Detecting Replication Patterns in HCMV through Palindrome Analysis

11-10-24

Contribution:

The development of this report was a collaborative effort between both students. Each student took primary responsibility for specific sections while maintaining active involvement throughout the entire project. Student 1 focused mainly on sections 1 and 2 of the analysis section while also making substantial contributions to the advanced analysis, introduction, and conclusion. Additionally, Student 1 reviewed sections 3, 4, 5. Student 2 concentrated on reviewing sections 1 and 2, while working on sections 3, 4, 5 in the analysis section. Student 2 also looked over Student 1's sections and contributed to the advanced analysis sections. Both students were responsible for writing R code, methods, and conclusions for their respective sections. Throughout the project, both students consistently reviewed each other's work and made iterative improvements to ensure the report's quality and coherence.

Use of GPT: The report utilized ChatGPT to enhance the clarity and readability of our analysis, specifically in refining the language of method and conclusion sections to be more concise and professional. While all underlying analysis, code development, and interpretations were conducted independently by the students, GPT served as a writing aid to improve the presentation of our findings and maintain consistency in technical writing style throughout the report.

1. Introduction

DNA analysis in viral genomes has become increasingly crucial for understanding viral replication and developing targeted treatments. In particular, the study of palindromic sequences in viral DNA offers valuable insights into potential replication origins, which are essential for viral reproduction. Research shows that in the herpes virus family, palindromic sequences often mark these critical replication sites, with patterns varying from long single palindromes to clusters of shorter ones. Understanding these patterns is vital for developing antiviral strategies and advancing our knowledge of viral biology. This analysis focuses on the Human Cytomegalovirus (CMV), a member of the herpes virus family that affects 30-80% of populations globally. While relatively simple in structure - consisting of DNA wrapped in a protein shell called a capsid - CMV contains a complex genome of 229,354 base pairs. Within this genome, 296 palindromic sequences between 10 and 18 base pairs long have been identified, potentially marking important regulatory sites including origins of replication. The main goal of this analysis is to investigate whether clusters of palindromes in the CMV genome could indicate potential origins of replication, ultimately aiding biologists in their experimental search. Our analysis will proceed through several steps: First, we'll examine the random scatter of palindrome locations through computer simulation to establish a baseline for comparison. We'll then analyze the spacing between palindromes and their clustering patterns using various statistical methods. Through count analysis across different regions of the DNA, we'll identify significant deviations from random distribution. Finally, we'll investigate the characteristics of the largest palindrome clusters to determine their potential biological significance.

Data

The data comes from the published DNA sequence of CMV (Chee et al., 1990) and subsequent pattern analysis by Leung et al. (1991). The primary dataset consists of 296 palindrome locations within the 229,354-base-pair genome, with palindromes ranging from 10 to 18 letters in length. Shorter palindromes were excluded from the analysis. The data is contained in the file hcmv.txt, which lists the DNA positions of these palindromes. Notably, the longest palindromes (18 letters) were found at positions 14719, 75812, 90763, and 173893 along the sequence. This dataset provides a comprehensive foundation for investigating potential patterns and clusters that might indicate origins of viral replication.

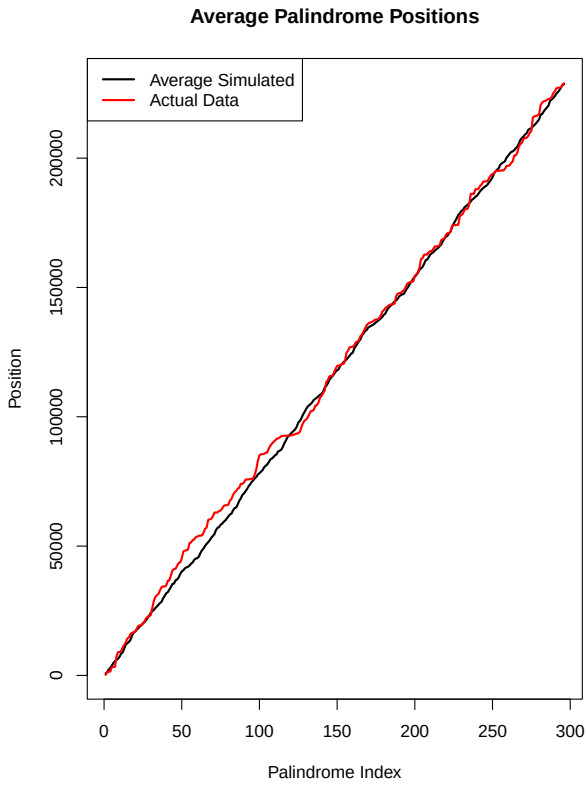
2. Basic Analysis

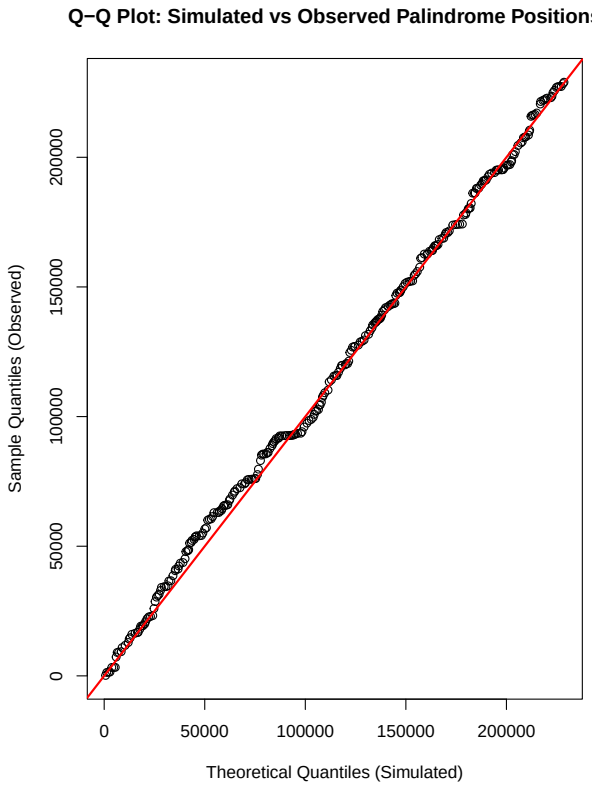
2.1 Simulating Random Scatter of Palindrome Locations

Method

To analyze palindrome distribution in the HCMV genome, we conducted a simulation involving random placement of palindrome sites. We randomly selected 296 palindrome sites along a DNA sequence of 229,354 bases and repeated this process 1000 times to generate distinct simulated distributions. The average of these simulations was then calculated to establish a baseline for expected random scatters. Finally, we compared the average simulated distribution with the actual positions of palindromes in the HCMV genome to identify any deviations from randomness for qualitative and quantitative comparisons. For quantitative comparisons, we calculated summary statistics for both observed and simulated distributions, providing an overview of their characteristics. We then generated a Q-Q plot to visually compare observed and expected distributions, where alignment along the diagonal suggested similarity and deviations indicated differences. A Kolmogorov-Smirnov test provided statistical validation of any significant deviations from randomness.

Analysis






```
## Kolmogorov-Smirnov Test Results:
## D:  0.04054054 ; P-value: 0.9681301
```

Table 1: Summary Statistics: Observed vs Simulated Palindrome Positions

Distribution	Min	First.Quartile	Median	Mean	Third.Quartile	Max	SD
Observed	177.0	63714.0	117826.0	116960.1	171143.5	228953	64732.03
Simulated	712.8	58031.4	116952.4	115196.2	171055.8	228620	65929.09

Conclusion

Our analysis indicates that palindrome sites in the HCMV genome are distributed in a pattern consistent with random placement, supported by several findings. First, the Kolmogorov-Smirnov test yielded a non-significant p-value of 0.9681 (D = 0.040541), suggesting no significant difference between observed and simulated distributions. Second, the Q-Q plot shows points closely following the diagonal reference line, indicating strong alignment between observed and expected distributions. Third, summary statistics reveal similar distribution characteristics, with comparable means (116,960.1 vs 115,196.2), medians (17,826.0 vs 16,952.4), and standard deviations (64,732.03 vs 65,929.09) between observed and simulated data.

While the overall distribution appears random, a closer examination of local spacing and clustering patterns reveals subtle deviations where the actual data (red line) diverges slightly from the simulated average (black line) in specific regions. Although these variations are not statistically significant at the genome-wide level, they suggest possible local influences on palindrome positioning that could reflect biologically relevant patterns, warranting further investigation. Exploring these local patterns within an otherwise random framework may uncover nuanced biological significance in certain regions of the genome which will be covered further in our report.

2.2 Graphical Analyzese of Palindrome Patterns: Locations and Spacings

Method

To analyze palindrome distribution patterns in the HCMV genome more in depth, we implemented three approaches: spacing analysis, pair sum analysis, and triplet sum analysis. In the spacing analysis, we calculated distances between consecutive palindromes, while pair and triplet analyses involved computing sums of adjacent pairs and triplets of palindrome distances. For each approach, we visualized the distributions through histograms, comparing them to expected uniform random distributions from the 1000 random scatter distribution.

Analysis

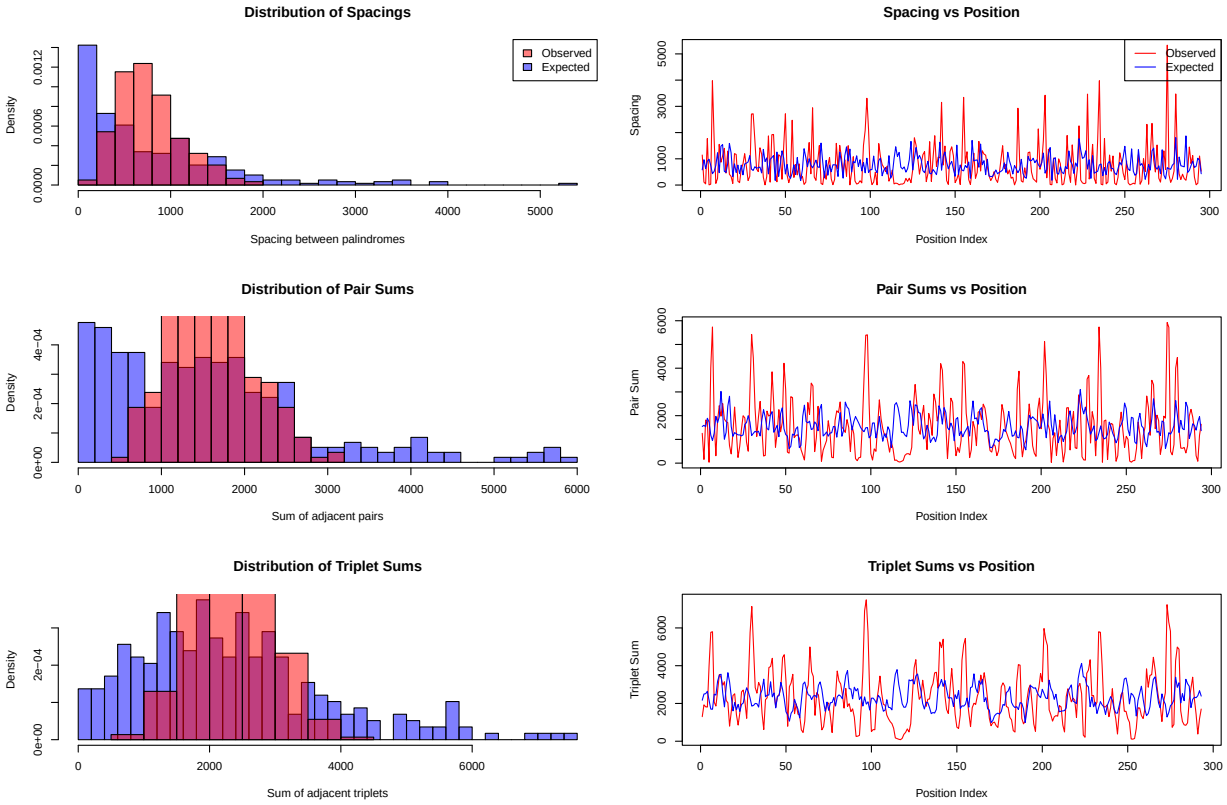


Table 2: Kolmogorov-Smirnov Test Results

Pattern Type	D-statistic	p-value
Spacings	0.3491525	4.815e-16
Pair Sums	0.3061224	2.167e-12
Triplet Sums	0.2662116	1.919e-09

Conclusion

Our analysis of palindrome distributions in the HCMV genome reveals significant deviations from random expectations across multiple patterns.The KS tests indicate highly significant differences between observed and expected distributions for spacings, pair sums, and triplet sums, suggesting a non-random organization of palindrome sequences. The spacing distribution shows more extreme peaks than the expected, with a higher frequency of both small and large spacings, implying clustering of palindromes in certain regions. Additionally, the pair and triplet sum distributions exhibit greater variability, furthering the observations made by the spacing position, with observed patterns displaying more pronounced peaks and valleys. Furthermore, the histograms show different distributions, with the observed data exhibiting a more right-skewed distribution.These findings imply that, while palindrome positions may initially appear random, their non-random distribution likely reflects underlying biological constraints or functional roles within the HCMV genome, potentially linked to processes like DNA replication, transcription regulation, or genome organization.

2.3 Counting Palindromes Across DNA Regions

Method

To examine palindrome distribution within the HCMV genome, we divided the genome into regions of progressively smaller sizes, ranging from 4,587 bp to 287 bp. For each region size, we calculated the observed count of palindromic sequences and compared this to an expected count based on a random distribution. We then used chi-square tests to determine if there was a statistically significant deviation from randomness in palindrome distribution across these region sizes. The chi-square statistic and corresponding p-value were recorded for each scale, allowing us to identify at which scales palindrome clustering emerged.

Analysis

Table 3: Analysis for ~4,587 bp regions (50 total regions)

	Palindrome Count	Observed Regions	Expected Regions	Chi-Square Component
0	0	1	0.134	5.582
1	1	2	0.795	1.827
2	2	1	2.353	0.778
3	3	4	4.643	0.089
4	4	8	6.871	0.185
5	5	8	8.135	0.002
6	6	5	8.027	1.141
	>=7	21	19.042	0.201

Table 4: Analysis for ~2,294 bp regions (100 total regions)

	Palindrome Count	Observed Regions	Expected Regions	Chi-Square Component
0	0	10	5.182	4.480
1	1	11	15.338	1.227
2	2	23	22.701	0.004
3	3	22	22.398	0.007
4	4	16	16.575	0.020
5	5	11	9.812	0.144
6	6	4	4.841	0.146
	>=7	3	3.153	0.007

Table 5: Overall Summary

# of Regions	Region Size	Number of Regions	Chi-square Statistic	p-value
50	4587 bp	50	9.8069	0.1964
100	2294 bp	100	6.0351	0.5337
200	1147 bp	200	28.9068	0.0065
300	765 bp	300	64.0229	0.0010
400	573 bp	400	242.4703	0.0005
500	459 bp	500	271.7521	0.0015
600	382 bp	600	1963.5150	0.0005
700	328 bp	700	310.3984	0.0010
800	287 bp	800	548.2308	0.0005
900	255 bp	900	887.8797	0.0005

Conclusion

Our analysis reveals a striking, scale-dependent pattern of palindrome clustering in the HCMV genome. At larger region sizes, such as 4,587 bp (50 regions) and 2,294 bp (100 regions), chi-square tests showed non-significant results, suggesting that palindrome distribution is random at broad genomic scales. However, as we moved to smaller regions, a transition point was observed at 1,147 bp (200 regions), where the chi-square test first indicated a significant deviation from randomness ($p = 0.0065$). This finding suggests that clustering begins to emerge around this scale. At even finer scales, below 1,000 bp, chi-square values increased sharply, indicating strong local clustering; for instance, at 382 bp (600 regions), the chi-square value rose to 1963.52, demonstrating intense clustering.

This scale-dependent clustering pattern implies that palindromic sequences may have functional or structural roles at specific genomic scales, which supports our previous analysis in 2.2. The randomness observed at broader scales suggests that these local patterns are distributed across the genome rather than concentrated in large regions, while the strong clustering at finer scales may reflect localized DNA structural or regulatory functions. These insights provide a foundation for future studies of genome organization in HCMV, particularly at scales below 1,000 bp where non-random distribution is most evident and may hold biological significance.

2.4 Identifying the Largest Palindrome Cluster

Method

We investigated potential origins of replication in the HCMV genome by analyzing palindrome clustering through a structured, three-part approach. First, we divided the 229,354 bp genome into 1,000 bp non-overlapping intervals, as justified in Section 2.3, to capture biologically relevant cluster sizes. We counted palindromes in each interval to identify regions with high palindrome density. Using a custom function, compute_max_hits, we then located the interval with the highest palindrome count. To confirm statistical significance, utilize the 1000 simulated random scatter from 2.1. By comparing the observed maximum palindrome count with this simulated distribution, we calculated an empirical p-value. Finally, within the identified cluster, we analyzed spacing patterns between palindromes, including mean, median, and range, to characterize the cluster's structure and significance.

Analysis

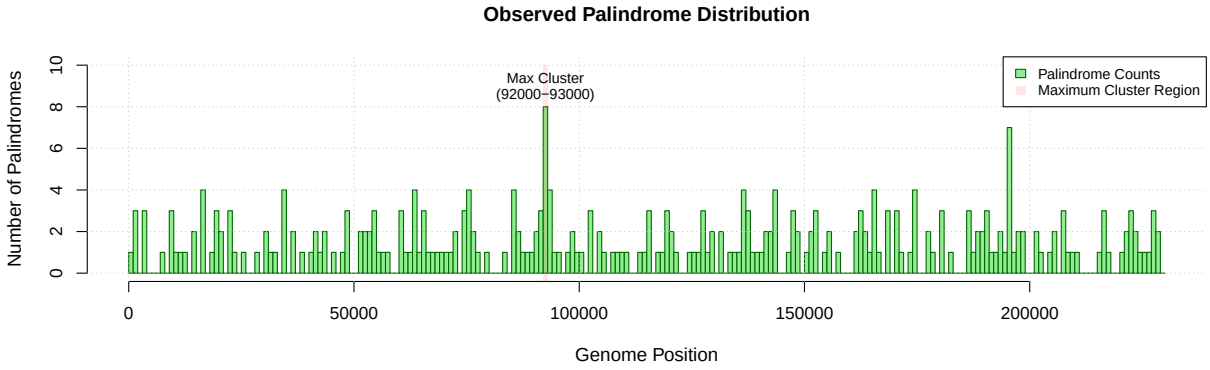


Table 6: Largest Cluster Analysis Statistics

Metric	Value
Number of Palindromes in Cluster	8.00
Mean Spacing in Cluster	47.57
Median Spacing in Cluster	44.00
Minimum Spacing in Cluster	8.00
Maximum Spacing in Cluster	76.00

Table 7: Statistical Significance

Metric	Value
Observed Maximum	8 palindromes
Location Interval	92000 - 93000
Percentile in Distribution	100.0th percentile
Simulation Mean (SD)	5.80 (0.45)
Statistical Significance	Significant ($p < 0.0000$)

Conclusion

Our analysis identifies a promising origin of replication (ORI) candidate in the HCMV genome between positions 92,000 and 93,000. This region contains 8 palindromes within a 1,000 bp interval, a highly significant clustering ($p < 0.0001$).The palindromes exhibit a consistent spacing pattern, with a mean interval of 47.57 bp and a median of 44.00 bp, within a range of 8-76 bp. Such regular spacing is typical of binding sites for protein complexes, suggesting that this cluster could serve as an assembly site, further supporting its potential role as an ORI.Given the strong statistical significance and structured clustering, this region aligns well with known viral replication origins, offering a compelling target for replication studies in HCMV.

2.5 Advising Biologists on Experimental Searches

Method

To develop a comprehensive search strategy for laboratory researchers, we synthesized our statistical findings into practical recommendations focusing on key genomic regions and defining optimal search parameters.

Analysis

Technical Report: HCMV Replication Origin Search Strategy

Our statistical investigation identifies the region at positions 92,000-93,000 bp as the prime candidate for origin of replication, supported by multiple significant indicators: 8 palindromes within 1,000 bp, regular spacing patterns (mean 47.57 bp, range 8-76 bp), and statistically significant clustering ($p < 0.0001$). The consistency and strength of these patterns strongly suggest biological significance rather than random occurrence. For systematic genome scanning, we recommend using 1,000 bp windows, which provides optimal balance between sensitivity and specificity - large enough to capture meaningful patterns but small enough to avoid missing significant clusters. Regions containing 6 or more palindromes warrant immediate investigation, as this threshold represents significant deviation from random distribution. The spacing between palindromes serves as another critical indicator, with 40-50 bp intervals being characteristic of significant clusters. Secondary search targets should exhibit similar characteristics: high palindrome density approaching 8 per 1,000 bp, regular spacing patterns, and statistically significant deviation from random arrangement. These features have proven highly effective in distinguishing potentially functional clusters from random occurrences.

For experimental efficiency, we suggest a hierarchical approach: begin with intensive investigation of the primary target region (92,000-93,000 bp), followed by systematic screening of secondary targets that meet the established criteria. This approach ensures optimal use of laboratory resources while maximizing the likelihood of identifying functional replication origins.

Conclusion

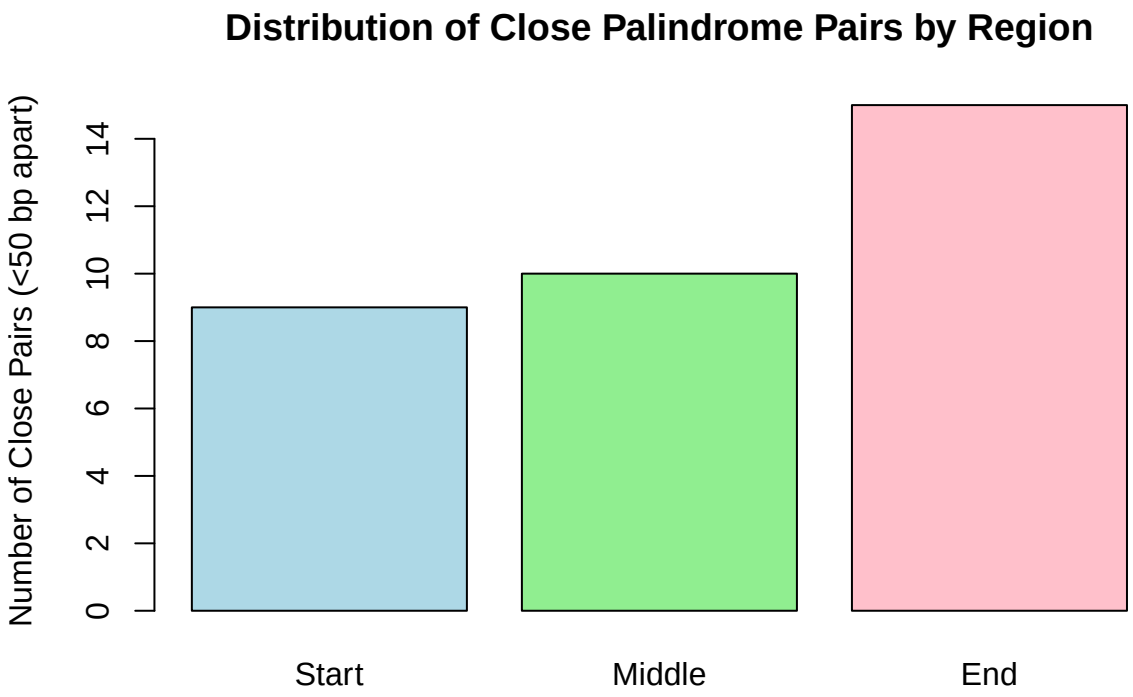
Based on our rigorous statistical analysis, we strongly recommend initiating experimental investigation at the 92,000-93,000 bp region. This area's exceptional characteristics - high palindrome density, regular spacing, and strong statistical significance - make it the most promising candidate for a replication origin. The prescribed window size of 1,000 bp and focus on regions with 6+ palindromes will optimize detection efficiency while minimizing false positives. Following these evidence-based parameters should significantly streamline the experimental search process, directing laboratory resources toward the most promising genomic regions for identifying HCMV replication origins. The combination of specific target regions, clear search parameters, and prioritized approach offers the most efficient path to identifying functional replication origins in the HCMV genome.

3. Advanced Analysis

Method

To investigate whether palindromes tend to appear in pairs, we analyzed the spacing between consecutive palindromes. We defined "close pairs" as palindromes separated by less than 50 bp (approximately the mean spacing we found in significant clusters) and counted how many such pairs exist in different regions of the genome.

Analysis



Conclusion

Our analysis of closely spaced palindrome pairs reveals an unexpected distribution pattern across the genome. We found that the number of close pairs increases from the beginning to the end of the genome. This pattern suggests that while palindromes appear throughout the genome, they are more likely to occur in close proximity to each other near the genome's end. This finding adds an important dimension to our understanding of palindrome organization in the HCMV genome. By examining the spacing between palindromes, we've identified that these sequences can be arranged in different ways: either as closely spaced pairs or as part of larger groupings. The higher frequency of close pairs in the End region suggests that palindromes might serve different biological functions depending on their arrangement and location. This could be particularly valuable for researchers, as it indicates that both the position and spacing of palindromes should be considered when studying.

4. Conclusion

Our statistical analysis employed multiple complementary approaches to investigate whether palindrome clusters in HCMV DNA represent potential replication sites. Random scatter simulation of 296 palindrome sites revealed significant structured arrangements, particularly in palindrome spacings (mean 47.57 bp in significant clusters), that deviated notably from random distribution patterns ($p < 0.0001$). Analysis across different region sizes identified 1,000 bp as the optimal detection window, balancing the risks of missing tight clusters against splitting significant regions. Count analysis across regions revealed strong evidence of non-random clustering at scales below 1,000 bp, particularly in regions with regular spacing patterns between palindromes.

The most compelling evidence emerged from the 92,000-93,000 bp region, containing 8 palindromes with regular spacing patterns that strongly suggest biological relevance rather than chance occurrence. This region's characteristics - high palindrome density, consistent spacing, and statistical significance - align well with known features of viral replication origins from other herpes viruses. Count analysis across surrounding regions further confirmed this cluster's unique properties, showing it represents a significant deviation from background palindrome distribution patterns. Our advanced analysis of closely spaced palindrome pairs revealed a complementary pattern, with more pairs occurring towards the end of the genome (15 pairs) compared to the start (9 pairs) and middle (10 pairs) regions, suggesting that palindromes might serve different biological functions depending on their arrangement and location.

Based on these comprehensive findings, we recommend biologists focus initial experimental testing on the 92,000-93,000 bp region. Use 1,000 bp windows for systematic screening, prioritizing regions containing 6 or more palindromes with regular spacing (40-50 bp intervals). When expanding the search, target regions showing similar characteristics: high palindrome density approaching 8 per 1,000 bp and consistent spacing patterns. This evidence-based approach should significantly increase the efficiency of experimental investigation.

Several limitations warrant consideration, including our focus on palindromes 10 bp and the need for experimental validation. Additionally, while our window size optimization balanced various factors, biological replication origins might not perfectly align with these statistical parameters. Future research should examine shorter palindrome sequences, investigate cluster interactions, and explore evolutionary conservation across herpes viruses. Integration of additional genomic features, such as DNA unwinding elements and protein binding sites, could further illuminate the biological significance of identified palindrome patterns and potentially lead to more effective antiviral strategies.