MBE Advance Access published January 13, 2011

Chromosome Size in Diploid Eukaryotic Species Centers on the Average Length with a Conserved Boundary

Research Article (Open Accession)

Xianran Li,[1,*] Chengsong Zhu,[1,*] Zhongwei Lin,[1,*] Yun Wu,[1] Dabao Zhang,[2] Guihua Bai,[1,3] Weixing Song,[4] Jianxin Ma,[5] Gary J. Muehlbauer,[6] Michael J. Scanlon,[7] Min Zhang,[2,†] and Jianming Yu[1,†]

[1]Department of Agronomy, Kansas State University, Manhattan, KS 66506.

[2]Department of Statistics, Purdue University, West Lafayette, IN 47907.

[3]USDA-ARS, Manhattan, KS 66506.

[4]Department of Statistics, Kansas State University, Manhattan, KS 665067.

[5]Department of Agronomy, Purdue University, West Lafayette, IN 4790.

[6]Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108.

[7]Department of Plant Biology, Cornell University, Ithaca, NY 14853.

[*]These authors contributed equally to this work.

[†] To whom correspondence should be addressed. E-mail: jyu@ksu.edu or minzhang@purdue.edu

Jianming Yu

2004 Throckmorton Hall, Manhattan, KS 66506, Phone: 785-532-3397, Fax: 785-532-6094,

Emial: jyu@ksu.edu

Key words: Chromosome size, genome evolution, evolutionary modeling

Running head: Chromosome size variation

ABSTRACT

Understanding genome and chromosome evolution is important for understanding genetic inheritance and evolution. Universal events comprising DNA replication, transcription, repair, mobile genetic element transposition, chromosome rearrangements, mitosis, and meiosis underlie inheritance and variation of living organisms. Although the genome of a species as a whole is important, chromosomes are the basic unit subjected to genetic events that coin evolution to a large extent. Now as many complete genome sequences are available, we can address evolution and variation of individual chromosomes across species. For example, "*how are the repeat and nonrepeat proportions of genetic codes distributed among different chromosomes in a multi-chromosome species*?" "*Is there a general rule behind the intuitive observation that chromosome lengths tend to be similar in a species, and if so, can we generalize any findings in chromosome content and size across different taxonomic groups*?" Here we show that chromosomes within a species do not show dramatic fluctuation in their content of mobile genetic elements as the proliferation of these elements increases from unicellular eukaryotes to vertebrates. Furthermore, we demonstrate that, notwithstanding the remarkable plasticity, there is an upper limit to chromosome size variation in diploid eukaryotes with linear chromosomes. Strikingly, variation in chromosome size for 886 chromosomes in 68 eukaryotic genomes (including 22 human autosomes) can be viably captured by a single model, which predicts that vast majority of the chromosomes in a species are expected to have a basepair length between 0.4035 and 1.8626 times the average chromosome length. This conserved boundary of chromosome size variation, which prevails across a wide taxonomic range with few exceptions, indicates that cellular, molecular, and evolutionary mechanisms, possibly together, confine the chromosome lengths around a species-specific average chromosome length.

INTRODUCTION

Genome sequencing has revealed detailed information on the genetic content of genomes and chromosomes for more than a hundred species across different phyla. It is now not only possible to answer questions concerning metagenomics of environmental samples and the molecular and evolutionary basis of speciation, but also to ask many more questions in biology and evolution (Tringe and Rubin 2005; Misteli 2007; Metzker 2010; Presgraves 2010). Although the genome size of eukaryotes varies over five orders of magnitude, the distribution is skewed toward small values (Oliver et al. 2007). Overall genome size and complexity clearly have increased during evolution from archaea and bacteria to eukaryota (Lynch and Conery 2003), but the network of mechanisms of the many competing processes that either expand or shrink the genome remain to be discovered in detail (Lynch and Conery 2003; Whitney et al. 2010). Previous research based on estimated genome size across 20 eukaryotic clades found that variation of genome size within a clade increases with the average genome size of the clade (Oliver et al. 2007). Based on genome size values measured by flow cytometry, a recent study demonstrated that there is a significant correlation between genome size and meiotic recombination rate (Whitney et al. 2010). Given the relative abundance of completed genome sequences, we can address the evolutionary dynamics of genome size and variation of chromosome size across species with base pair numbers. In particular, detailed sequence information allows us to characterize features and variations of chromosomes across multiple species, which was not possible with previous overall genome size estimation. In this study, we specifically address the following major questions, "*how are the repeat and nonrepeat proportions of genetic codes distributed among different chromosomes in a multi-chromosome species*?" "*Is there a general rule behind the*

intuitive observation that chromosome lengths tend to be similar in a species, and if so, can we generalize any findings in chromosome content and size across different taxonomic groups?"

In eukaryota, DNA repeats increase chromosome size, as do intron size and gene duplication (Lynch and Conery 2003). Changes in chromosome number reflect the balance between forces that increase chromosome number (such as chromosome fission, chromosome mis-segregation, as well as allopolyploidization or autopolyplodization) and those that decrease it (such as chromosome fusion or mis-segregation). Some of these events also lead to changes in chromosome size. A systematic examination of repeat proportion at genome level and chromosome level across taxonomic groups should provide further insight into genome and chromosome evolution.

The transition from circular to linear chromosomes is one prerequisite for increases in individual chromosome size and chromosome number (Schubert 2007). In a seminal paper using field bean, it was demonstrated experimentally that there is an upper boundary of chromosome size for normal development of an organism (Schubert and Oud 1997). Sterility was mediated by chromosomes with arms exceedingly long via disturbance of meiotic division. This phenomenon was confirmed for barley, a monocot with a large genome (Hudakova et al. 2002). On the other hand, chromosomes of much smaller size than average frequently do not segregate correctly during meiosis (Schubert 2001; Murata, Shibata, and Yokota 2006). Taken together, experimental research in individual species suggested a limit of chromosome size variation, and a generalization of this finding to a wide range of species should provide insight regarding

genome and chromosome size evolution, mechanisms involved in mitosis and meiosis, and genetic stability of natural or artificial minichromosomes.

Many evolutionary alterations affect chromosome number and/or chromosome size including reciprocal translocations, deletions and insertions, unequal crossover, dispersion of repetitive sequences, genome duplication, and chromosome fusion and fission and mis-segregation (Schubert 2007). Among these factors, reciprocal translocations have been considered one of the major forces to shape chromosome size variation (Bickmore and Teague 2002; Schubert 2007) and were incorporated in previous evolutionary modeling studies (Sankoff and Ferretti 1996; De et al. 2001). These studies primarily considered individual species with specific numbers of chromosomes and the comparisons were made to chromosome size estimated from karyotpes.

Here we examined genome complexity by coupling information about evolutionary mechanisms and genome sequence information, thus revealing a general increase in genome size, chromosome size, and variability of chromosome characteristics from prokaryotes to unicellular eukaryotes, invertebrates, vascular plants, and vertebrates. Systematic analyses and computer simulations using genome sequence information from various species revealed that chromosome size expansion in the course of evolution follows a stochastic process constrained by an upper limit to chromosome size variation in many diploid eukaryotic genomes. Despite the dramatic differences in cellular and organismal complexity, the common pattern of chromosome size variation in different eukaryotic genomes suggests a conserved constraint to chromosome evolution.

MATERIALS AND METHODS

Genomes and chromosomes

Genome and chromosome data of 128 genomes (68 eukaryotes and 60 prokaryotes) with multiple chromosomes were obtained from different databases including GenBank, Ensembl, JGI, and Phytozome as well as individual species' genome databases (**Supplementary Tables 1 and 2**). Sequences unanchored to chromosomes were not included in tabulating the basepair length. For species with more than one strain sequenced, we randomly selected one strain to represent the species. Chromosome sizes within each species were listed in ascending order in basepair units. Common name groups were assigned using the literature and database information. Accession number or version of genome assembly was provided. The sex chromosomes of 14 species were excluded from the analysis because of their unique evolutionary processes (Charlesworth and Charlesworth 2005; Charlesworth, Charlesworth, and Marais 2005). For species without masked-ready genome sequence information, we identified the repetitive sequences with RepeatMasker 3.2.8 by using the library identified by RepeatScout 1.0.5 to mask the repetitive regions (Smit, Hubley, and Green verified on May 11, 2010). Because our focus was to obtain the general pattern of repeat proportion of the genomes and chromosomes rather than exact values for a certain species, we chose this more extensively used library-based program (Lerat 2010). Repeat and nonrepeat regions of chromosomes were obtained after the masking process.

The common theme of the current study was to examine genome size and chromosome size across different species. Variations of genome size increased as average genome size increased across different common name groups (i.e., prokaryotes, unicellular eukaryotes, invertebrates,

vascular plants, and vertebrates). For chromosome size in diploid eukaryotes, we further demonstrated that the standard deviation of chromosome size increased as average chromosome size increased and that a common coefficient of variation existed. Further model fitting and computer simulations revealed that common distribution of chromosome size variation can be modeled with a Gamma distribution (**Fig. 3**).

Data analysis and statistical modeling

Data of genome size and chromosome size were analyzed with SAS and R following standard procedures of correlation, regression, and plotting (**Fig. 1; Supplementary Fig. 1 and 2**). Because circular chromosomes in prokaryotes have different mechanisms for replication and separation in cell cycles (Schubert 2007), we focused only on eukaryotes with linear chromosomes. We used two approaches to conduct statistical modeling of chromosome size variation. In the first approach, we fit an intuitive cubic function to capture the relationship between chromosome size and chromosome index. Chromosome size was calculated as the ratio of basepair length of a chromosome to average basepair length of chromosome of the species, $Z_{i(j)} = L_{i(j)} / \bar{L}_i$, where $L_{i(j)}$ is the basepair chromosome length for the *j*-th chromosome of a species *i;* $\bar{L}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} L_{i(j)}$; n_i is the total chromosome number; and *i* = 1, 2, …, *n* species. Chromosome index was calculated as $(\frac{j-0.5}{n_i})$. The fitted function was

$$\hat{Z}_{i(j)} = 0.3920 + 2.2890(\frac{j-0.5}{n_i}) - 3.9141(\frac{j-0.5}{n_i})^2 + 3.0753(\frac{j-0.5}{n_i})^3$$

where $\hat{Z}_{i(j)}$ is the predicted chromosome size for the *j*-th chromosome of a species *i* and n_i is the total chromosome number. Subtracting 0.5 in chromosome index was justified because we used a continuous distribution to model the discrete chromosome number; this is a standard practice.

The second approach was more systematic and aimed to model chromosome size variation from statistical distributions. We used iteratively reweighted least square method to derive the parameter estimate. Four distributions commonly used in biology were considered: Gamma distribution, Normal distribution, Truncated Normal distribution (truncation at zero), and Lognormal distribution. Gamma distribution was chosen for four reasons. First, $Z_{i(j)}$ were all non-negative. Second, the histogram of $Z_{i(j)}$ was skewed right and can be modeled by a Gamma distribution. Third, unlike Lognormal distribution, Gamma distribution is a member of the exponential family and permits a generalized linear model (Schabenberger and Pierce 2002). Fourth, model fitting showed that Gamma distribution had the best model fit. Model fitting statistics were calculated for mean square error (MSE), R^2, and Akaike's information criterion (AIC) (**Supplementary Table 3**). $MSE = \sum_{k=1}^{n}(Z_k - \hat{Z}_k)^2/(n-p)$, where Z_k is the k-th observed data point; $\hat{Z}_k$ is the predicted value; k = 1, … , n; n = 886 chromosomes, and p is the number of parameters in the model. The original definition of R^2 was used, that is, $R^2 = 1 - \frac{SSE}{SST}$, where $SSE = \sum_{k=1}^{n}(Z_k - \hat{Z}_k)^2$, $SST = \sum_{k=1}^{n}(Z_k - 1.0)^2$, and $AIC = n\ln(SSE) - n\ln(n) + 2p$.

Although it is not possible to prove statistically that chromosome size must follow a Gamma distribution, our analysis proved that Gamma distribution was the best candidate of the distributions examined. We present the modeling steps for the Gamma distribution in supplementary materials; similar steps were derived for three other distributions.

For cross validation, the observed data were randomly split into two parts: model fitting and validation (**Supplementary Fig. 4**). We then conducted computer simulations to further prove that Gamma distribution viably describes chromosome size and that numbers drawn from the Gamma distribution with the identified parameter *Gamma* (7.0438, 1/7.0438) can reproduce the pattern from observed data (**Supplementary Fig. 5**). Details for these two sections were provided in supplementary materials.

Reciprocal translocation

Among many evolutionary events, reciprocal translocation is a good starting point for understanding the dynamics of chromosome size variation through modeling (Sankoff and Ferretti 1996; De et al. 2001; Imai, Satta, and Takahata 2001; Mazowita, Haque, and Sankoff 2006). Simulations tested whether reciprocal translocation is partly responsible for observed chromosome size variation. Numbers obtained through simulation (see supplementary materials for details) were then plotted against the chromosome index to show whether the resulting line approximates the predicted line from the inverse of the Gamma cumulative distribution function (**Fig. 4**).

Four simulation schemes were carried out: (1) no constraints on chromosome size, (2) a lower threshold, (3) an upper threshold, and (4) both lower and upper thresholds (Sankoff and Ferretti 1996; De et al. 2001; Imai, Satta, and Takahata 2001; Mazowita, Haque, and Sankoff 2006). We incorporated constraints on the smallest and largest chromosomes in the modeling process because (1) chromosome size below a certain threshold will prevent any translocation events; (2) at the cytogenetic level, viable and functional chromosomes must contain at least a centromere

and two telomeres to maintain purely structural basis; and (3) each chromosome must have a length sufficient for at least one crossover among the four aligned sister chromatids in meiosis. Moreover, as shown experimentally, if one arm of the chromosome is more than 21.7% of the total length of all chromosomes, most offspring are sterile (Schubert 2007). The lower threshold was set for the smallest observed chromosome size (Sankoff and Ferretti 1996), and the upper threshold was set using a fitness function (De et al. 2001). In addition, we implemented a constraint in all simulation that resulting chromosomes from reciprocal translocation must have a centromere (De et al. 2001).

Details for reciprocal translocation simulation, confirming outlier species with known reasons, and estimating genome sizes for a much large samples of vascular plants and vertebrates were given in supplementary materials.

RESULTS

Is average genome size of a taxonomic group related to variation within that group?

We collected information on genome size, chromosome number, individual chromosome size, repeat-masked chromosome size (without repeat proportion), and common name groupings for 128 species with sequenced genomes, including prokaryotes, unicellular eukaryotes, invertebrates, vascular plants, and vertebrates. (**Supplementary Tables 1 and 2**). Across all sequenced prokaryotic and diploid eukaryotic species, genome size correlated with chromosome number and average chromosome size. Genome size varied considerably among species with similar levels of cellular and organismal complexity, but there was a general increase in genome size from prokaryotes to unicellular eukaryotes to multicellular eukaryotes (**Fig. 1**). In addition,

continuities in the scale of genome size across different groups of organisms indicate that organismal differences in cell/tissue anatomical structure or metabolism are unlikely to be the primary forces driving the evolution of genomic architecture (Lynch and Conery 2003).

Using these basepair data for genome size, we tested whether variation in genome size within each group was proportional to average genome size of the group. Given the sample size of available genomes, we focused our analysis on five phylogenetic branches (*i.e.*, prokaryotes, unicellular eukaryotes, invertebrates, vascular plants, and vertebrates) rather than other finer taxonomic levels. Clearly, variation in genome size (measured as standard deviation) significantly correlated with the average genome size (**Fig. 1**). After we removed the dependency with Log_{10} transformation (a method to break the association between average of a group of numbers and the variation of these numbers) (Oliver et al. 2007), the variation within each group showed no correlation with the average genome size. Groups with a larger average genome size obviously also had a larger variation in genome size. Variation of genome size of each group is the numerator in the calculation of rate of genome size evolution and could provide an approximation if the denominator, evolutionary distance or time, does not differ across groups on the same order of magnitude as the numerator. Interestingly, our findings regarding genome size showed a similar pattern with the previous research in which rate of genome size evolution was found to be proportional to average genome size of a clade when estimated genome size based on C-value was examined across 20 eukaryotic clades and evolutionary distance was obtained from phylogenetic analysis of 18S rDNA (Oliver et al. 2007).

How are the repeat and nonrepeat proportions of genetic codes distributed among different chromosomes in a multi-chromosome species?

To further examine the role of repeats on genome size and chromosome size, repeat masking of the genome was obtained from either original publications of the sequenced genomes or repeat masking analysis (Lerat 2010; Smit, Hubley, and Green verified on May 11, 2010). In general, the repeat proportion of the genome increased from prokaryotes (mean: 0.04) to unicellular eukaryotes (0.08), invertebrates (0.14), vascular plants (0.35), and vertebrates (0.38), following the same trend as genome size (**Fig. 1**). For vascular plants with complete genome sequence, the repeat proportion of maize (82.5%) and sorghum (60.9%) skewed distribution to the right side. Overall, repeat proportion of chromosomes increases during evolution from prokaryotes to vertebrates and this trend may become more evident as large genomes of vascular plants and vertebrates are sequenced.

Following the similar logic in genome size analysis, we also tested whether the standard deviation of chromosome size (in basepair) within each species was proportional to mean of chromosome size. Because of the difference in response to repeat accumulation between circular and linear chromosomes, we considered only eukaryotes with linear chromosomes in this analysis. There was a significant positive correlation between standard deviation of chromosome size and average chromosome size of a species (**Fig. 2**). After we removed the magnitude effects with Log_{10} transformation, however, the standard deviation of chromosome size for all eukaryotic species was bounded in a much smaller region than that for the prokaryotic species. Because 68 diploid eukaryotic species were used and the signal of the relationship between standard deviation and average chromosome size was strong ($P = 1.3\times10^{-38}$), we then derived the

regression slope (0.3700) of standard deviation on average chromosome size across species. This regression slope provided an *ad hoc* estimate of a common coefficient of variation (CV = standard deviation/mean) for the underlying distributions of chromosome sizes in different species. Although large differences existed for average chromosome size and standard deviation of chromosome size across species, the proportional relationship between them approached a constant. This was further verified by plotting coefficient of variation, and any deviation was not unexpected because individual CV calculated for each species represented a sample (**Supplementary Fig. 1**). On the other hand, there was no significant correlation between variation of chromosome size and total chromosome number of a species (**Supplementary Fig. 1**).

Similar to the findings for chromosome size, the standard deviation of non-repeat size was proportional to average non-repeat size, and the standard deviation of repeat size proportional to average repeat size. Although the mechanisms by which non-repeat and repeat sequences were expanded in eukaryotic genomes is complicated (Lerat 2010), our results suggest that rate of expansion among chromosomes is proportional to the preceding chromosome size, which indicates a stochastic process (**Fig. 2**). Previous estimations of repeat proportions of the genomes have been species-specific or based on extrapolation from a smaller number of species (Lynch and Conery 2003; Lerat 2010) than included in the current study. Our general approach to studying repeat evolution across species with genome sequence data lays the groundwork for detailed studies on evolution of different classes of repeats and their composition among chromosomes, genomes, and taxonomic groups.

Is there a general rule behind the intuitive observation that chromosome lengths tend to be similar in a species?

We next examined chromosome size variation in eukaryotes in detail because data available on chromosome length across the sequenced genomes permitted systematic modeling of chromosome size (**Supplementary Fig. 2**). In addition to the common CV of chromosome size in eukaryotes, we noted that basepair sizes of the chromosomes within individual species usually have the same order of magnitude; this inspired further investigation of chromosome size variation. Two transformations made the modeling process statistically possible and biologically sound: relative chromosome size and chromosome index. Relative chromosome size is obtained by dividing chromosome size in basepair by the average chromosome size of the individual species. Using average chromosome size as the unit of measure standardized the original chromosome size (in basepair) in different orders of magnitude for different species into comparable numbers. Chromosome index is obtained by dividing the ascending ranked chromosome number (subtracting a continuity correction factor 0.5) by the total chromosome number of that particular species. For example, for a species with two chromosomes, instead of 1 and 2, the chromosome index becomes 0.25 and 0.75. For a species with five chromosomes, instead of 1 through 5, the chromosome index becomes 0.1, 0.3, 0.5, 0.7, and 0.9. Chromosome index is bounded between 0 and 1, which permits modeling of chromosome size across species with different chromosome numbers. Amazingly, the plot of chromosome size against chromosome index revealed a clear pattern and strongly suggested a common curve similar to a cubic function: The incremental change in chromosome size larger at both ends of the curve but smaller in the middle (**Fig. 3**).

Further investigation into the potential distribution from which the chromosome sizes (samples) were drawn suggested that a Gamma distribution was a more plausible candidate than other distributions (**Fig. 3**). Gamma distribution is widely used in engineering and science to model continuous variables that are non-negative but have right-skewed probability densities (Schabenberger and Pierce 2002) and provides a natural framework to model chromosome size that is non-negative. Indeed, a Gamma distribution approximated a histogram of all chromosomes sizes (with a mean of 1 and skewness of 1.0046) better than a Normal distribution. Histograms generated from data of individual species, from the pooled data of species with the same total number of chromosomes, and from the pooled data of each common group corroborated this finding. We then theoretically derived the approximate relationship function between chromosome size and chromosome index as an inverse of a Gamma cumulative distribution function, $G^{-1}_{(\alpha,1/\alpha)}$, where α is the parameter. Because no closed form exists for this nonlinear function, we used an iterative procedure (iteratively reweighted least square) that minimizes the influence of variance heterogeneity to obtain the parameter estimate $G^{-1}_{(7.0438,1/7.0438)}$ with a 95% confidence interval of $\hat{\alpha}$ as (6.6609, 7.4267). Model fitting statistics indicated a better fit with the Gamma distribution than with other distributions or the intuitive cubic function. Notice that the variance (and CV because mean=1) of $G^{-1}_{7.0438}$ is 0.3768, which is close to the previous *ad hoc* CV estimate 0.3700 obtained through simple regression analysis. On the basis of $G^{-1}_{(7.0438,1/7.0438)}$, 95% of the chromosomes in a species are expected to have a basepair length between 0.4035 and 1.8626 times the average chromosome length; this interval is applicable to chromosomes in diploid eukaryotic species. However, we admit that practically a Normal distribution is almost equally viable in capturing the chromosome size variation (**Fig. 3**

and **Supplementary Table 3**) and is a more general one. The major reason of not choosing Normal distribution is the possible negative values implicated.

Can prediction be made on chromosome size?

It follows that, for a given species, chromosome sizes can be predicted by chromosome number. Furthermore, given either genome size or average chromosome basepair length (genome size = average chromosome size × total chromosome number), we can predict the size range of all chromosomes of that species in basepair (**Fig. 3**). Chromosome size proportion was obtained by dividing chromosome size by genome size; the sum of chromosome size proportions equaled one. For example, for a species with 15 chromosomes, the shortest and longest chromosomes would be expected to account for 2.87% and 11.99%, respectively, of the genome. The predicted ratio of the longest to the shortest chromosome for a given species was 1.68 for a species with two chromosomes and 5.70 for a species with 38 chromosomes. We used this general prediction to confirm the cases in which exceptions occurred for a few outlier species for known reasons: three species known to have macrochromosomes and microchromosomes, one haploid species, and one species with one linear chromosome and one circular chromosome (**Supplementary Tables 1 and 2, Supplementary Fig. 3**).

To show the robustness of the prediction and ensure that we had used an adequate number of genomes (68 diploid eukaryotic genomes), we performed a series of cross-validation experiments using different proportions of the observed data for function derivation and the rest of the data for validation. Plots of mean square prediction error (MSPE) and parameter estimate indicated the original sample size was large enough to derive a robust prediction function (**Supplementary**

Fig. 4). The MSPE decreased as more data points were used to derive the prediction function. Likewise, the parameter estimate (α) approached the value from the whole data set. With about 50% of the data ($\approx$35 species), both MSPE and α started to level off, indicating an adequate sample size in the original data to derive the function and make prediction. In addition, simulation results reproduced the pattern of the observed data, indicating that Gamma distribution viably describes the chromosome size variation observed (**Supplementary Fig. 5**). Numbers representing chromosome sizes were drawn from Gamma distributions with specific parameters for species having a chromosome number from 2 to 38. Both the dispersion of the scattered points and the fitted curves of the simulated and observed data confirmed that the pattern discovered was reproducible.

Should other evolutionary alterations besides reciprocal translocation be considered in evolutionary modeling studies?

To verify whether reciprocal translocations can adequately model the chromosome size variation as suggested in previous evolutionary modeling studies (Sankoff and Ferretti 1996; De et al. 2001; Imai, Satta, and Takahata 2001; Mazowita, Haque, and Sankoff 2006), we ran a set of computer simulations to compare the pattern generated by simulations and by our empirical data. Four simulation schemes were carried out: (1) no constraints on chromosome size, (2) a lower threshold, (3) an upper threshold, and (4) both lower and upper thresholds (Sankoff and Ferretti 1996; De et al. 2001; Imai, Satta, and Takahata 2001; Mazowita, Haque, and Sankoff 2006). Notice that these thresholds are for individual chromosome size, not their variations. Simulated chromosome sizes based on the reciprocal translocation model without thresholds showed greater variation than we observed in these sequenced genomes, but simulations with both

thresholds had a better approximation (**Fig. 4, Supplementary Fig. 6**). Our results suggest that reciprocal translocation is likely to one of the major forces and future modeling procedures that consider other evolutionary alterations (e.g. genome duplications, chromosome fusion, secondary rearrangements) besides reciprocal translocation may lead to even better congruency (Schubert 2007; The Chimpanzee Sequencing and Analysis Consortium 2005). Unlike previous studies in which modeling was conducted for individual species and much smaller numbers of species were examined, the current study with empirical data analyses and computer simulations established a benchmark for future evolutionary modeling research in chromosome size.

DISCUSSION

Genome and chromosome complexity have been addressed from different perspectives including population genetics and evolution (Lynch and Conery 2003; Oliver et al. 2007), molecular biology and cytogenetics (Schubert 2007), and evolutionary modeling (Sankoff and Ferretti 1996; Ma et al. 2008). In this work, we systematically studied the dynamics of genome and chromosome size variation. Using a combination of bioinformatics and statistics approaches and available genome sequences across the evolutionary spectrum, we examined genome size evolution, repeat size evolution, chromosome size variation, and evolutionary modeling. Chromosome size tends to center around the average chromosome length within a species for most diploid eukaryotes and chromosome size variation across species can be adequately modeled with a Gamma distribution. Although it may seem to be intuitive or a common place, systematic proof across multiple species is lacking prior to our study. Our findings are in agreement with the long-standing karyotypes in which chromosomes are usually visualized in descending order (Sankoff and Ferretti 1996). This connection assumes that the higher-order

structures of linear DNA sequence do not lead to a different pattern of chromatin size (as captured in karyotype) from the chromosome size in basepair (Misteli 2007). In other words, a relatively constant folding ratio ensures that higher basepair length generally corresponds to longer chromatin size. In a cell cycle, the synchrony of chromosome separation must be precisely controlled to correctly separate homologous chromosomes or sister chromatids. Although the exact mechanism of such synchrony is not clear, chromosome size variation as a basic feature of chromosome architecture deserves more attention. Uniform chromosome length may facilitate the cell achieving synchronized DNA replication time with the same number of replication forks, correct chromosome configuration on equatorial plate, and accurate migration of homologous chromosomes or sister chromatids to opposite poles (Sharp, Rogers, and Scholey 2000; Misteli 2007).

In the current modeling of chromosome size variation across 68 eukaryotic species, species with different genome sizes were examined, e.g. *Bigelowiella natans* with 0.37Mb, *Zea mays* with 2.05Gb, *Homo sapiens* with 2.88 Gb for autosomes, and *Monodelphis domestica* 3.42 Gb for autosomes. In addition, resampling simulations demonstrated that the major finding in chromosome size variation based on available data is robust to sampling process. We realized that genome sequences of some vascular plants and vertebrates with very large genome sizes are not available (Whitney et al. 2010). However, with estimated genome sizes from C values of a much larger number of species in vascular plants (2757) and vertebrates (3140), the rate of genome size evolution as measured by standard deviation of genome size within each group remains to be positively correlated with average genome size (**Supplementary Fig. 7)**. The boundary discovered for chromosome size variation, on the other hand, is less likely to be biased

because the context is individual genomes. For example, karyotypes of wheat genome (~16 Gb) (Gill, Friebe, and Endo 1991; Sankoff and Ferretti 1996) and barley genome (~5 Gb) (Lee et al. 2000) strongly suggest a boundary in chromosome size variation for these two large genomes with high proportion of repeats, same as discovered in the current study. Taking the general strategies of this cross-species analysis, evidence supporting the current discovery is likely to be further uncovered with more genomes being sequenced. On the other hand, it would be interesting to study the mechanisms of genome and chromosome stability with a few outlier species with known reasons shown in our study.

An upper limit to chromosome size variation provides better evolutionary fitness because the limit of the cell dimension and spindle extension do not favor having chromosomes with significantly different length (Schubert and Oud 1997; Schubert 2001; Schubert 2007). Considering the number of cells and the mitosis events in an organism, the overall energy savings may also be a factor because ATP molecules are required for chromosome velocity (Nicklas 1965). Temporal control of kinetochore–microtubule dynamics may be a mechanism for maintaining genome stability (Bakhoum, Genovese, and Compton 2009; Bakhoum et al. 2009). Depolymerization of kinetochore microtubules may partly power chromosome movement during mitosis (Molodtsov et al. 2005). Under normal conditions, chromosomes of different sizes in a single cell have similar chromosome velocity in anaphase (Nicklas 1965; Raj and Peskin 2006). Large variations in chromosome length may decrease the evolutionary fitness of an organism; overly lengthy chromosomes will delay the separation of sister chromatids and homologous chromosomes during mitosis and meiosis, resulting in cell cycle prolongation, sterility, or even death (Schubert 2007). Moreover, meiotic recombination was experimentally demonstrated to

depend on chromosome size in *Saccharomyces cerevisiae* (Kaback et al. 1992) and in human (Lander et al. 2001). Therefore, chromosome size variation is a vital factor in cell biology and evolution.

Genome sequences of neopolyploid species have not been reported. After resolving the assembly hurdle, further sequencing of polyploid genomes would allow us to extend this hypothesis beyond diploid genomes. Many current diploid species have undergone a process of polyploidization and diploidization. Detailed examination of available genomes may also reveal the evolutionary significance of ancient genome duplications (Van de Peer, Maere, and Meyer 2009). In addition, the locations of centromeres have been studied in only a few species (Henikoff, Ahmad, and Malik 2001). It is interesting that although chromosome segregation machinery is highly conserved across all eukaryotes, research about DNA and protein components at centromeric chromatin has not been able to readily identify of centromeres in non-model species. Once the positions of centromeres have been identified in a wide range of species, further study of length variation of the chromosome arm may allow us to understand both the fine control and variation in chromosome segregation machinery.

SUPPLEMENTARY MATERIALS

Supplementary Materials, Supplementary figures S1 to S7, and supplementary tables S1 to S3, are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

ACKNOWLEDGEMENTS

We are very grateful to Dr. Ingo Schubert and Dr. Brandon Gaut for their critical comments of the manuscript. This work is supported by the National Science Foundation (DBI-0820610; IIS-0844945), the National Institute of Health (NIH/NCI U01-CA128535-01), the Department of Defense (W81XWH-08-1-0065), the Purdue University Discovery Park Seed Grant, the National Research Initiative of the USDA-CSREES (2006-03578), and the Targeted Excellence Program of Kansas State University.

LITERATURE CITED

The Chimpanzee Sequencing and Analysis Consortium 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**:69-87.

Bakhoum, S. F., G. Genovese, and D. A. Compton. 2009. Deviant kinetochore microtubule dynamics underlie chromosomal instability. Curr Biol **19**:1937-1942.

Bakhoum, S. F., S. L. Thompson, A. L. Manning, and D. A. Compton. 2009. Genome stability is ensured by temporal control of kinetochore-microtubule dynamics. Nat Cell Biol **11**:27-35.

Bickmore, W. A., and P. Teague. 2002. Influences of chromosome size, gene density and nuclear position on the frequency of constitutional translocations in the human population. Chromosome Res **10**:707-715.

Charlesworth, D., and B. Charlesworth. 2005. Sex chromosomes: evolution of the weird and wonderful. Curr Biol **15**:R129-131.

Charlesworth, D., B. Charlesworth, and G. Marais. 2005. Steps in the evolution of heteromorphic sex chromosomes. Heredity **95**:118-128.

De, A., M. Ferguson, S. Sindi, and R. Durrett. 2001. The equilibrium distribution for a generalized Sankoff-Ferretti model accurately predicts chromosome size distribution in a wide variety of species. J. Appl. Prob. **38**:324-334.

Gill, B. S., B. Friebe, and T. R. Endo. 1991. Standard karyotype and nomenclature system for description of chromosome bands and structural aberrations in wheat (*Triticum Aestivum*). Genome **34**:830-839.

Henikoff, S., K. Ahmad, and H. S. Malik. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. Science **293**:1098-1102.

Hudakova, S., G. Kunzel, T. R. Endo, and I. Schubert. 2002. Barley chromosome arms longer than half of the spindle axis interfere with nuclear divisions. Cytogenet Genome Res **98**:101-107.

Imai, H. T., Y. Satta, and N. Takahata. 2001. Integrative study on chromosome evolution of mammals, ants and wasps based on the minimum interaction theory. J Theor Biol **210**:475-497.

Kaback, D. B., V. Guacci, D. Barber, and J. W. Mahon. 1992. Chromosome size-dependent control of meiotic recombination. Science **256**:228-232.

Lander, E. S.L. M. LintonB. BirrenC. et al. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860-921.

Lee, J. H., K. Arumuganathan, Y. S. Chung, K. Y. Kim, W. B. Chung, K. S. Bae, D. H. Kim, D. S. Chung, and O. C. Kwon. 2000. Flow cytometric analysis and chromosome sorting of barley (*Hordeum vulgare* L.). Molecules and Cells **10**:619-625.

Lerat, E. 2010. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity **104**:520-533.

Lynch, M., and J. S. Conery. 2003. The origins of genome complexity. Science **302**:1401-1404.

Ma, J., A. Ratan, B. J. Raney, B. B. Suh, W. Miller, and D. Haussler. 2008. The infinite sites model of genome evolution. Proc Natl Acad Sci U S A **105**:14254-14261.

Mazowita, M., L. Haque, and D. Sankoff. 2006. Stability of rearrangement measures in the comparison of genome sequences. J Comput Biol **13**:554-566.

Metzker, M. L. 2010. Sequencing technologies - the next generation. Nat Rev Genet **11**:31-46.

Misteli, T. 2007. Beyond the sequence: cellular organization of genome function. Cell **128**:787-800.

Molodtsov, M. I., E. L. Grishchuk, A. K. Efremov, J. R. McIntosh, and F. I. Ataullakhanov. 2005. Force production by depolymerizing microtubules: a theoretical study. Proc Natl Acad Sci U S A **102**:4353-4358.

Murata, M., F. Shibata, and E. Yokota. 2006. The origin, meiotic behavior, and transmission of a novel minichromosome in *Arabidopsis thaliana*. Chromosoma **115**:311-319.

Nicklas, R. B. 1965. Chromosome velocity during mitosis as a function of chromosome size and position. J Cell Biol **25**:SUPPL:119-135.

Oliver, M. J., D. Petrov, D. Ackerly, P. Falkowski, and O. M. Schofield. 2007. The mode and tempo of genome size evolution in eukaryotes. Genome Res **17**:594-601.

Presgraves, D. C. 2010. The molecular evolutionary basis of species formation. Nat Rev Genet **11**:175-180.

Raj, A., and C. S. Peskin. 2006. The influence of chromosome flexibility on chromosome transport during anaphase A. Proc Natl Acad Sci U S A **103**:5349-5354.

Sankoff, D., and V. Ferretti. 1996. Karyotype distributions in a stochastic model of reciprocal translocation. Genome Res **6**:1-9.

Schabenberger, O., and F. J. Pierce. 2002. Contemporary Statistical Models for the Plant and Soil Sciences. CRC Press, Boca Raton, FL.

Schubert, I. 2001. Alteration of chromosome numbers by generation of minichromosomes -- is there a lower limit of chromosome size for stable segregation? Cytogenet Cell Genet **93**:175-181.

Schubert, I. 2007. Chromosome evolution. Curr Opin Plant Biol **10**:109-115.

Schubert, I., and J. L. Oud. 1997. There is an upper limit of chromosome size for normal development of an organism. Cell **88**:515-520.

Sharp, D. J., G. C. Rogers, and J. M. Scholey. 2000. Microtubule motors in mitosis. Nature **407**:41-47.

Smit, A. F. A., R. Hubley, and P. Green. verified on May 11, 2010. RepeatMasker Open-3.0. . http://www.repeatmasker.org.

Tringe, S. G., and E. M. Rubin. 2005. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet **6**:805-814.

Van de Peer, Y., S. Maere, and A. Meyer. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet **10**:725-732.

Whitney, K. D., E. J. Baack, J. L. Hamrick, M. J. Godt, B. C. Barringer, M. D. Bennett, C. G. Eckert, C. Goodwillie, S. Kalisz, I. J. Leitch, and J. Ross-Ibarra. 2010. A role for nonadaptive processes in plant genome size evolution? Evolution **64**:2097-2109.

FIGURE LEGENDS

Figure 1. (**A**) Genome size in Mb of sequenced prokaryotes, unicellular eukaryotes, invertebrates, vascular plants, and vertebrates. (**B**) Boxplot of genome size in Log_{10} scale. The *F* test for genome size in Log_{10} scale among groups is highly significant ($P = 2.3\times10^{-57}$), and all pairwise group comparisons are significant. (**C**) The standard deviation (SD) of genome size within each group positively correlates with genome size ($r = 0.92$; $P = 0.025$). Values are in Log_{10} scale for plotting. (**D**) After the dependency of SD on genome size is removed with Log_{10} transformation, the standard deviation of genome size within the groups shows no correlation ($r = -0.05$; $P = 0.93$) with genome size. **(E)** Boxplot of the repeat proportions of genomes. The overall *F* test for repeat proportions among groups is highly significant ($P = 3.0\times10^{-26}$), and all pairwise group comparisons are significant except prokaryotes-unicellular eukaryotes and vascular plants-vertebrates.

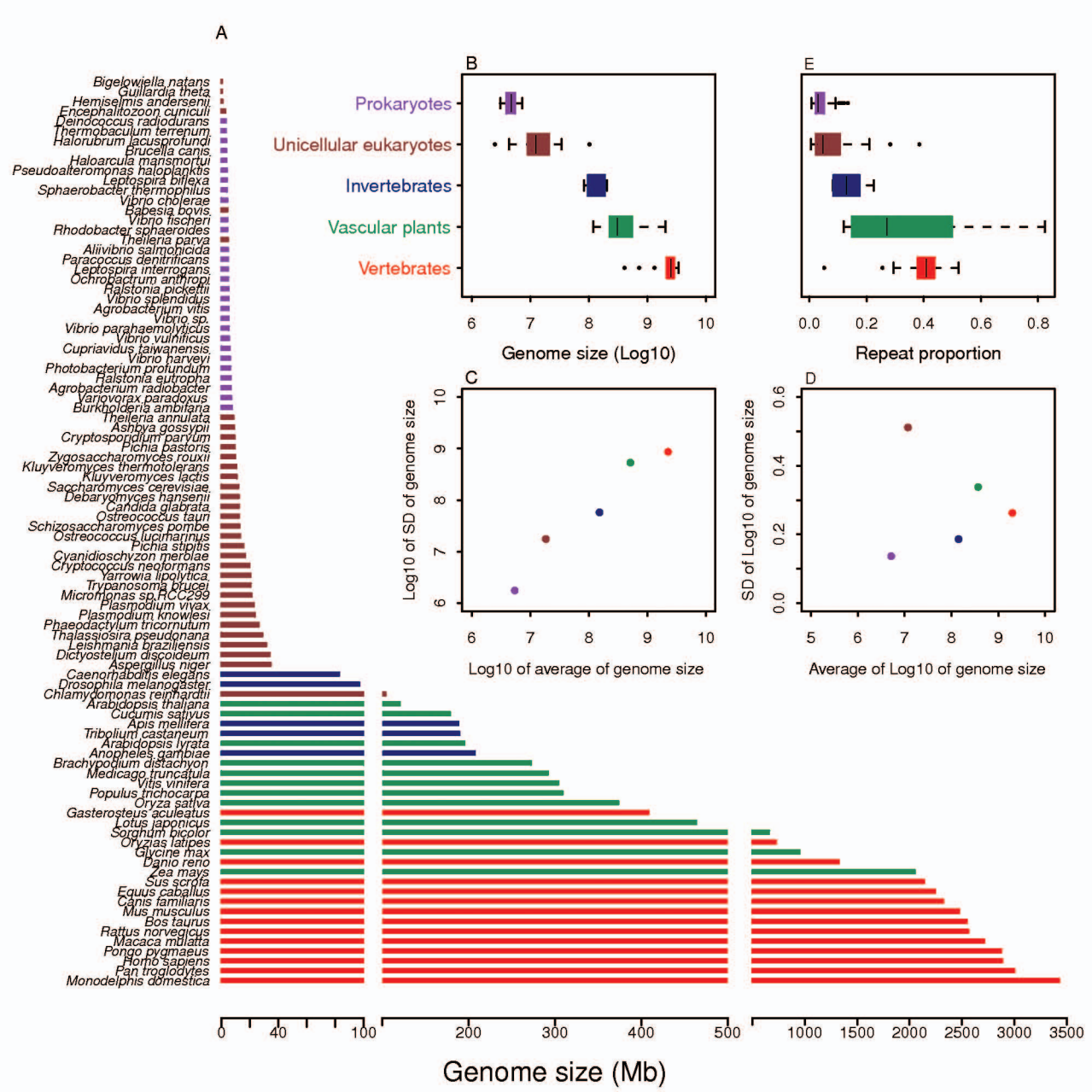
Figure 2. (**A**) Chromosome size variation as measured by standard deviation of chromosome size within species correlates positively with average chromosome size ($r = 0.96$, $P = 1.3\times10^{-38}$). Values are in Log_{10} scale for plotting. Estimate of a common coefficient of variation in original scale is 0.3700. (**B**) Absolute nonrepeat size variation ($r = 0.97$, $P = 5.8\times10^{-40}$). (**C**) Absolute repeat size variation ($r = 0.94$, $P = 4.8\times10^{-31}$). (**D**) After the dependency of absolute chromosome size variation on preceding chromosome size is removed with Log_{10} transformation, chromosome size variation within species shows no correlation ($r = -0.10$, $P = 0.43$) with average chromosome size. (**E**) Prior Log_{10} transformed nonrepeat size variation ($r = -0.11$, $P = 0.37$). (**F**) Prior Log_{10} transformed repeat size variation ($r = -0.02$; $P = 0.89$). Prokaryotic chromosomes are

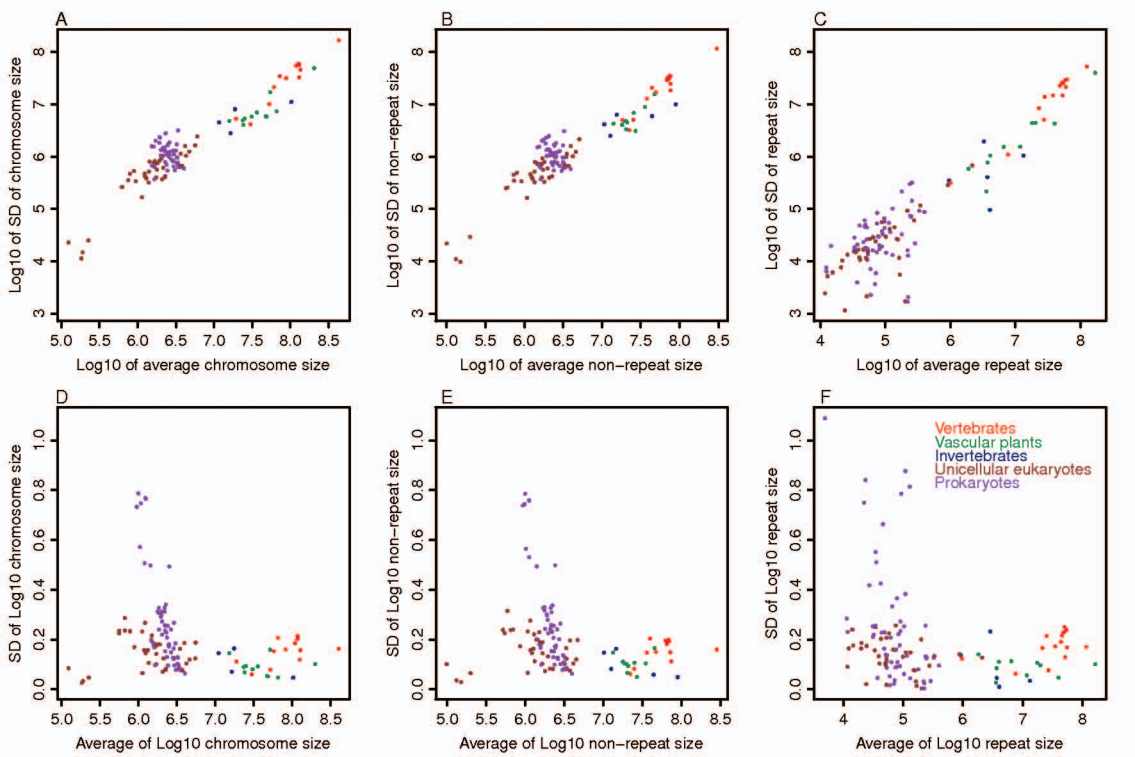
not included in the correlation calculation. Each color-coded dot represents the value for individual species.

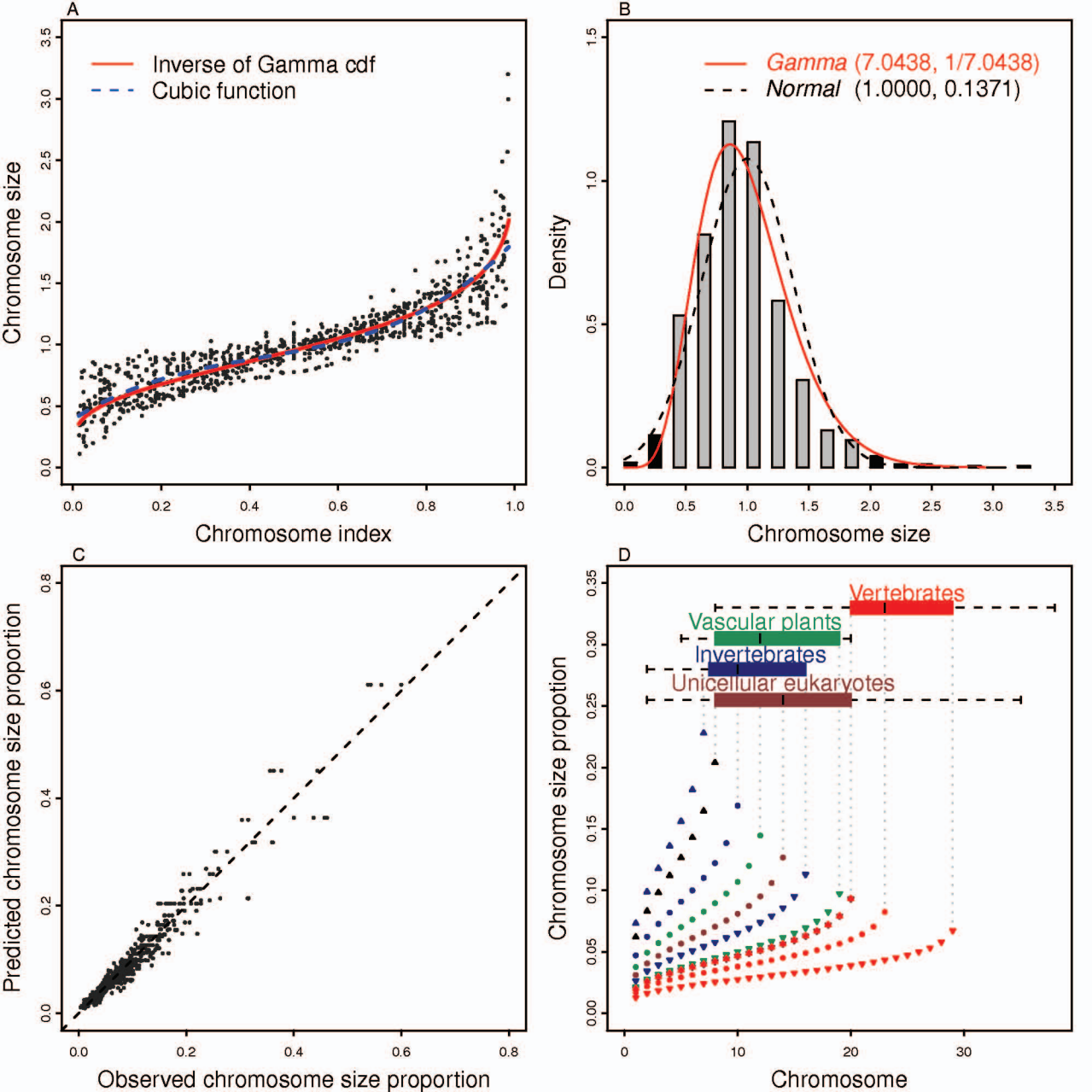
Figure 3. (**A**) Model fitting of chromosome size on chromosome index across 886 chromosomes from 68 diploid eukaryotic species. The blue dotted line is the fitted cubic function, and the red line is the fitted inverse of Gamma cumulative distribution function $\hat{Z}_{i(j)} = G_{\hat{\alpha}}^{-1}(\frac{j-0.5}{n_i}) / \hat{\alpha} = G_{7.0438}^{-1}(\frac{j-0.5}{n_i}) / 7.0438$, where $\hat{Z}_{i(j)}$ is the predicted chromosome size for the *j*-th ordered chromosome of a species *i* with a total of n_i chromosomes, and $G_{\hat{\alpha}}^{-1}$ is the inverse of Gamma cumulative distribution function with parameter $\hat{\alpha}$. (**B**) Histogram of chromosome size distribution with the overlaid probability density functions of *Gamma* (7.0438, 1/7.0438) and *Normal* (1.0000, 0.1371). The histogram has a mean of 1.0 and a skewness of 1.0046. Gray bars represent approximately 95% of the chromosome size between 0.3851 and 1.8608, and black bars represent the remaining 5% on both ends. *Gamma* (7.0438, 1/7.0438) has a means of 1.0 and a variance of 0.1420. Of the chromosome size from *Gamma* (7.0438, 1/7.0438), 95% lies between 0.4035 and 1.8626. (**C**) Predicted chromosome size proportion versus observed chromosome size proportion. (**D**) Predicted chromosome size proportion for a species with a given number of chromosomes. Predictions are plotted for the low hinge, median, and high hinge of the boxplot of individual common name groups: unicellular eukaryotes, invertebrates, vascular plants, and vertebrates.

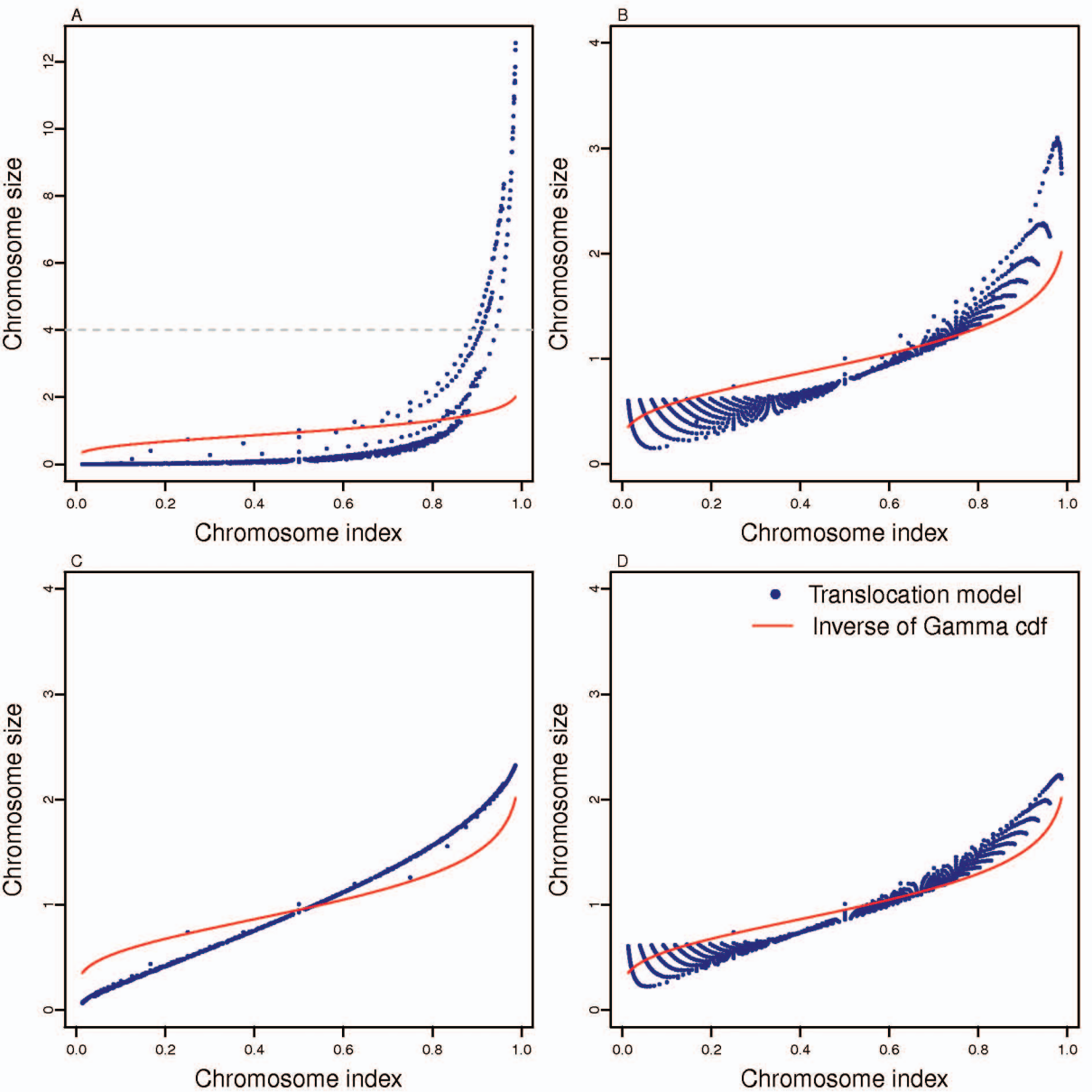
Figure 4. Simulation using the reciprocal translocation model to test whether it partly explains observed (red line) chromosome size variations. (**A**) No constraints on chromosome size. (**B**) A lower threshold. (**C**) An upper threshold. (**D**) Both lower and upper thresholds. Chromosome size

values are not expected to form a single line because the reciprocal translocation model predicts chromosome sizes independently for different total number of chromosomes.


A
Bigelowiella natans
Guillardia theta
Hemiselmis andersenii
Encephalitozoon cuniculi
Deinococcus radiodurans
Thermobaculum terrenum
Halorubrum lacusprofundi
Brucella canis
Haloarcula marismortui
Pseudoalteromonas haloplanktis
Leptospira biflexa
Sphaerobacter thermophilus
Vibrio cholerae
Babesia bovis
Vibrio fischeri
Rhodobacter sphaeroides
Theileria parva
Aliivibrio salmonicida
Paracoccus denitrificans
Leptospira interrogans
Ochrobactrum anthropi
Ralstonia pickettii
Vibrio splendidus
Agrobacterium vitis
Vibrio sp.
Vibrio parahaemolyticus
Vibrio vulnificus
Cupriavidus taiwanensis
Vibrio harveyi
Photobacterium profundum
Ralstonia eutropha
Agrobacterium radiobacter
Variovorax paradoxus
Burkholderia ambifaria
Theileria annulata
Ashbya gossypii
Cryptosporidium parvum
Pichia pastoris
Zygosaccharomyces rouxii
Kluyveromyces thermotolerans
Kluyveromyces lactis
Saccharomyces cerevisiae
Debaryomyces hansenii
Candida glabrata
Ostreococcus tauri
Schizosaccharomyces pombe
Ostreococcus lucimarinus
Pichia stipitis
Cyanidioschyzon merolae
Cryptococcus neoformans
Yarrowia lipolytica
Trypanosoma brucei
Micromonas sp.RCC299
Plasmodium vivax
Plasmodium knowlesi
Phaeodactylum tricornutum
Thalassiosira pseudonana
Leishmania braziliensis
Dictyostelium discoideum
Aspergillus niger
Caenorhabditis elegans
Drosophila melanogaster
Chlamydomonas reinhardtii
Arabidopsis thaliana
Cucumis sativus
Apis mellifera
Tribolium castaneum
Arabidopsis lyrata
Anopheles gambiae
Brachypodium distachyon
Medicago truncatula
Vitis vinifera
Populus trichocarpa
Oryza sativa
Gasterosteus aculeatus
Lotus japonicus
Sorghum bicolor
Oryzias latipes
Glycine max
Danio rerio
Zea mays
Sus scrofa
Equus caballus
Canis familiaris
Mus musculus
Bos taurus
Rattus norvegicus
Macaca mulatta
Pongo pygmaeus
Homo sapiens
Pan troglodytes
Monodelphis domestica
0
40
100
200
300
400
500
1000
1500
2000
2500
3000
3500
Genome size (Mb)
B
Prokaryotes
Unicellular eukaryotes
Invertebrates
Vascular plants
Vertebrates
6
7
8
9
10
Genome size (Log10)
E
0.0
0.2
0.4
0.6
0.8
Repeat proportion
C
Log10 of SD of genome size
6
7
8
9
10
Log10 of average of genome size
D
SD of Log10 of genome size
0.0
0.2
0.4
0.6
5
6
7
8
9
10
Average of Log10 of genome size


A
Log10 of SD of chromosome size
Log10 of average chromosome size
B
Log10 of SD of non−repeat size
Log10 of average non−repeat size
C
Log10 of SD of repeat size
Log10 of average repeat size
D
SD of Log10 chromosome size
Average of Log10 chromosome size
E
SD of Log10 non−repeat size
Average of Log10 non−repeat size
F
SD of Log10 repeat size
Average of Log10 repeat size
Vertebrates
Vascular plants
Invertebrates
Unicellular eukaryotes
Prokaryotes


A
Inverse of Gamma cdf
Cubic function
Chromosome size
Chromosome index
B
Gamma (7.0438, 1/7.0438)
Normal (1.0000, 0.1371)
Density
Chromosome size
C
Predicted chromosome size proportion
Observed chromosome size proportion
D
Vertebrates
Vascular plants
Invertebrates
Unicellular eukaryotes
Chromosome size propotion
Chromosome


A
B
C
D
Chromosome size
Chromosome index
Translocation model
Inverse of Gamma cdf