INCREASED SUBSTITUTION RATES IN DNA SURROUNDING LCRS

INCREASED SUBSTITUTION RATES IN DNA SURROUNDING LOW-COMPLEXITY REGIONS

By CAROLYN LENZ, B. Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University MASTER OF SCIENCE (2013) Hamilton, Ontario(Biology)

TITLE: Increased Substitution Rates in DNA Surrounding Low-Complexity Regions

AUTHOR: Carolyn Lenz, B. Sc.(University of Guelph)

SUPERVISOR: Brian Golding

NUMBER OF PAGES: xiii, 86

Abstract

Previous studies have found that DNA flanking low-complexity regions (LCRs) have an increased substitution rate. Here, the substitution rate was confirmed to increase in the vicinity of LCRs in several primate species, including humans. This effect was also found within human sequences from the 1000 Genomes Project. A strong correlation was found between average substitution rate per site and distance from the LCR, as well as between the proportion of genes with gaps in the alignment at each site and distance from the LCR. Along with substitution rates, d_N/d_S ratios were also determined for each site, and the proportion of sites undergoing negative selection was found to have a negative relationship with distance from the LCR.

Low-complexity regions in proteins often form and extend through the gain or loss of repeated units, a process that is dependent on the presence of a relatively pure string of repeats. Any interruption should disrupt the mechanisms of LCR extension and contraction, inhibiting LCR formation. Despite this, several examples have been found of LCR-coding DNA which are interrupted by introns. While many of these LCRs may be the result of two shorter LCRs forming on opposite sides of an intron, shuffling the order of exons showed that more intron-interrupted LCRs exist than would be expected to occur randomly. Another possible explanation for this phenomenon is the apparent movement of either the LCRs or introns, possibly through recombination or the appearance of new splice sites through the gain of repeat units.

Acknowledgements

I would like to thank Brian Golding, my supervisor, for his guidance and support throughout my research, as well as Ben Evans for his suggestions and advice. I am also grateful to Wilfried Haerty, whose research this thesis was a continuation of, for providing the SNP data from the Broad Institute and for his invaluable comments on my work in Chapter 2. The support of Golding and Evans lab members, past and present, was also a huge help and cannot be overstated.

Contents

List of Figures

List of Tables

Declaration of Academic Achievement

Through the course of my research, I have demonstrated that there is a strong relationship between proximity to an LCR and number of substitutions and indels. As well, these substitutions have sigificant evidence of negative selection. I have also shown that complexity increases with distance from the LCR, and, throughout proteins, number of substitutions increases as complexity decreases, indicating that the high mutation rate previously observed surrounding LCRs is linked to the low complexity of LCR flanking regions.

Intron-interrupted LCRs were also examined, and found to be more common than expected if they were the result of random processes. Measuring information content, I found these LCRs to be significantly less complex than non-interrupted LCRs. As well, the proteins containing these LCRs were significantly less complex than the average LCR-containing protein. As less complex proteins are more likely to give rise to LCRs in general, this could explain how intron-interrupted LCRs form.

Chapter 1

Introduction

1.1 Low-complexity Regions

Low-complexity regions, or LCRs, are protein-coding DNA sequences defined by their low information content. They are often extremely repetitive, and lack a stable, three dimensional structure in the translated protein (Simon and Hancock 2009). Due to this lack of structure, LCRs were initially thought to convey no benefit to the protein, and, in many cases, were found to actively disrupt the protein's function. A few often-cited examples of deleterious LCRs have significant impacts on human health, causing disorders such as Huntington's (Mangiarini et al. 1997) and myotonic dystrophy (Pearson and Sinden 1996). In general, LCRs can also have adverse effects through disrupting chromosome structure and silencing genes (Usdin 2008).

Despite the many examples of harmful LCRs, repetitive regions are a common trait in eukayotic genomes (Golding 1999, Romero et al. 2001, DePristo et al. 2006). This could be due to their high mutation rates, which can range from 100 to 10,000 times higher than the average nucleotide mutation rate across the rest of the genome (Vinces et al. 2009), especially as LCRs have a propensity to expand and contract (Lovell 2003). This happens as a result of the repetitive nature of LCRs, for example through slippage of DNA polymerase during replication (Levinson and Gutman 1987) or unequal recombination during cross-over events (Amos et al. 2008). LCRs can also accumulate point mutations, as genomes may be more tolerant of mutations occurring in LCRs, and the DNA repair mechanisms can be impeded by the formation of secondary structures (Moore et al. 1999). These substitutions may disrupt the repetitive nature of LCRs, decreasing the probability of slippage or unequal recombination, eventually increasing the information content of the LCR until they are no longer classified as low-complexity (Radó-Trilla and Albà 2012).

While some LCRs may retain their presence over long periods of time due to expansion (Lovell 2003), there is evidence that LCRs are more conserved than similarly repetitive regions that are located outside of protein-coding regions (Mularoni et al. 2010). A few even appear to have beneficial effects on phenotype. In domestic dogs, several direct relationships have been found between the lengths of certain LCRs and presence or absence of discrete traits, such as dewclaws, and even between LCR length and more quantitative traits, such as snout length (Fondon and Garner 2004). Another example of this

is seen in yeast, where the number of repeated units in various LCRs changes the structure of surface proteins, allowing the yeast to bind to different substrates (Verstrepen et al. 2005). LCRs have even been seen to affect behaviour, as the presence of a specific microsatellite in the vasopression 1a receptor gene in prairie voles appears to be related to social bonding. A comparable species without this microsatellite, the mountain vole, as well as prairie voles homozygous for shorter versions of this microsatellite, did not exhibit the same pair-bonding behaviour (Hammock and Young 2005). The fact that small changes to LCRs can cause minor adjustments to many phenotypic traits has led to speculation that LCRs may act as 'evolutionary tuning knobs' (Vinces et al. 2009).

Information content is used as a measure of how repetitive sequences are, because a lower number of amino acids represented limits the amount of information a protein sequence can convey. For the purposes of my study I used SEG, which determines the complexity of a sliding window of the minimum acceptable length for LCRs, and is typically used by BLAST (Altschul et al. 1990) to mask repeated sequences. SEG uses sliding windows to find regions with low complexity, as opposed to other programs, such as HMMER (Eddy 2009), that use hidden Markov models to determine the probability that two segments of a sequence are the result of repeat expansions. The complexity of each segment of the sequence is measured in bits, the most basic units of information content, and is defined as k, calculated as

$$k = -\sum_{i=1}^{20} \frac{n_i}{L} \left(\log \left(\frac{n_i}{L} \right) \right) \tag{1.1}$$

where L is the length of the sequence and n_i is the number of occurrences of each amino acid type i. Complexity of each window is compared to k_1, the trigger complexity. The default values for window length and k_1 are 12 and 2.2, respecitvely, however I used a window length of 15 and trigger complexity of 1.9. Although sequence complexity is a continuous measure, and determining whether a sequence falls into a discrete category such as low-complexity is difficult, these values were determined by trial and error by Huntley and Golding (2002) to identify the longer, more repetitive LCRs observed in eukaryotes. Once a window is identified where $k \leq k_1$, any overlapping windows of length L with a complexity of $k \leq k_2$ (which I kept at the default setting of 2.5) are merged with the window first identified as low complexity. SEG keeps extending the length of the LCR in this way until an overlapping window where $k \leq k_2$ cannot be found (Wootton 1994).

This report consists of two main chapters. The first chapter concerns the high number of mutations previously observed in regions flanking LCRs and the evidence for and against various mechanisms which could cause this effect. In the second chapter, I describe the phenomenon of intron-interrupted LCRs.

1.2 Flanking regions

Previous studies have found that DNA flanking LCRs has an increased substitution rate (Huntley and Clark 2007, Haerty and Golding 2011). Here, the substitution rate was confirmed to increase in the vicinity of LCRs in several primate species, including humans. This effect was also found within human sequences from the 1000 Genomes Project. A strong correlation was found between average substitution rate per site and distance from the LCR, as well as the proportion of genes with gaps in the alignment at

each site and distance from the LCR. Along with substitution rates, d_N/d_S ratios were also determined for each site, and the proportion of sites undergoing negative selection was found to have a negative relationship with distance from the LCR. As well, several mechanisms which could produce a high mutation rate in the vicinity of LCRs were investigated.

1.3 Intron-Interrupted LCRs

Low-complexity regions in proteins often form and extend through the gain or loss of repeated units, a process that is dependent on the presence of a relatively pure string of repeats. Any interruption should disrupt the mechanisms of LCR extension and contraction, inhibiting LCR formation. Despite this, several examples have been found of LCR-coding DNA which are interrupted by introns. While many of these LCRs may be the result of two shorter LCRs forming on opposite sides of an intron, others exhibit perfect repeats, and appear to be one continuous LCR. Conversely, introns are unlikely to have appeared in these regions after the formation of these LCRs. A possible explanation for this phenomenon is the apparent movement of either the LCRs or introns, possibly through recombination or the appearance of new splice sites through the gain of repeat units.

Bibliography

SF Altschul, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

W Amos, J Flint, and X Xu. Heterozygosity increases microsatellite mutation rate, linking it to demographic history. *BMC Genetics*, 9:72, 2008.

MA DePristo, MM Zilversmit, and DL Hartl. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene*, 378:19 – 30, 2006.

SR Eddy. A new generation of homology search tools based on probabilistic inference. *Genome Informatics*, pages 205 – 211, 2009.

JW Fondon and HR Garner. Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences*, 101:18058–18063, 2004.

GB Golding. Simple sequence is abundant in eukaryotic proteins. *Protein Science*, 8:1358–1361, 1999.

W Haerty and GB Golding. Increased Polymorphism Near Low-Complexity Sequences across the Genomes of Plasmodium falciparum Isolates. *Genome Biology and Evolution*, 3:539–550, 2011.

EAD Hammock and LJ Young. Microsatellite Instability Generates Diversity in Brain and Sociobehavioral Traits. *Science*, 308:1630–1634, June 2005.

MA Huntley and AG Clark. Evolutionary Analysis of Amino Acid Repeats across the Genomes of 12 Drosophila Species. *Molecular Biology and Evolution*, 24:2598–2609, 2007.

MA Huntley and GB Golding. Simple sequences are rare in the Protein Data Bank. *Proteins: Structure, Function, and Bioinformatics*, 48:134–140, 2002.

G Levinson and GA Gutman. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, 4:203–221, 1987.

SC Lovell. Are non-functional, unfolded proteins (junk proteins) common in the genome? *FEBS Letters*, 554:237 – 239, 2003.

L Mangiarini, K Sathasivam, A Mahal, R Mott, M Seller, and GP Bates. Instability of highly expanded CAG repeats in mice transgenic for the Huntington's disease mutation. *Nature genetics*, 15:197–200, 1997.

H Moore, PW Greenwell, CP Liu, N Arnheim, and TD Petes. Triplet repeats form secondary structures that escape DNA repair in yeast. *Proceedings of the National Academy of Sciences*, 96:1504–1509, 1999.

L Mularoni, A Ledda, M Toll-Riera, and MM Albá. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Research*, 20:745–754, 2010.

CE Pearson and RR Sinden. Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry*, 35:5041–5053, 1996.

N Radó-Trilla and M Albà. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. *BMC Evolutionary Biology*, 12, 2012.

P Romero, Z Obradovic, X Li, EC Garner, CJ Brown, and AK Dunker. Sequence complexity of disordered protein. *Proteins*, 42:38–48, 2001.

M Simon and J Hancock. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biology*, 10:R59, 2009.

K Usdin. The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. *Genome Research*, 18:1011–1019, 2008.

KJ Verstrepen, A Jansen, F Lewitter, and GR Fink. Intragenic tandem repeats generate functional variability. *Nature Genetics*, 37:986–990, 2005.

MD Vinces, M Legendre, M Caldara, M Hagihara, and KJ Verstrepen. Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability. *Science*, 324:1213–1216, 2009.

JC Wootton. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Computers & Chemistry*, 18(3):269 – 285, 1994.

Chapter 2

Increased substitution rates in DNA surrounding low-complexity regions

2.1 Focus

Although it's been found that DNA flanking low-complexity regions, or LCRs, has an increased substitution rate, this has only been observed in a few species. The availability of several primate genomes, as well as the 1000 Genomes Project (Consortium 2010), provides an excellent opportunity to confirm the increased mutation rate associated with low-complexity sequences within the protein coding regions of *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Gorilla gorilla* (gorilla), *Pongo pygmaeus* (orangutan) and *Macaca mulatta* (rhesus macaque), as well as between human individuals. The average branch lengths and substitution rate of the flanking regions were compared to those of regions further away from LCRs, or non-flanking regions, to confirm the effect of LCRs. As well, substitution rate was estimated for each individual site, to determine how distance from the LCR was affecting the number of substitutions.

A strong correlation was found between average substitution rate per site and distance from the LCR, as well as between distance from the LCR and the proportion of genes with gaps in the alignment at each site. Along with substitution rates, d_N/d_S ratios were also found for each site, and the proportion of sites which had evidence of positive selection was found to be higher in flanking regions than non-flanking regions. As well, the proportion of sites undergoing negative, or purifying selection was found to have a negative relationship with distance from the LCR. To measure selection seen within a single species, using the 1000 Genomes Project, Tajima's D was found for each codon, and a negative relatinship was observed between distance from the LCR and proportion of genes with negative values of Tajima's D at each site.

Several possible mechanisms that could potentially result in a high mutation rate around LCRs were investigated. In particular, complexity of non-LCR sequences seemed a likely candidate, as a strong relationship was found between distance from the LCR and complexity. As well, complexity had a negative correlation with number of substitutions even outside of LCRs.

2.2 Locating Flanking Regions

Representative *Homo sapiens* (human, GRCh37), *Pan troglodytes* (chimpanzee, CHIMP2.1.3), *Gorilla gorilla* (gorilla, gorGor2), *Pongo pygmaeus* (orangutan, PPYG2) and *Macaca mulatta* (rhesus macaque, MMUL_1) genomes were downloaded from Ensembl release 62 (Flicek et al. 2011). The sequences were searched for regions of low-complexity with SEG (Wootton and Federhen 1993), using a window size of 15 and a complexity threshold of 1.9, which finds longer and less complex sequences than SEG's default settings. The nucleotide sequence may be more informative than the translated protein sequence, as this is meant to identify sequences which are repetitive enough to be subject to the high mutation rates seen in LCRs and the protein sequence may appear more repetitive due to the degeneracy of the codon code. Complexity was calculated using both amino acids and codons to see how similar the two measurements were, and the translated protein sequence was found to be a close approximation of the complexity of the DNA sequence, with a nearly exact correlation (Pearson's correlation = 0.9928, $p < 10^{-16}$).

Using this method, 7387 LCRs were identified in humans. To verify these LCR sequences, they were searched for known repetitive protein-coding regions, such as the one found in Huntington's disease, to confirm that their identification was accurate. LCRs from the four other primate species were found and compared to find orthologs based on location within the protein as well as identity between the sequences, and 5584 LCRs (from 2425 genes) were found to be shared by all five species (Table 2.1). Data from the 1000 Genomes Project (Consortium 2010) was downloaded and LCRs that corresponded to the LCRs associated with the primate flanking regions were identified. The translated protein sequences containing these LCRs were aligned using Muscle (Edgar 2004).

Data from the 1000 Genomes Project (Consortium 2010) was downloaded and the protein-coding regions of each genome were searched for LCRs using SEG. Human LCRs that corresponded to the LCRs already identified in the primate genomes were found. The LCRs and surrounding DNA were aligned using Muscle (Edgar 2004) as before, to find flanking regions corresponding to the primate flanking regions described above.

In order to get a clear picture of how distance from an LCR encoding sequence affects substitution rate, flanking regions which did not overlap with those of another LCR and were not interrupted by either a transcription start site or stop codon were identified. For example, in the case of flanking regions up to 1500 nucleotides in length, this would mean the flanking region from an LCR could not be used if the LCR was less than 3000 nucleotides away from the nearest LCR, or twice the length of the flanking region. The termini of exons were not taken into account, as the DNA transcript was considered as a whole. As CodeML (Yang 2007) requires that sequences being compared do not contain gaps, flanking regions could only be used if sequences from all five primate species were present. Although using only flanking regions for which sequences both $5'$ and $3'$ of the LCR were usable would ensure the information on both sides of the LCR was comparable, the available number of flanking regions would be extremely small, leaving only 586 usable flanking regions 150 nucleotides in length and 79 examples 1500 nucleotides in length.

Since LCRs are often located near the termini of proteins (Huntley and Clark 2007), there were many LCRs for which only the $5'$ or $3'$ flanking region could be used, as the flanking region on the opposite site of the LCR was interrupted (see Figure 2.1 in Supplementary Material for distribution of maximum flanking region length). So that more sequences would be available for analysis, flanking

regions upstream and downstream of the LCR were instead considered separately to allow the use of uninterrupted sequences from one side of the LCR, even if the flanking region may have been cut off on the other side. This resulted in 1385 upstream flanking regions and 1399 downstream flanking regions of at least 300 nucleotides in length from 1136 genes, however the number of gaps in the vicinity of LCRs lowered the number of available examples surrounding the LCR.

2.3 Substitution Rates

To estimate the substitution rate for each site in both the primate and 1000 Genomes data, each flanking region was split into its constituent codons and CodeML (Yang 2007) was used to estimate the number of substitutions for each codon of the flanking regions, using model 2 (two or more d_N/d_S ratios for branches). The same was done for random samplings of regions which were at least 1500 nucleotides away from the nearest LCR, or non-flanking regions, to provide control regions that have not been influenced by nearby LCRs. As an accurate phylogenetic tree could not be provided for data from the 1000 Genomes Project, pairwise comparisons between all individuals were used instead. The mean of these pairwise comparisons was found for each site. Using the number of substitutions for each useable codon from each flanking region, an average number of substitutions was found based on distance from the LCR, i.e. the average number of substitutions was found between all useable codons adjacent to the LCR, codons that were separated from the LCR by three nucleotides, codons that were separated from the LCR by six nucleotides, etc (Figure 2.2). Pearson's product-moment correlation (R Development Core Team 2011) was used to test for a relationship between the average number of substitutions and the distance from the LCR. To ensure the substitutions detected were not the result of misalignments, the relationship between number of substitutions and distance from the LCR was also found only for flanking regions which were not likely to be inaccurately aligned (i.e. flanking regions which contained no gaps when aligned).

A wide view of the area surrounding the LCRs was obtained when considering flanking regions at least 300 nucleotides in length. Using these long flanking regions, LCR encoding sequences were found to have a profound effect on nearby DNA. A strong, statistically significant negative correlation was seen between the average number of substitutions and distance from the LCR (correlation coefficient = -0.7957, p-value $< 10^{-16}$, see Figure 2.3). As well, both synonymous and non-synonymous substitution rates had a significant relationship to distance from the LCR, although the relationship between synonymous substitution rate and distance was stronger than the relationship between non-synonymous substitution rate and distance (-0.8748 and -0.3787, respectively, p-value $< 10^{-16}$ for both, see Figures 2.4a and 2.4b).

Substitutions may be incorrectly identified if insertions or deletions cause a misalignment. As the DNA sequences studied have diverged over millions of years, some indels will most likely be present, especially in the presence of LCRs. To ensure that the relationship seen between distance from LCRs and the substitution rate was not due to misalignment, flanking regions were found which did not have any apparent indels (i.e., they did not have any gaps when aligned), and were not associated with LCRs containing apparent indels. The average number of substitutions per site was significantly lower, as compared using a t-test (R Development Core Team 2011), in flanking regions unaffected by indels (0.0802 vs 0.1419, $p < 2.2 \times 10^{-16}$). Although this indicates that at least some of the substitutions detected

may be due to misalignments caused by indels, the correlation between the number of substitutions per site and distance from the LCR was still strong (Pearson's correlation = -0.7432, $p < 2.2 \times 10^{-16}$). The association between LCRs and a high substitution rate is clearly not caused by inaccurate detection of substitutions through misalignements.

Using corresponding flanking regions from the 1000 Genome Project, a relationship similar to the one seen between primate sequences was found. The average number of substitutions per site had a significant negative correlation with distance from the LCR (correlation coefficient = -0.4999, p-value $< 10^{-16}$, Figure 2.5). This relationship was consistently seen, even when the number of substitutions was separated into synonymous and non-synonymous substitutions, although the correlation between distance and non-synonymous substitutions was slightly stronger than that for synonymous substitutions (correlation coefficients of -0.452 and -0.354, respectively, both with p-values of $< 10^{-16}$, Figures 2.6a and 2.6b).

Looking at flanking regions on a site-by-site basis, the relationship between LCRs and local substitution rate is obvious. Not only is the correlation statistically significant, but it is obvious simply through plotting the average number of substitutions that the substitution rate has increased in sequences near LCRs, and decreases rapidly as distance from the LCR increases. This demonstrates that the relationship between proximity to LCRs and substitution rate seen in previous studies (Huntley and Clark 2007, Haerty and Golding 2011) is also present in primates. It is also consistent with previous findings showing an increased number of mismatches and SNPs in the vicinity of microsatellites (Siddle et al. 2011, Vowles and Amos 2004), suggesting that the repetitiveness of these sequences is key to their increased mutation rate. This relationship is also apparent for substitutions segregating within humans, as shown using the data from the 1000 Genomes Project.

While the flanking regions clearly show a relationship between distance from the LCR and substitution rate, they also show asymmetry in that relationship. The average number of substitutions is higher on the $5'$ side of the LCR encoding sequence. Although it is not yet known for certain what mutational pressures are affecting DNA adjacent to LCRs, the skewness of the substitution rates could implicate a transcription related mechanism. Most of the LCRs used here were expressed in germ-line cells (expression patterns found using the Human Protein Atlas (Uhlen et al. 2012); 1108 upstream and 1119 downstream flanking regions were expressed in the germ-line and 277 upstream and 280 downstream flanking regions were not expressed in the germ-line), but the relationship between distance from the LCR and substitution rate remains skewed even if only cells not expressed in the germ-line are studied (Figures 2.7a - 2.8b in Supplementary Material). Polak et al. (2010) found a substitutional bias in DNA, where regions just downstream of the transcription start site are more prone to substitutions that lead to strong bonds (i.e. more CG base pairs, with three hydrogen-bonds, than TA pairs); an excess of C to T substitutions over G to A was also found near the transcription start site in some genes. Although a bias in substitution rate was not found, the difference in substitutions based on proximity to the transcription start site could be related to the asymmetry in the number of substitutions seen around LCRs. Because of the way LCRs are clustered at either end of proteins, there should be fewer upstream flanking regions near the start of the protein, and, conversely, fewer downstream flanking regions near the end of the protein. As these regions are biased in their locations relative to the transcription start site, it is possible that a bias in the substitution rate caused by proximity to the transcription start site is a factor.

2.4 Gap Rates

As CodeML is only capable of analyzing substitutions, but not insertions or deletions, any site with a gap in the alignment was ignored when calculating the average substitution rate. To study whether LCRs are associated with indels, the average number of indels per site was found using the aligned flanking regions. Where multiple sequences had gaps at aligned sites, the most parsimonious number of indels was estimated (i.e., if Human and Chimpanzee sequences had a gap at the same site, this is likely the result of one insertion or deletion, but if Human and Gorilla sequences have a gap that the Chimpanzee does not, this is more likely the result of two separate mutations). The average number of gaps had a significant, negative correlation with distance from the LCR (Pearson's correlation = -0.1959, $p < 2.2 \times 10^{-16}$, see Figure 2.9).

Both the high proportion of flanking sequences composed by gaps and the significant correlation between distance from the LCR and proportion of genes with a gap at each site demonstrate an increase in the number of insertions and deletions surrounding LCRs. Along with the number of substitutions, the number of indels in the aligned sequences appeared to have a relationship with distance from the LCR. From these results, it is clear that the frequency of both point mutations and indels increases in DNA immediately surrounding LCRs in primates. The mutation mechanisms which lead to a high substitution rate in LCR flanking regions clearly also contribute to the proliferation of indels.

Interestingly, in the sites immediately adjacent to the LCRs there was a drop in the proportion of genes with gaps. The decrease in indel mutations in close proximity to the LCR may indicate that DNA at the borders of LCRs is important for the correct alignment of these protein-coding regions during recombination. The LCRs themselves have variations in length that can cause inaccurate alignments; a slightly more stable region surrounding LCRs could mitigate the effects of this.

2.5 Selection

In order to test whether or not selection was occuring, CodeML was also used to estimate d_N/d_S, or ω, for each site for both flanking and non-flanking regions. An expected distribution of d_N/d_S ratios was found for each number of substitutions by simulating codon sequences under neutral selection using evolver (part of the PAML package; Yang (2007)). The d_N/d_S ratios were found for each of these sets of codons and used to determine whether estimated d_N/d_S ratios of the codons from flanking regions were expected under neutral selection, or indicated significant evidence of positive or negative selection. Although there were a larger number of aligned codons in proximity to the LCR, this should not affect the significance of d_N/d_S, as the sample size used when calculating the d_N/d_S ratio and generating expected distributions was always five. The proportion of genes which had evidence of positive selection (i.e. a significant d_N/d_S ratio greater than 1) or negative selection (i.e. a significant d_N/d_S ratio less than 1) was found for each site (Yang and Bielawski 2000). The average proportion of genes which had significant evidence of positive or negative selection was found for each aligned codon from the flanking regions, and the data was searched for a relationship between distance from the LCR and proportion of genes undergoing positive and negative selection.

Using the expected distributions of d_N/d_S for neutrally evolving codon alignments generated using evolver to identify only significant results and ignoring sites for which d_N/d_S could not be estimated (i.e.,

sites with gaps in the alignment or no substitutions), a strong, negative correlation was found between distance from the LCR and proportion of genes which had evidence for negative selection at each site (Pearson's product-moment correlation = -0.842, $p < 10^{-16}$, Figure 2.10). Like the findings of Huntley and Clark (2007), there was also a negative (although, in this case, non-significant) correlation between distance from the LCR and proportion of genes which had evidence for positive selection (Pearson's product-moment correlation, correlation coefficient = -0.008, $p = 0.7924$, Figure 2.11).

There is a significant skew in the proportion of genes with evidence for negative selection on either side of the LCR, with significantly more evidence of negative selection upstream of the LCR. This likely does not point to stronger selection upstream of the LCR, however. Since sites with no substitutions cannot provide any indication of whether selection is or is not happening, sequences with fewer substitutions should appear to have less evidence of either type of selection. The skew in evidence for negative selection is likely to be directly caused by the skew in number of substitutions.

To further explore the possible effects of selection, the value of Tajima's D (Tajima 1989) was calculated for each codon in the flanking regions of the data from the 1000 Genomes Project. To determine the significance of each D, ms (Hudson 2002) was used to generate a distribution of expected Tajima's D values for neutrally evolving fragments with the same length and number of polymorphic sites as each codon. The Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) was used to control for the false discovery rate, and the proportion of genes with evidence for positive or negative selection was found for each site.

Using only values of Tajima's D which were found to be significant, a negative correlation was found between distance from the LCR and proportion of genes with negative values of Tajima's D, which can indicate purifying selection (correlation coefficient = -0.393, $p < 10^{-16}$, Figure 2.12). As a negative value for Tajima's D can also indicate a bottleneck, a high number of negative values of Tajima's D at any distance from an LCR can be expected between human sequences due to demographic history, however the presence of an LCR clearly increases the proportion of sites undergoing negative selection. There was also a negative relationship between distance from the LCR and proportion of genes with positive values of Tajima's D, but this correlation was extremely weak and not significant (correlation coefficient = -0.010, $p = 0.7545$). Like the primate data, the proportion of genes with evidence for negative selection was non-symmetrical, although in this case there was more evidence of negative selection downstream of the LCR. There was a significant difference in the average proportion of genes with negative Tajima's D values per codon when both sets of data were compared with a t-test (0.2876 vs 0.3011, $p = 1.353 \times 10^{-12}$).

Another interesting result is the distinct dip immediately downstream of the LCR in both the proportion of genes with negative values of Tajima's D, suggesting that the DNA immediately adjacent to the LCRs is less conserved. This could be a result of the misidentification of the exact borders of the LCR, as these few sites might be under different selection pressures if they were part of the LCR itself.

2.6 SNP Density

The SNP calls, including the allele frequencies, for 178 individuals of African origin (Consortium 2010), were obtained, and the distance from the nearest LCR was found for all SNPs. Using the pairwise genome alignments downloaded from the UCSC database, the nucleotides were compared to homologous sites

in macaque and chimpanzee to determine the ancestral state of the SNPs using maximum parsimony. Because the most recent SNPs are expected to have low allelic frequency, the relationship between the alleleic frequency of each SNP and its distance to the LCR was assessed, in order to assess the mutagenic propensity of LCR encoding sequences on their flanking sequences. Only SNPs from protein-coding regions were used, since the more relaxed selection outside of coding regions could affect the relationship. Although a lower SNP frequency could also indicate more deleterious alleles, unless the presence of the LCR was increasing the substitution rate or causing a relaxation in selection, a distributional bias is not expected. As well, there is evidence against a relaxation in selection in regions flanking LCRs (see Section 2.5). Distance to the nearest protein-coding LCR, including introns and intergenic regions, was calculated.

Using the SNP calls from the Broad Institute, distance from the LCR was found to have a significant positive relationship with the frequency of derived SNPs (Spearman's $\rho = 0.0268$, $p < 10^{-16}$), although this relationship was quite weak due to the decreased influence of LCRs at greater distances. This confirms the high substition rate in the vicinity of LCRs, as SNPs which have lower frequencies (i.e. are more recent) are more likely to be found near LCRs.

2.7 Codon Composition

There are several possible causes for the high substitution rate in LCRs which have been previously studied (Levinson and Gutman 1987, Amos et al. 2008, Moore et al. 1999), but the reason why this effect should extend into flanking regions has not yet been determined. One possibility is unequal recombination. Due to the repetitive nature of LCRs and their variability in length, alignment during crossover events is sometimes innaccurate. This can lead to unequal DNA sequences being exchanged between two strands of DNA. These recombination events can also involve DNA adjacent to the LCR, which would explain the high substitution rate of LCR flanking regions. As well, more indels (as seen in Section 2.4) would be expected in the case of unequal recombination, due to the exchange of sequences of differing lengths.

There is also an alternative possibility that these flanking regions were previously part of the LCR, created through the LCR's expansion. Later point mutations, which can increase the complexity of an LCR to the point of causing LCR death (Radó-Trilla and Albà 2012), might then change the sequence enough to obscure its origin within an LCR. The high substitution rate in regions flanking LCRs would then be a result of the LCR's own high substitution rate. As well, the natural propensity for expansion and contraction of LCRs would then still be expected to have an effect on regions on the verge of gaining complexity through point mutations, leading to the relationship between number of gaps seen and distance from the LCR.

If this or unequal recombination were affecting regions flanking LCRs, a high correlation would be expected between the composition of LCRs and their flanking regions, as these mechanisms would be incorporating sequences from LCRs into nearby DNA. To test this, the proportion of each primate LCR and associated flanking region composed of each codon was calculated. The values for LCRs and their flanking regions were tested for a correlation.

The composition of the regions in direct contact with LCRs is compositionally biased towards codons

which appear within the LCR. The three stop codons, TAA, TAG and TGA, were not seen in any LCR, and could not be tested for a correlation. While an enrichment of LCRs has been observed at the termini of proteins in *Drosophila* (Huntley and Clark 2007), using only long flanking regions meant that LCRs located near the end of a protein would not be considered, explaining the absence of these codons. There were significant correlations between the proportion of each LCR composed of each codon and the proportion of its flanking regions composed of the same codon for 46 of the 61 codons that could be tested, or for 75.4% of all codons. As well, when all the results were combined, the correlation between codon use in the LCRs and flanking regions was extremely significant (see Table 2.2). This is similar to the compositional bias of the flanking regions of microsatellites described by Vowles and Amos (2004). This bias could be due to the SEG algorithm not identifying the true end of the LCRs, but many of the correlations were extremely significant. It is possible that these strong correlations are the result of the LCR extending, and then point mutations changing the identity of some of the amino acids, causing the LCR to shrink back (Kruglyak et al. 1998). If this were true, this could be the mechanism for the high substitution rate in flanking regions.

There were many cases, however, where a codon would not be represented at all in the LCR. These examples could skew the results, especially in the case of rarer codons. To control for this, codon composition data was only used if the codon was present in the LCR. Using this filtered data, the codon compositions were again tested for correlations between LCRs and their associated flanking regions. This drastically reduced the number of codons which had a correlation to 11.67% (see Table 2.3).

To resolve the effect of LCR composition on the composition of associated flanking regions, the nucleotide compositions were also found. As any bias could easily be due to the composition of the genes hosting these regions, the nucleotide composition of the protein-coding sequences of these genes were also found. The correlations between nucleotide composition in LCRs and their flanking regions were compared to the correlations between nucleotide composition in the flanking regions and the entire gene. While there were significant correlations for all nucleotides between LCRs and flanking regions, the relationships between the flanking regions and genes were much stronger (see Table 2.4). This indicates that the composition of LCRs and flanking regions are likely indirectly correlated, and a result of overall biases in the proteins containing these regions. Based on these results, the mechanism causing the increased mutation rate in flanking regions is unlikely to involve segments of the LCR being integrated into surrounding DNA, as unequal recombination and LCR expansion followed by an increase in complexity through point mutations would entail.

2.8 Effect of Indels on SNPs

There is some debate over whether indels themselves are mutagenic, stemming from observations that SNPs seem to often be clustered around the sites of insertions and deletions (Tian et al. 2008, Longman-Jacobsen et al. 2003, Zhang et al. 2008). This can be linked to the substitution rate surrounding LCRs, as LCRs are more prone to indels. As well, a higher mutation rate has been seen in LCRs which have more variation in length (Amos et al. 2008). Because of this, it has been suggested that the apparent association between indels and substitutions may be due to the fact that indels are likely to be seen within LCRs (McDonald et al. 2011).

To test whether there was a stronger association between indels and an increase in SNPs than between

LCRs and an increase in SNPs, the full protein-coding sequence was aligned for each LCR-containing gene and all SNPs and indels were located. The DNA surrounding indels was treated as flanking regions, and each SNP was assumed to be most closeley associated to the nearest indel. As before with the number of substitutions, the number of SNPs per site was found for all available distances from the indels, although distance was in this case measured in nucleotides rather than codons. The distance from each indel to the nearest LCR was also found, so that any difference between flanking and non-flanking regions could be seen.

Like substitutions in regions flanking LCRs, there was a significant negative relationship between distance from indels and the proportion of genes containing a SNP at each site (Pearson's correlation = -0.2753, $p < 2.2 \times 10^{-16}$). This correlation was stronger when only considering data from indels that were within an LCR or a flanking region (Pearson's correlation = -0.3291). Conversely, the correlation was weaker in non-flanking regions (Pearson's correlation = -0.2337), although both relationships were extremely significant ($p < 2.2 \times 10^{-16}$). From these three relationships, it appears unlikely that the apparent association between indels and SNPs is causing the high substitution rate around LCRs. Not only is the relationship between distance from the LCR and number of substitutions much stronger than the relationship between distance from the nearest indel and the number of SNPs, but the correlation between indels and SNPs is weaker when the indels are not located within LCRs or flanking regions. Although the relationship between polymorphisms and indels within flanking regions may be entirely due to the higher substitution rate caused by the LCRs, indels may still have a mutagenic effect, as a correlation was also seen in non-flanking regions, however it is difficult to determine how many of the apparent SNPs detected were only seen due to misalignments. From this data, it cannot be definitively stated that indels do not increase the probability of substitutions in nearby DNA, but it is evident that LCRs have a much stronger mutagenic effect.

2.9 Complexity of Flanking Regions

While low-complexity regions proliferate and expand through mutations caused by their repetitiveness, the formation of LCRs happens through substitution mutations, until a sequence loses enough complexity that mutations resulting in the gain of repeated units become extremely likely (Loire et al. 2013). Because of this, it might be expected that regions surrounding LCRs may have a lower complexity than regions further away, as less complex regions would need fewer substitutions to become LCRs, and thus would be more prone to spawning LCRs. However, the fact that sequences outside of LCRs are technically above the threshold complexity used in detecting LCRs does not necessarily mean they are not subject to the same effects of low complexity. This threshold for detecting LCRs is based on the complexity at which mutations caused by low-complexity become extremely likely, but even above this threshold there may be a relationship between complexity and mutation rate. A relationship between distance from the LCR and complexity could therefore increase the substitution rate in flanking regions.

Using equation 1.1 to calculate complexity, flanking regions were analysed in sliding windows. The windows were segments 20 base pairs in length, as this allows for sequences to have the maximum possible complexity, preventing the data from being artificially skewed to seem less complex. As these windows covered a range of nucleotides, the distance from the LCR of each window was calculated as the distance from the LCR to the nearest nucleotide from the sliding window. The average complexity was found for

all available distances along the flanking region, and these data were tested for a correlation.

The complexity of the flanking regions was lowest at short distances from the LCR. The average complexity of the segments grew with more distance from the LCR until reaching a plateau approximately 250 base pairs away (see Figure 2.13). This negative relationship was extremely strong (Spearman's ρ = 0.8670, p < 10^{-16}). On its own, however, this correlation does not point to a relationship between complexity and mutation rate outside of LCRs.

To see whether complexity increases the mutation rate, making this relationship a potential factor in increasing the mutation rate of flanking regions, the number of changes to the amino acid sequence and the number of substitutions to the DNA sequence per segment were found. In finding the number of amino acid changes, the translated protein sequence was used and compared across the five primate species to find the number of differences. As gaps in the sliding windows from aligned proteins would reduce the complexity, windows were only used if there were no gaps. The DNA coding for each protein segments was retrieved and the number of substitutions was found as above in Section 2.3 using CodeML. Because complexity is low in the vicinity of LCRs while the substitution rate is high, it is likely that a correlation between complexity and the number of amino acid differences or substitutions would be significant even if the low complexity is not directly causing the increase in substitutions (i.e. both effects are caused by the presence of an LCR, and are not directly related). The data sets were divided based whether the segments were part of a flanking region or had more distance from the LCR. This ensured that any relationship seen between complexity and number of mutations in non-flanking regions was not skewed by an LCR.

The number of amino acid changes in the segments of the full protein sequence had a significant negative correlation with the complexity of the segments (Spearman's ρ = -0.1169, p < 10^{-16}). Some association between low complexity and a high number of amino acid changes is expected, however, since these two traits are spatially related. Using the distance of each segment to the nearest LCR, the data on complexity and amino acid changes were split based on whether they formed part of an LCR, a flanking region, or a non-flanking region. In flanking regions, the correlation was very similar to what was seen throughout the protein (ρ = -0.1114, p < 10^{-16}). This relationship was weak in non-flanking regions, but still significant (ρ = -0.0836, p < 10^{-16}). Although the correlation between complexity and number of amino acids changes is not as strong in non-flanking regions, it still indicates that the substitution rate increases as complexity decreases, despite the location within the protein-coding sequence. The decrease in complexity in the vicinity of LCRs is a likely contributor to the increase in substitution rate, although the weaker relationship in non-flanking regions does potentially point to other factors, as a stronger correlation would be expected if complexity was the only factor driving substitution rates up. This correlation was similar, but stronger when using only segments from within LCRs (Spearmans ρ = -0.1355, p < 10^{-16}), suggesting that the mechanisms affecting LCRs are shared with the flanking regions.

Using CodeML to find the number of substitutions in the coding DNA for the sliding windows, there was a negative correlation between the number of substitutions and the complexity of the translated protein sequence (Spearman's ρ = -0.1055, p < 10^{-16}). This relationship was consistent when considering only non-synonymous substitutions (ρ = -0.1255, p < 10^{-16}), or synonymous substitutions (ρ = -0.0447, p = 3.704×10^{-5}). This relationship holds true when only segments from flanking regions (within 500 amino acids of the LCR) were used (ρ = -0.1160, p < 10^{-16}). Like the correlations described above, the relationship was slightly stronger for non-synonymous substitutions (ρ = -0.1055, p = 3.778×10^{-14})

than synonymous substitutions (ρ = -0.0920, p = 4.133×10^{-11}), although both were significant.

Conversely, the relationship between number of substitutions and complexity in non-flanking regions was not significant (Spearman's ρ = -0.0083, p = 0.6569). However, when considering only non-synonymous substitutions, a stronger correlation was seen (ρ = -0.1131, p = 1.095×10^{-9}). This increase in non-synonymous substitutions as complexity decreases is consistent with the relationship between number of amino acid changes and complexity that is described above. Unexpectedly, a positive relationship was seen when considering only synonymous substitutions (ρ = 0.0888, p = 1.773×10^{-6}). Despite this, the relationship between number of non-synonymous substitutions and complexity does provide evidence that the complexity of sequences surrounding LCRs could be contributing to the high mutation rate in flanking regions.

The fact that there is a negative correlation between non-synonymous substitutions and complexity while there is a positive correlation between synonymous substitutions and complexity could imply that selection relaxes when complexity decreases. This is contrary to the evidence presented in Section 2.5, which indicated strong negative selection surrounding LCRs.

It is also possible that, outside of the influence of LCRs, low complexity is not favoured in proteins. This could reflect a drive towards either less mutable DNA sequences or a more defined 3-dimensional protein structure. If this were the case, non-synonymous substitutions would be more favoured when complexity was low in order to promote higher sequence complexity. Less complex regions would therefore be more likely to be under positive selection, and the average complexity of sequences under positive selection would be lower than the average complexity of sequences under negative selection.

The average complexity was found for windows 20 amino acids in length with significant evidence of positive selection and significant evidence of negative selection. Significance of the d_N/d_S ratio was found by simulating coding sequences 20 codons in length under neutral selection. A permutation test was used to randomly shuffle the values and find whether the difference between the average complexities when positive and negative selection were seen was significant. In general (i.e., using all protein segments, regardless of their location relative to LCRs), the average complexity of segments under positive selection was lower than regions under negative selection. Although this would be expected if complexity was favoured, the difference was not significant (Table 2.5). When segments were separated into based on their location relative to the nearest LCR (i.e., within LCRs, flanking regions or non-flanking regions), no significant difference was found in complexity between regions under positive and negative selection.

2.10 Function of LCR-containing Genes

The various functional categories of the protein products of the genes were downloaded from Gene Ontology (Consortium 2000) to see if the genes studied tended to be associated with any particular function. Each gene's entry was located in the database, and the different associated functions were recorded. The number of times each function was seen across all the LCR-containing genes was found. This search revealed no major trends in the specific function of the genes studied. The majority of the terms are found only once. The functions seen most often are extremely common, and likely to be shared among many randomly selected genes.

2.11 Epigenetic Effects

The locations of methylated CpG dinucleotides, as well as sites associated with acetylated histones and methylated histones, specifically methylated lysine 4 on histone 3 (H3K4me), were downloaded from the University of California Santa Cruz Genome Browser (Meyer et al. 2013). These epigenetic modifications illustrate the expression level of different genes, as CpG methylation prevents binding of transcription factors (Razin 1998), and acetylation of histones and H3K4me methylation facilitates transcription-factor binding (Grunstein 1997, Strahl et al. 1999). The distance of each methylated or histone-associated site to the nearest LCR was found, and the density of these sites in relation to LCRs was plotted.

Despite the opposing effects of the modifications studied, CpG methylation, histone acetylation and H3K4me histones were all apparently clustered around LCRs (see Figures 2.14 - 2.16). To find the expected distribution, two random sets of locations were generated: one set was randomly sampled from throughout the entire genome, while the other was taken only from genes in which LCRs were found. The distance from the modifications to the nearest random location was found. While the distribution of CpG methylation, histone acetylation and H3K4me histones appeared completely random when using the locations from throughout the genome, when using locations from LCR-containing genes the distribution looked like what was seen with LCRs (see Figures 2.17 - 2.19). As there was such a large number of genes found to contain LCRs, with a wide variety of functions (see Section 2.10), it is likely that these genes also have a wide variety of expression patterns. The association of LCRs and LCR-containing genes with epigenetic modifications that both hinder and facilitate transcription indicates that LCRs are not linked to any particular level of expression.

2.12 Conclusions

It is clear that substitution rates are greatly increased not just in LCRs, but also in nearby DNA, and that this effect extends hundreds of base pairs beyond the LCR. A strong correlation between distance from an LCR and number of differences between sequences was seen not only when examining substitution rates, but also when studying indels. There is evidence that the propensity for mutations in flanking regions is due to unequal recombination in the relative codon composition of LCRs and their flanking regions. Average d_N/d_S ratios indicate that the regions around LCRs may be tightly controlled, as negative selection appears to be far more common than positive selection. Of the various mechanisms of mutation investigated, only the decrease in complexity surrounding LCRs seemed likely to be causing the increased substitution rate in proximity to LCRs.

2.13 Tables

	Human	Chimpanzee	Gorilla	Orangutan	Macaque
Human	7387	6908	6776	6635	6662
Chimpanzee	–	7208	6635	6611	6565
Gorilla	–	–	7098	6461	6412
Orangutan	–	–	–	6957	6414
Macaque	–	–	–	–	6454

Table 2.1: Number of 1-to-1 orthologous LCRs shared between each pair of species.

Codon	Correlation coefficient	p-value
AAA	0.1939	2.259×10^{-6}
AAC	-0.0273	0.5089
AAT	0.1173	0.004473
AAG	0.1214	0.003251
ACA	0.04914	0.2349
ACC	0.02069	0.6172
ACT	0.1011	0.01438
ACG	0.02228	0.5904
ATA	-0.05006	0.2262
ATC	0.0601	0.146
ATT	0.0315	0.4459
ATG	-0.0083	0.841
AGA	0.2393	4.43×10^{-9}
AGC	0.0920	0.02590
AGT	0.2676	4.575×10^{-11}
AGG	0.1854	6.271×10^{-6}
CAA	0.04918	0.2345
CAC	0.0559	0.1765
CAT	0.0379	0.3593
CAG	-0.0419	0.3108
CCA	0.0427	0.3025
CCC	-0.00779	0.8507
CCT	0.02419	0.5589
CCG	0.04347	0.2934
CTA	0.03596	0.3849
CTC	0.0735	0.07542
CTT	0.06680	0.1062
CTG	0.2157	1.338×10^{-7}
CGA	0.1058	0.01037
CGC	0.040687	0.3255
CGT	0.03525	0.3943
CGG	0.0954	0.02095
TAC	0.01086	0.793
TAT	0.0945	0.02209
TCA	0.1543	0.0001779
TCC	0.0982	0.01737
TCT	0.2415	3.212×10^{-9}
TCG	0.0964	0.01962
TTA	0.0962	0.01989
TTC	0.02121	0.6083
TTT	0.03726	0.368
TTG	0.0705	0.08802
TGC	0.0755	0.06795
TGT	0.145	0.0004296
TGG	0.08412	0.0418
GAA	0.145	0.0004252
GAC	0.0249	0.5467
GAT	0.1054	0.01067
GAG	0.0798	0.05339
GCA	-0.0170	0.681
GCC	-0.00275	0.947
GCT	-0.03803	0.3582
GCG	0.036006	0.3843
GTA	-0.007167	0.8625
GTC	0.03575	0.3877
GTT	0.08268	0.04544
GTG	-0.01265	0.7598
GGA	0.04837	0.2423
GGC	0.1228	0.002904
GGT	-0.01141	0.7828
GGG	-0.0194	0.6391

Table 2.2: Correlation between codon representation in LCRs and their associated flanking regions.

Codon	Correlation coefficient	p-value	Codon	Correlation coefficient	p-value
AAA	0.2535	0.1053	CGG	-0.0650	0.7106
AAC	-0.1758	0.5308	TAC	0.1193	0.7597
AAT	0.2676	0.2409	TAT	-0.002789	0.9935
AAG	0.20448	0.1380	TCA	0.2478	0.07357
ACA	0.06915	0.6886	TCC	0.03346	0.7897
ACC	0.011059	0.9425	TCT	0.3692	0.001309
ACT	0.1627	0.3735	TCG	0.2346	0.2388
ACG	-0.1576	0.5322	TTA	0.22847	0.5255
ATC	0.716	0.005895	TTC	0.3896	0.09918
ATT	-0.1913	0.5513	TTT	-0.02308	0.9463
ATG	0.273697	0.1855	TTG	0.2651	0.2105
AGA	0.0182	0.9266	TGC	0.084137	0.8057
AGC	0.1094	0.3707	TGT	0.13967	0.7004
AGT	0.4656	0.0005046	TGG	0.8328	0.002782
AGG	0.5301	0.002158	GAA	0.0336	0.792
CAA	0.3914	0.06479	GAC	0.2084	0.1694
CAC	0.3721	0.06703	GAT	0.1757	0.2913
CAT	0.32437	0.2579	GAG	0.1075	0.3519
CAG	0.1161	0.3879	GCA	0.019896	0.8932
CCA	0.0629	0.624	GCC	-0.02622	0.814
CCC	0.1376	0.2490	GCT	0.12286	0.3539
CCT	0.106	0.386	GCG	0.5281	0.0003248
CCG	0.1254	0.4172	GTA	0.2957	0.5196
CTA	0.20047	0.5114	GTC	0.13858	0.5184
CTC	0.308	0.05296	GTT	0.0309	0.9095
CTT	-0.25847	0.2122	GTG	0.0145	0.9195
CTG	0.3546	0.002413	GGA	0.0350	0.8192
CGA	0.0235	0.9215	GGC	0.3105	0.007089
CGC	-0.0045	0.9833	GGT	0.04756	0.7894
CGT	-0.064968	0.841	GGG	0.0658	0.6266

Table 2.3: Correlation between codon representation in LCRs and their associated flanking regions when disregarding data where a codon is not represented in either the LCR or in the surrounding DNA.

Nucleotide	Correlation between flanking region and LCR	Correlation between flanking region and protein
A	0.5070	0.6635
C	0.4152	0.6206
G	0.3343	0.5415
T	0.2726	0.5395

Table 2.4: Correlation between nucleotide representation in LCRs and their associated flanking regions, compared to the correlation between flanking regions and the rest of the protein. All correlations were extremely significant ($p < 2.2 \times 10^{-16}$).

	Average Complexity of Windows Under Positive Selection	Average Complexity of Windows Under Negative Selection	p-value
All windows	2.1847	2.2354	0.9993
Flanking regions	2.2096	2.2251	0.9997
Non-flanking regions	2.2488	2.3258	0.9995
LCRs	1.9097	1.7659	0.9981

Table 2.5: Comparison of the average complexity of protein sequence segments (20 bp windows) with significant evidence of positive or negative selection

2.14 Figures

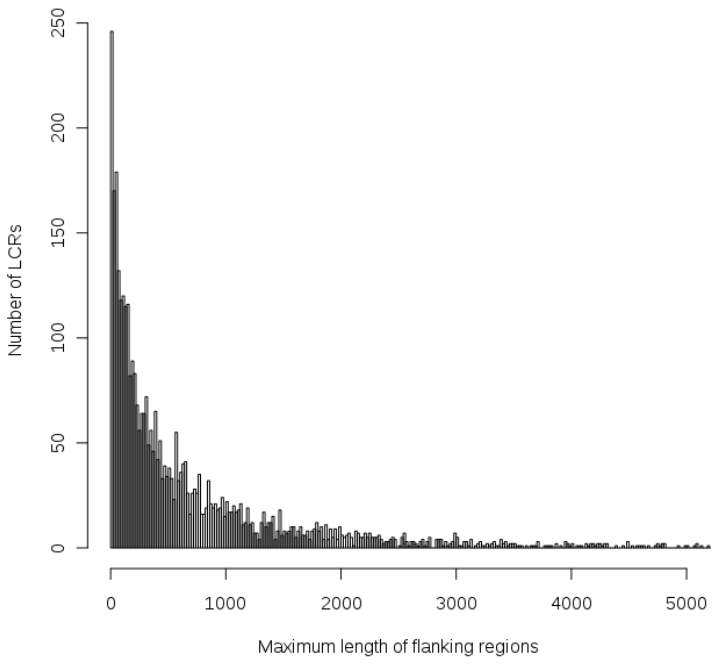


Figure 2.1: Maximum lengths of potential flanking regions of LCRs, accounting for the protein termini and midway point to the nearest LCR.

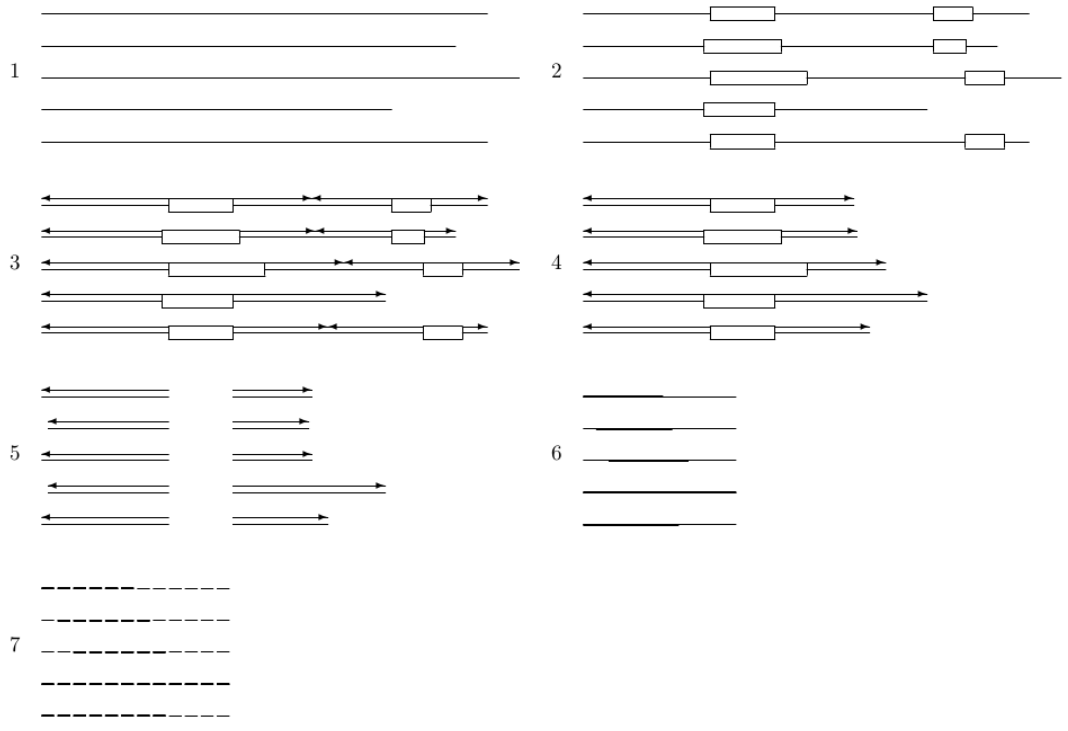


Figure 2.2: Data workflow — (1) Homologous proteins found for five primate species; (2) LCRs identified using SEG; (3) Maximum length of flanking sequences determined by protein termini and midpoints between two LCRs; (4) Flanking regions filtered for examples with homologous sequences from all five species. Since the second LCR in this example is present in only four species, its flanking regions are not used; (5) The 3′ and 5′ flanking sequences are considered separately, so that if all five homologous sequences are not available, the other can still be used; (6, continuing with the 3′ sequence) The upstream and downstream flanking regions are aligned separately; gaps are represented with thin lines; (7) Alignments of individual codons are used to find the number of substitutions at each site for each flanking region with CodeML. Codons with gaps (in this case codons 1, 2 and 7 through 12) are not useable, and are not considered when calculating the average number of substitutions per site. The number of substitutions was found for all useable sites of all genes found to contain LCRs, and the average across all useable sites was found for all positions relative to the LCR (i.e., codon 1, codon 2, etc.).

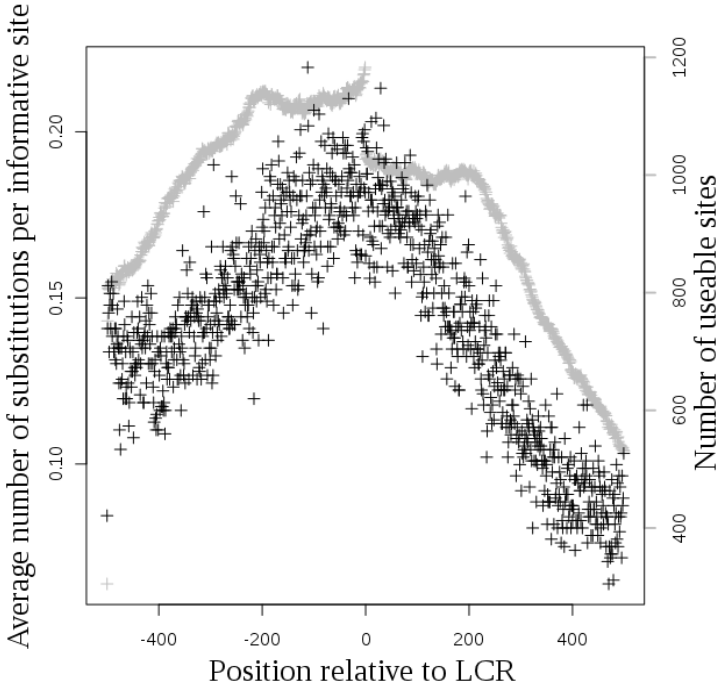
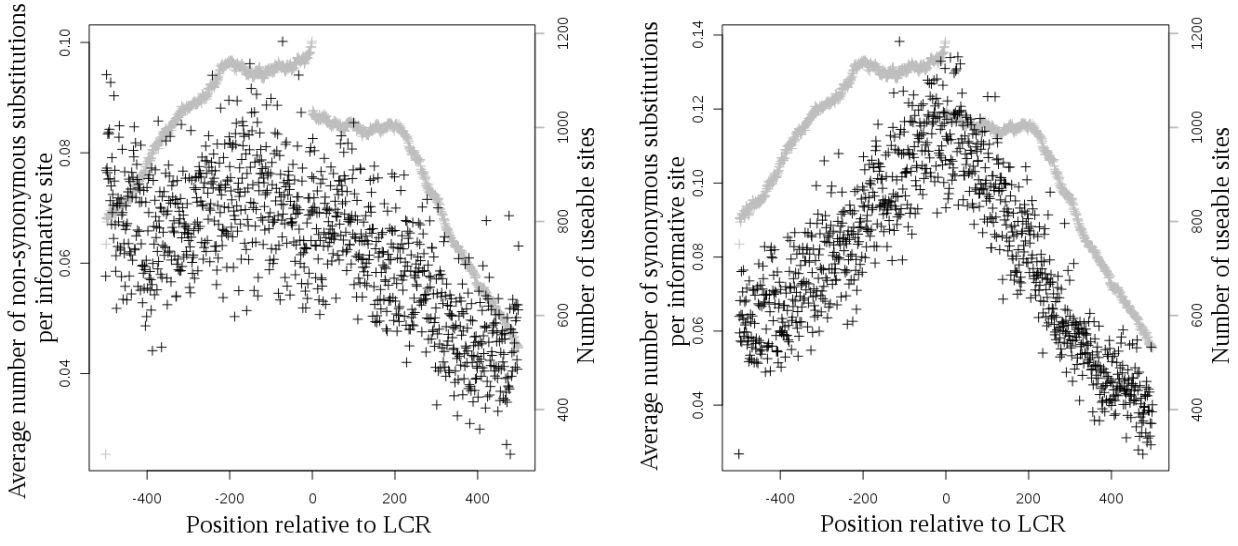


Figure 2.3: Distance for LCR vs. average substitution rate.



(a) Distance from LCR vs. average non-synonymous substitution rate

(b) Distance from LCR vs. average synonymous substitution rate

Figure 2.4: Effect of distance from LCR on average substitution rate of each codon in five primate species. Grey points indicate N, the number of genes which could provide information and were free of gaps at each site. Negative values are upstream of the LCR.

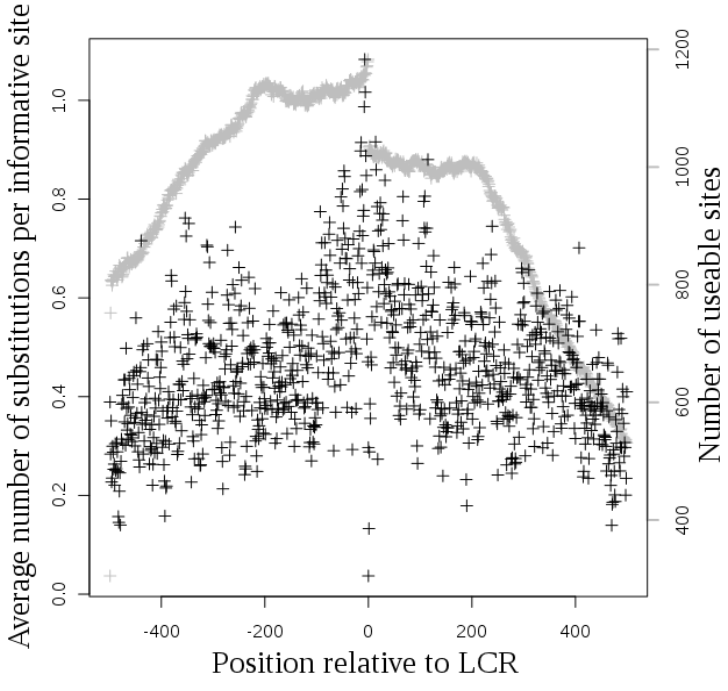
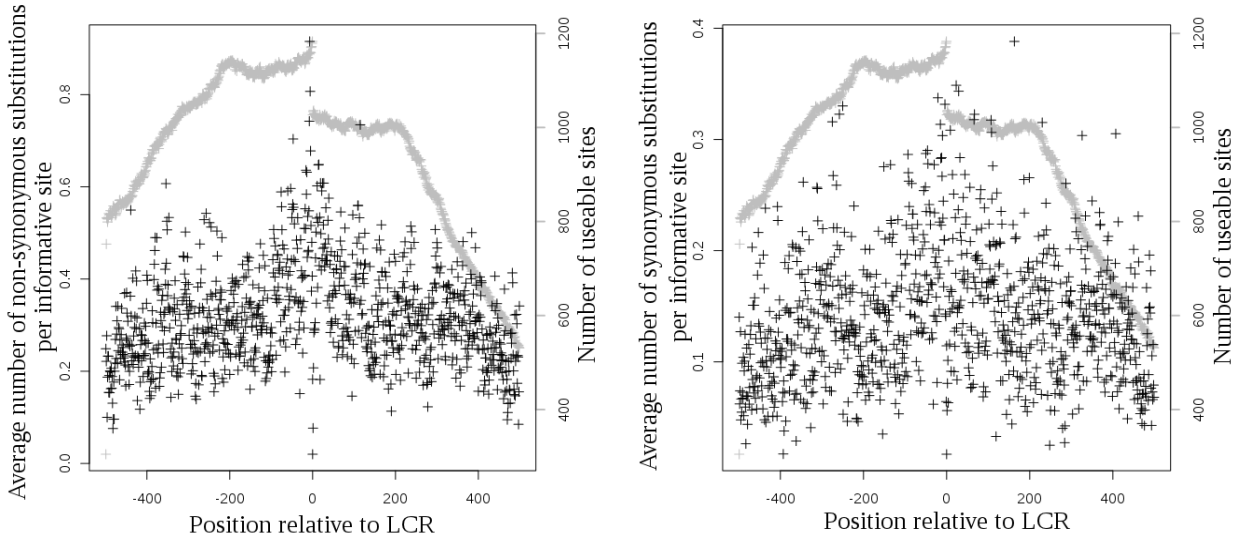


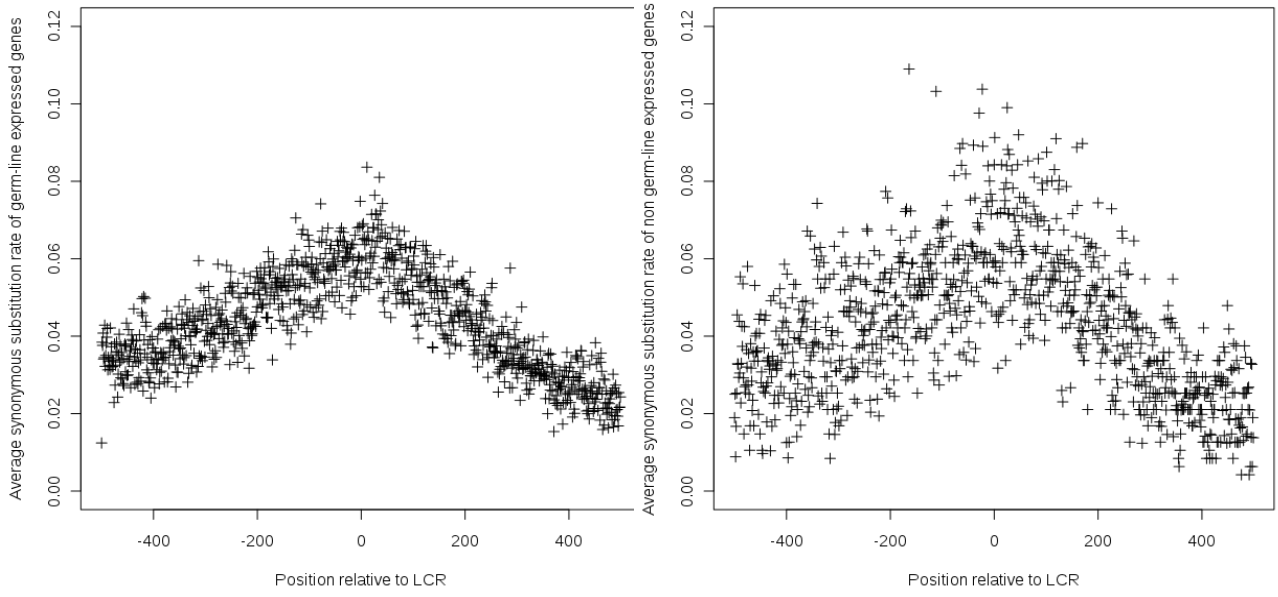
Figure 2.5: Effect of distance from LCR on average rate of substitutions per codon in humans. Grey points indicate N, the number of genes which could provide information and were free of gaps at each site. Negative values are upstream of the LCR.



(a) Distance from LCR vs. average non-synonymous substitution rate

(b) Distance from LCR vs. average synonymous substitution rate

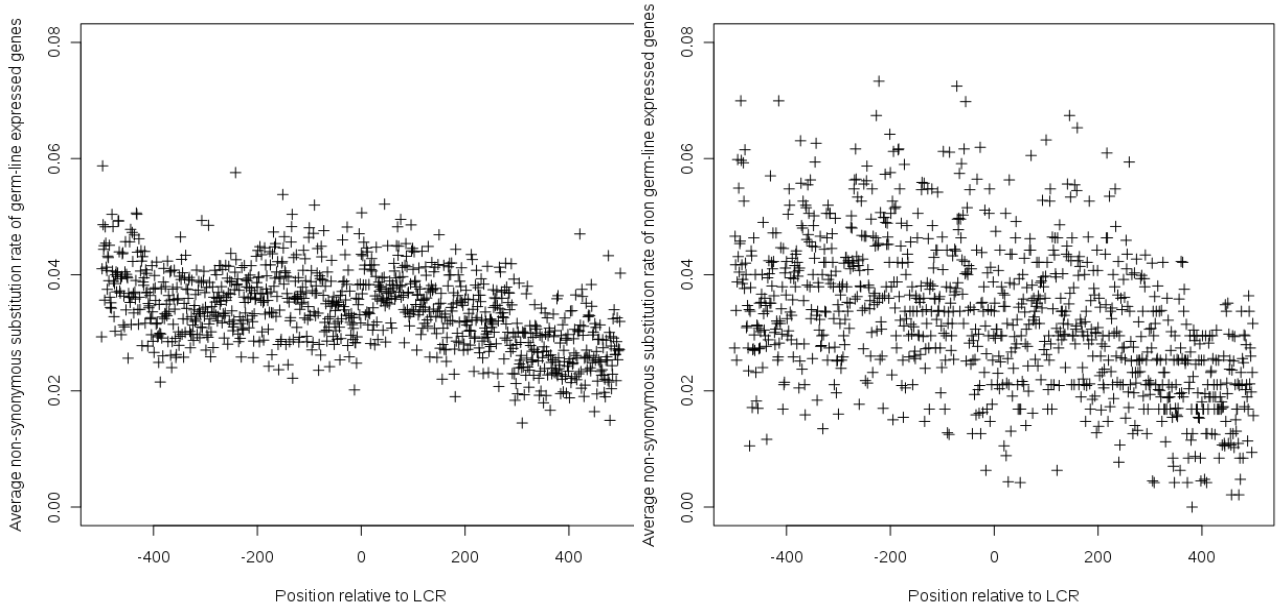
Figure 2.6: Effect of distance from LCR on average rate of synonymous and non-synonymous substitutions for each codon in humans. Grey points indicate N, the number of genes which could provide information and were free of gaps at each site. Negative values are upstream of the LCR.



(a) Distance from LCR vs. average synonymous substitution rate in germ-line expressed genes

(b) Distance from LCR vs. average synonymous substitution rate in non germ-line expressed genes

Figure 2.7: Effect of distance from LCR on average synonymous substitution rate of each codon in germ-line expressed and non germ-line expressed genes from five primate species. Negative values are upstream of the LCR, 1385 upstream flanking regions and 1399 downstream flanking regions were used.



(a) Distance from LCR vs. average non-synonymous substitution rate in germ-line expressed genes

(b) Distance from LCR vs. average non-synonymous substitution rate in non germ-line expressed genes

Figure 2.8: Effect of distance from LCR on average non-synonymous substitution rate of each codon in germ-line expressed and non germ-line expressed genes from five primate species. Negative values are upstream of the LCR, 1385 upstream flanking regions and 1399 downstream flanking regions were used.

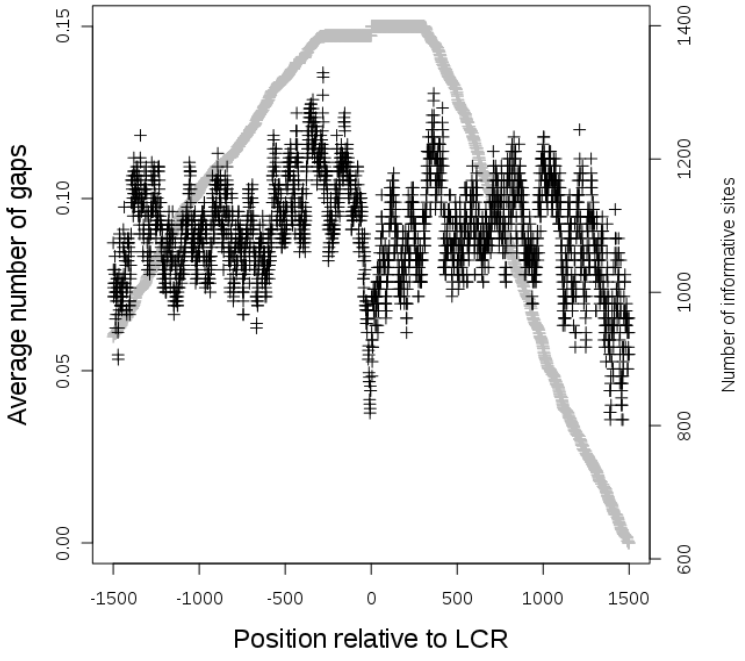


Figure 2.9: Effect of distance from the LCR on average number of indels at each nucleotide site in five primate species. Grey points indicate N, the number of genes which could provide information on each site. Negative values are upstream of the LCR.

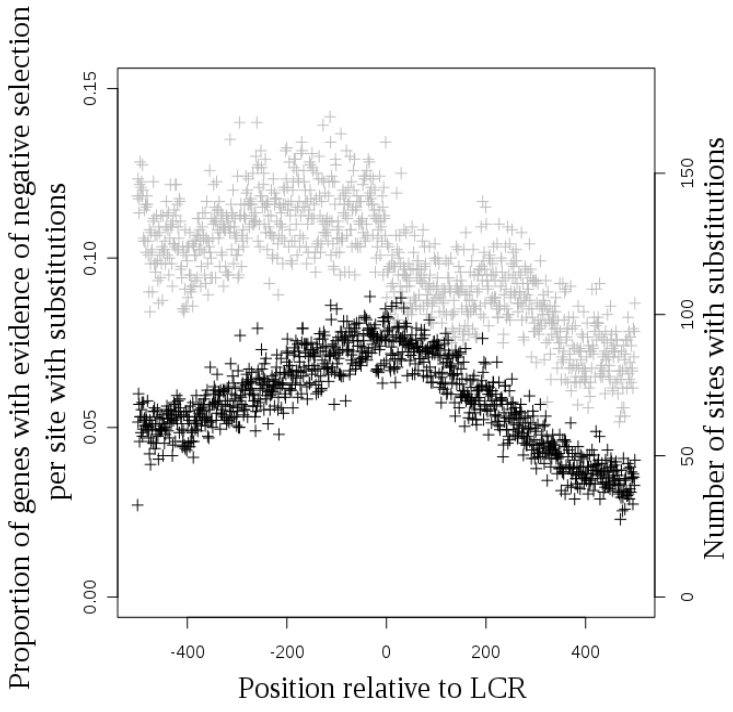


Figure 2.10: Effect of distance from LCR on proportion of flanking regions which had evidence for negative selection ($\omega < 1$). Grey points indicate the number of genes with had substitutions which could be used to calculate d_N/d_S. Negative values are upstream of the LCR.

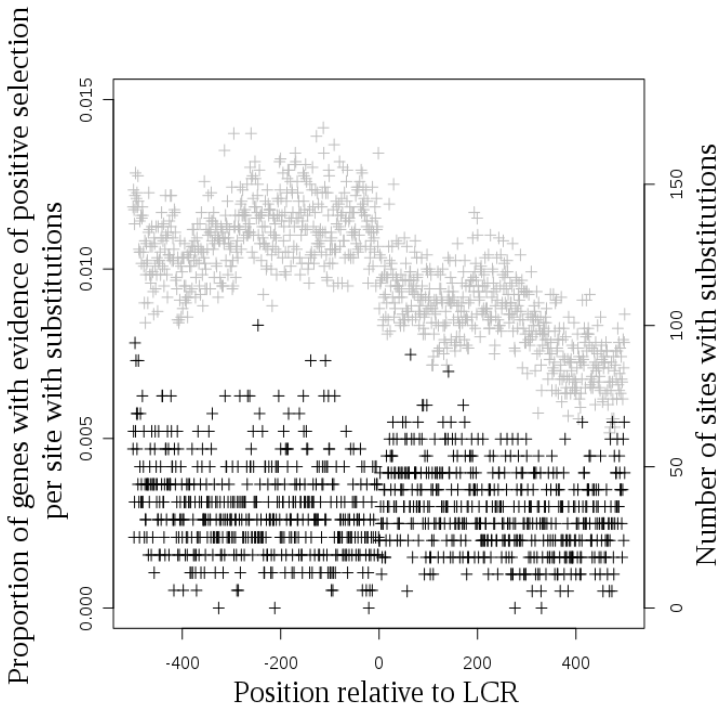


Figure 2.11: Effect of distance from LCR on proportion of flanking regions which had evidence for positive selection ($\omega > 1$). Grey points indicate the number of genes with had substitutions which could be used to calculate d_N/d_S. Negative values are upstream of the LCR.

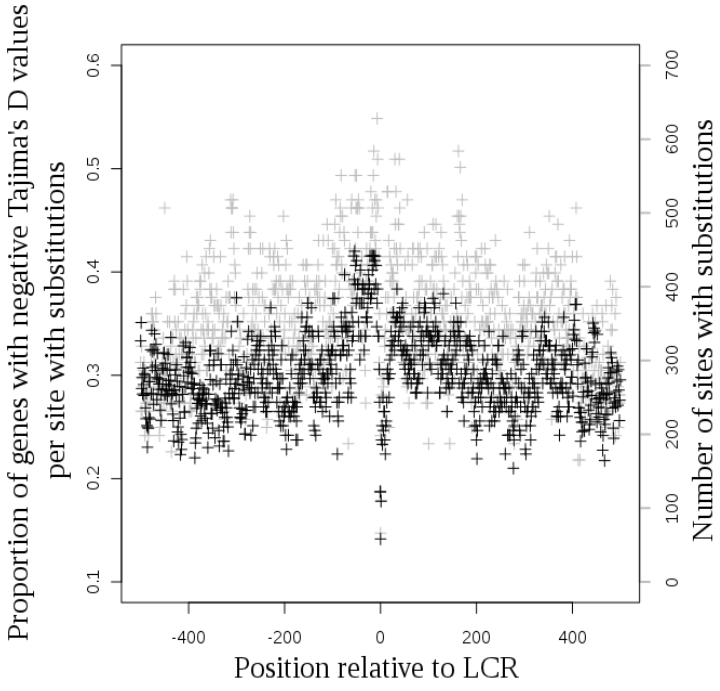


Figure 2.12: Effect of distance from the LCR on proportion of genes which had negative values for Tajima's D at each codon. Grey points indicate the number of genes with substitutions which could be used to calculate Tajima's D at each site. Negative values are upstream of the LCR.

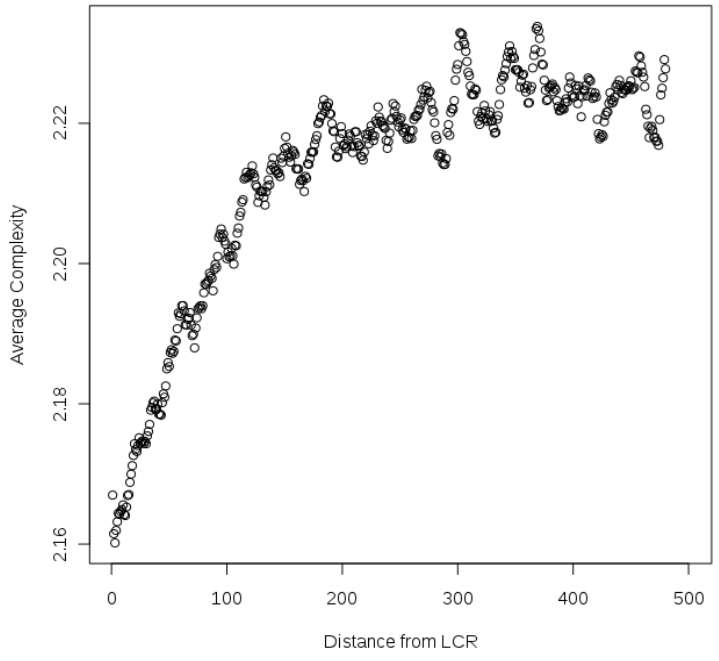


Figure 2.13: Effect of distance from the LCR on average complexity of short (20-base pair long) windows.

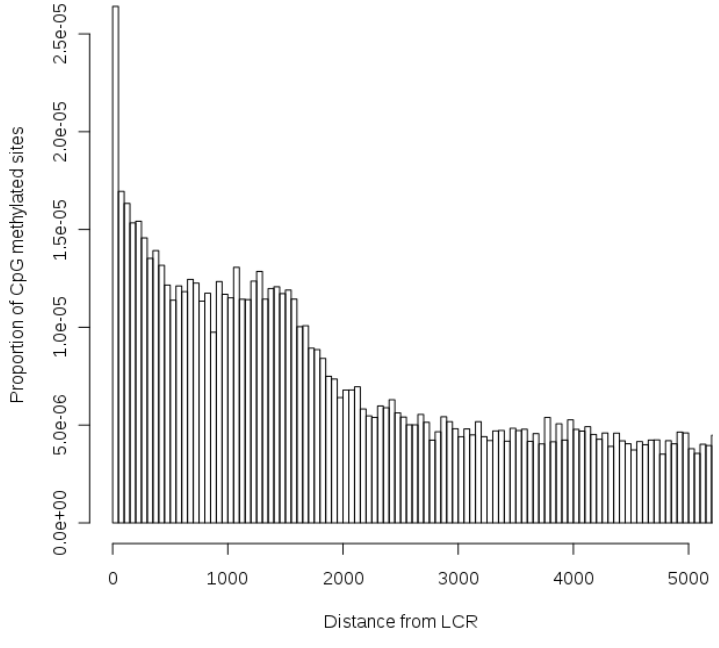


Figure 2.14: Distribution of methylated CpG dinucleotides around LCRs.

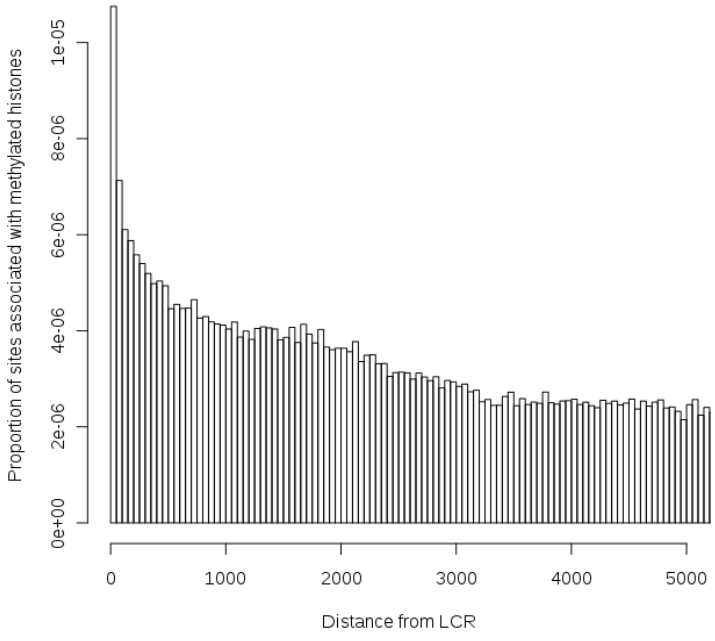


Figure 2.15: Distribution of sites associated with H3K4me histones around LCRs.

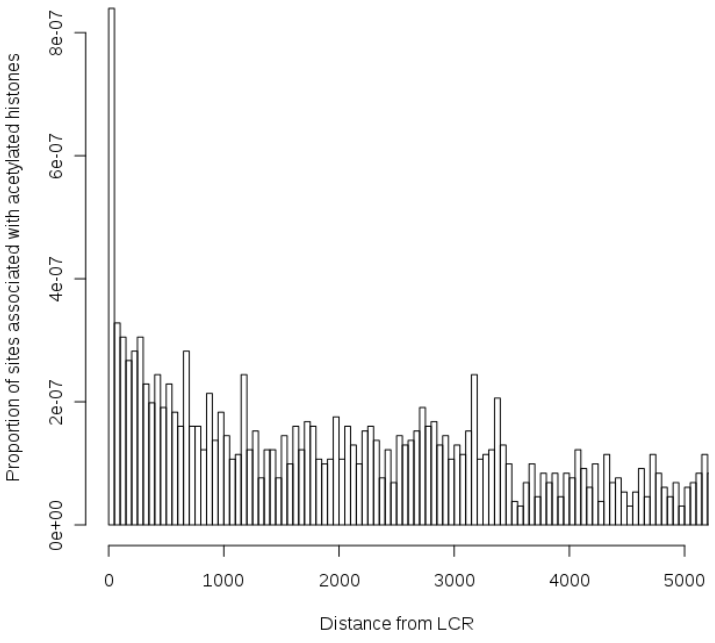


Figure 2.16: Distribution of sites associated with acetylated histones around LCRs.

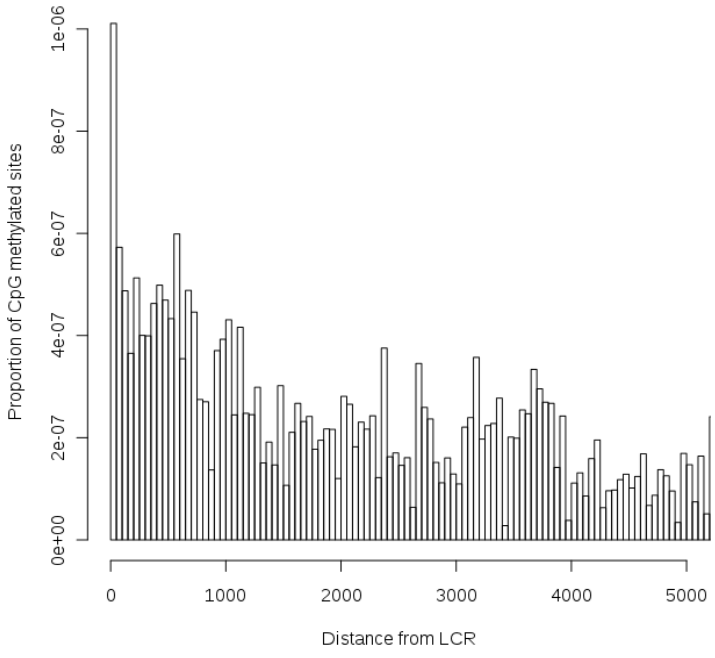


Figure 2.17: Distribution of methylated CpG dinucleotides around randomly chosen sites within LCR containing genes.

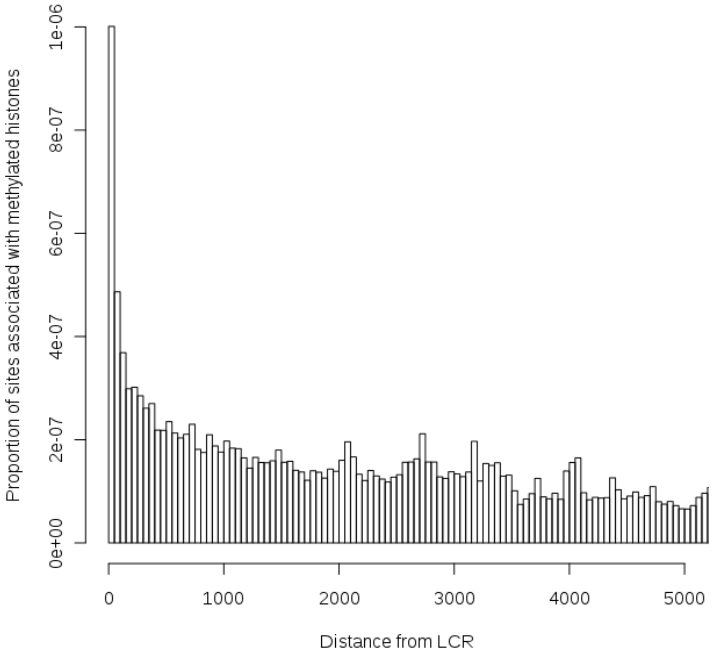


Figure 2.18: Distribution of sites associated with H3K4me histones around randomly chosen sites within LCR containing genes.

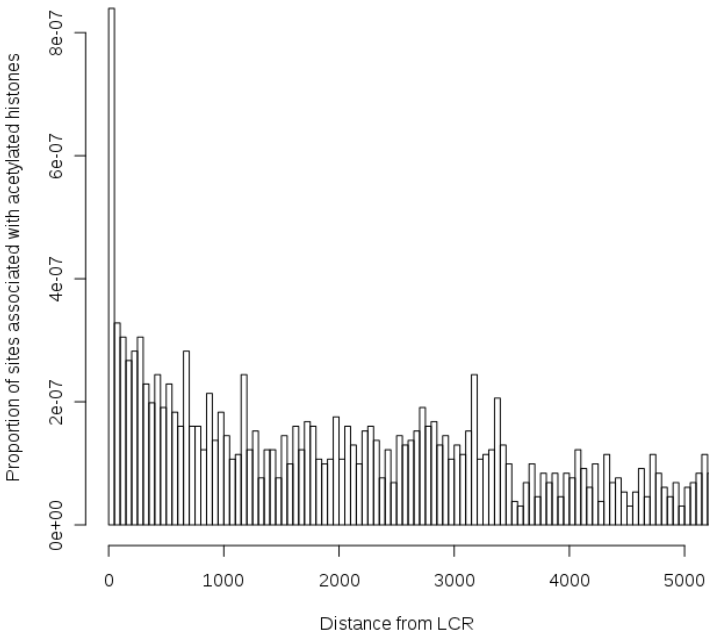


Figure 2.19: Distribution of sites associated with acetylated histones around randomly chosen sites within LCR containing genes.

Bibliography

W Amos, J Flint, and X Xu. Heterozygosity increases microsatellite mutation rate, linking it to demographic history. *BMC Genetics*, 9:72, 2008.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

RC Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32:1792–97, 2004.

P Flicek, MR Amode, D Barrell, K Beal, S Brent, Y Chen, P Clapham, G Coates, S Fairley, S Fitzgerald, L Gordon, M Hendrix, T Hourlier, N Johnson, A Kähäri, D Keefe, S Keenan, R Kinsella, F Kokocinski, E Kulesha, P Larsson, I Longden, W McLaren, B Overduin, B Pritchard, HS Riat, D Rios, GRS Ritchie, M Ruffier, M Schuster, D Sobral, G Spudich, YA Tang, S Trevanion, J Vandrovcova, AJ Vilella, S White, SP Wilder, A Zadissa, J Zamora, BL Aken, E Birney, F Cunningham, I Dunham, R Durbin, XM Fernández-Suarez, J Herrero, TJP Hubbard, A Parker, G Proctor, J Vogel, and SMJ Searle. Ensembl 2011. *Nucleic Acids Research*, 39:D800–D806, 2011.

M Grunstein. Histone acetylation in chromatin structure and transcription. *Nature*, pages 349–352, 1997.

W Haerty and GB Golding. Increased Polymorphism Near Low-Complexity Sequences across the Genomes of Plasmodium falciparum Isolates. *Genome Biology and Evolution*, 3:539–550, 2011.

R Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338, 2002.

MA Huntley and AG Clark. Evolutionary Analysis of Amino Acid Repeats across the Genomes of 12 Drosophila Species. *Molecular Biology and Evolution*, 24:2598–2609, 2007.

S Kruglyak, R T Durrett, M D Schug, and CF Aquadro. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 95:10774–10778, 1998.

G Levinson and GA Gutman. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, 4:203–221, 1987.

E Loire, D Higuet, P Netter, and G Achaz. Evolution of coding microsatellites in primate genomes. *Genome Biology and Evolution*, 2013.

N Longman-Jacobsen, JF Williamson, RL Dawkins, and S Gaudieri. In polymorphic genomic regions indels cluster with nucleotide polymorphism: Quantum Genomics. *Gene*, 312:257 – 261, 2003.

MJ McDonald, WC Wang, HD Huang, and JY Leu. Clusters of Nucleotide Substitutions and Insertion/Deletion Mutations Are Associated with Repeat Sequences. *PLoS Biol*, 9:e1000622, 2011.

LR Meyer, AS Zweig, AS Hinrichs, D Karolchik, RM Kuhn, M Wong, CA Sloan, KR Rosenbloom, G Roe, B Rhead, BJ Raney, A Pohl, VS Malladi, CH Li, BT Lee, K Learned, V Kirkup, F Hsu, S Heitner, RA Harte, M Haeussler, L Guruvadoo, M Goldman, BM Giardine, PA Fujita, TR Dreszer, M Diekhans, MS Cline, H Clawson, GP Barber, D Haussler, and WJ Kent. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Research*, 41:D64–D69, 2013.

H Moore, PW Greenwell, CP Liu, N Arnheim, and TD Petes. Triplet repeats form secondary structures that escape DNA repair in yeast. *Proceedings of the National Academy of Sciences*, 96:1504–1509, 1999.

P Polak, R Querfurth, and Arndt PF. The evolution of transcription-associated biases of mutations across vertebrates. *BMC Evolutionary Biology*, 10, 2010.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

N Radó-Trilla and M Albà. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. *BMC Evolutionary Biology*, 12, 2012.

A Razin. CpG methylation, chromatin structure and gene silencinga three-way connection. *The EMBO Journal*, pages 4905–4908, 1998.

KJ Siddle, JA Goodship, B Keavney, and MF Santibanez-Koref. Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome. *Bioinformatics*, 27:895–898, 2011.

BD Strahl, R Ohba, RG Cook, and CD Allis. Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in Tetrahymena. *Proceedings of the National Academy of Sciences*, 96:14967–14972, 1999.

F Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123:585–95, 1989.

D Tian, Q Wang, P Zhang, H Araki, S Yang, M Kreitman, T Nagylaki, R Hudson, J Bergelson, and JQ Chen. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature*, 455, 2008.

M Uhlen, P Oksvold, L Fagerberg, E Lundberg, K Jonasson, M Forsberg, M Zwahlen, C Kampf, K Wester, S Hober, H Wernerus, L Bjrling, and F Ponten. Towards a knowledge-based Human Protein Atlas. *Natural Biotechnology*, 28:1248–1250, 2012.

E.J. Vowles and W. Amos. Evidence for widespread convergent evolution around human microsatellites. *PLoS biology*, 2:e199, 2004.

JC Wootton and S Federhen. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry*, 17:149 – 163, 1993.

Z Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24: 1586–1591, 2007.

Z Yang and JP Bielawski. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15:496 – 503, 2000.

W Zhang, X Sun, H Yuan, H Araki, J Wang, and D Tian. The pattern of insertion/deletion polymorphism in Arabidopsis thaliana. *Molecular Genetics and Genomics*, 280:351–361, 2008.

Chapter 3

Intron-interrupted LCRs

3.1 Focus

As described in Chapter 1, the formation and proliferation of LCRs is likely often due to their repetitive sequences (Levinson and Gutman 1987). Despite this, several LCRs have been observed to have a introns within their underlying coding regions. As a relatively pure string of repeats is needed for extension through the gain of repeated units, the presence of an intron in the DNA of the LCR should have prevented the formation of the LCR by making slippage impossible over the long disruptions that introns would cause. As well, many of these intron-interrupted LCRs are nearly identical to LCRs which are not interrupted in other species, contradicting the idea that these LCRs could be the result of two short LCRs forming on opposite sides of an intron. It appears that either LCRs sometimes form despite the presence of an intron, or that introns can somehow change location to appear within an already-formed LCR.

LCRs containing introns within their coding sequences were located. The exons in all proteins studied were randomly re-ordered and searched for intron-interrupted LCRs to show that the number these LCRs that naturally occur is more than would be expected if they were the result of coincidence. The average complexity of these LCRs was compared to that of uninterrupted LCRs and found to be significantly lower; the average complexity of proteins that contained intron-interrupted LCRs was also found to be significantly lower than that of average LCR-containing proteins.

3.2 Identifying Intron-interrupted LCRs

Representative genomes of *Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla*, *Pongo pygmaeus* and *Macaca mulatta* were downloaded from Ensembl release 62 (Flicek et al. 2011). The coding sequences were translated into proteins which were searched for regions of low-complexity using SEG (Wootton and Federhen 1993), with a window size of 15 and complexity threshold of 1.9. 7387 LCRs were detected in total.

Locations of splice sites were downloaded from GenBank (Benson et al. 2006). The identified LCRs

Human
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS
Chimpanzee
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS
Gorilla
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS
Orangutan
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS
Rhesus macaque
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS

Figure 3.1: Low-complexity region from PRP4. Red sequence data is from exon 2, blue sequence data is from exon 3.

were searched to see if any contained splice sites. The genes hosting LCRs identified as having intron-interrupted coding sequences were scrutinized to make sure the splice site data from GenBank was accurate by aligning cDNA from Ensembl to the entire DNA sequence for each gene, to locate probable exons and introns.

Once intron-interrupted LCRs were identified, available homologous genes from a variety of mammal species were downloaded from Ensembl and searched for LCRs. The mammal LCRs were searched for splice sites using the same methods as the primate LCRs.

Initially 83 of the 5584 total LCRs (only including LCRs for which there were homologous examples in all five primate species) were found to have coding regions interrupted by introns, however many of these had a negligible amount of amino acids of the translated protein sequence separated from the majority of the LCR (see Section 5 starting on page 59, Figures S1 through S83 for a full list of all LCRs initially detected as intorn-interrupted). Although SEG detected these regions as one discrete LCR, the probability of seeing amino acids that appear, by chance, to be a continuation of an LCR onto a separate exon is not remote, especially if the portion of the sequence that appears within a separate exon is extremely short. Even if the coding region of the LCR is divided between multiple exons more evenly, its detection as an LCR may not indicate it should be considered as one unit, and instead may simply be the result of two separate LCRs that have coincidentally formed around the same intron. Although the different origins of the two LCRs should increase the complexity, making them less likely to be identified by SEG as one LCR, a bias in amino acid composition could very easily result in two independent, although apparently similar LCRs. This is supported by the fact that the 83 LCRs found were in only 59 unique genes, suggesting that genes that are especially prone to the formation of LCRs are more likely to contain interrupted LCRs. Even if the LCRs had little similarity between sequences, if they both have extremely low complexity, they may still have a complexity below SEG's threshold when conjoined.

Many LCRs, however, defy these explanations. While some examples appear discontinuous, as described above, other examples have been found where substantial segments of the LCR are on two or even three separate exons, and both sides clearly have repeated motifs present. One such LCR is found in pre-mRNA processing factor 4 homolog B (PRP4), as seen in Figure 3.1. Another example can be found in scaffold attachment factor B2 (SAFB2, see Figure 3.2).

Another interesting example is an extremely long and repetitive LCR found in transcription elongation factor 1. In humans and most of the other primates studied, this LCR is uninterrupted, however in

Human
EQRERERQRQREREIRETERRREREQREREQR
Chimpanzee
EQRERERQRQREREIRETERRREREQREREQR
Gorilla
EQRERERQRQREREIRETERRREREQREREQR
Orangutan
EQRERERQRQREREIRETERRREREQREREQR
Rhesus macaque
EQRERERQRQREREIRETERRREREQREREQR

Figure 3.2: Low-complexity region from SAFB2. Red sequence data is from exon 14, blue sequence data is from exon 15.

the orangutan and other various mammals the coding region for this LCR is on three separate exons. A sampling of the species with this LCR is shown in Figure 3.3. There is some variation not just in the presence and absence of splice sites but also in the location of these splice sites. In this case, this is probably due to the expansion and contraction of this LCR, as it is extremely repetitive and likely prone to gain and loss of repeat units. The appearance of splice sites in some species but not others could be the result of mistaken splice site identification or sequencing errors, however a similar pattern is shared by several different species. Although the translated protein sequence is very repetitive, the underlying DNA is composed of an array of codons, and the d_N/d_S ratio for this particular LCR, found using CodeML (Yang 2007), was less than one, indicating that this region may be under some selective pressure preventing significant changes to the sequence. If this LCR has some function within the gene, extensions to it that happened to form on the opposite side of an intron might be selected for, making it appear as though the LCR extends across several exons. Although transcription elongation factor 1 is a well documented gene, its listing in GenBank lacks any mention of this LCR as a feature, and its presence does not appear to have been remarked upon in literature concerning this gene.

3.3 Exon Shuffling

The probability of seeing intron-interrupted LCRs by chance was tested by repeatedly shuffling the order of exons within a protein, to find how often new LCRs were created through the splicing of two exons. Each protein was re-ordered 1000 times by randomly switching the placement of exons, before SEG was used to determine if any new LCRs were present.

The proportion of naturally occurring intron-interrupted LCRs (including LCRs for which homologous LCRs were not present in other species or were not interrupted by an intron) per protein studied was over twice as high as the proportion of interrupted LCRs found by randomly changing the order of exons (0.0724 vs 0.0311, respectively). While some intron-interrupted LCRs are expected to form through random chance, the actual number of intron-interrupted LCRs seen is much higher than would be expected if each of these LCRs were due to coincidence. As well, of the proteins which did not naturally contain an intron-interrupted LCR, the majority ($\sim$77%) produced no new intron-interrupted LCRs through randomly re-ordering exons. Most of the intron-interrupted LCRs were generated using proteins which had LCRs located at one end of an exon, making them extremely prone to giving rise to intron-interrupted LCRs.

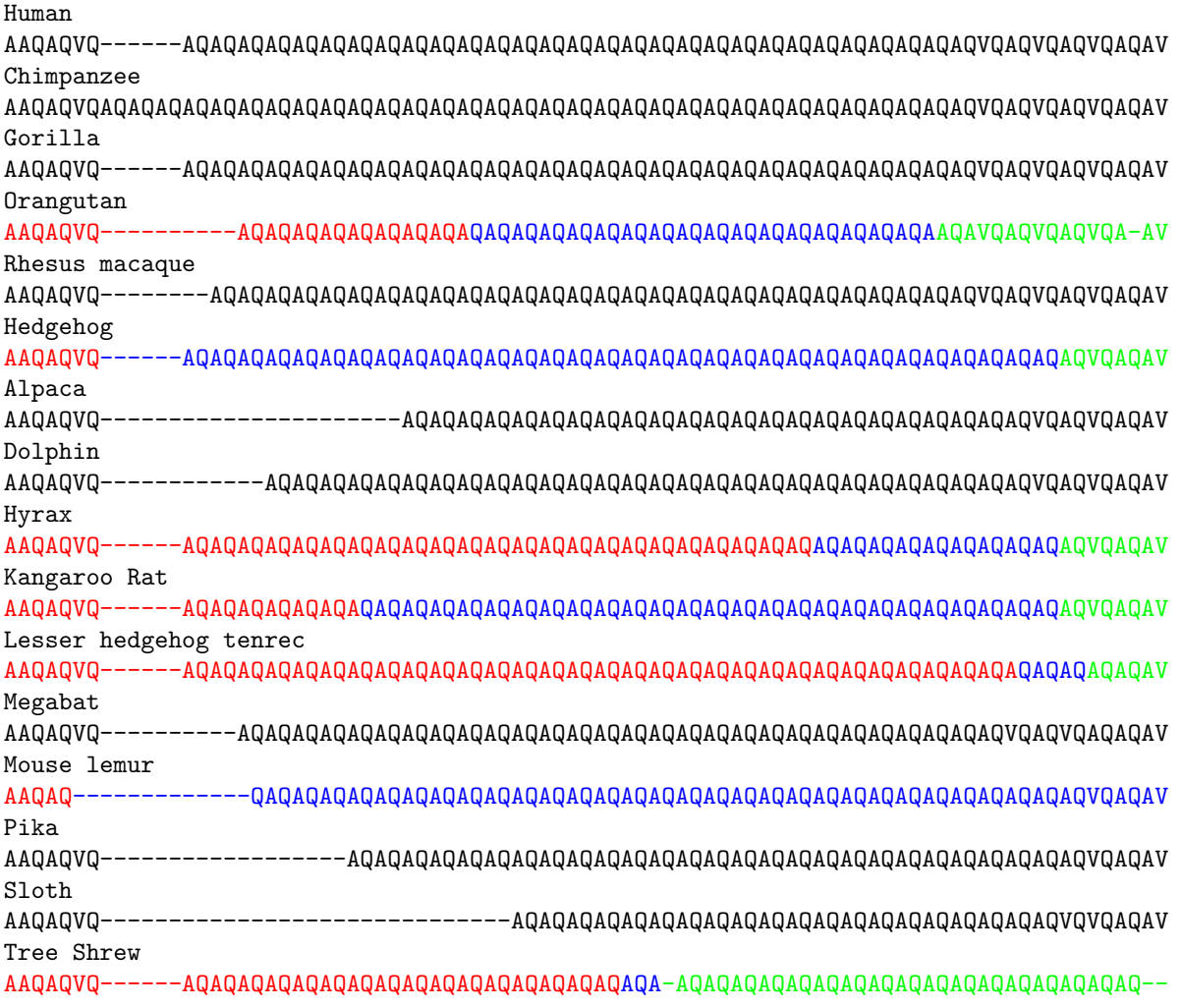


Figure 3.3: Low-complexity region from transcription elongation regulator 1. Black sequences are from uninterrupted LCRs, while red, blue and green indicate segments coded by different exons.

Human
QGTQQGQQQQQPQQQQHP---
Chimpanzee
QGTQQGQQQQQQQQQQHP---
Gorilla
QGTQQGQQQQQPQQQQHP---
Dolphin
QGTQQGQQQQQQQQQQQHPQH
Hyrax
PGTQQGQQQQQQQQQQHPPQH
Kangaroo Rat
QGTQQGQQQQQQQQQQHPPQH
Lesser hedgehog tenrec
QGSQQGQQQQQQQPQQQHPQH
Mouse lemur
QGTQQGQQQQQQQQQQHPSQH
Pika
PGAQQGQQQQQQQQQQHPAQH
Shrew
QGTQQGQQQQQQQQQQHPPQH
Tarsier
QGTQQGQQQQQQQQQQQPPQH
Tree shrew
QGTQQGQQQQQQQQQQQHPTQH

Figure 3.4: Low-complexity region from transcription factor 20. Black sequences are from uninterrupted LCRs, while red and blue indicate segments coded by different exons.

Additionally, intron-interrupted LCRs generated by randomly changing the order of exons were, on average, significantly more complex than naturally occurring interrupted LCRs (1.4257 vs 1.3822 respectively, $p < 10^{-7}$). Presumably, if an intron-interrupted LCR is formed by coincidence, the randomly included sequences from separate exons are not likely to have identical codon compositions, and should increase the complexity of the LCRs. As a higher complexity is expected in intron-interrupted LCRs which are truly the result of a random process, the significantly lower complexity of extant intron-interrupted LCRs is another indication that this phenomenon is not due to chance.

3.4 Intron movement

Although the changing position of the intron in transcription elongation regulator 1 is likely due to gain or loss of repeated motifs, this explanation for the variation in intron location within an LCR is not always sufficient. An example of this is found in transcription factor 20 (see Figure 3.4), which has an LCR which is interrupted in many non-primate species. The location of the intron interrupting the coding sequence varies between species, despite the fact that the LCR remains the same length. As well, when the intron disrupts the less repetitive segments of the LCR, the sequence still remains almost exactly the same. More cases of apparent intron movement can be found in section 5, for example Figures S2 and S10. If the apparent movement of splice sites were due to expansion and contraction on either side of the intron, the more unique parts of the LCR would be expected to disappear, rather than appear on a different exon.

If the intron-interrupted LCRs are not the result of two coincidentally similar LCRs forming on opposite sides of an intron, then this implies that the LCR formed as one unit before an intron became present in its coding sequence. A study of novel intron positions in *Drosophila* found new introns could form from tandem duplications of proto-splice sites (Lehmann et al. 2009). As this was only seen in repetitive regions, this could be a factor in LCRs, however a comparison of a wide range of mammalian introns and exons showed no evidence that intron gain occurs in mammals (Coulombe-Huntington and Majewski 2007). As well, although there are several documented cases of intron movement (Brenner and Corrochano 1996, Lehmann et al. 2009), these all involve movement of a few base pairs, much shorter distances than the examples of intron interrupted LCRs would indicate. Another possibility is recombination, which could displace part of an LCR, however the evolutionary history of intron interrupted LCRs has yet to be understood.

3.5 Complexity and tree lengths

All homologous LCRs were compared, not only for whether they were interrupted by introns at the same sites, but also to estimate complexity, k, calculated as

$$k = -\sum_{i=1}^{20} \frac{n_i}{L} \left(\log \left(\frac{n_i}{L} \right) \right)$$

and branch lengths, estimated using protdist then Fitch (Felsenstein 1989). Branch lengths and complexity were also found for LCRs that were not interrupted by introns and used for comparison.

Complexity was found to be, on average, lower in intron-interrupted LCRs (1.3177 and 1.4633, respectively). A permutation test indicated that these results are extremely significant ($p < 10^{-7}$). As well, the LCRs interrupted by introns produced shorter trees on average than the uninterrupted LCRs (0.0011 and 0.0190, respectively, $p = 0.0070$ using a permutation test).

As the tree lengths estimated by Fitch only take substitutions into account, the ratio of gaps to the rest of the sequence was found for alignments of the LCRs, in order to account for expansion and contraction mutations. On average, $\sim 12.8\%$ of sites in the interrupted LCR sequence contained gaps, while only $\sim 4.1\%$ contained gaps in uninterrupted LCRs. Using a permuation test, these averages were significantly different ($p < 10^{-7}$).

Taken together with the tree length data, these results indicate that intron interrupted LCRs are more prone to indels, but less prone to substitutions than uninterrupted LCRs. This could be due to the lower complexity of the interrupted LCRs, as a less complex region will be more repetitive, increasing the chance of slippage mutations, and higher complexity may be the result of substitutions. When looking at uninterrupted LCRs, complexity was in fact found to have a significant positive correlation with tree length (correlation coefficient = 0.1231, $p = 8.187 \times 10^{-11}$) and a weaker, although still significant, negative relationship with proportion of gaps (correlation coefficient = -0.0482, p = 0.01122) using Pearson's product-moment correlation (R Development Core Team 2011).

Neither of these relationships, however, are significant in intron interrupted LCRs. As several of the mechanisms that lead to the high mutation rate in LCRs are due to their repetition, the intron could

easily be disrupting these processes. For example, mutations through replication slippage should be less common in interrupted LCRs, as there is likely to be two short segments of DNA with a small number of repeated units separated by an intron, rather than one long string of a larger number of repeated units. This is supported by the fact that, when only segments of LCRs that can be found on one exon are considered (i.e. continuous portions of intron interrupted LCRs), there is a nearly significant negative relationship between complexity and number of gaps (correlation coefficient = -0.167, p = 0.07685). However, while the correlation between complexity and tree length was slightly stronger, it was still not significant.

3.6 Amino Acid Composition

Although some of the intron-interrupted LCRs found have obvious differences between the sequences on either side of the intron, many of the intron-interrupted LCRs have sequences which are very similar between fragments. The proteins containing LCRs could be more prone to giving rise to certain types of repetitive sequences if they had a bias towards certain amino acids in their sequence.

Using all 5584 LCRs for which there were five homologous sequences (including but not limited to intron-interrupted LCRs), the proportion of each LCR which was made up of each amino acid was calculated. The amino acid proportions of each protein containing an LCR were also found. The proportions were compared for a correlation between the amino acid compositions of LCRs and the proteins containing them, to check if the composition of LCRs was likely to be determined by the composition of the proteins in which they were found.

Overall, there was a significant correlation between the composition of LCRs and proteins (Pearson's correlation = 0.3545, $p < 10^{16}$). However, separating the data based on the type of amino acid, significant relationships were found for only 2 amino acids, phenylalanine and threonine (Table 3.1), suggesting that the composition of proteins does not have a strong influence on the composition of LCRs that may form within them. However, significant correlations between amino acid composition of LCRs and proteins was found using only proteins containing intron-interrupted LCRs for 18 out of 20 amino acids (Table 3.2).

The fact that a correlation in amino acid composition was only found for proteins which contained intron-interrupted LCRs could indicate that these proteins have a greater bias towards some amino acids. This can be confirmed by comparing the average complexity of proteins containing intron-interrupted LCRs and proteins containing non-interrupted LCRs, as an over-representation of some amino acids would decrease the complexity. Calculated using equation 1.1, the average complexity was significantly lower for proteins containing intron-interrupted LCRs than non-interrupted LCRs (2.849 vs 2.878, respectively, $p = 1.6 \times 10^{-4}$ using a permutation test). Since these proteins have a lower complexity, they may be more prone to giving rise to LCRs. This would increase the probability that two of these LCRs will be located on opposite sides of an intron.

3.7 Alternative Splicing

Since the length of some LCRs has been linked to morphological characteristics in dogs (Fondon and Garner 2004) and behavioural characteristics in voles, chimpanzees and humans (Hammock and Young 2005), some variation over the length of an LCR between different protein products of the same gene could be advantageous. Intron-interrupted LCRs could allow some control over the length of LCRs through alternative splicing.

The amino acid sequences of all protein products of each gene that contained an intron-interrupted LCR was downloaded from Ensembl (Flicek et al. 2011) for each of the five primate species. These protein sequences were searched to see if they contained the LCR in its entirety, or only a short fragment. 2708 transcripts of these coding sequences were found. Of these, 1268 contained the entire LCR encoding sequence, while 1420 did not contain any fragments of the LCR or did not produce a protein product. 6 of the coding transcripts which contained only part of the LCR encoding region were flagged on Ensembl as affected by nonsense mediated decay.

The remaining 14 protein products which contained only a segment of the intron-interrupted LCR were examined. Several proteins contained LCRs where the sequences that were seen on separate exons do not appear to be very similar, for example Figures S7 and S8. Other transcripts ended with one segment of the LCR; these were all labelled on Ensembl as having incomplete coding sequences. In all, the protein products which contained only a partial intron-interrupted LCR did not appear to have variation in the length of the LCR between different alternative splicings. Rather, they are likely the result of a few sequences which perhaps should not be labelled as one LCR and truncated protein transcripts.

3.8 Conclusions

Intron-interrupted LCRs are a fascinating and unusual phenomenon. Since the mutation mechanisms that affect LCRs and allow them to expand and proliferate are related to the repetition of the LCR coding regions, they should not be able to extend across an intron. This raises the question of whether the LCRs formed first and introns somehow became part of the coding region, or the LCRs coincidentally formed around the introns.

As there are no known cases of intron formation in mammals, it is highly unlikely that new splice sites were formed within the LCRs. However, a few cases have been noted where introns have changed position. In all documented cases of intron movement, the splice sites have only migrated by a few base pairs, but presumably larger distances are possible given long periods of time. Conversely, as LCRs are prone to expansion and contraction, the apparent movement of intron could be due to the gain of repeat units on one side of the intron followed by the loss of repeat units on the other side, or vice versa.

In many cases it appears fairly likely that intron-interrupted LCRs are the result of two separate LCR formation events. Several examples of these LCRs have quite obvious differences in composition between the two segments separated by the intron. As well, in some cases where the separated fragments of the LCR appear more similar, the proteins themselves may have a bias in their amino acid composition. At the same time, the average complexity of these LCRs is significantly lower than the average for all

LCRs studied. This would seem to suggest that the intron-interrupted LCRs cannot all be considered coincidental. At the same time, however, the lower complexity, on average, of proteins which contain intron-interrupted LCRs could make these proteins more likely to give rise to LCRs, increasing the probability of LCR formation. If a protein contains more LCRs, it is more likely that two regions of low complexity would coincidentally be located on either side of one intron.

3.9 Figures

Amino Acid	Pearson's correlation coefficient	p-value
Alanine (A)	0.0.0239	0.09748
Arginine (R)	-0.0023	0.872
Asparagine (N)	-0.0074	0.6074
Aspartic acid (D)	0.0243	0.09257
Cysteine (C)	0.0053	0.7126
Glutamic acid (E)	0.0127	0.3806
Glutamine (Q)	0.0279	0.05369
Glycine (G)	-0.0137	0.3429
Histidine (H)	-0.0100	0.4874
Isoleucine (I)	-0.0157	0.2776
Leucine (L)	0.0246	0.08883
Lysine (K)	0.0114	0.4295
Methionine (M)	0.0205	0.1555
Phenylalanine (F)	0.1467	$< 10^{-16}$
Proline (P)	0.0250	0.0828
Serine (S)	0.0129	0.3724
Threonine (T)	0.0324	0.02479
Tryptophan (W)	-0.0049	0.7328
Tyrosine (Y)	0.0033	0.8183
Valine (V)	0.0144	0.3175
Combined	0.3545	$< 10^{-16}$

Table 3.1: Correlations between composition of uninterrupted LCRs and the proteins in which they were found

Amino Acid	Pearson's correlation coefficient	p-value
Alanine (A)	0.3051	$< 10^{-16}$
Arginine (R)	0.5389	$< 10^{16}$
Asparagine (N)	0.1127	0.00042
Aspartic acid (D)	0.1144	0.00035
Cysteine (C)	0.1196	0.00018
Glutamic acid (E)	0.2953	$< 10^{-16}$
Glutamine (Q)	0.3765	$< 10^{-16}$
Glycine (G)	0.5419	$< 10^{-16}$
Histidine (H)	0.1821	$1.045 \times 10{-}8$
Isoleucine (I)	0.0316	0.3238
Leucine (L)	0.3590	$< 10^{-16}$
Lysine (K)	0.2915	$< 10^{-16}$
Methionine (M)	0.5050	$< 10^{-16}$
Phenylalanine (F)	0.0906	0.004665
Proline (P)	0.4405	$< 10^{-16}$
Serine (S)	0.5200	$< 10^{-16}$
Threonine (T)	0.0785	0.01429
Tryptophan (W)	0.0012	0.9702
Tyrosine (Y)	0.3769	$< 10^{-16}$
Valine (V)	0.2802	$< 10^{-16}$
Combined	0.4852	$< 10^{-16}$

Table 3.2: Correlations between composition of intron-interrupted LCRs and the proteins in which they were found

Bibliography

DA Benson, I Karsch-Mizrachi, DJ Lipman, J Ostell, and Wheeler DL. GenBank. *Nucleic Acids Research*, 2006.

S Brenner and LM Corrochano. Translocation events in the evolution of aminoacyl-tRNA synthetases. *Proc Natl Acad Sci USA*, pages 8485–8489, 1996.

J Coulombe-Huntington and J Majewski. Characterization of intron loss events in mammals. *Genome Research*, 17:23–32, 2007.

J Felsenstein. PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics*, 5:164–166, 1989.

P Flicek, MR Amode, D Barrell, K Beal, S Brent, Y Chen, P Clapham, G Coates, S Fairley, S Fitzgerald, L Gordon, M Hendrix, T Hourlier, N Johnson, A Kähäri, D Keefe, S Keenan, R Kinsella, F Kokocinski, E Kulesha, P Larsson, I Longden, W McLaren, B Overduin, B Pritchard, HS Riat, D Rios, GRS Ritchie, M Ruffier, M Schuster, D Sobral, G Spudich, YA Tang, S Trevanion, J Vandrovcova, AJ Vilella, S White, SP Wilder, A Zadissa, J Zamora, BL Aken, E Birney, F Cunningham, I Dunham, R Durbin, XM Fernández-Suarez, J Herrero, TJP Hubbard, A Parker, G Proctor, J Vogel, and SMJ Searle. Ensembl 2011. *Nucleic Acids Research*, 39:D800–D806, 2011.

JW Fondon and HR Garner. Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences*, 101:18058–18063, 2004.

EAD Hammock and LJ Young. Microsatellite Instability Generates Diversity in Brain and Sociobehavioral Traits. *Science*, 308:1630–1634, June 2005.

J Lehmann, C Eisenhardt, PF Stadler, and Krauss V. Novel intron positions in Drosophila are mostly caused by intron sliding and tandem duplications. *German Conference on Bioinformatics 2009*, pages 25 – 30, 2009.

G Levinson and GA Gutman. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, 4:203–221, 1987.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

JC Wootton and S Federhen. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry*, 17:149 – 163, 1993.

Z Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24: 1586–1591, 2007.

Chapter 4

Summary

4.1 Flanking Regions

The site-by-site analysis of protein-coding regions flanking LCRs revealed a clear relationship between distance from the nearest LCR and the number of substitutions and gaps. This confirms the effect of LCRs on flanking regions seen in *Drosophila* by Huntley and Clark (2007) and in *Plasmodium falciparum* by Haerty and Golding (2011), as well as demonstrating that this is likely a common characteristic across many genomes, rather than a trait only seen in some species. It also demonstrates that LCRs do not need to be as repetitive as homopolymers, which Siddle et al. (2011) showed to have a similar effect on adjacent DNA, for this to be seen. Not only does the presence of LCRs have a profound effect on the number of indels and substitutions in primates, but there was significant evidence that regions flanking LCRs were more conserved. The increased number of point mutations was also confirmed through the relationship between distance from the nearest LCR and the frequency of derived SNPs.

Although it is known some LCRs are linked to disease and can have damaging effects (Mangiarini et al. 1997, Pearson and Sinden 1996, Usdin 2008), the mutability of the regions flanking LCRs has not been studied for its potential relationship to disease. Because the effect of LCRs was found to extend for hundreds of base pairs, the ramifications of a high rate of indels and substitutions in these areas could be far-reaching. This major source for change within genes might generate disease alleles, but it would also mean that LCR-containing proteins are evolving at a much faster rate than proteins without LCRs. It has been theorized that LCRs allow some flexibility within the genome in order to quickly adapt (Vinces et al. 2009); if this is the case, the effect of LCRs on nearby DNA provides a significant source of adaptability as well.

Of the several mechanisms outlined which could cause this effect, the lower complexity in the DNA surrounding LCRs appears the most likely. Other potential mechanisms were contradicted by different aspects of the flanking regions. For example, the weaker correlations between composition of LCRs and their associated flanking regions refutes any mechanism, such as unequal recombination, that involves sequence from the LCR being incorporated into flanking regions. As well, the evidence for negative selection around LCRs makes relaxed selection highly unlikely.

Because other mechanisms seem unlikely, and because complexity has a significant relationship with

the number of substitutions throughout the protein, the high mutability of flanking regions could be due to the effects of low-complexity, albeit at a weaker level than within the LCRs. Although these regions are not likely to be repetitive enough for slippage mutations to have an effect, the low complexity could result in the formation of secondary structures which prevents DNA repair mechanisms. If the lower complexity of flanking regions is the cause of their instability, then the high substitution rate is not caused by the LCRs themselves, but rather is a lingering effect of the processes underlying LCR formation. As well, it suggests that, even in proteins which do not contain LCRs, similar effects may be seen as mutability varies with complexity.

4.2 Intron-Interrupted LCRs

The random re-ordering of exons indicates that, although some intron-interrupted LCRs are expected to form by chance, their prevalence in primates is higher than expected. This random shuffling also suggested that some proteins are more prone to the formation of intron-interrupted LCRs. As proteins which contain intron-interrupted LCRs have an overall lower complexity than the average, the biased composition of these proteins may cause a general greater likelihood of LCR formation, increasing the probability that LCRs will form at or near the borders of exons.

Although intron-interrupted LCRs appear more conserved than other LCRs, the intron within their coding regions is likely to disrupt the mutation mechanisms which commonly affect LCRs, making them more static. Despite this, the conservation of the sequence of intron-interrupted LCRs in cases where the intron has shifted position is still unexplained. The conserved sequence is usually only a few base pairs, however it may suggest evidence for intron movement, a phenomenon which is sometimes seen (Brenner and Corrochano 1996, Lehmann et al. 2009), but usually only consists of movement by a few base pairs. The examples where intron-interrupted LCRs have homologous uninterrupted LCRs in separate species are also interesting. Intron formation has not been observed in mammals (Coulombe-Huntington and Majewski 2007), which would suggest that this is either the result of extreme cases of intron movement, or misidentification of introns, however alignments between cDNA and the DNA sequence would seem to contradict the latter.

4.3 Future Directions

The increased probability of mutations seen in proximity to LCRs has been documented in several species, but the relationship between complexity and distance from an LCR described here has not. As well, the correlation between substitution rate and complexity, even above the complexity threshold at which the mechanisms of mutation which affect LCRs become much more active, has also not been commented upon. The complexity of regions flanking LCRs may cause their increased substitution rate, but it must be better studied in order to be confirmed. The relationship between complexity and substitution rate also raises the question of the mechanism through which complexity affects coding sequences outside of LCRs.

Another interesting question is whether complexity is advantageous outside of LCRs. Since the number of non-synonymous substitutions has a more significant relationship with complexity than syn-

onymous substitutions, it may be possible that less complex sequences are under positive selection, pushing them towards higher complexity. Alternatively, it may simply be that more complex sequences are more conserved.

The existence of intron-interrupted LCRs seems likely to be coincidence in many cases, but the higher number of these LCRs than expected, as well as the low complexity of proteins containing these introns, indicates that some proteins are more prone to the formation of intron-interrupted LCRs than others. This could be confirmed by investigating whether intron-interrupted LCRs are more common in species with less complex genomes, such as *Plasmodium falciparum*. As well, the apparent movement of introns is still unexplained. The strong conservation of amino acid sequences, even when the positions of splice sites are changed, merits a closer look in order to see whether this provides evidence for intron movement in mammals. A closer examination of the coding regions of LCRs where the intron presence or position is variable between species may provide an explanation. For example, splice sites may be similar to the repeat units of some intron-interrupted LCRs; expansion could then result in new repeat unit being mistaken for a splice site.

Bibliography

S Brenner and LM Corrochano. Translocation events in the evolution of aminoacyl-tRNA synthetases. *Proc Natl Acad Sci USA*, pages 8485–8489, 1996.

J Coulombe-Huntington and J Majewski. Characterization of intron loss events in mammals. *Genome Research*, 17:23–32, 2007.

W Haerty and GB Golding. Increased Polymorphism Near Low-Complexity Sequences across the Genomes of Plasmodium falciparum Isolates. *Genome Biology and Evolution*, 3:539–550, 2011.

MA Huntley and AG Clark. Evolutionary Analysis of Amino Acid Repeats across the Genomes of 12 Drosophila Species. *Molecular Biology and Evolution*, 24:2598–2609, 2007.

J Lehmann, C Eisenhardt, PF Stadler, and Krauss V. Novel intron positions in Drosophila are mostly caused by intron sliding and tandem duplications. *German Conference on Bioinformatics 2009*, pages 25 – 30, 2009.

L Mangiarini, K Sathasivam, A Mahal, R Mott, M Seller, and GP Bates. Instability of highly expanded CAG repeats in mice transgenic for the Huntington's disease mutation. *Nature genetics*, 15:197–200, 1997.

CE Pearson and RR Sinden. Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry*, 35:5041–5053, 1996.

KJ Siddle, JA Goodship, B Keavney, and MF Santibanez-Koref. Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome. *Bioinformatics*, 27:895–898, 2011.

K Usdin. The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. *Genome Research*, 18:1011–1019, 2008.

MD Vinces, M Legendre, M Caldara, M Hagihara, and KJ Verstrepen. Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability. *Science*, 324:1213–1216, 2009.

Chapter 5

Supplementary Material

ENSG00000017797
DDEKDHGKKKGKFKKKEKRTEG
ENSPTRG00000009859
DDEKDHGKKKGKFKKKEKRTEG
ENSGGOG00000023120
DDEKDHGKKKGKFKKKEKRTEG
ENSPPYG00000009010
DDEKDHGKKKGKFKKKEKRTEG
ENSMMUG00000019521
DDEKDHGKKKGKFKKKEKRTEG

Figure S1: Intron interrupted LCR in ralA binding protein 1

ENSG00000010244
----------IMMPMGGMMPPGPGIPPLMPGMPPGMPPPVPRPGIPP
M
ENSPTRG00000008989
----------IMMPMGGMMPPGPGIPPLMPGMPPGMPPPVPRPGIPP
M
ENSGGOG00000024003
----------IMMPMGGMMPPGPGIPPLMPGMPPGMPPPVPRPGIPP
M
ENSPPYG00000008161
----------IMMPMGGMMPPGPGIPPLMPGMPPGMPPPVPRPGIPP
M
ENSMMUG00000016875
PPLMPGVPPLMPGMPPVMPGMPPGMMPMGGMMPPGPGIPPLMPGMPPGMPPPVPRPGIPP
M

Figure S2: Intron interrupted LCR in zinc finger protein 207

ENSG00000029363
RSNSRSHSSRSKSRSQSSSRSRSRSHSRKKRYSSRSRSRTYSRSRSR
ENSPTRG00000018635
RSNSRSHSSRSKSRSQSSSRSRSRSHSRKKRYSSRSRSRTYSRSRSR
ENSGGOG00000001824
RSNSRSHSSRSKSRSQSSSRSRSRSHSRKKRYSSRSRSRTYSRSRSR
ENSPPYG00000017032
RSNSRSHSSRSKSRSQSSSRSRSRSHSRKKRYSSRSRSRTYSRSRSR
ENSMMUG00000003710
RSNSRSHSSRSKSRSQSSSRSRSRSHSRKKRYSSRSRSRTYSRSRSR

Figure S3: Intron interrupted LCR in BCL2-associated transcription factor 1

ENSG00000032219
KGGPKKKQKKKAKNK
ENSPTRG00000006393
KGGPKKKQKKKAKNK
ENSGGOG00000027092
KGGPKKKQKKKAKNK
ENSPPYG00000005857
KGGPKKKQKKKAKNK
ENSMMUG00000003514
KGGPKKKQKKKAKNK

Figure S4: Intron interrupted LCR in AT rich interactive domain 4A (RBP1-like)

ENSG00000033867
KKKKEDDKKKKEKEE
ENSPTRG00000014704
KKKKEDDKKKKEKEE
ENSGGOG00000011232
KKKKEDDKKKKEKEE
ENSPPYG00000014056
KKKKEDDKKKKEKEE
ENSMMUG00000019327
KKKKEDDKKKKEKEE

Figure S5: Intron interrupted LCR in solute carrier family 4, sodium bicarbonate cotransporter, member 7

ENSG00000047188
SSSYPSPCASPSPPSSGKGSKSPSP
ENSPTRG00000017143
SSSYPSPCASPSPPSSGKGSKSPSP
ENSGGOG00000014407
SSSYPSPCASPSPPSSGKGSKSPSP
ENSPPYG00000015691
SSSYPSPCASPSPPSSGKGSKSPSP
ENSMMUG00000001376
SSSYPSPCASPSPPSSGKGSKSPSP

Figure S6: Intron interrupted LCR in YTH domain containing 2

ENSG00000048052
EVTESSVSSSSPGSGPSSPNNGPTGS
ENSPTRG00000018959
EVTESSVSSSSPGSGPSSPNNGPTGS
ENSGGOG00000013656
EVTESSVSSSSPGSGPSSPNNGPTGS
ENSPPYG00000017755
EVTESSVSSSSPGSGPSSPNNGPTGS
ENSMMUG00000011170
EVTESSVSSSSPGSGPSSPNNGPTGS

Figure S7: Intron interrupted LCR in histone deacetylase 9

ENSG00000055917
SALSGFGSSVGSSASSSA
ENSPTRG00000011692
SALSGFGSSVGSSASSSA
ENSGGOG00000011374
SALSGFGSSVGSSASSSA
ENSPPYG00000012617
SALSGFGSSVGSSASSSA
ENSMMUG00000023181
SALSGFGSSVGSSASSSA

Figure S8: Intron interrupted LCR in pumilio homolog 2 (Drosophila)

ENSG00000055917
SLTPPPSLSSHGSSSSLHLGGLT
ENSPTRG00000011692
SLTPPPSLSSHGSSSSLHLGGLT
ENSGGOG00000011374
SLTPPPSLSSHGSSSSLHLGGLT
ENSPPYG00000012617
SLTPPPSLSSHGSSSSLHLGGLT
ENSMMUG00000023181
SLTPPPSLSSHGSSSSLHLGGLT

Figure S9: Intron interrupted LCR in pumilio homolog 2 (Drosophila)

ENSG00000058272
----------KENEREGEKREEEKEGE
ENSPTRG00000005252
----------KENEREGEKREEEKEGE
ENSGGOG00000014574
----------KENEREGEKREEEKEGE
ENSPPYG00000004796
----------KENEREGEKREEEKEGE
ENSMMUG00000019214
GEKREEEKEGKENEREGEKREEEKEG-

Figure S10: Intron interrupted LCR in protein phosphatase 1, regulatory subunit 12A

ENSG00000112739
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS
ENSPTRG00000017683
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS
ENSGGOG00000006422
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS
ENSPPYG00000016187
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS
ENSMMUG00000014017
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS

Figure S11: Intron interrupted LCR in PRP4 pre-mRNA processing factor 4 homolog B (yeast)

ENSG00000061936
---TTTAPPPPGTTPPPPPTTTETSSGATSTTTTTS---
ENSPTRG00000005640
---TTTAPPPPGTTPLPPPTTAETSSGATSTTTTTSALA
ENSGGOG00000014877
PGVTTTAPPPPGTTPPPPPTTAETSSGATSTTTTTSALA
ENSPPYG00000005129
PGVTTTAPPPPGTTPPPPPTTAETSSGATSTTTTTSALA
ENSMMUG00000020343
PGVTTTAPPPPGTTPPPPPTTAETSSGATSTTTTTSALA

Figure S12: Intron interrupted LCR in splicing factor, suppressor of white-apricot homolog (Drosophila)

ENSG00000061936
RRSRSRSRSPRRR
ENSPTRG00000005640
RRSRSRSRSPRRR
ENSGGOG00000014877
RRSRSRSRSPRRR
ENSPPYG00000005129
RRSRSRSRSPRRR
ENSMMUG00000020343
RRSRSRSRSPRRR

Figure S13: Intron interrupted LCR in splicing factor, suppressor of white-apricot homolog (Drosophila)

ENSG00000065526
SRPTRSPSGSGSRSRSSSSDSISSSSSTSSDSSDSSSSSSDDSPARS
ENSPTRG00000000210
SRPTRSPSGSGSRSRSSSSDSISSSSSTSSDSSDSSSSSSDDSPARS
ENSGGOG00000012041
SRPTRSPSGSGSRSRSSSSDSISSSSSTSSDSSDSSSSSSDDSPARS
ENSPPYG00000001834
SRPTRSPSGSGSRSRSSSSDSISSSSSTSSDSSDSSSSSSDDSPARS
ENSMMUG00000008026
SRPTRSPSGSGSRSRSSSSDSISSSSSTSSDSSDSSSSSSDDSPARS

Figure S14: Intron interrupted LCR in spen homolog, transcriptional regulator (Drosophila)

ENSG00000071626
GGYGPPPAGRGAPPPPPP
ENSPTRG00000010199
GGYGPPPAGRGAPPPPPP
ENSGGOG00000002814
GGYGPPPAGRGAPPPPPP
ENSPPYG00000009316
GGYGPPPAGRGAPPPPPP
ENSMMUG00000003047
GGYGPPPAGRGPPPPPPP

Figure S15: Intron interrupted LCR in DAZ associated protein 1

ENSG00000076108
KSLKQKEAKKKSKAEKEKGKTKQEKLKEKVKREKKEKVKMKEKEEVTK--
ENSPTRG00000005106
KSLKQKEAKKKSKAEKEKGKTKQEKLKEKVKREKKEKVKMKEKEEVTK--
ENSGGOG00000011580
KSLKQKEAKKKSKAEKEKGKTKQEKLKEKVKREKKEKVKMKEKEEVTK--
ENSPPYG00000004667
KSLKQKEAKKKSKAEKEKGKTKQEKLKEKVKREKKEKVKMKEKEEVTK--
ENSMMUG00000019332
KSLKQKEAKKKSKAEKEKGKTKQEKLKEKVKREKKEKVKMKEKEEVAKAK

Figure S16: Intron interrupted LCR in bromodomain adjacent to zinc finger domain, 2A

ENSG00000080503
KQAQAAKEKKKRRRRKKKAEE
ENSPTRG00000020731
KQAQAAKEKKKRRRRKKKAEE
ENSGGOG00000000513
KQAQAAKEKKKRRRRKKKAEE
ENSPPYG00000019258
KQAQAAKEKKKRRRRKKKAEE
ENSMMUG00000015279
KQAQAAKEKKKRRRRKKKAEE

Figure S17: Intron interrupted LCR in SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2

ENSG00000080503
SDSEESDSDYEEEDEEEESSRQETEE
ENSPTRG00000020731
SDSEESDSDYEEEDEEEESSRQETEE
ENSGGOG00000000513
SDSEESDSDYEEEDEEEESSRQETEE
ENSPPYG00000019258
SDSEESDSDYEEEDEEEESSRQETEE
ENSMMUG00000015279
SDSEESDSDYEEEDEEEESSRQETEE

Figure S18: Intron interrupted LCR in SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2

ENSG00000084463
PPMPGPPPLGPPPAPPLRPPGPPTGLPPGPPPGAPPFLRPPGMPGLRGPLPRLLPPGPPP
GRPPGPPPGPPPGLPPGPPPRGPPPRLPPPAPPGIPPPRPGMMRPPL
ENSPTRG00000004724
PPMPGPPPLGPPPAPPLRPPGPPTGLPPGPPPGAPPFLRPPGMPGLRGPLPRLLPPGPPP
GRPPGPPPGPPPGLPPGPPPRGPPPRLPPPAPPGIPPPRPGMMRPPL
ENSGGOG00000004448
PPMPGPPPLGPPPAPPLRPPGPPTGLPPGPPPGAPPFLRPPGMPGLRGPLPRLLPPGPPP
GRPPGPPPGPPPGLPPGPPPRGPPPRLPPPAPPGIPPPRPGMMRPPL
ENSPPYG00000004323
PPMPGPPPLGPPPAPPLRPPGPPTGLPPGPPPGAPPFLRPPGMPGLRGPLPRLLPPGPPP
GRPPGPPPGPPPGLPPGPPPRGPPPRLPPPAPPGIPPPRPGMMRPPL
ENSMMUG00000019299
PPMPGPPPLGPPPAPPLRPPGPPTGLPPGPPPGAPPFLRPPGMPGLRGPLPRLLPPGPPP
GRPPGPPPGPPPGLPPGPPPRGPPPRLPPPAPPGIPPPRPGMMRPPL

Figure S19: Intron interrupted LCR in WW domain binding protein 11

ENSG00000084676
QLRLQLQQRLQGQQQL
ENSPTRG00000011715
QLRLQLQQRLQGQQQL
ENSGGOG00000011279
QLRLQLQQRLQGQQQL
ENSPPYG00000012598
QLRLQLQQRLQGQQQL
ENSMMUG00000020838
QLRLQLQQRLQGQQQL

Figure S20: Intron interrupted LCR in nuclear receptor coactivator 1

ENSG00000091009
TETEEEEVKEEETET
ENSPTRG00000017373
TETEEEEVKEEETET
ENSGGOG00000013312
TETEEEEVKEEETET
ENSPPYG00000015909
TETEEEEVKEEETET
ENSMMUG00000004476
TETEEEEVKEEETET

Figure S21: Intron interrupted LCR in RNA binding motif protein 27

ENSG00000095015
SNSHTLSSSSTSTSSSENS
ENSPTRG00000016891
SNSHTLSSSSTSTSSSENS
ENSGGOG00000014603
SNSHTLSSSSTSTSSSENS
ENSPPYG00000015478
SNSHTLSSSSTSTSSSENS
ENSMMUG00000012387
SNSHTLSSSSTSTSSSENS

Figure S22: Intron interrupted LCR in mitogen-activated protein kinase kinase kinase 1, E3 ubiquitin protein ligase

ENSG00000096746
GGDGYDGGYGGFDDYGGYNNYGYGN
ENSPTRG00000002553
GGDGYDGGYGGFDDYGGYNNYGYGN
ENSGGOG00000025378
GGDGYDGGYGGFDDYGGYNNYGYGN
ENSPPYG00000002397
GGDGYDGGYGGFDDYGGYNNYGYGN
ENSMMUG00000009792
GGDGYDGGYGGFDDYGGYNNYGYGN

Figure S23: Intron interrupted LCR in heterogeneous nuclear ribonucleoprotein H3 (2H9)

ENSG00000096746
GLGGYGRGGGGSGGYYGQGGMSGGGWRG
ENSPTRG00000002553
GLGGYGRGGGGSGGYYGQGGMSGGGWRG
ENSGGOG00000025378
GLGGYGRGGGGSGGYYGQGGMSGGGWRG
ENSPPYG00000002397
GLGGYGRGGGGSGGYYGQGGMSGGGWRG
ENSMMUG00000009792
GLGGYGRGGGGSGGYYGQGGMSGGGWRG

Figure S24: Intron interrupted LCR in heterogeneous nuclear ribonucleoprotein H3 (2H9)

ENSG00000099995
VPTAFVPAPPVAPVPAPAPMPPVHPPPP
ENSPTRG00000014240
VPTAFVPAPPVAPVPAPAPMPPVHPPPP
ENSGGOG00000012631
VPTAFVPAPPVAPVPAPAPMPPVHPPPP
ENSPPYG00000011707
VPTAFVPAPPVAPVPAPAPMPPVHPPPP
ENSMMUG00000005547
VPTAFVPAPPVAPVPAPAPMPPVHPPPP

Figure S25: Intron interrupted LCR in splicing factor 3a, subunit 1, 120kDa

ENSG00000100201
RGGGFGDRDRDRDRGGFGARGGGG
ENSPTRG00000014369
RGGGFGDRDRDRDRGGFGARGGGG
ENSGGOG00000008309
RGGGFGDRDRDRDRGGFGARGGGG
ENSPPYG00000011821
RGGGFGDRDRDRDRGGFGARGGGG
ENSMMUG00000004304
RGGGFGDRDRDRDRGGFGARGGGG

Figure S26: Intron interrupted LCR in DEAD (Asp-Glu-Ala-Asp) box helicase 17

ENSG00000100201
GGGGGGGKGG
ENSPTRG00000014369
GGGGGGGKGG
ENSGGOG00000008309
--GGGGGGGG
ENSPPYG00000011821
--GGGGGGGG
ENSMMUG00000004304
--GGGGGGGG

Figure S27: Intron interrupted LCR in DEAD (Asp-Glu-Ala-Asp) box helicase 17

ENSG00000100354
QFQLACQLLLQQQQQQQLLQNQ
ENSPTRG00000014406
QFQLACQLLLQQQQQQQLLQNQ
ENSGGOG00000008444
QFQLACQLLLQQQQQQQLLQNQ
ENSPPYG00000011849
QFQLACQLLLQQQQQQQLLQNQ
ENSMMUG00000021345
QFQLACQLLLQQQQQQQLLQNQ

Figure S28: Intron interrupted LCR in trinucleotide repeat containing 6B

ENSG00000100461
----------------SNRSRDRDRYRRRNSRSRSRGRQRRHRSRS------
ENSPTRG00000006151
SKKKRSRSHSKSRDRKRSRSRDRDRYRRRNSRSRSRGRQRRHRSRS------
ENSGGOG00000015707
KRKKRSRSHSKSRDRKRSRSRDRDRYRRRNSRSRSRGRQRRHRSRS------
ENSPPYG00000005649
SKKKRSRSHSKSRDRKRSRSRDRDRYRRRNSRSRSRDRQRRHRSRSWDRRHS
ENSMMUG00000018381
-----------SRDRKRSRSRDRDRYRRRNSRSRS----------------

Figure S29: Intron interrupted LCR in RNA binding motif protein 23

ENSG00000100650
SKRHSRSRSRSRSRTRSSSRSRSRSRSRSRSRKSYSRSRSRSRSRSRSRSKSRSVSRS
ENSPTRG00000006488
SKRHSRSRSRSRSRTRSSSRSRSRSRSRSRSRKSYSRSRSRSRSRSRSRSKSRSVSRS
ENSGGOG00000016462
SKRHSRSRSRSRSRTRSSSRSRSRSRSRSRSRKSYSRSRSRSRSRSRSRSKSRSVSRS
ENSPPYG00000005945
SKRHSRSRSRSRSRTRSSSRSRSRSRSRSRSRKSYSRSRSRSRSRSRSRSKSRSVSRS
ENSMMUG00000022407
SKRHSRSRSRSRSRTRSSSRSRSRSRSRSRSRKSYSRSRSRSRSRSRSRSKSRSVSRS

Figure S30: Intron interrupted LCR in serine/arginine-rich splicing factor 5

ENSG00000100813
RLEREAREAAELEEASAESE
ENSPTRG00000006159
RLEREAREAAELEEASAESE
ENSGGOG00000007403
RLEREAREAAELEEASAESE
ENSPPYG00000005657
RLEREAREAAELEEASAESE
ENSMMUG00000012979
RLEREAREAAELEEASAESE

Figure S31: Intron interrupted LCR in apoptotic chromatin condensation inducer 1

ENSG00000100813
SRSTSE--SRSRSRSRSRSASSNSRKSLS
ENSPTRG00000006159
SRSTSESRSRSRSRSRSRSASSNSRKSLS
ENSGGOG00000007403
SRSTSESRSRSRSRSRSRSASSNSRKSLS
ENSPPYG00000005657
SRSTSE--SRSRSRSRSRSASSNSRKSLS
ENSMMUG00000012979
SRSTSE--SRSRSRSRSRSASSNSRKSLS

Figure S32: Intron interrupted LCR in apoptotic chromatin condensation inducer 1

ENSG00000103723
DSDPESESESDSKSSSESGSGESSSESDNED
ENSPTRG00000007381
DSDPESESESDSKSSSESGSGESSSESDNED
ENSGGOG00000006777
DSDPESESESDSKSSSESGSGESSSESDNED
ENSPPYG00000006890
DSDPESESESDSKSSSESGSGESSSESDNED
ENSMMUG00000000617
DSDPESESESDSKSSSESGSGESSSESDNED

Figure S33: Intron interrupted LCR in adaptor-related protein complex 3, beta 2 subunit

ENSG00000104517
SSDQSSSSSQSQSS
ENSPTRG00000020482
SSDQSSSSSQSQSS
ENSGGOG00000003058
SSDQSSSSSQSQSS
ENSPPYG00000018797
SSDQSSSSSQSQSS
ENSMMUG00000020766
SSDQSSSSSQSQSS

Figure S34: Intron interrupted LCR in ubiquitin protein ligase E3 component n-recognin 5

ENSG00000104517
AASTAPSSTSTPAASSA
ENSPTRG00000020482
AASTAPSSTSTPAASSA
ENSGGOG00000003058
AASTAPSSTSTPAASSA
ENSPPYG00000018797
AASTAPSSTSTPAASSA
ENSMMUG00000020766
AASTAPSSTSTPAASSA

Figure S35: Intron interrupted LCR in ubiquitin protein ligase E3 component n-recognin 5

ENSG00000104859
RSPSESSSESRSRSRSP
ENSPTRG00000022583
RSPSESSSESRSRSRSP
ENSGGOG00000003384
RSPSESSSESRSRSRSP
ENSPPYG00000010104
RSPSESSSESRSRSRSP
ENSMMUG00000014155
RSPSESSSESRSRSRSP

Figure S36: Intron interrupted LCR in CLK4-associating serine/arginine rich protein

ENSG00000104859
SARRRSSSSSSSSSASRTSSSRSSSRSSSRSSRRGGGYYRSGRHARSRSRSWSRSRSRSRR
YSRSRSRGRRHSGGGSRDGHR
ENSPTRG00000022583
SARRRSSSSSSSSSASRTSSSRSSSRSSSRSSRRGGGYYRSGRHARSRSRSWSRSRSRSRR
YSRSRSRGRRHSGGGSRDGHR
ENSGGOG00000003384
SARRRSSSSSSSSSASRTSSSRSSSRSSSRSSRRGGGYYRSGRHARSRSRSWSRSRSRSRR
YSRSRSRGRRHSGGGSRDGHR
ENSPPYG00000010104
SARRRSSSSSSSSSASRTSSSRSSSRSSSRSSRRGGGYYRSGRHARSRSRSWSRSRSRSRR
YSRSRSRGRRHSGGGSRDGHR
ENSMMUG00000014155
SARRRSSSSTSSSSASRTSSSRSSSRSSSRSSRRGGGYYRSGRHARSRSRSWSRSRSRSRR
YSRSRSRGRRHSGGGSRDGHR

Figure S37: Intron interrupted LCR in CLK4-associating serine/arginine rich protein

ENSG00000104859
RKIRMKERERREKEREEWER
ENSPTRG00000022583
RKIRMKERERREKEREEWER
ENSGGOG00000003384
RKIRMKERERREKEREEWER
ENSPPYG00000010104
RKIRMKERERREKEREEWER
ENSMMUG00000014155
RKIRMKERERREKEREEWER

Figure S38: Intron interrupted LCR in CLK4-associating serine/arginine rich protein

ENSG00000104859
SRSPSPRYSREYSSSRRRSRSRSRSPHYR
ENSPTRG00000022583
SRSPSPRYSREYSSSRRRSRSRSRSPHYR
ENSGGOG00000003384
SRSPSPRYSREYSSSRRRSRSRSRSPHYR
ENSPPYG00000010104
SRSPSPRYSREYSSSRRRSRSRSRSPHYR
ENSMMUG00000014155
SRSPSPRYSREYSSSRRRSRSRSRSPHYR

Figure S39: Intron interrupted LCR in CLK4-associating serine/arginine rich protein

ENSG00000106299
EDEDEDDEEDFEDDDEWEDDEDEDEDDEEDFEDDDEWED
ENSPTRG00000019640
EDEDEDDEEDFEDDDEWEDDEDEDEDDEEDFEDDDEWED
ENSGGOG00000024699
EDEDEDDEEDFEDDDEWEDDEDEDEDDEEDFEDDDEWED
ENSPPYG00000017959
EDEDEDDEEDFEDDDEWEDDEDEDEDDEEDFEDDDEWED
ENSMMUG00000011544
EDEDEDDEEDFEDDDEWEDDEDEDEDDEEDFEDDDEWED

Figure S40: Intron interrupted LCR in Wiskott-Aldrich syndrome-like

ENSG00000106346
KKHKKSKKKKKSKDKHRDRDSR
ENSPTRG00000018906
KKHKKSKKKKKSKDKHRDRDSR
ENSGGOG00000015369
KKHKKSKKKKKSKDKHRDRDSR
ENSPPYG00000017345
KKHKKSKKKKKSKDKHRDRDSR
ENSMMUG00000020321
KKHKKSKKKKKSKDKHRDRDSR

Figure S41: Intron interrupted LCR in ubiquitin specific peptidase 42

ENSG00000108798
LSAASSASSLASAGSA
ENSPTRG00000009371
LSAASSASSLASAGSA
ENSGGOG00000001954
LSAASSASSLASAGSA
ENSPPYG00000008916
LSAASSASSLASAGSA
ENSMMUG00000002437
-SAASSAFSLASAGSA

Figure S42: Intron interrupted LCR in ABI family, member 3

ENSG00000108798
---LEELSPPPPDEELPLPLDLPPPPPLDGDELGLPPPPPG
ENSPTRG00000009371
---LEELSPPPPDEELPLPLDLPPPPPLDGDELGLPPPPPG
ENSGGOG00000001954
---LEELSPPPPDEELPLPLDLPPPPPLDGDELGLPPPPPG
ENSPPYG00000008916
---LEELSPPPPDEELPLPLDLPPPPPLDGDELGLPPPPPG
ENSMMUG00000002437
PPPLEELSPPPPDEELPLPLDLPPPPPLDGDELGLPPPPPG

Figure S43: Intron interrupted LCR in ABI family, member 3

ENSG00000109111
EEEDDDEEEEEENLDDQDE
ENSPTRG00000008929
EEEDDDEEEEEENLDDQDE
ENSGGOG00000016688
EEEDDDEEEEEENLDDQDE
ENSPPYG00000008105
EEEDDDEEEEEENLDDQDE
ENSMMUG00000004708
EEEDDDEEEDEENLDDQDE

Figure S44: Intron interrupted LCR in suppressor of Ty 6 homolog (S. cerevisiae)

ENSG00000109111
EEGDEEGEGDEAEDEE
ENSPTRG00000008929
EEGDEEGEGDEAEDEE
ENSGGOG00000016688
EEGDEEGEGDEAEDEE
ENSPPYG00000008105
EEGDEEGEGDEAEDEE
ENSMMUG00000004708
EEGDEEGEGDEAEDEE

Figure S45: Intron interrupted LCR in suppressor of Ty 6 homolog (S. cerevisiae)

ENSG00000111605
PGGDRFPGPTGPGGPPPPFPAG
ENSPTRG00000005205
PGGDRFPGPAGPGGPPPPFPAG
ENSGGOG00000008211
PGGDRFPGPAGPGGPPPPFP-G
ENSPPYG00000004756
PGGDRFPGPAGPGGPPPPFP-G
ENSMMUG00000002921
PGGDRFPGPAGPGGPPPPFP-G

Figure S46: Intron interrupted LCR in cleavage and polyadenylation specific factor 6, 68kDa

ENSG00000111642
KEEKKEEEEKKE
ENSPTRG00000004580
KEEKKEEEEKKE
ENSGGOG00000001930
KEEKKEEEEKKE
ENSPPYG00000004192
KEEKKEEEEKKE
ENSMMUG00000018685
KEEKKEEEEKKE

Figure S47: Intron interrupted LCR in chromodomain helicase DNA binding protein 4

ENSG00000111799
GPPGPPGPAGGPGAKGP
ENSPTRG00000018351
GPPGPPGPAGGPGAKGP
ENSGGOG00000009324
GPPGPPGPAGGPGAKGP
ENSPPYG00000016780
GPPGPPGPAGGPGAKGP
ENSMMUG00000019261
GPPGPPGPAGGPGAKGP

Figure S48: Intron interrupted LCR in collagen, type XII, alpha 1

ENSG00000111799
GTPGLPGPPGPMGPPGDRG
ENSPTRG00000018351
GTPGLPGPPGPMGPPGDRG
ENSGGOG00000009324
GTPGLPGPPGPMGPPGDRG
ENSPPYG00000016780
GTPGLPGPPGPMGPPGDRG
ENSMMUG00000019261
GTPGLPGPPGPMGPPGDRG

Figure S49: Intron interrupted LCR in collagen, type XII, alpha 1

ENSG00000111799
GPRGPPGPPGSPGSPGVTGPSG
ENSPTRG00000018351
GPRGPPGPPGSPGSPGVTGPSG
ENSGGOG00000009324
GPRGPPGPPGSPGSPGVTGPSG
ENSPPYG00000016780
GPRGPPGPPGSPGSPGVTGPSG
ENSMMUG00000019261
GPRGPPGPPGSPGSPGVTGPSG

Figure S50: Intron interrupted LCR in collagen, type XII, alpha 1

ENSG00000112280
GPPGEQGPPGPPGPPGVPGIDGIDGD
ENSPTRG00000018329
GPPGEQGPPGPPGPPGVPGIDGIDGD
ENSGGOG00000010221
GPPGEQGPPGPPGPPGVPGIDGIDGD
ENSPPYG00000016759
GPPGEQGPPGPPGPPGVPGIDGIDGD
ENSMMUG00000005577
GPPGEQGPPGPPGPPGVPGIDGIDGD

Figure S51: Intron interrupted LCR in collagen, type IX, alpha 1

ENSG00000112280
GPKGPPGPPGPAGEPGKPGAPGKPG
ENSPTRG00000018329
GPKGPPGPPGPAGEPGKPGAPGKPG
ENSGGOG00000010221
GPKGPPGPPGPAGEPGKPGAPGKPG
ENSPPYG00000016759
GPKGPPGPPGPAGEPGKPGAPGKPG
ENSMMUG00000005577
GPKGPPGPPGPAGEPGKPGAPGKPG

Figure S52: Intron interrupted LCR in collagen, type IX, alpha 1

ENSG00000112280
GSPGLPGKLGSLGSPGLPGLPGPPGLPG
ENSPTRG00000018329
GSPGLPGKLGSLGSPGLPGLPGPPGLPG
ENSGGOG00000010221
GSPGLPGKLGSLGSPGLPGLPGPPGLPG
ENSPPYG00000016759
GSPGLPGKLGSLGSPGLPGLPGPPGLPG
ENSMMUG00000005577
GSPGLPGKPGSLGSPGLPGLPGPPGLPG

Figure S53: Intron interrupted LCR in collagen, type IX, alpha 1

ENSG00000112659
EQEDEEEKRLEEEEEEEEEEAEKE
ENSPTRG00000018191
EQEDEEEKRLEEEEEEEEEEAEKE
ENSGGOG00000005599
EQEDEEEKRLEEEEEEEEEEAEKE
ENSPPYG00000016634
EQEDEEEKRLEEEEEEEEEEAEKE
ENSMMUG00000005434
EQEDEEEKRLEEEEEEEEEEAEKE

Figure S54: Intron interrupted LCR in cullin 9

ENSG00000112739
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS
ENSPTRG00000017683
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS
ENSGGOG00000006422
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS
ENSPPYG00000016187
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS
ENSMMUG00000014017
RRRSRSPIRRRSRSPLRRSRSPRRRSRSPRRRDRGRRSRSRLRRRSRSRGGRRRRSRS

Figure S55: Intron interrupted LCR in PRP4 pre-mRNA processing factor 4 homolog B (yeast)

ENSG00000113360
SRHRSYERSRERERERHRHRDNRRS
ENSPTRG00000016762
SRHRSYERSRERERERHRHRDNRRS
ENSGGOG00000011006
SRHRSYERSRERERERHRHRDNRRS
ENSPPYG00000015370
SRHRSYERSRERERERHRHRDNRRS
ENSMMUG00000012193
SRHRSYERSRERERERHRHRDNRRS

Figure S56: Intron interrupted LCR in drosha, ribonuclease type III

ENSG00000113645
SQLKSLSSSMQSLSSGSSPGSL
ENSPTRG00000017506
SQLKSLSSSMQSLSSGSSPGSL
ENSGGOG00000007877
SQLKSLSSSMQSLSSGSSPGSL
ENSPPYG00000016022
SQLKSLSSSMQSLSSGSSPGSL
ENSMMUG00000007457
SQLKSLSSSMQSLSSGSSPGSL

Figure S57: Intron interrupted LCR in WW and C2 domain containing 1

ENSG00000113649
EEEDPKEEPIKEIKEEPKEEEMTEEEK
ENSPTRG00000017375
EEEDPKEEPIKEIKEEPKEEEMTEEEK
ENSGGOG00000003329
EEEDPKEEPIKEIKEEPKEEEMTEEEK
ENSPPYG00000015911
EEEDPKEEPIKEIKEEPKEEEMTEEEK
ENSMMUG00000023288
EEEDPKEEPIKEIKEEPKEEEMTEEEK

Figure S58: Intron interrupted LCR in transcription elongation regulator 1

ENSG00000114857
RTRSVSYSHSRSRSRSSTSSYRSRSYSRSRSRGWYSRGR
ENSPTRG00000014797
RTRSVSYSHSRSRSRSSTSSYRSRSYSRSRSRGWYSRGR
ENSGGOG00000011584
RTRSVSYSHSRSRSRSSTSSYRSRSYSRSRSRGWYSRGR
ENSPPYG00000013978
RTRSVSYSHSRSRSRSSTSSYRSRSYSRSRSRGWYSRGR
ENSMMUG00000021242
RTRSVSYSHSRSRSRSSTSSYRSRSYSRSRSRGWYSRGR

Figure S59: Intron interrupted LCR in natural killer-tumor recognition sequence

ENSG00000114857
RSYKSHRTSSRSRSRSSS
ENSPTRG00000014797
RSYKSHRTSSRSRSRSSS
ENSGGOG00000011584
RSYKSHRTSSRSRSRSSS
ENSPPYG00000013978
RSYKSHRTSSRSRSRSSS
ENSMMUG00000021242
RSYKSHRTSSRSRSRSSS

Figure S60: Intron interrupted LCR in natural killer-tumor recognition sequence

ENSG00000114857
SRSRSYTYDSYYSRSRSRSRSQRSDSY
ENSPTRG00000014797
SRSRSYTYDSYYSRSRSRSRSQRSDSY
ENSGGOG00000011584
SRSRSYTYDSYYSRSRSRSRSQRSDSY
ENSPPYG00000013978
SRSRSYTYDSYYSRSRSRSRSQRSDSY
ENSMMUG00000021242
SRSRSYTYDSYYSRSRSRSRSQRSDSY

Figure S61: Intron interrupted LCR in natural killer-tumor recognition sequence

ENSG00000119689
AAPAKAKPAEAPAAAAPKAEPTAAAVPPAAAP
ENSPTRG00000006543
AAPAKAKPAEAPAAAAPKAEPTAAAVPPPAAP
ENSGGOG00000011591
AAPAKAKPAEAPAAAAPKAEPTAAAVPPPAAP
ENSPPYG00000005993
AAPAKAKPAEAPAAAAPKAEPTAAAVPPPAAP
ENSMMUG00000004260
AAPAKAKPAEAPAAAAPKAEP----------

Figure S62: Intron interrupted LCR in dihydrolipoamide S-succinyltransferase (E2 component of 2-oxo-glutarate complex)

ENSG00000121486
AARIVVAAVARAAAR
ENSPTRG00000001776
AARIVVAAVARAAAR
ENSGGOG00000011778
AARIVVAAVARAAAR
ENSPPYG00000000416
AARIVVAAVARAAAR
ENSMMUG00000013879
AARIVVAAVARAAAR

Figure S63: Intron interrupted LCR in tRNA methyltransferase 1 homolog (S. cerevisiae)-like

ENSG00000121741
PVPTTVPVPVPVPV------
ENSPTRG00000005680
PVPTTVPVPVPVPV------
ENSGGOG00000007409
PVPTTVPVPVPVPV------
ENSPPYG00000005188
PVPTTVPVPVPVPV------
ENSMMUG00000030812
PVPTTVPVPVPVPVFLPTPL

Figure S64: Intron interrupted LCR in zinc finger, MYM-type 2

ENSG00000124749
GFGHPGEQGPPGPPGPEGPPG
ENSPTRG00000018302
GFGHPGEQGPPGPPGPEGPPG
ENSGGOG00000034842
GFGHPGEQGPPGPPGPEGPPG
ENSPPYG00000016734
GFGHPGEQGPPGPPGPEGPPG
ENSMMUG00000007616
GFGYPGEQGPPGPPGPEGPPG

Figure S65: Intron interrupted LCR in collagen, type XXI, alpha 1

ENSG00000125107
TTTTTSTTPATNTTCTAT
ENSPTRG00000008183
TTTTTSTTPATNTTCTAT
ENSGGOG00000009547
TTTTTSTTPATNTTCTAT
ENSPPYG00000007410
TTTTTSTTPATNTTCTAT
ENSMMUG00000014625
TTTTTSTTPATNTTCTAT

Figure S66: Intron interrupted LCR in CCR4-NOT transcription complex, subunit 1

ENSG00000126254
VVPPMVGGPPFVGPVGFGPG
ENSPTRG00000010852
VVPPMVGGPPFVGPVGFGPG
ENSGGOG00000009972
VVPPMVGGPPFVGPVGFGPG
ENSPPYG00000009871
VVPPMVGGPPFVGPVGFGPG
ENSMMUG00000018954
VVPPMVGGPPFVGPVGFGPG

Figure S67: Intron interrupted LCR in RNA binding motif protein 42

ENSG00000127616
SEESGSEEEEEEEEE
ENSPTRG00000010488
SEESGSEEEEEEEEE
ENSGGOG00000009882
SEESGSEEEEEEEEE
ENSPPYG00000009568
SEESGSEEEEEEEEE
ENSMMUG00000012042
SEESGSEEEEEEEEE

Figure S68: Intron interrupted LCR in SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4

ENSG00000130254
EIKIEKTIIKKEEKIEKKEEKKPEDIKKEEK
ENSPTRG00000010334
EIKIEKTVIKKEEKIEKKEEKKPEDIKKEEK
ENSGGOG00000015093
EIKIEKTVIKKEEKIEKKEEK----------
ENSPPYG00000009427
EIKIEKTVIKKEEKIEKKEEKKPEDIKKEEK
ENSMMUG00000017968
EIKIEKTVIKKEEKIEKKEEKKPEDIKKEEK

Figure S69: Intron interrupted LCR in scaffold attachment factor B2

ENSG00000130254
EQRERERQRQREREIRETERRREREQREREQR
ENSPTRG00000010334
EQRERERQRQREREIRETERRREREQREREQR
ENSGGOG00000015093
EQRERERQRQREREIRETERRREREQREREQR
ENSPPYG00000009427
EQRERERQRQREREIRETERRREREQREREQR
ENSMMUG00000017968
EQRERERQRQREREIRETERRREREQREREQR

Figure S70: Intron interrupted LCR in scaffold attachment factor B2

ENSG00000132694
LDLHVLLLEDLLVLL
ENSPTRG00000024318
LDLHVLLLEDLLVLL
ENSGGOG00000014973
LDLHVLLLEDLLVLL
ENSPPYG00000000695
LDLHVLLLEDLLVLL
ENSMMUG00000012436
LDLHVLLLEDLLVLL

Figure S71: Intron interrupted LCR in Rho guanine nucleotide exchange factor (GEF) 11

ENSG00000132842
SESEEEEDSSDSSSDSESESGSESGEQGESGEEGDSNEDSSEDSSSEQDSES
ENSPTRG00000017015
SESEEEEDSSDSSSDSESESGSESGEQGESGEEGDSNEDSSEDSSSEQDSES
ENSGGOG00000001771
SESEEEEDSSDSSSDSESESGSESGEQGESGEEGDSNEDSSEDSSSEQDSES
ENSPPYG00000015580
SESEEEEDSSDSSSDSESESGSE----------
ENSMMUG00000001360
SESEEEEDSSDSSSDSESESGSESGEQGESGEEGDSNEDSSEDSSSEQDSES

Figure S72: Intron interrupted LCR in adaptor-related protein complex 3, beta 1 subunit

ENSG00000133226
KEEKESSREKRERSRSPRRRKSRSPSPRRRSSPVRRER
ENSPTRG00000000349
KEEKESSREKRERSRSPRRRKSRSPSPRRRSSPVRRER
ENSGGOG00000005209
KEEKESSREKRERSRSPRRRKSRSPSPRRRSSPVRRER
ENSPPYG00000001716
KEEKESSREKRERSRSPRRRKSRSPSPRRRSSPVRRER
ENSMMUG00000006645
KEEKESSREKRERSRSPRRRKSRSPSPRRRSSPVRRER

Figure S73: Intron interrupted LCR in serine/arginine repetitive matrix 1

ENSG00000133226
--RPRSRSRSKSRSRTRSRSPSHTRPRRRHRSRS-----RRRPSPRRRPSPRRRTPPKRM
PPPPRHRRSRSPVRRRRRSSASLSGSSSSSSSSRSRS
ENSPTRG00000000349
KTRPRSRSRSKSRSRTRSRSPSHTRPRRRHRSRSRSYSPRRRPSPRRRPSPRRRTPPRRM
PPPPRHRRSRSPVRRRRRSSASLSGSSSSSSSSRSRS
ENSGGOG00000005209
KTRPRSRSRSKSRSRTRSRSPSHTRPRRRHRSRSRSYSPRRRPSPRRRPSPRRRTPPRRM
PPPPRHRRSRSPVRRRRRSSASLSGSSSSSSSSRSRS
ENSPPYG00000001716
KTRPRSRSRSKSRSRTRSRSPSHTRPRRRHRSRS-----RRRPSPRRRPSPRRRTPPRRM
PPPPRHRRSRSPVRRRRRSSASLSGSSSSSSSSRSRS
ENSMMUG00000006645
----------RPRRRHRSRS-----RRRPSPRRRPSPRRRTPPRRM
PPPPRHRRSRSPVRRRRRSSASLSGSSSSSSSSRSRS

Figure S74: Intron interrupted LCR in serine/arginine repetitive matrix 1

ENSG00000133226
SSSDSGSSSSS
ENSPTRG00000000349
SSSDSGSSSSS
ENSGGOG00000005209
SSSDSGSSSSS
ENSPPYG00000001716
SSSDSGSSSSS
ENSMMUG00000006645
SSSDSGSSSSS

Figure S75: Intron interrupted LCR in serine/arginine repetitive matrix 1

ENSG00000133226
TSPRGRRRRSPSPPPTRRRRSPSPAPPPRRRRTPTPPPRRRTPSPPPRRRSPSPRRYSPP

ENSPTRG00000000349
-----RRRRSPSPPPTRRRRSPSPAPPPRRRRTPTPPPRRRTPSPPPRRRSPSPRRYSPP

ENSGGOG00000005209
-----RRRRSPSPPPTRRRRSPSPAPPPRRRRTPTPPPRRRTPSPPPRRRSPSPRRYSPP

ENSPPYG00000001716
-----RRRRSPSPPPTRRRRSPSPAPPPRRRRTPTPPPRRRTPSPPPRRRSPSPRRYSPP

ENSMMUG00000006645
-----RRRRSPSPPPTRRRRSPSPAPPPRRRRTPTPPPRRRTPSPPPRRRSPSPRRYSPP

Figure S76: Intron interrupted LCR in serine/arginine repetitive matrix 1

ENSG00000134186
RRRSRSPRRSLSPRRSPRRSRSRS
ENSPTRG00000001032
RRRSRSPRRSLSPRRSPRRSRSRS
ENSGGOG00000011288
RRRSRSPRRSLSPRRSPRRSRSRS
ENSPPYG00000001092
RRRSRSPRRSLSPRRSPRRSRSRS
ENSMMUG00000010155
RRRSRSPRRSLSPRRSPRRSRSRS

Figure S77: Intron interrupted LCR in PRP38 pre-mRNA processing factor 38 (yeast) domain containing B

ENSG00000134748
RYRRSRSRSPRRRSRSPKRRSPSPRRER
ENSPTRG00000000739
RYRRSRSRSPRRRSRSPKRRSPSPRRER
ENSGGOG00000003940
RYRRSRSRSPRRRSRSPKRRSPSPRRER
ENSPPYG00000001348
RYRRSRSRSPRRRSRSPKRRSPSPRRER
ENSMMUG00000014333
RYRRSRSRSPRRRSRSPKRRSPSPRRER

Figure S78: Intron interrupted LCR in PRP38 pre-mRNA processing factor 38 (yeast) domain containing A

ENSG00000135250
KPIGKISKNKKKKLKKKQK
ENSPTRG00000019564
KPIGKISKNKKKKLKKKQK
ENSGGOG00000023098
KPIGKISKNKKKKLKKKQK
ENSPPYG00000017889
KPIGKISKNKKKKLKKKQK
ENSMMUG00000022732
KPIGKISKNKKKKLKKKQK

Figure S79: Intron interrupted LCR in SRSF protein kinase 2

ENSG00000136754
SGSSGGSGSRENSGSSSIG
ENSPTRG00000024219
SGSSGGSGSRENSGSSSIG
ENSGGOG00000011664
SGSSGGSGSRENSGSSSIG
ENSPPYG00000002164
SGSSGGSGSRENSGSSSIG
ENSMMUG00000005078
SGSSGGSGSRENSGSSSIG

Figure S80: Intron interrupted LCR in abl-interactor 1

ENSG00000137073
RGKRARGRGFGRGRGRGAGRF-
ENSPTRG00000020871
RGKRARGRGFGRGRGRGAGRF-
ENSGGOG00000013320
RGKRARGRGFGRGRGRGAGRF-
ENSPPYG00000019128
RGKRARGRGFGRGRGRGAGRF-
ENSMMUG00000012597
RGKRARGRGFGRSRGRGAGRFS

Figure S81: Intron interrupted LCR in ubiquitin associated protein 2

ENSG00000137073
LTSSPLSQLSSSLSSQQSSLSSAHAALSSSTSHTHAS
ENSPTRG00000020871
--SSPLSQLSSSLSSHQSSL-SAHAALSSS-------
ENSGGOG00000013320
--SSPLSQLSSSLSSHQSSL-SAHAALSSS-------
ENSPPYG00000019128
--SSPLSQLSSSLSSHQSSL-SAHAALSSS-------
ENSMMUG00000012597
LSSSPLSQLSSSLSSHQSSLSSAHAALSSSTSHTHAS

Figure S82: Intron interrupted LCR in ubiquitin associated protein 2

ENSG00000138193
SSSNKSPSSAWSSSS
ENSPTRG00000002775
SSSNKSPSSAWSSSS
ENSGGOG00000007807
SSSNKSPSSAWSSSS
ENSPPYG00000002499
SSSNKSPSSAWSSSS
ENSMMUG00000001608
-SSNKSPSSAWSSSS

Figure S83: Intron interrupted LCR in phospholipase C, epsilon 1