MARKOV MODEL OF SEGMENTATION AND CLUSTERING: APPLICATIONS IN DECIPHERING GENOMES AND METAGENOMES

Ravi Shanker Pandey, M.S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2017

APPROVED:

Rajeev K Azad, Major Professor
Armin Mikler, Committee Member
Vladimir Shulaev, Committee Member
Pamela Padilla, Committee Member
Pudur Jagadeeswaran, Committee Member
Art Goven, Chair of Department of Biological Sciences
David Holdman, Dean of College of Arts and Sciences
Victor Prybutok, Dean of the Toulouse Graduate School

Pandey, Ravi. *Markov Model of Segmentation and Clustering: Applications in Deciphering Genomes and Metagenomes*. Doctor of Philosophy (Biology), August 2017, 135 pp., 6 tables, 35 figures, 151 numbered references.

Rapidly accumulating genomic data as a result of high-throughput sequencing has necessitated development of efficient computational methods to decode the biological information underlying these data. DNA composition varies across structurally or functionally different regions of a genome as well as those of distinct evolutionary origins. We adapted an integrative framework that combines a top-down, recursive segmentation algorithm with a bottom-up, agglomerative clustering algorithm to decipher compositionally distinct regions in genomes. The recursive segmentation procedure entails fragmenting a genome into compositionally distinct segments within a statistical hypothesis testing framework. This is followed by an agglomerative clustering procedure to group compositionally similar segments within the same framework. One of our main objectives was to decipher distinctive evolutionary patterns in sex chromosomes via unraveling the underlying compositional heterogeneity. Application of this approach to the human X-chromosome provided novel insights into the stratification of the X chromosome as a consequence of punctuated recombination suppressions between the X and Y from the distal long arm to the distal short arm. Novel "evolutionary strata" were identified particularly in the X conserved region (XCR) that is not amenable to the X-Y comparative analysis due to massive loss of the Y gametologs following recombination cessation. Our compositional based approach could circumvent the limitations of the current methods that depend on X-Y (or Z-W for ZW sex determination system) comparisons by deciphering the stratification even if only the

sequence of sex chromosome in the homogametic sex (i.e. X or Z chromosome) is available. These studies were extended to the plant sex chromosomes which are known to have a number of evolutionary strata that formed at the initial stage of their evolution, presenting an opportunity to examine the onset of stratum formation on the sex chromosomes. Further applications included detection of horizontally acquired DNAs in extremophilic eukaryote, *Galdieria sulphuraria*, which encode variety of potentially adaptive functions, and in the taxonomic profiling of metagenomic sequences. Finally, we discussed how the Markovian segmentation and clustering method can be made more sensitive and robust for further applications in biological and biomedical sciences in future.

ACKNOWLEDGEMENTS

It is great pleasure to acknowledge all who have helped me during my doctoral work. I extend my deepest regards to my mentor, Dr. Rajeev Azad, who provided support and guidance throughout the work. His feedback on my scientific approaches as well as on presentations has helped me to improve considerably. He has always given valuable opinion on my personal and professional problems. I would like to thank the members of my dissertation committee, Dr. Armin Mikler, Dr. Vladimir Shulaev, Dr. Pamela Padilla and Dr. Pudur Jagadeeswaran for being part of my committee and for providing valuable advice on my research. I would also like to thank my two previous committee members Dr. Qunfeng Dong and Dr. Xiang Gao. I am extremely thankful to my collaborators Dr. Melissa Wilson Sayres, Dr. Debashish Bhattacharya and Dr. Huan Qiu. I am thankful to the Toulouse Graduate School, the College of Arts and Sciences and the Department of Biological Sciences for providing me scholarship opportunities which have helped me to cover my tuition fees and attend conferences. I am also thankful to David Burks and Garima Saxena for working with me on my projects. It has been a great journey and it would have not been possible without love and support given by lab members and other graduate students in the department. My family has been important pillars of support as they gave the freedom and support to pursue my PhD. research. Their faith, love and confidence in my capacities have been the greatest strength of my life. Finally, I thank almighty God who brought me to a platform where I could understand the broader meaning of education. Section 3.2 and 3.3 in chapter 3 is reproduced with permission from Springer, and materials at page 76-79 and 84-88 in chapter 4 is reproduced with permission from John Wiley and Sons.

TABLE OF CONTENTS

Page

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

DATA TABLES AND FIGURES INCLUDED AS SUPPLEMENTARY MATERIALS

Supplementary Tables

Supplementary Table 1. GC content (%) in clustered regions on the human [42].

Supplementary Table 2. Boundaries of predicted strata and collinear region on the papaya X [88].

Supplementary Table 3. Boundaries of predicted pseudoautosomal region and strata on the *S. latifolia* X [88].

Supplementary Table 4. Boundaries of predicted strata and collinear region on the chromosome 19 of *P. trichocarpa* [88].

Supplementary Table 5. Summary of the clusters for the *Ectocarpus sp*. male sex chromosome (chromosome V) that were obtained by the application of the segmentation and clustering algorithm to a concatenated sequence of genes, and the annotated SDR and PARs from a previous study [88].

Supplementary Table 6. Boundaries of the clusters on the *Ectocarpus sp*. chromosome V that were obtained by the application of the segmentation and clustering algorithm to a concatenated sequence of genes [88].

Supplementary Table 7. Summary of the clusters for the *Ectocarpus sp*. chromosome V that were obtained by the application of the segmentation and clustering algorithm to the sequence of chromosome V, and the annotated SDR and PARs from a previous study [88].

Supplementary Table 8. Boundaries of the clusters on the *Ectocarpus sp*. chromosome V that were obtained by the application of the segmentation and clustering algorithm to the sequence of chromosome V [88].

Supplementary Table 9. Boundaries of the PARs and predicted strata on the mating type chromosome a2 of *M. lychnidis-dioicae* [88].

Supplementary Table 10. Boundaries of the PARs and predicted strata on the mating type chromosome a1 of *M. lychnidis-dioicae* [88].

Supplementary Table 11. Frequency of repeat elements in clusters (predicted strata and collinear region) on the chromosome 19 of *P. trichocarpa* [88].

Supplementary Table 12. Frequency of repeat elements in clusters (predicted strata and collinear region) on the papaya X [88].

Supplementary Table 13. Frequency of repeat elements in the clusters (predicted strata

and pseudoautosomal region) on the *S. latifolia* X [88].

Supplementary Figures

Supplementary Figure 1. Comparison with segmentation only analysis [42].

Supplementary Figure 2. Clusters on concatenation of X-linked genes using alternative parameter values [42].

Supplementary Figure 3. Clusters on whole X chromosome using alternative parameter values [42].

Supplementary Figure 4. The segmentation and clustering algorithm was applied to the repeat masked sequence of chromosome 19 of *P. trichocarpa* at the same parameter setting as for unmasked sequences of chromosome 19 of *P. trichocarpa* (Fig. 3.7) [88].

Supplementary Figure 5. The segmentation and clustering algorithm was applied to the repeat masked concatenated sequence of papaya X-linked genes at same parameter setting as for unmasked sequences of papaya X-linked genes (Fig. 3.5) [88].

Supplementary Figure 6. Evolutionary strata identified using repeat masked concatenated sequence of papaya X-linked genes at a less stringent segmentation and clustering parameter setting [88].

Supplementary Figure 7. Density of repetitive elements within clusters (predicted strata and collinear region) on the chromosome 19 of *P. trichocarpa* [88].

Supplementary Figure 8. Density of repetitive elements within clusters (predicted strata and collinear region) on the papaya X chromosome [88].

Supplementary Figure 9. Accuracy in classifying reads generated using Metasim with exact (no error) model. Performance of whole metagenome profiling methods, shown in different colors, was assessed as a function of sequence read length and taxonomic level (genus and higher) via cross-validation by masking the test-sequence originating species in the database. Accuracy was obtained as the percentage of reads classified correctly by a method.

Supplementary Figure 10. Same as in Supplementary Fig. 9, but for test datasets generated using Metasim with Sanger error model.

Supplementary Figure 11. Same as in Supplementary Fig. 9, but for test datasets generated using Metasim with 454 error model.

Supplementary Figure 12. Same as in Supplementary Fig. 9, but for Phymm test datasets.

Supplementary Figure 13. Same as in Supplementary Fig. 9, but for hybrids for Phymm and BLAST (PhymmBL), and SGM and BLAST (SGMBL). Accuracies for component methods are also shown.

Supplementary Figure 14. Same as in Supplementary Fig. 9, but for hybrids for Phymm and BLAST (PhymmBL), and SGM and BLAST (SGMBL) assessed on test datasets generated using Metasim with Sanger error model. Accuracies for component methods are also shown.

Supplementary Figure 15. Same as in Supplementary Fig. 9, but for hybrids for Phymm and BLAST (PhymmBL), and SGM and BLAST (SGMBL) assessed on test datasets generated using Metasim with 454 error model. Accuracies for component methods are also shown.

Supplementary Figure 16. Same as in Supplementary Fig. 9, but for hybrids for Phymm and BLAST (PhymmBL), and SGM and BLAST (SGMBL) assessed on Phymm test datasets. Accuracies for component methods are also shown.

CHAPTER 1

INTRODUCTION

The genome of an organism can either be linear (in eukaryotes) or circular (in prokaryotes) DNA molecule, which is composed of four distinct nitrogenous bases adenine, thymine, cytosine and guanine and can be represented as a string of the letters A, T, C, G respectively [1]. A genome harbors a variety of distinctive and diverse features, all of which taken together constitute the “blueprint” for an organism [2]. Some of these features could be functionally important such as protein coding sequences, regulatory sequences, operons, promoters. Large structures such as isochores in vertebrate genomes and evolutionary strata on sex chromosomes have special evolutionary significance. In addition, CpG islands in vertebrate genomes or horizontally acquired genomic islands in prokaryotic genomes have been reported to display atypical composition and have specific functional significance [1].

With the development of sophisticated and efficient sequencing technologies, the complete genome sequences of several thousand species are now available, providing opportunities to perform analysis and interpretation of genomes both at the individual and comparative level [3]. One of the goals is to uncover the underlying patterns of organization and elucidate the compositional structure of a genome. Seeking insights into the organization of nucleotides has been an important problem, and such analysis is possible with the present computational resources. Given a genome, is it possible to robustly decipher its heterogeneity and leverage this to uncover novel structural and functional entities? What do the large structures within genomes signify? These are some broader questions that motivated the research problems pursued in this thesis.

1.1 Heterogeneity in DNA Sequences

The bases are distributed non-uniformly in genomic sequences of any organism. However, segments of DNA sequences can be homogeneous within, with respect to specific statistical or biological properties, but heterogeneous with respect to each other.

Microbial genomes have evolved through acquisition of DNAs from different lineages and are often littered with large numbers of recently-acquired genes [4]. A major force of rapid genomic change is the genomic island, a group of tens to hundreds of genes acquired through the evolutionary process of horizontal gene transfer. Among the first kind of genomic islands to be described were pathogenicity islands, so named because virulence genes in many pathogens were found to be not only physically clustered within the chromosome but also bear signs of recent acquisition such as unusual nucleotide composition [4].

On the other hand, in the eukaryotic genomes, distinctive regions often consist of sequence repeats. There are mainly two types of repeats present in genome, the tandem repeats that are simple sequences repeated continuously and the interspersed repeats that are complex sequences repeated over a stretch of DNA interspersed with non-repetitive DNA. Both these regions are compositionally distinct from the non-repetitive DNA. Interspersed repeats usually contain elements like transposons and tRNA genes while tandem repeats harbor satellite DNA and repeated ribosomal gene arrays. These regions are functionally distinct from non-repetitive DNA that harbors genes [5].

With respect to the distribution of the repeats, a genome can be divided in two regions: low complexity regions (LCR) and high complexity regions (HCR). A majority of

the coding regions of genomes are localized in the high complexity regions. Sequence heterogeneity of eukaryotic genomes is exemplified by the large variety of compositionally distinct parts of the genomes.

1.2 Possible Reasons for Sequence Heterogeneity

Base composition in genomic DNA obeys two parity rules. The first, which is strictly obeyed, is that in a double–stranded DNA molecule globally %A = %T and %G = %C, which constitutes the basis of Watson-Crick base pairing. The second rule, which is often violated, is that in each of the strands of DNA, %A ≈ %T, and %G ≈ %C [6, 7]. Furthermore, there could be asymmetry in the nucleotide composition in either AT or GC. Such deviations, generally termed skewness or strand asymmetry, result from differences in mutational and selection pressures as a consequence of replication and transcriptional processes [8].

The substitution rates of the nucleotides are not uniform across a single genome [9]. Genes that are highly conserved will have a lower nucleotide substitution rate as compared to genes that are highly variable. In the human X chromosome, genes on the X conserved region (XCR) are more conserved than genes on the recently added regions (X added region, XAR). Genes that are functionally important, for example, gene encoding histone, have a very high conservation from yeast to mammals. Conversely, genes encoding IgG protein have a tremendous repertoire of variants.

In addition, evolutionary events such as DNA rearrangements, insertions, deletions and mutations are important factors that create compositional variations along the sequence through their actions; for example, the compositional differences that are

apparent in the laterally transferred DNAs against native sequences in bacterial organisms. In complex eukaryotic organisms like mammals, chromosomes have evolved through various evolutionary processes including DNA rearrangements over the course of time. DNA sequences residing in a particular location such as on a specific chromosome for a longer time will have a nucleotide composition that is distinct from DNA sequences that have been acquired recently by the same chromosome. In the human X chromosome, for example, genes on the XCR are more conserved than genes on recently added regions XAR. These observations suggest that methods that can detect and analyze these compositional variations in a meaningful manner may prove useful in uncovering the origin of the functional variation as well.

1.3 Biological Importance

A number of sequence features of biological relevance can be characterized by analyzing differences in nucleotide composition within stretches of the genomic DNA sequences. One of the most important measures of nucleotide composition is the G+C content of any sequence. A structural feature defined based on G+C content is the “isochore”. The isochores are defined as regions of the genome having a similar G+C percentage and have lengths greater than 300 Kbp [10]. However, differences in base composition are not restricted to isochores alone as regions of the genome that are functionally different often tend to have distinct compositions. This is evident from the correlation of small and medium sized genes within regions of genomes that have a high G+C content [11-13]. The frequency of occurrence of repeats like LINES, SINES, Alu, etc., is also known to be correlated with the G+C content [13, 14].

Another type of regions in the genome that are known to be compositionally distinct based on their G+C composition is the CpG Islands or CGIs. CpG islands are usually defined as regions with at least 200 bp of sequence that has a GC percentage above 50 and an observed/expected ratio of the dinucleotide CG greater than 0.6 [15]. It is well known that CpG islands or CGIs have a very distinct functional role in mammalian gene expression because these are the sites for DNA methylation, which further play a crucial role in deciding the inactivation status of gene. For biologically important reasons, methylation process is suppressed in short stretches of genome, such as around the promoters or start regions of many genes. In these regions, CpG islands are observed more than elsewhere. Lyon has proposed that long interspersed repeat elements (L1) could correlate with X-inactivation patterns. L1 concentration is inversely correlated with the proportion of genes that escape inactivation [16].Thus a study of sequence composition is a key to identifying biologically relevant sequence features.

1.4 Measures of Sequence Heterogeneity

Statistical tools have been used often to uncover the underlying patterns of nucleotide organization, specifically, to identify compositionally homogeneous regions of variable lengths from heterogeneous genomes. These regions could be related to specific biological features or attributes such as those relevant to CpG islands, isochores, etc.

A simple way to examine the base composition patterns across genomes is to compare relative abundances of G+C (sometimes the A+T content is studied instead;

this is, of course, entirely equivalent). There is wide variation in the GC content across genomes regardless of the kingdom. In particular, prokaryotes show wide variation, ranging from as little as 25% to as much as 75% [17]. The reasons for this wide variation are not well understood, although there are a number of hypotheses. One explanation is that high temperature conditions may favor high GC content to provide thermal stability of the genome (as G and C pairing involves three hydrogen bonds compared to A and T pairing which involves only two) [17]. Likewise, remarkable variations in GC content have also been observed within genomes of many prokaryotes and eukaryotes.

Beyond G+C content, genomic DNA also displays distinctive patterns in dinucleotide composition. One way to study this is by computing relative dinucleotide abundances [18], namely, the odds ratios:

$$\rho_{XY} = f_{XY}/f_X\, f_Y$$

Where f_{XY} is the relative frequency of dinucleotide XY observed in the sequence, f_X and f_Y are the relative frequencies of the nucleotides X and Y, X, Y ∈ {A, T, C, G}. Significant deviation of ρ_{XY} from 1 indicates under-representation or over-representation of XY. Although extreme dinucleotide under-representation or over-representation has been observed both in prokaryotes as well as eukaryotes, instances are typically specific to individual taxonomic groups [18]. In particular, under-representation of TA in most genomes and of CG in vertebrates is a striking observation. Analysis carried out for oligomers also show characteristic patterns in the occurrence of certain oligomers [18]. The ρ_{XY} values for fragments (of length ≥ 50Kb) from same genome or from an evolutionary closer genome were found to be either equal or very closer compared to

those from distantly related ones [18]. Based on this observation, it was proposed that the set of ρ_{XY} values constitutes a genomic signature which can provide an alternate measure of phylogenetic relationship.

In the recent decades, a number of approaches have been employed to obtain a set of optimal segments that reveal the details of genomic organization. Elementary techniques such as those based on a sliding window have been used to assess the compositional heterogeneity, however, this procedure is sensitive to the size of the window.

The more sophisticated and quite promising idea to assess heterogeneity is to segment heterogeneous DNA sequences into distinct homogeneous subsequences with respect to given statistical criteria. The objective of this approach is to segment a DNA sequence into biologically meaningful regions. Several statistical models of DNA sequence segmentation have been proposed, where the compositional homogeneity is determined via the analysis of oligonucleotides distributions. Models that have been used in different segmentation algorithms are mostly based on either a hidden Markov model (HMM) [19, 20] or a multiple change point approach [21].

A key concept is the Shannon entropy [22], which is a measure of level of uncertainty about the states of a system. DNA can be considered as a “text” represented as a string of the letters A, T, C, and G. When the sequence is comprised of only one kind of nucleotide, say TTTTTTT, this is clearly a state of maximum order while sequences generated in a purely random manner, say AGGTCACAGTCA, are perhaps highly disordered. Information theory finds applications in diverse biological problems such as characterizing DNA binding sites [23], studying genetic polymorphisms [24],

protein evolution [25], correlations between amino acid residues in protein sequences [26, 27, 28], informational complexity of ribozymes and its implication on the functional activity [29], augmenting combinatorial drug designing [30], information content of protein sequences [31], biological complexity and its evolution [32, 33], and so on.

Bernaola-Galvan et al. [34] employed the Jensen-Shannon divergence measure to obtain homogeneous regions in a DNA sequence by partitioning the sequence recursively. The segmentation procedure is carried out until certain halting criteria are met. This measure captures the extent to which the (constituent) subsequences differ from the combined sequence. If the divergence between the resulting subsequences is significantly higher, then the sequence is split into two subsequences. This method finds application in a multitude of problems including identification of CpG islands, characterization of isochores, location of replication origin, and location of between coding and noncoding regions.

Another segmentation model based on Bayesian statistics entails optimal partitioning based on a likelihood function [35]. The “optimal” segments are further filtered by merging the less significant boundaries to obtain a set of longer segments.

The results of different segmentation methods are not always in concordance; different methods divide a sequence into different segments, each showing uniformity of composition as measured under the chosen criteria and thus suggests that the use of any single method is probably insufficient in uncovering the organizational complexities of DNA sequences. Azad et al. [21] have shown that the entropic segmentation method based on higher order Markov models has helped considerably in uncovering the underlying structure of genomes. Kelker et al. [36] used this segmentation approach to

detect evolutionary strata boundaries, but restricting recursive segmentation to certain number of steps.

An important limitation of the genome segmentation methods has been that the biological significance of structures identified by them often remains unknown. Attempts have been made to extend the genome segmentation as a way of coarse graining to evolve it into a method that can detect structures with similar properties [37]. Introduction of non-hierarchical agglomerative clustering, after recursive segmentation, has allowed removal of undesirably numerous segment boundaries and grouping of segments with similar properties into distinct clusters [37]. The present work involves application of such a method [37], namely Markov model of segmentation and clustering based on Jensen-Shannon divergence, which is discussed in detail in Chapter 2.

1.5 Focus and Organization of the Dissertation

The main focus of this dissertation is the study of genome organization using mathematical and statistical methods for genome segmentation to uncover biologically significant features. The availability of genome sequence data has enabled such studies. The other emphasis is on deciphering metagenomes, specifically taxonomic profiling of metagenomic reads, via segmentation and clustering of genomes and metagenomes. The integrative framework of recursive segmentation and clustering decomposes the compositional heterogeneity present in genome sequences by generating compositionally homogenous clusters of apparently biologically significant features. Importantly, this procedure allows visualization of the underlying segmental structures at different hierarchical levels, making possible investigating disparate

evolutionary events that have shaped the genomes via segmentation implemented at different levels of stringency. We therefore adapted this methodology to uncover biologically significant genomic features via assessment of compositional heterogeneity and demonstrated how this approach enabled decoding the underlying genome structures at different scales. Where possible we compared this approach to other methods in the respective fields that are often constrained by sparse sampling of related organisms or unavailability of homologous sequences. The methodology for genome analysis that was adapted in these studies is described in Chapter 2.

One of our main objectives is to decipher distinctive evolutionary patterns in sex chromosomes via unraveling the underlying compositional heterogeneity. Identifying the regions of recombination suppression between homologous sex chromosomes (e.g. between X and Y or Z and W), namely the “evolutionary strata”, is central to understanding the history and dynamics of sex chromosome evolution. Precise delineation of stratum structure on human X-chromosome may also help in identification of X-inactivation specific motifs/oligomers, which may be enriched in older stratum where genes most likely undergo inactivation but could be absent in the younger strata. We have studied the human X-chromosome which is known to have a number of evolutionary strata arising from the suppression of recombination between X and Y at different times. These studies were extended to the plant sex chromosomes which are known to have a number of strata that formed at the initial stage of their evolution. Dioecious plants present a unique opportunity to examine the onset of strata formation on the sex chromosomes, as sufficient time may not have elapsed yet for the loss of Y-linked homologous genes subsequent to recombination suppression between X and Y

(for XY sex determination system, and similarly for other systems including the ZW system), one can study the early phases of sex chromosome evolution. Furthermore, the recently sequenced sex chromosome V of the brown alga *Ectocarpus sp*. that has a haploid sex determination system (UV system) and the mating-type chromosomes of an anther-smut fungus *Microbotryum lychnidis-dioicae* were explored through this approach. Our results from sex chromosome analysis are summarized in Chapter 3.

Because of the ability of the Markovian segmentation and clustering to robustly discern disparate segments in genomes, we explored the genome of an extremophilic eukaryote, *Galdieria sulphuraria*, for the presence of DNAs horizontally acquired from different lineages. Our study uncovered novel foreign genes of prokaryotic origin in *G. sulphuraria*, results that were supported by multiple lines of evidence including composition-based, comparative data, and phylogenetics. These genes encode a variety of potentially adaptive functions, from metabolite transport to DNA repair. These results are presented in Chapter 4.

In Chapter 5, we propose a segmental genome model, wherein a genome is represented by an ensemble of signatures derived from segments of apparently different ancestries or origins, for taxonomic profiling of metagenomic sequences, as the current methods are inherently limited in exemplifying the microbial dynamism that shapes the genomes, resulting in chimeras with segments of different ancestries or origins. Robust classification of metagenomic reads based on our proposed segmental model will enable better understanding of the microbial communities and their interactions with hosts. For example, seawater microbiome sequencing has provided opportunities to assess microbial diversity in marine environment; similarly, profiling

human microbiome will help understand the impact of evolving microbial communities on the human health.

The discussion and perspectives are presented in Chapter 6. In this chapter, we also discuss how the Markovian segmentation and clustering can be made more sensitive and robust for further applications in biology and other disciplines in future. We assessed the power of our Markov model based method in discriminating genomic sequences from different sources as a function of model order, sequence length and phylogenetic divergence. We discuss how this could be exploited within a variable order model framework to decipher, in particular, the chimeric microbial genomes.

CHAPTER 2

METHODOLOGY

2.1 Background

Rapidly accumulating genomic data as a result of high-throughput sequencing has necessitated development of efficient computational methods to decode the biological information underlying these data. DNA composition varies across functionally different regions of the genome as well as those of distinct evolutionary origins.

Multiple change point methods locate those positions in a given DNA sequence where the segments left and right to the position are maximally distinct from each other based on a chosen compositional measure [38]. A frequently used change point method utilizes an entropy based measure, namely, the Jensen-Shannon (JS) divergence [34], in order to construct the partition. This segmentation algorithm computes the JS divergence at every position in the sequence, thus obtaining the difference between the subsequences left and right to each position. The sequence is segmented is at the location where JS divergence is maximum.

Markov segmentation method generalizes the entropic segmentation approach within the framework of Markov chain models. Zeroth order Markov segmentation is similar to the method employed by Bernaola-Galvan et.al. [34], where the JS divergence between nucleotide distributions is computed at each sequence position and the sequence is segmented at position where the divergence is maximized. This method has often been used as an exploratory tool for deciphering the genome organization. It has also been used to determine regions that are homogeneous with respect to GC/AT or purine/pyrimidine composition, localize isochores in eukaryotic chromosomes,

identify protein-coding regions [39], detect CpG islands, and address many other aspects of genome complexities.

The issue that has often been raised is whether the statistical methods for genomic segmentation generate segments of known biological significance. Different segmentation strategies have focused on generating optimal or biologically relevant segments. Although techniques such as the entropic segmentation have been successful in identifying CpG islands and isochores, or HMM based segmentation methods which achieved success in discovering functionally important entities such as protein-coding genes, there are many other aspects of genome organization that have remained yet unexplored.

Markov model for segmentation (MMS) was developed to incorporate higher order correlations, which can characterize the inhomogeneities inherent in a given genomic sequence more robustly. Thakur *et al.* [1] have shown the advantage of higher-order Markov model based segmentation procedures in deciphering the compositional heterogeneity in more biologically meaningful ways [38]. Although this has improved the sensitivity in detecting change points, it did not address the issue of identifying distinct sequence types within a sequence of interest. The resulting compositionally homogeneous sequence segments are considered independent entities, which may not be true. In reality, many of these sequence segments may share similarities with other non-neighboring segments. Therefore, the number of sequence types could in fact be much less than the number of sequence segments. A meaningful interpretation of genomic data is feasible only within an integrated framework for change point detection, as well as source (sequence type) identification, the former through segmentation and

the latter through classification.

Here, we have adapted such an integrated framework proposed earlier by Azad and Li [37] and applied to address a number of biological problems. This algorithm first identifies compositionally distinct segments and then clusters segments that are compositionally similar, without any prior knowledge of the composition of the sequence being analyzed. The method is described in detail below.

2.2 Markov Model of Segmentation

Markov model of segmentation was developed to account for higher order correlations [1, 4] within genomic sequences. In a Markov model framework, the occurrence of a symbol at any location in a sequence depends on its just preceding symbol(s). The number of preceding symbols defines the order of Markov model. The probability of a symbol given the preceding symbol(s) is referred to as the transition probability.

Segmentation of a given DNA sequence involves partitioning it into two subsequences that are maximally distinct from one another based on a chosen statistical criterion [1, 4, 34, 40]. This entails recursive application to resulting subsequences until a halting criterion is met. A commonly used entropic segmentation strategy maximizes the JS divergence [38] in order to construct this partition. The segmentation algorithm computes JS divergence at each point in the given DNA sequence and then segments it into two at the position of where the JS divergence is maximum (D_{max} in Fig. 2.1). We leveraged the power of Markov model in measuring the compositional heterogeneity, replacing nucleotide distribution with higher order

oligonucleotide distribution in the JS divergence formulation. Previous studies have also suggested utilization of oligomer frequency in analysis of compositional heterogeneity [41].

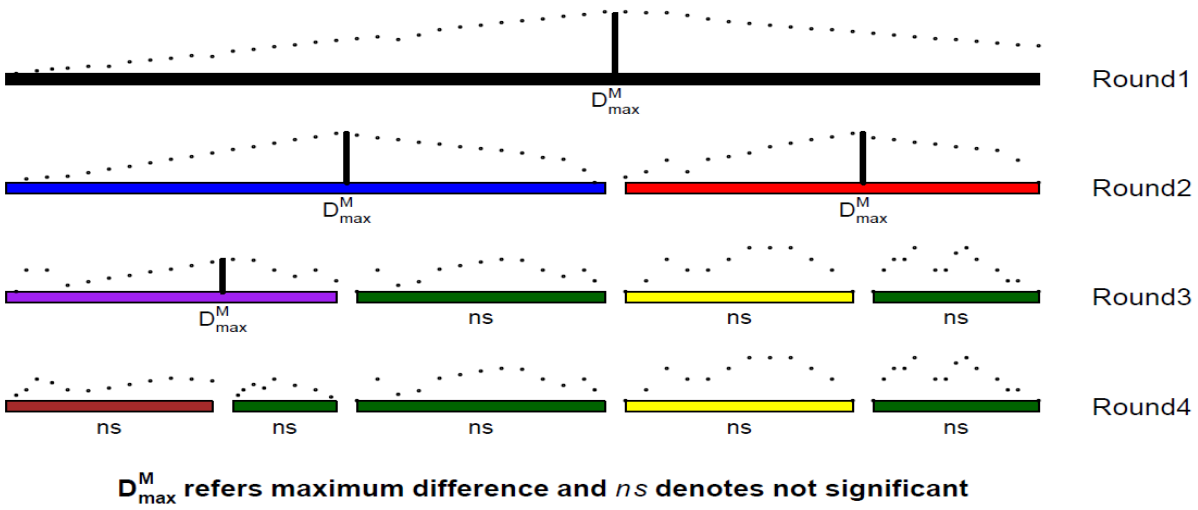

Figure 2.1. An illustration of Markov model of segmentation

Consider a sequence *S* of length *L* composed of symbols α_1, α_2, ...,α_L,

$$S = \alpha_1\ \alpha_2\ \alpha_3 \ldots \alpha_{L-1}\alpha_L,$$

the subscript to α denotes its position in *S*. The probability of this sequence using a zeroth order Markov model is given as,

$$P(S) = p(\alpha_1).p(\alpha_2).p(\alpha_3).p(\alpha_4) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots p(\alpha_L).$$

The Jensen-Shannon (JS) divergence between two subsequences S_1 and S_2 of length l_1 and l_2, generated by partitioning at any point in sequence *S* is obtained as [35],

$$D(S_1,S_2) = H(S_1 \oplus S_2) - \left(\frac{l_1}{l_1 + l_2} H(S_1) + \frac{l_2}{l_1 + l_2} H(S_2) \right), \quad (1)$$

where *H*(.) is the Shannon entropy,

$$H(.) = -\sum_{\alpha} p(\alpha) \log_2 p(\alpha)$$

$p(\alpha)$ denotes the probability of nucleotide α, estimated as N(α)/N where N denotes count. $L=l_1+l_2$.

The above measure of divergence is derived on the assumption of the independence of each nucleotide occurrence in *S*, which signifies the zeroth order Markov model. This can be generalized to higher order Markov models to account for short range correlations within nucleotide ordering patterns.

The transition probability of symbol α_2 given the preceding symbol α_1, corresponding to Markov model of order 1, is given as:

$$p(\alpha_2|\alpha_1) = \frac{frequency(\alpha_1\alpha_2)}{frequency(\alpha_1)}$$

The probability of the sequence *S*, for Markov model of order 1, is obtained as:

$$P(S) = p(\alpha_1).p(\alpha_2|\alpha_1).p(\alpha_3|\alpha_2).p(\alpha_4|\alpha_3) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots p(\alpha_N|\alpha_{N-1})$$

For the model of order 2, the transition probability of a symbol is computed based on its two preceding symbols. The number of preceding symbols used to predict the next symbol defines the order of a model.

Hence, if we generalize the above equation in the Markov chain model framework, the probability of this sequence using m^{th} order Markov model is obtained as,

$$P(S) = p(\alpha_1, \alpha_2, , , , \alpha_m) \prod_{i=m+1}^{N} p(\alpha_i | w = \alpha_{i-m}.\alpha_{i-m+1} \dots \dots \dots \dots \dots \dots . \alpha_{i-1}),$$

Where $p(w)$, w= α_1, α_2 , , ..., α_m, is probability of "word" w of size m symbols, and $p(\alpha_i|w)$ is the transition probability of symbol α_i given the preceding sequence of symbols or word w of length m.

The generalized JS divergence to account for short range correlations in the nucleotide ordering was obtained recently within the framework of Markov chain model of order m, defined for divergence between two subsequences S_1 and S_2 of length l_1 and l_2, generated by partitioning at any point in sequence S as [1, 4, 37],

$$D^m(S_1,S_2) = H^m(S_1 \oplus S_2) - \left(\frac{l_1}{l_1 + l_2} H^m(S_1) + \frac{l_2}{l_1 + l_2} H^m(S_2) \right). \qquad (2)$$

Here the $H^m( )$ denotes the conditional entropy function, defined as,

$$H^m( ) = -\sum_w p(w) \sum_b p(b \mid w) \log_2 p(b \mid w), \qquad (3)$$

where w denotes oligonucleotide of length m preceding the nucleotide b, $p(w)$ is the probability of oligonucleotide w, defined as,

$$P(w) \approx N(w)/(l_i - m + 1), \qquad (4)$$

and *p(b|w)* is the probability of nucleotide *b* given the preceding oligonucleotide *w*, defined as,

$$P(b|w) \approx N(wb)/N(w), \qquad (5)$$

N(.) denotes the count. The standard Jensen-Shannon divergence as given in equation 1 is recovered for *m*=0.

The probability distribution of D^m, that is, the probability of observing D^m or less by chance, has been shown to follow a chi-square distribution function.

$$P\left(D^m \le x\right) \approx \chi_\nu^2\left(2(l_1 + l_2)(\ln 2)x\right)$$

with $v = k^m(k-1)$ degrees of freedoms, *k* being the alphabet size (*k*=4 for DNA sequences derived from alphabet (A,T,C,G)) [1, 4].

The recursive segmentation procedure entails obtaining the maximum value of D^m; the statistical significance of the maximum value of D^m can be assessed from the probability distribution of D_{max}^m, which was shown to approximate a χ^2 distribution function with fitting parameters β and λ,

$$P(D_{\max}^m \le x) \approx \{\chi_\nu^2[2(l_1 + l_2)(\ln 2)x\beta]\}^\lambda;$$

β and λ were estimated by fitting the above analytic expression to the empirical distributions obtained via Monte Carlo simulations [1, 4].

A given DNA sequence is segmented at the point of maximal divergence if the p-value, Prob ($D > x$), is less than a preset significance level. The recursive segmentation

process is halted when none of the sequence segments can be segmented further within the statistical hypothesis testing framework. Cutoffs for segment length were also applied to avoid numerous small and biologically insignificant segments where needed. The final output from this procedure is thus a set of sequence segments that are homogeneous within, but heterogeneous between, according to a pre-specified criterion [42].

2.3 Markov Model of Clustering

As we allow hyper-segmentation at a relaxed stringency to detect the breakpoints with greater precision, we perform a non-hierarchical, agglomerative clustering to restore the segmental structure by identifying the compositionally similar contiguous segments ("Contiguous Clustering" in Fig. 2.2) [37]:

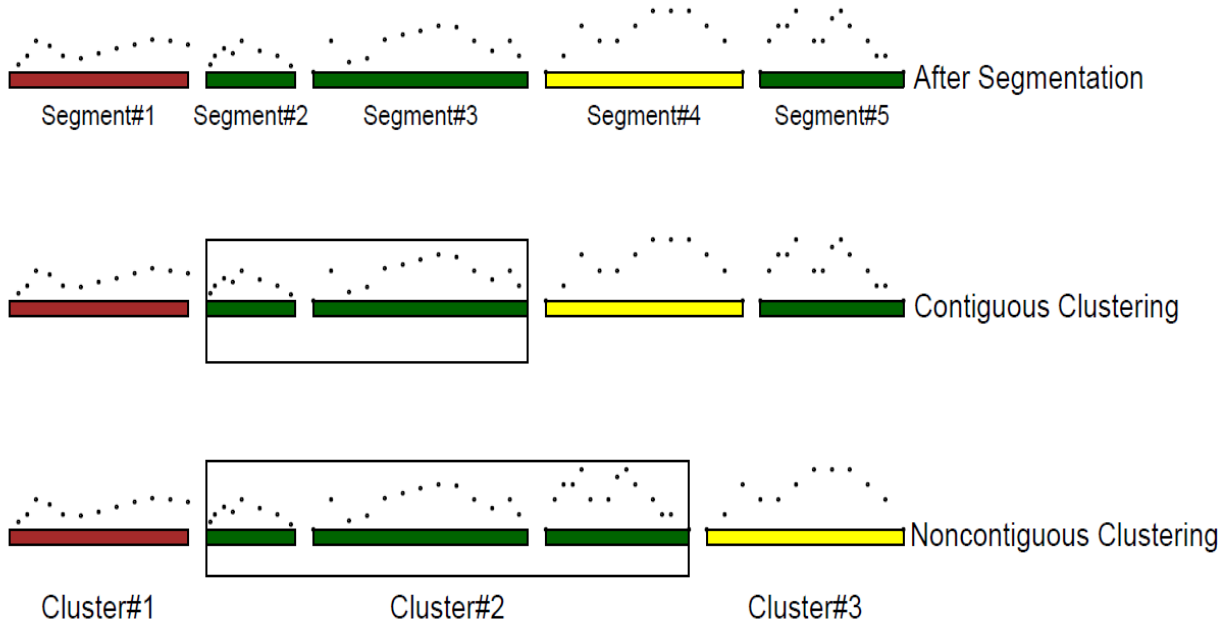

Figure 2.2. An illustration of Markov model of clustering

After identifying the compositionally similar contiguous segments and thus restoring the segmental structure, clustering of similar sequence segments is done recursively. Similar segments are grouped into distinct clusters until all resulting clusters are significantly different from each other (“Non-contiguous Clustering” in Fig. 2.2). This is performed within the same framework of statistical hypothesis testing, that is, if the p-value, Prob ($D^m > x$), computed for the JS divergence between two sequence segments, is greater than a preset significance level, the sequence segments or clusters are merged together, otherwise they are deemed statistically different and therefore, are not merged.

2.4 Advantages

Unraveling the intrusive level of organizational complexities within genomes is a challenging task. To fragment genomic sequences in biologically meaningful ways, numerous segmentation techniques have been developed that use different statistical criteria. Underlying many of these tools are the probabilistic models for deciphering structural and functional features via genome segmentation. Earlier studies have shown that Shannon entropy based segmentation is an effective technique for achieving this goal [1, 4].

In the binary segmentation procedure proposed by Bernaola-Galvan *et. al*., the boundaries obtained at previous rounds of segmentation are retained in the later segmentation steps irrespective of their significance at the later stages when local heterogeneities are measured and quantified. The Bayesian method of Ramansky *et al.* generates an optimal segmentation, however, it produces numerous small segments,

even of length one nucleotide, of doubtful biological significance.

HMM based method is undoubtedly promising and has successfully been applied in gene identification. HMM based methods output a single segmental structure underlying a DNA sequence. In contrast, entropic segmentation techniques were developed to explore multilayer inclusive complexities underlying genomic data. These are very flexible and can analyze genomes without any prior information or training data. Markov model of segmentation further improvised this technique. Higher-order Markov model based segmentation method provides enhanced sensitivity in deconstructing higher order organizational structures in genomes. Higher order Markov models are more sensitive to details of local patterns as well as provide a global view of genomic organization. This method was validated on chimeric sequences synthesized from a number of prokaryotic genomes [1].

The integrative methodology of Markov segmentation and clustering provides further advantage over the Markov segmentation and other methods that used recursive segmentation. It addresses the problems associated with both the recursive segmentation procedure for detecting the change points and the classification methods for detecting different feature types. This method showed promising results in deconstructing the mosaic organizational structure within genomic data [1, 4, 21, 36, 39]. One of the major problems of the recursive segmentation approach is often the difficulty in establishing a threshold that can result in precise detection of change points with fewer false positives. Relaxing the threshold may help in precise delineation, yet this may also generate many false positives. Azad and Li [37] have demonstrated that this can be circumvented in the integrated framework of recursive segmentation and

agglomerative clustering. The effectiveness of this method was shown in its application to solving a variety of biological problems, including alien gene prediction in bacterial genomes, structural variation quantification in human cancer genomes, and in the alignment-free genome comparisons. This method performed either as well or better than the sophisticated state-of-the-art methodologies, and thus emerged as a powerful statistical tool for deciphering the organizational complexities underlying genomic data. The method is not sensitive to the segmentation threshold and achieves high sensitivity without sacrificing the specificity.

We adapted the Azad and Li's generalized approach that consummates top-down and bottom-up information theoretic approaches yielding a robust integrative methodology for deconstructing genomic data. Importantly, the data heterogeneity was addressed by using multiple stringencies in the segmentation and clustering procedures, allowing hyper segmentation to detect precisely the change points followed by clustering in a non-hierarchical fashion to restore the inherent segmental structure of the data. In the next chapters, we present the application of this method to a host of biological problems and discuss the results and future directions.

CHAPTER 3

APPLICATION I: DETECTION OF EVOLUTIONARY STRATA

Sex chromosomes have evolved from a pair of homologous autosomes which differentiated into XY (male heterogametic) or ZW (female heterogametic) system. Sex chromosomes emerge likely through male-sterility mutations on the proto-X and female sterility-mutations on the proto-Y chromosome in the XY system (and analogously in the ZW system). Selection restricts recombination between the sex-determining regions (SDRs) of X and Y chromosomes, which otherwise would maladapt the individuals or hermaphrodites [43]. The initially established non-recombining regions may expand over most or all of the sex chromosomes due to the recruitment of sexually antagonistic (SA) alleles, beneficial in one sex but not so in the other, on the incipient sex chromosomes. Such a situation enforces selection for reduced recombination with the SDRs, because this increases the transmission of the male-beneficial alleles to males and reduces the transmission of these alleles to females, and vice versa [44, 45, 46]. Extension of the SDRs in this manner can lead to the creation of ‘‘evolutionary strata,’’ which are regions that have become non-recombining at different evolutionary times [47, 48, 49, 50].

Y-chromosomal genes experience smaller effective population size relative to X-chromosomal or autosomal genes, and are therefore more susceptible to accumulation of deleterious mutations or the replacement of wild-type alleles by genetic drift, leading to the degeneration of the heterozygous sex chromosome Y (or similarly, W) [51, 52].The accumulation of deleterious mutations, insertions of transposable elements (TEs), and the subsequent degeneration of Y (or W) chromosome are postulated to be

driven by at least four mechanisms [52]: 1) Background selection: Weakly advantageous mutations are eliminated if they co-occur with strongly deleterious mutations, which accelerate the fixation of weakly deleterious mutations and reduce the fixation of weakly advantageous mutations; 2) Muller's ratchet: Muller pointed out a "rachet" effect which could cause disadvantageous mutants to accumulate irreversibly in populations lacking recombination [53]. In a finite non-recombining population, due to stochastic effects, the chromosomes may accumulate deleterious mutations. The fixation of deleterious mutations on the non-recombining proto-Y chromosome occurs sequentially after each stochastic loss of ancient or proto-Y chromosome with fewest deleterious mutations, replaced by one with the next fewest deleterious mutations. The fixation results in gene function loss and eventually the gene loss from the proto-Y; 3) Weak selection Hill-Robertson effect: In the absence of recombination, the closely linked weakly selected beneficial alleles and deleterious alleles interfere with each other, which impedes the spread of beneficial alleles and results in their loss together with the linked deleterious genes; 4) Genetic hitchhiking: This entails fixation of deleterious mutations following selection of linked beneficial mutations on the non-recombining Y chromosome [52, 54].

Genomic regions with suppressed recombination often have higher density of TEs. Three models, namely, deleterious insertion model [55], ectopic exchange model [56] and deleterious transposition model [57] have been proposed to account for the deleterious effects of TEs. Dolgin and Charlesworth's simulation experiments showed that under the deleterious insertion model the fixation of TEs in the extremely low recombining region with rare excision and weak synergism between elements is a

consequence of Hill-Robertson effect in the form of Muller's ratchet [58]. However, if the above conditions are not met, then the alternative ectopic exchange model can be invoked to explain the accumulation of TEs in the recombination suppressed regions.

In mammals, birds, and other animals, such as, fishes and frogs, and in dioecious plants, either partial or almost the entire Y chromosome in XY system or W chromosome in ZW system ceases to recombine with its homologous partner (X or Z) due to successive suppression of recombination between their homologous regions. For example, in human, the recombining regions on X and Y have shrunk to small homologous segments known as "pseudoautosomal regions" (PARs) [59], while in asparagus, recombination is suppressed between the sex determining regions only [60]. The divergence level of X- and Y-linked homologous genes differs between successively suppressed regions along the X chromosome. The divergence level is higher where recombination ceased at earlier times and lower in recently suppressed regions [47]. Identifying the regions of recombination suppression, namely, the evolutionary strata, is central to understanding the history and dynamics of sex chromosome evolution. Evolutionary strata have been reported in organisms as diverse as smut fungi [61], human [42, 47-49, 59, 62], mouse [63], chicken [64], and dioecious plants, mainly, *S. latifolia* [65] and papaya [50].

One of the main mechanisms for stratum formation has been postulated to be serial inversions on Y chromosome, which results in suppression of X-Y recombination in males in the regions of the inversion (analogously in the ZW system) [47, 59, 62]. Following each inversion, the non-recombining regions on the X and Y chromosomes start to evolve independently, accumulating mutations and repetitive elements. This

results in the degeneration of the Y chromosome due to the accumulation of deleterious mutations. Because of this, in human, for example, the Y chromosome has become much smaller and gene poorer in comparison to the X chromosome, which unlike Y, recombines with its partner while in the females [48]. Similar to inversions on the mammalian Y chromosomes, two large inversions are reported to have occurred on the papaya Y chromosome and several other rearrangements within each inverted region on the Y chromosome are still ongoing [50], while in *S. latifolia* at least two large inversions, pericentric and paracentric, have been reported to have occurred on its Y chromosome [67].

The other plausible mechanisms for stratum formation include the emergence of sexually antagonistic mutations in PAR genes, followed by closer linkage between these sexually antagonistic genes and the male-specific Y regions (MSY), resulting in suppressed recombination and the extension of MSY. In this way, the formerly PARs shrink; this may be repeated several times, thus creating the evolutionary strata. The evolutionary strata can also form following the acquisition of autosomal DNA in an existing sex chromosome and the subsequent linkage between sexually antagonistic genes and the sex chromosome, resulting in suppression of the recombination [68]. Thus, whichever mechanisms suppress the recombination likely influence the formation of evolutionary strata on sex chromosomes [63].

It has been suggested that sex chromosome evolution happens in five stages [60, 69]. Initially, the suppression of recombination starts at and around the sex determining locus, which further spreads to neighboring regions and form a male-specific region on the Y chromosome (MSY) (or female-specific region on the W

chromosome (FSW)). Papaya sex chromosomes are reported to be at this stage of evolution. Then, MSY (and analogously, FSW) expands in its size due to duplications and accumulation of repetitive elements in male specific regions, resulting in expansion of the Y chromosome, and eventually the recombination suppression spreads to the entire Y chromosome. *S. latifolia* sex chromosomes represent this stage of the evolution, with its Y chromosome longer than the X chromosome. Finally, the Y chromosome begins to degrade due to the accumulation of deleterious mutations, which causes the loss of genes and reduction in size of the Y chromosome. Mammalian sex chromosomes are the best examples of this stage of evolution. Although the degeneration of the Y chromosome has also been reported in dioecious plants, such as, *S. latifolia,* it's not to an extent as in mammals [70]. The gymnosperm species *Cycas revoluta*'s sex chromosomes also bear such signs of degeneration. Further degradation may lead to the loss of Y chromosome and then the sex determination is believed to be controlled by X to autosome ratios [60, 69]. Notably, different plant species are at different early stages of sex chromosome evolution and thus are ideal model systems to study the mechanisms of recombination suppression, evolution of sex chromosomes, and the degeneration of the Y or W sex chromosome.

3.1 Detecting Evolutionary Strata on the Human X Chromosome

3.1.1 Background

The human X chromosome is comprised of distinct evolutionary strata, the boundaries of which have been characterized largely on the basis of substitution rates of select X-Y gene homologues [47]. The human X chromosome 155 Mbp long, with an

overall GC content of 40 percent and a total of about 1400 genes, corresponding to a low gene density of less than 10 genes per Mb on average (UCSC genome browser version GrCH19). One of the enduring problems in the study of mammalian sex chromosomes has to do with their origin and function.

As mentioned above, the sex chromosomes are believed to have evolved from a pair of autosomes; one of the major evidence for this has come from sequence comparison of the human chromosome X with the complete genome of *G. gallus* [70]. Ancestrally the sex chromosomes could recombine over their entire lengths, but following a series of recombination suppression events, including inversions on the Y chromosome [47, 59, 62], the human X and Y now only recombine in the small pseudoautosomal regions, found at both ends of the X and Y [59]. These recombination suppression events occurred serially, reflected in similar rates of X-Y divergence within, but different between, contiguous regions along the whole X chromosome, resulting in distinct “evolutionary strata” on the human X [47]. The eutherian (mammals excluding marsupials and monotremes) X chromosome is composed of an ancestral X-conserved region (XCR) that is shared with marsupials, like the opossum, but not with monotremes [71], and a more recently transposed X-added region (XAR) that is sex-specific in eutherians, but autosomal in marsupials [59, 72]. Genes within each stratum ceased to undergo homologous X-Y recombination around the same evolutionary time, and thus share a unique evolutionary history that is distinct from other strata. Some strata are shared across eutherian mammals, while others are lineage-specific [42].

Following each inversion on the Y chromosome [73, 74], which suppressed the X-Y recombination in males in the region of the inversion [47, 59, 62], the non-

recombining regions on the X and Y chromosomes have evolved and diverged from one another independently. Additionally, in the absence of recombination in males, the non-recombining regions may accumulate DNA elements, such as, transposable or repetitive elements, and sequences with shifts in GC content [59]. Furthermore, on the X chromosome, motifs/oligomers related to X-chromosome inactivation may also accumulate in response to loss of functional genes on the Y [75, 76], and thus the sequence composition of each stratum on the X diverges from neighboring regions that have ceased to recombine even before or are still undergoing the X-Y recombination. Studies have also suggested that genes on older strata are more likely to undergo inactivation, while genes on recently added region likely to escape from inactivation [16]. These observations suggest the presence of functionally distinct domains on the human X chromosome.

All current methods for stratum detection depend on X-Y comparisons but are severely limited by the paucity of X-Y gametologs. Synonymous substitution rates have been frequently used to estimate divergence between X-linked and Y-linked sequences [47, 48, 76], however, such studies may be biased due to the saturation of synonymous substitutions, gene conversion, or gene-specific conservation on the X and Y. Inversion analyses have been used to identify the younger strata [49, 59] but lose power to detect older strata boundaries due to a saturation of Y inversions. Phylogenetic methods use comparative genomics to estimate recombination suppression [62, 77], but are limited in their resolution by the number of species with available sequence. Thus, a comprehensive approach to identifying the evolutionary strata has not yet been realized due to the limitations imposed on *all* current methods of comparing the homologous X-Y

sequences because so few Y sequences remain following the loss or degradation of non-recombining ancestral Y sequences.

To circumvent the limitations of the current approaches to stratum detection, we implemented the recursive segmentation and agglomerative clustering algorithm to decipher compositionally distinct regions on the X, which reflect regions of unique X-Y divergence. We first tested the proposed method on the concatenated gene sequences of all 35 previously assayed human X-linked genes. Our method correctly classifies this set of genes and provides an alternative line of evidence supporting recent suggestions that the third stratum is actually composed of two distinct strata. Second, by applying our method to the entire X chromosome, we uncovered hitherto unknown strata, particularly in the XCR region, which is difficult to analyze using conventional, comparative methods due to lack of X-Y gametologs in this region. Our predictions reconciled well with predictions from existing methods that require Y homology, demonstrating that our method can be reliably used to identify evolutionary strata in the absence of sequence information from the heterogametic sex chromosome.

3.1.2 Materials: Genomic Sequences

The sequence of the human (hg19) X chromosome, the homologous regions in chicken (galGal3), as well as information about the repetitive elements, genic regions were downloaded from the UCSC Genome Browser [78]. The human X chromosomal sequences homologous to the chicken regions were defined as the human XCR, homologous to chicken chr4: 1-20 Mb, and the human XAR, homologous to chicken chr1: 103-123 Mb [42]. Definitions of previous strata were collected from the published

papers [47-49, 59, 62, 77].

3.1.3 Approach

Each stratum evolves independently following recombination suppression, with the oldest stratum at the extremity of the long arm of X being most diverged, the newest stratum near the tip of the short arm of X being least diverged, and with the divergence decreasing with distance from the distal long arm of the X. We posit that this differential divergence on the X will be reflected in the oligonucleotide compositional divergence between the strata along the X chromosome. As such, we took a multi-pronged algorithmic approach to first segment the X chromosome into regions of distinct oligonucleotide composition, and then segregate compositionally similar regions in distinct clusters using the agglomerative clustering algorithm. Here we globalized a Markovian segmentation approach used in a previous ad hoc application to detect strata boundaries [36] by making the recursive segmentation step-free and incorporating a posterior two-step clustering procedure as proposed by Azad and Li (2013), and thus introducing, for the first time, an unrestricted integrative algorithm to decipher segmental structures within long eukaryotic genomes. This algorithm first identifies compositionally distinct segments and then clusters segments that are compositionally similar, without any prior knowledge of the composition of the sequence being analyzed. In contrast to moving window or gene based methods, which are constrained to define stratum boundaries on the basis of window or gene boundaries, this approach can localize the stratum boundaries at any genomic region. Further, the segmentation and clustering algorithm’s parameter choices, including the significance thresholds, are preset at the

outset, and therefore the proposed procedure is free from the biases due to “artificial” human interventions, such as halting the algorithm at will to secure selected stratum boundaries as was done previously [36]. Our method thus detects the previously described strata without generating a plethora of segments of unknown significance, as is the case with a previous segmentation-only method [36] (Fig. S1). This approach thus provides an unbiased means to detect the presence of evolutionary strata, and their boundaries, within the framework of statistical hypothesis testing, and without gametologous Y sequences.

3.1.4 Comparison with Previous Analyses and Validation of Proposed Method

We first applied our segmentation and clustering algorithm to the concatenated sequence of the 35 X-linked genes that have been previously assayed using inversion [59, 62], phylogenetic [49] and substitution rate [48, 59] analyses. Protein-coding gene sequences are constrained in their sequence evolution to maintain functional gene products and so accumulate nucleotide differences (substitutions, deletions and insertions) slower than noncoding DNAs. As such, an analysis of the coding regions should be a proxy for the divergence rate between gametologous X-Y sequences, but is likely a conservative estimate of the sequence differences that have accumulated between the larger X and Y regions due to suppression of recombination between them. Similar to the expectations, when analyzing just the coding regions, the segmentation and clustering algorithm produces a conservative stratum structure (Fig. 3.1). Our method correctly classifies genes by previously defined stratum boundaries, and, consistent with recent suggestions [49, 62], we find evidence of two strata within the

previously described stratum 3 (Fig. 3.1). This analysis confirms that our method is able to recapitulate previous stratum definitions, but also highlights the challenges of relying only on X-linked coding sequences, which are necessarily more conserved than noncoding regions, with identifiable Y-linked gametologs.

3.1.5 Predicting Strata in the Absence of Y Sequence Information

Because protein-coding sequences are constrained in their evolution, and non-coding sequences may accumulate nucleotide heterogeneity faster, we next applied the segmentation and clustering algorithm to the entire DNA sequence of the human X chromosome. This approach takes advantage of the logic that, in the absence of recombination in males, DNA sequence on the human X chromosome is more likely to accumulate and retain DNA changes (substitutions, insertions, and deletions), than regions that undergo homologous recombination in both males and females. The longer the region has been evolving without recombination in males, the more changes it is likely to have accumulated. Further, in mammals, the drive towards dosage compensation, in response to gene loss on the Y [76], may result in the accumulation of specific sequences related to gene silencing or activation on the X [75]. Thus, we expect the oligonucleotide composition to be similar within strata on the X chromosome but differ between the strata; our segmentation and clustering algorithm identifies such differences.

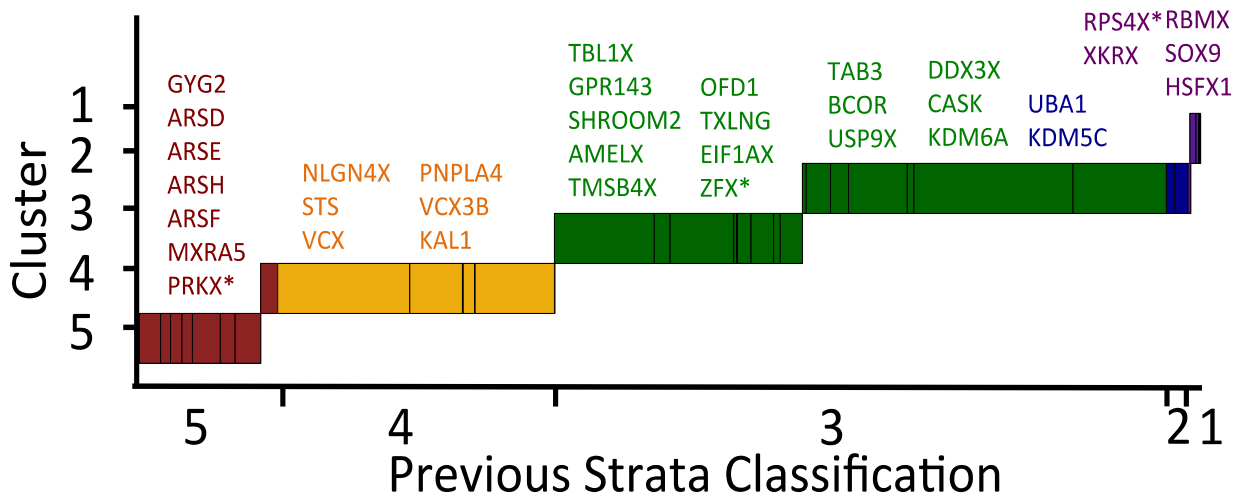

Figure 3.1. Strata identified using previously-assayed X-linked genes. Here we apply the segmentation and clustering algorithm to a concatenated string of the X-linked genes that have been previously assayed using inversion, phylogenetic and substitution rate analyses. Previous strata are colored: 5-Red, 4-Yellow, 3-Geen, 2-Blue, 1-Violet. Genes in each cluster are labeled above the cluster, similarly color-coded. 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 [42].

Thus, we expect the oligonucleotide composition to be similar within strata on the X chromosome but differ between the strata; our segmentation and clustering algorithm identifies such differences.

We make significant improvements to predicting strata by allowing all of the sequence on the X chromosome to be utilized, instead of relying only on genic regions, and also by not limiting to regions that retain Y homology. Our algorithm identifies twelve compositionally distinct regions on the human X chromosome. Three of these occur within the first pseudoautosomal region, PAR1. Given the strong evidence that the PAR1 is still recombining, we do not take this as evidence of recombination suppression. These three clusters also appear to be driven by repetitive elements

because when repetitive elements are masked out we observe only two clusters on the human X chromosome; one corresponding to the PAR1, and one corresponding to the rest of the X chromosome (which is almost entirely nonrecombining in males). Outside of the PAR, our method identifies nine clusters, which we expect to correspond to distinct evolutionary strata. First, our method identifies the previously defined boundaries of the human PAR1 and the two most recent strata (Fig. 3.2; Fig. S3; Table 1), which have been described in detail, and independently confirmed by different studies [47, 79]. We confirm, as others have hypothesized [43, 55], that the stratum previously described as stratum 3 is actually comprised of at least two compositionally distinct regions (Fig. 3.2; Table 1).

Finally, for the first time, we provide estimates of the positions and boundaries of strata in the oldest region of the sex chromosomes (Table 1). Specifically, in humans we show that rather than only one or two strata on the X-conserved region (XCR) [47, 48, 77], there were at least five independent recombination suppression events in the XCR (Fig. 3.2; Table 1). Given that species with very young sex chromosomes already have several observable strata (e.g., three strata have already been identified on the less than 10 million year old sex chromosomes of *Silene latifolia* [66]), the existence of many strata in the XCR is much more consistent with current theory of sex chromosome evolution than the likelihood of one or two extremely large inversions. Our method makes no assumption about the timing of any of the events, although other lines of evidence do suggest successive linear recombination suppressions. Specifically, we do not distinguish where the first recombination suppression event occurred. Current theory suggests that SOX3, from which the sex-determining SRY gene evolved, is in the

oldest stratum. We do not contest this; however, the new evidence from this study suggests that the earliest two inversions might have happened in quick succession, the inversion involving SOX3 (represented by the second cluster, Fig. 3.2) happening first, then recombination suppression proceeded in both directions along the chromosome, reaching the terminal end of Xq very quickly. Interestingly, the first cluster (~10 Mbp in size, Fig. 3,2) contains PAR2, which is only 320 kb long, was recently added to the X and Y, and reported to only occasionally undergo recombination, in contrast to PAR1 [80].

The X-added region, XAR, was added to the eutherian X chromosome approximately 105 million years ago, prior to the radiation of eutherian mammals, and is nearly one-third of the human X chromosome (from 0 to 46.88 Mb; [59]). We find one cluster spanning the region between the ancestral XCR, and the younger XAR (Fig. 3.2, Table 1). This is consistent with the hypothesis that the autosomal segment translocated to the ancestral X chromosome was added to an ancient pseudoautosomal region, PAR, that was still undergoing X-Y recombination in the common ancestor of eutherian mammals. Then, after the addition of the XAR, an event occurred to suppress recombination in a region spanning the ancestral PAR, and a portion of the XAR.

We also found that much of the compositional heterogeneity between strata on the human X chromosome can be explained by the presence of repetitive elements, because our method, when applied to the entire X chromosome with repetitive elements masked out, does not return the underlying strata structure (Fig. 3.2). Ross et al. [59] found that LINEs increase in frequency with increasing stratum age (at least when considering the most recent three strata), but did not observe a monotonic pattern with

respect to GC content. Similarly, we do not observe a striking pattern with respect to GC content (Supplementary Table 1).

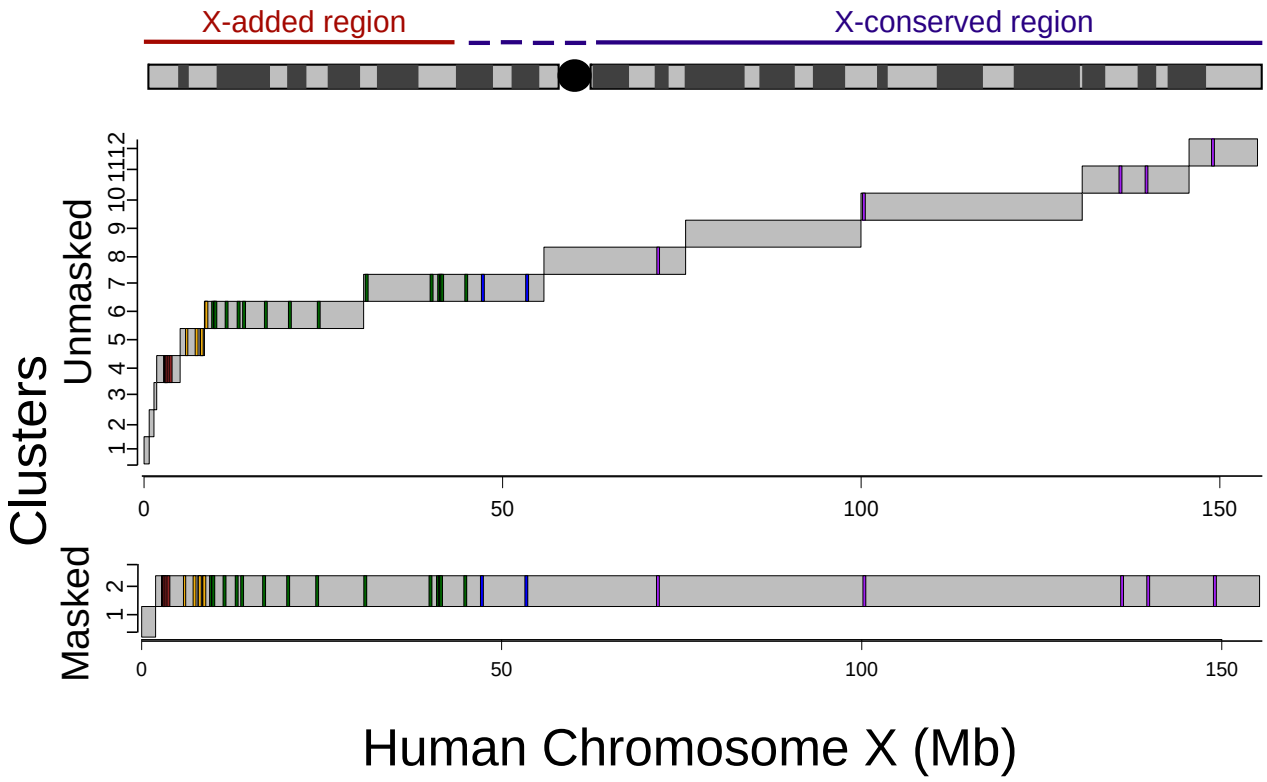

Figure 3.2. Strata identified across the whole X chromosome. Here we show the clusters that are determined using the entire sequence of the human X chromosome, either unmasked or masked for repetitive elements, as defined by Repeat Masker. We also plot the position and strata delineation of X-linked genes that have previously been assayed. Previous strata are colored: 5-Red, 4-Yellow, 3-Geen, 2-Blue, 1-Violet. We used Markov model of order 2 to perform segmentation and clustering at significance thresholds of 0.4 and 10-7 respectively [42].

Table 3.1. Summary of clusters identified from the segmentation and clustering algorithm, and comparison with previous definitions of strata. Here we show how clusters identified by our segmentation and clustering algorithm compare with previous efforts to identify evolutionary strata on the human X chromosome [42].

Cluster (Mb)	No.	Gene	Position (Mb)	[47]	[48]	[49]	[59]	[62]	[77]
0-0.73	12	-	-	-	-	-	-	-	-
0.73-1.40	11	-	-	-	-	-	-	-	-
1.40-1.78	10	-	-	-	-	-	-	-	-
1.78-5.04	9	XG	2.67	PAR	PAR	PAR	PAR	PAR	PAR
		GYG2	2.75	4	4	5	5	-	-
		ARSD	2.83	4	4	5	5	-	-
		ARSE	2.85	4	4	5	5	-	-
		ARSH	2.92	-	4	-	-	-	-
		ARSF	2.99	-	4	5	5	-	-
		MXRA5	3.23	-	4		5	-	-
		PRKX	3.53	4	4	5	5	5	-
5.04-8.43	8	NLGN4X	5.81	-	4	4	4	-	-
		VCX3B	6.45	-	4	4	-	-	-
		STS	7.14	4	4	4	4	-	-
		VCX	7.81	-	4	4	4	-	-
		PNPLA4	7.87	-	4	-	-	-	-
		*KAL1	8.50	4	4	4	4	-	-
8.43-30.62	7	TBL1X	9.62	-	3/4	3	3/4	4	-
		GPR143	9.69	-	3/4	-	-	-	-
		SHROOM2	9.75	-	3/4	-	3/4	-	-
		AMELX	11.31	4	3/4	3	3	4	-
		TMSB4X	12.99	3	3/4	-	3	4	-
		OFD1	13.75	-	3	-	-	-	-
		TXNLG	16.71	-	3	-	3	3/4	-
		EIF1AX	20.15	3	3	-	3	3/4	-
		ZFX	24.19	3	3	-	3	3/4	-
		*TAB3	30.85	-	-	-	3	-	-
30.62-55.78	6	BCOR	39.91	-	3	-	3	-	-
		USP9X	40.98	3	3	-	-	3	-
		DDX3X	41.19	3	3	-	-	3	-
		CASK	41.38	3	3	-	3	-	-
		KDM6A	44.73	3	3	-	3	3	-
		UBE1X	47.06	2	2	-	-	-	1
		KDM5C	53.22	2	2	-	-	-	1
55.78-75.53	5	RPS4X	71.49	1	1	-	-	-	1
75.53-99.98	4								

99.98-130.82	3	XKRX	100.17	-	-	-	-	-	1
130.82-145.73	2	RBMX	135.96	1	1	-	-	-	1
		SOX3	139.59	1	1	-	-	-	1
145.73-155.72	1	HSFX	148.67	-	-	-	-	-	1
		SPRY3	154.99	PAR	PAR	PAR	PAR	PAR	PAR

Further, we also observe a monotonically increasing trend in the density of repetitive elements in the strata with distance from Xpter across the entire XAR (Fig. 3.3 A). In this younger region, the stratum structure deciphered by our proposed method reinforces the earlier observation of increase in repetitive element densities after the suppression of X-Y recombination. This may also have contributed to the heterogeneity between strata.

In addition, we observe that this pattern breaks down in the older strata. Across the XCR, we observe that nearly all clusters share a similarly high level of repeat density, which may be attributed to the long evolutionary time that the XCR has been without homologous X-Y recombination, resulting in saturation in the density of repetitive elements across this region. Curiously, the cluster in the XAR that spans the recently X-transposed region has a higher repeat density than the remaining clusters in the XAR, which may be due to the inclusion of the centromeric sequence. When we break down the repetitive elements into different types, we observe that L1s (more than LTRs and Alus) account for the bulk of the variation in the density of repetitive elements between clusters (Fig. 3.3 B).

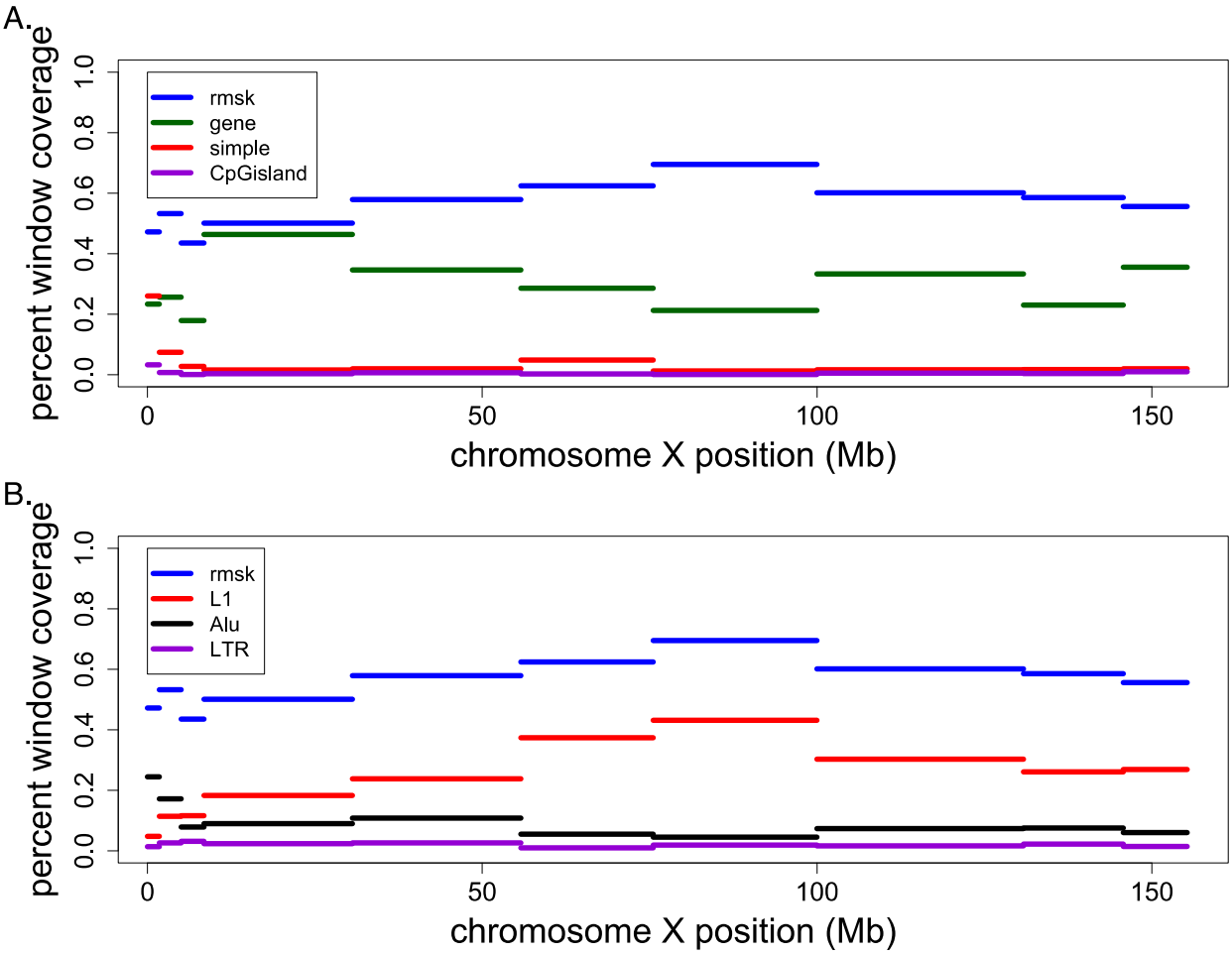

Figure 3.3. Density of repetitive elements or genes across the clusters identified on the X chromosome. Here we show the difference in the feature density between clusters: A) density of genes, repetitive elements, CpG islands, and simple repeats is plotted for each cluster; and, B) the repetitive element density is plotted for each cluster along with the L1, LTR, and Alu subsets of repetitive element (Performed by collaborator Melissa Wilson Sayres)[42].

To rule out the alternative hypothesis that the underlying structure of the X chromosome, derived from its autosomal ancestor, could be responsible for the distinct clusters observed on the X, we applied our method to the homologous autosomal sequence. In chicken, chromosome 4 (1-20Mb) is homologous to the mammalian XAR, and chromosome 1 (103-123Mb) is homologous to the XCR [59]. When applying the segmentation and clustering algorithm to the homologous chicken autosomal regions,

using the same parameters as for the human X, we do not observe any clustering in either region (Fig. 3.4).

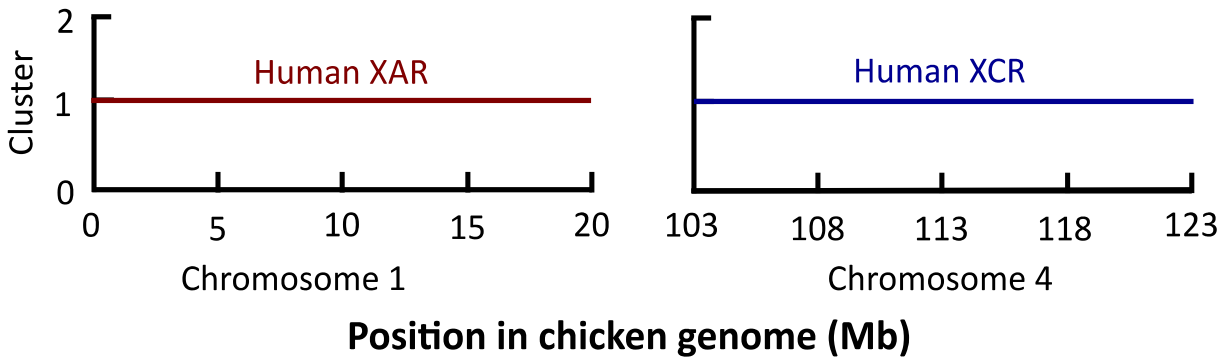

Figure 3.4. Segmentation and clustering of the homologous chicken autosomes. Here we plot the clusters determined from the regions in chicken that are homologous to the eutherian mammal X-added region, XAR (chicken chromosome 1), and X-conserved region, XCR (chicken chromosome 4) [42].

The absence of segmental structure within chicken autosomal sequences homologous to human X-linked sequences demonstrates that stratum formation fundamentally alters sequence composition, thus allowing identification of compositional signals unique to evolutionary strata by the proposed segmentation and clustering method.

3.1.6 Conclusions

In conclusion, we present a novel algorithm for the unbiased detection of the presence of evolutionary strata on human X chromosome and fine scale delineation of their boundaries. The ability of this approach to detect strata is not limited by the availability of gametologous sex chromosome sequences. We envisage the applicability

of this “unsupervised” method to sequences of the homogametic (X or Z) sex chromosome with yet unknown stratum history. Because of the complementary strengths of other approaches, wherever the gametologous sequences are available, we suggest using the integrative segmentation and clustering method in concert with substitution rate, inversion and phylogenetic methods. Where the heterogametic chromosomes are unavailable (Y or W), our algorithm is the only resource to investigating the sex chromosome evolution. Future efforts should focus on the development of a comprehensive approach that can exploit the complementary strengths of different methods for better understanding of novel sex chromosome systems.

3.2 Detecting Evolutionary Strata on the Plant Sex Chromosomes and Fungal Mating-type Chromosomes

3.2.1 Background

Dioecious plants are believed to have evolved from either ancestral hermaphrodites or monoecious plants that lacked separate sexes [60, 81]. The sex determination mechanism is governed by XY system in mammals and ZW system in birds, while in plants, both XY (male heterogametic) system as in white campion (*Silene latifolia*), papaya (*Carica papaya*) and asparagus (*Asparagus officianalis*), and ZW (female heterogametic) system as in willow (*Salix suchowensis*) and many other plants have been reported [60]. Plant species thus provide a unique opportunity to study the evolution of both types of sex chromosome determination systems (XY and ZW).

Evolution of sex chromosomes as a consequence of recombination suppression has been well studied in mammals and birds, but not so in plants, although 48 dioecious

land plants across 20 families have already been reported [68]. While the sex chromosomes in mammals (the placentals and maruspials) is reported to date back to approximately 165 million years ago [71], the sex chromosomes in plants are relatively much younger, such as the heteromorphic sex chromosomes in *S. latifolia* (~10-20 million years) [65] and papaya (~5-10 million years) [50]. Despite being so younger, the plant X chromosomes have already been stratified with several evolutionary strata as reported in several recent studies [65, 82-83]. These dioecious plants provide the opportunities to examine and understand the onset of strata formation on X or Z chromosomes, and thus are the ideal model systems to study the early phases of sex chromosome evolution. As enough time may not have elapsed yet since the onset of recombination suppression due to inversions on Y or other chromosomal rearrangements and the subsequent loss of Y-linked gametologous genes, different classes of methods could be readily applied for deciphering the plant sex chromosomal structure. However, only two plants *S. latifolia* [65, 82-83] and papaya [50] have been extensively studied for the presence of evolutionary strata on their X chromosomes. Other dioecious plants haven't been investigated yet because of the lack of gametologous sex chromosome sequences.

3.2.2 Materials: Genomic Sequences

The gene sequences of the papaya and *S. latifolia* X chromosomes, and the sequence of chromosome 19 of *P. trichocarpa* were downloaded from the NCBI ftp site (https://www.ncbi.nlm.nih.gov/Ftp/). Definitions of previous strata and names of genes were collected from the published papers [50, 65, 82-86]. Genome sequence of the

chromosome 15 of *S. suchowensis* was downloaded from the *S. suchowensis* website (115.29.234.170/willow) [86]. Genomic and gene sequences of *Ectocarpus* sp. sex chromosome V were downloaded from the OrcAE database (http://bioinformatics.psb.ugent.be/orcae/overview/Ectsi). The assemblies of the *M. Lychnidis-dioicae* mating type chromosome a1 and a2 were downloaded from EMBL-ENA (accession no. PRJEB79120) [84].

3.2.3 Approach

In the case of *S. latifolia* [65, 82] and papaya [50], synonymous substitution rates were earlier used to estimate divergence between X-linked and Y-linked sequences. As discussed earlier, the applicability of all current methods is, however, limited to organisms whose both sex chromosomes are partially or completely sequenced. Furthermore, the paucity of gametologous sequences due to the degradation of Y or W chromosome coupled with the inherent difficulties in sequencing the Y or W chromosome render these methods limited in their ability to reliably detect the evolutionary strata.

Because similar evolutionary processes have likely shaped the sex chromosomes of both plants and mammals, the genomics of mammalian sex chromosomes should be amenable to plant sex chromosomes as well [60]. We, therefore, adapted and implemented our algorithm, earlier tested on the human X chromosome [42], for the analysis of plant sex chromosomes. The overarching goal of this study was to understand the early processes in the evolution of sex chromosomes. Since the X (or Z) chromosomes have not been completely sequenced for any plant

species yet, we used the available gametologous gene or coding sequences on the X for delineating the strata on the plant X chromosomes. Additionally, we applied our method to the recently sequenced sex chromosome V of the brown alga *Ectocarpus* sp. that has a haploid sex determination system (UV system) and to the mating-type chromosomes of an anther-smut fungus *Microbotryum lychnidis-dioicae*. In what follows, we present the results from our analysis of the gene or protein-coding sequences of the papaya and *S. latifolia* X chromosomes [50, 83], incipient sex chromosome sequences of *Populus trichocarpa* and *Salix suchowensis*, and that of the sex chromosome V of *Ectocarpus* sp. and mating-type chromosomes of *M. lychnidis-dioicae*. We discuss the convergence of our predictions with those from the existing methods that require sequence homology and the applicability of our method to identifying the evolutionary strata even in the absence of sequence homology information.

3.2.4 Evolutionary Strata on Papaya X Chromosome

In papaya, three different sex types: female (XX), male (XY) and hermaphrodite (XY^h), are present. Male Y and hermaphrodite Y chromosome differ by only 1.2% in their DNA sequence content [87]. Wang et al. have reported 70 transcription units on the X chromosome that has homologous gene sequences on Hermaphrodite specific Y (HSY) chromosome and assayed these transcripts through synonymous substitution rates analysis and reported two evolutionary strata and one collinear region [50].

We applied our segmentation and clustering algorithm to the concatenated sequences of 63 X-linked genes including UTRs, exons and introns, using the gene

order reported in Wang et al. (2012). Our method correctly classified into clusters corresponding with the three previously inferred evolutionary strata [Fig. 3.5, Table 2, and Supplementary Table 2]. The first gene cluster includes 20 genes [genes 1–20, Fig. 3.5, Table 2], corresponding to the first inversion on the Y chromosome, which was estimated to have occurred around ~7 million years ago [50], and was named evolutionary stratum 1. Genes 21–42 were assigned to the second gene cluster [Fig. 3.5, Table 2], corresponding to the second evolutionary stratum that is thought to have arisen by a second HSY inversion ~2-3 million years ago [50]. Genes 43–45 were previously assigned to the second evolutionary stratum, but our algorithm groups them into the third cluster, comprised of 21 X-linked genes (43 to 63), most of which were previously reported to be in a collinear region where the X-linked genes still recombine with their Y-linked homologs [Fig. 3.5, Table 2].

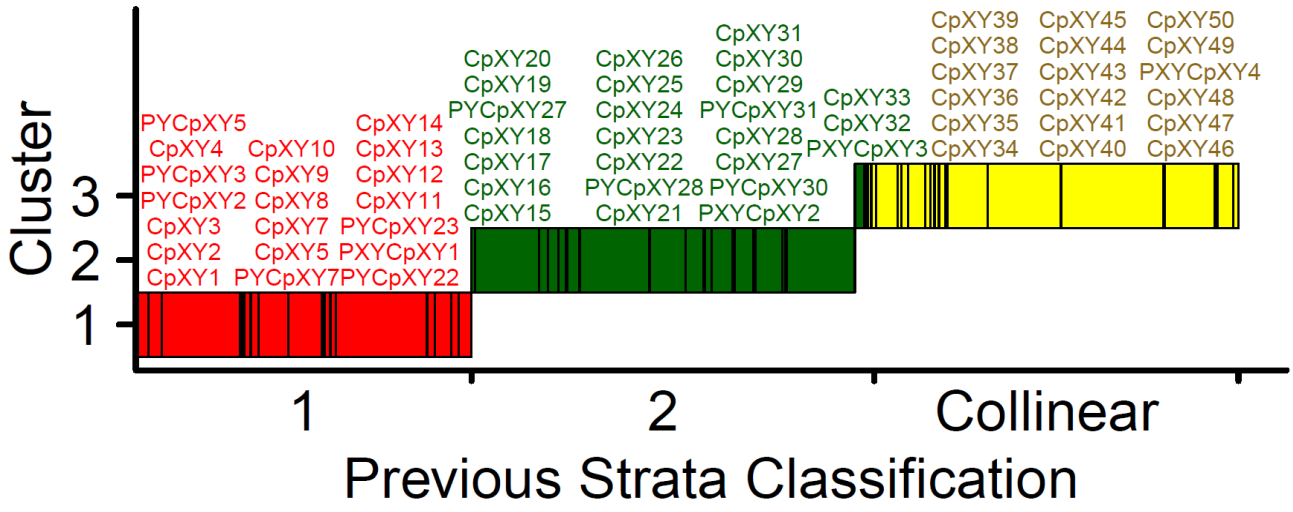

Figure 3.5. Strata on papaya's X identified using previously assayed X-linked genes. The segmentation and clustering algorithm was applied to a concatenated sequence of the papaya X-linked genes, which have been previously assayed using substitution rate analyses. Previous strata are shown in color, with first stratum in red, second in green and the collinear region in yellow. Genes in each cluster are labeled above the cluster, similarly color-coded. We used Markov model of order 2 to perform the segmentation and clustering at thresholds of 0.3 and 0.15 respectively [88].

Genes 43–45 were found compositionally more similar to genes in the collinear regions. There could be three possible reasons for this. First, after the second inversion, at least seven chromosomal rearrangements might have occurred on HSY [50], some of which might have resulted in realignment of the previously inverted genes to their homologs on X, and so these genes display similar X-Y divergence as the other genes in the collinear region. Note that while gene 43 is a pseudogene, the synonymous nucleotide divergence between the X and Y for genes 44 and 45 is closer to gene 46 of the collinear region than the genes on the second stratum. Second, about half of the gene pairs in the collinear region of the X and Y are also reported to have ceased to recombine [50], and therefore, their divergence level is expected to be higher than the other half of the collinear region and is expected to be more closer to that of the genes

43–45. As these genes are in an initial phase of recombination suppression, they could be potentially forming a new stratum on the papaya X chromosome. This also suggests that inversion is not the only mechanism responsible for recombination suppression [50]. Third, the genes 43–45 are actually part of the collinear region and have been undergoing recombination suppression together with the other non-recombining genes in this region. These data show that our composition based approach could recapitulate not just the previously reported stratum boundaries but could also provide hints for redefining novel stratum boundaries. Complete sequencing of papaya X chromosome will help test our hypotheses, validate predictions, and determine the stratum structure more robustly.

Table 3.2. Summary of the evolutionary strata on the papaya X chromosome predicted using the segmentation and clustering algorithm, and through the substitution rate analysis by Wang et al. [50, 88].

Gene number	Gene name	Strata identity from substitution rate analysis	Strata identity from segmentation and clustering analysis
1	CpXY1	1	1
2	CpXY2	1	1
3	CpXY3	1	1
4	PYCpXY2	1	1
5	PYCpXY3	1	1
6	CpXY4	1	1
7	PYCpXY5	1	1
8	PYCpXY7	1	1
9	CpXY5	1	1
10	CpXY7	1	1
11	CpXY8	1	1
12	CpXY9	1	1

(*table continues)*

Table 3.2 *(continued).*

Gene number	Gene name	Strata identity from substitution rate analysis	Strata identity from segmentation and clustering analysis
13	CpXY10	1	1
14	PYCpXY22	1	1
15	PXYCpXY1	1	1
16	PYCpXY23	1	1
17	CpXY11	1	1
18	CpXY12	1	1
19	CpXY13	1	1
20	CpXY14	1	1
21	CpXY15	2	2
22	CpXY16	2	2
23	CpXY17	2	2
24	CpXY18	2	2
25	PYCpXY27	2	2
26	CpXY19	2	2
27	CpXY20	2	2
28	CpXY21	2	2
29	PYCpXY28	2	2
30	CpXY22	2	2
31	CpXY23	2	2
32	CpXY24	2	2
33	CpXY25	2	2
34	CpXY26	2	2
35	PXYCpXY2	2	2
36	PYCpXY30	2	2
37	CpXY27	2	2
38	CpXY28	2	2
39	PYCpXY31	2	2
40	CpXY29	2	2
41	CpXY30	2	2

(table continues)

Table 3.2*(continued).*

Gene number	Gene name	Strata identity from substitution rate analysis	Strata identity from segmentation and clustering analysis
42	CpXY31	2	2
43	PXYCpXY3	2	Collinear
44	CpXY32	2	Collinear
45	CpXY33	2	Collinear
46	CpXY34	Collinear	Collinear
47	CpXY35	Collinear	Collinear
48	CpXY36	Collinear	Collinear
49	CpXY37	Collinear	Collinear
50	CpXY38	Collinear	Collinear
51	CpXY39	Collinear	Collinear
52	CpXY40	Collinear	Collinear
53	CpXY41	Collinear	Collinear
54	CpXY42	Collinear	Collinear
55	CpXY43	Collinear	Collinear
56	CpXY44	Collinear	Collinear
57	CpXY45	Collinear	Collinear
58	CpXY46	Collinear	Collinear
59	CpXY47	Collinear	Collinear
60	CpXY48	Collinear	Collinear
61	PXYCpXY4	Collinear	Collinear
62	CpXY49	Collinear	Collinear
63	CpXY50	Collinear	Collinear

3.2.5 Evolutionary strata on *Silene latifolia* X Chromosome

The dioecious plant, *Silene latifolia*, has been used as a model system for the study of plant sex chromosomes evolution. Initially, only five sex-linked gene pairs were used to assay the evolutionary strata on the *S. latifolia* X chromosome [82];

synonymous substitution rate analysis of these gene pairs indicated three evolutionary strata on the X chromosome (Table 3) [82]. Bergero et al. extended this list by adding three new gene pairs, which, however, suggested the presence of only two evolutionary strata on the X chromosome [65]. A more recent study based on a larger set of 29 fully sex-linked genes and 6 additional genes in the pseudoautosomal region identified a new stratum along with the previously identified two older strata [65] and a pseudoautosomal region [83]. However, this was inferred by assigning these genes to the strata based on their positions relative to the previously studied genes, rather than by performing *de novo* statistical analysis of these genes.

We performed the statistical analysis of 34 genes, whose coding sequences were previously made available. We applied our segmentation and clustering algorithm to the concatenated exonic or coding sequences of the 34 X-linked genes, which included 5 genes from the pseudoautosomal region. Our approach classified these genes into four distinct clusters [Fig. 3.6, Table 3, and Supplementary Table 3]. Cluster 1 contains genes 1–11 of the oldest stratum, evolutionary stratum 1 [83] [Fig. 3.6, Table 3]. Cluster 2 contains genes 12-19 of the previously described evolutionary stratum 2. Cluster 3 contains genes 20 – 28 and a part of gene 29, which lie on the *S. latifolia* Xp arm and have recently been reported to be forming the youngest stratum (evolutionary stratum 3 in [83]). Cluster 4 contains a part of gene 29 and genes 30−34; genes 30−34 were previously reported to be the pseudoautosomal genes while gene 29 (SlX6b in Fig. 3.6) was predicted to be in the youngest stratum (stratum 3 in Fig. 3.6; [83]). Interestingly, our method placed a part of SIX6b (~60%) in cluster 4 and the remaining part (~40%) in cluster 3. This region with SIX6b spanning the pseudoautosomal region and youngest

stratum is thus difficult to be characterized. Although Bergero et al. have placed SIX6b in stratum 3 [83], our results suggest a reassessment of this, also in the light of a recent report [83], on pseudoautosomal region having undergone significant evolutionary changes that included at least two events of expansion and one event of inversion. Cluster 4 thus represents the pseudoautosomal region of the *S. latifolia* X chromosome [Fig. 3.6].

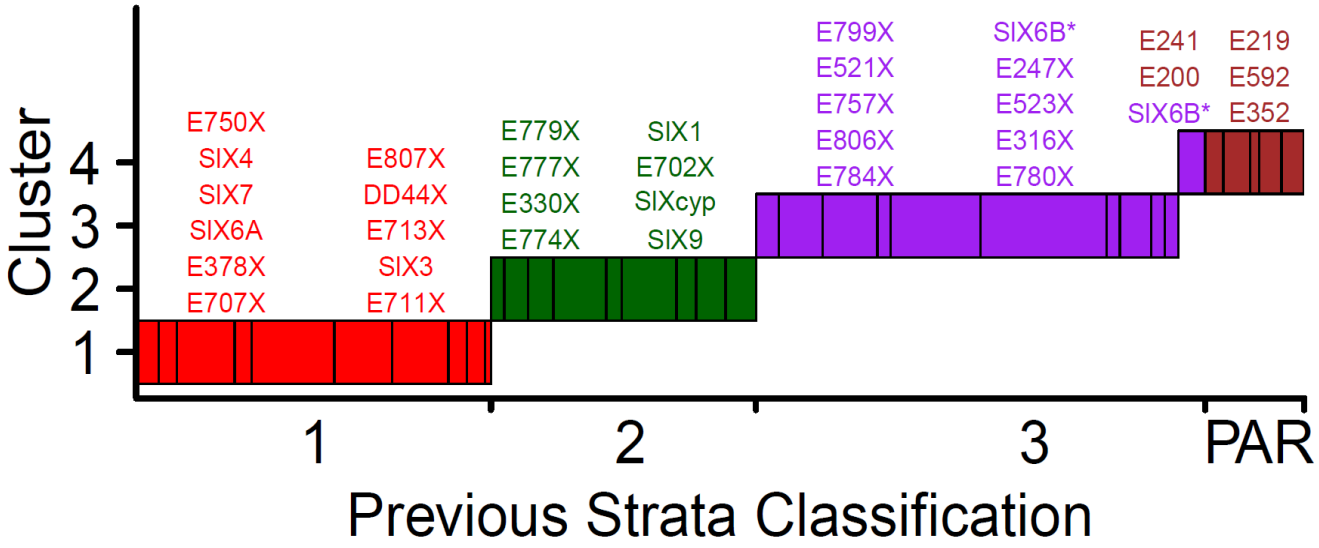

Figure 3.6. Strata identified using the coding sequences of previously-assayed *S. latifolia* X-linked genes. The segmentation and clustering algorithm was applied to a concatenated coding sequence of the *S. latifolia* X-linked genes, which have been previously assayed using substitution rate analyses and consensus map. Previous strata are shown in color, with the first stratum in red, second in green, third in violet and the PAR region in brown. Genes in each cluster are labeled above the cluster, similarly color-coded. We used Markov model of order 2 to perform the segmentation and clustering at thresholds of 0.7 and 0.15 respectively [88].

The above results further reinforce our method's unique capability to decipher evolutionary strata even in the absence of sequence information on the Y (or W) sex chromosome. Our method could recapitulate the previously identified evolutionary strata

on the *S. latifolia* X chromosome, predicting three evolutionary strata and one pseudoautosomal region. These findings suggest that despite its recent origin, ~10 - 20 mya [65, 82], many chromosomal rearrangements have already occurred on the *S. latifolia* sex chromosomes, which resulted in multiple recombination suppression events in a relatively short span of time. This rapid evolution is a consequence of at least two expansions and one inversion within the pseudoautosomal region [83] and two large inversions on the Y chromosome [66]. Accumulating data from various studies clearly indicate that inversions and other chromosomal changes, e.g. rearrangement, duplication, expansion or translocation, play a crucial role in the evolution of sex chromosomes in plants and in the formation of evolutionary strata on the X (or Z) chromosomes.

Table 3.3. Summary of the strata on the *S. latifolia* X chromosome predicted using the segmentation and clustering algorithm, and through substitution rate [65, 82] and consensus map analysis [65, 82, 83, 88].

Gene number	Gene identity	Substitution rate analysis		Consensus map analysis [83]	Segmentation and clustering
		[82]	[65]		
1	E707X	-	-	1	1
2	E378X	-	-	1	1
3	SlX6A	-	1	1	1
4	SlX7	-	1	1	1
5	SlX4	1	1	1	1
6	E750X	-	-	1	1
7	E711X	-	-	1	1
8	SlX3	1	1	1	1
9	E713X	-	-	1	1
10	DD44X	2	1	1	1

(table continues)

Table 3.3 *(continued).*

Gene number	Gene identity	Substitution rate analysis		Consensus map analysis [83]	Segmentation and clustering
		[82]	[65]		
11	E807X	-	-	1	1
12	E774X	-	-	2	2
13	E330X	-	-	2	2
14	E777X	-	-	2	2
15	E779X	-	-	2	2
16	SlX9	-	-	2	2
17	SlXcyp	-	2	2	2
18	E702X	-	-	2	2
19	SlX1	3	2	2	2
20	E784X	-	-	3	3
21	E806X	-	-	3	3
22	E757X	-	-	3	3
23	E521X	-	-	3	3
24	E799X	-	-	3	3
25	E780X	-	-	3	3
26	E316X	-	-	3	3
27	E523X	-	-	3	3
28	E247X	-	-	3	3
29	SlX6B	-	2	3	PAR/3
30	E200	-	-	PAR	PAR
31	E241	-	-	PAR	PAR
32	E352	-	-	PAR	PAR
33	E592	-	-	PAR	PAR
34	E219	-	-	PAR	PAR

3.2.6 Evolutionary strata on *Populus trichocarpa* Chromosome 19

After validation of our approach on model plants papaya and *S. latifolia*, we investigated the incipient sex chromosome of *Populus trichocarpa*, which has not yet

been analyzed for the stratum formation. Like *P. trichocarpa*, many plants have been reported to have evolved the sex chromosome system but their sex chromosomes have not been sequenced yet. Many of these are in the initial phase of sex chromosome evolution as in *Asparagus*. It was tempting to apply our method to the just sequenced sex chromosomes of these plants to decipher the early stages of recombination suppression between the gametologous sex chromosomes and identify the signatures of early formation of evolutionary strata on their X or Z sex chromosomes, which have not yet been probed due to lack of the sequences of the Y and W chromosomes.

Recently, the chromosome 19 of *P. trichocarpa* was reported to be an incipient sex chromosome due to the suppression of recombination and distorted segregation around the sex determining locus. The sex determining locus is located on peritelomeric end of chromosome 19 and the recombination suppression has spread up to 3−4 Mbp of this chromosome [89]. This region was also observed to have distinct features in comparison to the entire *P. trichocarpa* genome. It has been further reported that the frequency of simple sequence repeats (SSR) is significantly higher in the peritelomeric region of chromosome 19, which mapped to about 5.1 Mbp while the distal portion has lower SSR frequency than expected. There is an overabundance of nucleotide-binding-site-leucine-rich-repeat (NBS-LRR) genes in a 2.45 Mbp segment at the peritelomeric portion, along with higher occurrence of small microRNAs, located closer to the gender-determining locus [89-90]. It was speculated that these changes contributed towards the emergence and evolution of a nascent sex chromosome in *P. trichocarpa* [89]. Differential occurrence (and non-occurrence) of these evolutionary events may also have led to the fragmented compositional structure of this incipient sex chromosome,

resulting in the formation of evolutionary strata on this chromosome.

We implemented our stratum detection method to decipher the stratification structure on chromosome 19. In application to the complete DNA sequence (~16 Mbp) of this chromosome, our method generated three distinct clusters (Fig. 3.7, Supplementary Table 4). The first cluster ranges from 0–3 Mbp, which contains a region of recombination suppression, sex determining locus, and NBS-LLR genes in high abundance. Second cluster which ranges from 3–7 Mbp could be a region of reduced recombination, which has been reported to be a part of distorted regions in a previous study [90]. This region also has a high frequency of SSR. The third cluster ranges from 7–16 Mbp and is still recombining with its homologous partner as reported in Tuskan et al. 2008, which referred to this as the “non-distorted” region. This region has lower SSR frequency than expected. Thus, our method deciphered two, yet unknown, evolutionary strata on *P. trichocarpa*’s chromosome 19 [Fig. 3.7, Supplementary Table 4].

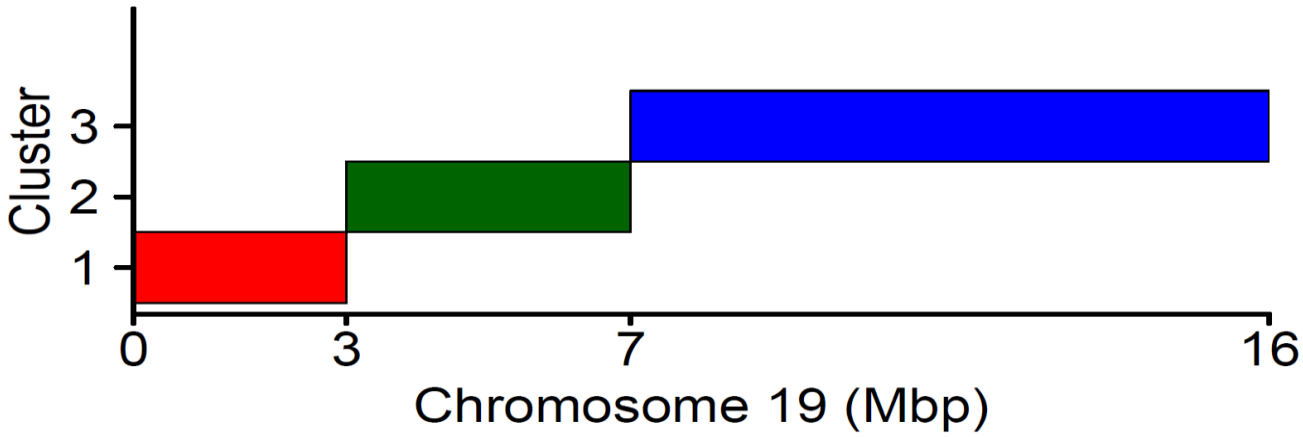

Figure 3.7. Strata identified on chromosome 19 of *P. trichocarpa*. Here we show the clusters that are determined using the entire sequence of the chromosome 19, an incipient sex chromosome, of *P. trichocarpa*. We used Markov model of order 2 to perform the segmentation and clustering at thresholds of 0.30 and 0.001 respectively [88].

Recent study based on 650 sex-linked SNPs, which are present in heterozygous state in males and homozygous state in females, suggests an XY sex-determining system in *P. trichocarpa* [91]. This contradicts a previous study that suggested ZW (female heterogametic) sex determination system in *P. trichocarpa* [90]. Previous studies have mapped the sex determining region to a single locus, the proximal telomeric end of chromosome 19 [89], while in the recent study by Geraldes et al., the SNPs associated with sex mapped to different genomic regions indicating assembly errors in the *P. trichocarpa* genome [91]. Although many sex-specific markers mapped to proximal telomeric end of Chromosome 19 substantiating previous studies that implicated this region as the locus of sex determination [90, 92], the estimated size of the sex-determining sequence is reported to be only around 100 Kbp [91].

Geraldes et al. also estimated the origin of sex chromosomes in *P. trichocarpa* to be around 6–7 Mya [91], and so it is possible that the sex determining region might have expanded. Taking cue from the papaya XY system that originated around 7 Mya but has its Y^h (Y chromosome in hermaphrodite) already expanded to more than twice of the original size [50] resulting in at-least two evolutionary strata observed on its X, this seems plausible that the actual size of the sex determining region could be larger than reported, possibly due to large tandem duplications and accumulation of transposable elements in the Y. Though only four small male-specific contigs were retrieved earlier, it is possible that more male-specific contigs, particularly in the regions of higher divergence to the female reference sequence, could be retrieved later enabling search for and confirmation of other sex-linked strata in *P. trichocarpa* [91].

3.2.7 Analysis of *Salix suchowensis* Chromosome 15

Salix and *Populus* are the sister genera in *Salicaceae*. The chromosomes of *Salicaceae* are typically metacentric and small. While the chromosome 19 has evolved as the sex chromosome in poplar, chromosome 15 is known to be the sex chromosome in willow (*S. suchowensis*). Absence of syntenic chromosomal segments in these two lineages suggests different origins for these sex chromosomes [86]. Comparative analysis by Hou et al. suggests that the transformation of autosomes into sex chromosomes happened after the divergence of *Salix* and *Populus* in these genera. Notably *S. suchowensis* was found to have the ZW sex determination system [86].

Application of our method to the chromosome 15 of *S. suchowensis* at the same parameter setting as was used for *P. trichocarpa* resulted in a single cluster. This suggests that the sex determining region is too small to be detected by our method, and is probably the only non-recombining region in this sex chromosome. The other reason could be the yet incomplete assembly of the chromosome 15 (personal communication with Dr. Tongming Yin).

3.2.8 Sex Determining Region on Brown Alga *Ectocarpus* sp. Sex Chromosome V

In *Ectocarpus* sp., sex is determined during the haploid phase. This sex determination system is referred to as the UV system, where U and V represent the female and male sex chromosomes respectively. In this system, recombination is suppressed in the sex determining regions of U and V haplotypes [85]. Gene degeneration as a consequence of recombination suppression at SDR loci has been observed in both male and female haplotypes to a similar level as expected. The male

and female SDRs are of similar size (~1 Mbp) and have been reported to be diverging for over 70 Mya. Pseudo-autosomal regions (PARs) that are still recombining flank the SDR. The total size of PAR in both the male and female sex chromosomes is around 4.08 Mbp. In the previous studies [85], evolutionary strata were not reported likely due to the limited expansion of SDR. However, recently the UV systems have been found to display patterns akin to evolutionary strata on their sex chromosomes, e.g. the divergence patterns resulting from at least two recombination suppression events in the UV system of the bryophyte [85].

On application of our stratum detection algorithm to a concatenated sequence of 357 genes retrieved from the scaffolds of male sex chromosome V, we were able to recover the boundaries precisely, both between PAR1 and SDR and between SDR and PAR2 (Fig. 3.8, Supplementary Tables 5 and 6). Application of our method to the concatenated genomic sequences that represent the current assembly of male sex chromosome V (strain Ec32; [85]) resulted in three clusters that corresponded to known PAR1, SDR and PAR2 respectively (Fig. 3.9). Though the boundary between PAR1 and SDR was detected with precision, the boundary between SDR and PAR2 was detected with less precision (Fig. 3.9, Supplementary Tables 7 and 8).

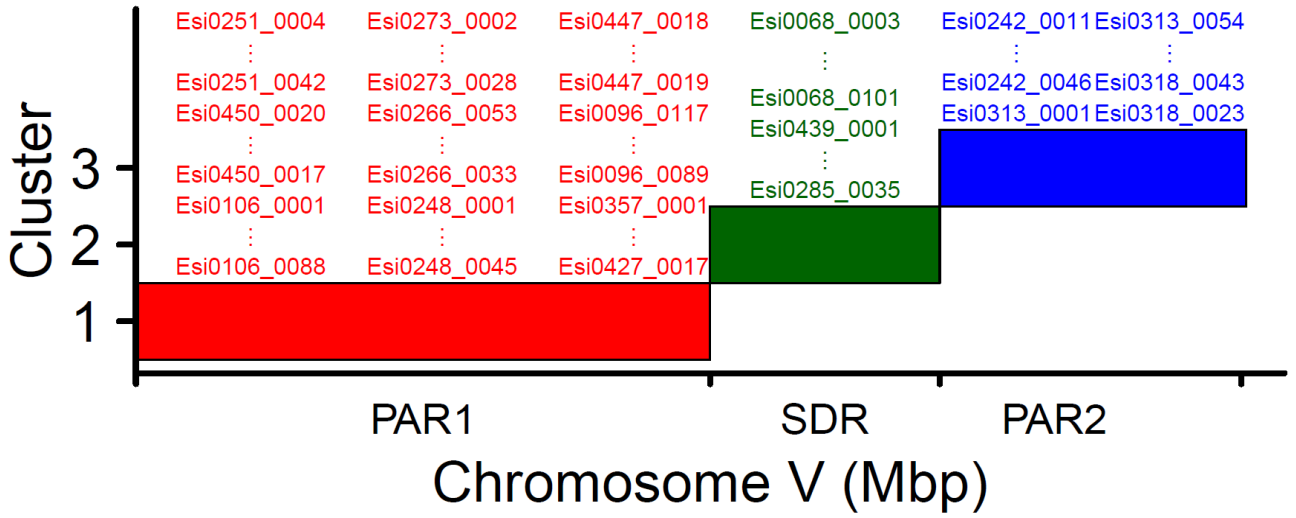

Figure 3.8. The sex determining region (SDR) and pseudoautomosal regions (PAR) identified using the gene sequences of male sex chromosome V of *Ectocarpus* sp. The segmentation and clustering algorithm was applied to a concatenated sequence of *Ectocarpus* V-linked genes. Previously reported SDR and PAR are shown in color, with PAR1 in red, SDR in green and PAR2 in blue. Genes in each cluster are labeled above the cluster, similarly color-coded. We used Markov model of order 2 to perform the segmentation and clustering at thresholds of 0.20 and 0.007 respectively [88].

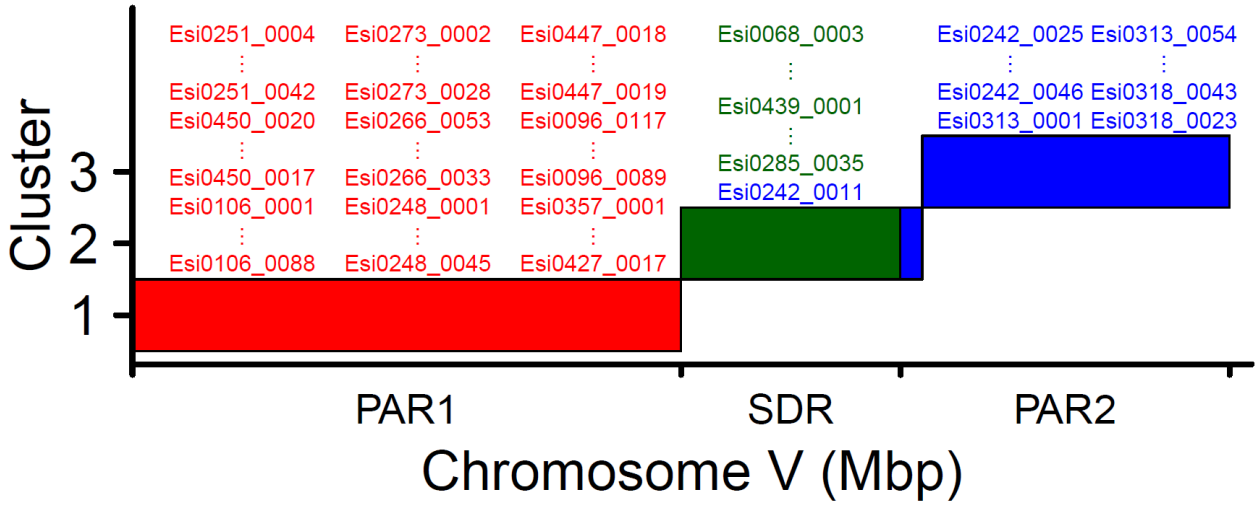

Figure 3.9. The sex determining region (SDR) and pseudoautomosal regions (PARs) identified on the sex chromosome V of *Ectocarpus* sp by the segmentation and clustering algorithm. Here we show the clusters determined using the entire sequence of the chromosome V of *Ectocarpus* sp. Previously reported SDR and PARs are shown in color, with PAR1 in red, SDR in green and PAR2 in blue. Genes in each cluster are labeled above the cluster, similarly color-coded. We used Markov model of order 2 to perform the segmentation and clustering at thresholds of 0.40 and 0.01 respectively [88].

3.2.9 Evolutionary Strata on Anther-Smut Fungus *Microbotryum lychnidis-dioicae* Mating-Type Chromosomes

The dimorphic *M. lychnidis-dioicae* mating-type chromosomes a1 and a2 are of size ~3.5 Mbp and 4 Mbp respectively. The majority of the mating-type chromosomes a1 and a2, ~3.08 Mbp and 3.67 Mbp respectively, are non-recombining [93-94]. These were confirmed by optical maps and marker segregation [93-94]. This is, however, in contrast to the other studies that reported the recombination suppression to be limited to small regions [61, 95]. The regions of suppressed recombination are flanked by PARs at both ends of the mating type chromosomes, spanning ~0.20 Mbp on either side [84].

As expected, the non-recombining regions harbored more repetitive elements and had lower gene density than the autosomes and PARs [84]. Furthermore, the higher divergence between a1 and a2 gametologous genes indicated rather ancient suppression of recombination between the two chromosomes. However, the previous studies couldn't find any stratification based on synonymous substitution rate analysis. Therefore, it was proposed that the frequent rearrangements within mating-type chromosomes may have obfuscated the stratum structure, or the absence of stratification may also be a consequence of gene conversion in some DNA segments [84].

We applied our integrated segmentation and clustering method to the a1 and a2 mating type chromosomes. While our method could recover the boundaries between pseudo-autosomal regions (pPAR and qPAR) and the non-recombining region and predicted five new strata within the non-recombining region in mating type chromosome a2 (Fig.3.10, Supplementary Table 9), it could recover the boundary between qPAR and the non-recombining region but not between pPAR and the non-recombining region in

mating type chromosome a1. Our method generated five clusters for the mating type chromosome a1 including a cluster that contains the previously defined pPAR region and another cluster that corresponds to qPAR (Fig. 3.11, Supplementary Table 10).

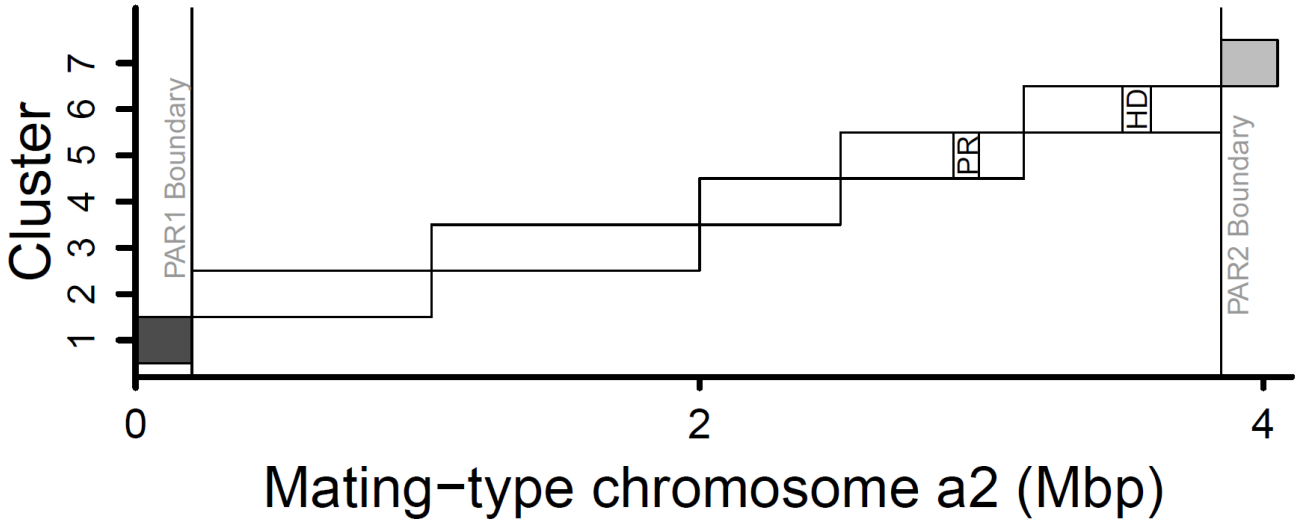

Figure 3.10. Strata identified on the mating type chromosome a2 of *M. lychnidis-dioicae* by the segmentation and clustering algorithm. Here we show the clusters that are determined using the entire sequence of the mating type chromosome a2. Boundaries between PARs (shown in shades of grey) and the non-recombining region (white) are indicated. Five strata are predicted in the non-recombining region, shown as five segments or clusters. We used Markov model of order 2 to perform the segmentation and clustering at thresholds of 0.40 and 0.1 respectively [88].

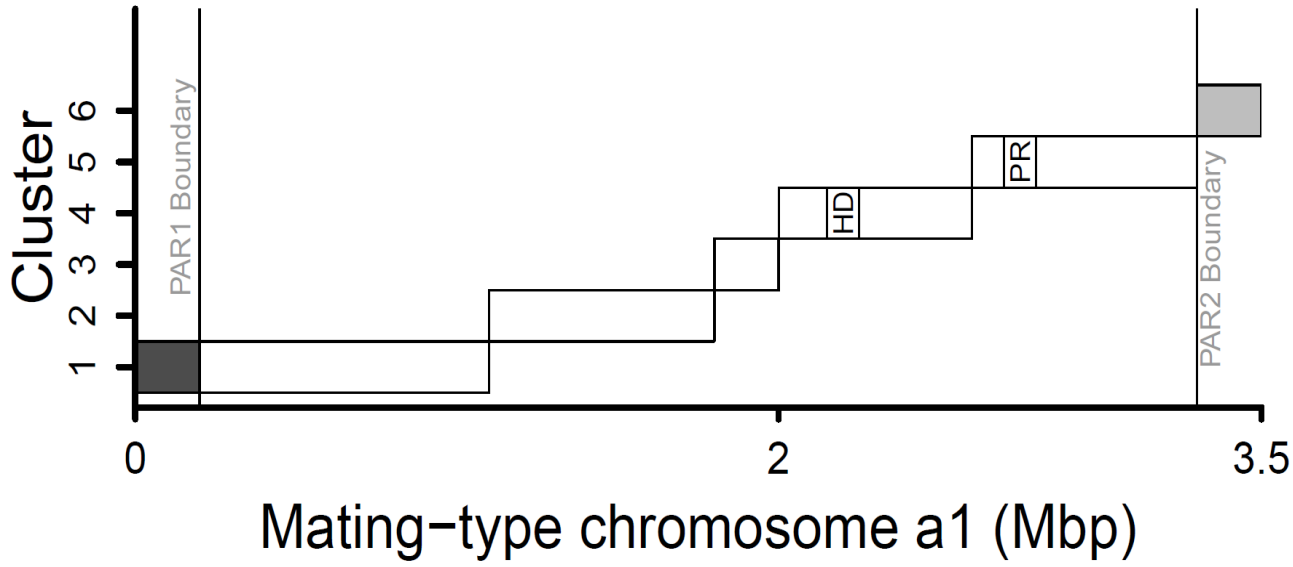

Figure 3.11. Strata identified on mating type chromosome a1 of *M. lychnidis-dioicae* by the segmentation and clustering algorithm. Here we show the clusters that are determined using the entire sequence of the mating type chromosome a1. Boundaries between PARs (shown in shades of grey) and the non-recombining region (white) are indicated. Five strata are predicted in the non-recombining region, shown as five segments or clusters. We used Markov model of order 2 to perform the segmentation and clustering at thresholds of 0.40 and 0.1 respectively [88].

Frequent rearrangements may be shaping the segmental structure of both mating type chromosomes. Notably, even the mating type loci HD (homeodomain) and PR (pheromone receptor) are inverted in both, which were detected in different clusters by our method (Fig. 3.10, Fig. 3.11). We thus predict five evolutionary strata on the dimorphic *M. lychnidis-dioicae* mating-type chromosomes a1 and a2 (Fig.3.10, Fig. 3.11). Previously, Votintseva and Filatov had proposed the presence of at least three strata using fourteen loci, however, the non-recombining region spanned only ~25% of the mating type chromosomes [61].

3.2.10 Conclusions

Our unsupervised composition based approach for the detection of evolutionary strata provides an exploratory tool for probing the incipient sex and mating-type chromosomes and understanding the early evolution of such chromosomes. Dioecious plants are ideal model systems for studying the early stages of sex chromosome evolution, and therefore, we assessed our method on plant sex chromosomes at different stages of evolution. In addition, we analyzed the recently sequenced sex chromosomes of a brown alga and mating-type chromosomes of an anther-smut fungus. In contrast to the frequently used methods including those based on substitution rate, inversion and phylogenetic analysis, our method could detect evolutionary strata without requiring the sex-linked gametologous sequences. Our method performed consistently well in identifying the known evolutionary strata on the X chromosomes of papaya and *S. latifolia*, and deciphered two, yet unknown, evolutionary strata on the *P. trichocarpa's* incipient sex chromosome. Furthermore, we could recover the previously reported SDR and PARs on the sex chromosome V of brown alga *Ectocarpus* sp., and most known PARs and the non-recombining regions in the mating-typing chromosomes of anther-smut fungus *M. lychnidis-dioicae*. We also predicted five strata in the non-recombining region of each of two mating-type chromosomes of *M. lychnidis-dioicae*.

In summary, our successful attempt in identifying the evolutionary strata in dioecious plants, the sex-determining region in a different sex determination system, the UV system in a brown alga, and the non-recombining regions in the mating-type chromosomes of an anther-smut fungus demonstrates the broader applicability of our comparison-independent method. Although there are very few plant, algal and fungal

species with completely or even partially sequenced sex chromosomes, we expect this to change soon due to advances in sequencing technology, which will likely provide us with plant, algal and fungal sex or mating-type chromosomes at different stages of evolution, and thus bring forth new opportunities to probe early events in the formation and evolution of sex and mating-type chromosomes.

3.3 Discussion

Our study revealed that the comparison-free integrated segmentation and clustering method could classify even a handful of genes by evolutionary strata as defined by the previous studies. It could accomplish this by utilizing the information on the sequence of X- (or Z-) linked genes or chromosomes in order to delineate the differential compositional biases arising because of time-lagged sequential formation of evolutionary strata, with each stratum accumulating mutations proportional to the time since the cessation of recombination. Notably, the segmentation and clustering algorithm’s parameter choices, including the significance thresholds, are preset at the outset, and therefore the proposed procedure is free from the biases due to “artificial” human interventions, such as, halting the algorithm at will to secure selected stratum boundaries as was done previously, and thus introducing, for the first time, an unrestricted integrative algorithm to decipher segmental structures within large eukaryotic genomes [42]. This study indicates that a more precise delineation of stratum boundaries is now made feasible within the integrated framework of segmentation and clustering, which works independently of the gene structure and Y (or W) chromosome sequence information.

Interestingly, the formation of several strata on the relatively recent *S. latifolia* sex chromosome, as revealed in previous studies and further substantiated in this study, belies the previous reports of only 4−5 strata on the 165 million year old human X chromosome whose youngest stratum was reported to have originated ~30 million years ago. This suggests that many more undetected inversion and chromosomal rearrangement events might have shaped our sex chromosomes. Indeed, our algorithm, when applied to the whole human X chromosome, detected four additional, novel strata in the older regions of the X chromosome [42]. This indicates more recombination suppression events might have accompanied the initial phase of sex chromosome evolution than previously reported, which is further supported by our analysis of the *S. latifolia* X chromosome. The older regions were hard to analyze for the stratum presence until recently because of the degradation of the Y (or W) chromosome and subsequent loss of Y- (or W-) linked gametologous genes in these regions. This critical barrier to progress is now dismantled by the composition based approach because of its ability to identify stratum without Y (or W) sequence information.

Two evolutionary strata and one collinear region had been previously deciphered on the papaya X chromosome, however, the number of strata could be greater, particularly, in the collinear region where many genes have been reported to have diverged, albeit not through inversions on the Y chromosome [50]. This also indicates that inversion is not the only mechanism for recombination suppression but other mechanisms may also be playing a role in the stratum formation. We predict more strata to be deciphered by our method once the whole X chromosomes of *S. latifolia* and papaya will be sequenced and made available for analysis. Indeed, this was borne out

in our study of the whole chromosome 19 of *P. trichocarpa*, an incipient sex chromosome, which is younger than *S. latifolia* and papaya sex chromosomes, yet has several strata already formed on it.

In addition to the paucity of Y or W sex chromosome sequences in the older regions, which is required by most methods that exploit the synonymous substitution rate to estimate the divergence between X-Y or Z-W gametologous sequences, saturation of synonymous substitutions could also potentially limit the power of these methods in detecting the older strata. The integrative segmentation and clustering method, which detects the strata via segmental heterogeneity within the sex chromosome of the homogametic sex (i.e. X or Z chromosome), is not sensitive to saturation of different sex determination systems including that of plant, fungal and mammalian systems. The differential accumulation of repetitive elements is well exploited by this method in deciphering the strata, which appears less affected by the saturation, thus augmenting its power in detecting signals differentiating the evolutionary strata by compositional biases.

Interestingly, the repetitive elements appear to contribute significantly to the distinct compositional biases of the strata. This was substantiated by our study of human X chromosome, *P. trichocarpa*'s chromosome 19 and gene sequences of papaya's X with repeats masked by the RepeatMasker program [96] ─ the stratification structure disappeared completely from the repeat-masked human X chromosome (Fig. 3.2) and *P. trichocarpa*'s incipient sex chromosome with repeats masked (Supplementary Fig. 4) while the two strata on the repeat masked papaya's X merged into one (Supplementary Fig. 5). In the latter, the merged strata was differentiated from

the still recombining collinear region, however, the two stratified regions were recovered when the segmentation stringency was relaxed slightly (Supplementary Fig. 6). Due to functional constraints on the genic or coding sequences, the repeats are expected to accumulate more in the non-coding regions. The availability of *P. trichocarpa*'s incipient sex chromosome sequence afforded the opportunity to exploit the full potential of repetitive elements, exemplified in their high density (Supplementary Table 11, Supplementary Fig. 7), in discriminating between the evolutionary strata. Indeed, in this case, the repeats seem to amplify the compositional biases of the strata to an extent that other signals, if any, are not required to be exploited for delineating the stratum boundaries. This demonstrates the overwhelming contribution of the repetitive elements in shaping the distinctive composition of the evolutionary strata. Apparently, the gene sequences, comprised of the sequences of exons, introns and UTRs, used in deciphering the strata in papaya have much lower repeat density (Supplementary Table 12, Supplementary Fig. 8), and therefore, the repetitive elements may only have partial contribution in shaping the compositional biases of strata that could be detected by our method. Indeed, masking the repeats didn't prevent the concatenated gene sequences from segmentation; the two resulting segments corresponded to the merged strata and the still recombining region. This required tuning the method's parameters (segmentation and clustering thresholds) to detect more subtle signals that could work in combination with the repetitive elements to aid in the detection of the papaya X's evolutionary strata and the recombining region. As expected, when the segmentation stringency was relaxed further, the stratification structure was deciphered even in repeat masked papaya X's concatenated gene sequences, although with slightly less precision

(Supplementary Fig. 6).

Our results highlight the role of repetitive elements in the stratification of sex chromosomes, and also the contributions of other subtle signals or factors in shaping the stratum structure within sex chromosomes, which are detected at more intrusive levels of segmentation. In fact, this was also borne out in our study of the *S. latifolia's* X sequences, which were merely coding sequences devoid of repeats (Supplementary Table 13) and therefore needed substantial relaxation of segmentation stringency to decipher the underlying stratum structure (Fig. 3.6). We believe the differential level of mutations in strata may also be contributing to stratum specific compositional biases, which are likely detected at more intrusive levels of segmentation. Our results demonstrate the power of the integrative segmentation and clustering method in detecting the evolutionary strata in whole sex chromosomal sequence independent of the gene structure information, or in a handful of gene or coding sequences irrespective of the presence or absence of the repetitive elements.

The studies performed so far on the plant, algal and mammalian sex chromosomes, and on the mating-type fungal chromosomes serve as the proof-of-concept of our comparison-free approach in delineating the evolutionary strata, SDRs and PARs, independent of sequence information on the heterozygous sex chromosomes. Therefore, our stratum detection algorithm provides an exploratory tool for understanding the evolution of sex and mating-type chromosomes, with minimal requirements, which makes it stand out among the state-of-the-art techniques in the field.

CHAPTER 4

APPLICATION II: IDENTIFICATION OF TRANS-DOMAIN GENE TRANSFERS IN THE EXTREMOPHILE *Galdieria sulphuraria*

4.1 Background

Horizontal gene transfer (HGT), the exchange of genetic material between organisms by means other than parent-to-offspring (vertical) inheritance [97-101], is recognized as a major force driving prokaryotic genome evolution [98-100]. The importance of HGT in the evolution of microbial eukaryotes is being assessed, or reassessed, in light of new evidence emerging from the rapidly accumulating genome sequence data. Of particular interest is how unicellular algae evolve and modulate their metabolic repertoire through horizontal gene transfer. This assumes special significance because of the biotechnological and commercial importance of these microbial eukaryotes. Numerous recent reports on HGT events involving algae have further galvanized this field, bringing gene flow among algae [101-103]. In contrast to other evolutionary mechanisms for genetic or genomic variations that have shaped the extant algal genomes, such as recombinations, mutations, duplications, deletions or inversions, HGT assumes a special place because of its ability to perpetuate variations and impart novel metabolic capabilities via genetic exchange among reproductively isolated species, that even crosses the domain boundaries as has been frequently observed with numerous algae.

Among the earliest examples of massive HGT involving eukaryotes are the endosymbioses of cyanobacteria into heterotrophic protists, which later transformed into the chloroplast, leading to the emergence of photosynthetic algae [101, 104-105]. The lineages of photosynthetic eukaryotes, Archaeplastida or Plantae, arose thereafter.

Evidence emerging from phylogenetic analysis supported the hypothesis of frequent endosymbiotic and horizontal gene transfer in the evolution of algal organisms [101, 104-109]. Comparative genomics studies have revealed frequent interdomain transfer, mainly from bacteria and archaea, to algae [110-111].

Recent studies have reported on the role of HGT in the evolution and adaptation of an extremophilic red alga, *Galdieria sulphuraria* [109]. The unicellular alga *G. sulphuraria* belongs to the division Rhodophyta whose members from genera, such as, *Cyanidium*, *Galdieria* and *Cyanidioschyzon*, are known to survive in thermo-acidophilic regions. Members of class *Cyanidiophyceae* are one of the major photosynthetic organisms that dwell in extreme environment like hot water springs. *G. sulphuraria* is known to occur naturally in hot sulfur springs in the volcanic areas. Being an extremophile, it exhibits tolerance to saline conditions and high temperatures. It is also known to be tolerant to toxic metals; perhaps it is one of the dominant eukaryotes found in toxic metal environments containing mercury, cadmium, arsenic and other heavy metals. *G. sulphuraria* harbors soluble ATPase with homologs in archaea but not in any other eukaryotes. Numerous pathways in *G. sulphuraria*, for instance, the methylmalonyl-coenzyme A pathway (odd-numbered chain fatty acids and leucine can be metabolized by this pathway) is known to occur in animals but is absent in photosynthetic organisms. *G. sulphuraria* can survive heterotrophically on a plethora of carbon sources owing to its versatile metabolic pathways. These characteristics make it a suitable candidate for processes like bioremediation and for applications in the field of biotechnology.

G. sulphuraria's adaptation in extreme environments where other algae and even

most of the living organisms cannot survive suggests a rather unusual path of evolution of this organism, which can be decoded by the analysis of its genome. Initial studies of the genome of *G. sulphuraria* have revealed the mechanisms and extent of the gene transfer from prokaryotes to the extremophile eukaryote. These studies have implicated 337 genes (~5% of the total genes) in *G. sulphuraria* to horizontal gene transfer [109]. A recent study was focused on the analysis of the organellar genomes of *G. sulphuraria* and comparison to the other members of Rhodophyta [112]. This study revealed that the *G. sulphuraria's* mitochondrial genome is greatly reduced and also displays the highest GC skew among all (nuclear and organellar) genomes [112]. This study also revealed the fastest rate of substitution among all algal members. The chloroplast genome was found to be similar to the other red algal strains with the exception of stem-loop structures present in the *G. sulphuraria* chloroplast.

Consistent with the role of the horizontally acquired genes in conferring novel metabolic capabilities to the recipient organism and helping it to adapt to a particular environment, the predicted "alien" genes in *G. sulphuraria* were reported to confer functions ranging from heavy-metal detoxification to glycerol uptake and metabolism, thus facilitating the adaptation of this species to high temperature, toxic metal-rich, acidic environment [109].

Notwithstanding the efforts made in understanding the algal genome evolution through HGT, our knowledge about the significance of this process remains lacking, because methods for detecting alien genes in algal genomes have primarily relied on detecting a phylogenetic distribution of genes inconsistent with the assumption of vertical inheritance. Phylogenetic approaches require adequate sampling of the

homologous genes for a reliable inference of alien genes [113]. Unexpected placement of genes within orthologous gene trees is an indicator of HGT; however, the lack of orthologous gene sequences from different lineages may render this unreliable for inferences. Furthermore, the genes with no homologs, the so called ‘orphan’ genes, cannot be analyzed using this approach.

Another class of phylogenetic method infers alien genes without making trees. This class of methods searches for unusual phyletic patterns by comparing the genomes of closely related organisms. If a gene is present in the genome of an organism but absent from several of its close relatives, it is inferred as alien [113]. Although high-throughput relative to tree-based methods, this approach requires a reliable sampling of multiple closely related genomes. The level of phylogenetic sampling needed for such an analysis is often elusive, and such analysis cannot be performed, in particular, on algal genomes with sparse phylogenetic sampling available to date. All *G. sulphuraria* studies until now have relied on sequence alignment or phylogenetic methods. However, the power of these methods is diminished by the lack of sequenced genomes of the close relatives of *G. sulphuraria.* The only members of *Cyanidiophyceae* family whose genomes have been completely sequenced are *Cyanidioschyzon merolae* and *Galdieria phlegrea*. While *C. merolae* and *G. sulphuraria* diverged from each other about one billion years ago [109], the similarity between the proteins of *G. sulphuraria* and *G. phlegrea*, the closest relative of *G. sulphuraria* whose genome sequence is available, was approximated to be between those of the human and teleost fish genome. This makes difficult assessing the extent and impact of HGT on the evolution of *G. sulphuraria,* or perhaps most algae isolated and sequenced to

date, through phylogenetic approaches.

To circumvent the limitations of phylogenetic methods and infer alien genes even in the absence of closely related genomes, alternative methods have been developed that can predict HGT without multiple comparisons [4, 37, 109, 113-114]. This class of methods, often called parametric or composition based methods, looks for atypical compositional patterns within the genome of interest [98, 113]. Because the horizontally acquired genes have evolved in genomic contexts different from the recipient genomic context, these genes appear atypical in the recipient genome background, and therefore can be detected by searching for anomalous compositional patterns such as the atypical nucleotide (G+C) composition or unusual dinucleotide or oligonucleotide composition. This approach is suited to identifying the recently transferred genes, as the acquired genes are subject to the host's directional mutational pressure and therefore slowly adjust their composition to the native genome composition. However, Lawrence and Ochman have shown that most alien genes at a snapshot of time are recent acquisitions [115]. Most of the acquired genes are not likely to provide long-term benefits and so are eventually lost. Therefore, the parametric or composition-based methods are powerful tools for quantifying the horizontally acquired DNAs.

While strongly typical ancestral genes and strongly atypical alien genes are identified easily by most parametric methods, compositionally ambiguous genes, that is, weakly typical native and weakly atypical alien genes, which constitute a significant proportion of the total genes in a microbial genome, present significant challenges to these methods [4, 114, 116]. Most parametric methods take a bottom up, gene by gene, approach and therefore incur high rates of false positives and false negatives [4, 113-

114]. Therefore, the acquired regions with clusters of alien genes appear fragmented when predicted using the bottom up methods [4]. To circumvent this limitation of the parametric methods, a top-down recursive segmentation approach that allows analysis of multiple genes simultaneously was proposed (Chapter 2; [4, 37]). Compositionally ambiguous genes are analyzed together with other compositionally distinct genes of a gene cluster (either a native genomic region or a genomic island) and thus are classified more precisely as native or alien. This procedure was earlier successfully applied in identifying genomic islands in bacterial genome [4]. Azad and Li (2013) further developed and implemented integrative framework of recursive segmentation and agglomerative clustering procedure to identify distinct classes of genes of apparently different ancestries in bacterial genomes, which we adapted in to address different biological problems presented in this dissertation.

Previous studies have used highly stringent criteria to obtain conservative estimates of alien genes in eukaryotes; one such recent study reported 75 transfer events with *G. sulphuraria* as the recipient and Bacteria or Archaea as the donor based on high phylogenetic bootstrap support and other stringent criteria [109]. This study, while shedding light on the role of alien genes in the adaptation of an extremophile eukaryote, also brought into focus the challenges in assessing the extent and impact of HGT on the evolution of microbial lineages. Whereas more sensitive and robust phylogenetic approaches aimed at maximizing the signal to noise ratio are clearly needed, particularly for sparse or noisy data, we argue that this bottleneck cannot be circumvented, at least to a reasonable extent, solely by tinkering with phylogenetic tree methods.

We posit that significant advances may result by exploiting the complementary strengths of parametric approaches, which could be used in concert with phylogenetics to more robustly catalog alien genes in eukaryotic genomes. We adapted the integrative framework of recursive segmentation and clustering method in combination with comparative genomics to search the *G. sulphuraria* genome for new, yet unreported alien genes, thus shining a light on the roles of new players in imparting versatility or adaptive capabilities that enable this organism to thrive in many different environments. The pipeline is described below, followed by discussion of the results.

4.2 Materials: Genome and Protein Sequences

The genome sequence of *G. sulphuraria* and the gene coordinates were obtained from the NCBI ftp site and *G. sulphuraria* database. The protein sequences were retrieved from the *G. sulphuraria* contigs based on the gene coordinate and structure information.

Table 4.1. Red algal genome and transcriptome data used in this study [123]

No.	Data	Taxa	Data source
1	Partial genome	*Calliarthron tuberculosum*	[117]
2	Partial genome	*Galdieria phlegrea*	[108]
3	Genome	*Chondrus crispus*	[118]
4	Genome	*Porphyridium purpureum*	[107]
5	Genome	*Cyanidioschyzon merolae*	[119]
6	Genome	*Pyropia yezoensis*	[120]
7	Transcriptome	*Compsopogon coeruleus*	[121]
8	Transcriptome	*Porphyridium aerugineum*	[121]
9	Transcriptome	*Erythrolobus australicus*	[121]
10	Transcriptome	*Rhodosorus marinus*	[121]
11	Transcriptome	*Erythrolobus madagascarensis*	[121]
12	Transcriptome	*Timspurckia oligopyrenoides*	[121]
13	Transcriptome	*Porphyra umbilicalis*	[122]

4.3 Approach

4.3.1 Parametric Method for Identifying Alien DNAs

As the first step in our analysis, we classified *G. sulphuraria* genes as being atypical or typical based on higher order compositional patterns. We implemented the integrative recursive segmentation and agglomerative clustering method to identify distinct classes of DNAs of apparent alien origin in the *G. sulphuraria* genome. The recursive segmentation method was employed to fragment the *G. sulphuraria* genome into compositionally distinct segments within the statistical hypothesis testing framework, followed by an agglomerative clustering procedure to group compositionally similar segments within the same framework. The backbone genome was identified as comprising the largest cluster, whereas atypical genomic DNAs were identified as residents of the smaller clusters. The compositionally atypical genes provided a first pass of putative alien genes, which were further processed using a comparative genomics approach as described below to obtain a highly conservative set of alien genes likely acquired from prokaryotes.

4.3.2 Comparative Genomics Approach

The next round entailed examining the phyletic pattern of the first pass putative alien genes (i.e., their presence or absence in the close relatives of *G. sulphuraria*). We considered 13 red algae whose genome or transcriptome assemblies are available (Table 4.1) to determine (using protein BLAST) the distribution of *G. sulphuraria* alien gene candidates. If an atypical gene (i.e., the encoded protein) was present in the majority (arbitrarily, > 7 taxa) of rhodophytes, then it was removed from the list of

putative alien genes. It should be noted that ancient transfers (i.e., ancestral to a subset of Rhodophyta and other algae/plants) from cyanobacterial or other prokaryotic sources were not of primary interest here, although our phylogenetic validation showed many such instances in the putative alien gene set.

Furthermore, we were interested in quantifying only inter-domain gene transfers from prokaryotes to *G. sulphuraria*. These putative alien genes with sporadic distribution in rhodophytes constituted the second pass *G. sulphuraria* alien genes. One can, however, argue that the atypical pattern may also arise because of the loss of compositionally atypical ancestral genes. To rule this out, we performed sequence comparison using BLASTp. If there are indeed bacterial or archaeal genes in the second pass alien gene-set, these should first hit the rhodophytes (where they are present) and the Bacteria or Archaea (i.e., the prokaryotic donor) in the BLAST search. Those satisfying this criterion constituted the final pass of putative alien genes, with donors identified among prokaryotes.

4.4 Results and Conclusions

Because the acquired DNAs from different lineages have compositional biases (e.g., short-range nucleotide ordering) that are distinct from each other and from those of the native genome, these could be identified by virtue of their compositional similarity within, but dissimilarity between, in a given genome. The segmentation and clustering approach, based on this premise, was therefore expected to unravel the fragmented evolution of *G. sulphuraria*, with segment (or cluster of similar segments) representing a distinct evolutionary trajectory. Application of the integrated segmentation and clustering

method to the *G. sulphuraria* genome indeed revealed a fragmented or divergent evolutionary histories underlying the genome. The segmental landscape of the *G. sulphuraria* genome revealed three distinct “alien” clusters harboring ~ 5%, 10% and 15% of atypical genes respectively against the genome backbone comprised of ~70% ancestral or native genes (Fig. 4.1).

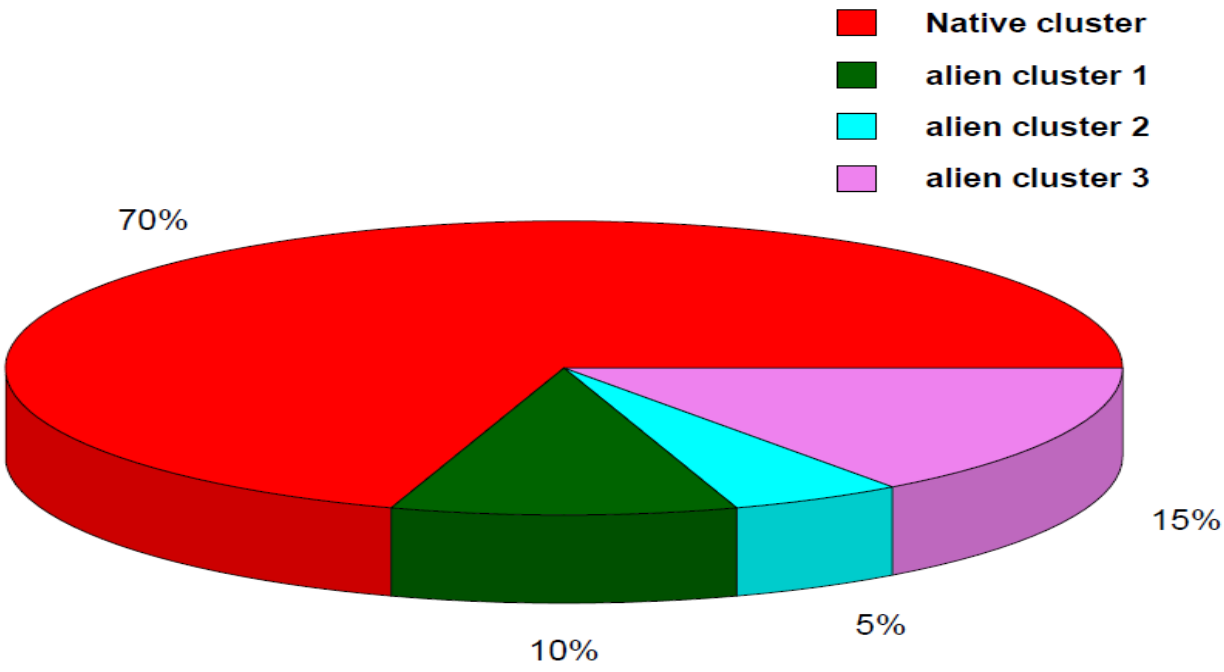

Figure 4.1. Distribution of *G. sulphuraria* genome segments into native cluster (red) and alien clusters (other colors), following application of Markov model of segmentation and clustering.

Sequence comparison using BLAST provided further support for our prediction (Fig. 4.2-4.4). Based on the BLAST support, we grouped the predicted alien genes into "most conservative estimate" (none of the top 5 BLAST hits, with redundancies removed, are algal hits, which includes archaeal ATPases and orphan genes),

"conservative estimate" (genes from “most conservative estimate” as well as the predicted alien genes with one algal hit out of the top 5 hits), and less conservative estimate (“conservative estimate” as well as the predicted alien genes with two algal hits out of the top 5 hits), for each cluster (Table 4.2). A significant fraction of the predicted alien genes, particularly the genes in the third cluster (Fig. 4.4), had no homologs in the database. These “orphans” genes are difficult to analyze using phylogeny based methods. In contrast, our study adds another dimension of information- most of these genes have anomalous compositional characteristics, indicating their likely origin in genomic contexts different from the recipient genome. We posit that these are likely horizontally acquired genes from organisms whose genomes are yet not sequenced. Furthermore, the predicted atypical genes that have algal hits show a distribution that can be more parsimoniously explained as a consequence of HGT than by other alternative scenarios including gene loss. This is substantiated by the fact that over 75% of the predicted alien genes have no algal hit or only one algal hit out of the top 5 BLAST hits (“conservative estimate” as defined above). This proportion reaches over 85% when cases with two algal hits out of the top 5 BLAST hits are also included, but this less conservative estimate require further study to rule out alternative scenarios in favor of HGT. Overall, the BLAST results show a good correlation of atypical phyletic patterns with atypical compositional patterns.

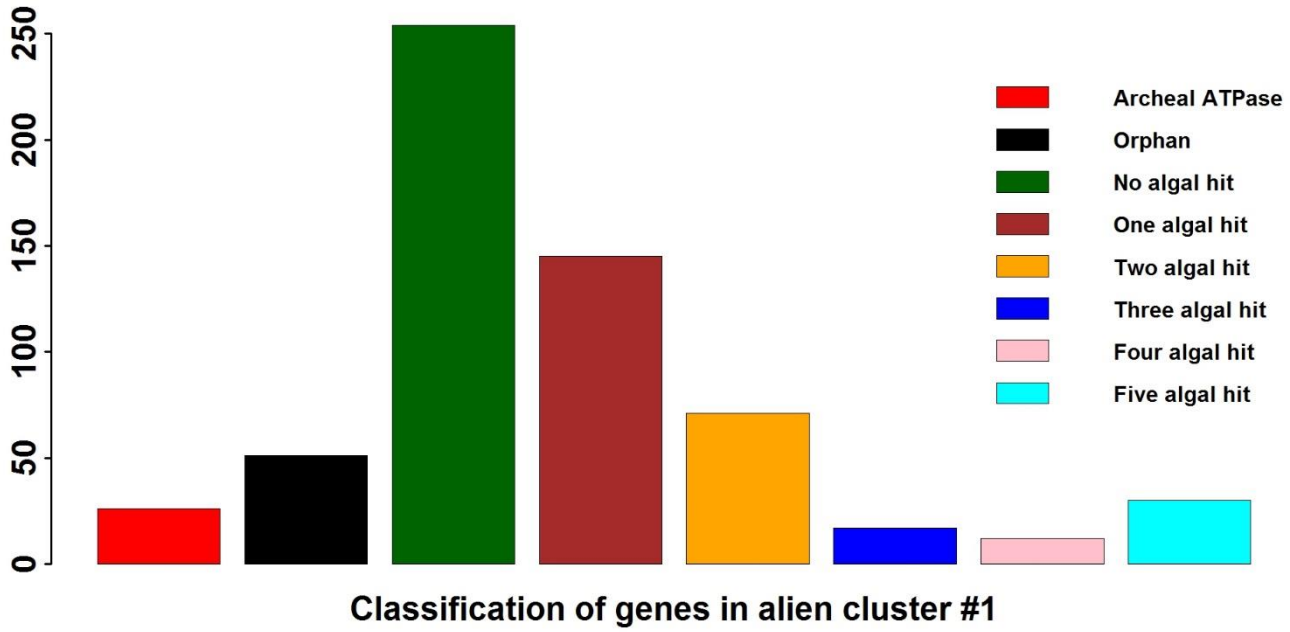

Figure 4.2. Distribution of *G. sulphuraria* genes in alien cluster#1: genes in cluster 1 were analyzed through Blastp program and classified based on top 5 hits obtained in blastp program.

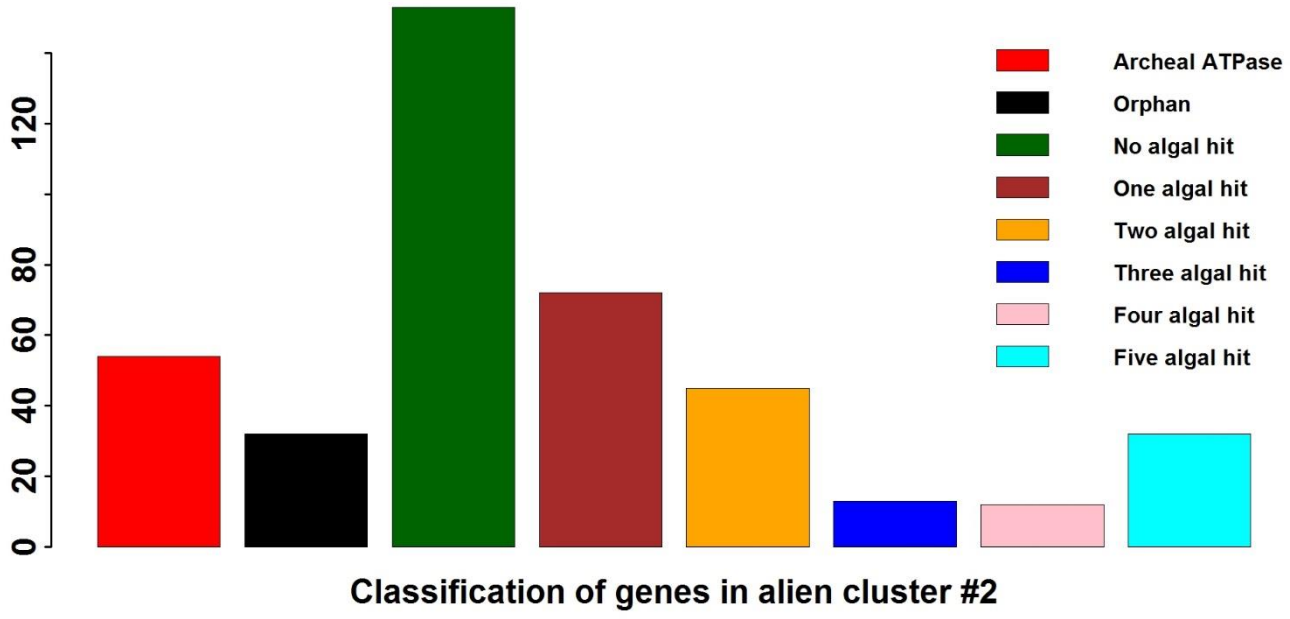

Figure 4.3. Distribution of *G. sulphuraria* genes in alien cluster#2: genes in cluster 2 were analyzed through Blastp program and classified based on top 5 hits obtained in blastp program.

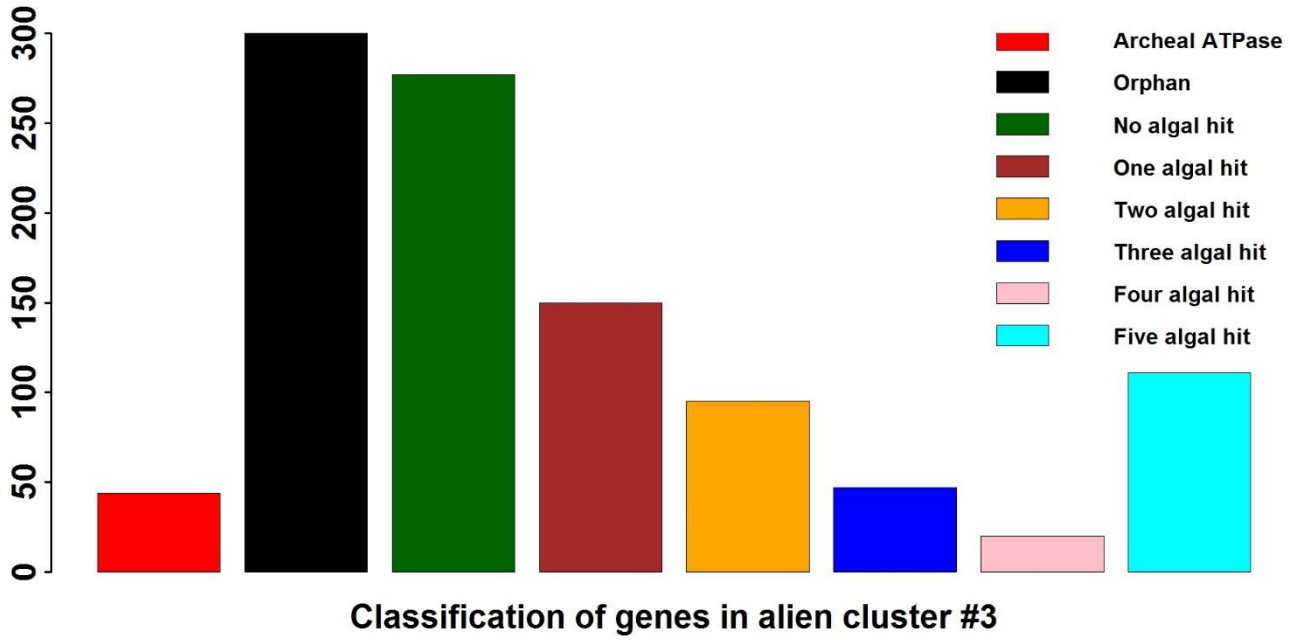

Figure 4.4. Distribution of *G. sulphuraria* genes in alien cluster#3: genes in cluster 3 were analyzed through Blastp program and classified based on top 5 hits obtained in blastp program.

Table 4.2. Most conservative estimate: genes having all non-algal, orphan and archeal ATPase hit; Conservative estimate: genes having all non-algal, orphan, archaeal ATPase and only one algal hit out of top 5 hits; Less conservative estimate: Conservative estimate and with two algal BLAST hit out of top 5 hits.

	Alien Cluster #1	Alien Cluster #2	Alien Cluster # 3
Total Number of genes	603	420	1012
Most Conservative Estimate	54.6%	63.5%	60 %
Conservative Estimate	78.5%	81%	74%
Least Conservative Estimate	90.2%	93.9%	82.5%

The genes of atypical composition constituted the first pass putative alien genes, which were subjected to phyletic pattern analysis, with their presence or absence examined in 13 other red algal genomic datasets (i.e., draft genomes or assembled

transcriptomes (see Table 4.1).

Of the 2,035 first pass alien genes, 458 genes were present in 8 or more of the rhodophytes (BLASTp query coverage >70%, Identity >30%, E-value < 1E-5) and were, therefore, eliminated from our list of alien genes. These second pass alien genes were subjected to sequence alignment against the non-redundant database to identify genes that had rhodophytes (where they are present) and prokaryotes (the potential donors) ranking higher than the other taxa among the BLAST hits. These 114 genes constituted the final pass alien genes of prokaryotic origin, of which 69 were not reported by Schönknecht et al. (2013; see Table 4.3, Supplementary excel file 1). 11 of the 114 genes were found to be plastid genes and therefore, these were not subjected to phylogenetic analysis as described below.

Maximum likelihood (ML) phylogenetic analysis was performed to validate these results (this analysis was performed by our collaborators Dr. Debashish Bhattacharya and Dr. Huan Qiu of Rutgers University). Here, each of the 69 novel *G. sulphuraria* HGT candidates was used to query a comprehensive local database comprising NCBI RefSeq version 58 and the rich collection of red algal sequence data discussed above using BLASTp (e-value cutoff = 1e-5). These sequences were retrieved with up to 10 sequences from each phylum.

Table 4.3. Summary of results from BLAST of *G. sulphuraria* genes of atypical composition against non-redundant database. Protein-products of *G. sulphuraria* atypical genes present in 6 or fewer of 13 rhodophytes (excluding *G. sulphuraria*) were blasted against the non-redundant protein database to identify those with best hits to prokaryotes in addition to rhodophytes where they are present [123].

Presence in other rhodophyte genomes (of total 13)	Number of genes	Number of genes with BLAST hits to only rhodophytes	Number of genes with top BLAST hits only to rhodophytes (where present) and prokaryotes	Previously reported as alien genes of prokaryotic origin by Schonknecht et al. (2013)
0	466	409	14	0
1	660	481	48	26
2	133	27	12	2
3	79	10	8	3
4	86	3	14	1
5	75	3	10	1
6	78	1	8	1
Total	1577	934	114	34

Each *G. sulphuraria* query together with the putative homologs were aligned using MUSCLE (version 3.8.31) [124] under the default settings. Each alignment was trimmed using trimAl [125] in the automated mode (-automated). ML trees were built using IQtree (version 0.9.6) [126] under the best amino acid evolutionary model selected using -m TEST with branch support values estimated using 1,500 ultrafast bootstrap replicates (-bb 1500). The resulting trees were manually inspected and this led to the validation of 17 novel HGT candidates in *G. sulphuraria* (see Supplementary excel file 1). The remaining trees showed that 13 candidates were of eukaryotic origin (i.e., false positives), 16 trees had topologies that could not be easily interpreted, and the remaining 23 trees represented ancient HGTs from prokaryotic sources that were shared by 7 or less red algae but present in other photosynthetic lineages. It is unclear why our method identified these ancient HGT events but they are of high interest with

regard to early algal evolution. These regions that encode foreign genes may show atypical composition due to other factors that are not yet clear to us.

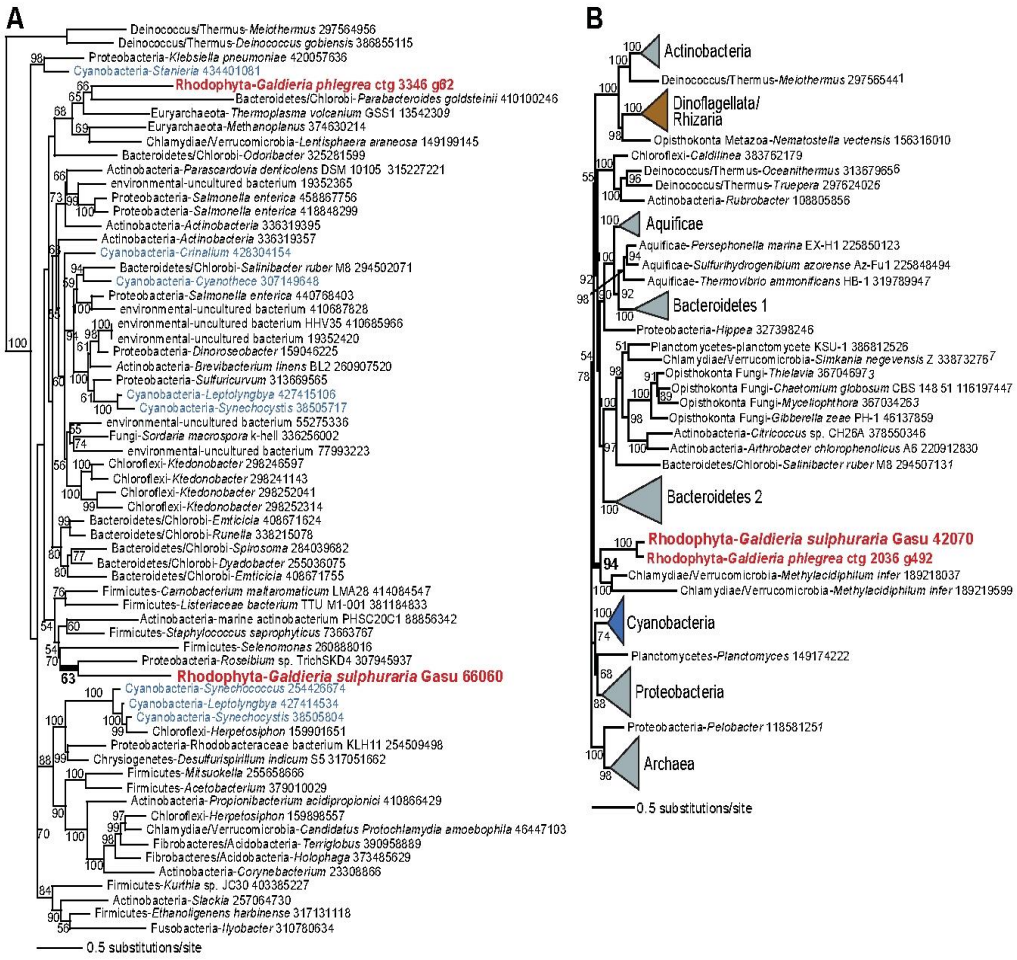

Figure 4.5. Maximum likelihood phylogenetic trees (built as described in the methods) showing two examples of HGT in *G. sulphuraria* from bacterial sources. (A) Tree of a resolvase gene belonging to the serine recombinase family. (B) Tree of pyruvate, phosphate dikinase that is involved in carbon metabolism. Red algae are shown in the red text, cyanobacteria in blue text, dinoflagellates/rhizarians with the brown rectangle, and all other prokaryotes and eukaryotes in black text. Monophyletic prokaryotic groups have been summarized with gray or blue triangles in panel B to save space. Bootstrap values from 1,500 ultrafast replicates are shown at the branches of both trees [123].

Examples of two strongly supported HGT candidates are shown in Figure 4.5. Interestingly, the 17 novel alien genes in the red alga *G. sulphuraria* encode a variety of functions that appear at first glance to be adaptive to a stressful, extreme environment. These include genes involved in DNA recombination (Gasu_66060; resolvase of the serine recombinase [SR] family [Fig. 4.5A]), DNA repair (Gasu_35170, shared with *G. phlegrea*; MutS involved in mismatch repair to combat DNA damage), and an ATP-binding cassette (ABC) sulphate transporter (Gasu_37370, shared with *G. phlegrea*). Other genes augment aspects of carbon metabolism (e.g., pyruvate, phosphate dikinase, PPDK; Fig. 4.5B). Interestingly, PPDK is widespread in bacteria where it catalyzes the reversible conversion of ATP and pyruvate to AMP and phosphoenolpyruvate (PEP). In plants, PPDK generates PEP from pyruvate in gluconeogenesis [127] and is also involved in the C4 photosynthetic pathway, under strict light regulation [128]. It should be noted that although many phylogenetic trees had apparently straightforward explanations, many others did not. This is due to the ancient origin of many prokaryotic genes, for example in the Archaeplastida ancestor, high divergence rate variation among putative homologs, and incomplete data sampling that make these results (and in other such studies; e.g., [129]) challenging to interpret.

4.5 Discussion and Conclusions

We highlighted here the importance of using complementary approaches for alien gene detection in microbial eukaryotes. Whereas the proposed pipeline deciphered dozens of novel alien genes of prokaryotic origin in *G. sulphuraria* eliciting atypical composition, atypical distribution among rhodophytes, and unusually high similarity in

distantly related prokaryotes, the number of alien genes of prokaryotic origin or otherwise in this extremophile could be much higher. This is reflected in the high numbers of *G. sulphuraria* genes of atypical composition that were unique to *G. sulphuraria* or were present in only one other red alga (466 and 660 genes respectively), most of which lacked homologs in other organisms in the database (409 and 481 genes respectively, Table 4.3). We anticipate many more alien genes of prokaryotic origin to be revealed in eukaryotes as the database grows further. The proposed pipeline can be used in concert with classical phylogenetic methods for comprehensively cataloging horizontally acquired genes in microbial eukaryotes. We strongly advocate a mixture of tree-dependent and tree-independent methods to address the extent and role of HGTs in algal evolution. Furthermore, it should be noted that the understanding of how nucleotide composition can be applied to putatively identify alien genes in compact algal (and eukaryotes in general) genomes such as *G. sulphuraria* is still in its infancy and this work is a start in that direction. These techniques would likely be more difficult to apply to large, repeat-rich eukaryotic genomes, but this idea remains to be tested in future work.

CHAPTER 5

APPLICATION III: METAGENOME PROFILING

5.1 Background

The emergence of high throughput sequencing platforms has led to the sequencing of hundreds of complete genomes of microorganisms (genome defined as complete set of genetic material represented by DNA sequence in an organism). Amongst the most recent developments is the sequencing of metagenomes—collections of incomplete, short genomic sequences obtained directly from environmental samples of microorganisms—by next generation sequencing machines [130]. Metagenomics involves collecting samples from an environment (e.g. water, soil, saliva, etc.) and then extracting, sequencing and studying the genetic material of the microorganisms present in these samples [131]. Many metagenomic studies are essentially community ecology studies, which seek to characterize communities statistically or dynamically in terms of composition, structure, abundance, demography, or succession, and sometimes with the consideration of other biotic or abiotic factors [130]. The metagenomic sequences are often in the form of unassembled reads whose average lengths (e.g. ~100 bp to ~1000 bp) in samples may vary by an order of magnitude depending on the sequencing technology used [130]. These unassembled collections of incomplete genomes from diverse microbial species co-inhabiting an environment present significant challenges to microbiologists. The foremost task is to identify the taxonomical group that each of these metagenomic sequence reads belongs to. The other important task is to identify all protein-coding genes in a metagenome. Microbial communities are ubiquitous, often referred to as

microbiomes, with implications in many health related problems. Unraveling microbiomes is thus of utmost importance in understanding the ecosystems, and also in addressing the health problems caused by microbial activities.

The basic strategy employed to classify these sequences is to compare the sequences of unknown origin to annotated sequences in public database (e.g. GenBank, NCBI, and Pfam). There are basically two approaches for taxonomic profiling of metagenomes: supervised learning and unsupervised learning. Supervised learning is attempt to classify sequences of unknown origin with previously labeled sequences, while unsupervised learning techniques use clustering approach to group similar sequences together, but these methods are useful if high-level characterization of the metagenome sample is sought (e.g. classification at the phylum rank).

The non-clustering techniques can be further subdivided into three main approaches that compare query sequences to database sequences in order to assign a taxonomic label: 1) Sequence similarity methods: These methods for taxonomical classification of short sequence reads have primarily relied on 16S ribosomal RNA sequencing (the highly conserved 16S rRNA sequences are often considered the most reliable marker for inferring the “Tree of Life”) [130-131]. Marker genes are ideally present in all organisms, and have a relatively high mutation rate that produces significant variation between species. However, there are limitations to these approaches; the marker genes may only account for a small percentage of a metagenomic sample. Other methods classify sequence reads based their similarity to known genomic sequences in the databases; the success of these methods depend on the breadth and depth of the sequence database. Most of these programs employ BLAST

(most commonly BLASTX), while some incorporate lowest common ancestor (LCA) algorithm which was first introduced by MEGAN [132]. Another class of methods aligns a query sequence against a database of reference sequences represented by profile hidden Markov models (pHMM) [133]. These alignment-based methods display good accuracy, even for short sequences, but suffer from two general shortcomings: a) since the reference databases are very large, it can take a long time to perform database search each query sequence; and b) these methods for the taxonomical profiling of metagenomic sequences rely on the microbial sequence database, which however represents a tiny fraction of microbes dwelling the earth, and if a query sequence does not have any close relative in the database, it could lead to false classification. 2) Sequence composition methods: In order to circumvent the limitations of marker gene based or alignment based methods, attempts have been made to systematically profile the metagenomic sequences, beginning with the species level towards higher taxonomic level until a “match” with the sequence of interest is found. Construction of signature models of sequenced microbial genomes underlies the whole metagenome sequence analyses and taxonomic profiling. These methods often make use of interpolated Markov models (IMMs) (e.g. Phymm) [134-135], naïve Bayesian classifiers (e.g. NBC) [136,137], k-means/k-nearest-neighbor algorithms (e.g. TACOA) [138], and support vector machine (e.g. PhyloPythiaS) [139] for classification. Extensive computing may be required in generating sequence models for various organisms, but once the models are built, these classification methods are generally faster than with alignment-based methods. Some of these methods choose to tradeoff accuracy with program runtime, e.g. CLARK [140] and RAIPhy [141]. Accuracy with these methods can be

improved by incorporating similarity search methods (e.g. BLAST), as implemented in PhymmBL [134-135], which achieves greater accuracy than the component methods. 3) Phylogenetic methods: These methods attempt to place a query sequence on a phylogenetic tree according to a model of evolution using maximum likelihood (ML), Bayesian methods and/or neighbor-joining (NJ) [130]. Most phylogenetic methods require the use of marker genes, as the first step in their workflows is to add the marker gene query sequence to a reference alignment (e.g. AMPHORA [142-143], Treephyler [144]). Phylogenetic methods are often computationally intensive as they utilize computationally intensive evolutionary models to achieve better accuracy. These methods are able to classify only a much smaller fraction of sample sequences in compared to similarity search and composition based methods.

Microbial dynamism, exemplified by chimeric genomes with DNAs of different ancestries or origins, belies the often invoked “one genome, one model” concept that underlie the current state-of-the-art in the field. We posit that these “static” models are inherently limited in exemplifying the microbial dynamism that shapes the genomes or metagenomes. Because microbial genomes are often chimeric, our hypothesis is that if metagenomic reads are assessed using models derived from regions of distinct lineages in genomes, rather than using single models of genomes that may ‘average out’ useful evolutionary signals for accurate classification or may not even represent any major evolutionary trends in genomes, a more robust classification of metagenomic reads could be achieved. We, therefore, propose a segmental genome model (SGM), wherein a genome is represented by an ensemble of signatures derived from segments of apparently different ancestries or origins. Within this framework, a genome is

subjected to a recursive segmentation procedure that generates segments of homogeneous composition, which are then grouped based on their compositional similarity using an agglomerative procedure, as described in Chapter 2. An ensemble of distinct signatures modeled on clusters on similar segments is obtained for each genome within a variable order Markov model framework. Metagenomic reads are queried against the database of ensembles of signature models of the prokaryotic genomes (Fig. 5.1). A probabilistic score for a read to be generated by a model is obtained for each model and a read is classified based on the model that yields the maximal score. Because only a small fraction of organisms in microbiomes is represented in the database, we performed metagenome profiling systematically from lower taxonomic levels (species or genus) to high levels (from family to phylum). Thus a read originating from an organism not represented at lower levels but at a higher taxonomic level in the database could be assigned a higher taxonomic identity with confidence.

Our proposed method SGM was first assessed on various synthetic test datasets of varying degree of complexities and variable read lengths, and then after optimization, was assessed on a well-studied acid mine drainage (AMD) metagenome dataset [145]. Our method significantly outperformed frequently used methods for metagenome classification, including Phymm [134-135], Naive Bayes Classifier (NBC) [136-137] and PhylopythiaS(+) [139]. Furthermore, a still better performance was achieved when our method was used in combination with BLAST, with classification accuracy significantly higher than either of the component methods. In what follows we present our results that validate our hypothesis that by incorporating segmental signature models within a

variable order Markov model framework for scoring metagenomics sequences, a microbiome is characterized more robustly.

To assess the gain in the power of discrimination of the SGM versus the whole genome model (WGM), we constructed a database of Markov chain models of different orders for the whole genomes along with the database for SGM as described above and in Section 5.3 below. Notably, our proposed method SGM works within a variable order Markov model framework allowing flexibility in selection of the order of a Markov chain model depending on the cluster size.

5.2 Materials

Phymm datasets and AMD data were downloaded from Phymm's website- http://www.cbcb.umd.edu/software/phymm. All bacterial and archaeal genome sequences were downloaded from the NCBI ftp server: ftp://ftp.ncbi.nih.gov/genomes/refseq/. Synthetic metagenome datasets were generated using Metasim software [146].

5.3 Approach

5.3.1 Segmental Genome Model

In a departure from a single signature model for prokaryotic genome, we propose and implement segmental genome model, an ensemble of models each derived from segments of apparently similar composition or shared ancestry. This is accomplished within the integrated framework of recursive genome segmentation and agglomerative genome segment clustering [37]. In contrast to the original version of this algorithm that

measures the composition of a segment or cluster in terms of Markov chain probabilistic model of order 2 regardless of the size of the segment or cluster [4, 42], here the composition is modeled within a variable order model framework with the order of a model now varying with the size of the segment or cluster. Within this framework, a model of higher order is preferred, provided the cluster size is at least four times the number of model parameters. The compositional difference between segments or clusters is thus computed as the difference between the models representing the segments or clusters, quantified using a generalized information-entropic based divergence measure [4, 42]. Segmental genome models were built for all completely sequenced prokaryotic genomes available in the NCBI genome database (for a total of 5127 chromosome and plasmid sequences representing 2753 strains, 1412 species, 681 genera of prokaryotes). For comparison, we also built whole genome model of order *m* (WGMm), *i.e. one* model of order *m* for the entire genome, for each organism. Whole genome models were constructed for different model orders, ranging from 0-8. Each model is comprised of marginal and transition probabilities that define a standard Markov chain model, with their values computed using the counts of nucleotides or oligonucleotides in the sequence used to build the model [147].

5.3.2 Taxonomic Classification

The database of SGMs was used to score metagenomic reads in the test datasets. The probabilistic score of a read *S* of length *N* given model *M* of order *m*, *P*(*S*|*M*), was computed as,

$$P(S \mid M) = p(\alpha_1 \alpha_2 \ldots \alpha_m) \prod_{i=m+1}^{N} p(\alpha_i \mid \alpha_{i-m} \alpha_{i-m+1} \ldots \alpha_{i-1}),$$

where α_i denotes nucleotide α at position *i* in *S*, $p(\alpha_1\alpha_2\cdots\alpha_m)$ is the probability of oligonucleotide $\alpha_1\alpha_2\cdots\alpha_m$ and $p(\alpha_i \mid \alpha_{i-m}\alpha_{i-m+1}\ldots\alpha_{i-1})$ is the transition probability of α_i given the preceding oligonucleotide $\alpha_{i-m}\alpha_{i-m+1}\cdots\alpha_{i-1}$ of length *m*.

The read *S* was inferred to have originated from a genome, represented by model *M* from the ensemble of signature models for this genome, if *P*(*S*|*M*) is maximum among all *M's* in the SGM database (Fig. 5.1). Taxonomic classification was performed similarly using WGMm, with *M* now representing the whole genome model.

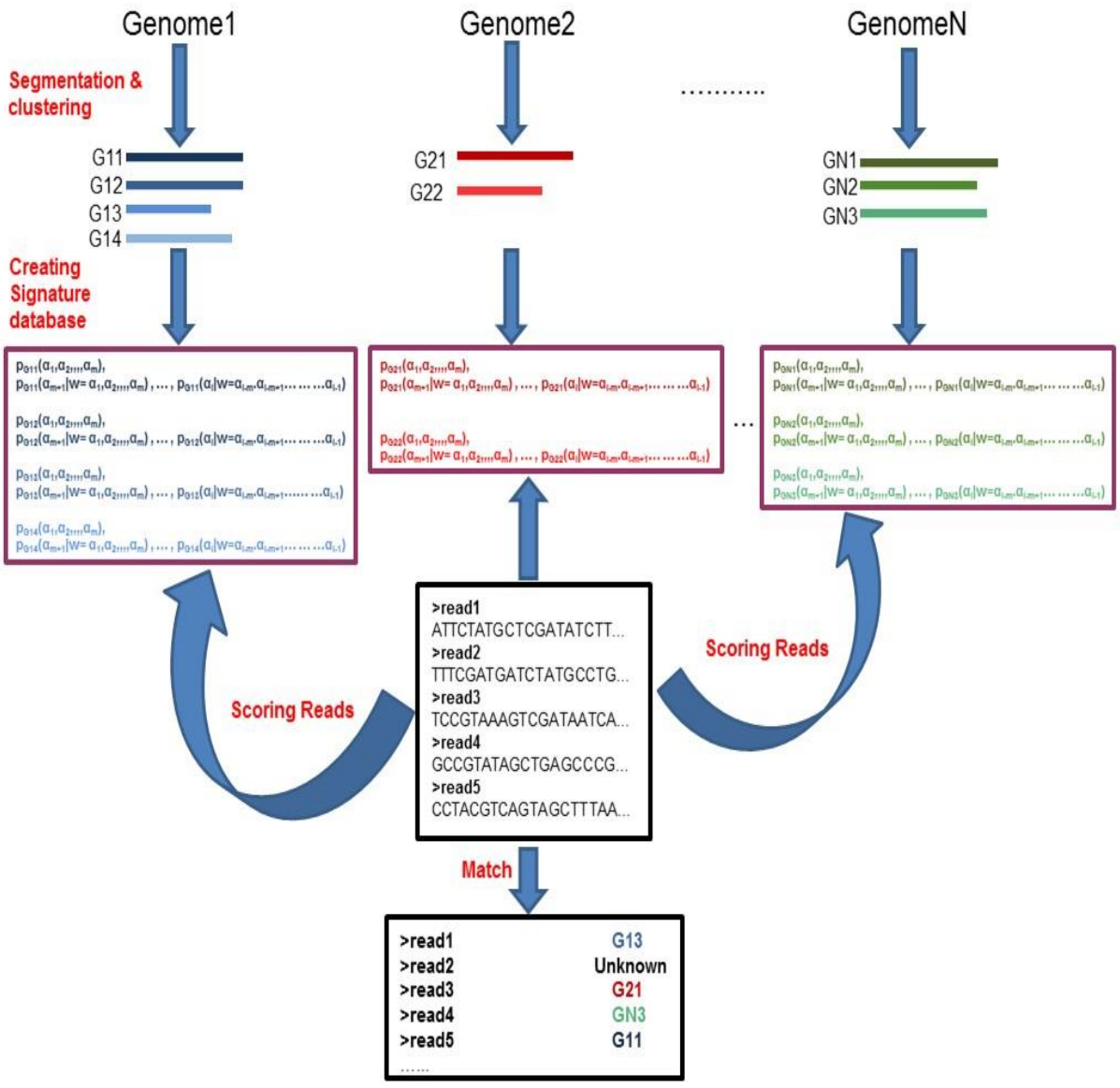

Figure 5.1. Proposed metagenome profiling pipeline based on segmented genome model.

5.3.3 BLAST Integration

This analysis was performed by David Burks, graduate student and a member of Dr. Azad's research group. Best blast hits for metagenomic reads were obtained using ncbi-blast+ 2.2.28 at the default setting [148]. In most instances, multiple top hits with

the same score and e-value representing different taxa were obtained; in these cases, the taxonomic identity was inferred based on the probabilistic scoring of reads by SGM (or WGM), i.e. the taxon M that yielded the highest probabilistic score P(S|M) for read S was inferred to be likely source of the read. Otherwise, BLAST was given the predictive priority, i.e. the read was assigned to a taxon based on the best blast hit. The taxonomic identity for a read with no blast hit was inferred based on the probabilistic scoring by SGM (or WGM). We also adapted the PhymmBL's integration scheme [134-135], where the e-value was first log-transformed and then combined with the probabilistic score using the formula: $PS+m(a-\log_x(\text{e-value}))$, PS is the probabilistic score, and m and a are the multiplicative and additive constants, respectively. The values of these parameters were determined through performance optimization on different test datasets, using a Python program that invoked a binary-stepping method [134-135].

5.4 Validation Using Synthetic Metagenomes

To assess the performance of SGM against WGM and other frequently used methods for whole metagenome analysis, we constructed four test datasets, three generated using exact (no error) model, sanger sequencing error model and 454 sequencing error model using Metasim [146] at the default setting with each containing sets of reads of length 100 bp, 250 bp, 500 bp and 1000 bp. The fourth test dataset with reads of lengths 100, 200, 400, 800 and 1000 bp was used in a previous study (Phymm dataset) [134]. Cross-validation was implemented by masking the test-sequence originating species in the database, and therefore, the performance was assessed at higher taxonomic levels to species, namely, genus, family, order, class, and phylum. We

selected genera with completely sequenced genomes for at least 3 species, one of these three was used for generating synthetic test dataset using Metasim and the other two were used to build models for the classification methods. The test datasets contained sequence reads from 505 chromosomes and plasmids representing 260 species, 62 genera, 54 families, 38 orders, 22 classes, and 13 phyla. We assessed the performance of SGM, WGM, and the current state-of-the-art in the field, namely NBC [136-137], CLARK [140], RAIPhy [141], TACAO [138], PhyloPhythiaS (+) [139], and Phymm [134-135]. In addition to assessment of different composition based methods, we evaluated BLAST augmented SGM, WGM and Phymm (SGMBL, WGMBL and PhymmBL [134-135] respectively).

Notably the accuracies generated by all methods as well as the SGM's margin of improvement over all other methods vary across all four test datasets used in this study. Genus level accuracy generated by SGM varied from ~15% - 57% for different read lengths for exact model datasets, which is an improvement of ~2% over Phymm and ~3 - 4% over NBC (top panel in Fig. 5.2). For test datasets generated using sanger error model (read sequences generated using sanger sequencing approach), SGM yields accuracy ranging from 14 - 55% at genus-level, which is an improvement of ~2 - 4% over Phymm & ~4 - 7% over NBC (center panel, Fig. 5.2), and for test datasets generated using 454 sequencing error model (read sequences generated by the pyrosequencing approach (Roche's 454)), SGM produced accuracy between ~13 - 51% at genus level and the improvement in accuracy over Phymm is ~ 3 - 9% and over NBC is ~4 - 9% (bottom panel, Fig. 5.2). Even for the Phymm test datasets, which were constructed by random sampling from database genomes [134], SGM has an accuracy

of ~25 - 63% at genus-level, outperforming Phymm and NBC by ~3% and ~10 -12% respectively (Supplementary Fig. 12). As expected, the accuracy increases for higher taxonomic level classifications (family and above), reaching ~90% for SGM, ~82% for Phymm and ~84% for NBC at the phylum level (Fig. 5.3, Supplementary Figs. 9-12).

We also compared SGM with alignment based comparison method BLAST; SGM outperformed BLAST_T (where, the top BLAST hit with lowest e-value was considered) by up to ~35% in accuracy for reads of length 250 bp and greater (except at genus and family level for 250 bp in 454 error model dataset), and similarly, outperformed BLAST_S (a single lowest e-value BLAST hit is obtained and therefore considered as the best match) by up to ~44% (Supplementary excel file 2). BLAST_T outperformed SGM on 100 bp read datasets at lower taxonomic levels (order and below) by up to ~10% in accuracy for the exact and error model datasets, and by ~5% at genus level for the Phymm dataset; in contrast, SGM outperformed BLAST_T on 100 bp reads by up to ~29% at class and phylum levels (Supplementary excel file 2).

Clark-l and RAIPhy, relatively recent programs, are faster to train and classify but performed much worse. PhylophythiaS (+) can’t handle reads smaller than 1000 bp; even at this resolution, the accuracy could reach only 57% at phylum level while SGM’s was 90% for the exact model, and 52% and 50% for the Sanger and 454 error model test sets respectively (88% and 86% respectively for SGM, Fig. 5.3). TACOA performs classifications at higher taxonomic ranks (order to phylum), it had the best accuracy of 17% on Phymm dataset (for 1000 bp reads) compared to 89.7% by SGM on this dataset (Supplementary Fig. 12). Among WGMs of different orders, WGM8 and WGM7 attained the best accuracies, and as hypothesized, they were indeed outperformed by

SGMs on all test datasets, except 1000 bp reads from exact model where both performed equivalently (Supplementary excel file 2); the most pronounced accuracy improvements were observed with shorter reads, often at lower taxonomic levels, for example, a 9% family-level accuracy improvement on 200 bp Phymm dataset reads, 5% family-level accuracy improvement on 250 bp exact model dataset reads, 5% phylum-level accuracy improvement on 100 bp Sanger error model dataset reads, and 6% class-level accuracy improvement on 100 bp 454 error model dataset reads (Supplementary excel file 2). Also, WGM7 outperformed WGM8 for reads longer than 250bp. Surprisingly, WGM often outperformed Phymm on exact and error model datasets with reads of 250 bp or greater, with WGM's accuracy as higher as 10% over Phymm (Supplementary excel file 2). The difference between Phymm and WGM lies in the difference in the underlying model structures; Phymm with interpolated Markov model was expected to outperform WGM with fixed order model, however, this was not observed with our tests on Metasim datasets.

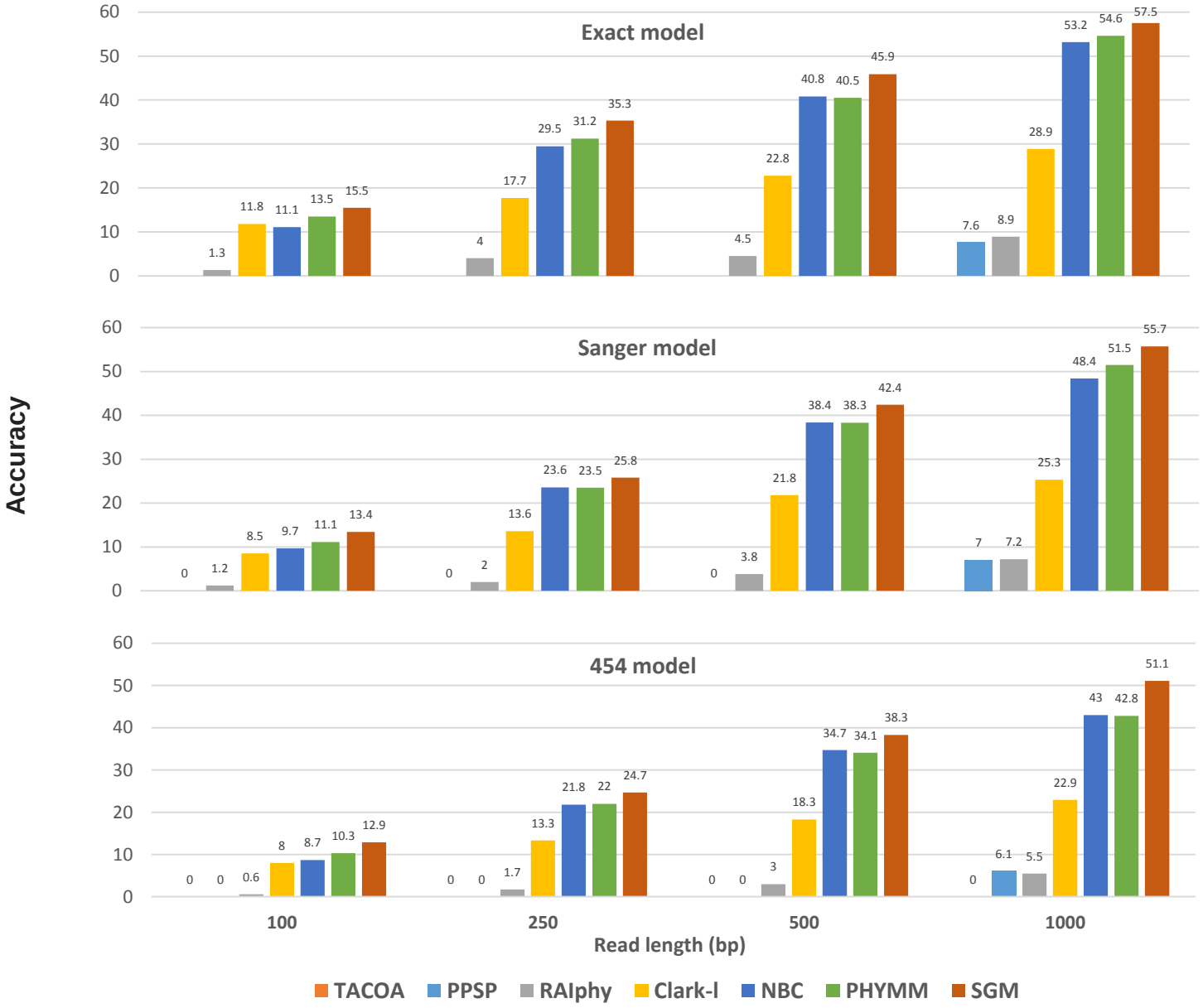

Figure 5.2. Accuracy in classifying reads generated using Metasim with exact (no error) model, Sanger and 454 error model. Performance of whole metagenome profiling methods, shown in different colors, was assessed as a function of sequence read length and genus level via cross-validation by masking the test-sequence originating species in the database. Accuracy was obtained as the percentage of reads classified correctly by a method. Accuracies were obtained at the default setting of the existing methods, and at a variable order model setting for SGM.

Because alignment based and alignment-free compositional approaches have complementary strengths, we developed a hybrid classifier by integrating BLAST with SGM (Section 5.3). As we used two different approaches to integrate BLAST with SGM, we assessed accuracy improvement by two variants of the hybrid of SGM and BLAST: SGMBL_F (using PhymmBL's formula integration scheme, [134]), and SGMBL_B (SGM gets precedence in classification only if multiple lowest e-value BLAST hits are obtained). SGMBL_B has 1-2% higher accuracy in comparison to SGMBL_F (Supplementary excel file 2). SGMBL_B outperforms the component methods SGM by 7-16% and BLAST by 8-12% in accuracy for all read lengths (genus and phylum level accuracies are shown in Fig. 5.4 and Fig. 5.5 respectively; see also Supplementary Figs. 13-16. As expected, SGMBL outperformed PhymmBL [134], a hybrid of Phymm and BLAST, by up to 10.5% in accuracy, (Fig. 5.4; Fig. 5.5; Supplementary Figs. 13-16). Overall, SGMBL outperformed all methods, including the hybrid methods.

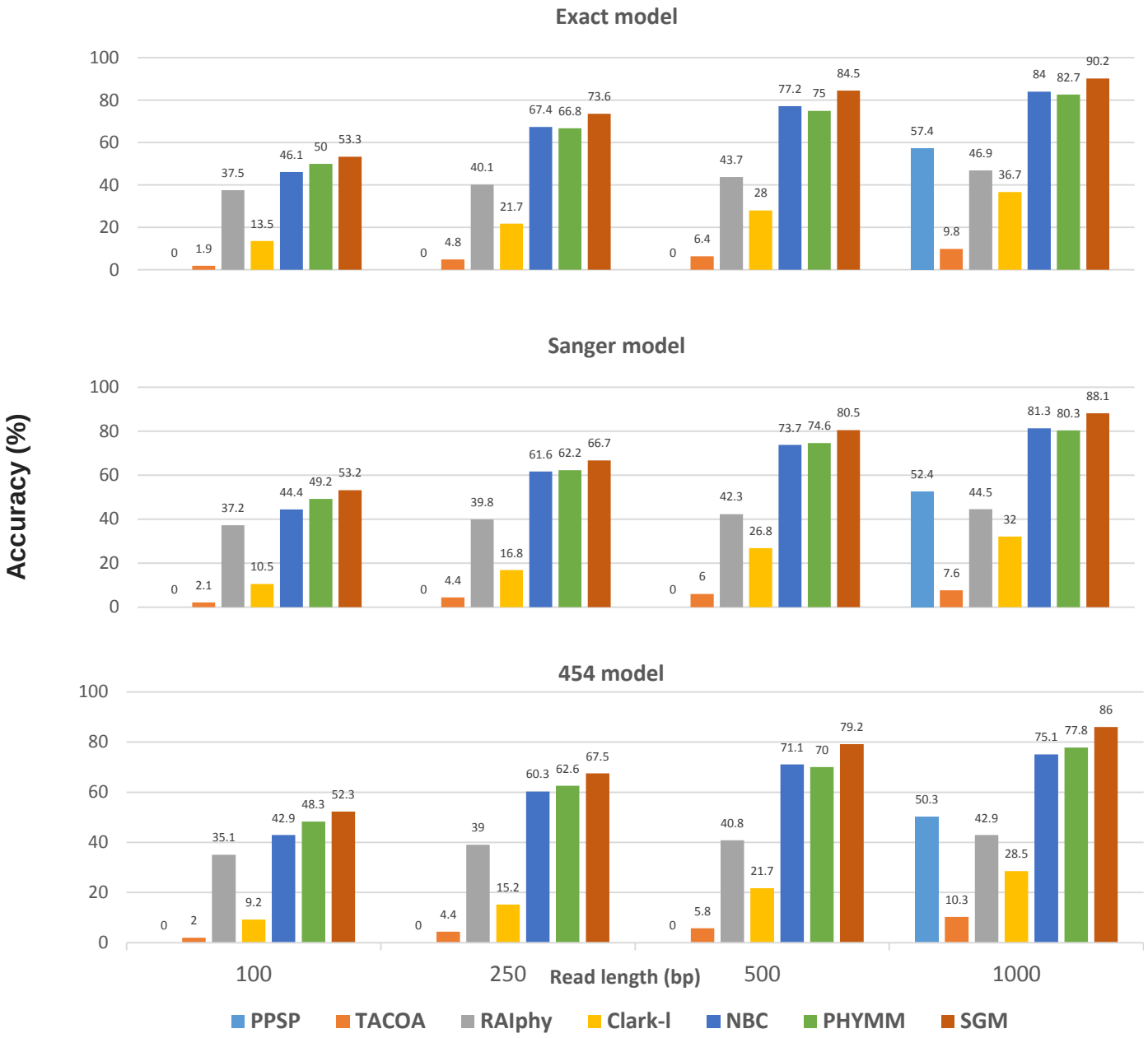

Figure 5.3. Same as in Figure 5.2, but at phylum-level.

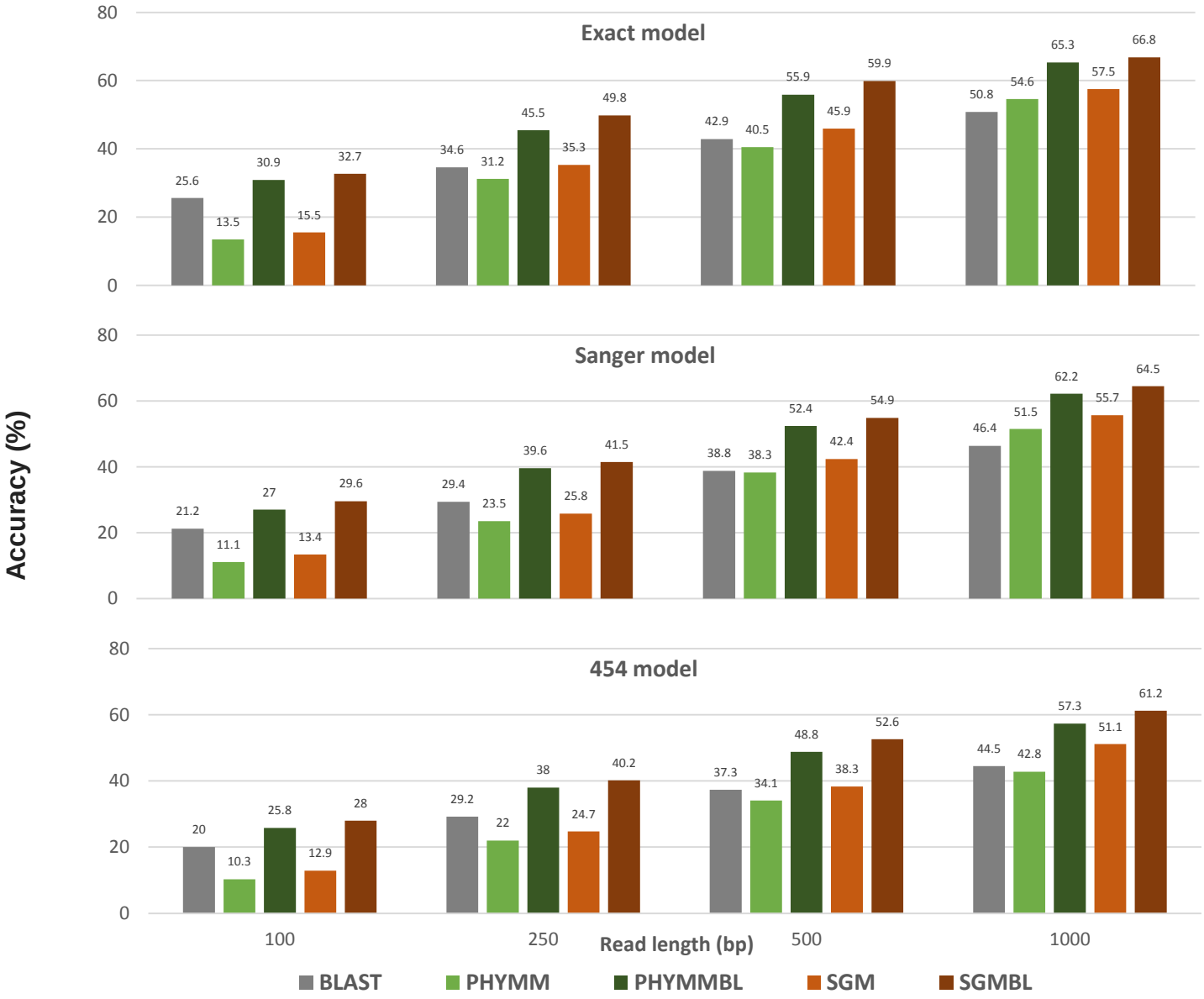

Figure 5.4. Same as in Figure 5.2, but for hybrids for Phymm and BLAST (PhymmBL), and SGM and BLAST (SGMBL). Accuracies for component methods are also shown.

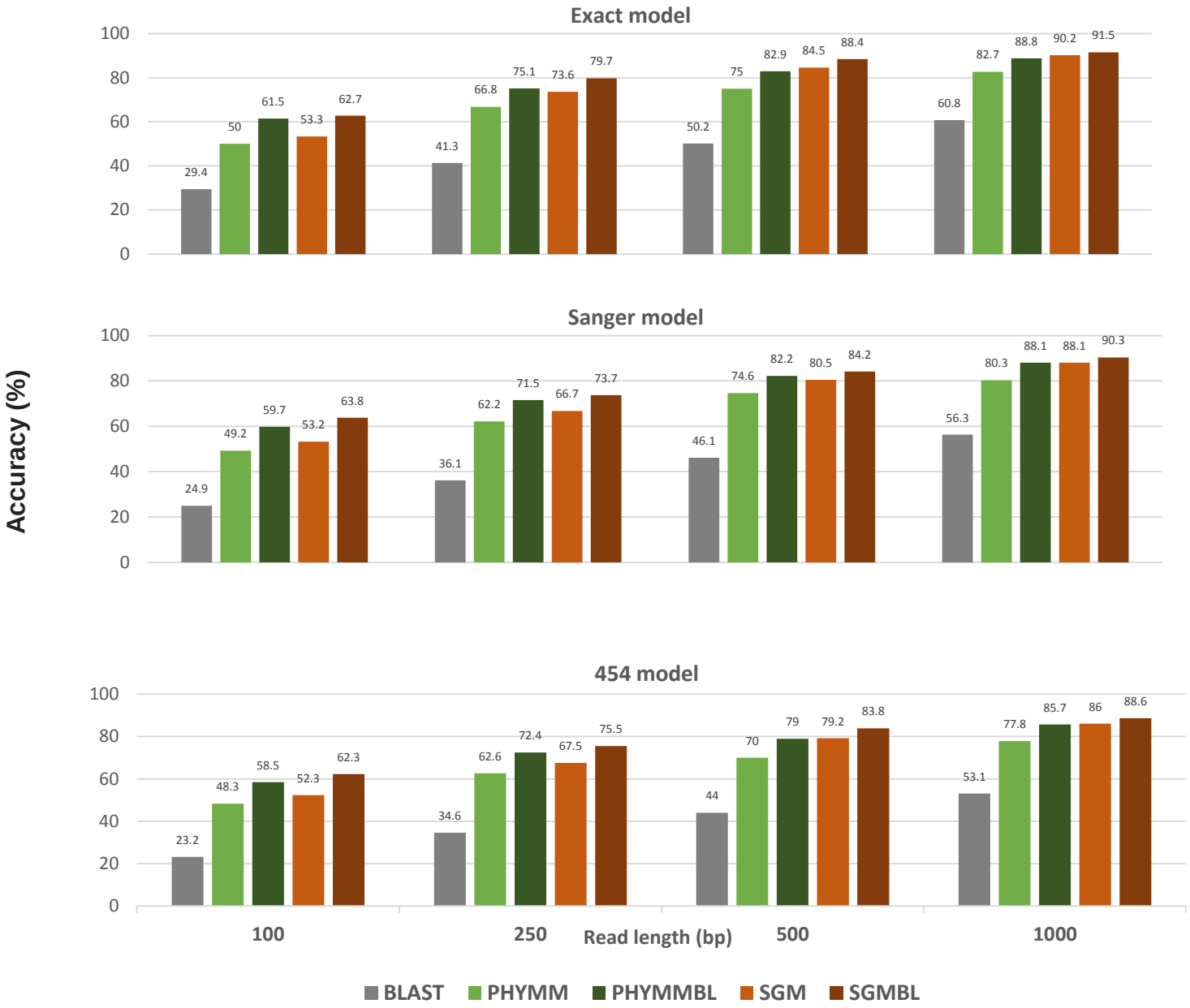

Figure 5.5. Same as in Figure 5.2, but for hybrids for Phymm and BLAST (PhymmBL), and SGM and BLAST (SGMBL) at phylum level. Accuracies for component methods are also shown.

5.5 Acid Mine Drainage Metagenome Analysis

Validation on real metagenomes is challenging as the true composition of the data cannot be determined with certainty. However, after assessment of SGM and SGMBL on synthetic metagenomes, we applied our methods to a well-studied acid mine drainage (AMD) metagenome that was reported to contain three dominant populations—the archaeon *Ferroplasma acidarmanus* and two groups of bacteria *Leptospirillum groups* II and III. *Ferroplasma acidarmanus* is at a relatively low abundance in comparison to *Leptospirillum* groups [145]. *Leptospirillum* groups II and III belong to phylum Nitrospirae , while *Ferroplasma acidarmanus* belongs to phylum Euryarchaeota. SGM and SGMBL assigned the 166,345 quality-filtered AMD reads to three major phyla Nitrospirae, Euryarchaeota and Proteobacteria (Figs. 5.4). SGM assigned more reads to Proteobacteria, which may be because some Nitrospirae were provisionally assigned to Proteobacteria [134].

With BLAST incorporated, more reads were inferred to have originated from Nitrospirae and Euryarchaeota by SGMBL (Fig. 5.4). The species-level classification assigned the AMD reads to *Leptospirillum*, *F. acidarmanus* and some proteobacterial species (Fig. 5.5), in agreement with the previous studies [145].

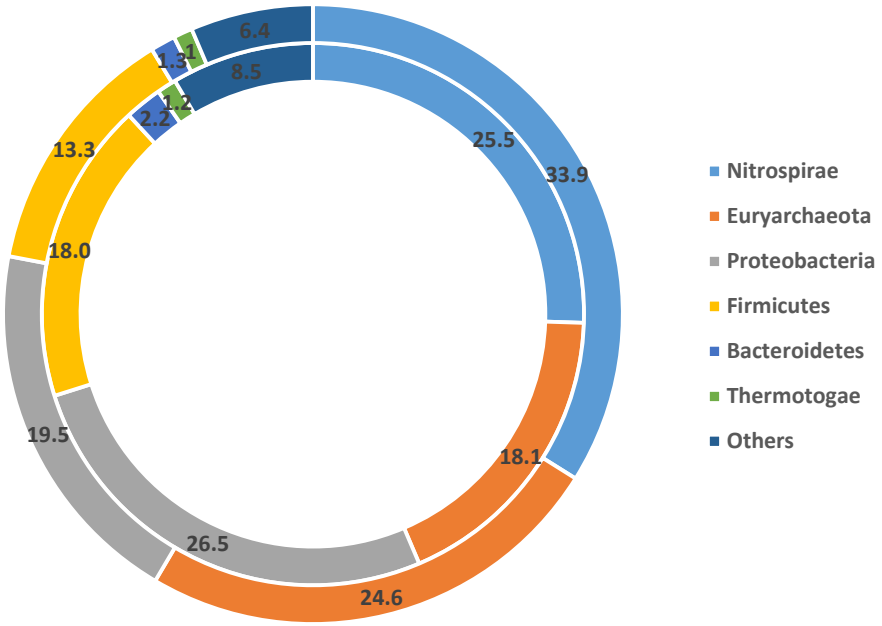

Figure 5.6. Phylum-level characterization of AMD data using segmental genome model with and without BLAST.

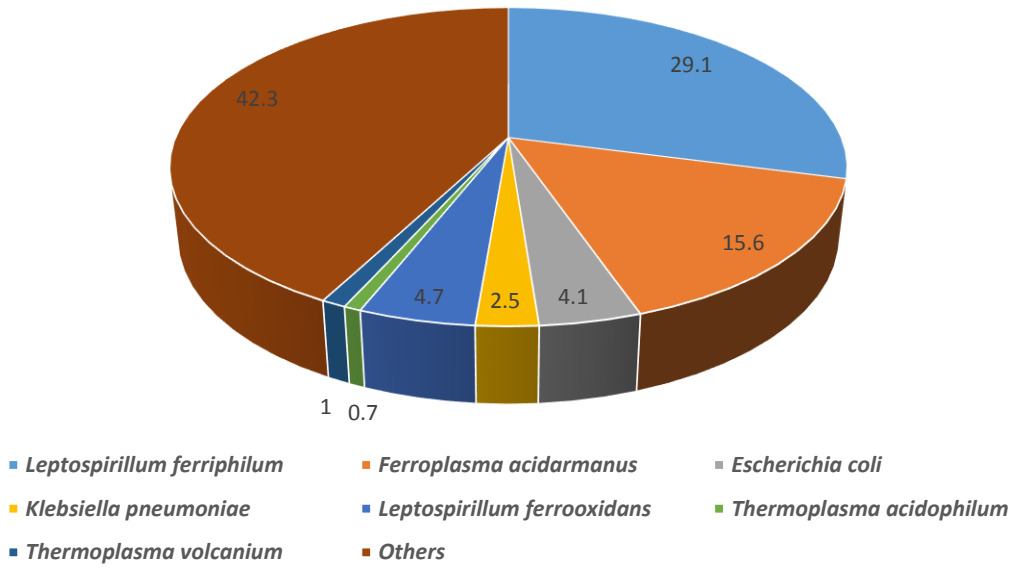

Figure 5.7. Species-level characterization of AMD data using SGMBL

5.6 Discussions

Our results demonstrate that the integrated segmentation and clustering method that generates compositionally distinct clusters to build multiple genome models provides a robust framework for taxonomic profiling of metagenomics reads. Our method SGM outperforms the frequently used Phymm, NBC, PhyloPythiaS(+) and other single genome models on various synthetic metagenome datasets. Cross-validation was performed by measuring the accuracy for species-masked test datasets by excluding reads originating from species that were used to build the models in first place.

Our method classifies reads at genus and higher taxonomic ranks, unlike PhyloPythiaS(+) or TACOA, which classifies the reads at only phylum or from order to phylum ranks. CLARK and Kraken require significant amount of memory (160-196 GB), and therefore may not be the methods of choice for researchers with low-memory computing resources. The light versions Clark-l and Minikraken can be run on low-memory computing environment, but at the cost of reduced standard database. Minikraken comes with a pre-built database, so we could not include this method for this study, as we need to create custom database for species-masked analysis, although Clark-l does allow us to create a custom database. Clark-l and RAIphy were the fastest programs to train and classify the sequences, but this was at the expense of accuracy.

Our method SGM showed more pronounced improvement in accuracy when the test datasets were more complex, namely the error model datasets with reads sampled with simulated errors, such as, substitution, insertion, deletion. This was to obtain read sequences modeled after the error-prone sequencing technologies. These data mimic

more closely the real metagenome data in the databases.

Sequence similarity based methods that often utilize BLAST for alignment and taxonomic labeling of sequences of unknown origin were previously reported to be more accurate than the composition based methods [119, 120], however, the accuracy decreases substantially if the query sequences are not represented in the database, which is indeed reflected in our cross-validation study; our method outperformed BLAST on species-masked database. BLAST, however, outperformed all composition based methods in classifying 100 bp reads.

The hybrid method SGMBL outperformed the component methods SGM and BLAST on all test datasets, highlighting the complementary strengths of methods. SGMBL also outperformed PhymmBL that combines BLAST with Phymm, as well as WGMBLs (hybrid of WGM7 and WGM8 combined with BLAST). Results from acid-mine drainage data are consistent with the previous studies [148]. Together, these results establish SGMBL as a powerful method for whole metagenome analysis.

CHAPTER 6

SUMMARY, DISCUSSION AND FUTURE SCOPE

6.1 Summary and Discussion

In this study, we adapted a genome mining method based on recursive segmentation and clustering and assessed its power in addressing several biological problems. Our analysis led to the development of a new program for the unbiased detection of evolutionary strata on sex chromosomes and fine scale delineation of their boundaries in human, plants and fungal-mating type chromosomes. Among its first applications in deciphering whole eukaryotic genomes, we assessed it ability to identify genomic islands in algal genomes. The segmental landscape generated by this method provided novel insights into the evolution of an extremophilic eukaryote. In contrast to frequently used classification methods that require specifying a priori the number of segments or the number of sources (sequence types) or both, this method generates the number of segments and their clusters corresponding to the inherent genomic heterogeneity.

Another interesting application is in the field of metagenomics. Because microbial genomes are often chimeric, we posited that these genomes cannot be characterized by a single model. Because of the inability of the current methods for metagenome profiling to exemplify the microbial dynamism that shapes the genomes, we proposed the segmental genome model that uses a probabilistic framework to model segments of apparently different ancestries in microbes that are thus represented as ensembles of compositional signatures. By incorporating segmental signature models, we achieved a more robust metagenome profiling.

The integrated segmentation and clustering method has a broad applicability in genome analysis [37]; the top-down hierarchical approach makes it possible to examine the sequence heterogeneity at different levels of hierarchy and correlate the segments with known biological features. In the studies of sex or mating-type chromosomes, we observed that at a certain hierarchical level, the segmental landscape in a sex or mating-type chromosome corresponds to the stratification structure resulting from recombination suppression. In the application to the human X chromosome [42], the X-conserved region (XCR) and X-added region (XAR) were recovered at a higher hierarchical level, while the stratum structure was recovered at a lower level as expected.

Although the segmentation and clustering algorithm for deciphering genomes and metagenomes works well in addressing many different problems, we discuss below its limitations and how this could be improved further. In our analysis of the sex chromosome stratification, we observed that the optimal parameter setting for delineating the stratum boundaries, mainly the segmentation and clustering thresholds, varies depending on lineages being examined or the sequence types- whole chromosome, genic or coding sequences. We determined the optimal setting based on the information available in the literature or based on the previously reported strata with strong support; this information allowed us to optimize our program by recapitulating the “knowns” and use the optimized program to find the yet unknown strata, particularly, in regions where gametologous sequence comparison is not possible due to the gene loss. No prior knowledge about the system or sequence could be a pitfall or limitation, in particular, for this analysis, however, we believe that this significant challenge can be

overcome by using this method in concert with other methods, e.g. synonymous substitution rate analysis, which can generate limited amount of “training data” from regions that are not yet substantially degenerated, which, in turn, can be used to optimize our prediction program. The parameter setting for the close relatives will likely be invariant and therefore, the program optimized for an organism can be reliably applied to its close relatives. Identifying small sex determining or non-recombining regions could be difficult, and was probably the reason for a single cluster observed for the chromosome 15 of *S. suchowensis*. Another limiting factor is the rigid model structure, e.g. using the same model order regarding of sequence length and composition. Therefore, future efforts could be directed towards making the underling model structure more flexible and sophisticated. First step in this direction could be to incorporate higher order Markov models in our algorithm, as the 2nd order Markov model proved to be more effective than 0th and 1st order Markov models in delineating stratum boundaries as well as in the identification of genome islands. Higher order Markov models provide greater predictive power and thus higher sensitivity, as we have also seen with the taxonomic profiling of metagenomes, where models of 7th and 8th order were found more effective. Therefore, we intend to generalize our algorithm to higher orders, up to the 8th order. Preliminary analyses along these directions are discussed in the next section. Currently, we recommend the users to use the default settings for different lineages or sequence types (provided with the source codes available at the journal websites or public repositories: https://github.com/pandeyravi15/SGMBL; https://doi.org/10.5281/zenodo.376782; [88]), however, the consensus setting will be refined as we obtain more sequence data in the

near future.

6.2 Augmentation of the Segmentation-Clustering Algorithm

We performed some “proof-of-concept” experiments to show that the segmentation-clustering algorithm could be augmented before its further applications in genome and metagenome analysis. And the model parameters, model order has a strong influence on the accuracy as observed in this study and also reported in earlier studies [149]. Finding an appropriate Markov model order for a given biological sequence is a non-trivial task. Questions such as- should model order be determined based on sequence size and/or based on the complexity of genomic sequences, are still open despite significant advances in this field. Complexity of genomic sequence could be a function of many different factors including evolutionary changes imparted by mutations, rearrangements, insertions, deletions, or acquisitions of foreign DNAs. It is also expected that DNAs from phylogenetically proximal organisms could be difficult to differentiate, while DNAs from phylogenetically distant species could be easier to differentiate even by utilizing lower order models. To this end, we evaluated different techniques and examined the effect of different order models in deciphering artificially constructed chimeric genome sequences with DNAs from phylogenetically proximal or distant donors, and of various lengths.

Chimeric sequences were generated by concatenating two sequences of equal or different lengths taken from two different species. The three phylum and corresponding species chosen for this study are as follows:

- Proteobacteria: Chimeric sequences were constructed by concatenating randomly sampled DNA sequence from *Escherichia coli* str. K-12 substr. MG1655

and sequence from a different proteobacterial species. These selected species were *Salmonella enterica* subsp. enterica serovar Typhi str. CT18, *Erwinia amylovora* ATCC 49946, *Yersinia pestis* A1122, *Vibrio cholerae* M66-2 chromosome I, *Haemophilus influenzae* 10810, and *Rickettsia typhi* str. B9991CWPP.

- Cyanobacteria: Chimeric sequences were constructed by concatenating randomly sampled DNA sequence from *Synechocystis* sp. PCC 6803 and sequence from a different cyanobacterial species. These selected species were, *Leptolyngbya* sp. PCC 7376 chromosome, *Microcystis aeruginosa* NIES-843, *Pleurocapsa* sp. PCC 7327, *Prochlorococcus marinus* str. MIT 9211, *Synechococcus elongatus* PCC 6301, and *Gloeobacter violaceus* PCC 7421.

- Actinobacteria: Chimeric sequences were constructed by concatenating randomly sampled DNA sequence from *Bifidobacterium longum* subsp. longum BBMN68 and sequence from a different actinobacterial species. These selected species were *Bifidobacterium adolescentis* ATCC 15703 DNA, *Gardnerella vaginalis* 409-05, *Mycobacterium abscessus* subsp. bolletii 50594, and *Conexibacter woesei* DSM 14684.

All genomic sequences were downloaded from NCBI ftp server: ftp://ftp.ncbi.nih.gov/genomes/refseq/.

For each phylum, one species was selected as a reference species: *Escherichia coli* str. K-12 for Proteobacteria, *Synechocystis* sp. for Cyanobacteria and *Bifidobacterium longum* for Actinobacteria, to make chimeric constructs by

concatenating sequence from the reference with the sequence from other selected species within same phylum. For each phylum, species are listed above in the order of their phylogenetic distance from the reference species (Supplementary excel file 3). Chimeric sequences of sizes 6kb, 12kb, 24kb, 36kb, 48kb and 96kb were constructed in five different ways:

- 1/4th reference sequence and 3/4th other species sequence from within the phylum.
- 1/3rd reference sequence and 2/3rd other species sequence from within the phylum.
- 1/2 reference sequence and 1/2 other species sequence from within the phylum.
- 2/3rd reference sequence and 1/3rd other species sequence from within the phylum.
- 3/4th reference sequence and 1/4th other species sequence from within the phylum.

Chimeric constructs allowed us to assess the power of the Markov model based segmentation method in discriminating genomic sequences from different sources as a function of model order, sequence length and phylogenetic divergence. Error was computed as the difference in the position of maximum divergence (from segmentation method) and the actual concatenation position in a chimeric construct. The mean errors (average over 1000 iterations) are presented in Supplementary excel file 3. This data provided insights into the model order appropriate for a given DNA sequence, and contrary to the expectation that higher orders should outperform lower orders for longer sequences, we observed that lower orders performed better in discriminating phylogenetically divergent sequence regardless of the sequence size.

We also assessed two model selection criteria, namely Akaike's information

criterion (AIC) and Bayesian Information Criterion (BIC) in selecting the “optimal” model order. This assessment was done on the datasets of chimeric constructs. These criteria are briefly explained below.

- Akaike's information criterion (AIC): The Akaike’s information criterion is defined as [150]:

$$\mathrm{AIC} = -2\log(\mathrm{L}) + 2\mathrm{K},$$

  where L is the maximum likelihood, K is the number of parameters in the model.

- Bayesian Information Criterion (BIC): The Bayesian Information Criterion is defined as:

$$\mathrm{BIC} = -2\log(\mathrm{L}) + \log(\mathrm{N})\mathrm{K},$$

  N is the length of sequence. BIC penalizes complex models more strongly than AIC. The smaller the BIC (or AIC), the better the model.

For our analysis, we measured the AIC and BIC for the two subsequence model, representing the two subsequences generated after segmentation of a sequence at the point of maximum divergence. The log likelihood for the two subsequence model for a given Markov model order *m* is obtained [151]:

$$\mathrm{Log}(\mathrm{L}) = \mathrm{N}_1\mathrm{E}_1 + \mathrm{N}_2\mathrm{E}_2,$$

where N_1 & N_2 are the lengths of the two subsequences and E_1 & E_2 are the respective entropies for the subsequences.

The number of free parameters K for two subsequence model for model order *m* is obtained as $2k^m(k-1)-1$, k being the alphabet size (k=4 for DNA sequences derived from alphabet (A,T,C,G)). Markov model order that results in smallest BIC or AIC the maximum number of times out of 100 trials (sequence constructs) was deemed optimal

in delineating the boundary of disparate segments in chimeric sequence constructs.

For proteobacterial group, higher model order (order 4-8, depending upon the size and composition of chimeras) yielded lower mean errors for chimeric DNA sequences assembled from phylogenetically proximal bacterial genomes (e.g. *E. coli* and *S. enterica* chimeric constructs). In contrast, usage of lower model orders (order 1-3) resulted in the lowest mean error for chimeric sequences assembled from phylogenetically distal genomes (e.g. order 1-3 for *E. coli* and *H. influenzae* chimeric constructs and order 0-3 for *E. coli* and *R. typhi* chimeric constructs)(Supplementary excel file 3).

In some instances, model order with lowest mean error increases with chimeric construct size. For *E. coli* and *S. enterica* chimeric constructs of 1/2 *E. coli* and 1/2 *S. enterica* composition, 5th order was found optimal for sequence length 6kb, and 8th order for 96kb, while 6th and 7th for intermediate lengths (12kb-48kb). For *E.coli* and *R. typhi*. chimeric constructs of 1/2 *E.coli* and 1/2 *R. typhi* composition, 0th order was found most appropriate for length 6kb, 2nd order for 12kb - 24kb, and 3rd order for 36kb - 96kb (Supplementary excel file 3). Note, however, that in the former case, 8th order was found optimal for 96 kb constructs but in the latter, it was only 3rd order that yielded the lowest mean error for 96 kb constructs. This suggests that model order cannot be chosen solely on the basis of sequence size, as has been done in several previous studies [149], but requires considering sequence composition as well.

Furthermore, the sizes of disparate components within a chimeric construct are also an important factor. Higher model orders performs better when both components are of similar size, while lower orders appear to be appropriate when one of the

components is substantially smaller than the other (Supplementary excel file 3). This was observed specifically for sequence constructs with components from phylogenetically proximal genomes.

Similarly, we carried out our analysis of sequences constructs obtained from species from two other phyla Actinobacteria and Cyanobacteria. We observed similar results for these phylum as well. In Actinobacteria, model orders between 4 - 8 yielded lowest mean error for chimeric constructs with disparate sequences from the same genus, Bifidobacterium (*B. longum* and *B. adolescentis*), while the model orders 1-3 were optimal for chimeric constructs with sequences from *B. longum* and *C. woesei*, representing different classes within Actinobacteria phylum. In Cyanobacteria, for sequence constructs with sequences from the *Synechocystis* and *Leptolyngbya*, which belong to same order but different families, model orders 3-6 yielded the lowest mean error (with a few exceptions). For *Synechocystis* and *G. violaceus* constructs, where G. violaceus is now phylogenetically more distant from *G. violaceus*, i.e. from different orders within the same class, model orders 1-2 were found optimal (Supplementary excel file 3). These results are consistent with those from other phylum studies, suggesting that the higher order Markov models are suitable for delineation of boundaries in sequence constructs with sequences from phylogenetically closer organisms, and vice-versa for phylogenetically distant organisms.

We wanted to investigate if the model order appropriate for a given chimeric DNA sequence could be determined using AIC or BIC. We did not observe any change in model order as a function of phylogenetic divergence between sequences in chimeric constructs. Model varied with construct length irrespective of the composition of

chimeric constructs. In Proteobacteria, for 6kb chimeric sequences, AIC selects model order 2 the maximum number of times irrespective of phylogenetic distance, while for chimeric sequences of length 12kb, AIC chooses model order 3 for *E. coli* and *S. enterica* constructs, but model order 2 for E.coli and distantly related *R. typhi* constructs. AIC chooses model order 3 for 24kb-48kb constructs, and model order 4 for 96kb constructs regardless of the phylogenetic distance. Similarly, BIC chooses model order 1 for 6kb-12kb; order 1-2 for 24kb, and order 2 for 36kb-96kb chimeric constructs (Supplementary excel file 4).

Further, in Actinobacteria and Cyanobacteria, we observed similar pattern with both AIC and BIC. For AIC, the model order varied from 2-4, and for BIC from 1-2, for smaller to larger chimeric constructs (Supplementary excel file 4). AIC and BIC do indicate that higher order Markov models are better choices for longer genomic sequences, but are not able to account for variable sequence composition that has a significant impact on model order used to deconstruct the chimeric sequences, as was observed with our error analysis above. Overall, the model order suggested by AIC or BIC does not result in lowest errors, and therefore, these criteria do not seem appropriate for chimeric genome analysis.

Together, this study suggests that the choice of Markov model order depends on both the sequence size and the composition. For longer constructs with sequences from closely related organisms, one can safely use higher order Markov models, while for shorter constructs with sequences from distantly related organisms, lower order Markov models could be more appropriate.

6.3 Future Directions

6.3.1 Development of an Evolutionary Platform for Assessing the Method

The integrated segmentation and clustering method predicted several novel strata on human X in addition to detection of known stratum boundaries. Validation of our predictions is not feasible due to the paucity of X-Y gametologous sequences. Future work could focus on developing a platform for assessing the method, as follows. Starting with an autosome with no strata, and assuming inversions on Y and subsequent recombination suppressions happening on X-Y regions, the autosomal sequence could be evolved into a stratified X using evolutionary models of nucleotide substitution (insertions and deletions) and mutations. Because regions of varying size on X would have to be evolved, from distal long arm to distal short arm, and by following time lapse in evolution in this order, strata on X mimicing the real strata would be created. Since the chromosome would be evolved using a known set of model parameter the boundaries and structure of strata would be known with certainty. This would serve as a valid test bed for evaluating different methods including the segmentation and clustering method.

6.3.2 Development of Integrative Framework for Stratum Detection

Similar to integration of parametric approach with comparative genomics for identification of foreign genes in G. sulphuraria, one can combine the complementary strengths of different stratum detection methods including substitution rate analysis [3, 11-12], inversion analysis [9, 10], phylogenetic methods [13, 14], and our segmentation and clustering algorithm [54]. A high confidence set of consensus predictions could thus

be obtained from regions with many X-Y homologous genes, otherwise, in the absence of X-Y sequences, the segmentation and clustering method should be relied upon.

6.3.3 Identification of Evolutionary Strata in Mammalian and Bird Sex Chromosomes

Evolutionary strata have been described in organisms across the kingdoms, from smut fungi, dioecious plants, birds, to mammals. However, in the absence of Y (or Z) chromosome or due to the severely degraded Y, a comprehensive comparative study is lacking. This study shows a way forward in this situation and allows detection and comparison of strata in different species. A comparative study could enhance our understanding of stratum evolution from fungus to human.

One may expect to identify similar evolutionary landscapes on the mammals which are closely related to human. Due to the shared evolutionary history, the older strata 1, 2 and 3 are thought to be shared between all eutherian mammals, while the oldest strata 1 and 2 are shared between eutherian mammals and marsupials. However, most recent stratum formation events should be unique to each mammalian species. This should enable the determination of the phylogenetic relationships of sex chromosomes from different mammals and thus understand their evolution.

6.3.4 Application in X-Chromosome Inactivation

Short motifs (sequences) unique to domains of genes which escape from inactivation or unique to domains that undergo inactivation have been previously observed. This indicates that the compositional biases in the two domains may differ, as also reported in a recent study [56], and therefore this could be exploited by our

segmentation and clustering approach to decipher the inactive domains on silenced X. We expect this method to delineate the boundaries between domains which contain genes that are subject to inactivation and domains that contain genes that escape from inactivation. Identification of these domains would may help to know the fate of new potential genes that will could be added in future, because if they are added within cluster of silenced or escapee genes, it they may be subject to the status of the region where they were added.

Future efforts could also be towards identification of the unique genomic signals present in domains that undergo inactivation but absent from domains which escape from inactivation and vice-versa using available bioinformatics tools. These signals might be playing a major role in regulating the inactivation signal. Distribution of such sequences in human autosome's and in X chromosome sequences of other closely related mammals could be examined, to find out whether they are specific to human X chromosome or not and their acquisition on the human X chromosome.

Disruption of X inactivation process causes many diseases like breast cancer, ovarian tumors, syndromes like Turner and Klinflinter, due to the aneuploidy of X chromosome and/or variable expression of different sex linked genes in different tissues. Functional analysis of genes that escape from inactivation, especially the genes which do not have Y homologs, could help in developing strategies for controlling the over-expression of the genes responsible for sex related diseases by blocking the genomic signals that inhibit inactivation. This should help into addressing many disease problems arising due to genes which escape from inactivation.

REFERENCES

1. Thakur V, Azad RK, Ramaswamy R (2007) Markov models of genome segmentation. Phys Rev E Stat Nonlin Soft Matter Phys 75: 011915.

2. Brown TA (2002) Genomes ‟BIOS Scientific Publishers, Oxford.

3. Graur D, Li WH (2000) Fundamentals of Molecular Evolution Sinauer, Sunderland.

4. Arvey AJ, Azad RK, Raval A, Lawrence JG (2009) Detection of genomic islands via segmental genome heterogeneity. Nucleic Acids Res 37: 5255-5266.

5. Guy-Franck R, Alix K, Bernard D (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev. 72(4):686.

6. Lobry JR (1995) Properties of a general model of DNA evolution under no-strand-bias conditions. J Mol Evol 40:326.

7. Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318.

8. Frank AC, Lobry JR (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238:65.

9. Oliver JL, Bernaola-Galvan P, Guerrero-Garcia J, Roman-Roldan R (1993) Entropic profiles of DNA sequences through chaos game derived images. J Theor Biol 160:457.

10. Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. Gene 241: 363.

11. Zoubak S, Clay O, Bernardi G (1996) The gene distribution of the human genome. Gene 174:95.

12. Duret L, Semo M, Piganeau G, Mouchiroud D, and Galtier N (1996) Statistical analysis of vertebrates sequences reveals that long genes are scarce in GC rich isochores. J Mol Evol 40:308.

13. Lander E, et al. (2001). Initial sequencing and analysis of human genome. Nature 409:860.

14. Smith ZE, Higgs DR (1999) The pattern of replication at a human telomeric region (16p13.3): its relationship to chromosome structure and gene expression. Hum Mol Genet 8:1373.

15. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196:261.

16. Carrel L, Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434:400-404.

17. Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structures and optimal growth temperatures in prokaryotes. J Mol Evol 44:632.

18. Karlin S, Campbell AM, Mrazek J (1988) Comparative DNA analysis across diverse genomes. Ann Rev Genet 32:185.

19. Churchill GA (1992) Hidden Markov chains and the analysis of genome structure. Comput Chem 16(2):107.

20. Nicolas P, et al. (2002) Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models, Nucleic Acid Res. 30(6):1418.

21. Azad K, Lawrence J, Thakur V, Ramaswamy R (2007) Advanced Computational Methods in Biocomputing and Bioimaging. (Nova Science Publishers, NY).

22. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379.

23. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. J Mol Biol 188:415-431.

24. Rogan PK, Schneider TD (1995) Using information content and base frequencies to distinguish mutations from genetic polymorphisms. Hum Mut 6:74-76.

25. Voigt CA, Mayo SL, Arnold FH, Wang Z (2001) Computational method to reduce the search space for directed protein evolution. Proc Natl Acad Sci 98:3778-3783.

26. Giraud BG, Lapedes A, Liu LC (1998) Analysis of correlations between sites of model protein sequences. Proc Natl Acad Sci 58:6312-6322.

27. Clarke ND (1995) Covariation of residues in the homeodomain sequence family. Prot Sci 4:2269-2278.

28. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AD (2000) Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. Mol Biol Evol 17:164-178.

29. Carothers JM, Oestreich SC, Davis JH, Szostak JW (2004) Informational complexity and functional activity. J Am Chem Society 126(16):5130.

30. Adami C (2002) Combinatorial drug design augmented by information theory. NASA Tech Briefs 26:52 .

31. Weiss O, Jimenez-Montano MA, Herzel H (2000) Information content of protein sequences. J Theor Biol 206:379-386.

32. Adami C, Ofria C, Collier TC (2000) Evolution of biological complexity. Proc Natl Acad Sci 97:4463-4468.

33. Adami C, Cerf NJ (2000) Physical complexity of symbolic sequences. Physica D 137:62-69.

34. Bernaola-Galvan P, Roman-Roldan R, Oliver JL (1996) Compositional segmentation and long-range fractal correlations in DNA sequences. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 53: 5181-5189.

35. Ramensky VE, Makeev VJ, Roytberg MA, Tumanyan VG (2000) DNA segmentation through the Bayesian approach. J Comp Biol 7(1):215.

36. Kelkar A, Thakur V, Ramaswamy R, Deobagkar D (2009) Characterisation of inactivation domains and evolutionary strata in human X chromosome through Markov segmentation. PLoS one 4:e7885.

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

38. Lin J (1991) Divergence measures based on the Shannon entropy. IEEE Trans Inform Theory 37:145-151.

39. Bernaola-Galvan P, Grosse I, Carpena P, Oliver JL, Roman-Roldan R, Stanley HE (2000) Finding borders between coding and noncoding DNA regions by an entropic segmentation method. Phys Rev Lett 85(6):1342.

40. Grosse I, et al. (2002) Analysis of symbolic sequences using the Jensen-Shannon divergence. Phys Rev E Stat Nonlin Soft Matter Phys 65:041905.

41. Peng CK, Buldyrev S, Goldberger A, Havlin S, Sciortino F, Simons M, Stanley HE (1992) Long-range correlations in nucleotide sequences. Nature 356:168.

42. 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 5:1863-1871.

43. Charlesworth D, Charlesworth B, Marais G (2005) Steps in the Evolution of Heteromorphic Sex Chromosomes. Heredity 95(2):118-128.

44. Rice WR (1984) Sex chromosomes and the evolution of sexual dimorphism. Evolution 38:735–742.

45. Rice WR (1996) Evolution of the Y sex chromosome in animals. Biosciences 46: 331–343.

46. Jordan C, Charlesworth D (2012) The potential for sexually antagonistic polymorphism in different genome regions. Evolution 66:505–516.

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

48. Skaletsky H, et al. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423:825-837.

49. Lemaitre C, Braga MD, Gautier C, Sagot MF, Tannier E, Marais GA (2009a) Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. Genome Biol Evol 1:56-66.

50. Wang J, et al. (2012) Sequencing Papaya X and Yh Chromosomes Reveals Molecular Basis of Incipient Sex Chromosome Evolution. Proc Natl Acad of Sci 109(34):13710-13715.

51. Nei M (1970) Accumulation of nonfunctional genes on sheltered chromosomes. Am Nat 104:311–322.

52. Charlesworth B, Charlesworth D (2000) The degeneration of Y chromosomes. Phil Trans R Soc Lond B Biol Sci 355:1563–1572.

53. Muller HJ (1964) The relation of recombination to mutational advance. Mutat Res 1:2-9.

54. Bachtrog D (2013) Y-chromosome evolution:emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet 14:113–124.

55. Finnegan DJ (1992) Transposable elements. The Genome of Drosophila melanogaster pp:1096–1107.

56. Montgomery EA, Charlesworth B, Langley CH (1987) A test for the role of natural selection in the stabilization of transposable element copy number in a population of Drosophila melanogaster. Genet Res 49:31–41.

57. Brookfield JFY (1991) Models of repression of transposition in P-M hybrid dysgenesis by P cytotype and by zygotically encoded repressor proteins. Genetics 128:471–486.

58. Dolgin ES, Charlesworth B (2008) The Effects of Recombination Rate on the Distribution and Abundance of Transposable Elements. Genetics 178(4):2169-2177.

59. Ross MT, et al. (2005) The DNA sequence of the human X chromosome. Nature 434:325-337.

60. Ming R, Moore P (2007) Genomics of Sex Chromosomes. Current Opinion in Plant Biology 10(2):123-130.

61. Votintseva AA, Filatov DA (2009) Evolutionary strata in a small mating-type-specific region of the smut fungus Microbotryum violaceum. Genetics 182:1391-1396.

62. Wilson MA, Makova KD (2009) Evolution and survival on eutherian sex chromosomes. Plos Genetics 5:e1000568.

63. Sandstedt SA, Tucker PK (2004) Evolutionary strata on the mouse X chromosome correspond to strata on the human X chromosome. Genome Res 14:267-272.

64. Nam K, Ellegren H (2008) The chicken (*Gallus gallus*) Z chromosome contains at least three nonlinear evolutionary strata. Genetics 180:1131-1136.

65. Bergero R, Forrest A, Kamau E, Charlesworth D (2007) Evolutionary Strata on the X Chromosomes of the Dioecious Plant *Silene Latifolia*: Evidence From New Sex-Linked Genes. Genetics 175(4):1945-954.

66. Hobza R, Eduard K, Boris V, Alex W (2007) The Role of Chromosomal Rearrangements in the Evolution of *Silene Latifolia* Sex Chromosomes. Molecular Genetics and Genomics 278(6):633-38.

67. Charlesworth D, Charlesworth B, Marais G (2005) Steps in the Evolution of Heteromorphic Sex Chromosomes. Heredity 95(2):118-28.

68. Ming R, Bendahmane A, Susanne S Renner (2011) Sex Chromosomes in Land Plants. Annual Review of Plant Biology 62(1):485-514.

69. Bergero R, Charlesworth D (2011) Preservation of the Y transcriptome in a 10MY old plant sex chromosome system. Current Biology 21:1470–1474.

70. Kohn M, Kehrer-Sawatzki H, Vogel W, Graves JA, Hameister H (2004) Wide genome comparisons reveal the origins of the human X chromosome. Trends Genet 20:12.

71. Veyrunes F, et al. (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Res 965-973.

72. Wilcox SA, Watson JM, Spencer JA, Graves JAM (1996) Comparative mapping identifies the fusion point of an ancient mammalian X-autosomal rearrangement. Genomics 35:66-70.

73. Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. Genetics 173:419-434.

74. Charlesworth B (1978) Model for evolution of Y chromsomes and dosage compensation. Proceedings of the National Academy of Sciences of the United States of America 75:5618-5622.

75. Carrel L, Park C, Tyekucheva S, Dunn J, Chiaromonte F, Makova KD (2006) Genomic environment predicts expression patterns on the human inactive X chromosome. Plos Genetics 2:1477-1486.

76. Wilson Sayres MA, Makova KD (2013) Gene survival and death on the human Y chromosome. Molecular Biology and Evolution 30:781-787.

77. Katsura Y, Satta Y (2012) No evidence for a second evolutionary stratum during the early evolution of mammalian sex chromosomes. PLoS One 7:e45488.

78. Fujita PA, et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Research 39:D876-882.

79. Lemaitre JF, Ramm SA, Barton RA, Stockley P (2009) Sperm competition and brain size evolution in mammals. Journal of Evolutionary Biology 22: 2215-2221.

80. Charchar FJ, Svartman M, El-Mogharbel N, Ventura M, Kirby P, Matarazoo MR, Ciccodicola A, Rocchi M, D'esposito M, Graves JA (2003) Complex events in the evolution of the human pseudoautosomal region 2 (PAR2). Genome Res 13:281-286.

81. Liu Z, et al. (2004) A primitive Y chromosome in papaya marks incipient sex chromosome evolution. Nature 427:348-352.

82. Filatov DA (2005) Evolutionary History of *Silene Latifolia* Sex Chromosomes Revealed by Genetic Mapping of Four Genes. Genetics 170(2):975-79.

83. Bergero R, Qiu S, Forrest A, Borthwick H, Charlesworth D (2013) Expansion of the Pseudo-autosomal Region and Ongoing Recombination Suppression in the *Silene Latifolia* Sex Chromosomes. Genetics 194(3):673-686.

84. Badouin H, Hood ME, Gouzy J, Aguileta G, Siguenza S, Perlin MH, Cuomo CA, Fairhead C, Branca A, Giraud T (2015) Chaos of Rearrangements in the Mating-Type Chromosomes of the Anther-Smut Fungus Microbotryum Lychnidis-dioicae. Genetics 200(4):1275-1284.

85. Ahmed S, et al. (2014) A Haploid System of Sex Determination in the Brown Alga *Ectocarpus Sp*.Current Biology 24(17):1945-1957.

86. Hou J, Ning Y, Defang Z, Yingnan C, Lecheng F, Xiaogang D, Tongming Y (2015) Different Autosomes Evolved into Sex Chromosomes in the Sister Genera of Salix and Populus. Sci Rep 5:9076.

87. Yu Q, et al. (2008) Low X/Y divergence of four pairs of papaya sex-liked genes. Plant J 53:124–132.

88. Pandey RS, Azad RK (2016) Deciphering evolutionary strata on plant sex chromosomes and fungal mating-type chromosomes through compositional segmentation. Plant Mol Biol 90(4):359-373.

89. Tuskan GA, et al. (2012) The Obscure Events Contributing to the Evolution of an Incipient Sex Chromosome in Populus: A Retrospective Working Hypothesis. Tree Genetics & Genomes 8(3):559-571.

90. Yin T, et al. (2008). Genome Structure and Emerging Evidence of an Incipient Sex Chromosome in Populus. Genome Res 18(3):422-30.

91. Geraldes A, et al. (2015) Recent Y Chromosome Divergence despite Ancient Origin of Dioecy in Poplars (Populus). Mol Ecol 24(13):3243-256.

92. Gaudet M, Jorge V, Paolucci I, Beritognolo I, Mugnozza G, Sabatti M (2008) Genetic linkage maps of Populus nigra L. including AFLPs, SSRs, SNPs, and sex trait. Tree Genetics and Genomes 4:25–36.

93. Hood ME, Petit E, Giraud T (2013) Extensive divergence between mating-type chromosomes of the anther-smut fungus. Genetics 193:309–315.

94. Fontanillas E, et al. (2015) Degeneration of the non-recombining regions in the mating type chromosomes of the anther smut fungi. Mol Biol Evol 32:928–943.

95. Whittle CA, Votintseva A, Ridout K, Filatov DA (2015) Recent and massive expansion of the mating-type specific region in the smut fungus Microbotryum. Genetics 199:809–816.

96. Smit AFA, Hubley R, Green P (2015) RepeatMakser Open-4.0.

97. Syvanen M, Kado CI (1998) Horizontal Gene Transfer. Chapman & Hall, London.

98. Ochman H, Lawrence JG, Groisman E (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299-304.

99. Kooni EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55:709-742.

100. Gogarten JP, Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol 3:679-687.

101. Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605-618.

102. Andersson JO (2005) Lateral gene transfer in eukaryotes. Cell Mol Life Sci 62:1182-1197.

103. Sanchez C (2011) Horizontal gene transfer: eukaryotes under a new light. Nat Rev Microbiol 9:228.

104. Chan C X, Bhattacharya D (2013). Analysis of horizontal genetic transfer in red algae in the post-genomics age. Mob Genet Elements 3:e27669.

105. Raymond J, Blankenship RE (2003). Horizontal gene transfer in eukaryotic algal evolution. Proc Natl Acad Sci USA 100:7419-7420.

106. Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ (2003) Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga Bigelowiella natans. Proc Natl Acad Sci USA 100:7678-7683.

107. Bhattacharya D, et al. (2013). Genome of the red alga Porphyridium purpureum. Nat Commun 4:1941.

108. Qiu H, Price DC, Weber AP, Reeb V, Yang EC, Lee JM, Kim SY, Yoon HS, Bhattacharya D (2013). Adaptation through horizontal gene transfer in the cryptoendolithic red alga Galdieria phlegrea. Curr Biol 23:R865-866.

109. Schonknecht G, et al. (2013) Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. Science 339:1207-1210.

110. Armbrust EV, et al. (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306:79-86.

111. Bowler C, et al. (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 456:239-244.

112. Jain K, Krause K, Grewe F, Gaven FM, Weber A, Alan C. Christensen, Mower JP (2015) Extreme Features of the *Galdieria sulphuraria* Organellar Genomes: A Consequence of Polyextremophily?. Genome Biol Evol 7(1):367-380.

113. Azad RK, Lawrence JG (2012) Detecting laterally transferred genes. Methods Mol Biol 855:281-308.

114. Azad RK, Lawrence JG (2007). Detecting laterally transferred genes: use of entropic clustering methods and genome position. Nucleic Acids Res 35:4629-4639.

115. Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44:383-397.

116. Azad RK, Lawrence JG (2011). Towards more robust methods of alien gene detection. Nucleic Acids Res 39:e56.

117. Chan CX, Yang EC, Banerjee T, Yoon HS, Martone PT, Estevez JM, Bhattacharya D (2011) Red and green algal monophyly and extensive gene sharing found in a rich repertoire of red algal genes. Curr Biol 21:328-333.

118. Collen J, et al. (2013) Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. Proc Natl Acad Sci USA 110:5247-5252.

119. Matsuzaki M, et al. (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature 428:653-657.

120. Nakamura Y, et al. (2013) The first symbiont-free genome sequence of marine red alga, Susabi-nori (*Pyropia yezoensis*). PLoS One 8:e57122.

121. Keeling PJ, et al. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol 12:e1001889.

122. Chan CX, et al. (2012) Porphyra (Bangiophyceae) transcriptomes provide insights into red algal development and metabolism. J Phycol 48:1328-42.

123. Pandey RS, Saxena G, Bhattacharya D, Qiu H, Azad RK (2017), Using complementary approaches to identify trans-domain nuclear gene transfers in the extremophile *Galdieria sulphuraria* (Rhodophyta). J. Phycol 53:7–11.

124. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

125. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.

126. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268-274.

127. Eastmond PJ, Astley HM, Parsley K, Aubry S, Williams BP, Menard GN, Craddock CP, Nunes-Nesi A, Fernie AR, Hibberd JM (2015) Arabidopsis uses two gluconeogenic gateways for organic acids to fuel seedling establishment. Nat Commun 6:6659.

128. Nakanishi T, Ohki Y, Oda J, Matsuoka M, Sakata K, Kato H (2004) Purification, crystallization and preliminary X-ray diffraction studies on pyruvate phosphate dikinase from maize. Acta Crystallogr D Biol Crystallogr 60:193-194.

129. Price DC, et. al. (2012) *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. Science 335:843-7.

130. Bazinet AL, Cummings MP (2012) A comparative evaluation of sequence classification programs. BMC Bioinformatics 13:92.

131. Peabody MA, Rossum TV, Lo R and Brinkman FSL (2015) Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC Bioinformatics 16:363.

132. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17(3):377–386.

133. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J (2008) Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res 36(7):2230–2239.

134. Brady A, Salzberg SL (2009) Phymm and PhymmBL: Phylogenetic classification with interpolated Markov models. Nat Methods 6:673–676.

135. Brady A, Salzberg SL (2011) PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat Methods 8:367.

136. Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B (2008) Metagenome fragment classification using N-mer frequency profiles. Adv Bioinformatics 2008:1-12.

137. Rosen GL, Reichenberger ER, Rosenfeld AM (2011) NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics. 27:127–129.
138. Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics 10:56.

139. Gregor I, Droge J, Schirmer M, Quince C, McHardy AC (2016) PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. Peer J 4:e1603.

140. Ounit R, Wanamaker S, Close TJ, Lonardi S (2015) CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 16:236.

141. Nalbantoglu OU, Way SF, Hinrichs SH, Sayood K (2011) RAIphy: Phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. BMC Bioinformatics 12:41.

142. Wu M, Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biol 9(10):R151.

143. Stark M, Berger SA, Stamatakis A, Von Mering C (2010) MLTreeMap–accurate MaximumLikelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. BMC Genomics 11:461.

144. Schreiber F, Gumrich P, Daniel R, Meinicke P (2010) Treephyler: fast taxonomic profiling of metagenomes. Bioinformatics 26(7):960–961.

145. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37-43.

146. Richter DC, Ott F, Auch AF, Schmid R, Huson DH (2008) MetaSim—A Sequencing Simulator for Genomics and Metagenomics. PLoS ONE 3:e3373.

147. Azad RK, Borodovsky M (2004) Effects of choice of DNA sequence model structure on gene identification accuracy. Bioinformatics 20:993-1005.

148. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421.

149. Re MA, Azad RK (2014) Generalization of Entropy Based Divergence Measures for Symbolic Sequence Analysis. PLoS ONE 9(4):E93532.

150. Tong H (1975) Determination of the Order of a Markov Chain by Akaike's Information Criterion. Journal of Applied Probability 12(3):488-497.

151. Li W (2001) New Stopping Criteria for Segmenting DNA Sequences. Phys Rev Lett 86(25):5815-5818.