Genome analysis

Negative correlation between compositional symmetries and local recombination rates

Liang Chen and Hongyu Zhao

1Department of Molecular, Cellular and Developmental Biology, 2Department of Epidemiology and Public Health and 3Department of Genetics, Yale University, New Haven, CT, USA

Received on June 6, 2005; revised on August 25, 2005; accepted on August 26, 2005
Advance Access publication August 30, 2005

ABSTRACT

Although still not much understood, the universal reverse complement symmetry in genomes may contain much information about the genome. In this article, under the hypothesis that recombination rate variations may be related to the high order DNA structure, we studied the association between local recombination rates and local symmetry levels in mouse, rat and human. We found significant negative correlations between recombination rates and reverse complement compositional symmetries in these three organisms. This negative correlation pattern also held at individual chromosome levels when data only from each individual chromosome was analyzed.

Contact: hongyu.zhao@yale.edu

INTRODUCTION

Chargaff’s first parity rule states that the frequency of A is equal to that of T and the frequency of C is equal to that of G in double-stranded DNA (Magasanik and Chargaff, 1951). Watson and Crick’s DNA helix model explained the first parity rule (Watson and Crick, 1953). Chargaff and colleagues also observed that for single-stranded DNA, the equalities are validated approximately (Rudner et al., 1968). That is, when only considering one strand of the double-stranded DNA, the frequency of A is equal to that of T and the frequency of C is equal to that of G. This intra-strand parity rule about a single nucleotide can be extended to longer oligonucleotides (Prabhu, 1993; Qi and Cuticchia, 2001). For example, under this parity rule, for single-stranded DNA, at order 2 (thus length 2), the frequency of GA is equal to that of TC (TC is the reverse complement of GA) and the frequency of CT is equal to that of AG (AG is the reverse complement of CT). Therefore, there is reverse complement symmetry for single-stranded DNA. Baisnee et al. (2002) conducted a comprehensive study of this single strand reverse complement symmetry. They measured the symmetry at orders 1–9 for a wide range of genomes including viruses, bacteria, archaea, mitochondria and eukaryota and demonstrated that the higher-order symmetry does not entirely result from the lower-order symmetry (Baisnee et al., 2002). The reason for this single strand reverse complement symmetry is still not well understood.

Forsdyke (1995) hypothesized that this symmetry results from the DNA stem–loop secondary structures. The single strand of the supercoiled duplex DNA may form stem–loop structures, which may facilitate the initiation of homologous recombination by way of ‘kissing’ between the tips of stem–loop structures. The recombination evolutionary advantage causes the selection of single strand reverse complement symmetry (Forsdyke, 1995). Baisnee et al. argued that the reverse complement symmetry does not result from point mutation or recombination, but from a combination effect of different mechanisms at different orders (Baisnee et al., 2002). Above all, the reverse complement symmetry may contain multiple levels of information about genome.

It is widely known that recombination rates vary along chromosomes with widespread recombination hotspots and coldspots (reviewed in Lichten and Goldman, 1995; Petes, 2001; Nachman, 2002). However, the reason for recombination rate variations is little known. Scientists have found that the cross-over hot-spot instigator (Chi) sequences locally increase recombination in Escherichia coli (Smith, 1988). But, for most of the recombination hotspots, no specific sequence motifs can be found. In yeast, double-strand DNA breaks (DSBs) initiate most, if not all, meiotic recombination. And these DSB sites usually are in deoxyribonuclease I sensitive regions (Wu and Lichten, 1994). All these suggest that DNA structure and accessibility may have an important role in recombination variation. GC content has been reported to be positively correlated with local recombination rates in the human genome (Fullerton et al., 2001). Other sequence features such as poly(A)/poly(T) fraction and CpG fraction also have significant correlations with recombination rates (Kong et al., 2002).

In this paper, we studied the association between local recombination rates and genome compositional reverse complement symmetry using publicly available genome-wide recombination rate data and genomic sequence data in Mus musculus (mouse), Rattus norvegicus (rat) and Homo sapiens (human). We found that local recombination rates are negatively correlated with compositional symmetries.

METHODS

Sequence data

We downloaded the genomic sequences of Mus musculus, Rattus norvegicus and Homo sapiens from the NCBI [ftp://ftp.ncbi.nih.gov/genomes/, April (May for mouse), 2005].

Measure of reverse complement symmetry

In this paper, we adopted a symmetry measure defined by $S_N = 1 - \frac{1}{N} \sum_{i=1}^{N} |f_i - f'_i| / (\sum_{i=1}^{N} |f_i| + |f'_i|)$ (Baisnee et al., 2002), where $f_i$ is the
frequency of the i-th N-mer oligonucleotide in a genomic region and $f'_i$ is the frequency of its reverse complement in the same region. We allow overlapping sequences in deriving these counts. SN is computed over the complete set of N-mers, and its value ranges from 0 (total asymmetry) to 1 (perfect symmetry). Baisnee and colleagues stated that this measure has some advantages over Pearson's correlation coefficient, e.g. Pearson's correlation coefficient is sensitive to outliers (Baisnee et al., 2002).

Other sequence features

In each non-overlapping genomic sequence window, after correcting for the number of ‘N’ s in the genome sequence, the fraction of G or C is the GC content. The fraction of CpG dinucleotides is the CpG fraction. The fraction of poly An or Tn where $n > 4$ is the poly(A)/poly(T) [(A)$_{n>4}$ and (T)$_{n>4}$] tract fraction.

Local recombination rates

The genome-wide recombination rates for Mus musculus, Rattus norvegicus and Homo sapiens were based on the paper of Jensen-Seaman et al., 2004. In that paper, the authors estimated the recombination rates using mouse OBxCAST F2 intercross genetic map (Dietrich et al., 1996) including 4880 markers, rat SHRSPxBN F2 intercross genetic map (Steen et al., 1999) including 2305 markers, and human Iceland pedigree map (Kong et al., 2002) including 5114 markers. Assuming a linear genetic distance across the immediately flanking genetic markers, they assigned each base a recombination rate. The average recombination rate of the bases within each non-overlapping 5 Mb window along the whole genome was shown in the paper’s Supplementary files. For human, they also estimated the sex-specific recombination rates which can be downloaded from the UCSC Genome Bioinformatics database, Table Browser (http://genome.ucsc.edu/). These sex-specific recombination rates correspond to non-overlapping 1 Mb windows. Except the sex difference study, the windows used are 5 Mb in the following. In our analyses, we calculated the reverse complement symmetry measures in these 5 Mb or 1 Mb windows. We also calculated CpG fraction, GC content fraction, and poly(A)/poly(T) [(A)$_{n>4}$ and (T)$_{n>4}$] tract fraction. Windows with >20% N bases and windows without estimated recombination rates were excluded from this study. For the 5 Mb non-overlapping windows, there were totally 426 windows for mouse, 467 windows for rat and 543 windows for human. For the 1 Mb windows for human, we only considered the autosomal chromosomes. There were totally 2563 windows for female-specific recombination rates and 2439 windows for male-specific recombination rates.

We used perl language to calculate the reverse complement symmetry measures, CpG fraction, GC content fraction and poly(A)/poly(T) [(A)$_{n>4}$ and (T)$_{n>4}$] tract fraction. R language was used in all the statistical analyses.

RESULTS

Reverse complement symmetry in mouse, rat and human

In Figure 1, we summarize the symmetry measures for oligonucleotide length 1–12 in mouse, rat and human. Symmetry measures correspond to non-overlapping 5 Mb windows along the genome (windows with >20% N bases and windows without estimated average recombination rates were excluded, there were a total of 426 windows for mouse, 467 windows for rat and 543 windows for human). A symmetry measure of 0 corresponds to total asymmetry and of 1 corresponds to perfect symmetry.
see that the symmetry measure is high for short oligonucleotides and drops slightly for long oligonucleotides. The variance of the symmetry measure for short oligonucleotides is very tiny ($\frac{1}{C_4} \times 10^{-5}$). In order to capture the potential relationship between local recombination rates and symmetry levels, we focused on oligonucleotide length 12 which has the highest variances for these three organisms ($1.04 \times 10^{-3}$ for mouse, $1.25 \times 10^{-3}$ for rat and $5.29 \times 10^{-4}$ for human) among the orders examined and reasonable symmetry levels (mean 0.41 for mouse, 0.40 for rat and 0.44 for human) in our study.

Negative correlation between recombination rates and symmetry measures

In Figure 2, the correlations between recombination rates and symmetry measures are plotted for oligonucleotide lengths 1–12. For short lengths, there was almost no correlation, which is not surprising because of the small variance of the calculated symmetry measures across different regions. For longer lengths, there were negative correlations for all three organisms.

The scatter plot between recombination rates and symmetry measures at order 12 for mouse is shown in Figure 3. The left panel shows all the data points. The regression coefficient for symmetry measure was $-3.67$ with $P$-value $8.05 \times 10^{-5}$. The two circled points may be considered possible outliers. After removing these two points, the regression coefficient was $-4.76$ with $P$-value $7.24 \times 10^{-10}$ (scatter plot in the right panel). For these two possible outlying points, the symmetry measure was relatively too high (0.77) or too low (0.27). The region with the high symmetry measure was on chromosome X, a 5 Mb region from 25 to 30 Mb. The number of Ns in this region was 770,513, which was high. The presence of these many Ns may affect the symmetry measure calculation. The region with the low symmetry measure was also on chromosome X, from 140 to 145 Mb. The number of Ns in this region was only 50,000. Because these special symmetry measures represent true biological observations, we kept these two points in the following study.

Figure 4 shows the scatter plot between recombination rates and symmetry measures at order 12 for rat. The regression coefficient for symmetry measure was $-3.37$ with $P$-value $3.54 \times 10^{-10}$.

Figure 5 is the scatter plot between recombination rates and symmetry measures at order 12 for human. The left panel is for all the data points. The regression coefficient for symmetry measure was $-11.85$ with $P$-value $<2 \times 10^{-16}$. After removing the circled point, the regression coefficient became $-14.01$ with $P$-value $<2 \times 10^{-16}$. For the circled point, the symmetry measure was high (0.68), and the corresponding region lay on chromosome 9 covering from
40 to 45 Mb. The number of Ns in this region was 659729, which was also high. We kept this point in the following study.

For mouse, Pearson’s correlation coefficient between recombination rates and symmetry measures at order 12 was $-0.27$ ($-0.21$ at order 10 and $-0.28$ at order 11). For rat, the correlation between recombination rates and symmetry measures at order 12 was $-0.29$ ($-0.15$ at order 10 and $-0.24$ at order 11). For human, the correlation between recombination rates and symmetry measures at order 12 was $-0.39$ ($-0.40$ at order 10 and $-0.47$ at order 11). These correlations are summarized in Table 1. We also list the correlations between recombination rates and three other sequence features: poly(A)/poly(T) $[(A)_{n>4}$ and $T_{n>4}]$ fraction, CpG fraction and GC content. poly(A)/poly(T) $[(A)_{n>4}$ and $T_{n>4}]$ fraction had a negative correlation with the recombination rate, whereas CpG fraction and GC content both had positive correlations with the recombination rate. In Table 2, we summarize the pairwise correlations among symmetry measure, poly(A)/poly(T) $[(A)_{n>4}$ and $T_{n>4}]$ fraction CpG fraction and GC content. For these three organisms, poly(A)/poly(T) $[(A)_{n>4}$ and $T_{n>4}]$ fraction, CpG fraction and GC content were highly correlated (absolute value of correlation $\geq 0.88$). It suggests that poly(A)/poly(T) $[(A)_{n>4}$ and $T_{n>4}]$ fraction, CpG fraction and GC content may capture similar information in genomic sequences. However, symmetry measure was much less correlated with these three DNA features. The absolute correlations between symmetry measure and three other DNA features were about 0.6 for mouse and 0.7 for rat. For human, the correlation between symmetry measure and CpG fraction was only $-0.08$, $-0.16$ for GC content, and 0.29 for poly(A)/poly(T) $[(A)_{n>4}$ and $T_{n>4}]$ fraction. Symmetry measure always had a negative correlation with GC content and CpG fraction and a positive correlation with poly(A)/poly(T) $[(A)_{n>4}$ and $T_{n>4}]$ fraction.

Multiple regressions were carried out between local recombination rate and symmetry measure, poly(A)/poly(T) $[(A)_{n>4}$ and $T_{n>4}]$ fraction, CpG fraction and GC content, and the results are summarized in Table 3. In order to capture potential interactions among sequence features, we performed backward stepwise regression with the Akaike information criterion (AIC) for model selection. Because these recombination rates were estimated for contiguous non-overlapping windows, they were possibly autocorrelated. The Durbin–Watson test was performed to test possible autocorrelations among regression residuals. The autocorrelation was $-0.04$ with $P$-value 0.5 for mouse, 0.08 with $P$-value 0.05 for rat and 0.26 with $P$-value 0 for human. So, for rat and human, we also fitted the generalized linear model incorporating autocorrelated residuals. Therefore, the coefficients and $P$-values were re-calculated. The final models are shown in Table 3. By using these sequence features, we can explain about 20% of the variance of the local recombination rates for mouse, 19% for rat and 49% for human. Symmetry measure had a significant negative effect on recombination rates for mouse and human, while it was not significant for rat. This difference may be due to less accurate estimation of recombination rates in rat. The results also show that there were significant interactions between symmetry measure, CpG fraction and poly(A)/poly(T) $[(A)_{n>4}$ and $T_{n>4}]$ fraction in human and mouse. GC content had a positive correlation with recombination rates (0.38 for mouse, 0.26 for rat and 0.44 for human). But in the multiple regression models, GC content had a significant negative effect for the recombination rate. This phenomenon was also noted.
Compositional symmetry and recombination

in previous papers (Jensen-Seaman et al., 2004; Kong et al., 2002), where the authors found that GC content was negatively correlated with the recombination rate after considering the CpG fraction and poly(A)/poly(T) [(A)_n>4 and (T)_n>4] tract fraction.

**Table 1.** Pearson’s correlation coefficients between recombination rate and symmetry measure at order 12, poly(A)/poly(T) [(A)_n>4 and (T)_n>4] tract fraction, CpG fraction and GC content fraction for mouse, rat and human

<table>
<thead>
<tr>
<th></th>
<th>Recombination rate (Mouse)</th>
<th>Recombination rate (Rat)</th>
<th>Recombination rate (Human)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symmetry measure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>at order 12</td>
<td>−0.27^a</td>
<td>−0.29^a</td>
<td>−0.39^a</td>
</tr>
<tr>
<td>Poly(A)/poly(T)</td>
<td>−0.39^a</td>
<td>−0.26^a</td>
<td>−0.48^a</td>
</tr>
<tr>
<td>CpG fraction</td>
<td>0.42^a</td>
<td>0.37^a</td>
<td>0.47^a</td>
</tr>
<tr>
<td>GC content</td>
<td>0.38^a</td>
<td>0.26^a</td>
<td>0.44^a</td>
</tr>
</tbody>
</table>

^P-value < 10^{-7}.

**Table 2.** Pairwise Pearson’s correlation coefficients among symmetry measure at order 12, poly(A)/poly(T) [(A)_n>4 and (T)_n>4] tract fraction, CpG fraction and GC content fraction

<table>
<thead>
<tr>
<th></th>
<th>Poly(A)/poly(T) fraction</th>
<th>CpG fraction</th>
<th>GC content</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mouse</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symmetry measure</td>
<td>0.63^a</td>
<td>−0.62^a</td>
<td>−0.61^a</td>
</tr>
<tr>
<td>at order 12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poly(A)/poly(T)</td>
<td>−0.90^a</td>
<td>−0.98^a</td>
<td></td>
</tr>
<tr>
<td>CpG fraction</td>
<td>0.94^a</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rat</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symmetry measure</td>
<td>0.74^a</td>
<td>−0.74^a</td>
<td>−0.71^a</td>
</tr>
<tr>
<td>at order 12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poly(A)/poly(T)</td>
<td>−0.88^a</td>
<td>−0.98^a</td>
<td></td>
</tr>
<tr>
<td>CpG fraction</td>
<td>0.92^a</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Human</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symmetry measure</td>
<td>0.29^a</td>
<td>−0.08</td>
<td>−0.16 (P-value = 0.06)</td>
</tr>
<tr>
<td>at order 12</td>
<td></td>
<td></td>
<td>(P-value = 0.0001)</td>
</tr>
<tr>
<td>Poly(A)/poly(T)</td>
<td>−0.88^a</td>
<td>−0.97^a</td>
<td></td>
</tr>
<tr>
<td>CpG fraction</td>
<td>0.95^a</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

^P-value < 10^{-10}.

**Table 3.** Multiple regression results between recombination rate and symmetry measure at order 12 (sym), poly(A)/poly(T) [(A)_n>4 and (T)_n>4] tract fraction (pApT), CpG fraction (CpG) and GC content fraction (GC)

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>SE</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mouse</strong> (adjusted R^2 = 0.20)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>150.57</td>
<td>43.07</td>
<td>0.00052</td>
</tr>
<tr>
<td>GC</td>
<td>966.74</td>
<td>465.26</td>
<td>0.038</td>
</tr>
<tr>
<td>pApT</td>
<td>−1229.79</td>
<td>414.69</td>
<td>0.0032</td>
</tr>
<tr>
<td>sym</td>
<td>−355.7</td>
<td>104.63</td>
<td>0.00074</td>
</tr>
<tr>
<td>GC:CpG</td>
<td>2504.67</td>
<td>1003.76</td>
<td>0.013</td>
</tr>
<tr>
<td>GC:sym</td>
<td>834.25</td>
<td>230.58</td>
<td>0.00033</td>
</tr>
<tr>
<td>pApT:sym</td>
<td>−2080.36</td>
<td>1162.69</td>
<td>0.074</td>
</tr>
<tr>
<td>GC:pApT:sym</td>
<td>3035.54</td>
<td>1050.73</td>
<td>0.0027</td>
</tr>
<tr>
<td><strong>Rat</strong> (adjusted R^2 = 0.19)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>17.40</td>
<td>7.27</td>
<td>0.017</td>
</tr>
<tr>
<td>GC</td>
<td>−26.27</td>
<td>20.02</td>
<td>0.19</td>
</tr>
<tr>
<td>pApT</td>
<td>−235.89</td>
<td>68.96</td>
<td>0.0007</td>
</tr>
<tr>
<td>sym</td>
<td>34.66</td>
<td>20.33</td>
<td>0.089</td>
</tr>
<tr>
<td>GC:CpG</td>
<td>1765.13</td>
<td>554.12</td>
<td>0.0015</td>
</tr>
<tr>
<td>GC:sym</td>
<td>464.94</td>
<td>155.80</td>
<td>0.003</td>
</tr>
<tr>
<td>pApT:sym</td>
<td>−108.86</td>
<td>60.82</td>
<td>0.074</td>
</tr>
<tr>
<td>GC:pApT:sym</td>
<td>1237.88</td>
<td>657.24</td>
<td>0.060</td>
</tr>
<tr>
<td><strong>Human</strong> (adjusted R^2 = 0.49)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>48.1</td>
<td>9.99</td>
<td>0.0000</td>
</tr>
<tr>
<td>GC</td>
<td>−33.72</td>
<td>11.92</td>
<td>0.0048</td>
</tr>
<tr>
<td>Cpg</td>
<td>694.17</td>
<td>650.67</td>
<td>0.29</td>
</tr>
<tr>
<td>pApT</td>
<td>−402.68</td>
<td>115.52</td>
<td>0.0005</td>
</tr>
<tr>
<td>sym</td>
<td>−86.87</td>
<td>21.44</td>
<td>0.0001</td>
</tr>
<tr>
<td>GC:CpG</td>
<td>−4406.10</td>
<td>991.74</td>
<td>0.0000</td>
</tr>
<tr>
<td>GC:sym</td>
<td>223.52</td>
<td>165.17</td>
<td>0.18</td>
</tr>
<tr>
<td>pApT:sym</td>
<td>8424.01</td>
<td>7586.72</td>
<td>0.27</td>
</tr>
<tr>
<td>GC:CpG:sym</td>
<td>5896.22</td>
<td>1289.09</td>
<td>0.0000</td>
</tr>
<tr>
<td>pApT:sym</td>
<td>750.63</td>
<td>229.81</td>
<td>0.0012</td>
</tr>
<tr>
<td>GC:pApT:sym</td>
<td>−49144.98</td>
<td>17908.79</td>
<td>0.0063</td>
</tr>
</tbody>
</table>

**Fig. 5.** The scatter plot of recombination rates versus symmetry measures at order 12 for human. In the left panel, all the data are plotted. The circled point is a possible outlier. After removing this point, the scatter plot is shown in the right panel. For the linear regression line, the regression coefficient for symmetry measure is −11.85 with P-value < 2 × 10^{-16} in the left panel, and −14.01 with P-value < 2 × 10^{-16} in the right panel.

**Table 4.** Pearson’s correlation coefficients between recombination rate and symmetry measure at order 12 (sym), poly(A)/poly(T) [(A)_n>4 and (T)_n>4] tract fraction (pApT), CpG fraction (CpG) and GC content fraction (GC)
Table 4. Pearson’s correlation coefficients between recombination rates and symmetry measures for individual chromosomes in mouse, rat, and human

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Pearson correlation (Mouse)</th>
<th>P-value (Mouse)</th>
<th>Number of regions (Mouse)</th>
<th>Pearson correlation (Rat)</th>
<th>P-value (Rat)</th>
<th>Number of regions (Rat)</th>
<th>Pearson correlation (Human)</th>
<th>P-value (Human)</th>
<th>Number of regions (Human)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.446</td>
<td>0.0046</td>
<td>33</td>
<td>-0.0622</td>
<td>0.33</td>
<td>51</td>
<td>-0.339</td>
<td>0.012</td>
<td>44</td>
</tr>
<tr>
<td>2</td>
<td>0.0642</td>
<td>0.64</td>
<td>32</td>
<td>-0.364</td>
<td>0.0064</td>
<td>46</td>
<td>-0.698</td>
<td>2.50 x 10^-6a</td>
<td>47</td>
</tr>
<tr>
<td>3</td>
<td>-0.185</td>
<td>0.18</td>
<td>26</td>
<td>-0.173</td>
<td>0.17</td>
<td>32</td>
<td>-0.634</td>
<td>9.69 x 10^-6a</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>-0.302</td>
<td>0.071</td>
<td>25</td>
<td>-0.215</td>
<td>0.11</td>
<td>34</td>
<td>-0.701</td>
<td>6.66 x 10^-6a</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>-0.348</td>
<td>0.0444</td>
<td>25</td>
<td>-0.164</td>
<td>0.18</td>
<td>34</td>
<td>-0.564</td>
<td>0.00021</td>
<td>35</td>
</tr>
<tr>
<td>6</td>
<td>-0.517</td>
<td>0.0041</td>
<td>25</td>
<td>-0.236</td>
<td>0.14</td>
<td>23</td>
<td>-0.702</td>
<td>3.80 x 10^-6a</td>
<td>32</td>
</tr>
<tr>
<td>7</td>
<td>-0.319</td>
<td>0.079</td>
<td>21</td>
<td>-0.372</td>
<td>0.030</td>
<td>26</td>
<td>-0.404</td>
<td>0.014</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>-0.258</td>
<td>0.12</td>
<td>23</td>
<td>0.0827</td>
<td>0.65</td>
<td>23</td>
<td>-0.551</td>
<td>0.0014</td>
<td>27</td>
</tr>
<tr>
<td>9</td>
<td>-0.491</td>
<td>0.014</td>
<td>20</td>
<td>0.13</td>
<td>0.20</td>
<td>0.309</td>
<td>0.081</td>
<td>0.08</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>-0.343</td>
<td>0.059</td>
<td>22</td>
<td>0.286</td>
<td>0.11</td>
<td>20</td>
<td>0.572</td>
<td>0.0011</td>
<td>26</td>
</tr>
<tr>
<td>11</td>
<td>-0.225</td>
<td>0.15</td>
<td>23</td>
<td>0.0985</td>
<td>0.63</td>
<td>13</td>
<td>-0.620</td>
<td>0.00049</td>
<td>25</td>
</tr>
<tr>
<td>12</td>
<td>-0.837</td>
<td>2.08 x 10^-6a</td>
<td>20</td>
<td>-0.850</td>
<td>0.0019</td>
<td>9</td>
<td>-0.847</td>
<td>4.73 x 10^-6a</td>
<td>25</td>
</tr>
<tr>
<td>13</td>
<td>-0.446</td>
<td>0.028</td>
<td>19</td>
<td>-0.494</td>
<td>0.0269</td>
<td>16</td>
<td>-0.812</td>
<td>2.15 x 10^-6a</td>
<td>18</td>
</tr>
<tr>
<td>14</td>
<td>-0.192</td>
<td>0.22</td>
<td>18</td>
<td>0.0925</td>
<td>0.64</td>
<td>18</td>
<td>-0.571</td>
<td>0.0084</td>
<td>17</td>
</tr>
<tr>
<td>15</td>
<td>-0.580</td>
<td>0.0092</td>
<td>16</td>
<td>-0.224</td>
<td>0.18</td>
<td>19</td>
<td>-0.481</td>
<td>0.030</td>
<td>16</td>
</tr>
<tr>
<td>16</td>
<td>-0.545</td>
<td>0.015</td>
<td>16</td>
<td>-0.328</td>
<td>0.28</td>
<td>15</td>
<td>-0.488</td>
<td>0.032</td>
<td>15</td>
</tr>
<tr>
<td>17</td>
<td>0.006</td>
<td>0.51</td>
<td>13</td>
<td>-0.404</td>
<td>0.054</td>
<td>17</td>
<td>-0.300</td>
<td>0.15</td>
<td>14</td>
</tr>
<tr>
<td>18</td>
<td>-0.388</td>
<td>0.069</td>
<td>16</td>
<td>-0.627</td>
<td>0.0047</td>
<td>16</td>
<td>-0.334</td>
<td>0.12</td>
<td>14</td>
</tr>
<tr>
<td>19</td>
<td>-0.315</td>
<td>0.173</td>
<td>11</td>
<td>-0.842</td>
<td>0.0011</td>
<td>10</td>
<td>-0.0774</td>
<td>0.42</td>
<td>10</td>
</tr>
<tr>
<td>20</td>
<td>-0.238</td>
<td>0.27</td>
<td>9</td>
<td>-0.183</td>
<td>0.30</td>
<td>11</td>
<td>-0.183</td>
<td>0.30</td>
<td>11</td>
</tr>
<tr>
<td>21</td>
<td>0.683</td>
<td>0.067</td>
<td>6</td>
<td>0.0607</td>
<td>0.55</td>
<td>6</td>
<td>0.683</td>
<td>0.067</td>
<td>6</td>
</tr>
<tr>
<td>22</td>
<td>0.0607</td>
<td>0.55</td>
<td>6</td>
<td>0.683</td>
<td>0.067</td>
<td>6</td>
<td>0.0607</td>
<td>0.55</td>
<td>6</td>
</tr>
</tbody>
</table>

The order of symmetry measure was 12. Each region was defined by a non-overlapping 5 Mb window. Windows with >20% N bases and windows without estimated average recombination rates were excluded.

* indicates that the one-sided P-value is <0.05.

Negative correlation exists at chromosome level

In order to know whether the negative association also holds at individual chromosome level, we calculated Pearson’s correlation coefficient between recombination rates and symmetry measures for each individual chromosome. The results are summarized in Table 4. For mouse, 8 out of 20 chromosomes had significant negative correlations with one-sided P-value <0.05. For rat, 8 out of 21 chromosomes had significant negative correlations. For human, 16 out of 23 chromosomes had significant negative correlations. For most of the chromosomes, the correlations were negative and no significant positive correlation was found.

In Figure 6, we plot 1-symmetry measure at order 12 (upper panel) and recombination rate (lower panel) along each individual chromosome in the human genome. Note that if there is a negative correlation between the recombination rate and symmetry measure, the correlation is positive between the recombination rate and (1—symmetry measure). We plot 1—symmetry measure instead of symmetry measure instead of symmetry measure because we have only one window, YM,i, window, YF,i window,

YM,i = sM,i + SM,i

YF,i = sM,i + SM,i

where YM,i is the female-specific recombination rate for the ith window, YF,i is the male-specific recombination rate for the ith window, sM,i is the symmetry measure at order 12 for the corresponding female’s ith window, SM,i is the symmetry measure at order 12 for the corresponding male’s ith window, β1 is the common symmetry effect, and β2 is the sex-specific symmetry effect and ε is the error term. Here n = 2563 and m = 2439. Because some windows did not have estimated female recombination rates and levels. We only considered autosomals, and the correlation between female-specific recombination rate and symmetry measure at order 12 was −0.12 (−0.24 at order 10 and −0.21 at order 11). The correlation between male-specific recombination rate and symmetry measure at order 12 was −0.14 (−0.32 at order 10 and −0.24 at order 11). To test the statistical significance of the observed sex difference, we used the following regression model to jointly consider sex-specific recombination rates and symmetry measure. The model is:

\[
\begin{pmatrix}
Y_{F,1} \\
Y_{F,2} \\
\vdots \\
Y_{F,n}
\end{pmatrix} =
\begin{pmatrix}
S_{F,1} \\
S_{F,2} \\
\vdots \\
S_{F,n}
\end{pmatrix}
\begin{pmatrix}
\beta_1 \\
\beta_2
\end{pmatrix} +
\begin{pmatrix}
e_1 \\
e_2 \\
\vdots \\
e_{(n+m)}
\end{pmatrix},
\]

where Y_{F,i} is the female-specific recombination rate for the ith window, Y_{M,i} is the male-specific recombination rate for the ith window, S_{F,i} is the symmetry measure at order 12 for the corresponding female’s ith window, S_{M,i} is the symmetry measure at order 12 for the corresponding male’s ith window, β1 is the common symmetry effect, and β2 is the sex-specific symmetry effect and ε is the error term. Here n = 2563 and m = 2439. Because some windows did not have estimated female recombination rates and
some windows did not have estimated male recombination rates, n was not equal to m. From the results, the sex-specific symmetry effect $b_d$ was estimated to be $-2.78$ with $P$-value $<2 \times 10^{-16}$. The common symmetry effect $b_s$ was $-2.59$ with $P$-value $1.13 \times 10^{-9}$. Therefore, symmetry measure had a negative effect on both sex-specific recombination rates. Compared to females, symmetry measure had an additional negative effect on male-specific recombination rates.

**DISCUSSION**

In this article, we have explored the negative correlations between local recombination rates and local symmetry levels. The negative correlation was significant for the three organisms studied using estimated local recombination rates. This negative correlation was not only observed at the genome level but also at the chromosome level. The results for rat were relatively less significant, which may be due to less reliable measured recombination rate estimates. For human, we note that the negative correlation was significantly stronger for males than females. Although there appeared to be some heterogeneity of variances in the regression analyses, this may not lead to a change of our conclusions due to the extreme $P$-values from these analyses. Here, we studied mouse, rat and human at the 5 Mb resolution. If we can get more accurate genetic maps, we may explore the association more correctly. As to the genome sequence windows, even after we removed the windows with >20% of N bases, there was still 8.6% Ns for the remaining windows for rat, 2.1% for mouse and 0.6% for human. The missing sequence information may affect the symmetry measure calculation. It may be another possible reason for the less significant results for rat. Although, we measured the symmetry level at order 12 in this article, the results and conclusions were similar for order 10 and order 11.

The reverse complement symmetry in many organisms has been known for a long time. However, it has not drawn much attention from scientists. Currently, there is little explanation for this universal symmetry phenomenon. Baisnee et al. (2002) argued that the symmetry results from a combination effect of different mechanisms at different orders. Unfortunately, they did not quantify the relative contribution of these different mechanisms. Forsdyke suggested that because the stem-loop structure in supercoiled DNA facilitates the initiation of recombination, there is evolutionary pressure to produce reverse complement DNA sequences (Forsdyke, 1995). If the local stem-loop structure is the only force for the reverse complement symmetry, the higher local symmetry levels should result in higher recombination rates. On the contrary, our analysis shows that there is a negative instead of positive correlation between the local symmetry levels and the local recombination rates. We hypothesize that although the reverse symmetry can cause stem–loop structure, the presence of symmetry may keep the stability of the chromatin. So, the high symmetry level can inhibit the occurrence of recombination events.
ACKNOWLEDGEMENTS

This work was supported in part by NSF grant DMS 0241160 and NIH grant GM 59507. We thank the reviewers for their constructive comments.

Conflict of Interest: none declared.

REFERENCES


