MODELING MITOCHONDRIAL POPULATION GENETICS

MODELING MITOCHONDRIAL POPULATION GENETICS

By

STEPHANIE SUN, B.Sc. (Hons.)

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science

McMaster University

MASTER OF SCIENCE (2009)
(Biology)

McMaster University
Hamilton, Ontario

TITLE: Modeling Mitochondrial Population Genetics

AUTHOR: Stephanie Sun, B.Sc.Hons. (McMaster University)

SUPERVISOR: Dr. G. Brian Golding

NUMBER OF PAGES: ix, 119

Abstract

Indirect tests have detected recombination in diverse animal mitochondrial DNA (mtDNA), including mammals. These results have far reaching implications for evolution and ecology, as virtually all animal population genetics studies assume mtDNA is clonally inherited. For the first time, we demonstrated that the molecular patterns detected by these tests could alternatively be explained by mutation rate heterogeneity, or clusters of sites with unusually low or high mutation rates. The false positive rates of six common tests for recombination were evaluated under models of mutation rate heterogeneity with theoretical and biologically estimated parameters. All tests produced an elevated level of false positives, casting serious doubts on the claim that animal mtDNA does not follow clonal inheritance.

With uniparental inheritance, a haploid genome, multiple copies within a cell, and a replication cycle that is independent of the cell cycle, mitochondria population genetics are markedly non-Mendelian. Numerous questions remain in mitochondrial population genetics theory, such as the effect of dominance in the context of unique mitochondrial biology. Using simulations, we determined the fixation probabilities of advantageous mtDNA mutations under various modes of dominance and levels of polyploidy within a cell. The effect of a bottleneck and multiple cell lines (somatic and germ) with different selective pressures was investigated. The effect of increasing drift on fixation probabilities depends on the mode of dominance: recessive mutations become more likely to fix, but dominant mutatinos become less likely to fix. These results support the theory that drift plays a fundamental role in maintaining evolutionary stability of mitochondria by increasing the genetic variation among offspring, but suggests that the efficiency of this mechanism involves a more complex interaction between dominance and drift than previously thought.

Acknowledgements

This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca) and Dr. Gwenaël Piganeau, who generously shared her $\texttt{LDr}^2$, $\texttt{LD}|\texttt{D}'|$, and $\texttt{Max}\ \chi^2$ scripts and provided assistance in using them.

I owe innumerable thanks to the many individuals, some of whom are not mentioned here, for taking my education beyond study and learning to one full of challenge, growth, and enlightenment. I will always look back fondly on my time at McMaster, but even more so on the people who made it into the unforgettable experience that it was.

Thank you to Wilson Sung, Melanie Lou, and all my Golding lab mates for their unending support, sage guidance, but most dear to me, their friendship.

Thank you to Ben Evans and J. P. Xu, who infused every graduate class, committee meeting, and casual conversation with insightful comments, humour and enthusiasm.

Thank you to my supervisor and teacher, Brian Golding. Always with the greatest of patience and humour, Brian has given me freedom to learn, challenge through questions, and encouragement after mistakes. In opening the door to his lab, he opened a door to new lifelong skills and and a mind stretched to new dimensions.

Last but never least, thank you to my family for inspiring me to make the most of any challenge; in research or in life.

Contents

List of Tables

List of Figures

Part I

INTRODUCTION

Chapter 1

Introduction to mitochondrial population genetics

Mitochondria have long since outgrown their nickname as the power-houses of the cell. With the exception of mature red blood cells and some protozoan, mitochondria can be found performing numerous essential functions in nearly all eukaryotic cells (Cummins, 1998). From ATP production and metabolism to apoptosis and detoxification, fertility and longevity to anthropology and conservation, mitochondria have become not just an important subject but an important tool in biological studies. These semi-autonomously self-replicating organelles contain a genome with physical characteristics, mechanisms of inheritance, and overall genetic behaviour that is distinct from the nuclear genome. Despite decades of research and the identification of major forces shaping mitochondrial evolution, many basic questions remain. How does the interplay between drift and selection in the context of non-Mendelian mitochondrial genetics? How certain can we be about fundamental assumptions made of mitochondrial population genetics? This introduction to mitochondrial population genetics is a springboard into these far-reaching questions.

1.1 An endosymbiotic origin

The connection between bacteria and the origin of mitochondria was first made in 1890 (Kutschera and Niklas, 2005) when it was reported that mitochondria shared similar staining properties with free-living microbes (Altmann, 1890). However, this suggestion was overlooked by the greater scientific community and it was not until almost a century later that Margulis (1970) revisited and updated the theory to the serial theory of endosymbiosis. This theory posits that multiple endosymbiosis events occurred to produce the mitochondrial and later, the plastid, lineages. The first endosymbiotic event, which lead to the origin of mitochondria, likely occurred between 2200 and 1500 MYA when α-proteobacteria capable of oxidative phosphorylation were engulfed by archaea-like host cells, producing an organism capable of cellular respiration. Evidence supporting this widely accepted theory is based upon numerous similarities between mitochondria and eubacteria DNA, ribosomes, and membranes.

1.2 Function and disease

Mitochondria play an essential role in metabolism and apoptosis (Wallace, 1999). In the final step in aerobic metabolism, oxidative phosphorylation, various metabolites are oxidized and their electrons are transported down a chain of proteins on the inner mitochondrial membrane, producing a proton gradient within the mitochondria. The energy stored in the proton gradient drives the production of ATP, which is used as energy throughout the cell. Apoptosis, or programmed cell death, involves the clustering of mitochondrial proteins to form pores at points where the inner and outer membranes meet. Upon opening, these pores release factors and enzymes into the cytoplasm, leading to cell death. The formation and opening of the pore can be induced by low energy output, excessive intake of calcium ions, or excessive exposure to reactive oxygen species produced during oxidative phosphorylation.

Proper function of mitochondria depends on both mitochondrial and nuclear genes, which have markedly different genetic characteristics. Since endosymbiosis, most genes encoding mitochondrial proteins and regulating mitochondrial processes have introgressed into the nucleus. This leads to interesting population dynamics, as the fitness of mitochondria is not just related to mtDNA, but the nuclear DNA background as well. Incompatibilities between the mitochondrial and nuclear genomes act as post-zygotic reproductive barriers, and probably play an important role in speciation (Cummins, 1998; Lee *et al.*, 2008).

Base substitutions and rearrangements in mtDNA are associated with a wide range of medical conditions; infertility (Reynier *et al.*, 2001; May-Panloup *et al.*, 2003), neurological degeneration (Howell, 1997), cardiovascular disease (DiMauro, 2001), and cancer (Wallace, 1999; Jakupciak *et al.*, 2008) are just a few examples. Mutations in genes encoding mitochondrial proteins lead to metabolic disorders in about one in every 10,000 live births (Smeitink, van den Heuvel and DiMauro, 2001). Individuals with disorders caused by mitochondrial mutations are generally *heteroplasmic*, or possess multiple mtDNA alleles (the alternate state of possessing all genetically identical mtDNA is *homoplasmy*). The penetrance of such disorders is typically correlated to the tissue or individual's level of heteroplasmy. Tissues that utilize a high level of oxidative phosphorylation (i.e. aerobic metabolism; muscles, nerves) are more sensitive to levels of mutant mtDNA (DiMauro, 2001). mtDNA may be a useful tool in medical diagnosis or screening for certain cancers (Jakupciak *et al.*, 2008) and many other disorders (Wong and Boles, 2005), but without a full understanding of mitochondrial population genetics, the utility of mtDNA as an accurate predictor is limited.

1.3 Mitochondrial DNA

In evolution and ecology

Animal mtDNA is well suited for high resolution evolutionary studies, such as those involving short time spans or within-species comparisons. This is owing to a high copy number, high substitution rate, and unusual pattern of segregation which produces low intra-population variation but high inter-population variation.

mtDNA has already facilitated the identification of cryptic animal species (Hebert *et al.*, 2004), exploration of human migration and origins (Wallace, Brown and Lott, 1999), and they have virtually unlimited applications in conservation and biodiversity monitoring, as well as in phylogenetics and ecology. In fact, numerous international projects are currently underway to catalog global biodiversity using standard extra-nuclear loci, or barcodes. The chosen barcode for animals is cytochrome *c* oxidase I (COI), and for land plants a 2-locus combination plastid barcode of a carboxylase and a tyrosine kinase (rbcL+matK) has been selected (Hollingsworth *et al.*, 2009). Selection of an appropriate marker for fungi is complicated by the presence of mobile introns; consensus on a fungi barcode has yet to be reached, although NADH dehydrogenase 6 has been proposed for Ascomycota (Santamaria *et al.*, 2009).

The utility of animal mtDNA as an evolutionary and ecological tool depends upon the integrity of certain assumptions made about mitochondrial genetics: that mitochondrial alleles segregate during mitosis, mtDNA are uniparentally inherited, and mtDNA loci are tightly linked. These non-Mendelian genetic characteristics are discussed in later sections.

Genome structure

The animal mitochondrial genome is a circular DNA molecule ranging from 15,000-17,000 base pairs in length. Each molecule contains 13 protein-coding genes (including subunits of NADH-ubiquinone oxireductase, cytochrome *c* oxidase, H^+-ATP synthase, and cytochrome *b*), 22 tRNA genes, 2 rRNA genes, some intergenic spacers (untranscribed regions), and the D-loop control region which regulates replication. The animal mitochondrial genome lacks splicisomal introns, with the exception of cnidarian mtDNA (van Oppen *et al.*, 2000), and with only small variations in genome size and gene order, is considered structurally and evolutionarily stable.

The mitochondrial genome of plants possesses much more structural variability; it can be linear or circular, ranges from 26,000 bp to 2,500,000 bp, has undergone frequent rearrangements, duplications, and deletions, and in some plants the genome is spread across multiple DNA molecules (subgenomic circles). Plant and animal mitochondrial genomes contain the same genes, but different plants may carry them in different numbers, and plant mtDNA also includes pseudogenes and introns (Graur and Li, 2000). From here onward, all references to mtDNA strictly refers to animal mtDNA.

Polyploidy

Mitochondria and cells contain multiple copies of haploid mtDNA. There are 2-10 genomes per mitochondrion (Nass, 1969a), and up to thousands (Nass, 1969b; Bogenhagen and Clayton, 1977) or hundreds of thousands of mitochondrial genomes in a cell (Piko and Taylor, 1987; Reynier *et al.*, 2001) depending on species, tissue, and cell type. Human somatic cells contain between 10^3-10^4 mtDNA copies per cell (Clayton, 1982), and about 10^5 in mature oocytes (Piko and Taylor, 1987).

Mitochondrial genomes are arranged within mitochondria in aggregates called nucleoids. The average mammalian nucleoid consists of 5-7 mtDNA which are packed into a compact bundle with DNA binding proteins, and tethered to the inner mitochondrial membrane (Iborra, Kimura and Cook, 2004). Nucleoids may move to other mitochondria as the organelles undergo fusion and fission.

Mutation rate

Generally, the mutation and substitution rates of animal mtDNA is higher than the nuclear rates. Although the evolution rate of mammalian mtDNA is often cited as being 5 to 10 times faster than that of nuclear DNA (Brown, George and Wilson, 1979), there is actually extensive heterogeneity among different regions of mtDNA; the ratio of mtDNA to nuclear DNA substitution rates range from about 3 in non-synonymous sites, to about 100 in tRNA. Mammalian mtDNA synonymous and non-synonymous sites, respectively, evolve at a rate of about 22 and 3 times that of nuclear DNA. While all mtDNA synonymous sites evolve uniformly, the rate of non-synonymous mtDNA evolution varies between genes. Cytochrome oxidase I (COI) is the slowest evolving gene, while ATP8 is the fastest evolving; ATP8 is about 6 times faster evolving than COI (Pesole *et al.*, 1999).

The high mutation rate in animal mtDNA may be due to a combination of reasons including exposure to a high concentration of mutagens, a lower fidelity DNA replication process, inefficient repair mechanisms, and reduced efficiency of selection. Ninety percent of reactive oxygen species (ROS) produced by the cell are generated by mitochondria during oxidative phosphorylation (Balaban, Nemoto and Finkel, 2005). Although mitochondria possess enzymes to convert the majority of ROS to safer materials, ROS undoubtedly create an environment conducive to mutations, particularly as mtDNA are not protected by histones. Indeed, numerous studies have reported increased mtDNA damage with age.

Continual replication, along with lower fidelity DNA replication, increase the mtDNA mutation rate. Mitochondria replicate continuously throughout the cell cycle, even in non-dividing cells. In addition, the error rate of mitochondrial DNA polymerase γ is 1 in 1 $\times 10^6$ to 20 $\times 10^6$, while the error rate in nuclear DNA is 1 in 10^9 to 10^{10} (Johnson and Johnson, 2001).

Unlike the nucleus, mitochondria do not use a nucleotide excision repair mechanism of for DNA repair. Mitochondria rely upon the less efficient mismatch repair mechanism, which in mammals, appears to be missing important components. Whether this means the mitochondrial repair mechanisms in mammals performs at a reduced level compared to other mitochondria, or whether mammals utilize an unconventional method of mismatch repair is unclear (Mason and Lightowlers, 2003).

Purifying selection may be reduced in mtDNA compared to nuclear DNA. Although mtDNA polyploidy is thought to protect against deleterious mutations by compensating for damaged mtDNA, which may be lost through drift or selection during cell division (Mason and Lightowlers, 2003), if the mutation is recessive, the mutation will be masked by the other mtDNA copies. Studies have shown that bottlenecks (Roze, Rousset and Michalakis, 2005) and linkage (Birky and Walsh,

1988) will also reduce the efficiency of selection.

1.4 Non-Mendelian genetics

Despite a high mutation rate, cells and mitochondria are typically homoplasmic. This is due to the non-Mendelian nature of mitochondrial genetics which generally acts to increase the influence of drift on mitochondrial evolution. Genetic drift (or simply "drift") refers to changes in allele frequency due to stochastic processes. That is, the allele frequency can change due to chance differences between mtDNA in replication, mitochondrial turnover and/or segregation. With each drift event, allele frequencies slightly increase or slightly decrease. An increase or a decrease are equally likely to happen, but whether drift produces fixation or loss is a function of the number of drift events, the initial allele frequency, and the population size. The effect of drift is most evident in small populations, where even small chance changes in frequencies can lead to rapid fixation or loss. Once an allele is fixed or lost, the allele frequency is constant until the next mutation.

Although there is variation in fixation rates among different animal species, fixation of mammalian mitochondrial mutations is rapid. For example, a mitochondrial mutation fixed in 2-3 generations in cows (Ashley, Laipis and Hauswirth, 1989), but in many more generations in insects (Rand and Harrison, 1986). The difference in segregation rate likely reflects a higher number of segregating mtDNA molecules in insects — an order of magnitude greater than the number of segregating mtDNA in mammals (Ashley, Laipis and Hauswirth, 1989), and hence, decreased sensitivity to drift. The number of segregating mtDNA (i.e. the effective population size, N_e, of mtDNA) is related to not only the number of mtDNA/cell, but the extent of mtDNA structuring in nucleoids, and bottlenecks. If nucleoids are homoplasmic, the N_e of mtDNA will decrease as the number of mtDNA/nucleoid increase (Khrapko, 2008) because mtDNA in the same nucleoid will segregate together. Differences in segregation rate could also be due to the presence of a tighter bottleneck in mammalian oogenesis (Ashley, Laipis and Hauswirth, 1989). Bottlenecks decrease the N_e of mtDNA, but how this affects mitochondrial evolution remains under debate. Although it is generally accepted that bottlenecks contribute to the genetic load of mitochondria, studies have shown that bottlenecks and other forms of within-generational drift can increase the efficiency of selection by increasing genetic variance among offspring (Bereiter-Hahn and Vöth, 1994; Takahata and Slatkin, 1983) when there is strict uniparental inheritance (Rokas, Ladoukakis and Zouros, 2003).

Drift within and between generations

Replication and segregation of the nuclear genome into daughter cells during mitosis is highly regulated compared to the mitochondrial genome (Birky, 2001). The nuclear genome is replicated once during interphase, and each daughter cell receives one copy. Since nuclear homologous alleles do not segregate during mitosis, heterozygous parents will faithfully produce heterozygous daughter cells. In contrast, mitochondria are continually splitting and being replaced on a schedule that is independent of the cell cycle, with some copies replicating more often due to chance or selection (Birky, 1983). Since partitioning of mitochondria during cytokinesis occurs stochastically, daughter cells will not necessarily inherit the parental genotype. As a result, heteroplasmic cells will often produce homoplasmic daughters. This pattern of within-generational drift is termed *vegetative segregation*, and occurs through both mitosis and meiosis. Since segregation is a random process in mitochondria, alleles do not necessarily segregate evenly into gametes. This is in contrast to nuclear alleles, which always segregate evenly during meiosis such that each gamete receives a copy of each allele ("Mendel's First Law of Segregation").

Uniparental inheritance

Mitochondria are maternally inherited in most eukaryotes. The mechanism can vary between organisms, but generally, maternal inheritance is ensured by the fact that oocytes contain about 1000 times more mtDNA than sperm, and by targeting of sperm mtDNA for degradation after fertilization. Although these mechanisms are not foolproof (e.g. paternal leakage, which according to Rokas, Ladoukakis and Zouros (2003), is more likely in inter-specific hybrid crosses), generally the only way for heteroplasmy to arise is through new mutations (Birky, 2001). However, with to random partitioning of parental mitochondria into daughter cells and low initial frequency of any new mutations, a heteroplasmic parent for a new mutation will likely produce homoplasmic daughters, leading to the loss of the mutation.

Though rare, instances of human individuals possessing multiple mitochondrial alleles due to paternal leakage have been recorded (Kraytsberg *et al.*, 2004). However, the extent of paternal leakage in mammals has yet to be determined.

Mussels (families Mytilidae and Unionidae) utilize doubly uniparental inheritance and are therefore an exception to the usual maternal inheritance. Male bivalve mussel oocytes do not eliminate paternal mtDNA after fertilization. The result is that male mussels retain both maternal and paternal mtDNA, which can differ by more than 20% (Hoeh *et al.*, 1997). Nevertheless, paternal and maternal mtDNA are rarely found in the same tissues or cells.

Complete linkage

Alleles of different nuclear genes will segregate independently of each other ("Mendel's Law of Independent Assortment"). This is due to physical separation, being on different chromosomes or via cross-overs during metaphase. Genes on different chromosomes are obviously physically separated, and genes on the same chromosome can be separated via cross-overs during metaphase. In contrast, the entire mitochondrial genome is contained on a single molecule. As a result, mitochondrial loci do not segregate independently of loci on the same mtDNA molecule. An exception to this rule is the bivalve mussel, in which recombinant mtDNA has been found (Ladoukakis and Zouros, 2001). Nevertheless, recombination is assumed not to occur in animal mtDNA.

1.5 Controversies

Although mitochondria fitness is closely tied to host fitness, the strength of selection will be reduced due low N_e. With a high mutation rate, low N_e, and a lack of recombination, it is unclear how mtDNA stability has been maintained since the first endosymbiotic event over 1 billion years ago. Perhaps the combination of evolutionary forces which would lead to melt-down of the nuclear genome do not affect the mitochondrial genome the same way. Or perhaps some aspect(s) of our current assumptions regarding mitochondrial genetics is wrong. Only further research into the unique biology of mitochondria and mtDNA will help address these puzzling issues.

Recombination in animal mtDNA

Virtually all animal models of population genetics in animal studies assume no recombination in mtDNA. Although recombination has been detected in animal mtDNA, empirical evidence of recombination remains rare. This may be because the m ajority of animal mtDNA does not recombine, or recombines at a level below levels of empirical detection. In response to the latter, numerous indirect tests have been developed, which utilize statistical methods of analyzing mtDNA to detect recombination. Using indirect tests, recombination has been detected in a wide range of animal mtDNA, from nematodes to birds, to fish and primates. Although mutation rate heterogeneity in the form of mutation hot spots (local regions with higher-than-average mutation rate) may be an alternative explanation to recombination, it remains overlooked in favour of recombination. In Part II, we evaluate the performance of indirect tests for recombination using simulations of mutation rate heterogeneity.

Muller's Ratchet

Recombination can improve the fitness of a population by combining sequences that each contain advantageous mutations, creating a recombinant sequence that may contain all advantageous mutations. Without recombination, the only way for a sequence to contain multiple mutations is if an advantageous mutation occurs multiple times on the same sequence. Otherwise, the two sequences will compete with each other, decreasing the fixation probability of both advantageous alleles. Recombination also provides a way of reducing the genetic load, or accumulation of deleterious mutations. Without recombination, the only way for a sequence to rid itself of a deleterious mutation is with recurrent mutations at that site. In other words, asexually reproducing entities such as mitochondria, which do not recombine, will continue to irreversibly accumulate deleterious mutations, inevitably leading to a mutational melt-down of the genome. This prediction is called Muller's Ratchet, a term first coined by Felsenstein (1974). How mitochondria escape Muller's Ratchet is the focus of virtually all studies in mitochondrial population genetics.

Numerous aspects of mitochondrial genetics are thought to contribute to Muller's Ratchet by decreasing N_e and consequently, decreasing the effectiveness of selection. For example, linkage decreases N_e, even if the linked loci do not interact. This effect, which is called the Hill-Robertson effect, has been demonstrated through mathematical and simulation studies where complete linkage decreases fixation of advantageous mutations but increases fixation of deleterious mutations, regardless of whether these alleles are linked to a deleterious mutation or advantageous mutation. Fixation of neutral mutations is determined by N_e but, being independent of selection, will be unaffected by linkage to selected sites, be they deleterious or advantageous (Birky and Walsh, 1988).

High levels of drift have been suggested as a mechanism by which mitochondria slow or even stop Muller's Ratchet. Simulations of within-generational drift (Takahata and Slatkin, 1983) and tight bottlenecks (Bereiter-Hahn and Vöth, 1994) increased the rate of fixation for advantageous mutations, while decreasing fixation of deleterious mutations. This is thought to happen by increasing the variation among offspring, thereby increasing the effectiveness of selection, despite the resulting decrease in the size of the highest fitness class.

How mitochondrial mutations spread through a cell, individual, or population has been studied theoretically but lacks quantitative answers. This is in part due to the difficulty in detecting mitochondrial heteroplasmy at extremely low frequencies, such as when the mutation initially arises. In Part III, we determined fixation probabilities of mitochondrial mutations under different population parameters. In particular we were interested in observing the effect of the mode of allelic interaction. Although haploid on the genome level, the degree of dominance of a mutation is an important consideration on the level of the organelle, cell, and individual, as this will affect the expression on a polyploid background. Drift affects fixation at every level of the mtDNA population and depending on the mode of allelic interaction, may be offset by selection. The purpose of these simulations, then, is to compare the influence of drift versus the influence of selection on mitochondrial mutation fixations, under different degrees of dominance.

Part II

RECOMBINATION

Chapter 2

Recombination in cercopithecine mtDNA

2.1 Abstract

Overlooking recombination during phylogenetic reconstruction will produce trees resembling those undergoing exponential population growth. Due to a unique population structure, cercopithecine monkey phylogenies will also resemble exponentially growing populations. Recombination has been detected in cercopithecine mtDNA using statistical tests, which have been shown to have elevated false positive rates under biologically extreme conditions. Here, we evaluate the performance of three common recombination tests under cercopithecine models of evolution. The number of segregating sites, nucleotide diversity, Γ-distributed rate heterogeneity, population growth, and population subdivision were estimated from cercopithecine mtDNA and simulated without recombination. When sequences are simulated without exponential population growth or population subdivision, most recombination tests (`Max` χ^2 Global, and GENECONV Global) do not return a significant level of false positives, although `Reticulate` is more susceptible to false positives at extreme rate heterogeneity (smaller Γ shape parameter, α) and high levels of divergence (Posada and Crandall, 2001). A cercopithecine mtDNA model of population structure and growth produce elevated false positives in one test, GENECONV Local.

2.2 Introduction

Different parts of a recombined sequence possess different evolutionary histories. If recombination is not accounted for, this contradiction leads to an overestimation of mutation rate heterogeneity (Galtier *et al.*, 2006; Schierup and Hein, 2000) and the number of substitution events (Schierup and Hein, 2000; Eyre-Walker, Smith and Maynard Smith, 1999). With longer terminal and total branch lengths, shorter basal branches, and an underestimated time to the most recent common ancestor, the phylogeny would resemble that of an exponentially growing population (Schierup and Hein, 2000).

Long terminal branch lengths and short internal branches are also characteristic of phylogenies built from highly diverged subpopulations, such as the mtDNA of cercopithecine monkeys. Cercopithecine monkeys, a sub-family of Old World monkeys, are organized by female philopatry, a form of population structure where females spend their lives in their natal groups while males disperse and migrate between groups. This restriction on mtDNA gene flow, coupled with a high mtDNA mutation rate (Brown, George and Wilson, 1979) results in very low levels of divergence within subpopulations but high levels of divergence between subpopulations and between species (Melnick and Hoelzer, 1992).

In a survey of mtDNA from 267 diverse animal species (Piganeau, Gardner and Eyre-Walker, 2004), five mammals were identified as being among those with the strongest evidence for recombination. Of the five mammals, three were cercopithecine monkeys: the Guinea baboon; *Papio papio*, the Pigtail macaque; *Macaca nemestrina*, and the Mandrill; *Mandrillus sphinx*. The support for recombination was obtained through indirect tests, which scan the alignment for signals left behind by recombination events, such as an uneven distribution of matches and mismatches, regions with high sequence similarity, and a correlation of linkage disequilibrium with distance. Given that overlooking recombination will produce data that resembles exponentially growing populations, and that recombination has been indirectly detected in a sub-family of animals whose mtDNA likely resembles that of exponentially growing populations, it may be that the indirect tests for recombination are producing false positives in the face of high sequence divergence and exponential growth.

Numerous statistical tests have been developed to detect signals left behind by recombination events, such as an uneven distribution of polymorphic sites (Maynard Smith, 1992; Posada and Crandall, 2001), regions with high sequence similarity (Sawyer, 1989), clustering of phylogenetically incompatible sites (Jakobsen and Easteal, 1996), or a high correlation of linkage disequilibrium with physical distance (Piganeau, Gardner and Eyre-Walker, 2004). When collecting indirect evidence for recombination, it is generally recommended that more than one test be used as the power of indirect tests can vary widely under different conditions (Posada and Crandall, 2001; Posada, 2002; Wiuf, Christensen and Hein, 2001). Six of the most widely used and powerful indirect tests for recombination (Posada and Crandall, 2001; Posada, 2002; Wiuf, Christensen and Hein, 2001; Bruen, Philippe and Bryant, 2006) are described in Section 2.3.

Previous studies have evaluated indirect recombination tests due to extreme rate heterogeneity, high levels of divergence (Posada and Crandall, 2001), and combinations of these conditions with exponential growth (Bruen, Philippe and Bryant, 2006), and have found conditions where indirect tests can be susceptible to a high false positive rate. However, the conditions where this occurs were not necessarily within the realm of biological reality. In this study, species-specific models of evolution were estimated from cercopithecine mtDNA using Bayesian likelihood: the proportion of segregating sites (Ss) and nucleotide diversity (π, the average number of site differences between two sequences), site-specific substitution rate heterogeneity, population structure, and exponential population growth. Recombination tests were evaluated against simulations built under these models, and the level of false positives returned was quantified.

2.2.1 Auto-correlated rate heterogeneity

Mutation hot spots, or localized regions with higher than average substitution rates, are often suggested as an alternative explanation to recombination in mtDNA. Yang (1995) proposed an auto-discrete-Γ model which goes a step beyond the traditional application of the Γ distribution, taking into account the spatial distribution of sites along a sequence. That is, in Yang (1995)'s "auto-discrete-Γ model", substitution rate heterogeneity is modeled upon a Γ distribution (which for computational purposes uses a discrete distribution to approximate the continuous one), but here, the rates of adjacent sites are correlated and Markov-dependent. Markov dependence is a characteristic where the rate of a site at position n is specified by the rate at site $n - 1$, but is independent of all other sites, including otherwise preceding sites (at position $n - 2$, $n - 3$, $n - 4$ ... *et cetera*), and for simplicity, any sites after position n ($n + 1$, $n + 2$, $n + 3$ ... *et cetera*).

Bruen, Philippe and Bryant (2006) showed that a 0.9 correlation parameter for the auto-discrete-gamma model, ρ, will produce a significantly elevated level of false positives from the tests `Max` χ^2 and `NSS` (also known as `Reticulate`) when either

1. nucleotide diversity is at least 25%, or
2. nucleotide diversity is at least 10% with extreme rate heterogeneity (shape parameter, $\alpha = 0.1$)

and either

1. the population is growing with rate $\beta = 5000N$ per generation, or
2. sample size is large ($N = 50$).

The recombination tests of linkage disequilibrium (as measured by r^2 and $|D'|$) did not falsely infer recombination with these conditions.

2.3 Method

2.3.1 Recombination tests

In this section the theory behind five tests for recombination will be briefly described. These five were chosen because they are among the most commonly used indirect tests of recombination, and have been used to detect recombination in animal mtDNA (Piganeau, Gardner and Eyre-Walker, 2004; Tsaousis *et al.*, 2005). The tests are `GENECONV` global, `GENECONV` local, `Max` χ^2 global, `Max` χ^2 local, and `Reticulate`.

`GENECONV` (GC) searches for evidence of gene conversion (a non-reciprocal form of recombination where a locus is copied over a locus in another sequence) by searching the alignment for highly similar pairs of sequences. To do this, GENECONV first removes all invariant sites from the alignment, leaving only polymorphic sites for consideration. Each pair of sequences is compared and regions of the sequences, or fragments, are identified which are either unusually long stretches of perfect matches, or are unusually similar. Similarity is based on a scoring system where a matching site adds 1 to the score, but a mismatch is penalized according to a user-defined mismatch-penalty. The optimal mismatch penalty depends on the rate of mutations relative to the rate of gene conversion. In terms of power, a higher mismatch penalty is a more conservative setting that is better suited for detecting recent gene conversion events or when the mutation rate is much lower than the recombination rate. The mismatch penalty is set by specifying a parameter called the gscale. In this study a gscale of 0 was used, which equates to an infinite mismatch penalty. That is, with a gscale of 0, GENECONV will only consider fragments of perfect matches. Data without recombination is created by permuting sites and scoring the resulting fragments. The significance of each fragment's similarity score is compared to the similarity scores of fragments from data without recombination, and is assessed in two contexts: globally or locally. The number of permutations, 10^4, and gscale, 0, are the same as those used in studies reporting widespread recombination in animal mtDNA (Piganeau, Gardner and Eyre-Walker, 2004; Tsaousis *et al.*, 2005).

In the global comparison, GENECONV Global (GCG), sites in the entire sequence alignment were permuted and fragments are compared to the fragments between all possible sequence pairs. The GENECONV Global test uses a built-in correction to correct for the multiple fragment comparisons between sequence pairs.

The local comparison, GENECONV Local (GCL), differed from GENECONV Global in three ways. The first is that analysis is conducted on three sequences at a time, instead of the entire alignment. A Bonferroni correction was used to correct for elevated false positives due to multiple comparisons within the same data. Although the Bonferroni correction is over-conservative and other methods of multiple-test corrections exist (Nakagawa, 2004), we used a Bonferroni correction because it was used to detect recombination in animal mtDNA (Tsaousis *et al.*, 2005). Secondly, significance is assigned using a BLAST-like scoring system rather than permutations. GENECONV Local is much more computationally intensive than GENECONV Global because it analyzes every combination of three sequences in the alignment. To mitigate this issue, Tsaousis *et al.* (2005) used a less accurate but faster modified BLAST scoring system (Altschul, 1993). The modification corrects for the fact that closely related sequences are more similar than distant sequences, and will therefore contain more significant fragments. GENECONV Local removes this bias by correcting fragment lengths and similarities by the evolutionary distance of sequence pairs. Thirdly, the gscale was changed from 0 to 1, allowing fragments to contain mismatches. This is justified in Tsaousis *et al.* (2005) by the drop in data available when comparing three sequences, rather than the whole alignment, at a time.

GENECONV analysis involves two types of fragments: inner and outer fragments. Inner fragments are interpreted as evidence for gene conversion between ancestors of two sequences present in the alignment. Up to this point, this explanation of GENECONV has only considered inner fragments. Outer fragments are used as evidence for gene conversion between ancestral sequences of a sequence in the alignment and one other that cannot be identified. An ancestral sequence cannot be identified if its descendant sequence is missing from the alignment, or if subsequent mutations and/or gene conversions have distorted the gene conversion sequence beyond recognition. Scoring outer fragments is analogous to scoring inner fragments. A site containing a unique base is like a match, adding 1 to the score, but a site containing a non-unique base is like a mismatch, and is penalized according to the gscale. When the gscale = 0, an outer fragment possesses a unique base at every polymorphic site. Outer fragments must be bounded on at least one side by a matching polymorphic site. In this study, cases when both inner and outer fragments were considered are abbreviated as GCG or GCL, and cases when only inner fragments were considered are abbreviated as GCI or GCLI. While some studies include outer fragments in the GENECONV criteria (Tsaousis *et al.*, 2005), others specifically do not (Piganeau, Gardner and Eyre-Walker, 2004).

Max χ^2 is a test to identify recombination breakpoints within a sequence alignment. This method was first proposed by Maynard Smith (1992), and is widely considered to be the best general test to detect recombination (Posada, 2002; Posada and Crandall, 2001). Max χ^2 assumes that in the absence of recombination, polymorphisms are evenly distributed among sites and between sequences. Invariant sites are first removed from the alignment. A sliding window is used to test each sequence pair by comparing the number of matches and mismatches in both halves of the window; that is, the middle of the window is the location of the potential breakpoint. To determine the significance of the difference in matches and mismatches on either side of the potential breakpoint, a $2 \times 2\chi^2$ value is calculated for each window, and is tested for significance using 100 permutations. If fewer than 10 permutations had a χ^2 value greater than or equal to the observed, the permutation test was repeated with 1000 permutations. This study used the same implementation and window settings as Piganeau, Gardner and Eyre-Walker (2004): a maximum window half-size of 5 variable sites, and a sliding window step size of 2bp. The implementation is a modification of Posada and Crandall (2001) where the entire length of the alignment is analyzed, and only windows with expected χ^2 greater than 2 are considered. According to Piganeau, Gardner and Eyre-Walker (2004), these modifications improve the power of Max χ^2 while maintaining the same false positive error rate. Max χ^2 local is identical to Max χ^2 global except that the analysis is conducted on three sequences at a time, rather than the whole alignment.

Reticulate (Jakobsen and Easteal, 1996), also known as NSS, compares pairs of informative sites to detect homoplasies. Pairs of sites are labelled as being incompatible (homoplastic) or compatible, and the significance of the clustering of incompatible sites is assigned by Neighbour Similarity Score (NSS) and permutation test. This method works best on relatively large alignments with well-diverged sequences and, according to the documentation, can detect recombination that changes tree topology but cannot detect recombination that only changes branch lengths. Reticulate first identifies all phylogenetically informative sites in the alignment. Informative sites are sites where the alignment contains at least 2 alleles, each of which is present in at least 2 samples. In this study, sites with more than 2 states were ignored. For each pair of informative sites the most parsimonious tree is generated. If the tree created from the two sites does not contain any recurrent or convergent mutations (homoplasies), the two sites are deemed compatible. Conversely, if the sites cannot produce the same tree without invoking homoplasy, they are labelled incompatible. A compatibility matrix is created where each row and column corresponds to an informative site in the order that they appear in the alignment. Each entry$_{ij}$ is coloured either black or white, indicating site$_i$ and site$_j$ are either incompatible or compatible. The Neighbor Similarity Score (NSS) is the average number of adjacent incompatible or compatible cells in the matrix and represents the degree of clustering of similar sites. The order of sites in the alignment was permuted 1000 times to calculate the significance of NSS.

The tests `LD|D'|` and `LD|D'|` (Piganeau, Gardner and Eyre-Walker, 2004) calculate the correlation between distance and $|D'|$ or r^2, respectively, then assign a level of significance according to a permutation test. Code provided in Piganeau, Gardner and Eyre-Walker (2004) was used to conduct the linkage disequilibrium tests `LDr`^{2} and `LD|D'|`. The original program included a correction for circularity of human mtDNA sequences. This correction was removed for our analysis because our sequences are non-human and cover 3 loci at the most. In hindsight, the removal of this correction is unnecessary, but should not affect the results because the sequence lengths do not nearly approach the complete length of mtDNA. These tests were included in the Sulawesi models with population structure and exponential growth.

2.3.2 Macaque-specific simulations

Ten partial *Macaca nemestrina* CYTB sequences, 859bp in length, were collected from Genbank; accession numbers AF350388 - AF350391 and AF350394 - AF350399. A distribution of parameters was estimated using MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) under a GTR+Γ+codon position model. Analysis ran for 5,000,000 generations, of which the first 1,000,000 generations were discarded as the burnin. Parameters were sampled every 100 generations. One-thousand replicate trees were simulated using `ms` (Hudson, 2002). Each tree contained 10 taxa, and each taxa contributed one 859bp sequence to the simulation; the same sample size and sequence length as the *Macaca nemestrina* CYTB data. Sequences were simulated using `Seq-Gen` (Rambaut and Grassly, 1997) under a GTR+Γ+codon position model of evolution.

A second dataset (Evans *et al.*, 1999) was gathered from Genbank, representing 7 macaque species from the Indonesian island of Sulawesi. This dataset (hereafter referred to as the Sulawesi macaques), consisted of 18 sequences including complete ND3, partial ND4L, and complete ND4. tRNA regions were included in `MrBayes` analysis but only parameters estimated from coding regions were used in simulations. Accession numbers of this group are AF091400 - AF091429. Population structure and exponential growth parameters for the Sulawesi macaques was modeled following the approach in Evans *et al.* (2008).

2.3.3 General growth model simulations

Random trees were generated using ms from which sequences were created using Seq-Gen. Sequence length was 1200bp, which is in the range of many mtDNA datasets. A range of θ and exponential population growth (β) was used to simulate sequences, which were then tested for recombination using Max χ^2, Reticulate, and GENECONV without outer fragments, and GENECONV with global fragments. β relates to population size as

$$N(t) = N_o e^{-\beta t} \tag{2.1}$$

where $N(t)$ is the population t generations before the present and N_o is the present population size. Time is measured in $4N_o$ generations. GENECONV local and Max χ^2 local were also tested, as these tests were used in Tsaousis *et al.* (2005). A sample size equal to Sulawesi macaque sample size ($N = 18$) was used, plus up to three others ($N = 9, 27$, or 54) which were evenly spaced to test the effect of sample size. When possible different sample sizes were tested, but due to time constraints the larger sample sizes were not always feasible to test.

2.3.4 Cercopithecine auto-correlated rate heterogeneity

Sequences were analyzed under the auto-discrete Γ model using MrBayes. Analysis ran for 5,000,000 generations with a burnin of 1,000,000 generations, and sampling every 100 generations. The total number of parameter samples used to form the distribution was therefore 8×10^4.

2.4 Results

2.4.1 *Macaca nemestrina* segregating sites and nucleotide diversity

The number of segregating sites (*Ss*) and nucleotide diversity (π) of *M. nemestrina* CYTB was calculated using a program written in Perl; 140 *Ss* and 0.0309 π. Through trial-and-error, a θ (defined as $4N\mu$) of 0.07 was used as it produced simulated datasets with approximately the same *Ss* and π as the *Macaca nemestrina* CYTB data; an average of 145.21 *Ss* and 0.0304 π for 1000 simulations. No tests detected an elevated level of false positives (Table 2.1). These results indicate that the specific *Ss* and π of *M. nemestrina* alone will not produce a significant level of false positives for recombination, although the level of false positives returned by GENECONV global is somewhat elevated at 8.9%.

2.4.2 *Macaca nemestrina* Γ-distributed rate heterogeneity

The shape parameter for a Γ-distributed rate heterogeneity of the *M. nemestrina* data set was determined by MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) using a GTR+Γ+codon position model. Analysis ran

Table 2.1: Percent false positives for recombination out of 1000 simulations of *Macaca nemestrina Ss* and π.

GENECONV global	Max χ^2 global	Reticulate
8.9	6.3	4.1

Table 2.2: Percent false positives out of 1000 simulations of *Macaca nemestrina Ss*, π and rate heterogeneity.

GENECONV global	Max χ^2 global	Reticulate
6.6	4.9	4.9

for 5,000,000 generations, of which the first 1,000,000 generations were discarded as the burnin. The average shape parameter for Γ-distributed rate heterogeneity was 24.146. Although the MrBayes estimated shape parameter was 24.146, a shape parameter of 24.0 with $\theta = 0.07$ better approximated the *Ss* and π of the *Macaca nemestrina* dataset, on average. This modification is necessary because the MrBayes estimated parameters are likelihood maximized based on the *Macaca nemestrina* consensus tree, but we are interested in parameters for a wide range of simulated trees. The average *Ss* for 1000 replicates was 156.161 and average π was 0.0294. Simulations of the *Macaca nemestrina* CYTB level of diversity and rate heterogeneity did not produce a significant level of recombination false positives (Table 2.2).

2.4.3 Sulawesi macaque population structure

Population structure was investigated as a potential source of false positives in Sulawesi macaque mtDNA. To approximate the sequence length and sample size of the Sulawesi macaque data, datasets 1200bp in length with 18 sequences/simulation were used. Using the approach in Evans *et al.* (2008), a 68 deme model matching the shape of Sulawesi with maximum likelihood estimated migration rate ($m = 1.9$) was used to simulate Sulawesi macaque mtDNA. The migration rate and deme structure were supplied to ms, which uses an infinite sites model to produce trees. These trees were then passed to Seq-Gen (GTR+I+Γ+codon position), which uses a finite sites model. Since Seq-Gen uses a finite sites model but ms uses an infinite sites model, the maximum likelihood estimated θ, once passed to Seq-Gen with the ms created infinite sites tree, produced *Ss* and π that was significantly higher than the Sulawesi *Ss* and π we wished to model. To remedy this, a $\theta = 0.0014$ was used, which produced an average *Ss* and π of 19.4% and 7.0% respectively. Although the simulations matched the Sulawesi on π, the average *Ss* was below the target *Ss* (24%). Since simply increasing θ to increase *Ss* would lead to an overestimated π, simulations were filtered based on π to test for false positives at a range of *Ss* and π close to the Sulawesi mtDNA values. GENECONV global and GENECONV lo-

Table 2.3: Percent false positives under Sulawesi macaque deme model, for *Ss* and π that approximate *Ss* and π of the Sulawesi macaque dataset. Labels are as follows: GCG; GENECONV global, GCI; GENECONV global inner fragments only, GCL; GENECONV local, GCLI; GENECONV local inner fragments only, MXG; Max χ^2 global, MXL; Max χ^2 local, RET; Reticulate, LDR; LDr^2, LDD; LD|D$'$|. Bottom rows summarize results of simulations filtered (Filt.) on π.

Avg. *Ss*	Avg. π	GCG	GCI	GCL	GCLI	MXG	MXL	RET	LDR	LDD
19.4	7	7.3	4.9	68.8	46.5	5.9	4.4	3.6	5.4	4.7
Avg. *Ss*	Filt. π	GCG	GCI	GCL	GCLI	MXG	MXL	RET	LDR	LDD
18.6	6	7.8	5.2	65.8	46.8	5.2	4.8	4.6	4.9	4.6
20.3	7	8.3	5.3	76.0	55.6	5.1	5.4	6.0	5.6	6.7
22.0	8	7.3	4.3	81.8	57.0	5.1	5.1	5.8	5.6	4.7

Table 2.4: Percent false positives under Sulawesi macaque growth model. Labels are as follows: GCG; GENECONV global, GCI; GENECONV global inner fragments only, GCL; GENECONV local, GCLI; GENECONV local inner fragments only, MXG; Max χ^2 global, MXL; Max χ^2 local, RET; Reticulate, LDR; LDr^2, LDD;LD|D$'$|.

GCG	GCI	GCL	GCLI	MXG	MXL	RET	LDR	LDD
6.9	3.8	39.5	24.8	4.1	4.2	4.9	5.8	5.3

cal produce elevated false positives under a model where Sulawesi macaques have population structure. GENECONV global consistently found over 7% false positives, although this can be reduced when outer fragments are ignored. GENECONV local found an extremely high number of false positives: between 45 - 82% (Table 2.3).

2.4.4 Sulawesi macaque growth model

One-thousand simulations were built from a Sulawesi macaque growth model with maximum likelihood estimated parameters of $\theta = 0.134$ and growth parameter, $\alpha = 4.5$ (Evans *et al.*, 2008). Again a significant level of false positives was not detected (Table 2.4) in any tests except GENECONV local. GENECONV local found 39.5% and 24.8% false positives, respectively, when outer fragments were and were not included. This level of false positives is not as high as the level produced under the deme model, but is significant nonetheless.

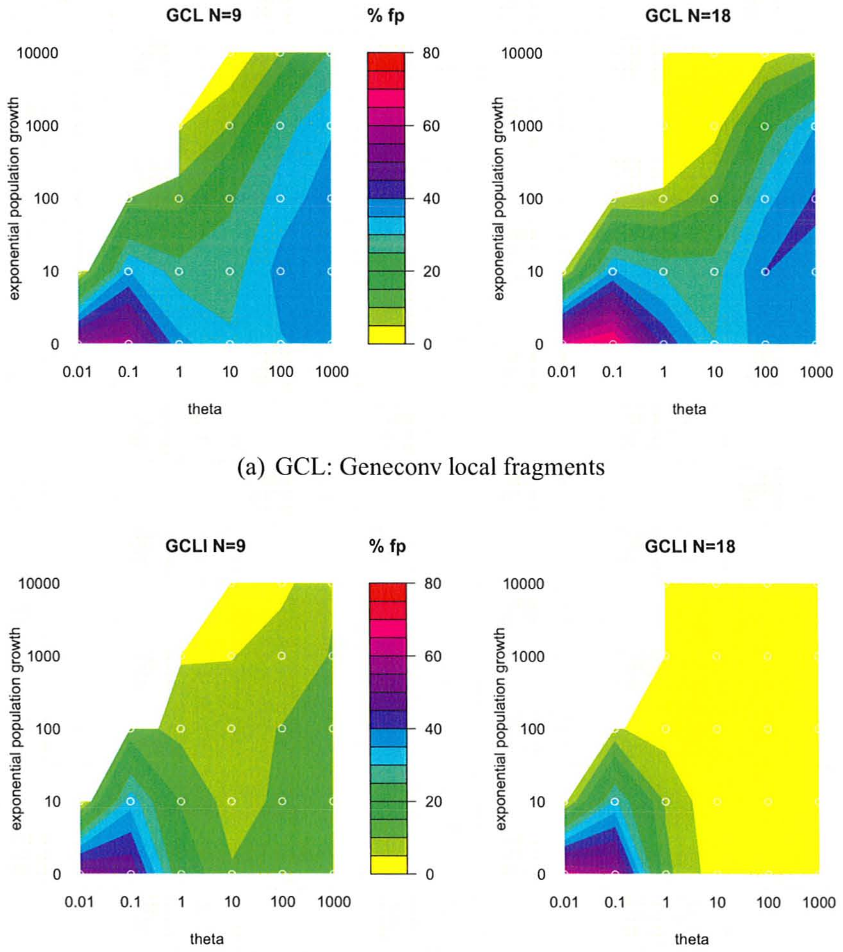
2.4.5 General growth model

To quantify the effect of exponential population growth on the number of false positives returned, simulations were created under a JC model of evolution. While theta was varied from 0.01 to 1000, the exponential growth parameter was varied from 0 (no growth) to 10000. Results for GENECONV local, GENECONV global, Max χ^2 global and local, and Reticulate can be found in Figures 2.1, 2.2, 2.3, and 2.4, respectively. Except GENECONV local (Figure 2.1), none of the tests produced a significant level of false positives, although GENECONV global (GCG) appears to produce between 5-10% false positives almost all the time, regardless of population growth (Figure 2.2). This suggests it has a higher intrinsic false positive rate, which is consistent with our previous results. The false positive rate of GENECONV local is improved by considering inner fragments only, particularly at high values of θ. In some cases, outer fragments are responsible for an increase of up to 40% false positives. This is likely because high levels of divergence increase the likelihood that a sequence will appear significantly different from the rest of the alignment. The level of false positives increases as the growth parameter, β, and θ increase. Areas where simulations did not produce enough sequence diversity for analysis to be run are indicated by white space in Figure 2.1. Larger values of β require larger values of θ to maintain the same level of diversity. As θ is increased, however, there comes a point when sites are so saturated with substitutions that any recombination signal is disrupted.

2.4.6 Cercopithecine auto-correlated rate heterogeneity

The distribution of *Macaca nemestrina* CYTB dataset parameters under an auto-discrete-Γ model are summarized in Table 2.5. Given that none of the correlation parameters are near 0.9 and the low nucleotide diversity of the CYTB dataset, an auto-discrete Γ model does not contradict evidence for recombination in *M. nemestrina* CYTB.

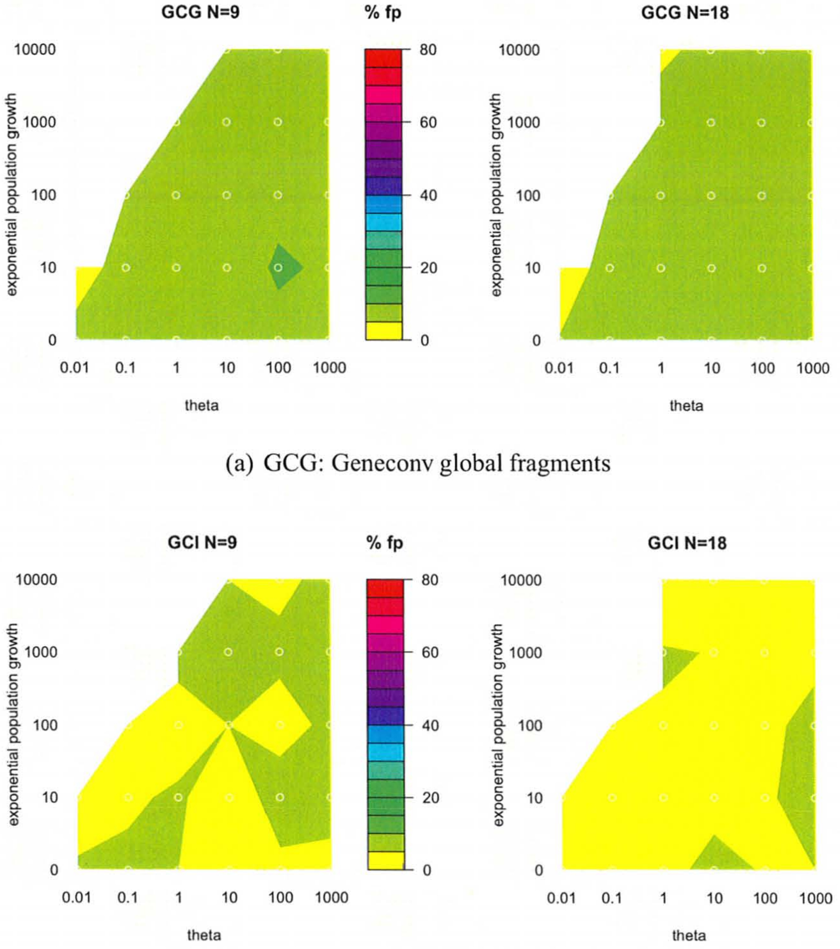
Parameters of the auto-discrete Γ model were also estimated for the third codon positions of *Mandrillus sphinx*, another cercopithecine CYTB dataset in which recombination was detected by Max χ^2 and LD|D′| (Piganeau, Gardner and Eyre-Walker, 2004). This dataset consisted of 71 sequences, 267bp long. The number of segregating sites was 29 and the nucleotide diversity was 0.0078. Although the third codon positions of this dataset had a sufficiently high correlation parameter, the extremely low nucleotide diversity of the dataset suggests the auto-discrete Γ model is insufficient to discredit the detection of recombination by Max χ^2, unless the *M. sphinx* population was undergoing population growth. This statement is based on an extrapolation of the results of Bruen, Philippe and Bryant (2006). With sample size 50, population growth of 5000N individuals/generation, and extreme substitution rate heterogeneity ($\alpha = 0.1$), Max χ^2 finds approximately 20% false positives at a nucleotide diversity of 0.05 (Bruen, Philippe and Bryant, 2006). However, differences in the *M. sphinx* dataset from the Bruen, Philippe and Bryant (2006) simulations make extrapolating the level of false positives difficult. The diversity of the *M. sphinx* sequences is much lower than even the lowest diversity sequences tested by Bruen, Philippe and Bryant (2006), suggesting *M. sphinx* is less



(a) GCL: Geneconv local fragments

(b) GCLI: Geneconv local inner fragments

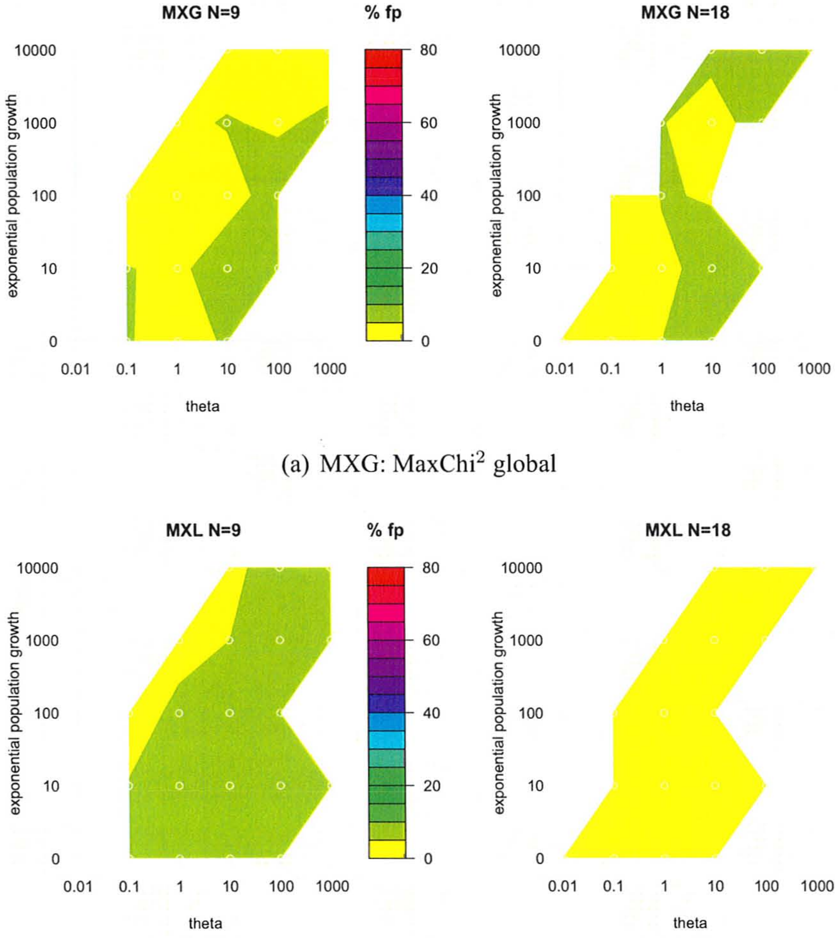
Figure 2.1: GENECONV local under two test criteria with sample size (N) 9 and 18. Colours correspond to percent false positives detected out of 1000 simulations. Grey circles demarcate batches of 1000 simulations.



(a) GCG: Geneconv global fragments

(b) GCI: Geneconv global inner fragments

Figure 2.2: GENECONV global under two test criteria with sample size (N) 9 and 18. Colours correspond to percent false positives detected out of 1000 simulations. Grey circles demarcate batches of 1000 simulations.



(a) MXG: MaxChi^2 global

(b) MXL: MaxChi^2 local

Figure 2.3: `Max` χ^2 global and local tests for sample size (N) 9 and 18. Colours correspond to percent false positives detected out of 1000 simulations. Grey circles demarcate batches of 1000 simulations.

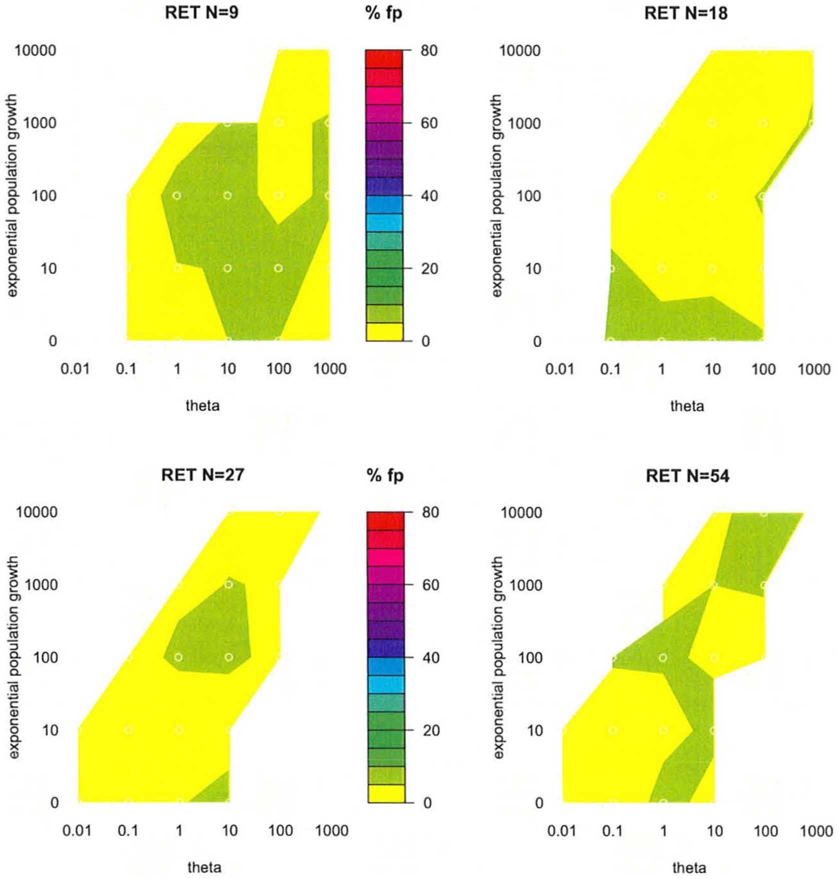


Figure 2.4: RET: `Reticulate` with sample size (N) 9, 18, 27, and 54. Colours correspond to percent false positives detected out of 1000 simulations. Grey circles demarcate batches of 1000 simulations.

Table 2.5: Summary of shape parameters (α) and correlation parameters (ρ) for *Macaca nemestrina* CYTB.

	All codon positions	
	α	ρ
Minimum	0.000401	-0.65612
1st Quartile	0.183802	-0.20423
Median	0.233785	-0.12854
Mean	0.246014	-0.12408
3rd Quartile	0.294702	-0.05018
Maximum	2.034663	1.00000
	3rd codon position	
	α	ρ
Minimum	0.3828	-0.999925
1st Quartile	0.4662	-0.587625
Median	28.0106	-0.005544
Mean	57.5077	-0.003252
3rd Quartile.	107.7423	0.580944
Maximum	199.9943	0.999998

Table 2.6: Summary of shape parameters (α) and correlation parameters (ρ) of *Mandrillus sphinx* collected.

	1st codon position		2nd codon position		3rd codon position	
	α	ρ	α	ρ	α	ρ
Minimum	0.01823	-1.0000	0.03393	-1.0000	0.05392	-1.00000
1st Quartile	0.04856	-0.9984	0.06859	-0.9994	0.34915	-0.99998
Median	0.06164	0.9988	0.14412	0.9977	0.37361	-0.99987
Mean	0.08133	0.3736	0.16478	0.1911	0.35410	-0.09876
3rd Quartile	0.10191	0.9997	0.25840	0.9999	0.39490	0.99994
Maximum	0.32725	1.0000	0.38107	1.0000	0.50884	1.00000

likely to produce false positives for recombination. However, the sample size of *M. sphinx* is almost 50% larger than the largest simulation sample size; larger sample sizes, with population growth, produced more false positives in Bruen, Philippe and Bryant (2006). Finally, Bruen, Philippe and Bryant (2006) tested linear population growth of 5000N individuals/generation. They demonstrated that higher growth rates will produce more false positives for Max χ^2 and Reticulate (but not LDr^2 or LD|D$'$|). With their highly diverged mtDNA populations, *M. sphinx* will appear to be undergoing exponential population growth, but the exact magnitude of the growth rate is unknown.

2.5 Discussion

With the exception of GENECONV Local, unique aspects of macaque population structure such as exponential population growth, migration, rate heterogeneity, or sample size did not produce elevated false positives. This is consistent with previous studies on Max χ^2 and Reticulate under linear population growth models (Wiuf, Christensen and Hein, 2001).

2.5.1 GENECONV Local

GENECONV Local conducts pairwise comparisons to find unusually long regions (or fragments) of high similarity. GENECONV analysis cannot run unless there is an adequate level of polymorphisms in the data. If π is low, there will be too few polymorphisms in most, if not all, pairs of sequences for GENECONV Local to analyze. As a result, GENECONV Local reports an extremely low level of false positives. When π is significantly high, GENECONV Local often detects an unreasonably high level of recombination.

In the survey of θ and β parameter space, GENECONV Local produced elevated false positives in two regions. The first, and most pronounced, is at low θ and low

β, where up to 85% false positives were returned. The greatest level of false positives occurs when there is no population growth ($\beta = 0$). As β increases, the level of false positives generally decreases. This is because β only affects the relative lengths of branches, not the total tree length. To maintain the same level of divergence (i.e. data for GENECONV), θ must be increased as well. As θ increases, the level of false positives decreases. This decrease is due to an over-saturation of substitutions in the sequences. GENECONV detects fragments of high similarity; when $\theta > 0.1$, fragments become shorter and less statistically significant. If one continues to increase θ, one observes a pronounced increase in the false positives for GCL, but a diminished or absent increase in GCLI. This indicates the second increase in false positives is due to increased detection of outer fragments. Outer fragments are evidence of gene conversion with a sequence that cannot be identified within the alignment. An outer fragment is identified when a sequence differs from other sites in the alignment (or triplet) at all (or nearly all) polymorphic sites. To create this signal without recombination would require many substitutions, and highly diverged sequences. In other words, more false positives should occur at high θ and high β, which is consistent with the results presented.

Sample size, N, affects the level of false positives, particularly at high levels of divergence. For example, when outer fragments are included, a larger dataset will produce many more false positives than a smaller one. However, when outer fragments are excluded, the smaller dataset produces more false positives than the larger one. Since θ is proportional to the product of N and mutation rate, for a given θ, three randomly selected sequences from the $N = 9$ population will, on average, be more diverged than three randomly selected sequences from the $N = 18$ population. Our results suggest there is a trade-off between the number of false positives returned by inner versus outer fragments. If sequences are very diverged, mismatches will outnumber matches, and fewer inner fragments will be detected. However, if mismatches involve unique bases, more outer fragments will be detected. If there are an extreme number of multiple mutations at a site such that mismatches do not involve unique bases, outer fragments will not be detected. When $N = 18$, the level of pairwise divergence is too high for inner fragments to be detected, but just high enough that mismatches involve unique bases. When $N = 9$, there are fewer unique bases and therefore fewer outer fragments, but enough mismatches that inner fragments begin to appear significantly similar in comparison. This last point is only possible because of the relaxed mismatch penalty parameter in GENECONV Local.

Although local tests have been used as means of detecting recombination in animal mtDNA (Tsaousis *et al.*, 2005), they are more so intended for identification of recombining sequences only after recombination has been detected via other more appropriate methods (Sawyer, 1999).

2.5.2 Auto-correlated rate heterogeneity

Whether an auto-discrete Γ model can discredit the detection of recombination in *M. sphinx* CYTB is unclear. Simulations from Bruen, Philippe and Bryant (2006) with a similar level of diversity as this dataset produced approximately 10% false positives in Max χ^2 and more than 50% false positives in Reticulate when $\alpha = 0.1$. When $\alpha = 1$, the level of false positives for both tests was less than 10%. Whether the *M. sphinx* data can be expected to produce similar levels depends on the relative importance of α and sample size. For example, the larger α of *M. sphinx* suggests the level of false positives in *M. sphinx* data would be less than 10% for Max χ^2 and less than 50% for Reticulate. Conversely, the larger *M. sphinx* sample size suggests the level of false positives can be extrapolated to be greater than the levels reported in Bruen, Philippe and Bryant (2006). Being a cercopithecine, the *M. sphinx* population is also structured by female philopatry, producing highly diverged populations that mimic exponentially growing populations. Based on the data of Bruen, Philippe and Bryant (2006), the level of false positives expected of *M. sphinx* CYTB, given sequence diversity, sample size, and population growth is at least 50% for Reticulate and at 20 - 40% for Max χ^2.

2.5.3 Implications of recombination

Although GENECONV Local is an unreliable test for recombination, the other four tests did not produce a significant level of false positives. These results demonstrate that macaque models of population structure and diversity, and general models of exponential growth do not produce most patterns of recombination. We therefore do not find convincing evidence against reports of recombination in widespread animal mtDNA.

Since most mutations are mildly deleterious, mtDNA is expected to accumulate many deleterious mutations over time. Without recombination, this increased genetic load would lead to an inevitable genetic melt-down. Some have suggested bottlenecks (Bereiter-Hahn and Vöth, 1994) and within-generational drift (Takahata and Slatkin, 1983) as mitochondrial alternatives to recombination, but exactly how mtDNA overcomes this issue remains unclear. If recombination is a common occurrence in animal mtDNA, an important aspect of mitochondrial evolution may be resolved. However, this creates issues in the use of mtDNA in population genetics studies, as virtually all animal population genetics models assume clonal inheritance.

Evidence for recombination in human mtDNA was taken by some to signal the necessity for a reevaluation of mtDNA studies, as the fundamental assumption of clonal inheritance had been violated (Awadalla, Eyre-Walker and Maynard-Smith, 1999). Although this recombination signal has since been determined to be an artifact of sampling size (Innan and Nordborg, 2002), it raised the question of how recombination would change the field of population genetics.

With recombination, different sites in the mtDNA genome may have different evolutionary histories. Generally, however, uniparental inheritance and vegetative segregation virtually guarantee all mtDNA within a cell are identical. The mtDNA of a recombining homoplasmic cell will be indistinguishable from mtDNA of a non-recombining homoplasmic cell. In other words, homoplasmy makes recombination evolutionarily arbitrary because clonal inheritance is maintained. Another argument against a re-evaluation of population genetics is that even if mtDNA sites have different histories, the histories of these sites will be closely correlated. This argument assumes recombination is rare, and occurs between closely related species. The first assumption, that recombination is rare, is doubtful given the detection of recombination in about 15% of animal mtDNA alignments tested (Tsaousis *et al.*, 2005; Piganeau, Gardner and Eyre-Walker, 2004). The second assumption is more promising. Since proper function of the mitochondria depends on both mtDNA and nuclear genes, recombination may occur between closely related species if these species have fewer incompatibilities between the mitochondrial and nuclear genomes.

Recombination becomes an issue for population genetics when mtDNA of different lineages are present in the same cell. This can occur through intra-species introgression, i.e. paternal leakage (Kraytsberg *et al.*, 2004), or interspecies introgression (Niki, Chigusa and Matsurra, 1989; Bernatchez *et al.*, 1995) through hybridization. Hybrid contact zones exist between all parapatric Sulawesi macaques (Watanabe *et al.*, 1991; Watanabe and Matsumura, 1991; Ciani *et al.*, 1988; Bynum, Bynum and Supriatna, 1997; Evans, Supriatna and Melnick, 2001); that is, all Sulawesi macaques included in this study except *M. brunescens*. In Evans, Supriatna and Melnick (2001), the genetic structure of the hybrid contact zone of *M. maura* and *M. tonkeana* was studied. Outside the hybrid contact zone, mtDNA of *M. maura* and *M. tonkeana* are monophyletic. Within the hybrid contact zone, however, mtDNA clusters with a parental group, *M. maura* or *M. tonkeana*, or with a third unresolved group. If recombination is occurring in Sulawesi macaques, as is supported by the results of this chapter, the third unresolved group may be an example of macaques with recombinant mtDNA, where loci with incongruent genealogies are leading the separation of recombinants from parentals.

Chapter 3

Mutation cold spots produce recombination false positives

3.1 Abstract

Indirect tests have detected recombination in diverse animal mitochondrial DNA (mtDNA), including mammals. Although the molecular patterns detected by these tests could alternatively be explained by mutation cold spots, or clusters of sites with low mutation rates, the effect of mutation cold spots on false positive rates has not been adequately explored. The false positive rates of six indirect tests for recombination were characterized using simulations of general models of mtDNA evolution with mutation cold spots but no recombination. All tests produced a high level of false positives under a general model, although the conditions producing the maximal level of false positives differed between tests.

3.2 Introduction

Clonal inheritance of animal mtDNA (Birky *et al.*, 2005) is an important assumption for many population-level studies. The generality of this assumption has been questioned with reports of recombination in various animal species, including bivalve mussels, crustaceans, amphibians (Ladoukakis and Zouros, 2001), lizards (Ujvari, Dowton and Madsen, 2007), scorpions (Gantenbein *et al.*, 2005), fish (Ciborowski *et al.*, 2007), birds, insects, nematodes, and many mammals, including non-human primates (Piganeau, Gardner and Eyre-Walker, 2004; Tsaousis *et al.*, 2005; Maynard Smith and Smith, 1999; White and Gemmell, 2009; Ladoukakis and Zouros, 2001) and humans (Kraytsberg *et al.*, 2004). Due to low levels of mtDNA heteroplasmy, most of these studies have relied upon "indirect" tests for recombination wherein recombination is inferred from patterns of molecular variation as opposed to direct tests for recombination where non-recombined parental sequences are compared to potentially recombined sequences from an offspring individual or cell. These statistical tests detect signals left behind by recombination events, such as an uneven distribution of polymorphic sites (Maynard Smith, 1992; Posada and Crandall, 2001), regions with high sequence similarity (Sawyer, 1989), clustering of phylogenetically incompatible sites (Jakobsen and Easteal, 1996), or a high correlation of linkage disequilibrium with physical distance (Piganeau, Gardner and Eyre-Walker, 2004). When collecting indirect evidence for recombination, it is generally recommended that more than one test be used as the power of indirect tests can vary widely under different conditions (Posada and Crandall, 2001; Posada, 2002; Wiuf, Christensen and Hein, 2001). Six of the most widely used and powerful indirect tests for recombination (Posada and Crandall, 2001; Posada, 2002; Wiuf, Christensen and Hein, 2001; Bruen, Philippe and Bryant, 2006) are described below.

GENECONV

GENECONV (GC) (Sawyer, 1989) searches for evidence of gene conversion, a nonreciprocal form of recombination where a locus is copied over a locus in another sequence, by searching the alignment for highly similar pairs of sequences. To do this, GENECONV first removes all invariant sites from the alignment, leaving only polymorphic sites for consideration. Each pair of sequences is compared and regions of the sequences, or fragments, are identified which are either unusually long stretches of perfect matches, or are unusually similar. Similarity is based on a scoring system where a matching site adds 1 to the score, but a mismatch is penalized according to a user-defined mismatch-penalty.

Data under the null hypothesis of no recombination is created by permutation and the significance of each fragment's similarity score is compared to the similarity scores from the permuted data. The number of permutations, 10^4, and gscale, 0, that are used below are the same as those used in studies reporting widespread recombination in animal mtDNA (Piganeau, Gardner and Eyre-Walker, 2004; Tsaousis *et al.*, 2005).

In GENECONV Global (GCG) sites in the entire sequence alignment were permuted and fragments are compared to the fragments between all possible sequence pairs. The GENECONV Global test uses a built-in correction to correct for the multiple fragment comparisons between sequence pairs. A global p-value is assigned to the fragments using a modified BLAST scoring system that corrects for the evolutionary distance of sequence pairs (Altschul, 1993). Assuming distantly related sequence pairs have more mismatches than closely related sequence pairs, a significant fragment from distantly related sequences will be less significant when compared to significant fragments from closely related sequences which contain fewer mismatches, and hence, longer fragments. GENECONV Global removes this bias by correcting fragment lengths and similarities by the evolutionary distance of sequence pairs.

GENECONV Global analysis considered two types of fragments: inner and outer fragments. Inner fragments are interpreted as evidence for gene conversion between ancestors of sequences present in the alignment. Outer fragments are considered evidence for gene conversion between ancestral sequences when one descendant sequence cannot be identified. An ancestral sequence cannot be identified if it is missing from the alignment, or if subsequent mutations and/or gene conversions have distorted the original gene conversion sequence beyond recognition. Scoring outer fragments is analogous to scoring inner fragments, where a site containing a unique base is like a match, and 1 is added to the score, but a site containing a non-unique base is like a mismatch, and is penalized. Outer fragments can also be identified between sequence pairs when two sequences share a region, bounded on at least one side by a matching polymorphic site, where all sites are mismatches.

In this study, cases when both inner and outer fragments were considered are abbreviated as GCG, and cases when only inner fragments were considered are abbreviated as GCI. While some studies include outer fragments in the GENECONV criteria (Tsaousis *et al.*, 2005), others specifically do not (Piganeau, Gardner and Eyre-Walker, 2004).

Max χ^2

Max χ^2 is a test to identify recombination breakpoints within a sequence alignment. This method was first proposed by Maynard Smith (1992), and is widely considered to be the best general test to detect recombination (Posada, 2002; Posada and Crandall, 2001).

Max χ^2 assumes that in the absence of recombination, polymorphisms are evenly distributed among sites and between sequences. Invariant sites are first removed from the alignment and a sliding window is used to test each sequence pair, comparing the number of matches and mismatches in both halves of the window; that is, testing whether the middle of the window is the location of the potential breakpoint. To determine the significance of the difference in matches and mismatches on either side of the potential breakpoint, a $2 \times 2\chi^2$ value is calculated for each window, and is compared to the χ^2 of 100 permutations. If fewer than 10 permutations had a χ^2 value greater than or equal to the observed, the permutation test was repeated with 1000 permutations. This study used the same implementation and window settings as Piganeau, Gardner and Eyre-Walker (2004): a maximum window half-size of 5 variable sites, and a sliding window step size of 2bp. The implementation is a modification of Posada and Crandall (2001) where the entire length of the alignment is analyzed, and only windows with expected χ^2 greater than 2 are considered. According to Piganeau, Gardner and Eyre-Walker (2004), these modifications improve the power of Max χ^2 while maintaining the same false positive error rate.

LDr^2 and LD$|$D$'|$

Polymorphic sites are in linkage disequilibrium when the frequency of observing alleles in the same haplotype does not match the expected haplotype frequency, which is the product of the allele frequencies. In animals, the entire mitochondrial genome is inherited as a single unit. Assuming no recombination, when a new allele arises in one mtDNA molecule, it is permanently linked to the existing alleles on that molecule. This is an example of extreme linkage disequilibrium. If recombination were to occur, it would be more likely to occur between physically distant loci where there is more space for a crossover to occur. This breaks down the physical linkage between mitochondrial loci. As the distance between two loci increases linkage disequilibrium will approach 0. How much the observed haplotype frequency differs from the expected haplotype frequency for a pair of loci is described by the linkage disequilibrium parameter, D. Since the range of D possible depends on the magnitude of the haplotype frequencies, which differs between pairs of loci, D is not a useful measure to compare the strength of linkage disequilibrium between different loci pairs. Two alternative measures of linkage disequilibrium are $|D'|$ and r^2. $|D'|$ is the normalization of D by the theoretical maximum, D_{max} and is calculated as

$$|D'| = |D|/D_{max}, \tag{3.1}$$

where D_{max} is the smaller of Ab or aB, where A and a are the allele frequencies at one diallelic locus, and B and b are allele frequencies at a second diallelic locus (Lewontin, 1964). r^2 (Hill and Robertson, 1968) is the square of the correlation between two alleles and is calculated as (McVean, 2001)

$$r^2 = \frac{D^2}{AaBb}. \tag{3.2}$$

The tests LD|D′| and LD|D′| (Piganeau, Gardner and Eyre-Walker, 2004) calculate the correlation between distance and $|D'|$ or r^2, respectively, then assign a level of significance according to a permutation test. Code provided in Piganeau, Gardner and Eyre-Walker (2004) was used to conduct the linkage disequilibrium tests LDr^2 and LD|D′|. The original program included a correction for circularity of human mtDNA sequences. This correction was removed for our analysis because our sequences are non-human and cover 3 loci at the most. In hindsight, the removal of this correction is unnecessary, but should not affect the results because the sequence lengths do not nearly approach the complete length of mtDNA.

Since r^2 has been suggested as more sensitive to allele frequencies and mutation rate variation than $|D'|$, LD|D′| is considered by some to be a better test for the correlation of linkage disequilibrium with distance (Jorde and Bamshad, 2000; Kumar *et al.*, 2000). On the other hand, LDr^2 has higher power than LD|D′| (Awadalla, Eyre-Walker and Maynard-Smith, 1999; White and Gemmell, 2009). Neither is sensitive to false positives due to sites with highly correlated rates (Bruen, Philippe and Bryant, 2006).

`Reticulate`

`Reticulate` (Jakobsen and Easteal, 1996) uses the Neighbour Similarity Score to detect significant clustering of incompatible informative sites. Informative sites are sites where the alignment contains at least 2 alleles, each of which is present in at least 2 samples. In this study, sites with more than 2 states were ignored. For each pair of informative sites the most parsimonious tree is generated. If the tree created from the two sites does not contain any recurrent or convergent mutations, or homoplasies, the two sites are deemed compatible. Conversely, if the sites cannot produce the same tree without invoking homoplasy, they are labelled incompatible. A compatibility matrix is created where each row and column corresponds to an informative site in order that they appear in the alignment. The Neighbor Similarity Score (NSS), or the average number of adjacent incompatible or compatible cells, is calculated for the matrix. Sites in the alignment were permuted 1000 times to calculate the significance of NSS. This method works best on relatively large alignments with well-diverged sequences and, according to the documentation, can detect recombination that changes tree topology but cannot detect recombination that only changes branch lengths.

PHI

PHI (Bruen, Philippe and Bryant, 2006) follows a similar method as Reticulate but with some important modifications. When comparing pairs of informative sites, Reticulate only considers two possibilities: the sites are either compatible or incompatible. All incompatible sites weighted the same, regardless of how many homoplasies are present on their respective trees. In contrast, PHI differentiates between incompatible sites based on their refined incompatibility score, or the minimum number of homoplasies on the tree inferred from the two sites. PHI uses the mean refined incompatibility score for sites up to w bases apart, Φ_w, as the test statistic. The probability of observing Φ equal to or less than Φ_w can be estimated using a permutation test. However, for computational efficiency, the distribution of Φ from permutations is approximated by assuming the distribution of permutation Φ values is normal; an assumption supported by simulation data (Bruen, Philippe and Bryant, 2006). Under this assumption, the probability of observing equal or fewer homoplasies than Φ_w by chance is calculated using a normal probability distribution function. In this study, as in Bruen, Philippe and Bryant (2006), w was set to 100.

3.2.1 Modeling Sulawesi macaque mtDNA

We explored the effect of uneven rates of evolution in a subset of closely related individuals and in a subset of their sequences. Hereafter we refer to this pattern of uneven rates as “heterotachy”. Heterotachy was first explored using simulations built from a general model of mutation cold spots, or clusters of sites with low mutation rates, shared by a subset of samples.

We begin with a detailed look at the evidence for recombination in mtDNA of Sulawesi macaques. Species within this group share a unique population structure that makes them an interesting model for potential recombination false positives, such as extreme population structure and high mutation rate (Brown, George and Wilson, 1979). However, we found that these characteristics are insufficient to produce elevated false positives for recombination (Chapter 2). Instead, we found evidence to support the presence of a mutation cold spot within a geographic sub-set of the macaque dataset. Using a general model of evolution, we describe the conditions under which mutation cold spots produce an elevated level of false positives from indirect recombination tests.

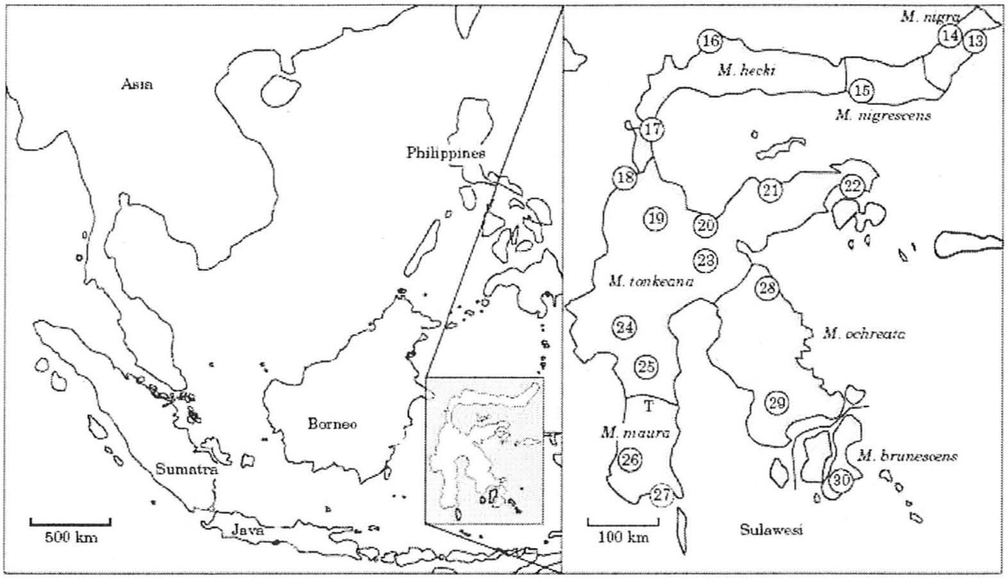


Figure 3.1: Map of Sulawesi macaque samples, adapted from Figure 1 in Evans *et al.* (1999). Samples are numbered as: *M. nigra*; 13-14, *M. nigrescens*; 15, *M. hecki*; 16-17, *M. tonkeana*; 18-25, *M. maura*; 26-27, *M. ocreata*; 28-29, *M. brunescens*; 30. Samples 13-25 constitute the *N/C* group, while samples 26-30 constitute the *So* group.

3.3 Methods

3.3.1 Which macaques are recombining and where?

Eighteen macaque mtDNA sequences collected from the Indonesian island of Sulawesi (Evans *et al.*, 1999) were tested for recombination using GENECONV, Max χ^2, and Reticulate. Since simulations were performed under a codon model, all tRNA sequences were removed such that the final sequences contained only coding sequence; two complete sequences (ND3 and ND4), and one partial sequence (ND4L). An invariant 9bp region containing a 7bp overlap between ND4 and ND4L was also removed. The final sequences were 1203bp long.

The 18 macaque mtDNA sequences were sampled broadly across the island of Sulawesi and represent 7 macaque species (Figure 3.1). Thirteen sequences were sampled from the North and Central regions of Sulawesi, and represent *M. nigra, M. nigrescens, M. hecki*, and *M. tonkeana*. From here onward these sequences will be collectively referred to as the *N/C* group. The five remaining sequences were sampled from Southern Sulawesi, and represent *M. ochreata, M. brunnescens*, and *M. maura*. These sequences will be referred to as the *So* group.

GENECONV was used to identify the nucleotide region and sequences seemingly undergoing gene conversion. GENECONV was used to test: the complete dataset, only the Northern and Central sequences, and only Southern sequences.

Max χ^2 was used to find the most significant breakpoint in the dataset. Max χ^2 tests the complete length of the sequence, but only reports the most significant breakpoint and the pair of sequences between which the breakpoint occurred. To find other significant breakpoints and to determine whether one sequence in particular was responsible for the recombination signal, one sequence from this pair was removed and the remaining sequences were re-tested. An additional sequence was then removed, and the remaining were re-tested, and so on. This was repeated until significant breakpoints were no longer detected.

`MrBayes` was used to build the phylogenies on either side of the most significant breakpoint. Analysis ran for 5×10^6 generations, with the first 10^6 generations discarded as the burnin. A visual check of the likelihoods showed that they had stabilized, and the potential scale reduction factor (PSRF) convergence diagnostic equalled 1.00. Based on these two criteria, stationarity was assumed to have been reached. A GTR+Γ+I+codon position model of evolution was specified.

3.3.2 Mutation cold spots and the distribution of segregating sites

A distribution of trees and tree parameters were estimated from the Sulawesi dataset using `MrBayes` (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) under a GTR+Γ+I+codon position model. An outgroup was specified to facilitate branch length manipulation in the cold spot simulations. A South macaque (*So*) was specified as the outgroup for this analysis. Selection of the outgroup was arbitrary except that the outgroup could not contain a cold spot; that is, the outgroup could be any *So* sequence except sample 704ochrea. *M. brunescens* (sample 707brune) was selected as the outgroup. As before, analysis ran for 5×10^6 generations, and the first 10^6 generations were discarded as the burnin. A visual check of the likelihood and PSRF suggested stationarity had been reached. Parameters from a GTR+Γ+I+codon position model of evolution were randomly selected from the `MrBayes` output. These parameters included a tree plus the following parameters as estimated from each codon position: a GTR substitution rate matrix, base frequencies, the proportion of invariant sites, and shape parameter for Γ-distributed site-specific rate heterogeneity.

To include a mutation cold spot in the simulation, a copy of the randomly selected tree was created and modified to model the cold spot region; internal and terminal branches of the copied tree were scaled by a "cold factor" of either 0.05, 0.2, 0.5, or 1. The cold factor represents the rate of evolution of the cold region relative to non-cold regions. For example, a cold factor of 0.05 means the substitution rate in the cold spot is reduced to 20% relative to the rest of the sequence, and a cold factor of 1 is equivalent to no cold spot. The `GENECONV` results suggested a cold spot was present in all *N/C* sequences and one *So* sequence, so for simplicity internal and terminal branches of the *N/C* clade were cooled, excluding the single *So* sequence from the cold clade. `GENECONV` also suggested that the cold spot was located over the middle third (400bp) of the sequence. Therefore, this modified cold tree was used to simulate the cold middle 400bp of the alignment, while the unmodified tree was used for non-cold regions. As these two trees possessed different branch lengths but identical topology, these simulations did not include the phenomenon of recombination. Simulation of nucleotide datasets off a single set of model parameters and two trees was completed using Seq-Gen (Rambaut and Grassly, 1997) and Seq-Gen's partition option.

The newly-generated dataset was retained if and only if it contained the same number of segregating sites as the natural data upon which it was based: 296 segregating sites in the Sulawesi macaque dataset. The simulation process from parameter selection to dataset simulation was repeated until a collection of 1000 simulated datasets, each containing exactly 296 segregating sites, was obtained. In total, 4 datasets containing 1000 simulations each were collected. Each dataset corresponded to a different cold factor tested.

In order to evaluate the degree to which observed sequences contained or did not contain a cold spot, we developed a metric based on the variance in the number of segregating sites in non-overlapping 100bp windows. For each simulated dataset, the variance in number of segregating sites between windows was counted and used to form a distribution of expected variance. The Sulawesi macaque variance (63.7197) was compared against these distributions. The number of simulations with a variance greater than or equal to the Sulawesi macaque variance is measure of the likelihood the Sulawesi data contains a cold spot with that distribution's cold factor.

3.3.3 Mutation cold spots produce recombination false positives.

Simulated nucleotide datasets were used to investigate the effect of mutation cold spots, or local regions with low mutation rates, on the level of false positives returned by indirect tests for recombination. Various cold spot characteristics were considered: coldness, or the mutation rate relative to non-cold regions; length of the cold spot, in proportion to total sequence length; size of the cold clade, or proportion of the total dataset sharing the cold spot; and location of the cold spot within the alignment, for example, in the middle surrounded by non-cold regions, or alternatively, at the periphery.

The main focus in this section was to explore how mutation cold spots affect the level of false positives. Unlike previous sections where parameters were estimated from real datasets, here, a more general model of evolution was used, and parameters were deliberately selected so a range of dataset characteristics could be explored. These parameters are summarized in Table 3.1.

The creation of a simulation, or nucleotide dataset, began by generating a random tree topology in ms with N sequences. From this tree a clade of n sequences was chosen to be the clade containing the cold spot, or the "cold clade". If no such clade existed or if this clade included the root, a new tree would be generated and checked. Once a tree with the desired clade size was found, a copy of the tree was created and modified to include a mutation cold spot. This was achieved by scaling the internal and terminal branches of the cold clade by the cold factor, c. Both the cold tree and unmodified tree were used to create a nucleotide alignment in Seq-Gen using a Jukes-Cantor model of evolution. The position, p, of the cold spot within the alignment was specified by splitting the alignment into sections, each of which was built off either the cold tree or unmodified tree. The section built off the cold tree was the section where the cold spot could be found in the cold clade. The

Table 3.1: Simulation parameters used in section 3.3.3.

Symbol	Description	Values
N	Number of sequences/alignment	15, 30
L	Sequence length	600bp, 1200bp
n	Number of sequences with cold spot	$\frac{1}{3}N$, $\frac{2}{3}N$
l	Length of cold spot	$\frac{1}{3}L$, $\frac{2}{3}L$
c	Cold factor	0.05, 0.2, 0.5, 1.0
p	Position of cold spot within alignment	middle, side
S	Number of segregating sites	$0.05L$, $0.1L$, $0.2L$, $0.5L$

nucleotide dataset was then screened for the total number of segregating sites, *S*. Unless the number of segregating sites exactly matched the target *S*, the entire process was repeated, beginning at the generation of tree topology. Finally, the dataset was tested for recombination using 6 tests: Reticulate, PHI, GENECONV with and without outer fragments, Max χ^2, LDr^2, and LD|D'|.

3.4 Results

3.4.1 Which macaques are recombining and where?

Both GENECONV and Max χ^2 conduct pairwise sequence analysis. With a dataset of 18 sequences there are 153 pairwise comparisons. Of the 153 pairs to compare, 50.98% will be pairs of North and Central sequences (*N/C - N/C*), 42.48% are combinations of North and Central with South sequences (*N/C - So*), and 6.53% are pairs of South sequences (*So - So*). These are the proportions expected to be detected by GENECONV and Max χ^2 if recombination is occurring broadly across Sulawesi, and will be helpful in determining whether recombination is occurring between all species, or if it is limited to only certain species.

Out of the 153 pairwise comparisons, GENECONV detected gene conversion in 13 pairwise comparisons (Table VI). With only one exception, gene conversion was detected between pairs of North/Central macaques (*N/C - N/C*). Gene conversion was also detected between a North/Central and Southern macaque (*N/C - So*).

These results suggest recombination occurred primarily between macaques from the North and Central regions of Sulawesi. To test this hypothesis all Southern macaque sequences were removed and the dataset was retested. In the North and Central only dataset, GENECONV detected a single global inner fragment (661nigra and 597tonkma) at a p-value of 0.0070; a value at least two orders of magnitude greater than the p-value reported for the same pair when the South sequences were included. In other words, recombination is still detected between macaques from North and Central Sulawesi, but the strength of this signal is much reduced when Southern Sulawesi sequences are not included in the analysis.

GENECONV detected gene conversion almost exclusively between *N/C - N/C* macaque pairs. This was confirmed visually by graphs of segregating sites (*Ss*) versus nucleotide position (Figure 3.2). *N/C* - *N/C* pairs of mtDNA contain a region of about 400bp which contains very few segregating sites. This potential "mutation cold spot" is followed by a section with an elevated number of segregating sites, which is a potential mutation hot spot. The cold spot is the fragment that GENECONV detects as gene conversion, and the border from cold spot to hot spot is roughly where Max-Chi detects breakpoints (Figure 3.2). A search in the Conserved Domain Database revealed this cold spot corresponds to the amino terminus of NADH ubiquinone oxidoreductase (complex I), a domain involved in electron transfer from NADH to ubiquinone.

Max χ^2 determined the most significant breakpoint to be at nucleotide position 901 between 661nigra and 648heckin, two macaques from the North/Central region of Sulawesi (*M. nigra* and *M. hecki*). To find other significant breakpoints in the dataset as well as to determine whether one sequence was responsible for the recombination signal, the dataset was jackknifed and re-tested. Of the 12 sequences that were removed (the first sequence names in the sequence pairs of Table VI), all but 2 were included in the recombination pairs found by GENECONV. Samples 648heckin and 655nigres were not included in any recombination pairs from GENECONV. Only two pairs of sequences found by Max χ^2 also showed evidence for recombination according to GENECONV. However, the list of sequence pairs found by Max χ^2 is not exhaustive, since the sequences Max χ^2 can test is affected by the order of sequence removal.

MrBayes was used to build the phylogenies on each side of the most significant breakpoint (Figure A.1). These trees do not support the Max χ^2 suggestion that samples 661nigra and 648heckin are recombining, but they do suggest recombination between samples 661nigra and wf128tonk.

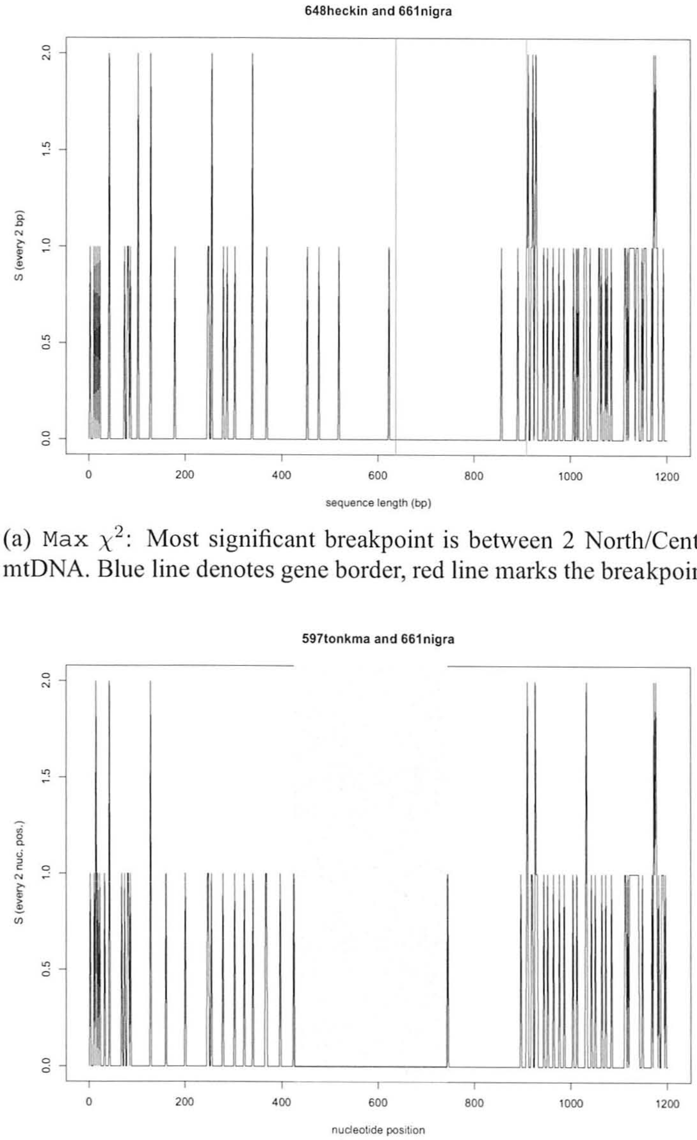
3.4.2 Mutation cold spots and the distribution of segregating sites

Although simulations matched the Sulawesi data on *Ss*, the simulation π was slightly higher than the observed. The observed Sulawesi π is 7%, and the simulation π ranged from 10.1% to 11% from no cold spot to a cold factor of 0.05, respectively.

Including mutation cold spots in the simulations increases the spread in the variance distributions (Figure 3.3). The variance in the number of polymorphic sites of the simulated data matched the Sulawesi data only with the inclusion of a mutation cold spot.

When the data does not contain a cold spot, there is < 0.001 probability that the Sulawesi macaque mtDNA variance will be observed (Figure 3.3). However, the probability of observing the Sulawesi macaque mtDNA variance increases to 0.09 when the substitution rate in the middle third of the North/Central clade is "cooled" or scaled by a factor of 0.05. In other words, this cold-spot model more accurately describes the distribution of polymorphisms in the Sulawesi macaque mtDNA than a model without a cold spot.

Based on the cold factors tested, the substitution rate within the cold spot is most



(a) `Max` χ^2: Most significant breakpoint is between 2 North/Central mtDNA. Blue line denotes gene border, red line marks the breakpoint.

(b) `GENECONV`: Most significant fragments are between 2 North/Central mtDNA. The fragment is highlighted in grey.

Figure 3.2: Recombination in Sulawesi macaque mtDNA visualized

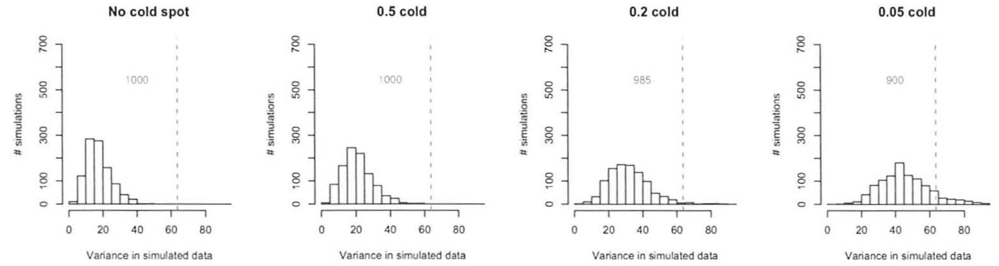


Figure 3.3: Variance in segregating sites along length of simulated sequences. Blue line marks Sulawesi macaque mtDNA variance. Red values are number of simulations, out of 1000, that have a variance equal to or less than the Sulawesi macaque mtDNA variance.

likely 0.05 times the rate outside the cold spot. However, this estimate is affected by the numerous factors and assumptions. The most obvious factor is the choice of possible cold factors. Here, only 4 were tested: 1 (no cold spot), 0.5, 0.2, and 0.05. In addition, certain assumptions were made regarding the number, length, location, and taxa size of the cold spot; the effect of one cold spot was modeled, 400bp long, situated in the middle of the alignment, and shared by the 13 out of 18 sequences constituting the North/Central clade. In other words, the main objective in this section was not to quantify the coldness of the cold spot, but to show that with all else being equal, cold spots are a plausible explanation for the recombination signal in Sulawesi macaque mtDNA.

3.4.3 Mutation cold spots produce recombination false positives.

The results of these studies are displayed in Figures 3.4-3.9. The size of the cold clade is included in brackets as (proportion of total taxa, length). The coloured contour map corresponds to the percentage of simulations positive for recombination, and was interpolated using the simulation data. In Figures 3.4-3.5, all sequences are 1200bp long with a cold spot in the middle third of n sequences. The strength of the cold spot, on the y-axis as the cold factor, is the substitution rate/site of the cold spot to non-cold spot. In Figures 3.6-3.7, all sequences are 1200bp long with the last third cold in n sequences (i.e. cold spot on the side), and in Figures 3.8-3.9 all sequences are 600bp long with the middle third cold in n sequences.

Without a cold spot, the tests returned approximately 5% false positives. The likelihood of the tests reporting a false positive increased as the cold spots became colder. This was particularly true when the sequences contained a high proportion of polymorphic sites. When cold spots were included, all tests produced an elevated level of false positives (34%, up to 99%), but not necessarily under the same conditions. The neighbour similarity tests, `PHI` and `Reticulate`, performed best, finding fewer false positives than the other tests, and in fewer cases. Nevertheless, these tests still returned a significant level of false positives (30% to 80%) when the cold spot covered a large area (two-thirds of the sequence length) and the cold

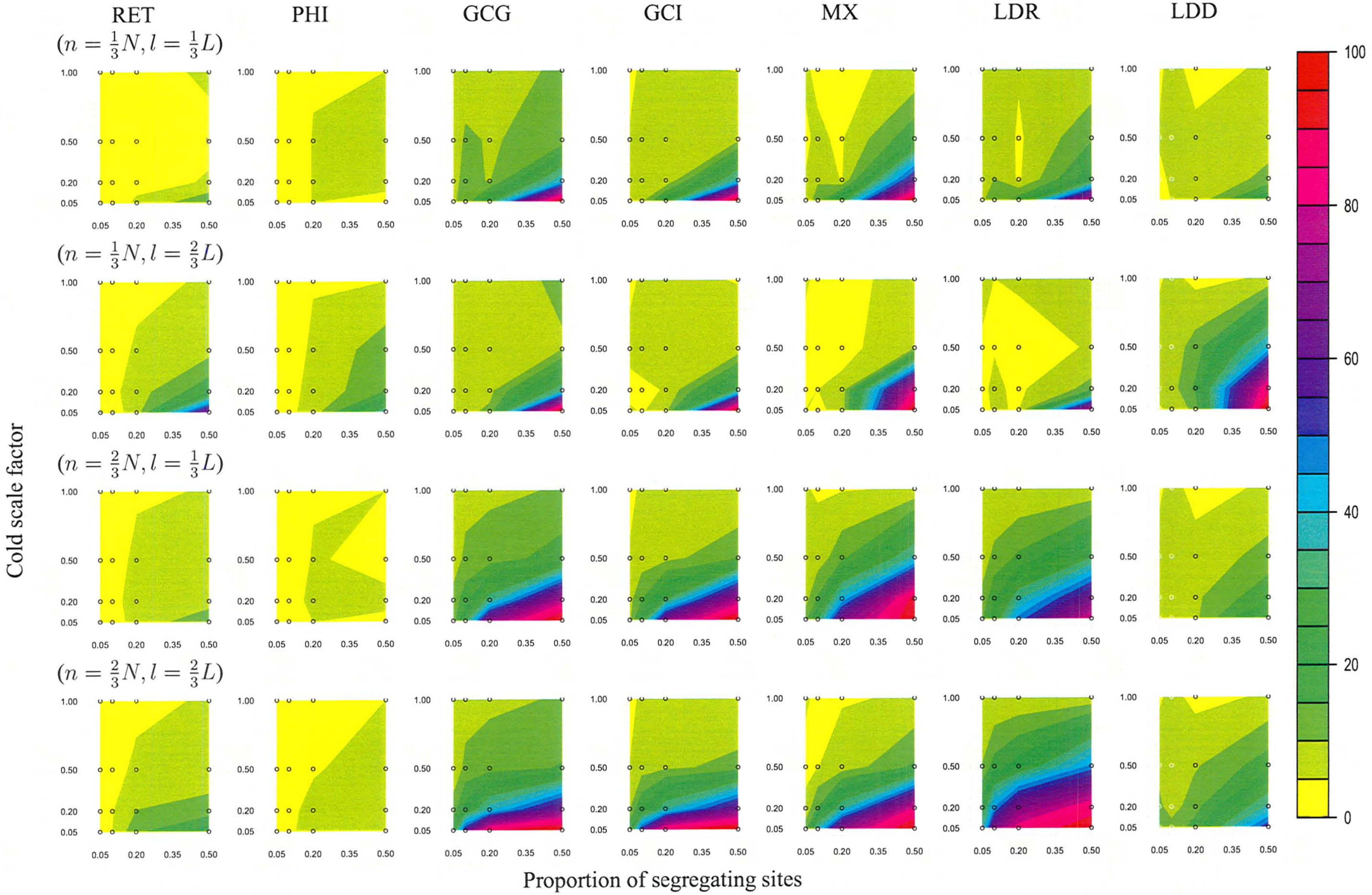


Figure 3.4: False positives in cold spot simulations: $N = 15, L = 1200bp, p =$ middle. n = number of sequence in cold clade, l = length of cold spot. Black circles represent 800 - 1000 simulations which were tested for recombination, and grey circles mark data collected from fewer than 800 simulations.

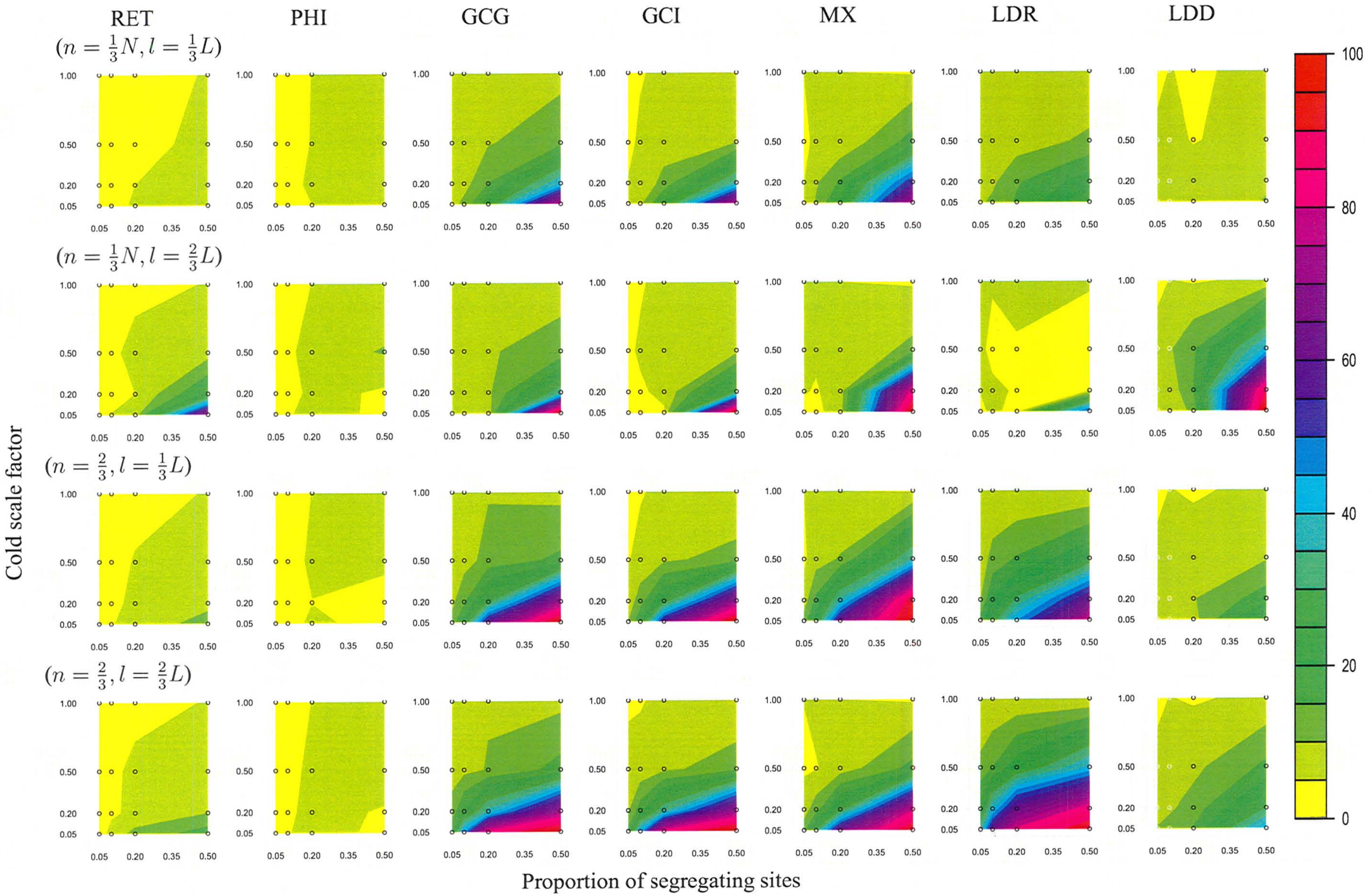


Figure 3.5: False positives in cold spot simulations: $N = 30, L = 1200bp, p =$ middle. n = number of sequence in cold clade, l = length of cold spot. Black circles represent 800 - 1000 simulations which were tested for recombination, and grey circles mark data collected from fewer than 800 simulations.

clade was small (one-third of total sample size). The polymorphism distribution tests (GENECONV, Max χ^2) performed the worst, detecting over 30% false positives in all cases. The GENECONV criteria of considering only inner fragments (GCI) is considered more conservative than the alternative of considering all global fragments detected (GCG). However, as is shown here, GCI finds the same high level of false positives as GCG. Although LDr^2 is a more powerful test than LD|D$'$| to detect recombination when it is present (White and Gemmell, 2009), LDr^2 found many more false positives than LD|D$'$|, suggesting the gain of power comes with a cost of accuracy.

The two most important factors in the level of false positives produced are the number of segregating sites (but not the simply proportion), and the coldness of the cold spot. Surprisingly, doubling the number of sequences in the dataset did not significantly improve the performance of the tests.

The level of false positives of larger datasets (30 sequences) was equal to or less than the level of false positives of smaller datasets (15 sequences). This contradicts a previous study, where an increase in number of sequences from 10 to 50 led to an increase in level of false positives detected, from 10% to over 50% in PHI, Reticulate, and Max χ^2 (Bruen, Philippe and Bryant, 2006). This could be due to additional factors included in the previous model, including exponential growth, extreme site-specific rate heterogeneity, and the method of simulating mutation hot spots, which would be shared across all sequences.

Moving the cold spot from the middle of the alignment to the edge lowered the maximal level of false positives for PHI from 34% to 14%, and by about 50% for LDr^2 and LD|D$'$|, virtually eliminating any significant level of false positives, except for the coldest and most polymorphic simulations. Reticulate, GENECONV, and Max χ^2 were unaffected.

For all tests, fewer false positives were detected when the overall sequence was shorter (600bp in Figure 3.8-3.9 versus 1200bp in Figure 3.4-3.5). Although GENECONV, Max χ^2, and LDr^2 still reported close to 100% false positives, the performance of PHI, Reticulate, and LD|D$'$| greatly improved, with these tests returning around 10% false positives, compared to the 30% - 100% false positives when the sequences were longer.

Without a cold spot, multiple 1000-simulation-datasets were simulated and tested for recombination. Generally, the most the number of false positives fluctuated was by 20 counts, or 2%.

Using 100 simulations, the power of the tests was briefly investigated. When each simulation contained at least 1 recombination event, the number of false negatives returned were: PHI 24, Reticulate 35, GENECONV Global 25, GENECONV Global (inner fragments only) 26, Max χ^2 24, LDr^2 22, and LD|D$'$| 34.

3.5 Discussion

3.5.1 Mutation hot spots and heterotachy

Recombination creates homoplasy by facilitating the exchange of genetic material between distantly related species or individuals, distorting the phylogenetic history

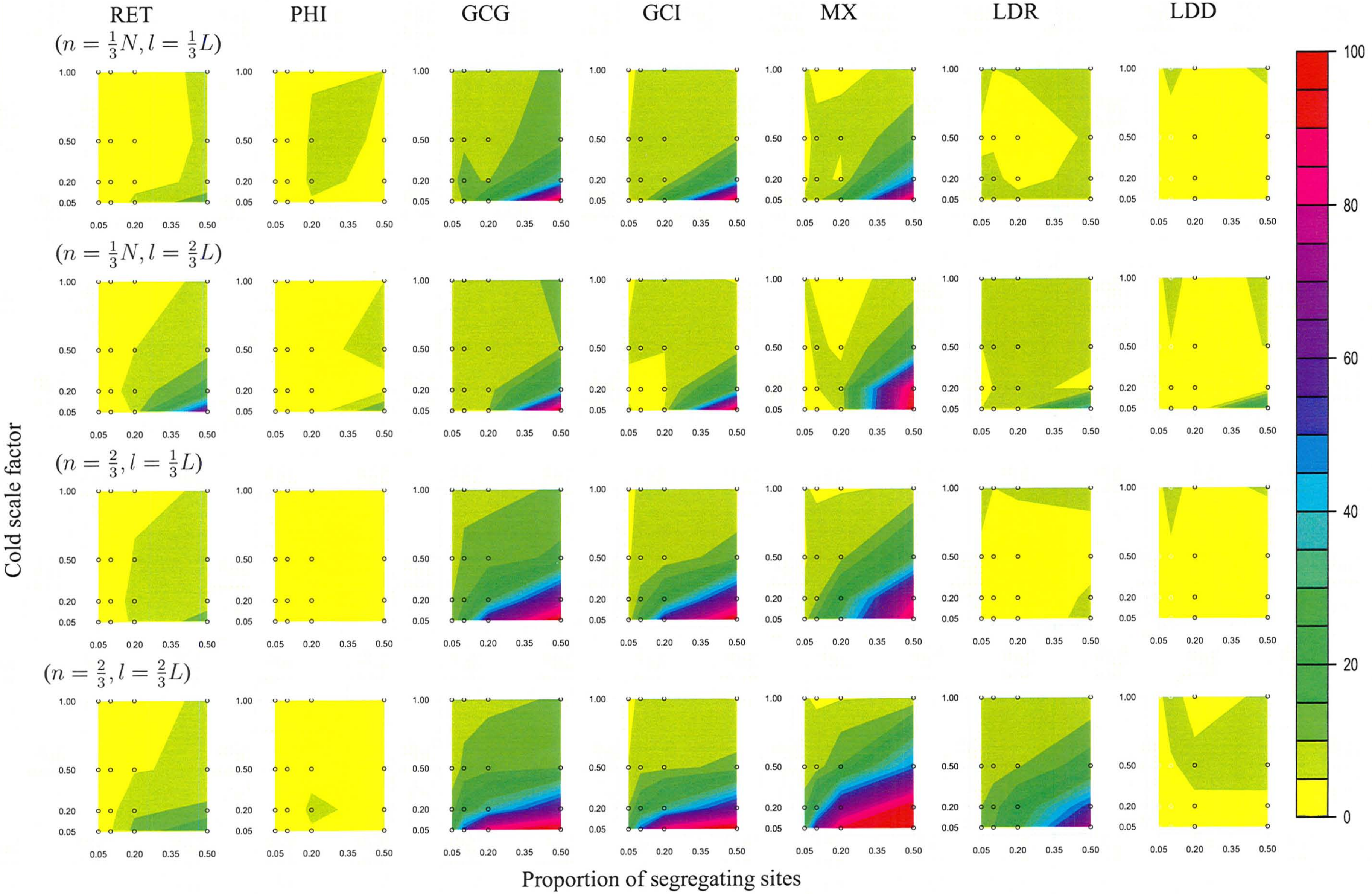


Figure 3.6: False positives in cold spot simulations: $N = 15, L = 1200bp, p =$ side. $n =$ number of sequence in cold clade, $l =$ length of cold spot. Black circles represent 800 - 1000 simulations which were tested for recombination, and grey circles mark data collected from fewer than 800 simulations.

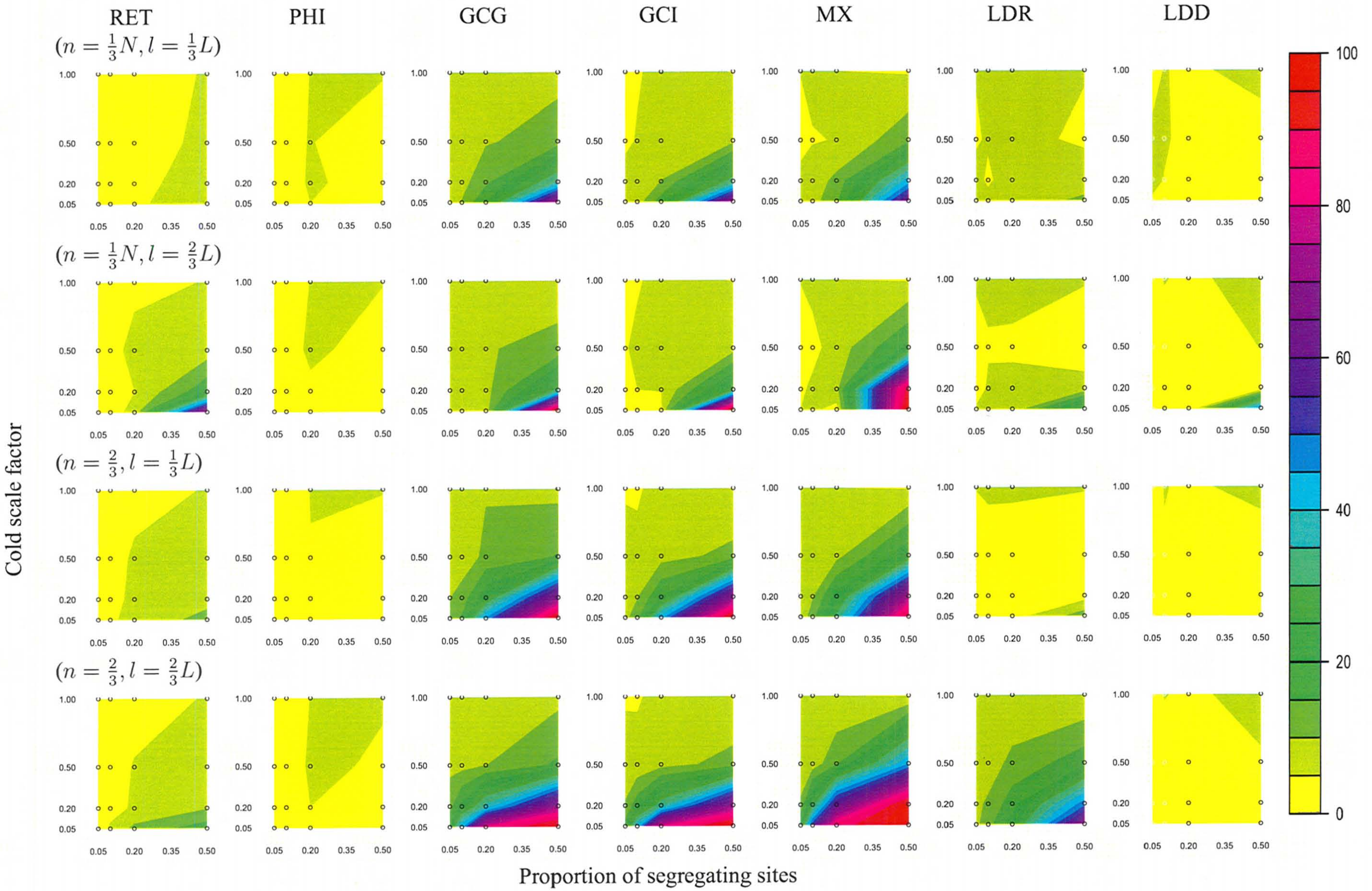


Figure 3.7: False positives in cold spot simulations: $N = 30, L = 1200bp, p =$ side. n = number of sequence in cold clade, l = length of cold spot. Black circles represent 800 - 1000 simulations which were tested for recombination, and grey circles mark data collected from fewer than 800 simulations.

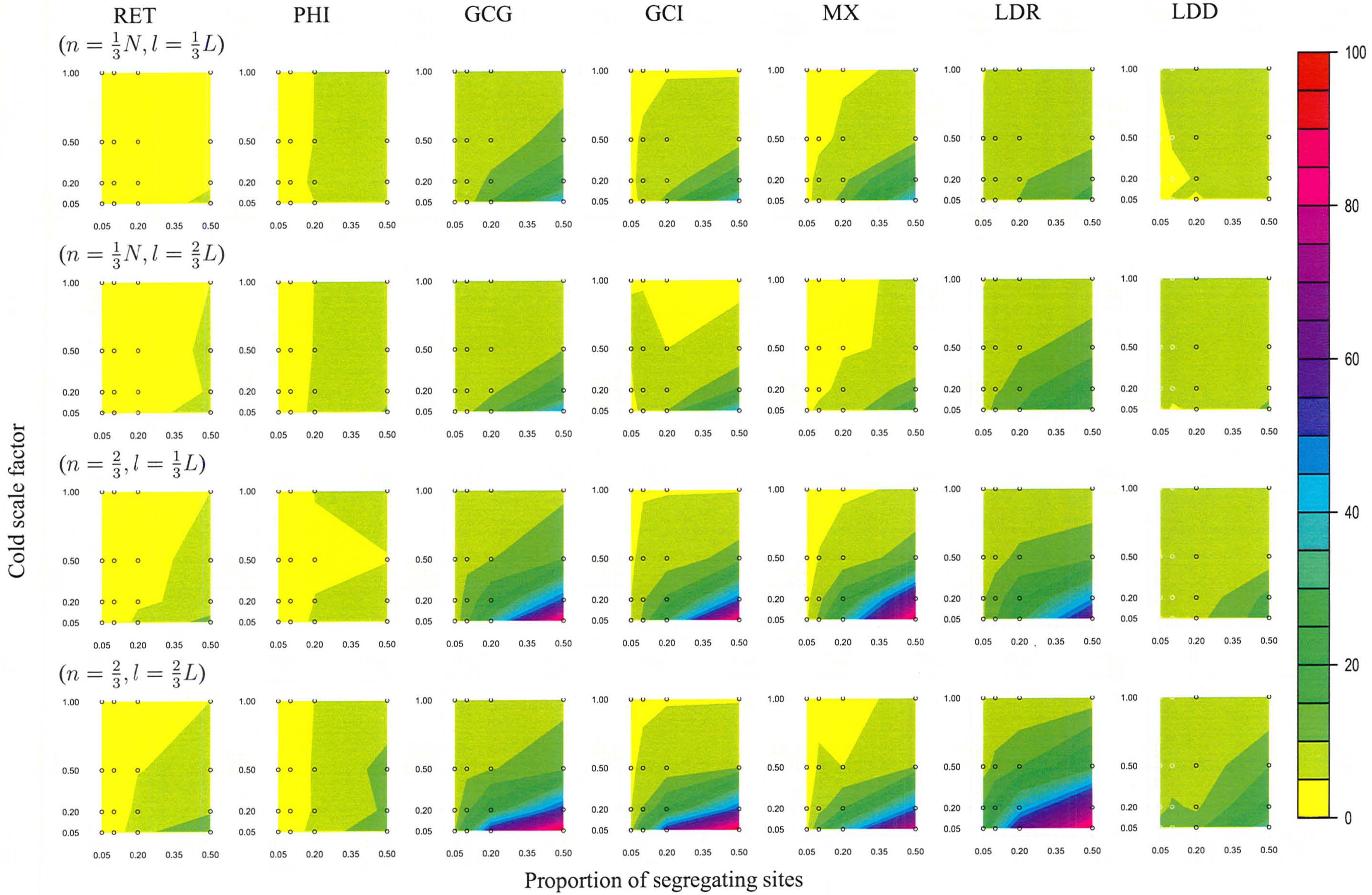


Figure 3.8: False positives in cold spot simulations: $N = 15, L = 600bp, p =$ middle. n = number of sequence in cold clade, l = length of cold spot. Black circles represent 800 - 1000 simulations which were tested for recombination, and grey circles mark data collected from fewer than 800 simulations.

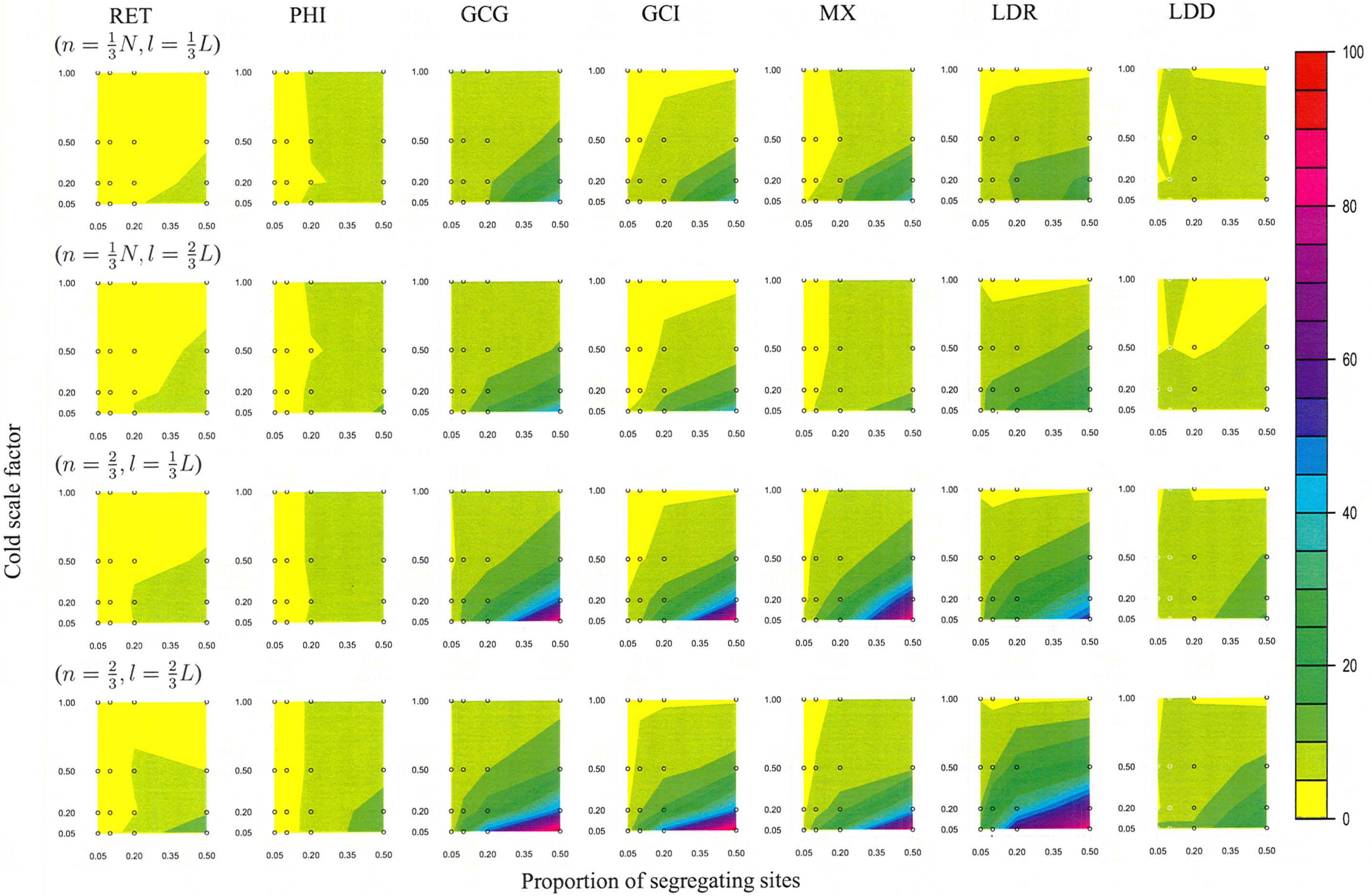


Figure 3.9: False positives in cold spot simulations: $N = 30, L = 600bp, p =$ middle. n = number of sequence in cold clade, l = length of cold spot. Black circles represent 800 - 1000 simulations which were tested for recombination, and grey circles mark data collected from fewer than 800 simulations.

of that sample. However, homoplasies can be generated by mechanisms other than recombination, such a hyper-variable sites (Galtier *et al.*, 2006).

The level of false positives returned by indirect recombination tests has been evaluated under various mutation hot spot models. Using simulations and a variety of evolutionary models, Posada and Crandall (2001) tested GENECONV, Max χ^2, and Reticulate for false positives given extreme rate heterogeneity and high levels of divergence. These tests did not produce an elevated level of false positives ($> 10\%$). Bruen, Philippe and Bryant (2006) found an elevated level of false positives ($> 30\%$) returned by Max χ^2 and Reticulate when sites had highly correlated rates, population growth, and high diversity, either with a large sample size of at least 50 or high level of nucleotide diversity (over 10%). With highly correlated rates between sites, Reticulate will also have elevated levels of false positives without population growth when testing large datasets with extreme rate variation and high nucleotide diversity. However, the biological relevance of the sometimes extreme correlation required is questionable (approximately 0.9 for Max χ^2, 0.3 or 0.6 for Reticulate, depending on degree of rate variation among sites). PHI, LDr^2, and LD$|$D$'|$ did not produce elevated false positives with population growth or high site-to-site correlation.

3.5.2 Cold spots versus hot spots

Alternative explanations for recombination have focused heavily on mutation hot spots rather than mutation cold spots for numerous reasons. Firstly, cold spots have been difficult to detect whereas the mtDNA control region has already been studied as a model mutation hot spot in mtDNA. Secondly, the likelihood of obtaining convergent mutations between distantly related individuals will be higher at a hyper-mutable site than at a site that rarely mutates. A third reason may be that the distinction between a hot or cold spot seems arbitrary. If the first half of an alignment contains a proportionately high density of hyper-mutable sites relative to the second half, whether the alignment has a hot spot or a cold spot depends on the frame (or sequence region) of reference. In the long term, it may be that mutation cold spots and mutation hot spots are two sides of the same coin. Mutation rates of non-synonymous sites are highly variable from region to region within mammalian mtDNA (Pesole *et al.*, 1999) and the mutation rates of specific mtDNA sites can change quickly over time, even between con-generic species (Galtier *et al.*, 2006). Although the location of cold and hot spots may be transient, we chose to model cold spots on a background of non-cold substitution rates, rather than a hot spot on a background of non-hot substitution rates. The genetic background sets the expectation upon which the mutational spot is compared. Considering the high average mutation rate of mtDNA, mutation cold spots will seem more significant to statistical tests than a hot spot. In short, whether a spot is hot or cold may be arbitrary on a genome-wide scale, but when evaluating the significance of a genetic signal at particular loci, the frame of reference can make all the difference.

In terms of modeling recombination false positives, a shift in focus from mutation hot spots to mutation cold spots could make an important difference when interpreting homoplasy in mtDNA. Consider three hypothetical alignments in which recombination has been detected using indirect tests: one containing a true recombinant region, one containing a hot spot, the other containing a cold spot. Indirect tests for recombination use statistics to determine whether sites in a putative recombinant region found within a subset of sequence samples appears significantly different from non-recombinant regions, yet are unusually similar to other sites within the recombinant region. For a mutation hot spot to mimic this pattern, there must be strong base composition bias to guide the homogenization of sites within the hot spot so that the spot appears to have undergone gene conversion (Awadalla, Eyre-Walker and Maynard-Smith, 1999). If there is no base composition bias, not only must there be convergent mutations in distantly related individuals, but the convergent mutations must occur in clusters according to the mutation's pseudo-phylogenetic histories. In comparison, a mutation cold spot can mimic recombination with fewer restrictions. Mutation cold spots will be indistinguishable from a gene conversion event, regardless of the base composition bias. On average, the hot spot alignment will have more homoplasies between pairs of sequences, but it will not necessarily have more homoplastic sites than the cold spot alignment.

3.5.3 Evidence for a mutational cold spot in Sulawesi macaque mtDNA

`Max` χ^2 and GENECONV detected an uneven distribution of segregating sites, and significantly long regions of perfect matches between samples collected from the North and Central regions of Sulawesi, as well as from one sample collected from the Southern region of Sulawesi. These signals could be interpreted as either recombination or mutation cold spots between North and Central, and one Southern Sulawesi macaque.

Similarly to recombination, mutation cold spots will produce an uneven distribution of segregating sites along the length of an alignment, as well as regions within an alignment with an unusually low number of mismatches. The significance of these signals can be assessed via permutation tests, as permutations will disrupt the clustering of cold sites, just as it would the clustering of recombinant sites.

3.5.4 Why are indirect tests susceptible to cold spots?

Max χ^2 and GENECONV

GENECONV assumes that without gene conversion, the distribution of bases at silent polymorphic sites will be determined by neutral mutations, which generally acts independent of site position. Degeneracy will affect the distribution of bases at particular site positions but this effect is probably independent of site position (Sawyer, 1989). Upon permutation, the distribution of bases will be retained, under the null hypothesis of no gene conversion. With gene conversion, however, the observed distribution of bases will be significantly different from distributions obtained upon permutation. When introducing his algorithm for detecting gene conversion, Sawyer (1989) states that GENECONV is robust to mutation hot and cold spots, provided the mutation spots are shared across all sequences. However, he acknowledges that GENECONV may be susceptible to false positives if a subset of samples contain a mutation cold spot, such as in the simulations tested here.

Max χ^2 and GENECONV assume an even distribution of segregating sites (permutation test) in the null hypothesis of no recombination. This assumption is violated by mutation cold spots. As shown here, mutation cold spots increase variation in the distribution of segregating sites along the length of a sequence. In fact, without recombination, the observed distribution in a supposedly recombining dataset (the Sulawesi macaques) could not be explained unless there was a cold spot.

Posada and Crandall (2001) classified these tests as substitution distribution methods; they look for significant clustering of substitutions. Other substitution distribution methods are the Homoplasy Test (Maynard Smith and Smith, 1998), Informative Sites Test (PIST) (Worobey, 2001), Chimaera (Posada and Crandall, 2001), the Runs Test (Takahata, 1994), and the Sneath Test (Sneath, 1995). Given that mutation cold spots violate the assumption that under a null hypothesis of no recombination, substitutions are evenly distributed across and between sequences, these substitution distribution methods will probably produce elevated false positives in a mutation cold spots model. This is especially true considering the performance of Max χ^2 and GENECONV, as they are the most powerful and robust substitution distribution methods. Indeed, two tests, the Homoplasy Test and PIST, have been found to produce high levels of false positives with extreme levels of rate variation (Posada and Crandall, 2001).

LD$|$D$'|$ and LDr^2

The null hypothesis, that without recombination, linkage disequilibrium does not correlate with distance fails on two counts. Firstly, mutation hot spots (or mutation cold spots, depending on the context) can produce a negative correlation between linkage disequilibrium and distance (Innan and Nordborg, 2002). Secondly, although recombination creates a negative correlation between linkage disequilibrium and distance in linear chromosomes (recombination breaks down linkage disequilibrium; there is more recombination over a greater distance), it is unclear how linkage disequilibrium and distance correlate in circular chromosomes (Wiuf, 2001). This is especially complicated when there are more than 2 breakpoints (for circular genomes, 2 breakpoints are required per recombination region to produce a recombinant circular chromosome); the relationship between linkage disequilibrium and distance is no longer monotonic. In short, regardless of whether linkage disequilibrium correlates with distance, we do not currently have a way to interpret it in circular mtDNA that can detect recombination or mutation rate heterogeneity, let alone a way to distinguish between the two. It makes sense, then, to look for evidence of recombination using knowledge of how recombination and mutation rate heterogeneity differ. For example, meaningful recombination cannot occur unless there is introgression of foreign mtDNA. Once there is introgression,there must be fusion of mitochondria, where the fusion rate is proportional to the recombination rate. In other words, until there is evidence of both introgression and fusion, mutation rate heterogeneity cannot be ruled out.

Assuming a breakdown in linkage disequilibrium with distance can be interpreted as evidence for recombination in circular mtDNA, mutation cold spots can produce a significant level of false positives for recombination, particularly when the cold spot is shared among the majority of sequences.

LDr^2 and LD|D$'$| perform similarly when evaluated using simulations with recombination and no mutation hot spots (Meunier and Eyre-Walker, 2001). With 0.5 recombination events per generation in the population, LDr^2 and LD|D$'$| correctly detects recombination in 0.28 and 0.30 of the simulations, respectively. In this study we have shown that a similar proportion of simulations can be detected from a wide range of mutation cold spot models.

LDr^2 produces the most false positives when the cold spot covers a larger proportion of sequence (two-thirds of total sequence length) and when the sample size of the cold clade is increased (two-thirds of total sample size). LD|D$'$| produces the most false positives when the cold spot covers a large proportion of the sequence but a smaller (one-third of total sample size) sampling of cold-spot sequences. Of all tests for recombination evaluated, LD|D$'$| was the most sensitive to alignments with a low proportion of segregating sites. Analysis by LD|D$'$| generally could not be conducted when the proportion of segregating sites was below 0.2 the total number of sites.

When the cold spot is situated at the periphery of the alignment or when the alignment is short (600bp), recombination is still falsely inferred in over 0.30 of the simulations, but under limited conditions; the cold spot is in one-third of the samples and covers two-thirds of the sequence length. Under these conditions, a similar level of recombination was inferred between LDr^2 and LD|D$'$| from simulations with over 0.2 polymorphic sites and a cold factor of 0.2. When the number of cold samples increases (to two-thirds of the total sample size), LDr^2 detects recombination whenever the cold factor was 0.2 or less, or when the proportion of segregating sites was greater than 0.2 times the total number of sites, and the cold spot was 0.5 the rate of the non-cold spot. In contrast, when the sample size of the cold clade was increased, LDr^2 no longer detected an elevated level of false positives.

Interestingly, the performance of LD|D$'$| is more similar to the compatibility methods Reticulate and PHI than to LDr^2, which is in turn more similar to the similarity comparison methods GENECONV and Max χ^2. The major difference between LD|D$'$| and LDr^2 is that LD|D$'$| analysis is only able to measure linkage disequilibrium when all four genotypes are present (the two parentals and two recombinants), and when allele frequencies are moderate to high (Awadalla, Eyre-Walker and Maynard Smith, 2000). It is not surprising, then, that LDr^2 has a higher level of false positives than LD|D$'$|, although LD|D$'$| will still produce an elevated level of false positives under certain conditions.

Reticulate and PHI

Reticulate and PHI are compatibility methods and of all the tests evaluated here, they are the most robust to mutation cold spots.

Reticulate and PHI produce elevated false positives at least 30% of simulations with a smaller sample size of cold-spot sequences (one-third over two-thirds of total sample size), and when the cold spot was longer (two-third over one-third of total sequence length). For PHI there is the added caveat that the total sample size is smaller (15 over 30 sequences) and that total sequence length is over 600bp. The improved false positive rate of PHI comes at a price; PHI is over-conservative when there are too few informative sites or too few incompatibilities. For PHI, this occurs when alignments have fewer than 15 samples or when nucleotide diversity is below 5% (Bruen, Philippe and Bryant, 2006).

Multiple hits increase the likelihood of false positives

Mutation cold spots most strongly affect the false positive rate of tests that assume polymorphisms are evenly distributed across the length of sequences. LD|D′|, Reticulate, and PHI are less susceptible to false positives by mutation cold spots. When these three tests do falsely infer recombination it is probably because sites have mutated multiple times in such a way that all four genotypes are present at a site (LD|D′|) and/or informative sites outside the cold spot have appear phylogenetically incompatible with sites inside the cold spot (Reticulate and PHI). Although the relative coldness of the cold spot's mutation rate plays a role in creating these recombination signals, the likelihood of these signals increases as the mutation rate per non-cold site, or background mutation rate, increases. This would explain why simulations with small cold clade but large cold spot produced the most false positives for LD|D′|, Reticulate, and PHI; a large cold spot forces the few remaining non-cold sites to accept more mutations per site, but if too many samples contain the cold spot, there will be too few informative sites to detect any signal. If these hypotheses are true, an elevated background mutation rate may produce false positives under widespread mutation cold spot models. Studies on the effect of linkage disequilibrium, albeit as measured by r^2, suggested that it was the contrast between mutation hot and cold spots that will produce false signals for recombination, not hot spots or cold spots on their own (Innan and Nordborg, 2002).

3.5.5 Regional mutation rate variation and heterotachy in animal mtDNA

Generally speaking, these tests are unreliable as methods for detecting recombination because they return an extremely high level of false positives in the presence of mutation cold spots, which have been documented within both coding and noncoding mammalian mtDNA (Pesole *et al.*, 1999). In terms of coding regions, where recombination has been detected, the rates at synonymous sites are generally uniform, while the rates at non-synonymous sites can vary widely depending on the gene. This non-synonymous substitution rate variation between genes supports a regional mutation rate heterogeneity model.

The problem in using these tests to detect recombination is not with the tests themselves, which are properly detecting the signals they have been written to detect, but in the interpretation of the test results as evidence for recombination. In other words, these signals could alternatively be interpreted as evidence for position-dependent mutation rate heterogeneity. Given the prerequisites for phylogenetically meaningful recombination, such as heteroplasmy through significant paternal leakage and/or the maintenance of low frequency mtDNA alleles despite extensive mitochondrial drift, mutation hot spots are arguably the more feasible explanation for these signals.

Perhaps the most hotly debated report of recombination in animal mtDNA was that detected in the human mtDNA genome by Awadalla, Eyre-Walker and Maynard-Smith (1999). This study reported a correlation between linkage disequilibrium and distance which was interpreted as indirect evidence for recombination. The methodology and interpretation was criticized on numerous fronts, the most convincing of which was the possibility of mutation hot spots as an alternative to invoking recombination. Eyre-Walker, Smith and Maynard Smith (1999) stated that there were too many homoplasies to be explained by hyper-variable sites, and that there was no evidence for mutation rate variation between sites, but the latter was based on the assumption that hyper-mutable sites would remain hyper-mutable over time. It was later determined that this is not so; even within the same genus, a hyper-mutable site in one species will not necessarily be hyper-mutable in another species (Galtier *et al.*, 2006). Ultimately, the evidence for recombination in human mtDNA was determined likely to have been a chance observation due to small sample size. However, in the process of showing this, Innan and Nordborg (2002) demonstrated that a correlation between linkage disequilibrium and distance could be produced by regions with a high density of fast-evolving sites (mutation hot spots) situated next to regions of slow-evolving sites (mutation cold spots), at least theoretically. This raises the question of whether mutation rate variation can lead to false positives for recombination under biologically realistic rates of evolution, and is the focus of Chapter 4.

Chapter 4

Heterotachy in animal mtDNA produces recombination false positives

4.1 Abstract

Twenty animal mtDNA datasets in which recombination has been detected were evaluated for false positives by allowing different regions of the sequences to evolve at different rates in different lineages. Simulations were created using biologically estimated rates, and tested for recombination. Through simulations, a re-analysis of these datasets shows that regional mutation rate heterogeneity can alternatively explain the detection of widespread recombination in animal mtDNA.

4.2 Introduction

4.2.1 Recombination in animal mtDNA

There is considerable evidence to support the possibility of recombination in animal mitochondria. Mammalian mitochondrial protein extracts can catalyze recombination (Thyagarajan, Padua and Campbell, 1996), and mtDNA genomes may mix as mitochondria are capable of forming dynamic networks within a cell through fusion and fission (Wilson, 1916; Bereiter-Hahn and Vöth, 1994) although the frequency of fusion is unknown (Birky, 2001), and some mitochondria may be fixed to a particular region of the cell, preventing them from fusing with other mitochondria. Nevertheless, when two parental cells are fused, mitochondrial recombinant genotypes can be detected (Birky, 2001), and recombination products have been directly observed from the gonads of heteroplasmic male bivalve mussels (Ladoukakis and Zouros, 2001) and in the heteroplasmic muscle cells of a human individual (Kraytsberg *et al.*, 2004). However, these examples remain exceptional cases in the otherwise clonal nature of animal mtDNA biology. They represent exceptions to two major features of mitochondrial genetics, maternal inheritance and vegetative segregation, which maintain the presence of only one mtDNA genotype in an individual

Table 4.1: Examples of tests used to detect recombination in animal mtDNA.

Test	Used in
GENECONV	Piganeau, Gardner and Eyre-Walker (2004); Tsaousis *et al.* (2005); Ujvari, Dowton and Madsen (2007); Gantenbein *et al.* (2005); Lawson and Zhang (2009)
Max χ^2	Piganeau, Gardner and Eyre-Walker (2004); Tsaousis *et al.* (2005); Bruen, Philippe and Bryant (2006); Ujvari, Dowton and Madsen (2007); White and Gemmell (2009); Gantenbein *et al.* (2005); Maynard Smith and Smith (1999)
LDr^2	Piganeau, Gardner and Eyre-Walker (2004); Awadalla, Eyre-Walker and Maynard-Smith (1999); White and Gemmell (2009); Gantenbein *et al.* (2005)
LD$\|D'\|$	Piganeau, Gardner and Eyre-Walker (2004); White and Gemmell (2009); Gantenbein *et al.* (2005)
Reticulate	Tsaousis *et al.* (2005); Bruen, Philippe and Bryant (2006); White and Gemmell (2009); Fitzgerald *et al.* (1996)
PHI	Bruen, Philippe and Bryant (2006); White and Gemmell (2009)

(Birky, 2001), a state known as homoplasmy. Whether mtDNA recombination is pervasive enough to require a serious re-evaluation of animal population studies remains contentious due to effective homoplasmy and the associated difficulty in collecting direct empirical evidence of recombination.

4.2.2 Heterotachy as an alternative to recombination

Evolutionary relationships are inferred by the genetic information stored in sequences, using assumptions or models which are also based on information from the samples. Differences between the sequences are interpreted in the context of the model to reconstruct an estimate of the ancestry, or evolutionary relationships between the sampled individuals. This process is confounded when different sites or regions in the sequences suggest conflicting evolutionary relationships, or homoplasy. One option is to accept the validity of all conflicting evolutionary relationships, meaning that different parts of the sequences have different evolutionary histories. This would require that an exchange of genetic material has occurred through recombination. On the other hand, sites may produce conflicting evolutionary relationships if there is a bias for some sites in some sequences to mutate in such a way that the true evolutionary history of the individuals is distorted by the chance mutations. Indeed, it has been demonstrated that hyper-variable sites, or sites with unusually high mutation rates, can generate homoplasies (Galtier *et al.*, 2006), and that the mutation rate at a certain site likely changes quickly over time, a phenomenon referred to as covarion or more generally as heterotachy (Fitch, 1971). Some indirect tests for recombination are liable for false positives under conditions that can approximate mutation hot spots, such as extreme rate heterogeneity among sites (Posada and Crandall, 2001) and high rate correlations between sites (Bruen, Philippe and Bryant, 2006). Most indirect tests, however, either do not produce elevated false positives under these conditions, produce false positives but under biologically extreme conditions, or have not been evaluated under these conditions.

In Chapter 3 we described the elevated level of recombination false positives detected due to mutation cold spots. In this chapter, we turn to twenty animal mtDNA datasets in which recombination has been inferred, to create simulation datasets with heterotachy under biologically relevant mtDNA parameters. These simulations were tested for recombination to determine the level of false positives expected from biologically estimated levels of heterotachy. The false positive rate for 6 tests of recombination were evaluated: GENECONV (Sawyer, 1989, 1999), Max χ^2 (Maynard Smith, 1992), LDr^2 (Piganeau, Gardner and Eyre-Walker, 2004; Hill and Robertson, 1968), LD$|$D$'|$ (Piganeau, Gardner and Eyre-Walker, 2004; Lewontin, 1964), Reticulate (Jakobsen and Easteal, 1996), and PHI (Bruen, Philippe and Bryant, 2006). These tests are among the most powerful methods of detecting recombination when direct identification of recombinants is not feasible, and have been used to screen for recombination in a wide range of animal mtDNA (Table 4.1).

4.3 Method

4.3.1 Partitioned models produce false positives

Twenty animal mtDNA datasets were collected from Genbank . These included the previously discussed Sulawesi macaque mtDNA plus 19 of the top 20 datasets with the strongest evidence for recombination, according to Piganeau, Gardner and Eyre-Walker (2004). *Mytilus galloprovincialis*, one of Piganeau, Gardner and Eyre-Walker (2004)'s top 20 datasets, was excluded as recombination has been empirically detected in this species (Ladoukakis and Zouros, 2001). The 20 animal mtDNA datasets are representative of a wide range of mitochondrial coding sequences, sequence lengths, level of polymorphism, and sample sizes. These characteristics are summarized in Table VI.

Each dataset was aligned using muscle (Edgar, 2004) and a consensus topology was obtained using MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). Each animal mtDNA dataset was simulated under a partition model without recombination. That is, although a single tree topology was used for the alignment, all other parameters were obtained by first partitioning the alignments into two sections then estimating maximum likelihood parameters off each section. Partition model names and patterns are described in Figure 4.1. Estimation of parameter values was accomplished using baseml from the PAML suite of programs (Yang, 1997, 2007), under a GTR+Γ model with 5 discrete Γ rate

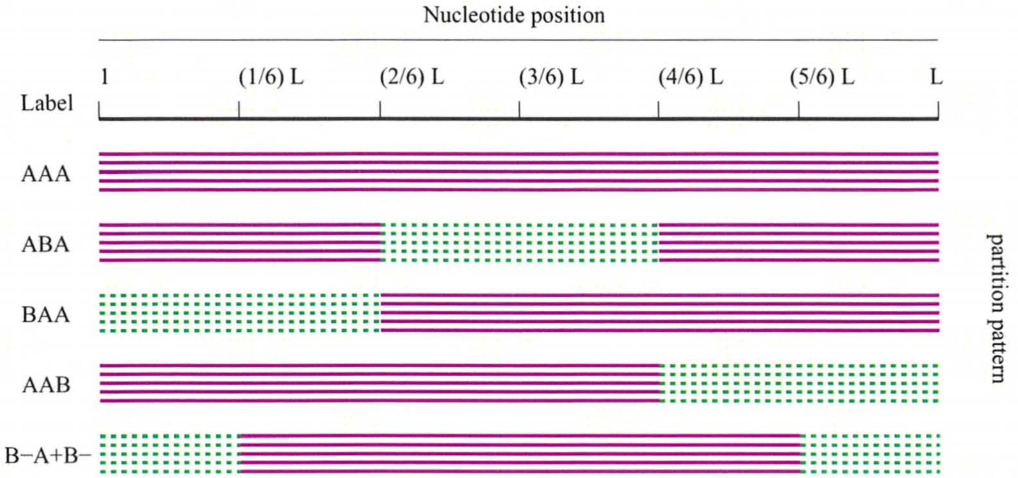


Figure 4.1: Five partition models for an alignment of L nucleotides. Section A is represented by solid lines, section B is represented by broken lines. Sections A and B are simulated off different models, and either A or B may contain the cold spot.

categories. The decision to partition the simulations into three sections, with the middle section covering roughly a third of the alignment, was based on the pattern of recombination in Sulawesi macaques, which suggested the presence of a cold spot in the middle third of the alignment (Chapter 3). Admittedly, this may be too few or too many partitions to detect cold spots when modeling other animals. If the number of partitions is too high, there may be too much noise in rates to distinguish a local patterns of rate heterogeneity. the overall picture of mutation rate heterogeneity will be lost. However, if the number of partitions is too low, mutation rate heterogeneity may be overshadowed by the averaging of rates among sites. In an attempt to resolve this, section length was set in proportion to total sequence length, rather than a set section length.

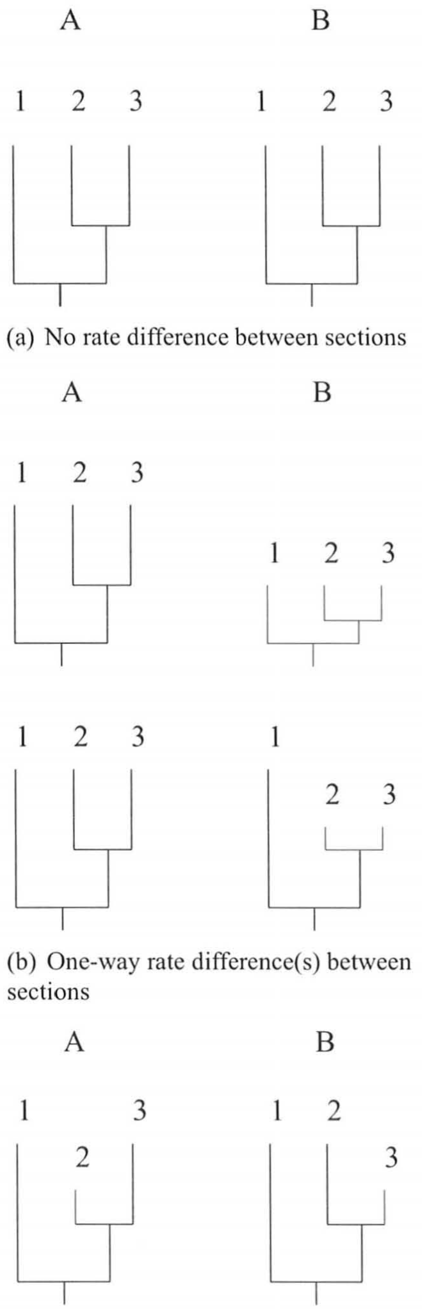
The likelihoods of the partitioned models (that is, *ABA*, *BAA*, *AAB*, or *B-A+B-*) were obtained by summing the lnL_{max} obtained for *A* and *B* sections, where L_{max} is the maximum likelihood estimate. The partitioned model with the greatest likelihood was used to create 1000 simulations. The `MrBayes` tree and 9 section-specific `baseml` estimated parameters were used to generate alignments in `Seq-Gen`, one section at a time. Sequences from section *A* and section *B* were then pieced back together and tested for recombination. Simulations were created in the same fashion from the no-partition model (*AAA*) except that the concatenation step was not necessary. Note that before supplying any trees to `Seq-Gen`, all multifurcations must be changed to zero branch lengths, as `Seq-Gen` only allows multifurcations at the root of the input tree.

At this step, maximum likelihood is an inappropriate means of comparing the models because the models have different degrees of freedom (d.f.). The Akaike Information Criterion (AIC) (Akaike, 1974) was used to determine whether a partitioned model (*ABA*, *BAA*, *AAB*, or *B-A+B-*) was significantly better than a no-partition model (*AAA*). The *AAA* model has 9 d.f. (5 rate ratios, 3 nucleotide frequencies, and 1 Γ shape parameter), while the other partition models have 18 d.f. (9 d.f. for partition *A* and 9 d.f. for partition *B*). Since maximum likelihood is designed towards models that are most fitted to the data, the model with more d.f. will always be favoured, even if the extra d.f. do not provide more information. A more appropriate means of comparing models with different d.f. is to use the AIC, which takes into consideration how much the model improves when degrees of freedom are added. The AIC was calculated as $2d.f. - 2lnL$ where there 9 d.f. in the *AAA* model, and 18 in all other models.

4.3.2 Types of partition rate heterogeneity

Branch lengths estimated from section *A* were compared to the branch lengths estimated from section *B* to determine whether the sections were evolving at different rates and if so, in what way they were different (Figure 4.2). For each branch, the ratio of section *A* branch length over section *B* branch length was recorded. Therefore, if the branch length ratio was greater than 1, section A appeared to be evolving faster than section *B*. Conversely, if the branch length ratio was less than 1, section *B* appears to be evolving faster along that branch.

Given a model of evolution, branch lengths are inferred by the number and type of substitutions occurring on that lineage. Assuming an even rate of evolution across the length of the sequence, one expects segregating sites to be evenly distributed along the length of the sequence. However, the stochastic nature of evolution ensures that the distribution of segregating sites will rarely ever be evenly distributed. With this in mind, a null distribution for branch length ratios was estimated. In this distribution, partitioned alignments were simulated as before, except that the same tree and parameters were supplied to both partitions. Therefore, any differences between the sections would be due to stochastic effects. Maximum likelihood estimated parameters from the *AAA* no-partition model were used to create 1000 simulations. The partitioned pattern with the greatest likelihood was used. Each branch was compared between partitions and linear regression of section *A* branches on section *B* branches was carried out. A 95% prediction interval was calculated off the linear regression, and is the interval within which 95% of observed data-points are expected to fall. If the branch length ratios of the animal mtDNA data fell within the 95% prediction interval, the animal mtDNA partitions were not significantly different from each other. If the branch lengths of the animal mtDNA fell outside the prediction bounds, section *A* and section *B* are inferred to be evolving at different rates. Whether this constituted a cold spot, hot spot, or tree-wide heterogeneity (for A relative to B) depended on whether the points fell below, above, or both below and above the prediction bounds, respectively.



(a) No rate difference between sections

(b) One-way rate difference(s) between sections

(c) Two-way rate difference(s) between sections

Figure 4.2: Examples of regional differences in substitution rates. Columns represent trees estimated from different sections of the alignment, each row describes possible outcomes and how they are interpreted.

4.4 Results

4.4.1 Partitioned models versus no-partition

The real data was partitioned lengthwise into two sections, and one set of parameters was estimated from each section. Partial sequences were simulated, then concatenated to form a complete simulation dataset. The maximum likelihood of simulated partitioned dataset was compared to the maximum likelihood of the sequence simulations without partitions (*AAA*). This comparison, which was done using AIC, was to judge whether the real sequences were better modeled with partitions or without partitions. For all but one animal mtDNA dataset, the maximum likelihood partitioned model had a significantly lower AIC than the *AAA* model, and hence is the more appropriate model. The exceptional case was *Mandrillus sphinx*, where a small difference in AIC between *AAA* and the maximum likelihood partitioned (< 1) supported *AAA* as the better model. We also compared the no partition (*AAA*) and partitioned model with highest likelihood (see Table B.2) using the likelihood ratio test (LRT). As expected, the partition models were significantly better than no partition model; even for the Mandrill dataset (p-value 0.0444).

4.4.2 Types of partition rate heterogeneity

The animal datasets were divided into three categories based on the differences in substitution rate between sections (Table 4.2), as inferred from corresponding branch lengths between sections. Datasets with no significant difference between branches from tree A and tree B are *Mandrillus sphinx*, *Alpheus lottini*, *Campylorhynchus brunneicap*, and *Gonatus onyx* (Figure 4.3). Datasets where either tree A or tree B contained a cold spot are *Micropterus salmoides*, *Vesicomya pacifica*, *Libellula quadrimaculata*, *Dendroica petechia*, *Apodemus sylvaticus*, *Gomphiocephalus hodgsoni*, *Papio papio*, *Passerella iliaca*, and *Merlangius merlangus* (Figure 4.4). Finally, datasets where both tree A and tree B contained cold spots are the Sulawesi macaques, *Bursaphelenchus conicaudatus*, *Macaca nemestrina*, *Microtus longicaudus*, *Macrodon ancylodon*, *Bradypodion occidentale*, and *Grus antigone* (Figure 4.5).

In Piganeau, Gardner and Eyre-Walker (2004)'s broad survey of recombination in 267 animal mtDNA, a Bonferroni correction was used to correct for testing multiple datasets. After this correction was implemented, only four datasets remained statistically significant – *Bursaphelenchus conicaudatus*, *Macaca nemestrina*, *Microtus longicaudus*, and *Micropterus salmoides*. Interestingly, three of these datasets – *Bursaphelenchus conicaudatus*, *Macaca nemestrina*, and *Microtus longicaudus*– display two-way heterogeneity, indicating the data are comprised of sites where one part of the sequence evolves faster in a subset of taxa whereas another part of the sequence evolves faster in a different subset of taxa. The fourth independently significant dataset is *Micropterus salmoides*, and falls in the one-way heterogeneity category. In other words, Piganeau, Gardner and Eyre-Walker (2004)'s strongest and most robust examples of recombination in animal mtDNA are among those with the most widespread heterogeneity.

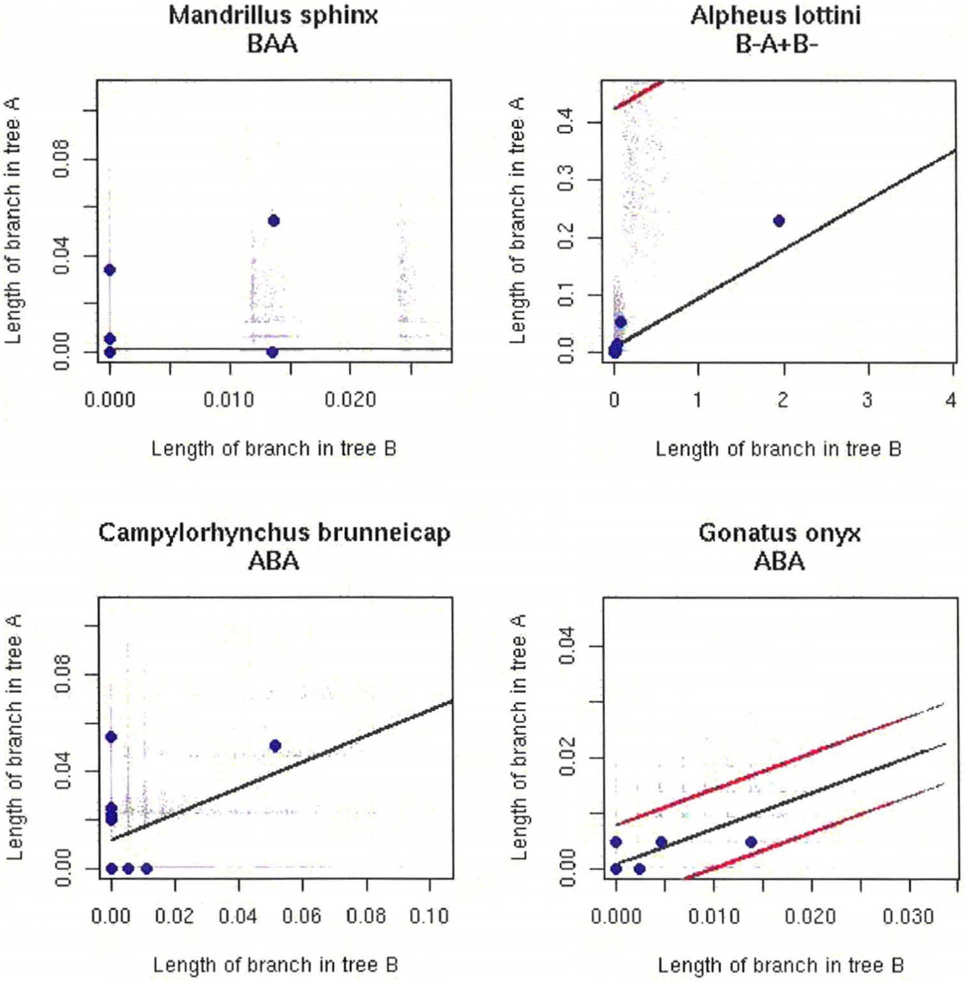


Figure 4.3: Datasets with no regional differences in animal mtDNA: $A = B$. Grey points represent branches from AAA simulations. Blue points represent branches from the animal mtDNA. Linear regression of branch A on branch B is represented by the black line, and red lines delineate the 95% prediction bounds. The graphs of *Mandrillus sphinx, Alpheus lottini*, and *Campylorhynchus brunneicap* were scaled to more closely show the location of observed data, consequently leaving some simulation data out of the picture. Graphs displaying all simulation data are can be found in Figure B.1.

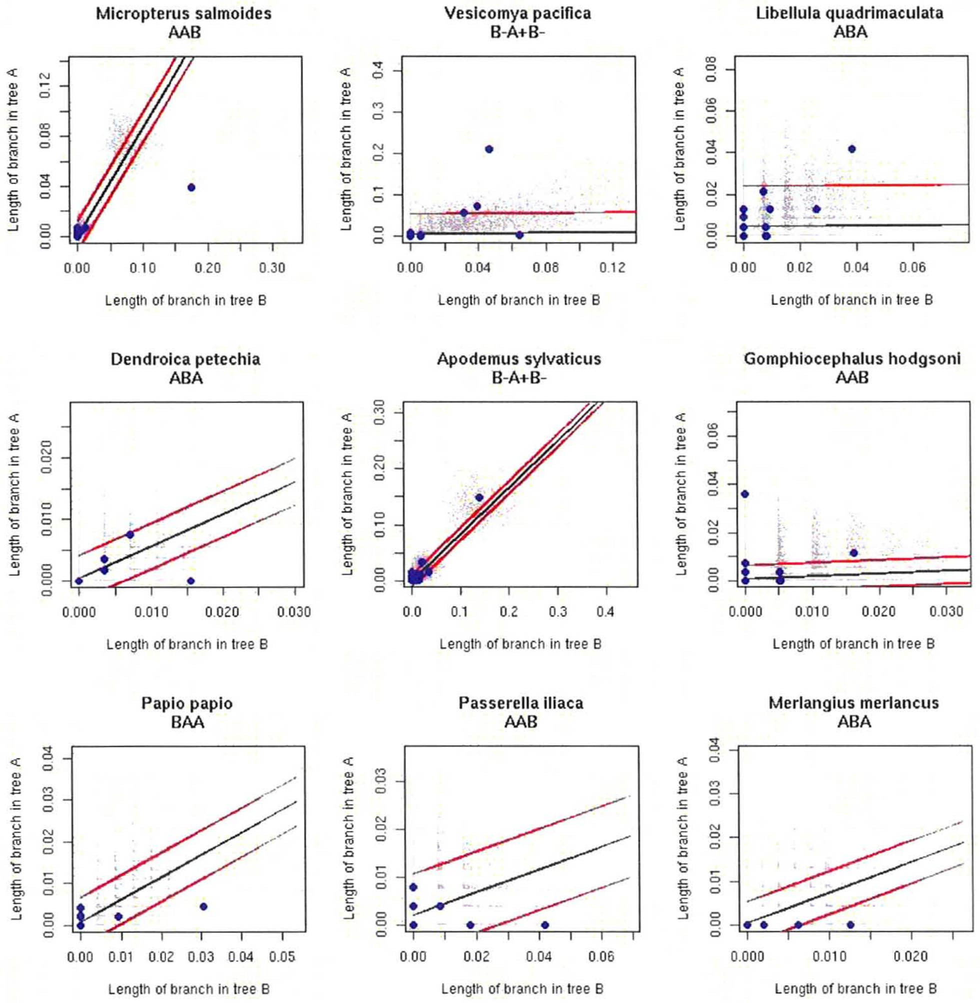


Figure 4.4: One-way heterogeneity in animal mtDNA: $A > B$ or $A < B$. Grey points are branches from AAA simulations. Blue points are branches from the animal mtDNA. Black line is linear regression of branch A on branch B, and red lines are the 95% prediction bounds. The graphs of *Vesicomya pacifica, Libellula quadrimaculata, Gomphiocephalus hodgsoni,* and *Macrodon ancylodon* were scaled in. See Figure B.1 for full picture.

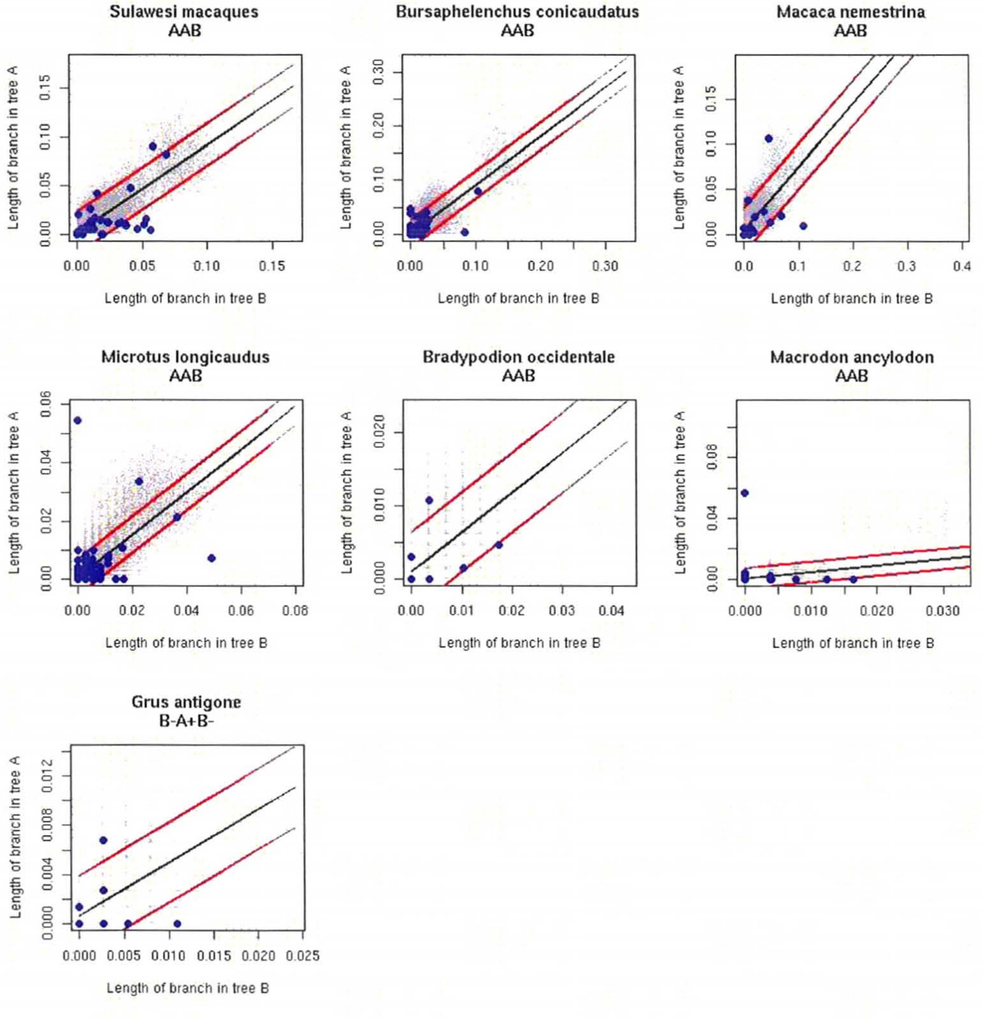


Figure 4.5: Two-way heterogeneity in animal mtDNA: $A > B$ and $A < B$. Grey points represent branches from AAA simulations. Blue points represent branches from the animal mtDNA. Linear regression of branch A on branch B is represented by the black line, and red lines delineate the 95% prediction bounds.

Table 4.2: Regional rate differences in animal mtDNA. A and B refer to the average substitution rate along a branch for sites from section A or section B, respectively. P-values are summarized from Piganeau, Gardner and Eyre-Walker (2004): $*** \leq 0.0005 < ** \leq 0.005 < * \leq 0.05 < -$. GC - GENECONV, MX - Max χ^2, LDR - LDr^2, LDD - LD$|$D$'|$.

Category	Dataset (common name)	GC	MX	LDR	LDD
No	*Mandrillus sphinx* (mandrill)	-	**	-	*
difference	*Alpheus lottini* (snapping shrimp)	-	**	-	-
$A = B$	*Campylorhynchus brunneicap* (wren)	-	**	*	-
	Gonatus onyx (squid)	-	-	**	na
One-way	*Micropterus salmoides* (bass)	***	***	***	-
difference	*Vesicomya pacifica* (bivalve)	**	*	*	-
$A < B$	*Libellula quadrimaculata* (dragonfly)	-	*	-	-
or	*Dendroica petechia* (warbler)	*	-	*	na
$B > A$	*Apodemus sylvaticus* (woodmouse)	-	**	*	-
	Gomphiocephalus hodgsoni (springtail)	**	**	-	-
	Papio papio (baboon)	*	-	*	-
	Passerella iliaca (sparrow)	*	-	*	*
	Merlangius merlangus (whiting)	*	-	*	-
Two-way	Sulawesi macaque	**	***	**	-
difference	*Bursaphelenchus conicaudatus* (Nematode)	***	***	***	**
$A < B$	*Macaca nemestrina* (Pig-tailed macaque)	-	***	***	-
and	*Microtus longicaudus* (Vole)	***	***	-	-
$B < A$	*Grus antigone* (Crane)	-	-	*	*
	Macrodon ancylodon (King weakfish)	-	-	*	**
	Bradypodion occidentale (Chameleon)	-	*	*	na

The three categories of partition rate heterogeneity are summarized in Table 4.2, along with a summary of Piganeau, Gardner and Eyre-Walker (2004)'s recombination test results.

4.4.3 Partitioned models produce false positives

Each simulation dataset consisted of 1000 alignments. The alignments were tested for recombination, and the number of times recombination was detected is presented in Figures 4.6-4.14. Since recombination was not included in the simulations, any detected recombination is a false positive.

Without exception, tests performed worse when the data was partitioned (*AAB*, *BAA*, *ABA*, *B-A+B-* partition models) than when only one set of parameters and branch lengths was provided for the whole sequence (*AAA*). Max χ^2, GENECONV, and $\texttt{LDr}^2$ performed the worst, sometimes returning up to 80% false positives. Generally, datasets with more mutation rate heterogeneity produced more false positives in more tests; datasets where cold spots were present in both trees produced the most recombination, while datasets without any cold spots produce the least.

The recombination results of datasets without mutation rate heterogeneity are summarized in Figures 4.6-4.8. The animal mtDNA without differences in rates between sections A and B (Figure 4.3) did not produce an elevated level of false positives ($> 10\%$), except in the case of *Mandrillus sphinx* in $\texttt{LDr}^2$ and *Campylorhynchus brunneicap* in $\texttt{LD}|\texttt{D}'|$ (Figure 4.8).

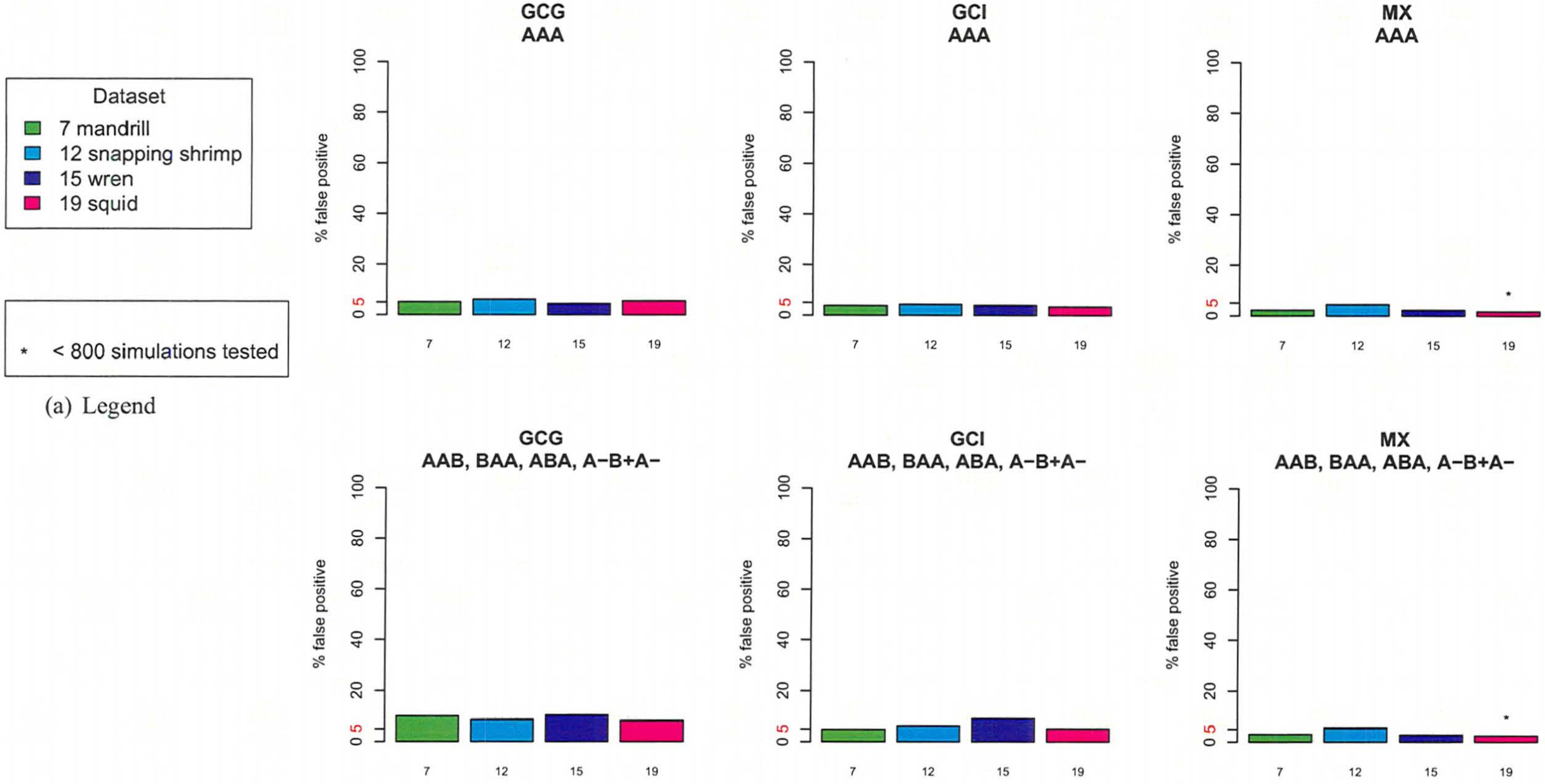
Most of the animal datasets tested (55%) had significant differences in evolutionary rates between section A and B, where branches in one section were evolving slower than or equal to the branches in the other section (Figure 4.4). That is, either section A contained cold spots relative to section B, or section B contained cold spots relative to section A. Results of the recombination tests are presented in Figures 4.9-4.11. Simulations from these datasets produced over 10% false positives in GENECONV, Max χ^2 (Figure 4.9), $\texttt{LDr}^2$, and $\texttt{LD}|\texttt{D}'|$ (Figure 4.11).

The third category also contained datasets where the evolutionary rates between section A and B differed, such that both section A and B contained slow evolving sites. That is, section A contained cold spots relative to B, but elsewhere in the tree section B contained cold spots relative to A (Figure 4.5). False positives returned by these datasets are presented in Figures 4.12-4.14. As a group, this category produced an elevated level of false positives from all tests. Two datasets, *Macrodon ancylodon* and *Bradypodion occidentale*, do not produce an elevated level of false positives in Reticulate or PHI (Figure 4.13). This is probably related to the fact that their cold spots in section A has a lower significance than the other datasets in this category.

4.5 Discussion

Heterotachy is widespread in the mitochondrial genome. For example, at least 28% to 95% polymorphic sites are heterotachous in mitochondrial coding regions (Lopez, Casane and Philippe, 2002), and the position of heterotachous sites does not appear to be tied to functional divergence nor spatial structure of the protein. Typical models of heterotachous evolution are defined by an evolutionary rate for a particular site that changes over time, or between lineages. Here, we consider a heterotachous model where the evolutionary rate of a cluster of sites changes over time. This model is similar to the likelihood model of Meade and Pagel (2008) which allows variation in rates at different sites considering multiple sets of independent branch lengths (i.e. trees with different branch lengths but the same topology). Likelihoods are summed over the sites for every set of branch lengths. In this way, heterotachous sites will favour different sets of branch lengths. Here, the data is partitioned into sections, and different sets of branch lengths are tested for each section. The likelihoods are then summed over each section.

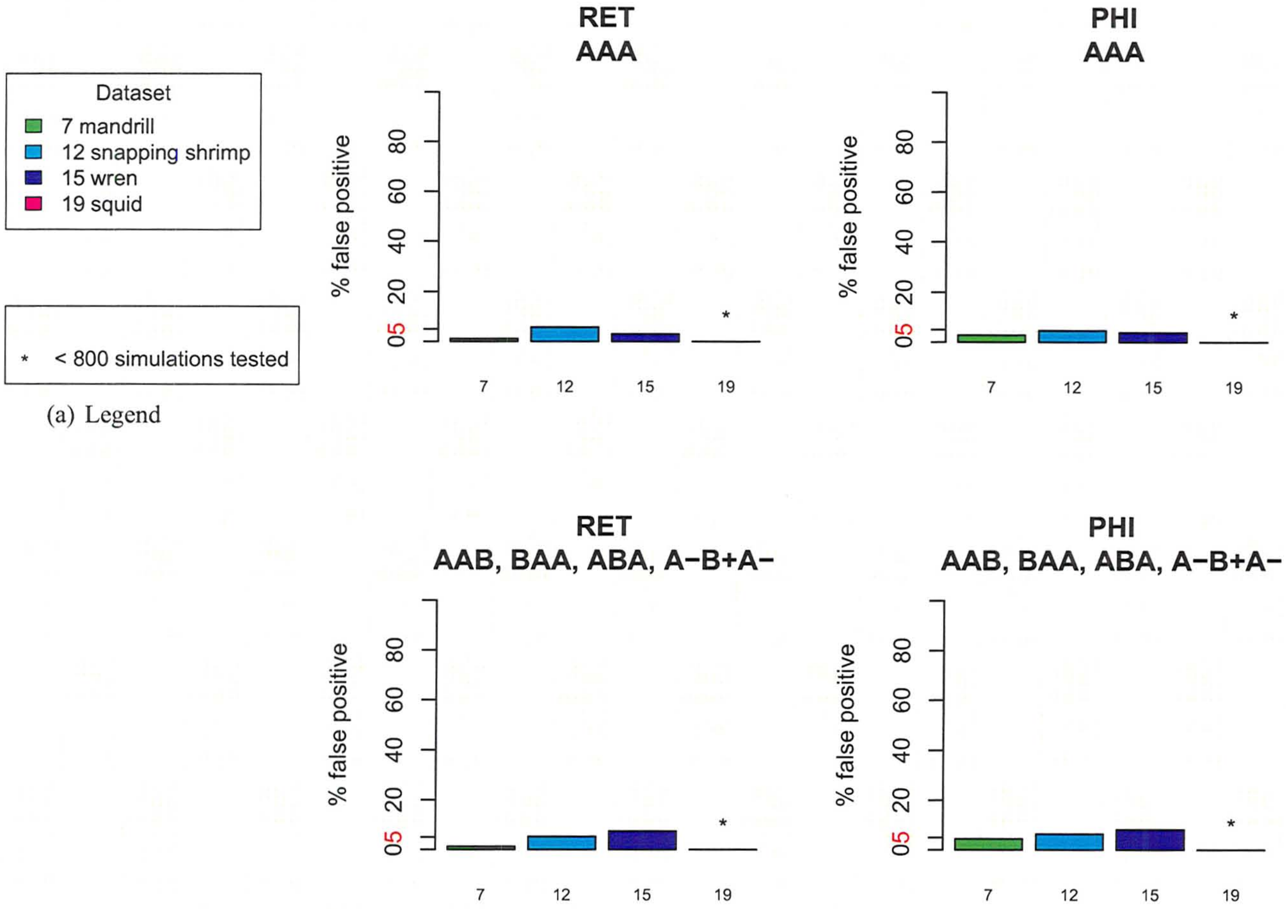
We have previously shown that mutation cold spots can produce false positives for recombination (Chapter 3). While Max χ^2, GENECONV, and $\texttt{LDr}^2$ were



(a) Legend

(b) GENECONV and Max χ^2

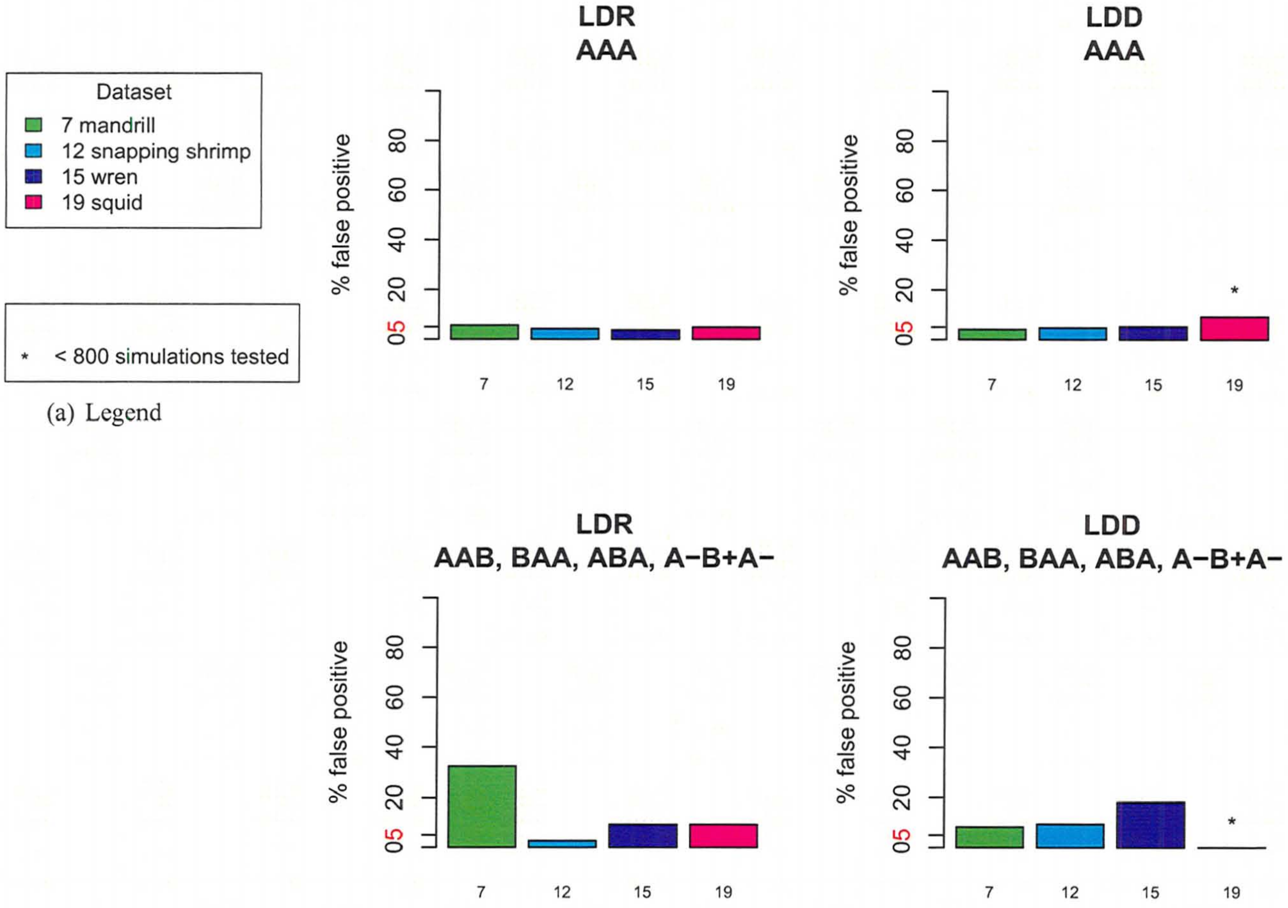
Figure 4.6: False positives for recombination in partition models of animal mtDNA. No heterogeneity ($A = B$). Unless otherwise noted (with *), 1000 simulations were tested per dataset.



(a) Legend

(b) RET and PHI

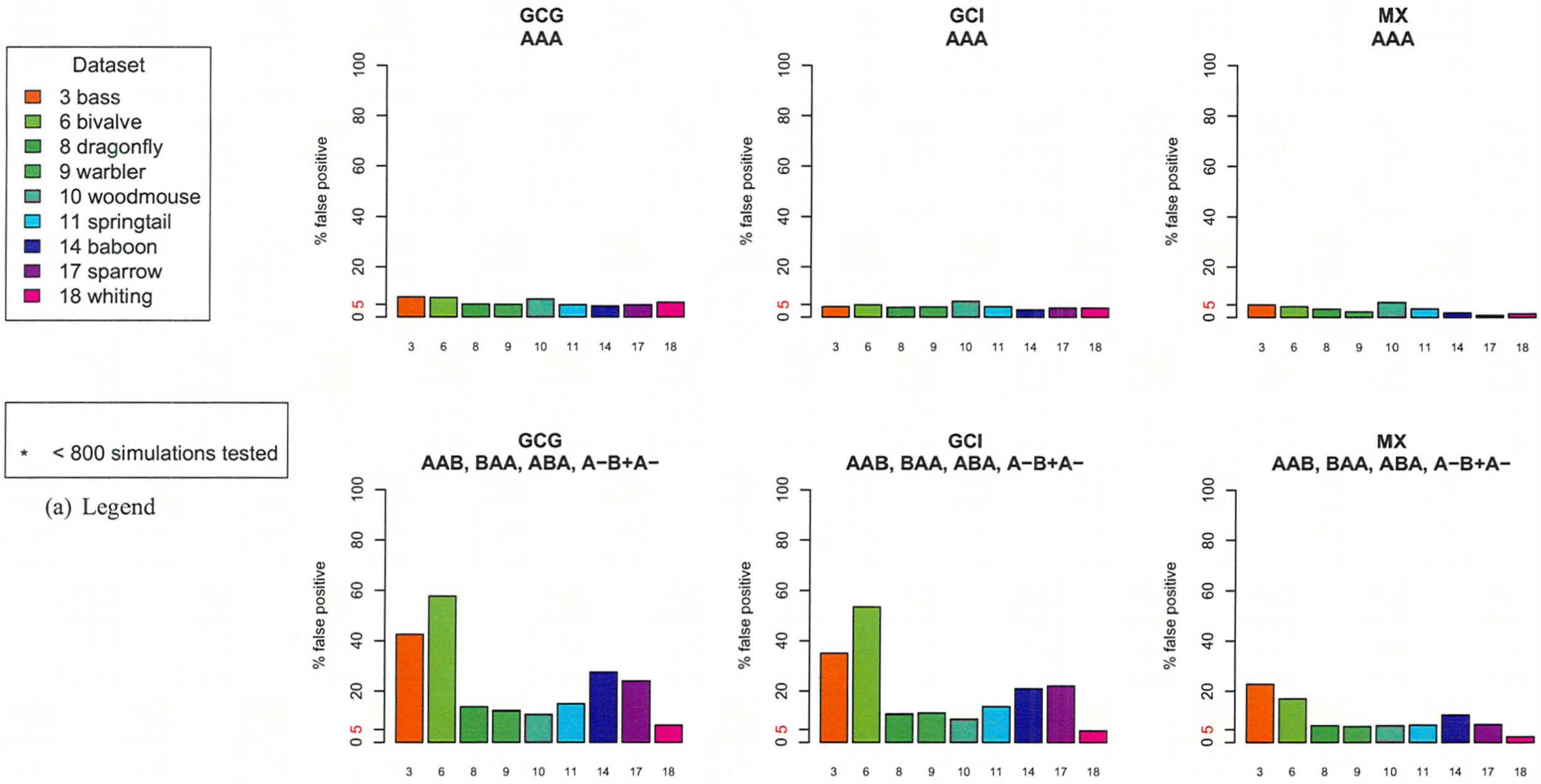
Figure 4.7: False positives for recombination in partition models of animal mtDNA. No heterogeneity ($A = B$). Unless otherwise noted (with *), 1000 simulations were tested per dataset.



(a) Legend

(b) $\mathtt{LDr}^2$ and $\mathtt{LD|D'|}$

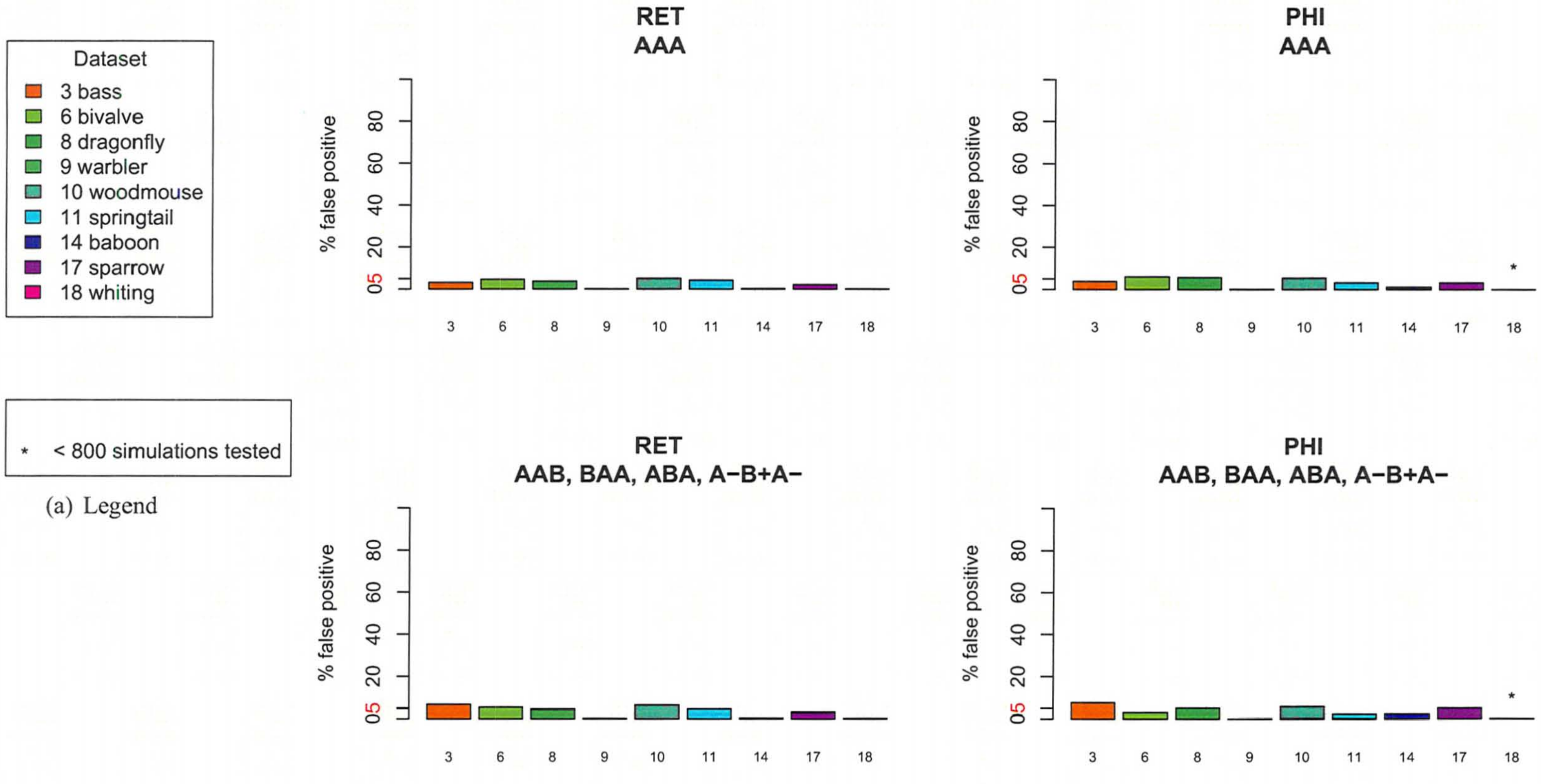
Figure 4.8: False positives for recombination in partition models of animal mtDNA. No heterogeneity ($A = B$). Unless otherwise noted (with *), 1000 simulations were tested per dataset.



(a) Legend

(b) GENECONV and Max χ^2

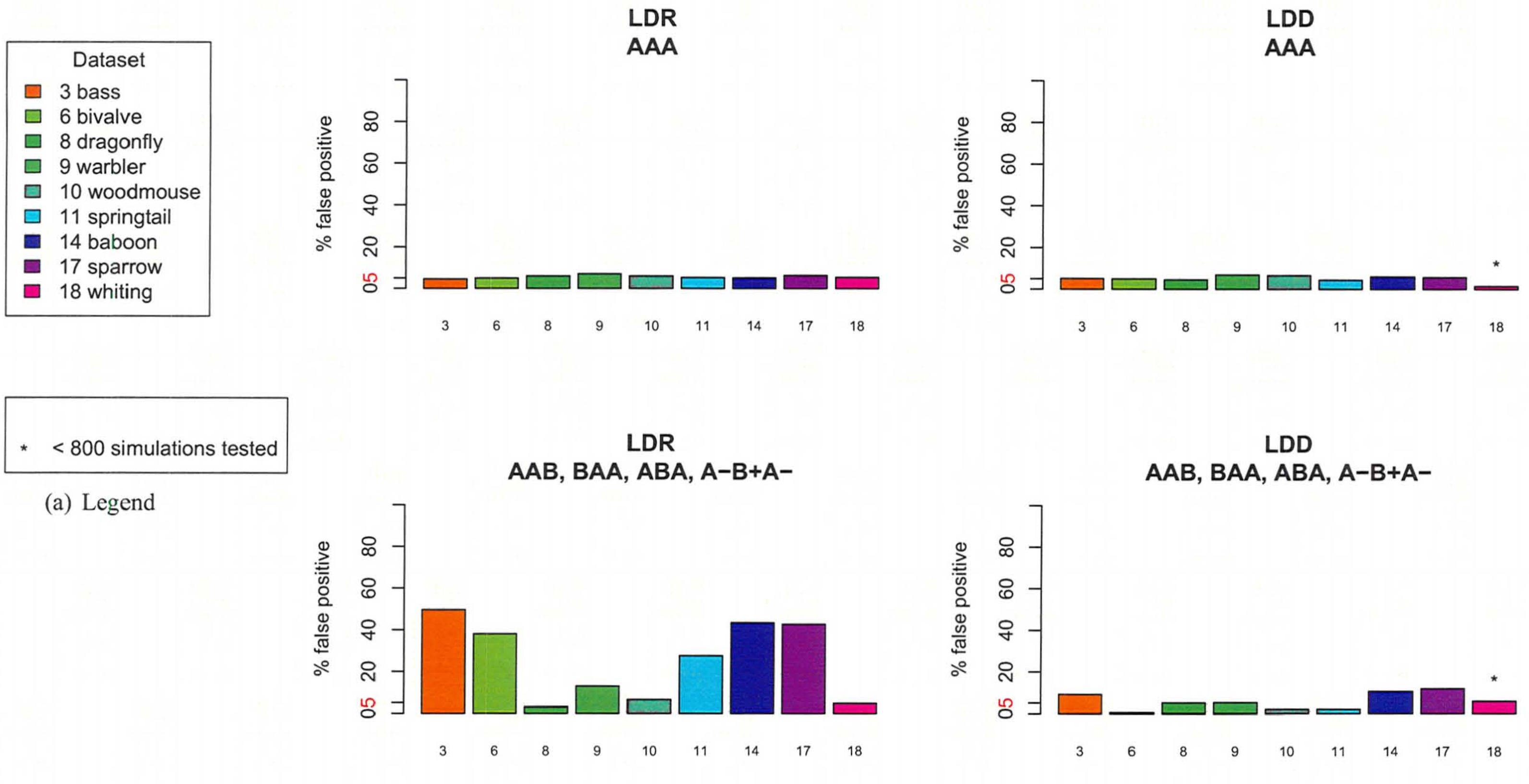
Figure 4.9: False positives for recombination in partition models of animal mtDNA. One-way heterogeneity ($A > B$ or $A < B$). Unless otherwise noted (with *), 1000 simulations were tested per dataset.



(a) Legend

(b) Reticulate and PHI

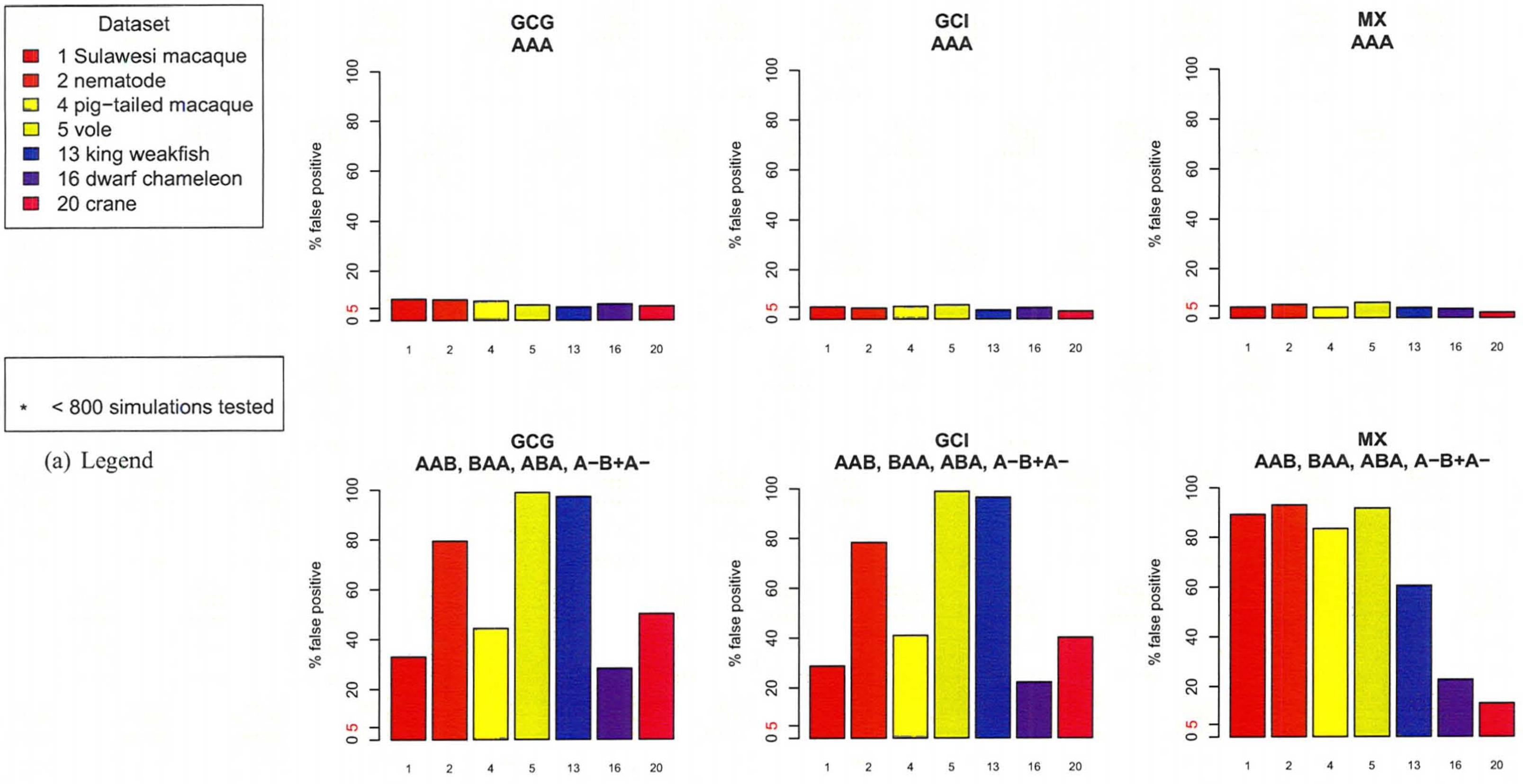
Figure 4.10: False positives for recombination in partition models of animal mtDNA. One-way heterogeneity ($A > B$ or $A < B$). Unless otherwise noted (with *), 1000 simulations were tested per dataset.



(a) Legend

(b) $\mathtt{LDr}^2$ and $\mathtt{LD}|\mathtt{D}'|$

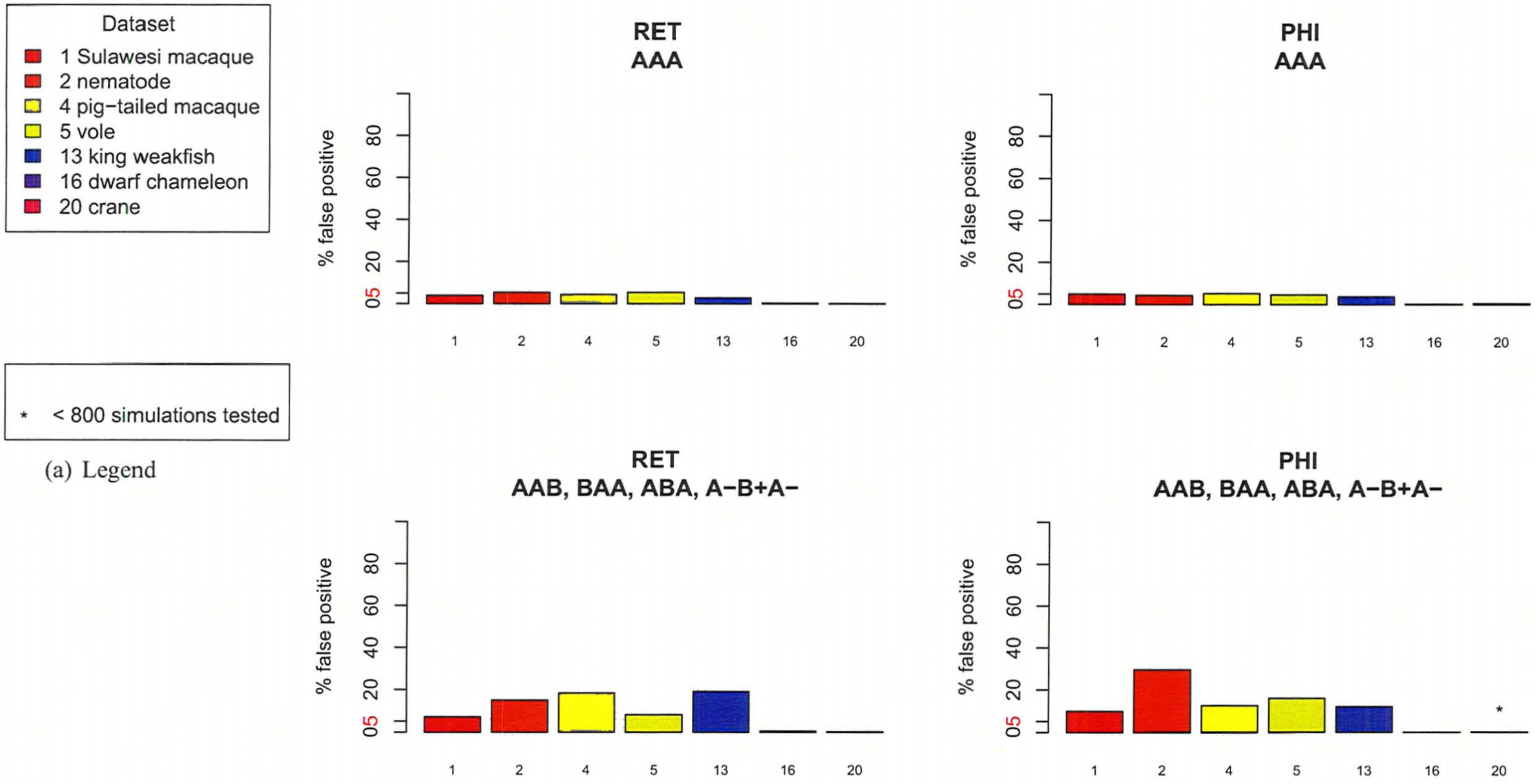
Figure 4.11: False positives for recombination in partition models of animal mtDNA. One-way heterogeneity ($A > B$ or $A < B$). Unless otherwise noted (with *), 1000 simulations were tested per dataset.



(a) Legend

(b) GENECONV and Max χ^2

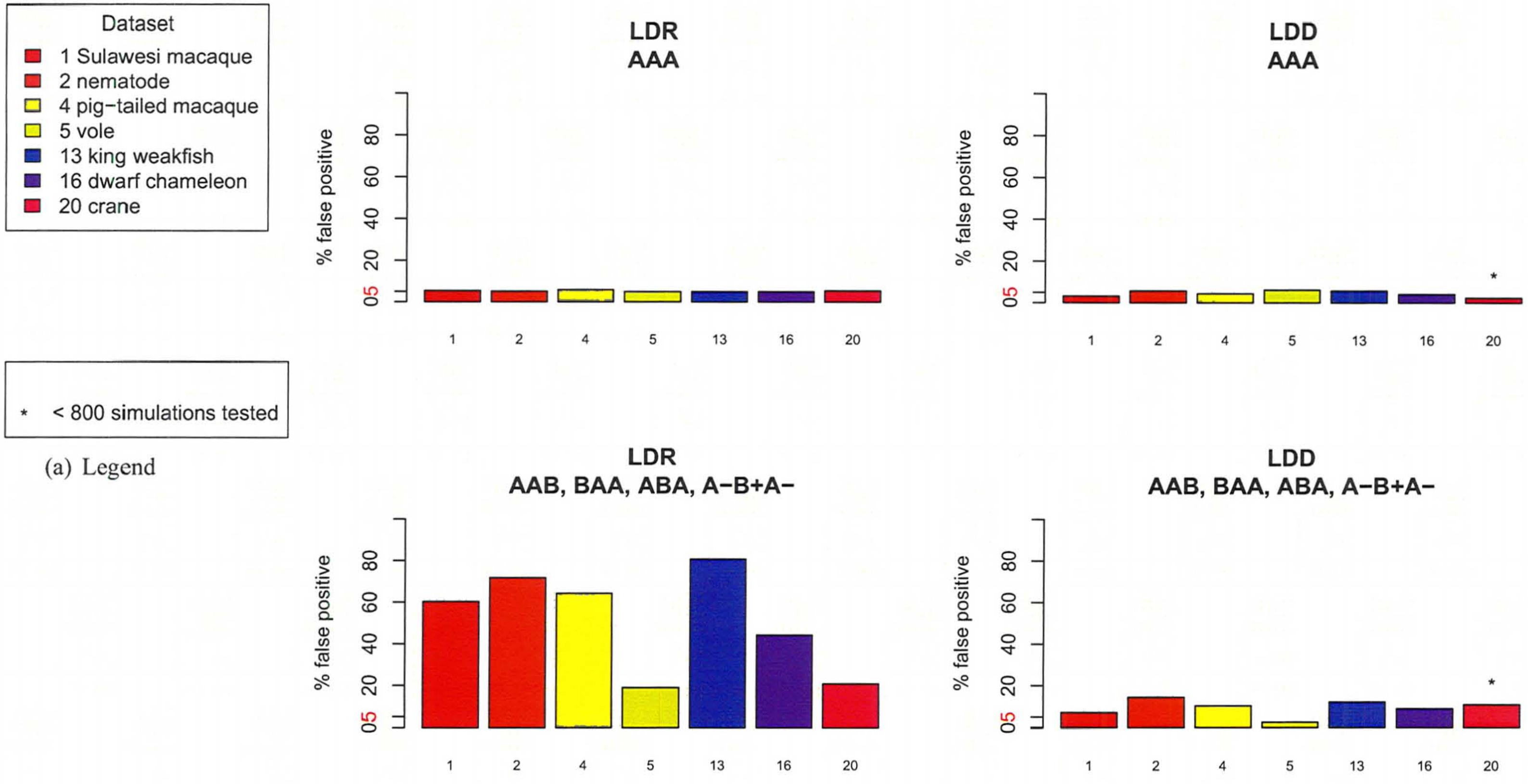
Figure 4.12: False positives for recombination in partition models of animal mtDNA. Two-way heterogeneity ($A > B$ and $A < B$). Unless otherwise noted (with *), 1000 simulations were tested per dataset.



(a) Legend

(b) `Reticulate` and `PHI`

Figure 4.13: False positives for recombination in partition models of animal mtDNA. Two-way heterogeneity ($A > B$ and $A < B$). Unless otherwise noted (with *), 1000 simulations were tested per dataset.



(a) Legend

(b) LDr^2 and LD$|$D$'|$

Figure 4.14: False positives for recombination in partition models of animal mtDNA. Two-way heterogeneity ($A > B$ and $A < B$). Unless otherwise noted (with *), 1000 simulations were tested per dataset.

highly susceptible to mutation cold spots, `Reticulate`, `PHI`, and `LD`$|$`D'`$|$ were sensitive to cold spots under certain limited conditions, such as a small proportion of cold-spot containing sequences, but large proportion of cold sites. We suggested `Reticulate`, `PHI`, and `LD`$|$`D'`$|$ would produce a significant level of false positives when mutation cold spots were present in a fast (or hot) mutation rate background or when a mutation hot spot was also present. The latter is illustrated in the animal datasets with two-way differences (Figure 4.12-4.14). In these datasets, section A is faster evolving than section B in some lineages, but section B is evolving faster than section A in other lineages. Consistent with our hypothesis, simulations off these datasets produced approximately 20% false positives in `Reticulate` and `PHI` and over 10% false positives in `LD`$|$`D'`$|$.

These indirect tests for recombination have been used in studies screening a wide range of animal mtDNA that we did not include in our simulation analysis here (Tsaousis *et al.*, 2005; Ujvari, Dowton and Madsen, 2007). However, it seems likely that simulations from these other datasets would also produce elevated false positives. Other indirect tests that have been used to detect recombination do not distinguish between regional mutation rate heterogeneity, and will likely produce as much or more false positives as the tests evaluated here.

The possibility of widespread recombination in animal mtDNA caught immediate attention as it implied important consequences for phylogenetic and population studies. We have shown that if mtDNA is not recombining, the signals detected by indirect recombination tests (Piganeau, Gardner and Eyre-Walker, 2004; Tsaousis *et al.*, 2005; Ujvari, Dowton and Madsen, 2007) are likely detecting regional mutation rates that change in time, or heterotachous clusters of sites. These results do not refute the possibility that recombination can and does occur in animal mtDNA. Rather, it casts serious doubts on the ability of indirect recombination tests to distinguish between recombination and regional mutation rate heterogeneity, or heterotachy. Ironically, this alternative to recombination also has important implications for phylogenetics, because heterotachy will generally reduce the accuracy of maximum parsimony methods (Philippe *et al.*, 2008), and will always decrease the accuracy of maximum likelihood methods (Kolaczkowski and Thornton, 2004) unless branch lengths of the tree are allowed to differ between sites (Meade and Pagel, 2008).

Part III

FIXATION

Chapter 5

Fixation probabilities of advantageous mtDNA mutations

5.1 Abstract

Analogous to a population within a population within yet another population, the fixation of new mitochondrial mutations is difficult to model given the high degree of polyploidy which complicates the interplay between drift and selection. Using simulations, the fixation probabilities of mitochondrial mutations were investigated under various modes of dominance and levels of polyploidy. Mutations in both single-celled and multi-cellular organisms were simulated under different numbers of mitochondria per cell, cell divisions, and selective pressures. Within-generational drift can either increase or decrease the fixation probability of advantageous mutations, depending on the mode of allelic interaction, or dominance. This is discussed in relation to the number of cell divisions per generation, mtDNA per cell, and heteroplasmy in a cell or individual. Bottlenecks enhance the effect of within-generational drift, but the addition of a somatic cell line, separate from the germ cell line but upon which selection acts, does not affect fixation probabilities.

5.2 Introduction

One fundamental way in which the fixation in mitochondrial genomes differs from fixation in the nuclear genome is the degree to which genetic drift (stochastic changes in allele frequencies in a population) plays a role. Genetic drift occurs in nuclear and mitochondrial genes due to random (i.e. stochastic) partitioning of alleles from one generation to the next during meiosis – a bottleneck event. However, mitochondrial genes undergo additional genetic drift events within generations as organelles are partitioned randomly into daughter cells during mitosis. In addition to within-generation drift, intracellular drift occurs via random replication and turnover of mitochondria and mitochondria genomes (Birky and Skavaril, 1984). Although genetic drift does not preferentially shift allele frequencies towards fixation or loss, alleles found at low frequencies, such as new mutations, are more vulnerable to being lost through these multiple levels of genetic drift than when found at higher frequencies. Genetic drift is more effective when the effective population size, N_e is smaller, leading to random changes in allele frequencies and a decrease in heterozygosity.

A new deleterious mitochondrial mutation does not have very good prospects of surviving to fixation. With hundreds of thousands of copies of mtDNA genomes in a cell, the new mutation will not only be out-replicated and out-sampled by the wild type allele, but selection will also be acting to remove it from the population. An advantageous mutation is also subject to a high level of drift, but if able to survive long enough for the allele frequency to increase, selection may become more predominant than drift and fix the mutation. For example, a dominant allele will rapidly spread through a mitochondria, cell, or population, provided it is not eliminated by drift early after it arises.

Allelic mode of dominance in the face of mtDNA polyploidy is another issue for new mutations. In the incomplete dominant case, selective advantage is proportional to the frequency of the mutant allele. For the recessive case, the mutation is effectively neutral until fixed. At low allele frequencies, most of the allele is in the form of heteroplasmic cells or mitochondria, which has lower or no selective advantage. Until the allele reaches a certain frequency, the effect of selection will be outweighed by the effect of drift.

On the other hand, within-generation drift and polyploidy within an organelle, cell, and individual have been suggested as methods of limiting the spread of deleterious mutations while promoting the spread of advantageous mutants (Takahata and Slatkin, 1983). This was suggested to occur through and increase in the amount of variation in the population on which selection acts, under a model where mutations have an additive effect on fitness – more mutations mean more effect on fitness.

We used computer simulations to investigate the rates of mitochondrial mutation fixation and the interplay between genetic drift and selection under three modes of dominance: complete recessivity, incomplete dominance, and complete dominance. We also explored the effect of bottlenecks and multiple cell lines on the probability of fixing a new mitochondrial mutation. Population dynamics of mitochondrial mutations were simulated under different selective pressures and assumed strict maternal inheritance. Uniparental inheritance of mitochondrial genomes is commonly observed in all animals, with the exception of bivalve mussels. Different organisms use different mechanisms to ensure uniparental inheritance and although these mechanisms are not foolproof (e.g. paternal leakage), generally when considering uniparental inheritance, the only way for heteroplasmy (the presence of more than one type of mitochondrial DNA in an organelle, cell, or individual) to arise is through new mutations.

5.2.1 Expectations

For a mitochondrial mutation, the probability of fixation is $\frac{1}{mN}$ where m is the number of mitochondria per cell, and N is the number of individuals in the population. As the number of mitochondria per cell increase, so does the number of genetic drift events. In terms of changes in allele frequency due to genetic drift, small frequency changes are more likely than large changes. If the initial frequency of a new mitochondrial mutation is very low (according to Piko and Taylor (1987), there can be up to hundreds of thousands of mitochondria per cell in certain tissues of some species), the probability of fixation due to genetic drift will also be very low, and will decrease as the number of mitochondria and cell divisions increases. However, as the number of mitochondria per cell increases, so will does the total number of alleles. This is analogous to an increased population size. With a larger population size, the probability of chance events decreases, so one would expect the fixation probability of a neutral allele to decrease with increased numbers of mitochondria per cell or increased population size.

Selection is expected to be the primary force of evolution in large populations. If this is so, the fixation probability of advantageous alleles should increase as the number of mitochondria per cell increases. However, newly arisen mutations are easily lost due to drift because of low initial allele frequencies. As more layers of drift are added, such as bottlenecks or additional cell lines, there are more opportunities for the alleles to be lost, and the fixation probabilities should decrease.

Bottlenecks

Conventionally, drift and the small N_e of mitochondria is expected to contribute to the overall accumulation and fixation of deleterious mutations in cells, leading to genetic melt-down. Although within-generational drift indeed increases and decreases fixation probability of mildly deleterious and advantageous mutations within a cell, respectively, studies have suggested that drift produces the reverse effect *at the population level*, provided the mitochondrial mutations do not affect fitness of individual mitochondria, but only affect the fitness of individuals (Takahata and Slatkin, 1983; Bergstrom and Pritchard, 1998), and that there is strict uniparental inheritance (Roze, Rousset and Michalakis, 2005). A similar effect has been observed in simulations of mildly deleterious mutations with an mtDNA bottleneck. Tight bottlenecks (20 segregating units or fewer), dramatically increased the relative fitness of individuals. As the bottleneck was tightened, the relative fitness of the most fit individuals increased (Bergstrom and Pritchard, 1998), outweighing the negative effects of having fewer "most fit" individuals, such that the net result was a reduced fixation of mildly deleterious mutations in the population. However, whether such a tight bottlenecks occur in biology has been questioned (Neiman and Taylor, 2009). Generally, bottlenecks are thought to be contributors to the high genetic load of mitochondria. Whether or not a bottleneck affects fixation of advantageous alleles under different modes of dominance has not yet been investigated.

5.3 Method

Life cycle: Overview

The fixation probability of advantageous mitochondrial mutations was simulated for a population with strict maternal inheritance and non-overlapping generations. Each simulation followed the mutant allele frequency in a population of 5000 females from an initial frequency of $\frac{1}{M}$ in one individual, where M is the the number of mtDNA/cell, through multiple generations, to a final outcome of allele fixation or loss. Simulations were repeated until 1000 fixations were observed. The probability of fixation was calculated as 1000 divided by the total number of simulations conducted.

Development: Within-generational drift

Each generation begins with a population of zygotes. Zygotes undergo c cell divisions to reach adulthood. At each cell division, mitochondria are randomly partitioned into daughter cells. This random partitioning is a source of within-generational drift, and was modeled by selecting a random binomial deviate at each division. The randomly selected binomial deviate represents how many mutant mtDNA the daughter cell contains, and is based on the allele frequency of the parent cell. The daughter cell inherits M mitochondria, each of which has a $\frac{m}{M}$ probability of containing a mutant allele, where m is the number of mutant mtDNA/cell. Using a binomial deviate to model drift during cell division is appropriate because inherited mitochondria is either mutant or wild-type so there are only two possible outcomes; mitochondria are partitioned independently of each other.

Adulthood: Selection and modes of dominance

Allele frequencies of the next generation were determined by multiplying allele frequencies by the appropriate relative fitness (according to the mode of dominance). Only beneficial mutations were considered and selection only acted on adults. Each individual possessed between 0 to M mutant mitochondria. Individuals were sorted into $M+1$ bins based on how many mutant mitochondria, m, they possessed. Each bin, therefore, represents the proportion of the population with m mutants, f_m. The value of f_m after selection was calculated according to the mode of allelic dominance. For dominant mutations, f_m when $m > 0$ was

$$f_m \times \frac{(1.0 + s)}{\bar{w_d}}, \tag{5.1}$$

and when $m = 0$:

$$f_0 \times \frac{1}{\bar{w_d}}. \tag{5.2}$$

For an incomplete dominant mutant, selection acts on the individual based upon the proportion of mitochondria with the mutation. The more mutations an individual possesses, the greater the fitness of that individual. For incomplete dominant mutations, f_m after selection was

$$f_m \times \frac{(1 + \frac{m}{M} \times s)}{\bar{w_{id}}}. \tag{5.3}$$

For recessive mutations, f_m when $m < M$ was calculated as

$$f_m \times \frac{1}{\bar{w_r}} \tag{5.4}$$

and when $m = M$,

$$f_M \times \frac{(1.0 + s)}{\bar{w_r}}. \tag{5.5}$$

Two selection coefficients, s, were considered: 0.01 and 1.00. Mean fitness was calculated according to the allelic mode of dominance: complete dominance, incomplete dominance, or complete recessive (Equations 5.6, 5.7, and 5.8, respectively).

$$\bar{w_d} = f_0 + \sum_{m=1}^{M} (f_m \times (1 + s)) \tag{5.6}$$

$$\bar{w_{id}} = \sum_{m=0}^{M} (f_m \times (1 + \frac{m}{M} \times s)) \tag{5.7}$$

$$\bar{w_r} = f_M \times (1 + s) + \sum_{m=0}^{M-1} f_m \tag{5.8}$$

Inheritance: Between-generational drift

Using the f_m values, a cumulative distribution of within-individual allele frequencies was created. This distribution represented the genetic profile of the population in the present generation. The allele frequency at the start of the next generation, f'_m, was determined by selecting a random number from the distribution of f_m after selection for each individual. In this way, f_m classes representing a larger proportion of the population are more likely to be selected but a stochastic element, representing random mating, also had an effect on f'_m. The number of mutants, m, associated with the selected f_m class was inherited by an individual in the next generation. Random selection of f_m classes continued for each individual. A new random number was selected for each individual in the next generation.

Bottleneck

We also considered the effect of bottlenecks on fixation probabilities. A bottleneck was added between each generation, modeling the changes in mtDNA/cell that occur from adults to oocytes to zygotes. The bottleneck restricted the number of mitochondria passed from parent to offspring to $\frac{M}{10}$. During recovery from bottleneck there is unbiased replication of mtDNA such that allele frequencies immediately after the bottleneck are the same upon recovery from the bottleneck. Although one can argue that this bottleneck model disregards changes in allele frequencies during cell divisions in development, observations in heteroplasmic bovine clones suggest mtDNA frequencies are relatively constant during prenatal development (Steinborn *et al.*, 2000).

Multicellularity: Germ + Soma

So far only germ cells have been considered. This is appropriate for modeling unicellular life, but otherwise assumes mitochondria in somatic cells have no effect on the fitness of multi-cellular organisms. This over-simplification was addressed by adding a somatic cell line to the germ only model. In this two cell line model, which is also referred to as the germ+soma model, simulations begin with a single mutant mtDNA in the germ cell line and a single mutation in the somatic cell line of one individual. Fitness is determined by the allele frequencies of somatic cells, but it is the germ line frequencies that are passed to the next generation. Allele frequencies of the two cell lines are otherwise independent, drifting along separate paths during zygote development. Each simulation continues until germ allele frequencies are fixed or lost. Simulations ran until 10000 fixations were observed.

5.4 Results

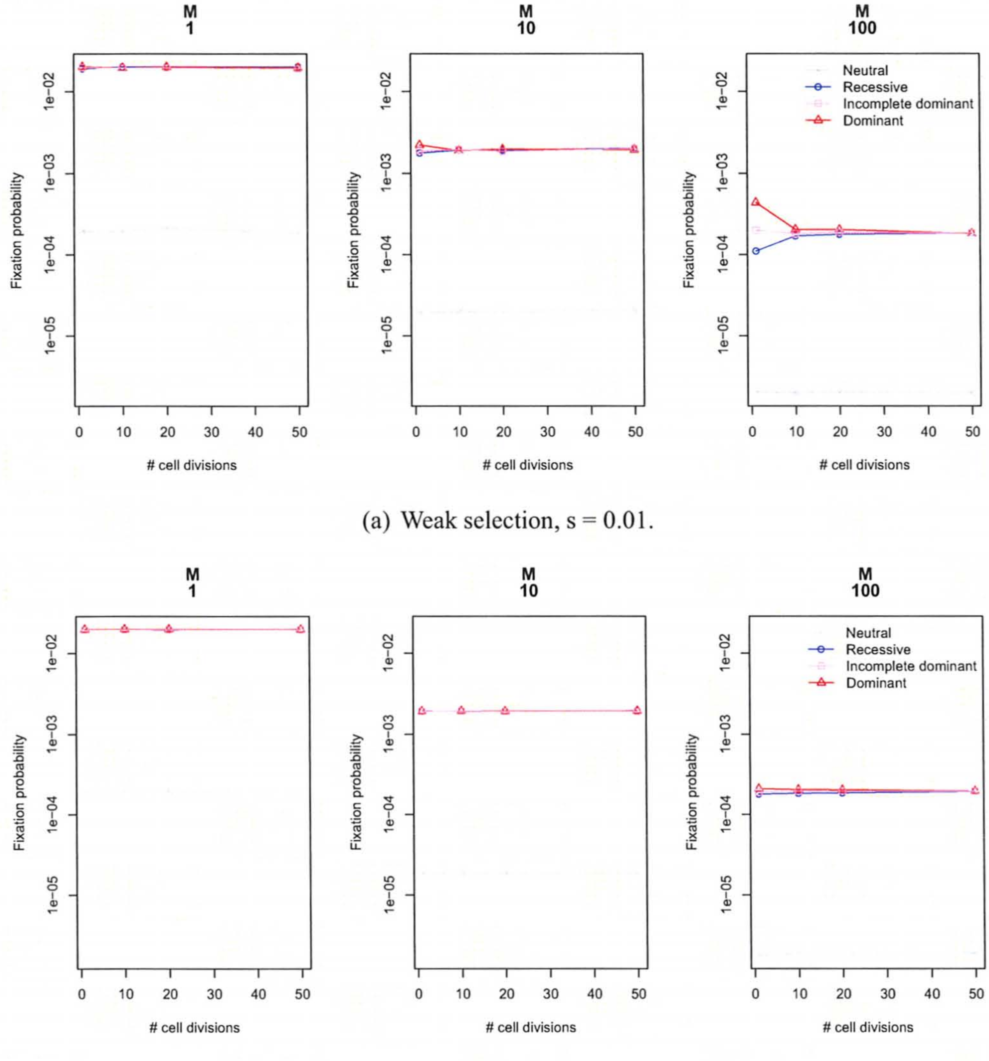
Regardless of selection or mode of allelic interaction, increasing the number of mtDNA/cell decreases fixation probabilities (Figure 5.1 and Figure 5.2). Since all simulations began with one mutant mtDNA in the population, the fixation probabilities of neutral mutants decreased in proportion to the increase in total mtDNA per cell, as expected.

The fixation probabilities of recessive, incomplete dominant, and dominant mutants decreased by varying amounts; the fixation probabilities of recessive mutations decrease the most, while dominant mutations decrease the least. In fact, when selection is strong ($s = 1.00$), the fixation probability of completely dominant mutations remained nearly constant when mtDNA/cell was increased from 10 to 100 (Figure 5.2). The incomplete dominant fixation probability fell between these two extremes.

As the number of cell divisions within a generation, c, increased, fixation probabilities: of completely dominant mutations decreased; of incomplete dominants decreased slightly; and of recessive mutations increased. Exceptions to this trend were when there was one mtDNA/cell, or when selection was weak ($s = 0.01$) and there were 10 mtDNA/cell. In these cases, the fixation probabilities of the selected mutants did not appear to fluctuate with increased cell divisions and appeared indistinguishable from one another. The fixation probabilities of neutral mutations were unaffected by the number of cell divisions per generation.

Under weak selection ($s = 0.01$), the fixation probability of completely recessive mutants is approximately equal to the fixation probability of completely dominant mutants when there are as few as 10 cell divisions/generation (Figure 5.1). Under strong selection ($s = 1.00$), fixation probabilities of recessive and dominant mutants converge between 10 to 20 cell divisions/generation when there are 10 mtDNA/cell, or at more than 50 cell divisions/generation when there are 100 mtDNA/cell.

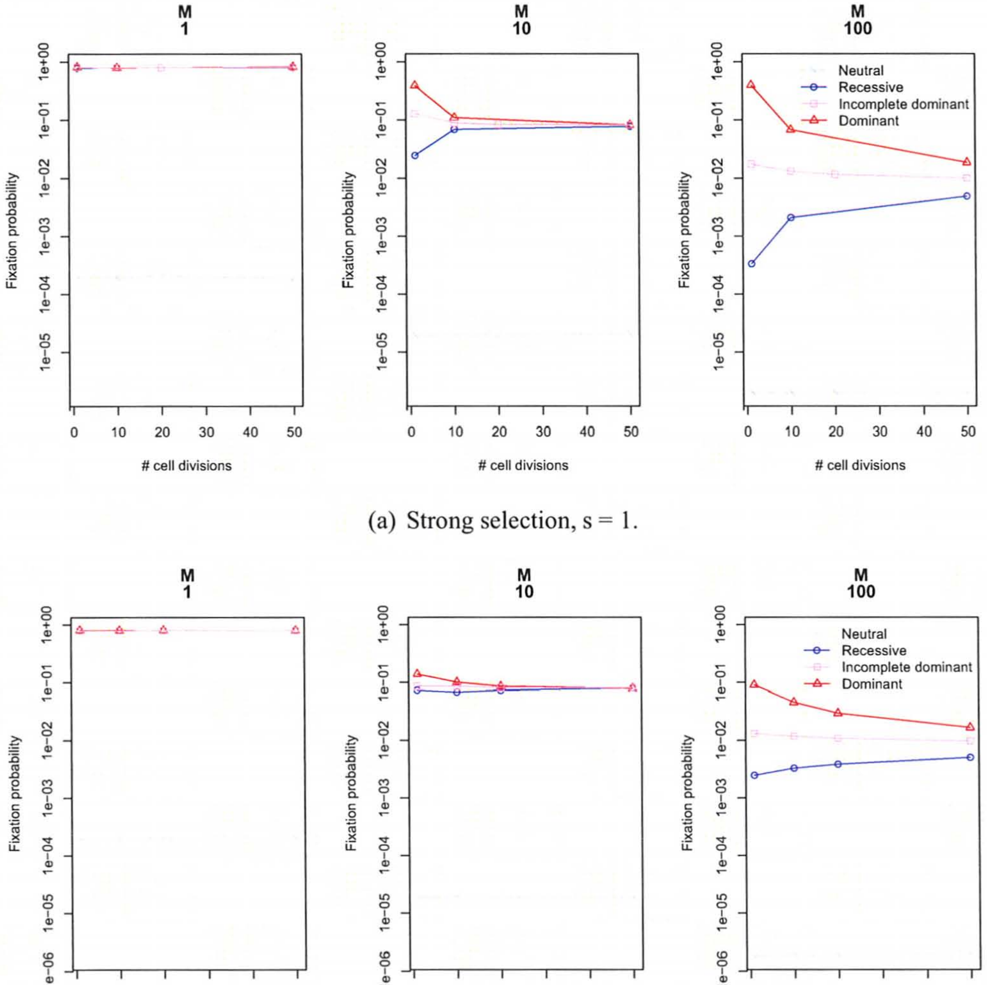
A bottleneck was created which reduced the number of mtDNA passed from adult to zygote from M to $\frac{M}{10}$. Addition of this bottleneck did not decrease the fixation probability of incomplete dominant mutations, nor the fixation probabil-



(a) Weak selection, s = 0.01.

(b) Weak selection, s = 0.01, with bottleneck.

Figure 5.1: Fixation probability of new mitochondrial mutants when $s = 0.01$. Initial frequency $\frac{1}{M}$ where M is the total number of mitochondria in an individual.



(a) Strong selection, s = 1.

(b) Strong selection, s = 1, with bottleneck.

Figure 5.2: Fixation probability of new mitochondrial mutants when $s = 1$. Initial frequency $\frac{1}{M}$ where M is the total number of mitochondria in an individual.

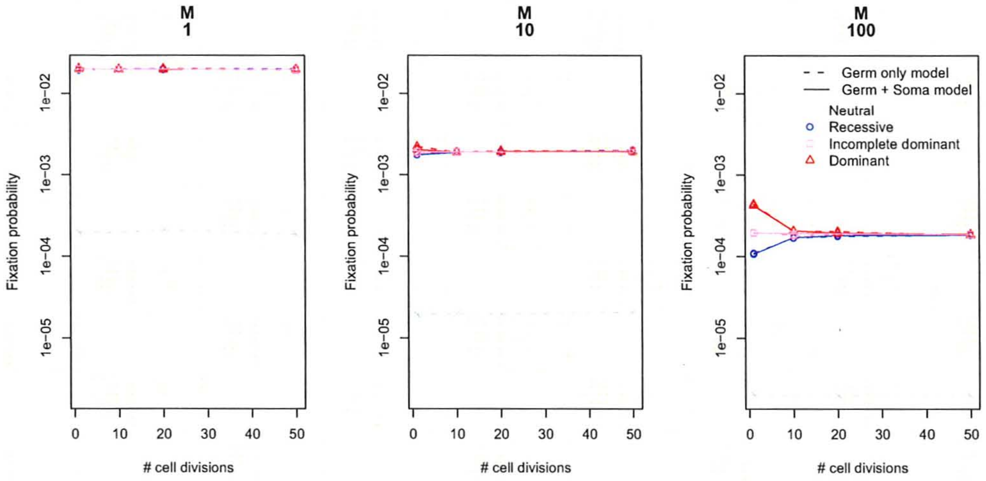


Figure 5.3: Fixation probabilities in a two cell line model. $s = 0.01$. Fitness is determined from somatic allele frequencies and germ allele frequencies are passed on to next generation. Initial frequency in each cell line was $\frac{1}{M}$ where M is the total number of mitochondria in an individual.

ity where mutations converged, but did decrease the spread of fixation probabilities between mutants with different modes of dominance (Figure 5.1 and Figure 5.2). With a bottleneck, the greatest difference between fixation probabilities between dominant to recessive mutants was 0.088 at 100 mtDNA/cell and 1 cell division/generation. Without a bottleneck, the greatest difference between fixation probabilities between dominant to recessive mutants at 100mtDNA/cell and 1 division/generation was 0.390665. In short, including a bottleneck increases the fixation probability of recessive mutations by 1.07×10^{-4} to 2.12×10^{-3} and decreases the fixation probability of dominant mutations by 1.54×10^{-3} to 3.01×10^{-1}. Fixation probabilities of neutral mutations were unaffected.

Addition of a second cell line was expected to introduce an additional level of drift, leading to lower fixation probabilities for new mutations. Interestingly, the additional cell line did not change fixation probabilities (Figure 5.3).

5.5 Discussion

Within-generational drift and dominance

Between-generational drift (drift due to random mating and bottlenecks) is constant because the size of the population does not change. There are two ways of increasing within-generation drift. Depending on the method of increasing drift, within-generational drift can increase or decrease fixation probabilities. The first way to increase within-generational drift is to increase the number of cell divisions per generation. Increasing the number of cell divisions per generation increases the fixation probability of completely recessive advantageous mutations, but decreases the fixation probability of completely dominant advantageous mutations. The point where fixation probabilities of recessive and dominant mutations converge depends on the number of cell divisions per generation, the number of genomes per cell, and the selection coefficient. The fixation probability where alleles with different modes of dominance will eventually converge can be estimated from fixation probability of incomplete dominant allele. The other method of increasing within-generational drift is to increase the number of mtDNA/cell. Since all mutations in this model are effectively neutral within-generations, this increasing the number of mtDNA/cell decreases the fixation probability by decreasing the initial frequency.

If selection is the primary driver of evolution, not only will the fixation probabilities of the advantageous mutant be much greater than the fixation probability of neutral mutations, but the fixation probabilities of mutants with different modes of allelic interaction will be equal. This is most clearly illustrated by simulations with one mitochondrial genome per cell. Within-generation drift is eliminated because there can be no sampling effects when the entire pool is retained between generations. Although there is still between-generation drift through random mating, fixation probabilities of mutants with different modes of allelic interaction are virtually indistinguishable.

This study demonstrates the interplay between drift and selection as exhibited by the fixation probabilities of new mitochondrial mutations. In some respects, mitochondrial mutations behave the same as nuclear mutations, such as when selection or drift is removed (neutral mutations or $M = 1$, respectively). As observed here, the probability of fixation for neutral mutations in the population of N females equals the initial mutation frequency, or $\frac{1}{M \times N}$. When $M = 1$ all cells carrying the mutant are immediately fixed so the mode of dominance is no longer a factor. Simulation fixation probabilities of all advantageous mutations equaled 0.02 when $s = 0.01$ and 0.8 when $s = 1$. These values follow the expectation for fixation probabilities of advantageous mutations:

$$\frac{1 - e^{-2s}}{1 - e^{-4Ns}}. \tag{5.9}$$

Fixation probabilities of mitochondrial mutants when $M > 1$ and when $s \neq 0$ become a question of the dynamics of fixation within a cell or individual, as well as the dynamics of fixation between cells. The probability of fixation due to drift within a cell equals the initial allele frequency, $\frac{1}{M}$; increasing M decreases the probability of fixation through drift alone. This is reflected in the observed decrease in fixation probability when M is increased. The decrease in fixation probability is proportional to the increase in M, with the exception of dominant mutations under strong selection, where the drop in fixation probability is less severe. In fact, when $s = 1$ and $c \leq 10$, fixation probabilities of dominant mutations between $M = 10$ and $M = 100$ are approximately equal. These conditions – strong selection and low c – decrease the effect of drift relative to selection.

Our results show that within-generation drift can both increase or decrease fixation probabilities of new mitochondrial mutations. This can be explained by the effect of within-generation drift on genetic diversity within the cell, or heteroplasmy. In these simulations, a mutation arises once and only once in a single individual. The allele frequency of the mutant is followed until the mutation is fixed in the en-

Table 5.1: Probability a cell is heteroplasmic after c cell divisions from zygote to adult, K_a. K_z is the probability of the cell being heteroplasmic as a zygote.

(a) K_a in the first generation ($K_z = \frac{1}{N}$, $N = 5000$ females)

M	$c = 1$	$c = 10$	$c = 20$	$c = 50$
10	1.8×10^{-4}	7.0×10^{-5}	2.4×10^{-5}	1.0×10^{-6}
100	2.0×10^{-4}	1.8×10^{-4}	1.6×10^{-4}	1.2×10^{-4}

(b) K_a in any generation.

M	$c = 1$	$c = 10$	$c = 20$	$c = 50$
10	$9.0 \times 10^{-1} K_z$	$3.5 \times 10^{-1} K_z$	$1.2 \times 10^{-1} K_z$	$5.2 \times 10^{-3} K_z$
100	$9.9 \times 10^{-1} K_z$	$9.0 \times 10^{-1} K_z$	$8.1 \times 10^{-1} K_z$	$6.1 \times 10^{-1} K_z$

tire population. Critical to fixation in the population is fixation within the original mutation-carrying individual. When the mutation is fixed in the cell, the individual is more likely to reproduce due to a fitness advantage conferred by the mutation. Before the mutation is fixed within the individual, however, fitness advantage is conferred depending on the mode of allelic interaction, or dominance, of the mutation. During the heteroplasmic state, dominant mutations will confer an advantage equal to the advantage conferred when the cell is homoplasmic for the mutant, incomplete dominant mutations confer an advantage that is proportional to the allele frequency within the cell, and recessive mutations confer no advantage.

Homoplasmy and fixation

When adult cells are homoplasmic, fixation probabilities of advantageous mutations are equal, regardless of the mode of dominance. Mutations with different modes of dominance, then, will fix at equal probabilities under conditions that facilitate homoplasmy, such as when $M = 1$, or with an increase in c. Birky, Maruyama and Fuerst (1983) showed that the probability an adult cell being heteroplasmic, K_a, is

$$K_a = (1 - \frac{1}{M})^c K_z, \tag{5.10}$$

where K_z is the probability a zygote cell is heteroplasmic. In other words, cells are increasingly likely to be homoplasmic when K_a is small, which occurs when M is low or c is high. This is supported by the fixation probabilities reported here, which converge as $K_a \rightarrow 0$ (Table 5.1). When M is small, K_a decreases dramatically with increasing c. When $K_a > 0$, dominant mutations still confer a fitness advantage, but recessive mutations, regardless of s, are effectively neutral. However, the fixation probabilities of dominant and recessive mutations converge as $K_a \rightarrow 0$. At this point, cells are almost always homoplasmic. Decreased heteroplasmy leads to increased fixation probability of recessive mutations because recessive mutants are essentially hidden in heteroplasmic cells. When $K_a \rightarrow 0$, the proportion of time cells are heteroplasmic could be shifted to cells being homoplasmic and lost, which does not change the playing field for recessive mutations, or could be shifted to cells being homoplasmic and fixed, in which case fitness advantage is conferred. In contrast, decreased heteroplasmy decreases fixation probabilities of dominant mutations because the missing heteroplasmy would have conferred a selective advantage, but it now the mutation is either fixed (no better off than before) or lost. In other words, as within-generation drift is increased, fixation probabilities of recessive mutations have nowhere to go but up, and dominant mutations have nowhere to go but down. Based on Table 5.1 and Figure 5.1 for $s = 0.01$, fixation probabilities of advantageous mutations converge when $K_a \leq 1.8 \times 10^{-4}$. When selection is stronger (increased s or M), mutations are more robust to decreased heteroplasmy and fixation probabilities will converge at a lower K_a (or higher c).

Bottlenecks

Conventionally, drift and the small N_e of mitochondria is expected to contribute to the overall accumulation and fixation of deleterious mutations in cells. Without recombination, this is expected to develop an unstable genetic load, leading to genetic melt-down; an expectation termed "Muller's Ratchet" by Felsenstein (1974). Although within-generational drift indeed increases and decreases fixation probability of mildly deleterious and advantageous mutations within a cell, respectively, studies have suggested that drift produces the reverse effect *at the population level*, provided the mitochondrial mutations do not affect fitness of individual mitochondria, but only affect the fitness of individuals (Takahata and Slatkin, 1983; Bergstrom and Pritchard, 1998), and that there is strict uniparental inheritance (Roze, Rousset and Michalakis, 2005). A similar effect has been observed in simulations of mildly deleterious mutations with an mtDNA bottleneck. Tight bottlenecks (20 segregating units or fewer), dramatically increased the relative fitness of individuals. As the bottleneck was tightened, the relative fitness of the most fit individuals increased (Bergstrom and Pritchard, 1998), outweighing the negative effects of having fewer "most fit" individuals, such that the net result was a reduced fixation of mildly deleterious mutations in the population. Whether such a tight bottlenecks occur in biology has been questioned (Neiman and Taylor, 2009), but even wide bottlenecks can contribute to mitochondrial stability by slowing the accumulation of mildly deleterious mutations *within a cell* (Bergstrom and Pritchard, 1998).

When $M = 10$, the bottleneck reduced the number of mtDNA/cell in zygotes to 1 mtDNA, and when $M = 100$, the bottleneck reduced the mtDNA/cell to 10 segregating units. We observed that the bottleneck did not change fixation probabilities of incomplete dominant mutations, or the fixation probability to which dominant and recessive mutations approached as $c \to \infty$, but the bottleneck did decrease the c where fixation probabilities converged. This can be explained by decreased genetic variability *within a cell* by the bottleneck; bottlenecks decrease K_a for a given c by decreasing N_e. Our results are supported by Roze, Rousset and Michalakis (2005), who looked at the effect of bottlenecks and parental inheritance on fixation probabilities under a more complex selection model, where selection can occur between mitochondria as well as between cells or individuals. They found that bottlenecks have no effect on fixation probabilities when there is strict uniparental inheritance, and limited effect when mitochondria are neutral within the cell.

Bottlenecks affect fixation probabilities in two ways. They may be beneficial to the population, increasing fixation of advantageous mutations while decreasing fixation of deleterious mutations increasing the amount of genetic variation between cells (Bergstrom and Pritchard, 1998). Conversely, they may lead in an opposite fashion, decreasing N_e and increasing fixation of deleterious mutations, or loss of advantageous mutations, through drift. Simulation studies of mutations that are selected for at the individual level but neutral at the mitochondrial level, have shown that bottlenecks increase the efficiency of selection against deleterious mutations (Bergstrom and Pritchard, 1998); the tighter the bottleneck, the more efficient the selection. These results are supported by simulations of within-generational drift, which was shown to encourage fixation of advantageous mutations, and the loss of deleterious mutations (Takahata and Slatkin, 1983). However, bottlenecks have a drastically different effect when biparental inheritance is allowed, even at low levels of less than 1:10 ratio of paternal to maternal mtDNA (Roze, Rousset and Michalakis, 2005). When there is paternal leakage, bottlenecks increase selection between mitochondria but decrease selection between cells. Fixation probabilities of mutations that are advantageous on the mitochondrial level will increase, while fixation probabilities of deleterious mutations decrease, regardless of the effect these mutations have on the individual's fitness. These results suggest that strict uniparental inheritance, combined with bottlenecks, do not contribute to Muller's Ratchet, but rather will actually help mitigate it.

Theoretical and empirical observations of mtDNA population dynamics support the benefits of within-generational drift in aiding mtDNA stability. In this study, the initial mutation arose at the beginning of development, just after the bottleneck. Other studies have investigated the importance of timing in fixation of new mutants. A mutation that arises just before a bottleneck must survive the bottleneck to continue to fixation, while a mutation that arises at the beginning of development must survive through c cell divisions as well as the bottleneck. Interestingly, Wahl and Gerrish (2001) found that advantageous mutations have a better chance of fixation if they arise early in the growth period long before the bottleneck, when the population is small, rather than if they arise late just before the bottleneck, when the population is large. This is particularly interesting considering early mutations (ones that arise early in the growth period) must survive more cell divisions (opportunities for drift), as well as the observation that more mutations tend to occur late in the growth period due to the larger population size. This suggests that the increased initial frequency of early-arising mutations (mutations that arise when the population is still small) the initial frequency of any allele is greater, thereby increasing the fixation probability through drift in the upcoming cell divisions. Frederiksen *et al.* (2006) followed the frequency of a pathogenic mtDNA point mutation (3243A→G) within human individuals and observed that the frequency generally decreased with age, except in the case of non-dividing cells. Mutation load decreased with age in blood, cheek cells (buccal mucosa), and urine epithelial cells, but not in muscle cells, which do not undergo mitosis (Frederiksen *et al.*, 2006). The authors suggested this was due to purifying selection during mitosis. That is, this mutation, which is deleterious on the host-level, may also be deleterious on the mitochondrial level. If this is so, the effect of the mutation within the cell must be deleterious on mutant mtDNA segregation, but neutral in relation to mutant mtDNA replication, as mitochondria continue to divide within muscle cells, despite the lack of mitosis. An alternative, or perhaps supplemental, explanation is that the decreased mutation load in dividing cells reflects the effect of within-generational drift in helping to ensure genetic stability of mtDNA (Takahata and Slatkin, 1983). Differences in the rate of mutation frequency decrease among the dividing tissue types are probably correlated with differences in the rate of cell division.

Germ + Somatic cells

The addition of a second cell line and the separation of germ frequencies from selection did not affect fixation probabilities. This is probably because the level of heteroplasmy in both somatic and germ cells was equally low; the effect of separating germ frequencies from selection, if any, were outweighed by the effect of heteroplasmy. In other words, in the germ only model, fixation probabilities were driven primarily by drift. The contribution from adding another level of drift did virtually nothing to change this balance between drift and selection.

Multiple levels of selection

Like a population within a population within a population, the fixation of mitochondrial mutants is determined by genetic drift and selection on three different fitness components (intracellular, intercellular, and individual) which can be affected in different ways by a single mutation. The mutant allele frequencies and substitution rates of each level may be correlated, but each will be determined by separate genetic drift events, and may have different selective pressures (Backer and Birky, 1985; Walsh, 1992; Birky and Walsh, 1992). Any non-neutral mutation can affect different components of fitness in different ways. For example, a mutant mitochondrial genome may affect the replication rate of the organelle, the fitness of the cell, or the fitness of the individual – or all three at once, possibly even such that one fitness component is increased while another is unaffected or decreased. The model studied here only implemented selection at the host or individual level. Simulation studies have suggested that the addition of selection at intercellular or intracellular level does not change fixation probabilities, so long as strict uniparental inheritance is maintained (Roze, Rousset and Michalakis, 2005). When uniparental inheritance is compromised (e.g. paternal leakage), fixation probabilities of advantageous and deleterious mutations become less sensitive to individual-level selection, and more sensitive to intra- and intercellular selection. That is, fixation of mutations advantageous to the individual but deleterious to the mitochondria will decrease, while fixation of mutations deleterious to the individual but advantageous to the mitochondria will increase. This effect is especially emphasized when bottlenecks are added to the model, underlining the importance of strict uniparental inheritance in countering Muller's Ratchet in the face of extensive within-generational drift in mitochondrial evolution. Nevertheless, as far as our model is concerned, the addition of inter-mitochondrial and inter-cellular selection should not change our results here, so long as strict uniparental inheritance is assumed.

Linkage

Complete linkage of loci on the mtDNA genome reduces the efficiency of selection. Neutral or mildly deleterious mutations may become fixed due to linkage (or "hitchhiking") with a strongly advantageous allele. The reverse may also be true: advantageous mutations linked to the more common deleterious mutations may be lost as the deleterious mutations are selected against ("background selection"). In the mitochondrial genome, where all loci are assumed as being completely linked (although, see Part II), hitchhiking and background selection no doubt play an important role in the fixation of mitochondrial mutations. The effects of linkage on new mutant fixation was investigated by Birky and Walsh (1988) with interesting results. While neutral mutations are unaffected by complete linkage to either advantageous or deleterious mutations, fixation of advantageous mutations is always decreased, and fixation of deleterious mutations is always increased, when there is complete linkage – even if linked to an advantageous, or deleterious mutation, respectively. They reason that linkage decreases N_e, leading to increased effect of drift relative to selection. These results were obtained in simulations and/or mathematical models of incomplete dominant mutations in diploid populations where selection acted upon the the loci. While this suggests linkage is detrimental to the survival of a diploid cell, linkage may produce different results in a mitochondrial system, which is polyploid and where there is extensive drift. As discussed previously, within-generational drift in a mitochondrial system seems to improve the efficiency of selection in the population, while decreasing the efficiency of selection within a cell. This within-generational effect should be even more emphasized if linkage on the mtDNA genome strengthens the effect of drift.

Heteroplasmy within mitochondria

There is yet another level of polyploidy that was not considered here, and that is the presence of multiple genomes within each mitochondria. There can be up to 10 mitochondrial genomes within each organelle (Nass, 1969a). So realistically, a new mitochondrial mutation may arise in one of the multiple mitochondrial genomes and one would have to account for fixation probability within an organelle. We assumed that when a new mitochondrial mutation arose, it was already fixed within an organelle. This simplification is often made in simulations of mitochondrial population genetics, and is supported by empirically by human germ-line cells, where mitochondria appear to effectively contain one mtDNA chromosome each (Jansen, 2000).

5.6 Conclusion

With high levels of drift, a high mutation rate, and a lack of recombination, evolutionary biologists have puzzled over the stability of mtDNA. Hypotheses to explain the persistence of this unique genome have included strong selection, or low levels of recombination, or a role for drift in the maintenance of mtDNA. Here, we used a dominance model to describe the interplay between selection and various levels of drift: within-generations, between-generations (bottleneck), and with both unicellular and multicellular models. Our results show that recessive mutations, or mutations that are effectively neutral until fixed, have a higher chance of fixation when drift is maximized, but dominant mutations, mutations that always confer the maximum fitness advantage, regardless of the allele frequency, have a decreased chance of fixation when drift is maximized; differences which are likely related to the effect of drift on levels of heteroplasmy. Our results are consistent with the hypothesis that within-generational drift can increase genetic stability under unique mitochondrial genetics, but demonstrate that this is so only under certain modes of dominance.

Part IV

CONCLUSION

Conclusion

This thesis addressed two fundamental aspects of mitochondrial population genetics. The first, clonal inheritance of mtDNA in animals, focused on alternatives to the detection of recombination in a wide range of animal datasets. It was demonstrated that while macaque-specific mtDNA models were insufficient, a general model of mutation rate heterogeneity, namely mutation cold spots, was a viable alternative to recombination. We then directly placed mutation cold spots in a biological context, using animal mtDNA to model cold spots in simulations without recombination. An extremely high level of recombination false positives was detected for all tests. These results directly demonstrate the limited ability of indirect tests to distinguish recombination from mutation rate heterogeneity, and cast doubt upon numerous reports of recombination in animal mtDNA.

This thesis also addressed the effect of dominance on the probability that a new mitochondrial mutation will become fixed in a population. Mitochondrial evolution is characterized by extensive drift, both between and within generations. We showed that whether within-generational drift increases or decreases fixation of advantageous mitochondrial alleles depends on the allele's the mode of dominance; as drift is increased, dominant mutations become less likely to fix but recessive mutations become more likely to fix. Adding a bottleneck decreased the difference in fixation probabilities between dominant and recessive mutations. Surprisingly, separating selection from allele frequencies in the germ line did not change the fixation probabilities. These results support the theory that drift may play an important role in maintaining mitochondrial evolutionary stability.

Part V

BIBLIOGRAPHY

Bibliography

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Altmann, R. (1890). *Die Elementaroganismen und ihre Beziehungen zu den Zellen.* Leipzig: Verlag von Veit & Comp.

Altschul, S. F. (1993). A protein alignment scoring system sensitive at all evolutionary distances. *Journal of Molecular Evolution 36*, 290–300.

Ashley, M. V., P. J. Laipis, and W. W. Hauswirth (1989). Rapid segregation of heteroplasmic bovine mitochondria. *Nucleic Acids Research 17*, 7325–7331.

Awadalla, A., A. Eyre-Walker, and J. Maynard Smith (2000). Questioning evidence for recombination in human mitochondrial DNA. *Science 288*, 1931a.

Awadalla, P., A. Eyre-Walker, and J. Maynard-Smith (1999). Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science 286*, 2524–2525.

Backer, J. S. and C. W. J. Birky (1985). The origin of mutant cells: mechanisms by which *Saccharomyces cerevisiae* produces cells homoplasmic for new mitochondrial mutations. *Curr. Genet. 9*, 627–640.

Balaban, R. S., S. Nemoto, and T. Finkel (2005). Mitochondria, oxidants, and aging. *Cell 120*, 483–495.

Bereiter-Hahn, J. and M. Vöth (1994). Dynamics of mitochondria in living cells: Shape changes, dislocations, fusion, and fission of mitochondria. *Microscopy Research and Technique 27*, 198–219.

Bergstrom, C. T. and J. Pritchard (1998). Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes. *Genetics 149*, 2135–2146.

Bernatchez, L., H. Glémet, C. C. Wilson, and R. G. Danzmann (1995). Introgression and fixation of Arctic char (*Salvelinus alpinus*) mitochondrial genome in an allopatric population of brook trout (*Salvelinus fontinalils*). *Can. J. Fish. Aquat. Sci. 52*, 179–185.

Birky, C. W. J. (1983). Relaxed cellular controls and organelle heredity. *Science 222*, 468–475.

Birky, C. W. J. (2001). The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annu. Rev. Genet. 35*, 125–148.

Birky, C. W. J., T. Maruyama, and P. Fuerst (1983). An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. *Genetics 103*, 513–527.

Birky, C. W. J. and R. V. Skavaril (1984). Random partitioning of cytoplasmic organelles at cell division: the effect of organelle and cell volume. *J. Theor. Biol. 106*, 441–447.

Birky, C. W. J. and J. B. Walsh (1988). Effects of linkage on rates of molecular evolution. *Proc. Natl. Acad. Sci. U. S. A. 85*, 6414–6418.

Birky, C. W. J. and J. B. Walsh (1992). Biased gene conversion, copy number and apparent mutation rate differences within chloroplast and bacterial genomes. *Genetics 130*, 677–683.

Birky, C. W. J., C. Wolf, H. Maughan, L. Herbertson, and E. Henry (2005). Speciation and selection without sex. *Hydrobiologia 546*, 29–45.

Bogenhagen, D. and D. A. Clayton (1977). Mouse L Cell Mitochondrial DNA Molecules Are Selected Randomly for Replication throughout the Cell Cycle. *Cell 11*, 719–727.

Brown, W. M., M. J. George, and A. C. Wilson (1979). Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. USA 76*, 1967–1971.

Bruen, T. C., H. Philippe, and D. Bryant (2006). A simple and robust statistical test for detecting the presence of recombination. *Genetics 172*, 2665–2681.

Bynum, E. L., D. Z. Bynum, and J. Supriatna (1997). Confirmation and location of the hybrid zone between wild populations of *Macaca tonkeana* and *Macaca hecki* in Central Sulawesi, Indonesia. *American Journal of Primatology 43*, 181–209.

Ciani, A. C., R. Stanyon, W. Scheffrahn, and B. Sampurno (1988). Evidence of gene flow between Sulawesi macaques. *American Journal of Primatology 17*, 257–270.

Ciborowski, K. L., S. Consuegra, C. García de Leániz, M. A. Beaumont, J. Wang, and W. C. Jordan (2007). Rare and fleeting: an example of interspecific recombination in animal mitochondrial DNA. *Biology Letters 3*, 554–557.

Clayton, D. A. (1982). Replication of animal mitochondrial DNA. *Cell 28*, 693–705.

Cummins, J. (1998). Mitochondrial DNA in mammalian reproduction. *Reviews of Reproduction 3*, 172–182.

DiMauro, S. (2001). Lessons from mitochondrial DNA mutations. *Cell & Developmental Biology 9*, 397–405.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research 32*, 1792–1797.

Evans, B. J., J. A. McGuire, R. M. Brown, N. Andayani, and J. Supriatna (2008). A coalescent framework for comparing alternative models of population structure with genetic data: evolution of Celebes toads. *Biology Letters 4*, 430–433.

Evans, B. J., J. C. Morales, J. Supriatna, and D. J. Melnick (1999). Origin of the Sulawesi macaques (Cercopithecidae: *Macaca*) as suggested by mitochondrial DNA phylogeny. *Biological Journal of the Linnean Society 66*, 539–560.

Evans, B. J., J. Supriatna, and D. J. Melnick (2001). Hybridization and population genetics of two macaque species in Sulawesi, Indonesia. *Evolution 55*, 1686–1702.

Eyre-Walker, A., N. H. Smith, and J. Maynard Smith (1999). How clonal are human mitochondria? *Proc. R. Soc. Lond. B. 266*, 477–483.

Felsenstein, J. (1974). The Evolutionary Advantage of Recombination. *Genetics 78*, 737–756.

Fitch, W. M. (1971). Rate of change of concomitantly variable codons. *Journal of Molecular Evolution 1*, 84–96.

Fitzgerald, J., H.-H. M. Dahl, I. B. Jakobsen, and S. Easteal (1996). Evolution of mammalian X-linked and autosomal *Pgk* and *Pdh* $E1\alpha$ subunit genes. *Molecular Biology and Evolution 13*, 1023–1031.

Frederiksen, A. L., P. H. Andersen, K. O. Kyvik, T. D. Jeppesen, J. Vissing, and M. Schwartz (2006). Tissue specific distribution of the 3243A→G mtDNA mutation. *Journal of Medical Genetics 43*, 671–677.

Galtier, N., D. Enard, Y. Radondy, E. Bazin, and K. Belkhir (2006). Mutation hot spots in mammalian mitochondrial DNA. *Genome Research 16*, 1–8.

Gantenbein, B., V. Fet, I. A. Gantenbein-Ritter, and F. Balloux (2005). Evidence for recombination in scorpion mitochondrial DNA (Scorpiones: Buthidae). *Proc. R. Soc. B. 272*, 697–704.

Graur, D. and W. H. Li (2000). *Fundamentals of Molecular Evolution* (second ed.). Sunderland, Massachusetts, USA: Sinauer Associates, Inc.

Hebert, P. D. N., E. H. Penton, J. M. Burns, D. H. Janzen, and W. Hallwachs (2004). Ten species in one: DNA barcoding reveals cryptic speacies in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc. Natl. Acad. Sci. U. S. A. 101*, 14812–14817.

Hill, W. G. and A. Robertson (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics 38*, 226–231.

Hoeh, W. R., D. T. Stewart, C. Saavedra, B. W. Sutherland, and E. Zouros (1997). Phylogenetic evidence for role-reversals of gender-associated mitochondrial DNA in *Mytilus* (Bivalvia: Mytilidae). *Soc. Biol. Evol. 14*, 959–967.

Hollingsworth, P. M., L. L. Forrest, J. L. Spouge, M. Hajibabael, S. Ratnasingham, M. van der Bank, M. W. Chase, R. S. Cowan, D. L. Erickson, A. J. Fazekas, S. W. Graham, K. E. James, K. J. Kim, W. J. Kress, H. Schnaeider, J. van AlphenStahl, S. C. H. Barrett, C. van den Berg, D. Bogarin, K. S. Burgess, K. M. Cameron, M. Carine, J. Chacón, A. Clark, J. J. Clarkson, F. Conrad, D. S. Devey, C. S. Ford, T. A. J. Hedderson, M. L. Hollingsworth, B. C. Husband, L. J. Kelly, P. R. Kesanakurti, J. S. Kim, Y. D. Kim, R. Lahaye, H. L. Lee, D. G. Long, S. Madriñán, O. Maurin, I. Meusnier, S. G. Newmaster, C. W. Park, D. M. Percy, G. Petersen, J. E. Richardson, G. A. Salazar, V. Savolainen, O. Seberg, M. J. Wilkinson, D. K. Yi, and D. P. Little (2009). A DNA barcode for land plants. *PNAS 106*, 12794–112797.

Howell, N. (1997). Leber hereditary optic neuropathy: Mitochondrial mutations and degeneration of the optic nerve. *Vision Research 37*, 3495–3507.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics 18*, 337–338.

Huelsenbeck, J. P. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics 17*, 754–755.

Iborra, F. J., H. Kimura, and P. R. Cook (2004). The functional organization of mitochondrial genomes in human cells. *BMC Biol. 2*, 9.

Innan, H. and M. Nordborg (2002). Recombination or Mutational Hot Spots in Human mtDNA? *Molecular Biology and Evolution 19*, 1122–1127.

Jakobsen, I. B. and S. Easteal (1996). A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *CABIOS 12*, 291–295.

Jakupciak, J. P., S. Maragh, M. E. Markowitz, A. K. Greenberg, M. O. Hoque, A. Maitra, P. E. Barker, P. D. Wagner, W. N. Rom, S. Srivastava, D. Sidransky, and C. D. O'Connell (2008). Performance of mitochondrial DNA mutations detecting early stage cancer. *BMC Cancer 8*, 285–295.

Jansen, R. P. (2000). Germline passage of mitochondria: quantitative considerations and possible embryological sequelae. *Human Reproduction 15*, 112–128.

Johnson, A. A. and K. A. Johnson (2001). Exonuclease proofreading by human mitochondrial DNA polymerase. *Journal of Biological Chemistry 276*, 38097–38107.

Jorde, L. B. and M. Bamshad (2000). Questioning evidence for recombination in human mitochondrial DNA. *Science 288*, 1931a.

Khrapko, K. (2008). Two ways to make an mtDNA bottleneck. *Nature Genetics 40*, 134–135.

Kolaczkowski, B. and J. W. Thornton (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature 431*, 980–984.

Kraytsberg, V., M. Schwartz, T. A. Brown, K. Ebralidse, W. S. Kunz, D. A. Clayton, J. Vissing, and K. K. (2004). Recombination of Human Mitochondrial DNA. *Science 304*, 981.

Kumar, S., P. Hedrick, T. Dowling, and M. Stoneking (2000). Questioning evidence for recombination in human mitochondrial DNA. *Science 288*, 1931a.

Kutschera, U. and K. J. Niklas (2005). Endosymbiosis, cell evolution, and speciation. *Theory in Biosciences 124*, 1–24.

Ladoukakis, E. D. and E. Zouros (2001). Direct evidence for homologous recombination in mussel *Mytilus galloprovincialis* mitochondrial DNA. *Molecular Biology and Evolution 18*, 1168–1175.

Lawson, M. J. and L. Zhang (2009). Sexy gene conversions: locating gene conversions on the X-chromosome. *Nucleic Acids Research*, 1–10. Advance Access published on May 31, 2009.

Lee, H. Y., J. Y. Chou, L. Cheong, N. H. Chang, S. Y. Yang, and J. Y. Leu (2008). Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. *Cell 135*, 1065–1073.

Lewontin, R. C. (1964). The interaction of selection and linkage. *Genetics 49*, 49–67.

Lopez, P., D. Casane, and H. Philippe (2002). Heterotachy, an Important Process of Protein Evolution. *Molecular Biology and Evolution 19*, 1–7.

Margulis, L. (1970). *Origin of Eukaryotic Cells*. New Haven: Yale University Press.

Mason, P. A. and R. N. Lightowlers (2003). Why do mammalian mitochondria possess a mismatch repair activity? *FEBS Letters 554*, 6–9.

May-Panloup, P., M. F. Chrétien, F. Savagner, C. Vasseur, M. Jean, Y. Malthièry, and P. Reynier (2003). Increased sperm mitochondrial DNA content in male infertility. *Human Reproduction 18*, 550–556.

Maynard Smith, J. (1992). Analyzing the Mosaic Structure of Genes. *Journal of Molecular Evolution 34*, 126–129.

Maynard Smith, J. and N. H. Smith (1998). Detecting recombination from gene trees. *Mol. Biol. Evol. 15*, 590–599.

Maynard Smith, J. and N. H. Smith (1999). Recombination in animal mitochondrial DNA. *Molecular Biology and Evolution 19*, 2330–2332.

McVean, G. A. T. (2001). What do patterns of genetic variability reveal about mitochondrial recombination? *Heredity 87*, 613–620.

Meade, A. and M. Pagel (2008). A phylogenetic mixture model for heterotachy. In *Evolutionary Biology from Concept to Application*. Springer Berlin Heidelberg.

Melnick, D. J. and G. A. Hoelzer (1992). Differences in male and female macaque dispersal lead to constrasting distributions of nuclear and mitochondrial DNA variation. *International Journal of Primatology 13*, 379–393.

Meunier, J. and A. Eyre-Walker (2001). The correlation between linkage disequilibrium and distance: Implications for recombination in hominid mitochondria. *Molecular Biology and Evolution 18*, 2132–2135.

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology 15*, 1044–1045.

Nass, M. M. K. (1969a). Mitochondrial DNA: Advances, Problems, and Goals. *Science 165*, 25–35.

Nass, M. M. K. (1969b). Mitochondrial DNA. I. Intramitochondrial distribution and structural relations of single- and double-length circular DNA. *Journal of Molecular Biology 42*, 521–528.

Neiman, M. and D. R. Taylor (2009). The causes of mutation accumulation in mitochondrial genomes. *Proc. R. Soc. B 276*, 1201–1209.

Niki, Y., S. I. Chigusa, and E. T. Matsurra (1989). Complete replacement of mitochondrial DNA in *Drosophila*. *Nature 341*, 551–552.

Pesole, G., C. Gissi, A. DeChirico, and C. Saccone (1999). Nucleotide substitution rate of mammalian mitochondrial genomes. *Journal of Molecular Evolution 48*, 427–434.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc (2008). Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology 5*, 50–58.

Piganeau, G., M. Gardner, and A. Eyre-Walker (2004). A broad survey of recombination in animal mitochondria. *Molecular Biology and Evolution 21*, 2319–2325.

Piko, L. and K. D. Taylor (1987). Amounts of mitochondrial DNA and abundance of some mitochondrial gene transcripts in early mouse embryos. *Dev. Biol. 123*, 364–374.

Posada, D. (2002). Evaluation of Methods for Detecting Recombination from DNA Sequences: Empirical Data. *Molecular Biology and Evolution 19*, 708–717.

Posada, D. and K. A. Crandall (2001). Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *PNAS 98*, 13757–13762.

Rambaut, A. and N. C. Grassly (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci. 13*, 235–238.

Rand, D. M. and R. G. Harrison (1986). Mitochondrial DNA transmission genetics in crickets. *Genetics 114*, 955–970.

Reynier, P., P. May-Panloup, M. F. Chrétien, C. J. Morgan, M. Jean, F. Savagner, P. Barrière, and Y. Malthièry (2001). Mitochondrial DNA content affects the fertilizability of human oocytes. *Molecular Human Reproduction 7*, 425–429.

Rokas, A., E. Ladoukakis, and E. Zouros (2003). Animal mitochondrial DNA recombination revisited. *TRENDS in Ecology and Evolution 18*, 411–417.

Ronquist, F. and J. P. Huelsenbeck (2003). MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics 19*, 1572–1574.

Roze, D., F. Rousset, and Y. Michalakis (2005). Germline bottlenecks, biparental inheritance and selection on mitochondrial variants: A two-level selection model. *Genetics 170*, 1385–1399.

Santamaria, M., S. Vicario, G. Pappadá, G. Scioscia, C. Scazzocchio, and C. Saccone (2009). Towards barcode markers in Fungi: an intron map of Ascomycota mitochondria. *BMC Bioinformatics 10*, S15–S27.

Sawyer, S. A. (1989). Statistical tests for detecting gene conversion. *Molecular Biology and Evolution 6*, 526–538.

Sawyer, S. A. (1999). GENECONV: A computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathematics, Washington University in St. Louis, available at `http://www.math.wustl.edu/~sawyer`.

Schierup, M. H. and J. Hein (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics 156*, 879–891.

Smeitink, J., L. van den Heuvel, and S. DiMauro (2001). The genetics and pathology of oxidative phosphorylation. *Nature Reviews Genetics 2*, 342–352.

Sneath, P. H. A. (1995). The distribution of the random division of a molecular sequence. *Binary 7*, 148–152.

Steinborn, R., P. Schinogl, V. Zakhartchenko, R. Achmann, W. Schernthaner, M. Stojkovic, E. Wolf, M. Muller, and G. Brem (2000). Mitochondrial DNA heteroplasmy in cloned cattle produced by fetal and adult cell cloning. *Nature Genetics 25*, 255–257.

Takahata, N. (1994). Comments on the detection of reciprocal recombination or gene conversion. *Immunogenetics 39*, 146–149.

Takahata, N. and M. Slatkin (1983). Evolutionary dynamics of extranuclear genes. *Genet. Res. 42*, 257–265.

Thyagarajan, B., R. A. Padua, and C. Campbell (1996). Mammalian mitochondria possess homologous DNA recombination activity. *Journal of Biological Chemistry 271*, 27536–27543.

Tsaousis, A. D., D. P. Martin, E. D. Ladoukakis, D. Posada, and E. Zouros (2005). Widespread recombination in published animal mtDNA sequences. *Molecular Biology and Evolution 22*, 925–933.

Ujvari, B., M. Dowton, and T. Madsen (2007). Mitochondrial DNA recombination in a free-franging Australian lizard. *Biol. Lett. 3*, 189–192.

van Oppen, M. J. H., J. Catmull, B. J. McDonald, N. R. Hislop, P. J. Hagerman, and D. J. Miller (2000). The mitochondrial genome of *Acropora tenuis* (Cnidaria; Scleractinia) contains a large Group I intron and a candidate control region. *Journal of Molecular Evolution 55*, 1–13.

Wahl, L. M. and P. J. Gerrish (2001). The probability that beneficial mutations are lost in populations with periodic bottlenecks. *Evolution 55*, 2606–2610.

Wallace, D. C. (1999). Mitochondrial diseases in man and mouse. *Science 283*, 1482–1488.

Wallace, D. C., M. D. Brown, and M. T. Lott (1999). Mitochondrial DNA variation in human evolution and disease. *Genetics 238*, 211–230.

Walsh, J. B. (1992). Intracellular selection, conversion bias, and the expected substitution rate of organelle genes. *Genetics 130*, 939–946.

Watanabe, K. and S. Matsumura (1991). The borderlands and possible hybrids between three species of macaques, *M. nigra, M. nigrescens*, and *M. hecki*, in the northern peninsula of Sulawesi. *Primates 32*, 365–370.

Watanabe, K., S. Matsumura, T. Watanabe, and Y. Hamada (1991). Distribution and possible intergradation between *Macaca tonkeana* and *M. ochreata* at the borderland of the species in Sulawesi. *Primates 32*, 385–389.

White, D. J. and N. J. Gemmell (2009). Can Indirect Tests Detect a Known Recombination Event in Human mtDNA? *Molecular Biology and Evolution 26*, 1435–1439.

Wilson, E. B. (1916). The distribution of the chondriosomes to the spermatozoa in scorpions. *Proc. Natl. Acad. Sci. USA 2*, 321–324.

Wiuf, C. (2001). Recombination in Human Mitochondrial DNA? *Genetics 159*, 749–756.

Wiuf, C., T. Christensen, and J. Hein (2001). A Simulation Study of the Reliability of Recombination Detection Methods. *Molecular Biology and Evolution 18*, 1929–1939.

Wong, L. J. C. and R. G. Boles (2005). Mitochondrial DNA analysis in clinical laboratory diagnostics. *Clinica Chimica Acta 354*, 1–20.

Worobey, M. (2001). A novel approach to detecting and measuring recombination: New insights into evolution in viruses, bacteria, and mitochondria. *Mol. Biol. Evol. 18*, 1425–1434.

Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics 139*, 993–1005.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in BioSciences 13*, 555–556.

Yang, Z. (2007). PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution 24*, 1586–1591.

Part VI

APPENDICES

Appendix A

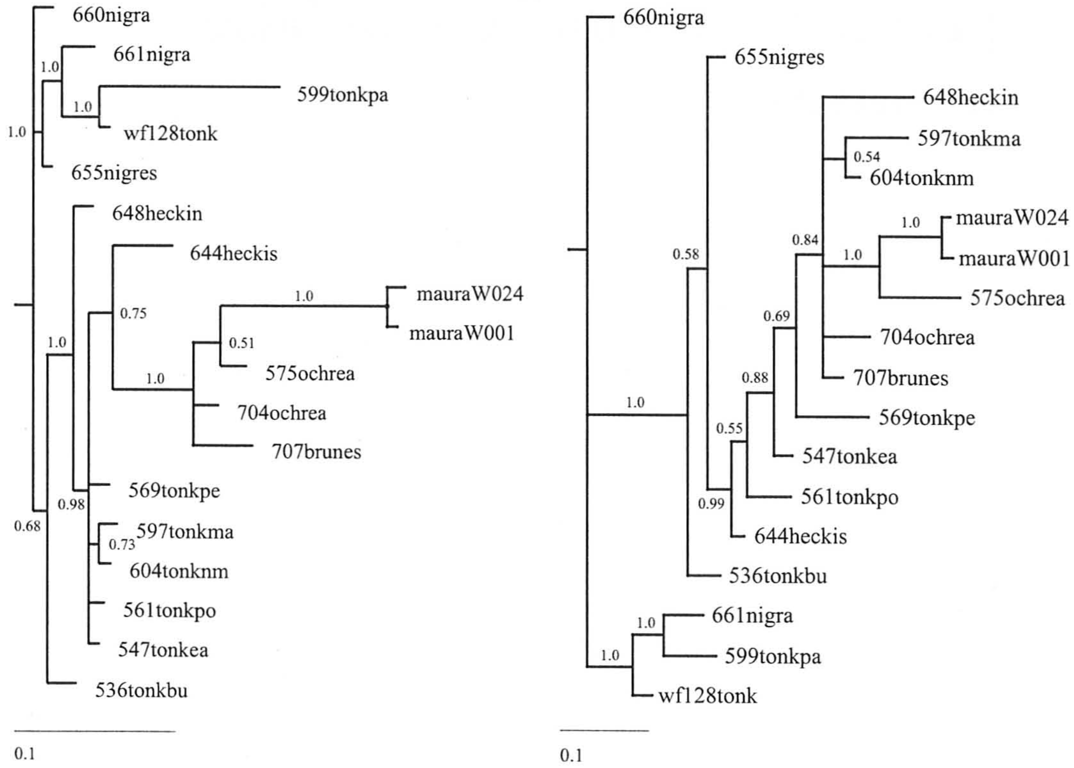
Mutation cold spots produce recombination false positives

Table A.1: Recombination detected by GENECONV. North/Central sequences are abbreviated to *N/C* and South sequences are abbreviated to So.

Sequence Pair	P-value	Pairing
661nigra; 597tonkma	< 0.0000	*N/C - N/C*
661nigra; 547tonkea	< 0.0000	*N/C - N/C*
599tonkpa; 597tonkma	0.0002	*N/C - N/C*
wf128tonk; 597tonkma	0.0002	*N/C - N/C*
599tonkpa; 604tonknm	0.0006	*N/C - N/C*
661nigra; 561tonkpo	0.0010	*N/C - N/C*
wf128tonk; 604tonknm	0.0015	*N/C - N/C*
660nigra; 597tonkma	0.0027	*N/C - N/C*
644heckis; 599tonkpa	0.0027	*N/C - N/C*
661nigra; 569tonkpe	0.0029	*N/C - N/C*
660nigra; 575ochrea	0.0144	*N/C - So*
661nigra; 536tonkbu	0.0231	*N/C - N/C*
569tonkpe; 536tonkbu	0.0431	*N/C - N/C*

Table A.2: Recombination detected by Max χ^2. The first sequence named in the sequence pair was removed from subsequent tests. The last column shows whether GENECONV inferred recombination in that pair. North/Central sequences are abbreviated to *N/C* and South sequences are abbreviated to So.

No. seqs.	Seq. Pair	P-value	Pairing	Breakpoint (nuc. pos.)	GENECONV
18	661nigra; 648heckin	0.000	*N/C - N/C*	909	N
17	660nigra; 648heckin	0.000	*N/C - N/C*	892	N
16	644heckis; 575ochrea	0.000	*N/C -So*	454	N
15	wf128tonk; 597tonkma	0.000	*N/C - N/C*	906	Y
14	604tonknm; 704ochrea	0.000	*N/C -So*	355	N
13	561tonkpo; 704ochrea	0.000	*N/C -So*	406	N
12	648heckin; 575ochrea	0.000	*N/C -So*	433	N
11	599tonkpa; 597tonkma	0.000	*N/C - N/C*	882	Y
10	597tonkma; 704ochrea	0.002	*N/C -So*	382	N
9	569tonkpe; 704ochrea	0.002	*N/C -So*	397	N
8	547tonkea; 704ochrea	0.006	*N/C -So*	404	N
7	655nigres; 575ochrea	0.016	*N/C -So*	421	N



(a) Nucleotides 1-909

(b) Nucleotides 910 - 1203

Figure A.1: Sulawesi macaque mtDNA `MrBayes` phylogenies from each side of the `Max` χ^2 predicted breakpoint. Posterior probabilities are listed on the branches.

Appendix B

Heterotachy in animal mtDNA produces recombination false positives

Table B.1: Description of data sets. Poly = polymorphic sites out of (# sites without gaps or ambiguous nuc). Following the Sulawesi macaques, datasets are presented in order of the strength of evidence for recombination (Piganeau, Gardner and Eyre-Walker, 2004)

Dataset	# sites	# samples	% poly (# sites)	Accession numbers
Sulawesi macaques ND3-ND4-ND4L	1203	18	24.63 (1202)	AF091400-2013AF091429
Bursaphelenchus conicaudatus COI	960	30	20.21 (960)	AB083711-39
Micropterus salmoides CYTB	1140	14	9.39 (1140)	AY115999; AY116000; AY225669-84
Macaca nemestrina CYTB	859	11	17.11 (859)	AF350388-91; AF350394-99
Microtus longicaudus CYTB	1140	68	18.86 (1140)	AF187160-230
Vesicomya pacifica COI	516	22	14.76 (515)	AF008287; AF008293-5;AF143290-304
Mandrillus sphinx CYTB	267	71	5.93 (253)	AF020423; AF301612-16; AY204763-827
Libellula quadrimaculata COI	416	15	8.79 (387)	AF228584-98
Dendroica petechia ATP8-ATP6	842	11	1.66 (841)	AF382957; AY115297-306Y
Apodemus sylvaticus CYTB	974	75	19.31 (875)	AF159395; AF60603; ASY98598; ASY98600; ASY98605; ASY311148; ASY511877-9; ASY511883-5; ASY511887; ASY511889-91; ASY511896-7; ASY511899; ASY511901; ASY511903-4; ASY511906-8; ASY511910-2; ASY511914-24; ASY511928-32; ASY511935-41; ASY511943-4; ASY511946-69; ASY511971-2
Gomphiocephalus hodgsoni COI	599	45	3.51 (598)	AY294562-606
Alpheus lottini CO1	564	42	17.05 (563)	AF107049-68; AF309910; ALU76428-489
Macrodon ancylodon CYTB	810	46	5.24 (801)	AY253604-9; AY253611-23; AY253625-32; AY253634-7; AY253639-50; AY253652; AY253655-6D
Papio papio ND4-ND5	696	8	2.16 (695)	AY212049-56
Campylorhynchus brunneicap ND2	298	60	8.90 (236)	AF291512-71
Bradypodion occidentale ND2	987	8	2.73 (951)	AF448728; AY289868; AY289888; AY289907-11
Passerella iliaca CYTB	432	9	3.57 (392)	U40162-70
Merlangius merlangus ATP6	878	13	1.59 (756)	AF526616-28
Gonatus onyx COI	657	7	1.98 (657)	AF000041; AF144718-23
Grus antigone CYTB	1143	9	1.49 (657)	U43618-25; U11060-1; U11064

Table B.2: AIC values for no partition model (AAA) and the highest likelihood partitioned model. $AIC = 2(d.f.) - 2ln(Likelihood)$. A difference of 1-2 is considered significant.

Dataset	Partition	AIC	Dataset	Partition	AIC
Sulawesi	AAA	6708.756682	Gomphiocephalus hodgsoni	AAA	1861.789882
	AAB	6412.71886		AAB	1850.25719
Bursa.	AAA	4879.692908	Alpheus lottini	AAA	2438.011198
	AAB	4668.515686		B-A+B-	2384.504476
Salmoides	AAA	4213.79893	Macrodon ancylodon	AAA	2645.606632
	AAB	4143.776244		AAB	2615.143714
Macaca nemestrina	AAA	3936.586504	Papio papio	AAA	2009.385272
	AAB	3814.159346		BAA	1997.05104
Microtus	AAA	5835.251176	Campylorhynchus brunneicap	AAA	843.035518
	AAB	5580.68436		ABA	835.074666
Vesico	AAA	2046.141962	Bradypodion occidentale	AAA	2801.848598
	B-A+B-	2004.126512		AAB	2782.667766
Mandrillus sphinx	AAA	842.102264	Passerella iliaca	AAA	1244.163596
	BAA	842.816848		AAB	1239.941976
Libellula quadrimaculata	AAA	1381.752308	Merlangius merlangus	AAA	2158.142204
	ABA	1360.668346		ABA	2147.732614
Dendroica petechia	AAA	2353.086366	Gonatus onyx	AAA	1932.35556
	ABA	2349.376928		ABA	1924.336196
Apodemus sylvaticus	AAA	4375.323012	Grus antigone	AAA	3261.565488
	B-A+B-	4202.847624		B-A+B-	3250.446496

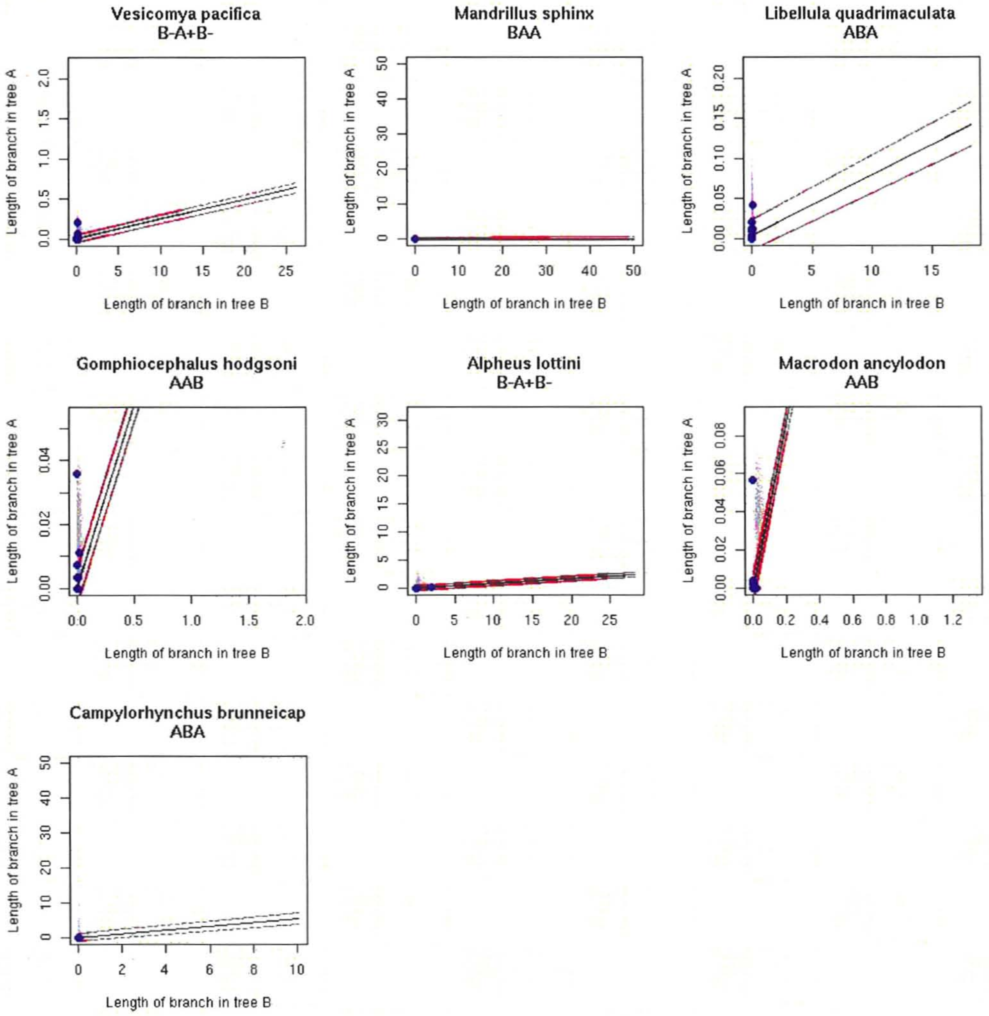


Figure B.1: Branch length graphs of *Vesicomya pacifica, Mandrillus sphinx, Libellula quadrimaculata, Gomphiocephalus hodgsoni, Alpheus lottini, Macrodon ancylodon, and Campylorhynchus brunneicap.* Grey points: branches from AAA simulations; Blue points: branches from the animal mtDNA; Black line: linear regression of branch A on branch B; Red line: 95% prediction bounds.

Appendix C

Fixation probabilities of advantageous mtDNA mutations

Table C.1: Probabilities of fixation for a new neutral mutant. Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. *c* is the number of cell divisions in one generation.

		Number of mitochondria/cell			
	c	1	10	50	100
Expected	1	2.00×10^{-4}	2.00×10^{-5}	4.00×10^{-6}	2.00×10^{-6}
Observed	1	1.95×10^{-4}	1.93×10^{-5}	3.87×10^{-6}	2.02×10^{-6}
	10	1.95×10^{-4}	1.99×10^{-5}	4.17×10^{-6}	1.98×10^{-6}
	20	2.11×10^{-4}	1.97×10^{-5}	3.94×10^{-6}	1.99×10^{-6}
	50	1.85×10^{-4}	2.08×10^{-5}	3.92×10^{-6}	1.96×10^{-6}

Table C.2: Probability of fixation for a new recessive mutant (simple selection based on an advantage of 0.01). Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. *c* is the number of cell divisions in one generation.

	Number of mitochondria/cell			
c	1	10	50	100
1	1.92×10^{-2}	1.75×10^{-3}	2.75×10^{-4}	1.10×10^{-4}
10	2.01×10^{-2}	1.90×10^{-3}	3.72×10^{-4}	1.69×10^{-4}
20	2.01×10^{-2}	1.90×10^{-3}	3.79×10^{-4}	1.77×10^{-4}
50	2.01×10^{-2}	2.00×10^{-3}	3.71×10^{-4}	1.84×10^{-4}

Table C.3: Probability of fixation for a new incomplete dominant mutant (simple selection based on an advantage of 1.0 + (x/total) × 0.01 where x is number of mutant mitochondria and total is total per cell). Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. c is the number of cell divisions in one generation.

	Number of mitochondria/cell			
c	1	10	50	100
1	1.98×10^{-2}	1.86×10^{-3}	4.16×10^{-4}	1.97×10^{-4}
10	1.95×10^{-2}	1.92×10^{-3}	3.70×10^{-4}	1.80×10^{-4}
20	1.95×10^{-2}	1.83×10^{-3}	3.79×10^{-4}	1.87×10^{-4}
50	1.90×10^{-2}	1.98×10^{-3}	3.86×10^{-4}	1.84×10^{-4}

Table C.4: Probability of fixation for a new dominant mutant (simple selection based on an advantage of 0.01). Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. c is the number of cell divisions in one generation.

	Number of mitochondria/cell			
c	1	10	50	100
1	2.02×10^{-2}	2.19×10^{-3}	6.16×10^{-4}	4.36×10^{-4}
10	1.95×10^{-2}	1.87×10^{-3}	3.79×10^{-4}	2.03×10^{-4}
20	2.00×10^{-2}	1.95×10^{-3}	3.64×10^{-4}	2.02×10^{-4}
50	1.95×10^{-2}	1.92×10^{-3}	3.76×10^{-4}	1.81×10^{-4}

Table C.5: Probability of fixation for a new recessive mutant (simple selection based on an advantage of 1.0). Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. c is the number of cell divisions in one generation.

	Number of mitochondria/cell		
c	1	10	100
1	7.79×10^{-1}	2.42×10^{-2}	3.35×10^{-4}
10	7.96×10^{-1}	6.81×10^{-2}	2.10×10^{-3}
50	7.88×10^{-1}	7.68×10^{-2}	4.92×10^{-3}

Table C.6: Probability of fixation for a new incomplete dominant mutant (simple selection based on an advantage of 1.0 + (x/total) × 1.0 where x is number of mutant mitochondria and total is total per cell). Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. *c* is the number of cell divisions in one generation.

	Number of mitochondria/cell		
c	1	10	100
1	8.02×10^{-01}	1.27×10^{-01}	1.72×10^{-02}
10	7.90×10^{-01}	8.78×10^{-02}	1.31×10^{-02}
20	7.92×10^{-01}	8.23×10^{-02}	1.16×10^{-02}
50	8.04×10^{-01}	8.13×10^{-02}	9.99×10^{-03}

Table C.7: Probability of fixation for a new dominant mutant (simple selection based on an advantage of 1.0). Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. *c* is the number of cell divisions in one generation.

	Number of mitochondria/cell		
c	1	10	100
1	8.06×10^{-1}	3.86×10^{-1}	3.91×10^{-1}
10	7.84×10^{-1}	1.08×10^{-1}	6.67×10^{-2}
50	8.16×10^{-1}	8.10×10^{-2}	1.82×10^{-2}

Table C.8: Probabilities of fixation for a new neutral mutant with bottleneck. Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. *c* is the number of cell divisions in one generation.

		Number of mitochondria/cell		
	c	1	10	100
Expected	1	2.00×10^{-4}	2.00×10^{-5}	4.00×10^{-6}
Observed	1	1.96×10^{-4}	1.85×10^{-5}	1.79×10^{-6}
	10	1.96×10^{-4}	1.89×10^{-5}	1.87×10^{-6}
	20	1.95×10^{-4}	1.90×10^{-5}	1.85×10^{-6}
	50	1.96×10^{-4}	1.89×10^{-5}	1.90×10^{-6}

Table C.9: Probability of fixation for a new recessive mutant (simple selection based on an advantage of 0.01) with bottleneck. Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. *c* is the number of cell divisions in one generation.

	Number of mitochondria/cell		
c	1	10	100
1	1.96×10^{-2}	1.94×10^{-3}	1.79×10^{-4}
10	1.96×10^{-2}	1.93×10^{-3}	1.84×10^{-4}
20	1.95×10^{-2}	1.92×10^{-3}	1.85×10^{-4}
50	1.98×10^{-2}	1.95×10^{-3}	1.94×10^{-4}

Table C.10: Probability of fixation for a new incomplete dominant mutant (simple selection based on an advantage of 1.0 + (x/total) × 0.01 where x is number of mutant mitochondria and total is total per cell) with bottleneck. Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. *c* is the number of cell divisions in one generation.

	Number of mitochondria/cell		
c	1	10	100
1	1.95×10^{-2}	1.93×10^{-3}	1.90×10^{-4}
10	1.97×10^{-2}	1.94×10^{-3}	1.95×10^{-4}
20	1.97×10^{-2}	1.92×10^{-3}	1.96×10^{-4}
50	1.96×10^{-2}	1.92×10^{-3}	1.96×10^{-4}

Table C.11: Probability of fixation for a new dominant mutant (simple selection based on an advantage of 0.01) with bottleneck. Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. *c* is the number of cell divisions in one generation.

	Number of mitochondria/cell		
c	1	10	100
1	1.98×10^{-2}	1.93×10^{-3}	2.09×10^{-4}
10	1.99×10^{-2}	1.95×10^{-3}	2.04×10^{-4}
20	1.99×10^{-2}	1.93×10^{-3}	2.03×10^{-4}
50	1.97×10^{-2}	1.95×10^{-3}	1.96×10^{-4}

Table C.12: Probability of fixation for a new recessive mutant (simple selection based on an advantage of 1.0) with bottleneck. Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. *c* is the number of cell divisions in one generation.

	Number of mitochondria/cell		
c	1	10	100
1	7.99×10^{-1}	7.36×10^{-2}	2.46×10^{-3}
10	8.01×10^{-1}	6.81×10^{-2}	3.27×10^{-3}
20	7.97×10^{-1}	7.34×10^{-2}	3.84×10^{-3}
50	7.97×10^{-1}	7.97×10^{-2}	5.03×10^{-3}

Table C.13: Probability of fixation for a new incomplete dominant mutant (simple selection based on an advantage of 1.0 + (x/total) × 1.0 where x is number of mutant mitochondria and total is total per cell) with bottleneck. Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. *c* is the number of cell divisions in one generation.

	Number of mitochondria/cell		
c	1	10	100
1	7.91×10^{-1}	8.58×10^{-2}	1.30×10^{-2}
10	7.91×10^{-1}	8.56×10^{-2}	1.17×10^{-2}
20	7.95×10^{-1}	8.15×10^{-2}	1.08×10^{-2}
50	7.95×10^{-1}	8.04×10^{-2}	9.67×10^{-3}

Table C.14: Probability of fixation for a new dominant mutant (simple selection based on an advantage of 1.0) with bottleneck. Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. *c* is the number of cell divisions in one generation.

	Number of mitochondria/cell		
c	1	10	100
1	7.99×10^{-1}	1.39×10^{-1}	9.00×10^{-2}
10	7.95×10^{-1}	1.01×10^{-1}	4.55×10^{-2}
20	8.00×10^{-1}	8.57×10^{-2}	2.95×10^{-2}
50	7.94×10^{-1}	7.87×10^{-2}	1.67×10^{-2}

Table C.15: Probabilities of fixation for a new neutral mutant in germ+soma model. Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. c is cell divisions/generation in the somatic and germ lines.

	Number of mitochondria/cell					
c	1	10	30	50	75	100
1	1.94×10^{-4}	9.47×10^{-5}	9.48×10^{-5}	9.57×10^{-5}	9.34×10^{-5}	9.66×10^{-5}
10	1.94×10^{-4}	3.55×10^{-5}	4.01×10^{-5}	3.90×10^{-5}	3.80×10^{-5}	3.77×10^{-5}
15	1.94×10^{-4}	2.68×10^{-5}	3.06×10^{-5}	2.95×10^{-5}	2.99×10^{-5}	3.03×10^{-5}
20	1.99×10^{-4}	2.16×10^{-5}	2.30×10^{-5}	2.45×10^{-5}	2.38×10^{-5}	2.27×10^{-5}
35	1.94×10^{-4}	1.99×10^{-5}	1.41×10^{-5}	1.47×10^{-5}	1.55×10^{-5}	1.57×10^{-5}
50	2.00×10^{-4}	1.91×10^{-5}	1.04×10^{-5}	1.03×10^{-5}	1.10×10^{-5}	1.16×10^{-5}

Table C.16: Probability of fixation for a new recessive mutant (simple selection based on an advantage of 0.01) in germ+soma model. Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. c is cell divisions/generation in the somatic and germ lines.

	Number of mitochondria/cell		
c	1	10	100
1	1.98×10^{-2}	1.75×10^{-3}	1.07×10^{-4}
10	1.97×10^{-2}	1.87×10^{-3}	1.71×10^{-4}
20	1.96×10^{-2}	1.89×10^{-3}	1.81×10^{-4}
50	1.98×10^{-2}	1.90×10^{-3}	1.83×10^{-4}

Table C.17: Probability of fixation for a new dominant mutant (simple selection based on an advantage of 0.01) in germ+soma model. Population size is 5000 females (10^4 individuals total) and is based on an observation of 1000 fixations. c is cell divisions/generation in the somatic and germ lines.

	Number of mitochondria/cell		
c	1	10	100
1	1.98×10^{-2}	2.01×10^{-3}	4.19×10^{-4}
10	1.98×10^{-2}	1.86×10^{-3}	2.06×10^{-4}
20	1.95×10^{-2}	1.90×10^{-3}	1.93×10^{-4}
50	1.97×10^{-2}	1.90×10^{-3}	1.88×10^{-4}