Detecting evolutionary strata on the human X chromosome: Markov Segmentation and Clustering Analysis

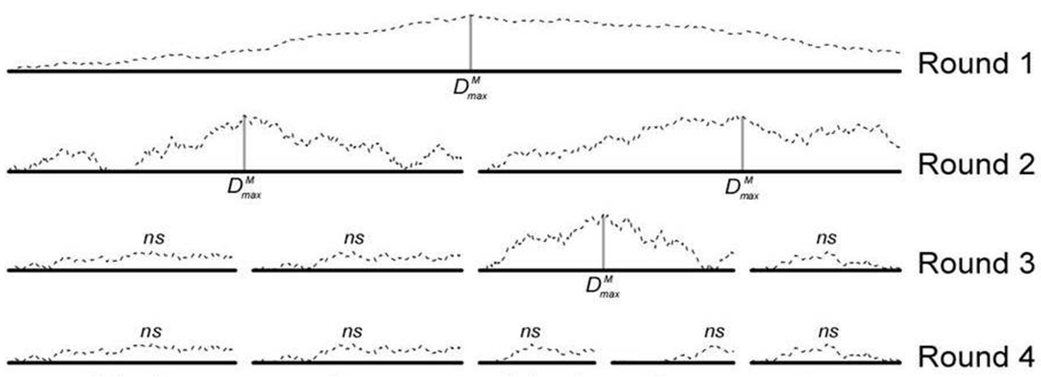
Melissa A. Wilson Sayres[1], Ravi Shanker Pandey[2], and Rajeev K. Azad[2]
[1]University of California-Berkeley, [2]University of North Texas

INTRODUCTION

In mammalian sex chromosomes, a stratum is formed on the X chromosome following the cessation of X-Y recombination. In the absence of homologous recombination, the non-recombining regions accumulate DNA elements, such as, transposable or repetitive elements, and sequences with shifts in GC content and thus the sequence composition of each strata diverges from the regions on the X that still undergo X-Y recombination. To delineate the stratum boundaries, we developed a multi-pronged strategy that applies a recursive segmentation and an agglomerative clustering algorithm within the framework of statistical hypothesis testing to identify compositionally distinct regions on the X chromosome.

SEGMENTATION AND CLUSTERING METHOD

1. Given a genome sequence, the compositional difference between subsequences to the left and right of each sequence position is measured. The sequence is segmented into two parts at the position where this difference is maximized, provided the split is statistically significant. This procedure is followed recursively for each of the resulting subsequences.



D_{max} refers to the point of maximum difference between the left and right sides and *ns* denotes "not significant".

2. Hyper-segmentation at high stringency is used to identify the segment boundaries more precisely. The segmental structure is restored by grouping contiguous similar segments, with a reduced stringency. Similar segments are recursively grouped together until all resulting clusters are significantly distinct from each other.

3. The Jensen-Shannon divergence, based on Shannon information entropy, is used to measure the compositional difference between DNA sequences S_1 and S_2 of length l_1 and l_2 as,

$$D(S_1, S_2) = H(S_1 \oplus S_2) - ((l_1/(l_1+l_2))H(S_1) + (l_2/(l_1+l_2))H(S_2)),$$

Where the Shannon entropy, $H(.) = -\Sigma_b p_b \log_2 p_b$, p_b denotes the probability of nucleotide, b.

4. A generalization of this measure to account for short range correlations in the nucleotide ordering was obtained recently within the framework of Markov chain model of order m, defined as,

$$D^m(S_1, S_2) = H^m(S_1 \oplus S_2) - ((l_1/(l_1+l_2))H^m(S_1) + (l_2/(l_1+l_2))H^m(S_2)),$$

Here, conditional probability function, $H^m(.) = -\Sigma_w p(w) \Sigma_b p(b|w) \log_2 p(b|w)$, w denotes oligonucleotide of length m preceding the nucleotide b, $p(w)$ is the probability of oligonucleotide w and $p(b|w)$ is the probability of nucleotide b given the preceeding oligonucleotide w.

5. The statistical significance of the maximum value of D^m can be assessed from the probability distribution of D_{max}, which was shown to approximate a X^2 distribution function with fitting parameters β and λ,

$$P(D^m{}_{max} \leq x) \approx \{X^2{}_v[2(l_1+l_2)(\ln 2)x\beta]\}^{\lambda}$$

β and λ were estimated by fitting the above analytic expression to the empirical distributions obtained via Monte Carlo simulations.

REFERENCES

Lahn BT, Page DC (1999) Four evolutionary strata on the human X chromosome. Science 286: 964-967.

Azad RK, Li J (2013) Interpreting genomic data via entropic dissection. Nucleic Acids Res 41: e23.

Pandey RS*, Wilson MA*, Azad RK (2013) Detecting evolutionary strata on the human X chromosome in the absence of gametologous Y-linked sequences. Genome Biol Evol.

RESULTS

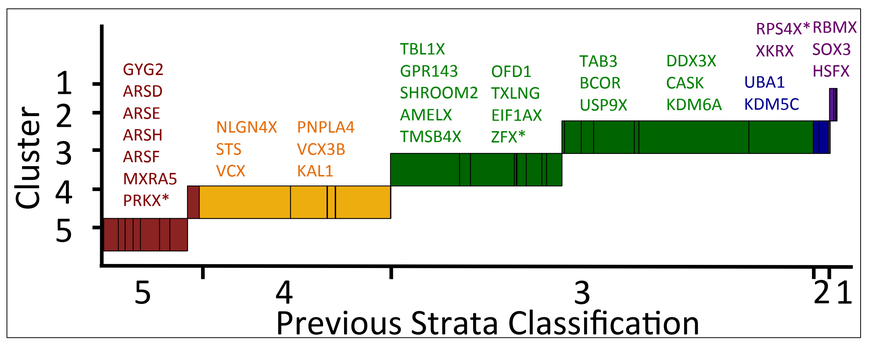


Figure 1. Strata identified using previously assayed X-linked genes. Here we apply the segmentation and clustering algorithm to a concatenated string of the X-linked genes that have been previously assayed using inversion, phylogenetic and substitution rate analyses. Previous strata are colored: 5-Red, 4-Yellow, 3-Geen, 2-Blue, 1-Violet. Genes that span cluster boundaries are marked with a star. We used Markov model of order 2 to perform segmentation and clustering at significance thresholds of 0.3 and 0.04 respectively.

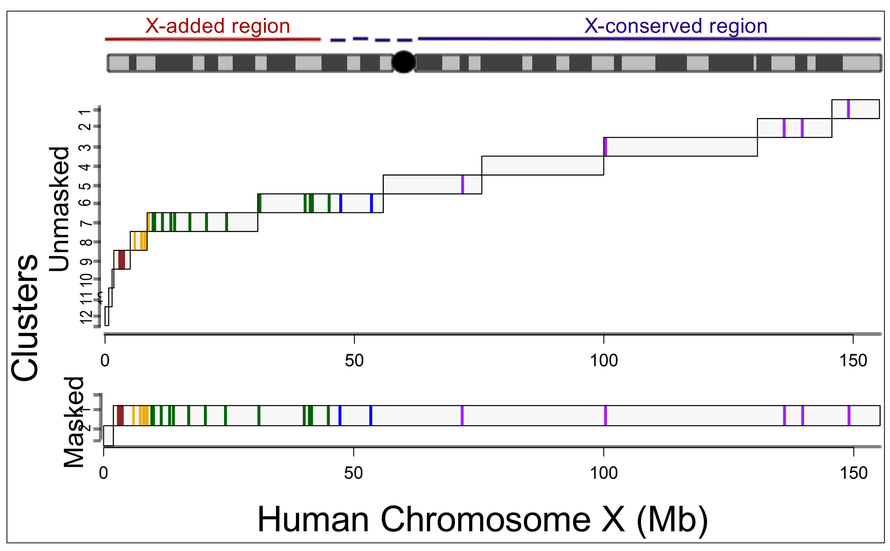


Figure 2. Strata identified across the whole X chromosome. Here we show the clusters that are determined using the entire sequence of the human X chromosome, either unmasked or masked for repetitive elements, as defined by RepeatMasker. We also plot the position and strata delineation of X-linked genes that have previously been assayed, as in Fig. 1. We used Markov model of order 2 to perform segmentation and clustering at significance thresholds of 0.4 and 10^{-7} respectively.

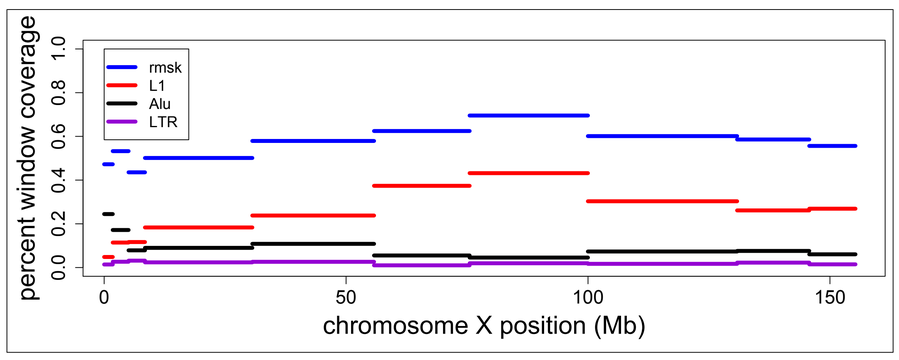


Figure 3. Density of repetitive elements or genes across the clusters identified on the X chromosome. Here we show the difference in the feature density between clusters: A) density of genes, repetitive elements, CpG islands, and simple repeats is plotted for each cluster; and, B) the repetitive element density is plotted for each cluster along with the L1, LTR, and Alu subsets of repetitive element.

CONCLUSIONS

We present a novel method for unbiased detection of evolutionary strata which is not limited by the availability of gametologous sex chromosome sequences. Our proposed approach does not depend on *a priori* information, and perhaps because of this, yields unbiased estimates of stratum boundaries. We envisage the applicability of our "unsupervised" method to sequences of the homogametic (X or Z) sex chromosome with unknown stratum history. Understanding where these strata begin and end can inform on the history of sex chromosome evolution and elucidate mechanisms driving dosage compensation.

ACKNOWLEDGEMENTS

This work was supported by Beth Baird graduate student scholarship to RSP, Miller fellowship from the Miller Institute for Basic Research in Science to MAWS and a faculty start up fund from the University of North Texas to RKA.