Article

Statistical Distributions of Genome Assemblies Reveal Random Effects in Ancient Viral DNA Reconstructions

Fernando Antoneli , Cristina M. Peter , Marcelo R.S. Briones *

Posted Date: 24 December 2024

doi: 10.20944/preprints202412.2065.v1

Keywords: Ancient DNA; genome assembly; ancient viruses; statistical distributions; power laws; log-normal laws

Statistical Distributions of Genome Assemblies Reveal Random Effects in Ancient Viral DNA Reconstructions

Fernando Antoneli, Cristina M. Peter and Marcelo R. S. Briones *

Center for Medical Bioinformatics, Escola Paulista de Medicina, Federal University of São Paulo (UNIFESP), São Paulo, SP 04039-032, Brazil
*Correspondence: marcelo.briones@unifesp.br.

Abstract: Ancient human viruses have been detected in ancient DNA (aDNA) samples ranging from Anatomically Modern Humans to Neanderthals. Reconstructing genomes from aDNA using reference mapping presents numerous problems due to the unique nature of ancient samples, their degraded state, smaller read sizes and limitations of current methodologies. Spurious alignments of reads to reference sequences (mapping) are a main source of false positives in aDNA assemblies and the assessment of signal-to-noise ratios is essential to differentiate *bona fide* reconstructions from random, noisy, assemblies. Here we analyzed the statistical distributions of viral genome assemblies, ancient and modern, and their respective random "mock" controls used to evaluate the signal-to-noise ratio. We tested if differences between real and random assemblies could be detected from their statistical distributions. Our analysis shows that the coverage distributions of: (1) real viral aDNA assemblies of adenovirus (ADV), herpesvirus (HSV) and papillomavirus (HPV) do not follow power laws nor log-normal laws, (ADV) and control aDNA assemblies are well approximated by log-normal laws, (3) negative control parvovirus B19 (real and random) follow a power law with infinite variance and (4) the mapDamage negative control with non-ancient DNA (modern ADV) and the mapDamage positive control (human mtDNA) are well approximated by the negative binomial distribution, consistent with the Lander-Waterman model. Our results show that the tails of the distributions of aDNA and their controls reveal the weight of random effects and can differentiate spurious assemblies, or false positives, from *bona fide* assemblies.

Keywords: Ancient DNA; genome assembly; ancient viruses; statistical distributions; power laws; log-normal laws

1. Introduction

The field of paleovirology research relies on detection of viral genomes embedded in DNA and raw sequencing data of its hosts. Because of smaller genomes and scarcity of integrated copies, sequence reads of these pathogens tend to be smaller than the average sequence reads of the hosts. Genome remnants of ancient viruses have been detected in ancient DNA (aDNA) samples ranging from the Middle Ages to the Paleolithic [1,2]. Reconstructing aDNA using genome mapping presents numerous challenges due to the unique nature of ancient samples, their degraded state, and limitations of current sequencing methodologies. These artifacts might produce spurious alignments in aDNA genome assemblies with even greater weight than in modern DNA assemblies.

Spurious alignments in genome assemblies occur when sequences are incorrectly aligned to the reference genome due to various technical or biological factors [3,4]. These misalignments can lead to errors in genome annotation, variant calling, or downstream analyses. Common causes and contexts for spurious alignments are: (1) repetitive sequences, such as highly repetitive regions (e.g., transposable elements, satellite DNA) can cause reads to align to multiple loci, leading to ambiguous or incorrect placements, (2) paralogous regions, or sequences that are similar due to gene duplication events (paralogs) can align to incorrect paralogous loci instead of their true origin, (3) low-complexity regions such as regions with simple sequence repeats (e.g., homopolymers, di-/tri-nucleotide repeats) often cause misalignments because they lack unique sequence context, (4) errors introduced during sequencing, such as substitutions, insertions, or deletions, that can distort the sequence and lead to incorrect alignments, (5) poor reference quality, such as incomplete or inaccurate reference genomes can result in reads aligning to incorrect locations or being mapped to scaffold gaps, (6) cross-species contamination, when reads originating from contaminant DNA (e.g., symbionts, pathogens, or laboratory contamination) may spuriously align to the closest matching sequences in the reference genome and (7) inversions, translocations, or structural variants, when large structural rearrangements can mislead mapping algorithms, causing reads from one genomic context to align to a different one [5].

False positives in variant calling can be caused by spurious alignments, when misalignments create the appearance of SNPs, indels, or structural variants that are not truly present in the sample [6]. Also, misannotation of genes due to incorrect alignment of reads leads to errors in gene prediction or expression quantification and assembly gaps and chimeric contigs can be produced by misplaced reads that contribute to assembly errors, such as artificial contigs or scaffolds [7]. Minimization of artifacts caused by spurious alignments can be obtained by improvements in mapping algorithms, masking repetitive elements, filtering of low-quality reads, stringent parameters, alternative reference genomes and post-mapping quality control. In cases of extreme complexity, performing *de novo* assembly can help reconstruct genomic regions without reliance on a reference genome and *de novo* assembly [8].

In the case of ancient DNA, the challenge of genome assembly is even greater. Reconstructing aDNA using genome mapping presents numerous problems due to the unique nature of ancient samples, their degraded state, and limitations of current methodologies [3]. The main problems with aDNA are: (1) DNA degradation by fragmentation, often into short pieces (~30-100 base pairs), making it difficult to map accurately to the reference genome, (2) chemical damage such as cytosine deamination causing C-to-T or G-to-A transitions, particularly at fragment ends, introducing errors in alignments and variant calling and (3) low complexity, where some degraded regions lose complexity and are difficult to align uniquely [9].

All mainstream methods of DNA sequencing rely on reading fragments of DNA (reads), that are usually much smaller than the genome to be sequenced and assembled by mapping to a reference. The common abstraction to these methods is that of a mathematical covering problem. In 1988, Lander and Waterman published a study examining the covering problem which is still used as a guideline to estimate the desired sequencing coverage [10]. In the Lander-Waterman model, the basic statistical assumption is that reads are generated uniformly, at random, from the genome, known as the homogeneity assumption. In the homogeneous model the coverage of each base pair follows a *Poisson distribution*. This distribution, however, imposes a severe restriction because it excludes the possibility of overdispersed coverage distributions.

When heterogeneity is considered, the coverage of each base pair follows a Poisson mixture with a latent distribution belonging to the *gamma distribution* family. Then, the number of reads covering a base pair follows a *Poisson-gamma distribution*, also known as a *negative binomial distribution*. This is a family of distributions parameterized by two positive real numbers (r, μ), where r is the dispersion parameter and μ is the mean value. When $\mu/(\mu+r)$ tends to 0 and r tends to infinity, in such a way that μ tends to a fixed limit $\mu 0$, the negative binomial distribution approximates a Poisson distribution with rate $\mu 0$. Therefore, the negative binomial distribution is the simplest generalization of the Poisson distribution that allows for over-dispersion. Finally, it is important to note that the Poisson distribution and the resulting negative binomial distribution are light-tailed distributions, that is, far from *power laws* [11]. Based on these considerations it can be proposed that the problem of quality in genome assemblies by mapping to a reference can be, at least in part, examined from the perspective of the distributions of the reads mapped to the reference. It seems that the parameters of these distributions can be affected by the randomness caused by spurious mapping of reads (or the other problems affecting genome assemblies, as discussed above). The comparison and analysis of distributions, and their properties, might reveal the level of randomness and assess the quality of genome assemblies.

Accordingly, here we analyzed the statistical distributions of viral genome assemblies, ancient and modern, and their respective random "mock" controls as defined previously to evaluate the signal-to-noise in aDNA assemblies [2]. We conclude that the tails of the distributions of aDNA and their controls reveal the weight of random effects in assemblies and can differentiate false positive assemblies, caused by spurious alignments, from *bona fide* aDNA genome assemblies.

2. Material and Methods

2.1. Theoretical Background

2.1.1. Distributions with Heavy Tails

The univariate distributions can be divided into two classes: the *heavy-tailed* and the *light-tailed*. The *heavy-tailed distributions* are characterized by the property that their tails decay more slowly than exponentially. The *light-tailed distributions* are characterized by the property that their tails decay at an exponential rate, or faster. This distinction is important in modeling real world phenomena because a heavy-tailed distribution (such as Cauchy distribution, a power law) has a greater probability of rare events (larger deviations from the mean) than a light-tailed distribution (such as a Gaussian and exponential) (**Figure 1**) [12,13]. This means that the heavier the tail the larger the random effects.

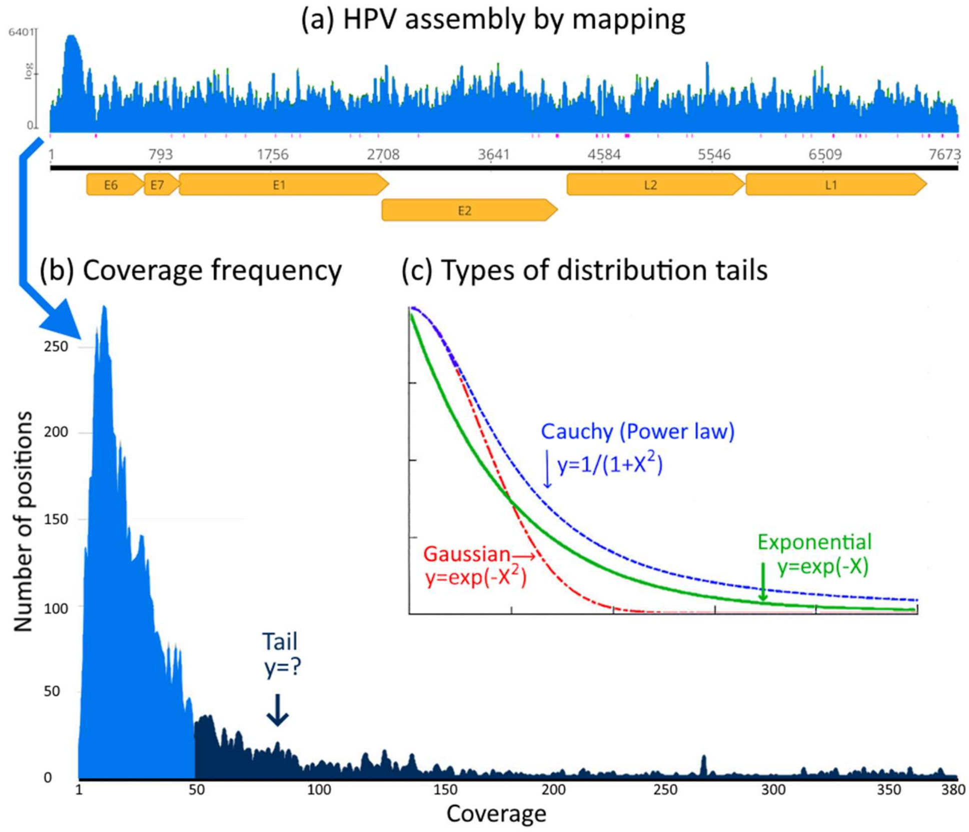

Figure 1. Distribution tails of genome assemblies. In (a) the coverage (number of reads per position) along the HPV genome (ref.). In (b) the coverage distribution (number of positions with given coverage). In (c) the different types of distribution tails show the difference between Gaussian (light-tailed), exponential (light-tailed) and a power law (Cauchy) a heavy-tailed distribution. Heavier tails indicate more random effects.

doi:10.20944/preprints202412.2065.v1

The exponential distribution, the gamma distribution and the normal (or Gaussian) distribution are examples of light-tailed distributions. The log-normal, the Pareto distribution and the Cauchy distribution are examples of heavy-tailed distributions.

The class of heavy-tailed distributions is quite vast and general which makes it difficult to work with. Therefore, many different narrower and more tractable subclasses of heavy-tailed distributions have been introduced. The two most important such subclasses are the *sub-exponential distributions* and the *regularly varying distributions*. The term “fat-tailed” in the literature does not have any rigorous definition. Depending on the research community, the terms fat-tailed and heavy-tailed are synonymous or that the fat-tailed is a subset of heavy-tailed. Here, we consider “fat-tailed” as a synonym of *regularly varying* [14,15].

The large majority of commonly used heavy-tailed distributions are, in fact, sub-exponential, including the log-normal and the Pareto distributions. However, what distinguishes these two examples is that the Pareto distribution is *regularly varying* whereas the log-normal is not. The Pareto distribution is an example of a continuous *power law probability distribution*, that is, it describes a quantity whose probability density decreases as a power of its magnitude. Power laws are the distributions with the heaviest tails and have the important property of *scale invariance*.

For the practical purpose of determining if a given real-world empirical distribution is a power law, there is an advantage in considering not only the pure power laws, but their “perturbations”, as well. The class of regularly varying distributions is very convenient to work with because it not only contains the “pure power laws”, such as the Pareto distributions, but is much larger. Particularly, it contains all the distributions that deviate from pure power laws by means of a *slowly varying function*, that is, a function that varies slowly at infinity, classic examples including functions converging to constants or powers of logarithmic functions. This definition allows the distribution to deviate from a pure power law arbitrarily but without affecting the power-law tail exponent [14].

Complex stochastic processes driving evolution of many different real-world phenomena can hardly produce perfect power-law dependencies without any deviation from a pure power law [16]. Searches for pure power law dependencies in real-world data revealed that this distribution is exceedingly rare [17]. Therefore, it is important to consider the full class of regularly varying distributions instead of the pure power laws.

2.1.2. Power Law Estimation

Proper estimation of the tail exponent under the assumption that a given empirical distribution is a regularly varying distribution is a hard problem. This problem has attracted extensive attention in probability, statistics, physics, engineering, and finance, where a variety of estimators have been developed for this task, all based on Extreme Value Theory [18].

We adopted the method of Voitalov et al. which consists of 3 estimators: Adjusted Hill (H), Moments (M) and Kernel (K) [19]. These are, currently, the only existing estimators that satisfy the following criteria: (1) are applicable to any regularly varying distribution, (2) are statistically consistent, i.e., have been proven to converge to the true tail exponent, if applied to increasing length sequences sampled from any regularly varying distribution and (3) can be fully automated by the means of the ‘double bootstrap method’ that has been proven to yield the optimal estimation of the tail exponent for any finite sequence of numbers sampled from any regularly varying distribution.

It is important to stress that based on any given finite sample, there is absolutely no way to tell how likely the hypothesis is that it was sampled from a regularly varying distribution. In view of this impossibility, the best strategy is to simply rely on the estimates of the Adjusted Hill (H), Moments (M) and Kernel (K) estimators.

If the estimator results are all positive, for a given sample, then it might be the case that the empirical distribution comes from a regularly varying distribution. Yet if these estimates are negative or close to zero, then the chances of that are vanishingly small. However, there is no, and cannot be any, rigorous way to quantify these chances, using hypothesis testing or any other methods. In view of these considerations, Voitalov et al. [19] take the conservative approach, and propose the following definition of an empirical power law distribution, based on the values of the 3 estimators above: (1)

an empirical distribution is 'Not Power Law' (NPL) if at least one estimator returns a negative or zero value, (2) an empirical distribution is 'Hardly Power Law' (HPL) if all the estimators return positive values, and if at least one estimator returns a value ≤1/4, (3) an empirical distribution is 'Power Law' (PL) if all the estimators return values >1/4, (4) Power-law distributions having divergent second moments, meaning that the tail exponent is <3, i.e., infinite variance, are of particular interest and (5) a power law empirical distribution has a 'Divergent Second Moment' (DSM) if all the estimators return values >1/2. Finally, it is important to note that there are no restrictions on how close to each other, the estimated values must be in the definitions above.

2.1.3. Empirical Coverage Distributions

The empirical coverage distribution is a discrete probability distribution $P(k)$ defined on the non-negative integers and is obtained from a genome assembly by mapping to a reference sequence by counting, for each $k = 0, 1, 2, ...$, how many bases are covered by k reads. The expected number of covered bases is the mean of this distribution. It is convenient to consider the log-transformed distribution (LTD), obtained by replacing k by (log k) (natural logarithm) in the above. This allows one to compare with a (discretized) Normal Distribution by a quantile-quantile (Q-Q) plot. It is also common to consider the log-log representation of distribution, given by (log k, log $P(k)$). Finally, one defines the complementary cumulative distribution function (CCDF) associated with $P(k)$ by $\dot{F}(k) = 1 - F(k)$, where $F(k)$ is the cumulative distribution function (CDF) associated with $P(k)$.

2.2. Genome Assembly Data

Genome data here analyzed are fully characterized elsewhere [2]. BAM files of the assemblies in [2] were used for coverage calculation using Geneious Prime 2024 software (https://www.geneious.com accessed on 22 March 2024). Coverage data were exported to a csv file and reordered as "number of sites as a function of coverage".

We considered the four ancient DNA (aDNA) virus assemblies from [2] and their corresponding random "mock" reference assemblies: (a) Neanderthal adenovirus virus reference (ADV) and random reference (ADV-R), (b) Neanderthal herpesvirus reference (HPV) and random reference (HPV-R), (c) Neanderthal papillomavirus reference (HSV) and random reference (HSV-R), and (d) the negative control parvovirus B19 reference (PB19) and its corresponding random reference (PB19-R).

Two mapDamage controls consisting of human mitochondrial DNA reference (mtDNA) mapped to aDNA reads (positive control) and modern adenovirus (MADV) reads mapped to ADV reference (negative control) from [2] were also analyzed.

2.3. Analysis of Coverage Distributions

For each of the four ancient virus assemblies described in item 2.2. the empirical coverage distribution was computed and analyzed by the program TIE. Subsequently, we used the program PLFit to estimate, in the cases where a power law was excluded, if the empirical coverage distribution can be modeled as a Log-Normal Distribution. For each of the two non-ancient assemblies described in item 2.2. the empirical coverage distribution was computed, and it was compared, using a two sample Kolmogorov-Smirnov test, with a simulated Negative Binomial Distribution, as predicted by the Lander-Waterman model. We used the R package 'KSgeneral', that allows for comparison of discrete distributions, i.e., ties (repeated observations) are allowed [20].

2.4. Software

The computation of the estimators for the classification of power law distributions is performed by a Python program called, 'Tail Index Estimation' (TIE) (https://github.com/ivanvoitalov/tail-estimation accessed on 30 September 2024) [19].

The program 'PLFit Algorithm' [17,21], was used as implemented in R package 'poweRlaw' [22], to test for the possibility of a non-power law distribution to be well approximated by another heavy tailed distribution. There are two tests implemented in the R package 'poweRlaw': Vuong's test [21] and a Bootstrap test [22].

3. Results

3.1. Coverage Distributions

The coverage distributions obtained from BAM files of all assemblies analyzed (ADV, HSV, HPV, PB19 and respective random controls) are depicted with their corresponding curve fittings (**Figure 2**). The corresponding histograms and Q-Q plots of the log transformed coverage are shown in **Figure 3**.

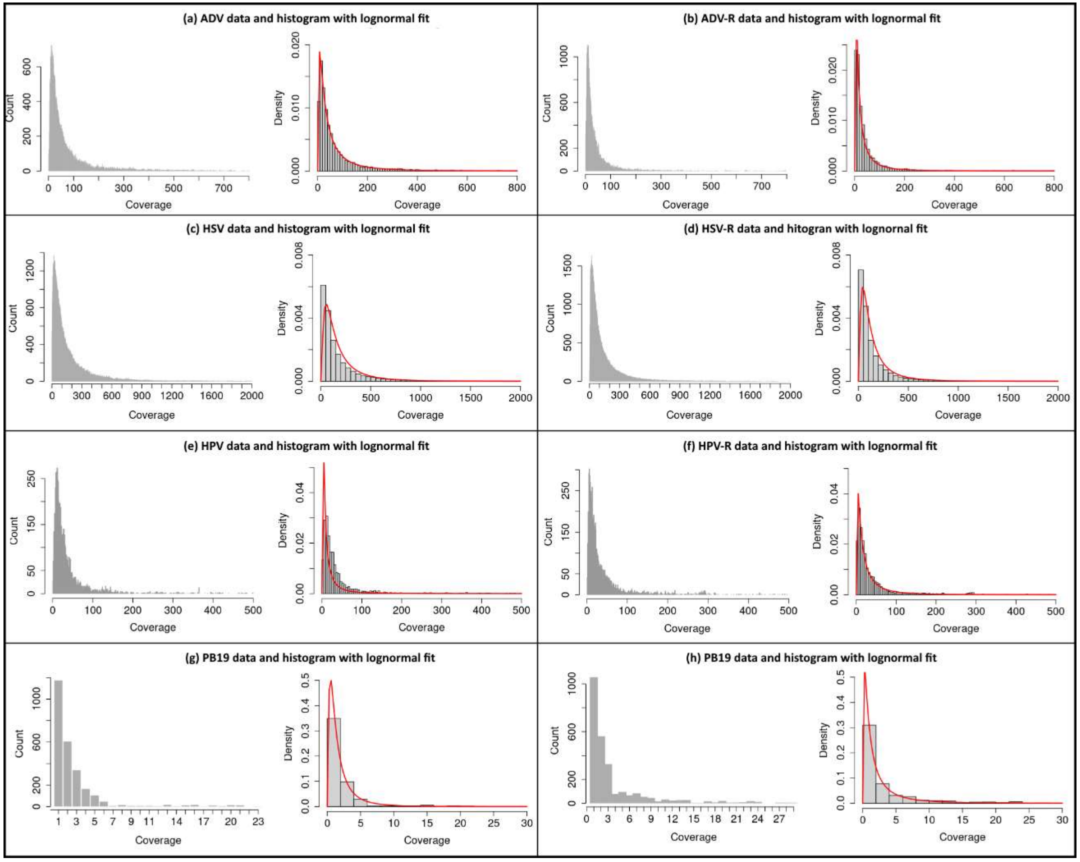

Figure 2. Coverage distributions of assemblies (BAM files) with corresponding lognormal fit. In (a) ADV and (b) the corresponding random "mock" control ADV-R. In (c) HSV and (d) the corresponding random "mock" control HSV-R. In (e) HPV and (f) the corresponding random "mock" control HPV-R. In (g) the negative control PB19 and (h) the corresponding random "mock" control PB19-R.

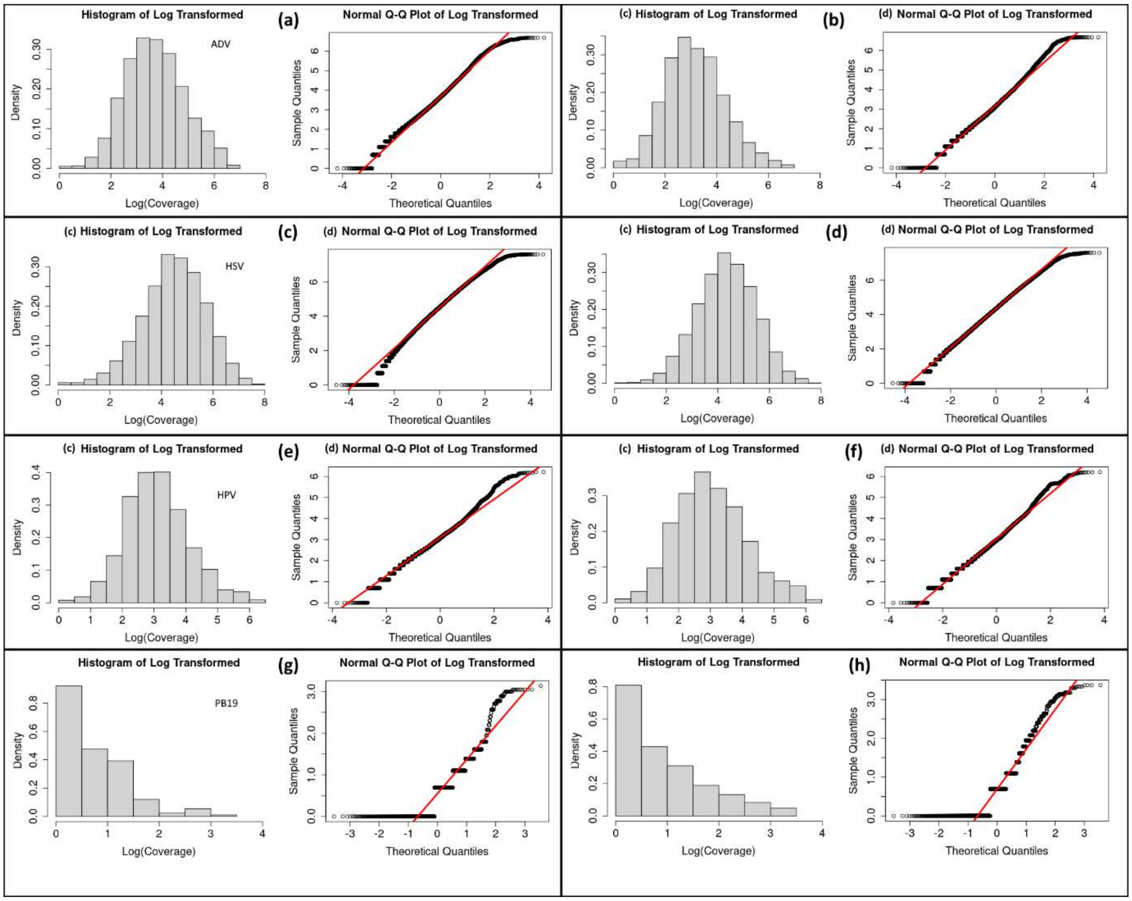

Figure 3. Histograms of log transformed, and Q-Q plots of log transformed coverage distributions of ADV (a) and ADV random control (b), HSV (c) and HSV random control (d), HPV (e) and HPV random control (f), negative control PB19 (g) and PB19 random control (h).

The coverage distributions were used to calculate the Log-log plots of the distributions (log k, log P(k)) and the Log-log plots of the complementary cumulative distribution functions (log k, log Ḟ(k)) (**Figure 4**).

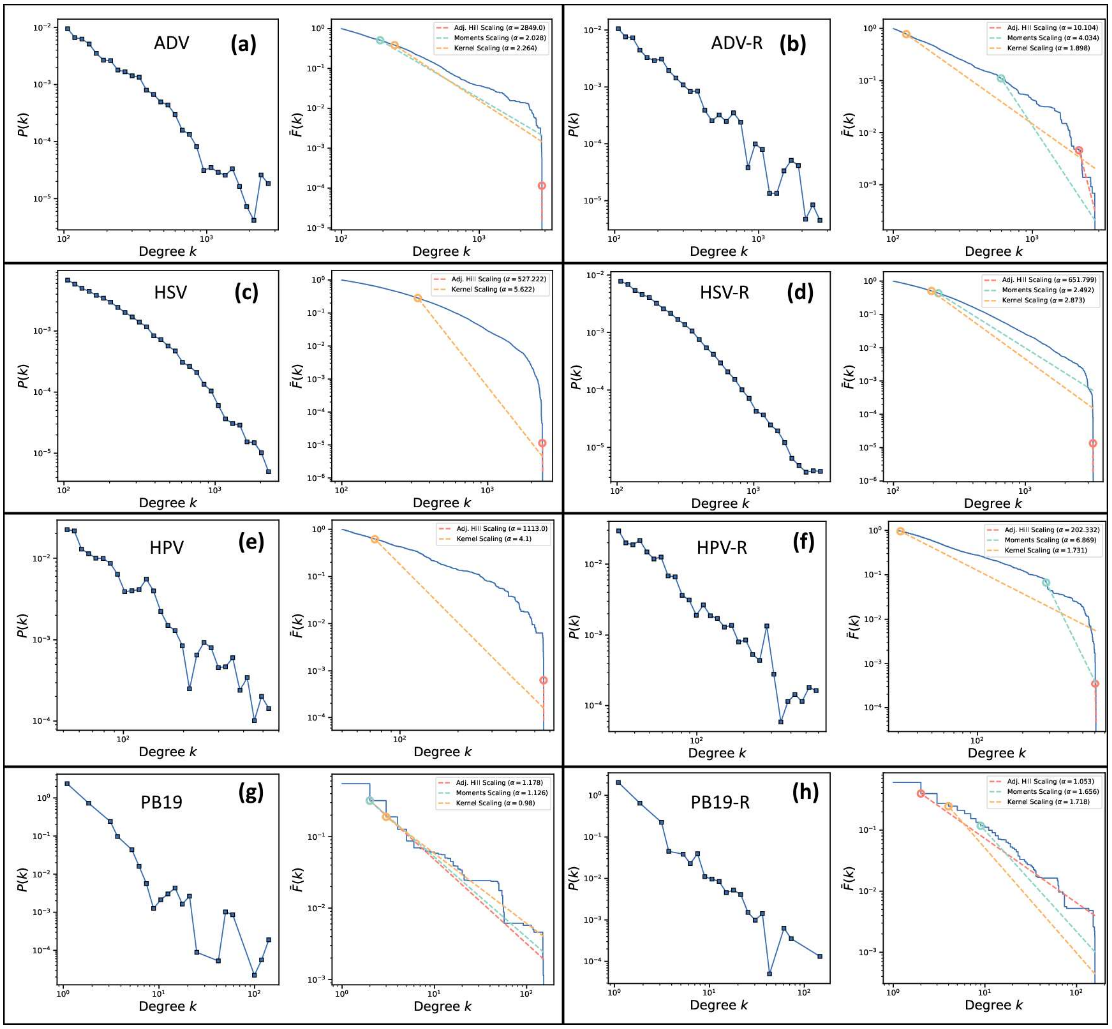

Figure 4. Log-log plots of the distribution (log k, log P(k)) and corresponding log-log plot of the complementary cumulative distribution function (log k, log F̄(k)) of ADV (a), ADV random (b), HSV (c), HSV random (d), HPV (e), HPV random (f), PB19 (g) and PB19 random (h), respectively.

The plots of the estimators are shown in **Supplementary Figures S1-S8**. The Adjusted Hill (H) estimator and its smoothed version are depicted in the original scale and in log-scale for the number of simulation steps (number of order statistics k). The Moments (M) and Kernel (K) estimators are depicted in their original scales and in the log-scale for the number of simulation steps (number of order statistics k) (**Supplementary Figures S1-S8**).

3.2. Basic Statistical Parameters of the Coverage Distributions

The basic statistical parameters of the coverage distributions depicted in **Figure 1** (mean, median, standard deviation and the number of mapped reads of each assembly) are shown in **Table 1**. These parameters are used in the Welch test for estimating the signal-to-noise as previously described [2].

doi:10.20944/preprints202412.2065.v1

Table 1. Basic statistical parameters of the empirical coverage distributions. Columns 2 to 5 contain the mean, median, standard deviation (SD) and number of reads (N) of the empirical coverage distributions, respectively. Columns 6 to 8 contain the mean, median and standard deviation of the log-transformed distributions (Log-Mean, Log-Median, Log-SD), respectively.

Assembly	Mean	Median	SD	N (reads)	Log-Mean	Log-Median	Log-SD
ADV	102.2	40	249.4	180,419	3.7	3.6	1.1
ADV-R	62.1	22	180.2	126,613	3.2	3.1	1.2
HSV	171.2	92	274.3	1,224,713	4.4	4.5	1.2
HSV-R	154.7	77	704.2	1,166,326	4.3	4.3	1.1
HPV	115.3	22	609.1	23,998	3.2	3.1	1.2
HPV-R	50.9	20	115.9	22,682	3.1	3.0	1.1
PB19	2.0	0	9.2	714	0.7	0.7	0.9
PB19-R	2.7	0	10.0	975	0.9	0.7	1.0

The log-transformed distribution (LTD) with its basic statistical parameters (Log-Mean, Log-Median, Log-SD) and are shown (**Table 1**).

3.3. The Tail Index Estimation (TIE)

The Results of the TIE program are shown in **Table 2**. For each assembly the TIE program computed the 3 estimators – the Adjusted Hill (H) Estimator, the Moments (M) Estimator and the Kernel (K) Estimator – as detailed in Material and Methods. The conclusion of the analysis is given by the combined results of the three estimators given four possibilities: 'Not Power Law' (NLP), 'Hardly Power Law' (HLP), 'Power Law' (PL) and 'Power Law with Divergent Second Moment' (PL-DSM).

Table 2. Results of the TIE program. The first three columns contain the values of the three estimators, the Adjusted Hill Estimator (H), the Moments Estimator (M) and the Kernel Estimator (K). The last column contains the results of the analysis: 'Not Power Law' (NLP), 'Hardly Power Law' (HLP), and 'Power Law with Divergent Second Moment' (PL-DSM), see 2.1.2.

Assembly	Hill (H)	Moments (M)	Kernel (K)	Conclusion
ADV	0.000	0.493	0.449	NPL
ADV-R	0.099	0.248	0.517	HPL
HPV	0.001	-0.531	0.236	NPL
HPV-R	0.005	0.146	0.576	HPL
HSV	0.002	-0.431	0.172	NPL
HSV-R	0.002	0.401	0.348	HPL
PB19	0.849	0.888	1.034	PL-DSM
PB19-R	0.949	0.604	0.613	PL-DSM

3.4. Positive and Negative Controls

The empirical coverage distributions of the two mapDamage control assemblies were computed: (1) human mitochondrial DNA (mtDNA) mapped with aDNA reads and (2) modern adenovirus reference mapped with present-day reads (MADV), for the purpose of comparison with the aDNA viral assemblies (**Figure 5**). Assuming that the Lander-Waterman model is a good approximation for these sequencing projects, it is expected that the empirical coverage distribution follows a negative binomial distribution. We tested this hypothesis using a two sample Kolmogorov-Smirnov test comparing the empirical coverage distribution with a simulated / bootstrapped negative binomial distribution with the same sample size as the empirical coverage distribution.

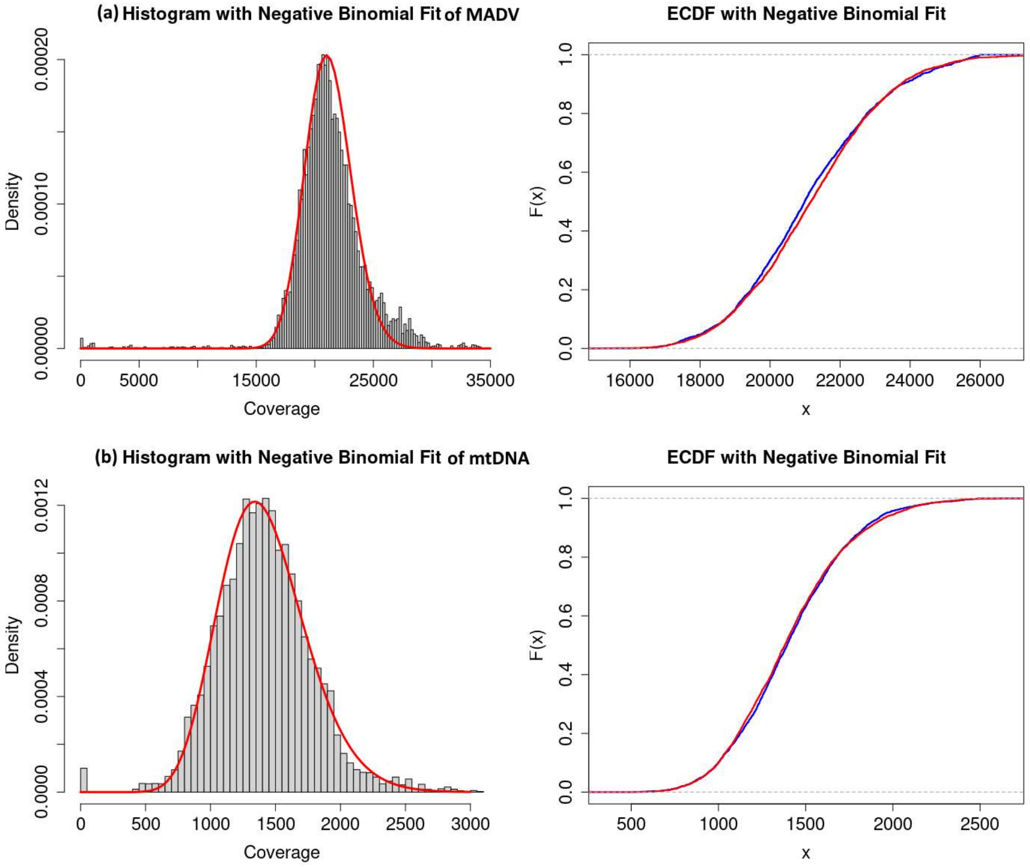

Figure 5. Histogram of coverage distributions and the empirical cumulative distribution function (ECDF) with negative binomial fit of modern ADV (MADV) (a) and mitochondrial DNA reference assembled with aDNA reads (b) used in (ref.) as mapDamage controls.

The results show that for the MADV a negative binomial distribution with parameters (r,μ)=(117;21125) and KS p-value 0.126 was obtained. For the mtDNA a negative binomial distribution with parameters (r,μ)=(19;1410) and KS p-value 0.232 was obtained. In both cases, the test did not reject the null hypothesis with significance level α=0.01 (99%).

3.5. Lognormal Fitting Using the PLFit Program

We considered the cases that were classified as 'Not Power Law' (NPL) and 'Hardly Power Law' (HPL) and tested if, in some of these cases, the empirical coverage distribution could be well approximated by a (discretized) Log-Normal Distribution. We perform the comparison by first applying Vuong's test to a Power Law fitting versus a Log-Normal fitting to the same empirical distribution. Then we apply the Bootstrap test to both fittings, to obtain two estimates for the goodness-of-fit (GOF). The comparison between the obtained p-values complementing the information of the estimators computed by the TIE program. The results are summarized in **Table 3**. First, we note that the p-values of the Bootstrap test are relatively high, in accordance with the TIE results. Second, due to the simulations needed to compute these p-values, there is some degree of uncertainty of about 0.01 in all cases. Therefore, to ensure a safe margin of error, we set the significance level to α=0.1 (90%). The conclusion of these tests indicates that the random assemblies are well approximated by Log-Normal Distributions, with very good agreement in the cases of HPV random and HSV random.

Table 3. PLFit results of comparison between Power Law fitting versus a Log-Normal fitting. The first column shows the *p*-values of Vuong's test. A small *p*-value indicates that one of the distributions is closer to the true distribution. Columns 2 and 3 show the *p*-values of the Bootstrap test. Column 2 shows the *p*-values for the Power Law (PL) GOF and column 3 shows the *p*-values for the Log-Normal (LN) GOF. Significance level α=0.1 (*one or both *p*-values are very close to the threshold α, giving a marginal rejection / non-rejection).

Assembly	Vuong's	Bootstrap PL	Bootstrap LN	Conclusion
ADV	0.0	0.25	0.04	Not Reject PL (PL>LN)
ADV-R	1.7×10^{-6}	0.09	0.15	*Not Reject LN (LN>PL)
HPV	1.4×10^{-7}	0.04	0.12	*Reject both (LN>PL)
HPV-R	1.7×10^{-8}	0.16	0.37	*Not Reject LN (LN>PL)
HSV	7.6×10^{-4}	0.00	0.10	Reject both (LN>PL)
HSV-R	1.8×10^{-9}	0.00	0.21	Not Reject LN (LN>PL)

4. Discussion

The basic assumption of genome assembly by mapping to a reference sequence is that it follows a Poisson distribution [10]. In a Poisson distribution the mean and variance are assumed to be equal. When the variance exceeds the mean, this indicates overdispersion. Overdispersion refers to a situation where the observed variance of a data set is greater than what is expected under a particular statistical model. In general, overdispersion can be caused by several factors, such as: unobserved heterogeneity, clustering or correlation between observations, misspecification of the distribution, measurement errors and external covariates [23,24]. Overdispersion in genome assemblies occurs when the variability in the number of reads mapped to each genomic position exceeds what would be expected under a simple Poisson model which assume that the mean and variance of the read coverage are equal, however in many biological contexts, the variance often exceeds the mean. In genome assemblies, overdispersion can arise from several sources, including: (1) uneven sequencing coverage, (2) repetitive sequences, (3) spurious mapping, (4) PCR amplification bias, (5) sequencing errors, (6) randomness in sampling reads, (7) Variation in Gene Copy Number, (8) Fragmentation Bias, (9) reference genome inaccuracies [25,26].

In the current study we show that aDNA assemblies are overdispersed as compared to modern DNA assemblies by reference mapping (**Figures 2 and 5**). The size of the reads is to be taken into consideration in the case o aDNA because a very stringent filter for size might discard precious information, as is the case with smaller reads associated with smaller pathogen genomes, in particular, viral genomes [1,2]. The comparison of assemblies with real reference sequences versus random references, shows that although both follow heavy tailed distributions, random assemblies are well approximated by Log-Normal Distributions, with very good agreement in the cases of HPV random and HSV random (**Table 3**). This might explain, at least in part, real reference assemblies, even with smaller read sizes, provided assemblies that passed the Welch's t test as shown by Ferreira et al. [2]. This also suggests that random assemblies are even more overdispersed than real assemblies and removal of reads might not be necessary to obtain statistically significant results, in other words, assemblies where the signa-to-noise ratio is acceptable.

Our analysis shows that the coverage distributions of the real ancient assemblies (ADV, HSV and HPV) do not follow power laws nor log-normal laws and that the coverage distributions of the *random controls* are well approximated by log-normal laws (**Tables 2 and 3**). On the other hand, the coverage distributions of the negative control parvovirus B19 (real and random) follow a power law with infinite variance (**Figure 3g and 3h**) while the coverage distributions of the mapDamage negative control with non-ancient DNA (modern ADV) and the mapDamage positive control (human mtDNA) (Figure 5) are well approximated by the negative binomial distribution which are consistent with predictions of the Lander-Waterman model [10].

doi:10.20944/preprints202412.2065.v1

Our present work addresses the problem of overdispersion in aDNA assemblies, particularly in what concerns the tails of distributions. This analysis might contribute to future research by providing statistical methods to help in research leading to identification of viral remnants in aDNA samples. Overdispersion has a significant impact on the tails of statistical distributions, particularly in the context of genomic data analysis, where the distribution of read counts often exhibits higher variability than expected under simpler models like the Poisson distribution. In general, overdispersion can lead to heavier tails in the statistical distribution, meaning that extreme events (very high or very low values) occur more frequently than predicted by distributions without overdispersion. This has important consequences in a variety of fields, including genomics, epidemiology, and ecology, where understanding the behavior of the tails of distributions is crucial for modeling rare events or extreme observations. In general, overdispersion impacts the tails of statistical distributions leading to (1) heavier tails, (2) higher variability and extreme values, (3) challenges in modeling, (4) robustness of tail predictions, and especially (5) power law behavior. Overdispersion can give rise to power law behavior in the tails of the distribution, where extreme values follow a power law decay rather than exponential decay. This is particularly relevant in contexts like biological networks or genomic data, where certain highly expressed genes or abundant sequences may appear disproportionately often [27].

5. Conclusion

In summary, the analysis described above provides a classification of the empirical coverage distributions: (1) the coverage distributions of the real aDNA assemblies (ADV, HSV and HPV) do not follow power laws nor log-normal laws, (2) the coverage distributions of the *random controls* of aDNA assemblies are well approximated by log-normal laws, (3) the coverage distributions of the negative control parvovirus B19 (real and random) follow a power law with infinite variance and (4) the coverage distributions of the mapDamage negative control with non-ancient DNA (Modern ADV) and the mapDamage positive control (human mtDNA) are well approximated by the negative binomial distribution. We conclude that the tails of distributions of reads in a genome assembly by reference mapping can reveal the level of random effects and assess the quality of the assemblies. In addition to non-parametric tests for signal-to-noise ratio, the statistical distributions, as studied here, can contribute to the mitigation of problems related to spurious alignments in aDNA reconstructions and inference.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author Contributions: Conceptualization, M.R.S.B. and F.A.; methodology, F.A. and C.M.P.; performed the experiments, F.A. and C.M.P.; analyzed the data, F.A. and M.R.S.B.; writing-original draft preparation, M.R.S.B. and F.A.; writing-review and editing, M.R.S.B., C.M.P. and F.A.. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, Brazil) grant 20/08943-5 to M.R.S.B. and F.A. and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil) grant 311154/2021-2 to M.R.S.B.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data produced by analyses in this study are available upon request after acceptance of the manuscript for publication.

Acknowledgments: We thank Renata C. Ferreira and Marcello Ramon for their continuous support.

Conflicts of Interest: The authors have no conflict of interests to declare.

References

1. Guellil, M.; van Dorp, L.; Inskip, S.A.; Dittmar, J.M.; Saag, L.; Tambets, K.; Hui, R.; Rose, A.; D'Atanasio, E.; Kriiska, A.; et al. Ancient Herpes Simplex 1 Genomes Reveal Recent Viral Structure in Eurasia. *Science Advances* **2022**, *8*, eabo4435, doi:10.1126/sciadv.abo4435.
2. Ferreira, R.C.; Alves, G.V.; Ramon, M.; Antoneli, F.; Briones, M.R.S. Reconstructing Prehistoric Viral Genomes from Neanderthal Sequencing Data. *Viruses* **2024**, *16*, 856, doi:10.3390/v16060856.
3. Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform. *Bioinformatics* **2009**, *25*, 1754–1760, doi:10.1093/bioinformatics/btp324.
4. de Filippo, C.; Meyer, M.; Prüfer, K. Quantifying and Reducing Spurious Alignments for the Analysis of Ultra-Short Ancient DNA Sequences. *BMC Biol* **2018**, *16*, 121, doi:10.1186/s12915-018-0581-9.
5. Goel, M.; Sun, H.; Jiao, W.-B.; Schneeberger, K. SyRI: Finding Genomic Rearrangements and Local Sequence Differences from Whole-Genome Assemblies. *Genome Biology* **2019**, *20*, 277, doi:10.1186/s13059-019-1911-0.
6. Koboldt, D.C. Best Practices for Variant Calling in Clinical Sequencing. *Genome Medicine* **2020**, *12*, 91, doi:10.1186/s13073-020-00791-w.
7. Salzberg, S.L.; Phillippy, A.M.; Zimin, A.; Puiu, D.; Magoc, T.; Koren, S.; Treangen, T.J.; Schatz, M.C.; Delcher, A.L.; Roberts, M.; et al. GAGE: A Critical Evaluation of Genome Assemblies and Assembly Algorithms. *Genome Res* **2012**, *22*, 557–567, doi:10.1101/gr.131383.111.
8. Baker, M. De Novo Genome Assembly: What Every Biologist Should Know. *Nat Methods* **2012**, *9*, 333–337, doi:10.1038/nmeth.1935.
9. Ma, X.; Shao, Y.; Tian, L.; Flasch, D.A.; Mulder, H.L.; Edmonson, M.N.; Liu, Y.; Chen, X.; Newman, S.; Nakitandwe, J.; et al. Analysis of Error Profiles in Deep Next-Generation Sequencing Data. *Genome Biology* **2019**, *20*, 50, doi:10.1186/s13059-019-1659-6.
10. Lander, E.S.; Waterman, M.S. Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis. *Genomics* **1988**, *2*, 231–239, doi:10.1016/0888-7543(88)90007-9.
11. Deng, C.; Daley, T.; Brandine, G.D.S.; Smith, A.D. Molecular Heterogeneity in Large-Scale Biological Data: Techniques and Applications. *Annual Review of Biomedical Data Science* **2019**, *2*, 39–67, doi:10.1146/annurev-biodatasci-072018-021339.
12. Resnick, S.I. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*; Springer Science & Business Media, 2007; ISBN 978-0-387-24272-9.
13. Embrechts, P.; Klüppelberg, C.; Mikosch, T. *Modelling Extremal Events: For Insurance and Finance*; Springer Science & Business Media, 2013; ISBN 978-3-540-60931-5.
14. Bingham, N.H.; Goldie, C.M.; Teugels, J.L. *Regular Variation*; Encyclopedia of Mathematics and its Applications; Cambridge University Press: Cambridge, 1987; ISBN 978-0-521-37943-4.
15. Foss, S.; Korshunov, D.; Zachary, S. *An Introduction to Heavy-Tailed and Subexponential Distributions*; Springer Series in Operations Research and Financial Engineering; Springer: New York, NY, 2013; ISBN 978-1-4614-7100-4.
16. Holme, P. Rare and Everywhere: Perspectives on Scale-Free Networks. *Nat Commun* **2019**, *10*, 1016, doi:10.1038/s41467-019-09038-8.
17. Broido, A.D.; Clauset, A. Scale-Free Networks Are Rare. *Nat Commun* **2019**, *10*, 1017, doi:10.1038/s41467-019-08746-5.

18. Stumpf, M.P.H.; Porter, M.A. Critical Truths About Power Laws. *Science* **2012**, *335*, 665–666, doi:10.1126/science.1216142.
19. Voitalov, I.; van der Hoorn, P.; van der Hofstad, R.; Krioukov, D. Scale-Free Networks Well Done. *Phys. Rev. Res.* **2019**, *1*, 033034, doi:10.1103/PhysRevResearch.1.033034.
20. Vuong, Q.H. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica* **1989**, *57*, 307–333, doi:10.2307/1912557.
21. Clauset, A.; Shalizi, C.R.; Newman, M.E.J. Power-Law Distributions in Empirical Data. *SIAM Rev.* **2009**, *51*, 661–703, doi:10.1137/070710111.
22. Gillespie, C.S. Fitting Heavy Tailed Distributions: The poweRlaw Package. *Journal of Statistical Software* **2015**, *64*, 1–16, doi:10.18637/jss.v064.i02.
23. Aitchison, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)* **1982**, *44*, 139–160, doi:10.1111/j.2517-6161.1982.tb01195.x.
24. Aschard, H.; Vilhjálmsson, B.J.; Joshi, A.D.; Price, A.L.; Kraft, P. Adjusting for Heritable Covariates Can Bias Effect Estimates in Genome-Wide Association Studies. *The American Journal of Human Genetics* **2015**, *96*, 329–339, doi:10.1016/j.ajhg.2014.12.021.
25. Treangen, T.J.; Salzberg, S.L. Repetitive DNA and Next-Generation Sequencing: Computational Challenges and Solutions. *Nat Rev Genet* **2011**, *13*, 36–46, doi:10.1038/nrg3117.
26. Mardis, E.R. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics* **2008**, *9*, 387–402, doi:10.1146/annurev.genom.9.081307.164359.
27. Hinde, J.; Demétrio, C.G.B. Overdispersion: Models and Estimation. *Computational Statistics & Data Analysis* **1998**, *27*, 151–170, doi:10.1016/S0167-9473(98)00007-3.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.