MODELING CORRELATION IN BINARY COUNT DATA

WITH APPLICATION TO FRAGILE SITE IDENTIFICATION

A Dissertation

by

CHRISTOPHER JERRY HINTZE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2005

Major Subject: Statistics

MODELING CORRELATION IN BINARY COUNT DATA

WITH APPLICATION TO FRAGILE SITE IDENTIFICATION

A Dissertation

by

CHRISTOPHER JERRY HINTZE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	P. Fred Dahm
Committee Members,	F. Michael Speed
	Henrik Schmiediche
	Ira F. Greenbaum
Head of Department,	Simon Sheather

August 2005

Major Subject: Statistics

ABSTRACT

Modeling Correlation in Binary Count Data With Application
to Fragile Site Identification. (August 2005)
Christopher Jerry Hintze, B.S., Brigham Young University;
M.S., Brigham Young University
Chair of Advisory Committee: Dr. P. Fred Dahm

Available fragile site identification software packages (FSM and FSM3) assume that all chromosomal breaks occur independently. However, under a Mendelian model of inheritance, homozygosity at fragile loci implies pairwise correlation between homologous sites. We construct correlation models for chromosomal breakage data in situations where either partitioned break count totals (per-site single-break and double-break totals) are known or only overall break count totals are known. We derive a likelihood ratio test and Neyman's C(α) test for correlation between homologs when partitioned break count totals are known and outline a likelihood ratio test for correlation using only break count totals. Our simulation studies indicate that the C(α) test using partitioned break count totals outperforms the other two tests for correlation in terms of both power and level. These studies further suggest that the power for detecting correlation is low when only break count totals are reported. Results of the C(α) test for correlation applied to chromosomal breakage data from 14 human subjects indicate that detection of correlation between homologous fragile sites is problematic due to sparseness of breakage data. Simulation studies of the FSM and FSM3 algorithms using parameter values typical for fragile site data demonstrate that neither algorithm is significantly affected by fragile site correlation. Comparison of simulated fragile site misclassification rates in the presence of zero-breakage data supports previous studies (Olmsted 1999) that suggested FSM has lower false-negative rates and FSM3 has lower false-positive rates.

ACKNOWLEDGMENTS

First, I would like to thank Amber, my wife, for her endless love and support of my graduate studies. I could not have done this without her. I am grateful to my children (Braden, Aubrielle, and McKenzie) for their happy smiles greeting me each day. I am also grateful to the Hintze and Orme families for all that they have done for me and my family. I would like to express appreciation to Dr. Fred Dahm, Dr. Ira Greenbaum, Dr. Michael Speed, and Dr. Henrik Schmiediche for their input and support in creating this manuscript. I would also like to thank Stacy Denison for providing chromosomal breakage data. I am also thankful to all the faculty and staff of the Statistics Department at Texas A&M University for making my work here enjoyable. Finally, I acknowledge the Lord's help in completing this work.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 DNA and Fragile Sites

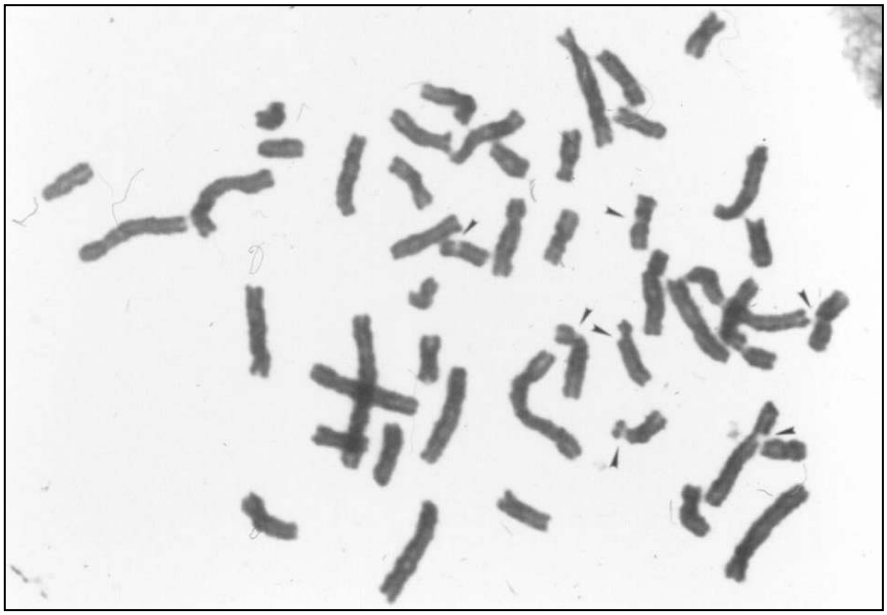
Deoxyribonucleic Acid (DNA) provides the genetic blueprint by which multi-cellular organisms are formed and maintained. The DNA contained within each cell is continuously accessed and interpreted by intricate intracellular networks that recognize, and respond to, environmental stimuli. Survival of an organism depends on correctly-specified, viable genes along the length of its DNA blueprint. Aberrations in certain key genes are associated with developmental abnormalities and cancer.

DNA strands replicate and segregate prior to cell division. This process is not immune to error, including the occurrence of breaks in the replicated DNA molecules. We define a break in the same way as Olmsted (1999); by chromosomal break, we mean a break or a gap in one or both sister chromatids (if in both, at the same place). Such breaks are microscopically visible at metaphase of mitosis, when the DNA is tightly coiled and condensed into familiar chromosome structures. The location of these breaks can be determined using a differential staining technique known as G-banding. Figure 1 displays stained chromosomes, some of which exhibit chromosomal breaks. The presence of a break (McAllister and Greenbaum 1997) is first determined by looking at uniformly stained chromosomes (Figure 1(a)). When a break is found, the uniform stain is removed, the chromosomes are G-banded (Figure 1(b)), and the break is mapped to a particular site. All chromosomal sites can be categorized into one of three groups:

1. Zero-Breakage Sites – Sites whose probability of breakage under specific conditions is equal to zero (i.e., sites at which breaks are never observed).
2. Non-Fragile Sites – Sites that experience rare, random breakage under specific conditions.
3. Fragile Sites – Sites that experience frequent, non-random breakage under specific conditions.

This dissertation follows the style of the *Journal of the American Statistical Association*.

(a)



(b)

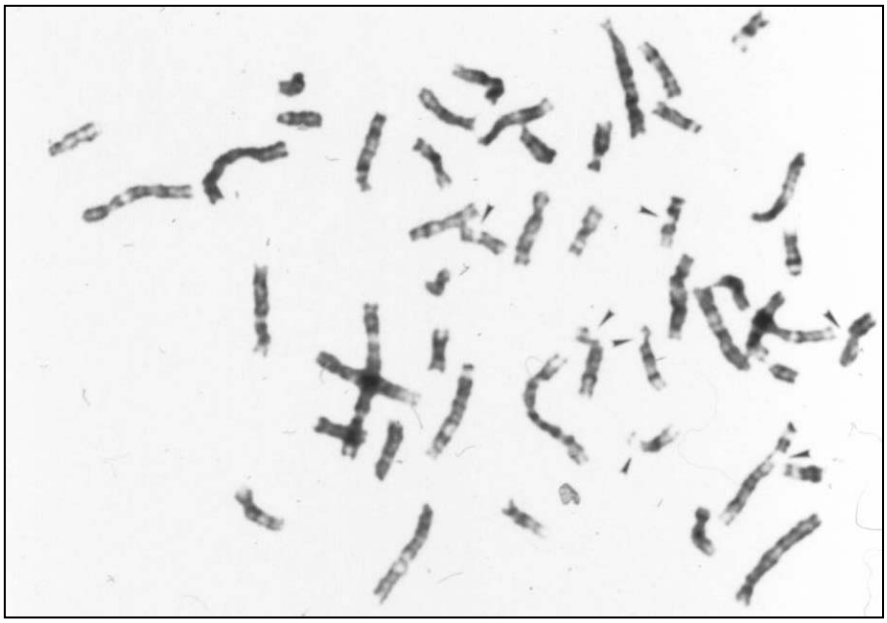


Figure 1. (a) Uniformly Stained and (b) Differentially Stained (G-banded) Metaphase Chromosomes From a Deer Mouse. Breaks in the chromosomes are indicated by arrows. The presence of a break is first determined by looking at the uniformly stained chromosomes in (a). When a break is found, the uniform stain is removed, and the chromosomes are G-banded as in (b). G-banding allows the location of each break to be mapped to a particular site. (Photos provided by Dr. Ira F. Greenbaum of Texas A&M University.)

Denison et al. (2003) defined fragile sites generally as "chromosomal loci that experience non-random breakage when challenged under appropriate tissue culture conditions." Jordan et al. (1990) defined fragile sites as "nonrandom, heritable sites on chromosomes that can be induced to form gaps, breaks and rearrangements under specific conditions." While it is possible to observe rare, random breaks in non-fragile regions, fragile sites are characterized by high relative break frequencies.

One well-characterized fragile site is known as FRAXA and has been linked to fragile X syndrome, the most common cause of inherited mental retardation. FRAXA is located near the end of the long arm of the human X chromosome (Xq27.3) and encompasses a CCG-trinucleotide repeat sequence immediately adjacent to a gene called FMR1 (for fragile X mental retardation 1, Verkerk et al. 1991). Many other less-studied fragile sites are believed to exist and are the subject of intense research in cytogenetics. Fragile sites also have been hypothesized to be associated with cancer (Yunis 1984; Yunis and Soreng 1984; Hecht and Glover 1984; Hecht and Sutherland 1984; Popescu 2003).

1.2 Fragile Site Identification Methods

Many methods, both statistical and non-statistical, have been developed in an attempt to identify fragile sites. We summarize these methods.

1.2.1 Ad Hoc Methods

Fragile site identifications began with ad-hoc, non-statistical methods. Initially, researchers used an arbitrarily-set threshold frequency of breaks (4%) to distinguish fragile sites from non-fragile sites (Rao et al. 1988). Olmsted (1999) noted, "The performance of this 4% rule depends very much on the expected break count for the non-fragile bands, which depends on the number of cells observed." Olmsted (1999) used a simulation study of Poisson counts to demonstrate that as the expected break count among non-fragile sites increases, the Type I error rate (associated with the 4% rule) decreases. When the expected number of breaks at non-fragile sites was 0.25, 23% of the

simulated non-fragile sites were declared fragile. When 0.5 breaks were expected, 0.2% of non-fragile sites were misclassified as fragile. For a constant non-fragile breakage probability, the expected number of breaks can only be increased by raising the number of metaphases observed. Thus, in terms of controlling Type I error, this simulation study suggests that the 4% rule is not reliable as a method for fragile site identification.

1.2.2 Probability-Based Methods

Several researchers have recognized the need for probability-based models in the identification of fragile sites. De Braekeleer and Smith (1988) and Vasarhelyi and Friedman (1989) proposed binomial models for fragile site identification. Dahm and Greenbaum (1994), however, showed that the binomial distribution does not provide an appropriate fit to breakage data. Mariani (1989) suggested that for breakage data the binomial distribution of break counts can be adequately approximated by the Poisson distribution. Mariani (1989) modeled the total number of random breaks observed at band i, N_i, as following a $Poisson(2c\pi_i)$ distribution, where c is the number of metaphases analyzed, and π_i is the probability of a break occurring at band i, $i = 1, \ldots, k$. The method of Mariani (1989) would flag a band as fragile if the total number of breaks for that band were greater than $h_{0.05}$, the least integer h such that $kP(Y \geq h) < 0.05$, where $Y \sim Poisson(2c\hat{\pi})$, and $\hat{\pi} = \overline{n}/2c$ is the MLE of π. Jordan et al. (1990) used a negative-binomial model to fit breakage data and contended that the negative-binomial distribution provides a better fit than the Poisson to each of three sets of data. Böhm et al. (1995) pointed out that "fitting chromosomal breakage data to these particular models does not address the question of whether the data provide evidence for fragile sites." The methods of Mariani (1989) and Jordan et al. (1990) declare sites as fragile if they fall within a prescribed proportion (e.g. the upper 5% tail) of the assumed distribution. Thus, even in the cases where no fragile sites are present, the methods of Mariani (1989) and Jordan et al. (1990) would declare some sites to be fragile if data points exist within the prescribed proportion. Tai et al. (1993) assumed that, given the sum of k break counts equals n, random (non-fragile) breaks counts follow a multinomial

distribution with all cell probabilities equal to $1/k$. Thus, under the null hypothesis of no fragile sites, the break count for each band follows a *Binomial*(n, $1/k$) distribution. Tai et al. (1993) suggested testing the hypotheses H_0: $\pi_i \leq 1/k$ vs. H_1: $\pi_i > 1/k$ to determine whether or not band i should be deemed fragile. The method of Tai et al. (1993), however, also suffers from the use of the binomial distribution as shown in Dahm and Greenbaum (1994).

Böhm et al. (1995) developed the FSM multinomial statistical model which addresses many of the shortcomings of preceding methods. For instance, FSM utilizes the Poisson distribution. Böhm et al. (1995) showed that when no fragile sites are present, the Poisson distribution is the appropriate model for induced-breakage frequencies. FSM employs an iterative stepwise procedure of fitting a multinomial distribution to arrive at a maximal subset of non-fragile sites. All other sites are deemed fragile. Therefore, in the case where no fragile sites are present, FSM is more likely than previous methods (e.g., Mariani (1989) and Jordan et al. (1990)) to correctly identify sites as non-fragile. (Section 1.5 contains details of the FSM algorithm.) Olmsted (1999) adapted the FSM algorithm to account for the presence of zero-breakage sites and named this algorithm the FSM3 algorithm. Zero-breakage sites are chromosomal regions for which random breaks are never observed (i.e., have probability of breakage equal to zero). Zero-breakage sites may occur either because the sites are actually resistant to breakage or because a break results in cell death before metaphase of mitosis, making it impossible to observe that break. The FSM3 algorithm makes use of the positive-Poisson distribution in determining site fragility in order to eliminate the adverse effects of zero-breakage site contamination in the data. (Section 1.6 contains details of the FSM3 algorithm.)

The common assumption of all of the methods mentioned in this section is that observed breaks are independent of one another. We consider a case where this assumption is violated; more specifically, we investigate the impact of correlation between fragile sites on homologous chromosomes on the FSM and FSM3 algorithms.

To facilitate the discussions that will follow, we first present the notation used in recording and summarizing breakage data.

1.3 Data Structure and Summaries

For a single individual, suppose that c somatic cells in metaphase are examined to determine whether there is a break in any of the k bands on either (or both) of the two homologous chromosomes. Use the following notation to record the presence (or absence) of a break. Define

$$X_{ijh} = \begin{cases} 1 & \text{if a break is observed at the } i^{\text{th}} \text{ band of the } j^{\text{th}} \text{ cell on the } h^{\text{th}} \text{ homolog,} \\ 0 & \text{otherwise,} \end{cases} \tag{1.1}$$

where $i = 1, 2,\ldots, k$, $j = 1, 2,\ldots, c$, and $h = 1, 2$. We do not distinguish between the maternal and paternal homologs because it is typically not possible to distinguish parentage based solely on karyotipic information such as that obtained when recording chromosomal breaks. The only exceptions to this would be the sex chromosomes, X and Y.

1.3.1 Break Count Totals (BCT)

The data can be summarized by determining the total number of breaks for a given band. The total number of breaks observed at band i is

$$N_i \equiv \sum_{j=1}^{c} \sum_{h=1}^{2} X_{ijh} \ . \tag{1.2}$$

We will hereafter refer to $N_1, N_2, \ldots, N_k$ as the *Break Count Totals* (BCT).

1.3.2 Partitioned Break Count Totals (PBCT)

It is also possible to summarize the data in a manner which provides more information than do just the total break counts. Define

$$X_{ij\bullet} \equiv \sum_{h=1}^{2} X_{ijh}, \tag{1.3}$$

and define

$$M_{0_i} \equiv \sum_{j=1}^{c} I_0\left(X_{ij\bullet}\right), M_{1_i} \equiv \sum_{j=1}^{c} I_1\left(X_{ij\bullet}\right), \text{ and } M_{2_i} \equiv \sum_{j=1}^{c} I_2\left(X_{ij\bullet}\right), \tag{1.4}$$

where

$$I_a\left(X_{ij\bullet}\right) = \begin{cases} 1 & \text{if } X_{ij\bullet} = a, \\ 0 & \text{otherwise,} \end{cases} \tag{1.5}$$

satisfy $M_{0_i} + M_{1_i} + M_{2_i} = c$ .Thus, M_{0_i} represents the number of metaphases examined for which no breaks are observed at band *i*, M_{1_i} represents the number of metaphases for which only one break is observed at band *i*, and M_{2_i} represents the number of metaphases examined for which a break is observed at band *i* in both homologs. We will hereafter refer to M_{0_i}, M_{1_i}, and M_{2_i} as the *Partitioned Break Count Totals* (PBCT). Note that the BCT can be computed from the PBCT as

$$N_i = M_{1_i} + 2M_{2_i} . \tag{1.6}$$

1.4 Independence Assumptions

The probability-based fragile site identification models mentioned in Section 1.2, including the FSM and FSM3 algorithms, are based on the following three independence assumptions:

1. Breaks occurring at different bands are independent, whether on the same chromosome or on non-homologous chromosomes. Thus, X_{ijh} and $X_{i'jh}$ are independent, where *i* and *i'* correspond to different bands.
2. Breaks occurring in different cells (i.e., different metaphases) are independent. Thus, X_{ijh} and $X_{ij'h}$ are independent, where *j* and *j'* correspond to different cells.
3. Breaks occurring at the same bands on homologous chromosomes in the same cell are independent. Thus, X_{ij1} and X_{ij2} are independent.

Taken together, these assumptions imply that X_{ijh} is independent of any other $X_{i'j'h'}$ when $(i, j, h) \neq (i', j', h')$.

1.5 FSM Algorithm

The FSM algorithm described in Böhm et al. (1995) is an iterative procedure that determines the maximal subset of non-fragile sites and identifies all other sites as fragile. Suppose that c metaphases are examined from each of r individuals, each having the same set of k chromosomal sites per haploid chromosomal complement. (The sex chromosomes are ignored, and the distinction between maternal and paternal homologs is not incorporated in the formulation of the FSM algorithm.) If we consider a single individual and denote the probability of breakage at band i as π_i, then based on the three independence assumptions of Section 1.4, the total number of breaks observed at site i is binomially distributed as

$$N_i \sim Bin(2c, \pi_i) \text{ for all } i = 1,\ldots, k. \tag{1.7}$$

Since the expected number of breaks, $2c\pi_i$, is small, the distribution of the total number of breaks at band i can be approximated with the Poisson distribution as

$$N_i \dot{\sim} Poisson(2c, \pi_i) \text{ for all } i = 1,\ldots, k. \tag{1.8}$$

The independence of the N_i's, i = 1, 2, …, k, and (1.8) allowed Böhm et al. (1995) to write the distribution of the vector of breaks, $\boldsymbol{N} = (N_1,..., N_k)$, conditional on the total number of breaks, $N = \sum_i N_i = n$, as multinomial

$$\boldsymbol{N} \mid N = n \sim Mult(n, k, \boldsymbol{p}), \tag{1.9}$$

where $\boldsymbol{p} = (p_1,..., p_k)$ with $p_i = \pi_i / \sum_i \pi_i$ for $i = 1,\ldots, k$.

The FSM algorithm begins by ordering the sites based on the number of break counts observed for each site and then testing the null hypothesis that there are no fragile sites present, i.e., that all sites have the same breakage probability equal to $1/k$. If this hypothesis is rejected, then the site with the highest observed breakage is excluded and the remaining sites are tested for homogeneity. This process continues until the hypothesis of homogeneity is not rejected, creating the maximal set of non-fragile sites. The excluded sites are deemed fragile. The test statistic used at each iteration is

$$X_{k_l}^2 = \sum_{i=1}^{k_l} N_i \left(\frac{N_i}{n_l / k_l} - 1 \right), \tag{1.10}$$

where l designates the iteration at which the test is performed, $n_l = \sum_{i=1}^{k_l} N_i$, and k_l is the number of non-fragile bands remaining at the l^{th} iteration. Because of the sparse nature of the data, the statistic of (1.10) cannot be assumed to follow a χ^2 distribution. Böhm et al. (1995) used the standardized test statistic

$$X_s^2 = \frac{X_{k_l}^2 - (k_l - 1)}{\sqrt{2(k_l - 1)}}, \tag{1.11}$$

which is asymptotically distributed as *Normal*(0,1). A Bonferroni-type adjustment of the significance level is used at each iteration, i.e., the adjusted level is equal to $\alpha/(l + 1)$. For the initial test of homogeneity, l is equal to zero.

1.6 FSM3 Algorithm

The FSM3 algorithm developed in Olmsted (1999) is similar in nature to the FSM algorithm but accounts for zero-breakage site contamination of the data. Zero-breakage sites are bands for which breaks are never observed, the inclusion of which violates the underlying assumptions of the FSM algorithm. Data containing zero-breakage sites display a greater number of bands with zero breaks than would be expected if the data were a random sample from a Poisson distribution (I.F. Greenbaum, personal communication, May 6, 2005). The FSM3 algorithm removes zero-breakage contamination from the data by eliminating all sites with zero observed breaks from the analysis.

When sites with zero observed breaks are ignored, the k_1 remaining break totals follow a positive-Poisson distribution. That is

$$\begin{aligned} p(n_i) &= P(N_i = n_i \mid n_i > 0) = \frac{e^{-\lambda_{k_1}} \lambda_{k_1}^{\ n_i}}{(1 - e^{-\lambda_{k_1}}) n_i!} \\ &\equiv Poisson^{+}(\lambda_{k_1}), \end{aligned} \tag{1.12}$$

for $i = 1, 2, \ldots, k_1$, $n_i = 1, 2, 3, \ldots$, and $\lambda_{k_1} > 0$. Johnson, Kotz, and Kemp (1993) showed that the MLE of λ_{k_1} is the value $\hat{\lambda}_{k_1}$ that satisfies

$$\overline{n} = \frac{\hat{\lambda}_{k_1}}{1 - e^{\hat{\lambda}_{k_1}}}, \tag{1.13}$$

which can be solved numerically.

The FSM3 algorithm begins at iteration zero by testing the null hypothesis of no fragile bands using the Rao-Robson statistic (Rao and Robson 1974; Olmsted 1999)

$$X^2_{RR} = \boldsymbol{X}\left(\hat{\lambda}_{k_1}\right)^T \boldsymbol{V}\left(\hat{\lambda}_{k_1}\right)^- \boldsymbol{X}\left(\hat{\lambda}_{k_1}\right), \tag{1.14}$$

where $\boldsymbol{X}\left(\hat{\lambda}_{k_1}\right)$ is the vector of Pearson residuals whose i^{th} element is equal to

$$\frac{n_i - k_1 p_i\left(\hat{\lambda}_{k_1}\right)}{\sqrt{k_1 p_i\left(\hat{\lambda}_{k_1}\right)}} \tag{1.15}$$

and $\boldsymbol{V}\left(\hat{\lambda}_{k_1}\right)^-$ is the generalized inverse of the estimated null asymptotic covariance matrix of $\boldsymbol{X}\left(\hat{\lambda}_{k_1}\right)$. The asymptotic distribution of X^2_{RR} is χ^2_{df}, where *df* equals the number of cells minus one. The number of cells is determined by the fact that each cell must have an expected count greater than or equal to *e*, where *e* is some specified minimum value. The default value of *e* for FSM3 is one.

If the null hypothesis of no fragile sites is rejected, the algorithm continues iteratively, setting the value of *m* to two (or incrementing *m* by one on subsequent iterations), fitting a one- and *m*-truncated Poisson distribution, and testing for goodness of fit using a Pearson-type statistic which follows a chi-square distribution. The algorithm terminates upon rejection the null hypothesis that the data follow a truncated Poisson distribution or if $m = f$, where *f* is the highest break count such that the expected number of break counts greater than or equal to *f* is greater than or equal to one. FSM3 then uses $\hat{\lambda}$, the MLE of λ associated with the truncated Poisson distribution from the previous iteration (or from the one- and *f*-truncated Poisson distribution if $m = f$), to estimate the expected number of non-fragile bands having each break count in the range

of the data. Bands for which these expected frequencies are less than a preset threshold are flagged as fragile. The default threshold for FSM3 is set to 0.10. (See Olmsted (1999) for details of the FSM3 algorithm.)

1.7 Correlated Binary Variables

As mentioned previously, the FSM and FSM3 algorithms are based on three independence assumptions listed in Section 1.4. The third independence assumption states that X_{ij1} and X_{ij2} are independent, i.e., that breaks occurring in the same bands on homologous chromosomes in the same cell are independent. In this section we give a biologically plausible explanation of why this assumption is likely violated by the presence of positive correlation between homologs. We also provide a review of statistical methods applied to correlated binary variables.

1.7.1 Biological Explanation for Correlation

Our consideration of correlation between homologs is motivated by the possibility of an underlying codominant, Mendelian model of fragile site inheritance. In investigating the heritability of fragile sites, researchers have found evidence that some fragile sites follow a Mendelian mode of inheritance (Sherman and Sutherland 1986). In a Mendelian model, we assume that unlinked alleles segregate independently of one another as they pass from one generation to the next. For purposes of discussion, let *F*

Table 1. Breakage Probabilities in a 2 x 2 Contingency Table

		X_{ij2} = 0	X_{ij2} = 1	
X_{ij1}	0	p_{00_i}	p_{01_i}	$1-\pi_i$
	1	p_{10_i}	p_{11_i}	π_i
		$1-\pi_i$	π_i	1

denote a fragile allele at some site, and let *f* indicate the presence of a non-fragile allele. All individuals inherit one allele from their mother and one from their father. Thus, an individual can be either homozygous fragile, *FF*, heterozygous, *Ff*, or homozygous non-fragile, *ff*. In reality, the genotype and parentage, however, are unknown.

Under a Mendelian model for the expression of fragility and if an individual is homozygous fragile (*FF*) at a certain locus, then the third assumption of Section 1.4 is tenuous. For an individual homozygous at band *i*, observing a break in the maternal homolog would logically raise the likelihood of a break in the paternal homolog (or vice versa) if band *i* expresses fragility. Mathematically, this can be expressed as $P(X_{ij1}=1 \mid X_{ij2}=1) > P(X_{ij1}=1 \mid X_{ij2}=0)$. Consider a single pair of homologous chromosomes (X_{ij1}, X_{ij2}) with breakage probabilities represented in a two by two contingency table (Table 1). With $p_{01_i} = p_{10_i}$, write

$$
\begin{aligned}
& P(X_{ij1}=1 \mid X_{ij2}=1) > P(X_{ij1}=1 \mid X_{ij2}=0) \\
& \Leftrightarrow \frac{p_{11_i}}{\pi_i} > \frac{p_{10_i}}{1-\pi_i} \\
& \Leftrightarrow p_{11_i}(1-\pi_i) > p_{10_i}\pi_i \\
& \Leftrightarrow p_{11_i}\left(p_{10_i} + p_{00_i}\right) > p_{10_i}\left(p_{11_i} + p_{01_i}\right) \\
& \Leftrightarrow p_{11_i} p_{00_i} > {p_{10_i}}^2 \\
& \Leftrightarrow p_{11_i}(1-2p_{10_i} - p_{11_i}) > {p_{10_i}}^2 \\
& \Leftrightarrow p_{11_i} - {p_{10_i}}^2 - 2p_{11_i}p_{10_i} - {p_{11_i}}^2 > 0 \\
& \Leftrightarrow p_{11_i} - \left(p_{10_i} + p_{11_i}\right)^2 > 0 \\
& \Leftrightarrow p_{11_i} - {\pi_i}^2 > 0 \\
& \Leftrightarrow Cov(X_{ij1}, X_{ij2}) > 0 \\
& \Leftrightarrow Corr(X_{ij1}, X_{ij2}) > 0.
\end{aligned}
$$

Thus, $P(X_{ij1}=1 \mid X_{ij2}=1) > P(X_{ij1}=1 \mid X_{ij2}=0)$ implies that the correlation between homologs must be positive. Individuals who are homozygous non-fragile (*ff*) or heterozygous (*Ff*) would not be expected to exhibit the same type of correlation since breaks at non-fragile sites are, by definition, random events. Therefore, we investigate the effects of positive correlation between homologs on the FSM and FSM3 algorithms and investigate this type of correlation for fragile sites only. We do concede that the Mendelian model described may not, in fact, be the correct model for fragile site inheritance, but for the purposes of this research, we will assume that correlation arises from Mendelian inheritance of fragile sites.

1.7.2 Methods for Modeling Correlated Binary Variables

Correlated binary data arise in a wide variety of applications. In ophthalmologic studies, for example, measurements are often made on both eyes from a single individual. These observations are highly correlated. Correlated binary data also arise in reproductive toxicity studies involving teratogenic, mutagenic or carcinogenic chemicals administered to laboratory animals. Responses measured on several littermates are frequently binary in nature, i.e. each animal is determined to be dead or alive, affected or normal, etc. Much of the research in correlated binary methodology has been motivated by experiments in reproductive toxicology. A review of methods for analyzing dichotomous response data from toxicological experiments can be found in Haseman and Kupper (1979).

Originally, for treatment groups indexed by *i* and litters indexed by *j*, the number of successes (affected fetuses), x_{ij}, in litters of size n_{ij} were assumed to follow either a $Binomial(n_{ij}, \pi_{ij})$ distribution (Rosenzweig and Blaustein 1970; Zawoiski 1975; Krüger 1970; Salsburg 1973) or a $Poisson(\lambda_{ij})$ distribution (Epstein et al. 1970, 1972; Dean, Doak and Somerville 1975). For each treatment group, the underlying probability of success, π_{ij}, in the former was generally assumed to be equal, i.e., $\pi_{ij} = \pi_i$ for all *j*. Similarly, the expected Poisson count of the latter, λ_{ij}, was generally assumed to be

equal to λ_i for all j. A number of investigators, however, have shown that the empirical distribution of fetal mortality often departs from a binomial model (Röhrborn 1968; McCaughran and Arnold 1976; Haseman and Soares 1976; Aeschbacher, Vuataz, Sotek and Stalder 1977) and from a Poisson model (Haseman and Soares 1976; McCaughran and Arnold 1976).

Williams (1975) proposed a beta-binomial model for data from toxicological experiments that accounts for extra-binomial variation observed in the data caused by correlation between responses within each litter. In general, for x_{ij} successes among n_{ij} observations in the j^{th} litter of the i^{th} treatment group, the model assumes that x_{ij} follows a *Binomial*(n_{ij}, π_{ij}) distribution where π_{ij} is a random variable from a *Beta*(α_i, β_i) distribution, so that the marginal distribution of x_{ij} is

$$\Pr\left(X_{ij} = x_{ij} \mid n_{ij}\right) = \binom{n_{ij}}{x_{ij}} \frac{B\left(\alpha_i + x_{ij}, n_{ij} + \beta_i - x_{ij}\right)}{B\left(\alpha_i, \beta_i\right)}, \tag{1.16}$$

where $B\left(\alpha_i, \beta_i\right) = \Gamma\left(\alpha_i\right)\Gamma\left(\beta_i\right)/\Gamma\left(\alpha_i + \beta_i\right)$. The model is reparameterized as $\mu_i = \alpha_i\left(\alpha_i + \beta_i\right)^{-1}$ and $\psi_i = \left(\alpha_i + \beta_i\right)^{-1}$. Estimates of μ_i and ψ_i are obtained using maximum likelihood. ψ_i determines the shape of the distribution whose variance is $\mu_i\left(1 - \mu_i\right)\psi_i\left(1 + \psi_i\right)^{-1}$. When $\psi_i = 0$, the model reduces to the simple, independent binomial model. Williams used this model for calculating asymptotic likelihood ratio tests for differences between treatment groups. In the case relevant to chromosomal breakage data, i.e., when $n_{ij} = 2$, the beta-binomial model is identical to the additive and correlated-binomial models of Altham (1978) and Kupper and Haseman (1978), respectively, described below. Tests for homogeneity of proportions against beta-binomial alternatives have been derived when proportions are known (Potthoff and Whittinghill 1966; Paul et al. 1989) and unknown (Gart 1970).

Kupper and Haseman (1978) used results from Bahadur (1961) to create the "correlated binomial" model applicable to modeling data from toxicological experiments involving littermates. The model has the form

$$P(X_{ij} = x_{ij} \mid n_{ij}) = \binom{n_{ij}}{x_{ij}} \pi_i^{x_{ij}} (1-\pi_i)^{n_{ij}-x_{ij}} \times \left\{ 1 + \frac{\theta_i}{2\pi_i^2 (1-\pi_i)^2} \left[(x_{ij} - n_{ij}\pi_i)^2 + x_{ij}(2\pi_i - 1) - n_{ij}\pi_i^2 \right] \right\}, \tag{1.17}$$

where θ_i is the pairwise covariance between responses in the same litter. The correlated binomial model is described in detail in Section 1.7.3.

Altham (1978) derived two generalizations of the binomial distribution that account for intragroup correlation. Altham called these the additive model (which is identical to the correlated binomial model of Kupper and Haseman (1978) given in (1.17)) and the multiplicative binomial model. The multiplicative binomial model has the form

$$\Pr\left(X_{ij} = x_{ij} \mid n_{ij}\right) = \frac{\binom{n_{ij}}{x_{ij}} \pi_i^{x_{ij}} (1-\pi_i)^{n_{ij}-x_{ij}} \delta_i^{x_{ij}(n_{ij}-x_{ij})}}{\sum_j \binom{n_{ij}}{x_{ij}} \pi_i^{x_{ij}} (1-\pi_i)^{n_{ij}-x_{ij}} \delta_i^{x_{ij}(n_{ij}-x_{ij})}}. \tag{1.18}$$

If $\delta_i = 1$, this generalization reduces to the binomial distribution. When $\delta_i > 1$, the distribution is (strongly) unimodal and is more sharply peaked than the binomial; this roughly corresponds to a negative association between the responses from the littermates. When $0 < \delta_i < 1$, the distribution is more diffuse than the binomial distribution and the responses are positively related. According to Altham (1978), the multiplicative model is less tractable and less interpretable than the additive (correlated binomial) model. Tarone (1979) derived C(α) tests for the goodness of fit of the binomial distribution which are asymptotically optimal against the generalized binomial alternatives given in Altham (1978) and Kupper and Haseman (1978).

McCaughran and Arnold (1976) suggested the use of a negative-binomial generalization to the Poisson model. The negative-binomial model is obtained by assuming that the Poisson parameter λ_{ij} follows a gamma distribution. Thus, the unconditional distribution of successes is negative binomial. McCaughran and Arnold

(1976) used the method of moments to estimate parameters, transformed the data in a manner that depends on the parameter estimates, and used analysis of variance techniques to test for differences between treatment and control groups.

Various other methods for modeling correlated binary data have been proposed. Paul (1982) suggested a jackknife method for modeling data from teratological experiments. Rosner (1982) studied ophthalmologic data and proposed an ANOVA model for detecting group differences that accounts for pairwise-correlated observations. Crowder (1985) suggested a Gaussian estimation technique for correlated binomial data. Others (Pack 1986; Makuch et al. 1989; Rudolfer 1990; Lipsitz et al. 1991; Rao and Scott 1992; George and Bowman 1995; Brooks et al. 1997) have also contributed methods for modeling correlated binary data applicable to various scientific disciplines.

While the research in modeling correlated binomial data is extensive, there is no method that can be directly applied to chromosomal breakage data. Previous research on data of this type focused on determining differences between treatment groups with litters as the experimental unit. Our purpose is to determine the effect of correlation on the sum of break counts from independent pairs of homologous chromosomes. We use results from the correlated binomial model of Kupper and Haseman (1978) and Altham (1978) in the special case where $n_{ij} = 2$ to determine the probabilities of observing zero, one, and two breaks in the presence of correlation. We have derived all other results.

1.7.3 The Correlated Binomial Model

Results from the additive correlated binomial model of Altham (1978) and Kupper and Haseman (1978) are applicable to chromosomal breakage data. The general model first will be presented following an example of Kupper and Haseman (1978), and then the model will be specialized to fragile site data.

Suppose that there are l_i litters in the i^{th} group ($i = 0, 1$), with the j^{th} litter in the i^{th} group being of size $n_{ij}, j = 1, 2,\ldots, l_i$. Let

$$X_{ij} = \sum_{h=1}^{n_{ij}} X_{ijh} , \tag{1.19}$$

where X_{ijh} takes on value 1 with probability π_i and 0 with probability $(1-\pi_i)$. When the assumption of independence is not reasonable, Badahur (1961) has shown that the correct expression for $P(X_{ij} = x_{ij})$ is given by

$$P(X_{ij} = x_{ij}) = \binom{n_{ij}}{x_{ij}} \pi_i^{x_{ij}} (1-\pi_i)^{n_{ij}-x_{ij}} \times f(x_{ij1}, x_{ij2}, ..., x_{ijn_{ij}}), \tag{1.20}$$

where $f(x_{ij1}, x_{ij2}, ..., x_{ijn_{ij}})$ corrects for the lack of mutual independence among the X_{ijh}'s. Kupper and Haseman (1978) demonstrated that if

$$Cov(X_{ijh}, X_{ijh'}) \equiv \theta_i, \tag{1.21}$$

such that

$$Corr(X_{ijh}, X_{ijh'}) \equiv \rho_i = \frac{\theta_i}{\pi_i(1-\pi_i)}, \tag{1.22}$$

then

$$\begin{aligned} P(X_{ij} = x_{ij}) = & \binom{n_{ij}}{x_{ij}} \pi_i^{x_{ij}} (1-\pi_i)^{n_{ij}-x_{ij}} \\ & \times \left\{ 1 + \frac{\theta_i}{2\pi_i^2 (1-\pi_i)^2} \left[(x_{ij} - n_{ij}\pi_i)^2 + x_{ij}(2\pi_i - 1) - n_{ij}\pi_i^2 \right] \right\}. \end{aligned} \tag{1.23}$$

In the special case where $n_{ij} = 2$ (which would be applicable to per-individual breakage data) the expression in (1.23) becomes

$$\begin{aligned} P(X_{ij} = x_{ij}) = & \binom{2}{x_{ij}} \pi_i^{x_{ij}} (1-\pi_i)^{2-x_{ij}} \\ & \times \left\{ 1 + \frac{\theta_i}{2\pi_i^2 (1-\pi_i)^2} \left[(x_{ij} - 2\pi_i)^2 + x_{ij}(2\pi_i - 1) - 2\pi_i^2 \right] \right\}, \end{aligned} \tag{1.24}$$

whence

$$\begin{aligned} P(X_{ij} = 0) &= (1-\pi_i)^2 + \theta_i, \\ P(X_{ij} = 1) &= 2\pi_i(1-\pi_i) - 2\theta_i, \\ P(X_{ij} = 2) &= \pi_i^2 + \theta_i. \end{aligned} \tag{1.25}$$

The three statements of (1.25) demonstrate how the presence of correlation between maternal and paternal homologs affects the probability of observing zero, one, and two breaks for a single homologous pair of chromosomes. Positive correlation, which corresponds to $\theta_i > 0$, results in a greater number of zero-breaks and double-breaks and fewer single-breaks than would be expected under a fully-independent binomial or Poisson model.

1.8 Overview

The main goal of this research is to determine the effect of pairwise correlation between bands on maternal and paternal homologs on the current FSM and FSM3 algorithms. The FSM and FSM3 algorithms are based on the assumption of complete independence between each observed break. In Chapter II we construct models for both BCT and PBCT that include the possibility that correlation between homologs exists. In Chapter III we derive statistical tests to detect correlation, provide simulation studies that compare the power and level achieved by each test, and apply the tests to a chromosomal breakage dataset. In Chapter IV we present simulation studies which determine the effect of various degrees of correlation on the FSM and FSM3 algorithms. Finally, in Chapter V we discuss the results of these simulation studies and state our conclusions.

CHAPTER II

CORRELATED BERNOULLI TRIALS MODEL

In the Correlated Bernoulli Trials (CorrBT) model, we retain all independence assumptions of FSM and FSM3 except the third assumption of Section 1.4. We consider the situation where correlation between identical bands on homologous chromosomes may exist, i.e., X_{ij1} and X_{ij2} are not necessarily independent. With the data structure as set forth in Section 1.3, define the covariance between X_{ij1} and X_{ij2} as

$$Cov\left(X_{ij1}, X_{ij2}\right) \equiv \theta_i \tag{2.1}$$

and the correlation between X_{ij1} and X_{ij2} as

$$Corr(X_{ij1}, X_{ij2}) \equiv \rho_i = \frac{\theta_i}{\sigma_{X_{ij1}} \sigma_{X_{ij2}}}, \tag{2.2}$$

where $\sigma_{X_{ijh}}$ is the standard deviation of X_{ijh}. If the probability of breakage at band i is π_i, then $\sigma_{X_{ijh}} = \sqrt{\pi_i\left(1-\pi_i\right)}$ regardless of the homolog, h. Therefore,

$$Corr(X_{ij1}, X_{ij2}) \equiv \rho_i = \frac{\theta_i}{\pi_i\left(1-\pi_i\right)}. \tag{2.3}$$

From (1.25) the probabilities of observing zero, one, and two breaks at band i in a single metaphase are, respectively,

$$\begin{aligned} P_{0_i} &\equiv P(X_{ij\bullet} = 0) = \left(1-\pi_i\right)^2 + \theta_i, \\ P_{1_i} &\equiv P(X_{ij\bullet} = 1) = 2\pi_i\left(1-\pi_i\right) - 2\theta_i, \\ P_{2_i} &\equiv P(X_{ij\bullet} = 2) = \pi_i^{\,2} + \theta_i. \end{aligned} \tag{2.4}$$

Using (2.4), the covariance can then be computed as

$$\theta_i = P_{2_i} - \pi_i^{\,2}. \tag{2.5}$$

If $\theta_i = 0$, then the probabilities in (2.4) reduce to binomial probabilities that are equivalent to those modeled by FSM. Hence, the binomial model used in the derivation of FSM is a special case of the CorrBT model.

In general, we assume that the observed bands can be divided into subsets based on their probability of breakage and correlation. This makes it possible to model various degrees of correlation for sites with different probabilities of breakage. A subset may include only a single band or consist of the entire set of observed sites. The derivations that follow will be presented using the generalized notation with the subscript *b* indicating that the results apply to a subset of sites. Simulation studies in Chapter III will be presented where the results derived in this chapter are applied to subsets of various sizes.

In this chapter we present the CorrBT model adapted to four different cases. The cases arise based on the type of breakage data reported (PBCT or BCT) and whether or not sites with break counts equal to zero are included in the analysis. The four cases are listed below.

1. PBCT and sites with observed break counts equal to zero included.
2. BCT and sites with observed break counts equal to zero included.
3. PBCT and sites with positive break counts only (i.e., sites with observed break counts equal to zero are excluded).
4. BCT and sites with positive break counts only (i.e., sites with observed break counts equal to zero are excluded).

The first two cases are motivated by FSM, where all sites (including those with no observed breaks) are included in the analysis. The latter two cases are motivated by FSM3, where the possible presence of zero-breakage sites leads to the exclusion of all sites for which no breaks were observed. The models for cases involving only positive break counts would be useful when the data potentially contains zero-breakage site contamination. As stated in Section 1.7.1, however, correlation is believed to exist for fragile sites only. Therefore, the CorrBT model should be used to detect correlation only in subsets of sites declared fragile. This would eliminate the possibility of including a zero-breakage site in the subset of fragile sites since a site must display some breaks to be declared fragile. Thus, the latter two cases are not applicable to identifying correlation in fragile site breakage data. We present them only for their potential application to a modified FSM3 algorithm, should modeling correlation prove to be useful in

identification of fragile sites. The distributions for the latter two cases are derived in this chapter and are briefly mentioned in Chapter III when deriving likelihood ratio test statistics for detecting correlation. The latter two cases are not included in any simulation studies and are not applied to breakage data.

2.1 Distribution of Partitioned Break Count Totals (PBCT) and Maximum Likelihood Estimators Under the CorrBT Model

We first consider the case where zero-, single-, and double-break counts are reported for each band. This type of information is usually observed by fragile-site researchers but is usually summarized into break count totals, which are reported and analyzed.

2.1.1 Distribution of PBCT, $\boldsymbol{M} = (\boldsymbol{M_0}, \boldsymbol{M_1}, \boldsymbol{M_2})$

Recall from (1.4) that M_{0_i} represents the number of metaphases examined for which no breaks are observed at band *i*, M_{1_i} represents the number of metaphases for which only one break is observed at band *i*, and M_{2_i} represents the number of metaphases examined for which a break is observed at band *i* in both homologs. For *c* metaphases with P_{0_i}, P_{1_i}, and P_{2_i} as in (2.4), the distribution of $\boldsymbol{M}_i = (M_{0_i}, M_{1_i}, M_{2_i})$ is multinomial with pmf

$$f_{\boldsymbol{M}_i}(\boldsymbol{m}_i) = \frac{c!}{m_{0_i}!m_{1_i}!m_{2_i}!} P_{0_i}^{\,m_{0_i}} P_{1_i}^{\,m_{1_i}} P_{2_i}^{\,m_{2_i}}. \tag{2.6}$$

Since $P_{0_i} = 1 - (P_{1_i} + P_{2_i})$ and $m_{0_i} = c - (m_{1_i} + m_{2_i})$, the pmf of (2.6) can be expressed in terms of P_{1_i} and P_{2_i} as

$$f_{M_{1_i}, M_{2_i}}(m_{1_i}, m_{2_i}) = \frac{c!}{(c-(m_{1_i}+m_{2_i}))!m_{1_i}!m_{2_i}!} \left(1 - (P_{1_i} + P_{2_i})\right)^{c-(m_{1_i}+m_{2_i})} P_{1_i}^{\,m_{1_i}} P_{2_i}^{\,m_{2_i}}. \tag{2.7}$$

Marginally,

$$\begin{aligned} M_{1_i} &\sim Binomial\left(c, P_{1_i}\right), \\ M_{2_i} &\sim Binomial\left(c, P_{2_i}\right). \end{aligned} \tag{2.8}$$

The PBCT likelihood for k bands is the product of the individual multinomial likelihoods of (2.7). That is,

$$L_{\boldsymbol{M_1},\boldsymbol{M_2}}\left(\boldsymbol{P_1}, \boldsymbol{P_2} \mid c, \boldsymbol{m_1}, \boldsymbol{m_2}\right) = \prod_{i=1}^{k} \frac{c!}{(c-(m_{1_i}+m_{2_i}))! m_{1_i}! m_{2_i}!} \left(1-\left(P_{1_i}+P_{2_i}\right)\right)^{c-(m_{1_i}+m_{2_i})} P_{1_i}^{\ m_{1_i}} P_{2_i}^{\ m_{2_i}}. \tag{2.9}$$

2.1.2 PBCT Maximum Likelihood Estimators of P_1, P_2, π, θ and ρ

Suppose a subset of bands, b, of size k_b, $k_b = 1, 2, \ldots, k$, are assumed to have the same probability of breakage, i.e. $\pi_1 = \pi_2 = \ldots = \pi_{k_b} \equiv \pi_b$, and the same covariance between maternal and paternal homologs, i.e. $\theta_1 = \theta_2 = \ldots = \theta_{k_b} \equiv \theta_b$. This subset may include only a single band or consist of the entire set of k observed sites. Further, let the members of the subset be indexed by i, where $i = 1, 2, \ldots, k_b$.

With $\boldsymbol{M}$ multinomially distributed, it is known (Johnson, Kotz, and Balakrishnan 1997, p. 51) that the MLEs of P_{1_b} and P_{2_b} are

$$\hat{P}_{1_b} = \frac{\sum_{i=1}^{k_b} m_{1_i}}{ck_b} \text{ and } \hat{P}_{2_b} = \frac{\sum_{i=1}^{k_b} m_{2_i}}{ck_b}, \tag{2.10}$$

respectively. By the invariance property of MLEs (Casella and Berger 2002), the maximum likelihood estimators of π_b and θ_b can then be computed using (2.4), $\hat{P}_{1_b}$ and $\hat{P}_{2_b}$ using the identity

$$\begin{aligned} P_1 + 2P_2 &= 2\pi(1-\pi) - 2\theta + 2\left(\pi^2 + \theta\right) \\ &= 2\pi - 2\pi^2 - 2\theta + 2\pi^2 + 2\theta \\ &= 2\pi. \end{aligned} \tag{2.11}$$

The MLE of π_b is

$$\hat{\pi}_b = \frac{\hat{P}_{1_b} + 2\hat{P}_{2_b}}{2ck_b} = \frac{\sum_{i=1}^{k_b}\left(m_{1_i} + 2m_{2_i}\right)}{2ck_b} = \frac{\sum_{i=1}^{k_b} n_i}{2ck_b}, \tag{2.12}$$

where n_i can be computed as in (1.2). Note that (2.12) implies the MLE of π_b under the CorrBT model is exactly the same that of the fully-independent Bernoulli trials model used by FSM and FSM3. That is, under the CorrBT model and using the partitioned break counts, modeling correlation does not affect the MLE of π_b. The MLE of θ_b using (2.5) with (2.12) is

$$\hat{\theta}_b = \hat{P}_{2_b} - \hat{\pi}_b^{\,2} = \frac{4ck_b\sum_{i=1}^{k_b} m_{2_b} - \left(\sum_{i=1}^{k_b} n_i\right)^2}{4c^2k_b^{\,2}}. \tag{2.13}$$

Using (2.3) and (2.13), the MLE of ρ_b is

$$\hat{\rho}_b = \frac{\hat{\theta}_b}{\hat{\pi}_b\left(1-\hat{\pi}_b\right)} = \frac{4ck_b\sum_{i=1}^{k_b} m_{2_i} - \left(\sum_{i=1}^{k_b} n_i\right)^2}{2ck_b\sum_{i=1}^{k_b} n_i - \left(\sum_{i=1}^{k_b} n_i\right)^2}. \tag{2.14}$$

2.2 Distribution of Break Count Totals (BCT) and Maximum Likelihood Estimators Under the CorrBT Model

We now consider the case where only break count totals for each band are known. This is the type of data most often reported by fragile site researchers (Barbi et al. 1984; Glover et al. 1984; Craig-Holmes et al. 1987; Yunis et al.1987; Nagesh Rao et al. 1988; Böhm et al. 1995; Denison et al. 2003).

2.2.1 Distribution of BCT, $\boldsymbol{N} = (N_1, ..., N_k)$

To obtain the distribution of N_i under the CorrBT Model, we use the distribution given in (2.7), $N_i = M_{1_i} + 2M_{2_i}$, and define

$$S \equiv M_1. \tag{2.15}$$

Therefore,

$$\begin{aligned} M_{1_i} &= S_i, \\ M_{2_i} &= \frac{N_i - S_i}{2}. \end{aligned} \tag{2.16}$$

Inserting the values from (2.16) into (2.7) and completing the change of variables we get

$$\begin{aligned} f_{S_i,N_i}(s_i, n_i) &= f_{M_{1_i},M_{2i}}\left(s_i, \frac{n_i - s_i}{2}\right) = \frac{c!\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^{c-\left(s_i+\frac{n_i-s_i}{2}\right)} P_{1_i}{}^{s_i} P_{2_i}{}^{\frac{n_i-s_i}{2}}}{\left(c - s_i + \left(\frac{n_i - s_i}{2}\right)\right)! s_i! \left(\frac{n_i - s_i}{2}\right)!} \\ &= \frac{c!\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^{c-\left(\frac{s_i+n_i}{2}\right)} P_{1_i}{}^{s_i} P_{2_i}{}^{\frac{n_i-s_i}{2}}}{\left(c - \frac{s_i + n_i}{2}\right)! s_i! \left(\frac{n_i - s_i}{2}\right)!}, \end{aligned} \tag{2.17}$$

where $n_i = 0, 1, 2, ..., 2c,$ and

$$s_i = \begin{cases} 1, 3, 5, ..., \min(n_i, 2c - n_i) \text{ if } n_i \text{ is odd.} \\ 0, 2, 4, ..., \min(n_i, 2c - n_i) \text{ if } n_i \text{ is even.} \end{cases} \tag{2.18}$$

The distribution of N_i is obtained from (2.17) by summing over all possible values of S_i for a fixed value of N_i. That is

$$\begin{aligned} f_{N_i}(n_i) = &\sum_{s_j\{even\}=0}^{\min(n_i, 2c-n_i)} \frac{c!\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^{c-\frac{s_j+n_i}{2}} P_{1_i}{}^{s_j} P_{2_i}{}^{\frac{n_i-s_j}{2}}}{\left(c - \frac{s_j + n_i}{2}\right)! s_j! \left(\frac{n_i - s_j}{2}\right)!} I_{\{even\}}(n_i) \\ &+ \sum_{s_j\{odd\}=1}^{\min(n_i, 2c-n_i)} \frac{c!\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^{c-\frac{s_j+n_i}{2}} P_{1_i}{}^{s_j} P_{2_i}{}^{\frac{n_i-s_j}{2}}}{\left(c - \frac{s_j + n_i}{2}\right)! s_j! \left(\frac{n_i - s_j}{2}\right)!} I_{\{odd\}}(n_i), \end{aligned} \tag{2.19}$$

where $n_i = 0, 1, 2, \ldots, 2c$, $s_j\{even\} = 0, 2, 4, \ldots, \min(n_i, 2c - n_i)$, $s_j\{odd\} = 1, 3, 5, \ldots, \min(n_i, 2c - n_i)$,

$$I_{\{even\}}(n_i) = \begin{cases} 1 & \text{if } n_i \text{ is even,} \\ 0 & \text{otherwise,} \end{cases} \tag{2.20}$$

and

$$I_{\{odd\}}(n_i) = \begin{cases} 1 & \text{if } n_i \text{ is odd.} \\ 0 & \text{otherwise.} \end{cases} \tag{2.21}$$

The mean of the distribution of N_i is

$$\begin{aligned} E(N_i) &= E(M_{1_i} + 2M_{2_i}) \\ &= cP_{1_i} + 2cP_{2_i} \\ &= 2c\pi_i. \end{aligned} \tag{2.22}$$

The variance of the distribution of N_i is

$$\begin{aligned} Var(N_i) &= Var(M_{1_i} + 2M_{2_i}) \\ &= Var(M_{1_i}) + 4Var(M_{2_i}) + 4Cov(M_{1_i}, M_{2_i}) \\ &= cP_{1_i}(1 - P_{1_i}) + 4cP_{2_i}(1 - P_{2_i}) - 4cP_{1_i}P_{2_i} \\ &= cP_{1_i}(1 - P_{1_i}) + 4cP_{2_i}(1 - (P_{1_i} + P_{2_i})). \end{aligned} \tag{2.23}$$

The mean and variance of the distribution of N_i are given for informational purposes only and will not be used in further derivations or calculations. The mean and variance formulas are derived based on properties of the multinomial distribution (Johnson, Kotz, and Balakrishnan 1997, p. 34). The likelihood for *k* bands is

$$L_{\boldsymbol{N}}(\boldsymbol{P_1}, \boldsymbol{P_2} \mid \boldsymbol{n}) = \prod_{i=1}^{k} f_{N_i}(n_i). \tag{2.24}$$

2.2.2 BCT Maximum Likelihood Estimators of P_1, P_2, π, θ and ρ

Suppose a subset of bands, b, of size k_b, $k_b = 1, 2, \ldots, k$, are assumed to have the same probability of breakage, i.e. $\pi_1 = \pi_2 = \ldots = \pi_{k_b} \equiv \pi_b$, and the same covariance between maternal and paternal homologs, i.e. $\theta_1 = \theta_2 = \ldots = \theta_{k_b} \equiv \theta_b$. This subset may include only a single band or consist of the entire set of k sites observed. Further, let the members of the subset be indexed by i, where $i = 1, 2, \ldots, k_b$.

The MLEs of P_{1_b}, P_{2_b}, θ_b, and ρ_b can be determined by first solving for the MLE of π_b, which is the same as that given in (2.12). By the invariance property of MLEs (Casella and Berger 2002), we can solve for the MLEs of P_{1_b} and P_{2_b} sequentially by solving for the MLE of P_{2_b} conditional on $\hat{\pi}_b$ and then using the identity in (2.11) to compute the MLE of P_{1_b}. Noting that $P_1 = 2\pi - 2P_2$, the likelihood for the k_b bands is

$$
\begin{aligned}
&L_{\boldsymbol{N}}\left(P_{2_b} \mid \boldsymbol{n}, \hat{\pi}_b\right) \\
&\quad = \prod_{i=1}^{k_b} f_{N_i}\left(n_i \mid \hat{\pi}_b\right) \\
&\quad = \prod_{i=1}^{k_b} \left[\sum_{s_j\{even\}=0}^{\min(n_i, 2c-n_i)} \frac{c!\left(1-2\hat{\pi}_b + P_{2_b}\right)^{c-\frac{s_j+n_i}{2}} \left(2\hat{\pi}_b - 2P_{2_b}\right)^{s_j} P_{2_b}^{\frac{n_i-s_j}{2}}}{\left(c-\frac{s_j+n_i}{2}\right)! s_j! \left(\frac{n_i-s_j}{2}\right)!} I_{\{even\}}(n_i) \right. \\
&\qquad \left. + \sum_{s_j\{odd\}=1}^{\min(n_i, 2c-n_i)} \frac{c!\left(1-2\hat{\pi}_b + P_{2_b}\right)^{c-\frac{s_j+n_i}{2}} \left(2\hat{\pi}_b - 2P_{2_b}\right)^{s_j} P_{2_b}^{\frac{n_i-s_j}{2}}}{\left(c-\frac{s_j+n_i}{2}\right)! s_j! \left(\frac{n_i-s_j}{2}\right)!} I_{\{odd\}}(n_i) \right].
\end{aligned}
\tag{2.25}
$$

In order to simplify the notation, define

$$
A_{ij} \equiv c - \frac{s_j + n_i}{2},\ B_{ij} \equiv \frac{n_i - s_j}{2},\ \text{and } C_{ij} \equiv \frac{B_{ij}}{P_{2_b}} - \frac{2s_j}{\left(2\hat{\pi}_b - 2P_{2_b}\right)} + \frac{A_{ij}}{\left(1-2\hat{\pi}_b + P_{2_b}\right)} \tag{2.26}
$$

The derivative of the log-likelihood from (2.25) with respect to P_{2_b} is

$$\frac{d\log L_{\boldsymbol{N}}\left(P_{2_b} \mid \boldsymbol{n}, \hat{\pi}_b\right)}{dP_{2_b}} = \sum_{i=1}^{k_b} \frac{\left(\frac{df_{N_i}\left(n_i \mid \hat{\pi}_b\right)}{dP_{2_b}}\right)}{f_{N_i}\left(n_i \mid \hat{\pi}_b\right)}, \tag{2.27}$$

where

$$\begin{aligned}\frac{df_{N_i}\left(n_i \mid \hat{\pi}_b\right)}{dP_{2_b}} = &\sum_{s_j\{even\}=0}^{\min(n_i,2c-n_i)} \frac{c!\left(1-2\hat{\pi}_b+P_{2_b}\right)^{A_{ij}}\left(2\hat{\pi}_b-2P_{2_b}\right)^{s_j} P_{2_b}{}^{B_{ij}} C_{ij}}{A_{ij}!s_j!B_{ij}!} I_{\{even\}}\left(n_i\right) \\ &+ \sum_{s_j\{odd\}=1}^{\min(n_i,2c-n_i)} \frac{c!\left(1-2\hat{\pi}_b+P_{2_b}\right)^{A_{ij}}\left(2\hat{\pi}_b-2P_{2_b}\right)^{s_j} P_{2_b}{}^{B_{ij}} C_{ij}}{A_{ij}!s_j!B_{ij}!} I_{\{odd\}}\left(n_i\right).\end{aligned} \tag{2.28}$$

The MLE of P_{2_b}, $\hat{P}_{2_b}$, solves

$$\sum_{i=1}^{k_b} \frac{\left(\frac{df_{N_i}\left(n_i \mid \hat{P}_{2_b}, \hat{\pi}_b\right)}{d\hat{P}_{2_b}}\right)}{f_{N_i}\left(n_i \mid \hat{P}_{2_b}, \hat{\pi}_b\right)} = 0. \tag{2.29}$$

The MLE of P_{1_b}, $\hat{P}_{1_b}$, is then

$$\hat{P}_{1_b} = 2\hat{\pi}_b - 2\hat{P}_{2_b}. \tag{2.30}$$

The MLE of θ_b based on (2.12) and (2.29) is

$$\hat{\theta}_b = \hat{P}_{2_b} - \hat{\pi}_b{}^2. \tag{2.31}$$

Using the estimators in (2.12) and (2.31), the MLE of ρ_b is

$$\hat{\rho}_b = \frac{\hat{\theta}_b}{\hat{\pi}_b\left(1-\hat{\pi}_b\right)}. \tag{2.32}$$

The R code written to calculate the MLEs of P_{1_b}, P_{2_b}, and π_b given by (2.30), (2.29), and (2.12), respectively, is presented in Appendix A.

2.3 Distribution of Partitioned Break Count Totals (PBCT) and Maximum Likelihood Estimators Under the CorrBT Model Where Only Positive Break Counts Are Included

2.3.1 Distribution of PBCT, $\boldsymbol{M} = (\boldsymbol{M_0}, \boldsymbol{M_1}, \boldsymbol{M_2})$, With Positive Counts Only

The conditional likelihood for positive counts is obtained by dividing the likelihood in (2.9) by the probability that the number of breaks is positive (which is equal to one minus the probability that $M_{0_i} = c$, $M_{1_i} = 0$, and $M_{2_i} = 0$). That is

$$L^{+}_{\boldsymbol{M_1},\boldsymbol{M_2}}(\boldsymbol{P_1}, \boldsymbol{P_2} \mid c, \boldsymbol{m_1}, \boldsymbol{m_2}) = \prod_{i=1}^{k_p} \frac{c!\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^{c-(m_{1_i}+m_{2_i})} P_{1_i}^{\,m_{1_i}} P_{2_i}^{\,m_{2_i}} I_{\{m_{1_i}>0 \text{ or } m_{2_i}>0\}}\left(m_{1_i}, m_{2_i}\right)}{(c-(m_{1_i}+m_{2_i}))!\, m_{1_i}!\, m_{2_i}!\left(1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^{c}\right)}, \tag{2.33}$$

where k_p is the number of sites with positive break counts.

2.3.2 PBCT Maximum Likelihood Estimators of P_1, P_2, π, θ and ρ With Positive Counts Only

Suppose a subset of bands with positive break counts, b_+, of size k_{b_+}, k_{b_+} = 1, 2, …, k_p, have the same probability of breakage, i.e. $\pi_1 = \pi_2 = ... = \pi_{k_{b_+}} \equiv \pi_{b_+}$, and the same covariance between homologs, i.e. $\theta_1 = \theta_2 = ... = \theta_{k_{b_+}} \equiv \theta_{b_+}$. This subset may include only a single band or consist of the entire set of the k_p sites with positive break counts. Further, let the members of the subset be indexed by i, where i = 1, 2, …, k_{b_+}.

The MLEs of $P_{1_{b_+}}$, $P_{2_{b_+}}$, θ_{b_+}, and ρ_{b_+} can be determined by first solving for the MLE of π_{b_+}. The positive binomial MLE of π_{b_+} given by Johnson, Kotz, and Kemp (1993) is the solution, $\hat{\pi}_{b_+}$, to the equation

$$\frac{\sum_{i=1}^{k_{b_+}} n_i}{2ck_{b_+}} = \frac{2c\hat{\pi}_{b_+}}{1-\left(1-\hat{\pi}_{b_+}\right)^{2c}}. \tag{2.34}$$

By the invariance property of MLEs (Casella and Berger 2002), we can solve for the MLEs of $P_{1_{b_+}}$ and $P_{2_{b_+}}$ sequentially by first solving for the MLE of $P_{2_{b_+}}$ conditional on $\hat{\pi}_{b_+}$ from (2.34) and then using the identity in (2.11) to compute the MLE of $P_{1_{b_+}}$. If we note that $P_1 = 2\pi - 2P_2$, then the likelihood for the k_{b_+} bands is

$$L^+_{\boldsymbol{M}_1,\boldsymbol{M}_2}\left(P_{2_{b_+}} \mid c, \boldsymbol{m_1}, \boldsymbol{m_2}, \hat{\pi}_{b_+}\right) = \prod_{i=1}^{k_{b_+}} \frac{c!\left(1-2\hat{\pi}_{b_+}+P_{2_{b_+}}\right)^{c-(m_{1_i}+m_{2_i})}\left(2\hat{\pi}_{b_+}-2P_{2_{b_+}}\right)^{m_{1_i}} P_{2_{b_+}}{}^{m_{2_i}} I_{\{m_{1_i}>0 \text{ or } m_{2_i}>0\}}\left(m_{1_i}, m_{2_i}\right)}{\left(1-\left(1-2\hat{\pi}_{b_+}+P_{2_{b_+}}\right)^c\right)(c-(m_{1_i}+m_{2_i}))!\, m_{1_i}!\, m_{2_i}!}. \quad (2.35)$$

The derivative with respect to $P_{2_{b_+}}$ of the log-likelihood in (2.35) is

$$\frac{d\log L^+_{\boldsymbol{M_1},\boldsymbol{M_2}}\left(P_{2_{b_+}} \mid c, \boldsymbol{m_1}, \boldsymbol{m_2}, \hat{\pi}_{b_+}\right)}{dP_{2_{b_+}}} = \frac{ck_{b_+} - (\sum_{i=1}^{k_{b_+}} m_{1_i} + \sum_{i=1}^{k_{b_+}} m_{2_i})}{\left(1-2\hat{\pi}_{b_+}+P_{2_{b_+}}\right)} - \frac{2\sum_{i=1}^{k_{b_+}} m_{1_i}}{\left(2\hat{\pi}_{b_+}-2P_{2_{b_+}}\right)} + \frac{\sum_{i=1}^{k_{b_+}} m_{2_i}}{P_{2_{b_+}}} + \frac{ck_{b_+}\left(1-2\hat{\pi}_{b_+}+P_{2_{b_+}}\right)^{c-1}}{1-\left(1-2\hat{\pi}_{b_+}+P_{2_{b_+}}\right)^c}. \quad (2.36)$$

So, the MLE of $P_{2_{b_+}}$, $\hat{P}_{2_{b_+}}$, solves

$$\frac{ck_{b_+} - (\sum_{i=1}^{k_{b_+}} m_{1_i} + \sum_{i=1}^{k_{b_+}} m_{2_i})}{\left(1-2\hat{\pi}_{b_+}+\hat{P}_{2_{b_+}}\right)} - \frac{2\sum_{i=1}^{k_{b_+}} m_{1_i}}{\left(2\hat{\pi}_{b_+}-2\hat{P}_{2_{b_+}}\right)} + \frac{\sum_{i=1}^{k_{b_+}} m_{2_i}}{\hat{P}_{2_{b_+}}} + \frac{ck_{b_+}\left(1-2\hat{\pi}_{b_+}+\hat{P}_{2_{b_+}}\right)^{c-1}}{1-\left(1-2\hat{\pi}_{b_+}+\hat{P}_{2_{b_+}}\right)^c} = 0. \quad (2.37)$$

The MLE of $P_{1_{b_+}}$, $\hat{P}_{1_{b_+}}$, is then

$$\hat{P}_{1_{b_+}} = 2\hat{\pi}_{b_+} - 2\hat{P}_{2_{b_+}}. \quad (2.38)$$

The MLE of θ_{b_+} based on (2.34) and (2.37) is

$$\hat{\theta}_{b_+} = \hat{P}_{2_{b_+}} - \hat{\pi}_{b_+}{}^2. \quad (2.39)$$

Using (2.34) and (2.39), the MLE of ρ_{b_+} is

$$\hat{\rho}_{b_+} = \frac{\hat{\theta}_{b_+}}{\hat{\pi}_{b_+}\left(1-\hat{\pi}_{b_+}\right)}. \quad (2.40)$$

The R code written to calculate the MLEs of $P_{1_{b_+}}$, $P_{2_{b_+}}$, and π_{b_+} given by (2.38), (2.37), and (2.34), respectively, is presented in Appendix A.

2.4 Distribution of Break Count Totals (BCT) and Maximum Likelihood Estimators Under the CorrBT Model Where Only Positive Break Counts Are Included

2.4.1 Distribution of BCT, $\boldsymbol{N} = (N_1,...,N_k)$, With Positive Counts Only

The conditional distribution of positive break counts is obtained by dividing the distribution of N_i in (2.25) by the probability that $N_i \neq 0$. That is,

$$f^+_{N_i}(n_i) = \frac{f_{N_i}(n_i) I_{\{n_i>0\}}(n_i)}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c}. \tag{2.41}$$

The mean of the positive counts distribution is

$$\begin{aligned} E(N_i \mid N_i>0) &= \sum_{n_i=1}^{2c} n_i f^+_{N_i}(n_i) = \frac{\sum_{n_i=0}^{2c} n_i f_{N_i}(n_i)}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c} \\ &= \frac{E(N_i)}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c} = \frac{cP_{1_i}+2cP_{2_i}}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c} \\ &= \frac{2c\pi_i}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c}. \end{aligned} \tag{2.42}$$

The variance of the positive counts distribution is

$$\begin{aligned} Var(N_i \mid N_i>0) &= E\left(N_i^{\,2} \mid N_i>0\right) - E(N_i \mid N_i>0)^2 \\ &= \sum_{n_i=1}^{2c} n_i^{\,2} f^+_{N_i}(n_i) - \left(\frac{cP_{1_i}+2cP_{2_i}}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^m}\right)^2 \\ &= \sum_{n_i=1}^{2c} n_i^{\,2} \frac{f_{N_i}(n_i \mid n_i>0)}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c} - \left(\frac{cP_{1_i}+2cP_{2_i}}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c}\right)^2 \end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_{n_i=0}^{2c} n_i^{\,2} f_{N_i}(n_i)}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c} - \left(\frac{cP_{1_i}+2cP_{2_i}}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c}\right)^2 \\
&= \frac{E\left(N_i^{\,2}\right)}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c} - \left(\frac{cP_{1_i}+2cP_{2_i}}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c}\right)^2 \\
&= \frac{Var\left(N_i\right)+E\left(N_i\right)^2}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c} - \left(\frac{cP_{1_i}+2cP_{2_i}}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c}\right)^2 \\
&= \frac{cP_{1_i}\left(1-P_{1_i}\right)+4cP_{2_i}\left(1-\left(P_{1_i}+P_{2_i}\right)\right)+\left(cP_{1_i}+2cP_{2_i}\right)^2}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c} \\
&\quad - \left(\frac{cP_{1_i}+2cP_{2_i}}{1-\left(1-\left(P_{1_i}+P_{2_i}\right)\right)^c}\right)^2 .
\end{aligned}
\tag{2.43}
$$

Again, the mean and variance of the positive counts distribution are given for informational purposes only and will not be used in further derivations or calculations. The mean and variance formulas are derived based on properties of the multinomial distribution (Johnson, Kotz, and Balakrishnan 1997, p. 34). The likelihood for k_p positive break count totals becomes

$$
L_N^+\left(\boldsymbol{P_1}, \boldsymbol{P_2} \mid \boldsymbol{n}\right) = \prod_{i=1}^{k_p} f_{N_i}^+\left(n_i\right), \tag{2.44}
$$

where k_p is the number of sites with positive break counts and $f_{N_i}^+\left(n_i\right)$ is given by (2.41).

2.4.2 BCT Maximum Likelihood Estimators of P_1, P_2, π, θ and ρ With Positive Counts Only

Suppose we select a subset of bands with positive break counts, b_+, of size k_{b_+}, k_{b_+} = 1, 2, …, k_p, with the same probability of breakage, i.e. $\pi_1 = \pi_2 = ... = \pi_{k_{b_+}} \equiv \pi_{b_+}$, and the same covariance between homologs, i.e. $\theta_1 = \theta_2 = ... = \theta_{k_{b_+}} \equiv \theta_{b_+}$. This subset may include only a single band or consist of the entire set of the k_p sites with positive

break counts. Further, let the members of the subset be indexed by i, where $i = 1, 2, \ldots, k_{b_+}$.

The MLEs of $P_{1_{b_+}}$, $P_{2_{b_+}}$, θ_{b_+}, and ρ_{b_+} can be determined by first solving for the MLE of π_{b_+}, which is the same as that given in (2.34). By the invariance property of MLEs (Casella and Berger 2002), we can solve for the MLEs of $P_{1_{b_+}}$ and $P_{2_{b_+}}$ sequentially by first solving for the MLE of $P_{2_{b_+}}$ conditional on $\hat{\pi}_{b_+}$ from (2.34) and then using the identity in (2.11) to compute the MLE of $P_{1_{b_+}}$. If we note that $P_1 = 2\pi - 2P_2$, then the likelihood for the k_{b_+} bands with positive break counts is

$$
\begin{aligned}
&L_{\boldsymbol{N}}^{+}\left(P_{2_{b_+}} \mid \boldsymbol{n}, \hat{\pi}_{b_+}\right) \\
&= \prod_{i=1}^{k_{b_+}} \left[\sum_{s_j\{even\}=0}^{\min(n_i, 2c-n_i)} \frac{c!\left(1-2\hat{\pi}_{b_+}+P_{2_{b_+}}\right)^{c-\frac{s_j+n_i}{2}} \left(2\hat{\pi}_{b_+}-2P_{2_{b_+}}\right)^{s_j} P_{2_{b_+}}^{\frac{n_i-s_j}{2}}}{\left(1-\left(1-\left(P_{1_{b_+}}+P_{2_{b_+}}\right)\right)^{c}\right)\left(c-\frac{s_j+n_i}{2}\right)!\, s_j!\left(\frac{n_i-s_j}{2}\right)!} I_{\{even\}}(n_i) \right. \qquad (2.45) \\
&\left. + \sum_{s_j\{odd\}=1}^{\min(n_i, 2c-n_i)} \frac{c!\left(1-2\hat{\pi}_{b_+}+P_{2_{b_+}}\right)^{c-\frac{s_j+n_i}{2}} \left(2\hat{\pi}_{b_+}-2P_{2_{b_+}}\right)^{s_j} P_{2_{b_+}}^{\frac{n_i-s_j}{2}}}{\left(1-\left(1-\left(P_{1_{b_+}}+P_{2_{b_+}}\right)\right)^{c}\right)\left(c-\frac{s_j+n_i}{2}\right)!\, s_j!\left(\frac{n_i-s_j}{2}\right)!} I_{\{odd\}}(n_i) \right].
\end{aligned}
$$

The derivative with respect to $P_{2_{b_+}}$ of the log-likelihood in (2.45) is

$$
\frac{d \log L_{\boldsymbol{N}}^{+}\left(P_{2_{b_+}} \mid \boldsymbol{n}, \hat{\pi}_{b_+}\right)}{dP_{2_{b_+}}} = \sum_{i=1}^{k_{b_+}} \frac{\left(\frac{df_{N_i}\left(n_i \mid \hat{\pi}_{b_+}\right)}{dP_{2_{b_+}}}\right)}{f_{N_i}\left(n_i \mid \hat{\pi}_{b_+}\right)} + \frac{c k_{b_+}\left(1-2\hat{\pi}_{b_+}+P_{2_{b_+}}\right)^{c-1}}{1-\left(1-2\hat{\pi}_{b_+}+P_{2_{b_+}}\right)^{c}}, \qquad (2.46)
$$

where $\frac{df_{N_i}\left(n_i \mid \hat{\pi}_{b_+}\right)}{dP_{2_{b_+}}}$ is defined as in (2.28). So, the MLE of $P_{2_{b_+}}$, $\hat{P}_{2_{b_+}}$, solves

$$\sum_{i=1}^{k_{b_+}} \frac{\left(\frac{df_{N_i}\left(n_i \mid \hat{P}_{2_{b_+}}, \hat{\pi}_{b_+}\right)}{d\hat{P}_{2_{b_+}}} \right)}{f_{N_i}\left(n_i \mid \hat{\pi}_{b_+}\right)} + \frac{ck_{b_+}\left(1-2\hat{\pi}_{b_+}+\hat{P}_{2_{b_+}}\right)^{c-1}}{1-\left(1-2\hat{\pi}_{b_+}+\hat{P}_{2_{b_+}}\right)^{c}} = 0. \tag{2.47}$$

The MLE of $P_{1_{b_+}}$, $\hat{P}_{1_{b_+}}$, is then

$$\hat{P}_{1_{b_+}} = 2\hat{\pi}_{b_+} - 2\hat{P}_{2_{b_+}}. \tag{2.48}$$

The MLE of θ_{b_+} based on (2.34) and (2.47) is

$$\hat{\theta}_{b_+} = \hat{P}_{2_{b_+}} - \hat{\pi}_{b_+}{}^2. \tag{2.49}$$

Using (2.34) and (2.49), the MLE of ρ_{b_+} is

$$\hat{\rho}_{b_+} = \frac{\hat{\theta}_{b_+}}{\hat{\pi}_{b_+}\left(1-\hat{\pi}_{b_+}\right)}. \tag{2.50}$$

The R code written to calculate the MLEs of $P_{1_{b_+}}$, $P_{2_{b_+}}$, and π_{b_+} given by (2.48), (2.47), and (2.34), respectively, is presented in Appendix A.

CHAPTER III

HYPOTHESIS TESTS FOR CORRELATION

With a model which accounts for correlation, we can now construct and evaluate hypothesis tests to detect non-zero correlation. Recall that correlation in this context is assumed to be between identical sites on homologous chromosomes (i.e., for homozygous fragile sites). We present two hypothesis tests which use the PBCT, Neyman's C(α) test and a Likelihood Ratio (LR) test, and a LR test which uses the BCT. These three tests are evaluated in a comprehensive simulation study of type I error rates and power for various degrees of correlation and simulation parameters based on practical experimental situations. Neyman's C(α) test, which will be shown to be the most powerful, is then applied to fragile site data obtained from 14 human subjects (Denison et al. 2003).

3.1 Neyman's C(α) Test for Correlation Using PBCT

Rao (1963), among others, proposed that the efficient score vector be used in testing a composite hypothesis about parameters in a likelihood-based model. For a hypothesis about the true values of the parameter vector $\boldsymbol{\beta}$, such as $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, Lachin (2000) shows that the efficient scores test statistic for testing the null hypothesis is

$$X_S^2 = \boldsymbol{U}(\boldsymbol{\beta}_0)^T \boldsymbol{I}(\boldsymbol{\beta}_0)^{-1} \boldsymbol{U}(\boldsymbol{\beta}_0), \tag{3.1}$$

where $\boldsymbol{U}(\boldsymbol{\beta}_0)$ is the score vector and $\boldsymbol{I}(\boldsymbol{\beta}_0)$ is the expected information matrix, both evaluated using the null hypothesized values of $\boldsymbol{\beta}$. The test statistic, X_S^2, is asymptotically distributed as χ^2 with degrees of freedom equal to the dimension of $\boldsymbol{\beta}$.

When constructing hypothesis tests, it is often necessary to estimate nuisance parameters whose values affect the outcome of the test. Neyman (1959) proposed an efficient score test, called the C(α) test, for a sub-hypothesis regarding elements of the parameter vector $\boldsymbol{\beta}$, where α denotes nuisance parameters that must be estimated. We

derive the C(α) test following the presentation of Lachin (2000) for the case where $\boldsymbol{\beta} = [\pi, \theta]^T$ in the notation of Chapter II. Though we derive the test for only two parameters, the theory extends to tests involving more than two parameters. We wish to test $H_0 : \theta = \theta_0$, but in doing so we must estimate π. Therefore, the composite null hypothesis becomes $H_{\boldsymbol{\beta}_0} : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = [\pi, \theta_0]^T$, where the value of π is unrestricted. If $L(\boldsymbol{\beta})$ is the likelihood of $\boldsymbol{\beta}$, then under the composite null hypothesis, the bivariate score vector is

$$\boldsymbol{U}(\boldsymbol{\beta}_0) = [U(\boldsymbol{\beta})_\pi \quad U(\boldsymbol{\beta})_\theta]^T, \tag{3.2}$$

where

$$U(\boldsymbol{\beta})_{\beta_i} = \frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_i}. \tag{3.3}$$

Because the nuisance parameter, π, is unrestricted, we must estimate π under the null hypothesis that $\theta = \theta_0$. We designate this estimate as $\hat{\pi}_0$, which is the solution to

$$U(\boldsymbol{\beta}_0)_\pi = 0. \tag{3.4}$$

Thus, the resulting parameter vector under $H_{\boldsymbol{\beta}_0}$ is $\hat{\boldsymbol{\beta}}_0 = [\hat{\pi}_0, \theta_0]^T$. The resulting bivariate score vector then becomes

$$\boldsymbol{U}(\hat{\boldsymbol{\beta}}_0) = [0 \quad U(\hat{\boldsymbol{\beta}}_0)_\theta]^T, \tag{3.5}$$

since by definition $U(\hat{\boldsymbol{\beta}}_0)_\pi = 0$. The corresponding estimated information matrix is

$$\boldsymbol{I}(\hat{\boldsymbol{\beta}}_0) = \boldsymbol{I}(\boldsymbol{\beta})_{|\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_0} = -E\left[\begin{array}{cc} \dfrac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \pi^2} & \dfrac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \pi \partial \theta} \\ \dfrac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \theta \partial \pi} & \dfrac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \theta^2} \end{array}\right]_{|\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_0}. \tag{3.6}$$

The general form of the C(α) test statistic is

$$X^2_{C(\alpha)} = \boldsymbol{U}(\hat{\boldsymbol{\beta}}_0)^T \boldsymbol{I}(\hat{\boldsymbol{\beta}}_0)^{-1} \boldsymbol{U}(\hat{\boldsymbol{\beta}}_0), \tag{3.7}$$

which has a form similar to that of (3.1), but differs in its use of $\hat{\boldsymbol{\beta}}_0$ rather than $\boldsymbol{\beta}_0$.

In order to apply (3.7) to fragile site data, we first select a subset of bands as outlined in Section 2.1.2 and assume that all k_b bands have the same probability of breakage, π_b, and the same covariance, θ_b. To determine the exact form of the C(α) test statistic for the test of the null hypothesis, $H_0 : \theta_b = 0$, we first write the likelihood of (2.9) in terms of π_b and θ_b using the identities in (2.4). That is,

$$L_{\boldsymbol{M_1},\boldsymbol{M_2}}(\pi_b, \theta_b \mid c, \boldsymbol{m_1}, \boldsymbol{m_2}) = \prod_{i=1}^{k_b} \frac{c!\left((1-\pi_b)^2 + \theta_b\right)^{c-(m_{1_i}+m_{2_i})} \left(2\pi_b(1-\pi_b) - 2\theta_b\right)^{m_{1_i}} \left({\pi_b}^2 + \theta_b\right)^{m_{2_i}}}{(c-(m_{1_i}+m_{2_i}))! m_{1_i}! m_{2_i}!}. \quad (3.8)$$

Thus, the log-likelihood is proportional to

$$\log L_{\boldsymbol{M_1},\boldsymbol{M_2}}(\pi_b, \theta_b \mid c, \boldsymbol{m_1}, \boldsymbol{m_2}) \propto \left(\sum_{i=1}^{k_b} \left(c-(m_{1_i}+m_{2_i})\right)\right) \log\left((1-\pi_b)^2 + \theta_b\right) + \left(\sum_{i=1}^{k_b} m_{1_i}\right) \log\left(\pi_b(1-\pi_b) - \theta_b\right) + \left(\sum_{i=1}^{k_b} m_{2_i}\right) \log\left({\pi_b}^2 + \theta_b\right). \quad (3.9)$$

Taking derivatives, it follows that

$$U(\pi_b, \theta_b)_{\pi_b} = -\frac{2(1-\pi_b)\sum_{i=1}^{k_b}\left(c-(m_{1_i}+m_{2_i})\right)}{\left((1-\pi_b)^2 + \theta_b\right)} + \frac{(1-2\pi_b)\left(\sum_{i=1}^{k_b} m_{1_i}\right)}{\left(\pi_b(1-\pi_b) - \theta_b\right)} + \frac{2\pi_b\left(\sum_{i=1}^{k_b} m_{2_i}\right)}{\left({\pi_b}^2 + \theta_b\right)} \quad (3.10)$$

and

$$U(\pi_b, \theta_b)_{\theta_b} = \frac{\sum_{i=1}^{k_b}\left(c-(m_{1_i}+m_{2_i})\right)}{\left((1-\pi_b)^2 + \theta_b\right)} - \frac{\left(\sum_{i=1}^{k_b} m_{1_i}\right)}{\left(\pi_b(1-\pi_b) - \theta_b\right)} + \frac{\left(\sum_{i=1}^{k_b} m_{2_i}\right)}{\left({\pi_b}^2 + \theta_b\right)}. \quad (3.11)$$

Using (3.4) together with (3.10), the MLE of π_b when $\theta_b = 0$ is the solution to the equation

$$-\frac{2(1-\hat{\pi}_{b0})\sum_{i=1}^{k_b}\left(c-(m_{1_i}+m_{2_i})\right)}{\left((1-\hat{\pi}_{b0})^2\right)} + \frac{(1-2\hat{\pi}_{b0})\left(\sum_{i=1}^{k_b} m_{1_i}\right)}{\left(\hat{\pi}_{b0}(1-\hat{\pi}_{b0})\right)} + \frac{2\hat{\pi}_{b0}\left(\sum_{i=1}^{k_b} m_{2_i}\right)}{\left({\hat{\pi}_{b0}}^2\right)} = 0, \quad (3.12)$$

which is identical to the estimator given in (2.12), that is,

$$\hat{\pi}_{b0} = \frac{\sum_{i=1}^{k_b}\left(m_{2_i} + 2m_{2_i}\right)}{2ck_b} = \frac{\sum_{i=1}^{k_b} n_i}{2ck_b}. \tag{3.13}$$

The bivariate score vector becomes

$$\boldsymbol{U}\left(\hat{\pi}_b, \theta_b = 0\right) = \left[0 \quad U\left(\hat{\pi}_b, \theta_b = 0\right)_{\theta_b}\right]^T, \tag{3.14}$$

where the second element is defined in (3.11). Taking second derivatives we have

$$\begin{aligned}
\frac{\partial^2 \log L(\pi_b, \theta_b)}{\partial {\pi_b}^2} &= \frac{\partial U(\pi_b, \theta_b)_{\pi_b}}{\partial \pi_b} \\
&= \frac{2\left(\left(1-\pi_b\right)^2 + \theta_b\right)\left(\sum_{i=1}^{k_b}\left(c - (m_{1_i} + m_{2_i})\right)\right) - 4\left(1-\pi_b\right)^2 \sum_{i=1}^{k_b}\left(c - (m_{1_i} + m_{2_i})\right)}{\left(\left(1-\pi_b\right)^2 + \theta_b\right)^2} \\
&\quad - \frac{2\left(\pi_b\left(1-\pi_b\right) - \theta_b\right)\left(\sum_{i=1}^{k_b} m_{1_i}\right) + \left(1-2\pi_b\right)^2 \sum_{i=1}^{k_b} m_{1_i}}{\left(\pi_b\left(1-\pi_b\right) - \theta_b\right)^2} \\
&\quad + \frac{2\left({\pi_b}^2 + \theta_b\right)\left(\sum_{i=1}^{k_b} m_{2_i}\right) - 4{\pi_b}^2 \sum_{i=1}^{k_b} m_{2_i}}{\left({\pi_b}^2 + \theta_b\right)^2},
\end{aligned} \tag{3.15}$$

$$\begin{aligned}
\frac{\partial^2 \log L(\pi_b, \theta_b)}{\partial {\theta_b}^2} &= \frac{\partial U(\pi_b, \theta_b)_{\theta_b}}{\partial \theta_b} \\
&= -\frac{\sum_{i=1}^{k_b}\left(c - (m_{1_i} + m_{2_i})\right)}{\left(\left(1-\pi_b\right)^2 + \theta_b\right)^2} - \frac{\left(\sum_{i=1}^{k_b} m_{1_i}\right)}{\left(\pi_b\left(1-\pi_b\right) - \theta_b\right)^2} - \frac{\left(\sum_{i=1}^{k_b} m_{2_i}\right)}{\left({\pi_b}^2 + \theta_b\right)^2},
\end{aligned} \tag{3.16}$$

and

$$\frac{\partial^2 \log L(\pi_b,\theta_b)}{\partial \pi_b \partial \theta_b} = \frac{\partial^2 \log L(\pi_b,\theta_b)}{\partial \theta_b \partial \pi_b} = \frac{\partial U(\pi_b,\theta_b)_{\pi_b}}{\partial \theta_b}$$

$$= \frac{2(1-\pi_b)\sum_{i=1}^{k_b}\left(c-(m_{1_i}+m_{2_i})\right)}{\left((1-\pi_b)^2+\theta_b\right)^2} + \frac{(1-2\pi_b)\left(\sum_{i=1}^{k_b} m_{1_i}\right)}{\left(\pi_b(1-\pi_b)-\theta_b\right)^2} - \frac{2\pi_b\left(\sum_{i=1}^{k_b} m_{2_i}\right)}{\left({\pi_b}^2+\theta_b\right)^2}. \quad (3.17)$$

Substituting the second derivatives of (3.15), (3.16), and (3.17) into (3.6) and finding the expectations, we obtain

$$\boldsymbol{I}(\hat{\pi}_{b0},\theta_b=0) = -E\begin{bmatrix} \frac{\partial^2 \log L(\pi_b,\theta_b)}{\partial {\pi_b}^2} & \frac{\partial^2 \log L(\pi_b,\theta_b)}{\partial \pi \partial \theta_b} \\ \frac{\partial^2 \log L(\pi_b,\theta_b)}{\partial \theta_b \partial \pi_b} & \frac{\partial^2 \log L(\pi_b,\theta_b)}{\partial {\theta_b}^2} \end{bmatrix}_{|\hat{\pi}_{b0},\theta_b=0}$$

$$= \begin{bmatrix} \frac{2ck_b}{\hat{\pi}_{b0}(1-\hat{\pi}_{b0})} & 0 \\ 0 & \frac{ck_b}{{\hat{\pi}_{b0}}^2(1-\hat{\pi}_{b0})^2} \end{bmatrix}, \quad (3.18)$$

and

$$\boldsymbol{I}(\hat{\pi}_{b0},\theta_b=0)^{-1} = \begin{bmatrix} \frac{\hat{\pi}_{b0}(1-\hat{\pi}_{b0})}{2ck_b} & 0 \\ 0 & \frac{{\hat{\pi}_{b0}}^2(1-\hat{\pi}_{b0})^2}{ck_b} \end{bmatrix}. \quad (3.19)$$

Using the PBCT and (3.7) along with (3.14) and (3.19), the C(α) test statistic for testing the null hypothesis, $H_0 : \theta_b = 0$, is

$$X^2_{C(\alpha)} = \left(\frac{\sum_{i=1}^{k_b}\left(c-(m_{1_i}+m_{2_i})\right)}{\left((1-\hat{\pi}_{b0})^2\right)} - \frac{\left(\sum_{i=1}^{k_b} m_{1_i}\right)}{\left(\hat{\pi}_{b0}(1-\hat{\pi}_{b0})\right)} + \frac{\left(\sum_{i=1}^{k_b} m_{2_i}\right)}{\left({\hat{\pi}_{b0}}^2\right)} \right)^2 \frac{{\hat{\pi}_{b0}}^2(1-\hat{\pi}_{b0})^2}{ck_b}$$

$$= \frac{\left(ck_b{\hat{\pi}_{b0}}^2 - \hat{\pi}_{b0}\sum_{i=1}^{k_b} n_i + \sum_{i=1}^{k_b} m_{2_i}\right)^2}{ck_b{\hat{\pi}_{b0}}^2(1-\hat{\pi}_{b0})^2}. \quad (3.20)$$

The statistic $X^2_{C(\alpha)}$ is asymptotically distributed as χ^2 with one degree of freedom.

Lachin (2000) gives the form of the statistic in (3.7) for the general case where $\boldsymbol{\beta}$ has more than two elements. This result can be applied to fragile-site modeling when simultaneously testing q subsets, $q = 1, 2, \ldots, k$, each with different assumed correlations and breakage probabilities. The generalized C(α) test statistic takes the form

$$X^2_{C(\alpha)} = \sum_{b=1}^{q} \frac{\left(ck_b \hat{\pi}_{b0}{}^2 - \hat{\pi}_{b0} \sum_{i=1}^{k_b} n_i + \sum_{i=1}^{k_b} m_{2_i} \right)^2}{ck_b \hat{\pi}_{b0}{}^2 \left(1 - \hat{\pi}_{b0}\right)^2}, \tag{3.21}$$

which is asymptotically distributed as χ^2 with q degrees of freedom.

3.2 Likelihood Ratio Test for Correlation Using PBCT

Derivation of the Likelihood Ratio (LR) test of the null hypothesis of zero correlation (or equivalently zero covariance) is straight-forward. The LR test statistic is simply computed using the ratio of the likelihood evaluated under the null hypothesis to the likelihood evaluated under the alternative hypothesis. More specifically, for a subset of k_b bands, direct evaluation of (3.8) yields

$$\begin{aligned} X^2_{LR(PBCT)} &= -2 \log \frac{L_{\boldsymbol{M_1},\boldsymbol{M_2}} \left(\hat{\pi}_{b0}, \theta_b = 0 \mid c, \boldsymbol{m_1}, \boldsymbol{m_2} \right)}{L_{\boldsymbol{M_1},\boldsymbol{M_2}} \left(\hat{\pi}_b, \hat{\theta}_b \mid c, \boldsymbol{m_1}, \boldsymbol{m_2} \right)} \\ &= 2 \sum_{f=1}^{3} \left(\sum_{i=1}^{k_b} m_{f_i} \right) \log \left(\frac{\sum_{i=1}^{k_b} m_{f_i}}{\hat{E}m_{fb}} \right), \end{aligned} \tag{3.22}$$

where

$$\begin{aligned} \hat{E}m_{0b} &= ck_b \left(1 - \hat{\pi}_{b0}\right)^2, \\ \hat{E}m_{1b} &= 2ck_b \hat{\pi}_{b0} \left(1 - \hat{\pi}_{b0}\right), \\ \hat{E}m_{2b} &= ck_b \hat{\pi}_{b0}{}^2, \end{aligned} \tag{3.23}$$

with $\hat{\pi}_{b0}$ defined as in (3.13). The statistic $X^2_{LR(PBCT)}$ is asymptotically distributed as χ^2 with one degree of freedom. When simultaneously testing q subsets, q = 1, 2, …, k, where each subset is assumed to have unique values of π_b and θ_b, the LR test statistic takes the form

$$X^2_{LR(PBCT)} = \sum_{b=1}^{q} X^2_{LR(PBCT)_b}, \tag{3.24}$$

with $X^2_{LR(PBCT)_b}$ defined as in (3.22) for a single subset. The test statistic in (3.24) is asymptotically distributed as χ^2 with q degrees of freedom.

It is possible to adapt this LR test to the situation where only sites with positive break counts are used to eliminate the effects of zero-breakage site contamination. The test statistic for a single subset using only positive counts becomes

$$X^2_{LR(PBCT)^+} = -2\log\frac{L^+_{\boldsymbol{M_1},\boldsymbol{M_2}}\left(\hat{\pi}_{b_+0}, \theta_{b_+} = 0 \mid c, \boldsymbol{m_1}, \boldsymbol{m_2}\right)}{L^+_{\boldsymbol{M_1},\boldsymbol{M_2}}\left(\hat{\pi}_{b_+}, \hat{\theta}_{b_+} \mid c, \boldsymbol{m_1}, \boldsymbol{m_2}\right)}, \tag{3.25}$$

where $L^+_{\boldsymbol{M_1},\boldsymbol{M_2}}\left(\pi, \theta \mid c, \boldsymbol{m_1}, \boldsymbol{m_2}\right)$ is defined as in (2.33). Furthermore, $\hat{\pi}_{b_+0}$ is equal to $\hat{\pi}_{b_+}$ as given in (2.34). The estimator $\hat{\theta}_{b_+}$ is calculated as in (2.39). The test statistic $X^2_{LR(PBCT)^+}$ is asymptotically distributed as χ^2 with one degree of freedom. As discussed at the beginning of Chapter II, since all sites declared fragile have positive break counts and there is no potential for zero-breakage sites among subsets of sites declared fragile, this test statistic based on positive counts will not receive further attention.

3.3 Likelihood Ratio Test for Correlation Using BCT

When only BCT are known, the LR test for correlation does not have an explicit form. As with the test in Section 3.2, the likelihood ratio test involves ratio of the BCT likelihood under the null hypothesis to that under the alternative hypothesis which allows non-zero correlation. Using (2.19) and (2.24), for a subset of bands as defined in

the previous section with homogeneous covariance, θ_b, and breakage probability, π_b, the likelihood of BCT can be written as

$$L_{\boldsymbol{N}}(\pi_b,\theta_b \mid \boldsymbol{n})$$

$$= \prod_{i=1}^{k_b}\left[\sum_{s_j\{even\}=0}^{\min(n_i,2c-n_i)} \frac{c!\left((1-\pi_b)^2+\theta_b\right)^{A_{ij}}\left(2\pi_b(1-\pi_b)-2\theta_b\right)^{s_j}\left(\pi_b{}^2+\theta_b\right)^{B_{ij}}}{A_{ij}!s_j!B_{ij}!} I_{\{even\}}(n_i) \right. \tag{3.26}$$

$$\left. + \sum_{s_j\{odd\}=1}^{\min(n_i,2c-n_i)} \frac{c!\left((1-\pi_b)^2+\theta_b\right)^{A_{ij}}\left(2\pi_b(1-\pi_b)-2\theta_b\right)^{s_j}\left(\pi_b{}^2+\theta_b\right)B_{ij}}{A_{ij}!s_j!B_{ij}!} I_{\{odd\}}(n_i)\right],$$

with A_{ij} and B_{ij} as defined in (2.26). The LR test statistic for testing the null hypothesis of no correlation is

$$X^2_{LR(BCT)} = -2\log\frac{L_{\boldsymbol{N}}(\hat{\pi}_{b0},\theta_{b0}=0 \mid \boldsymbol{n})}{L_{\boldsymbol{N}}(\hat{\pi}_b,\hat{\theta}_b \mid \boldsymbol{n})}. \tag{3.27}$$

The estimators $\hat{\pi}_{b0}$ and $\hat{\pi}_b$ are equivalent and defined in (3.13). The estimate of covariance, $\hat{\theta}_b$, is given by (2.31). The test statistic $X^2_{LR(BCT)}$ is asymptotically distributed as χ^2 with one degree of freedom. A simultaneous test of q subsets has a test statistic of the form

$$X^2_{LR(BCT)} = \sum_{b=1}^{q} X^2_{LR(BCT)b} \tag{3.28}$$

and is asymptotically distributed as χ^2 with q degrees of freedom.

When only positive break counts are analyzed, the likelihood ratio test statistic in (3.27) takes on a slightly different form. This test statistic is

$$X^2_{LR(BCT)^+} = -2\log\frac{L^+_{\boldsymbol{N}}(\hat{\pi}_{b_+0},\theta_{b_+0}=0 \mid \boldsymbol{n})}{L^+_{\boldsymbol{N}}(\hat{\pi}_{b_+},\hat{\theta}_{b_+} \mid \boldsymbol{n})}, \tag{3.29}$$

where $L^+_{\boldsymbol{N}}(\pi,\theta \mid \boldsymbol{n})$ is given by (2.44). The estimators $\hat{\pi}_{b_+0}$ and $\hat{\pi}_{b_+}$ are both computed as in (2.34). The estimator $\hat{\theta}_{b_+}$ is calculated as in (2.49). The statistic $X^2_{LR(BCT)^+}$ is asymptotically distributed as χ^2 with one degree of freedom. As discussed at the beginning of Chapter II, since all sites declared fragile have positive break counts and

there is no potential for zero-breakage site contamination among subsets of sites declared fragile, this test statistic will not receive further attention.

3.4 Simulation Studies

Simulation is a useful tool for studying complex situations in statistics. For model-based inference, such as that described in this research, Monte Carlo simulation is useful for estimating the true alpha level and power of statistical tests using empirical results. If the simulated data can be assumed to be random (or at least semi-random), the empirical estimates of power, alpha level, etc. based on simulated data are unbiased estimators of the true power, alpha level, etc. In our simulation study, we simulated data based on the number of metaphases analyzed and the multinomial probabilities of observing zero, one and two breaks (which depend on the assumed correlation) using a *Uniform*(0,1) random number generator. Since we know the true distribution from which the random numbers were generated, we can determine whether or not an inference made based on a given test is correct. The empirical estimates of alpha and power based on simulation follow a *Binomial*(s, p) distribution, where s is the number of simulated datasets and p is the proportion of null hypotheses rejected. Thus, confidence intervals can be constructed to portray the error associated with the empirical estimate. We construct 83% confidence intervals based on the normal approximation as

$$empirical\ estimate \pm 1.37\sqrt{\frac{p(1-p)}{s}}\,. \tag{3.30}$$

The individual confidence intervals are based on an alpha of 0.17 instead of an alpha of 0.05 because multiple confidence intervals are being plotted together. When making a pairwise comparison using two 83% confidence intervals, the effective level of the test for a difference is 0.05. If one is to make multiple comparisons, the confidence intervals should be further adjusted to control the overall error rate.

With the three test statistics described, we now present simulation studies to estimate the alpha level and power of our tests for various parameter combinations. Simulation parameters were chosen based on the fragile-site data and experimental

protocol given in Böhm et al. (1995). In this section we present only the graphs necessary to discuss the overall outcome of the simulation studies. We have included results for simulations where π_b is either 0.01 or 0.05 since the results depend on the probability of breakage (see discussions that follow). All simulations were performed using R (R Core Development Team (2003)) version 1.8.1. Complete simulated results are presented in Appendix B.

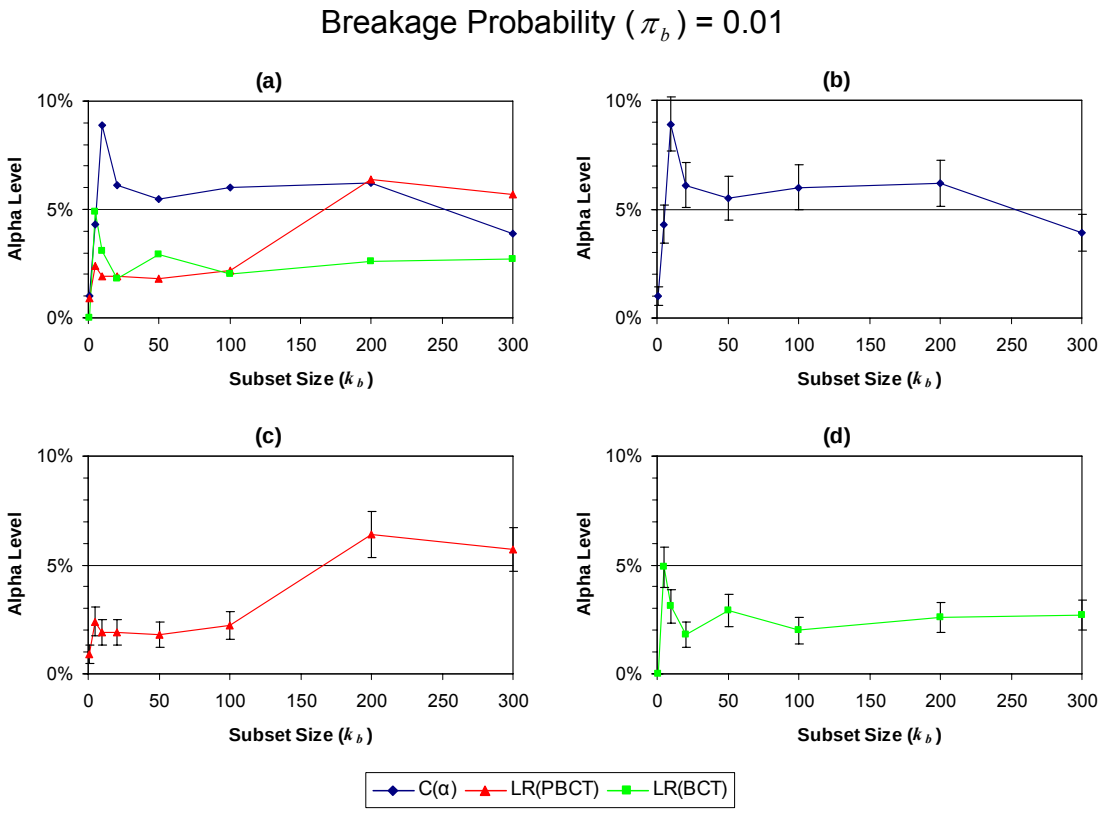


Figure 2. Simulated Alpha Level of Three Tests for Correlation Where the Breakage Probability Is 0.01. Simulated results are based on 1,000 Monte Carlo samples of chromosomal breakage data from 100 metaphases where correlation is equal to zero. The alpha level was computed as the percentage of simulations for which the null hypothesis of zero correlation was rejected. The curves in (a) are separated into plots (b), (c) and (d) and include 83% confidence intervals based on 1,000 simulations.

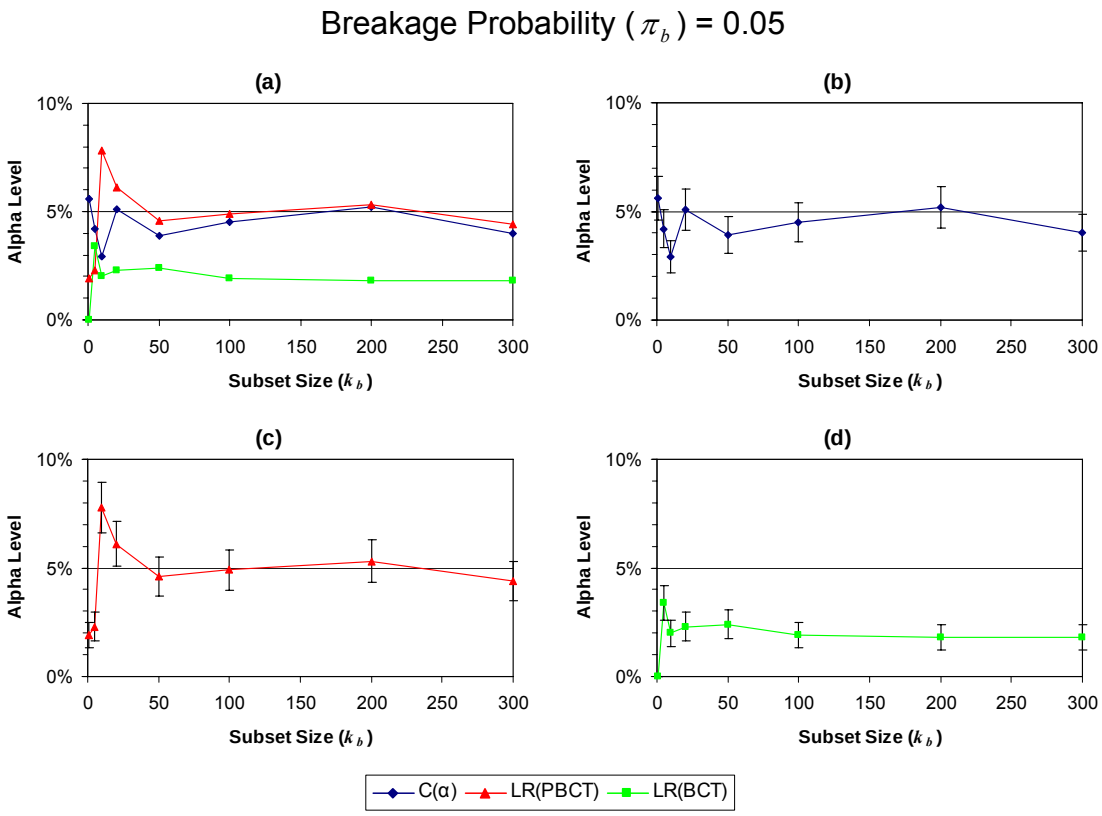


Figure 3. Simulated Alpha Level of Three Tests for Correlation Where the Breakage Probability Is 0.05. Simulated results are based on 1,000 Monte Carlo samples of chromosomal breakage data from 100 metaphases where correlation is equal to zero. The alpha level was computed as the percentage of simulations for which the null hypothesis of no correlation was rejected. The curves in (a) are separated into plots (b), (c) and (d) and include 83% confidence intervals based on 1,000 simulations.

3.4.1 Type I Error Rates (Alpha Level) in Detecting Correlation for Subsets of Size k_b

Prior to estimating of the power of our tests for correlation, we must estimate the type I error rate (or alpha level) of each test to ensure that power comparisons among tests are fair. Simulated results for a nominal alpha level of 5% are presented in Figure 2 and Figure 3. Each simulated result in this section is based on 1,000 Monte Carlo samples of induced chromosomal breakage data from 100 metaphases. No more than 1,000 Monte Carlo samples were taken because of the extensive time involved in maximizing the BCT likelihood given by (2.24).

Overall, the likelihood ratio test for correlation using only the break count totals (LR(BCT)) performed poorly. Regardless of the breakage probability or number of bands in a subset, Figure 2 and Figure 3 indicate that the true alpha level for LR(BCT) was consistently below the nominal 5% level. The only exception was found for a subset of five bands with breakage probability equal to 0.01 (Figure 2(d), first confidence interval).

Estimated levels for the C(α) and LR(PBCT) tests using partitioned break count totals were much closer to the nominal 5% level than was the case with the LR(BCT) test. For both C(α) and LR(PBCT) most of the simulated 83% confidence intervals for the true level covered the nominal 5% level when π_b is 0.05 and the subset size, k_b, is large (Figure 3), suggesting that the true alpha level for both C(α) and LR(PBCT) is close to the nominal level. Furthermore, the simulated 83% confidence interval for the true level of the C(α) test nearly contained the nominal level even when π_b is 0.01 and k_b is small, with exception of the test for a subset of ten bands (Figure 2(b), highest point). The LR(PBCT) test for correlation with π_b equal to 0.01 consistently demonstrated an estimated error rate significantly lower than the nominal level for subsets with 100 bands or less (Figure 2(c)), as demonstrated by the fact that the 83% confidence intervals do not cover the nominal level. Based on these results, we conclude that the C(α) test for correlation out-performs the others in achieving the nominal alpha level for nearly all subset sizes.

Note in Figure 2 that when testing for correlation in a single band, the estimated alpha level for all three tests was significantly below 5% when the breakage probability is 0.01 (Figure 2). Only the C(α) test had an estimated level not significantly different from the nominal level for a single site when the breakage probability is 0.05 (Figure 3). Sections 3.4.3 and 3.4.4 include further discussion on alpha levels and power for testing for correlation at a single band.

3.4.2 Power Curves for Detecting Correlation Using Subsets of Size k_b

We now present results of simulations designed to estimate the power of the C(α), LR(PBCT), and LR(BCT) tests for detecting correlation. Again, each simulation was performed using a nominal alpha level of 5%, 100 metaphases, and a Monte Carlo sample size of 1,000. A comprehensive set of the power plots is given in Appendix B.

Figure 4 and Figure 5 indicate that the C(α) test is significantly more powerful than its two competitors for both breakage probabilities examined. Moreover, we observe an increase in estimated power with larger breakage probabilities; greater

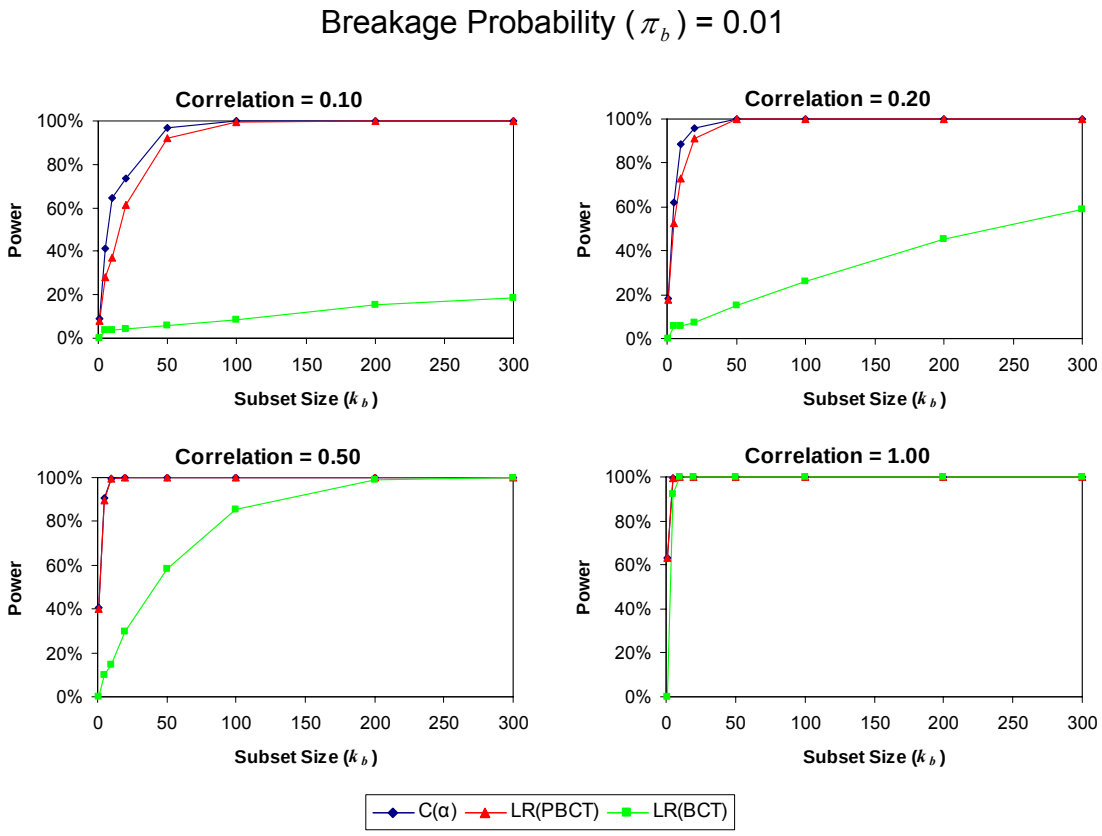


Figure 4. Simulated Power Curves of Three Tests for Correlation When the Probability of Breakage Is 0.01. Simulated results are based on 1,000 Monte Carlo samples with 100 metaphases. Correlations range from 0.10 to 1.00. Power was computed as the percentage of simulations for which the null hypothesis of no correlation was rejected.

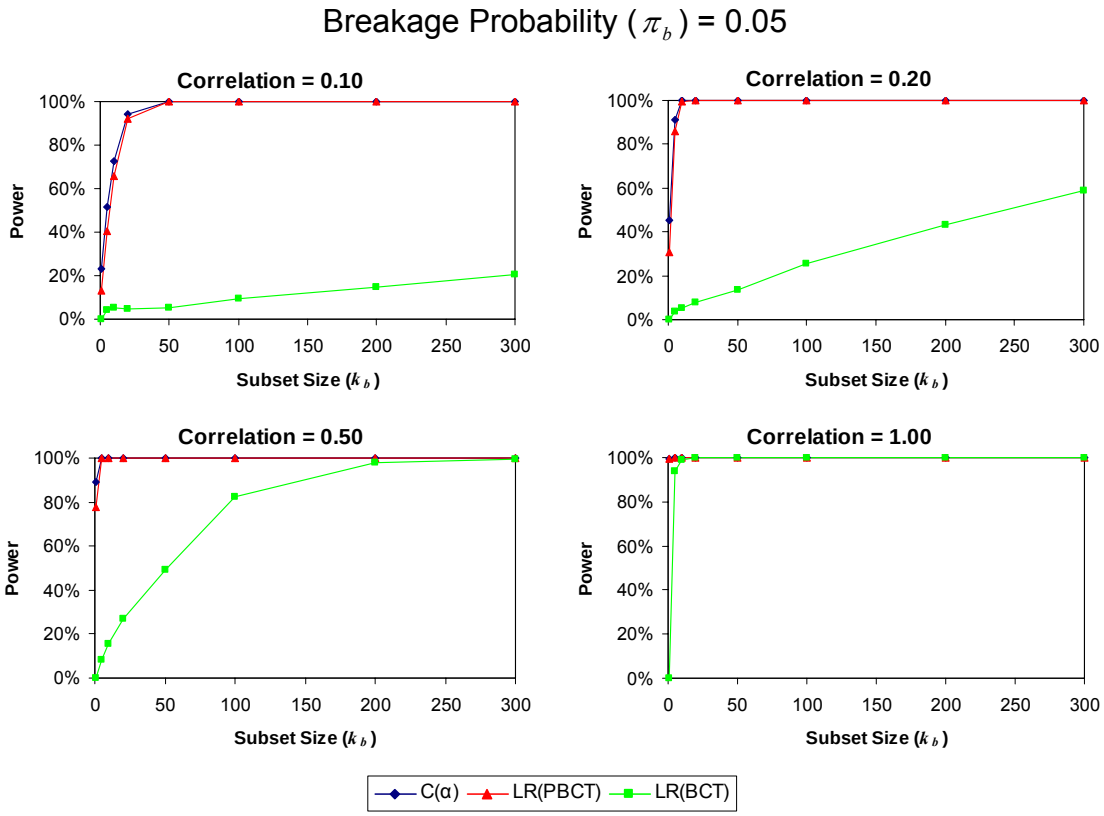


Figure 5. Simulated Power Curves of Three Tests for Correlation When the Probability of Breakage Is 0.05. Simulated results are based on 1,000 Monte Carlo samples with 100 metaphases. Correlations range from 0.10 to 1.00. Power was computed as the percentage of simulations for which the null hypothesis of no correlation was rejected.

breakage probabilities likely results in more observed breaks and, therefore, yield more information for detecting correlation. The rate of detection of correlation increases with the number of sites included in a subset in both Figure 4 and Figure 5.

Perhaps a more important inference to be made from both Figure 4 and Figure 5 is that the LR(BCT) test, as a practical matter, has insufficient power to detect correlations less than 0.50. The estimated power never gets above 20% for detecting a correlation of 0.10 with the LR(BCT) test. At least 50 sites with correlation equal to 0.50 must be present to detect the correlation with estimated power of about 60%. The LR(BCT) test detects correlation of one existent in five sites having the same breakage probability with over 90% power (Figures 4 and 5, Correlation = 1.00), an unrealistic

situation for fragile site data. Based on these results, we eliminate the LR(BCT) test from consideration as a useful test for detecting correlation between identical sites on homologous chromosomes. Furthermore, because of the poor performance of the LR(BCT) test, we decided that it was unnecessary to derive and compute Neyman's C(α) test for BCT data.

Partitioned break count totals are required to detect correlation with sufficient power. The C(α) test performs at least as well or better than the LR(PBCT) test in terms of power for all subset sizes and for both breakage probabilities. In Figure 4 and Figure 5, the estimated power for the C(α) test is always greater than equal to estimated power of the LR(PBCT) test. Also, the C(α) test generally out-performs the LR(PBCT) test in comparisons of estimated levels. Consequently, we conclude that the C(α) test is the best test of those compared for detecting all levels of correlation between identical sites on homologous chromosomes for subset sizes and breakage probabilities represented in our simulation study.

In application of this methodology to breakage data, each site is individually analyzed for the presence of correlation. Without more knowledge about the true breakage probabilities than is currently available, validation of the assumption that sites with equivalent observed break totals have the same breakage probability is problematic. Results of Olmsted (1999), Böhm et al. (1995), McAllister and Greenbaum (1997), and Denison et al. (2003) indicate that data from each individual should be analyzed separately because of substantial variation in per-site breakage from individual to individual. Thus, the pooling of chromosomal breakage data across individuals in an inappropriate method of achieving sufficient sample sizes to detect correlation. We recommend that the C(α) test for correlation between identical sites on homologous chromosomes should only be applied to individual sites with appropriate adjustments to the experiment-wise alpha level made to control overall type I error. We now present simulation studies of the C(α) test using PBCT for a single site.

In the simulation studies that follow we present the results as they relate to the breakage probability and number of metaphases observed. Greenbaum et al. (1997)

pointed out that the critical parameter involved in consideration of adequate sample size is not the number of metaphases observed, but the number of breaks observed. The expected number of breaks observed at a given site is a function of the number of metaphases observed and the site's probability of breakage. Thus, in our simulation study where the parameters are fixed, the expected number of breaks observed can be easily computed from the number of metaphases (c) and the probability of breakage (π) as $2c\pi$. We have presented the results based on the probability of breakage and the number of metaphases in order to see the effects of changing c and π individually and to maintain a consistent scale for comparison.

3.4.3 Type I Error Rates (Alpha Level) for Detecting Correlation at a Single Site Using the C(α) Test

We now present simulation studies of the alpha level for the C(α) test for correlation at a single site. Each simulated result in this section is based on 10,000 Monte Carlo samples. (We were able to increase the number of Monte Carlo samples over the number used in sections 3.4.1 and 3.4.2 since we were free of the computational constraints accompanying calculation of the LR(BCT) test.) All simulations were performed with a nominal alpha level of 5%. The number of metaphases varied from 10 to 200 and the breakage probabilities varied from 0.01 to 0.11. Denison et al. (2003) used FSM to analyze breakage data from 20 humans with numbers of metaphases ranging from 60 to 123. Thus, the number of metaphases used in our simulation covers a plausible range of metaphases that might be characteristic of actual experimental situations.

Figure 6 illustrates that the C(α) test for a single site apparently does not always achieve the nominal alpha level (5%). For numbers of metaphases greater than 100, the estimated alpha level achieved by the C(α) test is generally less than 5%. If the fragile-site breakage probability is equal to 0.05, the C(α) test comes the closest to achieving the correct level for large numbers of metaphases. For the lowest simulated breakage probability, 0.01, the C(α) test for a single site consistently achieves a level much lower

than the target 5%. The estimated alpha level generally gets closer to the nominal level as both the breakage probability and the number of metaphases increase. (The failure of the estimated level to achieve the nominal level is likely due to the discrete nature of the test statistic at these parameter values.) Thus, the performance in terms of type I error of the C(α) test for correlation at a single site appears to depend on the breakage probability and number of metaphases analyzed (hence, the expected number of breaks), but there exists no identifiable pattern for this dependence.

3.4.4 Power Curves for Detecting Correlation at a Single Site Using the C(α) *Test*

Using the same simulation parameters of Section 3.4.3 (10,000 Monte Carlo samples, 10 to 200 metaphases, alpha level of 5%), with the exception that correlation is now non-zero, we present simulation studies of the power associated with the C(α) test for correlation at a single fragile site. Correlation values range from 0.1 to 1.0. Simulated breakage probabilities ranged from 0.01 to 0.11, but only results for a breakage

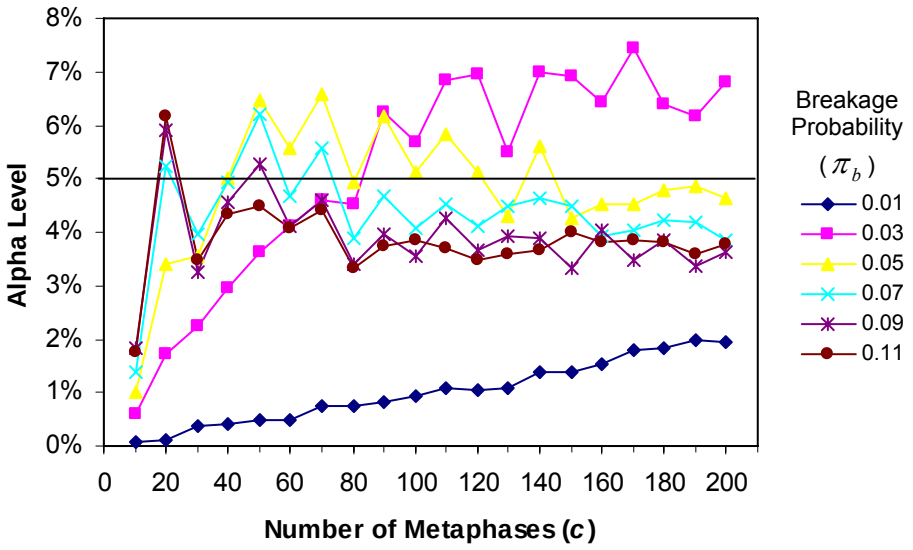


Figure 6. Simulated Alpha Level of the C(α) Test for Correlation at a Single Site. Simulated results are based on 10,000 Monte Carlo samples where correlation equals zero. The alpha level was computed as the percentage of simulations for which the null hypothesis of no correlation was rejected. Confidence interval bars are not included since the interval width is less than the height of the points on the graph.

probability of 0.05 (typical of the other breakage probabilities in terms of overall trend) are included in Figure 7. Complete simulation results are presented in Appendix B.

Figure 7 indicates that power increases with the correlation and with the number of metaphases analyzed. This might be expected since the amount of breakage data (i.e., the expected number of breaks) increases with the number of analyzed metaphases for a fixed breakage probability. If 60 metaphases were analyzed (corresponding to six expected breaks), which is the fewest metaphases examined by Denison et al. (2003), then the correlation had be about 0.6 or greater in order for us to detect it with at least 80% power. If 120 metaphases were analyzed (12 expected breaks), a correlation of 0.4 or greater was detected with about 85% or more power. For correlations equal to 0.1 and 0.2, the estimated power never reached 80%. These results indicate that the true correlation must be relatively high in order for us to detect it with the range of

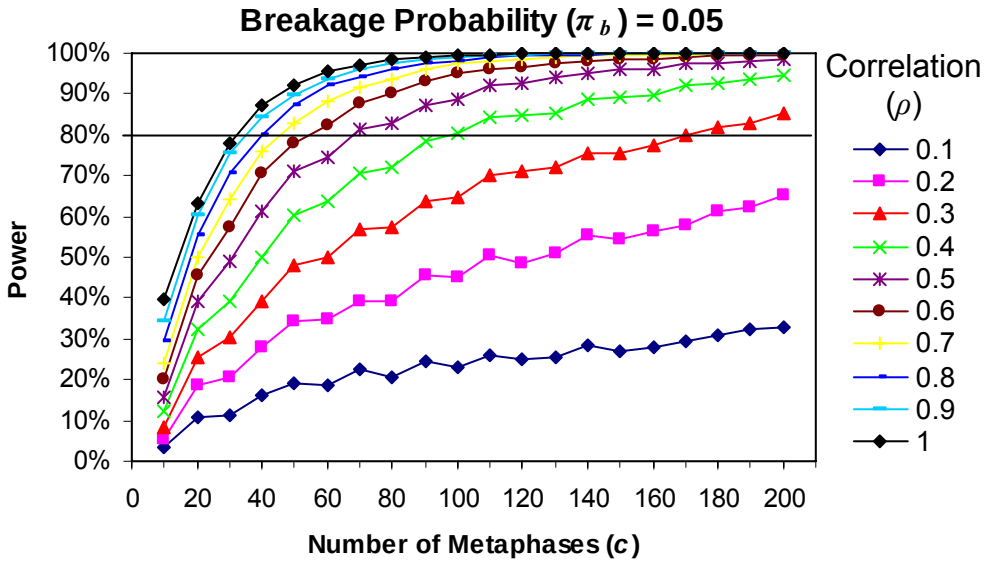


Figure 7. Simulated Power Curves for Detecting Various Levels of Correlation at a Single Site When the Probability of Breakage is 0.05. Simulated results are based on 10,000 Monte Carlo samples, where correlations range from 0.1 to 1.0 and the number of metaphases range from 10 to 200. Power was computed as the percentage of simulations for which the null hypothesis of no correlation was rejected. Confidence interval bars are not included since the interval width is less than the height of the points on the graph. See Appendix B for similar power curves with other breakage probabilities between 0.01 and 0.11.

metaphases commonly used. Also, the power curves change significantly with the breakage probability. As the breakage probability increased from 0.01 to 0.11, the power to detect correlation also increased (Appendix B).

3. 5 Application of the C(α) Test to Breakage Data

Having concluded that the C(α) test is preferable to the LR(BCT) and LR(PBCT) tests for detecting correlation between homologous chromosomes, we now apply the C(α) test to human chromosomal breakage data. Breakage data from 20 individuals (humans) were collected and reported in Table 2 of Denison et al. (2003). Denison et al. (2003) do not report the occurrence of single-breaks and double-breaks since the FSM algorithm only requires the break count totals. However, the original data for 14 of the 20 individuals were obtained (Denison, personal communication, Feb. 25, 2005), enabling determination of the partitioned break count totals for each reported site. Table 2 (below) presents a summary of the sites for which double-breaks were observed and the associated C(α) test for per-site correlation.

In total, 15 of 58 sites (25.9%) with double-breaks were found to have statistically significant non-zero correlation (Table 2). However, closer inspection of this table reveals that no sites with more than six total breaks displayed significant positive correlation and that the five sites with the highest estimated correlation (0.66) had one single-break and one double-break. The lowest estimated correlation that was also statistically significant was 0.317 at the 6p21 site of Individual 2. Only one site with multiple double-breaks had statistically-significant estimated correlation. Furthermore, all but two sites for which the null hypothesis of no correlation was rejected were also declared non-fragile by FSM3, the more conservative of the two FSM algorithms (i.e., FSM3 declares fewer sites to be fragile than does FSM). Four sites (each with 3 total breaks) with significant correlation were declared non-fragile by both FSM and FSM3 (Table 2).

Table 2. Summary of $C(\alpha)$ Tests for Correlation in Human Chromosomal Breakage Data

$X^2_{C(\alpha)}$	Indiv	Site	c	m_1	m_2	n	$\hat{\pi}$	$\hat{\rho}$	FS.FSM	FS.FSM3
46.88	12	4p11-q11	107	1	1	3	0.014	0.662*	no	no
43.77	10	7q32	100	1	1	3	0.015	0.662*	yes	no
31.32	9	1q24	72	1	1	3	0.021	0.660*	no	no
31.32	9	8q24	72	1	1	3	0.021	0.660*	no	no
25.10	1	1p11-q11	58	1	1	3	0.026	0.658*	no	no
41.30	13	1p21	96	2	2	6	0.031	0.656*	yes	yes
29.74	2	Xq22	123	2	1	4	0.016	0.492*	yes	no
23.99	16	1p21	100	2	1	4	0.020	0.490*	yes	no
22.99	13	5p14	96	2	1	4	0.021	0.489*	yes	no
21.49	14	2p11-q11	90	2	1	4	0.022	0.489*	yes	no
16.99	9	7p11-q11	72	2	1	4	0.028	0.486*	yes	no
16.99	9	9q12	72	2	1	4	0.028	0.486*	yes	no
18.47	2	1p21	123	3	1	5	0.020	0.388*	yes	no
13.67	3	7q31	93	3	1	5	0.027	0.383*	yes	no
12.33	2	6p21	123	4	1	6	0.024	0.317*	yes	yes
8.98	1	3p14	58	7	3	13	0.112	0.394	yes	yes
8.07	1	14q24	58	3	1	5	0.043	0.373	yes	no
9.78	10	1p21	100	4	1	6	0.030	0.313	yes	yes
9.33	13	11p14	96	4	1	6	0.031	0.312	yes	yes
8.22	17	14q23	86	4	1	6	0.035	0.309	yes	no
5.12	1	16q23	58	4	1	6	0.052	0.297	yes	yes
7.26	20	3p14	101	12	3	18	0.089	0.268	yes	yes
6.83	20	1p21	101	5	1	7	0.035	0.260	yes	yes
6.42	13	7q22	96	5	1	7	0.036	0.259	yes	yes
6.42	13	7q31	96	5	1	7	0.036	0.259	yes	yes
4.47	15	2p11-q11	72	5	1	7	0.049	0.249	yes	yes
5.22	12	2q33	107	6	1	8	0.037	0.221	yes	yes
4.79	10	14q23	100	6	1	8	0.040	0.219	yes	yes
3.76	13	3p14	96	15	3	21	0.109	0.198	yes	yes
2.19	1	2q33	58	6	1	8	0.069	0.194	yes	yes
2.04	1	7q32	58	10	2	14	0.121	0.188	yes	yes
3.25	13	2q33	96	7	1	9	0.047	0.184	yes	yes
2.76	17	2q33	86	7	1	9	0.052	0.179	yes	yes
3.11	12	3p14	107	28	7	42	0.196	0.171	yes	yes
2.09	9	2q33	72	7	1	9	0.063	0.170	yes	yes
2.09	15	16q23	72	7	1	9	0.063	0.170	yes	yes
1.95	19	3p14	79	21	5	31	0.196	0.157	yes	yes
2.17	3	3p14	93	23	5	33	0.177	0.153	yes	yes
1.80	4	2q33	82	8	1	10	0.061	0.148	yes	yes
1.85	17	3p14	86	17	3	23	0.134	0.147	yes	yes
1.83	20	16q23	101	9	1	11	0.054	0.135	yes	yes
1.80	16	Xp22	100	9	1	11	0.055	0.134	yes	yes
1.58	3	1p21	93	9	1	11	0.059	0.130	yes	yes
1.24	14	3p14	90	15	2	19	0.106	0.117	yes	yes
0.94	15	7q32	72	9	1	11	0.076	0.114	yes	yes
1.19	13	Xp22	96	10	1	12	0.063	0.111	yes	yes
0.83	19	16q23	79	15	2	19	0.120	0.103	yes	yes
0.83	4	14q23	82	10	1	12	0.073	0.101	yes	yes
0.38	17	16q23	86	12	1	14	0.081	0.067	yes	yes
0.32	4	7q22	82	12	1	14	0.085	0.063	yes	yes
0.27	14	16q23	90	13	1	15	0.083	0.055	yes	yes
0.18	15	3p14	72	12	1	14	0.097	0.051	yes	yes
0.00	9	3p14	72	23	3	29	0.201	0.007	yes	yes
0.00	12	16q23	107	19	1	21	0.098	-0.003	yes	yes
0.01	4	3p14	82	29	4	37	0.226	-0.012	yes	yes
0.05	10	3p14	100	20	1	22	0.110	-0.021	yes	yes
0.20	2	3p14	123	25	1	27	0.110	-0.040	yes	yes
0.17	16	3p14	100	22	1	24	0.120	-0.042	yes	yes

NOTE: A Bonferroni-type adjustment for 58 tests was used to control the overall experimental type I error rate. Individual correlations marked with a * are significantly different from zero at the α = 0.05/58 = 0.000862 level. The column titled "Indiv" refers to the individual given in Denison et al. (2003), Table 2. The columns titled "FS.FSM" and "FS.FSM3" refer to whether or not each site was declared fragile by FSM and FSM3, respectively.

These results from Table 2 suggest that the declaration of significant correlation at a site is closely tied to the total number of breaks observed. Only sites with relatively few total breaks displayed significant correlation. Because of the sparse nature of the data, it is difficult to believe that the correlation at any site could truly be as high as 0.66. Furthermore, the theory upon which the C(α) test is based assumes that the sample size is large enough for the true distribution of the C(α) test statistic to be approximated by a χ^2 distribution. With only three to six observed breaks for sites with significant correlation, the validity of this large-sample approximation is doubtful. The sparseness in the data, together with the fact that most sites with correlation were declared non-fragile by FSM3 and fragile by FSM, indicates that these sites likely represent moderately fragile sites exhibiting unusually low breakage numbers. Based on a Mendelian model for fragile site inheritance, however, it is likely that correlation does exist to some degree if an individual is homozygous fragile. The Bonferroni adjustment is very conservative when a large number of tests are performed, so the rejection of the null hypothesis of no correlation indicates that at least some significant correlation is present.

3.6 Chapter Summary

We have concluded that the C(α) test using PBCT is preferable to the LR(BCT) and LR(PBCT) test for detecting correlation between identical sites on homologous chromosomes (Figures 2 through 5). The estimated level of the LR(BCT) test was consistently below the nominal alpha level for nearly all subset sizes and for both breakage probabilities presented (Figures 2 and 3). Furthermore, the LR(BCT) test has insufficient power to detect correlations less than 0.50 (Figures 4 and 5); for practical applications, the PBCT must be reported in order to detect correlation. We have presented power curves useful for determining the number of metaphases (or expected number of breaks) needed to detect specific levels of correlation at various breakage probabilities (Figure 7 and Appendix B). We found that the C(α) test consistently had higher power for detecting correlation than the LR(BCT) test. Finally, we have shown

using actual breakage data that the detection of correlation is closely tied to the total number of breaks observed (Table 2). Without a substantial increase in the amount of data collected for each individual, i.e., without increasing the total number of breaks observed for each individual at each site, even the C(α) test has insufficient power to detect correlation less than 0.5 at a single site.

CHAPTER IV

FSM AND FSM3 SIMULATION STUDIES

An important goal of this research is to establish whether or not one should model correlation between fragile sites on homologous chromosomes when attempting to identify fragile sites. Both FSM and FSM3, which are the only two computer software packages available (to our knowledge) for fragile site identification, assume total independence of all sites observed. To determine the effects of correlation on FSM and FSM3, we simulate data with correlation present at varying degrees and study the effects of that correlation on the ability of FSM and FSM3 to distinguish fragile sites from non-fragile sites. We consider the case where correlation exists only for fragile bands. As described in Section 1.7.1, our assumed Mendelian model for fragile site inheritance suggests that correlation would exist for an individual who is homozygous at some particular fragile-site locus. There would be no reason to believe that non-fragile sites would display any degree of correlation since the breaks, by definition, in non-fragile sites are assumed to be random. Thus, we only simulate the case where correlation exists at fragile sites.

The breakage probabilities used in our simulations are based on parameters used by Olmsted (1999) in her performance tests of the FSM and FSM3 algorithms. Olmsted (1999) presented the results for ten simulated datasets, each with 100 metaphases, 44 non-fragile bands having a breakage probability of 0.005, and six fragile bands having breakage probabilities equal to 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. Note that Olmsted presented the simulation parameters in terms of expected numbers of breaks (λ) instead of breakage probabilities. We computed the breakage probabilities, π, using λ as $\pi = \lambda / 2c$, where c is the number of metaphases (100 in our simulations).

Olmsted (1999) also presented results demonstrating the effects of an excess of zero-breakage sites (i.e., sites with breakage probability equal to zero) on the FSM and FSM3 algorithms. Although performing a comparison of the FSM and FSM3 algorithms

is not the main objective in this research, we present an auxiliary simulation study of the performance of FSM and FSM3 in the presence and absence of zero-breakage sites.

In our simulation study, band resolutions of 300 and 400 were assumed, consistent with the band resolutions for G-banding chromosomes from deer mice and humans, respectively (Olmsted, 1999). Six, 12 or 18 bands were assumed fragile. Breakage probabilities for the case of six fragile sites were those used by Olmsted (1999); for the cases of 12 and 18 fragile sites, we assumed an equal number of fragile bands for each breakage frequency. For example, in the case of 18 fragile sites, there would be three bands with breakage probability equal to 0.022, three bands with probability of 0.024, and so forth. For 20 individuals, Denison et al. (2003) found a range of 7 to 20 fragile sites in each individual. For simplicity, we used 6, 12, and 18 fragile sites. In addition, we looked at fragile-site breakage probabilities equal to and half those of Olmsted (1999) and included a scenario in which 20% of all sites are zero-breakage sites. Each simulated result is based on 1,000 Monte Carlo samples.

We present for discussion only results needed to illustrate effects of correlation on FSM and FSM3. Consequently, not all of the cases we simulated are given in the following sections. Complete simulation results are presented in Appendix C.

4.1 FSM Simulation

The Fortran-compiled FSM697 executable version of FSM described in Greenbaum and Dahm (1995) was used in our simulation study of the effects of correlation on FSM. One thousand Monte Carlo samples of breakage data were generated using R (R Core Development Team (2003)) version 1.8.1. A Perl (Perl Programming Language (freeware) (2004)) script was used to parse the FSM output and determine the percentage of false positives and false negatives. The percentage of false positives (% FP) is defined as

$$\%\,\mathrm{FP} = \frac{\sum_{i=1}^{1,000} \frac{(Number\ of\ Non\text{-}Fragile\ Sites\ Identified\ as\ Fragile)_i}{(Total\ Number\ of\ Non\text{-}Fragile\ Sites)}}{1,000} \times 100\,. \quad (4.1)$$

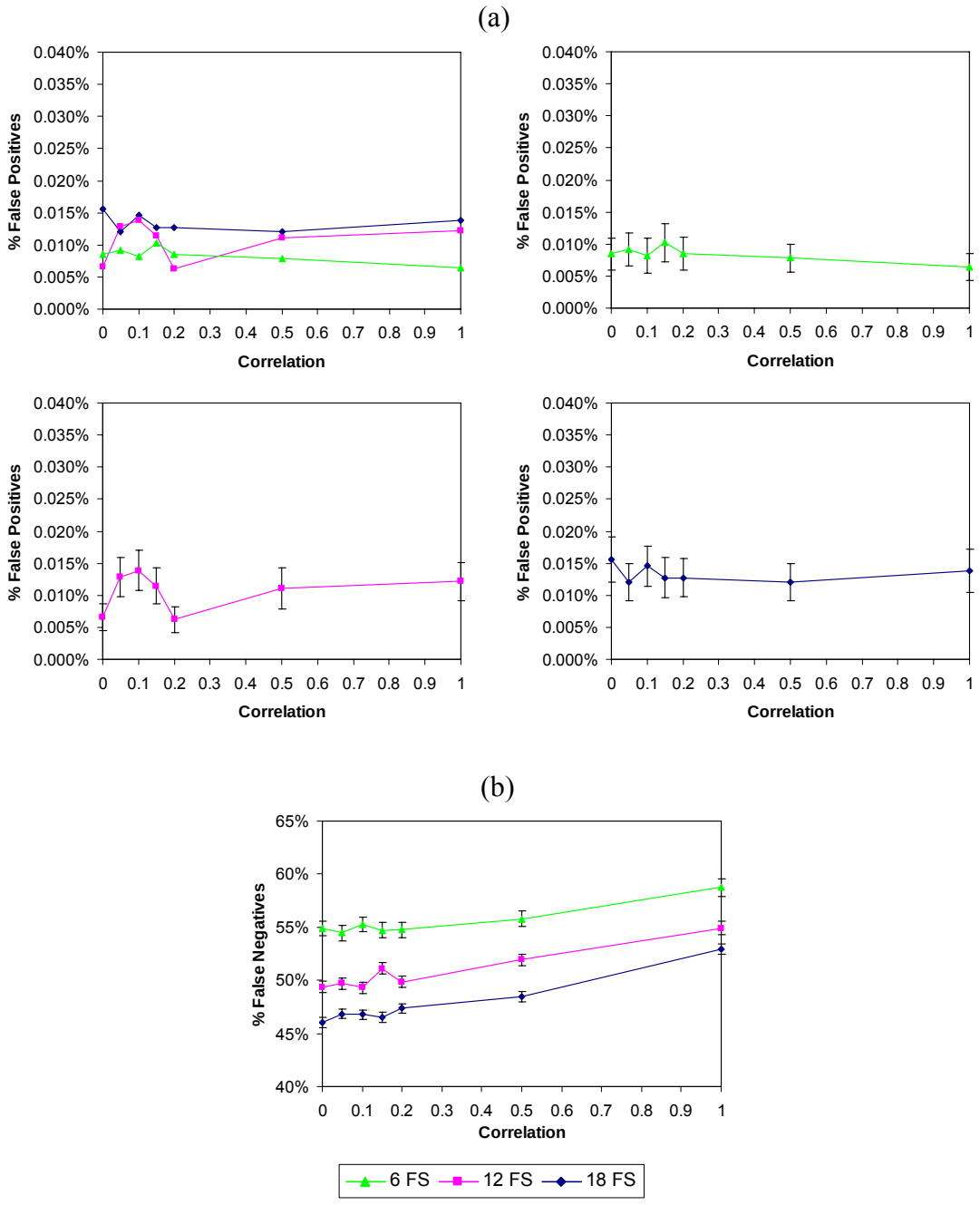


Figure 8. FSM (a) False-Positive and (b) False-Negative Rates for 6, 12, and 18 Fragile Sites. These graphs are based on 300 total bands and breakage probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. There are an equal number of fragile sites for each breakage frequency, e.g. for 18 fragile sites, there are three bands with breakage probability equal to 0.022, three bands at 0.024, and so forth. The upper left-hand graph in (a) contains curves for 6, 12 and 18 fragile sites; these same curves are then plotted separately in the other graphs of (a) to make it possible to see the individual 83% confidence intervals. Results are based on 1,000 Monte Carlo simulations.

The percentage of false negatives (% FN) is defined as

$$\%\,\mathrm{FN} = \frac{\sum_{i=1}^{1,000} \frac{(Number\ of\ Fragile\ Sites\ Identified\ as\ Non\text{-}Fragile)_i}{(Total\ Number\ of\ Fragile\ Sites)}}{1{,}000} \times 100\,. \quad (4.2)$$

Representative results for the FSM simulation with the fragile-site breakage probabilities derived from those in Olmsted (1999) are presented in figures 8 through 10. The reader is referred to Appendix C to view the complete simulation results.

Correlation between maternal and paternal homologs does not appear to have significant effects on either false-positive rate or false-negative rate of FSM. As the level of correlation present increases from 0.0 to 1.0, there is not a statistically significant change in the false-positive rate (Figure 8(a)), as demonstrated by the overlapping 83% confidence intervals moving from left to right in each graph. (For example, for 18 fragile sites and correlations of zero and 0.5 (Figure 8(a), upper right-hand graph), the 83% confidence intervals for the percentage of false positives are (0.0120%, 0.0192%) and (0.0092%, 0.0150%), respectively. These two confidence intervals overlap, indicating that there is not a significant difference between the two corresponding false-positive rates. Similar comparisons can be made for all confidence intervals in the graph.) This is not surprising since correlation is only introduced for fragile sites; the false-positive rate relates to the number of non-fragile sites declared fragile. Recall that FSM determines the maximal set of non-fragile sites and declares the others fragile, so it makes sense that the false-positive rate would not be grossly affected by correlation in fragile sites. The false-negative rate is only slightly affected by the introduction of correlation into the model. We observe an increase of only about 3% to 7% in the false-negative rate when correlation changes from 0.0 to 1.0 (Figure 8(b)). This 7% change represents the worse case scenario. For correlation between 0.1 and 0.5, the change in false-negative rate is less than 3%. In reality, our data (Table 2) suggest that correlation is likely to be less than 0.5. Only six out of 58 sites with double breaks had estimated correlation over 0.5; all of these six sites had fewer than seven total breaks, indicating that these six estimates are probably based on insufficient information. In the absence of correlation, the FSM algorithm performs with an overall false-negative rate of somewhere between 45% and

56%, depending on the number of fragile sites present (Figure 8(b)). Our study indicates that the introduction of correlation does little to raise the already high false-negative rate. Thus, we contend that correlation in fragile sites is unlikely to have a significant adverse effect on the performance of the FSM algorithm as applied to breakage data for mice and humans.

Figure 9 demonstrates the effect of decreasing the breakage probability of the fragile sites. When the breakage probabilities are halved to 0.011, 0.0132, 0.0165, 0.0198, 0.022, and 0.0275, while maintaining a homogeneous non-fragile breakage probability of 0.005, the false-negative rate for 18 fragile sites rises from just below 50% to between 80% and 90% (Figure 9(b)). For most correlations, the false-positive rate nearly doubles from between 0.01% and 0.02% to between 0.02% and 0.035% (Figure 9(a)). The performance of the FSM algorithm is closely tied to the magnitude of the

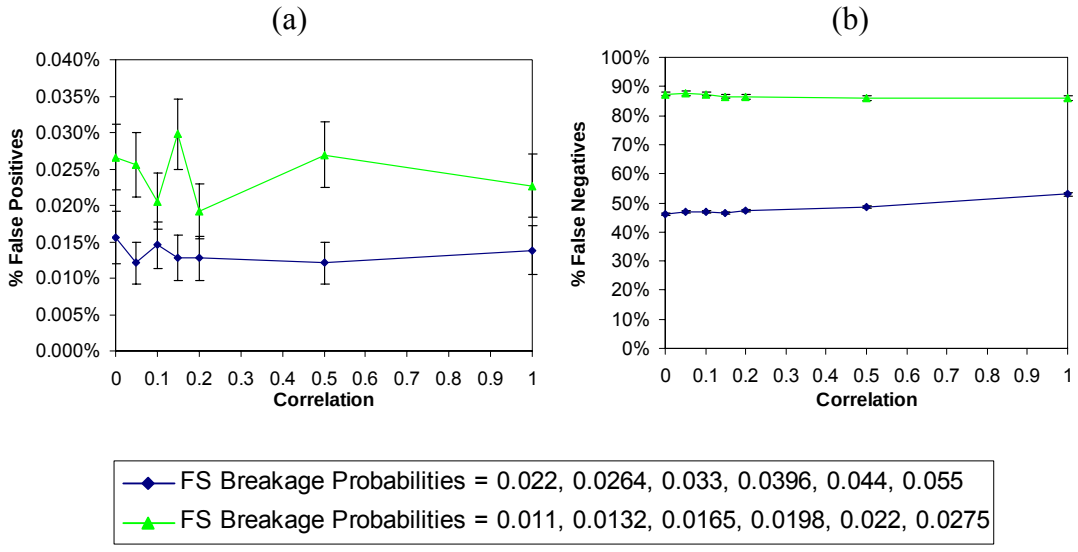


Figure 9. FSM (a) False-Positive and (b) False-Negative Rate Comparison for Different Breakage Probabilities. The green curve represents results where breakage probabilities are exactly half those for the blue curve. The results for 18 fragile sites are displayed with 83% confidence intervals. There are an equal number of fragile sites for each breakage frequency, e.g. for the blue curve, three bands have breakage probability equal to 0.022, three bands have probability of 0.024, and so forth. Results are based on 1,000 Monte Carlo simulations.

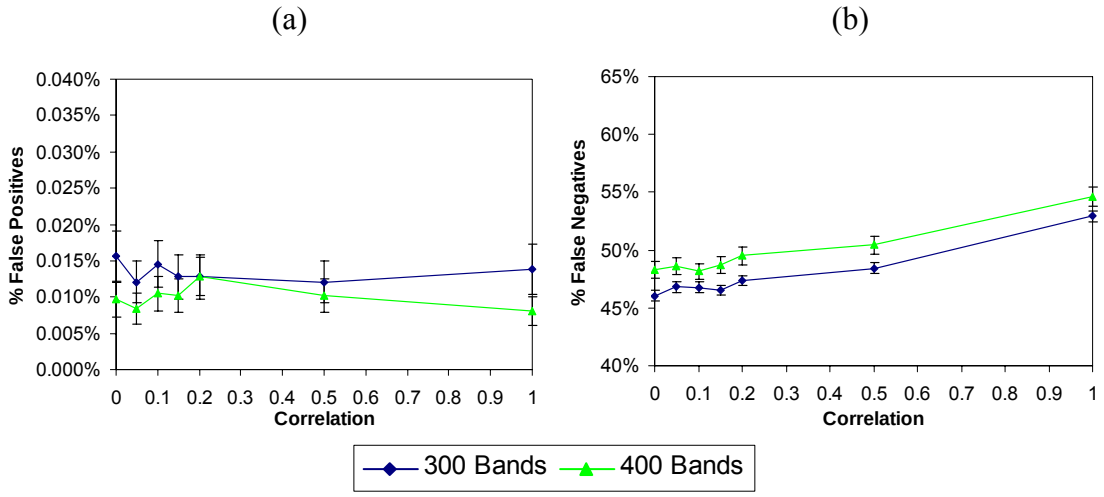


Figure 10. FSM (a) False-Positive and (b) False-Negative Rate Comparison Between 300- and 400-Band Resolutions. Fragile site breakage probabilities are equal to 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. The results for 18 bands are displayed. There are an equal number of fragile sites for each breakage frequency, i.e., three bands have breakage probability equal to 0.022, three bands have 0.024, and so forth. Results are based on 1,000 Monte Carlo simulations.

breakage probabilities of the fragile sites in relation to that of the non-fragile sites. Lower fragile-site breakage probabilities tend to yield break counts closer to those generated by non-fragile sites, making it difficult to distinguish fragile sites from non-fragile sites given such sparse information.

This simulation study supports the claim that the FSM algorithm is conservative (i.e., controls the false-positive rate and has a high false-negative rate) in fragile-site identification (Olmsted (1999)). About half of the 18 fragile sites with breakage probabilities ranging from 0.022 to 0.055 were misclassified by FSM as non-fragile, while only 0.01% to 0.02% of non-fragile sites were misidentified as fragile (Figure 9). In contrast, Section 4.2 results will indicate that in the presence of an excess of zero-breakage sites, the FSM algorithm becomes very liberal (i.e., has a relatively low false-negative rate and a high false-positive rate) in its classification of fragile sites.

Several other observations can be made based on this FSM simulation study. An increase in the total number of sites, both fragile and non-fragile, seems to be accompanied by a small increase in the false-negative rate (Figure 10(b)). For several

correlations in Figure 10(b), the 83% confidence intervals based on 300 bands do not overlap the 83% confidence intervals based on 400 bands. All of the 83% confidence intervals for the false-positive rate, however, do overlap (Figure 10(a)). Furthermore, a decrease in the number of fragile sites present is accompanied by a significant increase in the false-negative rate (Figure 8(b); for any correlation, the 83% confidence intervals do not overlap from curve to curve). There is, however, no significant increase in the FSM false-positive rate with an increase in the number of fragile sites (Figure 8(a); for any correlation, the 83% confidence intervals overlap from curve to curve).

4.2 FSM3 Simulation

The FSM3 simulations were performed using parameter combinations identical to those in the FSM simulations. The FSM3j.exe Fortran-compiled, executable version of FSM3 (Olmsted (1999)) was used in our simulation study. For a detailed description of the FSM3 algorithm, see Olmsted (1999). Again, we simulated the data in R (R Core Development Team (2003)) and used a Perl (Perl Programming Language (freeware) (2004)) script to parse the FSM3 output and determine false-positive and false-negative rates as defined in (4.1) and (4.2), respectively. Complete simulated results, including those for the scenario where zero-breakage sites are present, are given in Appendix C. All results are based on 1,000 Monte Carlo samples.

The FSM3 simulation results are similar to results of the FSM simulation presented in Section 4.1. Figure 11(a) suggests that correlation does not have any significant effect on the false-positive rate for FSM3; for each curve, the 83% confidence intervals overlap with the increase in correlation (i.e., left to right). Similarly, neither the number of fragile sites (Figure 11(a)) nor the total number of sites (Figure 12(a)) seems to have a significant effect on the FSM3 false-positive rate. In contrast to the results for FSM, neither the number of fragile sites (Figure 11(b)) nor the total number of bands (Figure 12(b)) affects the false-negative rate of FSM3, as demonstrated by the overlapping 83% confidence intervals at each level of correlation. An estimated

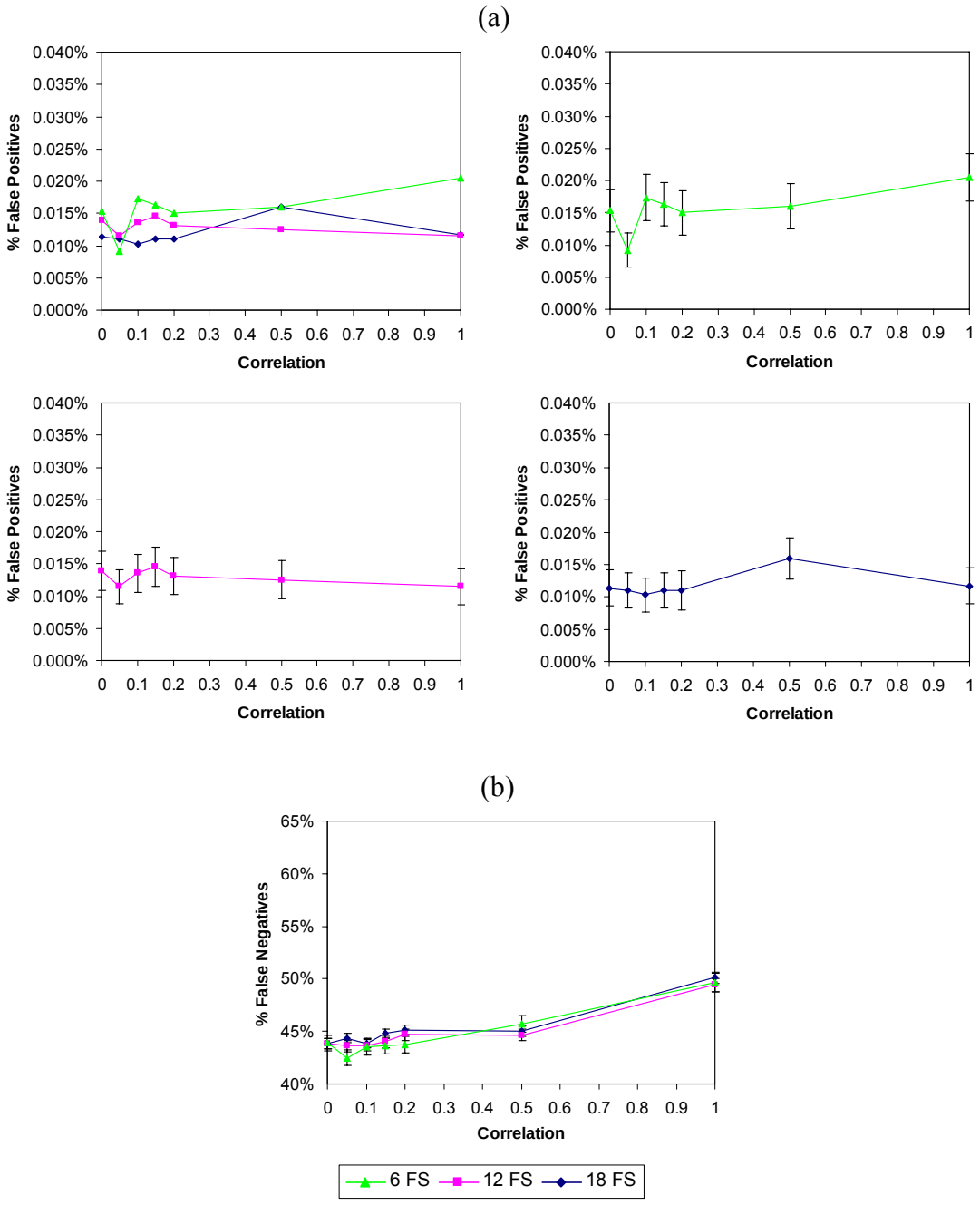


Figure 11. FSM3 (a) False-Positive and (b) False-Negative Rates for 6, 12, and 18 Fragile Sites. These graphs are based on 300 total bands and breakage probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. There are an equal number of fragile sites for each breakage frequency, e.g. for 18 fragile sites, there are three bands with breakage probability equal to 0.022, three bands at 0.024, and so forth. The upper left-hand graph in (a) contains curves for 6, 12 and 18 fragile sites; these same curves are then plotted separately in the other graphs of (a) to make it possible to see the individual 83% confidence intervals. Results are based on 1,000 Monte Carlo simulations.

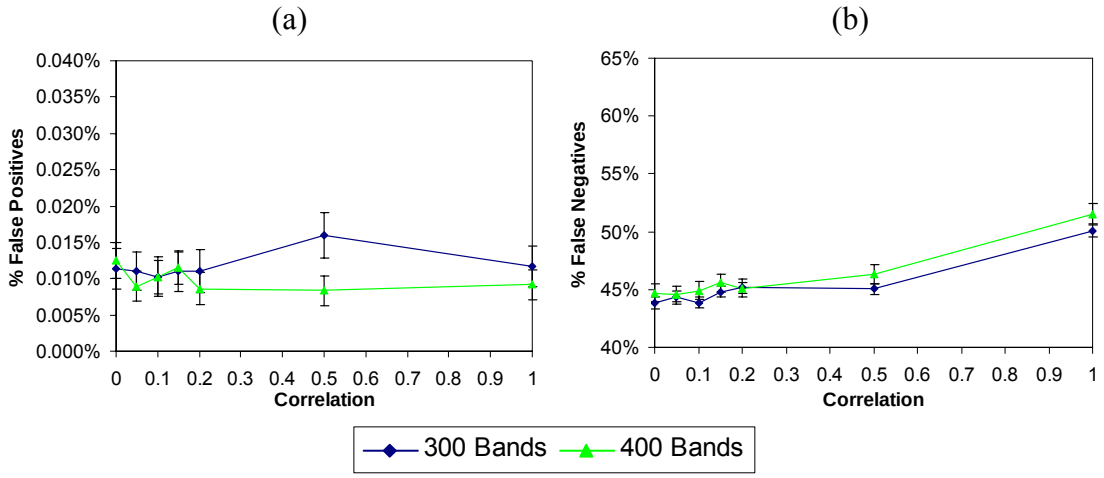


Figure 12. FSM3 (a) False-Positive and (b) False-Negative Rate Comparison Between 300- and 400-Band Resolutions. Fragile site breakage probabilities are equal to 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. The results for 18 bands are displayed. There are an equal number of fragile sites for each breakage frequency, i.e. three bands have breakage probability equal to 0.022, three bands have probability of 0.024, and so forth. Results are based on 1,000 Monte Carlo simulations.

increase of only about 4% to 7% in the percentage of false negatives accompanies an increase in correlation from zero to one; the estimated increase is less than 2% for correlation less than 0.5. No significant change in the false-positive rate occurs as correlation varies from zero to one (Figures 11(b) and 12(b)). The estimated false-negative rate for FSM3 ranges between 40% and 50%, indicating that the FSM3 algorithm is conservative. The estimated false-positive rate for FSM3 is between about 0.005% and 0.025%. (Figure 11). Figure 13 indicates that the estimated percentage of false positives for FSM3 remains constant (Figure 13(a)), while the estimated FSM3 false-negative rate nearly doubles when the fragile-site breakage probabilities are halved (Figure 13(b)). In conclusion, correlation between homologous chromosomes itself does not appear to significantly affect either the false-positive or false-negative rate of the FSM3 algorithm.

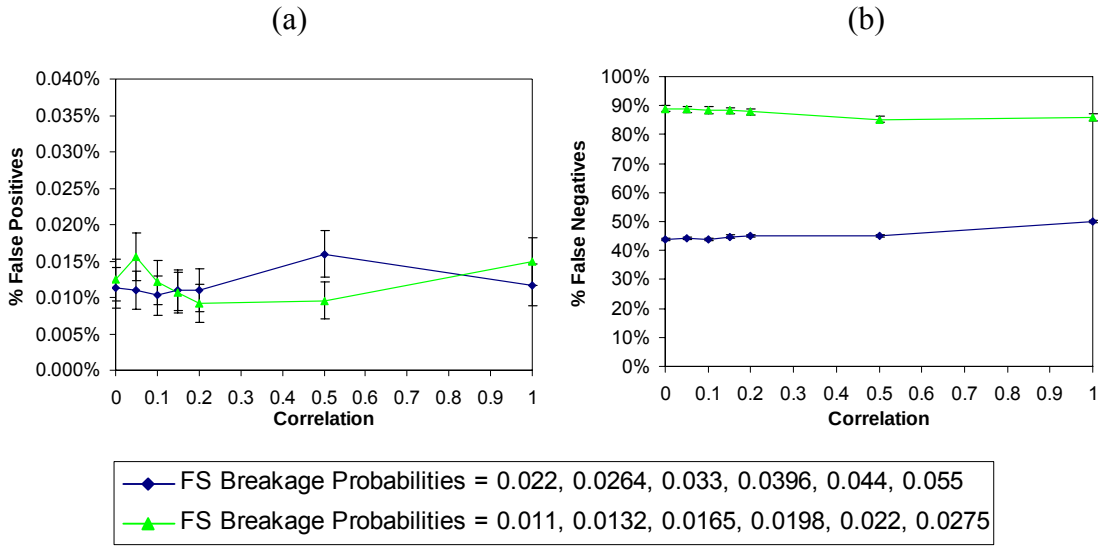


Figure 13. FSM3 (a) False-Positive and (b) False-Negative Rate Comparison for Different Breakage Probabilities. The green curve represents results where breakage probabilities are exactly half those of the blue curve. The results for 18 fragile sites are displayed. There are an equal number of fragile sites for each breakage frequency, e.g. for the blue curve, three bands have breakage probability equal to 0.022, three bands have probability of 0.024, and so forth. Results are based on 1,000 Monte Carlo simulations.

4.3 Comparison of the FSM and FSM3 Algorithms in the Presence of Zero-Breakage Sites

We now consider the case where 20% of all sites have zero probability of breakage (i.e., are zero-breakage sites). Specifically, at a 300-band resolution, 60 sites will never break and at a 400-band resolution, 80 sites will never break. We still consider a range of 6, 12, and 18 fragile sites. Results for 300 bands and 18 fragile sites with probabilities of breakage equal to 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055 are presented in Figure 14 and Figure 15 as representative of results for all parameter combinations. Complete simulation results are presented in Appendix C.

Figure 14 compares error rates of the FSM and FSM3 algorithms when no zero-breakage sites are present. The FSM3 algorithm appears to outperform the FSM algorithm to a very small degree in the absence of zero-breakage sites in terms of false-negative rate. FSM3 declares significantly fewer fragile sites to be non-fragile than does the FSM algorithm (Figure 14(b)) while maintaining a false-positive rate that is not

significantly different from that of FSM (Figure 14(a)). When 20% of all sites have zero probability of breakage, FSM3 significantly outperforms the FSM algorithm in terms of the false-positive rate (Figure 15). Zero-breakage sites cause an increase in the total number of sites being declared fragile by FSM. Thus, for FSM we see a significant drop in the estimated false-negative rate (Compare Figure 14(b) and Figure 15(b)) with nearly a ten-fold increase in the estimated false-positive rate (Compare Figure 14(a) and Figure 15(a)). The FSM algorithm is more liberal in declaring sites as fragile in the presence of zero-breakage sites than in the absence of such sites. The false-positive and false negative rates for the FSM3 algorithm, on the other hand, do not significantly change in the presence or absence of zero-breakage sites (Compare Figure 14 and Figure 15). These results indicate that the FSM3 algorithm is preferable to the FSM algorithm in terms of controlling the false-positive rate both with and without zero-breakage sites present.

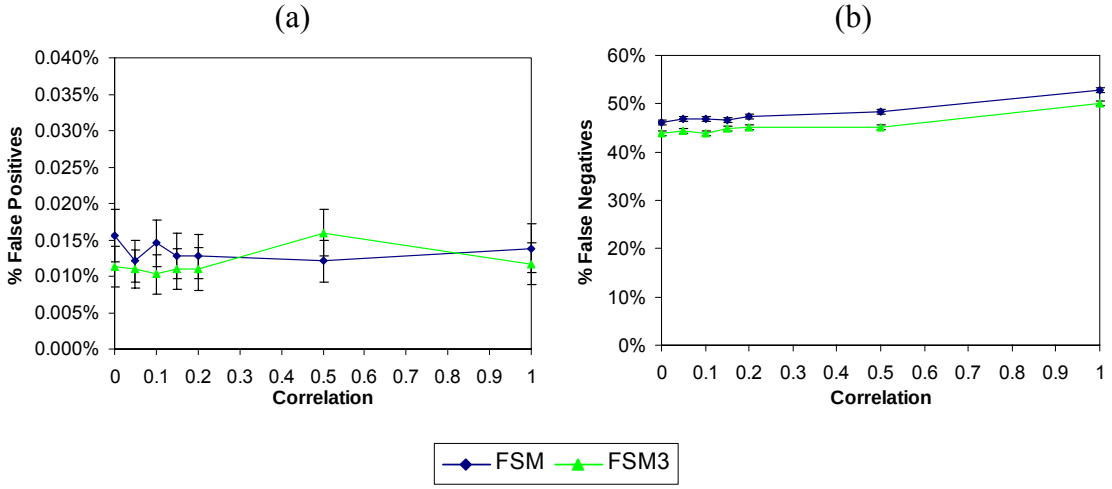


Figure 14. Comparison Between FSM and FSM3 (a) False-Positive and (b) False-Negative Rates With No Zero-Breakage Sites. These results are for 18 fragile sites, 300 total bands, and breakage probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. There are an equal number of fragile sites for each breakage frequency, i.e. three bands have breakage probability equal to 0.022, three bands have probability of 0.024, and so forth. Results are based on 1,000 Monte Carlo simulations.

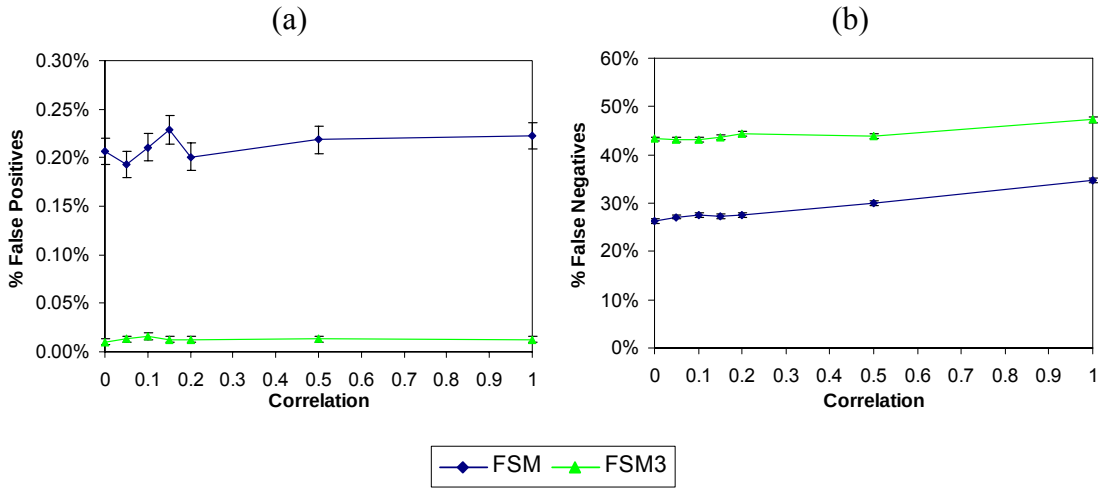


Figure 15. Comparison Between FSM and FSM3 (a) False-Positive and (b) False-Negative Rates With 20% Zero-Breakage Sites Present. These results are based on 18 fragile sites, 300 total bands, and breakage probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. There are an equal number of fragile sites for each breakage frequency, i.e. three bands have breakage probability equal to 0.022, three bands have probability of 0.024, and so forth. Results are based on 1,000 Monte Carlo simulations.

4.4 Chapter Summary

Our simulation study indicates that correlation does not significantly affect either the FSM or FSM3 algorithm. The observed 2% to 3% increase in false-negatives for correlations as high as 0.5 is relatively small compared to the false-negative rates of 40% to 60% for FSM and FSM3. The false-positive rate remained unchanged over the entire range of fragile-site correlations for both algorithms unless zero-breakage sites were present. The FSM3 algorithm performed significantly better than the FSM algorithm in the presence of zero-breakage sites by effectively controlling the false-positive rate. Both the FSM and FSM3 algorithms appear to be conservative, which is consistent with Olmsted (1999). Our results suggest that in the presence of 20% zero-breakage sites the FSM algorithm is more liberal than the FSM3 algorithm in declaring sites fragile.

CHAPTER V

SUMMARY AND CONCLUSIONS

Correlation between homologous chromosomes is characterized by the occurrence of more double-breaks than would be expected in the case of full independence. If an individual is homozygous fragile at a particular locus, the occurrence of a break on one homolog (at that locus) would be correlated with the occurrence of a break on the second homolog. Non-fragile sites, however, would not display this type of correlation, since breaks at non-fragile sites are, by definition, random events.

We first investigated detection of correlation using maximum likelihood-based hypothesis testing procedures. In Chapter II we derived the likelihood functions and maximum likelihood estimators for correlation and single-break and double-break probabilities under two scenarios: when only the total number of breaks per site (BCT) is known, and when partitioned break count totals (PBCT) are known. In Chapter III we derived Neyman's C(α) and likelihood ratio tests to detect correlation when the PBCT are known (LR(PBCT)) and when only the BCT are given (LR(BCT)). For realistic sample sizes based on experimental constraints, a simulation study of the three hypothesis tests for correlation demonstrated that the LR(BCT) test for correlation fails to achieve the correct level and has low power except in cases either of high correlation ($\rho > 0.5$) or large fragile site subset sizes ($k_b > 100$) with equal breakage probabilities and equal correlation. In the extreme case wherein each site is assumed to have a unique breakage probability and correlation, the LR(BCT) test has no power for detecting correlation. Thus, break count totals alone provide insufficient information for detecting correlation, i.e., known partitioned break count totals are required in order to achieve adequate power for detection of correlation via the likelihood ratio test. Simulation studies suggested that the C(α) test is the most powerful test of the three were investigated for detecting correlation at any of the breakage probabilities we considered. We provide power curves for different breakage probabilities and correlations that

indicate sample size requirements for application of the C(α) test for correlation to a single site (Chapter III and Appendix B). Application of the C(α) test to human breakage data (Denison et al. 2003) revealed that the rejection of the null hypothesis of no correlation is strongly influenced by the total number of observed breaks. The null hypothesis was never rejected for sites with more than six total breaks. The sparse nature of chromosomal breakage data appears to impose severe limitations on the ability to detect correlation.

Significant adverse effects of correlation on the FSM and FSM3 fragile site identification algorithms were not evident (Chapter IV). The false-negative error rate of FSM and FSM3 in the absence of zero-breakage sites is estimated to be between 40% and 60% for the fragile-site breakage probabilities of Olmsted (1999). The false-negative error rate exhibits an increase of only about 2% to 3% as correlation changes uniformly in all fragile sites from 0.0 to 0.5. The false-positive rate, estimated to range from about 0.005% to 0.025% for both algorithms, was not significantly affected by changes in correlation. Our empirical studies suggest that both the FSM and FSM3 algorithms are conservative, which is consistent with the results of Olmsted (1999). Restriction of 20% of all sites to have zero probability of breakage resulted in nearly a ten-fold increase in the percentage of false positives for the FSM algorithm. In contrast, the FSM3 algorithm maintains the observed range of 0.005% to 0.025% false positives in the presence of zero-breakage sites. Our results suggest that if one desires to control the rate of false positives, FSM3 should be used instead of FSM for fragile site identification because of the possibility that zero-breakage sites contaminate the data. In contrast, if the goal is to identify all sites that may be fragile, our results indicate that the FSM algorithm would be preferable. Since the FSM and FSM3 algorithms were not significantly affected by any degree of correlation, we believe that the two algorithms do not need to be modified to account for the presence of correlation between fragile sites on homologous chromosomes.

REFERENCES

Aeschbacher, H. U., Vuataz, L., Sotek, J., and Stalder, R. (1977), "Use of the Beta-Binomial Distribution in Dominant-Lethal Testing for 'Weak Mutagenic Activity,'" *Mutation Research*, 44, 369-390.

Altham, P. E. (1978), "Two Generalizations of the Binomial Distribution," *Applied Statistics*, 27, 162-167.

Bahadur, R. R. (1961), "A Representation of the Joint Distribution of Responses to *n* Dichotomous Items," in *Studies in Item Analysis and Prediction*, ed. H. Solomon, Stanford, CA: Stanford University Press, pp. 158-168.

Barbi, G., Steinbach, P., and Vogel, W. (1984), "Nonrandom Distribution of Methotrexate-Induced Aberrations on Human Chromosomes. Detection of Further Folic Acid-Sensitive Fragile Sites," *Human Genetics*, 68, 290-294.

Böhm, U., Dahm, P. F., McAllister, B. F., and Greenbaum, I. F. (1995), "Identifying Chromosomal Fragile Sites From Individuals: A Multinomial Statistical Model," *Human Genetics*, 95, 249-256.

Brooks, S. P., Morgan, B. T., Ridout, M. S., and Pack, S. E. (1997), "Finite Mixture Models for Proportions," *Biometrics*, 53, 1097-1115.

Casella, G., and Berger, R. L. (2002), *Statistical Inference* (2nd ed.), Pacific Grove, CA: Duxbury Press.

Craig-Holmes, A. P., Strong, L. C., Goodacre, A., and Pathak, S. (1987), "Variation in the Expression of Aphidicolin-Induced Fragile Sites in Human Lymphocyte Cultures," *Human Genetics*, 76, 134-137.

Crowder, M. (1985), "Gaussian Estimation for Correlated Binomial Data," *Journal of the Royal Statistical Society B*, 47, 229-237.

Dahm, P. F., and Greenbaum, I. F. (1994), "Reconsideration of the Binomial and *F* Distribution Relationship as a Test of Nonrandomness of Chromosomal Breakage," *Cytogenetics and Cell Genetics*, 66, 214-215.

De Braekeleer, M. and Smith, B. (1988), "Two Methods for Measuring the Non-Randomness of Chromosome Abnormalities," *Annals of Human Genetics*, 52, 63-67.

Dean, B. J., Doak, S. M. A., and Somerville, H. (1975), "The Potential Mutagenicity of Dieldrin (HEOD) in Mammals," *Food and Cosmetics Toxicology*, 13, 317-323.

Denison, S. R., Simper, R. K. and Greenbaum, I. F. (2003), "How Common are Common Fragile Sites in Humans: Interindividual Variation in the Distribution of Aphidicolin-Induced Fragile Sties," *Cytogenetic Genome Research*, 101, 8-16.

Epstein, S. S., Arnold, E., Andrea, J., Bass, W., and Bishop, Y. (1972), "Detection of Chemical Mutagens by the Dominant Lethal Assay in the Mouse," *Toxicology and Applied Pharmacology*, 23, 288-325.

Epstein, S. S., Arnold, E., Steinberg, K., MacKintosh, D., Shafner, H., and Bishop, Y. (1970), "Mutagenic and Antifertility Effects of TEPA and METEPA in Mice," *Toxicology and Applied Pharmacology*, 17, 23-40.

Gart, J. J. (1970), "Some Simple Graphically Oriented Statistical Methods for Discrete Data," in *Random Counts in Models and Structures*, I, ed. G. P. Patil, University Park, PA: Pennsylvania State University Press, pp. 171-191.

George, E. O., and Bowman, D. (1995), "A Full Likelihood Procedure for Analysing Exchangeable Binary Data," *Biometrics*, 51, 512-523.

Glover, T. W., Berger, C., Coyle, J., and Echo, B. (1984), "DNA Polymerase α Inhibition by Aphidicolin Induces Gaps and Breaks at Common Fragile Sites in Human Chromosomes," *Human Genetics*, 67, 136-142.

Greenbaum, I. F., and Dahm, P. F. (1995), "A User's Guide for FSM: An MS-DOS Program for the Statistical Identification of Chromosomal Fragile Sites," Technical Report No. 230, Texas A&M University, Statistics Department.

Greenbaum, I. F., Fulton, J. K., White, E. D., and Dahm P. F. (1997), "Minimum Sample Sizes for Identifying Chromosomal Fragile Sites From Individuals: Monte Carlo Estimation," *Human Genetics*, 101, 109-112.

Haseman, J. K., and Kupper, L. L. (1979), "Analysis of Dichotomous Response Data From Certain Toxicological Experiments," *Biometrics*, 35, 281-293.

Haseman, J. K., and Soares, E. R. (1976), "The Distribution of Fetal Death in Control Mice and Its Implications on Statistical Tests for Dominant Lethal Effects," *Mutation Research*, 41, 277-288.

Hecht, F., and Glover, T. W. (1984), "Cancer Chromosome Breakpoints and Common Fragile Sites Induced by Aphidicolin," *Cancer Genetics and Cytogenetics*, 13, 185-188.

Hecht, F., and Sutherland, G. R. (1984), "Fragile Sites and Cancer Breakpoints," *Cancer Genetics and Cytogenetics*, 12, 179-181.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1997), *Discrete Multivariate Distributions*, New York: John Wiley & Sons, Inc.

Johnson, N. L., Kotz, S., and Kemp, A. W. (1993), *Univariate Discrete Distributions* (2^{nd} ed.), New York: John Wiley & Sons, Inc.

Jordan, D. K., Burns, T. L., Divelbliss, J. E., Woolson, Robert F., and Patil, S. R. (1990), "Variability in Expression of Common Fragile Sites: In Search of a New Criterion," *Human Genetics*, 85, 462-466.

Krüger, J. (1970), "Statistical Methods in Mutation Research," in *Chemical Mutagenesis in Mammals and Man*, eds. F. Vogel and G. Röhrborn, Heidelberg, Germany: Springer-Verlag, pp. 460-502.

Kupper, L. L., and Haseman, J. K. (1978), "The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments," *Biometrics*, 34, 69-76.

Lachin, J. M. (2000), *Biostatistical Methods*, New York: John Wiley & Sons, Inc.

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1991), "Generalized Estimating Equations for Correlated Binary Data: Using the Odds Ratio as a Measure of Correlation," *Biometrika* 78, 153-160.

Makuch, R. W., Stephens, M. A., and Escobar, M. (1989), "Generalised Binomial Models to Examine the Historical Control Assumption in Active Control Equivalence Studies," *The Statistician*, 38, 61-70.

Mariani, Tullio (1989), "Fragile Sites and Statistics," *Human Genetics*, 81, 319-322.

McAllister, B. F., and Greenbaum, I. F. (1997), "How Common are Common Fragile Sites: Variation of Aphidicolin-Induced Chromosomal Fragile Sites in a Population of the Deer Mouse (*Peromyscus maniculatus*)," *Human Genetics*, 100, 182-188.

McCaughran, D. A., and Arnold, D. W. (1976), "Statistical Models for Numbers of Implantation Sites and Embryonic Deaths in Mice," *Toxicology and Applied Pharmacology*, 38, 325-333.

Nagesh Rao, P., Heerema, N. A., and Palmer, C. G. (1988), "Fragile Sites Induced by FUdR, Caffeine, and Aphidicolin – Their Frequency, Distribution, and Analysis," *Human Genetics*, 78, 21-26.

Neyman, J. (1959), "Optimal Asymptotic Tests of Composite Statistical Hypotheses," in *Probability and Statistics*, ed. U. Grenander, Stockholm, Sweden: Almqvist and Wiksell, pp. 213-234.

Olmsted, A. (1999), "Algorithms Using Chi-Squared and Other Goodness of Fit Tests to Identify a High-Expectation Subset of Independent Poisson Random Variables, or a Subset of Multinomial Cells Having Relatively High Probabilities, With Applications in Chromosomal Fragile Site Identification," unpublished Ph.D. dissertation, Texas A&M University, Dept. of Statistics.

Pack, S. E. (1986), "Hypothesis Testing for Proportions With Overdispersion," *Biometrics*, 42, 967-972.

Paul, S. R. (1982), "Analysis of Proportions of Affected Foetuses in Teratological Experiments," *Biometrics*, 38, 361-370.

Paul, S. R., Liang, K. Y., and Self, S. G. (1989), "On Testing Departure From the Binomial and Multinomial Assumptions," *Biometrics*, 45, 231-236.

Perl Programming Language (freeware) (2004), available at http://www.perl.org, accessed February 2005.

Popescu, N. C. (2003), "Genetic Alterations in Cancer as a Result of Breakage at Fragile Sites," *Cancer Letters*, 192, 1-17.

Pothoff, R. F., and Whittinghill, M. (1966), "Testing for Homogeneity, I. The Binomial and Multinomial Distributions," *Biometrika*, 53, 167-182.

R Core Development Team (2003), "R: A Language and Environment for Statistical Computing," ISBN 3-900051-00-3, Vienna, Austria: *R Foundation for Statistical Computing*, available at http://www.R-project.org.

Rao, C. R. (1963), "Criteria of Estimation in Large Samples," *Sankhya, A*, 25, 189-206.

Rao, J. K., and Scott, A. J. (1992), "A Simple Method for the Analysis of Clustered Binary Data," *Biometrics*, 48, 577-585.

Rao, K. C., and Robson, D. S. (1974), "A Chi-Square Statistic for Goodness-of-Fit Tests Within the Exponential Family," *Communications in Statistics – Theory and Methods*, 3, 1139-1153.

Rao, P. N., Heerema, N. A., and Palmer, C. G. (1988), "Fragile Sites Induced by FUdR, Caffeine, and Aphidicolin – Their Frequency, Distribution, and Analysis," *Human Genetics*, 78, 21-26.

Röhrborn, G. (1968), "Mutagenicity Tests in Mice. I. The Dominant Lethal Method and the Control Problem," *Humangenetik*, 6, 345-361.

Rosenzweig, S. and Blaustein, F. M. (1970), "Cleft Palate in A/J Mice Resulting From Restraint and Deprivation of Food and Water," *Teratology*, 3, 47-52.

Rosner, B. (1982) "Statistical Methods in Ophthalmology: An Adjustment for the Interclass Correlation Between Eyes," *Biometrics*, 38, 105-114.

Rudolfer, S. M. (1990), "A Markov Chain Model of Extrabinomial Variation," *Biometrika*, 77, 255-264.

Salsburg, D. S. (1973), "Statistical Considerations for Dominant-Lethal Mutagenic Trials," *Environmental Health Perspectives, Experimental Issue No. 6*, 51-58.

Sherman, S. L., and Sutherland, G. R. (1986), "Segregation Analysis of Rare Autosomal Fragile Sites," *Human Genetics*, 72, 123-128.

Tai, J. J., Hou, C.-D., Wang-Wuu, S., Wang, C.-H., Leu, S. Y., and Wuu, K.-D. (1993), "A Method for Testing the Nonrandomness of Chromosomal Breakpoints," *Cytogenetics and Cell Genetics*, 63, 147-150.

Tarone, R. E. (1979), "Testing the Goodness of Fit of the Binomial Distribution," *Biometrika*, 66, 585-590.

Vasarhelyi, K., and Friedman, J. M. (1989), "Analysing Rearrangement Between Breakpoint Distributions by Means of Binomial Confidence Intervals," *Annals of Human Genetics*, 53, 375-380.

Verkerk A. J., Pieretti, M., Sutcliffe, J. S., Fu, Y. H., Kuhl, D. P., et al. (1991), "Identification of a Gene (FMR-1) Containing a CGG Repeat Coincident With a Breakpoint Cluster Region Exhibiting Length Variation in Fragile X Syndrome," *Cell*, 65, 905-914.

Williams, D. A. (1975), "The Analysis of Binary Responses From Toxicological Experiments Involving Reproduction and Teratogenicity," *Biometrics*, 31, 949-952.

Yunis, J. J. (1984), "Fragile Sites and Predisposition to Leukemia and Lymphoma," *Cancer Genetics and Cytogenetics*, 12, 85-88.

Yunis, J. J., and Soreng, A. L. (1984), "Constitutive Fragile Sites and Cancer," *Science*, 226, 1199-1204.

Yunis, J. J., Soreng, A. L., and Bowe, A. B. (1987), "Fragile Sites are Targets of Diverse Mutagens and Carcinogens," *Oncogene*, 1, 59-69.

Zawoiski, E. J. (1975), "Prevention of Trypan Blue-Induce Exencephaly and Otocephaly in Gestating Albino Mice," *Toxicology and Applied Pharmacology*, 31, 191-200.

APPENDIX A

R CODE FOR CALCULATING NON-TRIVIAL MAXIMUM LIKELIHOOD ESTIMATORS

The R code used to find the MLEs for P_1, P_2, and π given in Chapter II is found in this Appendix. The functions that determine the MLEs also return the value of the log-likelihood evaluated at the corresponding MLE values. The functions that return the MLEs are given first, followed by a list of necessary functions which are called during the maximization routines.

PBCT Maximum Likelihood Estimators of P_1, P_2, and π With Positive Counts Only

```
## Usage: posMult.MLE.P1.P2.Pi(M,Band,cutoff)
## M = # of metaphases(c)
## Band = k x 3 Matrix whose first column contains zero-break counts
##        (M0), second column contains single-break counts (M1), and
##        third column contains double-break counts (M2)
## T = Vector of k break counts
## t = Single break count
## P1,P2,P = Values of P1,P2,Pi for calculating likelihood
## cutoff = Precision for ML estimates (1E-9 works fine)
## Returns vector (P1.hat,P2.hat,Pi.hat,log-Likelihood)
##
## The function Exact.MLE.P1.P2.Pi(M,T,Cutoff) requires the functions
## posBinomial.MLE(M,T)
## F.PMultpos<-function(M,T,P,P2)
## logLikelihood.posExact(M,T,P)
## posExact.MLE.above(M,T,cutoff)
## posExact.MLE.below(M,T,cutoff)

posMult.MLE.P1.P2.Pi(M,Band,cutoff)
```

BCT Maximum Likelihood Estimators of P_1, P_2, and π

```
## Usage: Exact.MLE.P1.P2.Pi(M,T,Cutoff)
## M = # of metaphases(c)
## T = Vector of k break counts
## t = Single break count
## P1,P2,P = Values of P1,P2,Pi for calculating likelihood
## Cutoff = Precision for ML estimates (1E-9 works fine)
```

```
## Returns vector (P1.hat,P2.hat,Pi.hat,log-Likelihood)
##
## The function Exact.MLE.P1.P2.Pi(M,T,Cutoff) requires the functions
## Binomial.MLE(M,T)
## Sum.f.of.P.t(M,t,P1,P2)
## Sum.dP2.h.of.P.t(M,t,P1,P2)
## F.P<-function(M,T,P,P2)
## logLikelihood.Exact(M,T,P)
## Exact.MLE.above(M,T,cutoff)
## Exact.MLE.below(M,T,cutoff)

Exact.MLE.P1.P2.Pi(M,T,cutoff)
```

BCT Maximum Likelihood Estimators of P_1, P_2, and π With Positive Counts Only

```
## Usage: posExact.MLE.P1.P2.Pi(M,T,cutoff)
## M = # of metaphases(c)
## T = Vector of k break counts
## t = Single break count
## P1,P2,P = Values of P1,P2,Pi for calculating likelihood
## cutoff = Precision for ML estimates (1E-9 works fine)
## Returns vector (P1.hat,P2.hat,Pi.hat,log-Likelihood)
##
## The function Exact.MLE.P1.P2.Pi(M,T,Cutoff) requires the functions
## posBinomial.MLE(M,T)
## Sum.f.of.P.t(M,t,P1,P2)
## Sum.dP2.h.of.P.t(M,t,P1,P2)
## F.Ppos<-function(M,T,P,P2)
## logLikelihood.posExact(M,T,P)
## posExact.MLE.above(M,T,cutoff)
## posExact.MLE.below(M,T,cutoff)

posExact.MLE.P1.P2.Pi(M,T,cutoff)
```

Necessary Functions

```
######################### Binomial.MLE ###################################
## Returns MLE of Pi & log-Binomial Likelihood
Binomial.MLE<-function(M,T){
   b<-length(T)
   Bin.MLE<-mean(T)/(2*M)
   Bin.Like<-0
   for (i in 1:b){
      Bin.Like<-Bin.Like + log(dbinom(T[i],2*M,Bin.MLE))
   }
c(Bin.MLE,Bin.Like)
}

######################### Sum.f.of.P.t ###################################
## Returns sum f(n)
```

```
Sum.f.of.P.t <- function(M,t,P1,P2){
   #EVEN t
   if(floor(t/2) == ceiling(t/2)){
      Sum <- 0
      for (s in 0:(min(t,2*M-t))){
         #EVEN s
         if(floor(s/2) == ceiling(s/2)){
            A<-s*log(P1) + sum(log(seq(1,M)))
            B<-((t-s)/2)*log(P2)
            C<-(M-(s+t)/2)*log(1-P1-P2)
            if(s == t){
               D<-0
            }else{
               D<-sum(log(seq(1,(t-s)/2)))
            }
            if(M == (s+t)/2){
               E<-0
            }else{
               E<-sum(log(seq(1,M-(s+t)/2)))
            }
            if(s == 0){
               F<-0
            }else{
               F<-sum(log(seq(1,s)))
            }
            log.sum <- A+B+C-D-E-F
            Sum <- Sum + exp(log.sum)
         }
      }
   }
   #ODD t
   if(floor(t/2) != ceiling(t/2)){
      Sum <- 0
      for (s in 1:(min(t,2*M-t))){
         #ODD s
         if(floor(s/2) != ceiling(s/2)){
            A<-s*log(P1)+ sum(log(seq(1,M)))
            B<-((t-s)/2)*log(P2)
            C<-(M-(s+t)/2)*log(1-P1-P2)
            if(s == t){
               D<-0
            }else{
               D<-sum(log(seq(1,(t-s)/2)))
            }
            if(M == (s+t)/2){
               E<-0
            }else{
               E<-sum(log(seq(1,M-(s+t)/2)))
            }
            if(s == 0){
               F<-0
            }else{
               F<-sum(log(seq(1,s)))
            }
```

```
            log.sum <- A+B+C-D-E-F
            Sum <- Sum + exp(log.sum)
         }
      }
   }
Sum
}

######################### Sum.dP2.h.of.P.t ##############################
##Returns derivative wrt P2 of h(P,n)
Sum.dP2.h.of.P.t <- function(M,t,P1,P2){
   #EVEN t
   if(floor(t/2) == ceiling(t/2)){
      Sum <- 0
      for (s in 0:(min(t,2*M-t))){
         #EVEN s
         if(floor(s/2) == ceiling(s/2)){
            A<-s*log(P1)+ sum(log(seq(1,M)))
            B<-((t-s)/2)*log(P2)
            C<-(M-(s+t)/2)*log(1-P1-P2)
            if(s == t){
               D<-0
            }else{
               D<-sum(log(seq(1,(t-s)/2)))
            }
            if(M == (s+t)/2){
               E<-0
            }else{
               E<-sum(log(seq(1,M-(s+t)/2)))
            }
            if(s == 0){
               F<-0
            }else{
               F<-sum(log(seq(1,s)))
            }
            G<-((t-s)/2)/P2 + (M-(s+t)/2)/(1-P1-P2)
            if (s>0){
               G<-G - 2*s/P1
            }
            if (G<0){
               log.sum <- A+B+C-D-E-F+log(-G)
               Sum<-Sum - exp(log.sum)
            } else {
               log.sum <- A+B+C-D-E-F+log(G)
               Sum<-Sum + exp(log.sum)
            }
         }
      }
   }
   #ODD t
   if(floor(t/2) != ceiling(t/2)){
      Sum <- 0
      for (s in 1:(min(t,2*M-t))){
         #ODD s
```

```
         if(floor(s/2) != ceiling(s/2)){
            A<-s*log(P1)+ sum(log(seq(1,M)))
            B<-((t-s)/2)*log(P2)
            C<-(M-(s+t)/2)*log(1-P1-P2)
            if(s == t){
               D<-0
            }else{
               D<-sum(log(seq(1,(t-s)/2)))
            }
            if(M == (s+t)/2){
               E<-0
            }else{
               E<-sum(log(seq(1,M-(s+t)/2)))
            }
            if(s == 0){
               F<-0
            }else{
               F<-sum(log(seq(1,s)))
            }
            G<-((t-s)/2)/P2 - (2*s)/P1 + (M-(s+t)/2)/(1-P1-P2)
            if (G<0){
               log.sum <- A+B+C-D-E-F+log(-G)
               Sum<-Sum - exp(log.sum)
            } else {
               log.sum <- A+B+C-D-E-F+log(G)
               Sum<-Sum + exp(log.sum)
            }
         }
      }
   }
Sum
}

################################ F.P ####################################
## Returns Sum Ratio of (dlogL/dP2)/f(P,n)
F.P<-function(M,T,P,P2){
   b<-length(T)
   P1<-2*P-2*P2
   Ratio<-0
   for(i in 1:b){
      t<-T[i]
      Denominator<-Sum.f.of.P.t(M,t,P1,P2)
      Numerator<-Sum.dP2.h.of.P.t(M,t,P1,P2)
      Ratio<-Ratio + Numerator/Denominator
   }
Ratio
}

######################## logLikelihood.Exact ############################
## Returns Likelihood Evaluated for k break totals
logLikelihood.Exact <- function(M,T,P){
   P1<-P[1]
   P2<-P[2]
   if (P1+P2<1 && P1>=0 && P2>=0){
```

```
      b<-length(T)
      Prob<-0
      for (i in 1:b){
         t<-T[i]
         Prob <- Prob + log(Sum.f.of.P.t(M,t,P1,P2))
      }
   }else{Prob <- -1E308}
   if (Prob == "NaN" | Prob == -Inf) {Prob <- -1E308}
(Prob)
}

########################## Exact.MLE.above ##############################
## Returns MLEs of P1, P2, and Pi Searching for P2.hat from above
Exact.MLE.above <- function(M,T,cutoff){
   P<-Binomial.MLE(M,T)[1]
   precision<-0.1
   P2<-P-precision
   while(P2<=0){
      P2<-P2+precision
      precision<-precision*0.1
      P2<-P2-precision
   }

   while (precision>cutoff){
      out<-F.P(M,T,P,P2)
      if (out<0 | out=="NaN") {
         P2<-P2-precision
         while(P2<=0){
            P2<-P2+precision
            precision<-precision*0.1
            P2<-P2-precision
         }

      } else {
         P2<-P2+precision
         precision<-precision*0.1
         P2<-P2-precision
      }
   }
   P1<-2*P-2*P2
c(P1,P2,P,logLikelihood.Exact(M,T,c(P1,P2)))
}

########################## Exact.MLE.below ##############################
## Returns MLEs of P1, P2, and Pi Searching for P2.hat from below
Exact.MLE.below <- function(M,T,cutoff){
   P<-mean(T)/(2*M)
   precision<-0.1
   P2<-precision
   while(P2>=P){
      P2<-P2-precision
      precision<-precision*0.1
      P2<-P2+precision
   }
```

```
   while (precision>cutoff){
      out<-F.P(M,T,P,P2)
      if (out>0 | out=="NaN") {
         P2<-P2+precision
         while(P2>=P){
            P2<-P2-precision
            precision<-precision*0.1
            P2<-P2+precision
         }

      } else {
         P2<-P2-precision
         precision<-precision*0.1
         P2<-P2+precision
      }
   }
   P1<-2*P-2*P2
c(P1,P2,P,logLikelihood.Exact(M,T,c(P1,P2)))
}

###################### Exact.MLE.P1.P2.Pi ##############################
## Finds MLE of P1,P2, and Pi & evaluates likelihood.
Exact.MLE.P1.P2.Pi<-function(M,T,cutoff){
   Above<-Exact.MLE.above(M,T,cutoff)
   Below<-Exact.MLE.below(M,T,cutoff)
   if(Above[4]<Below[4]){
      Exact.MLE<-Below
   }else{Exact.MLE<-Above}
Exact.MLE
}

####################### posBinomial.MLE ##############################
## Returns Pi.hat & Likelihood for positive binomial distribution
posBinomial.MLE <- function(M,T,cutoff){
   b<-length(T)
   mean.T<-mean(T)
   phat<-mean(T)/(2*M)
   F.L<-function(phat){
      ratio=2*M*phat/(1-(1-phat)^(2*M))
   ratio
   }
   diff<-F.L(phat) - mean.T
   precision<-0.1
   while (precision>cutoff){
      out<-F.L(phat)
      diff<-out-mean.T
      if (diff>0) {
         phat<-phat-precision
         while(phat<=0){
            phat<-phat+precision
            precision<-precision*0.1
            phat<-phat-precision
         }
      } else {
```

```
         phat<-phat+precision
         precision<-precision*0.1
         phat<-phat-precision
         while(phat<=0){
            phat<-phat+precision
            precision<-precision*0.1
            phat<-phat-precision
         }
      }
   }
   logPosBin.Like<-0
   for (i in 1:b){
      new<-dbinom(T[i],2*M,phat) / (1-(1-phat)^(2*M))
      logPosBin.Like<-logPosBin.Like + log(new)
   }
c(phat,logPosBin.Like)
}

############################## F.Ppos ###################################
## Returns Sum Ratio of (dlogL/dP2)/h(P,n) for positive data
F.Ppos<-function(M,T,P,P2){
   b<-length(T)
   P1<-2*P-2*P2
   Ratio<-b*M*((1-2*P+P2)^(M-1))/(1-((1-2*P+P2)^M))
   for(i in 1:b){
      t<-T[i]
      Denominator<-Sum.f.of.P.t(M,t,P1,P2)
      Numerator<-Sum.dP2.h.of.P.t(M,t,P1,P2)
      Ratio<-Ratio + Numerator/Denominator
   }
Ratio
}

######################## logLikelihood.posExact ########################
## Returns Likelihood Evaluated for k positive break totals
logLikelihood.posExact <- function(M,T,P){
   P1<-P[1]
   P2<-P[2]
   if (P1+P2<1 && P1>=0 && P2>=0){
      b<-length(T)
      Prob<- -b*log(1-(1-P1-P2)^M)
      for (i in 1:b){
         t<-T[i]
         Prob <- Prob + log(Sum.f.of.P.t(M,t,P1,P2))
      }
   }else{Prob <- -1E308}
   if (Prob == "NaN" | Prob == -Inf) {Prob <- -1E308}
(Prob)
}

########################## posExact.MLE.above ##########################
## Returns MLEs of P1, P2, and Pi Searching for P2.hat from above for
## positive count data.
posExact.MLE.above <- function(M,T,P,cutoff){
```

```
   precision<-0.1
   P2<-P-precision
   while(P2<=0){
      P2<-P2+precision
      precision<-precision*0.1
      P2<-P2-precision
   }

   while (precision>cutoff){
      out<-F.Ppos(M,T,P,P2)
      if (out<0 | out=="NaN") {
         P2<-P2-precision
         while(P2<=0){
            P2<-P2+precision
            precision<-precision*0.1
            P2<-P2-precision
         }

      } else {
         P2<-P2+precision
         precision<-precision*0.1
         P2<-P2-precision
         while(P2<=0){
            P2<-P2+precision
            precision<-precision*0.1
            P2<-P2-precision
         }

      }
   }
   P1<-2*P-2*P2
c(P1,P2,P,logLikelihood.posExact(M,T,c(P1,P2)))
}

########################## posExact.MLE.below ############################
## Returns MLEs of P1, P2, and Pi Searching for P2.hat from below for
## positive count data.
posExact.MLE.below <- function(M,T,P,cutoff){
   precision<-0.1
   P2<-precision
   while(P2>=P){
      P2<-P2-precision
      precision<-precision*0.1
      P2<-P2+precision
   }
   while (precision>cutoff){
      out<-F.Ppos(M,T,P,P2)
      if (out>0 | out=="NaN") {
         P2<-P2+precision
         while(P2>=P){
            P2<-P2-precision
            precision<-precision*0.1
            P2<-P2+precision
         }
```

```
      } else {
         P2<-P2-precision
         precision<-precision*0.1
         P2<-P2+precision
         while(P2>=P){
            P2<-P2-precision
            precision<-precision*0.1
            P2<-P2+precision
         }

      }
   }
   P1<-2*P-2*P2
c(P1,P2,P,logLikelihood.posExact(M,T,c(P1,P2)))
}

##################### posExact.MLE.P1.P2.Pi ############################
## Finds MLE of P1,P2, and Pi & evaluates likelihood for positive total
## break count data.
posExact.MLE.P1.P2.Pi<-function(M,T,cutoff){
   i<-1
   j<-0
   T.pos<-c(1)
   while(i<=length(T)){
      if (T[i] != 0){
         if (j == 0){
            T.pos[1]<-T[i]
            j=1
         }else{
            T.pos<-c(T.pos,T[i])
         }
      }
      i=i+1
   }
   P<-posBinomial.MLE(M,T.pos,cutoff)[1]
   Above<-posExact.MLE.above(M,T.pos,P,cutoff)
   Below<-posExact.MLE.below(M,T.pos,P,cutoff)
   if(Above[4]<Below[4]){
      posExact.MLE<-Below
   }else{posExact.MLE<-Above}
posExact.MLE
}

############################### F.PMultpos ############################
## Returns Sum Ratio of (dlogL/dP2)/h(P,n) for positive partitioned
## data
F.PMultpos<-function(M,Band,P,P2){
   m0<-sum(Band[,1])
   m1<-sum(Band[,2])
   m2<-sum(Band[,3])
   b<-length(Band[,1])
   P1<-2*P-2*P2
   Ratio<- b*M*((1-2*P+P2)^(M-1))/(1-((1-2*P+P2)^M))
```

```
   Ratio<-Ratio + m0/(1-2*P+P2) - 2*m1/P1 + m2/P2
Ratio
}

######################## Multinom.logLikelihood ########################
## Returns Likelihood Evaluated for k partitioned break totals
Multinom.logLikelihood <- function(M,Band,P1,P2){
   P0<-1-P1-P2
   M1<-Band[,2]
   M2<-Band[,3]
   M0<-Band[,1]
   b<-length(M1)
   if (P1+P2<1 && P1>=0 && P2>=0){
      A<-b*sum(log(seq(1,M)))
      B<-sum(M1)*log(P1)
      C<-sum(M2)*log(P2)
      D<-sum(M0)*log(P0)
      E<-0
      for (i in 1:b){
         if (M1[i]>1){
            E<-E + sum(log(seq(1,M1[i])))
         }
         if (M2[i]>1){
            E<-E + sum(log(seq(1,M2[i])))
         }
         if (M0[i]>1){
            E<-E + sum(log(seq(1,M0[i])))
         }
      }
      F<-b*log(1-P0^M)
      log.sum<-A+B+C+D-E-F
      Like<-log.sum
   }else{Like <- -1E308}
   if (Like=="NaN" | Like=="-Inf") {Like<--1E308}
Like
}

######################### posMult.MLE.above ##########################
## Returns MLEs of P1, P2, and Pi Searching for P2.hat from above for
## partitioned positive count data.
posMult.MLE.above <- function(M,Band,P,cutoff){
   precision<-0.1
   P2<-P-precision
   while(P2<=0){
      P2<-P2+precision
      precision<-precision*0.1
      P2<-P2-precision
   }
   while (precision>cutoff){
      out<-F.PMultpos(M,Band,P,P2)
      if (out<0 | out=="NaN") {
         P2<-P2-precision
         while(P2<=0){
            P2<-P2+precision
```

```
            precision<-precision*0.1
            P2<-P2-precision
         }

      } else {
         P2<-P2+precision
         precision<-precision*0.1
         P2<-P2-precision
         while(P2<=0){
            P2<-P2+precision
            precision<-precision*0.1
            P2<-P2-precision
         }

      }
   }
   P1<-2*P-2*P2
c(P1,P2,P,Multinom.logLikelihood(M,Band,P1,P2))
}

########################## posMult.MLE.below ############################
## Returns MLEs of P1, P2, and Pi Searching for P2.hat from below for
## partitioned positive count data.
posMult.MLE.below <- function(M,Band,P,cutoff){
   precision<-0.1
   P2<-precision
   while(P2>=P){
      P2<-P2-precision
      precision<-precision*0.1
      P2<-P2+precision
   }
   while (precision>cutoff){
      out<-F.PMultpos(M,Band,P,P2)
      if (out>0 | out=="NaN") {
         P2<-P2+precision
         while(P2>=P){
            P2<-P2-precision
            precision<-precision*0.1
            P2<-P2+precision
         }

      } else {
         P2<-P2-precision
         precision<-precision*0.1
         P2<-P2+precision
         while(P2>=P){
            P2<-P2-precision
            precision<-precision*0.1
            P2<-P2+precision
         }

      }
   }
   P1<-2*P-2*P2
```

```
c(P1,P2,P,Multinom.logLikelihood(M,Band,P1,P2))
}

##################### posMult.MLE.P1.P2.Pi ############################
## Finds positive MLE of P1,P2, and Pi & evaluates likelihood for
## partitioned break count data.
posMult.MLE.P1.P2.Pi <-function(M,Band,cutoff){
   i<-1
   j<-0
   T<-Band[,2]+2Band[,3]
   Band.pos<-matrix(1,1,3)
   while(i<=length(T)){
      if (T[i] != 0){
         if (j == 0){
            Band.pos[1,]<-Band[i,]
            j=1
         }else{
            Band.pos<-rbind(Band.pos,Band[i,])
         }
      }
      i=i+1
   }
   T.pos<-Band.pos[,2]+2Band.pos[,3]
   P<-posBinomial.MLE(M,T.pos,cutoff)[1]
   Above<-posMult.MLE.above(M,Band.pos,P,cutoff)
   Below<-posMult.MLE.below(M,Band.pos,P,cutoff)
   if(Above[4]<Below[4]){
      posMult.MLE<-Below
   }else{posMult.MLE<-Above}
posMult.MLE
}
```

APPENDIX B

ALPHA LEVEL PLOTS AND POWER CURVES FOR DETECTING CORRELATION IN BINARY COUNT DATA

Here we present the full array of alpha level plots and power curves for detecting correlation discussed in Chapter III. Simulation parameters were chosen based on the fragile-site data and experimental protocol given in Böhm et al. (1995). All results are based on a nominal alpha level of 5%, 100 observed metaphases, and 1,000 Monte Carlo samples unless otherwise indicated. All simulations were performed using R (R Core Development Team (2003)) version 1.8.1.

B.1 Type I Error Rate (Alpha Level) in Detecting Correlation for Subsets of Size k_b

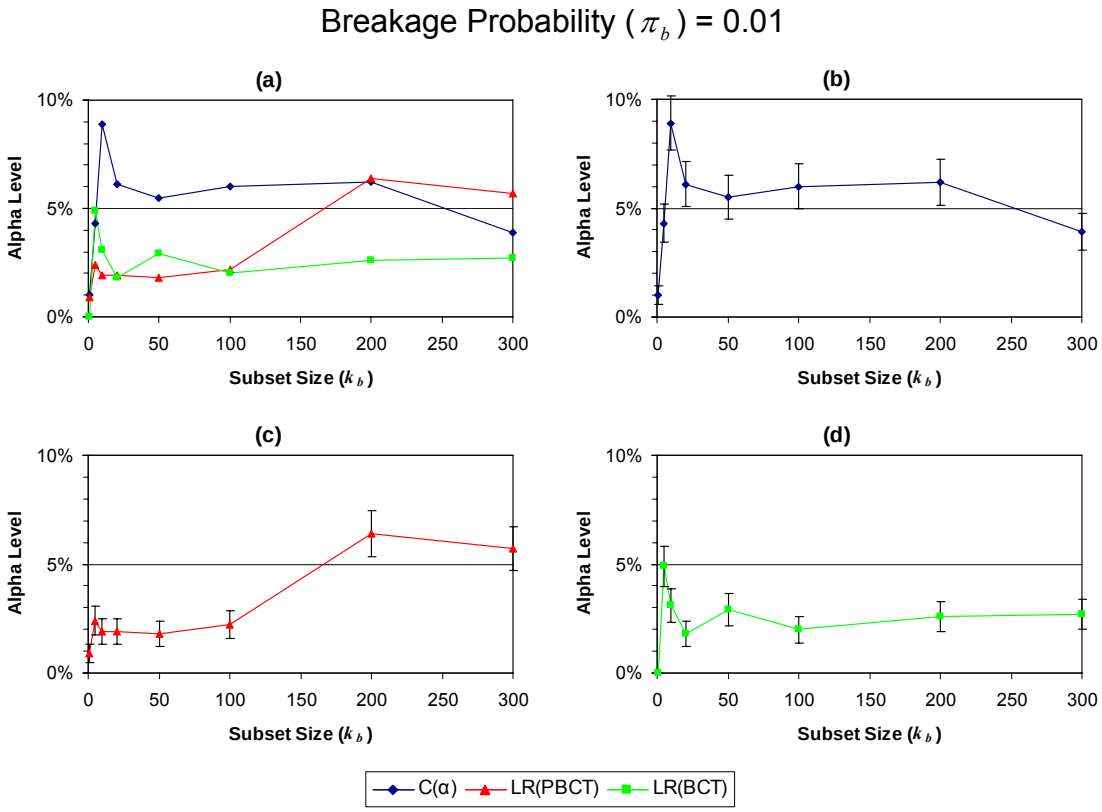


Figure B-1. Simulated Alpha Level of Three Tests for Correlation Where the Breakage Probability is 0.01. Simulated results are based on 1,000 Monte Carlo samples of chromosomal breakage data from 100 metaphases where correlation is equal to zero. The alpha level was computed as the percentage of simulations for which the null hypothesis of zero correlation was rejected. The curves in (a) are separated into plots (b), (c) and (d) and include 83% confidence intervals based on 1,000 simulations.

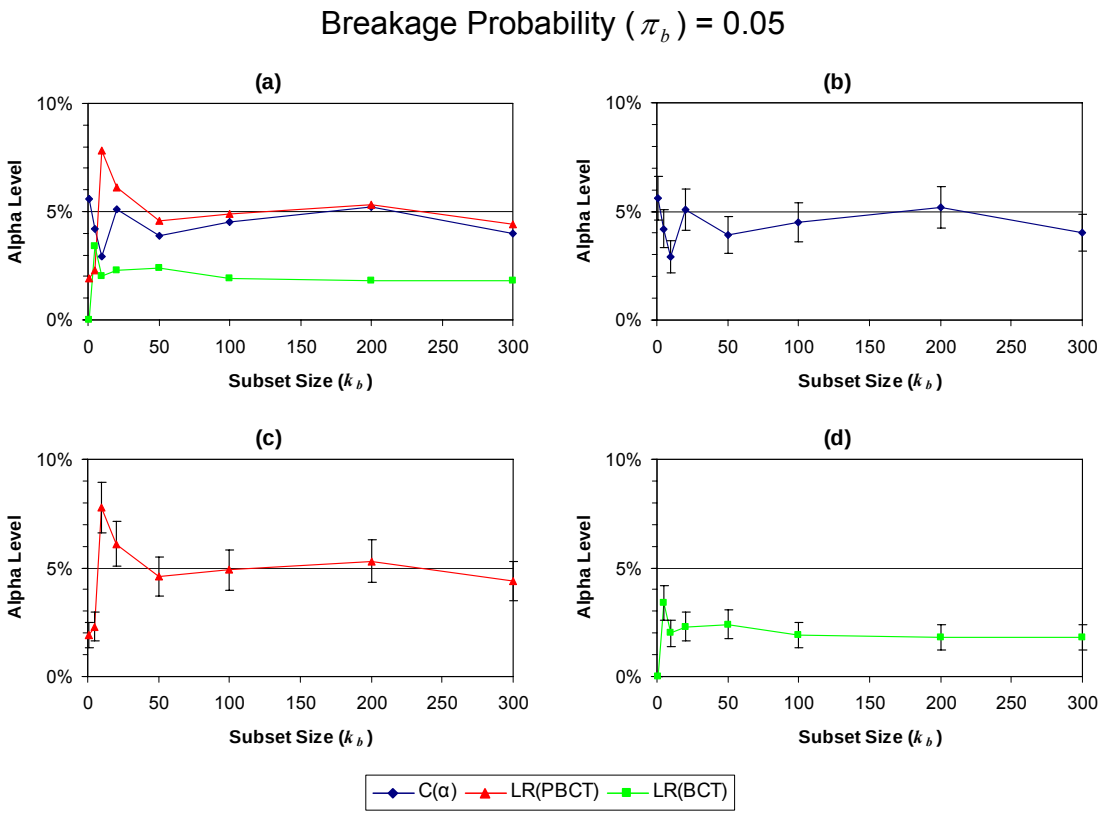


Figure B-2. Simulated Alpha Level of Three Tests for Correlation Where the Breakage Probability is 0.05. Simulated results are based on 1,000 Monte Carlo samples of chromosomal breakage data from 100 metaphases where correlation is equal to zero. The alpha level was computed as the percentage of simulations for which the null hypothesis of no correlation was rejected. The curves in (a) are separated into plots (b), (c) and (d) and include 83% confidence intervals based on 1,000 simulations.

B.2 Power Curves for Detecting Correlation Using Subsets of Size k_b

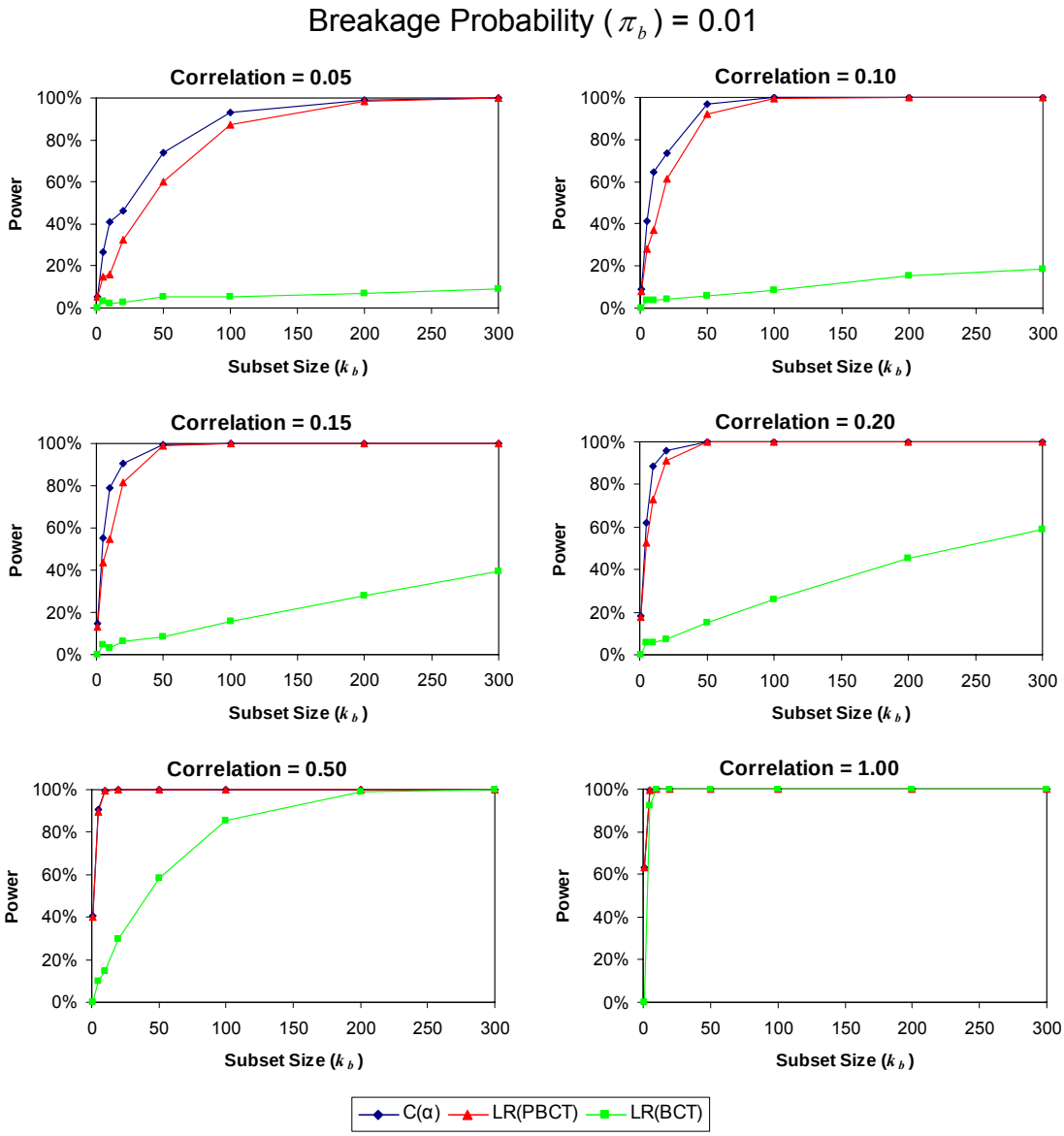


Figure B-3. Simulated Power Curves of Three Tests for Correlation When the Probability of Breakage is 0.01. Simulated results are based on 1,000 Monte Carlo samples with 100 metaphases. Correlations range from 0.10 to 1.00. Power was computed as the percentage of simulations for which the null hypothesis of no correlation was rejected.

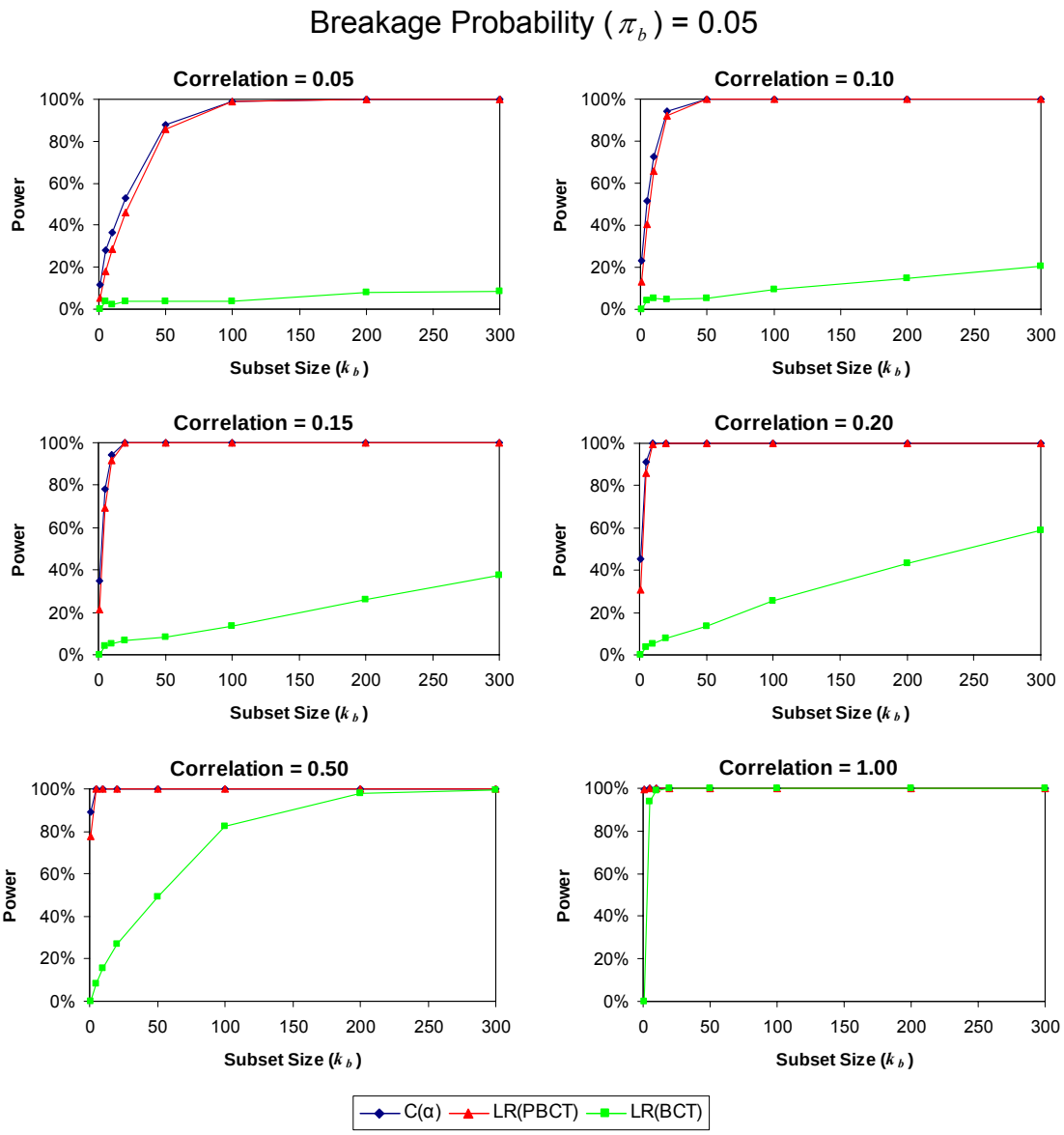


Figure B-4. Simulated Power Curves of Three Tests for Correlation When the Probability of Breakage is 0.05. Simulated results are based on 1,000 Monte Carlo samples with 100 metaphases. Correlations range from 0.10 to 1.00. Power was computed as the percentage of simulations for which the null hypothesis of no correlation was rejected.

B.3 Type I Error Rates (Alpha Level) for Detecting Correlation in a Single Site Using the C(α) Test

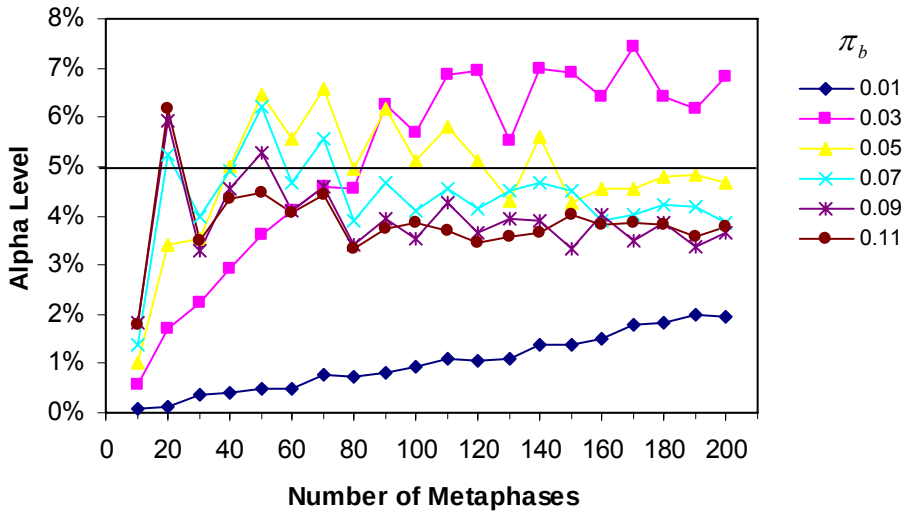


Figure B-5. Simulated Alpha Level of the C(α) Test for Correlation at a Single Site. Simulated results are based on 10,000 Monte Carlo samples where correlation equals zero. The alpha level was computed as the percentage of simulations for which the null hypothesis of no correlation was rejected. Confidence interval bars are not included since the interval width is less than the height of the points on the graph.

B.4 Power Curves for Detecting Correlation in a Single Site Using the C(α) Test

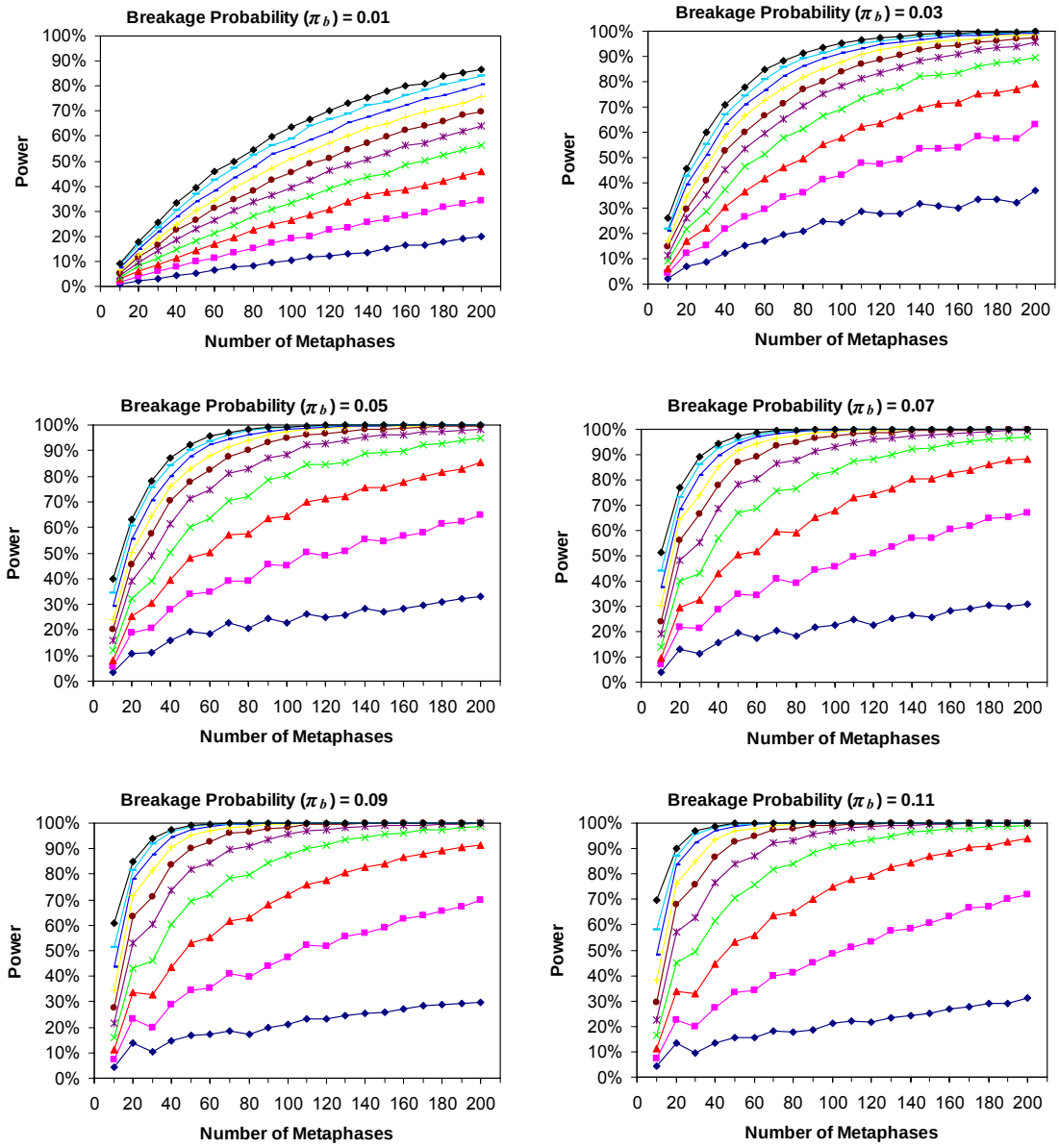


Figure B-6. Simulated Power Curves for Detecting Various Correlations When the Probability of Breakage Ranges From 0.01 to 0.11. Simulated results are based on 10,000 Monte Carlo samples, where correlations range from 0.1 to 1.0 and the number of metaphases range from 10 to 200. Power was computed as the percentage of simulations for which the null hypothesis of no correlation was rejected. Confidence interval bars are not included since the interval width is less than the height of the points on the graph.

APPENDIX C

FSM AND FSM3 SIMULATION RESULTS

The full complement of FSM and FSM3 simulation results discussed in Chapter IV are given here. All simulations are based on 1,000 Monte Carlo samples where the non-fragile breakage probability is 0.005 and 100 metaphases are observed. There are an equal number of fragile sites for each breakage probability. For example, if the breakage probabilities are 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055, then for 18 fragile sites, there are three bands with breakage probability equal to 0.022, three bands with probability equal to 0.024, and so forth.

C.1 FSM Simulation Results

C.1.1 FS Breakage Probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055

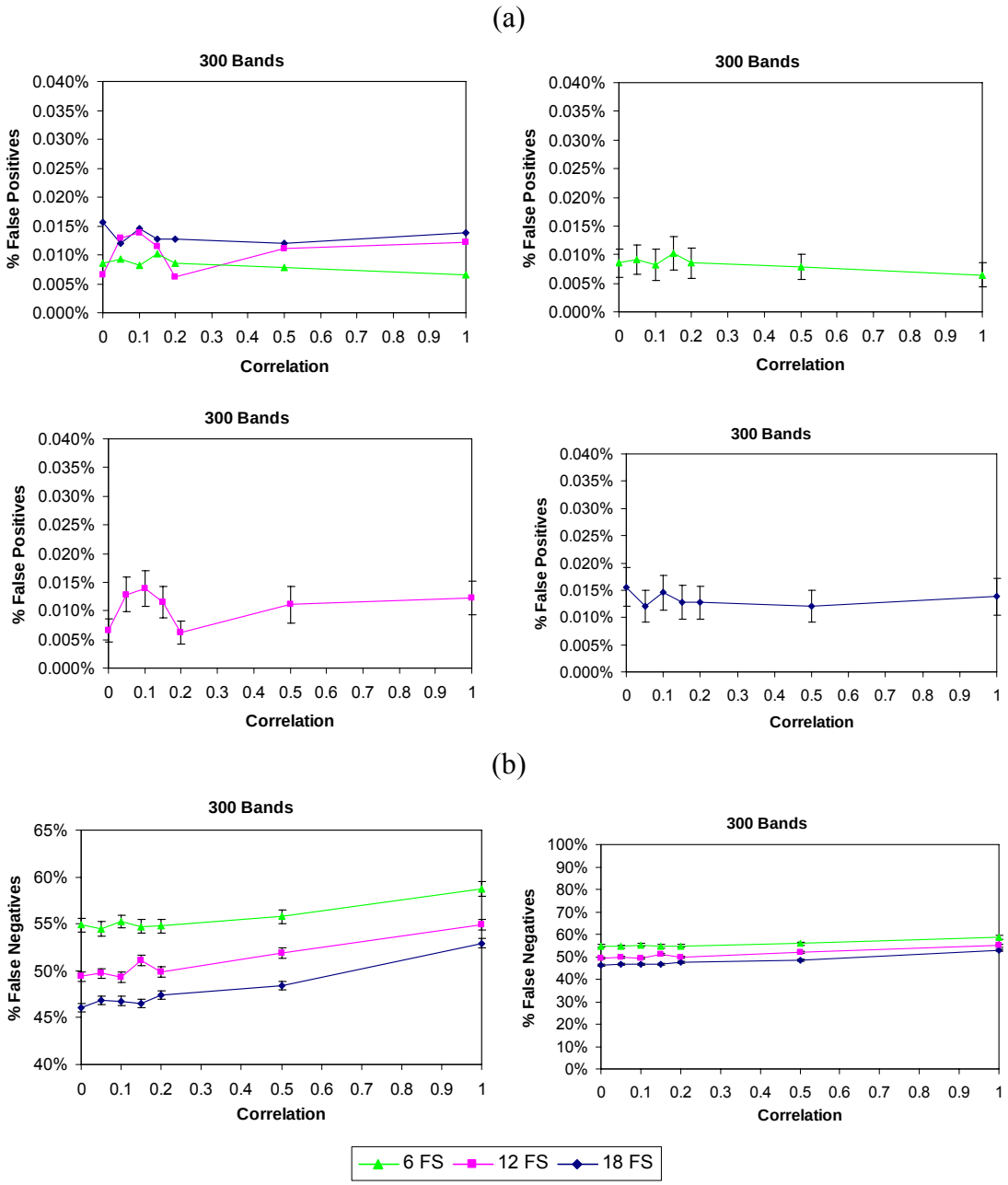


Figure C-1. FSM (a) False-Positive and (b) False-Negative Rates for 300 Bands and 6 to 18 Fragile Sites With Breakage Probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. The curves in (a) are first plotted together and then separately to make it possible to see the individual 83% confidence intervals. The curves in (b) are plotted on two different Y-axis scales.

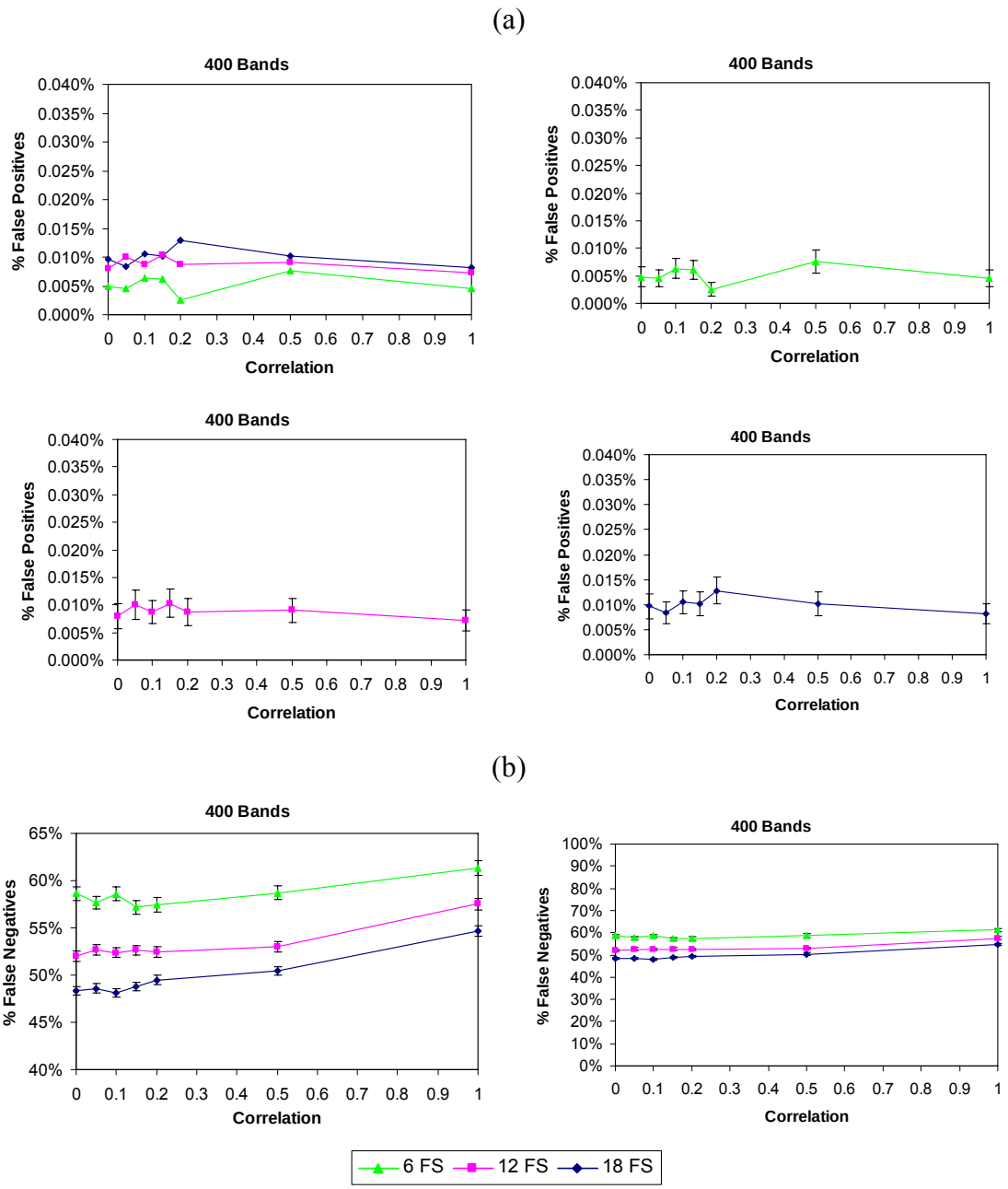
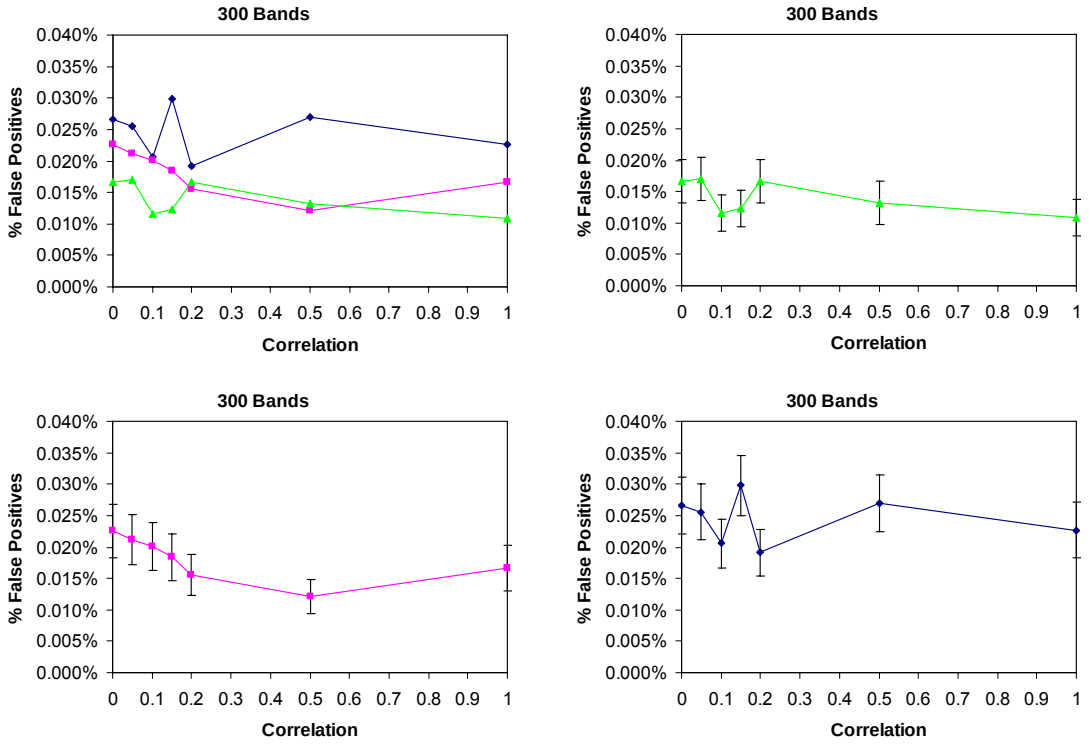


Figure C-2. FSM (a) False-Positive and (b) False-Negative Rates for 400 Bands and 6 to 18 Fragile Sites With Breakage Probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. The curves in (a) are first plotted together and then separately to make it possible to see the individual 83% confidence intervals. The curves in (b) are plotted on two different Y-axis scales.

C.1.2 FS Breakage Probabilities of 0.011, 0.0132, 0.0165, 0.0198, 0.022, and 0.0275

(a)



(b)

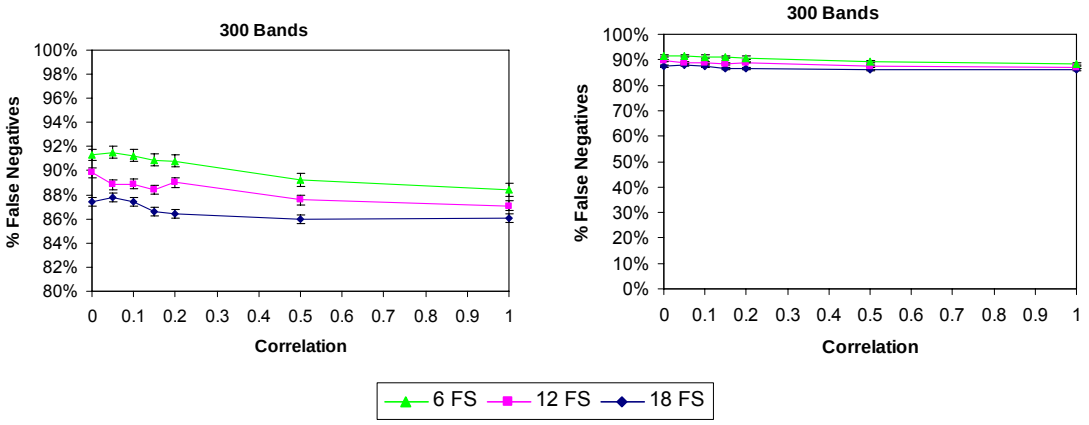


Figure C-3. FSM (a) False-Positive and (b) False-Negative Rates for 300 Bands and 6 to 18 Fragile Sites With Breakage Probabilities of 0.011, 0.0132, 0.0165, 0.0198, 0.022, and 0.0275. The curves in (a) are first plotted together and then separately to make it possible to see the individual 83% confidence intervals. The curves in (b) are plotted on two different Y-axis scales.

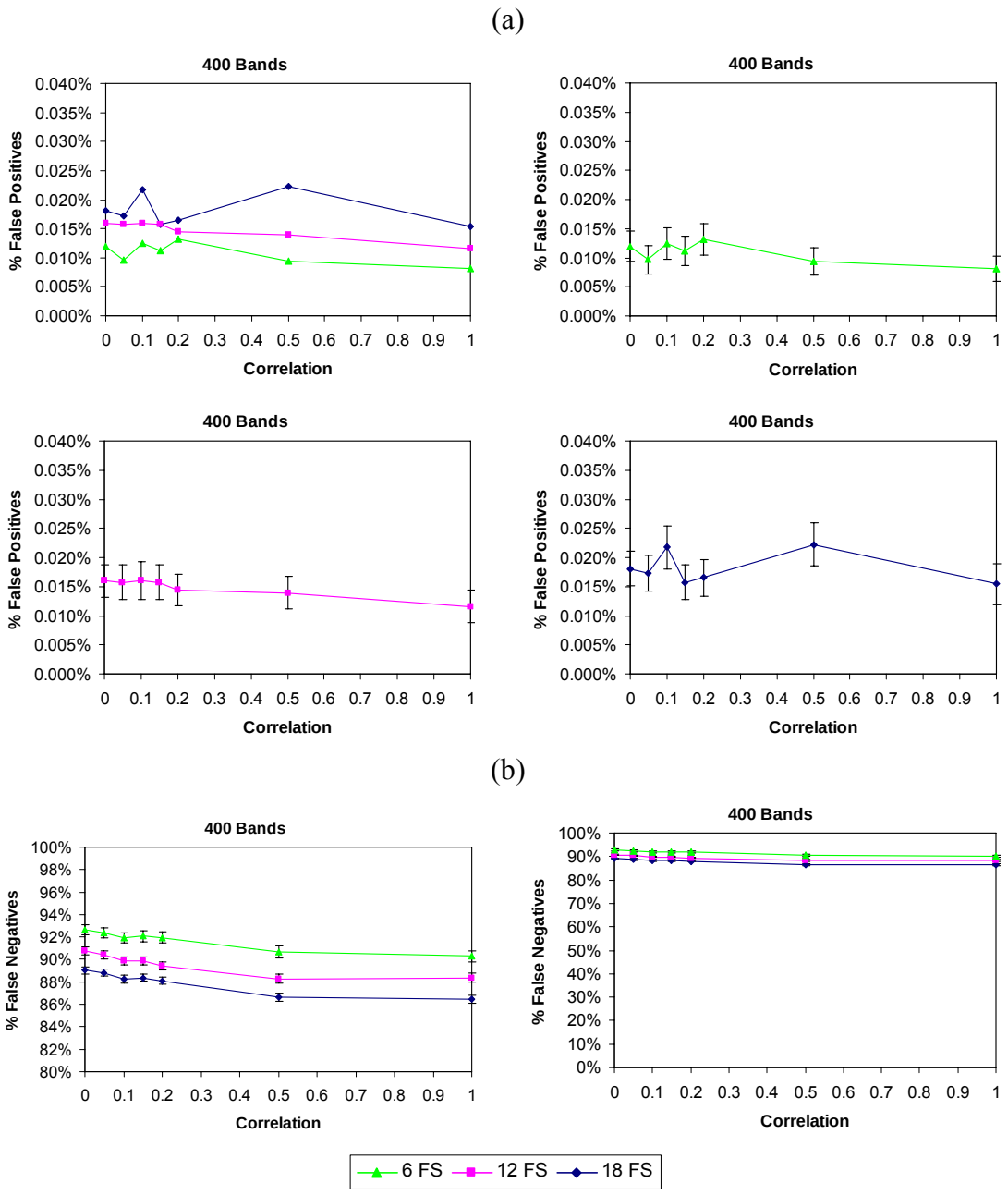


Figure C-4. FSM (a) False-Positive and (b) False-Negative Rates for 400 Bands and 6 to 18 Fragile Sites With Breakage Probabilities of 0.011, 0.0132, 0.0165, 0.0198, 0.022, and 0.0275. The curves in (a) are first plotted together and then separately to make it possible to see the individual 83% confidence intervals. The curves in (b) are plotted on two different Y-axis scales.

C.1.3 FS Breakage Probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055 with 20% Zero-Breakage Sites

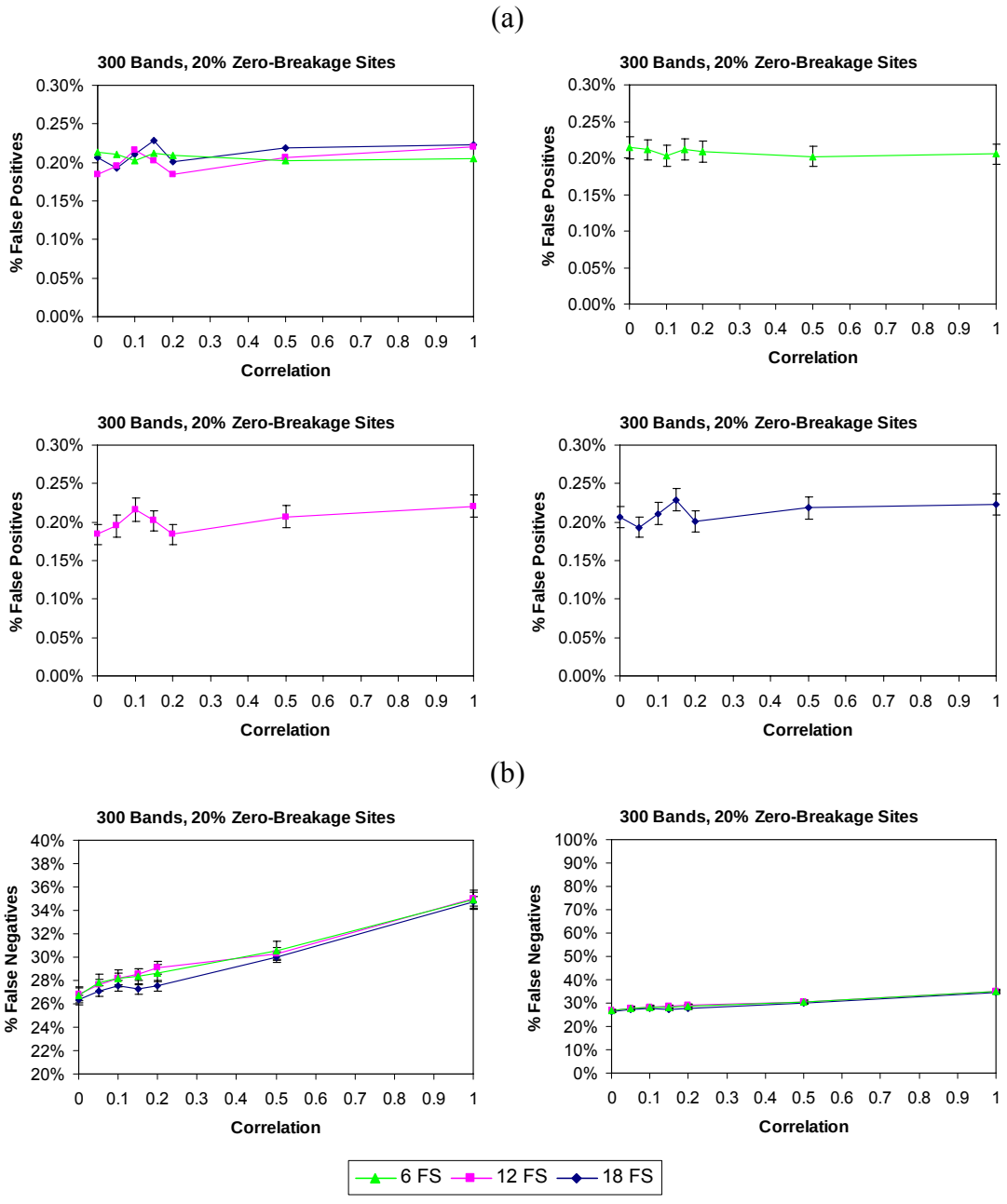


Figure C-5. FSM (a) False-Positive and (b) False-Negative Rates With 20% Zero-Breakage Sites for 300 Bands and 6 to 18 Fragile Sites With Breakage Probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. The curves in (a) are first plotted together and then separately to make it possible to see the individual 83% confidence intervals. Note that the Y-axis scales of (a) are different than those for the case where zero-breakage sites are not present. The curves in (b) are plotted on two different Y-axis scales.

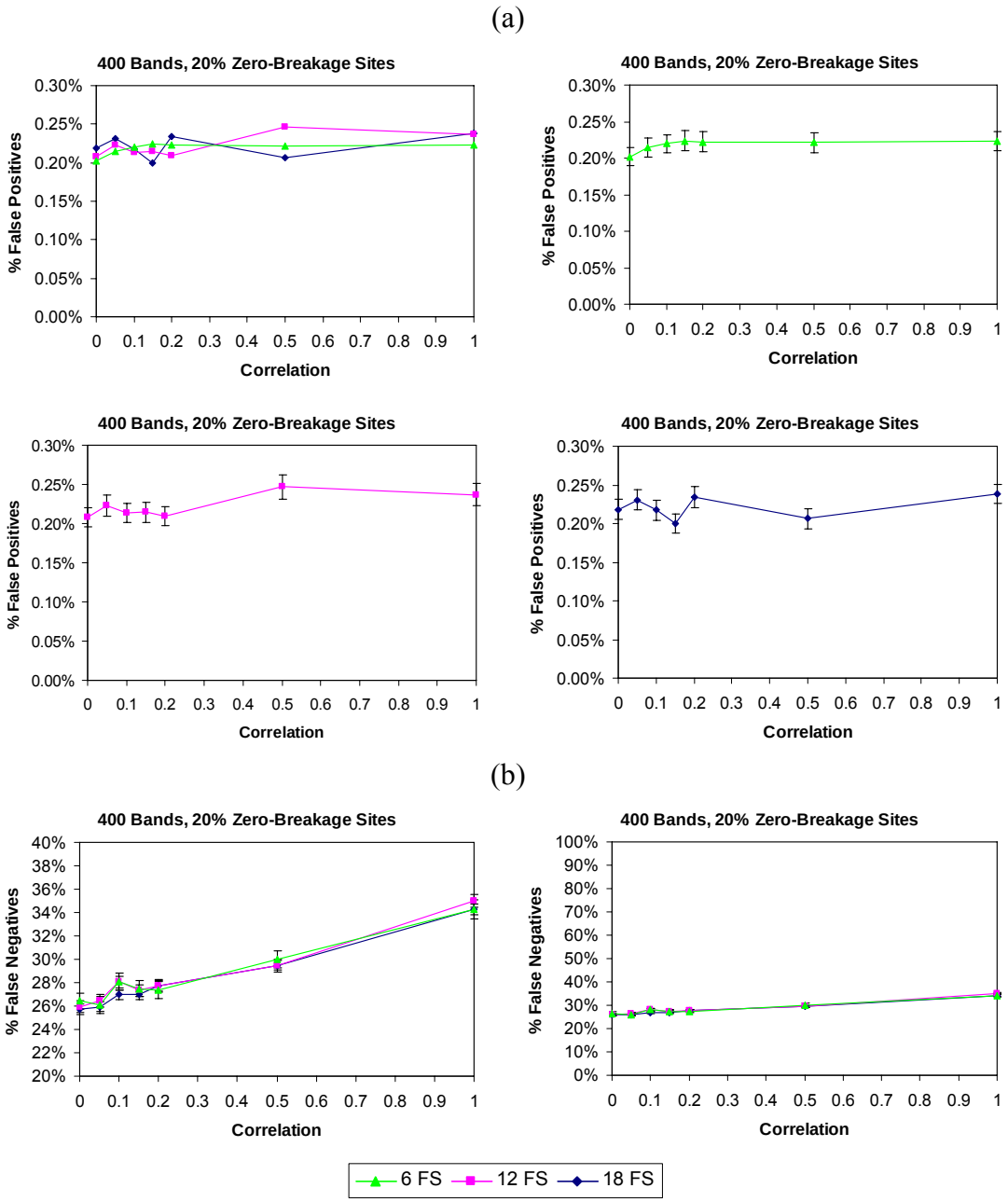


Figure C-6. FSM (a) False-Positive and (b) False-Negative Rates With 20% Zero-Breakage Sites for 400 Bands and 6 to 18 Fragile Sites With Breakage Probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. The curves in (a) are first plotted together and then separately to make it possible to see the individual 83% confidence intervals. Note that the Y-axis scales of (a) are different than those for the case where zero-breakage sites are not present. The curves in (b) are plotted on two different Y-axis scales.

C.2 FSM3 Simulation Results

C.2.1 FS Breakage Probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055

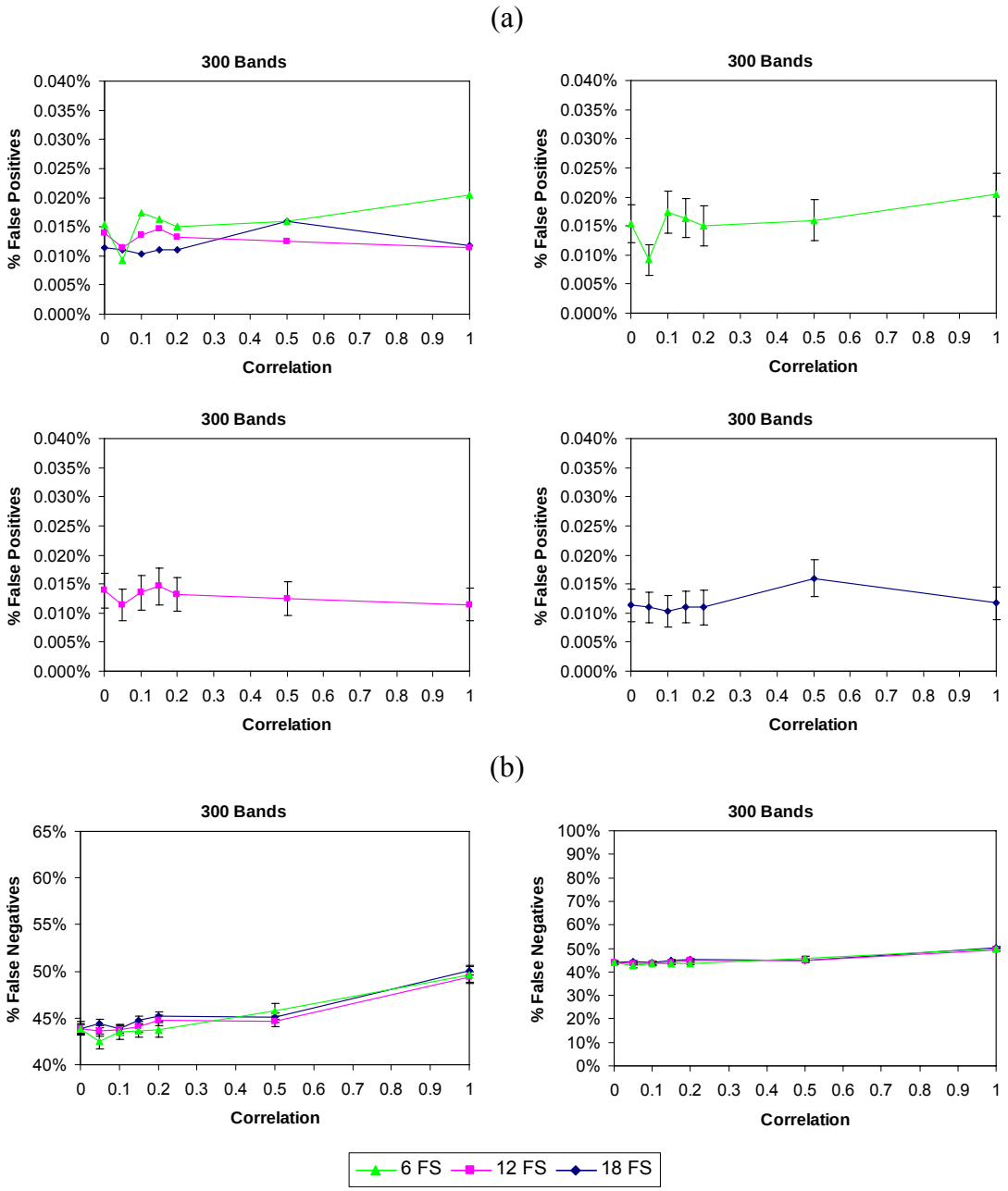


Figure C-7. FSM3 (a) False-Positive and (b) False-Negative Rates for 300 Bands and 6 to 18 Fragile Sites With Breakage Probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. The curves in (a) are first plotted together and then separately to make it possible to see the individual 83% confidence intervals. The curves in (b) are plotted on two different Y-axis scales.

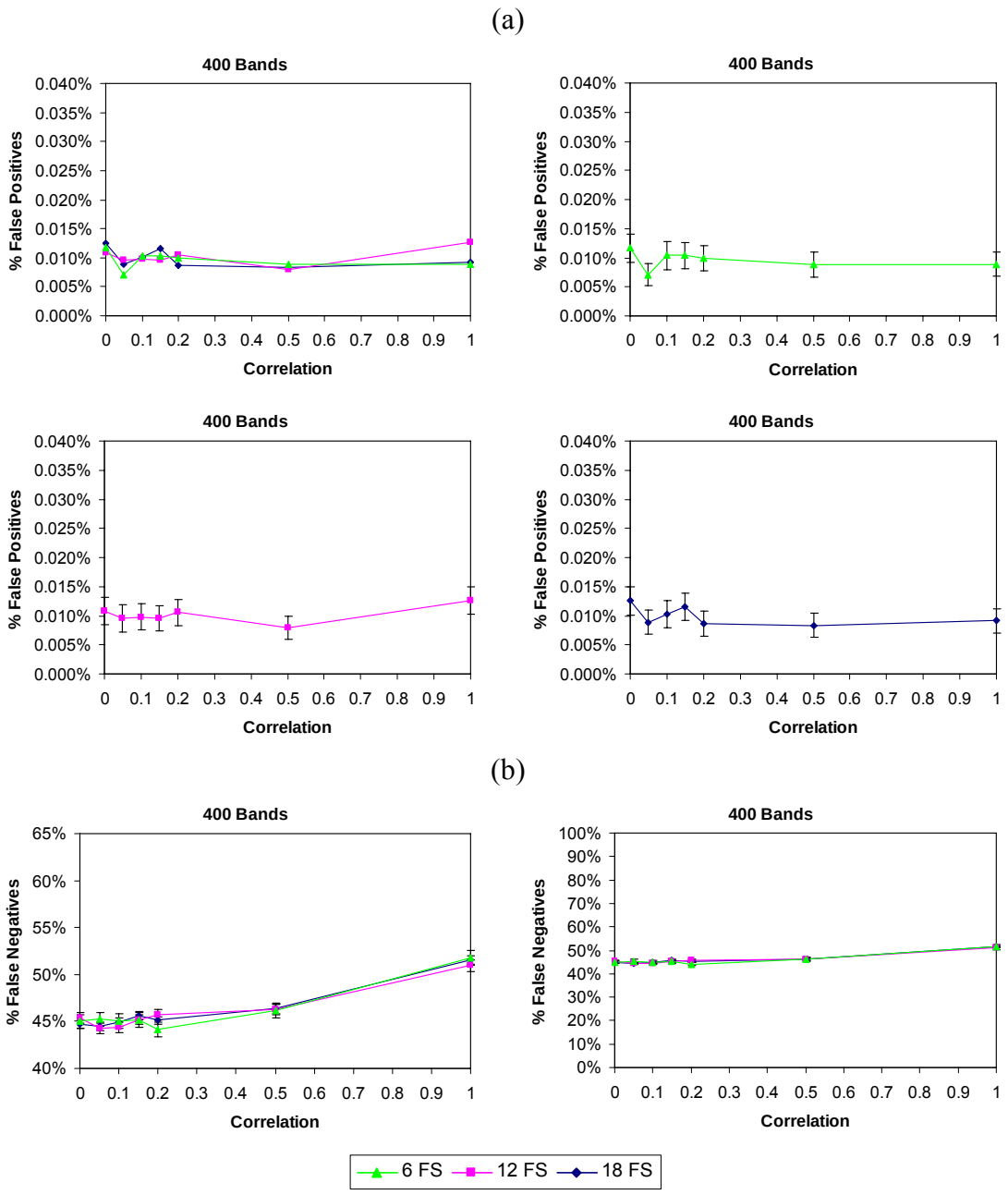


Figure C-8. FSM3 (a) False-Positive and (b) False-Negative Rates for 400 Bands and 6 to 18 Fragile Sites With Breakage Probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. The curves in (a) are first plotted together and then separately to make it possible to see the individual 83% confidence intervals. The curves in (b) are plotted on two different Y-axis scales.

C.2.2 FS Breakage Probabilities of 0.011, 0.0132, 0.0165, 0.0198, 0.022, and 0.0275

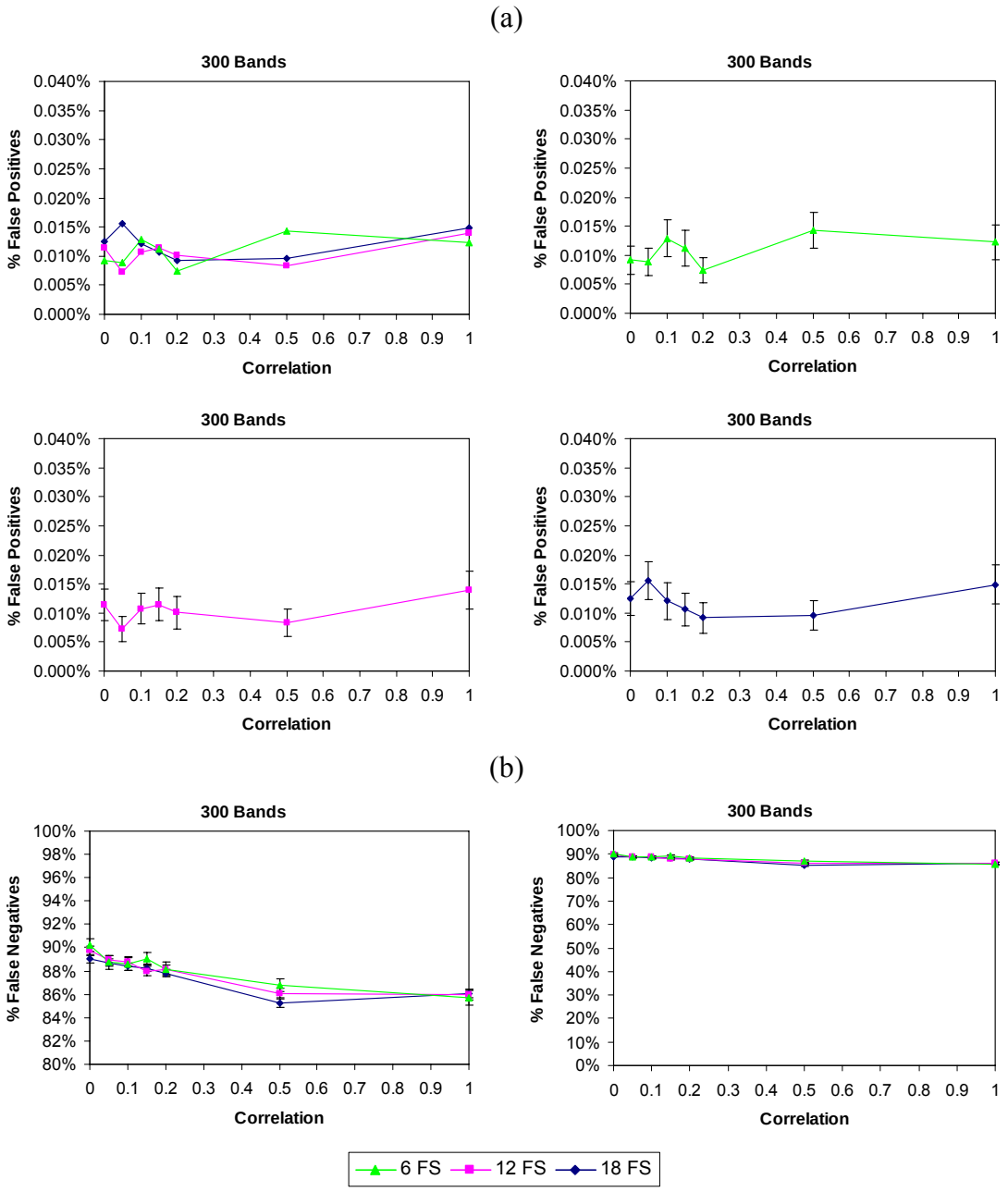


Figure C-9. FSM3 (a) False-Positive and (b) False-Negative Rates for 300 Bands and 6 to 18 Fragile Sites With Breakage Probabilities of 0.011, 0.0132, 0.0165, 0.0198, 0.022, and 0.0275. The curves in (a) are first plotted together and then separately to make it possible to see the individual 83% confidence intervals. The curves in (b) are plotted on two different Y-axis scales.

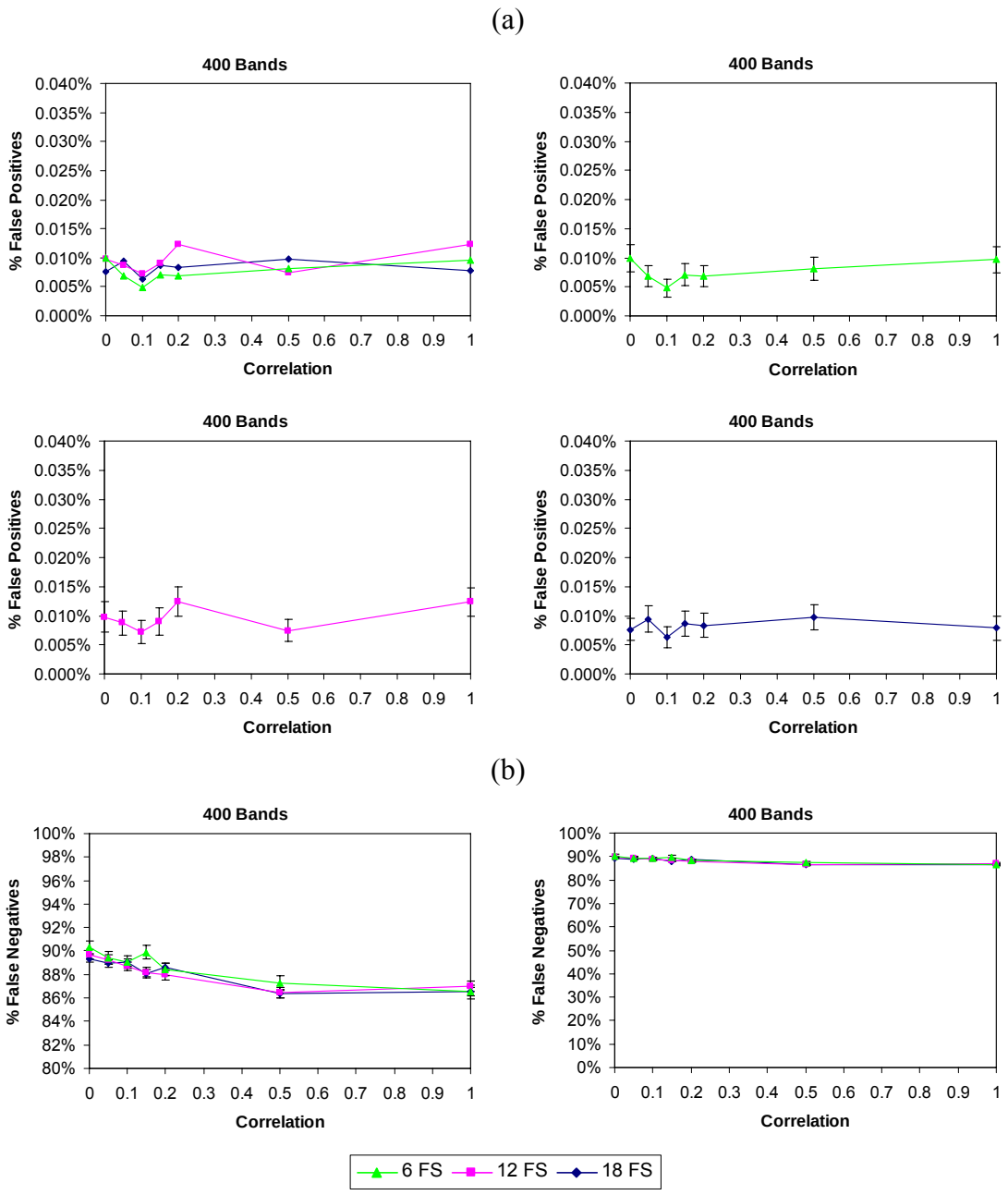


Figure C-10. FSM3 (a) False-Positive and (b) False-Negative Rates for 400 Bands and 6 to 18 Fragile Sites With Breakage Probabilities of 0.011, 0.0132, 0.0165, 0.0198, 0.022, and 0.0275. The curves in (a) are first plotted together and then separately to make it possible to see the individual 83% confidence intervals. The curves in (b) are plotted on two different Y-axis scales.

C.2.3 FS Breakage Probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055 with 20% Zero-Breakage Sites

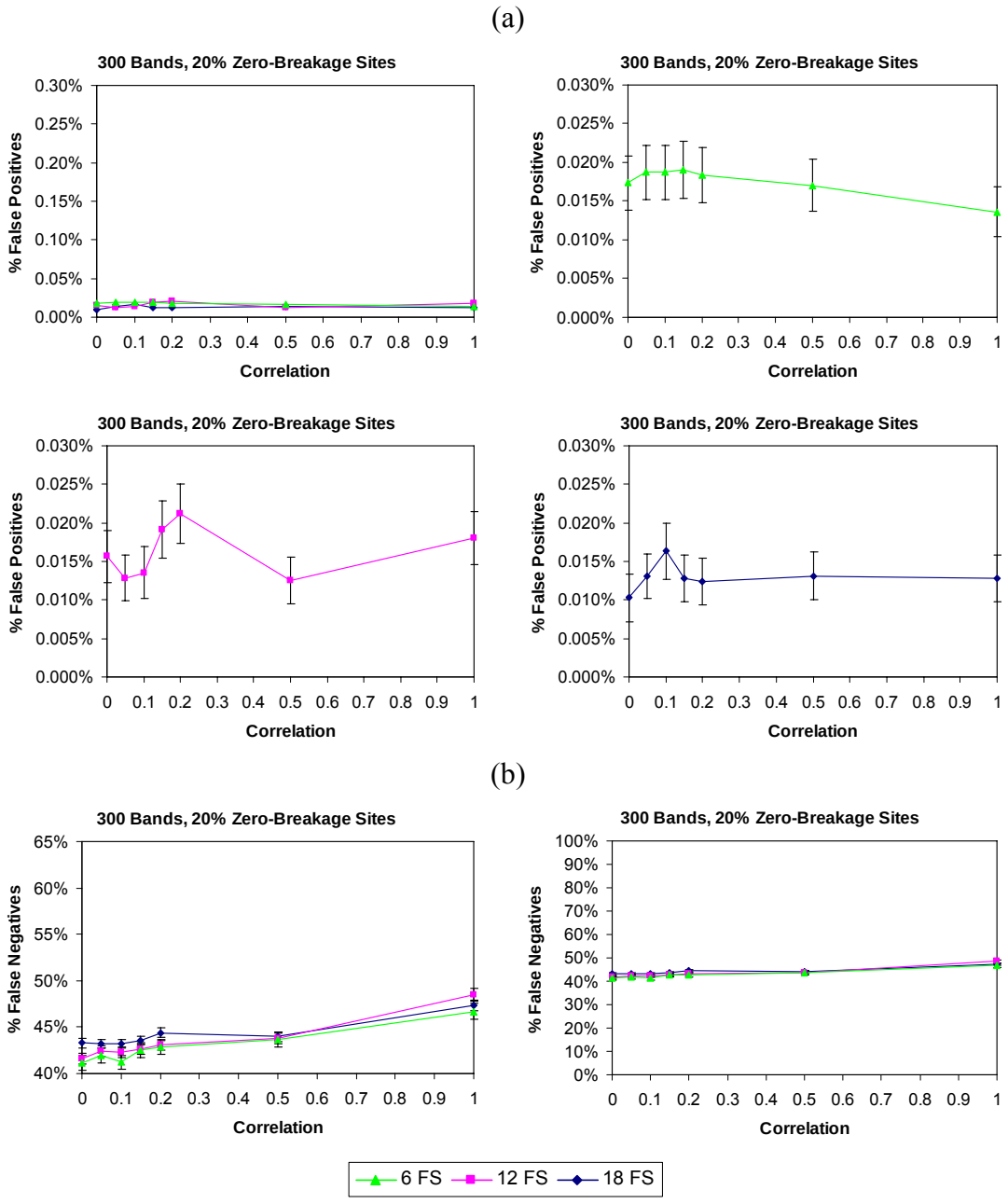


Figure C-11. FSM3 (a) False-Positive and (b) False-Negative Rates With 20% Zero-Breakage Sites for 300 Bands and 6 to 18 Fragile Sites With Breakage Probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. The curves in (a) are plotted together and then separately to make it possible to see the individual 83% CIs. Note that the Y-axis scale of the upper-left plot in (a) is different than those for the case where zero-breakage sites are not present. The curves in (b) are plotted on two different Y-axis scales.

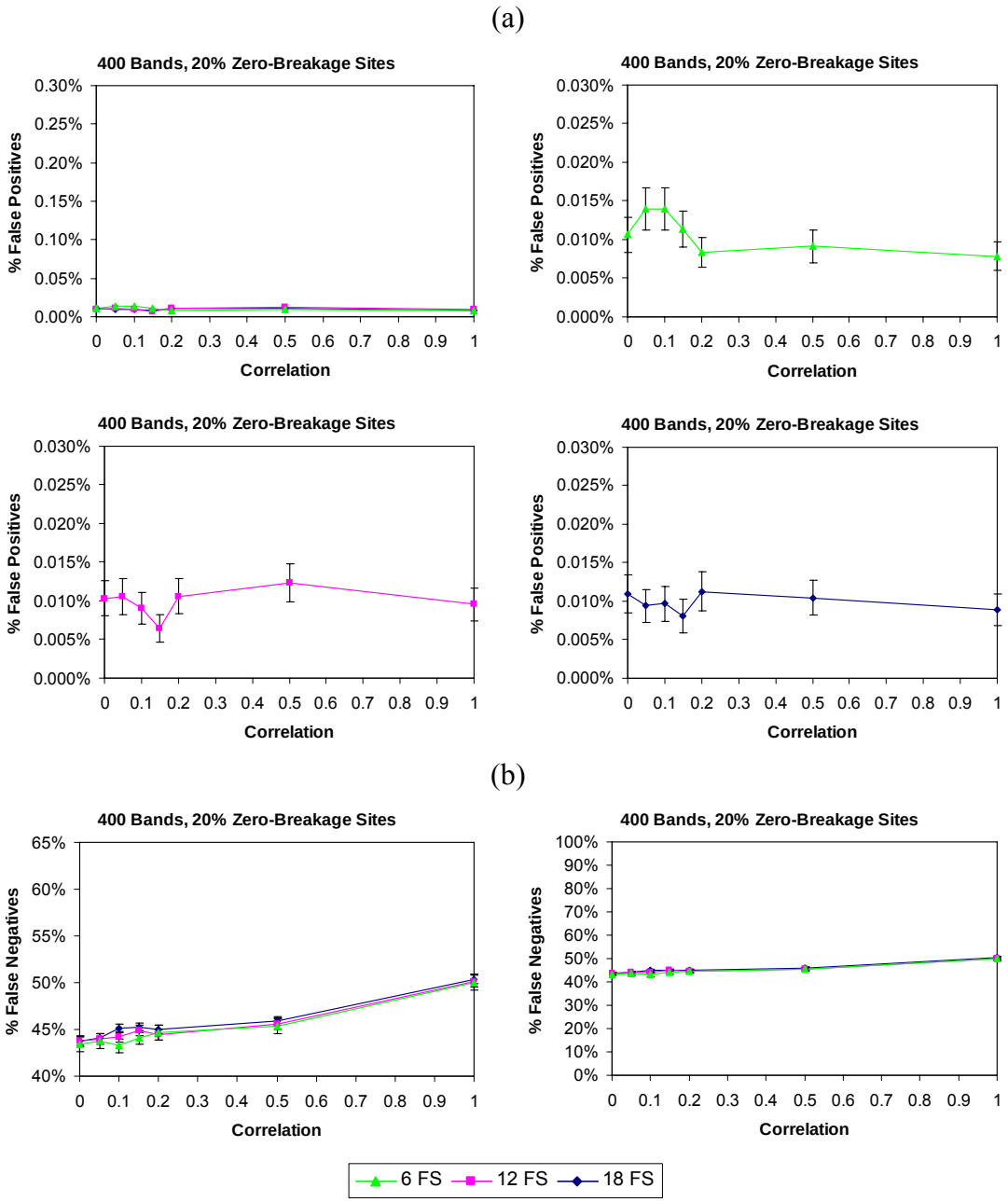


Figure C-12. FSM3 (a) False-Positive and (b) False-Negative Rates With 20% Zero-Breakage Sites for 400 Bands and 6 to 18 Fragile Sites With Breakage Probabilities of 0.022, 0.0264, 0.033, 0.0396, 0.044, and 0.055. The curves in (a) are first plotted together and then separately to make it possible to see the individual 83% confidence intervals. Note that the Y-axis scale of the upper-left plot in (a) is different than those for the case where zero-breakage sites are not present. The curves in (b) are plotted on two different Y-axis scales.

VITA

CHRISTOPHER JERRY HINTZE

329 N. 1000 E., Kaysville, UT 84037

Education

Aug 2005	Ph.D., Statistics, Texas A&M University Major Advisor: Dr. P. Fred Dahm.
May 2001	M.S., Microbiology, Brigham Young University
April 1999	B.S., Microbiology, Brigham Young University

Experience

2004-2005	Graduate Assistant, Teaching (Stat 301), Texas A&M University, Department of Statistics
2003-2004	Graduate Research Assistant, Texas A&M University, Department of Animal Science
2001-2003	Graduate Assistant, Non-Teaching, Texas A&M University, Department of Statistics
1999-2000	Graduate Teaching Assistant, Brigham Young University, Department of Microbiology
1998-2000	Graduate Research Assistant, Brigham Young University, Department of Microbiology

Awards

2002	Graduate Enhancement Fund Fellowship, Texas A&M University
2001	Regents' Fellowship, Association of Former Students Fellowship, College of Science Fellowship, and Graduate Enhancement Fund Fellowship, Texas A&M University

Presentations

1. "Increased Thymidine Kinase 1 Activity Correlates with a Decrease in Apoptosis Induced by Hyperthermia," (Poster) American Association for Cancer Research Annual Meeting, April 2000, and American Society of Microbiology Intermountain Branch Meeting, April 2000.
2. "Differential Overexpression of Bax in MCF-7 Cells Using the Thymidine Kinase 1 Promoter," (Talk) American Society of Microbiology Intermountain Branch Meeting, April 2000.