The Sex-Specific Impact of Meiotic Recombination on Nucleotide Composition

Alexandra Popa[1,2,3,4], Paul Samollow[5], Christian Gautier[1,2,3,4], and Dominique Mouchiroud[1,2,3,4,*]

[1]Université de Lyon, F-69000, France

[2]Université Lyon 1, France

[3]CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France

[4]Equipe BAMBOO, INRIA Rhône-Alpes, 655 avenue de l'Europe, F-38330, Montbonnot Saint-Martin, France

[5]Department of Veterinary Integrative Biosciences, Faculty of Genetics, Texas A&M University

*Corresponding author: Email: dominique.mouchiroud@univ-lyon1.fr.

Abstract

Meiotic recombination is an important evolutionary force shaping the nucleotide landscape of genomes. For most vertebrates, the frequency of recombination varies slightly or considerably between the sexes (heterochiasmy). In humans, male, rather than female, recombination rate has been found to be more highly correlated with the guanine and cytosine (GC) content across the genome. In the present study, we review the results in human and extend the examination of the evolutionary impact of heterochiasmy beyond primates to include four additional eutherian mammals (mouse, dog, pig, and sheep), a metatherian mammal (opossum), and a bird (chicken). Specifically, we compared sex-specific recombination rates (RRs) with nucleotide substitution patterns evaluated in transposable elements. Our results, based on a comparative approach, reveal a great diversity in the relationship between heterochiasmy and nucleotide composition. We find that the stronger male impact on this relationship is a conserved feature of human, mouse, dog, and sheep. In contrast, variation in genomic GC content in pig and opossum is more strongly correlated with female, rather than male, RR. Moreover, we show that the sex-differential impact of recombination is mainly driven by the chromosomal localization of recombination events. Independent of sex, the higher the RR in a genomic region and the longer this recombination activity is conserved in time, the stronger the bias in nucleotide substitution pattern, through such mechanisms as biased gene conversion. Over time, this bias will increase the local GC content of the region.

Key words: GC biased gene conversion, recombination, substitution pattern, GC-content, sex, meiosis.

Introduction

Homologous recombination is a virtually universal feature of gametogenesis. Indeed, in most species, meiotic recombination is believed to be essential for the proper alignment of homologous chromosomes and their successful segregation at anaphase I (Baker et al. 1976; Hunt and Hassold 2002; Petronczki et al. 2003), and its failure can generate aneuploid gametes and offspring. Furthermore, by reshuffling chromosome segments, meiotic recombination produces novel multilocus haplotypes, some of which may be favorable under certain conditions and serve as potential selective alternatives for adaptive evolution (Otto and Lenormand 2002; Marais and Charlesworth 2003).

Another important evolutionary role of recombination is its influence on the nucleotide composition of the genomes, both as a whole, and at the subchromosomal level. Within-genome variation in the intensity of recombination events correlates strongly with regional differences in nucleotide composition. In taxa ranging from yeast to mammals, higher recombination rates (RRs) have been found to correspond to regions enriched in guanine and cytosine (GC) (Gerton et al. 2000; Marais et al. 2001; Meunier and Duret 2004; Webster et al. 2005; Duret and Arndt 2008; Berglund et al. 2009). There are two opposing causal interpretations of this correlation. Several studies have proposed that high GC content promotes recombination, such that GC-rich regions represent sites that favor a chromatin structure that is open to the recombination machinery (Gerton et al. 2000; Petes 2001; Blat et al. 2002; Petes and Merker 2002; Marsolier-Kergoat and Yeramian 2009). The other

interpretation is that higher RRs are a direct cause of increased GC content. Meunier and Duret (2004) showed that in humans, the RR correlates more strongly with equilibrium GC content (GC*, defined below) rather than with the observed GC content. GC* is the GC content toward which a DNA sequence is expected to evolve under a constant base substitution pattern and rate. Based on their analysis, it seems likely that the RR has a direct influence on the pattern of nucleotide substitution, leading to an increased GC content in regions of high recombination, through the mechanism known as GC-biased gene conversion (gBGC) (Eyre-Walker 1993). Meiotic recombination requires the pairing of homologous chromosomes before exchange. If this pairing were to overlap a polymorphic site, then the resulting heteroduplex would contain a base pair mismatch which could be repaired by converting one of the alleles into the other. It has been shown experimentally that repair of A- or T-containing heteroduplexes in mammals is slightly biased toward conversion to a GC pair (Brown and Jiricny 1989; Bill et al. 1998). Therefore, high RRs can account for an increase in GC content over an evolutionary time scale (Galtier et al. 2001).

Importantly, the effect of recombination on the genomic landscape seems to be sex specific. Sex-specific differences in the intensity and distribution of recombination events (heterochiasmy) have been reported for many, although not all, animal and plant species. Some species have no recombination at all in one sex; for example, there is no double-strand break-mediated recombination in male *Drosophila melanogaster* (Morgan 1912, 1914) or in the females of some lepidopterans (Miao et al. 2005; Yamamoto et al. 2006). More often both sexes have recombination, but one sex exhibits more recombination overall than the other. In the majority of these cases, the observed heterochiasmy is toward greater whole-genome RRs in females. By contrast, linkage data from the domestic sheep *Ovis aries* (Maddox et al. 2001; Maddox and Cockett 2007) and three metatherian mammals, including the fat-tailed dunnart, *Sminthopsis crassicaudata* (Bennett et al. 1986), the tammar wallaby *Macropus eugenii* (Zenger et al. 2002), and the gray short-tailed opossum, *Monodelphis domestica* (Samollow et al. 2007), exhibit the reverse pattern, with male RRs exceeding those in females in the regions mapped. Nevertheless, pronounced heterochiasmy is not a universal feature of all vertebrates, as overall RRs in cattle, *Bos taurus* (Ihara et al. 2004), some galliform birds (Reed et al. 2005; Groenen et al. 2009), and some fishes (Walter et al. 2004) display no or very small differences in map lengths (inferred RRs) between sexes. Furthermore, studies in humans and mice have shown that, at a local scale, the ratio between female and male RRs (the F/M ratio) is <1.0 near telomeres and >1.0 in centromeric regions (Kong et al. 2002; Shifman et al. 2006). For sheep, this ratio is <1.0 in both of these regions (Maddox and Cockett 2007).

Given these sex differences in RR, it was expected that because the female sex has more crossovers (COs) than

the male in the majority of eutherian mammals, variation in female RR would correlate better with variation in the GC content across the genome. However, by analyzing the substitution patterns in *Alu* elements in humans, Webster et al. (2005) found that the GC* content of a region is more strongly correlated with male RR than with female RR. These findings were later confirmed by inferring the substitution pattern in 1 Mb of noncoding sequences across the human genome (Duret and Arndt 2008). Moreover, AT → GC biased substitution hot spots in humans have also been found to correlate more strongly with male than with female recombination patterns (Dreszer et al. 2007).

In the present study, we address the question of the differential impact of sex-specific RRs on the nucleotide composition landscape of several vertebrate genomes: five eutherians (human, mouse, dog, sheep, and pig), a metatherian (opossum), and a bird (chicken), which are subject to different heterochiasmy patterns. Based on our findings, we propose a new explanation for the impact of heterochiasmy on the nucleotide landscape: specifically, that at a local level, regions with temporally stable, especially high RRs, independent of sex, will experience more gBGC events and, thus, have a stronger influence on regional base composition than regions with low RRs.

Materials and Methods

Main variables: RRs, equilibrium GC (GC*), and observed GC contents were calculated using appropriately sized windows (described below) along each of the genomes of seven species: human, mouse, dog, sheep, pig, opossum, and chicken having available sex-specific genetic maps as well as genome assemblies. GC* was inferred from alignments between present transposable element (TE) sequences and their consensus sequence as retrieved from RepeatMasker (Arndt et al. 2003).

Transposable Elements

Whole-genome alignments of TEs to their consensus sequences were obtained from the RepeatMasker site http://www.repeatmasker.org/PreMaskedGenome.html for human (version HG19—March 2006), mouse (version mm9—July 2007), chicken (version galGal3—May 2006), and opossum (monDom5—October 2006) genomes. Comparable data were not available for dog, sheep, and pig so we launched RepeatMasker on the whole dog (CanFam 2.0—May 2006), sheep (Oarv2.0—March 2011), and pig (Sscrofa9—April 2009) genome assemblies using the slow option. Genomic sequences for each sheep chromosome were obtained from http://www.livestockgenomics.csiro.au/sheep/.

For dog and pig, these sequences were retrieved from Ensembl database, http://www.ensembl.org/index.html (Flicek et al. 2011). Following the protocol of Arndt et al. (2003),

we considered only TEs of type LINE and SINE, with alignment lengths longer than 250 and 100 bp, respectively. TEs that overlapped in these alignment files were discarded to avoid ambiguities in the inference of substitution rates. All TEs overlapping exons were also eliminated because selection on exon sequences could have confounding effects on the evolution of their GC content. Exon positions for all genes were retrieved from version 57 of the Ensembl database (Flicek et al. 2011), except for sheep for which CDS positions were downloaded from http://www.livestockgenomics.csiro.au/sheep/ (ISGC et al. 2010). Information regarding the number of TEs used for each species, before and after the application of the overlap and exon filters, is available in supplementary table S-1 (Supplementary Material online).

Windows

We restricted the analysis to autosomal sequences only. In the case of human, mouse, dog, and chicken, each chromosome was divided into nonoverlapping windows of 1 Mb. All TEs within a window were considered. Windows containing no TEs were discarded. The TEs overlapping the limits of consecutive windows were reassigned to the window for which the TE had the greatest overlap.

Due to the low number of genetic markers available for sheep, pig, and opossum, windows for these species were defined by the positions of two consecutive genetic markers on their genetic maps; http://www.livestockgenomics.csiro.au/sheep/, Vingborg et al. (2009), and Samollow et al. (2007), respectively. For these species, windows smaller than 5 kb were discarded from the analysis. The assignment of TEs to each window was made according to the same principle described above.

For all species, the distributions of window lengths were based on the distances between the first and the last TE assigned to each window. These are presented in supplementary figs. S-1 (for human, mouse, dog, and chicken) and S-2 (for sheep, pig, and opossum) (Supplementary Material online).

Recombination

RR data were calculated from published sex-specific and sex-averaged genetic maps: human (Matise et al. 2007), mouse (Cox et al. 2009), dog (Wong et al. 2010) (http://www.vgl.ucdavis.edu/dogmap/), pig (Vingborg et al. 2009), opossum (Samollow et al. 2007), and chicken (Groenen et al. 2009). For sheep, linkage map version 5.0 was downloaded from http://rubens.its.unimelb.edu.au/~jillm/jill.htm (11 October 2010).

In the case of human, mouse, dog, and chicken, each window was characterized by four genetic markers: the two closest genetic markers flanking the left window boundary and two markers flanking the right boundary of the window (supplementary fig. S-3, Supplementary Material online). The RRs,

expressed in centimorgans per megabase (cM/Mb), for each window were computed as the average of the RR between each adjacent pair ($n = 3$ pairs) of the above four consecutive genetic markers weighted by their overlap to the window. Only windows defined by at least three genetic markers were analyzed. Details of the formula used for the calculation of RR are presented in supplementary fig. S-3 (Supplementary Material online).

The genetic maps of sheep, pig, and opossum have 2371, 462, and 150 markers, respectively. The physical positions of the sheep markers were obtained from http://www.livestockgenomics.csiro.au/sheep/. In the case of the pig markers, we obtained the physical positions of the gene-associated single nucleotide polymorphisms from the Ensembl database (Flicek et al. 2011). In the case of the opossum markers, in order to find their corresponding physical position, we mapped their forward and reverse primers on the assembled opossum genome. The positions of the primers on the genome were obtained using nucleotide basic local alignment search tool (BlastN) (Altschul et al. 1990) at National Center for Biotechnology Information. We retained only genetic markers for which both genetic and physical positions mapped to appropriate positions on the same chromosome. All markers with inverted positions on the genetic versus physical maps were also discarded. In the end, we were left with 1,320 sheep, 227 pig, and 115 opossum markers, defining 1,294, 209, and 107 interlocus intervals, respectively.

Equilibrium GC (GC*) and Observed GC Contents

The maximum-likelihood method of Arndt et al. (2003) was used to compute substitution rates for individual nucleotides accounting for CpG hypermutability. The substitutions are inferred between the ancestral sequence as predicted by RepeatMasker and the current observed sequences of TEs. The consensus sequence of TEs is assumed to approximate the ancestral sequence. We retrieved the alignment between the TE and its consensus sequence inside each window, using RepeatMasker. Based on these alignments, the method infers seven substitution rates, supposing strand symmetry (e.g., $A \rightarrow G = T \rightarrow C$): four transversions, two transitions, and the CpG transition rate. GC* was thus calculated on all non-CpG substitutions in each window according to the model of Sueoka (1962) as the percentage of $AT \rightarrow GC$ substitutions among all $AT \rightarrow GC$ and $GC \rightarrow AT$ substitutions. In order to obtain high precision in the estimation of the substitution frequency, we eliminated from the analysis any windows containing concatenated alignments that had less than 100 kb of uninterrupted unambiguous nucleotide sequences (no indels and no N bases) (Duret and Arndt 2008). Because TEs are much less numerous in the chicken genome than in those of the mammals studied (supplementary table S-1, Supplementary Material online), we eliminated

windows containing alignments with less than 20 kb of informative nucleotide sequences (the value of the first quartile) for this species.

For each window, the observed GC content was computed from the genomic sequences after eliminating exons. The positions of predicted (annotated) exons were retrieved from version 63 of the Ensembl database (Flicek et al. 2011) for all species except sheep. The sheep genomic sequence, as well as its CDS positions were obtained from http://www.livestockgenomics.csiro.au/sheep/.

Centromere Positions

The centromere positions for human, mouse, dog, and chicken were retrieved from http://genome.ucsc.edu/. For sheep, all chromosomes are acrocentric except chromosomes 1, 2, and 3, which are metacentric. The approximate centromere positions of these three metacentric chromosomes were inferred from the data of Maddox et al. (2001). The centromere positions for pig were not available. The centromere positions for opossum were assigned according to the cytological data of Duke et al. (2007).

Statistics

We quantified the strength of the correlation between any two variables using Pearson's ρ correlation coefficient. In the case of two correlations sharing a common variable, we applied a Hotteling–William's t-test (Williams 1959), with the null hypothesis $r_{XY} = r_{XZ}$, in order to test which was stronger.

$$t = |r_{XY} - r_{XZ}| \sqrt{\frac{(N-1)(1+r_{YZ})}{2\frac{N-1}{N-3}|R| + \frac{(r_{XY}+r_{XZ})^2}{4}(1-r_{YZ})^3}},$$

where $|R| = 1 - r_{XY^2} - r_{XZ^2} - r_{YZ^2} + 2r_{XY}r_{XZ}r_{YZ}$ (the determinant of the 3×3 correlation matrix) and N is the number of observations in each variable X and Y. This ratio follows a Student's t distribution with $N - 3$ degrees of freedom.

P values of the Hotteling–William's t-test are reported in brackets in the main text.

Results

To study the sex-specific effect of recombination on nucleotide composition, we analyzed the correlations between equilibrium GC-content (GC*) and sex-specific RRs for seven vertebrate species having available sex-specific genetic maps as well as genome assemblies. We computed the GC* from TEs grouped in windows, using the maximum-likelihood method proposed by Arndt et al. (2003).

Various Influences of Sex on the RR/GC* Correlation

For the first time, we show that in mouse and dog, consistent with previous studies in human (Webster et al. 2005; Duret and Arndt 2008), and despite a longer genetic map in females (Kong et al. 2002; Matise et al. 2007; Cox et al. 2009; Wong et al. 2010), it is male local RR that is more strongly correlated with the nucleotide composition landscape (fig. 1 and table 1). Interestingly, while exhibiting the reverse global heterochiasmy pattern, with a longer genetic map in male than female, the GC content in the sheep genome is also more strongly correlated with male than with female local RR (fig. 1 and table 1). To compare the overlapping dependent correlations of sex-specific RRs with the common variable GC*, we performed a Hotteling–William's t-test (Williams 1959). The results of these tests showed that in human, mouse, dog, and sheep, the correlation of GC* with local RR was significantly stronger for males than for females (table 1).

However, for the other three species analyzed in this study, we find different results. For the pig (P value = 0.0002) and the opossum (P value = 0.028), we find that female rather than male local RR correlates more strongly with nucleotide composition (fig. 2 and table 1). In chicken, we see no difference between the sexes in the strength of this correlation (P value = 0.1434) (fig. 2 and table 1).

Heterochiasmy and Chromosome Localization

In figures 1 and 2, we represent the correlation of GC* with female and male RRs, considering individually regions close to telomeres (red) and regions close to centromeres (blue). We observe in figure 1 that the male component of the RR/GC* correlation, especially in human, is driven mainly by windows situated less than 5 Mb away from telomeres (red points). This strong regional impact in the male data set is the result of a highly preferential localization of male recombination hot spots in telomeric and subtelomeric regions as opposed to their more uniform distribution in females (Kong et al. 2002; Shifman et al. 2006; Cheung et al. 2007; Maddox and Cockett 2007).

To examine the impact of chromosome position on RR/GC* correlation, we eliminated windows situated within 5 Mb of telomeres and centromeres (table 1). In humans, local RR in the remaining interstitial regions correlates more strongly with the GC* for females (0.41), than for males (0.28) (table 1). In sheep, we observe no difference between sexes in the RR/GC* relationship in the interstitial regions (table 1). Contrastingly, elimination of the subtelomeric and centromeric regions does not abolish the stronger male RR/GC* correlation in mouse or dog (table 1). Elimination of telomeric and centromeric windows has no effect in chicken where the difference between sex-specific recombination pattern and GC* correlations remain small and nonsignificant (P value = 0.062). Sufficient telomeric and subtelomeric genetic markers are lacking for pig and opossum to enable similar examinations in these species.

As expected from the foregoing observations (reviewed by Backstrom et al. 2010), the distance to telomeres
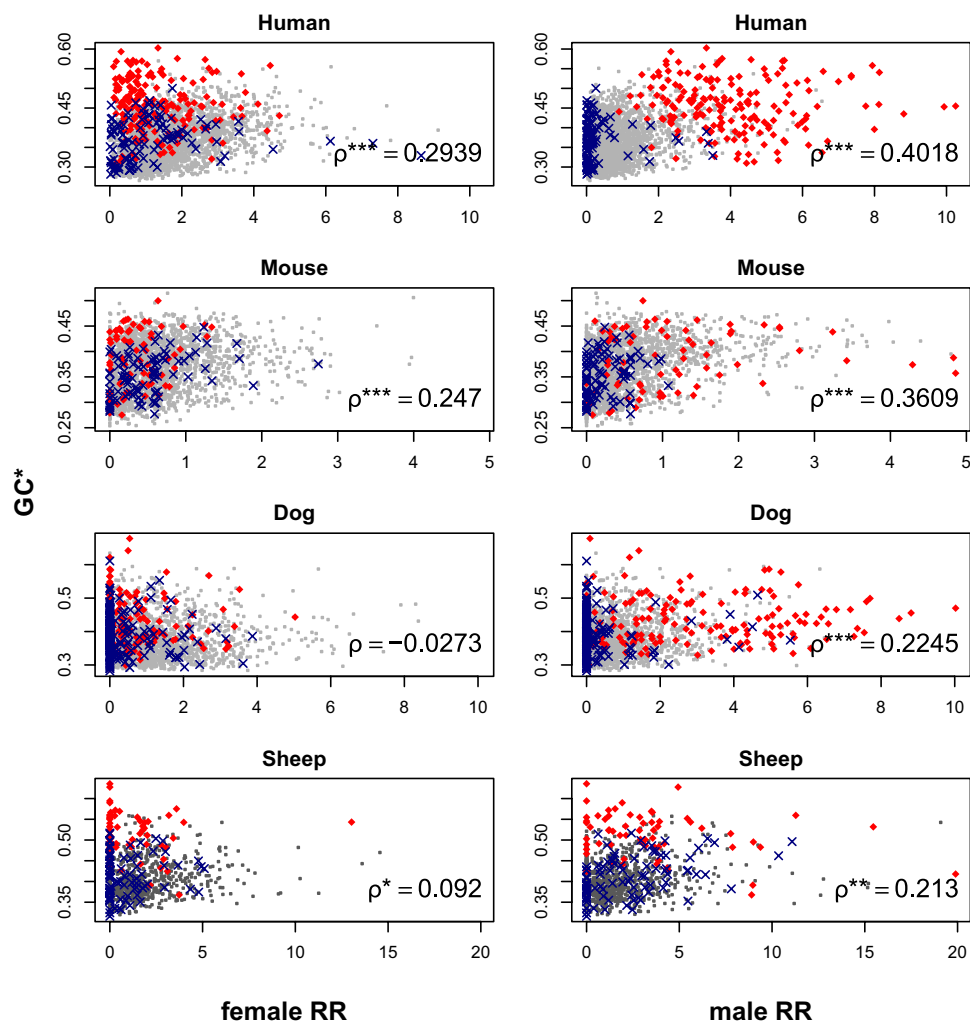
Fig. 1.—The correlations between RRs (male and female) and equilibrium GC-content (GC*) in human, mouse, dog, and sheep. Each point represents the value of the variables in a 1 Mb window, except for sheep for which the length of the window is defined between two consecutive genetic markers. Blue x-shaped points represent windows within 5 Mb (left and right) of the centromere. Red diamonds are windows within 5 Mb of the telomeres. Gray points represent windows that are not localized near centromeres or telomeres. The value of Pearson's ρ correlation coefficient is reported for each graph. Three, two, and one asterisks near these values stand for P values of the correlation $\leq 10^{-16}$, $\leq 10^{-10}$, and ≤ 0.05, respectively.

correlates strongly and negatively with RR. We confirm this result, and in agreement with a preferential localization of male hot spots in telomeric regions, we demonstrate a stronger correlation between RR and log distance to telomeres (LDT) in males than in females for human, mouse, and dog (supplementary table S-2, Supplementary Material online). In opossum, contrary to the other species analyzed in this study, LDT is more strongly negatively correlated with female than with male RR (supplementary table S-2, Supplementary Material online). There was no statistically significant sex-difference in the correlation between RR and LDT for pig or chicken (supplementary table S-2, Supplementary Material online).

Our results show that the sex with the higher RR in regions close to telomeres contributes a stronger effect on the RR/GC* correlation, probably through the process of

gBGC (fig. 1 and table 1). For example, both sheep and opossum have more recombination events in male than in female (Maddox and Cockett 2007; Samollow et al. 2007). However, in sheep, regions close to telomeres have elevated rates of recombination in males (Maddox and Cockett 2007), whereas in opossum, it is females that show elevated RR in these regions (Sharp and Hayman 1988). This would explain why having the same global heterochiasmy pattern, we find a stronger male local RR/GC* in sheep and stronger female local RR/GC* correlation in opossum (table 1).

Overall, we have identified three variables that have an influence on GC*: RR, sex, and LDT. We used partial correlations to try to quantify the contribution of each one of these factors (table 2). We define RR_{max} as the RR of the sex having the highest RR/GC* correlation (table 1). To test

Table 1

Pearson's ρ Correlation Coefficient between RR and GC* for Three Data Sets: 1. All Windows along the Chromosomes, 2. Only Windows 5 Mb Away from Telomeres, and 3. Only the Interstitial Windows (5 Mb away from telomeres and centromeres)

$\rho_{RR\,-\,GC*}$	Human ♀	Human ♂	Mouse ♀	Mouse ♂	Dog ♀	Dog ♂	Sheep ♀	Sheep ♂	Pig ♀	Pig ♂	Opossum ♀	Opossum ♂	Chicken ♀	Chicken ♂
All data	0.29*** <	0.40***	0.25*** <	0.36***	−0.03 <	0.22**	0.09* <	0.21**	0.37* >	0.16*	0.47* >	0.31*	0.54*** ≈	0.58***
	4×10^{-7}		5×10^{-7}		4×10^{-18}		0.002		0.0002		0.028		0.14	
No telomeres	0.39*** >	0.26***	0.25*** <	0.37***	0.02 <	0.15**	0.21* ≈	0.25*	—		—		0.56*** ≈	0.60***
	4×10^{-11}		3×10^{-7}		2×10^{-7}		0.36						0.22	
Interstitial regions	0.41*** >	0.28***	0.25*** <	0.37***	0.03 <	0.18**	0.16* ≈	0.16*	—		—		0.50*** ≈	0.57***
	4×10^{-11}		1×10^{-7}		2×10^{-7}		0.919						0.062	

NOTE.— >, <, and ≈ represent statistical difference between female and male ρ values (Hotteling–William's t-test) reported as inequalities. For each of the two data sets, the P values for the Hotteling–William's t-test are also reported in the bottom row. —, No available data.

*P values of the correlation test ≤0.05, **P values of the correlation test ≤10^{-10}, ***P values of the correlation test ≤10^{-16}.

---

the effect of sex and recombination on GC*, independent of the telomere effect, we performed the partial correlation of GC* to RR_{max} controlling for LDT. These analyses suggest that the RR_{max}/GC* correlations remain significant for all species even after controlling for the distance to telomeres (table 2). Although we observe a reduction in the strength of correlation between the GC* ~ RR_{max} and GC* ~ RR_{max}|LDT models, the amplitude of this reduction varies among species (table 2). Interestingly, the effect of LDT on GC*, independent of sex and recombination, remains significant (table 2) for all species except the opossum for which we lack data close to telomeres. We conclude that both sex and LDT are likely to be good predictors of GC*, in all these and probably other vertebrate species.

Different Map Resolution

Given the low resolution of the sheep, pig, and opossum genetic maps, we defined windows as the regions between pairs of consecutive genetic markers, thus, generating windows with median sizes of 1.7, 4.4, and 18.3 Mb, respectively, but with high variance (supplementary fig. S-2, Supplementary Material online). Differences in the length and variability of window sizes between these and the other species (windows of 1 Mb) could have had a confounding effect on our interpretation of the relationship between sex, RR, and GC*. Previous studies in humans (Duret and Arndt 2008) and yeast (Marsolier-Kergoat and Yeramian 2009), as well as our own work, show that the RR/GC* correlation coefficients increase with the size of the windows (supplementary table S-3, Supplementary Material online). Notably, at the scale of a few kilobase, the locations of CO hot spots are known to vary strongly among individuals of the same species (Neumann and Jeffreys 2006; Jeffreys and Neumann 2009), whereas it has been proposed that recombination across larger Mb-scale regions is more uniform among individuals and stable over time (Myers et al. 2005). Therefore, we tested the effect of window size (from 0.5 to 20 Mb) on the strength of correlations between female RR and GC* (ρ_{RR_F-GC*}) and male RR and GC* (ρ_{RR_M-GC*}) in human and mouse. In both species, the stronger male effect on RR/GC* correlation is detectable for small window sizes, but as the size of the windows increases, the sex-specific difference diminishes and ultimately disappears (supplementary table S-3, Supplementary Material online). Contrastingly, the stronger female RR/GC* correlation persists in opossum, even for windows with a median size of 18.3 Mb. Moreover, when dividing the opossum data set into windows smaller and larger than 20 Mb, the greater female versus male impact is still apparent (supplementary fig. S-4, Supplementary Material online). Similarly, the division of the sheep and pig data sets in windows with length less than 5 Mb yields the same conclusions as the original data sets (supplementary figs. S-5 and S-6, Supplementary Material online), stronger
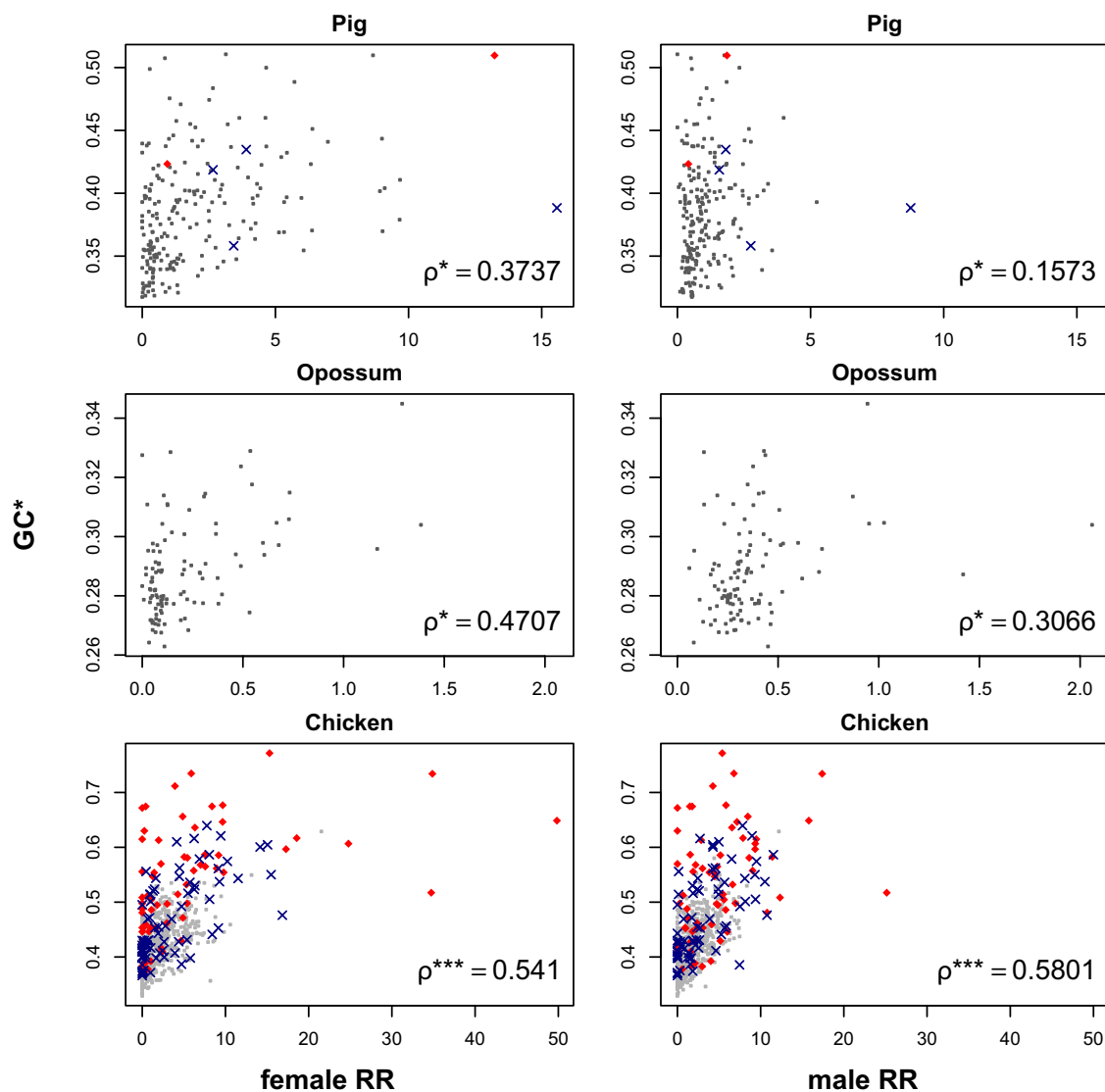
FIG. 2.—The correlations between RRs (male and female) and GC* in pig, opossum, and chicken. The value of Pearson's ρ correlation coefficient is reported for each graph. One asterisk near these values stands for *P* values ≤0.05 for the correlation.

RR/GC* correlation for male sheep and female pig. The outcomes suggest that the RR/GC* correlations detected for sheep, pig, and opossum are genuine and not artifacts of large window size or high window size variance.

Different TE Divergence

The focus on TEs has enabled us to conduct robust RR/GC* correlation analyses in other vertebrates than human, in the absence of multiple-species whole-genome alignments. TEs have been generated by bursts of transpositional insertions at different times during vertebrate evolution. Thus, the TEs present in a genome today have different ages, and the substitution patterns they generate are indicative of multiple substitution processes taking place over long periods of time, perhaps even prior to insertion into their present

locations. In contrast, the RRs inferred from genetic maps correspond to current recombination processes, which are dynamic, with ongoing birth and death of recombination hot spots (Ptak et al. 2005; Winckler et al. 2005). For this reason, a more accurate description of the RR/GC* relationship might be expected by focusing only on recently diverged sequences.

In order to reduce the risk of analyzing TE substitution patterns accumulated prior to insertion, we reduced our data set in human to those AluY and L1PA elements that have been found as conserved between human and chimpanzee (Kvikstad and Makova 2010). These conserved elements are assumed stable as they come from homologous regions in the two species. Although the data set and the length of the alignment per window are both greatly

Table 2

Pearson's ρ Correlation Coefficient and Partial Correlation Coefficients Explaining the Evolution of GC*

Species	Sex-Max	Correlation and Partial Correlations of GC* with		
		RR$_{max}$	RR$_{max}$|LDT	LDT|(RR♂, RR♀)
Human	♂	0.40***	0.07*	−0.337***
Mouse	♂	0.36***	0.30***	−0.16**
Dog	♂	0.22**	0.16**	−0.1*
Sheep	♂	0.21**	0.18*	−0.40***
Pig	♀	0.37*	0.34*	−0.254*
Opossum	♀	0.47*	0.36*	−0.07
Chicken	♂	0.58***	0.53***	−0.38***

NOTE.—The second column contains the sex for which the overall RR/GC* correlation has the highest ρ value, although for the chicken there is no statistical difference in ρ between sexes.

*P values of the correlation test ≤ 0.05, **P values of the correlation test $\leq 10^{-10}$, ***P values of the correlation test $\leq 10^{-16}$.

reduced, the results, shown in supplementary table S-4 (Supplementary Material online), are consistent with our previous observations on the whole TE data set: the stronger female effect on the RR/GC* correlation outside of the telomeric regions. We next computed the human GC* based on the triple alignment in 1-Mb windows (as described in Duret and Arndt 2008) and the RR from the deCODE 2002 human genetic map (Kong et al. 2002). This analysis showed that, as in the case of TEs, GC* is more strongly correlated with female than with male RR outside of subtelomeric regions (supplementary table S-5, Supplementary Material online). Moreover, conclusions regarding sex-specific effects on the RR/GC* correlation are robust to different divergence filters. When filtering the TE families with mean divergence >20% and standard deviation >5%, as well as copies with >20% divergence (supplementary table S-6, Supplementary Material online), or when focusing separately on the three Alu families (supplementary table S-7, Supplementary Material online), the conclusions remain unchanged. Overall, these results suggest that the use of older TE sequences leads to conclusions similar to those based on the less diverged triple alignments between human, chimpanzee, and macaque data set in Duret and Arndt (2008) (supplementary table S-5, Supplementary Material online). It appears that the use of TEs provides a reliable tool for the study of neutral substitution patterns in different vertebrate genomes.

Discussion

We analyzed the relationship between local RR, sex, and nucleotide composition across the autosomes of seven vertebrates: five eutherian mammals (human, mouse, dog, sheep, and pig), a metatherian mammal (opossum), and a bird (chicken). The analysis of this large panel of species reveals a broad range in this relationship, previously

characterized only for humans (Webster et al. 2005; Duret and Arndt 2008).

Novel Insights in Human

In this study, we confirm previous results of human data showing stronger male, rather than female, impact on the RR/GC* correlation (Webster et al. 2005; Duret and Arndt 2008). Moreover, RR and the LDT have been found to be the main predictors of GC* (Duret and Arndt 2008). Our results clarify and expand these interpretations by demonstrating that LDT is the main component of the stronger male RR/GC* correlation in human (tables 1 and 2). Specifically, removing telomeric and subtelomeric regions from the analyses drastically reduces the genome-wide variation in male RR in humans (table 1). Thus, the sex influence on GC* becomes reversed, as the female RR/GC* correlation becomes stronger in interstitial regions (table 1). Moreover, the correlation of GC* to RR greatly decreases (from 0.4 to 0.07) in male human after controlling for the impact of LDT (table 2).

Our results indicate that in human, observed sex-specific differences in RR/GC* correlation are largely the consequence of sex-specific strategies for the distribution of recombination events along chromosomes (heterochiasmy). Regions close to telomeres have a higher RR in male human than in female (Matise et al. 2007) (supplementary table S-2, Supplementary Material online). Moreover, high recombination seems to be a conserved feature of telomeric and subtelomeric regions, at least for the species vertebrate investigated in this study. The probability of AT → GC substitutions becoming fixed in a genomic region through gBGC depends both on the rate of recombination and on the conservation of recombination activity for a long period of time. In view of our results, we propose a new explanation of the relation between sex, recombination, and nucleotide composition. Regions with temporally stable high RRs, independent of the sex generating them, will experience more gBGC events and thus generate a stronger influence on the regional nucleotide composition than those with lower rates.

A previous hypothesis explaining the stronger male impact in human was that the CO rate in male is a better estimator of the total double-strand break rate than in female (Duret and Arndt 2008). Importantly, COs are not the only double-strand breakage processes that are expected to have a preferential subtelomeric localization. An unknown fraction of double-strand breakage events are resolved to non CO products (Blitzblau et al. 2007; Buhler et al. 2007; Barton et al. 2008), which could also be subject to gBGC. The RR measured through linkage mapping is characteristic only of those double-strand breaks that are resolved to recombinant chromosomes. If the ratio between COs and non-COs were more variable in human females, it would

generate a weaker correlation between the analyzed CO rate and the total RR and thus account for a less marked impact of COs on the GC* in this sex. Data on the number and distribution of non-CO DSB events are needed to enable further testing of this hypothesis. In the present study, we demonstrate that the impact of sex on the RR/GC* correlation depends greatly on the differential chromosomal localization. Although we do not propose that non-CO DSBs are the main factor explaining the stronger correlation of male RR to GC* in human, we do suggest they play a substantial role in increasing the GC content through gBGC. The strong contribution of the distance to telomeres with GC*, after controlling for both sex and RR, is indirect evidence of the non-CO DSBs influence (table 2). Therefore, we hypothesize that the distance to telomeres is indicative not of the CO rate alone but rather of the total DSB rate and thus provide a potential supplementary explanation for variation in GC*.

Heterochiasmy and Nucleotide Composition in Other Vertebrates

Can we extrapolate the results obtained in human to other vertebrates? We show that although the male RR is a better predictor of the local equilibrium GC content (GC*) in mouse, dog, and sheep, this correlation is stronger for female RR in pig and opossum (figs. 1 and 2 and table 1), and we observe no sex-differential impact in chicken (fig. 2 and table 1). Moreover, the role played by telomeres in the RR/GC* correlation is variable between species (tables 1 and 2). In sheep, we detect no sex-differential impact in interstitial chromosome regions (table 1). Similar to human, this change in outcome for sheep is the consequence of male recombination hot spots being localized primarily in telomeric and subtelomeric regions in these species compared with a more uniform distribution of female COs along the chromosomes (fig. 1) (Maddox et al. 2001; Maddox and Cockett 2007).

However, in sheep, as opposed to human, the impact of telomeres on the RR/GC* is weaker as the male RR conserves a strong correlation with GC* even after controlling for the LDT (table 2). For male sheep, the value of the GC*/RR correlation is 0.21 (table 2). The value of this correlation decreases to 0.18 when controlling for LDT (table 2). However, when performing an additional partial correlation in male sheep, between GC*/RR but controlling for both LDT and the distance to centromeres, the correlation greatly decreases to 0.11 (P value = 0.0002). In male sheep, centromeres, similar to telomeres, have higher male RR than female RR (fig. 1) (Maddox et al. 2001), and this sexual dimorphism is a conserved feature, as it is also present in a close relative, the bighorn sheep (Poissant et al. 2010). The availability in sheep of another conserved region with high male RR, centromeres, has allowed us to further confirm our hypothesis that regions with conserved high

RRs are major determinants of the RR/GC* correlation, independent of sex.

Despite establishing a sex-specific distribution of recombination hot spots similar to human, in the other two eutherians, mouse and dog, the stronger male effect is conserved outside the telomeric and subtelomeric regions (fig. 1) (Cox et al. 2009; Wong et al. 2010). This might be due to the more evolutionary derived state of these genomes, with multiple chromosomal rearrangements compared with that of the boreoeuthrerian common ancestor, relative to the less derived human genome (O'Brien et al. 1999; Nash et al. 2001; Wienberg 2004; Murphy et al. 2005; Kemkemer et al. 2009). This recent shuffling of the genomic sequence would result in younger chromosome ends as well as a distribution of latent telomere- and subtelomere-like structures along the interiors of chromosomes in these species (Meyne et al. 1990). This condition would influence the distribution of hot spots of recombination and nucleotide substitution patterns on a wider chromosomal range. This possibility is consistent with the observation in mouse that a substantial proportion of hot spots is shared between the two sexes outside of subtelomeric regions (Paigen et al. 2008).

The lack of genetic markers in telomeric and subtelomeric regions prevents us from determining the sex-specific correlation of RR and GC* according to chromosomal localization in pig and opossum. The stronger female RR/GC* correlation in pig (table 1), as well as the lack of genetic markers close to telomeres, could indicate a situation similar to that in human. It is essential to have genetic markers in subtelomeric regions to measure the influence of sex on GC* in this species. However, in opossum, even though the analyses are based on a small number of loci, we find that the distance to telomeres is more strongly negatively correlated with female than with male RR (supplementary table S-2, Supplementary Material online).

This result is also supported by cytologic studies of meiotic cells of the opossum (Hayman et al. 1988), which revealed that chiasmata are concentrated near the ends of chromosomes in female metaphase I nuclei, whereas those of males are much more evenly distributed. To the extent that this physical pattern reflects the actual distribution of chromosomal exchange events, we expect the female RR to be greater than male RR in subtelomeric regions. Consistent with this prediction, recent addition of new linkage map markers in subtelomeric regions beyond the previous map ends has had greater impact on increasing the length of female linkage map length than that of the male map (Samollow 2010). Thus, because subtelomeric regions in opossum are GC-rich relative to the genome average (Mikkelsen et al. 2007), we expect the correlation of female RR with GC* in opossum to become even clearer as more linkage data become available.

The impact of telomeres on the RR/GC* correlation is variable among species (tables 1 and 2). This variability might

depend on factors such as the chromosomal rearrangement rate as discussed above for mouse and dog. Moreover, telomeres might not be the only regions with a conserved recombinational landscape; consider, for example, the impact of centromeres, with high RR in male sheep (Maddox and Cockett 2007), on the GC*/RR correlation. The hypothesis of additional regions of conserved high RRs along the chromosomes is in agreement with the result in table 2, as the impact of sex on the RR/GC* correlations remains significant after controlling for the telomere effect. However, this analysis cannot exclude that sex per se might have other effects on GC*, independent of the chromosomal localization.

In conclusion, although the molecular mechanisms responsible for the sex-differential distribution of recombination along chromosomes are not yet understood, we show in this study that sex-specific differences in the distribution and frequency of recombination events are strong determinants in the evolution of nucleotide composition. Moreover, despite millions of years of divergence and resultant high sequence diversity among TE sequences, analyses of nucleotide composition characteristics based on these abundant and widely dispersed genomic elements provide substantial power for understanding patterns of genome evolution. Consequently, the use of TEs has, for the first time, allowed us to reveal a wide variability of the relation between sex, recombination, and nucleotide substitutions among vertebrate species.

Supplementary Material

Supplementary figures S-1 to S-6 and tables S-1 to S-10 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Arndt PF, Petrov DA, Hwa T. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. Mol Biol Evol. 20:1887–1896.

Backstrom N, et al. 2010. The recombination landscape of the zebra finch Taeniopygia guttata genome. Genome Res. 20:485–495.

Baker BS, Carpenter AT, Esposito MS, Esposito RE, Sandler L. 1976. The genetic control of meiosis. Annu Rev Genet. 10:53–134.

Barton AB, Pekosz MR, Kurvathi RS, Kaback DB. 2008. Meiotic recombination at the ends of chromosomes in Saccharomyces cerevisiae. Genetics 179:1221–1235.

Bennett JH, Hayman DL, Hope RM. 1986. Novel sex differences in linkage values and meiotic chromosome behaviour in a marsupial. Nature 323:59–60.

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. PLoS Biol. 7:e26.

Bill CA, Duran WA, Miselis NR, Nickoloff JA. 1998. Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese hamster ovary cells. Competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches. Genetics 149:1935–1943.

Blat Y, Protacio RU, Hunter N, Kleckner N. 2002. Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. Cell 111:791–802.

Blitzblau HG, Bell GW, Rodriguez J, Bell SP, Hochwagen A. 2007. Mapping of meiotic single-stranded DNA reveals double-stranded-break hotspots near centromeres and telomeres. Curr Biol. 17:2003–2012.

Brown TC, Jiricny J. 1989. Repair of base-base mismatches in simian and human cells. Genome 31:578–583.

Buhler C, Borde V, Lichten M. 2007. Mapping meiotic single-strand DNA reveals a new landscape of DNA double-strand breaks in Saccharomyces cerevisiae. PLoS Biol. 5:e324.

Cheung VG, Burdick JT, Hirschmann D, Morley M. 2007. Polymorphic variation in human meiotic recombination. Am J Hum Genet. 80:526–530.

Cox A, et al. 2009. A new standard genetic map for the laboratory mouse. Genetics 182:1335–1344.

Dreszer TR, Wall GD, Haussler D, Pollard KS. 2007. Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. Genome Res. 17:1420–1430.

Duke SE, Samollow PB, Mauceli E, Lindblad-Toh K, Breen M. 2007. Integrated cytogenetic BAC map of the genome of the gray, short-tailed opossum, Monodelphis domestica. Chromosome Res. 15:361–370.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4:e1000071.

Eyre-Walker A. 1993. Recombination and mammalian genome evolution. Proc Biol Sci. 252:237–243.

Flicek P, et al. 2011. Ensembl 2011. Nucleic Acids Res. 39:D800–D806.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907–911.

Gerton JL, et al. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 97:11383–11390.

Groenen MA, et al. 2009. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res. 19:510–519.

Hayman D, Moore H, Evans E. 1988. Further evidence for novel differences in chiasma distribution in marsupials. Heredity 61:455–458.

Hunt PA, Hassold TJ. 2002. Sex matters in meiosis. Science 296:2181–2183.

Ihara N, et al. 2004. A comprehensive genetic map of the cattle genome based on 3802 microsatellites. Genome Res. 14:1987–1998.

[ISGC] International Sheep Genomics Consortium, et al. 2010. The sheep genome reference sequence: a work in progress. Anim Genet. 41:449–453.

Jeffreys AJ, Neumann R. 2009. The rise and fall of a human recombination hot spot. Nat Genet. 41:625–629.

Kemkemer C, et al. 2009. Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. BMC Evol Biol. 9:84.

Kong A, et al. 2002. A high-resolution recombination map of the human genome. Nat Genet. 31:241–247.

Kvikstad EM, Makova KD. 2010. The (r)evolution of SINE versus LINE distributions in primate genomes: sex chromosomes are important. Genome Res. 20:600–613.

Maddox JF, Cockett N. 2007. An update on sheep and goat linkage maps and other genomic resources. Small Rumin Res. 70:4–20.

Maddox JF, et al. 2001. An enhanced linkage map of the sheep genome comprising more than 1000 loci. Genome Res. 11:1275–1289.

Marais G, Charlesworth B. 2003. Genome evolution: recombination speeds up adaptive evolution. Curr Biol. 13:R68–R70.

Marais G, Mouchiroud D, Duret L. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci U S A. 98:5688–5692.

Marsolier-Kergoat MC, Yeramian E. 2009. GC content and recombination: reassessing the causal effects for the Saccharomyces cerevisiae genome. Genetics 183:31–38.

Matise TC, et al. 2007. A second-generation combined linkage physical map of the human genome. Genome Res. 17:1783–1786.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol. 21:984–990.

Meyne J, et al. 1990. Distribution of non-telomeric sites of the (TTAGGG)n telomeric sequence in vertebrate chromosomes. Chromosoma 99:3–10.

Miao XX, et al. 2005. Simple sequence repeat-based consensus linkage map of Bombyx mori. Proc Natl Acad Sci U S A. 102:16303–16308.

Mikkelsen TS, et al. 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447:167–177.

Morgan T. 1912. Complete linkage in the second chromosome of the male of Drosophila. Science 36:719–720.

Morgan T. 1914. No crossing over in the male of drosophila of genes in the second and third pairs of chromosomes. Biol Bull. 26:195–204.

Murphy WJ, et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science 309:613–617.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324.

Nash WG, Menninger JC, Wienberg J, Padilla-Nash HM, O'Brien SJ. 2001. The pattern of phylogenomic evolution of the Canidae. Cytogenet Cell Genet. 95:210–224.

Neumann R, Jeffreys AJ. 2006. Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation. Hum Mol Genet. 15:1401–1411.

O'Brien SJ, et al. 1999. The promise of comparative genomics in mammals. Science 286:458–462, 479–481.

Otto SP, Lenormand T. 2002. Resolving the paradox of sex and recombination. Nat Rev Genet. 3:252–261.

Paigen K, et al. 2008. The recombinational anatomy of a mouse chromosome. PLoS Genet. 4:e1000119.

Petes TD. 2001. Meiotic recombination hot spots and cold spots. Nat Rev Genet. 2:360–369.

Petes TD, Merker JD. 2002. Context dependence of meiotic recombination hotspots in yeast: the relationship between recombination activity of a reporter construct and base composition. Genetics 162:2049–2052.

Petronczki M, Siomos MF, Nasmyth K. 2003. Un menage a quatre: the molecular biology of chromosome segregation in meiosis. Cell 112:423–440.

Poissant J, et al. 2010. Genetic linkage map of a wild genome: genomic structure, recombination and sexual dimorphism in bighorn sheep. BMC Genomics 11:524.

Ptak SE, et al. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet. 37:429–434.

Reed KM, Chaves LD, Hall MK, Knutson TP, Harry DE. 2005. A comparative genetic map of the turkey genome. Cytogenet Genome Res. 111:118–127.

Samollow PB. 2010. Marsupial genetics and genomics. Dordrecht, The Netherlands: Springer. p. 75–99.

Samollow PB, et al. 2007. A microsatellite-based, physically anchored linkage map for the gray, short-tailed opossum (Monodelphis domestica). Chromosome Res. 15:269–281.

Sharp PJ, Hayman DL. 1988. An examination of the role of chiasma frequency in the genetic system of marsupials. Heredity 60:77–85.

Shifman S, et al. 2006. A high-resolution single nucleotide polymorphism genetic map of the mouse genome. PLoS Biol. 4:e395.

Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci U S A. 48:582–592.

Vingborg RK, et al. 2009. A robust linkage map of the porcine autosomes based on gene-associated SNPs. BMC Genomics 10:134.

Walter RB, et al. 2004. A microsatellite genetic linkage map for Xiphophorus. Genetics 168:363–372.

Webster MT, Smith NG, Hultin-Rosenberg L, Arndt PF, Ellegren H. 2005. Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. Mol Biol Evol. 22:1468–1474.

Wienberg J. 2004. The evolution of eutherian chromosomes. Curr Opin Genet Dev. 14:657–666.

Williams E. 1959. The comparison of regression variables. J R Stat Soc Ser B (Methodological). 21:396–399.

Winckler W, et al. 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308:107–111.

Wong AK, et al. 2010. A comprehensive linkage map of the dog genome. Genetics 184:595–605.

Yamamoto K, et al. 2006. Construction of a single nucleotide polymorphism linkage map for the silkworm, Bombyx mori, based on bacterial artificial chromosome end sequences. Genetics 173:151–161.

Zenger KR, McKenzie LM, Cooper DW. 2002. The first comprehensive genetic linkage map of a marsupial: the tammar wallaby (Macropus eugenii). Genetics 162:321–330.

Associate editor: Laurence Hurst