In-silico monitoring of directed evolution convergence to unveil best performing variants with credibility score

Takahiro Nemoto[a,b,1], **Tommaso Ocari**[a], **Arthur Planul**[a], **Muge Tekinsoy**[a], **Emilia A. Zin**[a], **Deniz Dalkara**[a], **and Ulisse Ferrari**[a]

[a]Institut de la Vision, Sorbonne Université, INSERM, CNRS, 17 rue Moreau, 75012, Paris, France; [b]Graduate School of Informatics, Kyoto University, Yoshida Hon-machi, Sakyo-ku, Kyoto, 606-8501, Japan

This manuscript was compiled on December 15, 2022

Directed evolution (DE) is a versatile protein-engineering strategy applicable to proteins like enzymes, antibodies, or viral vectors. Starting from a library containing billions of random variants, DE screens them against a chosen task over multiple rounds until they converge onto a few variants with desired properties. Today, DE benefits from deep sequencing, allowing us to monitor millions of distinct variants through the screening iterations. Nevertheless deep sequencing data are noisy and it is often difficult to understand if the selection process has yet converged onto reliable results. As a consequence, experimentalists are compelled to increase the number of selection rounds or even to replicate the experiments. Here, we propose ACIDES (Accurate Confidence Intervals to rank Directed Evolution Scores), a combination of statistical inference and in-silico simulations to: (i) reliably estimate the selectivity of individual variants and its statistical error using the data from all available rounds; (ii) identify which variants among the best ones are worth keeping for further tests; (iii) quantify if DE has converged onto the few variants with desired function or if additional experimental rounds/replicates or deep sequencing are needed. The latter point allows for integrating ACIDES into the experimental pipeline to avoid unnecessary, costly and time-consuming experimental efforts. To showcase our method we re-analyze public datasets ranging from DE of viral vectors to phage-display or yeast competitive growth experiments. Lastly, we benchmark our method against previously proposed approaches in the literature.

directed evolution | phage display | deep mutational scanning | protein engineering | next generation sequencing

Introduction

Directed evolution (DE) (1–3) is a versatile protein engineering strategy to conceive and optimize proteins like enzymes (4–6), antibodies (7, 8), or viral vectors (9–15). DE starts from a massive library of random mutants, screens it against a given task over multiple rounds and searches for the variants with the highest performance. As the iteration continues, the best performing variants get enriched and emerge from the bulk, while ineffective ones are instead weeded out. Nowadays, we can rely on next generation sequencing (NGS) (16, 17) to sample millions of variants within the library and monitor their concentrations over multiple rounds or time-points. In this approach, the enrichment of the screened variants is measured to rank the variants depending on their performance. In a similar flavor, Deep mutational scanning (DMS) experiments (18–20) combine extensive mutagenesis with NGS to study the properties of proteins (21–25), promotors (26, 27), small nucleolar RNA (28), or other amino-acid chains. It uses similar techniques to DE and requires similar analysis. In this article, we do not distinguish DMS from DE, and focus on their common issues and needs.

The analysis of NGS data of multiple selection rounds presents several difficulties. First, variants need to be robustly scored based on their enrichment rates, so-called selectivities (29, 30). This task is complicated by the large noise in the NGS counts introduced by polymerase chain reaction (PCR) amplification during amplicon preparations (31–33). This noise needs to be taken into account in the analysis. Second, in order to rank the variants and to identify the best performing ones, the score should come with a precise estimation of its statistical error. As a consequence of the noise in the counts, some irrelevant variants might appear to be highly enriched. This would be anticipated if properly estimated credibility scores are available. Third, when running DE over multiple rounds, it is hard to know when to end the experiment: performing too few rounds could lead to selection of weak variants, nor representative of their true ranking. On the other hand, performing too many rounds is costly, time-consuming and even ethically questionable when working with *in-vivo* selections (14, 34). Similarly, it would be useful to understand the best NGS depth for a given DE experiment, as deepening the NGS by increasing reads results in better data, but adds an extra expense to the experiment.

In order to account for these issues and needs, we present ACIDES, Accurate Confidence Intervals for Directed Evolution Scores, a computational method to empower the analysis of

Significance Statement

Directed evolution (DE) is a protein engineering strategy that mimics Darwinian evolution within laboratory environments. Using NGS (Next Generation Sequencing: massively paralleled assay of DNA sequencing), candidate proteins are monitored during DE experiments to understand the selection process and to identify the fittest proteins. However, the number of evolving proteins is often larger than the number of proteins that NGS measurements can read, leading to noisy signals. Here, in order to harness the noise, we develop an in-silico monitoring framework to account for the noise and estimate the performance of each variant with an accurate credibility score. Our method can be used *on the fly* during the experiments and will help DE experimentalists with decision making.

Please provide details of author contributions here.

The authors declare no competing interests.

[1]To whom correspondence should be addressed. E-mail: takahiro.nemoto@i.kyoto-u.ac.jp, deniz.dalkara@inserm.fr, ulisse.ferrari@inserm.fr.

DRAFT

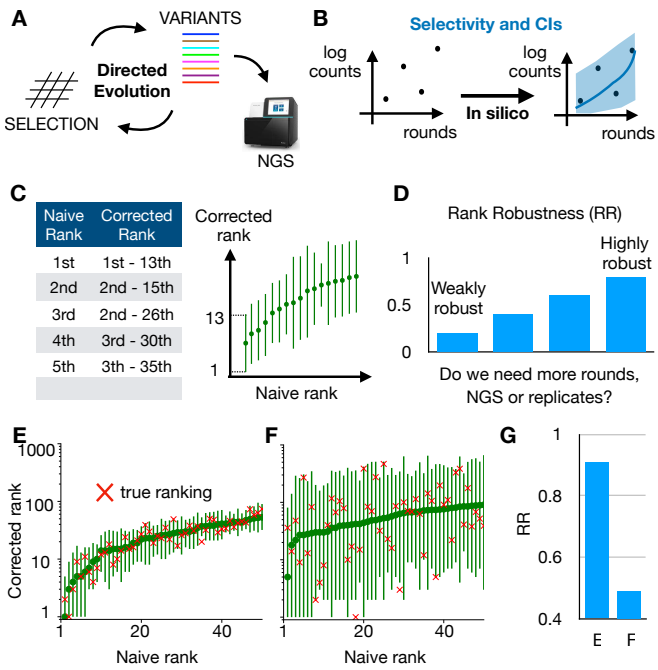

Fig. 1. ACIDES framework. (A) We consider directed evolution (DE) experiments, where protein variants are screened over multiple rounds, and massive NGS datasets are collected. (B) From the obtained count data, we estimate a score (selectivity) for each variant. The higher the score, the better the variant for the task. Each score is estimated together with its 95%-confidence interval (CI). (C) Sorting the scores of all variants in descending order, we obtain a variant rank (naive rank). Due to statistical errors in the scores, the obtained rank is biased in general. To correct for this, using *in-silico* simulations based on the CIs of the scores, we re-estimate the rank with 95%-CI (corrected rank), shown as green error bars. (D) From the obtained corrected rank, we compute Rank Robustness (RR), which is the ratio at which the identified top-50 variants in the naive rank are correct. (E,F) Examples of rank graphs for two synthetic datasets with different depths of NGS (per round) and numbers of unique variants (respectively, E: $10^7, 5 \times 10^4$; F: 10^6 and 10^6). The true rank is shown as red crosses. In both cases, most red crosses are within the 95%-CI of the corrected rank. (G) RR for the two synthetic datasets. Note that RR multiplied by 50 (E:~ 45.3; F: ~ 24.6) roughly provides the number of the correct top-50 sequences, which are 46 and 23, respectively. (See Fig. S4 for more systematic comparison.)

DE experiments. We focus on screening experiments on highly diverse libraries where massive NGS data are collected over multiple rounds or multiple time-points (Fig. 1A). Our goal is to develop a method to extract maximal information from noisy NGS data, and allows for scoring and ranking variants with accurate statistical confidence scores. Our approach can be applied to different kinds of experiments, such as *in-vivo* DE (13, 14, 35), phage-display (22, 29, 36), yeast two-hybrid (22) and DE of small nucleolar RNA (28). It is possible to apply ACIDES either *a posteriori* over data collected previously, or along the course of the experiment as soon as the NGS data become available. The latter strategy allows for monitoring the selection convergence *on the fly*, and to understand when the experiment can be ended.

Results

The first step of ACIDES estimates the selectivity of each individual variant present in the dataset (Fig. 1B) and its 95% confidence interval (95%-CI). In this study the term selectivity means the rate at which each variant increases its concentration with respect to the others, a measure of the fitness in the selection task (Materials and Methods). Compared with previous approaches (18, 19, 21–23, 26, 27, 29, 30, 37), our approach combines a robust inference framework with a better quantification of the NGS sampling noise (31–33). For this scope, our approach benefits from a negative binomial loss (38–41) (Fig. S1) in which the variance of the noise is overdispersed and grows as $\lambda + \lambda^{2-\alpha}/\beta$. Here λ is the expected mean count, and α, β are parameters to be inferred (Materials and Methods). Using novel data from a plasmid library, we observed that our negative binomial model realizes a 50- to 70-fold improvement over the Poisson model in the predictive ability of the NGS sampling noise (Fig. S1). The second step of ACIDES uses the estimation of the selectivities and their statistical errors to rank the variants. The rank obtained by sorting the selectivities in descending order (naive rank) is biased due to statistical fluctuations of the selectivities. We correct this bias using *in-silico* simulations (Fig. 1C). The third and last step of ACIDES uses simulations to quantify a Rank Robustness (RR), a measure of the quality of the selection convergence (Fig. 1D). Specifically, RR is the ratio at which the top-50 variants identified in the naive rank are correct (Materials and Methods). RR ranges from 0 to 1: a low value points out that the variants have not been selected enough, and therefore calls for the necessity to perform more rounds, deeper NGS sampling or possibly more replicates. Conversely, a large value confirms that the selection has properly converged, and suggests that the experiment can be ended without performing additional experimental steps.

Before focusing on experimental data, we apply ACIDES to two synthetic datasets (Materials and Methods) describing two opposite scenarios: data-rich case (more NGS reads with fewer unique variants) and data-poor case (less NGS reads with more considered variants). In the data-rich case, we first verify that our method reaches high performance in recovering the ground-truth values of the selectivities ($R^2 \simeq 0.92$, Fig.S2) in a teacher-student setting. In this first case, selection convergence is reached and the different variants can be robustly ranked (Fig. 1E). In the data-poor case, instead, CI-bars are large and the ranking is uncertain (Fig. 1F). Consistently, the estimated RRs are high and low for, respectively, the data-rich and -poor examples (Fig. 1G). Note that, once multiplied by 50, RR roughly provides the number of the correct top-50 variants in both cases (caption of Fig. 1G). Furthermore, we observe that most true rank values (red crosses) fall within the 95%-CI in both examples. These observations show that our approach can quantify statistical errors even in the data-poor regime.

Analysis of directed evolution experiments. In order to showcase ACIDES, we apply it to several DE datasets, where various proteins (and one RNA molecule) are screened using different experimental techniques (Table 1). Specifically, we consider three phage-display screening experiments targeting different proteins, such as the breast cancer type 1 susceptibility protein (BRCA1) for Data-A, human yes-associated protein 65 (hYAP65) for Data-F and immunoglobulin heavy chain (IgH) for Data-G, two *in-vivo* DEs of adeno-associated virus type 2 (AAV2) vectors targeting canine eyes for Data-C and murine lungs for Data-D, a multiplexed yeast two-hybrid assay targeting BRCA1 for Data-B and a yeast competitive growth screen measuring the fitness of mutant U3 gene for Data-E. For each of these experiments, we rank variants (naive rank) and compute the confidence interval of their ranks (corrected rank in Fig. 2A-G). The degree of convergence of the

DRAFT

Table 1. Next generation sequencing datasets of directed evolution experiments

Label	Experiment	Target	Time-points	Reads / round	# of unique variants	Replicates	Ref
A	Phage display	BRCA1	T0 → T5	13.6 M	35 k	2 x 3	Starita 2015 (22)
B	Yeast two-hybrid	BRCA1	T0 → T3	13.5 M	27 k	2 x 3	Starita 2015 (22)
C	*in-vivo* DE (dog eye)	AAV2-7mer	T1 → T5	17 M	5 k	1 x 1	Byrne 2018 (35)
D	*in-vivo* DE (murine lung)	AAV2-7mer	T0 → T5	6.2 M	0.5 k	1 x 1	Korbelin 2016 (13)
E	Yeast competitive growth	U3 snoRNA	T0 → T4	8 M	24 k	2 x 1	Puchta 2016 (28)
F	Phage display	hYAP65 WW	T0 → T3	5 M	470 k	2 x 1	Araya 2012 (29)
G	Phage display	Ab IgH	T1 → T3	0.1 M	29 k	1 x 1	Boyer 2016 (36)

List and properties of experiments considered in this study. First column introduces dataset label and corresponds to the panels of Fig. 2. Reads/round corresponds to the average NGS counts per time points. # of unique variants is the number of variants that is detected in the NGS at least once during whole DE experiments. In Replicates, $x \times y$ means that there are x replicates that do not share the same initial library, each of which has y technical replicates (that shares the same initial library).

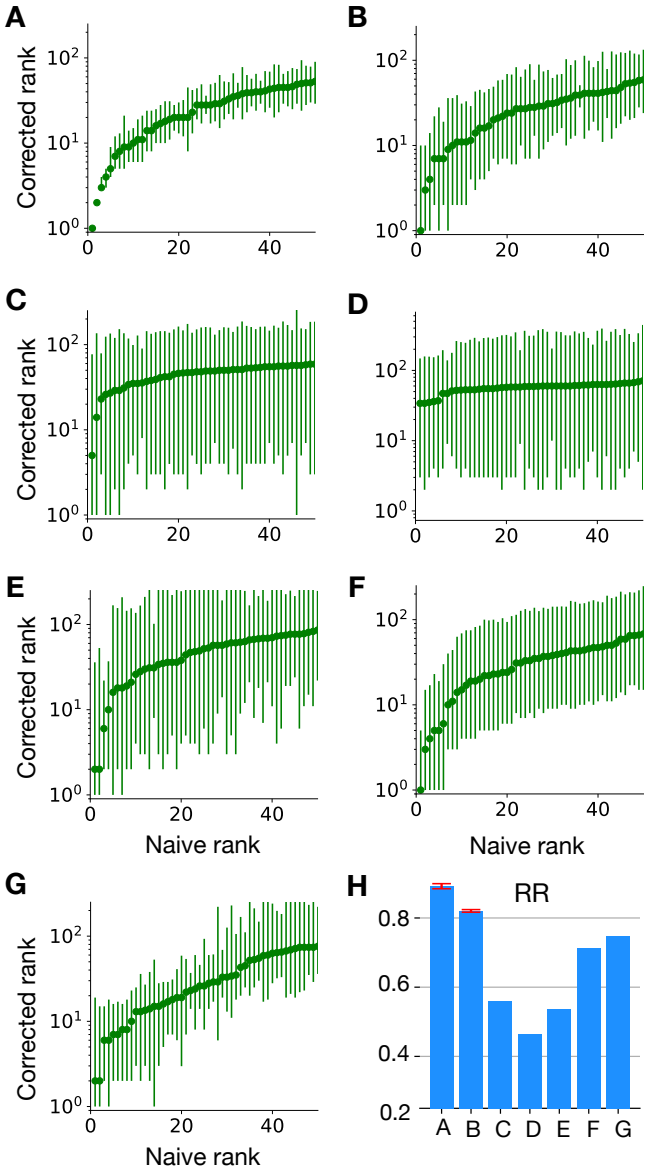

Fig. 2. Rank graph for various DE datasets. The panel labels A-G correspond to the experiments listed in table 1. (H) Rank robustness (RR) for each experiment. When technical replicates are available (Data-A and -B), the mean and standard deviation are shown.

selection is quantified by RR (2H). When technical replicates are available (Data-A and Data-B), we compute RR over all of them and obtained consistent results (shown by the small error-bars in Fig. 2H).

We classify the observed RRs into three groups depending on the quality of the selection convergence: high (Data-A and Data-B), intermediate (Data-F and Data-G), and low (Data-C, Data-D and Data-E) convergence groups. The high group seems to behave similarly to the data-rich synthetic data in Fig. 1E. Consistently, RR, NGS depth and the number of unique variants are indeed of the same order (Table 1). In these cases, the obtained naive rank is robust, as indicated by the value of RR (RR>0.8). In the intermediate group, the value of RR ranges between 0.6 and 0.8. The experimental techniques used in these datasets are similar to those in the high group, but the NGS depths (or the numbers of unique variants) are smaller (or larger), which could be the reason why they result in lower RRs. The low group suffers from the noise in the data. In Data-C and Data-D, the numbers of unique variants are lower than those of Data-A and Data-B, which would normally help these datasets with having higher RR, given the same NGS depth. As this is not the case, we see that some experiments are *intrinsically* more difficult than the others, *i.e.*, in-vivo DE (Data-C and Data-D) and RNS based DE (Data-E) will result in lower RRs than the other experiments if the NGS depth and number of variants are similar.

In datasets with low RRs, some variants seem to perform better than the others, but the difference between their scores is marginal compared with their statistical errors. This means that we cannot distinguish if the obtained variants are selected because of their ability to perform the task (fitness) or just there due to noise. In these cases, experimentalists have two possibilities: (i) based on the noisy identified variants, perform further tests in addition to DE (13, 14), as for example, study infective ability of viral vectors using single-cell RNA-seq (42). Or (ii) increase the quality of the DE datasets, by performing further selection rounds, increasing NGS depths, or replicating the experiments under the same conditions. This second possibility is explored in the next section. Overall our rank-analysis of the different experiments shows how our approach can provide an overview of the selection convergence, informing about the state of the experiment and eventually pointing out the necessity of more experimental efforts.

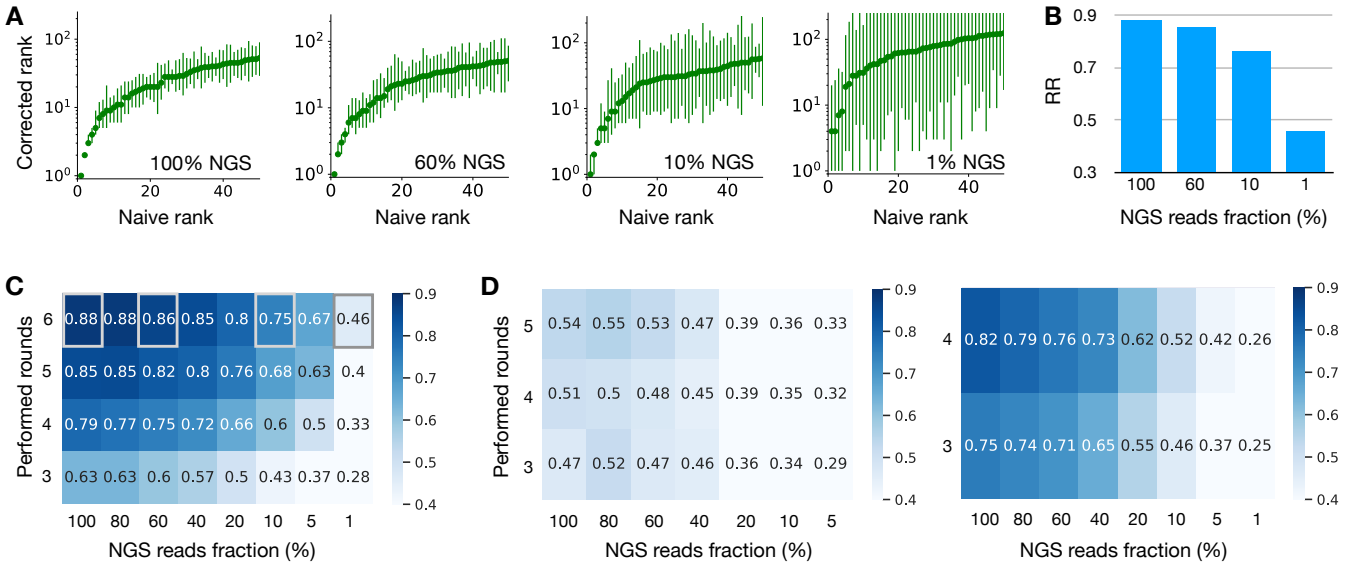

Fig. 3. How the rank robustness depends on the experimental protocol. (A) Rank graphs for different NGS depths in Data-A (Table 1). Different NGS-depth data are generated using downsampling (Materials and Methods). x% means the dataset where the number of NGS reads per round is reduced to x% (100% is the original dataset). (B) RR for the rank graphs in the panel A. Note that RR is higher than 0.7 even with the 10% NGS-depth. (C) The heat map showing RR for various NGS depths and performed rounds in Data-A. RR is larger than 0.7 for the data with (i) the 4 performed rounds with the NGS depth larger than or equal to 40% or with (ii) the 5 performed rounds with the NGS depth larger than or equal to 20%. This indicates that the data quality was already high with less experimental efforts. The four grey squares correspond to the four rank graphs in the panel A, respectively. (D) The same graphs as the panel C, but for different datasets. Data-E is used in the left panel, where RR is low and more NGS and/or DE rounds would be useful. Data-B is used in the right panel, where RR takes high values and seems to saturate in NGS depths. Further experimental efforts would probably not be necessary in this dataset.

DRAFT

Integration into the experimental pipeline. Noise in DE data can be reduced by performing additional selection rounds involving experiments, but in general these are expensive, time-consuming and, in case of experiments involving animal use, ethically problematic (34). For these reasons, it is important to choose accurately the number of rounds and the NGS depth. For this scope, ACIDES can be integrated into experimental pipelines to obtain an overview on how RR depends on these factors. This is to help DE experimentalists with making informed decisions about additional experimental efforts.

ACIDES can estimate RR after each selection round (or any time new data become available). This allows us to examine the data's behavior and to quantify the degree of convergence in terms of the selection rounds. Similarly, for each round, ACIDES can be run on downsampled NGS data to compute RR with smaller NGS depth (Materials and Methods). Using these two techniques, we monitor the need for more selection rounds or deeper NGS: a slow increase of RR (or no change in RR) upon improving data-quality implies that convergence is reached and suggests that the experiment can be ended. If, on the other hand, RR increases rapidly when improving the rounds and/or NGS depth, it is probably worth making further experimental efforts.

In order to showcase our approach, we study how RR depends on the number of DE rounds and NGS depth in previous experiments. We start by measuring RR in Data-A for different NGS depths. 95%-CI on corrected ranks gets larger as the NGS depth becomes smaller (Fig.3A). At 1% NGS depth, the variant ordering seems largely unreliable: RR is smaller than 0.5 (Fig.3B). Importantly, RR does not decrease smoothly as the NGS depth decreases, but it remains roughly constant at the beginning, and falls only at a very small NGS depth. This result suggests that the actual NGS depth of this experiment largely exceeds what was necessary (10% of the depth would have been sufficient). Next, we quantify how RR depends on both the number of performed rounds and NGS depth (Fig.3C). RR grows from 0.28 (3 performed rounds with 1% NGS depth) to 0.88 (6 performed rounds with 100% NGS depth). Saturation of RR seems to be observed for RR>0.7, which corresponds to 5 performed rounds with the NGS depth larger than 20%, or 4 performed rounds with the NGS depth larger than 40%. This again indicates that the experiment could have been stopped earlier (less rounds and/or lower sequence coverage) without much affecting the outcome. Note that different datasets show different behaviors. For Data-E more selection rounds with a higher number of NGS reads is expected to improve RR, while for Data-B they seem to have just reached the saturation point (Fig.3D).

Overall these results show how our approach can be implemented along experimental pipelines. By estimating RR while collecting new data, we can understand if we should continue/stop adding more rounds or increasing NGS depth. This could avoid unnecessary, costly and time-consuming experimental efforts. Similar analyses can be done on the number of replicate experiments (Fig.S6).

Comparison with previous work. We compare the performance of ACIDES with *Enrich2*, the state of the art for estimating variant scores (selectivities) in DE experiments (30). *Enrich2* is based on a weighted linear fitting of the log-count change along rounds, and the first step of ACIDES should be seen as an upgrade for this fitting. In order to compare these two approaches we take advantage of replicate datasets. We first investigate if the scores in each method are consistent over replicates. For this, we plot the scores obtained from one replicate against the scores obtained from the other (Fig. 4A,

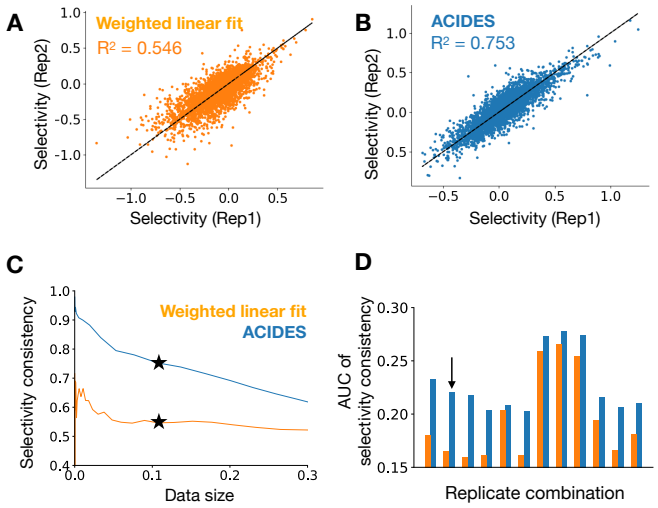

Fig. 4. Comparison of our approach with the state of the art. Using technical replicates in Data-A, we compare ACIDES with a weighted linear least squares method (*Enrich2*) (30). For both methods (*Enrich2* (A) and ACIDES (B)), the inferred selectivities from one replicate are plotted against the selectivities in the other replicate. The coefficient of determination (R^2), which quantifies the consistency between two replicates, is also shown. (C) We next examine how the comparison in the panels A and B depend on data quality. We consider a set of variants in which the estimated statistical errors (Materials and Methods) are smaller than a given threshold. Varying this threshold, sets of variants are systematical selected, where larger/smaller sets include variants with larger/smaller estimated statistical errors. For each set, we estimate R^2 between two replicates, and plot it as a function of the set size. The panel A and B correspond to the stars ⋆ in C (data size 0.11). (D) In order to test both methods more systematically, we perform the same analysis (as those in the panels A-C) for all possible 12 combinations of technical replicates in Data-A and Data-B. We define the area under curve of R^2 (in the panel C) and plot it for these combinations (D) . Our method systematically outperforms the weighted linear fitting method. The replicate combination used for the panels A-C is indicated by the arrow in the panel D.

B). The correlation between replicates is estimated using the coefficient of determination (R^2). The correlation quantifies the quality of the method, as higher (or lower) correlations imply that the estimated scores are more (or less) robust and fewer (or more) replicates are needed to obtain reliable results. The figure shows that ACIDES outperforms *Enrich2*. Next, we test how the comparison depends on the data quality. To this goal, we systematically select a set of variants based on the magnitude of predicted statistical score-errors (Materials and Methods). (Smaller/larger sets include variants with smaller/larger predicted statistical errors.) For each set, we measure the correlation between two replicates as in (Fig. 4A, B), and plot it as a function of the set size (Fig. 4C). We observe ACIDES's correlation becomes more dominant as the set size decreases, suggesting the better quality of both the estimated scores and statistical errors. In order to generalize these results, we perform the comparison for all possible 12 pairs of technical replicates in Data-A and Data-B (Table 1). In all cases our approach outperforms the competitor (Fig. 4D). We also perform an additional test to quantify the consistency of the predicted statistical errors (Supp. Fig. S7).

Discussion

In this work we have presented ACIDES, a method to quantify DE selectivities (fitness), rank variants with accurate credibility scores and measure the degree of experimental convergence. ACIDES can be used *on the fly* to offer an overview of the progress of DE experiments, which would help experimentalists with making informed decisions on whether new experimental efforts are needed. In this way, ACIDES can save significant experimental time and resources. We have applied ACIDES to several DE datasets where a number of different target proteins and one set of target RNA molecules have been 'evolved' using different experimental protocols. The heterogeneity of these datasets shows that ACIDES is a method of general use applicable to most DE experiments.

The first step of ACIDES estimates the score (selectivity) of each observed variant. This is a necessary step in DE experiments, and several alternative methods have been proposed in the past. In many applications, such scores are computed as the variant enrichment that is defined as the logarithmic ratio between the variant frequencies in the last and second to last round (13) or between the last and first round (14, 18, 19, 21, 26, 37). These approaches thus make use of data from only two rounds and disregard all the others. For this reason, this strategy is suboptimal and may lead to noisy score estimations. A more sophisticated approach that uses all the data consists in inferring the slope of a linear line fitted to the log-frequencies of variants over all the DE rounds/time points (22, 23, 27, 29). This method gives the same importance to log-frequencies in all the rounds. Yet as variant counts in the first rounds are typically small and noisy, assuming the same weight on them could result in an overfitting. To fix this effect, Enrich2 (30) uses the variance of the count data - estimated via a Poisson distribution assumption - as the weights in a linear least squares fitting. ACIDES' first step comes with a three-fold improvement over this last approach. First, instead of relying on the linear least squares fitting, we estimate the score by log-likelihood maximization. A major improvement happens for variants whose log-frequencies do not grow linearly with the rounds, and a simple linear weighted fit may struggle in identifying the correct slope. This is particularly visible in the bulk variants with intermediate scores (Fig.4 A, B). Secondly, instead of a simple exponential growth of the counts, we included a softmax non-linear function (Materials and Methods), where the denominator is inferred from data (43). This change improves the score estimation when the wildtype (if any) and/or few variants have a large fraction of the total counts and bend the exponential growth of the log-frequencies. Lastly, ACIDES uses a negative binomial distribution to model the count variability (38–41). This distribution accounts for the large dispersion of deep sequencing data (31–33) far better than the Poisson distribution (Fig. S1). Additionally, the negative binomial loss in the likelihood maximization allows us to better estimate statistical errors for the inferred scores. Thanks to all these improvements, our approach realizes a more robust and accurate estimation of the variant scores and outperforms the previous method (Fig. 4).

In case of noisy data, the estimated scores of variants come with statistical errors. This means that the rank obtained from the scores (naive rank in our figures) is in general biased: top ranked variants are overvalued, and vice-versa. This simple statistical effect was not taken into account in previous analyses related to DE experiments. The second step of ACIDES uses a bootstrap method to account for the bias and recover both the corrected rank and its 95%-CI. The deviation between this 95%-CI and the naive rank shows us how much we can trust

DRAFT

the naive rank. To quantify it, as a third step of ACIDES, we introduce RR that describes how many of the top-50 variants in the naive rank are correct. RR measures how stable and robust are the ranks of the variant selectivities. As such, it quantifies the degree of convergence of the experimental selection, providing an insightful overview of the state of the experiment.

Although ACIDES demonstrates advantages over the other methods, it has several limitations that may be addressed in the future. First of all, ACIDES does not account for changes in the selection pressure over rounds. This can potentially be included, but has not been done here, as the selection pressure is constant in most datasets we analyzed in this article. Second, ACIDES uses a negative binomial model to describe the dispersion of count data by assuming that the count variance depends only on the frequency of the variant. Although this assumption proves useful to describe NGS count errors (Fig. S1) and is used elsewhere (41), it is possible that dispersions induced by a sequence-dependent procedure, such as error-prone PCR (14, 35, 44), may not be taken into account by our method (Note that Data-C includes an error-prone PCR after the third round of selections, indicating that the estimated results for Data-C may contain biases). We would need to analyze more data from DE experiments using error-prone libraries to address this question. Third, statistical errors due to the replicates that do not share the same initial library cannot be described by ACIDES, provided that the model is only trained on a single series of DE rounds. To account for this, we would need a framework that generalizes ACIDES for different sources of variability.

Finally, using machine learning techniques, several studies have aimed at estimating selectivities from the amino-acid sequences of variants. Most of these methods rely on supervised algorithms, which are trained to predict the selectivity (output) from the sequence of a variant (input) (44–52). Because the performance of these methods depends on how the selectivity is estimated from data, ACIDES can potentially be incorporated in their pipelines to improve the overall performance. We leave such analysis for future developments. Other methods use instead unsupervised approaches to predict selectivities from the sequences of variants (43, 53–56). Even if these methods do not use any sequence scores for their training, they often use it to validate and/or test the model. Our approach would therefore be useful also in these cases.

Materials and Methods

Model. For a given series of samples over DE rounds, we perform NGS and denote by n_t^i the obtained count of the i-th variant ($i = 1, 2, ..., M$) at round (time-point) $t \in T$. We denote by N_t the total count $N_t = \sum_i n_t^i$ at t. For each sample, we define ρ_t^i as the *expected* value of frequency of the i-th variant at t. (Note that "expected" means that ρ_t^i itself does not fluctuate due to the noise in the experiment.) For each variant, an initial frequency ρ_0^i and a growth rate a^i are assigned, by which the expected frequency is computed as

$$\rho_{t+\Delta t}^i = C_t \rho_t^i \exp(a^i \Delta t), \qquad [1]$$

where Δt is the round- (or time-) difference between two consecutive NGSs. C_t is a normalization constant, defined as $C_t = 1/\sum_i [\rho_t^i \exp(a^i \Delta t)]$. We call this model (Eq. (1)) an exponential model.

We use a negative binomial distribution $\mathrm{NB}(n_t^i|\lambda, r)$ with two parameters λ and r to model the noise distribution of counts n_t^i. Here λ is the expected value of count n_t^i given as $N_t \rho_t^i$, while r is the dispersion parameter that describes the deviation of the negative binomial distribution from the Poisson distribution. (The negative binomial distribution is a generalization of the Poisson distribution with a variance equal to $\lambda(1 + \lambda/r)$: the Poisson distribution is recovered in the large r limit.) Here, based on Fig. S1 and (41), we assume r is a power-law function of λ: $r(\lambda) = \beta\lambda^\alpha$ (with $\alpha, \beta > 0$), where α and β are parameters that are common for all the variants in the experiment. (The variance is thus $\lambda + \lambda^{2-\alpha}/\beta$.) Model parameters α, β as well as ρ_0^i, a^i ($i = 1, 2, ..., M$) are inferred from the count data n_t^i ($i = 1, 2, ..., M$, $t \in T$) by maximizing the following likelihood function:

$$L\left(\alpha, \beta, (\rho_0^i)_{i=1}^M, (a^i)_{i=1}^M\right) = \prod_{i,t} \mathrm{NB}\left(n_t^i | \rho_t^i N_t, \beta\left(\rho_t^i N_t\right)^\alpha\right). \qquad [2]$$

The 95%-CIs of the estimated parameters are computed from the curvature of the log-likelihood function at the maximum.

Synthetic data. Synthetic count data n_t^i ($i = 1, 2, ..., M$, $t \in T$) are generated from the model (Eq. (2)) for a given parameter set $\alpha, \beta, \rho_0^i, a^i$ ($i = 1, 2, ..., M$). For Fig.1, we use $\alpha, \beta = 0.69, 0.8$ with $(a^i, \log \rho_0^i)$ generated from the normal distribution with the expected values $(-1, 1)$ and the standard deviations $(0.25, 1)$. (M, N_t) are $(5 \times 10^4, 10^7)$ for the data-rich case (Fig.1E) and $(10^6, 10^6)$ for the data-poor one (Fig.1F).

Model inference. To maximize the likelihood function, we develop a two-step algorithm. The first step infers (ρ_0^i, a^i), while the second (α, β) and then we iterate the two steps until convergence is reached. All inferences are done with a gradient descent algorithm, and to reach convergence 10-30 iterations are usually sufficient. The first step is itself iterative, and loops between the inference of (ρ_0^i, a^i) and C_t by treating C_t as a parameter. Here we also introduce a gauge choice because of the redundancy between ρ_0^i, a^i and C_t (the caption of Fig. S2 for more details). In the second step, the inference of (α, β) with a straightforward gradient method produces a bias (Fig. S2E). In order to correct this, at each iteration the algorithm adopts a teacher-student framework, runs a simulation of the count data with the current parameters to obtain an estimation of the bias, which is then used to correct the real inference and update the parameters.

In order to reduce computational time and to increase the stability of the algorithm, we first run the inference algorithm on a subset of variants to estimate α, β. We then compute (ρ_0^i, a^i) of the excluded variants using the estimated α, β. For this subset, we use the variants that satisfy the following two criterions: (i) their counts are larger than 0 more than twice in the selection rounds and (ii) whose total NGS count (as summed over all the rounds) is above a threshold. We set this threshold to 100 for all the datasets except for Data-C -D, where 10000 is used. This is because the noise in these experiments is larger than the others. Results are stable by changing the threshold value (Fig. S2F).

Simulated rank and rank robustness (RR). Using the standard deviations δa^i ($i = 1, ..., M$) of estimated scores a^i, we discard the variants with higher estimated errors. We keep 5000 variants for further analysis and denote by A their indices. We then rearrange the variant index in A in descending order of a^i to define a *naive rank* (the x-axis of Fig 2A-G). To obtain a *corrected rank* (the y-axis of Fig 2A-G), we first generate synthetic scores using the normal distribution with the expected value $(a^i)_{i \in A}$ and the standard deviation $(\delta a^i)_{i \in A}$. Based on the generated scores, we rearrange the variant index in descending order and define a synthetic naive rank. Repeating this estimation 3000 times, we then compute the median and 95%-CI of the obtained synthetic naive ranks. This 95%-CI is defined as the corrected rank.

To estimate RR, we compare the top-50 variants in the naive rank and each synthetic naive rank. We count the number of overlaps between them and average it over the 3000 estimations. RR is computed by dividing the obtained overlap by 50.

DRAFT

NGS-Downsampling for Fig.3. To obtain downsampled count data $\tilde{n}_t^i$ $(i = 1, 2, ..., M,\ t \in T)$ by a factor ϵ, we sample synthetic data from the likelihood function (Eq. (2)) with a reduced number of the total counts ϵN_t $(t \in T)$ and with the estimated parameters ρ_0^i, a^i, α, β $(i = 1, 2, ..., M)$. To obtain a downsampled RR in Fig. 3, we first re-estimate a^i $(i = 1, 2, ..., M)$ from $\tilde{n}_t^i$ $(i = 1, 2, ..., M,\ t \in T)$ using the values of (α, β) that are already known, and then perform the estimation of RR described above. Using the synthetic data, we show that this downsampling method captures well the RR of actual NGS-read-reduced data (Fig. S5).

Pre-processing of dataset D. In their original datasets, Data-C and -D contain a large number of variants whose total counts are very low (but not zero). In order to speed up the analysis and make the analysis more robust we removed the variants whose total counts are smaller than 1000 (Data-C) and than 100 (Data-D). The NGS depth and the number of unique variants shown in Table 1 are after this preprocessing.

ACKNOWLEDGMENTS. The authors would like to thank A. Rubin and D. Fowler for kindly providing them with datasets (Data-A and Data-B) and a working code for enrich2. The authors also would like to thank L. C. Byrne and T. Mora for useful comments and discussions, M. Desrosiers and C. Robert for their technical assistance with the production of plasmids and viral vectors, and O. Marre for facilitating and initiating the collaboration and helpful discussions. This work was supported by ERC Starting Grant (REGENETHER 639888), European Research Council (ERC) Horizon 2020 Framework Programme Project: 863214 – NEUROPA, UNADEV, the Institut National de la Santé et de la Recherche Médicale (INSERM), Sorbonne Université, The Foundation Fighting Blindness, Agence National de Recherche (ANR) RHU Light4Deaf, LabEx LIFESENSES (ANR-10-LABX-65), IHU FOReSIGHT (ANR-18-IAHU-01), and JSPS KAKENHI Grant Number 22K17994.

1. FH Arnold, Design by directed evolution. *Accounts chemical research* **31**, 125–131 (1998).
2. PA Romero, FH Arnold, Exploring protein fitness landscapes by directed evolution. *Nat. reviews Mol. cell biology* **10**, 866–876 (2009).
3. MS Packer, DR Liu, Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).
4. K Chen, FH Arnold, Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin e for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci.* **90**, 5618–5622 (1993).
5. NJ Turner, Directed evolution drives the next generation of biocatalysts. *Nat. chemical biology* **5**, 567–573 (2009).
6. OK Tawfik, D S., Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010) PMID: 20235827.
7. RE Hawkins, SJ Russell, G Winter, Selection of phage antibodies by binding affinity: Mimicking affinity maturation. *J. Mol. Biol.* **226**, 889–896 (1992).
8. ET Boder, KS Midelfort, KD Wittrup, Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proc. Natl. Acad. Sci.* **97**, 10701–10705 (2000).
9. L Perabo, et al., In vitro selection of viral vectors with modified tropism: the adeno-associated virus display. *Mol. Ther.* **8**, 151–157 (2003).
10. N Maheshri, JT Koerber, BK Kaspar, DV Schaffer, Directed evolution of adeno-associated virus yields enhanced gene delivery vectors. *Nat. biotechnology* **24**, 198–204 (2006).
11. S Michelfelder, M Trepel, Adeno-associated viral vectors and their redirection to cell-type specific receptors. *Adv. genetics* **67**, 29–60 (2009).
12. D Dalkara, et al., In vivo–directed evolution of a new adeno-associated virus for therapeutic outer retinal gene delivery from the vitreous. *Sci. Transl. Medicine* **5**, 189ra76–189ra76 (2013).
13. J Körbelin, et al., Pulmonary targeting of adeno-associated viral vectors by next-generation sequencing-guided screening of random capsid displayed peptide libraries. *Mol. Ther.* **24**, 1050–1061 (2016).
14. LC Byrne, et al., In vivo–directed evolution of adeno-associated virus in the primate retina. *JCI insight* **5** (2020).
15. M Tabebordbar, et al., Directed evolution of a family of aav capsid variants enabling potent muscle-directed gene delivery across species. *Cell* **184**, 4919–4938.e22 (2021).
16. S Behjati, PS Tarpey, What is next generation sequencing? *Arch. Dis. Child. - Educ. Pract.* **98**, 236–238 (2013).
17. SE Levy, RM Myers, Advancements in next-generation sequencing. *Annu. Rev. Genomics Hum. Genet.* **17**, 95–115 (2016) PMID: 27362342.
18. DM Fowler, et al., High-resolution mapping of protein sequence-function relationships. *Nat. methods* **7**, 741–746 (2010).
19. RT Hietpas, JD Jensen, DNA Bolon, Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci.* **108**, 7896–7901 (2011).
20. DM Fowler, S Fields, Deep mutational scanning: a new style of protein science. *Nat. methods* **11**, 801–807 (2014).
21. A Melnikov, P Rogov, L Wang, A Gnirke, TS Mikkelsen, Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic acids research* **42**, e112–e112 (2014).
22. LM Starita, et al., Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200**, 413–422 (2015).
23. S Matuszewski, ME Hildebrandt, AH Ghenu, JD Jensen, C Bank, A Statistical Guide to the Design of Deep Mutational Scanning Experiments. *Genetics* **204**, 77–87 (2016).
24. NJ Rollins, et al., Inferring protein 3d structure from deep mutation scans. *Nat. genetics* **51**, 1170–1176 (2019).
25. JM Schmiedel, B Lehner, Determining protein structures using deep mutagenesis. *Nat. genetics* **51**, 1177–1186 (2019).
26. RP Patwardhan, et al., High-resolution analysis of dna regulatory elements by synthetic saturation mutagenesis. *Nat. biotechnology* **27**, 1173–1175 (2009).
27. MS Rich, et al., Comprehensive Analysis of the SUL1 Promoter of Saccharomyces cerevisiae. *Genetics* **203**, 191–202 (2016).
28. O Puchta, et al., Network of epistatic interactions within a yeast snorna. *Science* **352**, 840–844 (2016).
29. CL Araya, et al., A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci.* **109**, 16858–16863 (2012).
30. AF Rubin, et al., A statistical framework for analyzing deep mutational scanning data. *Genome biology* **18**, 1–15 (2017).
31. JM Kebschull, AM Zador, Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* **43**, e143–e143 (2015).
32. K Best, T Oakes, JM Heather, J Shawe-Taylor, B Chain, Computational analysis of stochastic heterogeneity in pcr amplification efficiency revealed by single molecule barcoding. *Sci. reports* **5**, 1–13 (2015).
33. V Potapov, JL Ong, Examining sources of error in pcr by single-molecule sequencing. *PloS one* **12**, e0169774 (2017).
34. S Festing, R Wilkinson, The ethics of animal research. *EMBO reports* **8**, 526–530 (2007).
35. L Byrne, et al., Directed evolution of aav for efficient gene delivery to canine and primate retina - raw counts of variants from deep sequencing. *Dryad, Dataset* (2018) https://doi.org/10.6078/D1895R.
36. S Boyer, et al., Hierarchy and extremes in selections from pools of randomized proteins. *Proc. Natl. Acad. Sci.* **113**, 3482–3487 (2016).
37. DM Fowler, CL Araya, W Gerard, S Fields, Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* **27**, 3430–3431 (2011).
38. S Anders, W Huber, Differential expression analysis for sequence count data. *Nat. Preced.* pp. 1–1 (2010).
39. DJ McCarthy, Y Chen, GK Smyth, Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
40. MI Love, W Huber, S Anders, Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15**, 1–21 (2014).
41. M Puelma Touzel, AM Walczak, T Mora, Inferring the immune response from repertoire sequencing. *PLOS Comput. Biol.* **16**, e1007873 (2020).
42. BE Öztürk, et al., scaavengr, a transcriptome-based pipeline for quantitative ranking of engineered aavs with single-cell resolution. *eLife* **10**, e64175 (2021).
43. J Fernandez-de Cossio-Diaz, G Uguzzoni, A Pagnani, Unsupervised Inference of Protein Fitness Landscape from Deep Mutational Scan. *Mol. Biol. Evol.* **38**, 318–328 (2020).
44. Z Wu, SBJ Kan, RD Lewis, BJ Wittmann, FH Arnold, Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci.* **116**, 8852–8858 (2019).
45. RJ Fox, et al., Improving catalytic function by prosar-driven enzyme evolution. *Nat. biotechnology* **25**, 338–344 (2007).
46. PA Romero, A Krause, FH Arnold, Navigating the protein fitness landscape with gaussian processes. *Proc. Natl. Acad. Sci.* **110**, E193–E201 (2013).
47. J Otwinowski, DM McCandlish, JB Plotkin, Inferring the shape of global epistasis. *Proc. Natl. Acad. Sci.* **115**, E7550–E7558 (2018).
48. F Cadet, et al., A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. *Sci. reports* **8**, 1–15 (2018).
49. CN Bedbrook, et al., Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat. methods* **16**, 1176–1184 (2019).
50. KK Yang, Z Wu, FH Arnold, Machine-learning-guided directed evolution for protein engineering. *Nat. methods* **16**, 687–694 (2019).
51. Y Xu, et al., Deep dive into machine learning models for protein engineering. *J. chemical information modeling* **60**, 2773–2790 (2020).
52. DH Bryant, et al., Deep diversification of an aav capsid protein by machine learning. *Nat. Biotechnol.* **39**, 691–696 (2021).
53. C Bank, RT Hietpas, A Wong, DN Bolon, JD Jensen, A bayesian mcmc approach to assess the complete distribution of fitness effects of new mutations: uncovering the potential for adaptive walks in challenging environments. *Genetics* **196**, 841–852 (2014).
54. J Otwinowski, Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function. *Mol. Biol. Evol.* **35**, 2345–2354 (2018).
55. L Sesta, G Uguzzoni, J Fernandez-de Cossio-Diaz, A Pagnani, Amala: Analysis of directed evolution experiments via annealed mutational approximated landscape. *Int. J. Mol. Sci.* **22** (2021).
56. A Di Gioacchino, et al., Generative and interpretable machine learning for aptamer design and analysis of in vitro sequence selection. *PLOS Comput. Biol.* **18**, 1–31 (2022).

DRAFT

Supporting Information for

In-silico monitoring of directed evolution convergence to unveil best performing variants with credibility score

Takahiro Nemoto, Tommaso Ocari, Arthur Planul, Muge Tekinsoy, Emilia A. Zin, Deniz Dalkara, and Ulisse Ferrari

To whom correspondence should be addressed. E-mail: takahiro.nemoto@i.kyoto-u.ac.jp, deniz.dalkara@inserm.fr, ulisse.ferrari@inserm.fr.

This PDF file includes:

Figs. S1 to S7

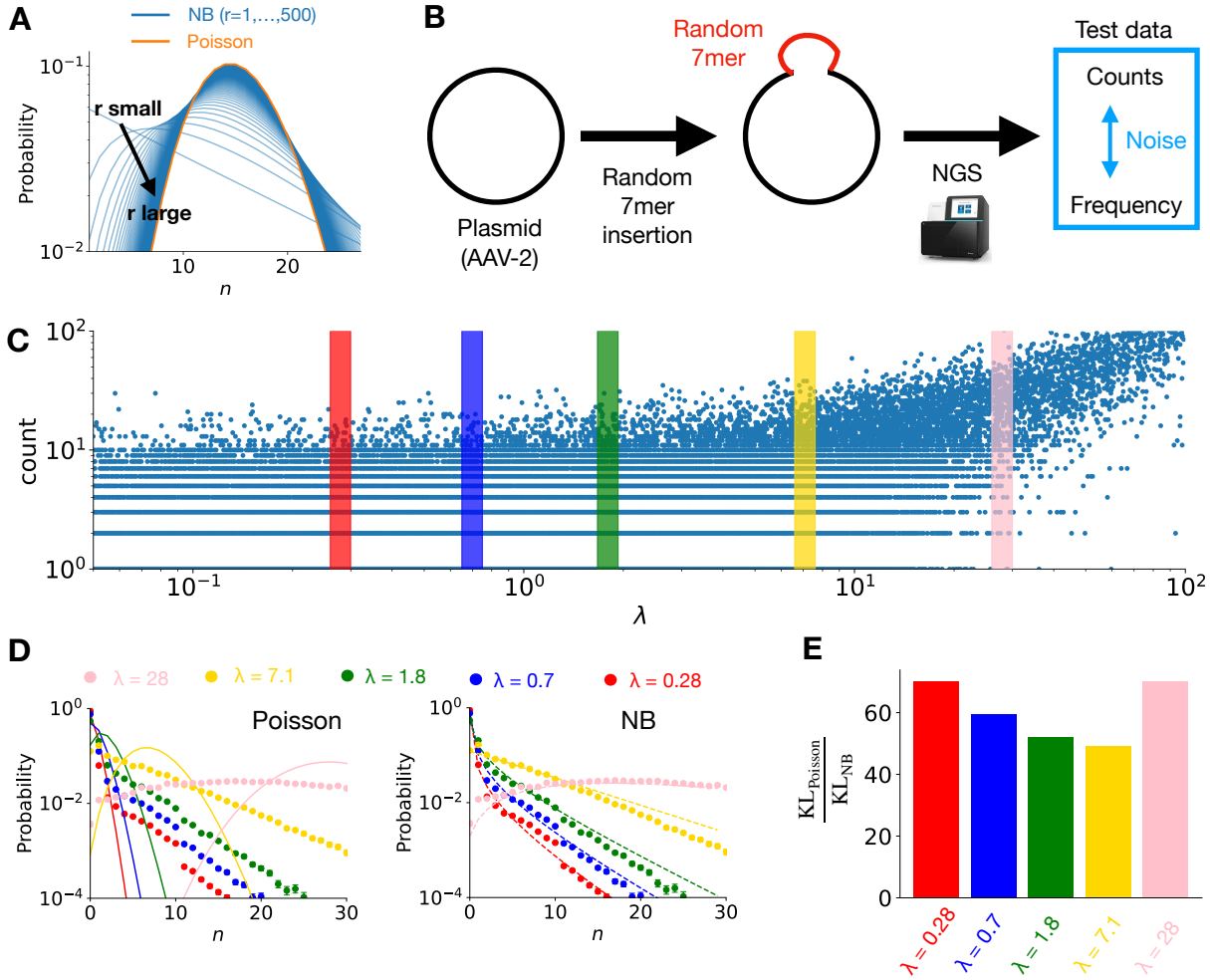

Fig. S1. Negative binomial model accounts for NGS count noise better than Poisson model. (A) The poisson distribution (orange) and the negative binomial distribution (tableau blue) with the expected value $\lambda = 15$. The dispersion parameter r for the negative binomial distribution is set to $1, 2, \ldots, 500$. The negative binomial distribution generalizes the Poisson distribution, allowing for large variances by decreasing r. It converges to the Poisson distribution in the large r limit. (B) In order to test the predictive ability of the negative binomial distribution, we performed the following experiment. Using a random peptide (of size 21 corresponding to a 7mer) as a barcode, we first barcoded a plasmid extracted from adeno-associated virus 2 (AAV2) wild type. The obtained 7-mer inserted library was then sent to NGS facility and the corresponding barcoded region was sequenced. Since these 21 nucleotides of barcode are randomly and independently generated, we can use a position weight matrix model to predict the frequency of each variant in the sample. Comparing the predicted frequency with the actual NGS reads, we investigate the noise distribution of NGS counts. (C) The graph showing the obtained counts (n) against the predicted frequencies multiplied by the total NGS reads (λ), where each point corresponds to a variant. We observe that the counts are largely dispersed. (D) The probability distribution of counts for a fixed value of λ together with the model predictions by Poisson distribution (left) and by the negative binomial distribution (right). The probability distribution is estimated in the following way: (i) picking up all the variants within 5 different colored rectangles in the panel (C). (ii) Using the variants corresponding to each rectangle, we then make a histogram of counts, which is plotted in the panel (D) as dots. For the Poisson prediction, we simply use the Poisson distribution with the mean $\lambda = 0.28, 0.7, 1.8, 7.1, 28$. For the negative binomial prediction, for each value of λ, we infer the dispersion parameter r via a maximum likelihood inference and fitted a power-law function $r = \beta\lambda^{\alpha}$ to the obtained estimations. The result is $r = 0.21\lambda^{0.744}$. Using this relation, the negative binomial distribution is then plotted for each $\lambda = 0.28, 0.7, 1.8, 7.1, 28$. (E) Comparison of predictive abilities between the poisson model and the negative binomial model. To quantify the predictive ability of each model, we use Kullback-Leibler divergence (KL) defined as $\mathrm{KL} = \sum_n P_{\mathrm{data}}(n) \log(P_{\mathrm{data}}(n)/P_{\mathrm{model}}(n))$. The ratio between KL for the poisson model and KL for the negative binomial model is plotted for each value of λ. $\mathrm{KL}_{\mathrm{Poisson}}$ itself is 0.37, 0.59, 0.87, 1.59, 3.29 for $\lambda = 0.28, 0.7, 1.8, 7.1, 28$, while $\mathrm{KL}_{\mathrm{NB}}$ is 0.0053, 0.0099, 0.017, 0.032, 0.047.

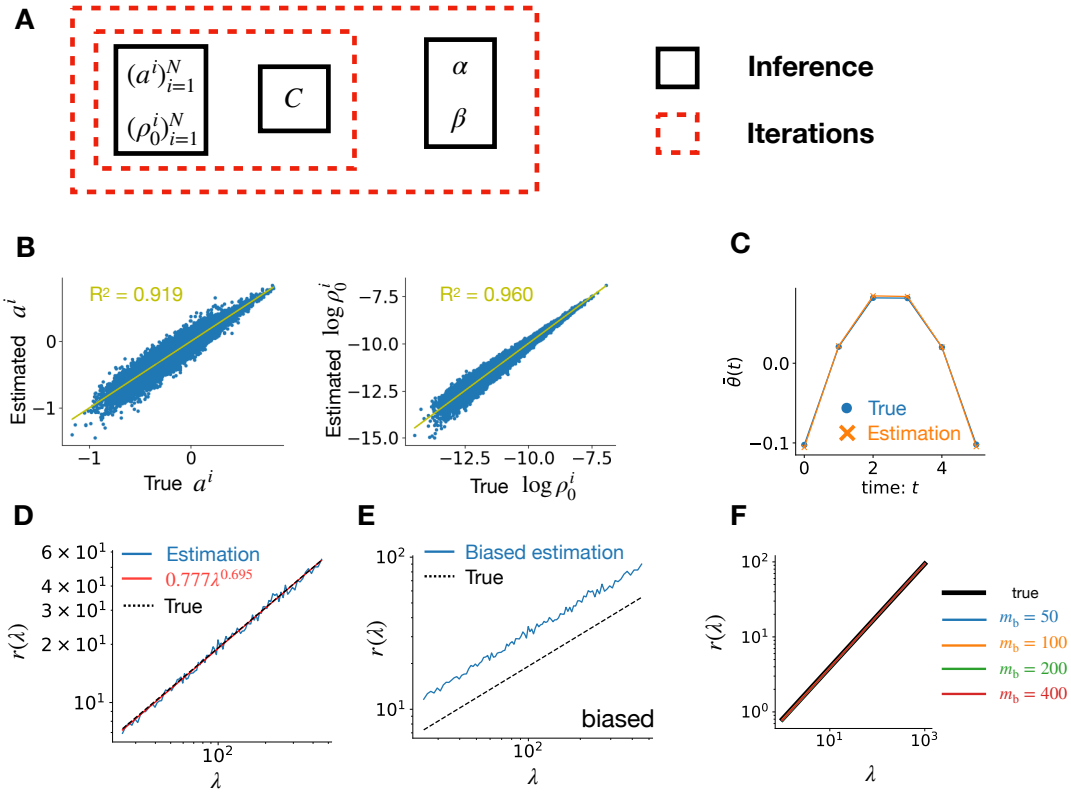

Fig. S2. Inference algorithm and synthetic teacher-student examples. (A) graphical illustration of the inference algorithm with its double loop structure. The internal loop accounts for the parameters of the exponential model (see Materials and Methods) and iterates between the inference of $(a^i, \rho_0^i)_{i=1}^M$ and C, while the external loop iterates between this internal loop and the inference of the negative binomial parameters (α, β). For the internal loop, we first infer $(a^i, \rho_0^i)_{i=1}^M$ for a fixed C, α, β by maximizing the likelihood function (2). C^* is then calculated from the obtained $(a^i, \rho_0^i)_{i=1}^M$ via $C^* = 1/\sum_i \rho_0^i \exp(a^i t)$. The linearly increasing part of $\log C^*$ is next subtracted as $\log C^* - (x^* t + y^*)$, where $x^*, y^* = \mathrm{argmax}_{x,y} \sum_t [\log C - xt - y]^2$ (fixing a gauge). The obtained quantity is $\log C$ for the next iteration. For the external loop, to obtain (α, β), we first infer the dispersion parameter r for a list of different values of λ. To do so, for a fixed value of λ, we select a set of index i and the time t by the condition $\lambda < \rho_t^i N_t < \lambda + \epsilon$ (with a small parameter ϵ). Only using these i and t, we then maximize the likelihood function (2) and determine the value of $r(\lambda)$. After obtaining the function $r(\lambda)$ for several values of λ, we fit a linear function $r(\lambda) = \alpha\lambda^\beta$ to it and determine α and β. (B,C,D) Ground truth comparisons for the synthetic data-rich dataset (Fig. 1E) after 30th iterations demonstrates that the algorithm can recover the generating parameters. In the panel B, we plot the estimated parameters $(a^i)_{i=1}^M$ (left) and $(\log \rho_0^i)_{i=1}^M$ (right) against their ground truths. The coefficient of determination R^2 is also shown. In the panel C, the normalization coefficient ($\bar{\theta}(t) \equiv \log C$) is plotted together with its ground truth. In the panel D, the estimated $r(\lambda)$ with a fitted line $\beta\lambda^\alpha$ and its ground truth are shown. (E) While estimating $r(\lambda)$ for a fixed λ, maximizing the likelihood function (2) results in a biased estimation as shown in the panel E. For fixing this, we generate a synthetic data probe using the current estimation of α_0 and β_0 with $(\rho_t^i)_{i,t}$ and use it to unbias the r estimation. More precisely, denoting the biased estimation by $r_{\text{bias}}(\lambda)$ (panel E) and the estimation of the r in the probe by $r_1(\lambda)$, the unbiased estimator plotted in the panel D is obtained as $r(\lambda) = r_{\text{bias}}(\lambda)\beta_0\lambda^{\alpha_0}/r_1(\lambda)$. (F) To determine α and β, we use only representative variants, as described in Materials and Methods. For the representative variants, we select the variants whose total counts are larger than a threshold value m_{b}. In (F), by using the synthetic data, we show that the inference results (α and β) are robust against the change of this parameter m_{b}.

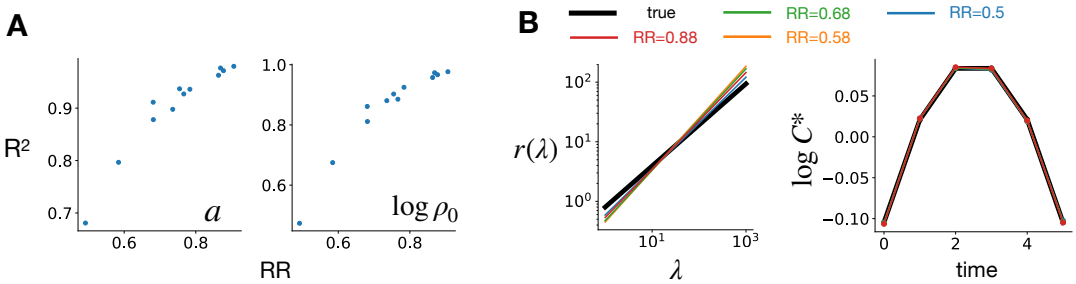

Fig. S3. RR offers a proxy for the accuracy of the estimated scores. Here we use synthetic datasets, ranging from data-poor to data-rich regimes, to show that the empirical quantity RR correlates with our ability to recover the true values of scores a^i and initial frequencies ρ_0^i. For this, we generate synthetic datasets with different values of total NGS reads N and number of variants M: $(N, M) = (10^7, 5 \times 10^5); (8 \times 10^6, 5 \times 10^5); (6 \times 10^6, 5 \times 10^5); (4 \times 10^6, 5 \times 10^5); (2 \times 10^6, 5 \times 10^5); (10^6, 5 \times 10^5); (10^7, 10^6); (8 \times 10^6, 10^6); (4 \times 10^6, 10^6); (2 \times 10^6, 10^6); (10^6, 10^6)$. In these datasets, the parameters (α, β) are the same as those used in Fig.1E, F. (A) Coefficients of determination R^2 between inferred and true values for a^i (left) and ρ_0^i (right) are plotted against RR. This demonstrates the correlation between R^2 and RR. (B) The inference of $r(\lambda)$ and C^* is robust across all synthetic datasets.

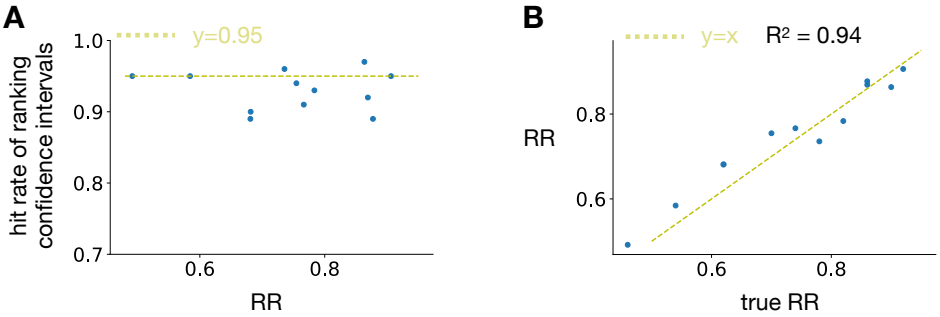

Fig. S4. Ranking CI and rank robustness are accurately estimated also in the data-poor regime. A) For all the synthetic datasets introduced in Fig. S3. hit rates of the confidence interval of ranking graphs are plotted against their RR, where the hit rate is defined as the number of true ranking (red crosses in Fig.1E, F for example) that are within the 95%-confidence intervals (green lines in Fig.1E, F), divided by 50. The hit rates fluctuate around 0.95 as expected, demonstrating the quality of our estimation of ranking CI. B) Rank robustness (RR) estimated from inferred parameters is plotted against the true value (ground truth) for the synthetic datasets. The obtained high coefficient of determination shows that ACIDES can estimate RR also in the data-poor regime.

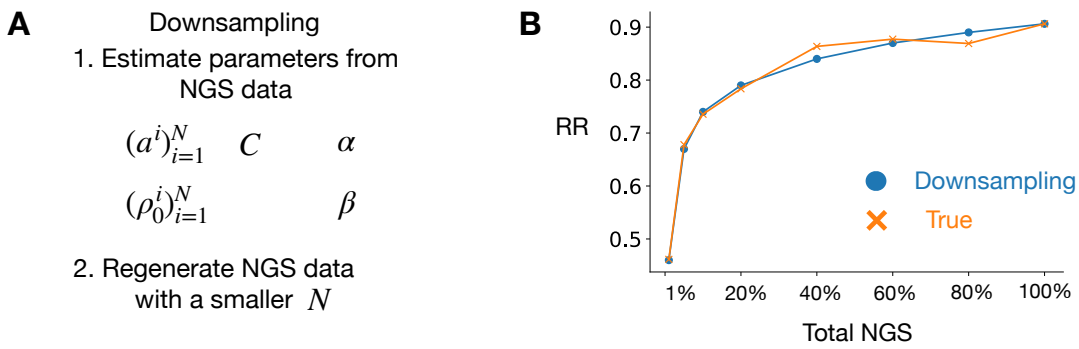

Fig. S5. Downsampling NGS counts. (A) For a given dataset of a DE experiment, we downsample the data, $i.e.$, create a synthetic dataset that corresponds to the dataset of the same DE experiments but with smaller values of the total NGS reads. For this, we first estimate the parameters of ACIDES from the dataset. We then generate synthetic NGS data using the likelihood function (2) with smaller numbers of the total NGS reads. For example, if we downsample the data to 40$\%$, we set N_t to be $0.4N_t$. We then estimate RR using ACIDES for this downsampled dataset without reinferring α, β. (B) We show the validity of this down sampling method on the synthetic data. The parameters for the synthetic data are $(N_t, M) = (10^7, 5 \times 10^5)$, $(\alpha, \beta) = (0.69, 0.8)$ and $(a^i, \log \rho_0^i)_{i=1}^M$ generated from the normal distribution with the expected values (-1,1) and the standard deviation (0.25, 1). We plot RRs obtained from this downsampling method (blue circles) and from a standard sampling method (orange crosses) as a function of the total number of NGS (where 100% means the original data). Here the standard sampling method means using ACIDES directly on the dataset with the total number of NGS $0.01xN_t$, where x is the percentage of the total NGS reads (x-axis in the panel B). We observe that our downsampling method estimates well the RR.

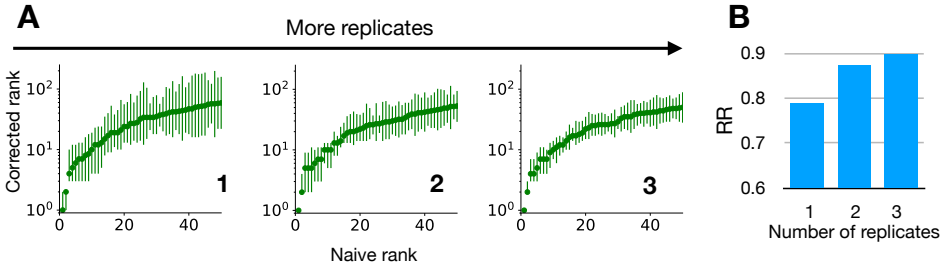

Fig. S6. Multiple replicates can be combined to increase RR. We use the first 4 rounds of Data-A (so that RR is relatively small for a single replicate), and perform ACIDES for each replicate. (A) Ranking graphs for one (left), two (middle), three (right) replicates. To combine variant scores of two replicates (denoted by a_1, a_2 with standard deviation $\delta a_1, \delta a_2$), we use $(a_1 \delta a_2^2 + a_2 \delta a_1^2)/(\delta a_1^2 + \delta a_2^2)$ for the combined score and $\sqrt{\delta a_1^2 \delta a_2^2}/(\delta a_1^2 + \delta a_2^2)$ for the combined standard deviation. (B) RR for the three cases.

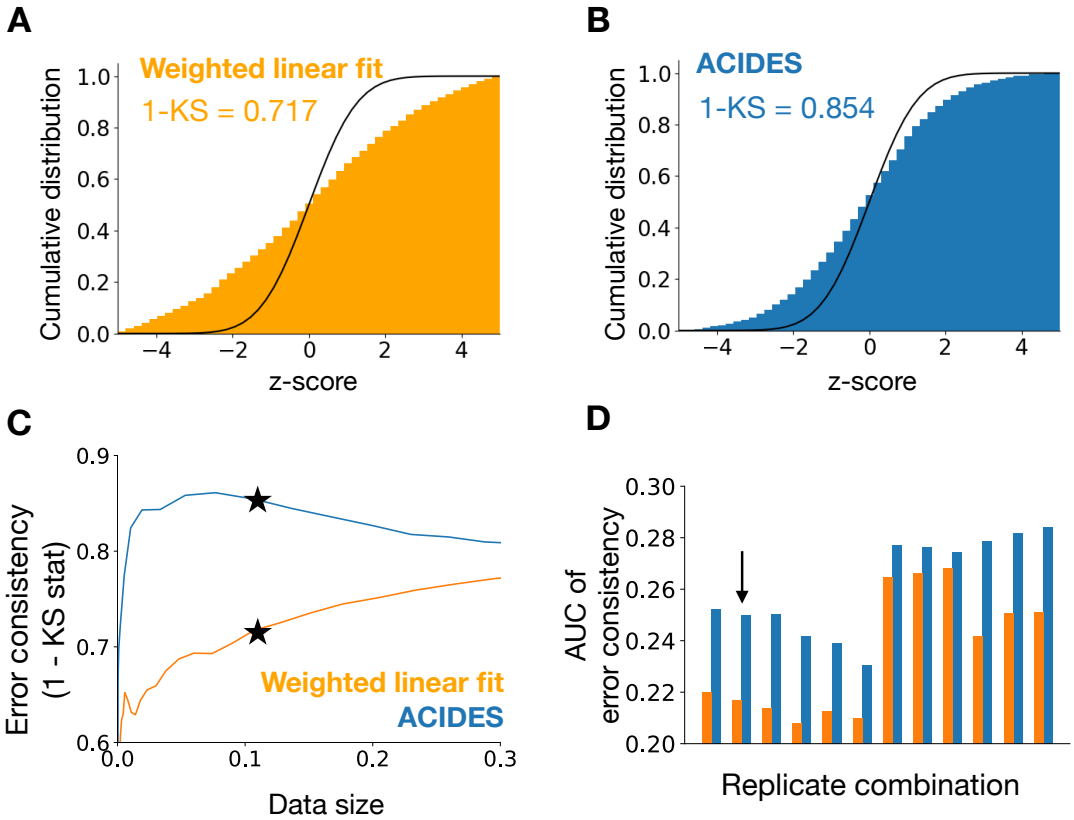

Fig. S7. ACIDES outperforms previous methods in the estimation of score's statistical errors. Using replicates in Data-A and Data-B (Table 1), we study the consistency of error bars in ACIDES and in $Enrich2$. Denoting by a_1, a_2 the scores of a variant in replicate 1 and 2 (similarly by δa_1, δa_2 the standard deviations of the scores), we study the following quantity $z = (a_1 - a_2)/\sqrt{\delta a_1^2 + \delta a_2^2}$ and compute the histogram of this quantity over different variants. Under the assumption that both scores are distributed following the normal distribution, this obtained histogram is approximated by the standard normal distribution. (A,B) Cumulative distributions of the histograms for $Enrich2$ (A) and ACIDES (B) together with the cumulative standard normal distribution. We use $1-$ Kolmogorov–Smirnov (KS) statistics (the maximum distance between two distributions measured in the y-direction) to quantify the distance between the histogram and the normal distribution. (C) To study $1 -$ KS more systematically, we introduce a threshold for the score statistical errors (Materials and Methods) by which we reduce the amount of data. For each fraction of the data, we estimate $1 -$ KS and plot them in the panel C. The stars in the panel correspond to panels A and B. (D) Finally, using all possible combinations of technical replicates in Data-A and Data-B, we compare ACIDES and $Enrich2$. We compute the area under curve (AUC) of $1 -$ KS (the panel C) for all combinations. ACIDES always shows better performance than $Enrich2$. The arrow in the panel D indicates the replicate combination used in the panels A-C.