Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Systems Biology

From sequence to function and back – High-throughput sequence-function mapping in synthetic biology

Simon Höllerer[2], Charlotte Desczyk[1,a],
Ricardo Farrera Muro[1,a] and Markus Jeschek[1,2]

Abstract

How does genetic sequence give rise to biological function? Answering this question is key to our understanding of life and the construction of synthetic biosystems that fight disease, resource scarcity and climate change. Unfortunately, the virtually infinite number of possible sequences and limitations in their functional characterization limit our current understanding of sequence-function relationships. To overcome this dilemma, several high-throughput methods to experimentally link sequences to corresponding functional properties have been developed recently. While all of these share the goal to collect sequence-function data at large scale, they differ significantly in their technical approach, functional readout and application scope. Herein, we highlight recent developments in the aspiring field of high-throughput sequence-function mapping providing a critical assessment of their potential in synthetic biology.

Addresses

[1] Institute of Microbiology, Synthetic Microbiology Group, University of Regensburg, Regensburg, D-93053, Germany

[2] Department of Biosystems Science and Engineering, Swiss Federal Institute of Technology – ETH Zurich, Basel, CH-4058, Switzerland

Corresponding author: Jeschek, Markus (markus.jeschek@ur.de)

[a] These authors contributed equally.

Current Opinion in Systems Biology 2024, **37:**100499

This review comes from a themed issue on **Synthetic biology (2023)**

Edited by **Sang Yup Lee** and **Byung-Kwan Cho**

For a complete overview see the Issue and the Editorial

Available online 13 December 2023

https://doi.org/10.1016/j.coisb.2023.100499

Keywords

Sequence-function mapping, Sequence-activity landscapes, Next-generation sequencing, DNA recording, DNA recorder, RNA-Seq, Flow-Seq, Sort-Seq, Recombinases, Integrases, Machine learning in biology, Data-driven modelling, Synthetic biology..

Introduction – High-throughput (HTP) sequence-function mapping

Synthetic biology targets the design of biological systems following user-defined functional specifications, which has far-reaching socioeconomic implications and the potential to address critical global challenges of our time [1]. Recent advancements in DNA synthesis and assembly allow us to rewire the genetic make-up of entire organisms [2,3], which opens up tremendous opportunities in this context. However, we are only beginning to understand which genetic sequences we need to “write” to build biosystems with desired functions, which can, at least to a certain extent, be compensated for by trial-and-error experimentation. Crucially, to advance towards truly rational biosystems design and thus unlock a much wider potential of synthetic biology, it is imperative to develop a substantiated understanding of sequence-function relationships on all levels of the central dogma of molecular biology (Figure 1). Unfortunately, such knowledge is extremely difficult to obtain due to the enormous complexity of even the simplest biosystems as most vividly embodied by the extremely large number of possible sequences, which cannot be exhaustively tested in experiments (i.e. combinatorial explosion) [4].

This complexity can only be addressed by experimentally linking large numbers of sequence variants to their corresponding function (i.e. HTP sequence-function mapping) to at least cover a representative subsample of the vast sequence space. Furthermore, the resulting data must be capitalized on to model the underlying sequence-function relationship and thus enable the *in silico* prediction and forward design of untested sequence variants with high accuracy (Figure 1). For the latter, machine learning (ML) provides an extremely powerful toolbox to model highly non-linear sequence-function interdependencies in a data-driven fashion even in the absence of *a priori* mechanistic knowledge as reviewed elsewhere [5,6]. However, these data-intense modelling approaches critically rely on access to high volumes of sequence-function data thus shifting the bottleneck of the development pipeline to the functional testing of sequence variants through experiments [7,8]. In conventional approaches, variants are individually cultivated or produced (e.g. in deep-well plates) and subsequently subjected to a functional assay. Crucially, sequence information about the variants is obtained via separate experimental procedures and devices imposing a need to retroactively link each variant's sequence back to the corresponding function, which becomes prohibitively challenging for variant numbers beyond 10^4 [4,8—10].

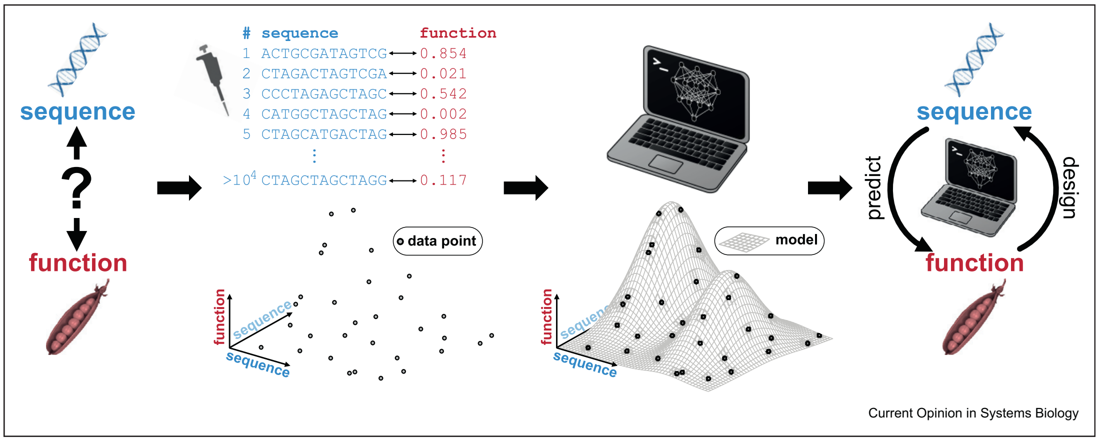
Approaches for HTP sequence-function mapping

In the past decade, different methods have been developed to overcome this bottleneck in HTP sequence-function data generation. This development has been largely fueled by the rapid advancement and cost reduction of next-generation sequencing (NGS) techniques [11], which all of these methods capitalize on. While nowadays it is rather straightforward to collect sequence information via NGS, methods for HTP sequence-function mapping manage to also convert the functional trait(s) of interest into an NGS-readable and ideally quantitative output (Figure 2). Consequently, both sequence and function can be read in NGS for extremely large numbers of variants (up to several hundred million per NGS run). To this end, methods for HTP sequence-function mapping can be grouped according to the strategy of how the functional information is converted into an NGS-readable output, which represents the key distinctive feature of each method. For the purpose of this review, we distinguish methods based on i: cell sorting, ii: RNA sequencing, iii: DNA recorders, and iv: competitive enrichment. Crucially, the groups differ significantly in terms of several practical features and application scope, which shall be highlighted in this review in a comparative fashion as a user-oriented guideline for method selection. Specifically, we focus herein on methods that can obtain more than 10^4 sequence-function data pairs per experiment and their use in the context of synthetic biology and large libraries of parts or genetic elements. This is discussed herein providing recent examples without raising the claim of comprehensiveness. For instance, we omit general omics methods (e.g., transcriptome- or proteome-wide approaches), which exceed the scope of this short review but are likewise of high importance as reviewed elsewhere [12,13].

Cell sorting-based methods

These methods rely on a combination of flow-cytometric cell sorting and NGS, and are thus commonly referred to as Sort-Seq or Flow-Seq [14]. Here, the functional trait of interest is coupled to a fluorescent reporter system such as the expression of a fluorescent protein (Figure 3a). Consequently, a library of functionally diverse sequence variants can be constructed and subsequently sorted into bins depending on each variant's fluorescence, for instance by fluorescence-activated cell sorting (FACS, Figure 3b). After, DNA is extracted from each bin, and target DNA fragments are amplified and ligated with NGS adapters containing bin-specific indices. Samples are pooled and collectively subjected to NGS to obtain the number of reads for each library member and fluorescence bin. Finally, a quantitative functional readout is statistically inferred for each library

Figure 1

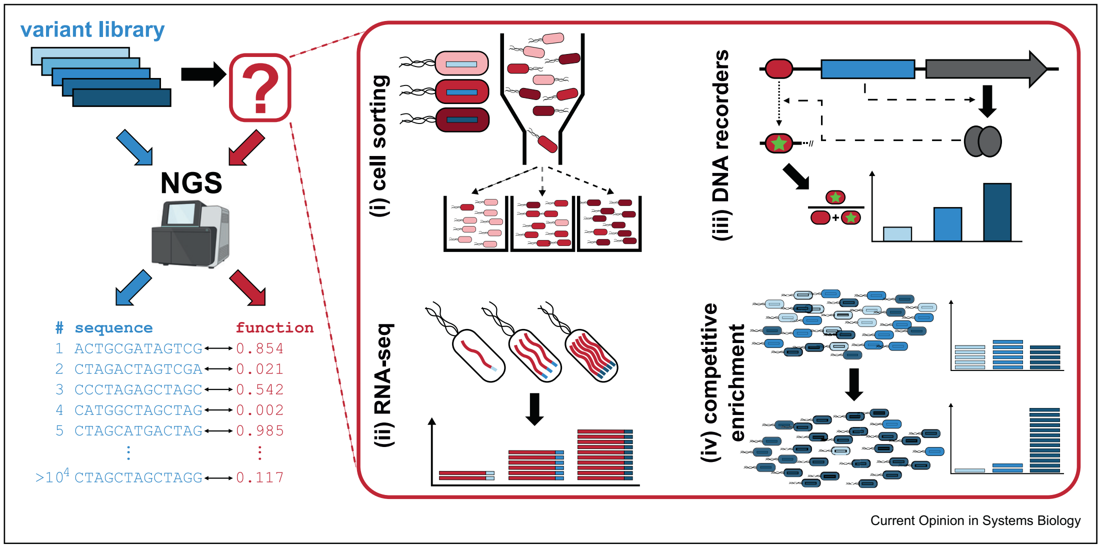


```
# sequence           function
1 ACTGCGATAGTCG⟷0.854
2 CTAGACTAGTCGA⟷0.021
3 CCCTAGAGCTAGC⟷0.542
4 CATGGCTAGCTAG⟷0.002
5 CTAGCATGACTAG⟷0.985
      ⋮                ⋮
>10^4 CTAGCTAGCTAGG⟷0.117
```

Current Opinion in Systems Biology

HTP sequence-function mapping enables data-driven modelling of biosystems. While of high importance for synthetic biology, our understanding of the relationship between genetic sequence and corresponding function is limited. In order to overcome this dilemma, experimental datasets linking large numbers of sequence variants to the corresponding quantitative functional trait(s) of interest are required. Such datasets can then be used to model sequence-function relationships *in silico*, e.g. by machine learning, which enables to predict and design functional properties of untested sequences in a straightforward fashion.

Figure 2



Schematic overview of HTP sequence-function mapping approaches. The distinctive feature of the approaches is the principle of how functional information is converted into an NGS-readable, quantitative output. To this end, methods relying on cell sorting (i), RNA sequencing (ii), DNA recorders (iii) and competitive enrichment (iv) can be distinguished.

member based on the relative per-bin read counts of each variant. More precisely, the relative read frequency of each variant in a given bin is multiplied by the mean fluorescence of this bin and subsequently averaged across all bins. The resulting statistically inferred, weighted fluorescence represents a quantitative readout for each variant's function. First described by Kinney and coworkers in 2010 [15], cell sorting-based approaches for HTP sequence-function mapping have been used for a large variety of different applications. Examples to that end include genetic elements governing DNA replication [16], transcription [15,17–22], translation [19,20,23–26], RNA/protein stability [27–29] and protein function [30,31] in both pro- and eukaryotes. In a noteworthy recent study in that context, Regev and coworkers reported on the characterization of transcription from over 100 million randomly generated sequences in yeast [18]. Here, the authors show that even at an extremely low sequencing coverage (most variants were represented by a single sequencing read only), the resulting big sequence-function datasets can be used to reliably predict transcriptional behavior by ML.

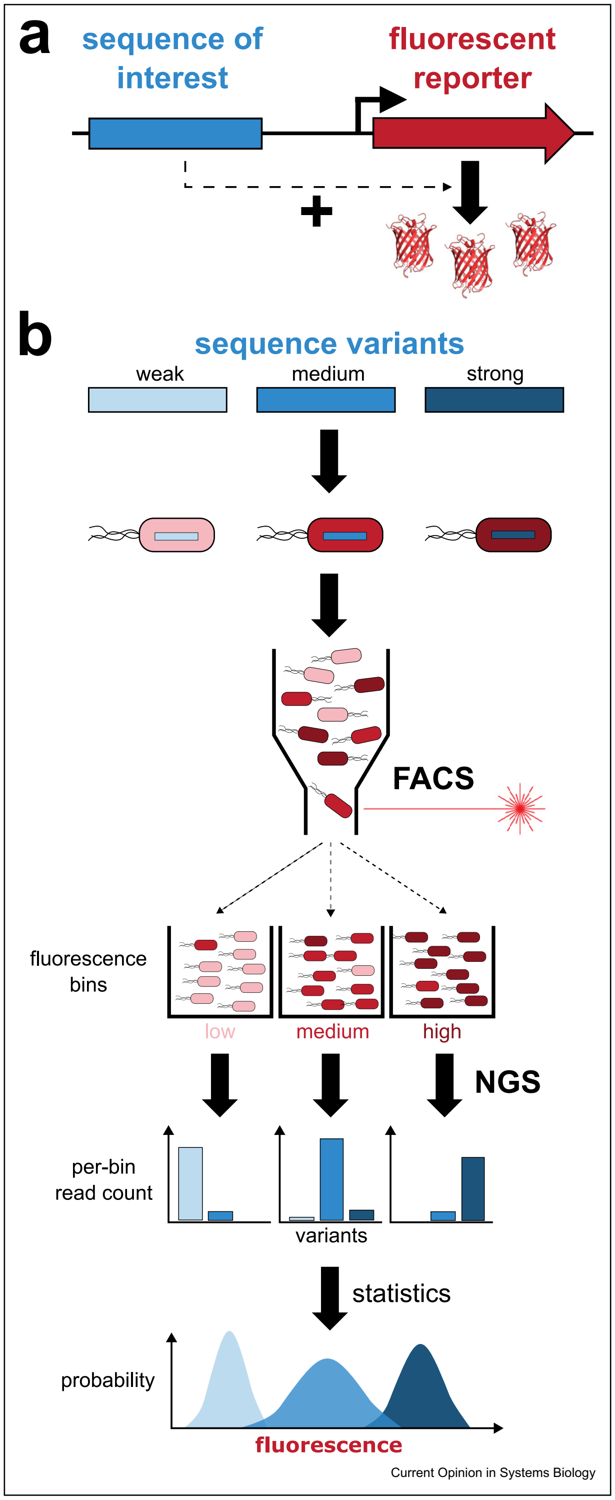
RNA sequencing

A second group of methods for HTP sequence-function mapping relies on RNA sequencing (RNA-Seq) [32] in order to capture how genetic elements of interest affect RNA (mostly mRNA) levels. RNA-Seq was originally developed for whole-cell or whole-organism transcriptomic analyses [33], which has recently advanced to the readout of RNA profiles down to single-cell resolution [34]. Furthermore, it has been adapted for numerous applications in synthetic biology to study the effect of synthetic libraries of genetic elements on transcription. Briefly, a library of the investigated element is introduced into the host organism of choice leading to a variant-dependent, differential RNA expression (Figure 4). The latter can be quantified via RNA extraction followed by reverse transcription into cDNA and subsequent NGS. The relative frequency of NGS reads obtained for each RNA is used as a functional readout, which directly correlates with the cellular RNA level. To this end, the number of sequencing reads per RNA variant can differ up to several orders of magnitude [35]. Furthermore, barcoding (so-called unique molecular identifiers, UMIs) can be used to enable HTP testing of genetic elements not encoded on the RNA itself (e.g. promoters, etc.) or population phenomena such as cell-to-cell variability [36,37]. RNA-Seq has been exploited to assess libraries of promoters in pro- and eukaryotes as well as *in vitro* [36,38–41] and different genetic elements affecting RNA stability or degradation [42,43]. In a recent study, Hossain et al. used RNA-Seq to characterize 4350 bacterial and 1722 yeast promoters designed to be non-repetitive, which reduced recombination and enhanced genetic stability [39].

DNA recorders

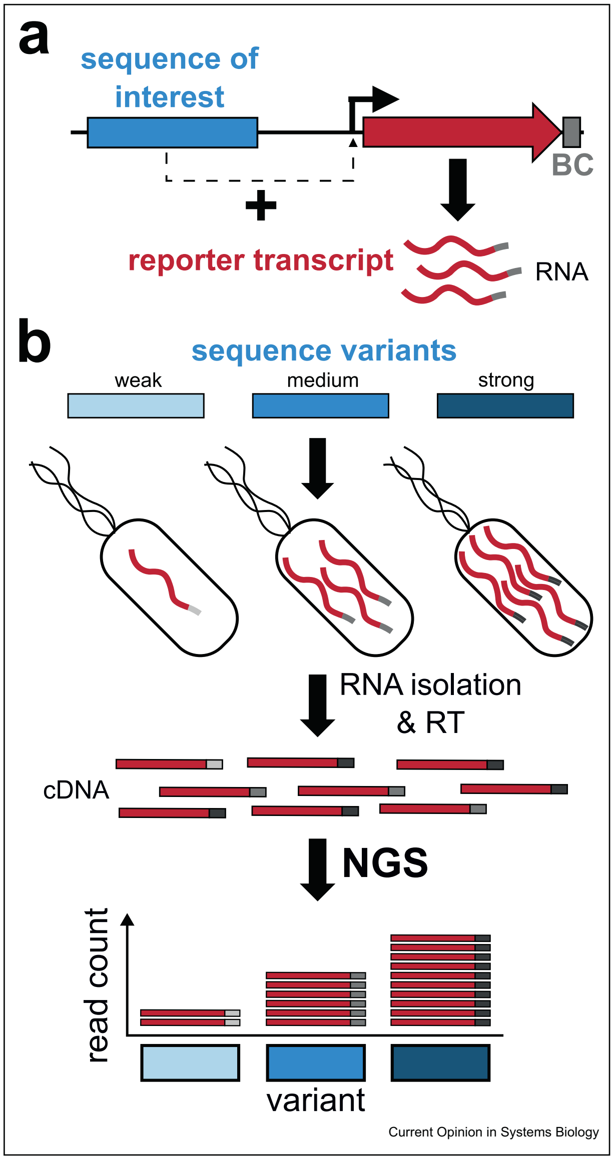
A third approach to HTP sequence-function mapping uses DNA-modifying enzymes (DMEs) as reporters.

Figure 3



Principle of cell sorting-based methods for HTP sequence-function mapping. a: The function of a genetic element of interest is coupled to a fluorescent signal such as the expression of a fluorescent reporter. **b**: This allows for functional sorting of genetic element variants into bins of different mean fluorescence, e.g. by FACS. NGS of DNA extracted from the different fluorescence bins is then used to obtain the relative per-bin frequency of each variant, from which a fluorescence distribution for each candidate can be reconstructed through statistical inference.

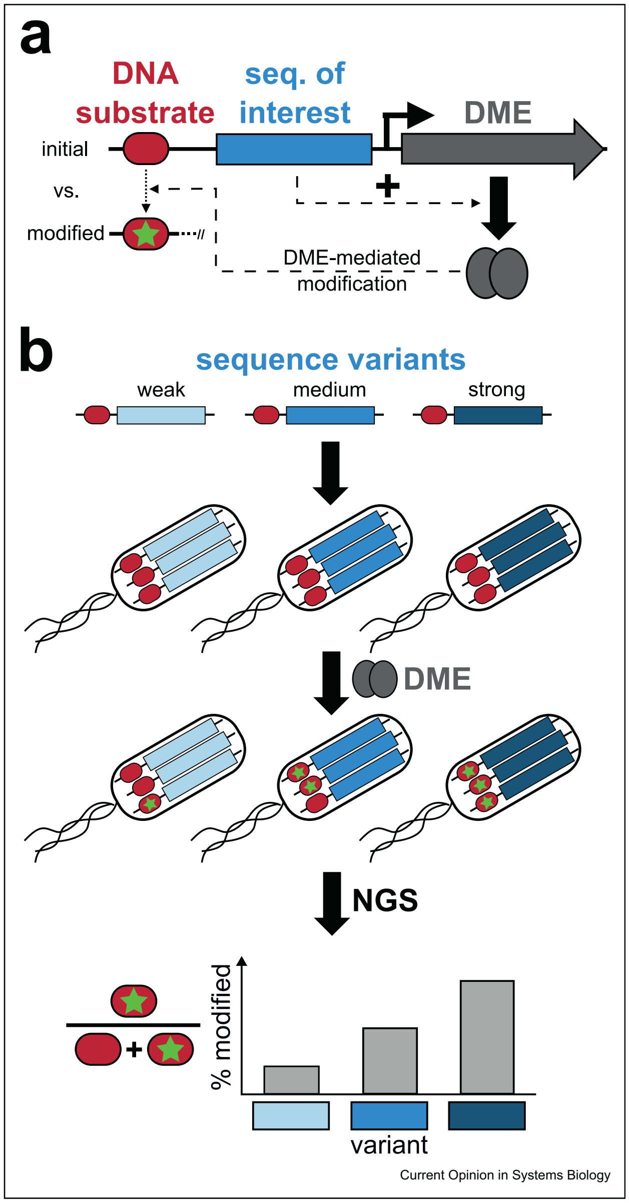
Figure 4



Principle of RNA sequencing-based methods for HTP sequence-function mapping. a: The function of a genetic element of interest is coupled to the level of a reporter transcript linked to a variant-specific barcode (BC). **b**: The stronger a given variant of the genetic element, the higher the resulting transcript levels will be. The latter can be quantified through RNA isolation followed by reverse transcription (RT) and NGS, where the relative frequency of each variant's BC is determined as proxy for the quantitative function of each candidate. Note that barcoding is required in cases where the genetic element is not encoded on the reporter transcript itself (e.g. for promoters, trans-regulatory elements, etc.).

Once activated, DMEs alter their DNA substrate either in a sequence-specific fashion or at random (Figure 5a) [44]. If coupled to DME expression, the function of a genetic element of interest can be recorded in DNA with the frequency of DNA modifications being proportional to the quantitative function of this element (Figure 5b). This allows to determine both sequence

Figure 5



Principle of HTP sequence-function mapping by DNA recorders. a: The function of a genetic element of interest is coupled to the expression of a DNA-modifying enzyme (DME), which can convert its cognate DNA substrate from an initial into a modified state (highlighted by green star). **b**: The stronger a given variant of the genetic element invokes expression, the more DNA modification will occur, which can be quantified by NGS determining the fraction of modified DNA substrates amongst all reads for each tested variant.

and quantitative function of genetic elements at high throughput solely using NGS, rendering additional functional assays obsolete. Congenially, it was shown that e.g. integrases and CRISPR-Cas systems can be used to record intra- and extracellular stimuli in DNA resolved across time and space [45–47], which is not discussed herein but nicely reviewed elsewhere [44].

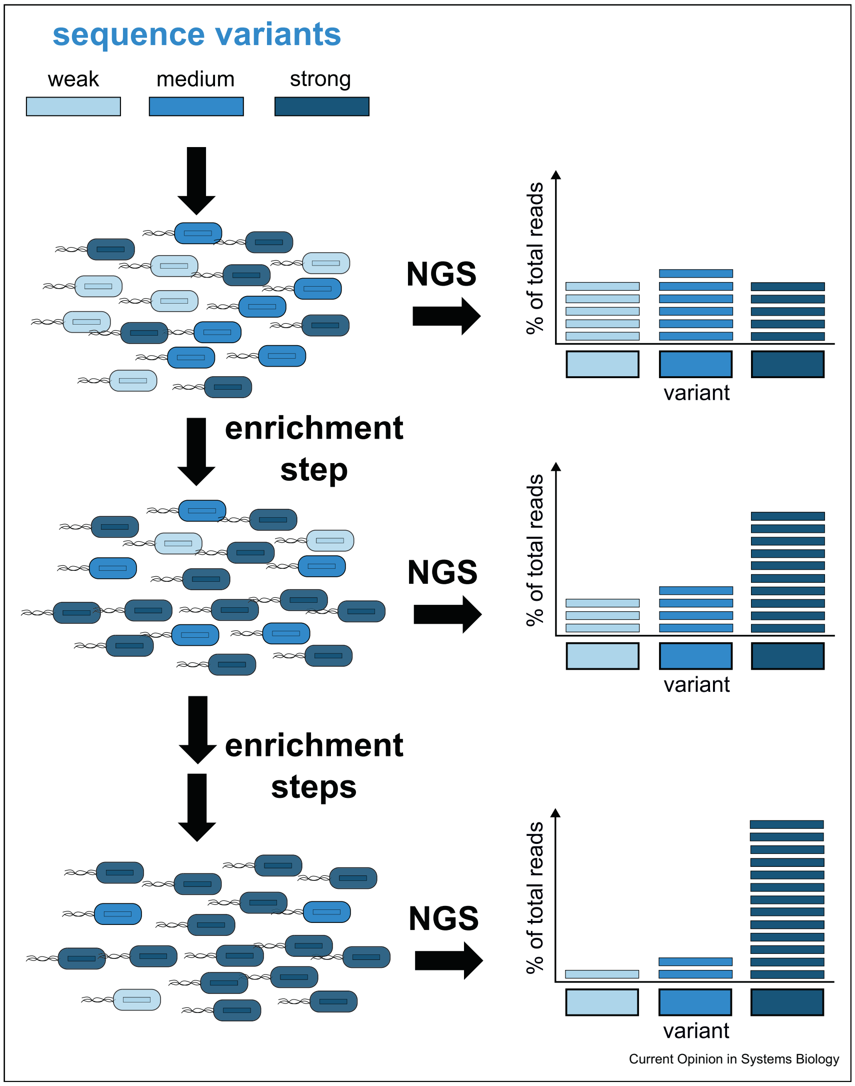
Anderson and coworkers developed an enzyme-coupled assay to record methylation activity in DNA *in vitro* [48]. The method relies on depletion of the methyl donor S-adenosyl methionine (SAM) by candidate methyltransferases. If a candidate is active for a tested substrate, SAM is depleted in the reaction mixture which translates to a reduced DNA-methylation activity upon subsequent addition of a DNA methylase reporter. DNA methylation is finally assessed by digestion with a methylation-sensitive restriction enzyme leading to differential fragmentation depending on the candidate's methyltransferase activity, which can be assessed by NGS. Combined with compartmentalized *in vitro* transcription and translation, this DNA recorder could in principle be used to perform HTP sequence-function mapping for methyltransferases, which, however, remains to be demonstrated. A DNA-methylation recorder amenable for *in vivo* application was later described by Yus et al. [49]. The corresponding approach termed "Expression Level Monitoring by DNA methylation" (ELM-Seq) uses the *Escherichia coli* DNA adenine methyltransferase (Dam), which methylates GATC sites yielding GA^{m}TC. In the study, approximately 250000 promoters and 5′-untranslated regions (5′-UTRs) were employed to control expression of Dam in *Mycoplasma pneumoniae*. The function of these gene-regulatory elements was recorded in an array of four GATC sites placed upstream of the diversified region. To enable readout of Dam activity via NGS, genomic DNA is extracted and separately treated with two methylation-sensitive enzymes, MboI and DpnI, which digest only GATC and GA^{m}TC sites, respectively. After, only uncut DNA fragments from both reactions are PCR-amplified and subjected to NGS to obtain the ratio of methylated versus unmethylated arrays for each library member as a metric directly correlating with function (here gene expression). Notably, ELM-Seq was later also used to study the impact of C-terminal amino acid composition on protein expression in *M. pneumoniae* [50], while application in other hosts remains to be demonstrated.

Besides DNA methylases, site-specific recombinases (so-called "integrases") have been employed as DMEs for HTP sequence-function mapping. In particular the large serine recombinase Bxb1 from the homonymous mycobacteriophage has proven to be a versatile and efficient recombination tool for diverse applications across the life sciences and different species [51,52]. Capitalizing on these features, we have recently developed a DNA-recording technique for HTP sequence-function mapping termed "ultradeep Acquisition of Sequence-Phenotype Interrelations (uASPIre) [53]. In uASPIre, the function of a genetic element of interest is coupled to the expression of Bxb1, which converts its cognate DNA substrate flanked by attachment sites *attB* and *attP* from an initial to a recombined state. If placed on the same DNA molecule as Bxb1's substrate, both the sequence of the genetic element and its function (imprinted in Bxb1's substrate) can be read via NGS at extremely high throughput. Further, reading of multiple copies of the DNA substrate for each candidate genetic element allows to determine the fraction of both states (initial vs. modified) for each variant as a quantitative, internally normalized readout of function. In a proof-of-concept study, we demonstrated feasibility of uASPIre measuring kinetic translation from over 300000 ribosome binding sites (RBSs) across nine time points in *E. coli* in a single experiment [53], which was later expanded to more than 1.2 million combinations of 5′-UTRs and coding sequences (CDSs) [54]. Moreover, the resulting big sequence-function data enabled us to train a deep learning model that accurately predicts translational strength directly from RBS sequence and unveiled a quadruplet base-pairing between initiator tRNA and mRNA through statistical analyses [53,54].

Competitive enrichment

A fourth group of approaches to obtain sequence-function data at large scale is based on the competitive enrichment (or depletion) of variants (Figure 6). Here, the target function of the genetic element of interest leads to differential growth, toxicity, proliferation or binding of variants and, in consequence, to a selective amplification of stronger over weaker sequences (or *vice versa*). The latter is tracked by NGS comparing the

Figure 6



Principle of HTP sequence-function mapping by competitive enrichment. The function of a genetic element of interest is coupled to its proliferation upon competing with all other variants in the library. This can be mediated through differential growth, toxicity or binding of variants leading to selective amplification of stronger over weaker sequences. The latter can be tracked by NGS before and after the different enrichment steps, where the relative abundance of reads is determined for each variant as a readout correlating with function.

relative abundance of variants in pre- and post-enrichment samples. While these methods share the concept of a direct competition amongst variants, they are very diverse in terms of the underlying molecular principles and experimental procedures. Therefore, the examples highlighted hereafter are to be understood merely as a representative yet non-comprehensive selection.

A widely used strategy to that end is the coupling of the functional trait of interest to growth or proliferation of the respective variant. To this end, both positive (e.g. antibiotic resistance gene, enzyme producing an essential metabolite) and negative (e.g. inhibitory/toxic gene product) markers can be employed to invoke differential growth under selective conditions. For instance, Seelig et al. have used an essential gene for histidine biosynthesis (HIS3) as a marker to characterize translation efficiency from approximately 490000 5′-UTRs in *Saccharomyces cerevisiae* through competitive growth in histidine-free media [55]. This led to the identification of several sequence features and motifs critical for translation as well as to a neural network model predicting the behavior of untested sequences with good confidence. Toxic gene products can likewise be expressed to apply selective pressure on variants in the library. This was for example used to assess the growth inhibitory effect of antimicrobial peptides in a highly parallelized manner using cell surface display or cytosolic expression in *E. coli* [56,57]. Lastly, functional traits of interest may also be coupled to the highly efficient proliferation of phages within continuous cultures of host bacteria, a strategy known as phage-assisted continuous evolution (PACE). However, obtaining large-scale sequence-function data from such “deep evolutionary” approaches remains challenging due to limitations in either maximum read length (Illumina) or read number and error rates (long-read techniques) of current NGS platforms. Notably, Liu and coworkers have recently used ML to reconstruct full-length genotypes from short-read NGS data of continuous evolution campaigns [58]. This allows to link sequence to function (or fitness) with high confidence and thus enables access to evolutionary trajectories for large variant numbers.

As an alternative to growth selection, competitive binding or coupling of RNA or DNA sequences to a desired target is used to enrich or deplete variants based on their function. Examples include polysome/ribosome profiling, chromatin immunoprecipitation (ChIP) and systematic evolution of ligands by exponential enrichment (SELEX), which can be combined with NGS for HTP sequence-function mapping. For instance, polysome profiling has been used to study the effects of mutations in the 5′-UTR and CDS on polysome loading in *E. coli* and human cells [20,59]. Further, the Seelig group has shown that the resulting sequence-function data enables the prediction and forward design of polysome loading through data-driven modelling [59]. Lastly, SELEX, a cyclic procedure to iteratively select for DNA or RNA aptamers with high-affinity for a desired target molecule, can be combined with NGS to obtain sequence-function data at large scale. While not strictly required to obtain superior binders, NGS provides insight into enrichment trajectories for millions of sequences during SELEX [60]. This can help to streamline and optimize the process of aptamer selection, for example by avoiding the loss of promising sequences in intermediate cycles or by guiding sequence design in a data-driven manner [61,62].

Considerations for method selection

While the presented methods share the same purpose, there is no “universal” approach to HTP sequence-function mapping. The available approaches differ in several technical aspects as well as their application scope, which has critical implications for method selection by the user. This shall be elaborated on in more detail hereafter and major aspects are summarized in Table 1.

Accessibility

While all methods use NGS including sample preparation at comparable degrees of complexity, protocols and instrumentation required for the generation of the functional readout differ significantly. Cell-sorting approaches based on fluorescence are well established and enable a high throughput, but they rely on sophisticated instruments and dedicated personnel, which is not available to many research teams and institutions. By contrast, the other methods largely rely only on comparably simple cultivation schemes from which samples can be directly drawn rendering them accessible to a wider user range. An exception to that end are long-term competitive selection schemes, which are technically challenging due to the risk of escape mutants (e.g. in biofilms) or contamination [63]. Furthermore, the effort and required expertise for the final sample preparation varies strongly. To this end, some of the approaches rely only on DNA extraction and amplification (e.g. cell sorting, recombination recorders, enrichment approaches) whereas for others more elaborate manipulation techniques are required (e.g. RNA extraction and cDNA generation, polysome/ribosome isolation, immunoprecipitation).

Kinetic measurements

Crucially, these aspects of “experimental simplicity” are not only important in terms of convenience and costs, but they also directly affect the possibility to measure function at high kinetic resolution. Dynamics are key to our understanding of biological processes, many of which occur within minutes such as transcription or translation. More intricate, multi-step sample

Table 1

Comparison of methods for HTP sequence-function mapping.

Method group	Variant distinction by	Application level	Kinetic resolution	Sources of error/bias	Functional readout characteristics	Exemplary references
Cell sorting (Sort-/Flow-Seq etc.)	fluorescence (mostly)	• DNA • RNA • protein	Low	• PCR/growth amplification • statistical inference	• quantitative • linear • distribution • high fold-change • transient	[15–31]
RNA sequencing (RNA-Seq)	RNA abundance	• RNA	High	• RNA extraction • RT • PCR amplification • barcoding	• semi-quantitative • linear • distribution (single-cell) • high fold-change • transient	[36,38–43,73,74]
DNA recording	DNA modification (e.g. methylation, recombination)	• DNA • RNA • protein	High	• PCR amplification • reporter-borne biases	• quantitative • non-linear • point estimate • medium fold-change • stable	[48–50,53,54]
Competitive enrichment	selectable phenotypes (e.g. growth, binding)	• DNA • RNA • protein	High	• PCR/growth amplification • pulldown biases	• qualitative • non-linear • point estimate • stable	[20,55–62,73]

RT: reverse transcription.

processing generally reduces the number of timepoints that can effectively be interrogated in a biologically meaningful way. As an example, even state-of-the-art cell-sorting devices need several hours of sorting to process large libraries at sufficient oversampling, which is further prolonged due to the need to sort into several bins of different fluorescence. This sequential analysis of variants effectively prohibits measurements in intervals down to few minutes. Further, it can introduce experimental error due to different processing times for each variant as fluorescence is likely to change over a few hours of incubation times even at low temperatures [64]. Therefore, fully parallelized methods with short post-processing procedures (e.g. where samples can be directly snap-frozen or quenched after withdrawal) can be viewed as superior in cases where kinetic measurements are essential. Here, DNA-recording can be advantageous since both sequence and functional information are stably recorded in DNA, which allows to defer post-processing by any required time without affecting the results.

Characteristics of the functional readout

Critical implications arise from the type of output that is generated by each method, which should be carefully considered by the user. For instance, RNA-Seq or enrichment-based methods allow for reliable *in situ* comparison of variants contained in a given library. However, the performance of each variant strongly depends on the performance of all other variants in the library. For instance, a promotor of intermediate strength will result in a high fraction of sequencing reads for its cognate mRNA/cDNA when tested against a library of weaker candidates whereas it will be strongly underrepresented in a library of strong promoters. Similar arguments may be made for enrichment-based approaches where variants directly compete for a limiting resource (e.g. growth-limiting nutrients or binding sites). This renders quantitative measurements of function difficult in these approaches, which can be compensated for to a certain degree by the deliberate introduction (or "spiking") of standard sequences with known quantitative function [65]. By contrast, Sort-Seq and DNA-recording approaches allow for a robust quantitative determination of each variants function and similarly standard sequences can be included to allow for cross-experiment normalization as we have previously shown [53]. Furthermore, they allow to assess also weakly active variants, which are commonly depleted in alternative approaches. In cases where population phenomena such as cell-to-cell variability are of interest, Sort-Seq offers another critical advantage since it delivers a distribution of the functional readout for each variant. By contrast, the other methods commonly deliver point estimates corresponding to the average readout for each variant. However, UMIs can be used to also measure variability in the functional readout, which has for instance been demonstrated for RNA-Seq down to the level of single cells [36,37]. Lastly, the accessible resolution and fold-change of the functional readout are an important consideration. Enrichment-based methods deliver hardly any information to that end, whereas RNA-Seq and sorting-based approaches are capable to span multiple orders of magnitude in fold-changes at high resolution [35]. DNA-recording approaches are currently inferior to that end since the number of modifiable DNA-substrate molecules is limited, even when using multi-copy plasmid systems. This can be compensated for by kinetic measurements with high temporal resolution (see above) [53,54]. Furthermore, DME-substrates with multiple modification sites can be envisioned, which would further increase the throughput and dynamic range of these methods.

Sources of error and bias

While none of the presented methods can be considered as bias- or error-free, substantial differences in the sources and types of these undesired effects exist. A critical step in this context is the amplification of DNA, which is in most cases required to obtain sufficient material for NGS. This is commonly done by PCR, which is a well-known (yet often underappreciated) source of biases since different sequence variants are amplified with different efficiency [66,67]. Therefore, unnecessary PCR steps and cycles should be avoided and UMIs can be used in primers to correct for such biases [68]. As an alternative, clonal expansion by growth after the actual experiment can be used, which, however, bears a similar risk due to variant-specific differences in growth rates (e.g. due to metabolic burden). This can lead to the underrepresentation or loss of promising variants. Thus, if possible, any amplification steps should be avoided. To this end, we have recently shown that an entirely PCR-free protocol completely removed bias compared to PCR-based sample preparation in a recombinase-based DNA recording approach [53]. However, such sample preparation approaches are not possible in many cases due to a very low amount of starting material (e.g. for RNA- or Flow-Seq). Secondly, in cases where *in situ* barcoding (UMIs) is required, potential effects introduced by the barcodes themselves must be considered. As an example, RNA-Seq requires barcoding of the transcript itself in many cases (e.g. for testing of synthetic promoter libraries), which can affect transcript abundance and also subsequent amplification (see above). In these cases, the use of multiple different barcodes for each variant can be used to average out barcode-specific errors as previously suggested [36]. Last, statistical errors must be considered during data analysis. Here, the number of NGS reads obtained per variant (i.e. coverage) as well as the occurrence of sequencing errors are critical parameters for all of the introduced approaches to HTP sequence-function mapping. In our own experience, read errors can occur quite frequently even in Illumina NGS. To this end, we have observed the same “false” (i.e. physically inexistent) sequence appearing in several dozens of reads. The latter can happen if a single variant is strongly overrepresented (e.g. the library parent), which increases the likelihood that the same read error occurs for this variant many times. Applying stringent minimal read coverage thresholds, ideally at least 100 reads per variant, can help to minimize these artifacts. Furthermore, clustering based on sequence similarity can be used to map such errors back to the actual variants only in highly diverse libraries. Crucially, claims made on the basis of few or even single sequencing reads must be carefully put into perspective. Unfortunately, there are currently no commonly applied “good practices” for read-count thresholds, which are hard to establish. While a variety of bioinformatic tools for processing of NGS data exists (see Ref. [69] for a recent overview), post-sequencing steps often rely on customized algorithms and are thus hard to fully standardize. Lastly, method-specific biases must be considered. Examples to that end include biases in RNA-extraction and reverse transcription [67], errors in gating and statistical inference in cell sorting-based methods [14], reporter-borne global biases such as toxicity and non-linear relationships between readout and function of interest in DNA recording approaches [53,70], and biases introduced through pulldown procedures [71]. To mitigate such effects, various tools for normalization and statistical analysis of HTP sequencing data exist, which exceeds the scope of this review but is reviewed elsewhere (e.g. Ref. [72]).

Application scope

Lastly, the presented methods for HTP sequence-function mapping differ in terms of the genetic elements and corresponding biological processes that can be interrogated. Approaches based on cell sorting, DNA recording or competitive enrichment can, in principle, be applied to questions on all levels of the central dogma. However, they rely on coupling of the trait of interest to the expression of fluorescent or DNA-modifying reporter, or to phenotypes that can be selected for by growth or binding, which is difficult to achieve in many cases. Furthermore, in these approaches it is not always possible to unambiguously attribute observed changes to the underlying mechanistic causes. As an example, mRNA mutations may affect function through transcription, translation, transcript/polypeptide stability, or specific activity of the protein, which cannot be disentangled assaying only reporter protein activity. This can be compensated for to a certain extent through smart, systematic library design [20] and deep statistical analyses enabled through sufficiently large and diverse datasets [54]. By contrast, other methods allow to interrogate cellular processes more specifically, which, however restricts their application scope to, for instance, processes affecting RNA levels (RNA-Seq) or translation (ribosome/polysome profiling).

Conclusion and outlook

In this review, we described and compared methods for HTP sequence-function mapping and grouped them according to the strategy with which function is converted into an NGS-readable output. This criterion was selected since it allows to highlight and distinguish the available methods in a practical, user-oriented fashion, but arguably other criteria could be applied. Importantly, the differences in technical aspects and application scope described herein are an opportunity for mutual complementation between methods. Not coincidentally, several studies have combined different approaches in order to reduce biases and shed more light into the mechanistic reasons for observed changes in function [20,42,43,73,74]. For example, Gorochowski et al. combined RNA-Seq and ribosome profiling to quantitatively determine the effect of different genetic elements on transcription and translation in *E. coli* [73]. In another noteworthy recent study, Hwa and coworkers combined RNA-Seq with genome-scale transcriptomics and proteomics to quantitatively assess bacterial gene regulation on both transcriptional and translational level to a previously unattainable degree [74]. Irrespective of the method, efficient exploitation of acquired sequence-function data is key to improve our ability to predict and design biological function from sequence, which was only briefly touched upon here. In principle, this can be either achieved by hypothesis-driven, mechanistic modelling or in a data-driven fashion relying on ML. In particular data-driven approaches have recently gained substantial momentum in biology owed to their ability to precisely model non-linear dependencies in complex datasets even in the absence of any prior mechanistic knowledge [5]. However, ML models are intrinsically difficult to interpret, which currently hinders gaining of a deeper mechanistic understanding upon using strictly data-driven approaches. Therefore, means to interpret successfully developed ML models are urgently required [75]. Nonetheless, in view of the current rapid development of NGS and HTP sequence-function mapping technology one can expect transformative new possibilities for the smart design of biosystems with new-to-nature properties in the upcoming years. This will be a major pillar for the urgently required transformation to a circular economy and a powerful toolbox to fight the socioeconomic challenges of our time.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Brooks SM, Alper HS: **Applications, challenges, and needs for employing synthetic biology beyond the lab**. *Nat Commun* 2021, **12**:1390, https://doi.org/10.1038/s41467-021-21740-0.

2. Zürcher JF, Robertson WE, Kappes T, Petris G, Elliott TS, Salmond GPC, *et al.*: **Refactored genetic codes enable bidirectional genetic isolation**. *Science* 2022, **378**:516–523, https://doi.org/10.1126/science.add8943.

3. Zürcher JF, Kleefeldt AA, Funke LFH, Birnbaum J, Fredens J, Grazioli S, *et al.*: **Continuous synthesis of E. coli genome sections and Mb-scale human DNA assembly**. *Nature* 2023, https://doi.org/10.1038/s41586-023-06268-1.

4. Jeschek M, Gerngross D, Panke S: **Combinatorial pathway optimization for streamlined metabolic engineering**. *Curr Opin Biotechnol* 2017, **47**:142–151, https://doi.org/10.1016/j.copbio.2017.06.014.

5. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ: **Next-generation machine learning for biological networks**. *Cell* 2018, **173**:1581–1592, https://doi.org/10.1016/j.cell.2018.05.015.

6. Freschlin CR, Fahlberg SA, Romero PA: **Machine learning to navigate fitness landscapes for protein engineering**. *Curr Opin Biotechnol* 2022, **75**, 102713, https://doi.org/10.1016/j.copbio.2022.102713.

7. Radivojević T, Costello Z, Workman K, Garcia Martin H: **A machine learning Automated Recommendation Tool for synthetic biology**. *Nat Commun* 2020, **11**:4879, https://doi.org/10.1038/s41467-020-18008-4.

8. Vanella R, Kovacevic G, Doffini V, Fernández de Santaella J, Nash MA: **High-throughput screening, next generation sequencing and machine learning: advanced methods in enzyme engineering**. *Chem Commun* 2022, **58**:2455–2467, https://doi.org/10.1039/d1cc04635g.

9. Dietrich JA, McKee AE, Keasling JD: **High-throughput metabolic engineering: advances in small-molecule screening and selection**. *Annu Rev Biochem* 2010, **79**:563–590, https://doi.org/10.1146/annurev-biochem-062608-095938.

10. Vornholt T, Christoffel F, Pellizzoni MM, Panke S, Ward TR, Jeschek M: **Systematic engineering of artificial metalloenzymes for new-to-nature reactions**. *Sci Adv* 2021, **7**, https://doi.org/10.1126/sciadv.abe4208.

11. • Foox J, Tighe SW, Nicolet CM, Zook JM, Byrska-Bishop M, Clarke WE, *et al.*: **Performance assessment of DNA sequencing platforms in the ABRF next-generation sequencing study**. *Nat Biotechnol* 2021, **39**:1129–1140, https://doi.org/10.1038/s41587-021-01049-5.

Study assessing reproducibility, accuracy and utility of massively parallel DNA sequencing technologies.

12. van den Berge K, Hembach KM, Soneson C, Tiberi S, Clement L, Love MI, *et al.*: **RNA sequencing data: Hitchhiker's guide to expression analysis**. *Annu Rev Biomed Data Sci* 2019, **2**:139–173, https://doi.org/10.1146/annurev-biodatasci-072018-021255.

13. Shuken SR: **An introduction to mass spectrometry-based proteomics**. *J Proteome Res* 2023, https://doi.org/10.1021/acs.jproteome.2c00838.

14. Peterman N, Levine E: **Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations**. *BMC Genomics* 2016, **17**:206, https://doi.org/10.1186/s12864-016-2533-5.

15. Kinney JB, Murugan A, Callan CG, Cox EC: **Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence**. *Proc Natl Acad Sci U S A* 2010, **107**:9158–9163, https://doi.org/10.1073/pnas.1004290107.

16. Batrakou DG, Müller CA, Wilson RHC, Nieduszynski CA: **DNA copy-number measurement of genome replication dynamics by high-throughput sequencing: the sort-seq, sync-seq and MFA-seq family**. *Nat Protoc* 2020, **15**:1255–1284, https://doi.org/10.1038/s41596-019-0287-7.

17. Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, *et al.*: **Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters**. *Nat Biotechnol* 2012, **30**:521–530, https://doi.org/10.1038/nbt.2205.

18. •• Boer CG de, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A: **Deciphering eukaryotic gene-regulatory logic with 100 million random promoters**. *Nat Biotechnol* 2020, **38**:56–65, https://doi.org/10.1038/s41587-019-0315-8.

Sort-Seq approach for the assessment of 100 million randomly generated yeast promoters. Sequence-function data is used to model promoter activity by ML.

19. Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP, *et al.*: **Composability of regulatory sequences controlling transcription and translation in Escherichia coli**. *Proc Natl Acad Sci U S A* 2013, **110**:14024–14029, https://doi.org/10.1073/pnas.1301301110.

20. •• Cambray G, Guimaraes JC, Arkin AP: **Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli**. *Nat Biotechnol* 2018, **36**: 1005–1015, https://doi.org/10.1038/nbt.4238.

Smart factorial sequence design is combined with different HTP sequence-function mapping approaches. This allowed to disentangle the influence of various sequence determinants of translation in *E. coli* through statistical analyses.

21. Kim NM, Sinnott RW, Rothschild LN, Sandoval NR: **Elucidation of sequence-function relationships for an improved biobutanol in vivo biosensor in E. coli**. *Front Bioeng Biotechnol* 2022, **10**, 821152, https://doi.org/10.3389/fbioe.2022.821152.

22. Rohlhill J, Sandoval NR, Papoutsakis ET: **Sort-Seq approach to engineering a formaldehyde-inducible promoter for dynamically regulated Escherichia coli growth on methanol**. *ACS Synth Biol* 2017, **6**:1584–1595, https://doi.org/10.1021/acssynbio.7b00114.

23. Noderer WL, Flockhart RJ, Bhaduri A, Diaz de Arce AJ, Zhang J, Khavari PA, *et al.*: **Quantitative analysis of mammalian translation initiation sites by FACS-seq**. *Mol Syst Biol* 2014, **10**:748, https://doi.org/10.15252/msb.20145136.

24. Goodman DB, Church GM, Kosuri S: **Causes and effects of N-terminal codon bias in bacterial genes**. *Science* 2013, **342**: 475–479, https://doi.org/10.1126/science.1241934.

25. Komarova ES, Dontsova OA, Pyshnyi DV, Kabilov MR, Sergiev PV: **Flow-Seq method: features and application in bacterial translation studies**. *Acta Naturae* 2022, **14**:20–37, https://doi.org/10.32607/actanaturae.11820.

26. Schmitz A, Zhang F: **Massively parallel gene expression variation measurement of a synonymous codon library**. *BMC Genomics* 2021, **22**:149, https://doi.org/10.1186/s12864-021-07462-z.

27. Peterman N, Lavi-Itzkovitz A, Levine E: **Large-scale mapping of sequence-function relations in small regulatory RNAs reveals plasticity and modularity**. *Nucleic Acids Res* 2014, **42**: 12177–12188, https://doi.org/10.1093/nar/gku863.

28. Ortega AD, Takhaveev V, Vedelaar SR, Long Y, Mestre-Farràs N, Incarnato D, *et al.*: **A synthetic RNA-based biosensor for fructose-1,6-bisphosphate that reports glycolytic flux**. *Cell Chem Biol* 2021, **28**:1554–1568.e8, https://doi.org/10.1016/j.chembiol.2021.04.006.

29. Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, *et al.*: **Global analysis of protein folding using massively parallel design, synthesis, and testing**. *Science* 2017, **357**:168–175, https://doi.org/10.1126/science.aan0693.

30. Koberstein JN, Stewart ML, Mighell TL, Smith CB, Cohen MS: **A sort-seq approach to the development of single fluorescent protein biosensors**. *ACS Chem Biol* 2021, **16**:1709–1720, https://doi.org/10.1021/acschembio.1c00423.

31. Adams RM, Mora T, Walczak AM, Kinney JB: **Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves**. *Elife* 2016, **5**, https://doi.org/10.7554/eLife.23156.

32. Stark R, Grzelak M, Hadfield J: **RNA sequencing: the teenage years**. *Nat Rev Genet* 2019, **20**:631–656, https://doi.org/10.1038/s41576-019-0150-2.

33. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**:57–63, https://doi.org/10.1038/nrg2484.

34. Aldridge S, Teichmann SA: **Single cell transcriptomics comes of age**. *Nat Commun* 2020, **11**:4307, https://doi.org/10.1038/s41467-020-18158-5.

35. Price A, Gibas C: **The quantitative impact of read mapping to non-native reference genomes in comparative RNA-Seq studies**. *PLoS One* 2017, **12**, e0180904, https://doi.org/10.1371/journal.pone.0180904.

36. Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J: **High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis**. *Nat Biotechnol* 2009, **27**: 1173–1175, https://doi.org/10.1038/nbt.1589.

37. Ma P, Amemiya HM, He LL, Gandhi SJ, Nicol R, Bhattacharyya RP, *et al.*: **Bacterial droplet-based single-cell RNA-seq reveals antibiotic-associated heterogeneous cellular states**. *Cell* 2023, **186**:877–891.e14, https://doi.org/10.1016/j.cell.2023.01.002.

38. Ohuchi S, Mascher T, Suess B: **Promoter RNA sequencing (PRSeq) for the massive and quantitative promoter analysis in vitro**. *Sci Rep* 2019, **9**:3118, https://doi.org/10.1038/s41598-019-39892-x.

39. • Hossain A, Lopez E, Halper SM, Cetnar DP, Reis AC, Strickland D, *et al.*: **Automated design of thousands of nonrepetitive parts for engineering stable genetic systems**. *Nat Biotechnol* 2020, **38**:1466–1475, https://doi.org/10.1038/s41587-020-0584-2.

RNA-Seq is used to characterize thousands of synthetic bacterial and yeast promoters designed to be non-repetitive sequences to avoid recombination.

40. Vo Ngoc L, Huang CY, Cassidy CJ, Medrano C, Kadonaga JT: **Identification of the human DPR core promoter element using machine learning**. *Nature* 2020, **585**:459–463, https://doi.org/10.1038/s41586-020-2689-7.

41. Jores T, Tonnies J, Wrightsman T, Buckler ES, Cuperus JT, Fields S, *et al.*: **Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters**. *Nat Plants* 2021, **7**:842–855, https://doi.org/10.1038/s41477-021-00932-y.

42. Blumberg A, Zhao Y, Huang Y-F, Dukler N, Rice EJ, Chivu AG, *et al.*: **Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data**. *BMC Biol* 2021, **19**:30, https://doi.org/10.1186/s12915-021-00949-x.

43. Xiang JS, Kaplan M, Dykstra P, Hinks M, McKeague M, Smolke CD: **Massively parallel RNA device engineering in mammalian cells with RNA-Seq**. *Nat Commun* 2019, **10**:4327, https://doi.org/10.1038/s41467-019-12334-y.

44. Sheth RU, Wang HH: **DNA-based memory devices for recording cellular events**. *Nat Rev Genet* 2018, **19**:718–732, https://doi.org/10.1038/s41576-018-0052-8.

45. Roquet N, Soleimany AP, Ferris AC, Aaronson S, Lu TK: **Synthetic recombinase-based state machines in living cells**. *Science* 2016, **353**:aad8559, https://doi.org/10.1126/science.aad8559.

46. Tang W, Liu DR: **Rewritable multi-event analog recording in bacterial and mammalian cells**. *Science* 2018, **360**, https://doi.org/10.1126/science.aap8992.

47. Schmidt F, Zimmermann J, Tanna T, Farouni R, Conway T, Macpherson AJ, *et al.*: **Noninvasive assessment of gut function using transcriptional recording sentinel cells**. *Science* 2022, **376**, eabm6038, https://doi.org/10.1126/science.abm6038.

48. Raad M de, Modavi C, Sukovich DJ, Anderson JC: **Observing biosynthetic activity utilizing next generation sequencing and the DNA linked enzyme coupled assay**. *ACS Chem Biol* 2017, **12**:191–199, https://doi.org/10.1021/acschembio.6b00652.

49. • Yus E, Yang J-S, Sogues A, Serrano L: **A reporter system coupled with high-throughput sequencing unveils key bacterial transcription and translation determinants**. *Nat Commun* 2017, **8**:368, https://doi.org/10.1038/s41467-017-00239-7.

A DNA recorder based on a Dam methylase is used to characterize approximately 250,000 promoters and 5′-UTRs in *Mycoplasma pneumoniae*.

50. Weber M, Burgos R, Yus E, Yang J-S, Lluch-Senar M, Serrano L: **Impact of C-terminal amino acid composition on protein expression in bacteria**. *Mol Syst Biol* 2020, **16**, e9208, https://doi.org/10.15252/msb.20199208.

51. Xu Z, Thomas L, Davies B, Chalmers R, Smith M, Brown W: **Accuracy and efficiency define Bxb1 integrase as the best of fifteen candidate serine recombinases for the integration of DNA into the human genome**. *BMC Biotechnol* 2013, **13**:87, https://doi.org/10.1186/1472-6750-13-87.

52. Jusiak B, Jagtap K, Gaidukov L, Duportet X, Bandara K, Chu J, *et al.*: **Comparison of integrases identifies Bxb1-GA mutant as the most efficient site-specific integrase system in mammalian cells**. *ACS Synth Biol* 2019, **8**:16–24, https://doi.org/10.1021/acssynbio.8b00089.

53. • Höllerer S, Papaxanthos L, Gumpinger AC, Fischer K, Beisel C, Borgwardt K, *et al.*: **Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping**. *Nat Commun* 2020, **11**:3551, https://doi.org/10.1038/s41467-020-17222-4.

A recombinase-based DNA recorder was built and used to kinetically characterize more than 300,000 RBSs in *E. coli*. The resulting sequence-function data was used to model RBS behavior at high accuracy in a deep learning approach.

54. •• Höllerer S, Jeschek M: **Ultradeep characterisation of translational sequence determinants refutes rare-codon hypothesis and unveils quadruplet base pairing of initiator tRNA and transcript**. *Nucleic Acids Res* 2023, **51**:2377–2396, https://doi.org/10.1093/nar/gkad040.

A recombinase-based DNA recorder was employed to measure translation from over 1.2 million 5′-UTR-CDS combinations in *E. coli*, which allowed to disprove a standing hypothesis on codon bias and revealed quadruplet base-pairing between mRNA and initiator tRNA.

55. •• Cuperus JT, Groves B, Kuchina A, Rosenberg AB, Jojic N, Fields S, *et al.*: **Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences**. *Genome Res* 2017, **27**:2015–2024, https://doi.org/10.1101/gr.224964.117.

Competitive growth selection was combined with NGS to test approximately 500,000 random 5′-UTRs in yeast. Deep learning was used to extract critical sequence features for translation and to predict the behavior of 5′-UTRs *in silico*.

56. • Tucker AT, Leonard SP, DuBois CD, Knauf GA, Cunningham AL, Wilke CO, *et al.*: **Discovery of next-generation antimicrobials through bacterial self-screening of surface-displayed peptide libraries**. *Cell* 2018, **172**:618–628.e13, https://doi.org/10.1016/j.cell.2017.12.009.

Competitive depletion of surface-displayed peptides in *E. coli* is tracked via NGS and used to discover candidate peptides with antimicrobial activity.

57. Koch P, Schmitt S, Cardner M, Beerenwinkel N, Panke S, Held M: **Discovery of antimicrobials by massively parallelized growth assays (Mex)**. *Sci Rep* 2022, **12**:4097, https://doi.org/10.1038/s41598-022-07755-7.

58. •• Shen MW, Zhao KT, Liu DR: **Reconstruction of evolving gene variants and fitness from short sequencing reads**. *Nat Chem Biol* 2021, **17**:1188–1198, https://doi.org/10.1038/s41589-021-00876-6.

ML is used to reconstruct full genotypes from short-read NGS data obtained in long-term evolution experiments rendering sequence-function mapping amenable also for longer sequences in these approaches.

59. Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen IJ, Morris DR, *et al.*: **Human 5' UTR design and variant effect prediction from a massively parallel translation assay**. *Nat Biotechnol* 2019, **37**:803–809, https://doi.org/10.1038/s41587-019-0164-5.

60. Nguyen Quang N, Bouvier C, Henriques A, Lelandais B, Ducongé F: **Time-lapse imaging of molecular evolution by high-throughput sequencing**. *Nucleic Acids Res* 2018, **46**: 7480–7494, https://doi.org/10.1093/nar/gky583.

61. Komarova N, Barkova D, Kuznetsov A: **Implementation of high-throughput sequencing (HTS) in aptamer selection technology**. *Int J Mol Sci* 2020, **21**, https://doi.org/10.3390/ijms21228774.

62. Nguyen Quang N, Perret G, Ducongé F: **Applications of high-throughput sequencing for in vitro selection and characterization of aptamers**. *Pharmaceuticals* 2016, **9**, https://doi.org/10.3390/ph9040076.

63. Marlière P, Patrouix J, Döring V, Herdewijn P, Tricot S, Cruveiller S, *et al.*: **Chemical evolution of a bacterium's genome**. *Angew Chem Int Ed Engl* 2011, **50**:7109–7114, https://doi.org/10.1002/anie.201100535.

64. Hebisch E, Knebel J, Landsberg J, Frey E, Leisner M: **High variation of fluorescence protein maturation times in closely related Escherichia coli strains**. *PLoS One* 2013, **8**, e75991, https://doi.org/10.1371/journal.pone.0075991.

65. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, *et al.*: **Synthetic spike-in standards for RNA-seq experiments**. *Genome Res* 2011, **21**:1543–1551, https://doi.org/10.1101/gr.121095.111.

66. Pawluczyk M, Weiss J, Links MG, Egaña Aranguren M, Wilkinson MD, Egea-Cortines M: **Quantitative evaluation of bias in PCR amplification and next-generation sequencing derived from metabarcoding samples**. *Anal Bioanal Chem* 2015, **407**: 1841–1848, https://doi.org/10.1007/s00216-014-8435-y.

67. van Dijk EL, Jaszczyszyn Y, Thermes C: **Library preparation methods for next-generation sequencing: tone down the bias**. *Exp Cell Res* 2014, **322**:12–20, https://doi.org/10.1016/j.yexcr.2014.01.008.

68. Johnson MS, Venkataram S, Kryazhimskiy S: **Best practices in designing, sequencing, and identifying random DNA barcodes**. *J Mol Evol* 2023, **91**:263–280, https://doi.org/10.1007/s00239-022-10083-z.

69. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, *et al.*: **Next-generation sequencing technology: current trends and advancements**. *Biology* 2023, **12**, https://doi.org/10.3390/biology12070997.

70. Løbner-Olesen A, Skovgaard O, Marinus MG: **Dam methylation: coordinating cellular processes**. *Curr Opin Microbiol* 2005, **8**: 154–160, https://doi.org/10.1016/j.mib.2005.02.009.

71. Hussmann JA, Patchett S, Johnson A, Sawyer S, Press WH: **Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast**. *PLoS Genet* 2015, **11**, e1005732, https://doi.org/10.1371/journal.pgen.1005732.

72. Abbas-Aghababazadeh F, Li Q, Fridley BL: **Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing**. *PLoS One* 2018, **13**, e0206312, https://doi.org/10.1371/journal.pone.0206312.

73. • Gorochowski TE, Chelysheva I, Eriksen M, Nair P, Pedersen S, Ignatova Z: **Absolute quantification of translational regulation and burden using combined sequencing approaches**. *Mol*

Syst Biol 2019, **15**, e8719, https://doi.org/10.15252/msb.20188719.
RNA-Seq and ribosome profiling are combined to quantify the effect of different regulatory elements on transcription and translation in *E. coli*.

74. Balakrishnan R, Mori M, Segota I, Zhang Z, Aebersold R,
•• Ludwig C, *et al.*: **Principles of gene regulation quantitatively connect DNA to RNA and proteins in bacteria**. *Science* 2022, **378**, eabk2066, https://doi.org/10.1126/science.abk2066.
Combined genome-scale approach employing RNA-Seq with different transcriptomics and proteomics methods to quantify the contribution of transcription and translation to gene expression in bacteria.

75. Sidak D, Schwarzerová J, Weckwerth W, Waldherr S: **Interpretable machine learning methods for predictions in systems biology from omics data**. *Front Mol Biosci* 2022, **9**, 926623, https://doi.org/10.3389/fmolb.2022.926623.