# The Genomic Distribution and Local Context of Coincident SNPs in Human and Chimpanzee

Alan Hodgkinson* and Adam Eyre-Walker

Centre for the Study of Evolution, School of Life Sciences, University of Sussex, Brighton, United Kingdom

*Corresponding author: E-mail: alan.j.hodgkinson@gmail.com.

**Accepted:** 2 July 2010

## Abstract

We have previously shown that there is an excess of sites that are polymorphic at orthologous positions in humans and chimpanzees and that this is most likely due to cryptic variation in the mutation rate. We showed that this might be a consequence of complex context effects since we found significant heterogeneity in triplet frequencies around coincident single nucleotide polymorphism (SNP) sites. Here, we show that the heterogeneity in triplet frequencies is not specifically associated with coincident SNPs but is instead driven by base composition bias around CpG dinucleotides. As a result, we suggest that cryptic variation in the mutation rate is truly cryptic, in the sense that the mutation rate does not appear to depend on any specific primary sequence context. Furthermore, we propose that the patterns around CpG dinucleotides are driven by the mutability of CpG dinucleotides in different DNA contexts. We also show that the genomic distribution of coincident SNPs is nonuniform and that there are some subtle differences between the distributions of single and coincident SNPs. Furthermore, we identify regions that contain high numbers of coincident SNPs and suggest that one in particular, a region containing the gene PRIM2, may be under balancing selection.

**Key words:** human, mutation, cryptic, variation.

## Introduction

There is variation in the mutation rate over a number of different scales in the human genome; on a local scale, there are hypermutable sites (Blake et al. 1992; Zhao et al. 2003; Hwang and Green 2004; Hodgkinson et al. 2009), and more broadly, large genomic regions and whole chromosomes can vary in their mutation rate (Matassi et al. 1999; Williams and Hurst 2000; Lercher et al. 2001; Li et al. 2002; Gaffney and Keightley 2005). What makes a region or site have a higher or lower mutation rate is poorly understood, except in the case of CpGs where cytosine can become methylated and unstable, leading to a higher rate of mutation (Coulondre et al. 1978; Bird 1980). However, understanding the factors that dictate the mutation rate is important because they influence human disease and our understanding of evolution.

In a previous study, we showed that there is an excess of coincident single nucleotide polymorphisms (SNPs), sites that have a SNP in both humans and chimpanzees, and that this excess could not be explained by the known influence of adjacent nucleotides on the mutation rate (Hodgkinson et al. 2009). We also showed that the excess of coincident SNPs was not a consequence of ancestral polymorphisms as the result was conserved over more distantly related species (human and macaque), and it is unlikely that SNPs would be preserved over this time frame. Furthermore, the excess is not a result of selection; positive selection tends to remove variation from the population through rapid fixation of beneficial alleles and negative selection, in which removal of variation may result in a general clustering of single SNPs in noncoding regions, was not observed. Finally, the excess of coincident SNPs is not a consequence of us mis-inferring mutation rates in different parts of the genome; we show that mutation rates correlate in GC-rich and GC-poor regions of the genome, and we also observe a significant excess of coincident SNPs in both sequence contexts. We therefore proposed that there is cryptic variation in the mutation rate. However, despite the evidence that this variation in the mutation rate is not due to simple context effects (i.e., the adjacent nucleotides), we did show that triplet frequencies are significantly heterogeneous to approximately 80 bp either side of the coincident SNP (Hodgkinson et al. 2009).

Here, we investigate both the local and genomic context of human and chimp coincident SNPs. We now find that the heterogeneity in triplet frequencies is not specifically

associated with coincident SNPs but is instead associated with patterns of base composition around CpG dinucleotides. As a result, we suggest that cryptic variation in the mutation rate is complex, in the sense that the mutation rate does not appear to depend on any specific context. We also show that the genomic locations of coincident SNPs are nonuniform and that there are subtle differences in the distributions of single SNPs compared with coincident SNPs across the genome.

## Materials and Methods

In our original analysis, we investigated whether human and chimpanzee SNPs tended to occur at the same site in the genome by Blasting all chimpanzee SNPs found in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) against a data set of human SNPs from the same resource (Hodgkinson et al. 2009). We obtained more than three hundred thousand 81 bp alignments that contained both a human and a chimpanzee SNP, and in 11,571 cases, the human and chimpanzee SNPs occurred at orthologous positions. We showed that this number was significantly more than we would expect to occur by chance if SNPs are randomly distributed along our alignments, even after taking into account that certain sites are more likely to contain a SNP due to the effects of neighboring nucleotides on the mutation rate.

In order to investigate heterogeneity in triplet frequencies around coincident SNPs and dinucleotides, we analyzed the 200 bp either side of each of the coincident SNPs identified in our previous analysis (Hodgkinson et al. 2009), except for a few changes due to the random selection of Blast alignments, together with an equal number of randomly chosen instances of each dinucleotide. Random dinucelotides were obtained from the entire human genome sequence from the Ensembl database (http://www.ensembl.org/index.html—build 55). We split our data set of coincident SNPs into two groups, CpG and non-CpG coincident SNPs. A SNP was designated as CpG if the site, or any of the alleles at the site, would yield a CpG dinucleotide.

To investigate whether triplet frequencies are significantly heterogeneous around coincident SNPs and dinucleotides, we proceeded as follows. We tabulated the frequency of each triplet at each site relative to the coincident SNP or dinucleotide—for example, to investigate heterogeneity in triplets 10 bp upstream of coincident SNPs, we tabulated the frequency of triplets where the central nucleotide is 10 bp upstream of a coincident SNP across all our 11,571 sequences containing a coincident SNP. We then summed the number of each triplet across all sites and divided this by the total number of sites to yield the average expected number of triplets at all sites. Whether the observed values were significantly different to the expected values was assessed using a standard chi-square test. To investigate whether the heterogeneity in triplet frequencies could be explained in terms of trends in base composition, we calculated the expected frequency of each triplet from the average nucleotide composition at each site; for example, to calculate the expected frequency of CTG at position +10, we would multiply the frequency of C at position +9 by the frequency of T at position +10 and the frequency of G at position +11.

To analyze the sequence context around each type of dinucleotide, we obtained 2,000 instances of each type of dinucleotide, flanked by 5,000 bp either side, at random from the human genome. Any sequences that contained possible CpG islands were removed; CpG islands were identified by following the method as outlined by Takai and Jones (2002). For each data set, we lined up the sequences so that the dinucleotide of interest was present in the central position and then we calculated the GC content of each position across all 2,000 sequences. To calculate the width of the peak (or depression) in GC content around dinucleotides, we first used the median of the first 1,000 GC content values (positions −5000 to −4000 with respect to the central dinucleotide in alignments) as an indicator of the average GC content. In all cases, the peak in GC content begins after the first 1,000 nucleotides in the alignments and so we believe that this value acts as a good indicator of the average GC content in surrounding sequences. We then assign the start of the peak as the point at which greater than 47 of 50 base positions in a row have an average GC content across the 2,000 sequences that is higher than the median value, labeling the start of the 50 bp as the start of the peak. Similarly, the end of the peak is designated in the same way but running in the opposite direction from the end of the sequences. For a depression in GC content, we required the GC content to be below the median value under the same criteria as before.

Genomic data on telomere and centromere locations, GC content, gene density, nucleosome occupancy, and single SNP density was downloaded from the UCSC genome browser (http://genome.ucsc.edu/). Data were downloaded per 1 MB for comparison with coincident SNP data. Gene density was taken as the number of base pairs that were part of an exon in each megabase region. A365 values were used for nucleosome occupancy scores as they are comparable with other methods at identifying regions with high nucleosome occupancy (Gupta et al. 2008). Recombination rate data was also averaged across each MB, using data from Kong et al. (2002), as were replication-timing scores (S50), which were taken from Chen et al. (2010).

We estimated human and chimp divergence per MB as follows. Alignments between the NCBI36 version of the human genome and the PanTro2 version of the chimp genome were downloaded from the UCSC Web site (http://genome.ucsc.edu/). Nucleotides were masked if they were of low quality in the chimp genome (error rate above 1/10,000); quality scores were unavailable for chromosomes

21 and Y and so these were not masked. We then masked any sequences where divergence scores were too high; 100 bp windows with greater than 10% divergence were masked, with sliding windows every 10 bp. Finally, we masked any sequences of less than 20 bp that were flanked both sides by >40 bp of sequence gaps. The number of mutations per MB was calculated in regions containing >100 kb of unmasked sequence.

In order to investigate balancing selection on the PRIM2 gene and the 175 kb region on chromosome 4, we downloaded low-coverage pilot variation data from the 1,000 genomes project (http://www.1000genomes.org) that was released in April 2009. The data were split into three groups (CEU, YRI, JPT + CHB) at each locus and comprised of allele frequency data within each population and phased genotype data for each sampled site. We obtained phased haplotypes from the data set for each region and then used the Neighbor-Joining method in PHYLIP (Felsenstein 2005) to construct a phylogenetic tree within each population. For PRIM2, CEU was sampled across 57 individuals and contained 3,974 SNPs, YRI was sampled across 56 individuals and contained 1,548 SNPs, and the combined populations of JPT + CHB were sampled across 59 individuals and contained 1,660 SNPs. For the region on chromosome 4, CEU was sampled across 57 individuals and contained 1,070 SNPs, YRI was sampled across 56 individuals, and contained 318 SNPs, and the combined populations of JPT + CHB were sampled across 59 individuals and contained 284 SNPs.

To calculate the significance of the Tajima's *D* values, we performed a coalescent simulation using MS (Hudson 2002) with the same number of haplotypes as found in each population, assuming a stationary population size and either no recombination or a constant recombination rate that was calculated using the Pairwise program in LDhat (McVean et al. 2002). We repeated this procedure 1,000 times for each population at each locus and calculated the Tajima's *D* statistic in each case; the *P* value was the number of times the Tajima's *D* value generated in each of the 1,000 coalescent simulations was greater than the observed value for each population at each locus.

## Results

### Local Context of Coincident SNPs

In a previous study, we found a significant excess of SNPs in the human genome that also contained a SNP at the orthologous position in chimpanzee; we refer to these as coincident SNPs (Hodgkinson et al. 2009). Furthermore, we showed that there is significant heterogeneity in triplet frequencies that extends to about 80 bp either side of coincident SNPs. We did this by tabulating the frequency of each triplet at each site relative to the coincident SNP across our sequences containing coincident SNPs. The triplet frequencies at each site were then compared with the average frequencies across all sites using a chi-square test. To investigate this pattern further, we divided our original data set into CpG and non-CpG coincident SNPs and repeated the analysis as above. For non-CpG coincident SNP sequences (4,517 cases), the frequency of triplets within approximately 10 bp either side of the coincident SNP are significantly different to the average triplet frequencies across all positions in the sequences (fig. 1*a*). This pattern is entirely driven by runs of A and T nucleotides; if we remove sequences where the coincident SNP falls at the start or the end of a mononucleotide triplet of any kind, the peak in triplet heterogeneity disappears outside of the neighboring nucleotides (fig. 1*b*). It should be noted at this point that mononucleotide runs are not the cause of the excess of coincident SNPs because removing SNPs that form part of a run of three or more nucleotides still leaves a large excess of coincident SNPs

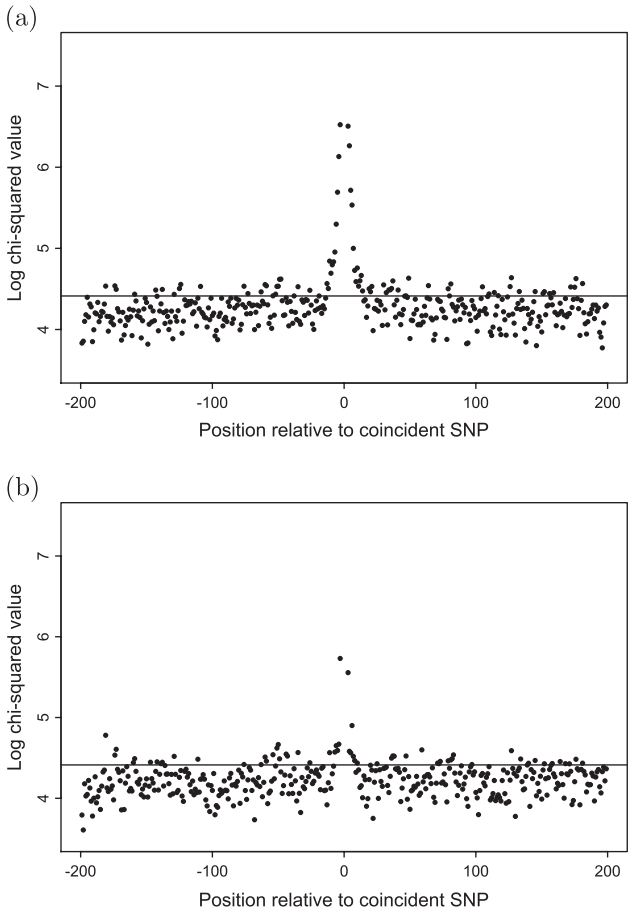


**Fig. 1.**—Heterogeneity in triplet frequencies around non-CpG coincident SNPs. Figure gives the log chi-squared value of the heterogeneity of triplet frequencies against the average triplet frequencies across the whole alignment for (*a*) all alignments containing a non-CpG coincident SNP and (*b*) alignments where the SNP is not part of a mononucleotide run of 3 or more nucleotides. The horizontal line marks the 5% significance value for the chi-square test.

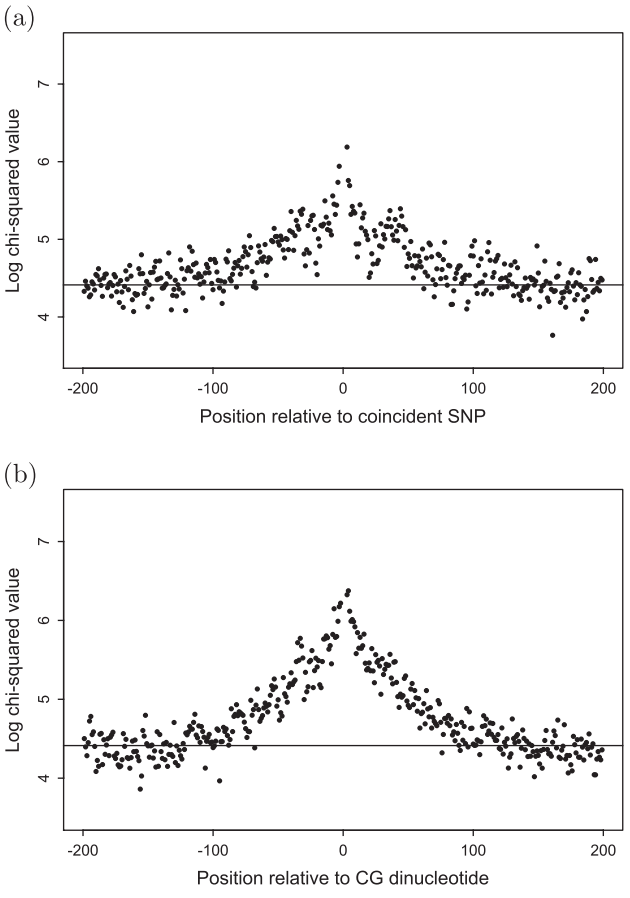


**Fig. 2.**—Heterogeneity in triplet frequencies. Figure gives the log chi-squared value of the heterogeneity of triplet frequencies against the average triplet frequencies across the whole alignment for (*a*) all alignments containing a CpG coincident SNP and (*b*) sequences that contain a CpG dinucleotide but no SNP. The horizontal line marks the 5% significance value for the chi-square test.

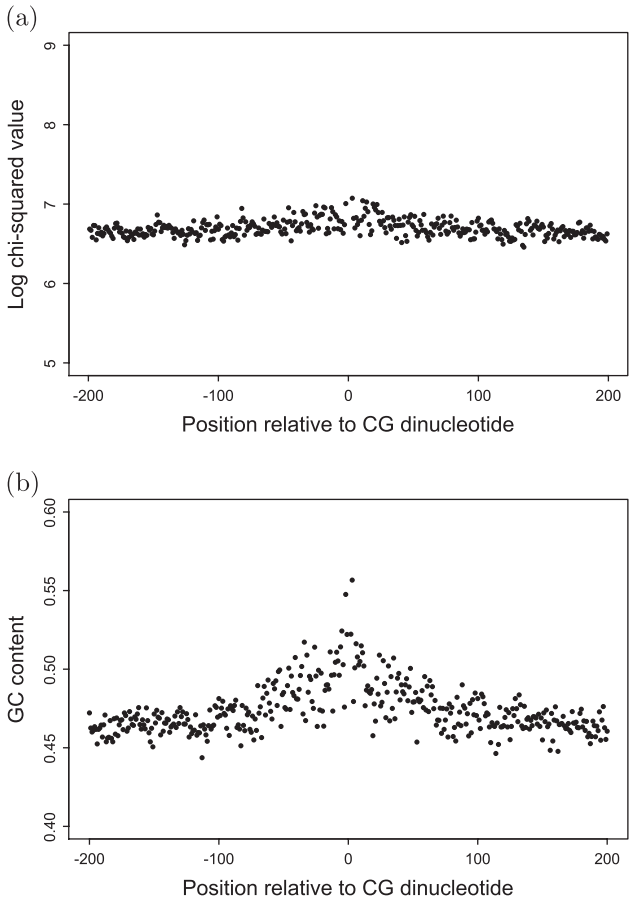


**Fig. 3.**—Nucleotide patterns around CpG dinucleotides that do not contain a SNP. Figure (*a*) gives log chi-squared values across alignments when single nucleotide frequencies are used to predict triplet frequencies at each position and (*b*) shows the GC content at each position in the alignment.

(Hodgkinson et al. 2009). There therefore appears to be no heterogeneity in triplet frequencies around non-CpG coincident SNPs when mononucleotide runs are removed.

For CpG coincident SNP sequences (5,930 cases), the heterogeneity of triplet frequencies extends up to ~100 bp either side of the coincident SNP (fig. 2*a*). However, if we take the same number of CpGs, which do not contain a SNP, from unique sequences at random from the human genome, we observe a very similar pattern (fig. 2*b*). This indicates that the pattern around CpG coincident SNPs is entirely dominated by the pattern around CpG dinucleotides, whether they contain a SNP or not, and thus, there are no local context effects associated specifically with coincident SNPs. As such, variation in the mutation rate in the human genome that was inferred from the excess of coincident SNPs is truly cryptic on a local scale, in the sense that there do not appear to be any nonrandom patterns of nucleotides in the surrounding sequence.

## Patterns Around CpG and Other Dinucleotides

Although we have shown that there are no local nucleotide contexts associated specifically with concident SNPs, it is interesting to consider what is driving the patterns in triplet heterogeneity around CpG dincleotides that do not contain a SNP. In order to investigate this, we estimated the frequencies of triplets we would expect to see given the single nucleotide compositions at each position in the sequences containing CpG dinucleotides. Interestingly, we found that the peak all but disappeared (fig. 3*a*). Furthermore, if we plot the GC content across alignments at each position, there is a similar peak around the central CpG dinucleotide (fig. 3*b*). This implies that there is a general increase in GC content around CpG dincleotides in the human genome, a pattern has also been previously observed in a study by Elango et al. (2008), and that this explains the pattern of triplet heterogeneity in regions around CpGs.

The question therefore arises as to whether there are patterns in nucleotide content around other non-SNP containing

dinucleotides in the human genome. To answer this, we selected 2,000 cases of each type of dinucleotide at random from the human genome, flanked by 5,000 bp either side. We then lined up the sequences so that the nucleotide of interest was present in the central positions and then considered the GC content across the sequences at each position. Sequences were only considered if they did not contain possible CpG islands (see Materials and Methods). However, as it is likely that some dinucleotides are part of mononucleotide and dinucleotide runs, we selected only dinucleotides that were not part of two or more of the same dinucleotides or where the first or second base of the dinucleotide was not part of a mononucleotide triplet. As expected, we find an increase in GC content around CpG dinucleotides that runs from −231 to 199 bp (with the dinucleotide at position zero) but also a peak in GC content around GpC dinucleotides that runs from −77 to 90 bp, which are shown in figure 4*a* and *b*, respectively. Furthermore, there is a decrease in GC content around TA dinucleotides that extends from −95 to 80 bp (fig. 4*c*). The specific widths of the peaks are clearly determined by the number of sequences used; however, they are useful in comparisons and all clearly show a context effect. Consequently, it appears that there are strong nucleotide patterns acting on a very local scale in the human genome. There are also peaks in GC content around GGs and CCs and troughs in GC content around AAs and TT; however, for GGs and CCs, this is caused by sequences in which a CpG or GpC is found immediately adjacent to the central dinucleotide, and for AAs and TTs where TA or AT is found immediately adjacent to the central dinucleotide. If these sequences are removed, the patterns disappear. There are no patterns in GC content around other dinucleotides.

## Genomic Distribution of Coincident SNPs

To investigate the genomic distribution of coincident SNPs, we split the human genome into regions of 1 MB and tallied the number of coincident and single (noncoincident) SNPs found in each region. Regions with no SNPs were excluded from further analysis; these are typically found in the heterochromatic regions near the centromere. On average, there were 8,014 and 3.91 simple and coincident SNPs per MB, respectively, 6,838 and 1.68 of which were non-CpG. If coincident SNPs occurred at random across the human genome, then we would expect the frequency of coincident SNPs in each 1 MB region to be Poisson distributed and to have a variance equal to the mean. However, the observed variance, 13.27, is far in excess of this, and using a chi-square test, we find that the number of coincident SNPs per MB is significantly overdispersed ($P < 0.001$); for example, the third quartile is 2.5-fold higher than the first quartile, whereas it would be expected to be 1.67-fold higher if coincident SNPs were distributed at random. Therefore, coincident SNPs are nonuniformly distributed across the human genome. It is possible that the distribution of coincident SNPs is not Poisson distributed across the human genome if there is variation in the level of sampling that has occurred between different regions. However, this would lead to an excess of coincident SNPs through general

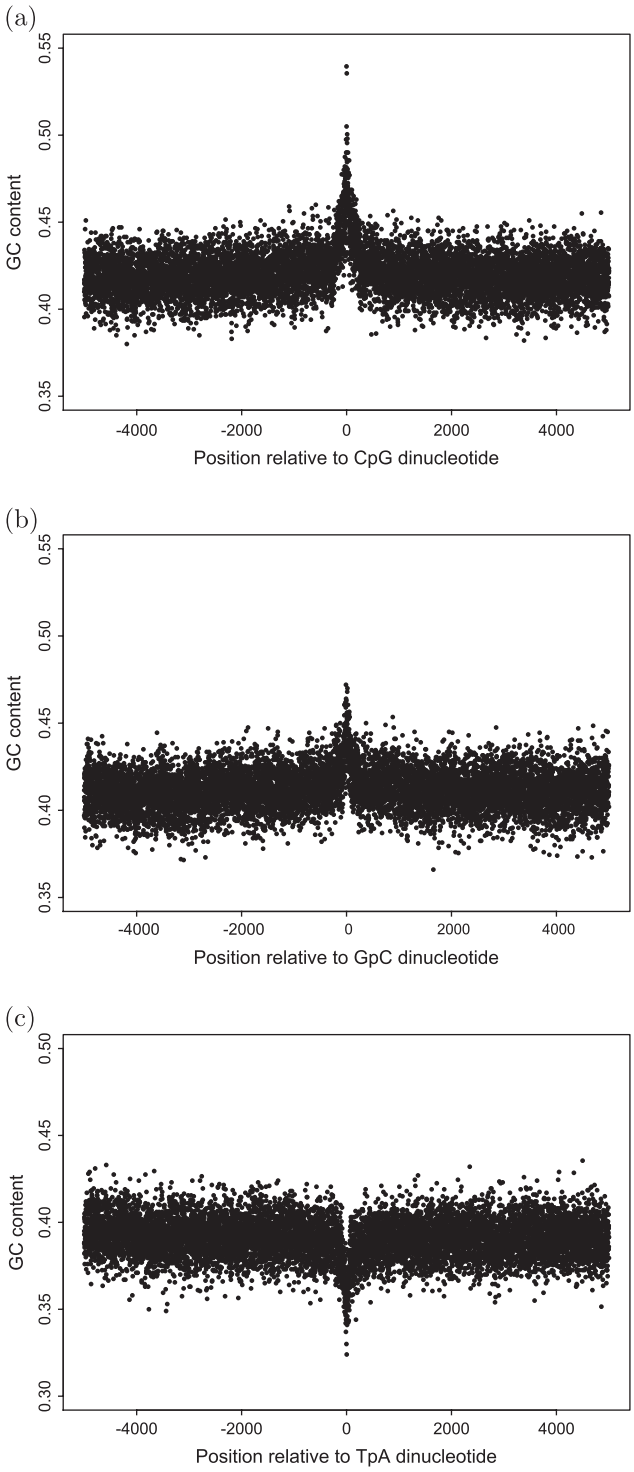


**Fig. 4.**—GC content around (*a*) CpG dinucleotides, (*b*) GpC dinucleotides, and (*c*) TpA dinucleotides.

**Table 1**
The Correlation between the Number of Coincident SNPs per MB and Various Genomic Features

| Feature | *r* | *P* |
|---|---|---|
| SNP density | 0.256 | <0.001 |
| Distance to telomere | −0.022 | 0.226 |
| Distance to centromere | 0.011 | 0.565 |
| Recombination rate | 0.107 | <0.001 |
| Nucleosome association | 0.004 | 0.832 |
| Gene density | −0.022 | 0.230 |
| GC content | −0.006 | 0.741 |
| Replication timing | 0.004 | 0.838 |

**Table 2**
The Correlation between the Number of Single SNPs per MB and Various Genomic Features

| Feature | *r* | *P* |
|---|---|---|
| Distance to telomere | −0.171 | <0.001 |
| Distance to centromere | −0.047 | 0.012 |
| Recombination rate | 0.234 | <0.001 |
| Nucleosome association | 0.187 | <0.001 |
| Gene density | 0.064 | 0.001 |
| GC content | 0.184 | <0.001 |
| Replication timing | 0.008 | 0.673 |

clustering of single SNPs, which we do not observe when considering the distribution of distances between human and chimpanzee SNPs in the original analysis (Hodgkinson et al. 2009).

The distribution of single SNPs is also known to be nonuniform (Venter et al. 2001), and we find that the density of coincident SNPs has a significantly positive correlation with the density of single SNPs ($r^2 = 0.065$, $P < 0.001$ for all SNPs and $r^2 = 0.037$, $P < 0.001$ for non-CpG SNPs); this is perhaps not surprising given that SNP densities must drive the locations of coincident SNPs to a certain extent because at least half of coincident SNPs are thought to be due to chance alone, as single SNPs coincide at random (Hodgkinson et al. 2009). However, the correlation between the density of single and coincident SNPs is not strong, and the lack of a strong correlation is not due to high sampling error in coincident SNP density; as we have shown above, the observed variance in the density of coincident SNPs is substantially greater than we expect from sampling error alone—that is, the distribution of coincident SNP density is overdispersed. We can estimate the approximate proportion of the variance in coincident SNP density that is due to sampling error as follows; because we expect the number of coincident SNPs in each genomic region to be Poisson distributed, the average sampling variance is likely to be of the order of the mean number of coincident SNPs per MB; this is approximately 30% of the total variance in the density of coincident SNPs. Given that the correlation between the density of single SNPs only explains 6.5% of the density in coincident SNPs, it is evident that the poor correlation is not due to sampling error in the density of coincident SNPs.

To investigate the variation in coincident SNP density in more depth, we compared the frequency of coincident SNPs in each 1 MB with some key genomic features (table 1). There is no significant correlation between the density of coincident SNPs and the distance to the centromere, the distance to the nearest telomere or the nucleosome association, the gene density, replication timing, or GC content of a region. There is, however, a significantly positive correlation between coincident SNP density and recombination rate ($r = 0.107$, $P < 0.001$). This may reflect the significant correlation that exists between the density of single SNPs and the rate of recombination ([Hellmann et al. 2003, 2005]; in our data set $r = 0.242$, $P < 0.001$), and the fact that approximately half of all coincident SNPs appear to be due to chance alone. To investigate this further, we performed a partial correlation of coincident SNP density against the rate of recombination controlling for the single SNP density and found that the correlation is still significantly positive ($r = 0.048$, $P = 0.011$). However, despite a significant correlation, very little variation in coincident SNP density is explained by recombination rates as the correlation has a very low $r^2$ value of 0.002. It is also interesting to note that there are significant correlations between single SNP densities and the same set of genomic features as mentioned above (table 2), suggesting that there are subtle differences between the distributions of coincident and single SNPs in the human genome.

It is puzzling that the density of single SNPs does not significantly correlate with replication timing because it has been previously shown that primate divergence rates are higher in late replicating regions, suggesting that they have a higher mutation rate (Stamatoyannopoulos et al. 2009; Chen et al. 2010). Furthermore, Stamatoyannopoulos et al. (2009) showed that replication also correlates with SNP density over a scale of 100 kb, although different SNP and replication timing data sets were used in this analysis. In an attempt to explain this discrepancy, we divided human SNP density by the average divergence between human and chimp per MB to give an estimation of the effective population size ($N_e$) for each region and compared this with replication timing; we find a significant negative correlation ($r = -0.200$, $P < 0.001$). We then performed a partial correlation between single SNPs and replication timing, whilst controlling for $N_e$, and we observe a significant positive correlation ($r = 0.276$, $P < 0.001$), suggesting that a negative relationship between $N_e$ and replication timing density may be canceling out a relationship between diversity and replication timing. Furthermore, if we perform a partial correlation between coincident SNP density and replication timing, whereas controlling for $N_e$, we observe a significant positive correlation ($r = 0.102$, $P < 0.001$), although

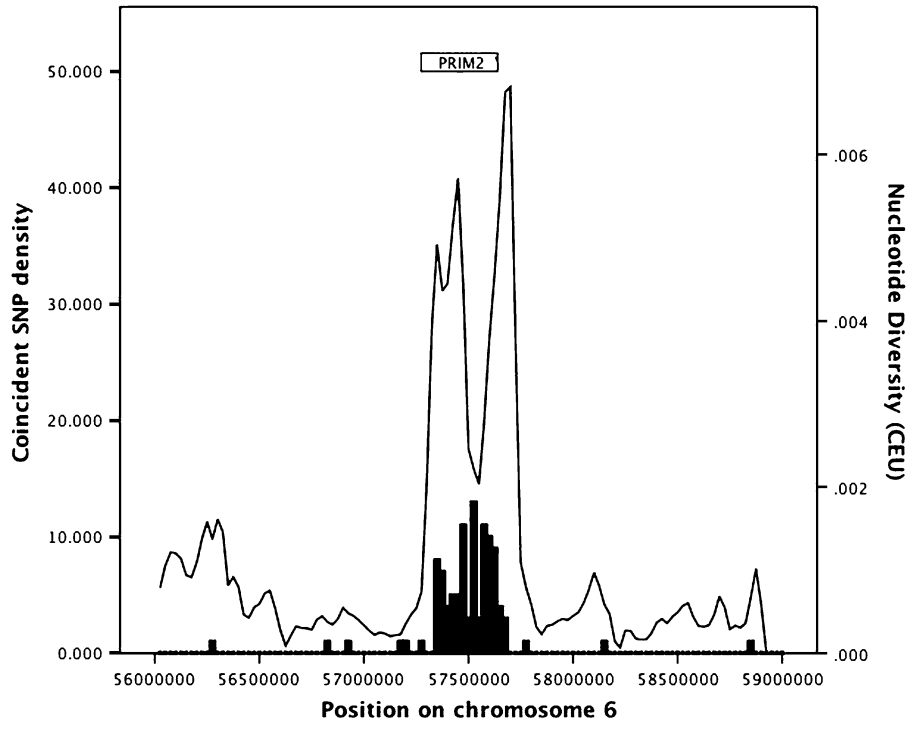


**Fig. 5.**—The nucleotide diversity across the region containing PRIM2 for the CEU population. The figure shows a sliding window of nucleotide diversity every 25 kb, with window size of 50 kb as a line graph corresponding to the right hand axis, with the coincident SNP density as a bar chart corresponding to the left-hand axis.

replication timing explains very little of the variance in the distribution of coincident SNPs (~1%).

## Regions in the Human Genome with High Numbers of Coincident SNPs

There are two 1 MB regions in the human genome that contain considerably more coincident SNPs than any other region and are outliers in the distributions of both all and non-CpG coincident SNPs. These regions are chromosome 4, 190–191 MB, which contains 57 coincident SNPs, and chromosome 6, 57–58 MB, which contains 100 coincident SNPs. The region on chromosome 4 falls very close to the end of the chromosome and contains no known genes; however, 53 coincident SNPs are found in the region running from 190712230 to 190887438 (www.hapmap.org), which is approximately 175 kb in length. A gene called PRIM2, which codes for the large DNA primase subunit, dominates the region on chromosome 6. The smaller primase subunit is encoded by PRIM1, and DNA primase is a polymerase that synthesizes small RNA primers for Okasaki fragments during discontinuous replication. PRIM2 is located at 57,290,381–57,621,334 (www.hapmap.org) and contains 86 coincident SNPs, all of which are intronic. The high number of coincident SNPs in PRIM2 and near the telomere of chromosome 4 could be due to a concentration of cryptically hypermutable sites; however, they could also be due to long-term balancing selection maintaining polymorphisms between species. It is also possible that the high concentration of coincident SNPs are a result of segmental duplications that increase the chances of chimpanzee SNPs being coincident with human SNPs if there are two or more almost identical regions on the human genome that the sequences surrounding SNPs could match to. It is important to note that this is not the cause of the significant excess of coincident SNPs in general, as this would result in a general clustering of SNPs in our original analysis, which we do not see. Similarly, balancing selection cannot explain the excess of coincident SNPs across the whole genome because we also showed that there is an excess of coincident SNPs between human and macaque; the large divergence between the two species would mean that balancing selection would be extremely unlikely (Hodgkinson et al. 2009).

To investigate the matter further, we downloaded SNP data from the 1000 genome project (http://www.1000genomes.org) for the three Hapmap populations that have already been sequenced. The region in the PRIM2 locus, which has a high concentration of coincident SNPs, also has a relatively high density of single SNPs in all three human populations (fig. 5 for the European population; supplementary fig. S1, Supplementary Material online for other populations); however, the region with the very highest density of coincident SNPs has a relatively low density of

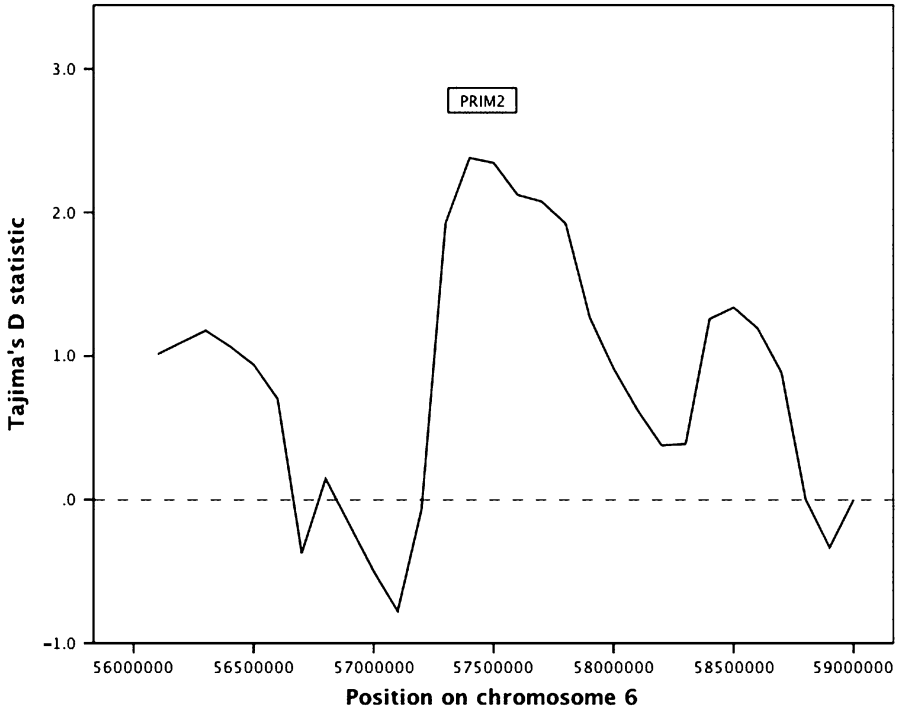


**Fig. 6.**—The Tajima's *D* statistic across the region containing PRIM2 for the CEU population. The figure shows a sliding window of nucleotide diversity every 100 kb, with window size of 200 kb.

single SNPs within the PRIM2 locus. Tajima's *D* is significantly positive in the PRIM2 locus in all three human populations ($D = 2.310$ in Europeans, 1.296 in Yorubans, and 1.253 in East Asians, $P < 0.0001$ assuming a constant rate of recombination). It could be argued that since SNP calling is conservative, many rare variants will have been missed, possibly leading to an artificially high Tajima's *D* statistic for the region. However, when we perform a sliding window analysis, calculating the Tajima's *D* statistic in each window of width 200 kb every 100 kb (overlapping windows), we clearly see a peak in the statistic above that observed in surrounding regions (fig. 6 for the European population; supplementary fig. S2, Supplementary Material online for other populations). There might therefore be some evidence of balancing selection acting in this locus, particularly, in the European population. However, under balancing selection, we might expect to see groups of divergent haplotypes, and this is not what we observe if we construct phylogenetic trees of inferred haplotypes in each population (results not shown).

In contrast, the region on chromosome 4 with a high concentration of coincident SNPs has one of the lowest densities of single SNPs in the region, especially for the African and East Asian populations (fig. 7 for the Yoruba population; supplementary fig. S3, Supplementary Material online for the other populations). Furthermore, Tajima's *D* is 1.654 for the European population, 0.827 for the East Asian population, and 0.647 for the Yoruban population in the region with the highest concentration of coincident SNPs, running from 190712230 to 190887438 on chromosome 4. Although the Tajima's *D* scores are all significantly positive assuming a constant rate of recombination ($P < 0.01$ for all populations), only the European population has a Tajima's *D* in excess of one and in all three populations, the Tajima's *D* statistic is considerably lower than those observed at PRIM2, possibly making PRIM2 a more likely candidate for balancing selection. Furthermore, as with PRIM2, if we construct phylogenetic trees of inferred haplotypes in each population, we do not see distinct groups of divergent haplotypes (results not shown).

The PRIM2 region is not part of any segmental duplications. However, for the 175 kb region on chromosome 4 that contains a high density of coincident SNPs, there are two regions that have undergone segmental duplication which contain a total of 20 coincident SNPs and span approximately 70 kb (Bailey et al. 2001, 2002). The average sequence identity between these two regions and their corresponding duplications is approximately 96.56%. Under our criteria for detecting coincident SNPs in homologous sequences between human and chimpanzee (requiring a match of at least 96 of 101 sites for alignments containing coincident SNPs) (Hodgkinson et al. 2009) and assuming that the number of mismatches per alignment is Poisson distributed, we expect that on average 46% of the chimpanzee sequences would not match to the associated duplicated region in humans. This allows us to conclude that approximately

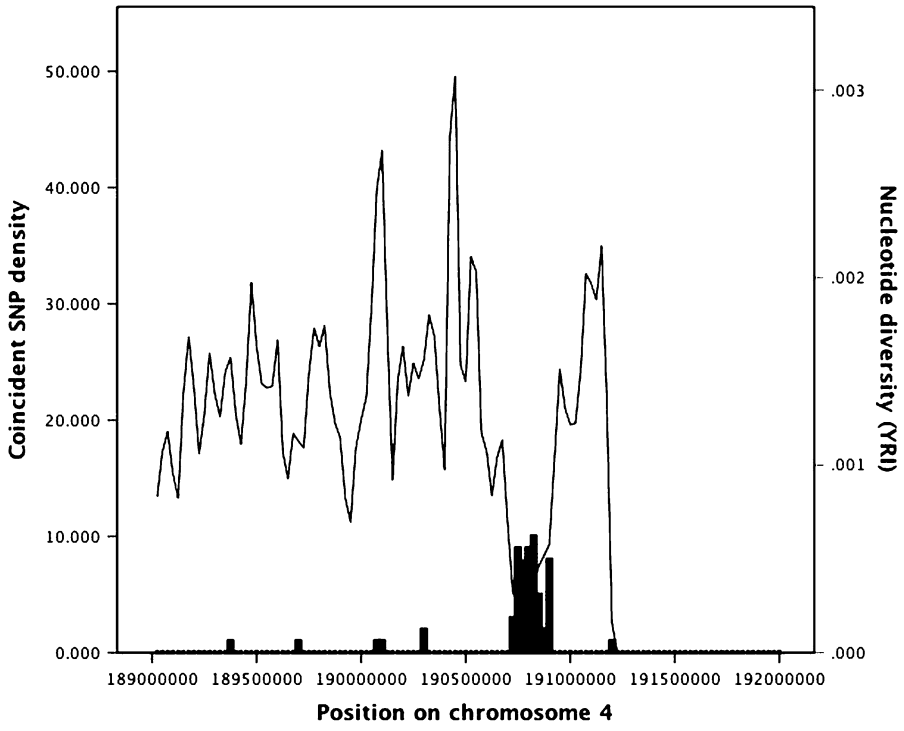


**Fig. 7.**—The nucleotide diversity across the region on chromosome 4 for the YRI population. The figure shows a sliding window of nucleotide diversity every 25 kb, with window size of 50 kb as a line graph corresponding to the right-hand axis, with the coincident SNP density as a bar chart corresponding to the left-hand axis.

42 of the 53 coincident SNPs found in the 175 kb region are correctly matched, which is still considerably above average, and therefore that segmental duplications are not having a great impact on the numbers of coincident SNPs in this region.

## Discussion

We have previously provided evidence for cryptic variation in the mutation rate in the human genome by showing that there is a significant excess of SNPs that are present at orthologous positions in human and chimp (Hodgkinson et al. 2009). We also suggested that there are complex context effects associated with coincident SNPs by showing that there was significant heterogeneity in triplet frequencies up to ~80 bp either side of coincident SNPs. However, we show here that this pattern is not specific to coincident SNPs and is instead present around CpG dinucleotides, regardless of whether the dinucleotide contains a SNP. To that end, it seems that cryptic variation in the mutation rate is not dependent upon specific local context effects that are associated with coincident SNPs and that the general excess of coincident SNPs is not driven by local context effects; indeed, no single context can explain more than a very small fraction of coincident SNPs.

The distribution of coincident SNPs across the genome is nonuniform such that some parts of the genome have higher densities of SNPs than others. However, this variation is not strongly correlated to any obvious genomic feature. There are a small number of regions that have very high numbers of coincident SNPs, and we studied these in more detail to investigate whether balancing selection might be involved. We obtained inconclusive results. Although the PRIM2 region has a high nucleotide diversity and Tajima's *D* is significantly positive, the region with the very highest density of coincident SNPs did not have a particularly high diversity. Furthermore, a phylogenetic tree of haplotypes did not reveal evidence for any deep branches, which might be indicative of long-term balancing selection. The region on chromosome 4 with a high density of coincident SNPs actually has rather low diversity for the genomic region in which it resides. Other studies have shown that genome-wide testing for balancing selection has thus far been fruitless, leading to suggestions that it is either rare or hard to find (Bubb et al. 2006). Interestingly, the MHC locus, a region that is thought to be undergoing strong balancing selection (Hughes and Nei 1988), has an average of 11.4 coincident SNPs per MB, which is far above the genome-wide average of 3.91 but markedly below the densities found at PRIM2 and the 175 kb region on chromosome 4.

Finally, the patterns around CpG dinucleotides are driven almost entirely by GC content, which increases from ~200 bp either side of the dinucleotide up to a peak immediately

adjacent to the dinucleotide. This distance-decaying pattern is almost identical to that seen in a study by Elango et al. (2008). We also observe a similar pattern in sequences surrounding GpC dinucleotides, albeit to a smaller extent, whereas the inverse is true of TpA dinucleotides, with the GC content decreasing toward the two central bases. The pattern in GpC dinucleotides is also evident in the work of Elango et al. (2008); however, they did not draw attention to this pattern because it operates over a limited scale, and they were interested in larger scale processes. It seems likely that the increase in local GC content around CpGs is caused by a process suggested by Fryxell and Zuckerkandl (2000) in which CpG dinucleotides mutate more rapidly in AT-rich regions due to an increased rate of DNA duplex melting. Cytosine deamination occurs ~143 times more often on ssDNA than it does on dsDNA (Frederico et al. 1990), thus the mutability of CpGs is closely linked to the melting temperature of the surrounding DNA. A 10% decrease in GC content reduces the melting temperature of a sequence by 4.1 $^{o}$C and thus increases the deamination of methylated cytosine by 2-fold (Fryxell and Zuckerkandl 2000). This process could also explain the increase in GC content around GpCs, as Fryxell and Moon (2005) note that unmethylated GpCs undergo deamination at lower rates in GC-rich regions due to reduced DNA melting. Furthermore, we may expect TpA dinucleotides to be present in AT-rich regions if they are remnants of former CpGs that have mutated at high rates in the past. An alternative explanation is that biased gene conversion (BGC) alters the base composition around CpG dincleotides (Elango et al. 2008). BGC is a mechanism in which base mismatches formed during recombination and the repair of double strand breaks in heterozygous individuals are preferentially repaired to GC over AT nucleotides (Marais 2003). This process may decrease the mutability of CpG dinucleotides near to recombination events if deaminated cytosines are preferentially repaired or if CpGs end up in less mutable GC-rich contexts as a result of BGC. However, this does not readily explain why the local increase in GC content is so much more conspicuous for CpG as opposed to GpC dinucleotides, and why patterns are absent around CC and GG dinucleotides. Furthermore, it is not obvious how BGC could generate a decrease in GC content around TpA dinucleotides, as regions between areas undergoing high levels of recombination, in which A and T nucleotides are relatively more likely to accumulate, are likely to be much larger than the ~200 bp over which the pattern appears to exist. In order to differentiate between the two potential mechanisms more formally we tested for the presence of peaks in GC content around CpGs in regions that do not undergo recombination on the Y chromosome in *Homo Sapiens*. As there is no recombination in these regions, and thus no BGC, we should see no differences in base composition across the sequences if the patterns around CpGs are caused by BGC. We repeated the procedure outlined above by obtaining 2,000 sequences from the Y chromosome and calculating the GC content at each position across the alignments. It has been reported that certain regions of the human Y chromosome undergo gene conversion (Skaletsky et al. 2003) and so these regions were not included in the analysis. For sequences containing CpG dinucleotides from the Y chromosome in *H. sapiens*, there is a peak of GC content that extends from −160 to 186 bp, showing that the pattern exists independently of recombination. It therefore seems that the melting temperature of different sequences can drive local biases in base composition around certain dinucleotides in the human genome.

We have shown that there are no obvious sequence contexts surrounding coincident SNPs, which we have inferred to be caused by cryptic variation in the mutation rate. Furthermore, we have failed to find any genomic feature that correlates strongly to the density of coincident SNPs. What then might cause some sites to have much higher mutation rates than others? It seems likely that it is caused by DNA topology and packaging but until we understand these processes in the germ line, we may struggle to understand this phenomenon further.

## Supplementary Material

Supplementary figures S1–S3 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

## Acknowledgments

A.H. and A.E.W. were funded by the BBSRC. Author contributions: A.H. and A.E.W. designed the analysis. A.H. collected the data and performed the analysis. A.H. and A.E.W. wrote the paper. Competing interests. The authors declare that they have no competing interests.

## Literature Cited

Bailey JA, et al. 2002. Recent segmental duplications in the human genome. Science. 297:1003–1007.

Bailey JA, et al. 2001. Segmental duplications: organization and impact within the current Human Genome Project assembly. Genome Res. 11:1005–1017.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 8:1499–1504.

Blake RD, Hess ST, Nicholson-Tuell J. 1992. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. J Mol Evol. 34:189–200.

Bubb KL, et al. 2006. Scan of human genome reveals no new loci under ancient balancing selection. Genetics. 173:2165–2177.

Chen CL, et al. 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res. 20: 447–457.

Coulondre C, et al. 1978. Molecular basis of base substitution hotspots in Escherichia coli. Nature. 274:775–780.

Elango N, et al. 2008. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. PLoS Comput Biol. 4:e1000015.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.
Frederico LA, Kunkel TA, Shaw BR. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry. 29:2532–2537.
Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol. 22:650–658.
Fryxell KJ, Zuckerkandl E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol Biol Evol. 17:1371–1383.
Gaffney DJ, Keightley PD. 2005. The scale of mutational variation in the murid genome. Genome Res. 15:1086–1094.
Gupta S, et al. 2008. Predicting human nucleosome occupancy from primary sequence. PLoS Comput Biol. 4:e1000134.
Hellmann I, et al. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet. 72:1527–1535.
Hellmann I, et al. 2005. Why do human diversity levels vary at a megabase scale? Genome Res. 15:1222–1231.
Hodgkinson A, Ladoukakis E, Eyre-Walker A. 2009. Cryptic variation in the human mutation rate. PLoS Biol. 7:e1000027.
Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.
Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class-I loci reveals overdominant selection. Nature. 335:167–170.
Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci U S A. 101:13994–14001.
Kong A, et al. 2002. A high-resolution recombination map of the human genome. Nat Genet. 31:241–247.
Lercher MJ, Williams EJ, Hurst LD. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. Mol Biol Evol. 18:2032–2039.
Li WH, Yi S, Makova K. 2002. Male-driven evolution. Curr Opin Genet Dev. 12:650–656.
Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet. 19:330–338.
Matassi G, Sharp PM, Gautier C. 1999. Chromosomal location effects on gene sequence evolution in mammals. Curr Biol. 9:786–791.
McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics. 160:1231–1241.
Skaletsky H, et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 423:825–837.
Stamatoyannopoulos JA, et al. 2009. Human mutation rate associated with DNA replication timing. Nat Genet. 41:393–395.
Takai D, Jones PA. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A. 99:3740–3745.
Venter JC, et al. 2001. The sequence of the human genome. Science. 291:1304–1351.
Williams EJ, Hurst LD. 2000. The proteins of linked genes evolve at similar rates. Nature. 407:900–903.
Zhao Z, et al. 2003. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene. 312:207–213.

**Associate editor:** Kenneth Wolfe