

## Genome analysis

# Negative correlation between compositional symmetries and local recombination rates

Liang Chen<sup>1</sup> and Hongyu Zhao<sup>2,3,\*</sup><sup>1</sup>Department of Molecular, Cellular and Developmental Biology, <sup>2</sup>Department of Epidemiology and Public Health and <sup>3</sup>Department of Genetics, Yale University, New Haven, CT, USAReceived 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, archae, 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.

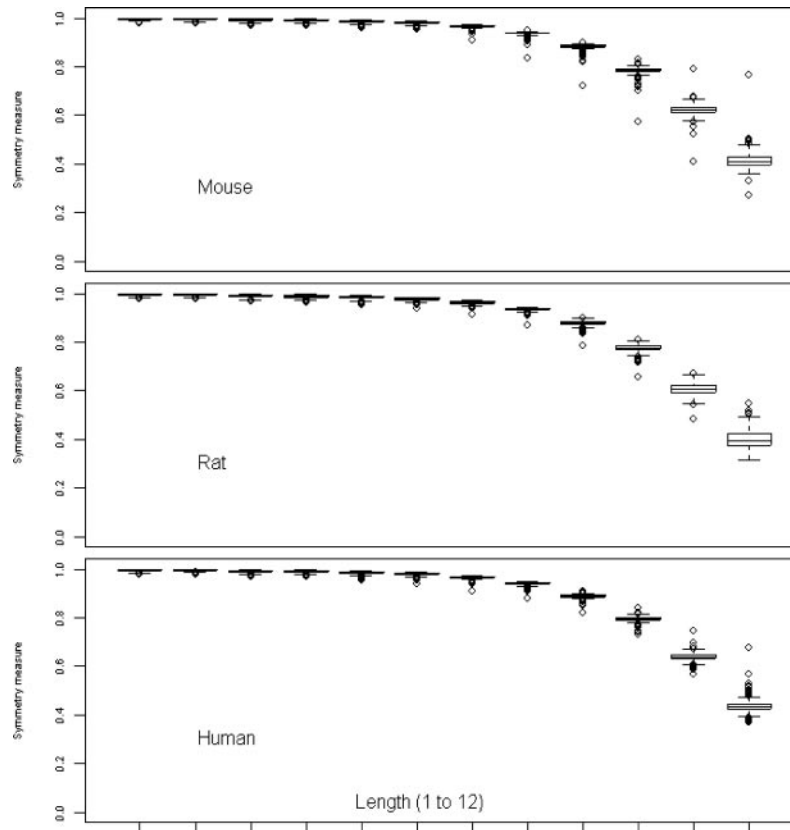
**METHODS****Sequence data**

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  $S_N = 1 - (\sum_i |f_i - f_i'|) / (\sum_i |f_i| + |f_i'|)$  (Baisnee *et al.*, 2002), where  $f_i$  is the

\*To whom correspondence should be addressed at 60 College Street, New Haven, CT 06520-8034, USA



**Fig. 1.** Box plots for the reverse complement 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.

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.  $S_N$  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 A<sub>n</sub> or T<sub>n</sub> where  $n \geq 4$  is the poly(A)/poly(T) [(A)<sub>n $\geq$ 4</sub> and (T)<sub>n $\geq$ 4</sub>] 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 for the sex difference study, the windows used are 5 Mb in the following.

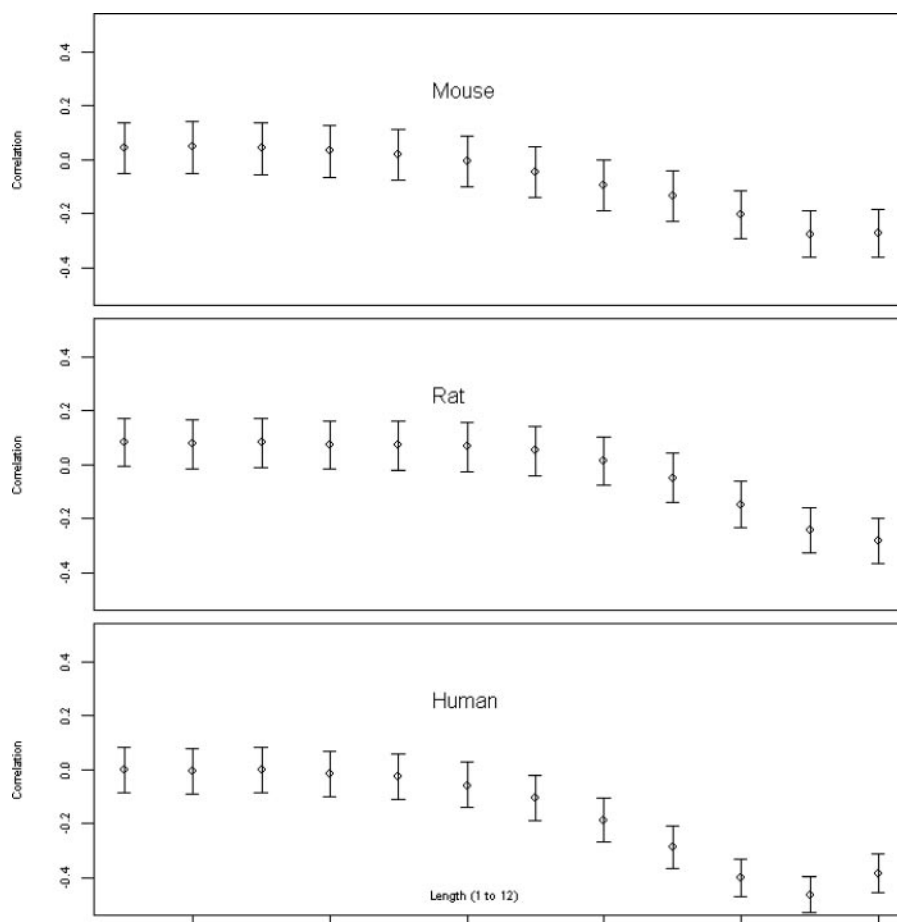
In our analyses, we calculated the reverse complement symmetry measures in these 5 Mb or 1 Mb windows. We also calculated CpG fraction, GC content fraction, and poly(A)/poly(T) [(A)<sub>n $\geq$ 4</sub> and (T)<sub>n $\geq$ 4</sub>] tract fraction. Windows with >20% N bases and windows without estimated 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)<sub>n $\geq$ 4</sub> and (T)<sub>n $\geq$ 4</sub>] 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 for non-overlapping 5 Mb windows. From these box plots of symmetry measures, we can



**Fig. 2.** Correlations between recombination rates and symmetry measures for oligonucleotide length from 1 to 12 for mouse, rat and human. The 95% confidence intervals are plotted together with point estimates.

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 ( $\sim 10^{-5}$ ). 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

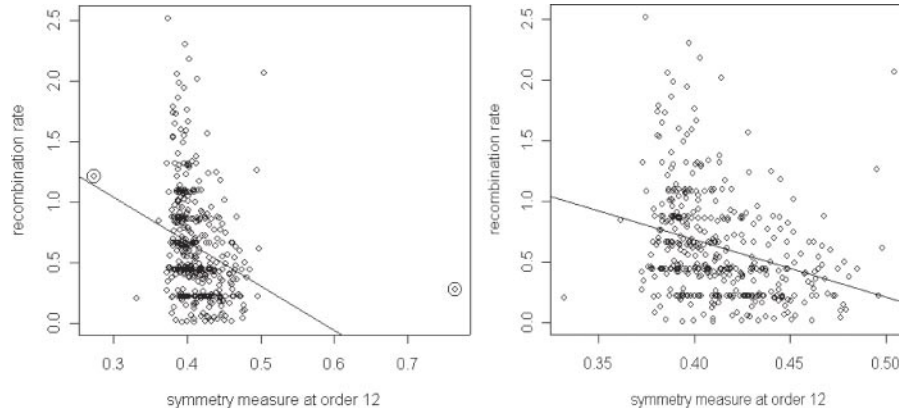
In Figure 2, the correlations between recombination rates and symmetry measures are plotted for oligonucleotide lengths 1–12. For short lengths, there was almost no correlation, which is not surprising because of the small variance of the calculated symmetry measures across different regions. For longer 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  $8.05 \times 10^{-9}$ . The two circled

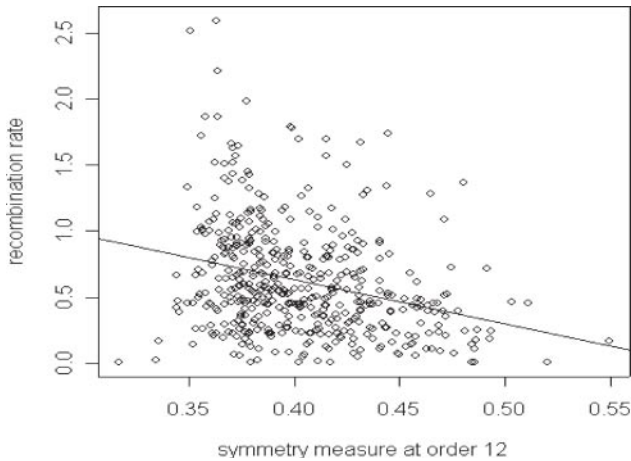
points may be considered possible outliers. After removing these two points, the regression coefficient was  $-4.76$  with  $P$ -value  $7.24 \times 10^{-10}$  (scatter plot in the right panel). For these two possible outlying points, the symmetry measure was relatively too high (0.77) or too low (0.27). 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 770 153, 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



**Fig. 3.** The scatter plot of recombination rates versus symmetry measures at order 12 for mouse. In the left panel, all the data are plotted. The two circled points are possible outliers. After removing these two points, the scatter plot is shown in the right panel. For the linear regression line, the regression coefficient for symmetry measure is  $-3.67$  with  $P$ -value  $8.05 \times 10^{-9}$  in the left panel,  $-4.76$  with  $P$ -value  $7.24 \times 10^{-10}$  in the right panel.



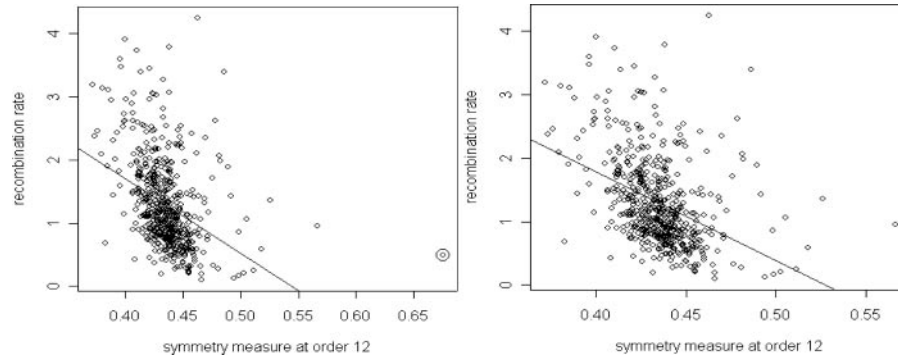
**Fig. 4.** The scatter plot of recombination rates versus symmetry measures at order 12 for rat. For the linear regression line, the regression coefficient for symmetry measure is  $-3.37$  with  $P$ -value  $3.54 \times 10^{-10}$ .

40 to 45 Mb. The number of Ns in this region was 659 729, 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)<sub>n≥4</sub> and (T)<sub>n≥4</sub>] fraction, CpG fraction and GC content. poly(A)/poly(T) [(A)<sub>n≥4</sub> and (T)<sub>n≥4</sub>] 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)<sub>n≥4</sub> and (T)<sub>n≥4</sub>] fraction CpG fraction and GC content. For these three organisms, poly(A)/poly(T) [(A)<sub>n≥4</sub> and (T)<sub>n≥4</sub>] fraction, CpG

fraction and GC content were highly correlated (absolute value of correlation  $\geq 0.88$ ). It suggests that poly(A)/poly(T) [(A)<sub>n≥4</sub> and (T)<sub>n≥4</sub>] 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)<sub>n≥4</sub> and (T)<sub>n≥4</sub>] fraction. Symmetry measure always had a negative correlation with GC content and CpG fraction and a positive correlation with poly(A)/poly(T) [(A)<sub>n≥4</sub> and (T)<sub>n≥4</sub>] fraction.

Multiple regressions were carried out between local recombination rate and symmetry measure, poly(A)/poly(T) [(A)<sub>n≥4</sub> and (T)<sub>n≥4</sub>] 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)<sub>n≥4</sub> and (T)<sub>n≥4</sub>] 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



**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  $< 2 \times 10^{-16}$  in the left panel, and  $-14.01$  with  $P$ -value  $< 2 \times 10^{-16}$  in the right panel.

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

	Recombination rate (Mouse)	Recombination rate (Rat)	Recombination rate (Human)
Symmetry measure at order 12	$-0.27^a$	$-0.29^a$	$-0.39^a$
Poly(A)/poly(T) fraction	$-0.39^a$	$-0.26^a$	$-0.48^a$
CpG fraction	$0.42^a$	$0.37^a$	$0.47^a$
GC content	$0.38^a$	$0.26^a$	$0.44^a$

<sup>a</sup> $P$ -value  $< 10^{-7}$ .

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

	Poly(A)/poly(T) fraction	CpG fraction	GC content
<b>Mouse</b>			
Symmetry measure at order 12	$0.63^a$	$-0.62^a$	$-0.61^a$
Poly(A)/poly(T) fraction		$-0.90^a$	$-0.98^a$
CpG fraction			$0.94^a$
<b>Rat</b>			
Symmetry measure at order 12	$0.74^a$	$-0.74^a$	$-0.71^a$
Poly(A)/poly(T) fraction		$-0.88^a$	$-0.98^a$
CpG fraction			$0.92^a$
<b>Human</b>			
Symmetry measure at order 12	$0.29^a$	$-0.08$	$-0.16$
Poly(A)/poly(T) fraction		$-0.88^a$	$-0.97^a$
CpG fraction			$0.95^a$

<sup>a</sup> $P$ -value  $< 10^{-10}$ .

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

	Coefficient	SE	$P$ -value
<b>Mouse (adjusted R<sup>2</sup> = 0.20)</b>			
(Intercept)	150.57	43.07	0.00052
GC	$-348.43$	94.06	0.00024
CpG	966.74	465.26	0.038
pApT	$-1229.79$	414.69	0.0032
sym	$-355.7$	104.63	0.00074
GC:pApT	2504.67	1003.76	0.013
GC:sym	834.25	230.58	0.00033
CpG:sym	$-2080.36$	1162.69	0.074
pApT:sym	3035.54	1005.73	0.0027
GC:pApT:sym	$-6402.96$	2457.18	0.0095
<b>Rat (adjusted R<sup>2</sup> = 0.19)</b>			
(Intercept)	17.40	7.27	0.017
GC	$-26.27$	20.02	0.19
CpG	$-1093.85$	419.29	0.0094
pApT	$-235.89$	68.96	0.0007
sym	34.66	20.33	0.089
GC:CpG	1765.13	554.12	0.0015
GC:pApT	464.94	155.80	0.003
GC:sym	$-108.86$	60.82	0.074
CpG:sym	1237.88	657.24	0.060
<b>Human (adjusted R<sup>2</sup> = 0.49)</b>			
(Intercept)	48.1	9.99	0.0000
GC	$-33.72$	11.92	0.0048
CpG	694.17	650.67	0.29
pApT	$-402.68$	115.52	0.0005
sym	$-86.87$	21.44	0.0001
GC:CpG	$-4406.10$	991.74	0.0000
GC:pApT	223.52	165.17	0.18
CpG:pApT	8424.01	7586.72	0.27
CpG:sym	5896.22	1289.09	0.0000
pApT:sym	750.63	229.81	0.0012
CpG:pApT:sym	$-49144.98$	17908.79	0.0063

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)<sub>n≥4</sub> and (T)<sub>n≥4</sub>] fraction.

**Table 4.** Pearson’s correlation coefficients between recombination rates and symmetry measures for individual chromosomes in mouse, rat, and human

Chromosome	Pearson correlation (Mouse)	P-value (Mouse)	Number of regions (Mouse)	Pearson correlation (Rat)	P-value (Rat)	Number of regions (Rat)	Pearson correlation (Human)	P-value (Human)	Number of regions (Human)
1	-0.446	0.0046 <sup>a</sup>	33	-0.0622	0.33	51	-0.339	0.012 <sup>a</sup>	44
2	0.0642	0.64	32	-0.364	0.0064 <sup>a</sup>	46	-0.698	2.50 × 10 <sup>-8a</sup>	47
3	-0.185	0.18	26	-0.173	0.17	32	-0.634	9.69 × 10 <sup>-6a</sup>	38
4	-0.302	0.071	25	-0.215	0.11	34	-0.701	6.66 × 10 <sup>-7a</sup>	37
5	-0.348	0.044 <sup>a</sup>	25	-0.164	0.18	34	-0.564	0.00021 <sup>a</sup>	35
6	-0.517	0.0041 <sup>a</sup>	25	-0.236	0.14	23	-0.702	3.80 × 10 <sup>-6a</sup>	32
7	-0.319	0.079	21	-0.372	0.030 <sup>a</sup>	26	-0.404	0.014 <sup>a</sup>	30
8	-0.258	0.12	23	0.0827	0.65	23	-0.551	0.0014 <sup>a</sup>	27
9	-0.491	0.014 <sup>a</sup>	20	-0.264	0.13	20	-0.309	0.081	22
10	-0.343	0.059	22	-0.286	0.11	20	-0.572	0.0011 <sup>a</sup>	26
11	-0.225	0.15	23	0.0985	0.63	13	-0.620	0.00048 <sup>a</sup>	25
12	-0.837	2.08 × 10 <sup>-6a</sup>	20	-0.850	0.0019 <sup>a</sup>	9	-0.847	4.73 × 10 <sup>-8a</sup>	25
13	-0.446	0.028 <sup>a</sup>	19	-0.494	0.026 <sup>a</sup>	16	-0.812	2.15 × 10 <sup>-5a</sup>	18
14	-0.192	0.22	18	0.0925	0.64	18	-0.571	0.0084 <sup>a</sup>	17
15	-0.580	0.0092 <sup>a</sup>	16	-0.224	0.18	19	-0.481	0.030 <sup>a</sup>	16
16	-0.545	0.015 <sup>a</sup>	16	-0.162	0.28	15	-0.488	0.032 <sup>a</sup>	15
17	0.006	0.51	13	-0.404	0.054 <sup>a</sup>	17	-0.300	0.15	14
18	-0.388	0.069	16	-0.627	0.0047 <sup>a</sup>	16	-0.334	0.12	14
19	-0.315	0.173	11	-0.842	0.0011 <sup>a</sup>	10	-0.0774	0.42	10
20				-0.238	0.27	9	-0.183	0.30	11
21							-0.683	0.067	6
22							0.0607	0.55	6
X	-0.233	0.148	22	-0.581	0.0091 <sup>a</sup>	16	-0.627	0.00018 <sup>a</sup>	28

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.

<sup>a</sup>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 for visual convenience. This figure clearly shows the variation of recombination rates and that of symmetry levels along chromosomes. Also the negative association between recombination rate and symmetry measure is apparent.

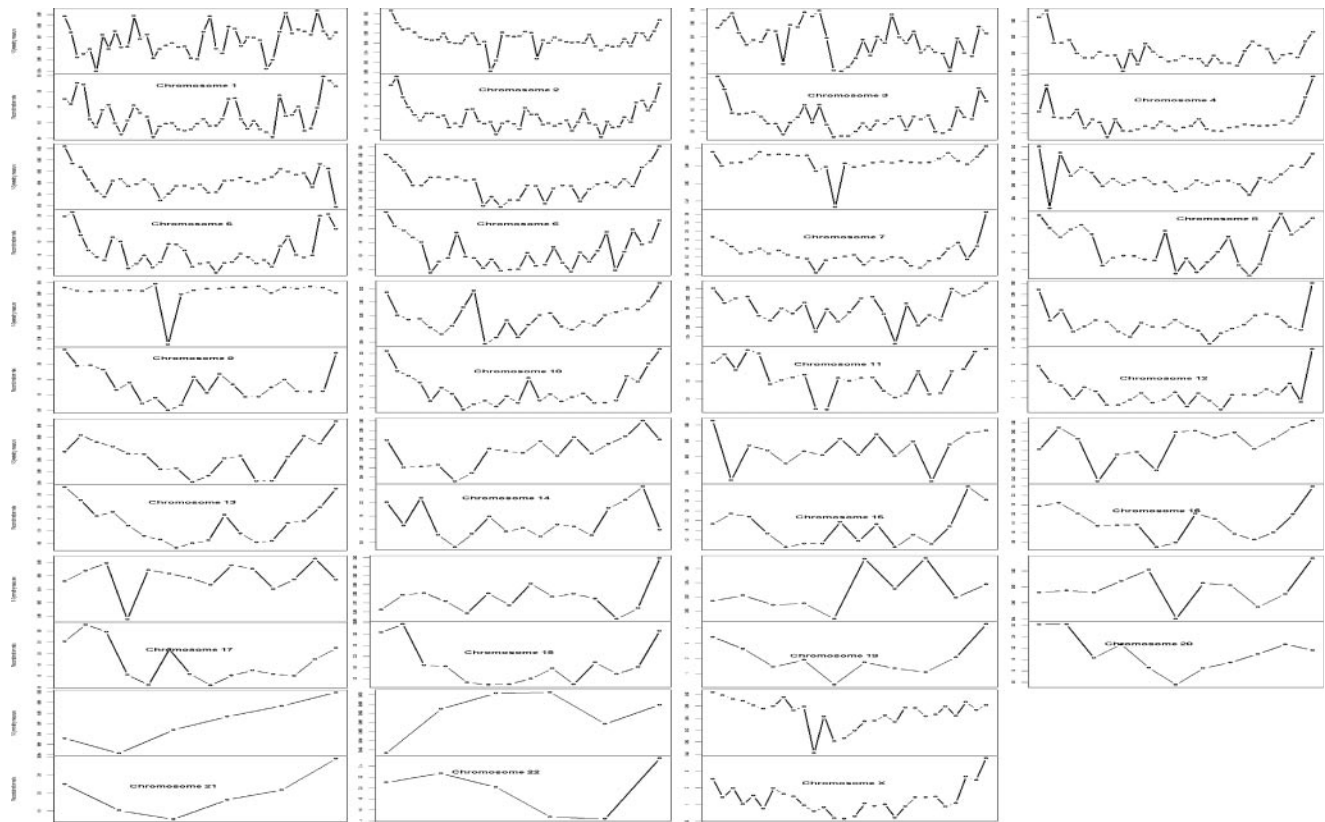
**Sex difference**

Since there are differences in recombination rates between males and females (Broman *et al.*, 1998), we also studied the negative correlations between sex-specific recombination rates and symmetry

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:

$$\begin{pmatrix} Y_{F,1} \\ Y_{F,2} \\ \dots \\ Y_{F,n} \\ Y_{M,1} \\ Y_{M,2} \\ \dots \\ Y_{M,m} \end{pmatrix} = \begin{pmatrix} S_{F,1} & 0 \\ S_{F,2} & 0 \\ \dots & \dots \\ S_{F,n} & 0 \\ S_{M,1} & S_{M,1} \\ S_{M,2} & S_{M,2} \\ \dots & \dots \\ S_{M,m} & S_{M,m} \end{pmatrix} \begin{pmatrix} \beta_s \\ \beta_d \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \varepsilon_{(n+m)} \end{pmatrix},$$

where  $Y_{F,i}$  is the female-specific recombination rate for the *i*th window,  $Y_{M,i}$  is the male-specific recombination rate for the *i*th window,  $S_{F,i}$  is the symmetry measure at order 12 for the corresponding female’s *i*th window,  $S_{M,i}$  is the symmetry measure at order 12 for the corresponding male’s *i*th window,  $\beta_s$  is the common symmetry effect, and  $\beta_d$  is the sex-specific symmetry effect and  $\varepsilon$  is the error term. Here  $n = 2563$  and  $m = 2439$ . Because some windows did not have estimated female recombination rates and



**Fig. 6.** The plots of estimated recombination rate versus 1—symmetry measure along each individual chromosome in human. For each chromosome, the upper panel plots '1—symmetry measure at order 12' and the lower panel plots the corresponding recombination rates. Symmetry measures and recombination rates were calculated across non-overlapping 5 Mb windows.

some windows did not have estimated male recombination rates,  $n$  was not equal to  $m$ . From the results, the sex-specific symmetry effect  $\beta_d$  was estimated to be  $-2.78$  with  $P$ -value  $< 2 \times 10^{-16}$ . The common symmetry effect  $\beta_s$  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

- Baisnee,P.F. *et al.* (2002) Why are complementary DNA strands symmetric? *Bioinformatics*, **18**, 1021–1033.
- Broman,K.W. *et al.* (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.*, **63**, 861–869.
- Dietrich,W.F. *et al.* (1996) A comprehensive genetic map of the mouse genome. *Nature*, **380**, 149–152.
- Forsdyke,D.R. (1995) A stem-loop ‘kissing’ model for the initiation of recombination and the origin of introns. *Mol. Biol. Evol.*, **12**, 949–958.
- Fullerton,S.M. *et al.* (2001) Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.*, **18**, 1139–1142.
- Jensen-Seaman,M.I. *et al.* (2004) Comparative recombination rates in the rat, mouse and human genomes. *Genome Res.*, **14**, 528–538.
- Kong,A. *et al.* (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**, 241–247.
- Lichten,M. and Goldman,A.S. (1995) Meiotic recombination hotspots. *Annu. Rev. Genet.*, **29**, 423–444.
- Magasanik,B. and Chargaff,E. (1951) Studies on the structure of ribonucleic acids. *Biochim. Biophys. Acta*, **7**, 396–412.
- Nachman,M.W. (2002) Variation in recombination rate across the genome: evidence and implications. *Curr. Opin. Genet. Dev.*, **12**, 657–663.
- Petes,T.D. (2001) Meiotic recombination hot spots and cold spots. *Nat. Rev. Genet.*, **2**, 360–369.
- Prabhu,V.V. (1993) Symmetry observations in long nucleotide sequences. *Nucleic Acids Res.*, **21**, 2797–2800.
- Qi,D. and Cuticchia,A.J. (2001) Compositional symmetries in complete genomes. *Bioinformatics*, **17**, 557–559.
- Rudner,R. *et al.* (1968) Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. *Proc. Natl Acad. Sci. USA*, **60**, 921–922.
- Smith,G.R. (1988) Homologous recombination in procaryotes. *Microbiol. Rev.*, **52**, 1–28.
- Steen,R.G. *et al.* (1999) A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. *Genome Res.*, **9**, AP1–8, insert.
- Watson,J.D. and Crick,F.H. (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **171**, 964–967.
- Wu,T.C. and Lichten,M. (1994) Meiosis-induced double-strand break sites determined by yeast chromatin structure. *Science*, **263**, 515–518.