of the human genome

The complete SSR-clouds annotation comprises 8,983,547 loci covering 221.6 Mb (7.15%) of the human genome. Of these loci, 46.92% intersect a transposable element, which includes poly-A regions annotated as part of the transposable element. A total of 3,085,675 of the loci, comprising 62 Mb (28.15% of all bases annotated by SSR-clouds) do not overlap with any previous repetitive element (including SSRs annotated by TRF), and thus represent novel repetitive sequence. Accounting for false discoveries (see Methods), we conclude that at least 6.77% of the genome is made up of SSRs or is SSR-derived.

The average false discovery rate is 5.31%, but the probability of being a false discovery varies widely among loci, depending on length. Most loci have a high positive predictive value (the inverse of the false discovery rate), but 3,423,735 loci covering 53.8 Mb (~25% of the SSR-clouds annotation) have a false discovery rate > 10% (maximum FDR= 0.175). The majority (3,020,997, or 88%) of these less certain SSR loci are either 16 bp or 17 bp in length, while the remainder are comprised of short perfect SSR loci under 13 bp in length. Although these loci have high false discovery rates because they are short, there are millions more of these loci than expected by chance based on dinucleotide frequencies. This abundance of short SSRs indicates that simple sequences of this length may often originate during evolution but die quickly through mutation accumulation before they have a chance to extend to create longer loci. It is also worth noting that regardless of their origin, these short loci are identical in sequence to areas that have potentiated SSR expansions and likely good spawning grounds for future SSRs.

Comparison of SSR-clouds detection to Tandem Repeats Finder

Although the purpose of this research was not to replace Tandem Repeats Finder (TRF), we nevertheless compared the SSR-cloud annotations with TRF annotations using the same parameters as in (2), which yielded the widely-quoted 3% SSR genomic estimation (2). Table 1 (see also Tables S3 and S4) highlights that SSR-clouds annotations of SSRs captures nearly all TRF SSR loci as well as millions of likely SSR-like loci that are not detected by TRF. The greatest increase in SSR-cloud loci occurs where the stringency of the SSR-cloud locus is low. These elements are likely missed by TRF because of their short length or divergence from a perfect SSR sequence. The discordance between the SSR-clouds and TRF annotation sets

highlights that previous estimations of SSRs in the genome are likely extremely conservative and frequently overlook SSR-derived regions of more ancient origin. This is conservative in the wrong direction for research questions that require eliminating as many SSR-derived regions as possible, for example if one is trying to identify low-copy regions of the genome or trying to discriminate sequences derived from specific types of TEs, which might themselves include SSRs.

Age characterization of SSR-derived sequences using *Alu* transposable elements

The approximate ages of poly-A SSR-derived sequences were determined by leveraging the relationship between *Alu* transposable elements and poly-A SSRs (15, 40, 44). *Alu* has over a million copies in the human genome, and their relative ages can be accurately determined (37). We divided *Alus* into three age groups approximately representing the main families of *Alu* and assessed how frequently poly-A loci detected by SSR-clouds of different stringencies could be found in the poly-A regions of *Alu* elements. While 63% of young poly-A tails tend to be annotated by uninterrupted poly-A clouds, older poly-A tails from the oldest group of *Alus* (42,125 loci, or ~50%) are unsurprisingly the most difficult to detect and are often annotated only by low stringency SSR-clouds (Figure 6). These results support the idea that lower-stringency SSR annotations are indeed derived from SSRs but are difficult to detect through other means because of their divergence from the original poly-A repeat.

About 25% of old loci were not detected by poly-A clouds of any stringency level, but an additional 11,821 annotations were found using SSR-clouds from any SSR family, not just poly-A. Thus, almost 90% of the oldest *Alus* (74,846 loci out of 84,346 total) had some sort of SSR-

derived locus in the expected poly-A region. It is possible that the 9,500 old *Alus* without detected SSR-clouds had their tails deleted or moved through genomic rearrangements over time or they degenerated to the point of being unidentifiable. The oldest group of *Alus* is 1.60 times older than the average age for all *Alus,* while the unannotated *Alus* are 1.64 times older (Welch two-sample t-test, $p < 2.2 \times 10^{-16}$), supporting the idea that loss of tails increases with age.

Discussion

SSR-clouds is a rapid, non-parametric method based on *P-clouds* for finding SSRs and SSR-derived regions in the genome. SSR-clouds finds numerous previously undiscovered SSR loci whose overlap with poly-A regions of known ancient transposable element loci provides compelling evidence that these loci are indeed SSRs or are SSR-derived. SSR-clouds analyses reveal that SSR-derived regions comprise a larger portion of the human genome than previously appreciated, increasing the SSR-derived percentage from about 3% to at least 6.77%. This increase is due to increased annotation length of previously annotated loci as well as newly annotated loci (Table 1). The output for SSR-clouds follows a standard bed file format (including the chromosome/scaffold and beginning and ending coordinates for a locus), with additional information about the SSR motif family present in the locus. As seen in Figure 7, different regions of a locus may be annotated by the clouds of multiple families, creating a complex locus. For complex loci, SSR-clouds gives information about each of the families present in the locus, including the average cloud stringency of that family's oligos in the locus and what percentage of the locus is covered by oligos from that family's clouds. We consider this output, which simultaneously considers all families that may be present in a locus, to more accurately reflect

the true nature of SSRs, given the propensity of SSRs to spawn different SSR motif families during their evolution.

By identifying millions of previously overlooked short and imperfect SSR loci, we provide evidence that the SSR life cycle is highly flexible and show that multiple paths to SSR death exist. While some of the short loci may be fossils of longer ancient loci that are no longer detectable, our analysis of *Alu* poly-A's suggests that only ~10% of mature SSR loci fall below detectability even after 65 million years. It thus seems reasonable that a substantial fraction of these short loci are more frequent than expected from point mutation processes and therefore created by some amount of slippage, but never reached SSR maturity where slippage events would have rapidly increased the locus size, and instead died in their infancy. Regardless of their precise origins, it is reasonable to think that these short loci may yet act as birthing grounds and nurseries for future SSRs, thus creating another alternate route through the SSR life cycle without ever passing through adulthood. The abundance of these short SSR-derived loci also indicates that SSRs may be born much more frequently than appreciated; with nearly 9 million separate loci, there is an average of one SSR for every 350 bp.

An important feature included in SSR-clouds that is lacking in standard SSR annotation software is the estimation of false discovery rates for each locus. Recently active SSR loci can be identified with high confidence because they have spent little time in the genomic churn caused by mutation and fragmentation, but this is not the case for millions of ancient SSR loci that we identified here. We note that even the short loci with high false discovery rates, although they may not be derived from mature SSR loci with high slippage rates, may be important to identify as potential sources of new SSR loci. Furthermore, loci with high false discovery rates can be

included or excluded in downstream analyses based on user-defined analysis-specific false discovery thresholds and the needs and tolerances of the researchers for both false discoveries and failure to detect relevant elements. Figure S5 illustrates the effect of different false discovery thresholds on the total number of base pairs identified as SSRs in the human genome.

Conclusions

We extend previous reports of sequence bias near SSR loci (29, 30) and show that the boundaries of this bias, though motif dependent, may extend for hundreds of base pairs to either side of an SSR locus. The length of sequence bias near SSR loci indicates that distinct boundaries on the distance of SSR spawning events exist, and the data presented here suggests that such events are generally limited to within several hundred base pairs of parent loci. Our characterization of similarity between clustered SSR loci supports this assertion and provides further evidence that the generation of new SSR loci is greatly influenced by the evolution of locally active SSRs.

Because the motif, purity, and length-dependent nature of SSR locus evolution is complex, the SSR-clouds approach presents an important and tractable method to improve studies of the different phases of the SSR life cycle that cannot be easily achieved through other approaches. The data presented here reveal unprecedented detail into the proposed SSR life cycle (15-18). The signals of highly biased sequence near SSR loci and clustered similar loci (see Figs 1-3) can be generated through repeated rounds of interrupting mutations within an SSR locus to isolate regions of the locus followed by expansion in regions that remain susceptible to slippage. This process of constant sloughing off of SSR detritus can be likened to simultaneous birth and

death processes, and creates natural boundaries at SSR loci, which we report here. This process also makes predictions about SSR sequence degeneracy over time possible; long dead SSR loci resemble the derived and most degenerate portions of active SSR loci that are near the boundaries of the SSR locus.

A large fraction of recent (4-6 million years old) *Alu* elements (~60%) have intact poly-A tails, and only a small fraction (<5%) have different motifs or no SSR at all in their poly-A tail region. Notably, the remaining nearly 40% have already begun to degenerate, even after relatively recent successful retrotransposition. However, although the poly-A appears to rapidly degenerate, these degenerate regions are detectable in many of even the oldest of *Alu* elements, demonstrating both a surprising longevity of SSR character in ancient simple repeats, and the sensitivity of SSR-clouds method.

The longevity of SSR loci is further highlighted by the fact that a substantial proportion (~15%) of poly-A's from the oldest group of *Alus* spawned new SSRs with different motifs (Figure 6). Spawning of SSRs has not been characterized in great detail (15), but this evidence, combined with the tendency of similar SSR repeats to cluster, presents a timeline for spawning events while also characterizing the expected motif bias for newly spawned loci.

The high degree of overlap between transposable elements and SSR loci we present here supports the hypothesis that transposable elements play a substantial role in the generation of SSR loci (27, 40, 44). About half (46.92%) of SSRs intersect with an easily-identifiable transposable element. Because about half the genome is made up of easily-identifiable transposable elements (1), this might suggest that SSR origins are similar in TE and non-TE

regions. However, evidence suggests that many transposable elements in the 'dark matter' portion of the genome are not-so-easily-identifiable (32, 33). It seems likely that a large fraction of the remaining SSRs were generated through the action of the hard-to-identify old and fragmented elements. Due to the ability of an SSR locus to maintain SSR character over long periods of time through constant slippage and spawning, the SSR loci identified by SSR-clouds may yet provide additional information in identifying the origins of 'dark matter' in the genome.

Methods

Annotation of perfect SSRs and surrounding regions

Oligonucleotide sequences representing all possible SSR sequences were created *in silico* using a *Perl* script that clusters alternate phases of the same SSR motif (ACT = CTA = TAC) and reverse complements of each phase into a single motif family. Perfect SSR repeat loci were defined as uninterrupted tandem repeats of a single motif family ≥ 12bp in length, and perfect stretches separated by 1 bp or more non-motif nucleotides were considered different loci. Perfect SSRs, as defined above, were annotated in an unmasked version of hg38.To identify sequence bias in regions near perfect SSR loci, each kmer *(k*-length oligonucleotide sequence) within 1,000 bp of a perfect repeat locus was compared with the kmers from different phases of the perfect motif. Mean similarities to the closest repeat kmer were calculated versus distance from locus boundaries, and distances between perfect SSR repeat loci were also recorded.

Constructing SSR-clouds

SSR-clouds were constructed similarly to cloud construction methods outlined in (32, 33) with modifications described here. To construct p-clouds from SSR-flanking regions we conservatively used 16-mer oligonucleotides and considered only 50 bp on either side of a perfect repeat locus as a template for cloud formation. P-clouds for each SSR motif family were constructed separately from one another using a training set that consisted of a randomly chosen subset of 90% of loci for each family, with the remaining 10% of loci used as annotation tests. Loci that were separated by fewer than 100 bp from other loci of the same family were merged into a single locus before cloud formation to prevent double counting oligos in the regions between the loci. Following standard *P-cloud*s formation protocol (32), p-clouds were organized around 16-mer core oligonucleotides, including every 16-mer oligo with count above the threshold that was within one nucleotide of the cloud core or any other oligo already in a cloud. For each motif family, we created nested oligonucleotide clouds using lower threshold counts for clouds of lower stringency, such that all oligonucleotides of higher stringency clouds were included in lower stringency clouds. Perfectly repeated 12-mer oligonucleotides were also automatically added to the highest stringency cloud. Different threshold counts were used as criteria for inclusion in p-cloud sets for each motif family depending on the total number of perfect loci used for cloud training, though motif families with fewer than 100 loci in the training set were not used in cloud building. These thresholds, the number of loci used in cloud formation, and the counts of unique oligonucleotides in each stringency level are specified in Table S1. Transposable elements (e.g., *Alu* in humans) were not our targets but are highly represented in regions flanking SSRs, and so all transposable elements annotated by

RepeatMasker (35) were removed prior to cloud formation. Because clouds were formed separately for each family, individual oligonucleotides, including those representing perfect repeats, can belong to cloud sets for multiple families.

Annotation with SSR-clouds was performed in an unmasked version of hg38 by simultaneously mapping oligonucleotide clouds from all motif families, and then merging loci within 5 bp of each other into a single locus. Annotations with merge distances of 0 bp and 30 bp were also performed and are presented as supplements. After annotation, loci were ranked and separated according to the highest stringency cloud found in the locus. In analyses presented here that use only single motif families, (poly-A and $(AC)_n$), annotation was performed in the same way except that only oligonucleotides created from that family were used.

Simulating genomes to obtain false positive rates

Fifteen simulated genomes were created from the human genome (hg38) using nucleotide and dinucleotide frequencies obtained from1 Mb windows along the genome. Prior to creation of the simulated genomes, all regions annotated as either a perfect SSR or annotated as transposable elements or other repeat regions by RepeatMasker were masked so that they would be representative of non-repetitive portions of the genome. The simulations proceeded by randomly selecting nucleotides conditional on the dinucleotide frequencies. When the previous nucleotide was absent or undetermined, a starting nucleotide was selected based on independent single nucleotide frequencies. SSR clouds were annotated in the simulated genomes exactly as done for the actual genome. False positive rates for each locus length (or longer) were calculated, for each cloud stringency setting, as the cumulative amount of simulated sequence annotated, divided by

the amount of sequence analyzed. Under a given stringency setting, the length of a locus was considered to be the longest stretch of the locus that was consecutively annotated. False positive rates for each locus length and cloud stringency category were calculated for hg38. False discovery rates were then calculated as the expected cumulative falsely annotated sequence, conservatively assuming the entire genome is not SSR, divided by the observed cumulative length annotated for each setting

Comparison with Tandem Repeats Finder Annotations

Tandem Repeats Finder (TRF) (31) version 4.07b was run under the two parameter sets described in Warren et al. 2008 that were applied to the human genome (hg38) with centromeres and telomeres masked. The two resulting annotation sets were merged to obtain the TRF annotation used here. TRF SSR annotations were segregated into groups by motif family and annotations within each family were merged using BEDTools (36). The BEDTools Intersect function was used to search for SSR-clouds annotations that overlapped with TRF SSR annotations and to determine the number of novel SSR-clouds annotations.

Intersection with Poly-A Regions of Alu Elements for Age Analysis

Full-length and non-concatenated *Alu* elements were obtained by filtering RepeatMasker *Alu* annotations from the hg38 assembly of the human genome. Relative ages of each element (measured in inferred number of substitutions since retrotransposition) were then estimated by applying the AnTE method to this dataset (37). We began with 823,789 individual full-length *Alu* elements, with each element having an estimated age or retrotransposition relative to the mean age of retrotransposition of all *Alu* elements. To maximize the chances that the *Alus* tested

still contained their poly-A tail, we removed all *Alus* that were < 275 bp or > 325 bp in length as well as those *Alus* that were within 50 bp of another TE. After filtering, 407,438 *Alus* remained.

The remaining *Alu* annotations were split into three groups by age and roughly based on the major expansions of *AluY*, *AluS*, and *AluJ*. The youngest group consisted of 57,873 *Alu* elements, ~97% of which are classified as *AluY* by RepeatMasker, with a mean age of 0.51 relative to the mean age of all *Alus*. The second and largest group, 99% of which are classified as *AluS* elements, consisted of 265,219 elements with a mean age of 0.92 relative to the mean age of all *Alus*. The third group consisted of all *Alu* elements older than those included in the first two groups, 90% of which are classified as *AluJ* and 10% as *AluS*, and had 84,346 elements with a mean age of 1.6 relative to the mean age of all *Alus*.

To ensure detection of only the poly-A region of *Alu* rather than other SSR-rich regions in *Alu*, we used the 30 bp directly 3' to each *Alu* tested for intersection. We used the *intersect* function from BEDTools (36) to count the number of *Alu* elements that intersected each of the poly-A SSR annotations, beginning with the highest stringency poly-A annotations and proceeding to the lowest stringency annotations.

Declarations

Availability of data and materials

The SSR-clouds software and datasets generated and/or analyzed during the current study are available either in the GitHub repository, https://github.com/popgengent/SSRclouds, or from the corresponding author on reasonable request. The P-clouds SSR package was written and

implemented in *Perl*. The program parameters can be easily modified for different applications and sensitivity via either the command line or a control file.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Institutes of Health (NIH; GM083127 and GM097251 to DDP).

Authors' contributions

JAS developed code, designed, performed, and interpreted analyses, and was a major contributor in writing the manuscript; RPR contributed code, designed, and interpreted analyses; CC developed *Perl* version of *P-clouds* used in SSR-clouds code; ACW contributed the library of *Alu* elements and determined their times of retrotransposition; DDP designed, consulted, and interpreted analyses, and was a major contributor in writing the manuscript.

Acknowledgements

Not applicable.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.

2. Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, et al. Genome analysis of the platypus reveals unique signatures of evolution. Nature. 2008;453(7192):175-83.

3. Weber JL, Wong C. Mutation of human short tandem repeats. Human Molecular Genetics. 1993;2(8):1123-8.

4. Ellegren H. Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics. 2004;5(6).

5. Mirkin SM. Expandable DNA repeats and human disease. Nature. 2007;447(7147):932-40.

6. Kashi Y, King DG. Simple sequence repeats as advantageous mutators in evolution. Trends in Genetics. 2006;22(5).

7. Duitama J, Zablotskaya A, Gemayel R, Jansen A, Belet S, Vermeesch JR, et al. Large-scale analysis of tandem repeat variability in the human genome. Nucleic Acids Research. 2014;42(9):5728-41.

8. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annual review of genetics. 2010;44:445-77.

9. Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability. Science. 2009;324(5931):1213-6.

10. King DG. Evolution of simple sequence repeats as mutable sites. Adv Exp Med Biol. 2012;769:10-25. PubMed PMID: 23560302. eng.

11. Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, et al. Microsatellite Tandem Repeats Are Abundant in Human Promoters and Are Associated with Regulatory Elements. PLoS ONE. 2013;8(2).

12. Bilgin Sonay T, Carvalho T, Robinson MD, Greminger MP, Krützen M, Comas D, et al. Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. Genome research. 2015;25(11):1591-9.

13. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nature Genetics. 2016;48(1).

14. Nazaripanah N, Adelirad F, Delbari A, Sahaf R, Abbasi-Asl T, Ohadi M. Genome-scale portrait and evolutionary significance of human-specific core promoter tri- and tetranucleotide short tandem repeats. Hum Genomics. 2018 Apr;12(1):17. PubMed PMID: 29622039. PMCID: PMC5887250. Epub 2018/04/05. eng.

15. Kelkar YD, Eckert KA, Chiaromonte F, Makova KD. A matter of life or death: How microsatellites emerge in and vanish from the human genome. Genome Research. 2011;21(12).

16. Buschiazzo E, Gemmell NJ. The rise, fall and renaissance of microsatellites in eukaryotic genomes. BioEssays. 2006;28(10):1040-50.

17. Taylor JS, Durkin JM, Breden F. The death of a microsatellite: a phylogenetic perspective on microsatellite interruptions. Mol Biol Evol. 1999 Apr;16(4):567-72. PubMed PMID: 10331282. eng.

18. Messier W, Li SH, Stewart CB. The birth of microsatellites. Nature. 1996 Jun;381(6582):483. PubMed PMID: 8632820. eng.

19. Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Research. 2008;18(1):30-8.

20. Ellegren H. Heterogeneous mutation processes in human microsatellite DNA sequences. Nature Genetics. 2000;24(4):400-2.

21. Sun JX, Helgason A, Masson G, Ebenesersdóttir SSS, Li H, Mallick S, et al. A direct characterization of human mutation based on microsatellites. Nature genetics. 2012;44(10):1161-5.

22. Baptiste BA, Ananda G, Strubczewski N, Lutzkanin A, Khoo SJ, Srikanth A, et al. Mature microsatellites: mechanisms underlying dinucleotide microsatellite mutational biases in human cells. G3 (Bethesda, Md). 2013;3(3):451-63.

23. Bacon AL, Farrington SM, Dunlop MG. Sequence interruptions confer differential stability at microsatellite alleles in mismatch repair-deficient cells. Human molecular genetics. 2000;9(18):2707-13.

24. Ananda G, Hile SE, Breski A, Wang Y, Kelkar Y, Makova KD, et al. Microsatellite interruptions stabilize primate genomes and exist as population-specific single nucleotide polymorphisms within individual human genomes. PLoS genetics. 2014;10(7).

25. Goldstein DB, Pollock DD. Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. The Journal of heredity. 1997;88(5):335-42.

26. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nature Reviews Genetics. 2002;3(5):370-9.

27. Ahmed M, Liang P. Transposable Elements Are a Significant Contributor to Tandem Repeats in the Human Genome. International Journal of Genomics. 2012;2012.

28. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Research. 2013;41(D1).

29. Vowles EJ, Amos W. Evidence for widespread convergent evolution around human microsatellites. PLoS biology. 2004;2(8).

30. Webster MT, Hagberg J. Is there evidence for convergent evolution around human microsatellites? Molecular biology and evolution. 2007;24(5):1097-100.

31. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research. 1999;27(2):573-80.

32. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS genetics. 2011;7(12).

33. Gu W, Castoe TA, Hedges DJ, Batzer MA, Pollock DD. Identification of repeat structure in large genomes using repeat probability clouds. Analytical biochemistry. 2008;380(1):77-83.

34. Maumus F, Quesneville H. Impact and insights from ancient repetitive elements in plant genomes. Curr Opin Plant Biol. 2016 04;30:41-6. PubMed PMID: 26874965. Epub 2016/02/09. eng.

35. RepeatMasker Open-4.0 [Internet]. 2013-2015.

36. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2014;47.

37. Wacholder AC, Cox C, Meyer TJ, Ruggiero RP, Vemulapalli V, Damert A, et al. Inference of Transposable Element Ancestry. PLoS Genetics. 2014;10(8).

38. Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, Makova KD. What Is a Microsatellite: A Computational and Experimental Definition Based upon Repeat Mutational Behavior at A/T and GT/AC Repeats. Genome Biology and Evolution. 2010;2(0).

39. Ananda G, Walsh E, Jacob KD, Krasilnikova M, Eckert KA, Chiaromonte F, et al. Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. Genome Biol Evol. 2013;5(3):606-20. PubMed PMID: 23241442. PMCID: PMC3622297. eng.

40. Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA. Alu repeats: a source for the genesis of primate microsatellites. Genomics. 1995 Sep;29(1):136-44. PubMed PMID: 8530063. eng.

41. Willems T, Gymrek M, Highnam G, Genomes Project C, Mittelman D, Erlich Y. The landscape of human STR variation. Genome research. 2014;24(11):1894-904.

42. Roy-Engel AM, Salem AH, Oyeniran OO. Active Alu element “A-tails”: size does matter. Active Alu element “A-tails”: size does matter. 2002.

43. Dewannieux M, Heidmann T. Role of poly(A) tail length in Alu retrotransposition. Genomics. 2005;86(3):378-81.

44. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nature reviews Genetics. 2009;10(10):691-703.

Figures and Tables

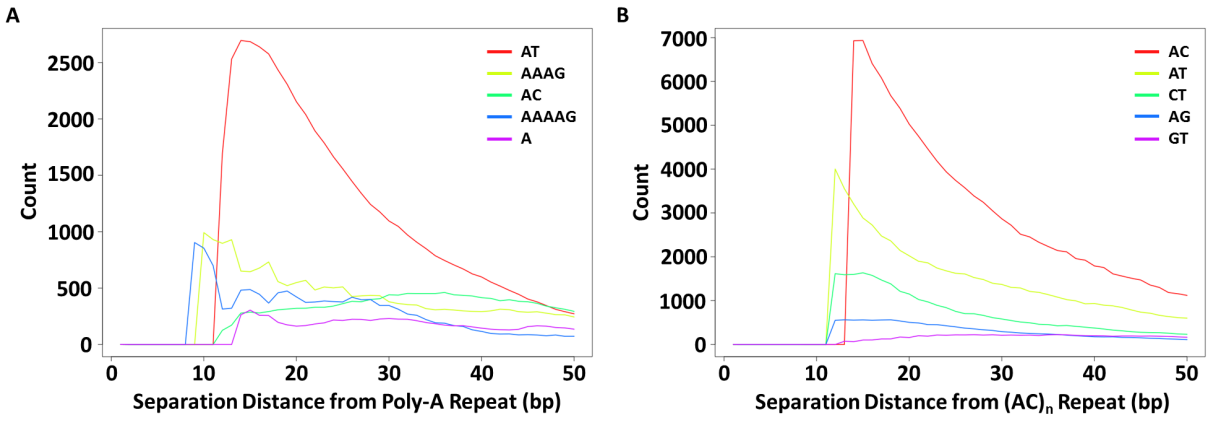

Fig. 1. Clustering of SSR loci depending on motif similarity. All perfect SSRs (≥12 bp) were annotated in a transposable-element masked version of the human genome (hg38) and the count of nearby SSR motifs were recorded as a function of distance from the repeat. Here, we show the 5 motifs that are most frequently found near (A) perfect poly-A SSRs (n=350,763); and (B) perfect $(AC)_n$ SSRs (n=85,161). The motifs of nearby SSRs often differ from the repeated motif by simple mutations. To allow for overlapping non-reference motif families (i.e., a compound locus comprised of two or more different motif families), x=0 begins 11 bp within the perfect reference motif repeat. Flat curves at x=0 reflects that the first several bases are still part of the perfect repeat and thus can only be annotated by another family to the extent that their motifs overlap.

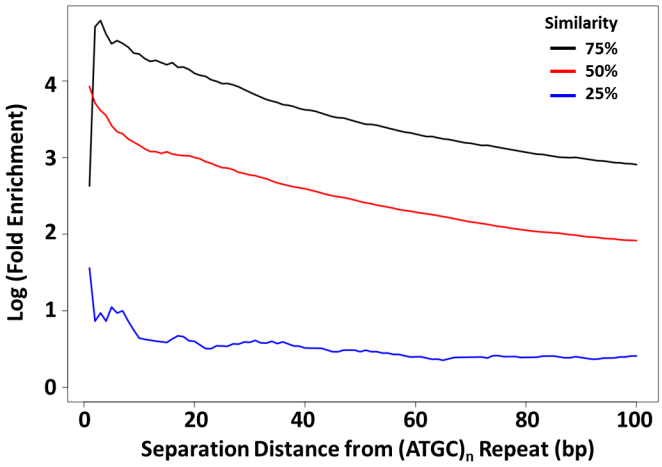

Fig. 2. Enrichment of similar SSR loci near ATGC repeat loci. The average enrichment levels of perfect SSR loci within 100 bp of a perfect ATGC repeat locus are shown for SSR families with motifs with 1 (75%, black), 2 (50%, red), or 3 (25%, blue) differences from the 'ATGC' motif. Enrichment for SSR motifs was determined relative to the genomic average for all possible motifs with the given difference.

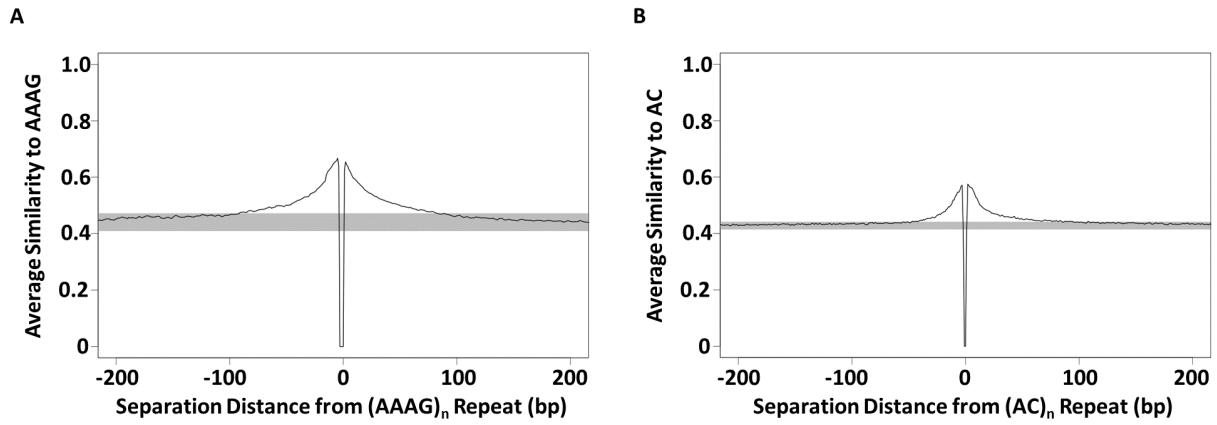

Fig. 3. Decay of sequence similarity with distance from perfect SSR repeats. Average similarities were calculated for short segments within 200 bp of perfect SSR repeats with a given motif. Similarity was measured as the proportion of identical nucleotides at each position for a segment of the same length and read direction as the repeated motif shown, $(AAAG)_n$ in A, $(AC)_n$ in B. For example, a segment reading "ATAG" would have a similarity of 0.75 with the repeat motif "AAAG". Average similarities were calculated for segments beginning at every nucleotide separation distance within 200 bp of the perfect repeat beginning or end. The black line shows the average similarity to each repeat, while the gray box shows a range of 3 standard deviations from the mean similarities calculated in 700 bp windows from 300-1,000 bp away from both ends of the perfect repeat loci. The dips near x=0 reflect that a non-motif base must precede and follow the perfect region of the repeat at the start and end of the perfectly repeated segment.

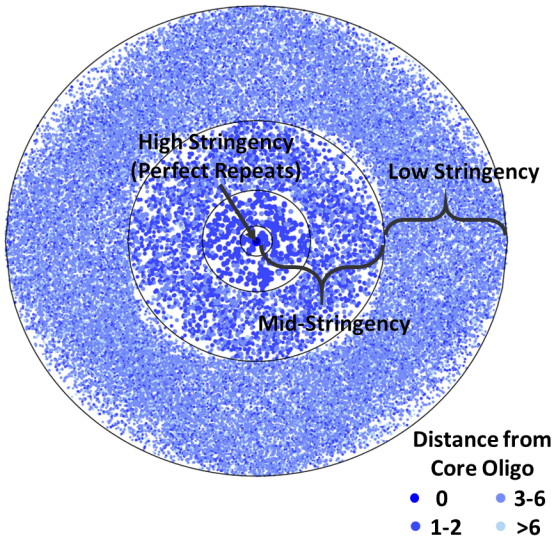

Fig. 4. Visual of numbers of poly-A cloud oligos with different similarities from poly-A. Each point represents a 16-mer oligo built into the cloud set for the poly-A SSR family, with oligos clustered into concentric rings depending on distance from the core oligo (poly-A). Darker shades of blue near the center represent higher similarity cloud oligos, and lighter shades represent lower similarity cloud oligos, as indicated in the legend.

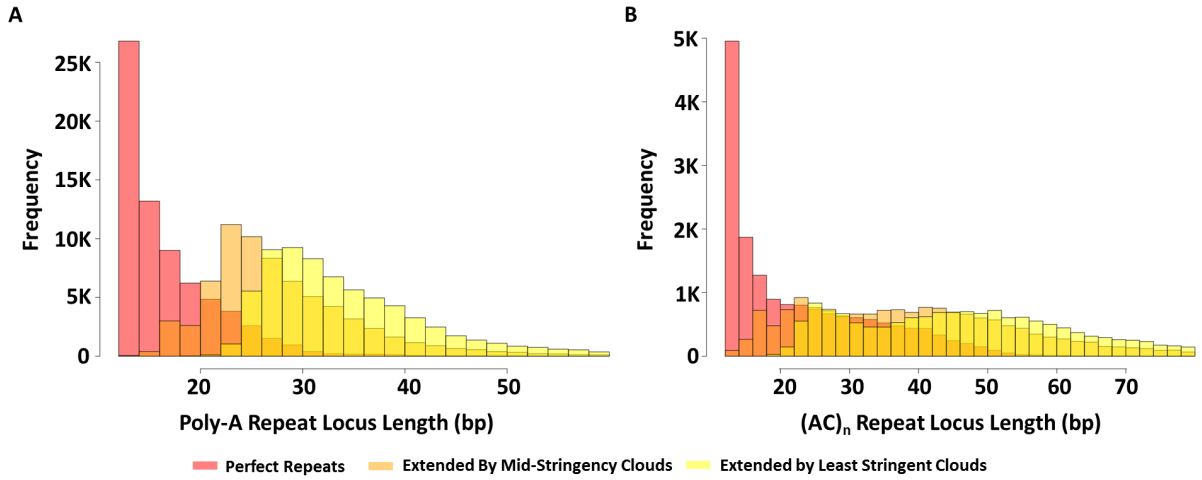

Fig. 5. Length distribution of perfect SSR loci annotations expanded using SSR-derived oligos. The length count distributions are shown for: (A) poly-A SSRs; and (B), $(AC)_n$ SSRs. Perfect repeat annotations are shown in red, mid-stringency annotations in light orange, and low-stringency annotations in yellow, with darker regions indicating overlap.

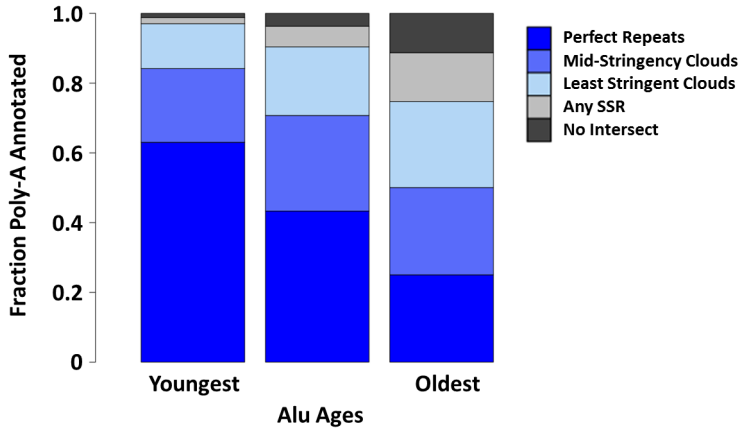

Fig. 6. SSR-cloud annotation of poly-A regions adjacent to annotated *Alu*s. Full length *Alus* (275-325 bp) were divided into three groups based on their age (roughly corresponding to the three major expansions of *Alu, AluJ, AluS, and AluY*) and 5' overlap with poly-A SSR-cloud annotated regions was evaluated. The region expected to carry the poly-A tail was defined as within 30 bp of the *Alu* terminus. Different cloud stringency extensions are colored with dark blue indicating highest stringency poly-A annotations found, and light blue lowest-stringency poly-A annotations. If no poly-A annotations were found, other SSR-cloud loci found are shown in light gray, and no intersecting SSR annotations found shown in dark grey.

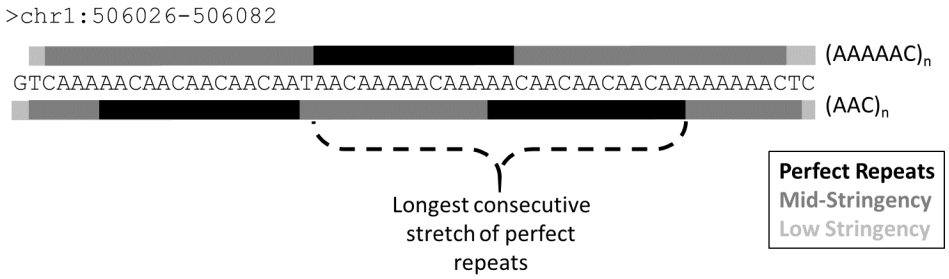

Fig. 7. Anatomy of a complex SSR locus and its annotation by SSR-clouds. The sequence for an SSR locus found at bp 506026-506082 on chromosome 1 in hg38 is shown. Regions annotated by the two most prevalent families, AAAAAC (top) and AAC (bottom), are shown, with perfect repeats indicated with a black bar, mid-stringency cloud annotations with a dark gray bar, and the lowest stringency cloud annotations with a light gray bar. The longest stretch of perfect repeats of any kind (26 bp) is indicated, and was used to determine the false discovery rate of the locus (see Methods).

Table 1. SSR-clouds recovery of Tandem Repeats Finder (TRF) loci

		Highest Cloud Stringency of Locus						FDR ≤ %5	
		Perfect Repeats		Mid-stringency		Low Stringency			
		Loci	bp	Loci	bp	Loci	bp	Loci	bp
Poly-A	P-Clouds TRF Intersection	453,128	11,518,426	615,893	16,085,955	556,794	17,373,114	660,469	17,272,038
	Total P-Cloud Recovery of TRF	67.73%	62.37%	92.06%	87.10%	83.23%	94.07%	98.72%	93.52%
	Novel Clouds	244,269	13,490,320	889,630	36,272,378	2,282,559	65,260,452	1,552,401	53,363,205

Total TRF Poly-A Loci = 669,020

Total TRF Poly-A bp = 18,468,468

		Highest Cloud Stringency of Locus						FDR ≤ %5	
		Perfect Repeats		Mid-stringency		Low Stringency			
		Loci	bp	Loci	bp	Loci	bp	Loci	bp
$(AC)_n$	P-Clouds TRF Intersection	120,498	4,813,795	143,941	5,989,636	148,027	6,301,466	148,027	6,301,466
	Total P-Cloud Recovery of TRF	81.09%	65.02%	96.86%	80.90%	99.61%	85.11%	99.61%	85.11%
	Novel Clouds	28,365	3,444,295	724,496	25,393,739	1,621,096	44,746,021	1,621,096	44,746,021

Total TRF $(AC)_n$ Loci = 148,607

Total TRF $(AC)_n$ bp = 7,403,867

		Highest Cloud Stringency of Locus						FDR ≤ %5	
		Perfect Repeats		Mid-stringency		Low Stringency			
		Loci	bp	Loci	bp	Loci	bp	Loci	bp
All SSRs	P-Clouds TRF Intersection	1,741,873	59,642,996	1,965,320	67,616,136	2,119,405	71,906,834	1,946,410	68,221,956
	Total P-Cloud Recovery of TRF	78.73%	67.40%	88.83%	76.41%	95.80%	81.26%	87.98%	77.10%
	Novel Clouds	2,046,914	58,749,285	2,690,429	75,993,192	6,702,981	149,673,223	2,008,354	70,732,930

Total TRF SSR Loci = 2,212,414

Total TRF SSR bp = 88,485,889

SSR-clouds loci with a merge distance of 5 bp were divided into 3 nested sets based on the most stringent oligo used to annotate each locus. Cells in the table report the number of loci or bp that overlap with TRF loci as well as the number of novel SSR-clouds loci and bp. Comparisons were also made for SSR-clouds loci with FDR ≤ 5%.

Supplemental Figure and Table Legends

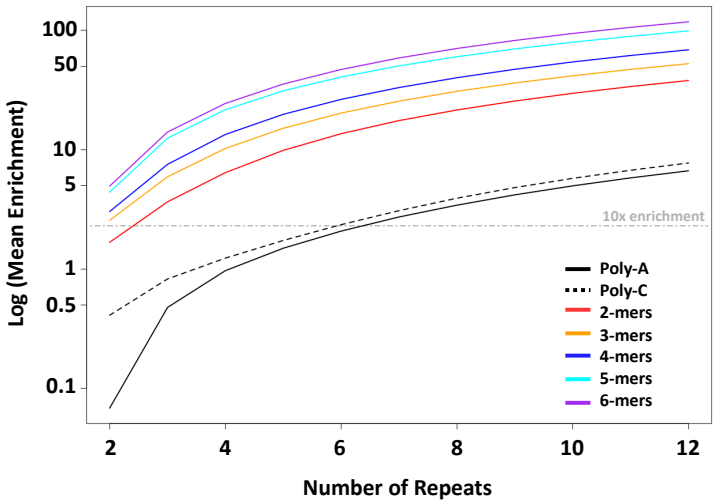

Fig. S1. Enrichment of SSRs in the human genome. The mean enrichment of perfect repeats is shown relative to expectation from single nucleotide frequencies. All SSR motifs of a given length were clustered into groups, except that the Poly-A and poly-C single nucleotide repeats are shown as separate lines. The enrichment is shown for the number of repeats of a given size observed in tandem, and the gray dashed lines indicate 10x, 100x, and 1000x enrichments.

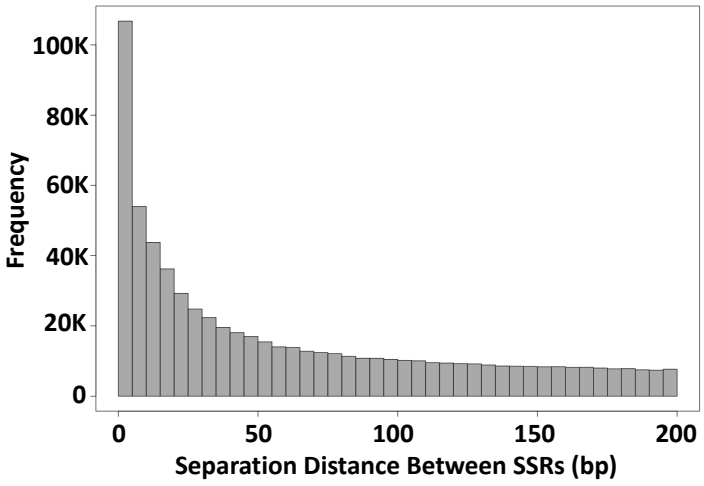

Fig. S2. Separation distance between perfect SSRs in the human genome. The frequency of pairs of perfect SSRs ≥ 12 bp long with a given separation distance is shown. The separation distances were binned into groups of 5. The results in A) are for a masked version of the human genome, while B) shows results for an unmasked genome, demonstrating the strong effect and particular features of transposable element SSRs.

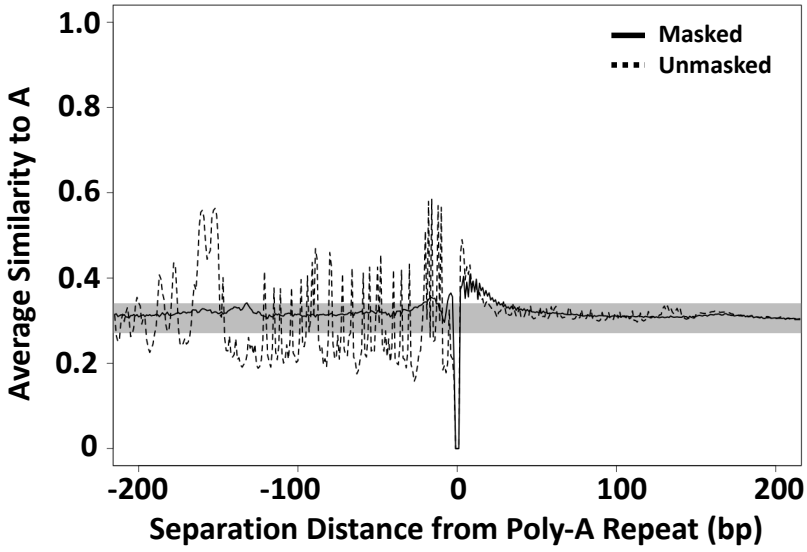

Fig. S3. Asymmetric similarity to poly-A. The frequency of adenine nucleotides (A) at every site within 200 bp of perfect poly-A repeats. The solid line shows the frequency of A in a human genome where all transposable elements have been masked and the dotted line shows the frequency in an unmasked human genome. As a reference, the gray box represents a range of 3 standard deviations from the mean frequencies of A calculated in 700 bp windows from 300-1,000 bp away from both ends of all perfect repeats. The strongly varying frequencies in the unmasked genome are mostly a symptom of the high copy number of retroelements such as Alu and Line1. The asymmetric frequency of A's adjacent to perfect A repeats in the masked genome likely reflects incomplete masking of transposable elements and the existence of other unmasked retrotransposed sequences in what would have been the 5' region of the retrotransposed poly-A mRNAs.

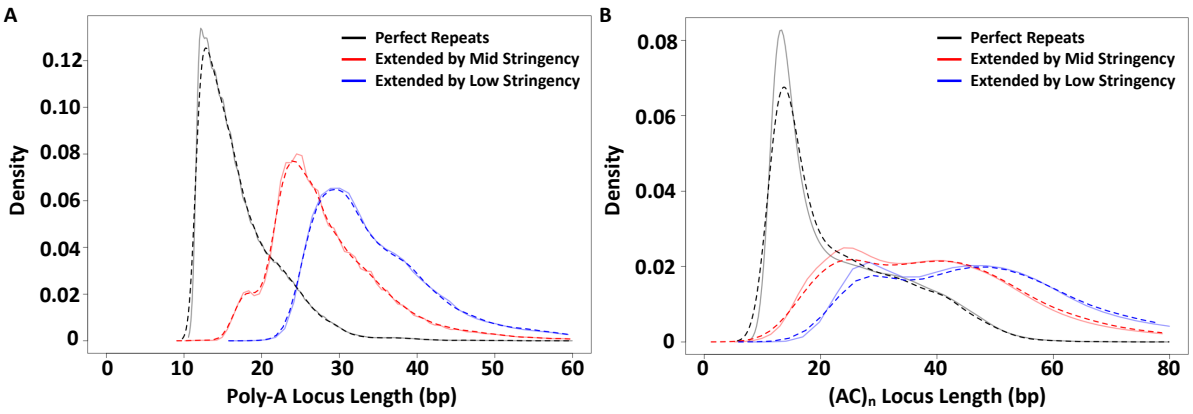

Fig. S4. Cloud extension length distributions of training and test loci. Locus length density plots of SSR loci containing perfect repeats (black) and lengths after extension by mid- (red) and low-stringency (blue) cloud sets. Solid lines depict the distributions of lengths for training loci and dashed lines depict the almost perfectly overlapping distributions of lengths for test loci.

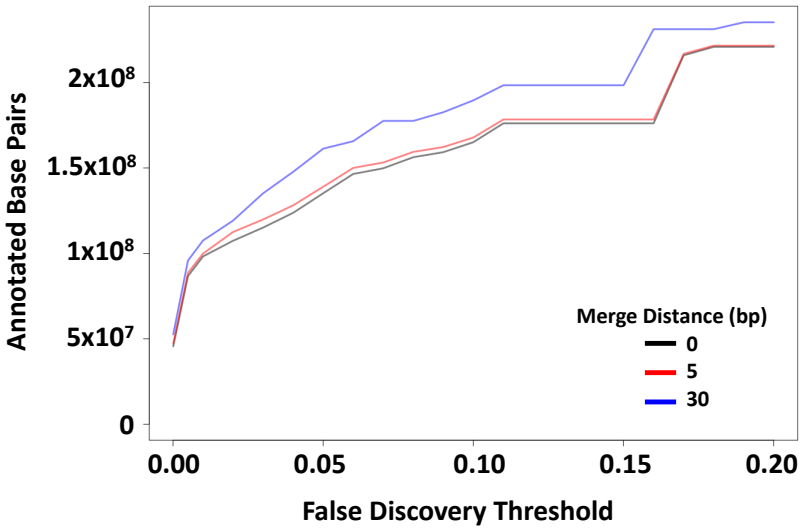

Fig. S5. Genomic SSR content annotated with different merge distances and false discovery thresholds. The number of bp in the human genome that were annotated by SSR-clouds under various conditions are shown. with different merge distances and false discovery thresholds. Three lines are shown for merge distances of 0 bp (black), 5 bp (red), and 30 bp (blue), with the per-locus maximum false discovery criterion on the X axis.

Tables S1 and S2 are included at the very end of the manuscript due to their length

Table S3. SSR-clouds recovery of Tandem Repeats Finder (TRF) loci

		Highest Cloud Stringency of Locus						FDR ≤ %5	
		Perfect Repeats		Mid-stringency		Low Stringency			
		Loci	bp	Loci	bp	Loci	bp	Loci	bp
Poly-A	P-Clouds TRF Intersection	452,967	11,456,790	615,785	16,008,805	665,794	17,354,944	660,151	17,209,029
	Total P-Cloud Recovery of TRF	67.71%	62.03%	92.04%	86.68%	99.52%	93.97%	98.67%	93.18%
	Novel Clouds	244,608	13,318,976	35,515,654	903,315	2,348,525	65,094,178	1,590,055	52,714,179

Total TRF Poly-A Loci = 669,020

Total TRF Poly-A bp = 18,468,468

		Highest Cloud Stringency of Locus						FDR ≤ %5	
		Perfect Repeats		Mid-stringency		Low Stringency			
		Loci	bp	Loci	bp	Loci	bp	Loci	bp
$(AC)_n$	P-Clouds TRF Intersection	120,385	4,673,530	143,915	5,871,478	148,027	6,271,366	148,027	6,271,366
	Total P-Cloud Recovery of TRF	81.01%	63.12%	96.84%	79.30%	99.61%	84.70%	99.61%	84.70%
	Novel Clouds	28,691	3,188,467	742,448	24,760,007	1,680,113	44,595,076	44,595,076	1,680,113

Total TRF $(AC)_n$ Loci = 148,607

Total TRF $(AC)_n$ bp = 7,403,867

		Highest Cloud Stringency of Locus						FDR ≤ %5	
		Perfect Repeats		Mid-stringency		Low Stringency			
		Loci	bp	Loci	bp	Loci	bp	Loci	bp
All SSRs	P-Clouds TRF Intersection	1,738,346	57,867,165	1,964,039	66,587,760	2,119,405	71,596,437	1,939,140	67,143,121
	Total P-Cloud Recovery of TRF	78.57%	65.40%	88.77%	75.25%	95.80%	80.91%	87.65%	75.88%
	Novel Clouds	2,070,465	57,053,674	2,736,626	74,441,046	6,841,941	149,260,686	2,014,804	68,007,767

Total TRF SSR Loci = 2,212,414

Total TRF SSR bp = 88,485,889

SSR-clouds loci with a merge distance of 0 bp were divided into 3 nested sets based on the most stringent oligo used to annotate each locus. Cells in the table report the number of loci or bp that overlap with TRF loci as well as the number of novel SSR-clouds loci and bp. Comparisons were also made for SSR-clouds loci with FDR $\leq$ 5%.

Table S4. SSR-clouds recovery of Tandem Repeats Finder (TRF) loci

		Highest Cloud Stringency of Locus						FDR ≤ %5	
		Perfect Repeats		Mid-stringency		Low Stringency			
		Loci	bp	Loci	bp	Loci	bp	Loci	bp
Poly-A	P-Clouds TRF Intersection	454,655	11,649,451	616,750	16,209,662	665,794	17,446,377	660,870	17,373,267
	Total P-Cloud Recovery of TRF	67.96%	63.08%	92.19%	87.77%	99.52%	94.47%	98.78%	94.07%
	Novel Clouds	241,932	14,420,787	862,842	39,084,443	2,151,156	67,376,882	1,484,085	56,519,523

Total TRF Poly-A Loci = 669,020

Total TRF Poly-A bp = 18,468,468

		Highest Cloud Stringency of Locus						FDR ≤ %5	
		Perfect Repeats		Mid-stringency		Low Stringency			
		Loci	bp	Loci	bp	Loci	bp	Loci	bp
$(AC)_n$	P-Clouds TRF Intersection	120,822	5,311,956	144,057	6,394,349	148,027	6,573,040	148,027	6,573,040
	Total P-Cloud Recovery of TRF	81.30%	71.75%	96.94%	86.36%	99.61%	88.78%	99.61%	88.78%
	Novel Clouds	27,486	4,207,203	696,337	27,533,146	1,524,044	46,227,577	1,524,044	46,227,577

Total TRF $(AC)_n$ Loci = 148,607

Total TRF $(AC)_n$ bp = 7,403,867

		Highest Cloud Stringency of Locus						FDR ≤ %5	
		Perfect Repeats		Mid-stringency		Low Stringency			
		Loci	bp	Loci	bp	Loci	bp	Loci	bp
All SSRs	P-Clouds TRF Intersection	1,762,767	64,154,932	1,974,700	70,907,679	2,119,427	74,288,802	1,956,493	71,532,092
	Total P-Cloud Recovery of TRF	79.68%	72.50%	89.26%	80.13%	95.80%	83.96%	88.43%	80.84%
	Novel Clouds	1,922,645	73,344,350	2,504,656	91,070,015	6,072,473	160,927,330	1,889,528	89,725,301

Total TRF SSR Loci = 2,212,414

Total TRF SSR bp = 88,485,889

SSR-clouds loci with a merge distance of 30 bp were divided into 3 nested sets based on the most stringent oligo used to annotate each locus. Cells in the table report the number of loci or bp that overlap with TRF loci as well as the number of novel SSR-clouds loci and bp. Comparisons were also made for SSR-clouds loci with FDR ≤ 5%.

Table S1. SSR-clouds construction summary

Motif	Number of Training Loci	Cloud Threshold Fold-Reduction	Threshold Counts					Stringency specific Kmer Counts			
			Cloud 1 (High Stringency)	Cloud 2 (Mid Stringency)	Cloud 3 (Mid Stringency)	Cloud 4 (Mid Stringency)	Cloud 5 (Low Stringency)	High Stringency (cloud 1 16-mers and perfect repeat 12-mers)	Mid-Stringency (clouds 2-4)	Low Stringency (cloud 5)	All Stringencies
A	632,710	10.96	158,178	14,432	1,317	121	11	4	3,774	44,500	48,278
AC	153,656	8.26	38,414	4,652	564	69	9	8	5,708	48,027	53,743
AAAAAT	139,645	8.10	34,912	4,309	532	66	9	24	1,716	22,333	24,073
AAAT	128,396	7.97	32,099	4,029	506	64	8	16	1,778	20,572	22,366
AAAAT	106,065	7.67	26,517	3,458	451	59	8	20	1,744	20,537	22,301
AAAAAG	104,414	7.64	26,104	3,415	447	59	8	24	2,230	24,838	27,092
AAAAC	99,142	7.57	24,786	3,277	434	58	8	20	1,754	17,925	19,699
AAAAG	97,506	7.54	24,377	3,233	429	57	8	20	2,474	22,699	25,193
ACCTCC	96,599	7.53	24,150	3,209	427	57	8	24	12	1,466	1,502
AAAAAC	92,317	7.46	23,080	3,095	415	56	8	24	1,290	16,369	17,683
AT	84,257	7.32	21,065	2,877	393	54	8	6	4,932	31,869	36,807
AAAC	82,417	7.29	20,605	2,826	388	54	8	16	1,394	11,207	12,617
AAAG	82,179	7.29	20,545	2,820	387	54	8	16	3,994	24,636	28,646
AG	68,928	7.04	17,232	2,450	349	50	8	8	4,238	29,703	33,949
AAT	62,928	6.91	15,732	2,278	330	48	7	12	1,363	12,762	14,137
AAGG	44,951	6.46	11,238	1,740	270	42	7	16	5,184	29,681	34,881
AATGG	43,036	6.40	10,759	1,681	263	41	7	20	7,576	37,242	44,838
AAC	42,645	6.39	10,662	1,669	262	41	7	12	1,226	7,328	8,566
AAAATT	38,889	6.27	9,723	1,550	247	40	7	24	898	12,061	12,983
AAAGAG	35,527	6.16	8,882	1,442	234	38	7	24	2,596	18,489	21,109
AGGG	32,136	6.04	8,034	1,331	221	37	7	16	3,800	23,127	26,943
AAATAT	31,778	6.03	7,945	1,319	219	37	7	24	1,778	16,294	18,096
AATG	31,161	6.00	7,791	1,298	217	37	7	16	934	7,612	8,562
ACATAT	26,373	5.81	6,594	1,136	196	34	6	24	2,899	20,525	23,448
AAAATG	23,965	5.70	5,992	1,052	185	33	6	24	182	9,109	9,315
AGCCTC	23,287	5.66	5,822	1,029	182	33	6	24	0	1,835	1,859
AACAG	23,095	5.65	5,774	1,022	181	32	6	20	16	1,051	1,087
ACCCCC	22,985	5.65	5,747	1,018	181	32	6	24	788	9,851	10,663

AAAAGG	22,632	5.63	5,658	1,005	179	32	6	24	450	9,565	10,039
AACTAG	21,214	5.56	5,304	955	172	31	6	24	0	39	63
AGG	20,610	5.53	5,153	933	169	31	6	12	1,900	13,837	15,749
ACACAT	20,268	5.51	5,067	921	168	31	6	24	2,166	15,253	17,443
AAATT	20,115	5.50	5,029	915	167	31	6	20	639	7,303	7,962
AAATGT	19,331	5.46	4,833	886	163	30	6	24	16	3,102	3,142
AAAGG	19,177	5.45	4,795	881	162	30	6	20	1,680	12,682	14,382
ATCC	18,896	5.43	4,724	870	161	30	6	16	2,823	17,735	20,574
AGGGG	18,819	5.43	4,705	868	160	30	6	20	1,605	13,622	15,247
AGAGGC	18,789	5.42	4,698	866	160	30	6	24	57	3,531	3,612
AGAGGG	18,468	5.41	4,617	855	158	30	6	24	1,224	10,377	11,625
AGAT	17,201	5.33	4,301	807	152	29	6	16	1,911	11,841	13,768
AAATG	16,628	5.29	4,157	786	149	29	6	20	1,041	8,707	9,768
AAGAGG	16,363	5.28	4,091	776	147	28	6	24	976	9,450	10,450
AAG	16,293	5.27	4,074	773	147	28	6	12	1,196	7,331	8,539
AAACAG	16,088	5.26	4,022	765	146	28	6	24	52	4,375	4,451
ACAT	15,848	5.24	3,962	756	145	28	6	16	1,796	9,504	11,316
AGGGGG	15,644	5.23	3,911	748	144	28	6	24	728	8,283	9,035
AATAT	15,579	5.23	3,895	746	143	28	6	20	2,041	11,801	13,862
ACCCC	15,132	5.19	3,783	729	141	27	6	20	574	5,439	6,033
ACAGAG	15,043	5.19	3,761	725	140	27	6	24	1,386	9,183	10,593
AAGGG	14,036	5.12	3,509	686	135	27	6	20	1,583	10,418	12,021
AAATAG	13,917	5.11	3,480	682	134	27	6	24	42	4,178	4,244
AAATTT	13,469	5.08	3,368	664	131	26	6	18	651	5,453	6,122
AAGAG	13,252	5.06	3,313	655	130	26	6	20	540	6,367	6,927
AAATAC	12,690	5.02	3,173	633	127	26	6	24	26	2,981	3,031
AGAGG	12,431	4.99	3,108	623	125	25	5	20	784	8,379	9,183
AAGGAG	12,368	4.99	3,092	620	125	25	5	24	1,091	9,701	10,816
ACC	12,366	4.99	3,092	620	125	25	5	12	1,292	8,807	10,111
AGGCGG	12,127	4.97	3,032	611	123	25	5	24	0	369	393
AAAATC	11,613	4.93	2,904	590	120	25	5	24	28	4,055	4,107
AGCCCC	11,487	4.92	2,872	585	119	25	5	24	256	6,484	6,764
ACTGC	11,423	4.91	2,856	582	119	25	5	20	0	102	122
AAGGGG	11,368	4.91	2,842	580	119	25	5	24	790	8,738	9,552
ACATGC	11,230	4.89	2,808	574	118	24	5	24	358	4,207	4,589
AAAACC	11,135	4.89	2,784	570	117	24	5	24	270	4,056	4,350
AAACAC	11,079	4.88	2,770	568	117	24	5	24	272	4,844	5,140
ACACAG	10,966	4.87	2,742	563	116	24	5	24	582	6,212	6,818
ATC	10,605	4.84	2,652	548	114	24	5	12	1,108	7,418	8,538

ACTCC	10,603	4.84	2,651	548	114	24	5	20	2,298	10,589	12,907
AGGC	10,520	4.83	2,630	545	113	24	5	16	255	3,743	4,014
AAAAGC	10,365	4.82	2,592	539	112	24	5	24	86	4,306	4,416
AAAGGG	10,252	4.81	2,563	534	111	24	5	24	450	6,925	7,399
AGGGGC	9,768	4.76	2,442	514	108	23	5	24	222	4,874	5,120
AAACTG	9,679	4.75	2,420	510	108	23	5	21	0	1,281	1,302
AGATAT	9,677	4.75	2,420	510	108	23	5	24	1,070	7,748	8,842
AGC	9,672	4.75	2,418	510	108	23	5	12	644	5,697	6,353
AAATGG	9,599	4.74	2,400	506	107	23	5	24	4	4,223	4,251
AATT	9,576	4.74	2,394	506	107	23	5	12	554	3,561	4,127
AAAAGT	9,563	4.74	2,391	505	107	23	5	24	34	4,046	4,104
AATATT	9,279	4.71	2,320	493	105	23	5	18	433	4,825	5,276
AAAGAC	9,092	4.69	2,273	485	104	23	5	24	22	2,274	2,320
AGGGC	9,064	4.69	2,266	484	104	22	5	20	372	5,044	5,436
AGCCC	8,789	4.66	2,198	472	102	22	5	20	372	4,926	5,318
AAATTC	8,726	4.65	2,182	469	101	22	5	24	10	3,828	3,862
AAAACT	8,604	4.64	2,151	464	100	22	5	24	26	2,740	2,790
ACCATC	8,512	4.63	2,128	460	100	22	5	24	1,392	8,887	10,303
AACAAT	8,171	4.59	2,043	445	97	22	5	24	164	2,910	3,098
ACACC	7,917	4.56	1,980	434	96	21	5	20	12	1,745	1,777
AAACAT	7,914	4.56	1,979	434	96	21	5	24	52	3,135	3,211
AAGCAG	7,856	4.56	1,964	432	95	21	5	24	40	3,373	3,437
AAATTG	7,803	4.55	1,951	429	95	21	5	24	2	2,766	2,792
AATTAT	7,659	4.53	1,915	423	94	21	5	18	467	5,105	5,590
ACAG	7,641	4.53	1,911	422	94	21	5	16	994	5,149	6,159
ACCC	7,434	4.51	1,859	413	92	21	5	16	1,054	6,729	7,799
AAATC	7,419	4.50	1,855	412	92	21	5	20	24	1,818	1,862
AGGGCC	7,219	4.48	1,805	403	90	21	5	24	204	4,757	4,985
AAGAAT	7,094	4.46	1,774	398	89	20	5	24	22	2,531	2,577
AAAGAT	7,039	4.46	1,760	395	89	20	5	24	24	2,325	2,373
AATGAT	7,008	4.45	1,752	394	89	20	5	24	146	2,769	2,939
AGGCCC	6,993	4.45	1,749	393	89	20	5	24	58	3,422	3,504
AAACC	6,932	4.44	1,733	390	88	20	5	20	298	2,509	2,827
AAAGC	6,709	4.41	1,678	380	87	20	5	20	160	2,826	3,006
AATAG	6,707	4.41	1,677	380	87	20	5	20	356	3,095	3,471
AGCTCC	6,483	4.38	1,621	370	85	20	5	24	60	2,734	2,818
AGCAGG	6,434	4.38	1,609	368	84	20	5	24	72	3,305	3,401
AAGATG	6,406	4.37	1,602	367	84	20	5	24	48	2,587	2,659
ACTC	6,386	4.37	1,597	366	84	20	5	16	657	3,606	4,279

AGGCC	6,385	4.37	1,597	366	84	20	5	20	14	3,007	3,041
AGATGG	6,336	4.36	1,584	363	84	20	5	24	12	1,854	1,890
AATGAG	6,157	4.34	1,540	355	82	19	5	24	18	2,235	2,277
ATCCCC	6,114	4.33	1,529	353	82	19	5	24	75	2,551	2,650
AAATGC	6,036	4.32	1,509	350	81	19	5	24	8	2,025	2,057
AATATG	5,914	4.30	1,479	344	80	19	5	24	12	2,317	2,353
CCCCCG	5,871	4.30	1,468	342	80	19	5	24	358	3,318	3,700
AATTC	5,772	4.28	1,443	337	79	19	5	20	229	2,929	3,178
AACATC	5,722	4.28	1,431	335	79	19	5	24	0	391	415
AATC	5,686	4.27	1,422	333	78	19	5	16	280	1,848	2,144
ACACCC	5,587	4.26	1,397	329	78	19	5	24	226	3,248	3,498
ACAGCC	5,548	4.25	1,387	327	77	19	5	24	22	2,501	2,547
AAAGTG	5,526	4.25	1,382	326	77	19	5	24	4	1,647	1,675
AATTAG	5,376	4.22	1,344	319	76	18	5	24	10	1,687	1,721
AGCC	5,349	4.22	1,338	317	76	18	5	16	1,289	4,253	5,558
AATCTG	5,287	4.21	1,322	314	75	18	5	24	2	654	680
AACATT	5,229	4.20	1,308	312	75	18	5	24	14	1,657	1,695
AAGTGG	5,177	4.19	1,295	309	74	18	5	24	12	1,103	1,139
ACTGCC	5,116	4.18	1,279	306	74	18	5	24	2	1,243	1,269
AACAAG	5,064	4.17	1,266	304	73	18	5	24	26	1,163	1,213
ACTCCC	4,997	4.16	1,250	301	73	18	5	24	16	1,567	1,607
AGAGAT	4,990	4.16	1,248	300	73	18	5	24	279	1,881	2,184
AGAGC	4,911	4.15	1,228	296	72	18	5	20	12	2,073	2,105
CCCCG	4,892	4.14	1,223	296	72	18	5	20	448	3,026	3,494
AATAGT	4,888	4.14	1,222	295	72	18	5	24	94	1,374	1,492
CCG	4,887	4.14	1,222	295	72	18	5	12	664	2,882	3,558
ACAGGG	4,856	4.14	1,214	294	71	18	5	24	56	1,804	1,884
AGATG	4,849	4.14	1,213	294	71	18	5	20	426	2,788	3,234
AATGTG	4,809	4.13	1,203	292	71	18	5	24	0	1,270	1,294
ACACTC	4,716	4.11	1,179	287	70	17	5	24	444	3,268	3,736
AATCAG	4,585	4.09	1,147	281	69	17	5	24	0	1,376	1,400
ACCCTC	4,500	4.08	1,125	277	68	17	5	24	73	1,728	1,825
AAACCC	4,497	4.08	1,125	276	68	17	5	24	2	584	610
ACAGC	4,475	4.07	1,119	275	68	17	5	20	16	1,173	1,209
ACAGGC	4,462	4.07	1,116	275	68	17	5	24	4	738	766
AAGGCC	4,457	4.07	1,115	274	68	17	5	24	20	707	751
AACAGC	4,442	4.07	1,111	274	68	17	5	24	130	1,109	1,263
AGCATC	4,414	4.06	1,104	272	67	17	5	24	148	1,078	1,250
AAAGGC	4,400	4.06	1,100	272	67	17	5	24	0	1,057	1,081

AGAGCC	4,340	4.05	1,085	269	67	17	5	24	6	1,389	1,419
AACCCC	4,299	4.04	1,075	267	66	17	5	24	76	1,653	1,753
ACTCTC	4,262	4.03	1,066	265	66	17	5	24	50	888	962
AGGATG	4,257	4.03	1,065	265	66	17	5	24	67	1,268	1,359
AATGGG	4,248	4.03	1,062	264	66	17	5	24	0	993	1,017
AAATCT	4,158	4.01	1,040	260	65	17	5	24	6	1,118	1,148
AAACT	4,126	4.01	1,032	258	65	17	5	20	94	945	1,059
AACAGG	4,088	4.00	1,022	256	64	16	4	24	0	1,261	1,285
AGCTC	4,029	3.99	1,008	253	64	16	4	20	113	2,197	2,330
AAATCC	3,970	3.98	993	250	63	16	4	24	0	1,507	1,531
AATAC	3,961	3.97	991	250	63	16	4	20	84	1,658	1,762
AATCAT	3,936	3.97	984	248	63	16	4	24	22	1,680	1,726
AAGAGC	3,919	3.96	980	248	63	16	4	24	18	1,307	1,349
AACCTC	3,911	3.96	978	247	63	16	4	24	0	46	70
AACC	3,903	3.96	976	247	63	16	4	16	356	2,086	2,458
AATTAC	3,868	3.95	967	245	62	16	4	24	4	1,544	1,572
AATGAC	3,845	3.95	962	244	62	16	4	24	18	1,628	1,670
AATATC	3,840	3.95	960	244	62	16	4	24	0	1,314	1,338
AATGT	3,825	3.95	957	243	62	16	4	20	76	2,323	2,419
AGGCCG	3,794	3.94	949	241	62	16	4	24	6	270	300
AGCTGC	3,785	3.94	947	241	62	16	4	18	78	2,587	2,683
AAGCTG	3,715	3.92	929	237	61	16	4	24	2	1,273	1,299
AAAGCC	3,695	3.92	924	236	61	16	4	24	0	1,197	1,221
AAGATC	3,629	3.90	908	233	60	16	4	24	12	541	577
AAGGTG	3,624	3.90	906	233	60	16	4	24	6	1,449	1,479
AAAGTT	3,623	3.90	906	233	60	16	4	20	4	1,241	1,265
AACATG	3,609	3.90	903	232	60	16	4	24	0	908	932
AAACTC	3,609	3.90	903	232	60	16	4	24	0	848	872
ACAGTG	3,551	3.89	888	229	59	16	4	24	2	1,347	1,373
AATTCC	3,530	3.88	883	228	59	16	4	24	2	865	891
AAACTT	3,522	3.88	881	227	59	16	4	19	0	845	864
ACCAGC	3,522	3.88	881	227	59	16	4	24	28	1,518	1,570
ACTGAG	3,481	3.87	871	225	59	15	4	24	0	637	661
AACAC	3,464	3.87	866	224	58	15	4	20	68	1,269	1,357
AAAGTC	3,439	3.86	860	223	58	15	4	23	0	1,112	1,135
ACAGAT	3,434	3.86	859	223	58	15	4	24	59	1,469	1,552
AGCCTG	3,425	3.86	857	222	58	15	4	24	0	947	971
AACTTG	3,406	3.86	852	221	58	15	4	24	0	767	791
ACCTGC	3,393	3.85	849	221	58	15	4	24	10	1,564	1,598

AAAGT	3,371	3.85	843	220	57	15	4	20	34	1,482	1,536
AACAT	3,320	3.84	830	217	57	15	4	20	84	1,608	1,712
AAGTCC	3,300	3.83	825	216	57	15	4	24	58	515	597
ATCCC	3,283	3.83	821	215	57	15	4	20	104	1,948	2,072
AAGC	3,268	3.82	817	214	56	15	4	16	391	2,326	2,733
C	3,219	3.81	805	212	56	15	4	4	446	1,780	2,230
AAGTAG	3,204	3.81	801	211	56	15	4	24	0	876	900
ACCCTG	3,160	3.80	790	209	55	15	4	24	70	1,666	1,760
AGATGC	3,150	3.80	788	208	55	15	4	24	0	1,141	1,165
ACTCTG	3,139	3.79	785	207	55	15	4	23	20	956	999
AAGCC	3,056	3.77	764	203	54	15	4	20	0	818	838
ACATCC	3,011	3.76	753	201	54	15	4	24	26	1,038	1,088
ACAGG	2,986	3.75	747	199	53	15	4	20	26	1,491	1,537
ACATC	2,982	3.75	746	199	53	15	4	20	0	741	761
ACACGC	2,982	3.75	746	199	53	15	4	24	639	3,102	3,765
ATATC	2,972	3.75	743	199	53	15	4	20	8	1,141	1,169
ACTAT	2,970	3.75	743	198	53	15	4	20	68	1,805	1,893
AAGGTC	2,952	3.75	738	197	53	15	4	24	12	770	806
AAGGAC	2,947	3.75	737	197	53	15	4	24	0	875	899
AAGGGC	2,937	3.74	735	197	53	15	4	24	22	927	973
AATACT	2,931	3.74	733	196	53	14	4	24	26	1,080	1,130
AAGTG	2,881	3.73	721	194	52	14	4	20	0	664	684
AAGAC	2,880	3.73	720	194	52	14	4	20	30	832	882
AACTG	2,875	3.73	719	193	52	14	4	20	2	897	919
AATGC	2,858	3.72	715	192	52	14	4	20	190	1,475	1,685
AACCAG	2,858	3.72	715	192	52	14	4	24	0	327	351
AACTGG	2,853	3.72	714	192	52	14	4	24	0	478	502
AAGATT	2,851	3.72	713	192	52	14	4	24	1	919	944
AATAGG	2,824	3.71	706	191	52	14	4	24	0	524	548
AAGGC	2,778	3.70	695	188	51	14	4	20	136	1,345	1,501
ACATGG	2,777	3.70	695	188	51	14	4	23	0	303	326
ACTCAG	2,744	3.69	686	186	51	14	4	24	0	383	407
ACATAG	2,730	3.69	683	186	51	14	4	24	50	783	857
ACCTC	2,700	3.68	675	184	50	14	4	20	0	866	886
AAAGGT	2,604	3.65	651	179	49	14	4	24	2	448	474
ACCCAG	2,596	3.65	649	178	49	14	4	24	12	766	802
AACTTC	2,589	3.65	648	178	49	14	4	20	0	236	256
ACAGTC	2,588	3.65	647	178	49	14	4	24	2	463	489
AACACT	2,588	3.65	647	178	49	14	4	24	62	712	798

AACTCC	2,580	3.65	645	177	49	14	4	24	0	181	205
AACAGT	2,578	3.65	645	177	49	14	4	24	0	593	617
AATGTC	2,574	3.65	644	177	49	14	4	24	0	311	335
AGCATG	2,529	3.63	633	175	48	14	4	24	0	74	98
AATAGC	2,521	3.63	631	174	48	14	4	24	0	563	587
ACTGGG	2,489	3.62	623	172	48	14	4	24	5	492	521
CCCCGG	2,486	3.62	622	172	48	14	4	24	142	1,685	1,851
AGATCG	2,473	3.62	619	171	48	14	4	14	0	7	21
ATGCCC	2,445	3.61	612	170	47	14	4	24	0	262	286
AATCAC	2,425	3.60	607	169	47	13	4	22	0	264	286
ATATGC	2,413	3.60	604	168	47	13	4	18	146	1,319	1,483
AAGACC	2,406	3.60	602	168	47	13	4	20	2	140	162
AACCCT	2,399	3.59	600	167	47	13	4	24	466	1,958	2,448
AAAGCT	2,394	3.59	599	167	47	13	4	24	12	322	358
ACCTCT	2,380	3.59	595	166	47	13	4	24	0	232	256
ACCAGG	2,372	3.59	593	166	47	13	4	24	4	581	609
AACACC	2,362	3.58	591	165	46	13	4	24	24	647	695
AACTAT	2,339	3.58	585	164	46	13	4	23	9	530	562
ACATGT	2,337	3.58	585	164	46	13	4	18	106	1,127	1,251
AATCT	2,337	3.58	585	164	46	13	4	20	4	675	699
ATGC	2,328	3.57	582	163	46	13	4	12	319	1,978	2,309
ACACTG	2,296	3.56	574	162	46	13	4	22	0	71	93
ACT	2,259	3.55	565	160	45	13	4	12	314	1,208	1,534
AATGGC	2,229	3.54	558	158	45	13	4	24	2	248	274
AATCTC	2,223	3.54	556	158	45	13	4	24	0	181	205
ACCT	2,210	3.54	553	157	45	13	4	16	395	1,661	2,072
AACTGT	2,201	3.53	551	156	45	13	4	24	4	265	293
ACTATC	2,199	3.53	550	156	45	13	4	24	76	873	973
AAGAGT	2,153	3.52	539	154	44	13	4	19	0	54	73
AAGCCC	2,152	3.52	538	153	44	13	4	24	0	275	299
ATATCC	2,139	3.51	535	153	44	13	4	24	55	560	639
AATGGT	2,124	3.51	531	152	44	13	4	24	8	523	555
AAACCT	2,086	3.50	522	150	43	13	4	24	0	112	136
AACCAC	2,076	3.49	519	149	43	13	4	24	23	417	464
AAGAT	2,044	3.48	511	147	43	13	4	20	16	481	517
ACTGGC	2,043	3.48	511	147	43	13	4	24	2	211	237
ACCCCT	2,036	3.48	509	147	43	13	4	24	12	566	602
AAGGAT	1,989	3.46	498	144	42	12	4	24	10	278	312
ACCTGG	1,926	3.44	482	140	41	12	4	24	0	231	255

AACTC	1,924	3.44	481	140	41	12	4	20	3	190	213
AATGCT	1,908	3.43	477	139	41	12	4	24	0	209	233
ACTGCT	1,897	3.43	475	139	41	12	4	24	40	429	493
AATTGC	1,895	3.43	474	139	41	12	4	14	0	14	28
ACTG	1,894	3.43	474	139	41	12	4	16	78	556	650
AAGCAC	1,889	3.43	473	138	41	12	4	24	0	105	129
ACATCT	1,879	3.42	470	138	41	12	4	24	12	292	328
ACAGT	1,878	3.42	470	138	41	12	4	20	10	428	458
ATGCC	1,876	3.42	469	138	41	12	4	20	40	266	326
CCCGG	1,875	3.42	469	138	41	12	4	20	167	937	1,124
AAGTAT	1,871	3.42	468	137	40	12	4	22	0	330	352
ACGCCC	1,866	3.42	467	137	40	12	4	20	12	107	139
AATCC	1,836	3.41	459	135	40	12	4	20	89	433	542
AACCTG	1,834	3.41	459	135	40	12	4	24	0	45	69
AACCC	1,832	3.41	458	135	40	12	4	20	4	475	499
AATACC	1,827	3.40	457	135	40	12	4	24	0	31	55
ACCACT	1,788	3.39	447	132	39	12	4	24	123	572	719
ACTGAT	1,760	3.38	440	131	39	12	4	16	0	33	49
ACCAG	1,739	3.37	435	130	39	12	4	20	38	250	308
ACATG	1,738	3.37	435	129	39	12	4	20	0	144	164
AAGCAT	1,738	3.37	435	129	39	12	4	15	0	41	56
AAGCTC	1,706	3.36	427	128	38	12	4	24	0	3	27
AATCCT	1,697	3.35	425	127	38	12	4	24	0	107	131
AGATCC	1,670	3.34	418	125	38	12	4	24	0	85	109
CCCG	1,668	3.34	417	125	38	12	4	16	200	1,063	1,279
AAGT	1,660	3.34	415	125	38	12	4	16	62	370	448
AATCCC	1,647	3.33	412	124	38	12	4	18	0	49	67
ACAGCT	1,638	3.33	410	123	37	12	4	24	0	50	74
AATGCC	1,619	3.32	405	122	37	12	4	22	0	47	69
AAGTGT	1,608	3.32	402	122	37	12	4	18	0	37	55
AACTGC	1,606	3.32	402	122	37	12	4	18	0	7	25
AACTAC	1,599	3.31	400	121	37	11	4	17	0	44	61
ACCATG	1,575	3.30	394	120	37	11	4	22	0	34	56
AGATC	1,573	3.30	394	120	37	11	4	20	0	27	47
ACTCCT	1,568	3.30	392	119	36	11	4	24	0	63	87
ACTGG	1,550	3.29	388	118	36	11	4	20	47	284	351
ACCTG	1,543	3.29	386	118	36	11	4	16	0	46	62
AAGTC	1,543	3.29	386	118	36	11	4	19	12	63	94
AACTCT	1,540	3.29	385	118	36	11	4	24	0	53	77

AAGTAC	1,490	3.27	373	114	35	11	4	18	0	6	24
AGGCAT	1,480	3.26	370	114	35	11	4	18	0	3	21
AGGATC	1,465	3.26	367	113	35	11	4	21	0	21	42
AGCAT	1,450	3.25	363	112	35	11	4	20	22	299	341
ACTATG	1,448	3.25	362	112	35	11	4	21	0	35	56
AACCAT	1,432	3.24	358	111	35	11	4	24	0	23	47
ACACT	1,405	3.23	352	109	34	11	4	20	21	185	226
AAGTCT	1,356	3.21	339	106	33	11	4	19	0	10	29
AACCTT	1,348	3.20	337	106	33	11	4	19	0	5	24
AAGACT	1,345	3.20	337	106	33	11	4	22	0	14	36
AGCCT	1,326	3.19	332	104	33	11	4	20	56	248	324
ACAGGT	1,296	3.18	324	102	33	11	4	24	0	55	79
ACTCAT	1,271	3.17	318	101	32	11	4	24	0	45	69
ATGGCC	1,252	3.16	313	100	32	10	4	18	10	236	264
AAGCT	1,234	3.15	309	99	32	10	4	19	10	111	140
AAGGT	1,228	3.14	307	98	32	10	4	20	0	62	82
AGAGCT	1,221	3.14	306	98	31	10	4	24	0	31	55
ACTCT	1,218	3.14	305	98	31	10	4	20	8	138	166
ACCTAT	1,191	3.12	298	96	31	10	4	24	32	197	253
AGCGGC	1,175	3.12	294	95	31	10	4	24	112	407	543
AGGGAT	1,125	3.09	282	92	30	10	4	24	0	26	50
ACCAT	1,119	3.09	280	91	30	10	4	20	15	146	181
AGGCGC	1,113	3.08	279	91	30	10	4	24	20	262	306
ACCCT	1,107	3.08	277	90	30	10	4	20	4	165	189
CCCGCG	1,096	3.07	274	90	30	10	4	24	44	570	638
AGGAT	1,073	3.06	269	88	29	10	4	20	15	218	253
AGCCAT	1,071	3.06	268	88	29	10	4	24	0	37	61
ATCATG	1,055	3.05	264	87	29	10	4	18	16	113	147
ACTAG	1,042	3.04	261	86	29	10	4	20	3	55	78
AGCCCT	1,018	3.03	255	85	28	10	4	24	8	144	176
ACCGCC	1,017	3.03	255	84	28	10	4	24	98	352	474
AACCT	1,014	3.03	254	84	28	10	4	20	2	59	81
AAGGGT	999	3.02	250	83	28	10	4	24	0	29	53
AACT	998	3.02	250	83	28	10	4	16	56	168	240
ACACCT	980	3.00	245	82	28	10	4	24	1	91	116
AAGGCT	961	2.99	241	81	27	9	3	18	0	6	24
ACCCAT	960	2.99	240	81	27	9	3	24	0	40	64
AGCCGC	957	2.99	240	80	27	9	3	24	21	551	596
CG	935	2.98	234	79	27	9	3	6	361	1,029	1,396

ACCGAG	918	2.97	230	78	27	9	3	15	0	1	16
ACTAGG	914	2.96	229	78	27	9	3	21	0	15	36
AAGCCT	897	2.95	225	76	26	9	3	16	2	13	31
AGCGGG	893	2.95	224	76	26	9	3	24	20	543	587
ACCTAG	887	2.95	222	76	26	9	3	19	0	15	34
CCGCG	865	2.93	217	74	26	9	3	20	66	656	742
ACGC	775	2.87	194	68	24	9	3	16	313	955	1,284
AGCCGG	739	2.84	185	66	23	9	3	24	17	380	421
AGCT	738	2.84	185	65	23	9	3	12	90	590	692
ACGGGG	732	2.83	183	65	23	9	3	24	24	344	392
AGATCT	732	2.83	183	65	23	9	3	18	8	121	147
AGAGCG	702	2.81	176	63	23	8	3	24	278	377	679
AGCTAT	691	2.80	173	62	23	8	3	18	4	153	175
AGGGCG	618	2.74	155	57	21	8	3	24	4	301	329
CCCGGG	603	2.73	151	56	21	8	3	18	8	410	436
AGCGCC	584	2.71	146	54	20	8	3	23	2	89	114
AAGCTT	583	2.71	146	54	20	8	3	15	0	94	109
ACGCGC	583	2.71	146	54	20	8	3	24	162	533	719
ACGGAG	578	2.70	145	54	20	8	3	24	69	223	316
ACGGCC	576	2.70	144	54	20	8	3	24	0	107	131
ACCCCG	542	2.67	136	51	20	8	3	24	11	194	229
AAAACG	532	2.66	133	51	19	8	3	16	0	48	64
AGGCCT	503	2.63	126	48	19	7	3	13	0	20	33
CCGCGG	501	2.63	126	48	19	7	3	16	14	238	268
AGCCCG	486	2.61	122	47	18	7	3	24	23	170	217
ACGTGC	468	2.59	117	46	18	7	3	17	20	450	487
ACGAGG	439	2.56	110	43	17	7	3	24	26	183	233
ACGTCC	417	2.53	105	42	17	7	3	24	0	21	45
ACTAGT	410	2.52	103	41	17	7	3	17	0	31	48
AGGCG	405	2.52	102	41	16	7	3	20	35	145	200
AGCG	405	2.52	102	41	16	7	3	16	104	241	361
ACCCGC	399	2.51	100	40	16	7	3	24	8	72	104
AAGCGG	385	2.49	97	39	16	7	3	22	1	23	46
CCGGCG	382	2.49	96	39	16	7	3	17	11	131	159
AACGG	381	2.49	96	39	16	7	3	20	492	747	1,259
ACCGGC	365	2.47	92	37	16	7	3	13	0	5	18
ACACGG	362	2.46	91	37	15	7	3	15	0	20	35
AGCGG	361	2.46	91	37	15	7	3	20	2	125	147
CCGG	358	2.46	90	37	15	7	3	12	11	200	223

ACGATG	342	2.43	86	36	15	6	3	16	45	74	135
ACGTAT	336	2.43	84	35	15	6	3	18	426	532	976
AGCGC	322	2.41	81	34	14	6	3	20	0	39	59
AGCCG	308	2.38	77	33	14	6	3	20	1	61	82
ACGGGC	286	2.35	72	31	13	6	3	14	0	7	21
ACCGGG	279	2.34	70	30	13	6	3	22	2	28	52
ACACGT	271	2.32	68	30	13	6	3	23	3	69	95
ACGG	270	2.32	68	30	13	6	3	16	240	297	553
AAGGCG	267	2.32	67	29	13	6	3	20	0	6	26
ACGGG	252	2.29	63	28	13	6	3	20	10	132	162
ACCGTC	251	2.29	63	28	12	6	3	18	8	54	80
AATCG	245	2.28	62	27	12	6	3	32	346	474	852
ACCCGG	243	2.27	61	27	12	6	3	24	1	7	32
ACGCC	236	2.26	59	27	12	6	3	20	0	31	51
ATCGCC	235	2.26	59	27	12	6	3	14	0	3	17
ACGCTC	227	2.24	57	26	12	6	3	18	0	20	38
AAGACG	225	2.24	57	26	12	6	3	21	24	83	128
ACAGCG	205	2.20	52	24	11	5	3	14	0	2	16
ACGCAG	204	2.20	51	24	11	5	3	15	0	5	20
AAAGCG	198	2.18	50	23	11	5	3	14	4	5	23
AACGGG	184	2.15	46	22	10	5	3	13	1	1	15
AGCGCG	183	2.15	46	22	10	5	3	18	0	10	28
AAACG	173	2.12	44	21	10	5	3	19	17	30	66
ACCGC	168	2.11	42	20	10	5	3	20	0	6	26
ACGGCG	155	2.08	39	19	9	5	3	21	26	57	104
AACGAC	140	2.04	35	18	9	5	3	16	39	5	60
ACG	132	2.01	33	17	9	5	3	12	102	81	195
AACCGG	126	1.99	32	16	8	4	2	18	0	6	24
ACGGC	122	1.98	31	16	8	4	2	15	0	6	21
AAATCG	121	1.98	31	16	8	4	2	16	0	2	18
ATCCCG	120	1.97	30	16	8	4	2	18	31	26	75
AACCCG	119	1.97	30	16	8	4	2	24	55	172	251
ACTCGC	118	1.97	30	15	8	4	2	14	0	3	17
ACCCG	112	1.95	28	15	8	4	2	15	1	50	66
AAGTCG	111	1.94	28	15	8	4	2	13	0	1	14
AACG	108	1.93	27	14	8	4	2	16	146	152	314

Table S2. Summary statistics of perfect SSR loci in hg38 for each SSR family

Motif	Motif Reverse Complement	# Nucleotides in Motif	%AT	%GC	Total Loci	Total Nucleotides	Average Locus Length	Longest Locus
A	T	1	100%	0%	703,012	12,047,125	17.136	90
AC	TG	2	50%	50%	170,729	3,797,349	22.242	161
AAAAAT	TTTTTA	2	100%	0%	155,162	2,141,311	13.8	49
AAAT	TTTA	2	100%	0%	142,663	2,529,416	17.73	72
AAAAT	TTTTA	2	100%	0%	117,851	1,808,375	15.345	388
AAAAAG	TTTTTC	2	83%	17%	116,016	1,618,531	13.951	80
AAAAC	TTTTG	2	80%	20%	110,158	1,799,309	16.334	71
AAAAG	TTTTC	2	80%	20%	108,341	1,583,986	14.62	128
ACCTCC	TGGAGG	3	33%	67%	107,333	1,301,510	12.126	43
AAAAAC	TTTTTG	2	83%	17%	102,575	1,515,515	14.775	56
AT	TA	2	100%	0%	93,619	1,725,227	18.428	600
AAAC	TTTG	2	75%	25%	91,575	1,542,896	16.848	52
AAAG	TTTC	2	75%	25%	91,311	1,764,832	19.328	334
AG	TC	2	50%	50%	76,587	1,314,024	17.157	166
AAT	TTA	2	100%	0%	69,920	1,140,522	16.312	79
AAGG	TTCC	2	50%	50%	49,946	949,977	19.02	303
AATGG	TTACC	3	60%	40%	47,818	798,469	16.698	341
AAC	TTG	2	67%	33%	47,384	768,242	16.213	56
AAAATT	TTTTAA	2	100%	0%	43,211	559,469	12.947	36
AAAGAG	TTTCTC	2	67%	33%	39,475	518,255	13.129	102
AGGG	TCCC	2	25%	75%	35,707	544,696	15.255	100
AAATAT	TTTATA	2	100%	0%	35,309	459,593	13.016	94
AATG	TTAC	3	75%	25%	34,624	513,584	14.833	100
ACATAT	TGTATA	3	83%	17%	29,304	404,082	13.789	277
AAAATG	TTTTAC	3	83%	17%	26,628	338,060	12.696	27
AGCCTC	TCGGAG	4	33%	67%	25,875	329,503	12.734	39
AACAG	TTGTC	3	60%	40%	25,662	314,741	12.265	87
ACCCCC	TGGGGG	2	17%	83%	25,539	337,075	13.198	54
AAAAGG	TTTTCC	2	67%	33%	25,147	318,189	12.653	60
AACTAG	TTGATC	4	67%	33%	23,572	284,908	12.087	29
AGG	TCC	2	33%	67%	22,900	329,949	14.408	218
ACACAT	TGTGTA	3	67%	33%	22,520	302,336	13.425	76
AAATT	TTTAA	2	100%	0%	22,351	299,546	13.402	77
AAATGT	TTTACA	3	83%	17%	21,479	264,529	12.316	26

AAAGG	TTTCC	2	60%	40%	21,308	298,571	14.012	193
ATCC	TAGG	3	50%	50%	20,996	389,884	18.569	533
AGGGG	TCCCC	2	20%	80%	20,911	302,366	14.46	164
AGAGGC	TCTCCG	3	33%	67%	20,877	267,020	12.79	138
AGAGGG	TCTCCC	2	33%	67%	20,520	300,374	14.638	431
ACCAGT	TGGTCA	4	50%	50%	19,786	238,483	12.053	38
AGAT	TCTA	3	75%	25%	19,113	479,956	25.111	191
AAATG	TTTAC	3	80%	20%	18,476	246,700	13.352	78
AAGAGG	TTCTCC	2	50%	50%	18,182	241,094	13.26	75
AAG	TTC	2	67%	33%	18,104	301,333	16.645	187
AAACAG	TTTGTC	3	67%	33%	17,876	224,596	12.564	30
ACAT	TGTA	3	75%	25%	17,609	268,447	15.245	87
AGGGGG	TCCCCC	2	17%	83%	17,383	233,042	13.406	77
AATAT	TTATA	2	100%	0%	17,310	238,497	13.778	200
ACCCC	TGGGG	2	20%	80%	16,814	220,090	13.09	52
ACAGAG	TGTCTC	3	50%	50%	16,715	228,843	13.691	83
AAGGG	TTCCC	2	40%	60%	15,596	225,246	14.443	306
AAATAG	TTTATC	3	83%	17%	15,464	195,129	12.618	30
AAATTT	TTTAAA	2	100%	0%	14,966	192,709	12.876	31
AAGAG	TTCTC	2	60%	40%	14,725	198,113	13.454	321
AAATAC	TTTATG	3	83%	17%	14,101	177,079	12.558	59
AGAGG	TCTCC	2	40%	60%	13,813	187,615	13.582	231
AAGGAG	TTCCTC	2	50%	50%	13,743	188,605	13.724	146
ACC	TGG	2	33%	67%	13,741	202,054	14.704	632
AGGCGG	TCCGCC	3	17%	83%	13,475	166,889	12.385	58
AAAATC	TTTTAG	3	83%	17%	12,904	162,215	12.571	28
AGCCCC	TCGGGG	3	17%	83%	12,764	166,563	13.049	69
ACTGC	TGACG	4	40%	60%	12,693	155,431	12.245	44
AAGGGG	TTCCCC	2	33%	67%	12,632	169,127	13.389	82
ACATGC	TGTACG	4	50%	50%	12,478	156,093	12.509	36
AAAACC	TTTTGG	2	67%	33%	12,373	162,403	13.126	58
AAACAC	TTTGTG	2	67%	33%	12,310	156,413	12.706	35
ACACAG	TGTGTC	3	50%	50%	12,185	157,486	12.925	64
ATC	TAG	3	67%	33%	11,784	185,414	15.734	370
ACTCC	TGAGG	3	40%	60%	11,782	168,774	14.325	510
AGGC	TCCG	3	25%	75%	11,689	153,938	13.169	77
AAAAGC	TTTTCG	3	67%	33%	11,517	146,840	12.75	29
AAAGGG	TTTCCC	2	50%	50%	11,392	146,498	12.86	166
AGGGGC	TCCCCG	3	17%	83%	10,854	140,375	12.933	68

AAACTG	TTTGAC	4	67%	33%	10,755	132,460	12.316	27
AGATAT	TCTATA	3	83%	17%	10,753	162,823	15.142	163
AGC	TCG	3	33%	67%	10,747	154,768	14.401	79
AAATGG	TTTACC	3	67%	33%	10,666	133,989	12.562	30
AATT	TTAA	2	100%	0%	10,641	156,658	14.722	38
AAAAGT	TTTTCA	3	83%	17%	10,626	133,545	12.568	26
AATATT	TTATAA	2	100%	0%	10,311	132,147	12.816	30
AAAGAC	TTTCTG	3	67%	33%	10,103	127,646	12.634	32
AGGGC	TCCCG	3	20%	80%	10,072	131,502	13.056	54
AGCCC	TCGGG	3	20%	80%	9,766	128,362	13.144	53
AAATTC	TTTAAG	3	83%	17%	9,696	121,823	12.564	26
AAAACT	TTTTGA	3	83%	17%	9,560	121,265	12.685	30
ACCATC	TGGTAG	3	50%	50%	9,458	135,031	14.277	72
AACAAT	TTGTTA	3	83%	17%	9,079	116,652	12.849	41
ACACC	TGTGG	2	40%	60%	8,797	118,846	13.51	59
AAACAT	TTTGTA	3	83%	17%	8,794	110,721	12.591	36
AAGCAG	TTCGTC	3	50%	50%	8,729	110,418	12.65	74
AAATTG	TTTAAC	3	83%	17%	8,670	108,340	12.496	29
AATTAT	TTAATA	2	100%	0%	8,511	108,451	12.742	37
ACAG	TGTC	3	50%	50%	8,490	128,095	15.088	51
ACCC	TGGG	2	25%	75%	8,260	112,058	13.566	47
AAATC	TTTAG	3	80%	20%	8,244	105,475	12.794	53
AGGGCC	TCCCGG	3	17%	83%	8,022	104,276	12.999	41
AAGAAT	TTCTTA	3	83%	17%	7,883	99,344	12.602	53
AAAGAT	TTTCTA	3	83%	17%	7,822	98,009	12.53	26
AATGAT	TTACTA	3	83%	17%	7,787	99,764	12.812	54
AGGCCC	TCCGGG	3	17%	83%	7,771	98,800	12.714	38
AAACC	TTTGG	2	60%	40%	7,703	107,196	13.916	220
AAAGC	TTTCG	3	60%	40%	7,455	95,827	12.854	94
AATAG	TTATC	3	80%	20%	7,453	101,875	13.669	191
AGCTCC	TCGAGG	4	33%	67%	7,204	92,611	12.855	55
AGCAGG	TCGTCC	3	33%	67%	7,149	90,733	12.692	33
AAGATG	TTCTAC	3	67%	33%	7,118	89,959	12.638	32
ACTC	TGAG	3	50%	50%	7,096	95,194	13.415	59
AGGCC	TCCGG	3	20%	80%	7,095	90,153	12.707	31
AGATGG	TCTACC	3	50%	50%	7,041	89,172	12.665	41
AATGAG	TTACTC	3	67%	33%	6,842	86,107	12.585	25
ATCCCC	TAGGGG	3	33%	67%	6,794	86,157	12.681	56
AAATGC	TTTACG	4	67%	33%	6,707	84,429	12.588	30

AATATG	TTATAC	3	83%	17%	6,572	82,943	12.621	30
CCCCCG	GGGGGC	2	0%	100%	6,524	87,198	13.366	93
AATTC	TTAAG	3	80%	20%	6,414	83,553	13.027	73
AACATC	TTGTAG	3	67%	33%	6,358	83,402	13.118	40
AATC	TTAG	3	75%	25%	6,318	96,356	15.251	44
ACACCC	TGTGGG	2	33%	67%	6,208	81,695	13.16	87
ACAGCC	TGTCGG	3	33%	67%	6,165	78,146	12.676	45
AAAGTG	TTTCAC	3	67%	33%	6,141	76,852	12.515	35
AATTAG	TTAATC	3	83%	17%	5,974	75,211	12.59	27
AGCC	TCGG	3	25%	75%	5,944	82,949	13.955	47
AATCTG	TTAGAC	4	67%	33%	5,875	72,906	12.41	27
AACATT	TTGTAA	3	83%	17%	5,810	72,661	12.506	28
AAGTGG	TTCACC	3	50%	50%	5,753	73,378	12.755	28
ACTGCC	TGACGG	4	33%	67%	5,685	72,586	12.768	31
AACAAG	TTGTTC	3	67%	33%	5,627	71,384	12.686	68
ACTCCC	TGAGGG	3	33%	67%	5,553	69,932	12.594	39
AGAGAT	TCTCTA	3	67%	33%	5,545	73,691	13.29	113
AGAGC	TCTCG	3	40%	60%	5,457	68,967	12.638	43
CCCCG	GGGGC	2	0%	100%	5,436	74,973	13.792	65
AATAGT	TTATCA	3	83%	17%	5,432	69,322	12.762	35
CCG	GGC	2	0%	100%	5,430	85,177	15.686	71
ACAGGG	TGTCCC	3	33%	67%	5,396	68,823	12.754	53
AGATG	TCTAC	3	60%	40%	5,388	70,998	13.177	142
AATGTG	TTACAC	3	67%	33%	5,344	66,992	12.536	25
ACACTC	TGTGAG	3	50%	50%	5,240	68,933	13.155	39
AATCAG	TTAGTC	4	67%	33%	5,095	63,960	12.553	25
ACCCTC	TGGGAG	3	33%	67%	5,001	64,038	12.805	53
AAACCC	TTTGGG	2	50%	50%	4,997	64,682	12.944	39
ACAGC	TGTCG	3	40%	60%	4,973	64,233	12.916	65
ACAGGC	TGTCCG	3	33%	67%	4,958	62,153	12.536	32
AAGGCC	TTCCGG	3	33%	67%	4,953	62,232	12.565	29
AACAGC	TTGTCG	3	50%	50%	4,936	63,220	12.808	41
AGCATC	TCGTAG	4	50%	50%	4,905	63,645	12.976	53
AAAGGC	TTTCCG	3	50%	50%	4,889	61,247	12.528	28
AGAGCC	TCTCGG	3	33%	67%	4,823	61,071	12.662	34
AACCCC	TTGGGG	2	33%	67%	4,777	61,128	12.796	52
ACTCTC	TGAGAG	3	50%	50%	4,736	59,704	12.606	29
AGGATG	TCCTAC	3	50%	50%	4,731	60,417	12.77	30
AATGGG	TTACCC	3	50%	50%	4,721	59,125	12.524	30

AAATCT	TTTAGA	3	83%	17%	4,620	58,354	12.631	25
AAACT	TTTGA	3	80%	20%	4,585	60,291	13.15	91
AACAGG	TTGTCC	3	50%	50%	4,543	56,770	12.496	28
AGCTC	TCGAG	4	40%	60%	4,477	57,445	12.831	33
AAATCC	TTTAGG	3	67%	33%	4,412	55,300	12.534	35
AATAC	TTATG	3	80%	20%	4,402	60,868	13.827	137
AATCAT	TTAGTA	3	83%	17%	4,374	56,327	12.878	46
AAGAGC	TTCTCG	3	50%	50%	4,355	55,209	12.677	24
AACCTC	TTGGAG	3	50%	50%	4,346	55,208	12.703	24
AACC	TTGG	2	50%	50%	4,337	64,981	14.983	61
AATTAC	TTAATG	3	83%	17%	4,298	54,043	12.574	50
AATGAC	TTACTG	4	67%	33%	4,273	53,899	12.614	32
AATATC	TTATAG	3	83%	17%	4,267	54,059	12.669	30
AATGT	TTACA	3	80%	20%	4,250	54,879	12.913	69
AGGCCG	TCCGGC	3	17%	83%	4,216	53,463	12.681	26
AGCTGC	TCGACG	4	33%	67%	4,206	53,108	12.627	24
AAGCTG	TTCGAC	4	50%	50%	4,128	51,701	12.524	24
AAAGCC	TTTCGG	3	50%	50%	4,106	51,382	12.514	31
AAGATC	TTCTAG	4	67%	33%	4,033	49,886	12.369	20
AAGGTG	TTCCAC	3	50%	50%	4,027	50,500	12.54	34
AAAGTT	TTTCAA	3	83%	17%	4,026	50,149	12.456	22
AACATG	TTGTAC	4	67%	33%	4,011	50,151	12.503	62
AAACTC	TTTGAG	3	67%	33%	4,010	50,293	12.542	29
ACAGTG	TGTCAC	4	50%	50%	3,946	49,511	12.547	30
AATTCC	TTAAGG	3	67%	33%	3,923	48,978	12.485	22
AAACTT	TTTGAA	3	83%	17%	3,914	48,570	12.409	28
ACCAGC	TGGTCG	3	33%	67%	3,914	49,515	12.651	38
ACTGAG	TGACTC	4	50%	50%	3,868	48,251	12.474	26
AACAC	TTGTG	2	60%	40%	3,849	49,638	12.896	61
AAAGTC	TTTCAG	4	67%	33%	3,822	47,782	12.502	26
ACAGAT	TGTCTA	4	67%	33%	3,816	48,873	12.807	47
AGCCTG	TCGGAC	4	33%	67%	3,806	47,501	12.481	27
AACTTG	TTGAAC	4	67%	33%	3,785	47,619	12.581	23
ACCTGC	TGGACG	4	33%	67%	3,771	47,577	12.617	43
AAAGT	TTTCA	3	80%	20%	3,746	47,233	12.609	63
AACAT	TTGTA	3	80%	20%	3,689	53,841	14.595	112
AAGTCC	TTCAGG	4	50%	50%	3,667	47,870	13.054	28
ATCCC	TAGGG	3	40%	60%	3,648	48,718	13.355	79
AAGC	TTCG	3	50%	50%	3,632	52,656	14.498	63

C	G	1	0%	100%	3,577	48,786	13.639	81
AAGTAG	TTCATC	3	67%	33%	3,561	44,563	12.514	29
ACCCTG	TGGGAC	4	33%	67%	3,512	45,964	13.088	72
AGATGC	TCTACG	4	50%	50%	3,500	43,868	12.534	26
ACTCTG	TGAGAC	4	50%	50%	3,488	43,795	12.556	48
AAGCC	TTCGG	3	40%	60%	3,396	42,933	12.642	28
ACATCC	TGTAGG	3	50%	50%	3,346	42,364	12.661	55
ACAGG	TGTCC	3	40%	60%	3,318	42,970	12.951	62
ACACGC	TGTGCG	3	33%	67%	3,314	43,815	13.221	36
ACATC	TGTAG	3	60%	40%	3,314	42,178	12.727	82
ATATC	TATAG	3	80%	20%	3,303	43,129	13.058	64
ACTAT	TGATA	3	80%	20%	3,300	44,928	13.615	83
AAGGTC	TTCCAG	4	50%	50%	3,280	41,114	12.535	22
AAGGAC	TTCCTG	3	50%	50%	3,275	41,090	12.547	23
AAGGGC	TTCCCG	3	33%	67%	3,264	41,519	12.72	29
AATACT	TTATGA	3	83%	17%	3,257	41,959	12.883	58
AAGTG	TTCAC	3	60%	40%	3,202	39,922	12.468	25
AAGAC	TTCTG	3	60%	40%	3,201	40,957	12.795	68
AACTG	TTGAC	4	60%	40%	3,195	40,224	12.59	29
AACCAG	TTGGTC	3	50%	50%	3,176	39,620	12.475	23
AATGC	TTACG	4	60%	40%	3,176	40,373	12.712	74
AACTGG	TTGACC	4	50%	50%	3,171	39,678	12.513	23
AAGATT	TTCTAA	3	83%	17%	3,168	39,663	12.52	26
AATAGG	TTATCC	3	67%	33%	3,138	39,239	12.504	53
AAGGC	TTCCG	3	40%	60%	3,087	40,842	13.23	52
ACATGG	TGTACC	4	50%	50%	3,086	38,305	12.413	23
ACTCAG	TGAGTC	4	50%	50%	3,049	38,216	12.534	25
ACATAG	TGTATC	4	67%	33%	3,034	38,969	12.844	98
ACCTC	TGGAG	3	40%	60%	3,000	37,728	12.576	31
AAAGGT	TTTCCA	3	67%	33%	2,894	36,133	12.485	24
ACCCAG	TGGGTC	3	33%	67%	2,885	36,518	12.658	48
AACTTC	TTGAAG	3	67%	33%	2,877	35,842	12.458	22
AACACT	TTGTGA	3	67%	33%	2,876	37,476	13.031	43
ACAGTC	TGTCAG	4	50%	50%	2,876	35,917	12.489	22
AACTCC	TTGAGG	3	50%	50%	2,867	35,833	12.498	31
AACAGT	TTGTCA	4	67%	33%	2,865	35,908	12.533	22
AATGTC	TTACAG	4	67%	33%	2,860	35,939	12.566	30
AGCATG	TCGTAC	4	50%	50%	2,811	35,027	12.461	22
AATAGC	TTATCG	4	67%	33%	2,802	35,248	12.58	31

ACTGGG	TGACCC	4	33%	67%	2,766	34,913	12.622	29
CCCCGG	GGGGCC	2	0%	100%	2,763	37,124	13.436	77
AGATCG	TCTAGC	4	50%	50%	2,748	33,256	12.102	23
ATGCCC	TACGGG	4	33%	67%	2,717	34,102	12.551	29
AATCAC	TTAGTG	3	67%	33%	2,695	33,687	12.5	22
ATATGC	TATACG	4	67%	33%	2,682	34,140	12.729	84
AAGACC	TTCTGG	3	50%	50%	2,674	33,270	12.442	20
AACCCT	TTGGGA	3	50%	50%	2,666	49,523	18.576	1305
AAAGCT	TTTCGA	4	67%	33%	2,660	33,557	12.615	21
ACCTCT	TGGAGA	3	50%	50%	2,645	33,193	12.549	61
ACCAGG	TGGTCC	3	33%	67%	2,636	33,034	12.532	31
AACACC	TTGTGG	2	50%	50%	2,625	33,390	12.72	43
AACTAT	TTGATA	3	83%	17%	2,599	32,628	12.554	32
AATCT	TTAGA	3	80%	20%	2,597	33,210	12.788	61
ACATGT	TGTACA	4	67%	33%	2,597	33,125	12.755	29
ATGC	TACG	4	50%	50%	2,587	34,047	13.161	27
ACACTG	TGTGAC	4	50%	50%	2,552	31,738	12.437	21
ACT	TGA	3	67%	33%	2,510	37,852	15.08	62
AATGGC	TTACCG	4	50%	50%	2,477	31,111	12.56	22
AATCTC	TTAGAG	3	67%	33%	2,471	30,984	12.539	42
ACCT	TGGA	3	50%	50%	2,456	37,153	15.127	57
AACTGT	TTGACA	4	67%	33%	2,446	30,591	12.507	30
ACTATC	TGATAG	3	67%	33%	2,444	30,635	12.535	24
AAGAGT	TTCTCA	3	67%	33%	2,393	29,795	12.451	23
AAGCCC	TTCGGG	3	33%	67%	2,392	29,877	12.49	26
ATATCC	TATAGG	3	67%	33%	2,377	30,338	12.763	44
AATGGT	TTACCA	3	67%	33%	2,361	29,696	12.578	30
AAACCT	TTTGGA	3	67%	33%	2,318	29,108	12.557	49
AACCAC	TTGGTG	2	50%	50%	2,307	29,387	12.738	29
AAGAT	TTCTA	3	80%	20%	2,272	28,611	12.593	76
ACTGGC	TGACCG	4	33%	67%	2,271	28,338	12.478	22
ACCCCT	TGGGGA	3	33%	67%	2,263	29,184	12.896	59
AAGGAT	TTCCTA	3	67%	33%	2,211	28,114	12.716	82
ACCTGG	TGGACC	4	33%	67%	2,140	26,810	12.528	47
AACTC	TTGAG	3	60%	40%	2,138	26,574	12.429	25
AATGCT	TTACGA	4	67%	33%	2,120	26,551	12.524	24
ACTGCT	TGACGA	4	50%	50%	2,108	26,818	12.722	51
AATTGC	TTAACG	4	67%	33%	2,106	25,760	12.232	25
ACTG	TGAC	4	50%	50%	2,105	27,627	13.124	42

AAGCAC	TTCGTG	3	50%	50%	2,099	26,355	12.556	24
ACATCT	TGTAGA	3	67%	33%	2,088	26,466	12.675	56
ACAGT	TGTCA	4	60%	40%	2,087	26,529	12.712	49
ATGCC	TACGG	4	40%	60%	2,085	26,401	12.662	31
CCCGG	GGGCC	2	0%	100%	2,084	27,917	13.396	70
AAGTAT	TTCATA	3	83%	17%	2,079	25,763	12.392	55
ACGCCC	TGCGGG	3	17%	83%	2,074	25,352	12.224	28
AATCC	TTAGG	3	60%	40%	2,041	25,740	12.611	65
AACCTG	TTGGAC	4	50%	50%	2,038	25,486	12.505	24
AACCC	TTGGG	2	40%	60%	2,036	25,611	12.579	32
AATACC	TTATGG	3	67%	33%	2,031	25,313	12.463	39
ACCACT	TGGTGA	3	50%	50%	1,987	25,228	12.697	22
ACTGAT	TGACTA	4	67%	33%	1,956	24,417	12.483	22
ACCAG	TGGTC	3	40%	60%	1,933	24,537	12.694	38
AAGCAT	TTCGTA	4	67%	33%	1,932	24,077	12.462	20
ACATG	TGTAC	4	60%	40%	1,932	24,395	12.627	39
AAGCTC	TTCGAG	4	50%	50%	1,896	23,719	12.51	23
AATCCT	TTAGGA	3	67%	33%	1,886	23,642	12.536	35
AGATCC	TCTAGG	4	50%	50%	1,856	23,287	12.547	25
CCCG	GGGC	2	0%	100%	1,854	24,951	13.458	30
AAGT	TTCA	3	75%	25%	1,845	24,845	13.466	47
AATCCC	TTAGGG	3	50%	50%	1,830	22,872	12.498	24
ACAGCT	TGTCGA	4	50%	50%	1,821	22,833	12.539	27
AATGCC	TTACGG	4	50%	50%	1,799	22,499	12.506	24
AAGTGT	TTCACA	3	67%	33%	1,787	22,178	12.411	19
AACTGC	TTGACG	4	50%	50%	1,785	22,242	12.461	22
AACTAC	TTGATG	3	67%	33%	1,777	22,279	12.537	19
ACCATG	TGGTAC	4	50%	50%	1,750	21,899	12.514	34
AGATC	TCTAG	4	60%	40%	1,748	21,930	12.546	24
ACTCCT	TGAGGA	3	50%	50%	1,743	21,838	12.529	27
ACTGG	TGACC	4	40%	60%	1,723	21,683	12.584	29
AAGTC	TTCAG	4	60%	40%	1,715	21,406	12.482	23
ACCTG	TGGAC	4	40%	60%	1,715	21,477	12.523	20
AACTCT	TTGAGA	3	67%	33%	1,712	21,992	12.846	556
AAGTAC	TTCATG	4	67%	33%	1,656	20,557	12.414	20
AGGCAT	TCCGTA	4	50%	50%	1,645	21,169	12.869	20
AGGATC	TCCTAG	4	50%	50%	1,628	20,403	12.533	19
AGCAT	TCGTA	4	60%	40%	1,612	21,782	13.512	107
ACTATG	TGATAC	4	67%	33%	1,609	20,179	12.541	29

AACCAT	TTGGTA	3	67%	33%	1,592	19,945	12.528	25
ACACT	TGTGA	3	60%	40%	1,562	20,276	12.981	89
AAGTCT	TTCAGA	4	67%	33%	1,507	18,677	12.393	21
AACCTT	TTGGAA	3	67%	33%	1,498	18,667	12.461	40
AAGACT	TTCTGA	4	67%	33%	1,495	18,675	12.492	21
AGCCT	TCGGA	4	40%	60%	1,474	18,299	12.415	26
ACAGGT	TGTCCA	4	50%	50%	1,440	18,165	12.615	31
AAGTGC	TTCACG	4	50%	50%	1,433	17,780	12.408	17
ACTCAT	TGAGTA	3	67%	33%	1,413	17,661	12.499	22
ATGGCC	TACCGG	4	33%	67%	1,392	17,484	12.56	23
AAGCT	TTCGA	4	60%	40%	1,372	17,375	12.664	23
AAGGT	TTCCA	3	60%	40%	1,365	17,146	12.561	29
AGAGCT	TCTCGA	4	50%	50%	1,357	17,075	12.583	27
ACTCT	TGAGA	3	60%	40%	1,354	17,364	12.824	68
ACCTAT	TGGATA	3	67%	33%	1,324	17,130	12.938	44
AGCGGC	TCGCCG	3	17%	83%	1,306	17,311	13.255	52
AGGGAT	TCCCTA	3	50%	50%	1,251	15,803	12.632	51
ACCAT	TGGTA	3	60%	40%	1,244	16,152	12.984	66
AGGCGC	TCCGCG	3	17%	83%	1,237	15,704	12.695	27
ACCCT	TGGGA	3	40%	60%	1,231	15,860	12.884	63
CCCGCG	GGGCGC	2	0%	100%	1,218	16,040	13.169	37
AGGAT	TCCTA	3	60%	40%	1,193	16,549	13.872	183
AGCCAT	TCGGTA	4	50%	50%	1,190	14,991	12.597	27
ATCATG	TAGTAC	4	67%	33%	1,173	14,923	12.722	37
ACTAG	TGATC	4	60%	40%	1,158	14,837	12.813	54
AGCCCT	TCGGGA	4	33%	67%	1,132	14,405	12.725	41
ACCGCC	TGGCGG	3	17%	83%	1,130	15,029	13.3	40
AACCT	TTGGA	3	60%	40%	1,127	14,258	12.651	54
AAGGGT	TTCCCA	3	50%	50%	1,111	13,916	12.526	21
AACT	TTGA	3	75%	25%	1,109	15,851	14.293	51
ACACCT	TGTGGA	3	50%	50%	1,089	13,773	12.647	35
AACTT	TTGAA	3	80%	20%	1,073	13,135	12.241	17
AAGGCT	TTCCGA	4	50%	50%	1,068	13,307	12.46	23
ACCCAT	TGGGTA	3	50%	50%	1,067	13,302	12.467	24
AGCCGC	TCGGCG	3	17%	83%	1,064	13,725	12.899	39
CG	GC	2	0%	100%	1,039	15,061	14.496	28
ACCGAG	TGGCTC	3	33%	67%	1,020	12,397	12.154	18
ACTAGG	TGATCC	4	50%	50%	1,016	12,656	12.457	26
AAGCCT	TTCGGA	4	50%	50%	997	12,449	12.486	20

AGCGGG	TCGCCC	3	17%	83%	993	12,901	12.992	35
ACCTAG	TGGATC	4	50%	50%	986	12,312	12.487	22
CCGCG	GGCGC	2	0%	100%	962	12,693	13.194	30
ACGC	TGCG	3	25%	75%	862	12,201	14.154	49
AGCCGG	TCGGCC	3	17%	83%	822	10,911	13.274	42
AGCT	TCGA	4	50%	50%	820	11,480	14	35
ACGGGG	TGCCCC	3	17%	83%	814	10,432	12.816	47
AGATCT	TCTAGA	4	67%	33%	814	10,546	12.956	54
AGAGCG	TCTCGC	3	33%	67%	780	10,490	13.449	37
AGCTAT	TCGATA	4	67%	33%	768	9,750	12.695	26
ACTAGC	TGATCG	4	50%	50%	717	8,907	12.423	21
AGGGCG	TCCCGC	3	17%	83%	687	8,983	13.076	35
CCCGGG	GGGCCC	2	0%	100%	670	8,541	12.748	23
AGCGCC	TCGCGG	3	17%	83%	649	8,244	12.703	36
AAGCTT	TTCGAA	4	67%	33%	648	8,072	12.457	19
ACGCGC	TGCGCG	3	17%	83%	648	8,694	13.417	117
ACGGAG	TGCCTC	3	33%	67%	643	8,518	13.247	47
ACGGCC	TGCCGG	3	17%	83%	640	8,136	12.713	26
ACCCCG	TGGGGC	3	17%	83%	603	7,847	13.013	58
AAAACG	TTTTGC	3	67%	33%	592	7,346	12.409	17
AGGCCT	TCCGGA	4	33%	67%	559	6,950	12.433	17
CCGCGG	GGCGCC	2	0%	100%	557	7,026	12.614	23
AGCCCG	TCGGGC	3	17%	83%	540	7,140	13.222	48
ACGTGC	TGCACG	4	33%	67%	520	6,565	12.625	23
ACGAGG	TGCTCC	3	33%	67%	488	6,310	12.93	24
ACGTCC	TGCAGG	4	33%	67%	464	5,782	12.461	28
ACTAGT	TGATCA	4	67%	33%	456	5,710	12.522	22
AGCG	TCGC	3	25%	75%	450	6,245	13.878	26
AGGCG	TCCGC	3	20%	80%	450	6,026	13.391	30
ACCCGC	TGGGCG	3	17%	83%	444	5,675	12.782	28
AAGCGG	TTCGCC	3	33%	67%	428	5,393	12.6	23
CCGGCG	GGCCGC	2	0%	100%	425	5,408	12.725	22
AACGG	TTGCC	3	40%	60%	424	6,313	14.889	54
ACCGGC	TGGCCG	3	17%	83%	406	5,236	12.897	19
ACACGG	TGTGCC	3	33%	67%	403	4,988	12.377	19
AGCGG	TCGCC	3	20%	80%	402	5,149	12.808	26
CCGG	GGCC	2	0%	100%	398	5,125	12.877	20
ACGATG	TGCTAC	4	50%	50%	381	4,859	12.753	23
ACGTAT	TGCATA	4	67%	33%	374	5,075	13.57	36

AGCGC	TCGCG	3	20%	80%	358	4,617	12.897	30
AGCCG	TCGGC	3	20%	80%	343	4,730	13.79	179
ATGCGC	TACGCG	4	33%	67%	339	4,163	12.28	21
ACGGGC	TGCCCG	3	17%	83%	318	3,987	12.538	20
ACCGGG	TGGCCC	3	17%	83%	311	3,953	12.711	23
ACACGT	TGTGCA	4	50%	50%	302	3,831	12.685	53
ACGG	TGCC	3	25%	75%	300	4,834	16.113	55
AAGGCG	TTCCGC	3	33%	67%	297	3,730	12.559	29
ACGGG	TGCCC	3	20%	80%	281	3,822	13.601	36
ACCGTC	TGGCAG	4	33%	67%	279	3,565	12.778	34
AATCG	TTAGC	4	60%	40%	273	3,770	13.81	25
ACCCGG	TGGGCC	3	17%	83%	270	3,439	12.737	29
AAACGG	TTTGCC	3	50%	50%	268	3,302	12.321	18
ACGCC	TGCGG	3	20%	80%	263	3,460	13.156	29
ATCGCC	TAGCGG	4	33%	67%	262	3,284	12.534	29
ACGCTC	TGCGAG	4	33%	67%	253	3,230	12.767	41
AAGCCG	TTCGGC	3	33%	67%	252	3,139	12.456	32
AAGACG	TTCTGC	3	50%	50%	250	3,240	12.96	37
ACTCCG	TGAGGC	4	33%	67%	246	3,083	12.533	22
ATCCGC	TAGGCG	4	33%	67%	243	3,104	12.774	27
ACCGTG	TGGCAC	4	33%	67%	235	2,922	12.434	20
AACGTG	TTGCAC	4	50%	50%	232	2,875	12.392	21
ACAGCG	TGTCGC	3	33%	67%	228	2,877	12.618	22
ACGCAG	TGCGTC	3	33%	67%	227	2,883	12.7	44
AAAGCG	TTTCGC	3	50%	50%	221	2,767	12.52	17
AACGGG	TTGCCC	3	33%	67%	205	2,582	12.595	19
AGCGCG	TCGCGC	3	17%	83%	204	2,628	12.882	23
AACTCG	TTGAGC	4	50%	50%	196	2,427	12.383	16
AAACG	TTTGC	3	60%	40%	193	2,599	13.466	39
AACGAG	TTGCTC	3	50%	50%	193	2,405	12.461	18
ACGCTG	TGCGAC	4	33%	67%	191	2,361	12.361	15
ACCTCG	TGGAGC	4	33%	67%	189	2,368	12.529	21
ACCGC	TGGCG	3	20%	80%	187	2,440	13.048	29
ACGGCG	TGCCGC	3	17%	83%	173	2,283	13.197	35
ACTCGG	TGAGCC	4	33%	67%	165	2,078	12.594	24
AAACGT	TTTGCA	4	67%	33%	157	1,931	12.299	16
AACGAC	TTGCTG	3	50%	50%	156	2,051	13.147	18
ACCACG	TGGTGC	3	33%	67%	156	1,977	12.673	23
AACACG	TTGTGC	3	50%	50%	154	1,901	12.344	18

AAGCGC	TTCGCG	3	33%	67%	153	1,925	12.582	21
ACACCG	TGTGGC	3	33%	67%	152	1,865	12.27	16
AAACCG	TTTGGC	3	50%	50%	149	1,840	12.349	18
ACCGGT	TGGCCA	4	33%	67%	147	1,784	12.136	17
ACG	TGC	3	33%	67%	147	2,263	15.395	53
AACCGG	TTGGCC	3	33%	67%	141	1,795	12.73	23
ACGGC	TGCCG	3	20%	80%	136	1,718	12.632	24
AAATCG	TTTAGC	4	67%	33%	135	1,670	12.37	23
ATCCCG	TAGGGC	4	33%	67%	134	1,744	13.015	30
AACCCG	TTGGGC	3	33%	67%	133	1,968	14.797	71
AACGGC	TTGCCG	3	33%	67%	133	1,643	12.353	19
ACTCGC	TGAGCG	4	33%	67%	132	1,640	12.424	17
ACCCG	TGGGC	3	20%	80%	125	1,609	12.872	24
AAGTCG	TTCAGC	4	50%	50%	124	1,549	12.492	18
ACGATC	TGCTAG	4	50%	50%	123	1,542	12.537	27
ACGAGC	TGCTCG	3	33%	67%	121	1,482	12.248	15
AAACGC	TTTGCG	3	50%	50%	120	1,469	12.242	14
AACG	TTGC	3	50%	50%	120	1,714	14.283	48
AACGCC	TTGCGG	3	33%	67%	116	1,444	12.448	21
ATCGGC	TAGCCG	4	33%	67%	115	1,425	12.391	17
ACGCAT	TGCGTA	4	50%	50%	114	1,407	12.342	17
ACGCGG	TGCGCC	3	17%	83%	110	1,377	12.518	17
AATACG	TTATGC	4	67%	33%	109	1,332	12.22	17
ACGCCG	TGCGGC	3	17%	83%	105	1,366	13.01	17
ACGAG	TGCTC	3	40%	60%	101	1,425	14.109	58
AATCGG	TTAGCC	4	50%	50%	95	1,157	12.179	15
ACCGCG	TGGCGC	3	17%	83%	94	1,210	12.872	20
ACGTC	TGCAG	4	40%	60%	94	1,383	14.713	43
AACCGC	TTGGCG	3	33%	67%	90	1,117	12.411	18
ACTGCG	TGACGC	4	33%	67%	90	1,113	12.367	17
AATCCG	TTAGGC	4	50%	50%	88	1,073	12.193	16
AACGAT	TTGCTA	4	67%	33%	87	1,073	12.333	17
ACATCG	TGTAGC	4	50%	50%	87	1,066	12.253	15
AGCTCG	TCGAGC	4	33%	67%	86	1,050	12.209	15
ACCGCT	TGGCGA	4	33%	67%	84	1,077	12.821	18
AACGTC	TTGCAG	4	50%	50%	83	1,020	12.289	15
ACGTAG	TGCATC	4	50%	50%	78	966	12.385	16
AATGCG	TTACGC	4	50%	50%	76	926	12.184	14
AACGCT	TTGCGA	4	50%	50%	71	884	12.451	17

ACGGAT	TGCCTA	4	50%	50%	71	867	12.211	17
ACGGCT	TGCCGA	4	33%	67%	71	901	12.69	18
AAGCG	TTCGC	3	40%	60%	66	840	12.727	19
ACCGG	TGGCC	3	20%	80%	66	837	12.682	16
ACGCCT	TGCGGA	4	33%	67%	64	808	12.625	23
ACGAT	TGCTA	4	60%	40%	63	980	15.556	84
AATTCG	TTAAGC	4	67%	33%	62	765	12.339	17
ACACG	TGTGC	3	40%	60%	62	795	12.823	24
AATCGT	TTAGCA	4	67%	33%	59	717	12.153	14
AACCG	TTGGC	3	40%	60%	54	705	13.056	34
AATCGC	TTAGCG	4	50%	50%	54	664	12.296	14
ATCG	TAGC	4	50%	50%	54	721	13.352	21
ACGACT	TGCTGA	4	50%	50%	53	667	12.585	23
AACGT	TTGCA	4	60%	40%	51	960	18.824	64
AACCGT	TTGGCA	4	50%	50%	50	608	12.16	14
ATCCGG	TAGGCC	4	33%	67%	49	614	12.531	16
AACGTT	TTGCAA	4	67%	33%	48	588	12.25	15
AACGGT	TTGCCA	4	50%	50%	47	581	12.362	14
AAGCGT	TTCGCA	4	50%	50%	45	563	12.511	18
ACCG	TGGC	3	25%	75%	45	608	13.511	31
ACGAGT	TGCTCA	4	50%	50%	44	545	12.386	15
AGCGAT	TCGCTA	4	50%	50%	42	513	12.214	14
ATCGC	TAGCG	4	40%	60%	42	583	13.881	53
ACCCGT	TGGGCA	4	33%	67%	40	492	12.3	17
ATATCG	TATAGC	4	67%	33%	36	442	12.278	18
AACGC	TTGCG	3	40%	60%	34	456	13.412	34
ACCGAT	TGGCTA	4	50%	50%	32	392	12.25	16
ACGCG	TGCGC	3	20%	80%	31	387	12.484	16
ATCCG	TAGGC	4	40%	60%	30	388	12.933	24
ACTCG	TGAGC	4	40%	60%	29	361	12.448	15
ACGT	TGCA	4	50%	50%	27	386	14.296	31
AACGCG	TTGCGC	3	33%	67%	25	308	12.32	15
ACGCT	TGCGA	4	40%	60%	24	329	13.708	24
ACCGT	TGGCA	4	40%	60%	14	190	13.571	33
AGCGCT	TCGCGA	4	33%	67%	14	177	12.643	15
ATCGCG	TAGCGC	4	33%	67%	10	122	12.2	13
ACGTCG	TGCAGC	4	33%	67%	8	99	12.375	13
ACGCGT	TGCGCA	4	33%	67%	2	24	12	12