Supplementary Online Material

A time-invariant principle of genome evolution

Subhajyoti De and M. Madan Babu

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK,
email: SD (sde@mrc-lmb.cam.ac.uk) and MMB (madanm@mrc-lmb.cam.ac.uk)

SM-1: Control calculations for substitution rate .................................................................2
  Figure SF1: Correlation between dS vs KS and dN vs KA ..................................................2
  Figure SF2: Distribution of KS and KA for different CGN bins .........................................2
  Table ST1: Calculation of statistical significance for the association of CGN score with KS, KA, KI and KP ... 2
  Table ST2: Calculation of statistical significance for the association of CGN score with dS and dN ..........2

SM-2: Dependence on GC content ..................................................................................3
  Figure SF3: Distribution of GC content for different CGN bins .......................................3
  Figure SF4: Distribution of ΔGC content for different CGN bins ........................................3
  Table ST3: Calculation of statistical significance for the association of CGN score with %GC and ΔGC content ..3

SM-3: Dependence on recombination rate ........................................................................4
  Figure SF5: Distribution of recombination rate for different CGN bins .......................4
  Table ST4: Calculation of statistical significance for the association of CGN score with recombination rate ....4

SM-4: Sequence divergence in other mammals and SNPs in humans ..........................5
  Figure SF6: Distribution of mouse-rat and dog-cow divergence rates for different CGN bins ........5
  Figure SF7: Distribution of SNPs and polymorphic species-level differences for different CGN bins ..........6

SM-5: Association between structural variation and SNPs in human populations ............7
  Figure SF8: Distribution of SNPs at various distances from site of structural variations .......7
  Figure SF9: Single nucleotide substitution pattern at various distances from the site of structural variation ..........8
  Figure SF10: Correlation between SNP density and InDel density in personal genomes ..........9

SM-6: Comparison and analyses of the different personal genomes ..............................10
  Figure SF11: SNP density for blocks with and without InDel using KorRef as a reference ........10
  Figure SF12: SNP density for blocks with and without InDel using Yoruban as a reference .....11
  Table ST5: SNP density (above median) and InDel occurrence for 100 Kb genomic blocks ..........11
  Table ST6: SNP density (above median) and InDel occurrence for 1Mb genomic blocks ...........12
  Table ST7: SNP density (above median) and InDel occurrence for 10 Kb genomic blocks ..........12
  Table ST8: SNP density (above 75 percentile) and InDel occurrence for 100 Kb genomic blocks ....12

SM-7: Single nucleotide mutations near sites of InDels/translocation in cancer and engineered cells ... 13
  Figure SF13: Specific examples from cancer cells ..........................................................13
  Figure SF14: Distribution of SNP density as a function of distance from the alteration site ..........14
  Figure SF15: Specific examples from cells genetically engineered using Zn finger nuclease ........15

SM-8: Analysis of structural alteration between different time-scales ..........................16
  Figure SF16: CGN score distribution and genes rearrangements in cancer .........................17
  Table ST9: Overlap between structural alteration in cancer and HuRef personal genome (genomic blocks) ....17
  Table ST10: Overlap between structural alteration in cancer and HuRef personal genome (genes) .........17

SM-9: Materials and Methods .............................................................................18

References .................................................................................................................19
Supplementary Module 1: Control calculations for substitution rate

We obtained dN and dS values for 13772 human genes in our dataset from Ensembl Release 48. In addition, for a subset of 9351 genes, KA, KS, KI and KP values were available from Khaitovich et al.(1).

First, we compared equivalent parameters from the two sources (dN vs KA; and dS vs KS) to get an estimate of the agreement between them. The results are shown in Figure SF1. The correlation between equivalent parameters was very high (Spearman Rank Correlation coefficient for dN vs KA: 0.81; dS vs KS: 0.84). It shows that (i) the observations of our analysis are independent of the Ensembl version used and (ii) data from the two sources can be used in our analysis within the same conceptual framework.

Figure SF1: Correlation between dS vs KS (left) and dN vs KA. Intensity of gray colour represents density of data-points in a region. Outliers are excluded only to improve visualisation of the graphs.

We further binned the genes that have one-to-one orthologous relationship between human and chimpanzee according to their CGN scores and then compared the synonymous and nonsynonymous substitution rate, measured by KS and KA respectively, among the bins. There were 9351 genes for which KA, KS and CGN scores were available. We found that the set of genes with low CGN score have both KA and KS values systematically higher than the genome-wide average (Figure SF2 and Table ST1). Our observation is qualitatively similar to that described in the main text using dS and dN (Figure 2A and table ST2), suggesting that our conclusions are independent of the datasets used.

Table ST1: Calculation of statistical significance involving distributions of KA, KS, KI and KP values

<table>
<thead>
<tr>
<th>Category</th>
<th>CGN=0</th>
<th>CGN&gt;0</th>
<th>CGN&lt;0.4</th>
<th>CGN&gt;=0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median No. obs</td>
<td>Median No. obs</td>
<td>P value</td>
<td>Median No. obs</td>
</tr>
<tr>
<td>KS</td>
<td>0.0145 21</td>
<td>0.0105 9330</td>
<td>0.54</td>
<td>0.0145 89</td>
</tr>
<tr>
<td>KA</td>
<td>0.0028 21</td>
<td>0.0018 9330</td>
<td>0.27</td>
<td>0.0024 89</td>
</tr>
<tr>
<td>KI</td>
<td>0.0142 21</td>
<td>0.0122 9330</td>
<td>1.19E-04</td>
<td>0.0135 89</td>
</tr>
<tr>
<td>KP</td>
<td>0.0106 21</td>
<td>0.0100 9330</td>
<td>0.11</td>
<td>0.0106 89</td>
</tr>
</tbody>
</table>

Figure SF2: Distribution of KS and KA values for orthologous genes between human and chimpanzee grouped accordingly to their CGN scores. The genome-wide median values are shown as horizontal lines. Outliers are excluded only to improve visualisation of the graphs.

Table ST2: Calculation of statistical significance involving distributions of dN and dS values

<table>
<thead>
<tr>
<th>Category</th>
<th>CGN=0</th>
<th>CGN&gt;0</th>
<th>CGN&lt;0.2</th>
<th>CGN&gt;=0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median No. obs</td>
<td>Median No. obs</td>
<td>P value</td>
<td>Median No. obs</td>
</tr>
<tr>
<td>dN</td>
<td>0.0045 57</td>
<td>0.0029 13715</td>
<td>1.06E-03</td>
<td>0.0038 191</td>
</tr>
<tr>
<td>dS</td>
<td>0.0145 57</td>
<td>0.0113 13715</td>
<td>7.44E-03</td>
<td>0.0126 191</td>
</tr>
<tr>
<td>dN/dS</td>
<td>0.3725 55</td>
<td>0.2682 12895</td>
<td>0.037</td>
<td>0.3380 184</td>
</tr>
</tbody>
</table>
Supplementary Module 2: Dependence on GC content

We collected the % GC content information for human and chimpanzee genes from Ensembl Release 48. There were 16448 genes for which both % GC content and CGN scores were available. We grouped the genes in our dataset according to their CGN scores and then compared the distribution of GC content for human and chimpanzee, among the groups. Results are shown in Figure SF3 and Table ST3. We found that in both the cases, % GC content of the genes with low CGN is systematically lower compared to that of the genes with high CGN score.

Figure SF3: Distribution of % GC content for human (left) and chimpanzee (right) for orthologous genes between human and chimpanzee grouped accordingly to their CGN scores. The genome-wide median value is shown as horizontal lines. Outliers are excluded only to improve visualisation of the graphs.

Next we calculated the difference in %GC content of the orthologous genes between human and chimpanzee. We investigated whether genes with altered genomic neighbourhood have high difference in the GC content between the orthologous gene-pairs in order to estimate the impact of biased gene conversion on the sequence-level evolution of such genes. We calculated change in GC content by two measures. We calculated (i) $\delta_{GC} = \text{abs}(G_{\text{human}} - G_{\text{chimp}})$ and (ii) $\Delta_{GC} = \text{abs}(G_{\text{human}} - G_{\text{chimp}})/\text{average}(G_{\text{human}}, G_{\text{chimp}})$ and then plotted the distribution of $\delta_{GC}$ and $\Delta_{GC}$ for the genes grouped according to their CGN scores. Results are shown in Figure SF4 and Table ST3. We found that although the range of $\delta_{GC}$ and $\Delta_{GC}$ for genes with low CGN are relatively large, overall there was no consistent and significant difference in median between the distributions of change in GC content ($\delta_{GC}$, $\Delta_{GC}$) among the CGN groups.

Figure SF4: Distribution of $\Delta_{GC}$ content for orthologous genes between human and chimpanzee grouped accordingly to their CGN scores. The genome-wide median value is shown as horizontal lines. Outliers are excluded only to improve visualisation of the graphs.

Table ST3: Calculation of statistical significance involving distributions of $G_{\text{human}}$, $G_{\text{chimp}}$ and $\Delta_{GC}$

<table>
<thead>
<tr>
<th>Category</th>
<th>CGN=0</th>
<th>CGN&gt;0</th>
<th>P value</th>
<th>CGN&lt;0.2</th>
<th>CGN&gt;=0.2</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median</td>
<td>No. obs</td>
<td>Median</td>
<td>No. obs</td>
<td>Median</td>
<td>No. obs</td>
</tr>
<tr>
<td>GC human</td>
<td>41</td>
<td>98</td>
<td>45</td>
<td>16746</td>
<td>1.22E-05</td>
<td></td>
</tr>
<tr>
<td>GC chimp</td>
<td>40.5</td>
<td>98</td>
<td>43</td>
<td>16746</td>
<td>6.61E-04</td>
<td></td>
</tr>
<tr>
<td>$\delta_{GC}$</td>
<td>1</td>
<td>98</td>
<td>1</td>
<td>16746</td>
<td>0.335</td>
<td></td>
</tr>
</tbody>
</table>
| $\Delta_{GC}$ | 0.029 | 98     | 0.024   | 16746   | 2.44E-02 |        | 0.024   | 266     | 0.027    | 16578    | 0.069
Supplementary Module 3: Dependence on recombination rate

We collected the recombination rate data for different regions in the human genome from the UCSC genome browser(2). There were 16,448 genes for which both recombination rate (cM/Mb) and CGN scores were available. We grouped the genes in our dataset according to their CGN scores and then compared the distribution of recombination rate, among the groups. The results are shown in Figure SF5 and Table ST4. We found that there is no consistent and statistically significant difference in the distribution of recombination rate between genes with low and high CGN.

Figure SF5: Distribution of recombination rate for orthologous genes between human and chimpanzee grouped accordingly to their CGN scores. The genome-wide median value is shown as horizontal lines. Outliers are excluded only to improve visualisation of the graphs.

Table ST4: Calculation of statistical significance involving distributions of recombination rate

<table>
<thead>
<tr>
<th>Category</th>
<th>CGN=0</th>
<th>CGN&gt;0</th>
<th>P value</th>
<th>CGN&lt;0.2</th>
<th>CGN&gt;=0.2</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recombination rate (cM/Mb)</td>
<td>Median 0.96</td>
<td>No. obs 78</td>
<td>Median 1.13</td>
<td>No. obs 16370</td>
<td>P value 0.322</td>
<td>Median 1.05</td>
</tr>
</tbody>
</table>
Supplementary Module 4: Sequence divergence in other mammals and SNPs in humans

To test how the genes with altered neighbourhood between human and chimpanzee are evolving during mammalian evolution, we obtained the dN, dS and dN/dS ratio for these genes for mouse-rat pair and cow-dog pair from ENSEMBL (3).

For a subset of 13,055 human genes we had both CGN score (calculated between human and chimpanzee) and dN, dS values calculated between mouse and rat orthologous pairs. We grouped the genes based on their CGN scores and plotted distribution of dS, dN and dN/dS. We repeated the analysis for a subset of 12,070 human genes for which both CGN score (calculated between human and chimpanzee) and dN, dS values calculated between mouse and rat were available. In both cases, unlike that in Figure 2, we did not observe high dN, dS and dN/dS ratio for the genes with low CGN (Figure SF6). Our observations show that the genes that have altered neighbourhood between human and chimpanzee diverge rapidly between human and chimpanzee, but not in the mouse-rat and dog-cow species-pairs. Therefore, our observation that genes with altered neighbourhood evolve rapidly at the sequence-level is not biased by presence of the genes that are anyway evolving rapidly among mammals, and evolution at the two levels appears to be linked temporally.

Next we analysed how the genes with low CGN are evolving within human populations. To investigate it, we estimated the extent of single nucleotide polymorphisms in the genes for which CGN score was calculated. In the absence of systematic resequencing-based SNPs map of human and chimpanzee genomes it is difficult to perform a methodical survey of how the coding, promoter and intronic regions of the genes with low CGN are evolving within human and chimpanzee populations. Despite the limitations, we performed a rough analysis taking only nonsynonymous SNPs (nsSNPs) that were deposited in dbSNP v127. The HapMap project aimed to capture nearly all of the common nsSNPs (minor allele frequency>0.05) and deposited them in dbSNP.

We collected the nsSNPs in protein-coding regions in humans that are reported in dbSBP v127. From the human-chimpanzee pairwise protein-sequence alignments we identified 82,463 species-level amino-acid differences between human and chimpanzees. We plotted the distribution of (i) the number of species-level mutations per peptide length, (ii) the number of nsSNPs per peptide length and (iii) the proportion of species-level nonsynonymous mutations that are polymorphic in human populations, for the genes grouped according to their CGN scores.
Figure SF7: Distribution of species-level aminoacid mutations (top), nsSNPs (middle) and proportion of species-level amino acid mutations that are also polymorphic (bottom) for genes grouped according to their CGN scores calculated between human and chimpaznee. The genome-wide median values are shown as horizontal lines. Outliers are excluded only to improve visualisation of the graphs.

While species-level mutations followed the pattern we have observed for dN (human-chimpanzee), the distribution of nsSNPs and proportion of species-level amino acid mutations that are also polymorphic in humans didn’t show the same trend (Figure SF7). Therefore, (i) not all of the genes with low CGN that diverge rapidly between human and chimpanzee also harbour high proportion of nsSNPs, and (ii) a considerable proportion of mutations accumulated at the species-level are not polymorphic in human populations.
Supplementary Module 5: Association between structural variation and SNPs in human populations

At the population divergence time scale (~tens of thousands of years), we investigated whether regions with structural alteration (InDels) are correlated with increased single nucleotide polymorphisms in the sequenced genomes of individuals from different populations that diverged several centuries ago. We repeated the analysis described in the main text for structural variations of size > 500 bp for the HuRef genome (4) and those of size > 20bp in the Yoruban (NA18507)(5) and the Korean (KorRef)(6) genomes. The choice of alternative window sizes was guided by the availability of larger insertions and deletions reported in the respective personal genomes; e.g. the HuRef genome had relatively more structural variations of a larger size compared to the Yoruban and KorRef genomes. In all the three cases we find results consistent with that in Figure 3 in the main text i.e. that the density of single nucleotide polymorphisms is high in a relatively narrow window immediately close to the structural variation, and decays with increasing distance away from the structural variation event (Figure SF8).

Figure SF8: Distribution of the density of single nucleotide change as a function of distance from the site of structural alteration in (A) HuRef genome, structural variations of size > 500bp, (B) Yoruban (NA18507) genome, structural variations of size > 20bp and (C) KorRef genome, structural variations of size > 20bp.
Next, for each of the SNPs analyzed we investigated whether the substitution falls into one of the following categories: AC, AG, AT, CG, CT, GT. Please note that while in one strand AG substitution will lead to CT substitution in the other strand, for each SNP we took the allele states as reported in the publication (i.e. either AG or CT). For windows of variable sizes (1kb, 10kb, 100kb, 1Mb) around structural variations (of size >500bp; n=673), we calculated the proportion of SNPs ($p_{ijk}$, where $i$: substitution category, $j$: window size and $k$: structural variation event) for each of the six substitution categories. Then we calculated the average proportion across all structural variation events ($p_{ij} = \frac{1}{n} \sum p_{ijk}$) and the average value for each substitution category for each window size (Figure SF9) was obtained. We found that there is no significant correlation between proportion of SNPs in each category and distance from the sites of structural variation, suggesting that biased gene conversion is unlikely to affect our observation.

The proportions of substitutions appear to be the same at different distances from the site of structural alteration (Figure SF9) suggesting no substitution bias in the vicinity. Please note that the ratio of transitions (AG and CT) to transversions (AT, AC, GT, GC) is consistent with what is expected of neutral mutations, which is 3:2.

![Figure SF9: Proportion of SNPs in different substitution categories (AC, AG, AT, CG, CT, GT) is shown for various distances from the site of structural alteration in the HuRef genome for structural alterations of size >500bp. Proportions are averaged across all structural alterations of size >500bp (n=673).](image)

We scanned the human genome using a 1Mb non-overlapping window and for each window we calculated the density of structural variations (InDel density) and the density of single nucleotide polymorphisms (SNP density). For InDel density calculation, we counted the number of base-pairs affected in the 1Mb windows. Since InDels of very large sizes (e.g. 1kb or larger) are rare and even one occurrence of such a large structural variation can skew our observation, we first compared the SNP density and InDel density by considering only the InDels of size 100bp or smaller (Figure SF10). We found that the SNP density and InDel density are correlated (Spearman correlation coefficient $=0.66; p < 2.2\times10^{-16}$). We repeated our analysis by (i) including InDels of 1kb or smaller (Spearman correlation coefficient $=0.51; p < 2.2\times10^{-16}$) and (ii) all InDels (Spearman correlation coefficient $=0.41; p < 2.2\times10^{-16}$), and found that our observation is consistent.

Next we repeated the analysis using SNPs and polymorphisms reported in the Yoruban (NA18507) and the Korean Reference (KorRef) genome using similar strategy (Figure SF10). Please note that in the NA18507 and KorRef genomes, few large structural variations were reported. Therefore we didn’t apply any InDel size restriction. In all the genomes, we found high correlation between the genome-wide patterns of SNP density and InDel density.
Figure SF10: Correlation between SNP density and InDel density for non-overlapping 1Mb windows across the HuRef genome (Spearman correlation coefficient: 0.66; \( p < 2.2 \times 10^{-16} \)), the Yoruban (NA18507) genome (Spearman correlation coefficient: 0.82; \( p < 2.2 \times 10^{-16} \)), and the Korean Reference (KorRef) genome (Spearman correlation coefficient: 0.83; \( p < 2.2 \times 10^{-16} \)). SNPs and InDels were calculated with respect to the reference human genome (NCBI 36).
Supplementary Module 6: Comparison across different personal genomes

We first divided the personal genomes (HuRef (4), Yoruban (5), and KorRef (6)) into non-overlapping 100kb genomic blocks and calculated the density of single nucleotide substitutions relative to the reference human genome (NCBI36). We found that the genome-wide single nucleotide substitution densities are highly correlated across the three personal genomes (Spearman correlation coefficient: 0.7-0.8). In other words, if a genomic block harbors high (or low) single nucleotide substitution density in one genome, it is likely to harbor high (or low) single nucleotide substitution density also in the second genome, suggesting that mutational hotspots and coldspots are mostly shared between the personal genomes.

6.1 Analysis of equivalent genomic regions between personal genomes

To investigate the possible causal relationship between InDels and SNPs, we compared equivalent regions from pairs of personal genomes. It should be noted that the experimental approach to sequence the distinct individuals are different for the three personal genomes and the absolute number of mutations identified (e.g., SNP densities) are themselves not directly comparable. Therefore, we only analysed the Yoruban and KorRef genomes that were comparable in terms of the number of structural alterations (of size >30bp, >20bp and >10bp) and single nucleotide substitution density.

We first used the Yoruban genome as a frame of reference and identified all genomic blocks that did not have any structural alteration (Figure SF11 and SF12). We then divided this set of genomic blocks into two subsets: the first subset contained blocks with at least one structural alteration in the equivalent segment of KorRef genome and the second subset contained blocks that did not contain a structural alteration in an equivalent segment of the KorRef genome. As a null hypothesis, if the local single nucleotide density depends on factors other than occurrence of structural alteration in the vicinity, then one should expect that these two sets of genomic blocks should show comparable single nucleotide substitution densities. In contrast, if the occurrence of structural alteration drives single nucleotide substitution events in the vicinity, one may expect the first subset of genomic blocks (those with at least one structural alteration) to have higher single nucleotide substitution density than the second subset (those with no structural alteration). Indeed we find that the genomic blocks with at least one structural alteration show significantly higher single nucleotide substitution density than those with no structural alteration in the KorRef genome (one tailed Mann Whitney test, p value: <2.2x10^{-16}). By performing a complementary analysis, using KorRef as the frame of reference, we found that the genomic blocks with at least one structural alteration tend to show significantly higher single nucleotide substitution density than those with no structural alteration in the Yoruban genome (one tailed Mann Whitney test, p value: <2.2x10^{-16}).

Figure SF11: (A) SNP density for genomic blocks (of size 100kb) in the Yoruban genome which have at least one structural alteration (small red rectangles) is significantly different from that of the genomic blocks which have no structural alteration, given that both sets of blocks have no InDels in the KorRef genome (p<2.2x10^{-16}). (B) SNP density (x10^{-5}) for genomic blocks (of size 100kb) in the KorRef genome which have at least one structural alteration is significantly different from that of the genomic blocks which have no structural alteration, given that both sets of blocks have no structural alteration in the Yoruban genome (p<2.2x10^{-16}).
We repeated the analyses by excluding mutational hotspots and by using only the genomic blocks that contained the lower 25 percentile single nucleotide substitution density in the reference genome. Even after such a filtering, we found that in both the genomes, the genomic blocks with at least one structural alteration show significantly higher single nucleotide substitution density than those with no structural alteration (Figure SF12, one tailed Mann Whitney test, KorRef genome: p value: 6.86x10^{-05}, Yoruban genome: p value: 1.51x10^{-10}).

Figure SF12: (A) SNP density for genomic blocks (of size 100kb) in the Yoruban genome which have at least one structural alteration (small red rectangles) is significantly different from that of the genomic blocks which have no structural alteration, given that both sets of blocks have no InDels in the KorRef genome (p=1.51x10^{-10}). (B) SNP density (x10^{-5}) for genomic blocks (of size 100kb) in the KorRef genome which have at least one structural alteration is significantly different from that of the genomic blocks which have no structural alteration, given that both sets of blocks have no structural alteration in the Yoruban genome (p=6.86x10^{-05}).

6.2 Analysis of SNP density and InDel occurrence in the personal genomes

We divided each of the three personal genomes into 100kb non-overlapping genomic blocks, and identified the blocks that have (i) at least one large structural variation (20bp, 30bp and 500bp), (ii) above median single nucleotide substitution density and (iii) both above median single nucleotide substitution density and at least one large structural variation (20bp, 30bp and 500bp). We then calculated (i) the conditional probability of finding genomic blocks with at least one structural alteration given above median single nucleotide substitution density and (ii) conditional probability of finding genomic blocks with above median single nucleotide substitution density given at least one structural alteration. We found that in all the three genomes, the conditional probability of finding genomic blocks with above median single nucleotide substitution density given at least one structural alteration is much higher than the converse, i.e., the conditional probability of finding genomic blocks with at least one structural alteration given above median single nucleotide substitution density. These results indicate that the presence of structural variation is one of the factors that can increase single nucleotide substitution density in the vicinity, while the converse (i.e., presence of high SNP density does not result in an increased likelihood of finding a structural alteration) is not true. Our results hold true for different sizes of genomic blocks, varying up to three orders of magnitude (window size 1Mb, 100kb and 10kb) and different thresholds for high single nucleotide substitution density (Table ST5-8, next page). This suggests that our observation is independent of the choice of thresholds used.
increased likelihood of finding a structural alteration) is likely to be a weak effect. Substitution density in the vicinity, while the converse (i.e., presence of high SNP density does not result in an increased probability of finding at least one structural variation is one of the factors that can increase single nucleotide substitution density in the vicinity, while the converse (i.e., presence of high SNP density does not result in an increased likelihood of finding a structural alteration) is likely to be a weak effect.

Table ST5: Association between high single nucleotide substitution density (above median) and occurrence of at least one structural alteration event in the personal genomes over non-overlapping 100kb genomic blocks.

<table>
<thead>
<tr>
<th>Personal Genomes</th>
<th>Size of structural variation</th>
<th># genomic blocks with at least one structural alteration</th>
<th># genomic blocks with &gt;median single nucleotide substitution density</th>
<th># genomic blocks with &gt;median single nucleotide substitution density AND at least one structural alteration</th>
<th>Conditional probability of &gt;median single nucleotide substitution density given at least one structural alteration</th>
<th>Conditional probability of at least one structural alteration given &gt;median single nucleotide substitution density</th>
</tr>
</thead>
<tbody>
<tr>
<td>HuRef</td>
<td>&gt;30bp</td>
<td>2409</td>
<td>1513</td>
<td>285</td>
<td>0.566</td>
<td>0.188</td>
</tr>
<tr>
<td></td>
<td>&gt;500bp</td>
<td>469</td>
<td>1513</td>
<td>127</td>
<td>0.668</td>
<td>0.083</td>
</tr>
<tr>
<td>KORREF</td>
<td>&gt;30bp</td>
<td>193</td>
<td>15104</td>
<td>325</td>
<td>0.690</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>&gt;500bp</td>
<td>471</td>
<td>15104</td>
<td>269</td>
<td>0.629</td>
<td>0.177</td>
</tr>
</tbody>
</table>

Table ST6: Association between high single nucleotide substitution density (above median) and occurrence of at least one structural alteration event in the personal genomes over non-overlapping 1Mb genomic blocks.

<table>
<thead>
<tr>
<th>Personal Genomes</th>
<th>Size of structural variation</th>
<th># genomic blocks with at least one structural alteration</th>
<th># genomic blocks with &gt;median single nucleotide substitution density</th>
<th># genomic blocks with &gt;median single nucleotide substitution density AND at least one structural alteration</th>
<th>Conditional probability of &gt;median single nucleotide substitution density given at least one structural alteration</th>
<th>Conditional probability of at least one structural alteration given &gt;median single nucleotide substitution density</th>
</tr>
</thead>
<tbody>
<tr>
<td>HuRef</td>
<td>&gt;30bp</td>
<td>5558</td>
<td>15104</td>
<td>4003</td>
<td>0.720</td>
<td>0.265</td>
</tr>
<tr>
<td></td>
<td>&gt;500bp</td>
<td>601</td>
<td>15104</td>
<td>427</td>
<td>0.710</td>
<td>0.028</td>
</tr>
<tr>
<td>KORREF</td>
<td>&gt;20bp</td>
<td>193</td>
<td>15104</td>
<td>152</td>
<td>0.787</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>&gt;500bp</td>
<td>471</td>
<td>15104</td>
<td>325</td>
<td>0.690</td>
<td>0.021</td>
</tr>
</tbody>
</table>

Table ST7: Association between high single nucleotide substitution density (above median) and occurrence of at least one structural alteration event in the personal genomes over non-overlapping 10kb genomic blocks.

<table>
<thead>
<tr>
<th>Personal Genomes</th>
<th>Size of structural variation</th>
<th># genomic blocks with at least one structural alteration</th>
<th># genomic blocks with &gt;median single nucleotide substitution density</th>
<th># genomic blocks with &gt;median single nucleotide substitution density AND at least one structural alteration</th>
<th>Conditional probability of &gt;median single nucleotide substitution density given at least one structural alteration</th>
<th>Conditional probability of at least one structural alteration given &gt;median single nucleotide substitution density</th>
</tr>
</thead>
<tbody>
<tr>
<td>HuRef</td>
<td>&gt;30bp</td>
<td>2409</td>
<td>1513</td>
<td>285</td>
<td>0.566</td>
<td>0.188</td>
</tr>
<tr>
<td></td>
<td>&gt;20bp</td>
<td>193</td>
<td>15104</td>
<td>127</td>
<td>0.668</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>&gt;500bp</td>
<td>471</td>
<td>15104</td>
<td>269</td>
<td>0.629</td>
<td>0.177</td>
</tr>
</tbody>
</table>

Table ST8: Association between high single nucleotide substitution density (above 75 percentile) and occurrence of at least one structural alteration in the personal genomes over non-overlapping 100kb genomic blocks.

<table>
<thead>
<tr>
<th>Genome</th>
<th>Size of structural variation</th>
<th># genomic blocks with at least one structural alteration</th>
<th># genomic blocks with &gt;75 percentile single nucleotide substitution density</th>
<th># genomic blocks with &gt;75 percentile single nucleotide substitution density AND at least one structural alteration</th>
<th>Conditional probability of &gt;75 percentile single nucleotide substitution density given at least one structural alteration</th>
<th>Conditional probability of at least one structural alteration given &gt;75 percentile single nucleotide substitution density</th>
</tr>
</thead>
<tbody>
<tr>
<td>HuRef</td>
<td>&gt;30bp</td>
<td>5558</td>
<td>7552</td>
<td>2318</td>
<td>0.417</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>&gt;500bp</td>
<td>601</td>
<td>7552</td>
<td>274</td>
<td>0.456</td>
<td>0.036</td>
</tr>
<tr>
<td>KORREF</td>
<td>&gt;20bp</td>
<td>193</td>
<td>7552</td>
<td>99</td>
<td>0.513</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>&gt;500bp</td>
<td>471</td>
<td>7552</td>
<td>199</td>
<td>0.422</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Please note that Levy et al. (4) reported a relatively larger number of InDels in the HuRef genome (compared to the other personal genomes). A vast majority (>90%) of the genomic blocks of size 1Mb had at least one InDel (size>30bp). Therefore, the conditional probability of finding at least one InDel (size>30bp) in genomic blocks with above median SNP density was high ($p=0.901$), but comparable to the conditional probability ($p=0.919$) of finding at least one InDel in the genomic blocks with above median SNP density in table ST6. Therefore, it is consistent with the proposal that the presence of structural variation is one of the factors that can increase single nucleotide substitution density in the vicinity, while the converse (i.e., presence of high SNP density does not result in an increased likelihood of finding a structural alteration) is likely to be a weak effect.
Supplementary Module 7: Association between structural alteration and SNPs in cancer tissues and in genetically engineered cells using zinc finger nucleases

7.1 Specific examples of the association in cancer cells

We collected information from the literature where sequence read map to regions of (i) gene fusions in prostate cancer and (ii) melanoma (Figure SF13)(7-9). Inspection of sequencing reads from the high-throughput sequencing approaches that map to the translocation or fusion sites in cancer cells clearly revealed instances of single nucleotide substitutions in certain sequenced reads in the vicinity of the structural alteration event (Figure SF13). This may reflect the underlying heterogeneity in the cell population, which might have appeared stochastically in individual cells in the cell population due to the structural alteration event (see possible mechanism in main text). Single nucleotide mutations observed in such reads are unlikely to be sequencing or mapping artifacts as (i) we observe them in independent studies and in different genomic loci and (ii) the locus with the structural alteration (fusion or translocation) mapped by the read will be unique and will only be found in the genome of the cancer cell investigated. Furthermore, the observed frequency of mutations appears to be higher compared to that expected due to sequencing or mapping artefacts. We hope that the availability of more cancer resequencing data will help us address this problem in a more systematic way.

Figure SF13: Single nucleotide mutations (substitutions marked in a red box) in the proximity of the site of translocation in different genomic locations from human prostate cancer cell line VcaP (A-D) and melanoma MeWo cells (E-F). (A and B) TMPRSS2–ERG fusion on chromosome 21 from Maher et al (7) and Maher et al (8) (C) HJURP–EIF4E2 fusion on chromosome 2 from Maher et al (7) and (D) USP10–ZDHHC7 fusion on chromosome 16 from Maher et al (7). (E) RECK–ALX3 fusion on chromosome 9 and 1 from Berger et al (9). (F) CDK2–RAB5B fusion on chromosome 12 from Berger et al (9).
7.2 Genome-scale evidence for the association between structural alteration and single nucleotide change in melanoma

We collected data on structural alterations and re-sequencing based somatic single nucleotide substitution data obtained by comparing malignant melanoma and lympho-blastoid cell-line from one cancer patient (10). We then investigated whether the genomic regions with de novo structural alterations also show increased single nucleotide substitutions at the time scale of multiple cell divisions in somatic tissue (~days). We considered only the structural alterations with quality score >4 for the analysis described in the main text and here. We repeated the analysis described in the main text for structural alterations of size > 10 bp and >20bp. The choice of alternative window sizes was guided by the availability of structural alterations reported in the melanoma genome. In both the cases we find results consistent with that described in Figure 4 in the main text i.e. that the density of single nucleotide polymorphisms is high in a relatively narrow window immediately close to the structural variation, and decays with increasing distance away from the structural variation event (Figure SF14).

Figure SF14: Distribution of the density of single nucleotide change as a function of distance from the site of structural alteration in the cancer (melanoma) genome (left) structural variations of size > 10bp and (right) structural variations of size > 20bp. Y axis represents the density of single nucleotide substitutions while X axis represents distance from the site of structural alteration event.

7.3 Specific examples of the association in genetically engineered cells using zinc finger nucleases

We systematically collected and analysed sequences of clones from normal individual cells that were obtained after insertion or deletion of the target sequences through genetic engineering by zinc finger nucleases from humans (11), Zea mays (12) and zebra fish (13). While the genetic engineering procedure correctly inserted or deleted the template sequence as expected in the right genomic location, we noticed a number of single nucleotide changes in the vicinity of the alteration site in certain clones from individual cells (Figure SF15, next page). This may again indicate the underlying heterogeneity in the engineered cell population that might have been introduced due to the engineered InDel event (see possible mechanism in the main text). Taken together, our observations suggest that genomic alterations involving insertion or deletion and single nucleotide changes may even be linked in normal and diseased somatic cells. Further experiments that involve insertion and deletion of genetic material in cultured cell lines or germ cells in whole organisms, followed by monitoring of sequence divergence over successive divisions or generations, could provide a more resolved understanding of this phenomenon.

Though the process of engineering using zinc finger nuclease is known to be imprecise (14) and the mechanism of zinc finger nuclease engineering has not been subjected to millions of years of evolutionary pressure, it uses the same repair pathway involving molecules that are evolutionarily conserved and have relatively low fidelity (i.e., homologous recombination repair pathway involving non-replicative polymerases). It is also clear that the imprecision of the error prone polymerase introduces genetic diversity in a population of cells that have been engineered. While it would be ideal to provide a statistical support for our observations, unfortunately, there is no data on sequences from gene targeting studies and no study (to our knowledge) provides information of the sequences around the engineered site in a systematic way. Our observation from three separate studies in three different organisms suggests that it is unlikely to be an artefact. However, we hope that future systematic studies on zinc finger nucleases will provide better insight and statistical support for our observations.
Figure SF15: Single nucleotide mutations (substitution, insertion or deletion marked in a red box) in the proximity of the sites of structural alterations in (A) CCR locus in humans from Lee et al (11) (B) IPK1 locus in *Zea mays* using the zinc-finger nuclease from Shukla et al (12) and (C) the gol locus in zebrafish using zinc-finger nuclease from Doyon et al (13).
Supplementary Module 8: Analysis of structural alteration between different times-scales

8.1 Species and population level analysis

We obtained the positions of human segmental duplications (SD) from the Segmental Duplication Database (http://humanparalogy.gs.washington.edu/) and identified the 318 genes that are located in SD regions and also had CGN scores calculated. We found that the genes that reside within SDs are more likely to have undergone a change in genomic neighbourhood, and have significantly low CGN score distribution as compared to that of all genes in the human genome (p value <1.0E-10) (see also supplementary material of De et al. (15)). Next, we collected data on copy number variations specific to human and chimpanzee lineages and those shared between the two species from Perry et al. (16). We observed that 337, 352 and 483 genes overlap with human-specific CNVs, Chimpanzee specific CNVs and shared CNVs, respectively, and also have their CGN score calculated. We compared the CGN score of genes in each group against the distribution of CGN score for all genes in the human genome using the Mann-Whitney test. We found that the genes, which overlap with CNVs shared between human and chimpanzee populations have significantly low CGN score (p value <0.01) as compared to the genome-wide background (De et al. (15) Supplementary Material). Finally, we collected data on inversion polymorphisms in the human HapMap populations from Bansal et al. (17). From that list, 105 genes also had their CGN scores available. We compared the CGN score of these genes against the distribution of CGN score for all genes in the human genome using the Mann-Whitney test. We found that the genes, which overlap with inversion polymorphisms in human HapMap populations have significantly low CGN score (p value <1.0E-4) as compared to the genome-wide background (De et al. (15) Supplementary Material). Taken together, our analyses indicate that regions that harbour segregating structural variations in the human population (population-divergence timescale) are more likely to have altered genomic neighbourhood between human and chimpanzee (speciation timescale). The structural variation data on Yoruban (5) and KorRef (6) were not suitable for similar analysis. The HuRef genome (4) had copy number variation data, but many of the CNVs were also common CNVs that were already present in the datasets of Bansal et al. and Perry et al., which we have already used in the analysis above. The number of de novo CNVs for which the genes also had CGN score available was very small to perform a statistically meaningful analysis.

8.2 Species and cancer level analysis

We obtained a list of 1282 Refseq genes that underwent somatic rearrangements in breast cancer genomes from Stephens et al. (18). The authors used a paired-end sequencing strategy to identify somatic rearrangements in 24 breast cancer samples. Of these genes, 1063 had also CGN scores available. We found that the genes that underwent somatic rearrangement in the breast cancer samples had significantly lower CGN score relative to other genes in the dataset of De et al. (15) (one tailed Mann Whitney test, p value: 1.55x10^{15}). Our result suggests that the genes in rearranged regions in the breast cancer samples are likely to be in an unstable genomic region that is prone to change in genomic neighborhood both in speciation time scale and in the cancer genome.

Of these rearranged genes found in the work of Stephens et al. (18), many are likely to be passenger mutation, while some will be driver cancer genes. Of the 1282 cases, 34 were known cancer genes. When we compared the distribution of CGN score of these known cancer genes with the CGN score of all other genes, we find that the cancer genes that underwent somatic rearrangement in the breast cancer samples had lower CGN score relative to other genes but the statistical significance is weak (one tailed Mann Whitney test, p value: 2.35x10^{7}). We also collected all the known cancer genes reported in COSMIC and compared the distribution of CGN score of these genes with all other genes in the dataset of De et al. (15) We found that the known cancer genes in Cancer Gene Census does not have a significantly different CGN score relative to other genes in the dataset of De et al. (15) (one tailed Mann Whitney test, p value: 0.15). Taken together, our results suggest that genes implicated in cancer are not more pronounced to change their genomic neighborhood in the speciation time scale. These are key developmental genes, and it is perhaps not surprising that in healthy organisms rearrangements leading to change in genomic neighborhood are not favored.
Figure SF16: (A) Boxplot showing distribution of CGN score of the genes that underwent somatic rearrangement in the breast cancer samples and all other genes in the dataset of De et al. (15). (B) Boxplot showing distribution of CGN score of the genes that are implicated in cancer in COSMIC and all other genes in the dataset of De et al (15).

8.3 Population and cancer level analysis

Next, we investigated whether the regions that undergo structural alteration (insertion or deletion of size >10bp) in a melanoma genome as reported by Pleasance et al. (10) are also enriched in structural alterations (insertion or deletion of size >10bp, >30bp and >500bp) in the HuRef genome in the population time-scale. We divided the genome in 100kb blocks and identified blocks that have at least one structural alteration in the melanoma (or the HuRef) genome. We counted the number of blocks that have at least one structural alteration in both the melanoma and HuRef genome and calculated statistical significance of the overlap using hyper-geometric distribution (Table ST9). We found that structural alterations are relatively rare in the cancer genome relative to the HuRef genome, and that the overlap is significantly higher than expected by chance, suggesting that many regions that undergo structural alteration in cancer are also likely to be unstable at the population timescale.

Table ST9: Overlap between structural variations in cancer and HuRef personal genomes.

<table>
<thead>
<tr>
<th>Type of gene in Stephens et al.</th>
<th># genes undergoing rearrangement in breast cancer</th>
<th># genes showing CNVs in the HuRef genome</th>
<th>Overlap</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rearranged genes</td>
<td>1153</td>
<td>129</td>
<td>5</td>
<td>1.37x10^{-4}</td>
</tr>
<tr>
<td>Rearranged cancer genes</td>
<td>36</td>
<td>129</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

We also investigated whether the genes that undergo rearrangement in cancer (and those that are also known cancer genes) show structural variation in the population time-scale. We first collected the list of the genes that are rearranged in breast cancer from Stephens et al. (18) and those that show CNVs in the HuRef genome (4). We counted the number of genes that underwent structural variation in both the datasets and calculated the statistical significance of the overlap using hyper-geometric distribution. We found that that there is a very weak overlap (p value: 1.37x10^{-1}). We repeated our analysis using only the subset of the genes in the dataset of Stephens et al (18), which are known to be implicated in cancer as per the Cancer Gene Census (www.sanger.ac.uk/genetics/CGP/Census/). We found no known cancer gene that overlap between the two datasets (ST10), which is not surprising given that HuRef genome represents an apparently healthy individual.

Table ST10: Overlap between structural variations in cancer and HuRef personal genomes.
Supplementary Module 9: Materials and methods

CGN score for the genes that have one-to-one orthologous genes between human and chimpanzee was obtained from De et al. (15) In brief, the authors obtained chromosomal location of the genes for human (Homo sapiens: NCBI 36), chimpanzee (Pan troglodytes: PanTro 2.1), and macaque (Macaca mulatta: MMUL1.0) genomes from Ensembl Release 48. Orthologs of human genes were obtained from Ensembl-Compara Release 48, which identifies orthologs by building phylogenetic trees of homologs using a maximum likelihood approach. The CGN score metric used in this study provides an estimate of how conserved the genomic neighbourhood of the gene of interest is. A value close to 1 indicates that the genomic neighbourhood within a window of 2Mb is mostly conserved with very little insertion or deletion of genetic material, while a value close to 0 suggests that the neighbourhood has been extensively altered due to insertion or deletion of genetic material. Synonymous (dS) and nonsynonymous (dN) divergence rates for the human-chimpanzee orthologous gene-pairs were obtained from Ensembl Release 48. 13,722 genes had CGN score, dN and dS available. Of them 12,950 genes had the dN/dS ratio calculated (dS>0). We obtained KS, KA and KI, the synonymous substitution rate, nonsynonymous substitution rate and the divergence rate of all interspersed repeats in the intergenic/intronic regions within 250 kb of the mid-point of each gene for a subset of these genes (9,351 genes) from Khaitovich et al. (1). Genes with different CGN scores were binned and the distribution of the values of the different parameters was compared using the Mann Whitney test.

Data on structurally altered regions and SNPs, compared to the reference human genome (Homo sapiens: NCBI36), for the Venter (HuRef)(4), Yoruban (NA18507)(5) and the Korean Reference (KorRef)(6) genome sequences were obtained from the respective publications. SNP density was calculated as the number of single nucleotide substitutions in a window (of 1 Kb, 2 Kb, upto 1 Mb) around the structural variation. The relationship between InDel density (number of InDels per Mb) and the SNP density (number of SNPs per Mb) was estimated by calculating the Spearman’s rank correlation coefficient. All statistical analyses were done using the R statistical package.

The cancer genome resequencing data was obtained from Pleasance et al. (10). The data on reads mapping to fusion or translocation in different cancer cell-types, and the information on sequences of distinct individual cells from cell populations that were genetically altered using Zinc-finger nucleases were obtained from relevant publications (see Supplementary Module 7).
References