Nucleic Acids Research, 2025, **53**, gkaf793
https://doi.org/10.1093/nar/gkaf793
Computational Biology

Optimal sequencing depth for measuring the concentrations of molecular barcodes

Tommaso Ocari[1,*], **Emilia A. Zin**[1], **Muge Tekinsoy**[1], **Timothé Van Meter**[1], **Mélissa Desrosiers**[1], **Chiara Cammarota**[2,3], **Deniz Dalkara**[1], **Takahiro Nemoto**[1,4,†], **Ulisse Ferrari**[1,†]

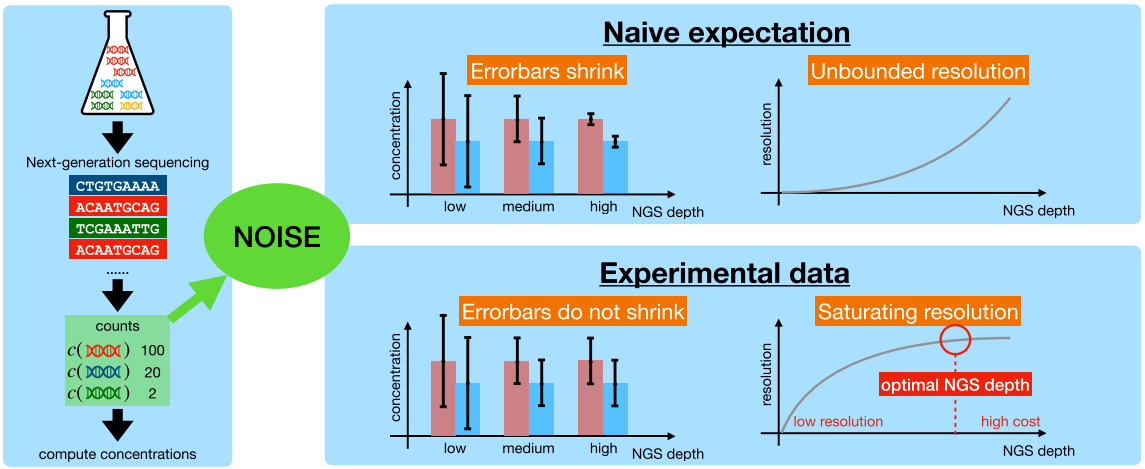
[1]Institut de la Vision, Sorbonne Université, INSERM, CNRS, 17 rue Moreau, 75012 Paris, France
[2]Physics department, University of Rome 'La Sapienza', Piazzale Aldo Moro 5, 00185 Rome, Italy
[3]INFN, sezione di Roma1, Piazzale Aldo Moro 5, 00185 Rome, Italy
[4]Premium Research Institute for Human Metaverse Medicine (WPI-PRIMe), Osaka University, Suita, Osaka 565-0871, Japan

*To whom correspondence should be addressed. Email: tommy.ocari@gmail.com
†The last two authors should be regarded as Joint Last Authors.

Abstract

In combinatorial genetic engineering experiments, next-generation sequencing (NGS) allows for measuring the concentrations of barcoded or mutated genes within highly diverse libraries. When designing and interpreting these experiments, sequencing depths are thus important parameters to take into account. Service providers follow established guidelines to determine NGS depth depending on the type of experiment, such as RNA sequencing or whole genome sequencing. However, guidelines specifically tailored for measuring barcode concentrations have not yet reached an accepted consensus. To address this issue, we combine the analysis of NGS datasets from barcoded libraries with a mathematical model taking into account the polymerase chain reaction amplification in library preparation. We demonstrate on several datasets that noise in the NGS counts increases with the sequencing depth; consequently, beyond certain limits, deeper sequencing does not improve the precision of measuring barcode concentrations. We propose, as rule of thumb, that the optimal sequencing depth should be about ten times the initial amount of barcoded DNA molecules before any amplification step.

Graphical abstract



Introduction

Since its commercialization in the early 2000s, next-generation sequencing (NGS) allows to sequence millions of DNA molecules in parallel [1, 2], revolutionizing nucleic acid research. Since then, NGS cost has constantly dropped [3], and today it represents a powerful and accessible tool for many research investigations. In protein engineering, NGS has radically changed experimental approaches to study the fitness landscape of mutant proteins, as in deep mutational scanning (DMS) experiments [4–8]. It has also shifted the paradigm of designing new proteins with augmented functions in systematic evolution of ligands by exponential enrichment [9–13] or in directed evolution (DE) [14–20]. In these experiments, highly diverse gene libraries are initially prepared,

Received: June 14, 2024. **Revised:** July 3, 2025. **Editorial Decision:** July 11, 2025. **Accepted:** August 20, 2025

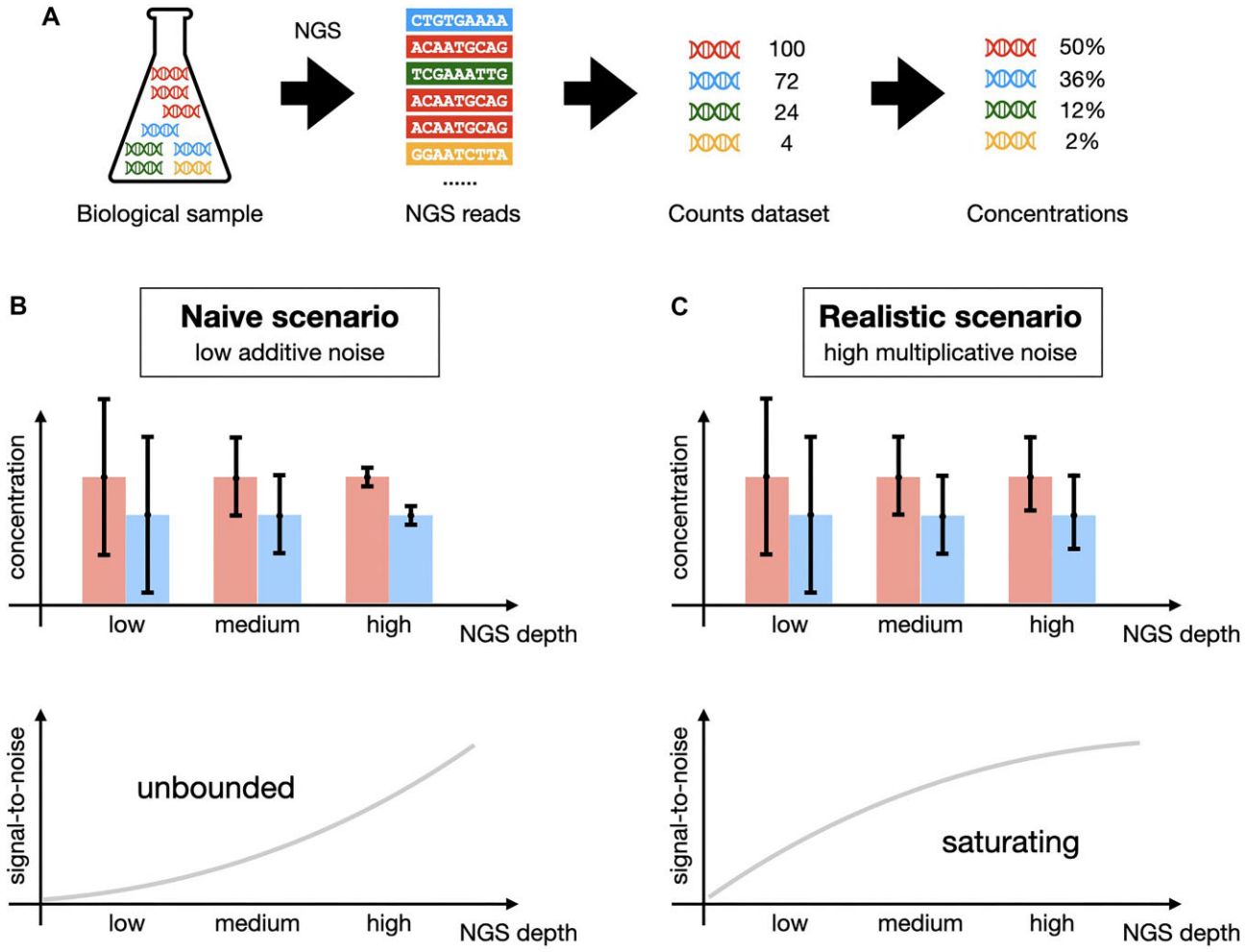


Figure 1. Statistical errors in sequencing analysis. (**A**) NGS is often used to compute concentrations of different species or molecular barcodes in highly diverse DNA libraries. (**B**) Naïve scenario underestimates noise in NGS counts, expects that statistical errors can be reduced by enlarging NGS depths, and therefore the signal-to-noise ratio be increased indefinitely. (**C**) In the realistic scenario evinced in this work, errors do not always shrink with deeper NGS, and signal-to-noise saturate after a certain depth.

where each gene is tagged with a unique molecular barcode or a distinctive mutation. Then, the fitness of each variant is assessed by measuring the relative concentration of each barcode using NGS (Fig. 1A), before and after one or multiple screening rounds. For example, in DMS experiments, a large library with thousands, if not millions of mutants undergo a screening [5], and NGS before and after a selection experiment distinguishes between beneficial and detrimental mutations, as those that increase or decrease their concentration [6, 7].

Being able to precisely measure the concentrations of the library content is therefore crucial. For this reason, the number of NGS reads (NGS depth) is an important parameter to both the design and interpretation of the experiments. In other applications of NGS, such as whole genome sequencing or RNA sequencing, the optimal NGS depth has been historically discussed intensively in terms of coverage [21], and guidelines have been established https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html. Yet, for the measurements of barcode concentrations, such guidelines are not yet available and the NGS depth is often pushed to the limit, resulting in non-negligible costs for the laboratories. Here, we point out that this is not beneficial for most cases. Increasing the NGS depth does not always reduce the impact of random sampling and therefore does not shrink the statistical errors (*naïve scenario*; Fig. 1B).

A standard workflow to prepare a DNA-sample library for NGS includes polymerase chain reaction (PCR) amplifications. Even highly reliable PCRs introduce errors during DNA replication, and they are of different nature. During a PCR amplification, some DNA molecules might not get replicated, or some wrong nucleotides may be inserted randomly [22]. If this happens at the beginning of the amplification protocol then the error will be amplified as well [23–27]. In this work we first analyze the structure of these amplification biases in NGS datasets, to then develop a mathematical model that accounts for it. Built on previous works [28–33], our model accounts for how PCR mistakes impact the whole NGS process. This analysis showcases that, by increasing the NGS depth, both signal and noise increase in the estimation of concentration in highly diverse libraries. As a consequence, beyond a certain depth—~10 times the initial amount of molecules of barcoded DNA (from now on simply called initial DNA amount)—deeper NGS does not increase data quality (*realistic scenario*, Fig. 1C), and this sets an optimal number of reads

Table 1. Summary of datasets

Dataset	Length	Unique sequences	NGS depth	Probability distribution
AAV dataset 1 [27]	7 amino acids	1252 032	13 653 491	Biased
AAV dataset 2 [34]	7 amino acids	3071 451	12 195 473	Nucleo-flat
AAV dataset 3	7 amino acids	98 468 090	240 537 294	Amino-flat
hYAP65 WW rep 1 [36]	99 nucleotides	1336 842	3588 061	Error-prone PCR
hYAP65 WW rep 2 [36]	99 nucleotides	1615 935	4931 094	Error-prone PCR
scRNAseq-UMI 1 [37]	16 nucleotides	14 512 602	75 820 168	Nucleo-flat
scRNAseq-UMI 2 [38]	16 nucleotides	16 020 053	173 326 275	Nucleo-flat
scRNAseq-UMI 3 [39]	16 nucleotides	16 429 852	312 213 184	Nucleo-flat
scRNAseq-UMI 4 [40]	16 nucleotides	14 974 774	182 330 818	Nucleo-flat

Nucleo-flat (amino-flat) means that the nucleotides (amino-acids) showed similar frequencies across different position. The first dataset was biased toward an high adenine content, resulting in an highly heterogeneous amino-acid content. The position-independent model is used to analyze AAV and scRNAseq datasets, whereas the mutation model is used for hYAP65 datasets.

that need to be chosen to obtain the most from the data, without wasting resources.

Materials and methods

Dataset characteristics

In this work we considered nine barcode libraries with different characteristics (Table 1). Three datasets (AAV dataset 1, 2, 3) used random 7-mer peptides as barcodes, inserted between amino acid 587 and 588 of the cap2 gene in the adeno-associated viruses serotype 2 (AAV2) plasmid. AAV dataset 1 used random 7-mer peptides composed of six VNN codons (N stands for any base, V stands for any base but thymine) and one VNG codon (G stands for guanine only) [27], where the distribution of amino acids is biased towards a higher abundance of lysine (Supplementary Fig. S1). AAV dataset 2 used random 7-mer peptides of seven NNK codons (K stands for guanine or thymine base), where all admitted nucleotides are designed to have the similar occurrence. This led to relatively flat amino-acid distributions [34] (Supplementary Fig. S1). AAV dataset 3 is our in-house dataset prepared using Twist Bioscience® as described below, and resulted in the most homogeneous occurrence of all amino-acids at all of the seven positions (Supplementary Fig. S1) among the three datasets. NGS was used to sequence the barcoded regions of these libraries, where the number of unique sequences and NGS depths are summarized in Table 1. Other two datasets (hYAP65 WW rep 1, 2) are made of barcodes of the mutated WW domain of the hYAP65 protein (99 nucleotides of variable region). These datasets were generated through error-prone PCR, so their counts decrease with the number of mutations from the wild-type sequence. The last four datasets (scRNAseq-UMI 1, 2, 3, 4) used as barcodes the unique molecular identifiers (UMI) of publicly available single-cell-RNA-sequencing datasets from 10X Genomics database. We observed that nucleotide distributions are relatively uniform, though Guanine-Cytosine ratio (GC-ratio) is <50% in all of scRNAseq-UMI datasets (Supplementary Fig. S1). Note that all nine datasets are made of oligos (barcodes). The different names are related to the template where the barcodes are inserted. However the templates do not influence the analysis which follows here.

For AAV datasets, cutadapt [35] was used to extract sequences that correspond to random 7-mer peptides from FASTQ files. The average error rates are estimated for each sequence from Q score and those that have >0.1 error rates are removed for further analysis. Translating nucleotide sequences into amino acid sequences, seven amino acids are obtained as barcodes. In this step, sequences containing stop codon and amino acids that cannot be formed due to the codon design are removed. For the hYAM65 datasets forward and reverse reads were compared and if the two reads match the entire sequence was kept in the dataset. For scRNAseq-UMI datasets, BAM files are downloaded from the 10X Genomics database, and UMI with UR tags consisting of 16 nucleotides are extracted. These UMIs are used as barcodes.

Amino-flat library preparation

We ordered 50 μg of a AAV plasmid library from Twist Bioscience® that used silicon-based DNA synthesis platform. A total of 10 amino acids were inserted at position 587/588 in the AAV2 Cap sequence, where three of them were the AA/A linkers and seven of them, flanked by AA/A, were random peptides consisting of a diversity of 10^8–10^9 variants. The random peptides were inserted in the way that each amino acid occurred at the same probability in each position. We requested that the restriction sites for restriction enzymes HindIII, NotI, and AscI be kept in the original AAV2 Rep/Cap plasmid sequence to permit future cloning.

Ten nanograms of the 7-mer library was PCR amplified using primers 5′-ATCAGGACAACCAATCCCGTGGCTA-3′ and 5′-TGTCCTGCCAGACCATGCCTG-3′ with Takara Bio PrimeSTAR GXL high-fidelity polymerases. After the PCR amplification, reactions were cleaned-up with a Macherey-Nagel Gel Extraction and PCR Clean-Up kit. Samples were submitted to Plateforme GENOM'IC at Institut Cochin for sequencing. They were sequenced with Illumina NextSeq 2000, P3 1*200 cycles, at 200 million reads per sample.

A position-independent model

To analyze the data AAV and scRNAseq, we use a position-independent or position weight model [41], which serves as a proxy for the original concentrations of each barcode in the samples. In this model, we assume that each barcode sequence $s = s^1 s^2 \ldots s^L$, where L is the length of the barcode, is generated from the following probability: $P(s) = \prod_{i=1}^{L} f_i(s^i)$, where $f_i(s^i)$ is the frequency of s^i at the position i estimated from each dataset. To analyze the AAV datasets, we use amino acids sequence: $s^i \in$ (20 amino acids) with $L = 7$, whereas for the scRNAseq-UMI datasets, we use nucleotides sequence: $s^i = T, A, G, C$ with $L = 16$.

A mutation model

To analyze the data of error-prone PCR we use a position-independent mutation model. In this model we assume that each barcode sequence is uniquely described by the number and the type of mutations from the wild-type barcode $\{(a_i, b_i)\}_{i=1}^{n}$ where n is the number of total mutations and a_i is the nucleotide of the wild-type which is mutated into the nucleotide b_i. The parameters of the models are the vector $p(n)$ which is the probability of having exactly n mutations and the matrix $M(a, b)$ which gives the probability of replacing a nucleotide $a = A, C, G, T$ with the nucleotide $b \neq a$. The probability of a barcode to be in the dataset is indeed $P\left(\{(a_i, b_i)\}_{i=1}^{n}\right) \propto p(n) \prod_{i=1}^{n} M(a_i, b_i)$.

Calculation of the noise extent (Fano factor) from data

Either using the position-independent model or the mutation model, the expected count $\lambda(s)$ of a barcode sequence s is computed by multiplying the sequence probability $P(s)$ by the sequencing depth D: $\lambda(s) = DP(s)$. Note that the expected count can be estimated for all possible barcode sequences, including those that have zero observed counts in the dataset. Next, barcode sequences are binned based on their expected count $\lambda(s)$ in N_b bins, such that each bin has the same number of barcode sequences with nonzero observed counts. Note that the binning would have been infeasible on the observed counts given their discrete nature. The average value and variance of the observed counts in each bin are denoted by $\langle c \rangle_{\text{bin}}$ and $\sigma^2_{c,\text{bin}}$. Similarly, those of the expected counts $\lambda(s)$ in each bin are denoted by $\langle \lambda \rangle_{\text{bin}}$ and $\sigma^2_{\lambda,\text{bin}}$. Using these quantities, the extent of noise in the data is computed as the ratio between variance and mean, commonly referred to as the Fano factor (FF). The FF is a quantity related to the amount of dispersion we have in the data, since a higher FF corresponds to a larger variance given the same mean. Here, taking into account the variance of λ in each bin, we estimate it as

$$\text{Noise extent } (\text{FF}_\text{E}) = \frac{\sigma^2_{c,\text{bin}} - \sigma^2_{\lambda,\text{bin}}}{\langle c \rangle_{\text{bin}}},$$

where the subscript E in FF_E indicates an empirical estimation. Note that the difference at the numerator is due to the fact that the expected count λ is also a random variable in the bin, so we needed to subtract the variance of this variable. Nevertheless, this is a minor correction, which is negligible for most of the bins. Note that, in AAV dataset 3 and scRNAseq-UMI 1, 2, 3, and 4, $P(s)$ is almost uniform because all amino acids (AAV dataset 3) or all nucleotides (scRNAseq-UMI 1, 2, 3, 4) have a similar probability (Supplementary Fig. S1). Therefore, we set $N_b = 1$ in these cases.

Mathematical model of the amplification process

The PCR amplification process is numerically simulated by computing the average number of DNA molecules along the cycles. The model comprises the following three steps:

- An amount N_0 of the DNA molecules is taken from the initial sample. Because this mimics pipetting a part of the sample, we assumed that this random selection is not correlated, and consequently we modeled it as Poisson sampling.
- The sample undergoes PCR amplification processes, where each DNA molecule in the sample is replicated with a probability $p_n < 1$ in each round n. Errors that change the sequences are not taken into account in this framework. We modeled this process as a series of binomial processes.
- Finally, D random molecules are sequenced using NGS. We modeled it again as a Poisson process.

Note that we use the following way of computing collective quantities at each amplification step without running the complete simulation of the amplification process, which is computationally heavy. We will focus on the case where $p_n = (1 + N_n/K_{\text{MM}})^{-1}$ where N_n is the number of DNA molecules at the n-th step, and K_{MM} is called Michaelis–Menten (MM) constant. The relation between the noise extent (FF), the NGS depth D and the initial DNA amount N_0 is computed as $\text{FF}_\text{M} = 1 + a_n D/N_0$ (see the 'Results' section, Eq. 1), where a_n depends *a priori* on the number of amplification rounds n and the subscript M in FF_M stands for an estimation based on the model. The method takes in input the initial reduced DNA amount N_0/K_{MM} and it outputs, for each of 100 rounds of amplification, the replication probability p_n, the reduced DNA amount N_n/K_{MM} and the values of the function a of equation 1, here labeled as a_n. It proceeds in an iterative way:

- The reduced DNA amount is initialized to N_0/K_{MM}.
- The function a_n is initialized to the unity $a_0 = 1$.
- The replication probability is computed iteratively as $p_n = (1 + N_{n-1}/K_{\text{MM}})^{-1}$.
- The function a_n is computed iteratively as $a_n = 1 + (1 + p_n) \times (a_{n-1} - 1) + [p_n(1 - p_n)/(1 + p_n)] \times [(N_0/K_{\text{MM}})/(N_n/K_{\text{MM}})]$.
- The reduced DNA amount is updated iteratively as $(N_{n+1}/K_{\text{MM}}) = (1 + p_n) \times (N_n/K_{\text{MM}})$.

Dataset resolution and optimal depth

We start defining the signal-to-noise (snr) ratio as the mean expected count of each variant divided by the square root of its variance: $\text{snr} = P(s)D/\sqrt{(P(s)D\,\text{FF}_\text{E})} = \sqrt{P(s)D/\text{FF}_\text{E}}$, where $P(s)$ is the probability (concentration) of the variant s, D is the sequencing depth, and FF_E is the empirical noise extent (FF). Note that snr is a variant dependent quantity, but the dependence is only via the term $\sqrt{P(s)}$. By normalizing with respect to it, we can define a resolution as $\text{res} = \sqrt{D/\text{FF}_\text{E}}$.

Replacing FF_E by FF_M in res as a working hypothesis, and using our formula for the noise extent (Eq. 1) with $a = 1$, the resolution can be computed as $\sqrt{DN_0/(N_0 + D)}$. We then define the exploited resolution as the ratio between the resolution and its large D limit ($\sqrt{N_0}$) to obtain $\text{exp-res} = \sqrt{D/(N_0 + D)} = \sqrt{D/N_0/(1 + D/N_0)}$. The exploited resolution varies between 0 and 1, depending only on the ratio D/N_0. The optimal depth is computed as the depth D for which the exploited resolution reaches 95%. Alternatively, we could have defined an exploited snr, by dividing the snr by its large D limit, obtaining $\text{exp-snr} = \sqrt{D/(N_0 + D)}$, which is equal to the exploited resolution.

Inference of the number of DNA molecules before amplification

We used Eq. 1 of the Result to estimate the number of initial DNA molecules (N_0) from the sequencing depth (D) and the noise extent (FF_M). In the case of replication probability logistically dependent on the amplification rounds, it is possible to

approximate $a_n \sim 1$. Given noise extent (FF_M) and NGS depth D, the DNA amount before amplification N_0 is estimated as $D/(FF_M - 1)$.

Results

Mean and variance of NGS counts are proportional

We started our analysis by focusing on three barcode datasets where a simple machine learning model provides accurate predictions of the concentrations, and compared them with empirical estimations from NGS counts. We first analyzed AAV datasets 1 and 2, whose NGS depths were ∼13.6 and 12.2 million reads, resulting in 1.2 and 3.0 millions different observed sequences (Table 1). For each dataset, we inferred a position-independent model (see the 'Materials and methods' section), and computed the predicted count for all possible sequences, which amounts to 0.18 and 1.28 billion for AAV datasets 1 and 2, respectively. In order to verify the accuracy of model predictions, we compared the measured counts with those predicted by the model (Fig. 2B). Strong positive correlations were observed in sequences with the empirical count higher than 100. In contrast, sequences with lower empirical counts showed no clear correlations, likely due to their inherently noisy nature. We then sorted the sequences into 200 bins based on their expected counts, grouping those with similar values together (see the 'Materials and methods' section). For each bin, the mean of the predicted counts was very close to that of the empirical counts (Fig 2C), showcasing the accuracy of the model predictions. The count variances within each bin, on the other hand, were approximately proportional to the corresponding count means (Fig 2D). We estimated the noise extent of the counts as the FF (FF_E; see the 'Materials and methods' section), which corresponds to the variance divided by the mean. The FF_E serves as a measure of noise, with higher values indicating higher variance for a given mean. The noise extent has been computed in each bin, and found that it was significantly larger than 1, indicating that the counts were overdispersed with respect to Poisson noise. The deviation from the proportionality law shown at very high count mean (Fig 2D) is due to the approximations of the model [42] and can at least be decreased by using a more complex model (Supplementary Fig. S3).

The analysis of the AAV dataset 2 provided similar results. Because empirical counts are lower, we did not observe clear correlations between the empirical and predicted counts (Fig. 2E). However, after sorting the sequences into 80 bins based on their expected counts, model predictions of the mean count were well aligned with the empirical values (Fig. 2F). Noise extent (FF_E) was also significantly larger than 1 (Fig. 2G), indicating that the count data were overdispersed.

We finally perfomed the same analysis on the dataset hYAP65 WW rep 1. To compute the predicted counts we inferred a mutation model based on the number and the type of mutations from the wild-type sequence (Fig. 2H; see the 'Materials and methods' section). Note that the predicted counts are sharply clustered by the number of mutations from the wild-type sequence (different color of Fig. 2H). The noise extent (FF_E) resulted only slightly higher than 1 (Fig. 2J and inset). Similar results came from the analysis of the second replicate (hYAP65 WW rep 2; Supplementary Fig. S2).

Mathematical model and simulations reproduce the mean-variance proportionality observed in data

In order to increase our understanding of the empirical results found in the previous section, we developed a three-step mathematical model (see the 'Materials and methods' section) that mimics the NGS protocol including the PCR amplification process, as detailed in Fig. 3A.

We first analyzed the model under the simplified assumption of a constant replication probability $p_n = p$ across the amplification rounds. Using the theory of Galton–Watson process [43], we computed the noise extent analytically (Supplementary Fig. S4 and Supplementary mathematics):

$$\text{noise extent } (FF_M) = \left(1 + a\frac{D}{N_0}\right), \tag{1}$$

where D is the NGS depth, N_0 is the initial number of DNA molecules (from now on also called initial DNA amount) before the PCR and here a is a constant. We use the subscript M in FF_M to distinguish it from the empirically estimated FF (FF_E). This simple mathematical model recovered the proportionality relationship between the mean and variance of the NGS count that we observed in experimental data (Fig. 2). Interestingly, here $a \approx 2/(1 + p)$ and the noise extent (FF_M) turned out to be independent from the number of amplification rounds. This observation suggests that reducing the number of amplification (such as PCR) rounds does not reduce the noise extent of the final NGS reads (see the 'Discussion' section).

In order to verify that our analytical results were robust, we relaxed the constant p assumption by accounting for the presence of limiting replication factors, such as the quantity of polymerases or primers in PCR. Inspired by previous enzymatic descriptions of PCR process [28–30], we adopted a replication probability following a MM discretized equation (Fig. 3B): $p_n = (1 + N_n/K_{MM})^{-1}$, where N_n is the total number of DNA molecules present at the beginning of the n-th amplification step (DNA amount) and K_{MM} is a parameter linked to the limiting factors, known in the literature as the MM constant.

We derived an analytical expression for the noise extent in the case of a general replication probability (Supplementary mathematics), and found that it has the same analytical form as before (Eq. 1). Numerical evaluation of the analytical solution with MM replication probability indicates that $a \approx 1$ holds for a wide range of realistic cases, and in particular for both K_{MM} much smaller or much larger than N_0 (Supplementary Fig. S5). The latter is consistent with the limiting case without replication, for which we obtained $a = 1$ exactly. In the first case, instead, the sample undergoes an initial noiseless replication ($p \approx 1$) that increases the initial available DNA and therefore shrinks any variability.

To verify our analytical results we conducted numerical simulations of our mathematical model. We initiated the process with a synthetic heterogeneous library containing 20^7 different sequences, with concentrations given by the inferred probabilities from AAV dataset 2 (Fig. 2E–G). We sampled approximately $N_0 \approx 1.5 \times 10^6$ DNA molecules, and we computed the replication probability and the DNA amount for each round up to 30 amplification rounds (Fig. 3B and C). We chose $K_{MM} \approx 1.5 \times 10^{11}$ for the MM constant, in order to have an initial *reduced* DNA amount of $N_0/K_{MM} \approx 10^{-5}$. This value of the MM constant lies in the range of realistic

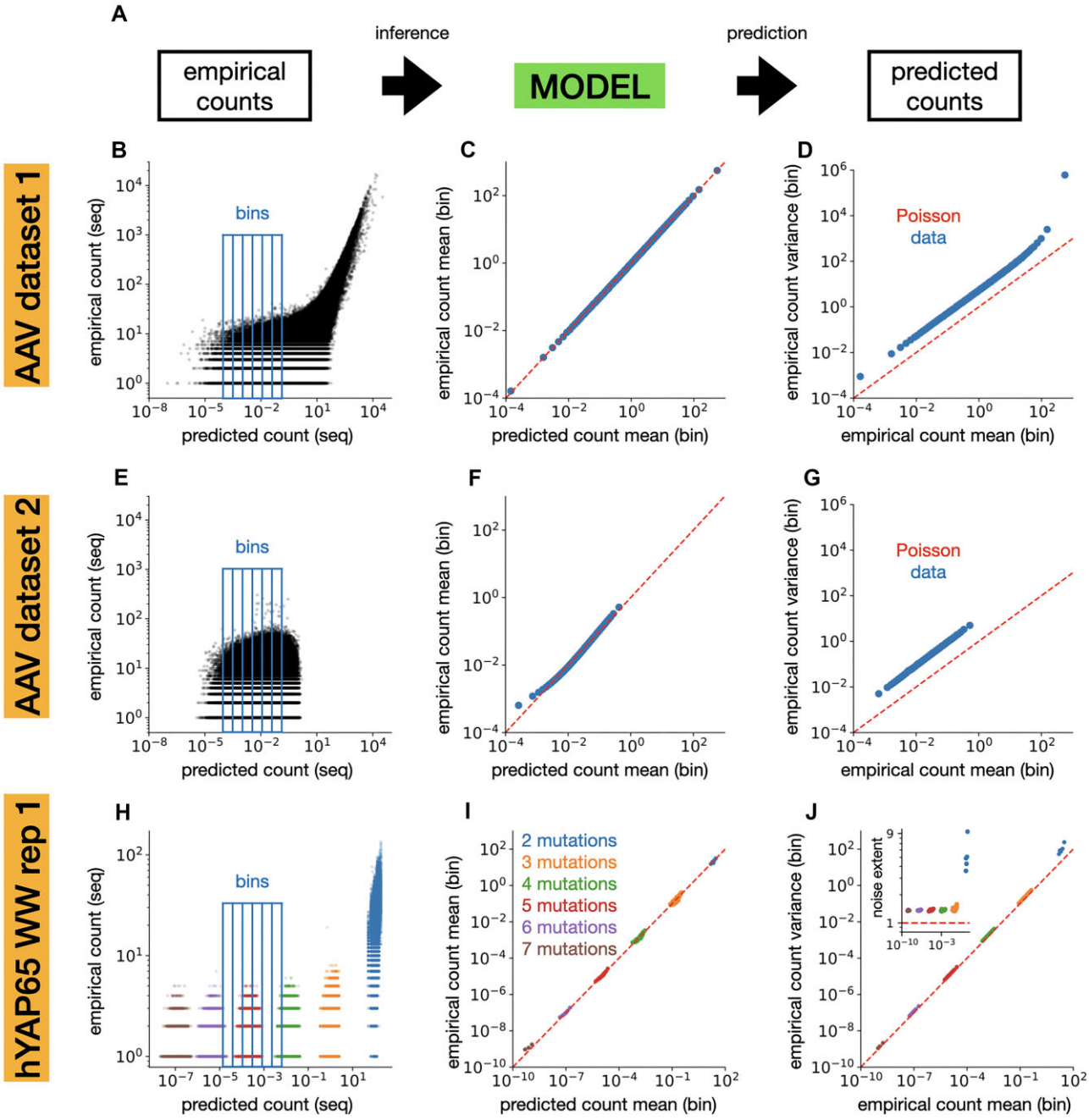


Figure 2. Mean and variance of the counts are proportional in NGS datasets. (**A**) Starting from the observed counts of molecular barcodes in each NGS dataset we fitted a statistical model to predict the expected count of each sequence. (**B**) AAV dataset 1. Empirical counts plotted against the expected counts for each sequence are shown (black points). Examples of bins grouping sequences with similar predicted counts are also shown as blue boxes. (**C**) AAV dataset 1. Empirical counts and predicted counts were averaged over each bin and are plotted against each other (blue dots). These points align along the equality line (red). (**D**) AAV dataset 1. Variance of the counts against the expected count (where both are averaged over each bin) is displayed by blue dots. In Poisson noise, count variance and expected count are equal to each other (red dashed line). The blue dots align to a line parallel to, yet above, the equality line, thus indicating overdispersion. (**F**)–(**G**) Same as panels (B)–(D), but use AAV dataset 2. (**H**) Same as panel (B) but use hYAP65 rep 1 and its corresponding mutation model. (**I**) Same as panel (F) but use hYAP65 rep 1 and its corresponding mutation model. Different colors correspond to different number of mutations from the wild-type sequences. (**J**) Same as panel (G) but use hYAP65 rep 1 and its corresponding mutation model. Different colors correspond to different number of mutations from the wild-type sequences. In the insight the FF_E (noise extent) for each bin plotted against the mean of the empirical counts.

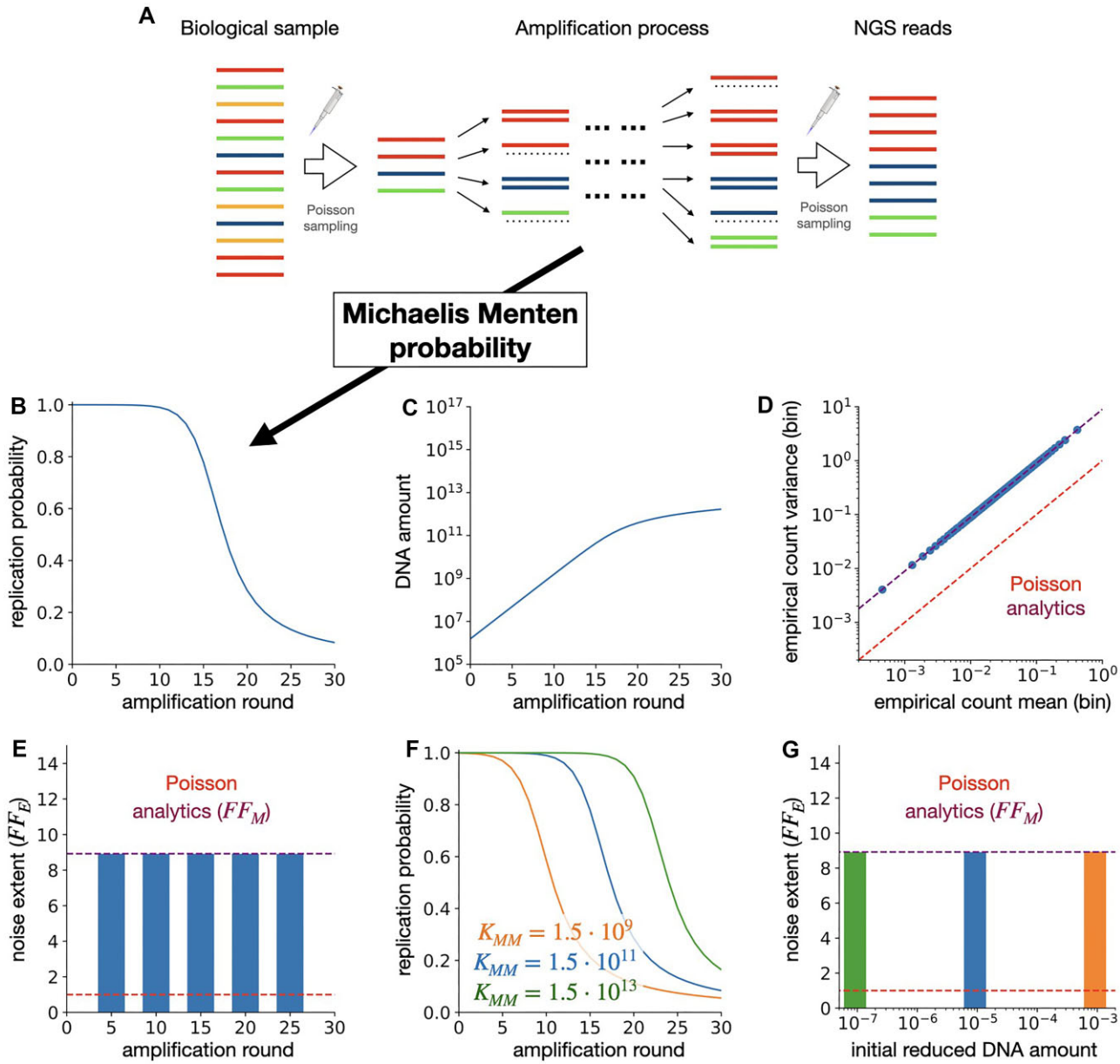


Figure 3. Mathematical model and simulations reproduce the mean-variance proportionality observed in data. (**A**) The three steps of the mathematical model describing the NGS procedure including PCR amplifications: a Poisson sampling is followed by multiple rounds of a binomial process and concluded by another Poisson sampling. (**B**) The replication probability against the number of rounds when the MM constant K_{MM} is 1.5×10^{11}. (**C**) The total number of DNA barcodes for each round of amplification was simulated, starting from $\sim 10^6$ DNA molecules at the 0th round. (**D**) The empirical count variance against the empirical count mean for each bin of a simulated dataset (blue points) shows overdispersion when compared with Poisson noise (dashed red line) and agrees with the analytical prediction (purple dashed line). (**F**) Noise extent (FF_E) against the number of amplification rounds confirms its independence on the amplification process. **(F)** The replication probability against the number of rounds for three different MM constant (1.5×10^9 in orange, 1.5×10^{11} in blue, 1.5×10^{13} in green). (**G**) The noise extent (FF_E) is plotted against the initial reduced DNA amount (N_0/K_{MM}) for the three different MM constants (as shown in the panel f). It suggests the negligible impact of the MM constant in this realistic range.

PCR, which has been studied to be in between 10^6 and 10^{15} [44]. Finally, the library was sequenced by sampling the counts with a NGS depth of $D \approx 1.4 \times 10^7$ reads. The resulting synthetic NGS data were binned over their expected counts - as done in the previous section on real data. We observed an overdispersed behavior with a count variance proportional to the mean, in agreement with analytical predictions (Fig. 3D).

To validate the independence of the noise extent (FF_M) from the amplification processes, we ran five simulations with the same parameters but varying the number of amplification rounds (5, 10, 15, 20, and 25), we computed the noise extent (FF_M) for each case (see the 'Materials and methods' section) and we found similar values for each simulation (Fig. 3E), according to our analytical results.

Finally, to better understand the impact of the MM constant K_{MM}, we ran three simulations with three different values ($K_{MM} \approx 1.5 \times 10^9$, 1.5×10^{11}, and 1.5×10^{13}). As expected, increasing the constant only shifted the behavior of the replication probability (Fig. 3f). Noise extent was instead constant, in agreement with our prediction (Eq. 1, Fig. 3G).

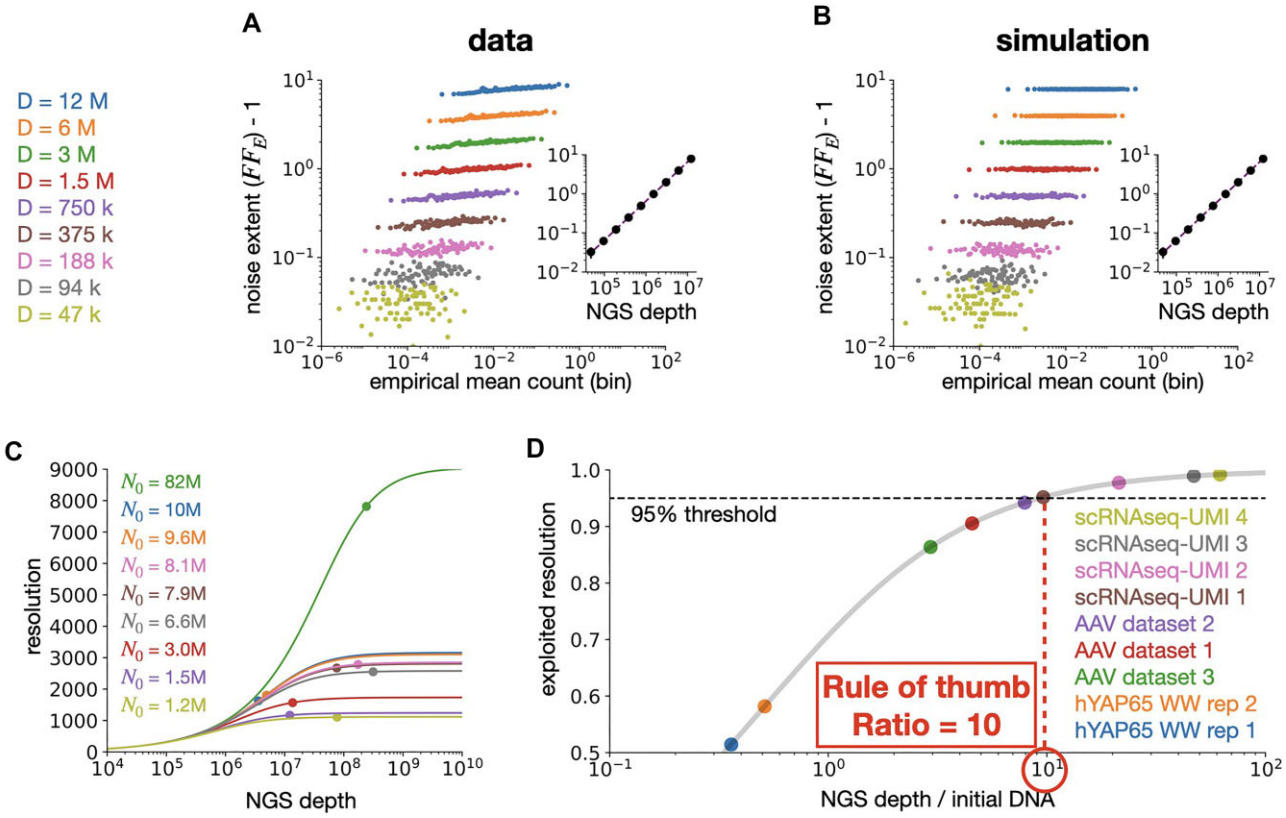


Figure 4. Which is the optimal NGS depth? (**A**) Eight synthetic datasets were generated from AAV dataset 2 by subsequently halving the NGS depth. Noise extent (FF_E) − 1 is plotted against the average expected counts per bin for each dataset (different colors). Inset: FF_E − 1 against the NGS depth aligns with the analytical predictions (purple dashed line). (**B**) The same figure as in panel (A), but obtained from simulating our mathematical model. The parameters for the simulation were tuned to reproduce AAV dataset 2 (see text for details). (**C**) Resolution against the NGS depth for nine different initial DNA amounts (N_0), chosen to match the nine datasets of Table 1. Color code is the same as in panel (D). (**D**) Exploited resolution against the ratio between the NGS depth and the initial DNA (gray curve). The values for the nine datasets (colored points) are added to compare their depth with the optimal one.

Optimal NGS depth

Our mathematical model suggests that our noise extent (FF) is approximately proportional to the NGS depth (Eq. 1). To validate this observation, we generated eight synthetic datasets by subsequently halving the NGS depth of AAV dataset 2. Computing the noise extent after binning (see the 'Materials and methods' section), we observed a linear relationship between FF_E − 1 and the NGS depth as predicted by the mathematical model (Fig. 4A). To further confirm this result, we ran nine simulations of our mathematical model, and analyzed the data with the same procedure. Simulations confirmed the expected linear behavior and reproduced the empirical analysis (Fig. 4B).

The observation that the noise extent increases with the NGS depth implies that increasing the latter might not always improve data quality, particularly the ability to measure the concentration of DNA barcodes. In order to quantify this ability, we estimated the resolution of a given NGS dataset, as the rescaled signal-to-noise ratio between expected count and its variability (see the 'Materials and methods' section and Fig. 4C). The resolution depends on the two quantities that most affect noise extent, namely the initial DNA copy number and the sequencing depth (N_0 and D in Eq. 1, respectively). As expected, increasing the amount of initial DNA always increases resolution, or at worst it does not affect it for very low sequencing depth. Interestingly, for a fixed initial amount of DNA, increasing sequencing depth always enhances resolution, but the improvement quickly saturates. Beyond a sequencing depth of ten times the amount of DNA, much deeper sequencing is needed to significantly increase resolution. For comparison, we estimated the resolution for the nine datasets in Table 1 (Fig. 4C). For each of them, we first estimated the noise extent by binning the sequences as we did before, to then estimate the initial DNA amount (see the 'Materials and methods' section) and eventually the resolution. For some datasets, depending on the initial DNA amount, the sequencing depth was close to or beyond saturation, while for others, further improvement could still be possible.

To further investigate the saturation effect, we computed the exploited resolution (see the 'Materials and methods' section) by normalizing with respect to the different initial DNA amount, so as to focus on the distance from the saturation point (Fig. 4D). Using the exploited resolution, we then defined an optimal sequencing depth D^{opt} as the depth at which the exploited resolution reaches 95%. This optimal depth can be computed as (see the 'Materials and methods' section)

$$D^{\text{opt}} \simeq 10N_0, \tag{2}$$

which means that the optimal NGS depth is about ten times larger than the amount of DNA before amplification (N_0). The main message of our article is that, for a given initial DNA amount N_0, increasing the sequencing depth beyond $10N_0$ offers no benefit due to the saturation of the exploited reso-

lution. Direct experimental verification suggests that an even stricter rule could be considered (see the next subsection).

We computed the exploited resolution and the optimal depth D^{opt} for the nine datasets in Table 1. We observed that NGS depth was too large for three of the datasets, and too small for two of them (Fig. 4D). Using our rule of thumb (2), we can now recommend the appropriate NGS depth for measuring DNA barcode density based on the estimate total number of DNA barcodes for each scenario.

Experimental comparison between FF_M and FF_E

To directly investigate our noise model, we conducted a series of experiments in which both the sequencing depth (D) and the initial number of DNA barcodes before amplification (N_0) were systematically varied (see Supplementary Fig. S6 for details across nine datasets). First of all, we found that $FF_M - 1$ depends primarily on the ratio D/N_0, increasing proportionally with this quantity (Supplementary Fig. S6, left panel). Since this ratio corresponds to $FF_E - 1$, these results confirm that FF_M and FF_E exhibit qualitatively similar behavior. Second, the empirical noise extent is consistently larger than the model noise extent ($FF_E \gtrsim FF_M$), likely due to additional noise sources not accounted for in our model (see discussion). This observation suggests that, in practice, the effective resolution can reach saturation at sequencing depths somewhat lower than our proposed rule of thumb (Supplementary Fig. S6, right panel), which can be expressed as

$$D^{opt} < 10N_0. \tag{3}$$

Although this inequality is less specific than Eq. (2), it conveys an important message: increasing sequencing depth beyond a certain point yields diminishing returns in accuracy. This, in fact, is our central message and the core warning we aim to emphasize in this article.

Discussion

In this work we addressed the problem of determining an optimal NGS depth for sampling highly diverse libraries with barcodes. Optimal depth is usually discussed in terms of coverage, which sets the target reading redundancy for complete sequencing [21]. These established practices and methods are however designed for sequencing long DNA fragments, or even entire genomes. They may not be suitable for estimating the abundances of short read sequences, such as molecular barcodes, because of the overwhelming diversity of the sample. In these cases, NGS depth is often determined heuristically, or simply pushed at maximum, with additional costs and resources. Thanks to our analyses, we understood that increasing the sequencing depth above certain values does not increase the dataset quality. Based on our results, we found that the optimal depth is about ten times the initial DNA amount before amplification (N_0, see the 'Results' section).

Our 'rule of thumb' for the sequencing depth does not depend on the properties of the barcode library, such as its diversity or flatness. This follows from the independence of the noise extent from those quantities: it is the same for all barcodes and depends only on the sequencing depth and the initial number of DNA molecules (Eq. 1). As a consequence, the signal-to-noise ratio has the same dependence on the sequencing depth for each barcode, and—after a rescaling—this allows us to focus on resolution and its exploited percentage. These two quantities are therefore the same for all barcodes. Eventually, the saturating behavior of the exploited resolution allowed us to derive our rule for optimal sequencing depth. Finally, although our rule of thumb does not depend directly on the number of unique barcodes, the latter still has an influence on it. Indeed, a highly heterogeneous library will require a larger number of initial barcodes to ensure sufficient coverage, which in turn necessitates a higher NGS depth according to our rule of thumb. In contrast, for highly biased libraries, a smaller number of initial barcodes is adequate, resulting in a lower NGS depth.

Increasing this DNA copy number always improves the resolution of an NGS dataset (Fig. 4C), yet for a given value, depths larger than the optimal value do not improve data quality (Fig. 4D), and in particular the ability to measure the concentration of DNA barcodes. For the application of our rule of thumb, the initial amount of DNA needs to be measured. Various methods are available for this purpose. One example is the NanoDrop, which enables rapid quantification of DNA molecules using only 1–2 μl of sample, with a relatively low error rate of 5% [45]. It is important to note that our rule of thumb is predicted on experimental conditions involving only two subsamplings: one prior to PCR amplifications and one subsequent to them (see Fig. 3A). Additional subsamplings may be necessary in certain experiments; in these cases, it is required to measure the concentration at each step (see Supplementary Mathematics for the derivation): $D^{opt} \simeq 10/(\sum_i 1/N_i)$. Based on this formula, one can derive recommended values for the NGS reads in any setting.

Our results come from a combination of data analysis of NGS datasets and mathematical modeling, which allowed us to quantify the over-dispersion in data (large noise extent), and to understand its nature. At first, for the NGS datasets of molecular barcodes (Table 1), we inferred the original concentrations of each barcode in the sample using a simple, yet precise model of the count probability distribution. This allowed us to study the count statistics in three datasets (Fig. 2), where variance and mean are proportional. The proportionality factor, identified as the noise extent of the dataset, is considerably higher (AAV datasets) or slightly higher (hYAP65 WW dataset) than 1 (Poisson noise). We then built a three-step mathematical model for NGS procedures, including the amplification process required to produce the necessary amount of genetic material [46]. We were able to solve the model analytically, obtaining a concise mathematical expression for the noise extent (Eq. 1). Our formula recovered the correct over-dispersion relation, and showed that reducing the number of amplification rounds has a limited impact on noise (Fig. 3E). From this we deduced that over-dispersion is an intrinsic property of NGS [46], that can at best be attenuated by accurate experimental protocols. Lastly, we observed that noise extent increases with sequencing depth. This is the reason for the small noise in the hYAP65 WW datasets and for the saturation of the resolution of count measurements (Fig. 4). From this last observation, we defined the optimal sequencing depth as the value at which the exploited resolution reaches 95% of its saturated value, and subsequently we derived an analytical expression for that optimal value. What makes it possible to define a consistent sampling depth at the desired resolution is that mean and variance of the counts are proportional and that the proportionality factor is the same for all barcodes in the sample and it scales linearly with the depth.

Previous works in the literature already discussed and modeled noise in NGS data. They however limited their analysis on the statistical properties of noise, and described it with over-dispersed negative binomial [23, 24, 26, 27] or beta-binomial [25] probability distributions. Here we built a mathematical model to understand the origin of such over-dispersion. Our amplification model is based on a series of mathematical works studying PCR [28–33], among which [30] was particularly relevant. We used their analytical frameworks to study the statistics of noise in NGS counts, thereby establishing a mathematical foundation for analyzing over-dispersion.

Three main limitations of our work can lead to further developments. First of all, the replication rate might be more complex than a MM with constant parameter. However, this is a minor limitation, because it will not much impact the properties of the noise that we need to derive our optimal depth. Also, noise extent in NGS data showed a small dependence on the expected count. In our analyses, we observed small deviations for large count values (Fig. 2D and G). Going beyond our simple position independent model slightly reduced this effect (Supplementary Fig. S2), and we expect that better models can reduce even more this deviation: errors in predicting the expected counts introduce some additional fluctuations in our binning procedure, and therefore increase the observed noise variance. In our analysis we did not try more flexible models because overfitting NGS counts lead to underestimation of noise extent, providing a trade off between model complexity and precision. It is also possible that different regimes appear for large count values [47] and the mechanism that generate these different regime remains unknown. In fact, previous work already noticed that the count variance grows faster than the mean [23–27]. Our mathematical analysis excluded these behaviors, but this might come from our simplified amplification process which considered only missed replications. Finally, other sources of errors might be present during PCR, such as those resulting in replacing one nucleotide for another during replication [46]. Even if these errors can be reduced by high-fidelity PCR and have in general smaller consequences in comparison to missed replication [22, 46], their integration into our framework is possible as a future perspective towards refinement of our method. These additional noise sources could partially explain why we observed $FF_E > FF_M$ (Supplementary Fig. S6). Another interesting avenue to explore would be about optimizing the absolute signal-to-noise ratio and not the exploted one (which is mathematically equivalent to the exploited resolution). In this case, to estimate the predicted count of 'rare' barcodes it would be necessary to increase the initial amount of DNA (N_0), and the sequencing depth should be adjusted following our rule of thumb. We leave this last interesting point of discussion for future developments.

Acknowledgements

The authors would like to thank L.C. Byrne for providing the AAV dataset 2. The authors would also like to thank J. Fernandez-de-Cossio-Diaz, G. Uguzzoni, and L.C. Byrne for useful comments and discussions. The authors would also like to thank Twist Bioscience for producing the plasmid library and the platform GENOM'IC at Institut Cochin for sequencing. This work has been done within the framework of the PostGenAI@Paris project and it has benefitted from financial support by the Agence Nationale de la Recherche (ANR) with the reference ANR-23-IACL-0007. This work was supported by ERC Starting Grant (REGE-NETHER 639888 to D.D.), European Research Council (ERC) Horizon 2020 Framework Programme Project (863214 - NEUROPA to D.D.), UNADEV, BpiFrance (Grant i-Demo - GEAR project to D.D. and U.F.), the Institut National de la Santé et de la Recherche Médicale (INSERM), Sorbonne Université (to D.D. and U.F.), The Foundation Fighting Blindness, Agence National de Recherche (ANR) RHU Light4-Deaf, LabEx LIFESENSES (ANR-10-LABX-65 to D.D.), IHU FOReSIGHT (ANR-18-IAHU-01 to D.D.), JSPS KAKENHI (Grant Number 22K17994 to T.N.), World Premier International Research Center Initiative (WPI), MEXT, Japan (to T.N.), and Paris Region Postdoctoral Fellowship (PRPF to E.Z.).

Author contribution: Conceptualization: T.O., C.C., D.D., T.N., U.F.; data curation: T.O., E.Z., M.D.; formal analysis: T.O., T.N., U.F.; funding acquisition: D.D., U.F.; investigation: T.O., E.Z., M.T., M.D., T.N., U.F.; methodology: T.O., T.N., U.F.; project administration: T.O., E.Z., D.D., U.F.; resources: M.D., D.D., U.F.; software: T.O., T.V.M., T.N.; supervision: D.D., T.N., U.F.; visualization: T.O., T.N., U.F.; writing—original draft: T.O., T.N., U.F.; writing—review and editing: T.O., E.Z., T.V.M., M.D., D.D., T.N., U.F.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

Agence Nationale de la Recherche (grant numbers ANR-10-LABX-65 and ANR-18-IAHU-01), H2020 European Research Council (grant number 863214 – NEUROPA), Bpifrance (grant number I-Démo - project GEAR), Union Nationale des Aveugles et Déficients Visuels, European Research Council Starting Grant (grant number REGE-NETHER 639888), Japan Society for the Promotion of Science (grant number 22K17994). Funding to pay the Open Access publication charges for this article was provided by Bpifrance - I-Démo - GEAR project

Data availability

The in-house dataset is available at https://doi.org/10.5281/zenodo.16417895. The code used for the analysis is available in GitHub, https://github.com/tommyocari/NGS, and in Zenodo, https://doi.org/10.5281/zenodo.16277405.

References

1. Behjati S, Tarpey PS. What is next generation sequencing? *Arch Dis ChildEduc Pract Ed* 2013;**98**:236–8. https://doi.org/10.1136/archdischild-2013-304340
2. Levy SE, Myers RM. Advancements in next-generation sequencing. *Annu Rev Genom Hum Genet* 2016;**17**:95–115. https://doi.org/10.1146/annurev-genom-083115-022413
3. Furlani B, Kouter K, Rozman D *et al.* Sequencing of nucleic acids: from the first human genome to next generation sequencing in COVID-19 pandemic. *Acta Chim Slov* 2021;**68**:268–78. https://doi.org/10.17344/acsi.2021.6691

4. Olson CA, Wu NC, Sun R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol* 2014;**24**:2643–51. https://doi.org/10.1016/j.cub.2014.09.072
5. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. *Nat Methods* 2014;**11**:801–7. https://doi.org/10.1038/nmeth.3027
6. Starita LM, Young DL, Islam M *et al.* Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* 2015;**200**:413–22. https://doi.org/10.1534/genetics.115.175802
7. Puchta O, Cseke B, Czaja H *et al.* Network of epistatic interactions within a yeast snoRNA. *Science* 2016;**352**:840–4. https://doi.org/10.1126/science.aaf0965
8. Diss G, Lehner B. The genetic landscape of a physical interaction. *Elife* 2018;**7**:e32472. https://doi.org/10.7554/eLife.32472
9. Darmostuk M, Rimpelova S, Gbelcova H *et al.* Current approaches in SELEX: An update to aptamer selection technology. *Biotechnol Adv* 2015;**33**:1141–61. https://doi.org/10.1016/j.biotechadv.2015.02.008
10. Zhou Y, Qi X, Liu Y *et al.* DNA-nanoscaffold-assisted selection of femtomolar bivalent human α-thrombin aptamers with potent anticoagulant activity. *ChemBioChem* 2019;**20**:2494–503. https://doi.org/10.1002/cbic.201900265
11. Lyu C, Khan IM, Wang Z. Capture-SELEX for aptamer selection: a short review. *Talanta* 2021;**229**:122274. https://doi.org/10.1016/j.talanta.2021.122274
12. Di Gioacchino A, Procyk J, Molari M *et al.* Generative and interpretable machine learning for aptamer design and analysis of *in vitro* sequence selection. *PLoS Comput Biol* 2022;**18**:e1010561. https://doi.org/10.1371/journal.pcbi.1010561
13. Zutterling C, Todeschini AL, Fourmy D *et al.* The forkhead DNA-binding domain binds specific G2-rich RNA sequences. *Nucleic Acids Res* 2023;**51**:12367–80. https://doi.org/10.1093/nar/gkad994
14. Körbelin J, Sieber T, Michelfelder S *et al.* Pulmonary targeting of adeno-associated viral vectors by next-generation sequencing-guided screening of random capsid displayed peptide libraries. *Mol Ther* 2016;**24**:1050–61. https://doi.org/10.1038/mt.2016.62
15. Byrne LC, Day TP, Visel M *et al.* *In vivo*-directed evolution of adeno-associated virus in the primate retina. *JCI Insight* 2020;**5**:e135112. https://doi.org/10.1172/jci.insight.135112
16. Fantini M, Lisi S, De Los Rios P *et al.* Protein structural information and evolutionary landscape by *in vitro* evolution. *Mol Biol Evol* 2020;**37**:1179–92. https://doi.org/10.1093/molbev/msz256
17. Stiffler MA, Poelwijk FJ, Brock KP *et al.* Protein structure from experimental evolution. *Cell Syst* 2020;**10**:15–24. https://doi.org/10.1016/j.cels.2019.11.008
18. Tabebordbar M, Lagerborg KA, Stanton A *et al.* Directed evolution of a family of AAV capsid variants enabling potent muscle-directed gene delivery across species. *Cell* 2021;**184**:4919–38. https://doi.org/10.1016/j.cell.2021.08.028
19. Goertsen D, Goeden N, Flytzanis NC *et al.* Targeting the lung epithelium after intravenous delivery by directed evolution of underexplored sites on the AAV capsid. *Mol Ther Methods Clin Dev* 2022;**26**:331–42. https://doi.org/10.1016/j.omtm.2022.07.010
20. Stanton AC, Lagerborg KA, Tellez L *et al.* Systemic administration of novel engineered AAV capsids facilitates enhanced transgene expression in the macaque CNS. *Med* 2023;**4**:31–50. https://doi.org/10.1016/j.medj.2022.11.002
21. Sims D, Sudbery I, Ilott NE *et al.* Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* 2014;**15**:121–32. https://doi.org/10.1038/nrg3642
22. Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res* 2015;**43**:e143.
23. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:R106.
24. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. https://doi.org/10.1186/s13059-014-0550-8
25. Illingworth CJ, Roy S, Beale MA *et al.* On the effective depth of viral sequence data. *Virus Evol* 2017;**3**:vex030. https://doi.org/10.1093/ve/vex030
26. Puelma Touzel M, Walczak AM, Mora T. Inferring the immune response from repertoire sequencing. *PLoS Comput Biol* 2020;**16**:e1007873. https://doi.org/10.1371/journal.pcbi.1007873
27. Nemoto T, Ocari T, Planul A *et al.* ACIDES: on-line monitoring of forward genetic screens for protein engineering. *Nat Commun* 2023;**14**:8504. https://doi.org/10.1038/s41467-023-43967-9
28. Schnell S, Mendoza C. Enzymological considerations for the theoretical description of the quantitative competitive polymerase chain reaction (QC-PCR). *J Theor Biol* 1997;**184**:433–40. https://doi.org/10.1006/jtbi.1996.0283
29. Schnell S, Mendoza C. Theoretical description of the polymerase chain reaction. *J Theor Biol* 1997;**188**:313–8. https://doi.org/10.1006/jtbi.1997.0473
30. Jagers P, Klebaner F. Random variation and concentration effects in PCR. *J Theor Biol* 2003;**224**:299–304. https://doi.org/10.1016/S0022-5193(03)00166-8
31. Lalam N, Jacob C, Jagers P. Modelling the PCR amplification process by a size-dependent branching process and estimation of the efficiency. *Adv Appl Probab* 2004;**36**:602–15. https://doi.org/10.1239/aap/1086957587
32. Chatterjee N, Banerjee T, Datta S. Accurate estimation of nucleic acids by amplification efficiency dependent PCR. *PLoS One* 2012;**7**:e42063. https://doi.org/10.1371/journal.pone.0042063
33. Best K, Oakes T, Heather JM *et al.* Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Sci Rep* 2015;**5**:14629. https://doi.org/10.1038/srep14629
34. Byrne L, Day T, Visel M *et al.* Directed evolution of AAV for efficient gene delivery to canine and primate retina-raw counts of variants from deep sequencing. *Dryad*. 2020. Berkeley: University of California, https://doi.org/10.6078/D1895R, (10 June 2018, date last accessed).
35. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;**17**:10–12. https://doi.org/10.14806/ej.17.1.200
36. Araya CL, Fowler DM, Chen W *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc Natl Acad Sci USA* 2012;**109**:16858–63. https://doi.org/10.1073/pnas.1209751109
37. 10x Genomics. Fresh Frozen Visium on CytAssist: Human Breast Cancer, Probe-Based Whole Transcriptome Profiling. *Avaden Biosciences*. 2023, https://www.10xgenomics.com/resources/datasets/fresh-frozen-visium-on-cytassist-human-breast-cancer-probe-based-whole-transcriptome-profiling-2-standard, (10 June 2024, date last accessed).
38. 10x Genomics. Fresh Frozen Visium on CytAssist: Mouse Brain, Probe-Based Whole Transcriptome Profiling. *Avaden Biosciences*. 2023, https://www.10xgenomics.com/resources/datasets/fresh-frozen-visium-on-cytassist-mouse-brain-probe-based-whole-transcriptome-profiling-2-standard, (10 June 2024, date last accessed).
39. 10x Genomics. Adult Mouse Brain Coronal Section (Fresh Frozen). *BioIVT Asterand*. 2023, https://www.10xgenomics.com/resources/datasets/adult-mouse-brain-coronal-section-fresh-frozen-1-standard, (10 June 2024, date last accessed).
40. 10x Genomics. 5k Human PBMCs, 3′ v3.1, Chromium Controller. *BioIVT Asterand*. 2022, https://www.10xgenomics.com/resources/datasets/5k-human-pbmcs-3-v3-1-chromium-controller-3-1-standard, (10 June 2024, date last accessed).

41. Cocco S, Feinauer C, Figliuzzi M *et al.* Inverse statistical physics of protein sequences: a key issues review. *Rep Prog Phys* 2018;**81**:32601. https://doi.org/10.1088/1361-6633/aa9965
42. Tubiana J, Cocco S, Monasson R. Learning protein constitutive motifs from sequence data. *Elife* 2019;**8**:e39397. https://doi.org/10.7554/eLife.39397
43. Watson HW, Galton F. On the probability of the extinction of families. *J Anthropol Inst Great Britain Ireland* 1875;**4**:138–44.
44. Chigansky P, Jagers P, Klebaner FC. What can be observed in real time PCR and when does it show? *J Math Biol* 2018;**76**:679–95. https://doi.org/10.1007/s00285-017-1154-1
45. Masago K, Fujita S, Oya Y *et al.* Comparison between fluorimetry (Qubit) and spectrophotometry (NanoDrop) in the quantification of DNA and RNA extracted from frozen and FFPE tissues from lung cancer patients: a real-world use of genomic tests. *Medicina* 2021;**57**:1375. https://doi.org/10.3390/medicina57121375
46. Potapov V, Ong JL. Examining sources of error in PCR by single-molecule sequencing. *PLoS One* 2017;**12**:e0169774. https://doi.org/10.1371/journal.pone.0169774
47. Lazzardi S, Valle F, Mazzolini A *et al.* Emergent statistical laws in single-cell transcriptomic data. *Phys Rev E* 2023;**107**:44403. https://doi.org/10.1103/PhysRevE.107.044403

Received: June 14, 2024. **Revised:** July 3, 2025. **Editorial Decision:** July 11, 2025. **Accepted:** August 20, 2025