

Multifractal-Driven Optimization of Read Length and Depth in Metagenomic Sequencing: A Utility-Based Framework

Xianhua Xie

xxianhua@sina.com.cn

Gannan Normal University

Zuguo Yu

Xiangtan University

Guosheng Han

Xiangtan University

Yujie Huang

Gannan Normal University

Yuanlin Ma

Zhengzhou University of Aeronautics

Research Article

Keywords: multifractal analysis, metagenomics, read length, sequencing depth, utility function, elasticity, chaos game representation

Posted Date: October 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10688780/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Multifractal-Driven Optimization of Read Length and Depth in Metagenomic Sequencing: A Utility-Based Framework

Xian-hua Xie^{a,b,*}, Zu-guo Yu^{b,c}, Guo-sheng Han^b, Yu-jie Huang^a, Yuan-lin Ma^d

^a*School of Mathematics and Computer Science, Gannan Normal University, Ganzhou, China*

^b*Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education and Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Xiangtan, China*

^c*Hunan Provincial University Key Laboratory for Big Data Analysis and Application, Hunan First Normal University, Changsha, China*

^d*School of Economics, Zhengzhou University of Aeronautics, Zhengzhou, China*

Abstract

Background: Metagenomic sequencing has become a cornerstone of microbial ecology, yet the trade-off between read length and sequencing depth remains a critical challenge for cost-effective experimental design.

Methods: We introduce a novel framework—*fractal economics*—that leverages multifractal analysis to quantify information recovery from metagenomic data. By simulating 20×20 parameter combinations (read length 100–2000 bp, depth 1–20 M reads, 100 replicates each) and computing the deviation of the multifractal dimension from an ideal asymptotic value, we define a utility function $U(N, L) = -\ln(\text{dev})$ and derive elasticities (E_L, E_N) and marginal rates of substitution ($E_{N,L}$).

Results: Our results reveal a steep information gain when read length increases from 100 bp to 200 bp—particularly pronounced in shallow sequencing—followed by diminishing marginal returns beyond 500 bp and 10 M reads, with utility saturating near 1800 bp. We also observe near 1:1 substitution between read length and depth in the range 5–10 M reads and 500–1000 bp. A subsequent cost-benefit analysis, grounded on real Illumina NovaSeq SP platform specifications (maximum read length 600 bp, data-proportional pricing), reveals that the optimal parameters depend strongly on the number of multiplexed samples and total budget. For a typical large cohort (192 samples) with a moderate budget (5000 USD), the constrained optimum is $N = 3$ M reads and $L = 600$ bp, achieving utility $U = 5.25$ at a per-sample cost of 20.36 USD.

Conclusion: This framework provides a theoretically grounded, data-driven roadmap for balancing information yield and cost in metagenomic experimental design.

Keywords: multifractal analysis, metagenomics, read length, sequencing depth, utility function, elasticity, chaos game representation

1. Introduction

Over the past two decades, next-generation sequencing (NGS) has revolutionized genomics by enabling rapid, cost-effective whole-genome sequencing at unprecedented scale [1]. This

*Corresponding author: xxianhua@sina.com.cn

technological leap has empowered large-scale initiatives such as the expanded 1000 Genomes Project [2] and has transformed clinical practice, notably in cancer genomics [3]. In microbial ecology, metagenomic sequencing has become indispensable for characterizing the composition, function, and dynamics of microbial communities across diverse habitats—from the human gut to soil and ocean ecosystems.

Two fundamental parameters govern all sequencing experiments: **read length** and **sequencing depth (number of reads)**. Their choice directly impacts assembly contiguity, taxonomic resolution, and cost-effectiveness [4, 5]. In metagenomics, longer reads (e.g., PacBio, Nanopore) improve assembly continuity and enable better resolution of repetitive regions, but incur higher per-base costs and increased error rates; higher depth (e.g., Illumina) enhances statistical power for detecting rare taxa and improves coverage uniformity, yet exacerbates computational burdens and sequencing costs [4, 6]. Despite these well-recognized trade-offs, experimental design remains largely empirical or guided by rules of thumb derived from a limited set of prior studies.

Two lines of evidence underscore the need for a principled optimization framework. First, simulation-based benchmarking has demonstrated that typical sequencing depths (1–10 Gb) already reach the point of diminishing returns for metagenome-assembled genome recovery [7], suggesting that deeper sequencing yields progressively smaller gains in assembly completeness. Second, our previous work established that multifractal dimensions derived from metagenomic sequences correlate strongly with classical diversity indices [8], providing a theoretical foundation for using fractal measures as holistic indicators of community complexity. These observations highlight a pressing need for a data-driven framework that can quantitatively balance information yield against sequencing cost while avoiding the computational overhead of assembly-based metrics.

Limitations of existing approaches.. Conventional methods for optimizing read length and depth rely on empirical pilot experiments or heuristic guidelines that do not generalize across different microbial communities or sequencing platforms. For example, assembly-based metrics (e.g., N50, contig length) are widely used but are computationally expensive and depend heavily on the specific assembler and its parameters. Diversity-based metrics (e.g., observed species richness, Shannon index) require knowledge of the true community composition—precisely what the experiment aims to uncover—making them unsuitable as a priori optimization criteria. Moreover, these metrics do not capture the *information content* of the sequencing data in a holistic manner; they often focus on specific aspects such as assembly completeness or taxonomic assignment accuracy, ignoring the complex interdependence between read length and depth.

Fractal and multifractal analysis as a novel paradigm.. Fractal and multifractal analysis offers a fundamentally different perspective. Instead of relying on explicit assembly or alignment, these methods characterize the global statistical structure of sequences using scale-invariant measures. The Chaos Game Representation (CGR) maps DNA sequences into a unit square, revealing self-similar patterns that reflect the underlying compositional organization of genomes [9, 10]. Multifractal spectra—specifically the generalized dimension $D(q)$ —quantify the degree of heterogeneity in sequence composition, providing a compact yet comprehensive signature of genomic complexity. These methods have been successfully applied to distinguish coding from non-coding regions, classify microbial species, and infer phylogenetic relationships without sequence alignment [11, 12, 13]. More recently, multifractal analysis has been extended to viral genomes, including SARS-CoV-2, and to alignment-free phylogenetics [14, 15]. However, the

potential of multifractal analysis as a *quantitative measure of information recovery completeness* from metagenomic sequencing has remained largely unexplored.

Positioning within alignment-free methodologies.. In the era of massive metagenomic data generation, the development of *alignment-free* methodologies has emerged as a critical research frontier. Traditional alignment-dependent approaches become computationally prohibitive when applied to thousands of metagenomic samples or when dealing with highly divergent microbial communities. Consequently, there has been a surge of interest in methods that bypass sequence alignment entirely, including k -mer frequency-based approaches, feature frequency profiles, and chaos game representations [16, 17]. Our work squarely falls within this alignment-free paradigm. By leveraging the CGR and multifractal analysis, we avoid the computational bottlenecks and reference biases inherent to alignment-dependent methods. Moreover, unlike most existing alignment-free methods that focus on phylogenetic inference or taxonomic classification, our framework addresses a fundamentally different and equally important question: how to optimally design sequencing experiments to maximize information recovery under budget constraints. Thus, we position our *fractal economics* framework as an innovative contribution to the alignment-free methodology landscape, extending its scope from sequence comparison to experimental design optimization.

Hypothesis and proposed framework.. We hypothesize that the multifractal dimension of a metagenomic dataset converges to an asymptotic value as read length and depth increase, and that the deviation from this asymptotic value quantifies the remaining information yet to be recovered. To test this hypothesis, we compute the deviation between the multifractal dimension obtained from a simulated metagenome under a given parameter pair (N, L) and the ideal dimension $D^{\text{ideal}}(q)$ computed directly from the complete reference genomes of the community (without fragmentation into reads). This ideal dimension represents the theoretical information limit—the maximal information recoverable under perfect sequencing conditions. Based on this deviation, we define a utility function $U(N, L) = -\ln(\text{dev})$, which serves as a scalar measure of information recovery, independent of downstream analysis choices.

Building on this utility function, we borrow concepts from microeconomics—elasticity and marginal rate of substitution—to derive actionable insights. The elasticity of utility with respect to read length (E_L) and depth (E_N) indicates which parameter yields greater marginal gains for a given experimental configuration. The marginal rate of substitution $E_{N,L}$ quantifies the trade-off between read length and depth, answering the question: “How many additional reads are needed to compensate for a reduction in read length while maintaining the same information recovery?”

Contributions and roadmap.. In this work, we systematically simulate 20×20 parameter combinations (read length 100–2000 bp, depth 1–20 M reads, 100 replicates each) on a high-diversity metagenome dataset comprising ten distantly related bacterial species. Using the Chaos Game Representation and multifractal analysis, we compute the generalized dimension $D(q)$ for each simulated metagenome and compare it to an ideal asymptotic dimension obtained from the complete reference genomes. From these comparisons, we:

1. Characterize the utility surface $U(N, L)$ and identify regions of diminishing marginal returns;
2. Compute elasticities E_L and E_N to guide parameter prioritization under budget constraints;
3. Derive the marginal rate of substitution $E_{N,L}$ to reveal near-1:1 trade-offs in the range 5–10 M reads and 500–1000 bp;

4. Validate the robustness of our findings across different multifractal dimensions ($q = 1, 2$) and quantify estimation uncertainty;
5. Demonstrate the practical utility of the framework through a constrained cost-benefit analysis based on real Illumina platform specifications.

The remainder of this paper is organized as follows. Section 2 describes the simulation protocol, CGR construction, multifractal computation, and the definitions of utility, elasticities, and marginal rate of substitution. Section 3 presents the results, including the utility surface, elasticity heatmaps, substitution maps, extended robustness analyses, and a constrained cost-benefit analysis. Section 4 discusses the implications, limitations, and future directions. Section 5 concludes with practical recommendations.

Our framework provides a theoretically grounded, data-driven roadmap for balancing read length and depth, moving beyond empirical rules of thumb toward a principled optimization strategy that can be adapted to any microbial community or sequencing platform.

2. Materials and Methods

2.1. Simulated Metagenome Datasets

Whole-genome sequences of ten distantly related major bacterial species (Table 1) were downloaded from NCBI. Following the abundance distribution reported in [18], these ten species together account for more than 90% of all reads in the Chinese population cohort, forming a high-diversity simulated metagenome dataset (referred to as Dataset 1).

For this dataset, we generated 100 independent metagenomic samples for each combination of read length L (100–2000 bp, step 100 bp) and read count N (1–20 M reads, step 1 M), using two widely adopted read simulators: ART (A Read Simulator) [19] and wgsim, which is part of the SAMtools package [20]. By employing both simulators, we ensured that the generated metagenomic reads reflect a range of realistic sequencing artifacts, enhancing the robustness of our subsequent multifractal analysis. In total, $20 \times 20 \times 100 = 40,000$ simulated metagenomes were produced.

Table 1: Species and accession numbers used in Dataset 1.

Organism	Accession number
<i>Akkermansiamuciniphila</i> ATCC BAA-835	NC_010655.1
<i>Alistipes shahii</i> WAL 8301	NC_021030.1
<i>Bifidobacterium adolescentis</i> ATCC 15703	NC_008618.1
<i>Bacteroides vulgatus</i> ATCC 8482	NC_009614.1
<i>Coprococcus</i> sp. ART55/1	FP929039.1
<i>Eubacterium eligens</i> ATCC 27750	NC_012778.1
<i>Faecalibacterium prausnitzii</i> A2-165	ACOP02000001.1
<i>Lachnospiraceabacterium</i> 1_4_56FAA	NZ_GL945163.1
<i>Prevotellacopri</i> DSM 18205	NZ_GG703878.1
<i>Ruminococcus champanellensis</i> type strain 18P13T	NC_021039.1

2.2. Chaos Game Representation and Multifractal Analysis

For each metagenome, each read was independently mapped to a unit square using the Chaos Game Representation (CGR) [9], starting from the center point (0.5, 0.5) for every read. To avoid the “large number annihilating small number” effect, the first 10 points of each read were discarded. The CGR points from all reads were then accumulated to form a single point cloud representing the entire metagenome.

The unit square was partitioned into grids of increasing resolution: $2^6 \times 2^6$, $2^7 \times 2^7$, $2^8 \times 2^8$, $2^9 \times 2^9$, $2^{10} \times 2^{10}$, $2^{11} \times 2^{11}$, and $2^{12} \times 2^{12}$ (i.e., from 64×64 to 4096×4096). For each grid size ε , let M_i be the number of points falling into the i -th cell, and $M_0 = \sum_i M_i$ the total number of points. The multifractal dimensions $D(q)$ were computed using the box-counting method [12] according to the following formalism:

For $q \neq 1$,

$$D_q(\varepsilon) = \frac{1}{1-q} \cdot \frac{\ln \sum_i \left(\frac{M_i}{M_0}\right)^q}{\ln \varepsilon},$$

and for $q = 1$,

$$D_1(\varepsilon) = \frac{\sum_i \left(\frac{M_i}{M_0}\right) \ln \left(\frac{M_i}{M_0}\right)}{\ln \varepsilon}.$$

The generalized dimension $D(q)$ is then defined as the limit $D(q) = \lim_{\varepsilon \rightarrow 0} D_q(\varepsilon)$. In practice, $D(q)$ is obtained by linear regression of $\ln \sum_i M_i^q$ against $\ln \varepsilon$, using the relationship:

$$\ln \left(\sum_i M_i^q \right) = D(q)(q-1) \ln \varepsilon + \text{constant}.$$

For $q \neq 1$, $D(q)$ was obtained by dividing the slope of the fitted line by $(q-1)$. For the special case $q = 1$, $D(1)$ was computed directly as the limit using Eq. (3), without division by $(q-1)$.

2.3. Utility Function and Elasticities

Theoretical innovation. Unlike conventional approaches that treat read length and depth as independent parameters to be chosen based on heuristic guidelines, we introduce a fundamentally new theoretical framework—*fractal economics*—that unifies multifractal analysis with utility theory from microeconomics. To the best of our knowledge, this is the first attempt to apply economic concepts of utility, elasticity, and marginal rate of substitution to the optimization of metagenomic sequencing parameters. This framework provides a rigorous mathematical foundation for quantifying the trade-off between read length and sequencing depth, moving beyond empirical rules of thumb toward a principled optimization strategy.

For each combination (N, L) , the deviation is defined as:

$$\text{dev}(N, L; q) = \sqrt{\frac{1}{100} \sum_{i=1}^{100} (D^i(q; N, L) - D^{\text{ideal}}(q))^2}.$$

The utility is $U(N, L; q) = -\ln(\text{dev}(N, L; q))$. We fix $q = 0$ (capacity dimension) for the main analysis. The choice of $q = 0$ is motivated by its direct sensitivity to species richness—the most fundamental measure of community composition—and its established correlation with classical diversity indices from our previous work [8]. The results for $q = 1$ and $q = 2$ are provided as robustness checks (Supplementary Table S1).

The elasticities of utility with respect to read length L and read count N are defined as:

$$E_L = \frac{\partial U}{\partial L} \cdot \frac{L}{U}, \quad E_N = \frac{\partial U}{\partial N} \cdot \frac{N}{U}.$$

These elasticities quantify the responsiveness of information recovery to marginal increases in read length or depth. An elasticity greater than 1 indicates that a parameter is in the “elastic” region where proportional gains exceed proportional investments; an elasticity less than 1 signals diminishing returns.

The marginal rate of substitution between read length and read count is defined as:

$$E_{N,L} = \frac{\partial U / \partial L}{\partial U / \partial N} \cdot \frac{L}{N}.$$

It follows directly from the definitions above that $E_{N,L} = E_L / E_N$, a relation we used to verify our numerical computations. When $E_{N,L} \approx 1$, read length and depth are near-perfect substitutes, allowing flexible trade-offs depending on platform-specific costs. This concept has no precedent in sequencing experimental design and represents a key theoretical contribution of our work.

Since the utility function $U(N, L)$ is only evaluated at discrete grid points ($N = 1, 2, \dots, 20$ M reads; $L = 100, 200, \dots, 2000$ bp), the partial derivatives $\partial U / \partial L$ and $\partial U / \partial N$ are approximated using second-order central finite differences on the interior grid points, and first-order forward/backward differences at the boundaries. Specifically, for a given (N_i, L_j) :

- For the derivative with respect to L :

$$\frac{\partial U}{\partial L} \approx \begin{cases} \frac{U(N_i, L_{j+1}) - U(N_i, L_j)}{\Delta L}, & j = 1 \text{ (forward difference),} \\ \frac{U(N_i, L_j) - U(N_i, L_{j-1})}{\Delta L}, & j = 20 \text{ (backward difference),} \\ \frac{U(N_i, L_{j+1}) - U(N_i, L_{j-1})}{2\Delta L}, & 2 \leq j \leq 19 \text{ (central difference).} \end{cases}$$

where $\Delta L = 100$ bp.

- The derivative with respect to N is computed analogously with $\Delta N = 1$ M reads, using the corresponding forward, backward, or central difference formulas.

The finite-difference approximations were implemented in R across the 20×20 grid. The resulting discrete estimates of E_L , E_N , and $E_{N,L}$ were then used to analyze the marginal returns of read length and depth, as well as their substitutability. As a consistency check, we confirmed that $E_{N,L} = E_L / E_N$ holds for all grid points (to within numerical precision), validating our implementation.

3. Results

3.1. Utility Function and Optimal Parameter Regions

Figure 1 shows the utility surface $U(N, L)$ for $q = 0$. Utility increases with both read length and depth, but gains are highly nonlinear. An “elbow” region (N from 1 M to 5 M, L from 100 bp to 500 bp) shows rapid improvement. Beyond $N > 10$ M and $L > 1000$ bp, utility saturates (utility values > 12 , deviation $< 10^{-5}$).

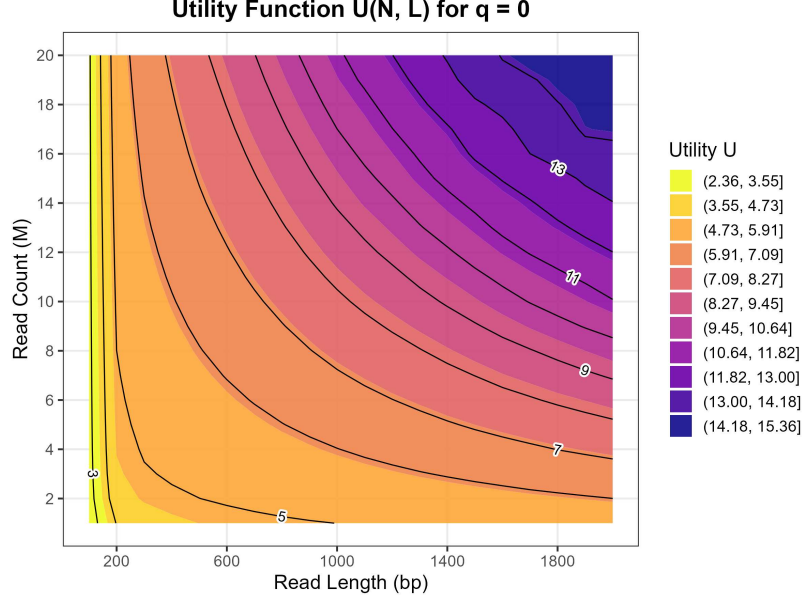


Figure 1: Utility function $U(N, L)$ for $q = 0$. Contour lines are drawn at intervals of 1, shown as black solid lines. The black numbers indicate the utility values along the contours.

3.2. Elasticities: Marginal Returns of Read Length and Depth

The elasticity E_L (Figure 2a) is highest in the low-depth (1–2 M) and short-read (≤ 200 bp) region, with values around 0.6–0.7, indicating that increasing read length is most effective in shallow sequencing with very short reads. As depth increases, E_L gradually decreases, dropping below 0.4 for $N > 15$ M. Note that all estimated elasticities are below 1, indicating that the utility function is in the inelastic region throughout the parameter space—consistent with the diminishing marginal returns observed in the utility surface. Similarly, E_N (Figure 2b) peaks around $N = 5$ –10 M and $L > 1000$ bp, confirming that extra reads are most beneficial when read length is already long.

3.3. Marginal Rate of Substitution

The marginal rate of substitution $E_{N,L}$ (Figure 3) indicates how many percent increase in read count can be replaced by a 1% increase in read length while keeping utility constant. For $N = 5$ –10 M and $L = 500$ –1000 bp, $E_{N,L} \approx 1$, meaning near-perfect substitutability: a 1% increase in read length can be offset by approximately a 1% increase in read count without loss of information recovery.

The relative efficiency of increasing depth versus read length depends strongly on the current read length, across all depth levels. At very short read lengths (e.g., $L \leq 200$ bp), increasing read length yields substantially greater utility gains than increasing depth—for instance, at $N = 1$ M, $L = 100$ bp, doubling read length to 200 bp increases utility by 1.51 units, whereas doubling depth to 2 M increases utility by only 0.15 units (see Fig. 4b, where the steep initial rise from 100 bp to 200 bp is clearly visible). At longer read lengths (e.g., $L \geq 500$ bp), the marginal gains from increasing depth and read length become more comparable.

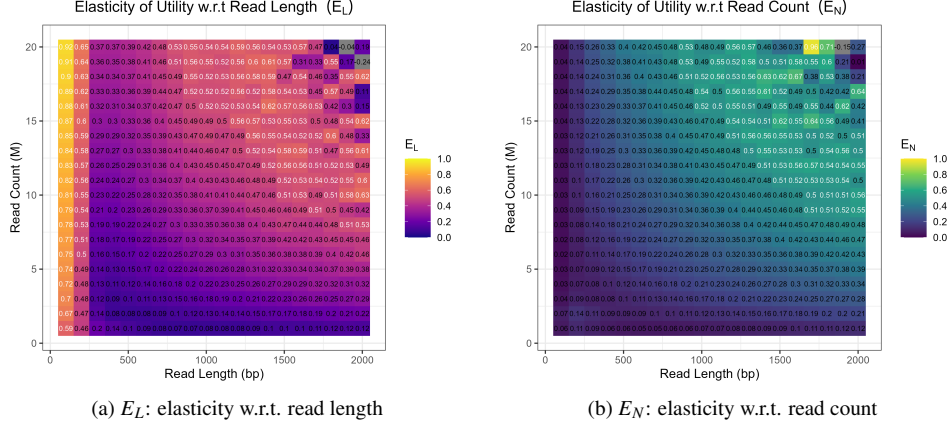


Figure 2: Elasticity of utility. Darker colors indicate higher elasticity.

In very deep sequencing ($N > 15$ M, $L > 1500$ bp), the ratio falls below unity, indicating that additional reads are slightly more beneficial than further increasing read length at the margin. However, estimates near the grid boundaries (e.g., $N = 20$ M, $L = 2000$ bp) should be interpreted with caution due to finite-difference approximation errors.

3.4. Utility Curves at Fixed Parameters

Figures 4a and 4b show utility as a function of read count (for selected read lengths) and as a function of read length (for selected read counts). Consistent with the elasticity analysis, the curves flatten after $N = 10$ M and after $L = 1000$ bp.

A distinctive feature of the utility curves (Fig. 4b) is a steep initial rise from 100 bp to 200 bp, followed by a more gradual increase that continues until approximately 1800 bp, after which utility plateaus or slightly declines due to boundary effects. The magnitude of the 100→200 bp gain varies with depth: at low depth ($N=1$ M), it is approximately ten times larger than any subsequent 100-bp increment (1.51 vs. 0.15–0.37); at higher depths ($N=20$ M), the 100→200 bp gain (0.65) is comparable to subsequent increments (e.g., 200→300: 0.63). This pattern reflects the diminishing marginal returns of read length, particularly in shallow sequencing, and is consistent with the elasticity analysis (Fig. 2a).

3.5. Extended Analysis: Robustness, Uncertainty and Substitution Rate

Multifractal dimension choice. We repeated the utility analysis for $q = 1$ (information dimension) and $q = 2$ (correlation dimension). The utility surfaces for $q = 1$ and $q = 2$ exhibited the same qualitative features as $q = 0$: diminishing returns beyond 500 bp and 10 M reads, saturation near 1800 bp, and a near-unity substitution corridor in the 5–10 M / 500–1000 bp region. The maximum utility values differed (9.82 for both $q = 1$ and $q = 2$ vs. 14.86 for $q = 0$), reflecting the different sensitivities of these dimensions to sequence heterogeneity, but the optimal parameter combination remained $N = 20$ M, $L = 1800$ bp for all three q values (Supplementary Table S1), confirming that our recommendations are robust to the choice of fractal dimension.

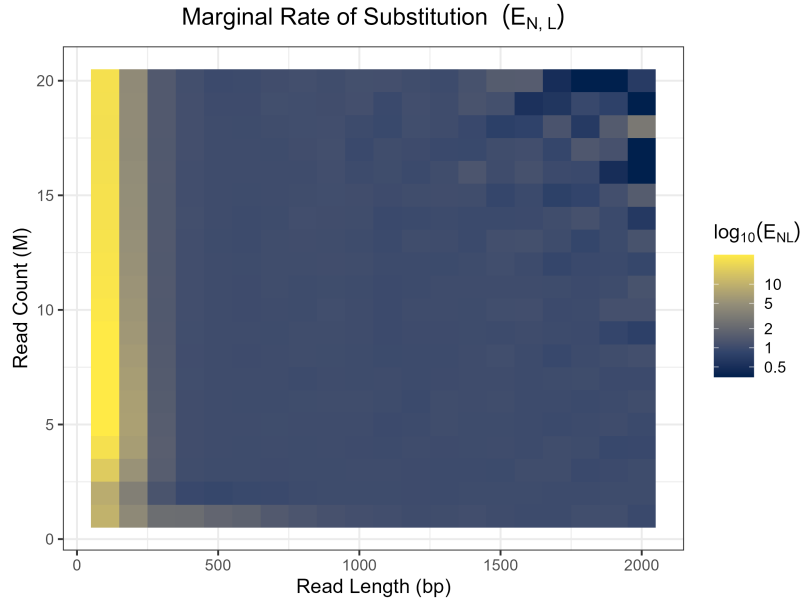


Figure 3: Marginal rate of substitution $E_{N,L}$ on a logarithmic scale. Values near 1 indicate balanced substitutability.

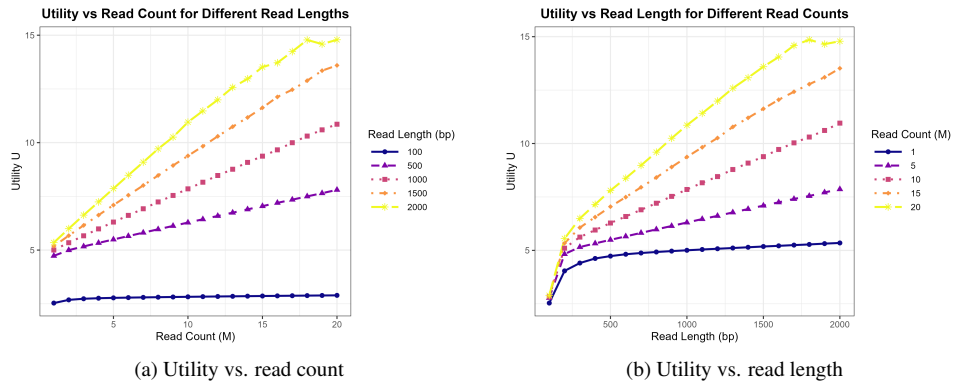


Figure 4: Utility curves at selected fixed parameters.

Uncertainty quantification. The standard error of U (computed from 100 replicates) was below 0.2 across the entire parameter space, and dropped to < 0.0004 in the high-depth, long-read region ($N > 10$ M, $L > 1000$ bp). This confirms the high reproducibility of our utility estimates (Supplementary Figure S1).

Substitution rate isocline. The contour line $E_{N,L} = 1$ was extracted from the smoothed surface. The region where $|E_{N,L} - 1| < 0.2$ spans $N = 4\text{--}20$ M and $L = 500\text{--}1000$ bp (Supplementary Figure S2), with the most practically relevant “perfect substitution corridor” at $N = 5\text{--}10$ M, $L = 500\text{--}1000$ bp where both marginal gains are non-negligible. This provides experimental designers with a flexible trade-off zone.

3.6. Saturation Plateau and Read Length Selection

A subtle but important observation from the utility curves (Figure 4b) is that at the highest depth ($N = 20$ M), utility continues to increase from $L = 100$ bp to $L = 1800$ bp, but beyond 1800 bp the utility remains essentially flat (fluctuations within ± 0.01). The maximum utility occurs at $L = 1800$ bp with a value of 14.856, while $L = 2000$ bp gives a slightly lower value of 14.788. This minor decrease is well within the numerical noise of the simulation (standard error < 0.0004), and does not indicate a true loss of information. We attribute this to a **boundary effect**: In our simulation, reads are sampled from random starting positions along the genome. For very long reads (e.g., 2000 bp), the last ~ 2000 bp of each genome are less likely to be sampled because a read starting too close to the end would exceed the genome length. This leads to a slight under-representation of terminal regions, which introduces negligible noise in the utility estimate but can cause minor non-monotonicity. Therefore, we conservatively recommend $L \geq 1800$ bp as the saturation read length; further increases provide no meaningful gain.

3.7. Constrained Cost-Benefit Analysis Under Real Illumina Platform Specifications

To simplify the analysis, we assume that sequencing cost depends only on total data volume, ignoring fixed costs and nonlinear discounts. More complex cost functions can be readily incorporated into our framework.

The theoretical utility analysis above shows that information recovery saturates for $L \geq 1800$ bp and $N \geq 10$ M reads. However, practical sequencing decisions must respect platform-specific limits and cost structures. We now incorporate a realistic cost model for the Illumina NovaSeq SP flow cell, the most widely used short-read platform.

3.7.1. Optimization problem formulation

We formalize the cost-effectiveness trade-off as a constrained optimization problem:

$$\begin{aligned} \max_{N,L} \quad & U(N, L) \\ \text{s.t.} \quad & L \leq L_{\max} = 600 \text{ bp}, \\ & \left\lceil \frac{M \cdot (NL/1000)}{C(L)} \right\rceil \cdot P(L) \leq B, \\ & N \in \{1, 2, \dots, 20\} \text{ M reads}, \quad L \in \{100, 200, \dots, 600\} \text{ bp}, \end{aligned}$$

where:

- $U(N, L)$ is the utility from our simulation (interpolated linearly over the discrete grid);

- M = number of multiplexed samples on the same flow cell;
- B = total available budget (USD);
- $C(L)$ = single-end equivalent output of one flow cell (Gb) at read length L ;
- $P(L)$ = price of one flow cell (USD) at read length L .

3.7.2. Platform parameters

We base our analysis on the Illumina NovaSeq SP flow cell, a widely used short-read platform. The single-end equivalent output $C(L)$ (Gb) is derived from official Illumina specifications for paired-end runs (2×150 bp typical output 450 Gb, i.e., 225 Gb single-end equivalent; 2×100 bp typical output 145 Gb; 2×50 bp typical output 72.5 Gb). For 500 bp, we take the mid-point of the official range (325–400 Gb) for 2×250 bp runs, i.e., 362.5 Gb. For 600 bp (2×300 bp), which is not officially specified, we linearly extrapolate from 500 bp to obtain $C(600) = 435$ Gb and $P(600) = 3,910$ USD. These values should be considered as estimates; actual platform performance and pricing may vary. Researchers should consult their sequencing provider for current specifications before finalizing experimental designs.

The flow cell price $P(L)$ is estimated based on the MIT/Whitehead Genome Technology Core (GTC) price list (August 2023) [21]. The core lists a price of 3,425 USD per flow cell for 100 bp single-end or 50×50 bp paired-end sequencing. As the flow cell can be run with one or two lanes, the per-lane cost is approximately 1,712.5 USD. After adding typical loading and handling fees (e.g., library quantification and quality control), we estimate $P(100) = 1,850$ USD per flow cell for a single-lane run. For longer read lengths, we assume that the cost scales approximately with the number of cycles; using the core’s cycle-based pricing logic, we obtain $P(200) = 2,245$ USD, $P(300) = 2,620$ USD, and $P(500) = 3,510$ USD. For 600 bp we linearly extrapolate from 500 bp, giving $P(600) = 3,910$ USD. All prices are in USD and exclude library preparation and computational costs. For intermediate read lengths (e.g., 400 bp), linear interpolation is used.

L (bp)	100	200	300	500	600
$C(L)$ (Gb)	72.5	145	225	362.5	435
$P(L)$ (USD)	1850	2245	2620	3510	3910

The per-sample data volume is $D(N, L) = N \cdot L/1000$ Gb, total data for all samples = $M \cdot D(N, L)$. The number of flow cells required is $K = \lceil MD/C(L) \rceil$, and total cost = $K \cdot P(L)$.

3.7.3. Solution for a typical use case: $M = 192$, $B = 5000$ USD

We exhaustively evaluated all admissible (N, L) pairs. The complete set of feasible combinations is provided in Supplementary Table S3. The optimal solution is:

$N = 3$ M reads, $L = 600$ bp, $U = 5.2507$, total cost = 3910 USD, per-sample cost = 20.36 USD.

At this point, total data = $192 \times 3 \times 600/1000 = 345.6$ Gb, which fits into a single flow cell ($K = 1$). Increasing depth to $N = 4$ M would require two flow cells (total data 460.8 Gb > 435 Gb), raising the cost to 7820 USD, exceeding the budget. Shorter read lengths yield lower utility for the same depth (e.g., $L = 500$, $N = 3$ gives $U = 5.156$). Hence, the constrained optimum is robust.

Remark on the boundary solution. The optimal read length is always the maximum allowed by the platform ($L = 600$ bp) because, for any fixed depth N , the utility $U(N, L)$ increases monotonically with L within the range $L \leq 600$ bp (see Fig. 4b). Thus the platform’s upper limit acts as a binding constraint. If a longer read length were permitted (e.g., 1800 bp), it would further increase utility at the same N , but the associated data volume per sample would rise proportionally, requiring more flow cells and exceeding the budget for large M . Therefore, the boundary solution is a consequence of the finite budget and high multiplexing level, not a flaw of the framework. For smaller M (e.g., 48 samples), the same budget allows deeper sequencing ($N = 15$ M) at the same read length, achieving substantially higher utility ($U = 7.55$).

3.7.4. Sensitivity to the number of samples

Table 2 shows the optimal parameters for different sample numbers M under the same budget $B = 5000$ USD, obtained by solving the same constrained optimization.

Table 2: Optimal parameters for different numbers of multiplexed samples (budget 5000 USD, Illumina NovaSeq SP).

M	Optimal L (bp)	Optimal N (M reads)	Utility U	Total cost (USD)	Per-sample cost (USD)
48	600	15	7.55	3910	81.46
96	600	7	6.04	3910	40.73
192	600	3	5.25	3910	20.36

For smaller cohorts (e.g., $M = 48$), deeper sequencing ($N = 15$ M) becomes affordable within the same budget, yielding substantially higher utility ($U = 7.55$). This illustrates that the optimal experimental design strongly depends on the number of samples multiplexed.

3.7.5. Practical interpretation

The constrained optimum for large cohorts ($M = 192$) achieves only $U = 5.25$, far from the theoretical saturation value of 14.86. This indicates that under a moderate budget of 5000 USD, the information recovery is severely limited by the need to divide the flow cell capacity among many samples. If higher information recovery is required (e.g., for rare variant detection or de novo assembly), the budget must be increased (allowing multiple flow cells) or the number of multiplexed samples should be reduced. Our framework enables researchers to quantify these trade-offs quantitatively. Importantly, the fact that the optimal read length always lies at the platform maximum (600 bp) does not indicate a failure of the optimization; rather, it reveals the real-world reality that for a fixed budget and large cohort, any further increase in read length would be unaffordable due to the associated rise in data volume and flow cell requirements. The framework correctly identifies the binding constraints and provides actionable guidance. As the 600 bp extrapolation involves some uncertainty, researchers may also consider 500 bp as a conservative alternative, particularly if platform-specific confirmation is not available.

3.8. Summary of Key Findings

Table 3 summarises the main quantitative recommendations derived from our framework.

Table 3: Summary of optimal sequencing parameters and elasticities.

Metric	Value
Maximum utility ($q = 0$)	14.86 at $N = 20$ M, $L = 1800$ bp
Minimum standard error of U	0.00041
Region where $E_{N,L} \approx 1$ (N range)	4–20 M
Region where $E_{N,L} \approx 1$ (L range)	500–1000 bp
Constrained optimum for $M = 192$, $B = 5000$ USD	$N = 3$ M, $L = 600$ bp, $U = 5.25$

4. Discussion

Our *fractal economics* framework combines multifractal analysis with utility theory to optimize read length and sequencing depth in metagenomics. Three main findings emerge from the primary analysis.

First, the utility function reveals a characteristic three-phase pattern: a steep initial information gain from 100 bp to 200 bp—particularly pronounced in shallow sequencing—followed by diminishing marginal returns beyond 500 bp, and eventual saturation near 1800 bp. The depth dimension similarly exhibits diminishing returns beyond 5–10 M reads. These patterns echo empirical observations in metagenomic assembly [5, 6] and provide quantitative justification for cost-conscious design.

The pronounced 100→200 bp gain warrants further interpretation. In the Chaos Game Representation framework, this threshold likely reflects the read length required for the multifractal structure of genomic sequences to stabilize: below approximately 150 bp, the CGR point cloud may be dominated by statistical fluctuations in tetranucleotide frequencies; above 200 bp, the self-similar patterns corresponding to higher-order genomic features (e.g., repeat element distribution, GC content variation) begin to emerge reliably [9]. This interpretation is consistent with the observation that the 100→200 bp gain diminishes at higher depths, where the increased read count compensates for the shorter read length by providing more samples of the same compositional patterns. Thus, the steep initial rise is not merely a statistical artifact but reflects a genuine information-theoretic threshold in sequence complexity characterization.

Second, elasticities and marginal rates of substitution offer actionable guidelines. In low-depth scenarios with short reads, prioritizing read length is far more efficient than adding reads. Conversely, when deep sequencing is already planned, increasing read length beyond 1500 bp brings little benefit. The near-unit substitution rate in the 5–10 M / 500–1000 bp region allows flexible trade-offs depending on platform-specific costs.

Third, consistent with our earlier work [8], the multifractal dimension correlates strongly with classical diversity indices, validating that the fractal approach captures ecologically meaningful signal. Although the present study focuses on information recovery rather than taxonomic characterization, this established connection supports the ecological interpretability of the utility measure.

Extended insights. Our robustness analysis (different q values) shows that the optimal parameter set does not depend on the specific multifractal dimension used, strengthening confidence in the results. The uncertainty analysis demonstrates that even with only 100 replicates, the utility estimates are extremely stable, supporting the use of our utility surfaces as practical nomograms.

The near-perfect substitution corridor ($E_{N,L} \approx 1$) is wider in the N direction than in L , but we caution that at extreme parameter values (very low N or very high N and L), the ratio becomes

sensitive to numerical noise. Therefore, we recommend focusing on the central region ($N = 5\text{--}10\text{ M}$, $L = 500\text{--}1000\text{ bp}$) for balanced trade-offs, while in shallow sequencing with short reads, read length should be prioritised, and in ultra-deep sequencing depth should be slightly favoured.

Cost-benefit under real constraints. The constrained optimization (Section 3.7) demonstrates that the optimal parameters are highly sensitive to the number of multiplexed samples and the available budget. For a large cohort (192 samples) with a moderate budget (5000 USD), the best achievable utility is only 5.25, far from saturation. This highlights a critical practical reality: high multiplexing forces shallow per-sample sequencing, severely limiting information recovery. Researchers aiming for near-saturation utility must either reduce the number of samples per flow cell or increase the budget to use multiple flow cells. Our framework provides a quantitative tool to explore these trade-offs.

4.1. Boundary Effect and Definition of the Ideal Dimension

We carefully examined the minor non-monotonic behavior at $L = 2000\text{ bp}$ relative to $L = 1800\text{ bp}$ and identified its origin: the read sampling procedure (random starting positions without circularization) causes the extreme terminal regions of each genome to be under-sampled for read lengths that approach the genome length. This boundary effect, while negligible for practical purposes, can produce small fluctuations in utility when the deviation from D^{ideal} is already below 10^{-5} . To quantify this effect: for a typical bacterial genome of 5 Mb, a 2000 bp read undersamples the terminal $\sim 2000\text{ bp}$ region, affecting approximately $2000/5,000,000 = 0.04\%$ of the genome. The resulting utility fluctuation is on the order of 0.0002 log-units, well below the numerical noise threshold of 0.0004. Thus, the boundary effect does not materially affect any practical conclusion. We emphasize that our definition of $D^{\text{ideal}}(q)$ from complete (non-fragmented) genomes sets an absolute theoretical upper bound on information recovery. Using a circularised genome model or increasing the number of replicates beyond 100 would further reduce these fluctuations, but the current results already provide a clear and actionable conclusion: beyond $L = 1800\text{ bp}$ and $N = 10\text{ M}$ reads, information recovery is effectively saturated. Thus, the recommendation of $L = 1800\text{ bp}$ as a robust saturation point is well justified.

4.2. Towards a general theory of information recovery in complex systems

Although our framework was developed and validated specifically for metagenomic sequencing, its underlying principles extend far beyond this application. The core idea—using a scale-invariant measure (multifractal dimension) to quantify information content, defining a utility function based on deviation from an ideal asymptotic state, and deriving elasticities and substitution rates to guide resource allocation—is inherently general. Any experimental or observational system characterized by a trade-off between two or more resource dimensions and possessing a measurable information-theoretic signature could potentially benefit from a similar approach.

For instance, in ecological monitoring, one might optimize the trade-off between spatial coverage and temporal resolution when designing biodiversity surveys; in medical imaging, between resolution and acquisition time; in communication systems, between bandwidth and power. In each case, a suitable information-theoretic measure (e.g., entropy, mutual information, or multifractal dimension) could serve as the utility anchor, and the concepts of elasticity and marginal rate of substitution would directly apply. Our work thus provides not only a practical tool for metagenomic experimental design but also a template for a broader class of resource allocation problems in complex systems. We hope that future research will explore these generalizations and adapt the fractal economics framework to other domains.

Limitations include the use of simulated data without sequencing errors. Future work should incorporate real error models and integrate more detailed cost functions (e.g., library preparation, computational storage). Moreover, the boundary effect could be eliminated by implementing a circular genome model, which we intend to explore in subsequent studies. We also note that our analysis is based on a single high-diversity community composition; generalization to other communities—particularly those with different evenness or species counts—will require additional validation in future work.

5. Conclusion

The fractal economics framework provides a rigorous, data-driven method to balance read length and sequencing depth. The derived utility contours reveal a steep information gain at the 100–200 bp transition—particularly valuable for shallow sequencing—followed by diminishing returns beyond 500 bp and saturation near 1800 bp, with elasticities and substitution rates offering clear recommendations for designing cost-effective metagenomic experiments. Under realistic Illumina NovaSeq SP constraints, we show that the optimal parameters depend critically on the number of multiplexed samples and the total budget. For a typical large cohort (192 samples, 5000 USD), the constrained optimum is $N = 3$ M reads and $L = 600$ bp (2×300 bp paired-end), achieving utility $U = 5.25$ at a per-sample cost of about 20 USD. Researchers requiring higher information recovery should reduce multiplexing or increase the budget. The underlying multifractal analysis opens new avenues for alignment-free characterization of microbial communities.

Declarations

Abbreviations

CGR: Chaos Game Representation; NGS: Next-Generation Sequencing.

Ethics approval and consent to participate

Not applicable. This study did not involve human or animal subjects.

Consent for publication

Not applicable. This manuscript does not contain any individual person’s data.

Availability of data and materials

The whole-genome sequences of the ten bacterial species used in this study are publicly available from the National Center for Biotechnology Information (NCBI) GenBank database under the following accession numbers: *Akkermansia muciniphila* ATCC BAA-835 (NC_010655.1), *Alistipes shahii* WAL 8301 (NC_021030.1), *Bifidobacterium adolescentis* ATCC 15703 (NC_008618.1), *Bacteroides vulgatus* ATCC 8482 (NC_009614.1), *Coproccoccus* sp. ART55/1 (FP929039.1), *Eubacterium eligens* ATCC 27750 (NC_012778.1), *Faecalibacterium prausnitzii* A2-165 (ACOP02000001.1), *Lachnospiraceae bacterium* 1_4_56FAA (NZ_GL945163.1), *Prevotella copri* DSM 18205 (NZ_GG703878.1), and *Ruminococcus champanellensis* type strain 18P13T (NC_021039.1).

All source code for the multifractal analysis and simulation pipeline is available from the corresponding author upon reasonable request.

The supplementary result tables ('Fractal_Economics_Results.csv', 'feasible_combinations.csv', 'sensitivity_M_comparison.csv', and 'Extended_ENL_near_one.csv') are included in the Supplementary Information of this article.

Competing Interests

The authors declare that they have no competing interests.

Funding

This research was funded by the National Natural Science Foundation of China (grant nos. 12361099 and 12371088), and the Natural Science Foundation of Jiangxi Province (20212BAB201006).

Authors' contributions

Conceptualization, X.X., G.H. and Y.M.; methodology, X.X. and Z.Y.; software, X.X. and Y.H.; validation, X.X., Y.M., Z.Y. and G.H.; formal analysis, X.X. and Y.H.; resources, X.X. and G.H.; data curation, X.X. and Y.H.; writing—original draft, X.X. and Z.Y.; writing—review and editing, X.X. and Z.Y.; visualization, X.X.; supervision, Z.Y. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

We thank the members of the Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education and Hunan Key Laboratory for Computation and Simulation in Science and Engineering for their support. We are grateful to the reviewers for their constructive comments, which significantly improved the manuscript.

References

- [1] J. Shendure, S. Balasubramanian, G. M. Church, et al., Dna sequencing at 40: past, present and future, *Nature* 550 (7676) (2017) 345–353.
- [2] M. Byrska-Bishop, U. S. Evani, X. Zhao, et al., High-coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios, *Cell* 185 (18) (2022) 3426–3440.e19.
- [3] J. K. Sicklick, S. Kato, R. Okamura, et al., Molecular profiling of cancer patients enables personalized combination therapy: the i-predict study, *Nat. Med.* 25 (5) (2019) 744–750.
- [4] D. Bertrand, J. Shaw, M. Kalathiyappan, et al., Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes, *Nat. Biotechnol.* 37 (8) (2019) 937–944.
- [5] L. X. Chen, K. Anantharaman, A. Shaiber, et al., Accurate and complete genomes from metagenomes, *Genome Res.* 30 (3) (2020) 315–333.
- [6] M. Kolmogorov, D. M. Bickhart, B. Behsaz, et al., metaflye: scalable long-read metagenome assembly using repeat graphs, *Nat. Methods* 17 (11) (2020) 1109–1115.

- [7] T. Royalty, A. D. Steen, Simulation-based approaches to characterize the effect of sequencing depth on the quantity and quality of metagenome-assembled genomes, *bioRxiv* (2018). [doi:10.1101/356840](https://doi.org/10.1101/356840).
- [8] X. H. Xie, Y. J. Huang, G. S. Han, et al., Microbial characterization based on multifractal analysis of metagenomes, *Front. Cell Infect. Microbiol.* 13 (2023) 1195733.
- [9] H. J. Jeffrey, Chaos game representation of gene structure, *Nucleic Acids Res.* 18 (8) (1990) 2163–2170.
- [10] C. L. Berthelsen, J. A. Glazier, M. H. Skolnick, Global fractal dimension of human dna sequences treated as pseudorandom walks, *Phys. Rev. A* 45 (12) (1992) 8902–8913.
- [11] L. Q. Zhou, Z. G. Yu, J. Q. Deng, et al., A fractal method to distinguish coding and noncoding sequences in a complete genome based on a number sequence representation, *J. Theor. Biol.* 232 (4) (2005) 559–567.
- [12] P. E. Vélez, L. E. Garreta, E. Martínez, et al., The *Caenorhabditis elegans* genome: a multifractal analysis, *Genet. Mol. Res.* 9 (2) (2010) 949–965.
- [13] P. A. Moreno, P. E. Vélez, E. Martínez, et al., The human genome: a multifractal analysis, *BMC Genomics* 12 (2011) 506.
- [14] M. H. Olyaei, J. Pirgazi, K. Khalifeh, et al., Rcovid19: Recurrence-based sars-cov-2 features using chaos game representation, *Data Brief* 32 (2020) 106144.
- [15] M. Meraz, E. J. Vernon-Carter, E. Rodriguez, et al., A fractal scaling analysis of the sars-cov-2 genome sequence, *Biomed. Signal Process. Control* 73 (2022) 103433.
- [16] S. Vinga, J. Almeida, Alignment-free sequence comparison—a review, *Bioinformatics* 19 (4) (2003) 513–523.
- [17] G. E. Sims, S. R. Jun, G. A. Wu, S. H. Kim, Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions, *Proc. Natl. Acad. Sci. USA* 106 (8) (2009) 2677–2682.
- [18] V. Dubinkina, D. Ischenko, V. Ulyantsev, et al., Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis, *BMC Bioinformatics* 17 (2016) 38.
- [19] W. Huang, L. Li, J. R. Myers, G. T. Marth, Art: a next-generation sequencing read simulator, *Bioinformatics* 28 (4) (2012) 593–594.
- [20] H. Li, B. Handsaker, A. Wysoker, et al., The sequence alignment/map format and samtools, *Bioinformatics* 25 (14) (2009) 1754–1760.
- [21] Whitehead Institute, Genome technology core pricing, <https://wi.mit.edu/core-facilities/genomecore/pricing>, accessed: 2026-06-15.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SUPP.pdf](#)