OPEN ACCESS Freely available online

PLOS | ONE

Insular Organization of Gene Space in Grass Genomes

Andrea Gottlieb[1◑], Hans-Georg Müller[1◑], Alicia N. Massa[2¤a], Humphrey Wanjugi[3¤b], Karin R. Deal[4], Frank M. You[4¤d], Xiangyang Xu[2¤c], Yong Q. Gu[3], Ming-Cheng Luo[4], Olin D. Anderson[3], Agnes P. Chan[5], Pablo Rabinowicz[6¤e], Katrien M. Devos[2◑], Jan Dvorak[4*◑]

1 Department of Statistics, University of California Davis, Davis, California, United States of America, **2** Institute of Plant Breeding, Genetics and Genomics (Department of Crop and Soil Sciences), Department of Plant Biology, University of Georgia, Athens, Georgia, United States of America, **3** USDA/ARS Western Research Center, Albany, California, United States of America, **4** Department of Plant Sciences, University of California Davis, Davis, California, United States of America, **5** The J. Craig Venter Institute, Rockville, Maryland, United States of America, **6** Institute for Genome Sciences, and Department of Biochemistry and Molecular Biology, University of Maryland School of Medicine, Baltimore, Maryland, United States of America

Abstract

Wheat and maize genes were hypothesized to be clustered into islands but the hypothesis was not statistically tested. The hypothesis is statistically tested here in four grass species differing in genome size, *Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor*, and *Aegilops tauschii*. Density functions obtained under a model where gene locations follow a homogeneous Poisson process and thus are not clustered are compared with a model-free situation quantified through a non-parametric density estimate. A simple homogeneous Poisson model for gene locations is not rejected for the small *O. sativa* and *B. distachyon* genomes, indicating that genes are distributed largely uniformly in those species, but is rejected for the larger *S. bicolor* and *Ae. tauschii* genomes, providing evidence for clustering of genes into islands. It is proposed to call the gene islands ''gene insulae'' to distinguish them from other types of gene clustering that have been proposed. An average *S. bicolor* and *Ae. tauschii* insula is estimated to contain 3.7 and 3.9 genes with an average intergenic distance within an insula of 2.1 and 16.5 kb, respectively. Inter-insular distances are greater than 8 and 81 kb and average 15.1 and 205 kb, in *S. bicolor* and *Ae. tauschii,* respectively. A greater gene density observed in the distal regions of the *Ae. tauschii* chromosomes is shown to be primarily caused by shortening of inter-insular distances. The comparison of the four grass genomes suggests that gene locations are largely a function of a homogeneous Poisson process in small genomes. Nonrandom insertions of LTR retroelements during genome expansion creates gene insulae, which become less dense and further apart with the increase in genome size. High concordance in relative lengths of orthologous intergenic distances among the investigated genomes including the maize genome suggests functional constraints on gene distribution in the grass genomes.

Citation: Gottlieb A, Müller H-G, Massa AN, Wanjugi H, Deal KR, et al. (2013) Insular Organization of Gene Space in Grass Genomes. PLoS ONE 8(1): e54101. doi:10.1371/journal.pone.0054101

Editor: Shyamal D. Peddada, National Institute of Environmental and Health Sciences, United States of America

Received August 7, 2012; **Accepted** December 6, 2012; **Published** January 11, 2013

Funding: This work was supported by the National Science Foundation Plant Genome Research Program Cooperative Agreements DBI-0077766 and DBI-0638558. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jdvorak@ucdavis.edu

◑ These authors contributed equally to this work.

¤a Current address: Department of Plant Biology, Michigan State University, East Lansing, Michigan, United States of America
¤b Current address: Monsanto Company, Saint Louis, Missouri, United States of America
¤c Current address: Division of Plant Sciences, University of Missouri, Columbia, Missouri, United States of America
¤d Current address: Cereal Research Center, Agriculture and Agri-Food Canada, Winnipeg, Manitoba, Canada
¤e Current address: DOE Office of Biological and Environmental Research, Washington, D. C., United States of America

Introduction

Genes in grass genomes are separated from each other by intergenic regions often containing transposable elements (TEs) [1]. The balance between the rate with which TEs are inserted into an intergenic region and the rate with which they are deleted determines whether the region is expanding or contracting [1–4]. Changes in this balance are almost certainly the primary cause of variation in genome size and gene density along chromosomes.

Gene density in many grass genomes increases overall from the proximal towards the distal regions of chromosome arms. This gradient is heterogeneous and regions of higher and lower gene density are superimposed on it. This pattern of gene distribution has been observed in all grass genomes sequenced to date [5–10] and was also inferred in wheat chromosomes by deletion mapping [11]. Heterogeneity in gene distribution along wheat chromosomes has been used to hypothesize that wheat genes are clustered into a limited number of very large gene islands separated by vast expanses of gene-empty space [12,13]. Gene distribution on BAC-based physical maps of wheat chromosome 3B and the seven chromosomes of *Aegilops tauschii*, the diploid source of the wheat D genome [14,15], failed to support this hypothesis [16,17]. Sequencing of single wheat BAC clones and BAC clone contigs suggested that gene clustering into many small islands is a more likely scenario [18–20]. Sequencing and annotation of BAC contigs from wheat chromosome 3B totaling 18.2 Mb revealed a chromosome-wide average gene density of 1 gene per 96 kb but a frequency distribution of intergenic distances showed a peak of very short intergenic distances [20]. This peak was used as evidence for gene clustering into small gene islands [20]. A N50 (median) distance between genes of 43 kb was arbitrarily chosen as the upper limit for intergenic distances within gene islands. Using this criterion, 75% of genes and pseudogenes were contained within 42 gene islands in the 18.2 Mb region of chromosome 3B. The islands contained on average 3.2 genes.

A similar pattern has also been reported in the maize genome. Analysis of two proximal regions in maize chromosome 3 revealed gene islands containing from 1 to 4 genes separated by clusters of long-terminal repeat (LTR) retroelements [21]. These LTR retroelement clusters ranged from 23 to 155 kb [21]. While the overall gene density was estimated to be 1 gene per 111 kb, a gene density of 1 gene per 16 kb was observed within gene islands. Analyses of a larger portion of the maize genome showed that about half of the intergenic distances in maize were shorter than 20 kb [22].

All studies of gene clustering in grass genomes have been intuitive and lacked statistical inference. It cannot be excluded in the absence of a formal statistical analysis that the distribution of distances between genes, such as that described by Choulet et al. [20], is typical for genes that are distributed uniformly along a chromosome.

In order to gain better understanding of the structure and evolution of gene space in grass genomes, we obtained statistical inference for the null hypothesis that gene locations in orthologous regions of the genomes of four grass (Poaceae) species, *Sorghum bicolor* (sorghum), *Oryza sativa* (rice), *Brachypodium distachyon* (false brome), and *Ae. tauschii* (goat grass), follow a homogeneous Poisson process. *Aegilops tauschii* belongs to the tribe Triticeae, and like all members of the tribe it has a large genome (4,020 Mb [23]). The sorghum genome at 730 Mb [7] is about a one-sixth of the *Ae. tauschii* genome. The remaining two species, rice and *B. distachyon,* have genomes more than an order of magnitude smaller than the *Ae. tauschii* genome; the rice genome is 389 Mb [5] and that of *B. distachyon* is 271 Mb [9]. By comparing four genomes differing in size, we assess the effects of genome expansion on the evolution of gene space and gene distribution in grass genomes.

Materials and Methods

Genome Sequences

Since the *Ae. tauschii* genome has not been sequenced, sequences of nine randomly selected megabase-size contigs of BAC clones totaling 9,796 kb were generated for this study [24]. The contigs were located in four of the seven *Ae. tauschii* chromosomes and originated from both distal and proximal chromosome regions. A total of 90 genes were manually annotated in the *Ae. tauschii* contigs suggesting an average gene density (excluding pseudogenes) of 1 gene/106 kb and predicting a total of 36,371 genes in the genome [24]. These estimates are close to those inferred for the wheat B genome on the basis of chromosome 3B BAC contig sequencing [20].

The orthologous regions in *B. distachyon,* rice and sorghum had been shown to harbor 62, 67 and 72 genes, respectively [24]. In all species analyzed, intergenic distances were measured as the distance between the start/end of the coding region of neighboring genes.

Because the *Ae. tauschii* BAC sequence consisted of, on average, 4.5 scaffolds per BAC [24] and the size of the gaps between scaffolds was unknown, we considered scaffolds as contiguous (that is, no extra bases were added to account for the sequence gaps) for the purpose of calculating the intergenic distances. This is unlikely to affect our analysis as the majority of small intergenic distances analyzed here (97% of intergenic distances ≤20 kb) were derived from genes located on the same scaffold. Also, although 64% of sequence scaffolds within a BAC clone could be ordered based on the overlap with neighboring BAC clones [24], some genes were located on unordered scaffolds. For those genes, we determined both the smallest and largest distance to the nearest gene considering all possible contig orders, and used the average between the two distance measurements in our statistical analyses (Table S1). For genes that were located on different BAC clones, intergenic distances were calculated as described above, but taking into account the overlap between neighboring BAC clones. To analyze the distribution of intergenic distances, all genes annotated in the orthologous regions in *Ae. tauschii*, *B. distachyon*, rice, and sorghum were taken into account (Tables S2, S3, S4). To analyze the level of gene island conservation across species, we only considered orthologous gene pairs that were present in all four species analyzed (Table S5).

Distribution of Intergenic Distances

An initial step in the analysis of gene density in the grass genomes is to formally test the pertinent null hypothesis that genes are randomly located along the chromosomes according to a uniform distribution. This distribution will arise if the locations of the genes along the chromosome follow a homogeneous Poisson process [25]. Under this model, gene locations are independent of one another and the expected number of genes in any one region is simply proportional to the length of that region.

A consequence of the gene locations following a homogeneous Poisson process is that the density of intergenic distances will follow an exponential distribution. This implies that rather than directly testing for the Poisson process or the uniform distribution conditional on the total number of genes, we can test for the exponential distribution using the aggregated sample of intergenic distances from the 9 sequenced contigs, viewing the collection as a random sample from the population of all intergenic distances in the genome.

For each of the four species, denoting the density function of the distribution of intergenic distances by f_0, we tested the null hypothesis.

H_0: The intergenic distances follow an exponential distribution, *i.e.*

$$f_0(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \quad \text{for a } \lambda > 0 \qquad (1)$$

vs H_A: The intergenic distances do not follow an exponential distribution.

$$f_0(x) \neq \lambda e^{-\lambda x}, \quad x \geq 0, \quad \text{for any } \lambda > 0 \qquad (2)$$

This test can be implemented with a chi-square goodness of fit test using 10 cells with varying bin widths. This leads to expected frequencies ranging from 5 to 10 per cell, depending on the species. Under H_0, the rate parameter λ was estimated using the maximum likelihood method. The applicable null distribution is a chi-square distribution with 8 degrees of freedom.

Intergenic Distances Density Estimation

An estimate of the true density of intergenic distances was achieved via non-parametric density estimation. We employed a binning approach, first binning the raw data then applying a local linear kernel smoother on the resulting histograms [26]. The initial histogram density estimate is defined in the following way: Given the random sample $x_1, x_2, \ldots, x_m \in (a,b)$ from the density of interest f, define an equally spaced partition of (a, b) by $(t_0, \dots, t_P)$ where $a = t_0 < t_1 < \dots < t_P = b$. The bin width Δ is determined by the number of bins so that $\Delta = \frac{b-a}{P}$. We use y_j to denote the count of sample points falling within bin $B_j = [t_{j-1}, t_j)$, and s_j to denote the midpoint of B_j for $j = 1, \dots, P$, so that $\sum_{j=1}^{P} y_j = m$. The histogram density estimate $\hat{f}_H$ of f is given by

$$\hat{f}_H(x) = \frac{y_i}{m\Delta}, x \in B_j, j = 1,\dots,P. \tag{3}$$

We obtain $\hat{f}$ from $\hat{f}_H$ via local linear smoothing methods [27]. Formally, the estimate $\hat{f}$ is obtained by minimizing the local weighted sums of squares,

$$\sum_{j=1}^{P} K\left(\frac{s_j - t}{h}\right)\left\{\hat{f}_H(s_j) - \left\{\alpha_0 + \alpha_1(s_j - t)\right\}\right\}^2 \tag{4}$$

for each t with respect to α_0 and α_1 from which we obtain $\hat{f}(t) = \hat{\alpha}_0$. We used a Gaussian kernel $K(x)$ and chose the bandwidth h manually.

Conservation of Gene Islands Across Species

Gene pairs were used to evaluate the preservation of gene islands across species. For the purposes of this analysis we defined a gene pair as two genes that were neighbors in the ancestral grass genome [24] and that were both present in the analyzed species *Ae. tauschii*, *B. distachyon*, rice and sorghum. For each pair, we obtained the distance between the neighbors in each of the four species. In addition to 28 gene pairs identified from the *Ae. tauschii* data set, we also identified 8 gene pairs from sequenced BAC clones of hexaploid wheat cultivar Chinese Spring (J.L. Bennetzen and K.M. Devos, unpublished data) with orthologs in *B. distachyon,* rice and sorghum. In order to evaluate the preservation of gene islands across species we calculated Kendall's coefficient of concordance W [28,29] using a correction for ties since the original data consists of raw distances rather than ranks. This statistic was then used to test significance of the coefficient of concordance by testing the null hypothesis $H_0 : W = 0$.

Correlation of Distances between Neighboring Genes and Location of Contig on the Centromere-telomere Axis

The relative location of each contig on the centromere-telomere axis, ranging from 0 (centromere) to 1.0 (telomere), was downloaded from the *Ae. tauschii* genetic maps reported by Massa et al. [24]. Distances from the end of a gene to the beginning of the neighboring gene in nucleotides were determined for all nine contigs. Only distances between genes were measured; distances from the first and last gene to the respective contig edge were disregarded. For each contig, the empirical quantile function of intergenic distances was computed and distances within the first quartile were considered intra-insular distances and distances within the fourth quartile were considered inter-insular distances. In order to determine if a relationship exists between location on the centromere-telomere axis and gene island structure, we implemented the regression model:

$$\mathrm{E}(Y_i | L_i, S_i) = \beta_0 + \beta_1 L_i + \beta_2 S_i + \beta_3 L_i S_i \tag{5}$$

where Y_i is the intergenic distance between the *i*th pair of genes, L_i is the location along the centromere-telomere axis of the *i*th gene pair and $S_i = 0$ if the pair is categorized as intra-insular and $S_i = 1$ if the pair is categorized as inter-insular by the above quantile approach. For intra-insular distances, $S_i = 0$, and the model is

$$\mathrm{E}(Y_i | L_i, S_i = 0) = \beta_0 + \beta_1 L_i \tag{6}$$

where β_0 represents the intercept and β_1 is the slope describing the linear relationship between intra-insular distances and location. For inter-insular distances, $S_i = 1$, and the model is

$$\mathrm{E}(Y_i | L_i, S_i = 1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) L_i \tag{7}$$

where $\beta_0 + \beta_2$ represents the intercept and $\beta_1 + \beta_3$ is the slope describing the linear relationship between inter-insular distances and location. This model allows for formal testing of the following two null hypotheses:

Hypothesis 1. H_0 : Intra-insular distances within islands are homogeneous along the *Ae. tauschii* chromosomes irrespective of location along the centromere-telomere axis ($\beta_1 = 0$).

Hypothesis 2. H_0 : Inter-insular distances are homogeneous along the *Ae. tauschii* chromosomes irrespective of location along the centromere-telomere axis ($\beta_1 + \beta_3 = 0$).

Repeat Elements in Intergenic Regions

All intergenic regions in the nine *Ae. tauschii* contigs that were 100 kb or less in length were analyzed for the presence of LTR retrotransposons by BLASTN searches against the Triticeae Repeat (TREP) database (http://wheat.pw.usda.gov/ITMI/Repeats/). We also conducted a BLASTN search of the intergenic regions against Triticeae sequences in the 'High Throughput Genomic Sequences' (HTGS) section of GenBank, which contains several hundred sequenced wheat and *Ae. tauschii* BAC clones. Intergenic regions with multiple hits were annotated according to the annotation found in the BAC clones deposited in GenBank. Hits with a minimum length of 50 bp and an e-value $\leq 1\mathrm{e}^{-10}$ were considered significant (Table S1). LTR retrotransposons in rice intergenic regions were identified using the same criteria by BLASTN searches against the repeat database ISU Cereal Repeat DB 3.1 (http://magi.plantgenomics.iastate.edu) (Table S2).

Results

Density Functions of Intergenic Distances

The graphs of the non-parametrically estimated density functions (dashed lines in Figure 1) show that short distances between neighboring genes are the most frequent distances in all four genomes. Short distances are also most frequent in the graphs of gene locations as captured by the null hypothesis of uniform locations of genes along chromosome arms given by Equation (1) (solid lines in Figure 1). In *B. distachyon* and rice, the non-parametric density functions do not significantly differ from the fitted exponential density for those species (Figures 1A and 1B and Table 1) and indicate that genes are not clustered in these species. In contrast, in sorghum and *Ae. tauschii,* non-parametric density functions significantly differ from the fitted exponential density and have several features in common (Figures 1C and 1D and Table 1). In both genomes, the non-parametric density graph is above the fitted exponential density graph for short intergenic distances, indicating the presence of too many short distances, and then below it, indicating too few intermediate distances. This pattern suggests that gene clustering is present. We propose to call these clusters "gene insulae" and the pattern of gene distribution observed in the sorghum and *Ae. tauschii* genomes as "insular gene distribution". We propose this new term because the English term "gene islands" has been used for other forms of gene clustering (see Introduction).

The sorghum and *Ae. tauschii* non-parametric density graphs display an initial interval of very short intergenic distances on which the estimated non-parametric density function is concave down (Figures 1C and 1D). This portion of the graphs intersects the density function based on the null hypothesis at 3 kb and 22 kb in sorghum and *Ae. tauschii,* respectively, passes through a local minimum at 8 and 81 kb, respectively, changes to concave up, again intersects the function based on the null hypothesis and reaches a local maximum, with one or more additional concavity changes as the intergenic distances increase (Table 2). We propose using the first local minimum to separate the intra-insular intergenic distances on the left from inter-insular distances on the right. This choice is motivated by cluster analysis based on non-parametric density estimation [30]. The mean intergenic distance shorter than this minimum (*i.e.* the mean intra-insular distance) is 2.1 kb in sorghum and the average insula contains 3.7 genes. The mean intergenic distance shorter than this minimum is 16.7 kb in *Ae. tauschii* and the average insula contains 3.9 genes. The average intergenic distance longer than the first local minimum (*i.e.* the mean inter-insular distance) is 15.1 kb in the sorghum genome but is significantly larger at 205.2 kb in the *Ae. tauschii* genome (*Mann-Whitney Test*, *P*-value = 0). The distances shorter than this critical point represent 70% and 63% of all intergenic distances in the sorghum and *Ae. tauschii* genomes, respectively.

Gene Density along the Centromere-telomere Axis

Gene density increases from the proximal, low-recombination regions towards the distal, high-recombination regions in the *Ae. tauschii* chromosomes [17]. This gradient could be caused by a change in gene density within gene insulae, a change in the distances between gene insulae, or both. In order to determine which of these three possibilities is most consistent with the data, we fit the regression model in (5) to *Ae. tauschii* data after removing a single outlier due to its high influence on the parameter estimates. The fitted regression model (5) indicates no relationship between intra-insular distances and gene location along the centromere-telomere axis based on the high *P* values for parameters β_0 and β_1 (Table 3). Since β_1 cannot be distinguished

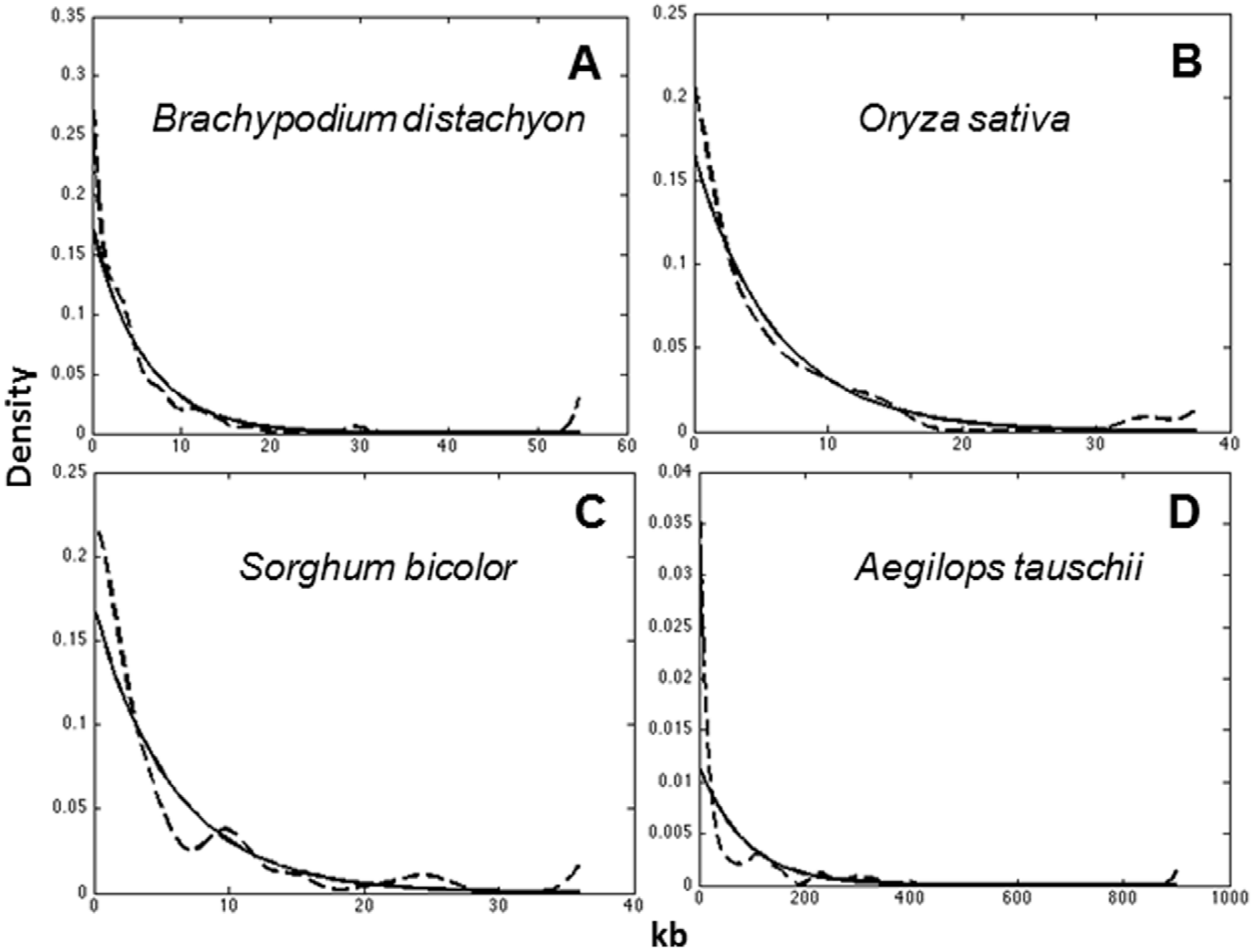

Figure 1. Density functions for intergenic distances in *B. distachyon* (A), rice (B), sorghum (C), and *Ae. tauschii* (D). Shown is the exponential density fitted by Maximum Likelihood (2) (solid) and the non-parametric density estimate (4) (dashed), with bandwidth $h = 1.25$ (**A**), $h = 1.75$ (**B**), $h = 1.50$ (**C**), and $h = 15.00$ (**D**).
doi:10.1371/journal.pone.0054101.g001

Table 1. Summary of test results for the null hypothesis that gene locations are uniformly distributed in the four species.

	No. intergenic distances	Estimated rate parameter $\hat{\lambda}$*	Estimated standard error of $\hat{\lambda}$	*P*-value**
B. distachyon	52	0.174	0.024	0.330
Rice	57	0.167	0.022	0.064
Sorghum	62	0.170	0.022	0.007
Ae. tauschii	81	0.012	0.001	0.000

*The estimated exponential rate parameter $\hat{\lambda}$ is the maximum likelihood estimator of λ as given in Equation (1).
**Based on a χ^2- goodness of fit test.
doi:10.1371/journal.pone.0054101.t001

from zero at the chosen significance level, the significance of β_3 indicates that while there is no significant relationship between intra-insular distances and insula location along the centromere-telomere axis there is a significant relationship between inter-insular distances and location along the centromere-telomere axis (Figures 2A and 2B). This relationship between inter-insular distances and gene location is quantified in equation (7) and results are shown in Table 4. The fitted regression model (7) indicates that there is a strong negative relationship between inter-insular distances and insula location along the centromere-telomere axis (Figure 2A). We therefore conclude that increase in gene density towards the ends of *Ae. tauschii* chromosomes is caused primarily by a decrease in the distances between insulae.

It should be noted that as an alternative to the quantile approach to classify intergenic distances as intra-insular or inter-insular used here, we could have used the first local minimum at 81 kb as the boundary between intra-insular and inter-insular distances. In fact, we implemented this method as well and it yielded identical inferences. We report the results of the quantile approach because this method is particularly robust. Similarly, we tried several variance stabilizing transformations prior to fitting the regression model. In all cases, the inferences were identical. We therefore report here the most parsimonious model. Clearly there is a strong signal in the data indicating a negative relationship between inter-insular distances and insula location along the centromere-telomere axis.

To determine if there is a relationship between the insula size and its location on the centromere-telomere axis in *Ae. tauschii*, we plot the number of genes per insula and insula location on the centromere-telomere axis for 15 *Ae. tauschii* insulae (Figure 2C). There is no relationship between the number of genes per insula and insula location ($r = 0.19$, *P*-value = 0.42). The situation does not change after an obvious outlier (insula with 13 genes) is removed, ($r = 0.06$, *P*-value = 0.8). We conclude that the sizes of insulae do not appreciably change along the centromere-telomere axis in the *Ae. tauschii* genome.

Table 2. Insular structure in sorghum and *Ae. tauschii.*

Statistic	Unit	Sorghum	*Ae. tauschii*
First local minimum	kb	8.0	81.0
Mean of distances shorter than the minimum	kb	2.1	16.7
Mean of distances longer than the minimum	kb	15.1	205.2
Shorter distances	%	70.0	63.0

doi:10.1371/journal.pone.0054101.t002

Conservation of gene Insulae

Kendall's coefficient of concordance W = 0.49, which indicates that there is a high level of concordance of distances between orthologous gene pairs in the four genomes. This means that gene pairs that are relatively close to one another in one genome are likely to be also close to one another in other genomes. Analogously, gene pairs that are relatively distant from each other in one genome are likely to be also far apart in the other three genomes. The null hypothesis that the intergenic distances are independent across genomes is rejected with $P < 0.001$.

Gene Insulae and LTR Retroelements

In the *Ae. tauschii* genome, all inter-insular intergenic distances (distances >81 kb) contain LTR retrotransposons but only 27% of intra-insular intergenic distances contain LTR retrotransposons or their remnants. As shown in Figures 1C and 1D, most of the intra-insular distances are overrepresented relative to the uniform gene distribution. In the *Ae. tauschii* genome, only 9% these intergenic distance (distances ≤22 kb, Figure 1D) contain LTR retrotransposon remnants (Table S6), which indicates that the absence of LTR retrotransposon insertions is the primary contributor to the overrepresentation of short intergenic distances and an important attribute of insulae in the *Ae. tauschii* genome.

Discussion

An important feature of the density functions based on the null hypothesis is a prominent peak of very short intergenic distances predicted for each of the four genomes. Thus, very short intergenic distances are expected to dominate density profiles of observed intergenic distances even if gene locations were distributed uniformly along chromosomes. This fact is exemplified by the *B. distachyon* and rice genomes in which short intergenic distances dominate the density profiles, yet genes are not clustered in those genomes. Observation of a prominent peak of short distances in an empirical density profile, such as that observed by Choulet et al. [20] is not in itself sufficient to draw a conclusion regarding the existence of gene clustering.

Comparisons of the density functions based on the null hypothesis with the non-parametric density functions estimated from observed distances between neighboring genes divided the four genomes into two contrasting pairs. The first pair consisted of the small *B. distachyon* and rice genomes. In those genomes the nonparametric and exponential density functions did not significantly differ from each other. The second pair consisted of the larger sorghum and *Ae. tauschii* genomes, for which the two density functions differed significantly.

Variation in genome size in grasses is principally caused by the accumulation or loss of TEs. Genome expansion takes place

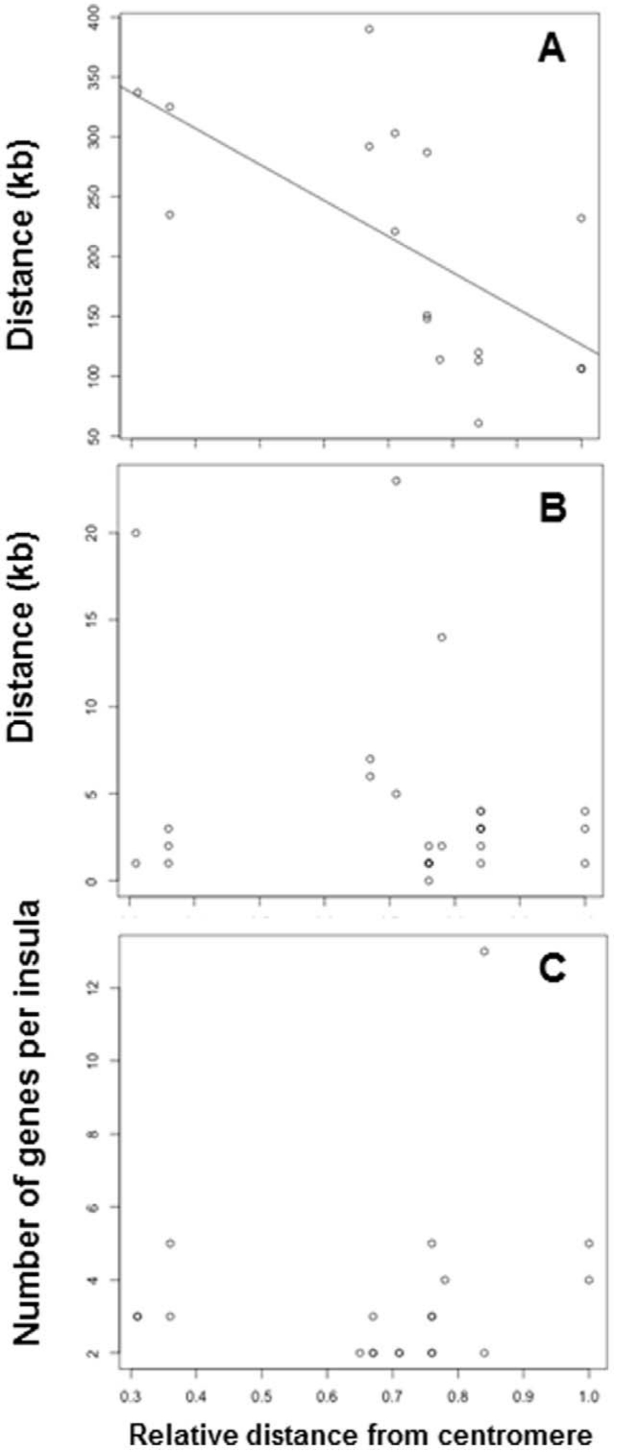

Figure 2. Relationships between inter-insular distances (A), intra-insular distances (B) and the number of genes per insula (C) and gene location along the centromere-telomere axes of *Ae. tauschii* chromosome arms. (**A**) shows a fitted regression line from Model (7) of the inter-insular distances and gene location on centromere-telomere axes of *Ae. tauschii* chromosome arms.
doi:10.1371/journal.pone.0054101.g002

Table 3. Fitting the full regression model specified in equation (5) to the *Ae. tauschii* data.

Parameter	Estimate	*P*-value	95% confidence interval
β_0	8.3	0.82	(−64.5, 81.1)
β_1	−5.5	0.92	(−102.8, 91.9)
β_2	387.8	~0.00	(279.9, 495.7)
β_3	−257.8	~0.00	(−403.0, −112.5)

doi:10.1371/journal.pone.0054101.t003

principally by the accumulation of LTR retroelements, which have the tendency to insert into LTR retroelements resident in the chromosomes [1,2,4]. Genome contraction takes place principally by DNA deletion caused by unequal homologous recombination between LTRs of the same or related LTR retrotransposons and by illegitimate recombination within TE and other non-essential DNA [3]. Most of the observations that emerged here can be attributed to TE dynamics and the nonrandom insertion of LTR retroelements. In small, gene-dense genomes, such as those of *B. distachyon* and rice, the overall retroelement content is lower, 21.4% in *B. distachyon* [9] and 26% in rice [5], than in the larger genomes, 54% in sorghum and at least 51% in the *Ae. tauschii* genome [31]. Genes are distributed uniformly and there is very little gene clustering in the small genomes. In the nine studied regions, LTR retrotransposons or remnants of LTR retrotransposons were present in only 12% of rice gene pairs (Table S2) compared to 54% of *Ae. tauschii* gene pairs (Table S1). When comparing only orthologous gene pairs, 12% of the rice gene pairs but 61% of the *Ae. tauschii* gene pairs were separated by sequences with homology to LTR retrotransposons (Table S6). All of the *Ae. tauschii* inter-insular spaces contained retrotransposons. This is consistent with the hypothesis that expansion of grass genomes, which takes place predominantly by the accumulation of LTR retroelements [2], occurs principally in regions already containing LTR retroelements. As a result of this nested insertion of retroelements, genes that are separated by LTR retroelements will be pushed further apart from each other during genome expansion, which will create large arrays of LTR retroelements characteristic of inter-insular space [20,21]. Genes that are not separated by LTR retroelements will tend to remain near each other and form insulae. Due to the LTR retroelement dynamics, gene distribution is largely homogeneous in small grass genomes but assumes insular organization as a consequence of genome expansion. As illustrated by the comparison of orthologous regions in the sorghum and *Ae. tauschii* genomes, insulae become less gene dense and separated by greater spans of nested TEs as a genome expands.

The distal chromosome regions are typically more gene-dense than the proximal chromosome regions in the Triticeae chromosomes. Regression analysis showed that gene number per insula and gene density within an insula were similar along *Ae. tauschii* chromosomes but that inter-insular distances were shorter in distal, high-recombination regions compared to proximal, low-recombination regions. Therefore, the increase in gene density toward the distal regions of *Ae. tauschii* chromosomes [17] is largely

Table 4. Fitting the regression model (7) to the *Ae. tauschii* data.

Parameter	Estimate	*P*-value	95% confidence interval
$\beta_0+\beta_2$	427.2	~0.00	(274.3, 580.1)
$\beta_1+\beta_3$	−300.8	0.01	(−502.9, −98.6)

doi:10.1371/journal.pone.0054101.t004

due to shortening of inter-insular distances.

Insulae in the distal gene-rich regions of wheat chromosome 3B were reported to contain larger numbers of genes than insulae in the proximal, gene-poor chromosome regions [16]. This was not observed in the *Ae. tauschii* genome. The two-fold larger threshold assumed to delimit intra- and inter-insular spaces in this study (81 kb) compared to the 3B study (43 kb), and the exclusion of pseudogenes from our analysis could have caused this difference. Our observation that insulae contain similar numbers of genes in distal chromosome regions compared to the less gene-dense proximal chromosome regions is paralleled by similar numbers of genes per insula in the genome of sorghum (3.7 genes per insula) compared to the less gene-dense genome of *Ae. tauschii* (3.9 genes per insula).

The observation that the greater gene density in the distal regions of chromosome 3B is due to genes that are not syntenic with rice and *B. distachyon* [16] is probably unrelated to the insular dynamics. It is more likely a reflection of a high proportion of novel genes in those regions [17,32,33] due to greater incidence of duplicated genes in the distal regions of wheat chromosomes and those of wheat diploid ancestors [17,32].

If the evolution of gene insulae was driven entirely by insertions of LTR retroelements, it would be counter-intuitive to expect conservation of insulae over long periods of time. This is particularly true if the remarkably high rate of turnover of intergenic spaces in grass genomes is taken into account. In the Triticeae genomes, for example, virtually the entire intergenic space equivalent to nearly 90% of the genome, is replaced within four to five million years [34]. The four species studied here belong to grass subfamilies Pooideae (*Ae. tauschii* and *B. distachyon*), Ehrhartoideae (rice) and Panicoideae (sorghum). The Panicoideae subfamily diverged about 52.5 million years ago (MYA) from the ancestor of the Pooideae and Ehrhartoideae subfamilies [9]. Within the subfamily Pooideae, *Ae. tauschii* and *B. distachyon* diverged 35.8 MYA and their common ancestor diverged from rice 47.3 MYA [9]. In spite of the antiquity of these divergence times, comparison of the four genomes suggested that heterogeneity in gene distribution is conserved to some degree. Neighboring genes that are close to each other in one genome tend to be close to each other in other genomes and, *vice versa*, neighboring genes that are far apart in one genome show a tendency to be far apart in other genomes. This observation remained true even when comparing intergenic distances in *Ae. tauschii* with those in maize (Table S5), a Panicoideae species with a haploid genome size of 2,500 Mb (*P*-value <0.005).

The evolutionary conservation of gene distribution over long spans of time argues for a functional factor playing a role in the evolution of insulae, in addition to the process based on the dynamics of LTR retroelement insertions. It is possible that some gene pairs do not tolerate separation by LTR retroelements. Co-location of functionally related or co-expressed genes or genes encoding proteins in the same biochemical pathways was reported in a number of plants [16,35–39]. Such functional constrains may represent another factor playing a role in the formation and conservation of insulae.

Supporting Information

Table S1 Locations of genes in nine *Ae. tauschii* contigs, the distances between neighboring genes, and homology of intergenic regions to LTR retrotransposons.
(XLSX)

Table S2 Locations of genes in the rice regions orthologous to the *Ae. tauschii* contigs, the distances between neighboring rice genes, and homology of the intergenic regions to LTR retrotransposons.
(XLSX)

Table S3 Locations of genes in the *B. distachyon* regions orthologous to the *Ae. tauschii* contigs and the distances between neighboring *B. distachyon* genes.
(XLSX)

Table S4 Locations of genes in the *S. bicolor* regions orthologous to nine *Ae. tauschii* contigs and distances between neighboring *S. bicolor* genes.
(XLSX)

Table S5 Orthologous gene pairs and their intergenic distances in *Ae. tauschii*, rice, *B. distachyon*, sorghum, and maize.
(XLSX)

Table S6 Orthologous gene pairs in *Ae. tauschii* and rice, their intergenic distances and homology of the intergenic sequences to known retrotransposons.
(XLSX)

Author Contributions

Conceived and designed the experiments: AG HGM KMD JD. Performed the experiments: ANM HW KRD XX YQG MCL ODA APC PR KMD JD. Analyzed the data: AG HGM KMD FMY JD. Contributed reagents/materials/analysis tools: KRD MCL. Wrote the paper: AG HGM KMD JD.

References

1. SanMiguel P, Tikhonov A, Jin Y-K, Moutchoulskaia N, Zakharov D, et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765–768.
2. Bennetzen JL, Kellogg EA (1997) Do plants have a one-way ticket to genomic obesity? Plant Cell 9: 1509–1514.
3. Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res 12: 1075–1079.
4. Bennetzen JL (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica 115: 29–36.
5. Consortium (2005) Sequence, annotation, and analysis of synteny between rice chromosome 3 and diverged grass species. Genome Res 15: 1284–1291.
6. Yu J, Hu S, Wang J, Wong GK, Li S, et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296: 79–92.
7. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457: 551–556.
8. Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, et al. (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326: 1112–1115.
9. International Brachypodium Genome Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463: 763–768.
10. Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, et al. (2012) Full genome sequence analysis of the model plant Setaria. Nature Biotech 30: 555–561.
11. Gill KS, Gill BS, Endo TR (1993) A chromosome region-specific mapping strategy reveals gene-rich telomeric ends in wheat. Chromosoma 102: 374–381.
12. Sandhu D, Gill KS (2002) Gene-containing regions of wheat and the other grass genomes. Plant Physiol 128: 803–811.
13. Erayman M, Sandhu D, Sidhu D, Dilbirligi M, Baenziger PS, et al. (2004) Demarcating the gene-rich regions of the wheat genome. Nucleic Acids Res 32: 3546–3565.
14. Kihara H (1944) Discovery of the DD-analyser, one of the ancestors of *Triticum vulgare* (Japanese). Agricult Hort (Tokyo) 19: 13–14.
15. McFadden ES, Sears ER (1946) The origin of *Triticum spelta* and its free-threshing hexaploid relatives. J Hered 37: 81–89, 107–116.
16. Rustenholz C, Choulet F, Laugier C, Safar J, Simkova H, et al. (2011) A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. Plant Physiol 157: 1596–1608.

17. Luo MC, Gu YQ, You FM, Deal KR, Ma YQ, et al. (2012) A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii,* the wheat D-genome progenitor. Proc Natl Acad Sci USA, submitted.
18. Feuillet C, Keller B (1999) High gene density is conserved at syntenic loci of small and large grass genomes. Proc Natl Acad Sci USA 96: 8265–8270.
19. Wicker T, Zimmermann W, Perovic D, Paterson AH, Ganal M, et al. (2005) A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley *Hv-eIF4E* locus: recombination, rearrangements and repeats. Plant J 41: 184–194.
20. Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, et al. (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22: 1686–1701.
21. Kronmiller BA, Wise RP (2009) Computational finishing of large sequence contigs reveals interspersed nested repeats and gene islands in the *rf1*-associated region of maize. Plant Physiol 151: 483–495.
22. Wei FS, Stein JC, Liang CZ, Zhang JW, Fulton RS, et al. (2009) Detailed analysis of a contiguous 22-Mb region of the maize genome. PLOS Genet 5(11): e1000728. doi:10.1371/journal.pgen.1000728.
23. Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Reporter 9: 208–218.
24. Massa AN, Wanjugi H, Deal KR, O'Brian K, You FM, et al. (2011) Gene space dynamics during the evolution of *Aegilops tauschii, Brachypodium distachyon, Oryza sativa,* and *Sorghum bicolor* genomes. Mol Biol Evol 28: 2537–2547.
25. Cox DR, Isham V (1980) Point processes. Chapman and Hall, London, England.
26. Zhang Z, Müller HG (2011) Functional density synchronization. Comput Stat & Data Analysis.55: 2234–2249.
27. Fan J, Gijbels I (1996) Local Polynomial Modelling and its Applications. Chapman and Hall, London, England.
28. Kendall MG, Smith BB (1939) The problem of m rankings. Annals Mathem Stat 10: 275–287.
29. Kendall MG (1945) The treatment of ties in ranking problems. Biometrika 33: 239–251.
30. Silverman BW (1986) Density Estimation for Statistics and Data Analysis: Chapman and Hall, London, England.
31. Li W, Zhang P, Fellers JP, Friebe B, Gill BS (2004) Sequence composition, organization, and evolution of the core Triticeae genome. Plant J 40: 500–511.
32. Dvorak J, Akhunov ED (2005) Tempos of deletions and duplications of gene loci in relation to recombination rate during diploid and polyploid evolution in the Aegilops-Triticum alliance. Genetics 171: 323–332.
33. See DR, Brooks S, Nelson JC, Brown-Guedira G, Friebe B, et al. (2006) Gene evolution at the ends of wheat chromosomes. Proc Natl Acad Sci USA 103: 4162–4167.
34. Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. Science 316: 1862–1866.
35. Lee JM, Sonnhammer ELL (2003) Genomic gene clustering analysis of pathways in eukaryotes. Genome Res 13: 875–882.
36. Williams EJB, Bowles DJ (2004) Coexpression of neighboring genes in the genome of Arabidopsis thaliana. Genome Res 14: 1060–1067.
37. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, et al. (2005) A gene expression map of *Arabidopsis thaliana* development. Nature Genetics 37: 501–506.
38. Xu ZY, Kohel RJ, Song GL, Cho JM, Alabady M, et al. (2008) Gene-rich islands for fiber development in the cotton genome. Genomics 92: 173–183.
39. Lou XL, Han B (2009) Evolutionary conservation of neighbouring gene pairs in plants. Gene 437: 71–79.