Decoding Human Cytomegalovirus DNA

Nathan Liittschwager[1], Arielle Thibeault[1], Katherine Yamamoto[1], Alec Guthrie[1], Alex Liebscher[1], and Mike Ona[1]

***Abstract*—**

I. INTRODUCTION

Human cytomegalovirus (CMV) is the leading betaherpesvirus to cause congenital birth defects in the Western world, affecting 1-2.5% of all live births (Stack & Stacey). Babies born with CMV may have brain, liver, spleen, or growth problems, with hearing loss being the most common symptom of congenital CMV contraction. CMV, however, may be contracted at any time during a person's life. While the method of transmission is unknown, it is suspected to be transmitted through bodily fluids such as saliva, sexual contact, or a mother's breast milk. After contraction, the virus lays dormant in a person's lymphocytes for the rest of their life until it goes through a period of reactivation, in which it reproduces and produces symptoms similar to that of mononucleosis (cdc.gov). The incidence of CMV varies globally, with 30% to 80% in most developed countries. In the United States, approximately 60% of people are infected by the age of six, and that number rises to approximately 85-90% of people by the age of 75-80 (Pawelec G, McElhaney JE, Aiello AE, Derhovanessian E, 2012).

II. DATA

A. Source

Our dataset, from Chee *et al.* (1990), contains 229,354 nucleotides, which were analyzed by Leung M *et al.* (2005) to extract complementary palindrome locations.

B. Data Summary

There are $N = 296$ palindromes reported with their locations. Locations are denoted by the index of the first nucleotide in the strand. On average, palindromes are 776 nucleotides apart. Palindromes shorter than 10 nucleotides were ignored and not reported. Random, uniform distribution of locations shows no significant mode or pattern (Fig 2), thus strengthening the suggestion that a deeper investigation into the strand should expose peculiarities within the CMV strand.

[1]University of California, Mathematics, San Diego, USA

III. BACKGROUND

A. Contextual Importance

Cytomegalovirus (CMV), belonging to the order Herpesvirales, is an obligate intracellular parasite found primarily in mammalian cells. Traditionally, once contracted, the virus is present in the host until host's termination. CMV may lie dormant or be expressed in the form of an infection, whose common symptoms include fever, sore throat, swollen glands, and fatigue (cdc.gov). Initial contraction may be undetected until the time of the infection, usually activated by a weakened immune system.

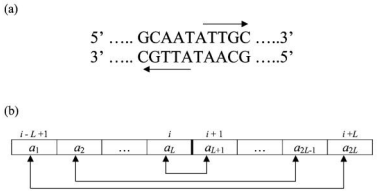
Like its Herpesvirales counterparts, the cytomegalovirus replicates lysogenically: infecting the host organism, integrating host DNA with the viral genome, and replicating through its assimilation in the genome of daughter cells. Since the DNA sequences are concatenated intermittently, it is unrealistic to manipulate the genome using the traditional methods scientists have used in the past, possible methods being reverse transcription or sythetic inhibition.

B. Approach

It is important to continue developing new methods that combat cytomegalovirus infections. A common approach to combating CMV is to prevent its replication by inhibiting the replication genome within infected host DNA. Once this genome is located effective drugs can be developed (Peter L *et al.*, 2008). Finding the location of this genome is of extreme importance. Thankfully, it is known that heavy clusters of complementary palindromes in DNA from the signify regions of the CMV replication genome (Leung M *et al.*, 2005). An example of such palindromes is provided in the figure below.

This report analyzes the complementary palindromes of a CMV infected DNA sequence in order to identify unusual clusters that may signify the location of CMV's replication genome. This report adopts the perspective that clustered palindromes will deviate from a random scatter best modeled by a Poisson process (Leung M et al., 2005). To test for this deviation, four statistical tests are conducted from different perspectives of the palindrome data.

Fig. 1: Visualization of a DNA complementary palindrome thanks to Leung M *et al.*, 2005. a) Shows that a complementary palindrome is a mirrored sequence on two strands of DNA. b) Shows that the location nucleotides defines the palindrome.



IV. Statistical Analysis

A. Question 1: Random Scatter Baseline

This report begins investigating the complementary palindrome clusters by observing how they compare to a baseline distribution. The baseline simulates a uniform random distribution which acts as a null hypothesis. If the original data appears to depart from the baseline, then there is good reason to follow up with more statistical investigations such as the χ^2 Goodness of Fit Test (See Theory) in order to determine if a null hypothesis should be rejected.

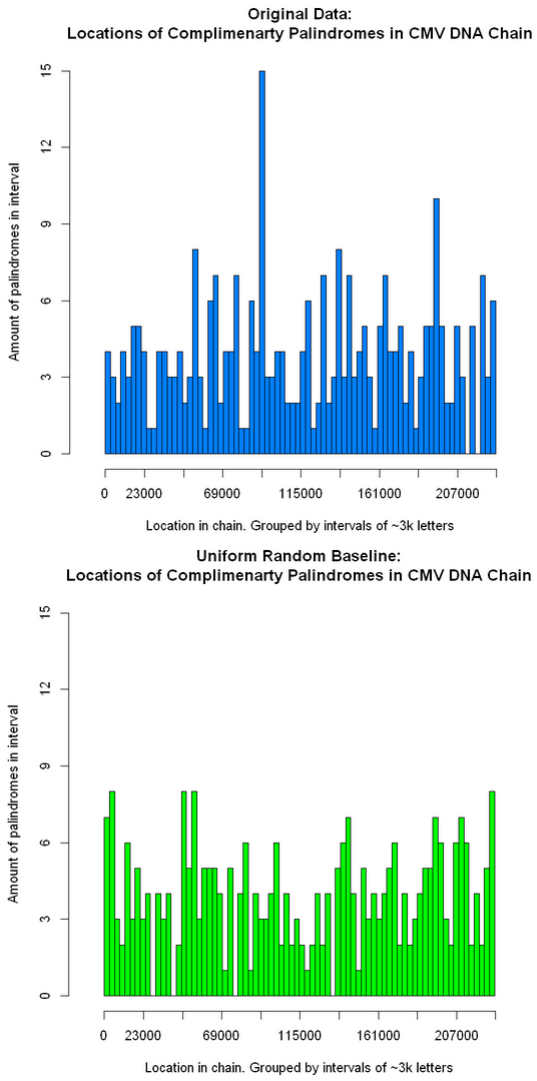
Two histograms are generated in Fig 2. The first shows the original 296 locations of palindromes. These locations are with respect to the entire CMV DNA chain that is 229,354 nucleotides long. Locations are grouped into intervals of 3,017.82 or roughly 3k nucleotides. There are 76 intervals. The second histogram shows the location of 296 hits from a uniform random distribution.

In observing the two histograms, it is difficult to see much difference between the original and simulated data. The original data does have one interval around nucleotide 92000 with a large number of palindromes, however this could be an outlier. Clearly more rigorous analysis is necessary to determine if there are any outliers in this data. The following three sections present this statistical analysis.

B. Question 2: Location and Spacing of Palindromes

The CMV DNA strand is partitioned into intervals of equivalent length in order to determine the location of palindromes, enabling us to locate any clusters in the strand. The locations of palindromes are expected to follow a uniform distribution. Thus, conducting a χ^2 Goodness of Fit test can help determine whether the observed frequency of palindromes differs from the theoretical frequency. A histogram comparing the expected and observed locations of palindromes can be found in Figure 3 along with a residual plot in Figure 4. Both show evidence of two unusual clusters of

Fig. 2: Histograms grouping the locations of complementary palindromes. Notice how the original data departs slightly from a uniform random baseline.



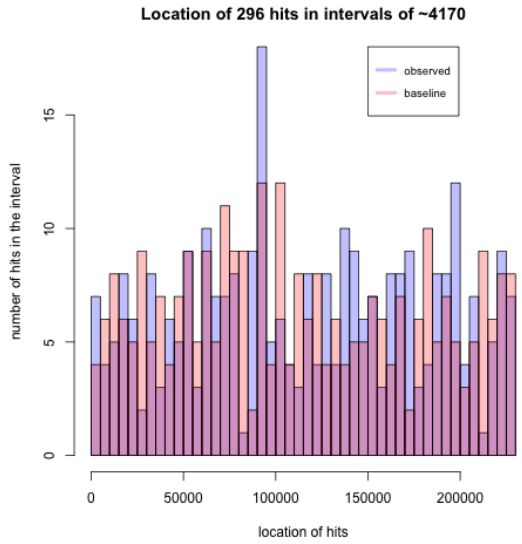
palindromes around the 90,000 and 195,000 base pairs of the CMV DNA sequence.

- Procedure:

The CMV DNA strand is split into 55 equal intervals of approximately 4170 base pairs and the number of palindromes in each interval is determined. The interval with the lowest number of palindromes is the second interval, base pairs 4340-8500, which contains only 1 palindrome. Alternatively, the 23rd interval, base pairs 91,700-95,800, contains the most with 14 palindromes located within the interval. The test statistic, χ^2, is calculated using the observed and expected frequencies of palindromes in each interval. As the locations are presumed to follow a uniform distribution, the expected frequency is 296/55 or 5.38 palindromes per interval and

thus $\chi^2 = 65.96$.

Fig. 3: Histogram comparing the observed and expected locations of 296 palindromes after partitioning CMV DNA sequence into 55 equal intervals of 4170.



- Null Hypothesis ($\alpha = 0.05$):

There does not exist a statistically significant difference between the observed and expected locations of palindromes.

- Results:

The resulting p-value from the χ^2 Goodness of Fit test with 54 degrees of freedom, 0.1275, is greater than the significance level of 0.05, so we fail to reject the null hypothesis. Thus, there does not exist statistically significant evidence that the locations of palindromes do not follow a uniform distribution which is an expected condition of the Poisson process. Furthermore, Table I demonstrates that we fail to reject the null hypothesis given different interval lengths.

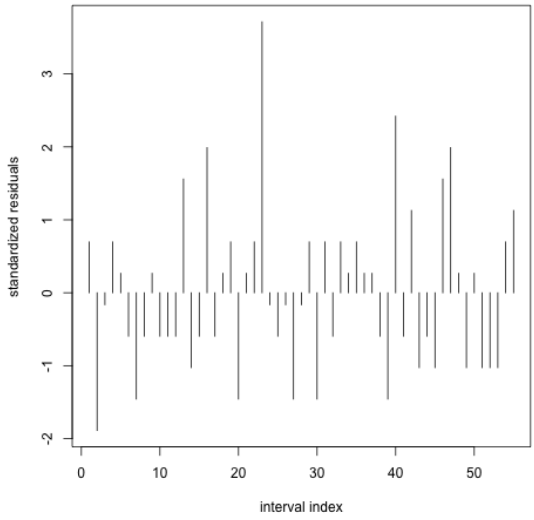
TABLE I: A table of the resulting p-values given the number of intervals. It can be observed that as the interval length increases and the number of intervals shrinks, the test statistic χ^2 decreases significantly while the p-value increases.

interval length	Subintervals	test statistic, χ^2	P-value
9174	25	18.70	0.7680
5734	40	33.19	0.7315
4170	55	72.15	0.1275

Goodness of Fit Testing Sequential Complementary Palindromes for Exponentiality

- Null Hypothesis

Fig. 4: Standardized residuals between the observed and expected locations of palindromes within each interval. Residuals greater than 3 or less than -3 indicate possible outliers that have a significant effect on the p-value.



The distances between sequential complementary palindromes locations follow an exponential distribution for α = 0.05.

- Results

Since the p-value from the χ^2 Goodness of Fit test with 19 degrees of freedom is less than the significance level of the test, we reject the null hypothesis and conclude that the distances between complementary palindrome locations are not exponentially distributed, under these conditions.

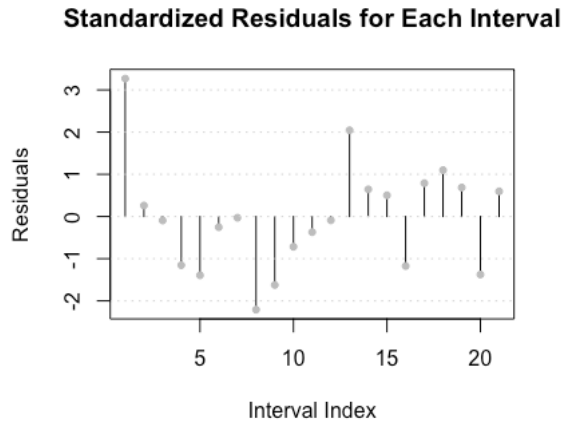
- Procedure:

For an observed phenomenon to be properly modeled by a Poisson Process, three conditions must be met over the entire domain: the locations of events must occur uniformly, the frequency of events must follow an exponential distribution, and the number of events must follow a Poisson distribution. To evaluate the frequency or the spatial distribution of the complementary palindrome locations for the CMV DNA segment, the difference was calculated from sequential data points to create a distance distribution (Fig. 6). To perform a χ^2 Goodness of Fit test for an exponential distribution, the maximum likelihood estimator for the parameter λ was calculated as 0.0013. Then, separating the locational data into 21 intervals, the χ^2 test was performed with 19 degrees of freedom, under the condition that the expected number of complementary palindromes, which yielded the p-value 0.0237. Note that the degrees of freedom for the test were adjusted to consider the estimated parameter $\hat{\lambda}$. The test was repeated with 9

degrees of freedom and produced similar results.

Goodness of Fit Testing the Sum of Sequential Pairs of Complementary Palindromes for the Gamma Distribution

Fig. 5: Plot of the standardized residuals to visualize the lack of fit for the exponential distribution. Note the variance of the distribution, residuals ranging from 0 to more than three units apart from the observed values.



- Null Hypothesis

The distances between the consecutive pairs of complementary palindrome locations follow a gamma(2,$\hat{\lambda}$) distribution for α = 0.05.

- Results

Since the p-value from the χ^2 test with 8 degrees of freedom is less than the significance level of the test, we reject the null hypothesis and conclude that the distances of sequential pairs of complementary palindrome locations do not follow a gamma(2,$\hat{\lambda}$) distribution, under these conditions.

- Procedure

Using the relationship that an exponential distribution is a special case of the gamma distribution when $k = 1$, it stands to reason that if the spacing of sequential complementary palindrome locations follows an exponential distribution, the distribution of the sum of consecutive pairs of complementary palindromes should follow a gamma distribution with $k = 2$. Note that since, the goodness of fit test for the exponential distribution rejected the null hypothesis, it is unlikely that the sum of consecutive pairs should follow a gamma distribution. A sample of 148 locational data points were extracted from the CMV DNA data for use in the analysis and the difference distribution was calculated. The maximum likelihood estimator for the parameter λ was calculated as 0.0006. The χ^2 test was performed with 8 degrees of freedom and yielded a p-value < 0.0001, as anticipated. The test was repeated on a disjoint subset of the data for the same degrees of freedom and reported similar results.

Goodness of Fit Testing the Sum of Sequential Triplets of Complementary Palindromes for the Gamma Distribution

- Null Hypothesis

The distances between the sequential triplets of complementary palindrome locations follow a gamma(3, $\hat{\lambda}$) distribution for α = 0.05.

- Results

Since the p-value from the χ^2 test with 5 degrees of freedom is less than the significance level of the test, we reject the null hypothesis and conclude that the distances of sequential triplets of complementary palindrome locations do not follow a gamma(3, $\hat{\lambda}$) distribution, under these conditions.

- Procedure

With similar reasoning to the prior test, if the spacing of consecutive complementary palindrome locations follows an exponential distribution, the distribution of the sum of sequential triplets should follow a gamma distribution with $k = 3$. Three samples of sizes 99, 99, and 98 elements were taken from the locational data and the difference distributions were calculated. For the first sample of 99 data points, the maximum likelihood estimator for λ was computed as 0.0004. The χ^2 test was performed on the sample with 5 degrees of freedom and produced a p-value < 0.0001. Since the p-value was arbitrarily close to zero, the test was not repeated on the additional samples.

C. Question 3: Counts of Palindromes in Intervals

We may partition the CMV DNA strand into nucleotide intervals of equivalent length 2,500 and observe the count of palindromes, k, in each interval. If we expect palindromes to be uniformly scattered throughout the DNA strand, we thus expect the probability of encountering an interval with k palindromes $P(X = k)$ to be modeled by the Poisson Distribution. Using a Maximum Likelihood Estimate, we deduce that the Poisson parameter λ is best estimated by $\hat{\lambda}$ which is equivalent to the expected value of the discrete observations (Fig 7). Should the observed values be estimated well by the Poisson Distribution, a χ^2 Goodness of Fit test should result in a p-value less than $\alpha = 0.05$.

Fig. 6: The distribution of the sequential complementary palindrome differences. The differences were calcuated from adjacent locations for each of the 296 complimentary palindromes for the CMV genome.

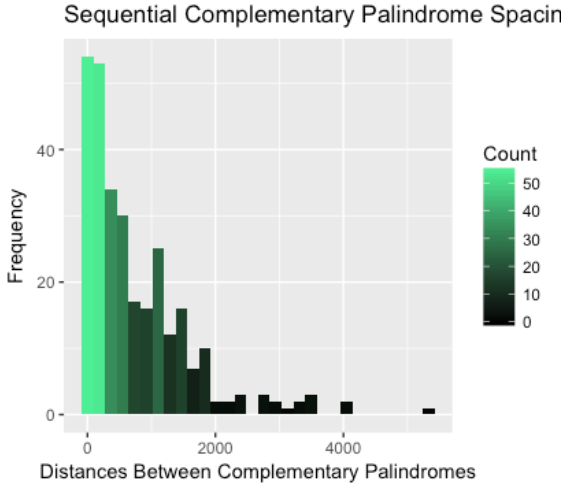
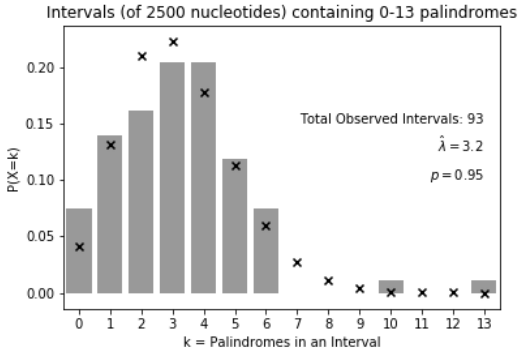


Fig. 7: The probability distribution of having an interval containing k palindromes, overlaid with an estimated Poisson Distribution ($\hat{\lambda} = 3.2$).



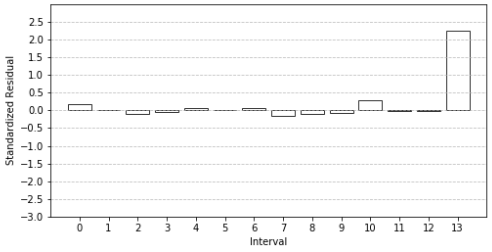
To assess the fit of the theoretical distribution, Fig 8 visualizes the standardized residuals, calculated by

$$\frac{N_i - \mu_i}{\sqrt{\mu_i}}$$

where N_i is the observed count and μ_i is the expected count for each interval. Residuals indicate poor model fit if the value is less than -3 or greater than 3. From the plot, it's clear that the 14th interval is far different from the other intervals.

We perform the χ^2 Goodness of Fit test to determine if the Poisson Distribution of complementary palindrome locations models the observed data well. The null hypothesis states that there is no preference for the number of palindromes in any interval. After performing the test, we do not reject the null hypothesis, $\chi^2(12, N = 93) = 5.19, p < 0.05$. Thus, we claim that there is no preference for how many palindromes are in any interval, despite the bias in the 14th interval. We thus assume the observed data is modeled by the Poisson Distribution.

Fig. 8: Normalized residuals from the observed and expected count of palindromes in each interval. Values less than -3 or greater than 3 indicate significant lack of fit.



D. Question 4: The Biggest Cluster

Since the number of palindromes in an interval are independent observations from a Poisson distribution under the Poisson process model, the interval with the most palindromes is the maximum of these independent random variables. The chance that k is the largest cluster size can be approximated with a given sample rate $\hat{\lambda}$. Hypothesis testing can determine if such a cluster size is unusual, providing evidence that the interval with the most palindromes is a potential site of replication.

- Procedure:

The CMV DNA strand is split into 76 intervals each of base pair length 3000. After observing the number of palindromes in each interval of equal length, the most palindromes in an interval comes out to be 15. The 296 palindrome locations divided by the number of subintervals equals the sample rate $\hat{\lambda}$. For 76 subintervals, $\hat{\lambda}$ = 3.89. We can then calculate the probability of 15 being the largest cluster size.

- Null Hypothesis ($\alpha = 0.05$):

There is no statistically significant evidence the cluster size k = 15 is larger than expected from the Poisson process.

- Results:

The resulting p-value of 0.0011 is less than the significance level of 0.05, allowing us to reject the null hypothesis. Thus, there is statistically significant evidence that the largest cluster size of 15 is larger than expected from the Poisson process. Table II provides similar results from intervals of other sizes.

TABLE II: A table of the resulting p-values given the number of subintervals.

Interval length	Subintervals	Max cluster size	P-value
1000	229	9	0.0020
3000	76	15	0.0011
6000	38	18	0.0445

E. Additional Hypothesis #1: Replication Study - Reconstruction of reported Confidence Intervals in Stockdale et al. using Bootstrap.

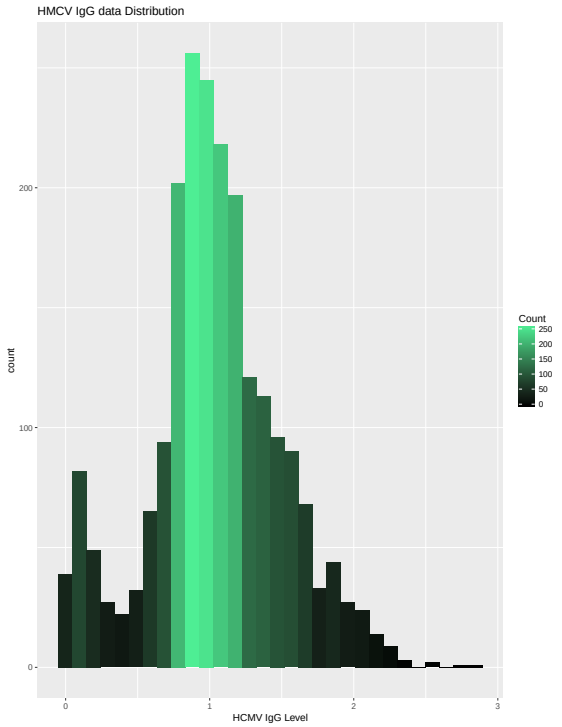
Before proceeding with the additional hypothesis, we cite the paper we are examining and provide a summary of the necessary background in order to follow the additional hypotheses. For a more detailed description of the study in question, please see the original publication, which will be cited in the Bibliography.

Background for Additional Hypotheses 1, 2, and 3: According to (Stockdale L, Nash S, Nalwoga A, Painter H, Fletcher H, Asiki G, Newton R, 2018), there is a dearth of information on the effects of CMV in low-income and developing countries, despite the fact that seropositivity is higher. In order to bridge this gap, some 2174 Ugandan Individuals (which includes fetuses within pregnant women) were sampled from the population. Cardiovascular biomarkers such as blood pressure, cholesterol, and d-lipids were taken when possible. For every indvidual, HMCV IgG antibody levels were measured and separated into tertiles of "high, medium, and low". It was noted if individuals were HIV positive or Tuberculosis (TB) positive. Upon collection of the data, various statistical tests were performed to describe the differences between groups and analyze the association of HCMV IgG levels with Cardiovascular Disease (CVD), HIV, and TB.

Of the statistical tests performed, of note was the Student's t-Test for a difference in means between independent groups, as well as a linear regression model fit to test the association of HCMV IgG levels and sex, tribe, TB positivity, and HIV positivity. Moreover, linear regression was used to rest association of HCMV IgG with continuous variables, such as cholesterol, HDL, LDL, BMI, and HbA1c.

Aim of Additional Hypothesis #1: In the paper given by Stockdale et al. the reported 99% confidence intervals for the means of HCMV IgG antibody levels between independent groups relied on normal approximation theory. While these confidence intervals give close to the nominal coverage under the Central Limit Theorem, they are sometimes sensitive to skew and departures from normality, which was present in

Fig. 9: Distribution of the HCMV IgG Antibody data. Note the extremely strong asymmetry, to the point of being bimodal and slight right tail skew. The data is decidedly not normal by inspection. Normal Approximation theory is not robust to skew and asymmetry, even with large sample sizes. We opt for a non-parametric approach.



the HIV positive group. Moreover, the HIV+ group had a relatively small sample size in comparison to the HIV negative group (n = 100 compared to m = 2074). As a supplement to the information provided by Stockdale et al., we provide 99% confidence intervals constructed by the percentile bootstrap as a non-parametric parallel to the normal approximation confidence intervals. We hope that the tabulation of the confidence intervals will provide a more robust estimation of the mean HCMV IgG antibody levels between groups. Moreover, since confidence intervals are equivalent to a hypothesis test, any disjoint 99% confidence intervals will indicate a statistically significant difference in means at the $\alpha = 0.01$ level.

Results: Table III tabulates the results of the bootstrap confidence interval construction. Notice that the confidence intervals between HIV negative and HIV positive groups are entirely disjoint, indicating a statistically significant difference between HIV negative and HIV positive groups. The female HIV positive cohort had the largest mean HCMV IgG level, while the male

(unadjusted) cohort had the lowest. moreover, females in general had higher levels of HCMV IgG antibodies, potentially indicating a higher degree of exposure. It remains to be seen whether this is by chance or a fundamental difference in human female physiology that makes them more susceptible to HCMV infection. Whether females have a statistically significant difference in HCMV IgG levels will be further analyzed in Additional Hypothesis #2.

TABLE III: The Percentile Bootstrap Confidence intervals for the mean HCMV IgG antibody levels in various independent Ugandan cohorts. Note that all the HCMV IgG levels are roughly similar, except for those HIV positive groups, which has a higher HCMV IgG level. The higher level of antibodies indicates a greater degree of exposure to HCMV. The HIV positive groups are immuno-compromised, which explains the greater degree of exposure to HCMV. It is unknown at present why females seemed to present higher levels of exposure than males, though it should be noted that females had a higher proportion of HIV positivity (62 to 38).

Group	Mean HCMV IgG	99% CI
All (unadjusted)	1.014	(1.012, 1.064)
All HIV +	1.544	(1.399, 1.680)
All HIV -	1.014	(0.989, 1.039)
Males (unadjusted)	1.009	(0.979, 1.043)
Females (unadjusted)	1.067	(1.030, 1.105)
Males (HIV +)	1.458	(1.235, 1.675)
Females (HIV +)	1.598	(1.400, 1.781)

F. Additional Hypothesis 2 - Reevaluating the use of Student's t-test in Stockdale et al.

Aim of Additional Hypothesis #2:

The Student's t-test, while robust to departures from normality, is an inappropriate test when equality of variances between groups cannot be assumed – in those cases, the Welch's t-test is preferred. Moreover, Student's (and Welch's) t-test is highly sensitive when there are large differences in sample sizes between independent groups. In the aforementioned paper, student's t-test was used to compare the mean levels of HCMV IgG antibody levels between sexes and between the HIV positive/negative groups. Since there is not convincing evidence that the variance between the groups to be compared is homogeneous (no statistical tests of variance was performed), we will repeat the comparison of means using Welch's t-test, which is the more appropriate test when equality of variances is not a reasonable assumption. Moreover, to account for the large differences in sample sizes between HIV positive ($n = 100$) and HIV negative groups($m = 2074$), a bootstrapped sub-sampling procedure will be enacted in order to approximate the p-value of the Welch's t-test. *Our aim is to see if the p-values from the more correctly applied Welch's t-test and bootstrapped t-test approaches that of the p-values reported in the paper.* Moreover, in Additional Hypothesis # 1, it was noted that females, on average, had higher levels of HCMV IgG levels. This difference will be examined more formally using Welch's t-test. It will also be seen if the majority of the observed differences in HCMV IgG antibody levels can be explained by the presence of immuno-compromised cohorts (those that presented HIV positive).

Results: Following Stockdale et al., the criteria for the rejection of the null hypothesis (no difference between groups) will be fixed at $p < 0.01$.

Welch's t-test was applied to compare the average HCMV IgG levels between women and men when the HIV group was included in the sample and also applied when the HIV group was not included. A statistically significant difference between groups was found when the HIV group was included in the sample, but *no* statistically significant difference was found when the HIV group was removed from the samples (Table IV). This implies that the majority of the difference between male and female HCMV IgG antibody levels was related to HIV-positivity.

TABLE IV: The mean difference of HCMV IgG antibody levels was compared between sex using the more appropriate Welch's t-test. Relatively similar confidence intervals of the observed difference were found when compared to Stockdale et al., but the acquired p-values were not as low as their results – ours are more conservative. In fact, upon controlling for HIV in men and women, no statistically significant difference was found between sex at the $\alpha = 0.01$ level. Please see Stockdale et al. for their reported p-values.

Welch's t-Test	HCMV IgG (mean diff.)	P-value	99% CI
Male vs. Female	0.058	0.003 (reject)	(0.007, 0.110)
Male vs. Female (HIV controlled)	0.042	0.029 (fail to reject)	(-0.008, 0.092)

Table V tabulates the difference between the HIV negative and HIV positive groups. A statistically significant difference between mean HCMV IgG antibody levels was found between HIV negative and HIV positive groups, with the p-value being approximately what was obtained by Stockdale et al. However, the given confidence interval is a result of bootstrap, as the disparity between sample size of the two groups would lead to bias within the normal approximation CI.

TABLE V: Result of the comparison between HIV positive groups and HIV negative groups. Because HIV positive had only a sample size of 100, and HIV negative consisted of 2074 individuals, a t-test is inappropriate without a bootstrapped subsample procedure. Despite this, the attained p-value remains close to Stockdale et al. However, the confidence interval obtained deviates from their paper. Stockdale et al. reported (0.40, 0.60) as their 99% CI.

Bootstrap Welch's Test	HCMV IgG (mean diff.)	P-value	99% CI (bootstrap)
HIV vs. no HIV	0.530	p<0.001	(.356, .713)

G. Additional Hypothesis #3: Difference in scale of HCMV IgG Distributions between Ugandan Cohorts.

Aim of Additional Hypothesis #3: It was shown in Additional Hypothesis #1 and Additional Hypothesis #2 that there was certainly a difference in *location* of the distributions of independent groups, such as males vs females, HIV+ vs HIV-, but may the distributions be treated as essentially the same aside from the shift in location? That is, does it make sense to think of the data as coming from a single distribution that is simply shifted in its center location? We answer this question by applying a bootstrap permutation test in order to determine the equality of distributions between two independent groups. That is, if F is the distribution of group 1, and G is the distribution of group 2, the null hypothesis of a permutation test is $F = G$ and the alternative hypothesis is $F \neq G$. *In particular, we examine the equality of distribution of HCVM IgG levels between sex, and HIV positive/negative groups*. Please see Methods and Theory for a full description of the test.

A difference in scale *and* location of two distributions implies that the data may not be treated as exchangeable under any circumstances, and that care needs to be taken when applying any tests that are not robust under departures from some assumed distribution. While the Welch's t-test used in Additional Hypothesis #2 is robust under the central limit theorem, statistical tests that do not share this trait will neccesarily be biased when comparing samples from different distributions. *Our aim is to answer the question: May we treat various independent groups as exchangeable?*

Results: The bootstrap permutation test was applied to compare Male vs Female when HIV was controlled, when it was not controlled, and when cohorts consisted of those with HIV positive status. Finally, HIV positive and HIV negative groups were compared. The test statistic used was $\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right)$ to reflect our interest in the shape of the distributions, as it's already known that center location differs.

Statistical significance was found in the difference of distribution between HIV positive and HIV negative groups. Statistical significance was also found between males and females only when HIV was not controlled. When HIV was controlled between sex, the p-value was small, but not enough to reject the null of $F = G$ at the $\alpha = 0.01$ level. Moreover, when comparing HIV positive males and HIV positive females, the test found absolutely no difference in distribution. Table V summarizes this.

TABLE VI: Results of the bootstrap test for equality of distribution. Note that at the $\alpha = 0.01$ significance level, the test rejected the null hypothesis that the distributions were equal when comparing HIV+ vs HIV-. It also rejected Male vs Female when HIV status was not controlled. The test statistic used was $\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right)$.

Group compared for equality of distribution F=G (HCMV IgG levels)	Ratio of $\frac{\sigma_1^2}{\sigma_2^2}$	Log -transformed Ratio	P-value ($\alpha = 0.01$)
1) Male vs Female	1.206	0.187	0.005
2) Male vs Female (HIV controlled)	1.150	0.140	0.039
3) Male vs Female (HIV positive)	1.113	0.107	0.710
4) HIV positive vs HIV negative	1.165	0.479	0.004

We can interpret the results of the bootstrap permutation test that the *majority of the observed differences in HCMV IgG levels between men and women are owed to HIV positivity*, and that HIV in conjunction with HCMV exposure affects men and women in approximately the same manner. With such strong evidence in the difference of distribution of HCMV IgG antibodies between HIV positive and HIV negative groups, we recommend that experimenters take care and control for any presence of HIV positivity in samples when examining the effects of HCMV on individuals, as the presence of HIV (and possibly any other immuno-suppressant disease) may lead to systemic bias in outcomes. However, HIV positive cohorts may be treated as exchangeable, regardless of whether they are male or female.

V. DISCUSSION

There are many improvements that can be made to this analysis in order to better guide biologists' search for the origin of replication in CMV DNA. For one, this analysis is dependent on the assumption that the Poisson Process is the appropriate model to use to determine the location of clusters of palindromes. Realistically, the Poisson Process may not be the most accurate model for analyzing DNA sequences. In fact, the analysis demonstrates that the CMV DNA sequence did not meet

all of the criteria of the process. Although the observed counts of palindromes followed a Poisson distribution and the observed locations of palindromes followed a uniform distribution, the spacing between palindromes did not follow an exponential distribution.

Lack of information may also have a confounding effect on the analysis. While the data set provides the locations of 296 palindromes in the CMV DNA sequence and specifies that they are between 10 and 18 base pairs long, the specific lengths of the individual palindromes are unknown. Furthermore, there are a few instances in the data where consecutive palindromes are separated by less than 10 base pairs meaning some palindromes may be overlapping thus affecting the location, spacing, and counts of the observed palindromes. For future analyses, it may be beneficial to either represent the location of the palindrome as the location of the midpoint or to determine the length of each palindrome to ensure that none overlap.

For the additional hypotheses, do to the relatively heavy skew and variance within the data set of Ugandan Cohorts, we recommend a non-parametric approach whenever possible and to be acutely aware of differing sample sizes, as many tests are not robust when there is a large disparity of sample sizes. However, normal approximation theory appears to hold relatively well when the sample sizes are sufficiently large. Nevertheless, there is no obvious advantage over non-parametric methods.

VI. CONCLUSIONS

The analysis demonstrated that the palindromes in the CMV DNA did not follow the Poisson process. The observed locations of the palindromes followed a uniform distribution and the counts of palindromes followed a Poisson distribution, both of which are characteristics of the Poisson process. On the other hand, the spacings between sequential palindromes were not exponentially distributed. Further analysis demonstrated that the spacing between sums of consecutive pairs and triplets of palindromes also failed to follow respective gamma distributions. Thus, the observed palindromes fail to meet all characteristics of the Poisson process. After breaking the CMV DNA sequence into intervals of 3000 base pairs and observing the number of palindromes in each interval, it was determined that an interval holds at most 15 palindromes. Hypothesis testing revealed that the cluster size $k = 15$ is larger than what is expected from the Poisson process.

It is recommended that biologists search for the interval with the **most palindromes** as the site of replication for CMV, under the assumption of a poisson process model. Our statistical results indicate that these sites are unusual under the poisson process model are worth further examination for a possible vaccine or cure.

In regard to the additional hypothesis, it was determined that Immuno-compromised individuals, in particular those who presented as HIV Positive, had higher degrees of exposure to HCMV as measured by the HCMV IgG antibody levels. Our non-parametric methods of analysis closely agreed with that of Stockdale et al. It was also shown that the majority of HCMV IgG antibody levels observed between cohorts was actually a function of HIV positivity. Furthermore, permutation tests indicated that HIV positive men and HIV positive women may be treated as approximately exchangeable, and that HIV positivity affect their HCMV IgG levels in approximately the same way. However, We do *not* recommend that experimenters treat immuno-compromised individuals as exchangeable with immuno-competent individuals when performing analysis or experimentation on cohorts that present as HCMV sero-positive. Instead, we recommend that Immuno-compromised cohorts be controlled for in any analysis or experimentation that is intended to generalize to the whole population.

VII. METHODS AND THEORY

In this section, the various procedures and theoretical tools used in this analysis will be stated and summarized. While there is insufficient space to provide fully rigorous proofs, some proofs might be sketched, and the reader will be directed to the appropriate source material.

Maximum Likelihood Estimation. Consider a random vector $(X_1, ..., X_n)$ from the joint probability density function $f(X_1, ..., X_n|\theta)$, where θ is some unknown parameter. Suppose that each $X_i, 1 \leq i \leq n$ is independent and that for each i, $X_i = x_i$, where x_i is the *observed* value of X_i. Then, the likelihood function of θ is given by

$$L(\theta) = f(x_1, x_2, ..., x_n|\theta) \tag{1}$$

$$= \prod_{i=1}^{n} f(X_i|\theta) \tag{2}$$

where the second equality follows because each X_i is independent and identically distributed. Then, to estimate θ, the likelihood function $L(\theta)$ must be maximized, which is equivalent to the maximization of a monotonic function. Hence, the *log likelihood* function is given by

$$\log L(\theta) = \sum_{i=1}^{n} \log[f(X_i|\theta)]. \tag{3}$$

Thus, to produce estimators for λ in regards to the exponential and Poisson distributions, simply replace $f(x_1, ..., x_n|\theta)$ with $f(x_1, ..., x_n|\lambda)$ so that $f(x_1, ..., x_n|\lambda)$ is the appropriate (mass) density function for the (Poisson) exponential distributions, and maximize the corresponding likelihood functions and solve for λ. Please see *Mathematical Statistics and Data*

Analysis, (Rice, 2007) for a more detailed description of theory and procedure. †

Method of Moments. The kth moment of a probability law is defined as

$$\mu_k = E(X^k) \tag{4}$$

where X is a random variable of that probability law. If $X_1, ..., X_n$ are i.i.d. random variables from that distribution, the kth *sample moment* is defined as

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \tag{5}$$

which may be seen as an estimator for μ_k. Thus, if there are k unknown parameters, $\theta_i, 1 \leq i \leq k$ may be expressed in terms of the k moments. That is, we may represent each unknown parameter θ_i as the system of equations:

$$\begin{aligned} \theta_1 &= f_1(\mu_1, ..., \mu_k) \\ &\vdots \\ \theta_k &= f_k(\mu_1, ..., \mu_k) \end{aligned}$$

but since $m\hat{u}_i$ is an estimator for μ_i for each $1 \leq i \leq k$, we may estimate θ_i with $\hat{\theta}$ given by the following system

$$\begin{aligned} \hat{\theta}_1 &= f_1(\hat{\mu}_1, ..., \hat{\mu}_k) \\ &\vdots \\ \hat{\theta_k} &= f_k(\hat{\mu}, ..., \hat{\mu}_k). \end{aligned}$$

Hence, solving the preceding system of equations for each $\hat{\theta}_i, 1 \leq i \leq k$ gives an estimator for each of the k unknown θ's. Thus, applying the method of moments to the gamma distribution allows estimation of its shape and scale parameters. (Rice, 2007). †

The Poisson Process. A stochastic process $\{N(t), t \geq 0\}$ is said to be a *counting process* if $N(t)$ represents the total number of "events" that have occurred up to time t. Hence, the counting process $\{N(t), t \geq 0\}$ is said to be a *Poisson Process having rate λ with $\lambda > 0$* if:

(i) $N(0) = 0$ (i.e., no events have occurred at time $t = 0$).
(ii) The process has independent increments – the number of events that have occurred by time t is independent of the events that have occurred between time t and $t + s$. That is, $N(t)$ is independent of the increment $N(t + s) - N(t)$.
(iii) The number of events in any interval of length t is Poisson distributed with mean λt. That is, for all $s, t \geq 0$,

$$P(N(t + s) - N(s) = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!},$$

where $n = 0, 1, 2, ...$ It follows from condition (iii) that the Poisson process has stationary increments, and that $E[N(t)] = \lambda t$.

By *stationary increments*, we mean that the distribution of the number of events that occur in any interval of time depends only on the length of the time interval. †

The Waiting Time Distribution. Consider a Poisson process $\{N(t), t \geq 0\}$, and let X_1 denote the time of the first event. For $n \geq 1$, let X_n denote the time between the $(n - 1)st$ event and the nth event. The sequence $(X_n)_{n=1}^{\infty}$ is called the *sequence of interarrival times*. Define S_n to be the waiting time of the nth event generated by a Poisson Process.Then, for $n \geq 1$

$$S_n = \sum_{i=1}^{n} X_i.$$

It easily follows that S_n has a *gamma distribution* with paramaters n and λ by observing that the nth event occurs prior or at a time t if and only if the number of events occuring by time t is at least n. That is,

$$N(t) \geq n \iff S_n \leq t.$$

For a complete proof, we direct the interested reader to *Stochastic Processes* by Sheldon M. Ross, 1983. †

Distribution of the Arrival Times. It will be shown, informally, that the distribution of arrival times of a Poisson process is uniformly distributed.

Consider a Poisson process $\{N(t), t \geq 0\}$ and suppose for some time $t > 0$, there is an interval of time, $[0, t]$ with a single event in it. Since a Poisson process has stationary and independent increments, the time of the event is uniformly distributed over $[0, t]$. To see this, note that for $s \leq t$,

$$
\begin{aligned}
&P(X_1 < s | N(t) = 1) \\
&= \frac{P(X_1 < 2, N(t) = 1)}{P(N(t) = 1)} \\
&= \frac{P(\textit{1 event in [0,s), 0 events in [s,t)})}{P(N(t) = 1)} \\
&= \frac{P(\text{1event in [0,s)})P(\text{0 events in [s,t)})}{P(N(t) = 1)} \\
&= \frac{\lambda s e^{-\lambda s} e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} \\
&= \frac{s}{t}.
\end{aligned}
$$

Since s is the time until the first event, over the length of time t, it follows that the distribution for the arrival time of an event in an interval $[0, t]$ is uniformly distributed. This result is easily generalized to n random variables, but requires the notion of Order Statistics, which is beyond the scope of this paper. The result is simply stated instead.

Given that $N(t) = n$, the n arrival times of $S_1, ..., S_n$ have the same distribution as the order statistics corresponding to the n independent random variables uniformly distributed on the interval (0,t).

For a full proof, we again direct the reader to *Stochastic Processes* by Ross. †

It is the view of preceding theoretical discussion that validates our choice of hypothesis tests regarding the Poisson, Exponential/Gamma and Uniform distributions in this paper.

The χ^2 Goodness of Fit Test. Let O_i be the number of observations in the ith interval, $1 \leq i \leq k$. Let E_i be the expected counts in the ith interval. Then, the Pearson χ^2 test statistic is

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

This will give some real valued number. Then, depending on α, compute $\chi^2_{\bar{v}}(1-\alpha)$, and compare T and $\chi^2_{\bar{v}}(1-\alpha)$. We reject the null hypothesis if

$$T \geq \chi^2_{\bar{v}}(1-\alpha)$$

where $\bar{v}$ is the degrees of freedom. In our case, $\bar{v} = k-2$, where k is the number of intervals we partitioned our data into. †

Hypothesis Test for Max Palindromes. Under the Poisson process model, the numbers of hits in a set of non-overlapping intervals of the same length are independent observations from a Poisson distribution. This implies that the greatest number of hits in a collection of intervals behaves as the maximum of independent Poisson random variables. If we suppose that there are m such intervals then

$$
\begin{aligned}
&P(\text{maximum count over m intervals} \geq k) \\
&= 1 - P(\text{maximum count over m intervals} < k) \\
&= 1 - P(\text{all interval counts} < k) \\
&= 1 - P(\text{first interval counts} < k)^m \\
&= 1 - \left[\lambda^0 e^{-\lambda} + \cdots + \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}\right]^m
\end{aligned}
$$

For a given estimate of λ, from the above expression, we ca find the approximate chance that the greatest number of hits is at least k. If this chance is unusually small, then it provides evidence for a cluster that is larger than the expected from the Poisson process. We can use the maximum palindrome counts as a test statistic, and the computation above provides the p-value for the test statistic. †

Welch's Two Sample t-test. Suppose that $X_1, ..., X_n$ are independent and normally distributed random variables with mean μ_X and variance σ_X^2 and that $Y_1, ..., Y_m$ are independent and normally distributed random variables with mean μ_Y and variance σ_Y^2 and that Y_i are independent of X_i. Then the statistics,

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \tag{6}$$

follows an approximate t distribution with v degrees of freedom, where v is

$$v = \frac{[s_X^2/n + s_Y^2/m]^2}{\frac{(s_X^2/n)^2}{n} + \frac{(s_Y^2/m)^2}{m}} - 2 \tag{7}$$

rounded to the nearest integer. †

Bootstrap Algorithm. The bootstrap is a simulation method that draws **with** replacement from the sample distribution in order to make inferences about the population. Suppose that $(x_1, ..., x_n)$ is a random vector representing data drawn from the distribution F. Define $\hat{F}$ to be the empirical distribution of $(x_1, ..., x_n)$ putting probability $1/n$ of drawing the data point $x_i, i = 1, ..., n$. We define the **bootstrap sample** to be the vector $\mathbf{x}^* = (x_1^*, ..., x_n^*)$, where each of the $x_j^*, j = 1, ..., n$ is drawn with replacement from $\hat{F}$, hence x_i^* need not equal x_i.

Let $s(\mathbf{x})$ be a test statistic, in particular, an estimator $\hat{\theta}$. Corresponding to the bootstrap sample $\mathbf{x}^*$ is the bootstrap replicate $\hat{\theta^*} = s(\mathbf{x}^*)$ so that $\hat{\theta^*}$ is the test statistic $s(\cdot)$ applied to the bootstrap sample. This processes is repeated many times, say B times times, and the resulting B bootstrap replicates is the collected. From there, the standard error of the estimator $\hat{\theta}$ may be approximated using the bootstrap replicates in order to construct confidence intervals and conduct hypothesis tests.

In order to conduct a hypothesis test, first note the observed test statistic $s(\mathbf{x})_{obs}$, then manipulate $\hat{F}$ so that it is the empirical distribution *under the null hypothesis* H_0. Then B replicates of the bootstrapped test statistic $s(\mathbf{x}^*)$ are computed and an approximate p-value is calculated by

$$p \approx \frac{1 + \sum_{i=1}^{B} I(s_i(\mathbf{x}^*) \geq s(\mathbf{x})_{obs})}{(B+1)} \tag{8}$$

Where I is the indicator function. †

Bootstrap Confidence Intervals. Let $\hat{\theta}(\mathbf{x})$ be an estimator function for an unknown parameter θ. By applying the bootstrapping algorithm described below, draw B samples of size n, where n is the size of the original data, and pass the B samples through the estimator $\hat{\theta}(\mathbf{x})$ in order to produce B bootstrap replicates of the estimator. That is, produce $\theta_b^*, b = 1, ..., B$ bootstrap replicates via passing the simulates through $\hat{\theta}$. Then, the approximate $100\%(1-\alpha)$ confidence interval for θ is given by

$$\left[\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2}\right]$$

where $\theta^*_{\alpha/2}$ is the $\alpha/2$ percentile of the B bootstrap replicates θ_b^*. This is equivalent to taking the $\alpha/2$ quantiles of the B bootstrap replicates if they were ordered from least to greatest. This procedure is fully described in (Efron, 2000). †

Subsampling Bootstrap Welch's t-test. If $\mathbf{x}$ consists of n observations and $\mathbf{y}$ consists of m observations, we do not resample with n and m bootstrap samples, instead we *subsample* from $\hat{F}$ and $\hat{G}$ while fixing the number of subsamples. In this case, choosing $n_s = m_s = \min n, m$, where n_s, m_s are the number of bootstrap subsamples from $\hat{F}$ and $\hat{G}$, respectively. Then, replace $s(\cdot)$ with $p(\cdot) = P(t(\mathbf{x}) \geq t_{\alpha/2}) + P(t(\mathbf{x}) \leq -t_{1-\alpha/2})$, where $t(\mathbf{x}, \mathbf{y})$ is given in (1) and where t_α is such that $P(T \leq t_\alpha) = 1 - \alpha$ with the degrees of freedom given in (2). Fixing at the $\alpha > 0$ significance level, our approximate p-value is calculated as being

$$p^* \approx \frac{1}{B} \sum_{i}^{B} p_i$$

where p_i is the p-value obtained from the ith bootstrap replicate. Under the central limit theorem and by the bootstrap principle, $p^* \to p$ as $B \to \infty$. For more information, see (Hinkley, 2006). †

Non-Parametric Permutation Test S. uppose having observed $X_1, ..., X_n \sim f_X(x)$ and $Y_1, ..., Y_m \sim f_Y(y)$, we wish to test the null hypothesis $F = G$, where F and G are the cumulative distribution functions of $f_X(x)$ and $f_Y(y)$ respectively. The equality $F = G$ means that F and G assign equal probabilities to all sets of data generated by the sampling process. If H_0 is true, there is no difference in the probabilistic behavior of random variables X_i or Y_j. Suppose that N makes up the concatenated size $n + m$. Then since the first sample has n data points, there are $\binom{N}{n}$ possible permutations of resampling from the concatenated sample of N size, if we choose n points at random. That is, if $\mathbf{g}$ is a vector made of of n resamples from the concatenated sample, we have the following lemma:

Permutation Lemma. Under $H_0 : F = G$, the vector $\mathbf{g}$ has probability $\frac{1}{\binom{N}{n}}$ of equaling any one of its possible values.

Thus, if $\hat{\theta} = S(\mathbf{g}, \mathbf{v})$ is the observed test statistic, we have $\mathbf{g} = (X_1, ..., X_n)$, and $\mathbf{v} = (Y_1, ..., Y_m)$. Then, let $\mathbf{g}^*$ indicate any one of the possible $\binom{N}{n}$ possible vectors of $X_i's$, and $Y_j's$, we may define the *permutation replication* of $\hat{\theta}$ by

$$\hat{\theta}^* = \hat{\theta}(\mathbf{g}^* = S(\mathbf{g}^*, \mathbf{v}) \tag{9}$$

and we may calculate a p-value for the test $F = G$ with

$$P_{perm} = \#\{\hat{\theta}^* \geq \hat{\theta} / \binom{N}{n}. \tag{10}$$

Since it is not possible to calculate all possible permutations when the sample size becomes large (as factorials grow extremely fast), we apply the bootstrap principle. The bootstrap algorithm is the same as what we have been using, with a minor difference. Instead of drawing n and m resamples separately from the two independent groups, the observed sample is concatenated into a pooled sample of $n + m$ size. Then, $n + m$ replications are drawn from the pooled sample. The first n are formed into vector g^* and the last m are formed into

vector $\mathbf{v}^*$. Upon passing the vectors through the statistic $\hat{\theta}$, the bootstrap replicate $\hat{\theta}^*(\mathbf{g}^*, \mathbf{v}^*)$ is created. Repeat this process B times. Then, the approximate p-value is obtained by

$$p^* = \frac{\#\hat{\theta}^*(b) \geq \hat{\theta} + 1}{B + 1} \tag{11}$$

where $\hat{\theta}^*(b)$ is the bth bootstrap replicate, $1 \leq b \leq B$.

Then, if $p^* < \alpha$, we reject the null hypothesis that $F = G$ and conclude that the probabilistic behavior of each distribution is not the same. For a more detailed treatment, see (Efron, 2000) and (Hinkley, 2006). †

VIII. ACKNOWLEDGMENTS

REFERENCES

[1] Centers for Disease Control and Prevention. *Cytomegalovirus*. https://www.cdc.gov/cmv/overview.html. Accessed May 2, 2018.

[2] Chee, M. S., Bankier, A. T., Beck, S., Bohni, R., Brown, C. M., Cerny, R., ... & Preddie, E. (1990). Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. In *Cytomegaloviruses* (pp. 125-169). Springer, Berlin, Heidelberg.

[3] Davison, A. C., and D. V. Hinkley. Bootstrap Methods and Their Application. *Cambridge University Press*, 2006.

[4] Efron, B. An Introduction to the Bootstrap. *Boca Raton, Florida*, 2000.

[5] Jiwa, N. M., Van, G. G., Raap, A. K., de Rijke Van, F. M., Mulder, A., Lens, P. F., ... & der Ploeg Van, M. (1989). Rapid detection of human cytomegalovirus DNA in peripheral blood leukocytes of viremic transplant recipients by the polymerase chain reaction. *Transplantation, 48*(1), 72-76.

[6] Lischka, P., Zimmermann, H. (2008). Antiviral strategies to combat cytomegalovirus infections in transplant recipients, *Current Opinion in Pharmacology, 8*(5), 541-548.

[7] Leung M-Y, Choi KP, Xia A, Chen LHY. (2005). Nonrandom Clusters of Palindromes in Herpesvirus Genomes. *Journal of computational biology,12*(3), 331-354.

[8] Pawelec, G., McElhaney, J. E., Aiello, A. E., & Derhovanessian, E. (2012). The impact of CMV infection on survival in older humans. *Current Opinion in Immunology, 4*(4), 507-511.

[9] Stack, Gabrielle, and Maria Stacey. *Human Cytomegalovirus (HCMV)*. British Society for Immunology, *Human Cytomegalovirus (HCMV)*.

[10] Stockdale L, Nash S, Nalwoga A, Painter H, Asiki G, et al. (2018) Human cytomegalovirus epidemiology and relationship to tuberculosis and cardiovascular disease risk factors in a rural Ugandan cohort.

IX. APPENDIX