NRC Research Press

ARTICLE

A novel application of ecological analyses to assess transposable element distributions in the genome of the domestic cow, *Bos taurus*[1]

Brent Saylor, Tyler A. Elliott, Stefan Linquist, Stefan C. Kremer, T. Ryan Gregory, and Karl Cottenie

Abstract: Transposable elements (TEs) are among the most abundant components of many eukaryotic genomes. Efforts to explain TE abundance, as well as TE diversity among genomes, have led some researchers to draw an analogy between genomic and ecological processes. Adopting this perspective, we conducted an analysis of the cow (*Bos taurus*) genome using techniques developed by community ecologists to determine whether environmental factors influence community composition. Specifically, each chromosome within the *Bos taurus* genome was treated as a "linear transect", and a multivariate redundancy analysis (RDA) was used to identify large-scale spatial patterns in TE communities associated with 10 TE families. The position of each TE community on the chromosome accounted for ~50% of the variation along the chromosome "transect". Multivariate analysis further revealed an effect of gene density on TE communities that is influenced by several other factors in the (genomic) environment, including chromosome length and TE density. The results of this analysis demonstrate that ecological methods can be applied successfully to help answer genomic questions.

Key words: genome ecology, transposon ecology, transposable elements, spatial patterns, gene density, chromosome length.

Résumé : Les éléments transposables (TE) sont parmi les composantes les plus abondantes au sein de plusieurs génomes eucaryotes. Les efforts visant à expliquer cette abondance et diversité des TE au sein des génomes ont amené certains chercheurs à voir une analogie entre les processus génomiques et évolutifs. C'est dans cette perspective que les auteurs ont réalisé une analyse du génome de la vache (*Bos taurus*) en employant des techniques développées par les écologistes des communautés pour déterminer si des facteurs environnementaux influencent la composition d'une communauté. Spécifiquement, chaque chromosome au sein du génome du *Bos taurus* a été traité comme un « transect linéaire » et une analyse de redondance (RDA) multivariée a été réalisée pour identifier les grandes lignes de la répartition de communautés de TE appartenant à 10 familles de TE. La position de chaque communauté de TE au sein du chromosome expliquait ~50 % de la variation le long d'un « transect » chromosomique. Une analyse multivariée a ensuite révélé un effet de la densité génique sur les communautés de TE qui est influencé par plusieurs autres composantes de l'environnement (génomique), dont la longueur du chromosome et la densité en TE. Les résultats de cette analyse montrent que les méthodes de l'écologie peuvent servir à répondre à des questions génomiques. [Traduit par la Rédaction]

Mots-clés : écologie du génome, écologie des transposons, éléments transposables, distribution spatiale, densité génique, longueur du chromosome.

Introduction

The biology of transposable elements

Transposable elements (TEs) are mobile segments of DNA capable of being cut or copied from one location in the genome and reinserted into another. TEs are generally divided into two major classes according to their mode of transposition. Class I TEs, or retrotransposons, replicate by using the cells machinery to copy themselves into RNA. Then, using a TE-encoded reverse transcriptase they are copied from RNA back into DNA and this new DNA strand is then inserted into a new location. Class II elements, or DNA transposons, replicate via a cut-and-paste mechanism without an RNA intermediate. TEs in both classes are further subdivided into families based on sequence similarity (Wicker et al. 2007).

Despite their potential to cause deleterious mutations (e.g., Houle and Nuzhdin 2004; Pasyuokva et al. 2004), TEs represent the most abundant type of DNA sequence within most eukaryotic genomes (Gregory 2005). It has been observed in many larger genomes that TEs are not uniformly distributed throughout the genome, but rather have a tendency to be clustered in heterochromatic regions with low gene density. This could be due to relaxed selection on these regions due to a paucity of host-functional sequences, such that TEs are able to accumulate without being eliminated by purifying selection (Langley et al. 1988; Maside and Bartolomé 2004; Fontanillas et al. 2007). On the other hand, some TEs (e.g., certain miniature inverted terminal repeat elements or MITEs) have been found to preferentially insert into, or perhaps to be less easily deleted from, regions nearby to protein-coding genes in rice, mice, and humans (Zhang et al. 2000; Naito et al. 2009). Similarly, the *Drosophila* gene disruption project (GDP) used a specific TE, the *P* element, to disrupt genes because of its known preference to insert into promoter regions (Bellen et al. 2011). Furthermore, the abundance of *Mutator* family elements in maize has been found to correlate positively with distance from the gene poor centromere (Liu et al. 2009). Taken together, these

Received 24 October 2012. Accepted 29 May 2013.

Corresponding Editor: Jillian Bainard.

B. Saylor, T.A. Elliott, T.R. Gregory, and K. Cottenie. Department of Integrative Biology, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada.
S. Linquist. Department of Philosophy, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada.
S.C. Kremer. School of Computer Science, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada.

Corresponding author: Brent Saylor (e-mail: bsaylor@uoguelph.ca).

[1]This article is one of a selection of papers published in this Special Issue on Genome Size Evolution.

observations suggest that transposon distribution is influenced, at least in part, by local interactions with the host genome that appear to depend on TE-specific properties.

An ecological perspective on TEs

To date, most research on TEs has focused on their molecular properties and the potential effects of their insertion on host (whole organism) fitness (eg. Sela et al. 2010*a*; 2010*b*; Stewart et al. 2011). A more recent proposal for the study of transposon behaviour shifts the focus to the TEs themselves by invoking concepts and models from the discipline of ecology (Kidwell and Lisch 2001; Brookfield 2005; Le Rouzic et al. 2007; Venner et al. 2009). From this "genome ecology" perspective, TEs are viewed as biological entities in their own right, with the surrounding genome being the environment in which (and with which) they interact.

Although the utility of this approach has been discussed by several authors, a detailed recent review of the genome ecology literature revealed a fundamental confusion between what constitutes ecology versus evolution at the TE level (Linquist et al. 2013). Thus, some studies purporting to examine TEs within an ecological framework were actually utilizing methods (e.g., population genetics) and (or) observing patterns (e.g., host-TE coevolution) that were evolutionary, rather than reflecting the shorter-term dynamics that would truly represent TE ecology (Linquist et al. 2013).

Community ecologists often utilize linear transects across an environmental gradient to investigate potential effects on community composition. For example, an ecologist might sample a species distribution along a mountain slope to test for an influence of altitude, moisture, sunlight exposure, or other environmental factors (Whittaker 1960). Secondarily, they may compare several transects drawn from similar environmental gradients but in different locations (Lajeunesse 2010). In a similar vein, Linquist et al. (2013) suggested that factors in the genomic environment, such as total genome size and CG content, could account for significant amounts of variation in TE abundance and distribution, especially among closely related host genomes. However, as yet, the methods of community ecology have not been brought to bear on questions regarding distributional patterns of TEs within genomes.

In this paper, we explore two major questions. The first is whether the distribution of TE "communities" varies systematically along a spatial (chromosomal) gradient. To identify a spatial gradient in an ecological context, one compares community composition along a series of sampling plots. At the genomic level, TE communities can be compared across consecutive windows of fixed width along the chromosome, with each chromosome treated as an ecologist might analyze a mountain within a mountain range. Finding patterns in the relative abundances of particular TE families at certain regions of the chromosome could suggest, as in the case of species, that this distribution is influenced by some "environmental" factor. This would be considered a large-scale spatial pattern because it affects the distribution of TEs across the entire chromosome. Alternatively, if particular TE species' (families) tend to co-occur within a chromosome, but not along a spatial gradient, this suggests the transposon-level equivalent of localized biotic interactions. This would be considered a small-scale pattern. In addition, we asked whether protein-coding gene density predicts the composition of TE communities in the way that the existence of specific habitat may influence ecosystem composition—for example, if particular TEs are excluded from those regions whereas others insert preferentially into them (Zhang et al. 2000; Naito et al. 2009). We used an intrachromosome analysis to test for each chromosome individually if TE community composition was a function of both the spatial location and gene density.

The second major question that we explore is whether the two primary factors outlined above—location and gene density—have equal effects across the different chromosomes, or whether this differs from chromosome to chromosome. For this, we use an interchromosome analysis that enabled us to detect significant patterns across all chromosomes, and to account for additional chromosomal factors potentially influencing community responses. Factors of potential significance include chromosome length, gene number per chromosome, and number of TEs per chromosome. These properties serve as potential predictors of the effects of gene density on the TE community composition.

In this study, we use the genome of the domestic cow (*Bos taurus*) as our model study system (Elsik et al. 2009). The cow genome has been nearly fully sequenced (~92% complete) to a depth of 7× coverage, with 90% of this sequenced DNA having been localized with confidence on specific chromosomes (Elsik et al. 2009). The cow genome is well annotated, and the missing sections of the sequence for each chromosome are provided as blank (N) bases in the available sequence, thereby making it possible to locate TE and gene positions without bias. The cow genome contains 29 pairs of autosomes and one pair of sex chromosomes, providing sufficient numbers of replicates to conduct our interchromosome analysis. Moreover, the chromosomes are known to be variable in size, gene number, and TE content (Table S1[2]). Finally, the cow genome is sufficiently large (~2.8 Gb) and complex to provide a suitable test of the proposed ecological approach but is not so large as to present insurmountable computational challenges.

Materials and methods

Sequence data

The complete sequence of each of the 30 *B. taurus* chromosomes was obtained from GenBank (accession Nos. CM000177.4–CM000206.4). The names and locations of the TEs on each chromosome were obtained using RepeatMasker software (Smit et al. 2004). Each chromosome was then completely divided up into nonoverlapping sampling windows. The selection of window size is important because it determines the scale at which patterns can be detected, as well as the size of the TE community in each window. If the windows are too large, then some of the small-scale patterns might be missed. However, the smaller the window size, the more memory is required to calculate all of the possible spatial relationships between individual windows (see explanation of principal components of neighbour matrices (PCNM)). The window size was set individually for each chromosome such that the average window would contain 100 TEs (though in practice any given window could contain more or fewer than 100 TEs). Therefore, each window size corresponded to a base pair size that was fixed for each chromosome but was variable between the different chromosomes (Table 1). Partitioning the windows in such a way allowed the PCNMs to be calculated on a computer with 64 GB of RAM, while permitting a robust sampling of the abundance of each TE within a window. To investigate the robustness of this window size selection mechanism, we performed three additional analyses. First, we increased and decreased the window size to an average of 50 and 200 TEs per window. This allowed us to assess the suitability of using an average of 100 TEs (compared with 50 and 200) per window for each chromosome. The analysis was also re-run using a constant window size across all chromosomes of 120 155 bp, which was the average of the window sizes for all chromosomes in the 100 TE analysis. In each analysis, the number of TEs for each family per window were tallied for each chromosome. This resulted in a TE

[2]Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/gen-2012-0162.

Table 1. Chromosome window information.

Chr.	Chr. length	No. of TEs	No. of windows	Base pairs per window
1	161 108 518	124 446	1244	129 461
2	140 689 454	75 260	753	186 938
3	127 865 071	115 519	1155	110 687
4	124 429 929	95 849	958	129 819
5	125 642 737	106 628	1066	117 833
6	122 646 612	95 884	959	127 911
7	111 948 904	96 416	964	116 110
8	116 941 315	98 408	984	118 833
9	108 100 207	79 215	792	136 464
10	106 310 658	93 156	932	114 121
11	110 261 590	93 035	930	118 516
12	85 442 768	60 657	607	140 862
13	84 433 115	77 487	775	108 964
14	81 409 064	68 367	684	119 077
15	84 800 091	75 492	755	112 330
16	77 906 053	65 805	658	118 389
17	76 519 030	62 447	624	122 534
18	65 948 816	57 906	579	113 889
19	65 317 834	57 902	579	112 808
20	75 862 687	59 423	594	127 666
21	69 307 001	56 216	562	123 287
22	61 892 535	86 624	866	71 450
23	53 331 164	48 309	483	110 396
24	65 017 658	45 192	452	143 870
25	44 044 338	43 125	431	102 132
26	51 861 200	39 816	398	130 252
27	48 749 334	45 377	454	107 432
28	46 105 694	63 667	637	72 417
29	52 131 757	34 816	348	149 735
X	88 516 663	80 128	801	110 469

Note: The size of each chromosome (chr.) in *Bos taurus*, the number of transposable elements (TEs) in that chromosome, and the number of windows and their respective sizes are shown.

community for each window location, consisting of counts of the number of TEs belonging to each family present at that window's location.

The gene positions for each chromosome were downloaded from GenBank in the gene table format. This format was chosen because it only includes proteins for which known transcripts have been found (to avoid annotation pseudo genes and annotation errors). Gene density was then calculated by counting the number of genes in each of the windows generated in the TE step.

Intrachromosome analysis

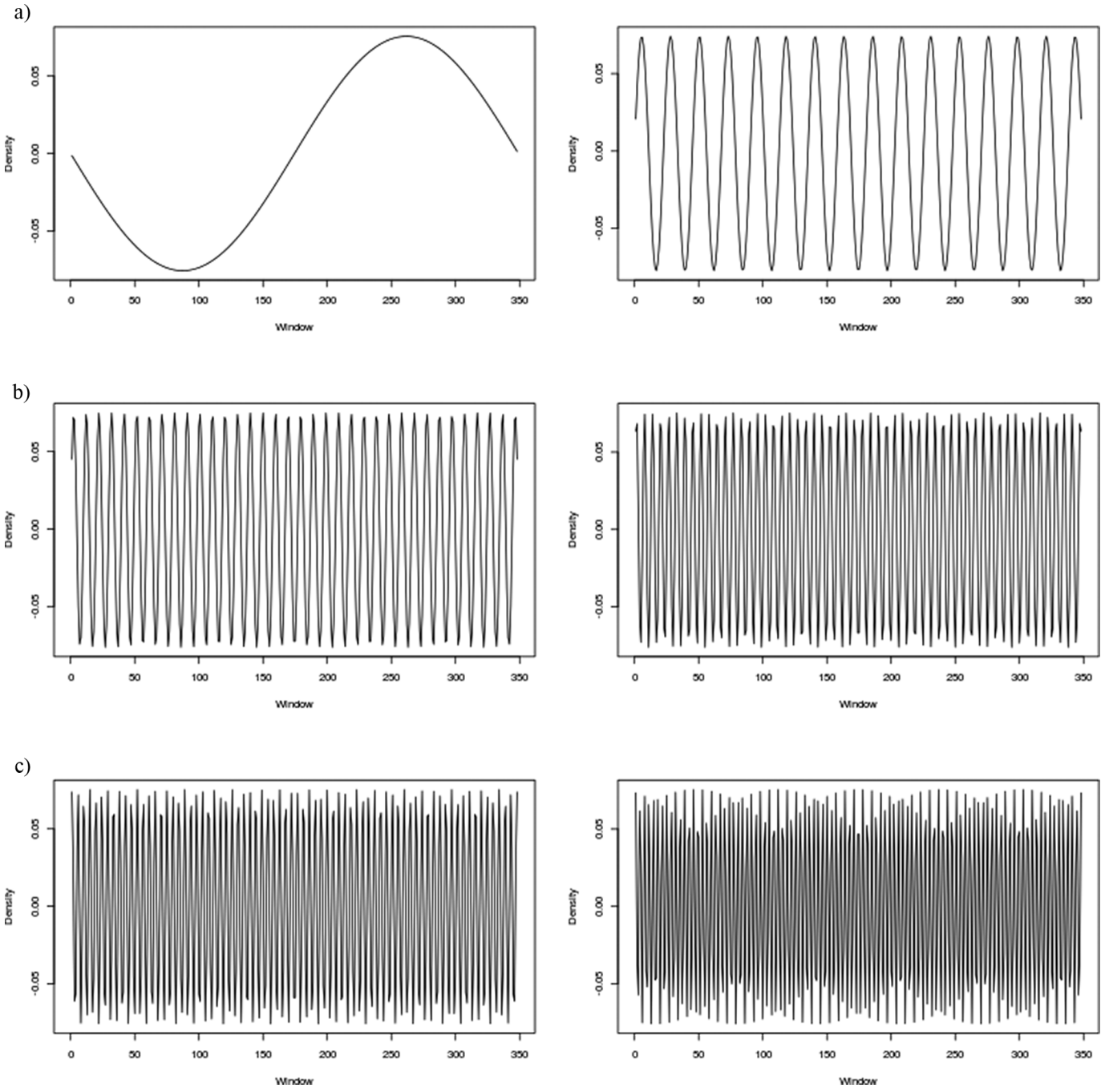
The independent variables used to explain variation within the TE community were spatial location of the window and gene density within each sampling window. The PCNM procedure was then run to generate all of the potential spatial distributions between the windows of a given chromosome (Dray et al. 2006; Griffith and Peres-Neto 2006). To represent the spatial relationship of each window in a given chromosome, a distance matrix was generated for each chromosome. This matrix represents the distance between each window and every other window on the same chromosome. This procedure resulted in graphs representing each of the possible spatial distributions of TEs across the windows of a given chromosome. This is equivalent to generating all of the potential spatial distributions between different plots along a mountain range. Following standard ecological practices (see e.g., Borcard et al. 2012), the sets of graphs for each chromosome were each divided into three equally sized groups. The first third of the potential spatial distributions for a given chromosome represent large-scale spatial patterns (Fig. 1*a*), the second third represent medium-scale spatial patterns (Fig. 1*b*), and the third set represent small-scale spatial patterns (Fig. 1*c*).

Once the potential spatial distributions were generated, the amount of variation they were able to explain was computed using a redundancy analysis (RDA; Legendre and Legendre 1998).

The ability of this procedure to extract ecological patterns from data despite the addition of noise was recently shown in Borcard and Legendre (2002). They did this by first generating an ecological type dataset made up of 4 underlying distribution components, a single central bump, 4 waves, 17 waves, a positive linear trend, and some noise (Fig. 2). By using the RDA procedure they were able to identify four submodels, made up of one or more potential spatial distributions (Fig. 3*a*–3*d*). As seen in Fig. 3*e*, these submodels, which show the feature of all of the components used to generate the pseudo-ecological dataset, combine to show a distribution very similar to the one generated in Fig. 2. By running this procedure on many TE families, we will be able to detect TE families that have similar patterns underlying their distribution along a chromosome.

The RDA procedure is constrained analysis, meaning that the final results only show the variation in the TE families that are associated by a group of predictor variables—other variation is not shown. This is convenient for identifying the particular families that are affected by the ecological variables selected in this study. This is a multivariate extension of multiple regressions, with more than one dependent variable (number of TEs of each family in each window, or TE community) and several independent variables, namely spatial position and gene density (Dray et al. 2006). Analogous to a multiple regression, one can compute the amount of variation explained by each group of explanatory variables (adjusted R^2) (Peres-Neto et al. 2006), and the unique variation associated with each group of explanatory variables (Dray et al. 2006). In our case, this means that the amount of

Fig. 1. Potential spatial distributions produced by the PCNM function in the R software package. Each graph illustrates the potential relative abundances a transposable element family could be distributed along the windows of a chromosome, given the number of windows in that chromosome. Examples of (*a*) large-, (*b*) medium-, and (*c*) small-scale spatial patterns.



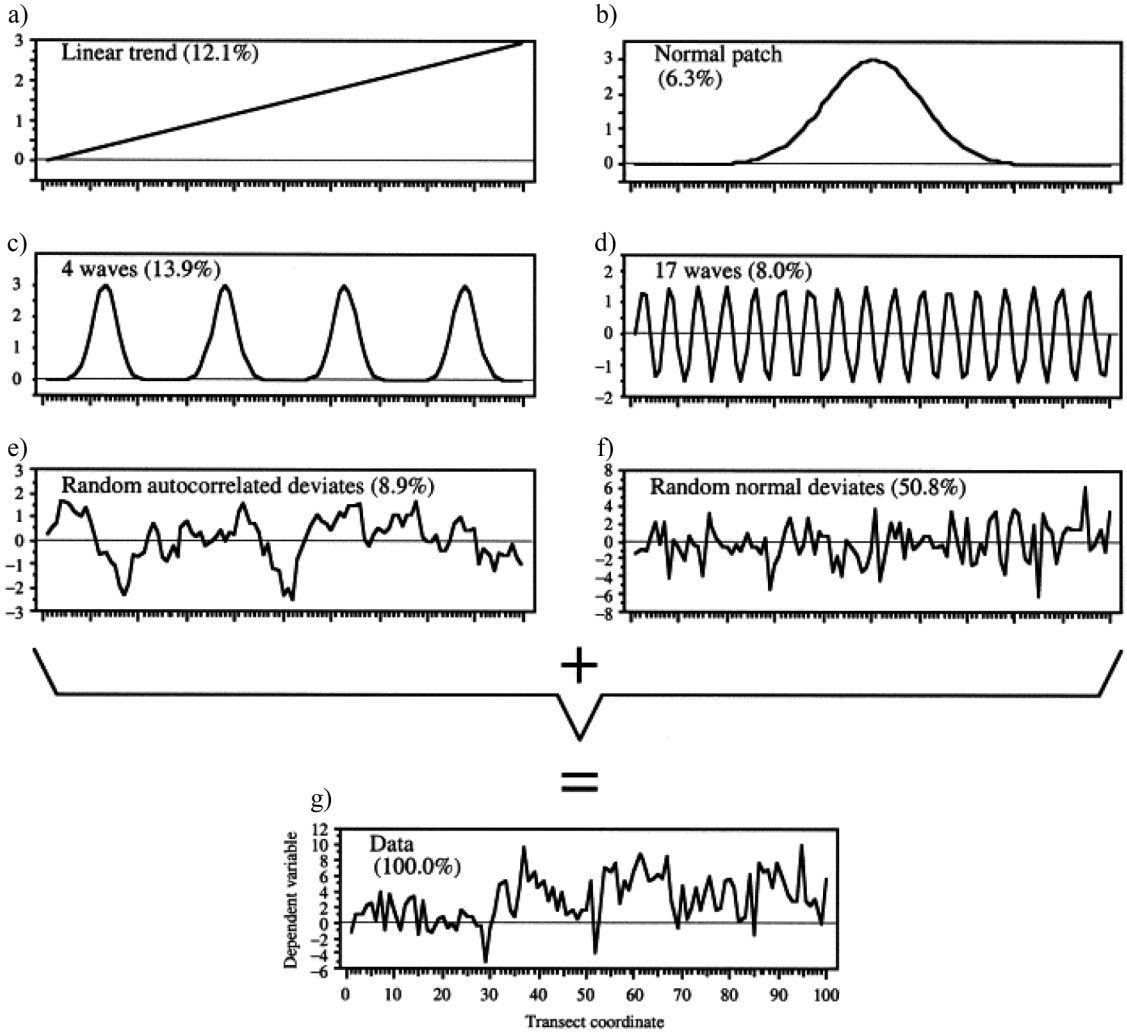
variation in abundance of each TE family can be explained by large-, medium-, and small-scale spatial patterns, as well as gene density. An RDA was run on the large-, medium-, and small-scale potential spatial distributions separately so that their significances could be tested separately. The significance of each of the three sets of potential spatial distributions was tested using a permutation procedure (Dray et al. 2006) to determine the amount of variation explained by the spatial location variables. The only difference between the RDA of gene density and the one used to analyze spatial variation is that the matrix of explanatory variables consisted of only one variable, the number of genes per window.

To simplify the task of interpreting the graphs, R scripts were written to extract the amount of variation explained by each analysis, the TE families for whom there was a significant amount of variation explained, and potential spatial distributions able to explain significant amounts of variation from the RDA results. After being extracted, the script then sorted the TE families, and the potential spatial distributions, by how much variation they were able to explain.

Interchromosome analysis

In a subsequent step, we performed an analysis to determine whether differences between chromosomes, such as chromosome

Fig. 2. Construction of an artificial pseudo-ecological dataset, taken from Borcard and Legendre (2002). The dataset consists of (*a*) a positive linear trend, (*b*) a normally distributed section in the middle of the distribution, (*c*) a distribution pattern consisting of four waves, and (*d*) a distribution containing 17 waves. These were all combined with *e* and *f*, which add an element of randomness to the data. Reproduced by permission of Elsevier.



length, gene number, and TE number, were able to account for differences in the results of RDAs performed at the chromosome level. Again, in this analysis each chromosome was treated in the same way that a different mountain within a mountain range would be treated in an ecological study. As we only have observational data, we performed the most basic type of comparisons: extracting a comparable statistic on how TE community structure responds to either spatial location or gene density at the intrachromosome analysis, and extracting predictor information at the chromosome level that could be associated with differences in these statistics. In this case, the statistics are results from the RDAs on the amount of variation in the TE community of a window explained by spatial predictor variables and gene density. A simple ANOVA determined if the results of the RDA could be explained by chromosome length, number of genes per chromosome, or number of TEs per chromosome. Considering there was a possibility that these three measures were correlated to each other, the interactions between the chromosome properties were also included in the analysis.

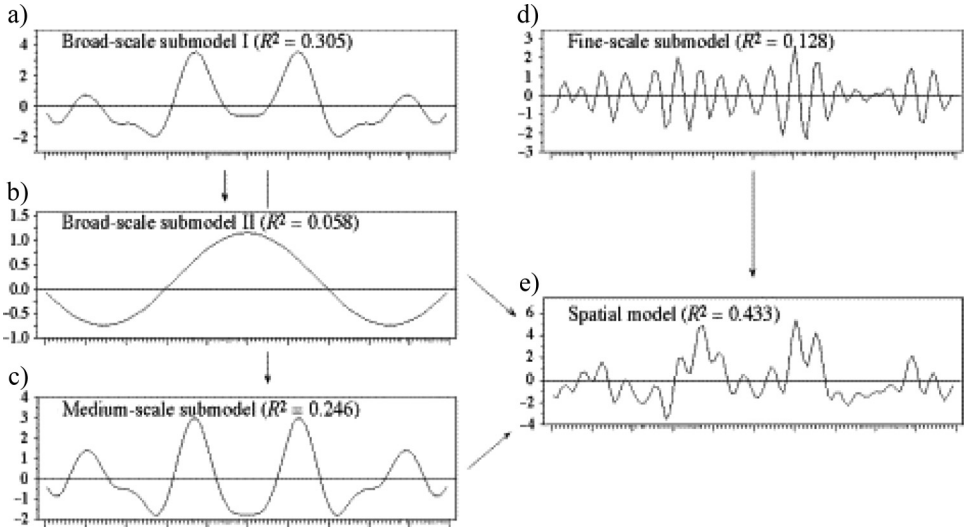
Results

Spatial intrachromosome analysis

Figure 4*a* shows the results of the RDA of large-scale spatial factors on X chromosome. Each arrow represents one of the potential large-scale spatial distributions generated by the PCNM procedure (Fig. 1*a*), and the length of the arrow corresponds to how much of the TE families distribution along the chromosome that potential spatial distribution is able to explain. Thus, longer arrows correspond to the potential spatial distributions that are able to explain the most of the distributions of a chromosomes TE families. The location of a TE family on the graph is determined by the amount of variation the potential spatial distributions corresponding to that arrow or combination of arrows are able to explain for that particular family. In this figure, the LINE/L1 family is located opposite to the arrow representing PCNM2 (Fig. 4*b*, whereas LINE/L2 and SINE/MIR are along the arrow representing PCNM8 (Fig. 4*c*). This means that the distribution of LINE/L1 on this chromosome is opposite the PCNM2, and there is indeed a small increase in the abundance of LINE/L1 elements in the center of chromosome 1 (Fig. 4*b*), and LINE/L2 and SINE/MIR should be opposite of PCNM8 (Fig. 4*c*). If the TE families had been in the same direction the arrows were pointing, we would expect them to match the PCNMs instead of being opposite to them. Figure 4*c* also shows that TE families close to each other in the graph produced by the RDA have similar distributions.

Using the RDAs of each chromosome, combined with the resulting graphs, we were able to determine that spatial patterns were

Fig. 3. Result of combining a redundancy analysis with the principal coordinates of neighbour matrices to conduct a spatial analysis on a pseudo-ecological dataset (Fig. 2), taken from Borcard and Legendre (2002). Distributions *a*–*d* illustrate submodels, made up of one or more potential spatial distributions, similar to those shown in Fig. 1. These submodels were identified by the analysis as being able to explain some of the variation in the pseudo-ecological dataset. When combined they show a distribution very similar to that of the pseudo-ecological dataset (e). Reproduced by permission of Elsevier.



found to explain a significant amount of variation within the TE community of a chromosome. Interestingly, only large-scale spatial patterns, those in which the TE community changes gradually over the many windows, were found to be significant, explaining 49.52% of the variation in the TE community across all 30 chromosomes. Neither small- nor medium-scale patterns, those in which the TE communities of neighbouring windows are not very similar, were found to be significant for any chromosome in the cow genome. To ensure the identification of all significant TE families, and spatial patterns from each analysis, the results were extracted and organized by amount of variation explained using the R scripts discussed in the Materials and methods. The script was run on the RDA of each of the *B. taurus* chromosomes, and the results can be found in Table 2.

Intrachromosome gene density analysis

RDAs were run on each of the *B. taurus* chromosomes to see if gene density had a significant impact on TE community composition. The gene density data was significantly able to explain some TE community variation in 29 of the 30 chromosomes (Table 3). In Fig. 5, the vector is pointing perpendicular to the axis along which the TE families are distributed, indicating that gene density is determining very little of the variation in the TE community, relative to the spatial analysis. This makes it very difficult to locate groupings along the gene density axis. As well as serving as a visual representation of the percent variation explained by gene density in each chromosome, Fig. 4 also illustrates that the amount of variation explained by gene density is at least one order of magnitude lower for each chromosome than that of the large-scale spatial patterns (Table 3).

Interchromosome analysis

An analysis was conducted to investigate chromosome properties that may influence the amount of variation explained by gene density and spatial location, on the composition of the TE community. First, the amount of variation explained by each analysis was extracted from the RDA results. This was done both individually, for each analysis, as well as for the proportion of variation explained by both analyses (Table 4). These values were compared with chromosome properties such as length, TE number per chromosome, and gene number per chromosome.

The amount of variation explained by the spatial analysis was positively related to the number of genes in a chromosome (P value = 0.036) such that as the number of genes in a chromosome increases, the analysis is able to explain more variation in that chromosome. None of the other chromosome properties had a significant effect on the explanatory power of the spatial analysis. The amount of variation explained by the gene density analysis was positively correlated to both gene and TE number per chromosome, but negatively correlated to chromosome length (P value = 0.038, 0.0001, and 0.0318, respectively). Many of the interaction terms between the genome properties were also significant, meaning that for instance, the effect on the explanatory power of TE number per chromosome and gene number per chromosome together was higher than the sum of their individual effects. The amount of variation explained by both the gene density analysis and the spatial analysis was also positively correlated to both gene and TE number per chromosome, and negatively related to chromosome length. Many of the interaction terms were also significant for all three genome properties (Table 5).

Alternate window sizes

To investigate the effect of varying window size on the above analysis, three additional window selection models were used. The first two generated the windows based on the number of TE in the chromosome in the same way as in the reported results. However, instead of using 100 TEs on average per window, they used 50 and 200 TEs per window. Additionally, a fixed window size of 120 155 bp per window, across all chromosomes, was also used.

The results of the spatial and gene density analysis were similar regardless of the method used to determine window size (Tables S2–S7). Each analysis reported largely the same TE families spatial distributions being explained among each chromosome, most often MIR, L1, or L2 elements. The potential spatial distributions were also very similar from model to model with the lower numbered spatial distributions, representing the largest scale patterns, explaining the most variation in TE communities.

The amount of the TE community variation explained by the spatial analysis ranged between models, with the 200 TE model explaining 55.23% of the TE community variation, followed by the 100 TE model, the fixed window model, and the 50 TE model,

Fig. 4. Results from the redundancy analysis (RDA) of the transposable element (TE) community of the X chromosome using large-scale spatial patterns as an explanatory variable. (*a*) RDA plot showing the TE families whose spatial distribution was able to be explained, with arrows representing potential spatial distributions (PCNMs). The magnitude of the arrow represents the amount of variation in the TE community, and the direction of the arrow is the axis along which that PCNM explains variation. The closer the TE families are found on the plot, the more similar their distributions are along the chromosome. RDA1 and RDA2 are unitless variables, similar to a principal component. (*b*) PCNM2 and the number of LINE/L1 elements in each window of the X chromosome. The LINE/L1 distribution is opposite to that of PCNM2 because it is along the same axis as the arrow for PCNM2, with the arrow pointing in the opposite direction. (*c*) PCNM8 and the distributions of LINE/L2 and SINE/MIR elements.

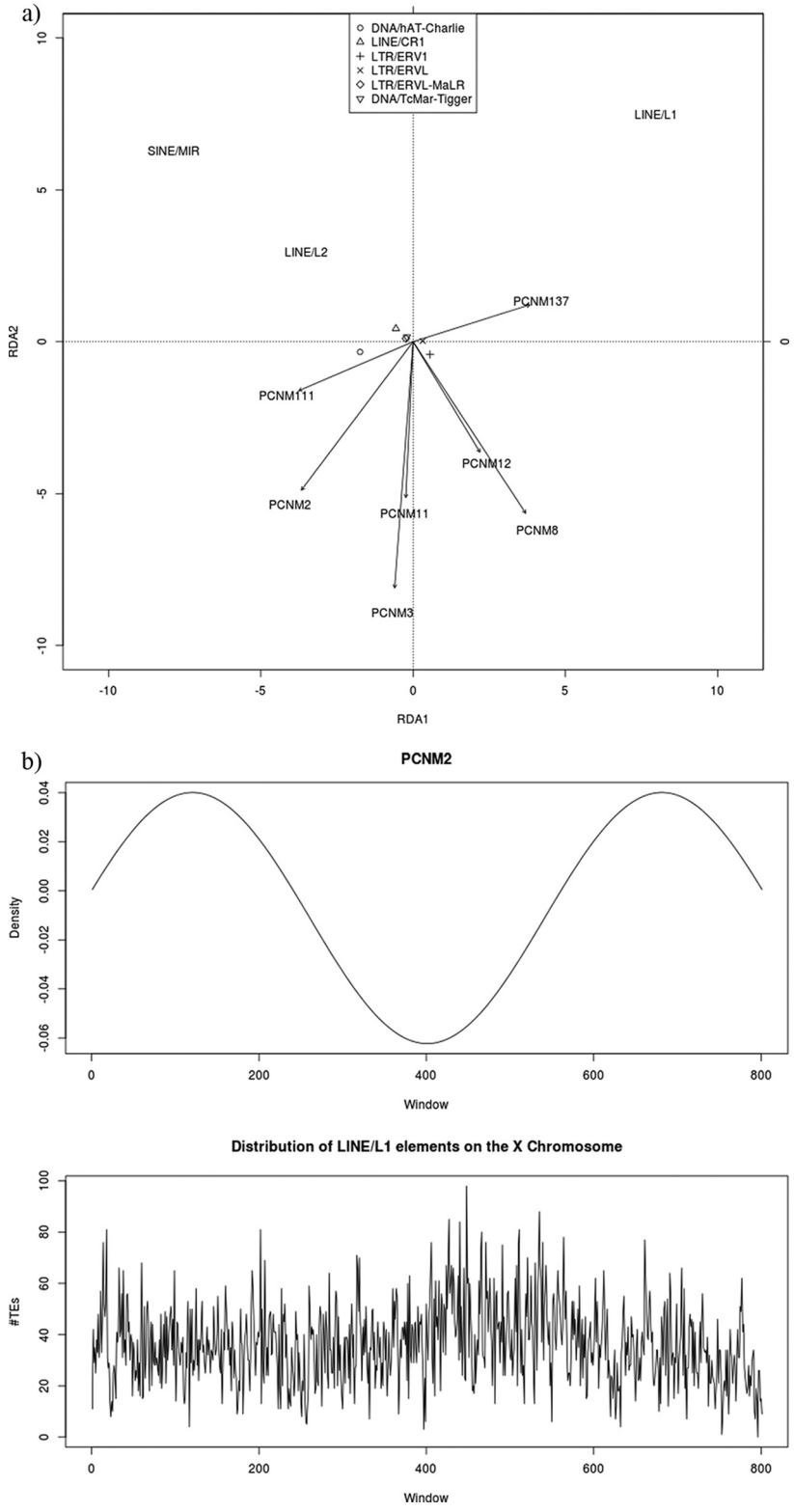
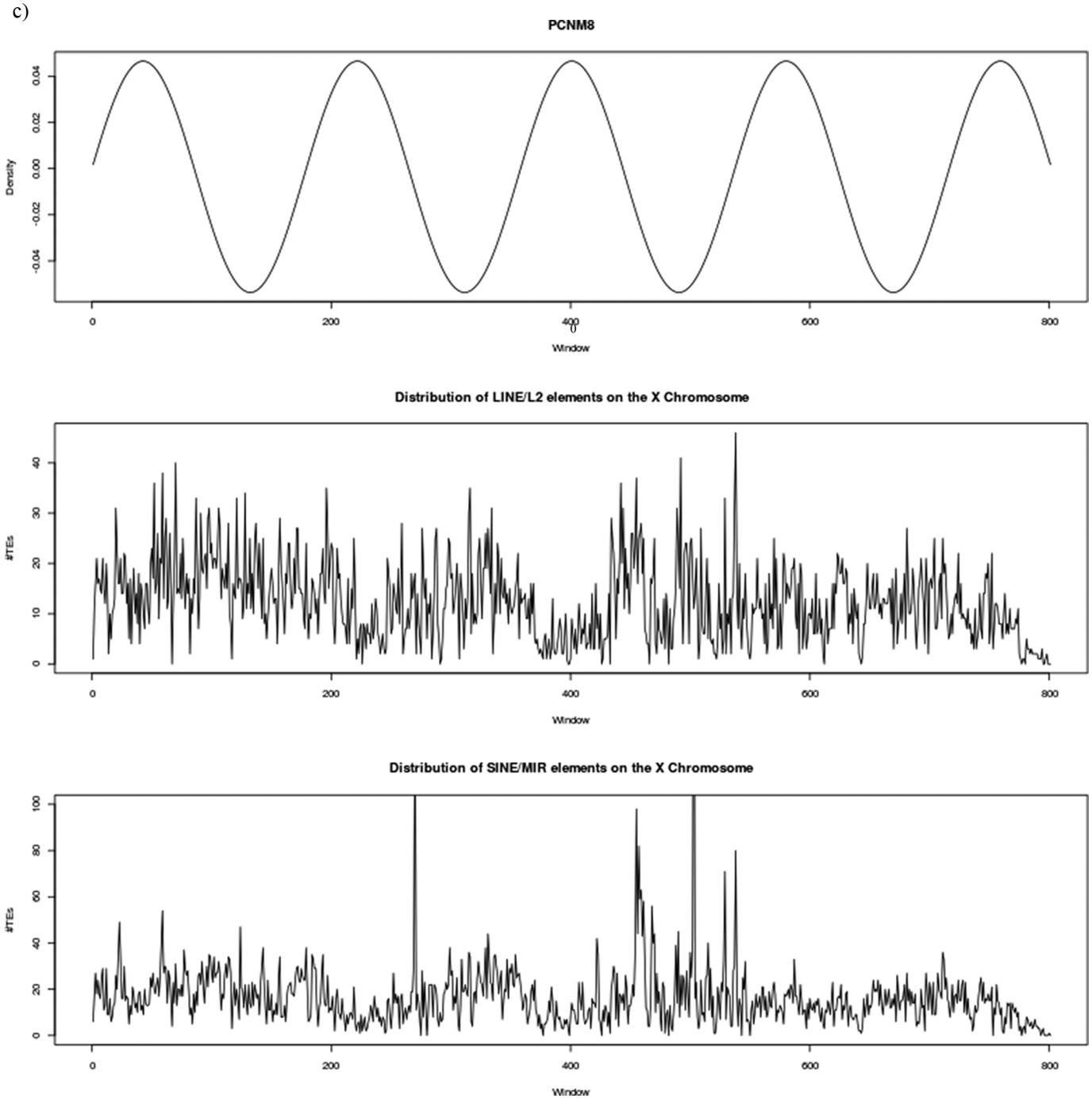


Fig. 4 (*concluded*).



explaining 48.07%, 47.66%, and 42.27% of the variation, respectively. The pattern among the gene density analysis was the same, with the 200 TE model explaining the most variation at 2.25% of the variation, followed by the 100 TE model, the fixed window model, and the 50 TE model, explaining 1.28%, 1.17%, and 0.75% of the variation, respectively.

Variation in the window selection method had more of an effect on the interchromosomal analysis (Tables S8–S10). Whereas decreasing windows did not have much of an effect on the results (Table S8), and fewer of the TE properties were found to be significant when the window size was increased (Table S9). Finally, when using a fixed window approach, far fewer of the chromosome properties were found to be significant (Table S10).

Discussion

Implications for the *Bos taurus* genome

There has been a growing level of enthusiasm for the application of ecological concepts and models at the level of the genome (Brookfield 2005; Abrusán and Krambeck 2006; Venner et al. 2009; Linquist et al. 2013). Ecology, at any level, investigates the factors influencing an entity's relationship to its environment. Previous results suggested that TE-ecological factors do in fact have an influence on TE abundance and distribution, especially over evolutionary short-time scales (Linquist et al. 2013). The current study builds upon this approach by adopting techniques traditionally used to identify environmental influences on multispecies commu-

Table 2. Redundancy analysis results of spatial analysis.

Chr.	Variation	TE families[a]	Potential spatial distribution[b]
1	42.62%	LINE/L1, SINE/MIR, LINE/L2, DNA/hAT-Charlie, LINE/CR1, LTR/ERV1, LTR/ERVL, LTR/ERVL-MaLR, DNA/TcMar-Tigger	3, 8, 2, 11, 12, 111, 137
2	61.74%	LINE/L1, SINE/MIR, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1, LTR/ERVL, DNA/TcMar-Tigger, LTR/ERV1, LINE/RTE	1, 4, 9, 2, 35, 8, 63, 21, 10, 13, 15, 37, 66, 33
3	52.01%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1, LTR/ERVL, DNA/TcMar-Tigger, LTR/ERV1, LINE/RTE, DNA/hAT-Tip100	1, 3, 15, 19, 2, 140, 112, 134
4	38.34%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LINE/CR1, LTR/ERVL-MaLR, LTR/ERVL, DNA/TcMar-Tigger, LTR/ERV1, LINE/RTE	10, 22, 1, 14, 21, 13, 19, 38, 59, 3, 20, 12, 15, 73, 4
5	48.42%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LINE/CR1, LTR/ERV1, LTR/ERVL-MaLR, LTR/ERVL, DNA/hAT-Tip100, DNA/hAT-Blackjack	1, 9, 12, 5, 46, 15, 56, 23, 4, 27, 58, 57, 44, 11, 39
6	45.13%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1, DNA/hAT-Tip100, LTR/ERVL, DNA/TcMar-Tigger, LTR/ERV1	4, 2, 3, 8, 18, 52, 6, 38, 43, 28, 67, 41, 53
7	52.70%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LTR/ERVL, DNA/TcMar-Tigger, LINE/CR1, DNA/hAT-Tip100	8, 2, 4, 14, 18, 26, 1, 13, 44, 17
8	51.72%	LINE/L1, SINE/MIR, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1, LTR/ERVL, DNA/TcMar-Tigger, DNA/hAT-Tip100, LTR/ERV1, LINE/RTE, DNA/hAT-Blackjack	1, 9, 4, 3, 49, 38, 2, 6, 105
9	32.29%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LTR/ERVL, LINE/CR1, DNA/hAT-Tip100, LTR/ERV1, LINE/RTE, DNA/hAT-Blackjack	4, 3, 1, 13, 6, 24, 18, 25
10	44.45%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LTR/ERVL, LTR/ERV1, LINE/CR1, DNA/hAT-Tip100	2, 5, 1, 9, 18, 52, 92, 118, 32, 38
11	48.04%	LINE/L1, SINE/MIR, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1, DNA/hAT-Tip100, DNA/TcMar-Tigger, LTR/ERV1	3, 10, 5, 9, 39, 15, 6, 54
12	45.84%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, DNA/TcMar-Tigger, LTR/ERVL, LINE/CR1, LTR/ERV1, DNA/hAT-Tip100	7, 1, 3, 4, 84, 5, 11, 8, 13, 51
13	58.49%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1, LTR/ERVL, DNA/TcMar-Tigger, DNA/hAT-Tip100, LTR/ERV1	4, 12, 5, 1, 6, 15, 20
14	52.87%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LTR/ERV1, DNA/hAT-Tip100, DNA/TcMar-Tigger, LINE/CR1, LTR/ERVL, LTR/ERVK	5, 7, 6, 13, 1, 4, 20, 56
15	57.64%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LTR/ERV1, DNA/hAT-Tip100, LINE/RTE, DNA/TcMar-Tigger, LINE/CR1, DNA/hAT	1, 2, 3, 12, 4
16	43.65%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1, LTR/ERVL, LTR/ERV1, DNA/TcMar-Tigger, DNA/hAT-Tip100, LINE/RTE	3, 4, 14, 13, 2, 9, 1, 32, 7, 19
17	58.43%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LTR/ERVL, LINE/CR1, LTR/ERV1, DNA/TcMar-Tigger, LINE/RTE, DNA/hAT-Tip100	3, 5, 2, 1, 9, 4, 40, 8, 10, 101, 46, 56
18	50.89%	SINE/MIR, LINE/L2, LINE/L1, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1, DNA/hAT-Tip100, DNA/TcMar-Tigger, LTR/ERVL, DNA/hAT, DNA/hAT-Blackjack, LTR/ERV1	1, 2, 3, 4, 24, 52, 12
19	48.27%	LINE/L1, SINE/MIR, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LTR/ERV1, LINE/CR1, LTR/ERVL, DNA/TcMar-Tigger, LINE/RTE, DNA/hAT-Tip100, DNA/hAT-Blackjack	8, 5, 25, 1, 4, 2, 9, 3, 11
20	55.21%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LTR/ERV1, LINE/CR1, LTR/ERVL, DNA/TcMar-Tigger, DNA/hAT-Tip100	1, 2, 17, 7, 3, 5, 28, 6, 26
21	49.32%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1, LTR/ERVL, DNA/TcMar-Tigger, DNA/hAT-Tip100, LTR/ERV1	8, 3, 5, 10, 6, 27, 2, 51, 4
22	45.05%	LINE/L1, SINE/MIR, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1, DNA/TcMar-Tigger, LTR/ERV1, LTR/ERVL, DNA/hAT-Tip100	4, 8, 18, 3, 1, 44, 22, 5, 14
23	38.19%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LINE/CR1, LTR/ERVL-MaLR, LTR/ERV1, LTR/ERVL, DNA/TcMar-Tigger	5, 1, 4, 13, 7, 6, 9, 29
24	45.10%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1, LTR/ERV1, LTR/ERVL, LINE/RTE, DNA/hAT-Tip100	2, 3, 1, 43
25	49.66%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LTR/ERV1, LTR/ERVL, LINE/CR1, DNA/TcMar-Tigger, DNA/hAT-Tip100, LINE/RTE, DNA/hAT, DNA/hAT-Blackjack	3, 9, 14, 5, 28, 35, 42
26	49.75%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1, LTR/ERVL, LTR/ERV1, DNA/TcMar-Tigger, DNA/hAT-Tip100	5, 10, 2, 3
27	38.89%	LINE/L1, SINE/MIR, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LTR/ERV1, LINE/CR1, DNA/TcMar-Tigger, LTR/ERVL, DNA/hAT-Tip100	13, 6, 2, 7, 14, 16, 9, 15, 51, 12
28	50.28%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LINE/CR1, LTR/ERVL-MaLR, LTR/ERV1, DNA/TcMar-Tigger, LTR/ERVL	1, 2, 6, 12
29	52.07%	SINE/MIR, LINE/L1, LINE/L2, DNA/hAT-Charlie, LTR/ERVL-MaLR, LTR/ERVL, LINE/CR1, LTR/ERV1, LINE/RTE, DNA/hAT-Tip100	5, 4, 7, 3, 9, 22, 8, 1, 2, 11, 23, 6, 13
X	34.83%	LINE/L1, SINE/MIR, LINE/L2, DNA/hAT-Charlie, LINE/CR1, LTR/ERVL, LTR/ERV1, DNA/TcMar-Tigger, LTR/ERVL-MaLR, DNA/hAT-Tip100	2, 3, 6, 1, 14, 16
Average	48.52%		

Note: Results from the redundancy analysis (RDA) of each of the 30 chromosomes. The RDA used transposable element (TE) community as the independent variable and the potential spatial distributions generated by the principal coordinates of neighbour matrices (PCNM) procedure for each chromosome as the dependent variables.

[a]Entries are ordered by how much of the distribution of each TE family it can explain (see, PCNMs from Fig. 4*a*).

[b]Entries are ordered by how much of the variation in that TE families distribution can be explained by the corresponding potential spatial distributions.

Table 3. Variation in transposable element community explained by gene density.

Chr.	Variation	TE families[a]
1	**1.47%**	LINE/L1, SINE/MIR, DNA/hAT-Charlie, LINE/L2, LTR/ERVL-MaLR, LTR/ERVL, LINE/RTE
2	**6.81%**	SINE/MIR, LINE/L2, LINE/L1, DNA/hAT-Charlie, LINE/RTE
3	**0.41%**	LTR/ERVL-MaLR, LINE/L2, LINE/L1, SINE/MIR, LTR/ERVL, DNA/hAT-Charlie
4	**0.19%**	SINE/MIR, DNA/hAT-Charlie, LINE/L2, LINE/L1
5	**1.21%**	SINE/MIR, LTR/ERVL-MaLR, LTR/ERVL, LINE/RTE
6	**1.07%**	LINE/L1, SINE/MIR, DNA/hAT-Charlie, LTR/ERVL, LTR/ERVL-MaLR, LINE/L2, DNA/TcMar-Tigger
7	**2.42%**	LINE/L2, SINE/MIR, LTR/ERVL, DNA/hAT-Charlie, LTR/ERV1, LINE/RTE, DNA/TcMar-Tigger, LINE/CR1
8	**0.31%**	LINE/L1, DNA/hAT-Charlie, LTR/ERVL, LTR/ERVL-MaLR, LINE/RTE, LINE/L2
9	**1.08%**	SINE/MIR, DNA/hAT-Charlie, LINE/L2
10	**0.57%**	LTR/ERVL-MaLR, LINE/L1, LINE/L2, LTR/ERVL, DNA/hAT-Charlie, SINE/MIR
11	**0.91%**	SINE/MIR, LTR/ERVL-MaLR, LINE/L2, LINE/L1
12	**1.82%**	LINE/L1, DNA/hAT-Charlie, SINE/MIR, LINE/L2, LTR/ERVL, LTR/ERV1, LINE/RTE, DNA/TcMar-Tigger
13	**2.31%**	SINE/MIR, LINE/L2, LINE/L1, LTR/ERVL-MaLR, DNA/hAT-Charlie
14	**0.43%**	SINE/MIR, LINE/L2, LTR/ERVL, LINE/L1, LTR/ERVL-MaLR, tRNA
15	**0.38%**	LTR/ERVL-MaLR, SINE/MIR, LINE/L2, LTR/ERVL, DNA/hAT-Charlie
16	**0.73%**	SINE/MIR, LINE/L1, LTR/ERVL-MaLR, LTR/ERVL, DNA/hAT-Charlie
17	**2.39%**	SINE/MIR, LTR/ERVL, LTR/ERVL-MaLR, LINE/L2, DNA/hAT-Charlie, LTR/ERV1
18	**1.60%**	LINE/L2, Satellite, SINE/MIR, LTR/ERVL, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/CR1
19	0.55%	LINE/L2, LINE/L1, LTR/ERVL-MaLR, SINE/MIR, DNA/hAT-Charlie
20	1.28%	LINE/L1, SINE/MIR, DNA/hAT-Charlie, LINE/L2, DNA/TcMar-Tigger, LTR/ERVL, LTR/ERVL-MaLR
21	**2.24%**	SINE/MIR, LINE/L2, DNA/hAT-Charlie, LINE/L1, LTR/ERVL, LTR/ERVL-MaLR, DNA/TcMar-Tigger
22	1.62%	LINE/L1, LTR/ERVL-MaLR, DNA/hAT-Charlie, SINE/MIR, LTR/ERVL
23	0.47%	LINE/L1, LTR/ERVL-MaLR, LINE/L2, LTR/ERV1, LTR/ERVL, SINE/MIR, DNA/TcMar-Tigger
24	1.06%	LINE/L1, DNA/hAT-Charlie, LTR/ERVL-MaLR, LINE/RTE
25	1.73%	LINE/L1, SINE/MIR, LTR/ERVL-MaLR, DNA/hAT-Charlie, LTR/ERVL, LINE/L2, DNA/TcMar-Tigger
26	0.84%	SINE/MIR, LINE/L1, LTR/ERVL-MaLR, DNA/hAT-Charlie, LINE/L2, LTR/ERVL
27	0.75%	LINE/L1, LTR/ERVL, DNA/hAT-Charlie, DNA/TcMar-Tigger, LTR/ERV1
28	0.38%	SINE/MIR, DNA/hAT-Charlie, LINE/L1, LINE/L2, DNA/TcMar-Tigger, LTR/ERVL-MaLR
29	**0.53%**	LINE/L1, LINE/L2, LTR/ERVL-MaLR, SINE/MIR
X	**0.80%**	SINE/MIR, LINE/L2
Average	1.28%	

Note: Percentage of the transposable element (TE) community variation that was explained by gene density as calculated by the redundancy analysis (RDA). Significant values are in bold.

[a]Entries are ordered by how much of the distribution of each TE family it can explain (see, PCNMs from Fig. 4*a*).

Fig. 5. Results from the redundancy analysis (RDA) of the transposable element (TE) community of the X chromosome using gene density as an explanatory variable. The arrow represents gene density along each window of the chromosome. TE families distributed perpendicular to the arrow are varying by a factor other than gene density. RDA1 is a unitless variable, similar to a principal component; PC1 is a principal component.

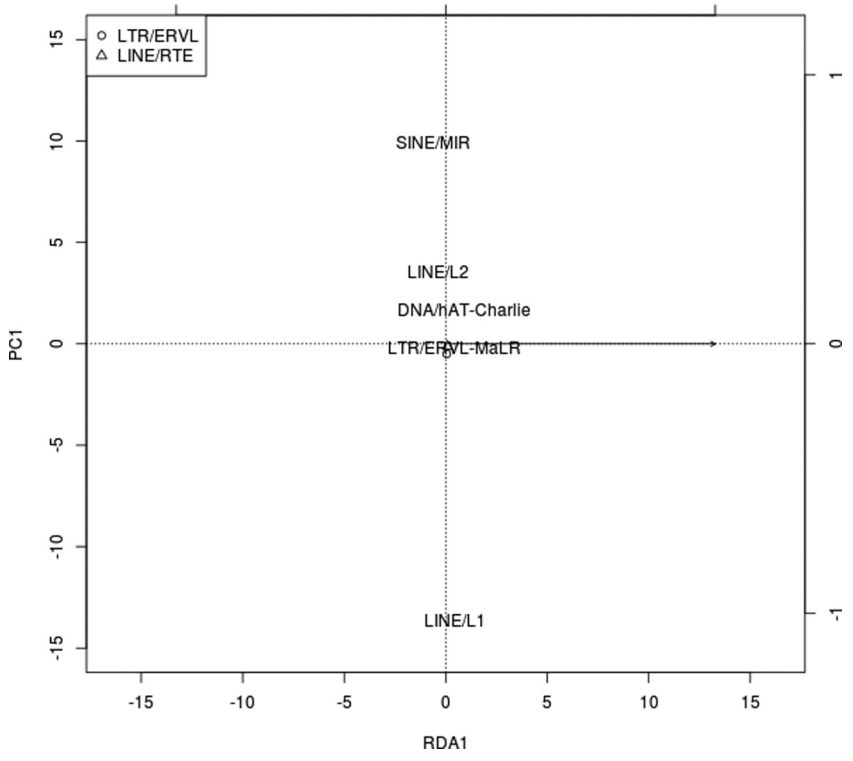


Table 4. Proportion of the variation in transposable element community explained by spatial and gene density analysis.

Chr.	Spatial	Combined	Gene density
1	0.426294757	0.014147199	0.014717876
2	0.617433888	0.0688884	0.068186268
3	0.520135096	0.000736762	0.004144985
4	0.383452654	0	0.0018518
5	0.48426803	0.005779276	0.012169762
6	0.451322783	0.01014396	0.010738951
7	0.52707503	0.022065685	0.024251873
8	0.517263148	0.003325016	0.003124366
9	0.322930005	0.008044735	0.010811862
10	0.444532609	0.001304887	0.005722629
11	0.480494065	0.008219626	0.009138669
12	0.458433578	0.015976445	0.018220236
13	0.584920517	0.020949355	0.023193524
14	0.528775383	0.005170558	0.004397368
15	0.576464256	0.004471257	0.003808855
16	0.436574357	0.007977225	0.007342175
17	0.584364957	0.023875651	0.02394532
18	0.50899372	0.014395521	0.016065077
19	0.482747832	0.006872255	0.005590254
20	0.552102841	0.013603691	0.012804755
21	0.493268535	0.018188904	0.022498538
22	0.450564691	0.016542598	0.016294668
23	0.381969333	0	0.00476579
24	0.451078642	0.010929556	0.010699995
25	0.496683019	0.012949679	0.017322606
26	0.497515117	0.007768834	0.008491701
27	0.388930227	0.008926153	0.0075839
28	0.502818591	0.004931733	0.003808852
29	0.520734818	0.007819161	0.00538416
X	0.348362084	0.002976043	0.008053451

Note: Summary of the 30 Venn diagrams, which displayed how much of the variation in transposable element (TE) community of each chromosome was partitioned. The Spatial and Gene Density columns report the variation in TE community explained by only that analysis, while the Combined column reports the amount of variation explained by both analysis.

nities. Here, we applied these techniques at the genomic level to identify features of the chromosomal environment that effect the composition of TE communities within the *B. taurus* genome.

Each chromosome was analyzed as if it were a linear transect, and each explanatory variable was treated as if it were an environmental gradient. A key difference between this application of transect analysis to genomes, compared with its traditional ecological application, is the high resolution afforded by genomic data. Our study analyzed each chromosome in its entirety, comparable to analyzing every individual tree found on each mountain over an entire mountain range. We thus have a complete description of TE communities—not a statistical sample as in the traditional ecological case. Using this method it was possible to assess which TE families are responsible for explaining the greatest amount of variation within a given chromosome and were also able to compare TE families between chromosomes. Interestingly, just 10 TE families (Table 2) explain over ~50% of the spatial distribution of TEs within each *B. taurus* chromosome. Among these 10 families, the LINE/L1, SINE, and LTR/ERVL families were identified as evolutionarily relevant by the *B. taurus* Genome Sequencing Consortium (Elsik et al. 2009). However, there were discrepancies between the evolutionarily relevant families identified by this consortium and those identified by our study as having an impact on TE community composition. It is likely that these other TE families influence the evolution of the *B. taurus* genome in ways that are not significantly related to the spatial distribution. SINE/MIR, LINE/L2, and LINE/L1 significantly explained some of the variation in each of the 30 chromosomes, and DNA/hAT-Charlie, LTR/ERVL-MaLR, LTR/ERVL, DNA/TcMar-Tigger, and LINE/CR1 were able to explain a significant amount of the spatial variation in at least one chromosome (Table 2). Such chromosome-specific influences of location on community composition raise interesting questions including the extent to which different chromosomes constitute distinct environments. This is a topic for future consideration.

One of the most striking findings of this study is that large-scale spatial patterns are able to explain more than half of the variation among TE communities within chromosomes. The predominant spatial pattern, as identified by the RDA, involved a gradual change in community composition as one moves from one chromosomal tip to the center, and then another gradual change from the center to the other tip (Fig. 4*b*). This spatial pattern is potentially explained by the distribution of highly heterochromatic regions, such as telomeres and centromeres, in which TEs are thought to be found (Wong and Choo 2004). Additionally, using the graphs produced by RDA, TE families that have similar distributions are easily identified, as they will be close to each other on the graph. This was the case for LINE/L1 and SINE/MIR elements on the X chromosome.

Given the mutagenic effects of transposable elements and previous evidence for the preferential insertion of particular TE families in both gene rich and gene poor regions (Pearce et al. 1996; Biessmann and Mason 1997; Takahashi et al. 1997; Zhang et al. 2000, 2011; Naito et al. 2009), it is surprising that gene density, although significant, was not able to explain more of the variation among TE communities. It was possible to identify other chromosomal properties that help explain when gene density has an effect on the TE community. We found that chromosome length, TE number, and gene number were all able to explain the degree to which gene density was informative. This result is not surprising, as the chromosome properties used in the intrachromosome analysis are related to each other, as evidenced by the significance of the interaction terms in the ANOVA.

Implications for future genome analyses

Using this method provides an alternative starting point for any genomic TE analysis—to get a first look into which TEs families are shaping the genome. It will also help to determine which of the many interacting genomic environmental factors are potentially responsible for shaping a genome's TE distribution. An ecological approach to analyzing this genome revealed that most of the spatial variation within TE communities is explained by variation within a limited number of TE families. This approach also identified the relative importance of gene density as an environmental variable, as well as detecting some of the properties of the chromosome that affect gene density. Additionally, depending on the number of windows in a given chromosome, the computing time for executing this method on a genome the size of *B. taurus* is less than 1 h, which makes this an efficient approach for analyzing whole genomes.

However, the application of an ecological framework to genomic data requires further refinement. For instance, the memory requirement for computing the PCNM goes up with the size of the distance matrix (i.e., the number of windows). Because of these requirements, the window size used in this analysis may not have been sufficiently small to capture small-scale interactions, such as a pair of TEs that always appear together. A possible solution to this is to add a measure of the average physical distance between each of the TE families in a window into the analysis. This would allow the detection of small-scale interactions that this analysis may have missed. It is also evident that the properties of the chromosome selected for the analysis were very highly correlated. To shed light on the relationship between these chromosome properties and the composition of the TE community, several methods of calculating the window size were used. These all produced similar results for the spatial and gene density analysis. Changing from a method of generating window size per chromosome to a similarly sized fixed window across all chromosomes produced a slight increase in the amount of variation explained; however, the overall results were very similar. By running the

Table 5. Interchromosomal ANOVA results, 100 transposable elements per window in each chromosome.

	Spatial					Gene					Combined				
	df	SS	MS	*F*	Pr >*F*	df	SS	MS	*F*	Pr >*F*	df	SS	MS	*F*	Pr >*F*
Length	1	0	1×10^{-6}	0	0.9863	1	0.00031	0.00031	5.25	0.03189*	1	0.00022	0.00022	3.8	0.0641
TE per chr.	1	0.00648	0.00648	1.462	0.2395	1	0.00125	0.00125	21.037	0.00014***	1	0.00196	0.00196	33.681	7.72×10^{-6}***
Genes per chr.	1	0.02195	0.02195	4.95	0.0367*	1	0.00029	0.00029	4.83	0.03879*	1	0.00032	0.00032	5.542	0.0279*
Length: TE per chr.	1	0.00166	0.00166	0.374	0.5474	1	0.00026	0.00026	4.324	0.04945*	1	0.00018	0.00018	3.119	0.0912
Length: Genes per chr.	1	0.00694	0.00694	1.565	0.2241	1	0.00013	0.00013	2.181	0.15393	1	9.9×10^{-5}	9.9×10^{-5}	1.709	0.2046
TE per chr. : genes per chr.	1	0.00529	0.00529	1.194	0.2864	1	0.00075	0.00075	12.59	0.0018**	1	0.0007	0.0007	12.004	0.0022**
Length: TE per chr. : genes per chr.	1	0.00507	0.00508	1.145	0.2963	1	0.00018	0.00018	3.047	0.09483	1	0.00016	0.00016	2.68	0.1159
Residuals	—	0.09754	0.00443	—	—	22	0.00131	5.9×10^{-5}	—	—	22	0.00128	5.8×10^{-5}	—	—

Note: Results of the ANOVA comparing the amount of variation explained by the spatial and genome size analysis, as well as the variation explained by both (combined variation), properties of each chromosome (chr.). This analysis was done with a fixed window size that would contain an average of 100 transposable elements (TEs) per window in a particular chromosome. df, degrees of freedom; SS, sum of squares; MS, mean square; *F*, *F* value; Pr >*F*, probability of obtaining a greater than *F* statistic value. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

analysis with different average numbers of TEs per window, we were also able to show that the amount of variation explained increases as window size goes up. This is most likely because all of the most significant potential spatial distributions are large-scale patterns, and increasing the window size has the effect of smoothing out the distribution across the chromosome. Varying the window size did have more of an impact of the interchromosomal analysis. Decreasing the window size did not have much of an effect on measuring how chromosome properties affect how much variation can be explained within a chromosome, while increasing the window size decreased the ability of the analysis to detect a relationship between the number of genes in a chromosome and the amount of variation explained. The fixed window analysis still detected a relationship between some of the chromosome properties in the gene density analysis, although at a much lower level of significance. This suggests that normalizing TE density across chromosomes may overestimate the effect of correlated chromosome properties.

The MIR1, L1, and L2 families possessed variation that was most significantly correlated with gene density and spatial distribution (Table 2). This means that these families tended to be inserted in specific patterns across a chromosome. The MIR1 and L2 families are both ancient components of the cow genome, and all are eutherian genomes, and have not actively transposed since the Mesozoic era (Smit and Riggs 1995). Analysis of the average size of all three families corroborated their ancientness, as few insertions were the size of the consensus repeat for each family. MIR elements have been found associated with genes in the human genome and have been implicated with contributing to human evolution through the exaptation of MIR-derived sequences in several genes (Murnane and Morales 1995; Tulko et al. 1997). This may be the case in the cow genome as well, although this would require further investigation. There is evidence for recent L1 activity in the cow genome (Girardot et al. 2006). Our current analysis did not allow us to break up families into subfamilies, so subfamilies had their variation explained best by gene density and spatial patterns could not be addressed, so whether active L1 subfamilies were responsible could not be addressed. The most recently active families (Bov-B and ART2A) identified by Elsik et al. (2009) did not show up as being heavily influenced by gene density or spatial distribution in our analysis.

Questions in transposable element ecology

In addition to the promise of methodological improvements to the implementations of this type of analysis, there are many other questions that could be answered using this method. One possible source of questions would be to gather more detailed information on the TEs that make up the TE community. For example, by adding information on TE age it would be possible to gain many insights into the evolutionary history of the TE community and the genome it resides in. This information could be of great help to genome biologists and sequencing consortia, many of which look at TEs and its history largely to connect them with the evolutionary history of the host organism (Bennetzen 2000; Kazazian 2004; Elsik et al. 2009). Age of the different families would also be crucial for establishing recently active families based on low intrafamily sequence divergence, along with supplementary information about known, active families and data concerning the integrity of repeats and protein coding regions of elements. Based on the operational definitions of ecology and evolution used by Linquist et al. (2013) currently or recently active TE families would be the most appropriate subjects when looking at the ecology of TE communities in the genome. Future work should focus on analyzing these active families in the methodological framework we have laid out to obtain a better picture of the dynamics in a current TE ecosystem. How the dynamics of the TE community have changed over time could also be investigated by restricting analysis to fossil TE families known to be active at the same time.

Another way to improve ecological analysis would be to use additional genomic environmental variables. TEs have been implicated in both the formation and the possible origin of heterochromatic regions (Miller et al. 1999). By adding data on indicators of heterochromatin formation, which is becoming more readily available (Rohde et al. 2008), it may be possible to gain many insights into the dynamics of this relationship. One of the best known examples of a host genome co-opting TEs for cellular function is that of V(D)J recombination, which is responsible for creating variation among antibodies and relies on a recombinase derived from transposable elements (Kapitonov and Jurka 2005; Panchin and Moroz 2008). Using the method above, combined with gene ontology (GO) terms, which provide functional classification for each gene, would make it possible to detect relationships between specific TEs and certain types of genes. This in turn, could lead to the detection of new instances of TEs being co-opted for cellular function.

Acknowledgements

This work was made possible by funding from the Natural Sciences and Engineering Research Council of Canada to K.C., T.R.G., and, S.C.K.; an Ontario Graduate Scholarship in Science and Technology to B.S.; and by access to computational resources provided by SHARCNET. The authors wish to thank the anonymous reviewers and the Editors whose comments greatly improved an earlier draft of the paper.

References

Abrusán, G., and Krambeck, H.-J. 2006. Competition may determine the diversity of transposable elements. Theor. Popul. Biol. **70**(3): 364–375. doi:10.1016/j.tpb.2006.05.001. PMID:16814337.

Bellen, H.J., Levis, R.W., He, Y., Carlson, J.W., Evans-Holm, M., Bae, E., et al. 2011. The Drosophila gene disruption project: progress using transposons with distinctive site specificities. Genetics, **188**(3): 731–743. doi:10.1534/genetics.111.126995. PMID:21515576.

Bennetzen, J.L. 2000. Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. **42**(1): 251–269. doi:10.1023/A:1006344508454. PMID:10688140.

Biessmann, H., and Mason, J.M. 1997. Telomere maintenance without telomerase. Chromosoma, **106**: 63–69. doi:10.1007/s004120050225. PMID:9215555.

Borcard, D., and Legendre, P. 2002. All-scale spatial analysis of ecological data by means of principal coordinated of neighbour matrices. Ecol. Modell. **153**: 51–68. doi:10.1016/S0304-3800(01)00501-4.

Borcard, D., Legendre, P., Avois-Jacquet, C., and Tuomisto, H. 2012. Dissecting the spatial structure of ecological data at multiple scales. Ecology, **85**(7): 1826–1832. doi:10.1890/03-3111.

Brookfield, J.F.Y. 2005. The ecology of the genome mobile DNA elements and their hosts. Nat. Rev. Genet. **6**(2): 128–136. doi:10.1038/nrg1524. PMID:15640810.

Dray, S., Legendre, P., and Peresneto, P. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecol. Model. **196**(3–4): 483–493. doi:10.1016/j.ecolmodel.2006.02.015.

Elsik, C.G., Tellam, R.L., and Worley, K.C. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science, **324**(5926): 522. doi:10.1126/science.1169588. PMID:19390049.

Fontanillas, P., Hartl, D.L., and Reuter, M. 2007. Genome organization and gene expression shape the transposable element distribution in the *Drosophila melanogaster* euchromatin. PLoS Genet. **3**: e210. doi:10.1371/journal.pgen.0030210. PMID:18081425.

Girardot, M., Guibert, S., Laforet, M.P., Gallard, Y., Larroque, H., and Oulmouden, A. 2006. The insertion of a full-length *Bos taurus* LINE element is responsible for a transcriptional deregulation of the Normande *Agouti* gene. Pigment Cell Res. **19**(4): 346–355. doi:10.1111/j.1600-0749.2006.00312.x. PMID:16827753.

Gregory, T.R. 2005. Synergy between sequence and size in large-scale genomics. Nat. Rev. Genet. **6**(9): 699–708. doi:10.1038/nrg1674. PMID:16151375.

Griffith, D.A., and Peres-Neto, P.R. 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. Ecology, **87**(10): 2603–2613. doi:10.1890/0012-9658(2006)87[2603:SMIETF]2.0.CO;2. PMID:17089668.

Houle, D., and Nuzhdin, S.V. 2004. Mutation accumulation and the effect of *copia* insertions in *Drosophila melanogaster*. Genet. Res. **83**: 7–18. doi:10.1017/S0016672303006505. PMID:15125062.

Kapitonov, V.V., and Jurka, J. 2005. RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. PLoS Biol. **3**(6): e181. doi:10.1371/journal.pbio.0030181. PMID:15898832.

Kazazian, H.H. 2004. Mobile elements: drivers of genome evolution. Science, **303**(5664): 1626–1632. doi:10.1126/science.1089670. PMID:15016989.

Kidwell, M.G., and Lisch, D.R. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution, **55**(1): 1–24. doi:10.1554/0014-3820(2001)055[0001:PTEPDA]2.0.CO;2. PMID:11263730.

Lajeunesse, M.J. 2010. Achieving synthesis with meta-analysis by combining and comparing all available studies. Ecology, **91**(9), 2561–2564. doi:10.1890/09-1530.1. PMID:20957949.

Langley, C.H., Montgomery, E., Hudson, R., Kaplan, N., and Charlesworth, B. 1988. On the role of unequal exchange in the containment of transposable element copy number. Genet. Res. **52**: 223–235. doi:10.1017/S0016672300027695. PMID:2854088.

Legendre, P., and Legendre, L. 1998. Numerical ecology. 2nd ed. Elsevier Science, Amsterdam.

Le Rouzic, A., Dupas, S., and Capy, P. 2007. Genome ecosystem and transposable elements species. Gene, **390**: 214–220. doi:10.1016/j.gene.2006.09.023. PMID:17188821.

Linquist, S., Saylor, B., Elliott, T.A., Kremer, S., Cottenie, K., and Gregory, T.R. 2013. Distinguishing ecological from evolutionary approaches to transposable elements. Biol. Rev. Camb. Philos. Soc. **88**(3): 573–584. doi:10.1111/brv.12017. PMID:23347261.

Liu, S., Yeh, C.-T., Ji, T., Ying, K., Wu, H., Tang, H.M., et al. 2009. *Mu* transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet. **5**(11): e1000733. doi:10.1371/journal.pgen.1000733. PMID:19936291.

Maside, X., and Bartolomé, C. 2004. The lack of recombination drives the fixation of transposable elements on the fourth chromosome of *Drosophila melanogaster*. Genet. Res. **83**: 91–100. doi:10.1017/S0016672304006755. PMID:15219154.

Miller, W.J., McDonald, J.F., Nouaud, D., and Anxolabéhère, D. 1999. Molecular domestication—more than a sporadic episode in evolution. Genetica, **107**(1–3): 197–207. doi:10.1023/A:1004070603792. PMID:10952213.

Murnane, J.P., and Morales, J.F. 1995. Use of a mammalian interspersed repetitive (MIR) element in the coding and processing sequences of mammalian genes. Nucleic Acids Res. **23**(15): 2837–2839. doi:10.1093/nar/23.15.2837. PMID:7659505.

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C.N., Richardson, A.O., et al. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature, **461**(7267): 1130–1134. doi:10.1038/nature08479. PMID:19847266.

Panchin, Y., and Moroz, L.L. 2008. Molluscan mobile elements similar to the vertebrate Recombination-Activating Genes. Biochem. Biophys. Res. Commun. **369**: 818–823. doi:10.1016/j.bbrc.2008.02.097. PMID:18313399.

Pasyuokva, E.G., Nuzhdin, S.V., Morozova, T.V., and MacKay, T.F.C. 2004. Accumulation of transposable elements in the genome of *Drosophila melanogaster* is associated with a decrease in fitness. J. Hered. **95**(4): 284–290. doi:10.1093/jhered/esh050. PMID:15247307.

Pearce, S.R., Pich, U., Harrison, G., Flavell, A.J., Heslop-Harrison, J.S., Schubert, I., and Kumar, A. 1996. The Ty1-*copia* group retrotransposons of *Allium cepa* are distributed throughout the chromosomes but are enriched in the terminal heterochromatin. Chromosome Res. **4**(5): 357–364. doi:10.1007/BF02257271. PMID:8871824.

Peres-Neto, P.R., Legendre, P., Dray, S., and Borcard, D. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology, **87**(10): 2614–2625. doi:10.1890/0012-9658(2006)87[2614:VPOSDM]2.0.CO;2. PMID:17089669.

Rohde, C., Zhang, Y., Jurkowski, T.P., Stamerjohanns, H., Reinhardt, R., and Jeltsch, A. 2008. Bisulfite sequencing data presentation and compilation (BDPC) web server—a useful tool for DNA methylation analysis. Nucleic Acids Res. **36**(5): 2–7. doi:10.1093/nar/gkn083. PMID:18296484.

Sela, N., Kim, E., and Ast, G. 2010*a*. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. Genome Biol. **11**: R59. doi:10.1186/gb-2010-11-6-r59. PMID:20525173.

Sela, N., Mersch, B., Hotz-Wagenblatt, A., and Ast, G. 2010*b*. Characteristics of transposable element exonization within human and mouse. PLoS One, **5**(6): e10907. doi:10.1371/journal.pone.0010907. PMID:20532223.

Smit, A.F.A., and Riggs, A.D. 1995. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Res. **23**(1): 98–102. doi:10.1093/nar/23.1.98. PMID:7870595.

Smit, A., Hubley, R., and Green, P. 2004. RepeatMasker. Available from http://www.repeatmasker.org [accessed September 2011].

Stewart, C., Kural, D., Strömberg, M.P., Walker, J.A., Konkel, M.K., Stütz, A.M., et al. 2011. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet. **7**: e1002236. doi:10.1371/journal.pgen.1002236. PMID:21876680.

Takahashi, H., Okazaki, S., and Fujiwara, H. 1997. A new family of site-specific retrotransposons, *SART1*, is inserted into telomeric repeats of the silkworm, *Bombyx mori*. Nucleic Acids Res. **25**(8): 1578–1584. doi:10.1093/nar/25.8.1578. PMID:9092665.

Tulko, J.S., Korotkov, E.V., and Phoenix, D.A. 1997. MIRs are present in coding regions of human genes. DNA Seq. **8**(1–2): 31–38. PMID:9522118.

Venner, S., Feschotte, C., and Biémont, C. 2009. Dynamics of transposable elements: towards a community ecology of the genome. Trends Genet. **25**(7): 317–323. doi:10.1016/j.tig.2009.05.003. PMID:19540613.

Whittaker, R.H. 1960. Vegetation of the Siskiyou mountains, Oregon and California. Ecol. Monogr. **30**(3): 279–338. doi:10.2307/1943563.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., et al. 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8**(12): 973–982. doi:10.1038/nrg2165. PMID:17984973.

Wong, L.H., and Choo, K.H.A. 2004. Evolutionary dynamics of transposable elements at the centromere. Trends Genet. **20**(12): 611–616. doi:10.1016/j.tig.2004.09.011. PMID:15522456.

Zhang, Y., and Mager, D.L. 2012. Gene properties and chromatin state influence the accumulation of transposable elements in genes. PLoS One: **7**: e30158. doi:10.1371/journal.pone.0030158. PMID:22272293.

Zhang, Q., Arbuckle, J., and Wessler, S.R. 2000. Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family *Heartbreaker* into genic regions of maize. Proc. Natl. Acad. Sci. U.S.A. **97**(3): 1160–1165. doi:10.1073/pnas.97.3.1160. PMID:10655501.

Zhang, Y., Romanish, M.T., and Mager, D.L. 2011. Distributions of transposable elements reveal hazardous zones in mammalian introns. PLoS Comput. Biol. **7**(5): e1002046. doi:10.1371/journal.pcbi.1002046. PMID:21573203.