Sequencing enabling design and learning in synthetic biology

Pierre-Aurelien Gilliot[1] and Thomas E. Gorochowski[1,2,*]

[1] School of Biological Sciences, University of Bristol, Life Sciences Building Tyndall Avenue, Bristol BS8 1TQ, UK

[2] BrisSynBio, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK

* Correspondence should be addressed to T.E.G. (thomas.gorochowski@bristol.ac.uk)

Keywords: sequencing; omics; synthetic biology; systems biology; machine learning; biological design.

Abstract

The ability to read and quantify nucleic acids such as DNA and RNA using sequencing technologies has revolutionized our understanding of life. With the emergence of synthetic biology, these tools are now being put to work in new ways – enabling *de novo* biological design. Here, we show how sequencing is supporting the creation of a new wave of biological parts and systems, as well as providing the vast data sets needed for the machine learning of design rules for predictive bioengineering. However, we believe this is only the tip of the iceberg and end by providing an outlook on recent advances that will likely broaden the role of sequencing in synthetic biology and its deployment in real-world environments.

Highlights

- Sequencing can capture detailed information about diverse biological processes.
- Synthetic biology is beginning to exploit sequencing to aid design.
- Large sequencing datasets are powering new machine learning approaches in biology.
- Emerging trends will see the application of sequencing in synthetic biology grow.

Introduction

Sequencing technologies allow us to read DNA [1], RNA [2,3], and even some protein molecules [4], and monitor their changing abundance and composition over time. Such information provides a window into the inner workings of cells, helping us to better understand both their underlying genetic code and the ways it is interpreted to produce and regulate biological processes.

Synthetic biology aims to develop the means to modify existing biological systems or create new ones from scratch with our own desired behaviors [5]. This has resulted in engineered microbes that can sense and report the presence of cancer [6], efficiently self-regulate heterologous protein production to optimize yields [7] and allowed for the creation of programmable living materials that dynamically respond to their environment [8]. While these applications are impressive, they are the exception and not the norm as development of even simple bioengineered systems remains a challenge. Our limited ability to predict how biological parts or systems will function in new contexts [9,10] and the physical limitations and trade-offs when tapping into the limited resources of a host cell [11–13] means that generally numerous designs have to be built before a partially working version is found. Then, time-consuming “tinkering” is often required to optimize the performance further.

Sequencing offers solutions to some of these issues and emerging capabilities could open up more effective approaches for bioengineering [3,4,14,15]. In this review, we focus on recent developments in sequencing and how they are being used to support design processes in synthetic biology and the computational learning of biological design rules [10,16]. We summarize the wealth of sequencing methods available today, the diversity of information that can be captured (**Table 1**) and explain how this insight can guide the design and optimization of new biological parts and systems. This perspective is different to how sequencing is commonly viewed – as a technology for purely reading biological substrates – and instead emphasizes the ability for this reading capacity to support the writing process in synthetic biology and bioengineering. Finally, we consider how new and emerging sequencing technologies might provide steps towards a more holistic spatiotemporal picture of engineered biological systems and support the more systematic design of synthetic biology across scales.

In search of new biological parts

In order to build novel living systems, synthetic biologists require parts that encapsulate simple functionalities. Examples include regulatory elements like promoters and terminators that control where transcription starts and stops, enzymes able to perform specific chemical conversions, and structural proteins that maintain the physical features of a cell. Existing organisms have a wealth of these parts encoded within their genomes and large DNA

sequence databases like GenBank are a treasure trove of parts for synthetic biologists to choose from. The term “part mining” has even become commonplace when speaking about searching these repositories for sequences with a desired functionality. This revolution has been made possible by the rapidly falling costs of DNA sequencing (DNA-seq) over the past few decades. This has resulted in DNA-seq becoming the go to method for part discovery, allowing for genetic information to be extracted from virtually any environment and organism, including those not even culturable in the lab (**Figure 1a**) [17].

While DNA-seq is able to uncover sequences that encode biological parts, it does not capture any information about how they might perform. For parts controlling gene expression, such information is vital because precise levels of expression are often required for a device or system to function correctly. Computational models have been developed to try and bridge this gap [18,19], but their reliability is questionable when used outside of model organisms like *Escherichia coli*. For some key parts, such as transcriptional promoters and terminators, RNA sequencing (RNA-seq) can be used to measure part performance directly, providing a snapshot of RNA abundance at a point in time [2]. Furthermore, RNA-seq is able to characterize all promoters and terminators present in a cell simultaneously, if the transcripts produced are unique [15,20]. More recently, a system for DNA Regulatory Element Analysis by Cell-Free Transcription and Sequencing (DRAFTS) was developed to enable rapid high-throughput measurements of regulatory sequences controlling transcription (**Figure 1b**) [21]. This method brings together cell-free expression systems with multiplexed RNA-seq to allow for *in vitro* characterization of regulatory parts in a wide range of different organisms. This approach has been shown to display a good correlation with *in vivo* part performance and will be able to expand not only the number of parts available to bioengineers, but also provide crucial information about which non-model organisms they can be effectively used within.

Design and optimization of genetic and molecular parts

Biological parts taken directly from sequence databases rarely function exactly as desired. Optimization is therefore necessary to refine their behavior for specific applications. One of the most common biological parts used in synthetic biology are transcription factors (TFs), which allow for gene expression to be controlled in response to their abundance. TFs are proteins that generally bind DNA sequences near a promoter and either sterically block or recruit RNA polymerase (RNAP) to repress or activate downstream genes, respectively. While DNA-binding proteins can be inferred from acid sequence, the precise DNA sequence that is bound is more difficult to predict. Chromatin immunoprecipitation (IP) assays followed by sequencing (ChIP-seq) allow for the identification of the binding site of DNA-associated proteins and has been applied to asses genome-wide binding motifs and regulatory networks

in cells (**Figure 2a**) [22]. ChIP-seq has also been used with cognate site identifier (CSI) arrays [23] to test the binding affinity of a TF to a library of possible DNA binding sequences *in vitro*. This allows for the DNA-binding motif to be accurately inferred and has been shown to have close correspondence to *in vivo* results. Such an approach has been used to exploit genome database mined repressors (TetR homologs) and to rapidly create synthetic regulated promoters based on CSI inferred binding motifs [24].

RNA-based parts and devices have become popular in recent years due to their portability across organisms and the potential for their *de novo* design using software able to predict RNA secondary structure [25,26]. Unfortunately, models struggle to capture the changes often necessary during the functioning of RNA-based devices making computational design a challenge. To support this effort experimentally, SHAPE-seq [27,28] applies selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) chemistry with multiplexed RNA sequencing to allow for the reactivity of each nucleotide to be calculated (**Figure 2b**). Low reactivities correspond to constrained nucleotides that are likely in secondary or tertiary structures. By performing SHAPE-seq for different conformations of a riboswitch or ribocomputing device, the experimentally measured structures for each state can be used to guide specific modifications for improved performance and enable a better understanding of structure-function relationships [29]. Recent developments have also extended this approach to capture nascent RNAs during transcription, allowing for cotranscriptional dynamics of a fluoride riboswitch to be observed at nucleotide resolution [30]. In addition to SHAPE-seq, standard RNA-seq has also been used to aid the design of riboswitches in mammalian cells where mRNA levels relate to switch activity [31].

Rather than optimizing a design through a sequential process of carefully chosen modifications, an alternative approach is to use high-throughput assays to build genotype-phenotype (GP) maps and extract design principles that can be readily exploited. Such approaches are supported by recent advances in massively parallel reporter assays (MPRAs) [32]. These combine chip synthesized oligos with combinatorial DNA assembly to generate a diversity of genetic constructs whose phenotype is linked to the expression of a fluorescent reporter protein. Due to the size of these libraries (often containing >10,000 designs), each member could not be assayed in isolation. However, by pooling the designs and using fluorescence activated cell sorting (FACS) followed by DNA sequencing of the separate fluorescent sorted bins, a GP map for the entire library can be generated (**Figure 2c**). This methodology termed Flow-seq [32] has allowed MPRAs to unravel design principles for numerous biological systems, including: the inference of promoter regulatory logic in eukaryotes [32,33], assessing the composability of regulatory elements and the potential contextual effects that can arise [10], and dissecting the influence of RNA secondary structure, codon usage and other factors on translational efficiency in bacteria [16].

An alternative to screening increasingly complex combinatorial libraries of designs using MPRAs is to directly predict function from sequence by learning key relationships in the growing corpus of datasets. Statistical models ranging from linear regression [34] to deep learning architectures [35] have been used to realize this mapping. Compared to traditional machine learning methods, where there is often a clear relationship between features of a data (*e.g.* linear regression), deep learning exploits vast datasets to derive informative representations in an unsupervised way [36]. While yielding excellent performance in many cases, these models are complex and difficult to interpret. This is due to the high dimensional and non-linear relationships they use to generate accurate predictions, which do not generally map to simple features or relationships in the underlying data. Opening up these “black box” models to better understand what they have learnt is still in its infancy and it is crucial that practitioners appreciate the fragility of their conclusions when looking at saliency maps [37] or applying other interpretive frameworks [38]. This is not to say that deep learning cannot be used to help infer simpler, mechanistic, links between genotype and phenotype that can be used for predictive design, but that care should be taken to thoroughly verify any interpretations. Due to the complexity and sensitivity of these models, it is possible for deep neural networks to learn features of an experiment, not the underlying biology. While it is clear that deep learning will play a crucial role in providing valuable predictions to guide biological design, we suggest caution in blinding following the learnt representations that make these predictions possible, without careful consideration as to whether they might have a biologically feasible underpinning.

Characterizing and debugging of large circuits and systems

The implementation of complex functions in living cells often requires the assembly of many biological parts and devices to create larger circuits and systems [9]. This is made possible by connecting parts whose inputs and outputs are based on a common signal [5,9,39,40]. For example, transcriptional devices often use RNAP flux as a signal and promoters to guide RNAP flux to specific points within a circuit [15,24]. These interconnections have to be carefully designed to ensure that regulatory signals have sufficient dynamic range to propagate correctly throughout an entire system. However, most of these internal signals and states are not directly observable, meaning that part failures are impossible to diagnose from the output alone.

To tackle this issue, RNA-seq has been applied to large genetic logic circuits to observe the steady-state concentration of mRNAs [20] and use these to infer the underlying RNAP flux present [15] (**Figure 3a**). This allows for transcriptional signals to be traced through the circuit and for each part/device to be characterized in the context of the larger system.

Using this information, the root cause of failures can be quickly identified, and targeted modifications made to accelerate the construction of a working system [15]. More recently, this approach has been supplemented with ribosome profiling (Ribo-seq) [41] that allows for translational signals (*i.e.* ribosome flux) to be monitored concurrently (**Figure 3a**) [14,16]. Furthermore, quantification of these signals in absolute units has become possible through the introduction of external RNA standards (*e.g.* ERCC RNA spike-ins [42]) and measurements of key cellular properties (*e.g.* cell mass and growth rate) enabling the read counts from the sequencer to be converted into absolute RNAP/s and ribosome/s units for RNA-seq and Ribo-seq data, respectively [14]. Moving to more quantitative and calibrated measurements of biological parts and systems is of growing importance in helping to improve both the accuracy and reusability of data [14,42,43].

The 3D spatial organization of genetic constructs within a cell can also play a major role in gene expression and regulation [44,45]. Although exploiting such mechanisms in synthetic biology is currently rare [46], genome synthesis efforts such as the Sc2.0 project [47] have started to explore the effect of major structural rearrangements of genomes by using the Synthetic Chromosome Rearrangement and Modification by LoxPsym-mediated Evolution (SCRaMbLE) system [48]. SCRaMbLE causes massive changes in genome content and organization that can alter gene expression and enable the optimization of heterologous biosynthesis pathways by modification of both pathway and host genome simultaneously [49,50]. The Hi-C methodology [51,52] can be used to better understand how the physical organization of chromosomes might underlie these beneficial changes (**Figure 3b**). Hi-C works by crosslinking genomic loci that are close in space, performing a digestion and then ligation of the chromosomal DNA to connect these distant loci into a DNA fragment, and then reverse crosslinking them before deep sequencing [51–53]. From this data, a contact map can then be created to guide 3D models of the chromosomes and estimate their physical arrangement within the cell. Recent advances in Hi-C have used long-read sequencing to enable contacts between more than two loci to be measured simultaneously (IMN Ulahannan et al. bioRxiv doi: 10.1101/833590).

When applied to *Saccharomyces cerevisiae* cells containing original and SCRaMbLE'ed Sc2.0 chromosomes, Hi-C was able to show that the redesigned chromosomes in most cases maintained similar structures to the native versions and that removal of highly repetitive regions in the synthetic variants led to better resolved contact maps [54]. However, two major changes were observed: the loss of an interaction due to a gene deletion, and the relocation of an array of ribosomal RNA repeats to vicinity of the centromere cluster causing genome-wide conformational changes due to physical constraints within the nucleus [54]. As more comprehensive modifications are made to living cells, the

importance of understanding their impact on the spatial organization of core cellular components and machinery will grow.

Looking to the future

Several new trends in sequencing have begun to emerge that further broaden the ways that sequencing can support biological design. From a technology perspective, the Oxford Nanopore Technologies (ONT) sequencing platform offers some interesting capabilities such as full-length reads of DNA and RNA molecules [3,55], as well as the ability to capture base modifications [56,57]. This opens up new avenues to support the development of regulatory mechanisms based on DNA/RNA modifying enzymes [58] and the application of long-read DNA and RNA sequencing to monitor genome replication dynamics [59]. It also is well positioned to help dissect the influence of long-range interactions between genetic parts in a circuit. Another interesting development is the potential to use the same ONT hardware to perform protein sequencing and identification, through the use of disordered protein tags. These are designed to naturally pass through a nanopore due to their charge and generate a unique disturbance in the measured electrical current (IMN Zhang et al. bioRxiv doi: 10.1101/837542) [4].

Capturing the variability between cells is a major limitation of many current sequencing approaches. Generally, nucleic acids are extracted from large populations of cells and thus data for population averages are measured. While single cell sequencing is becoming more accessible and economical [60], the precision of the measurements made and inherent noise in these techniques at present makes their data challenging to use for engineering tasks. Longer term, as synthetic biology moves beyond single cells to multi-cellular consortia, tissues, and even to synthetic ecosystems, spatial sequencing approaches will also become valuable [61]. These will allow for cell/species abundance and transcriptional states to be linked to spatial positions/niches within complex structured environments – information that will be vital for the effective deployment of synthetic microbial consortia into real-world environments [62].

Conclusions

Sequencing has historically been seen as a tool to read genetic information and observe its regulation in natural contexts. However, advances in the breadth of data that can be collected and the ability for this to help inform biological design decisions, highlights it potential to underpin new bioengineering workflows. Established engineering fields rely on advanced monitoring and debugging tools to enable the rational construction and verification of complex systems before their deployment. As the field of synthetic biology matures, we expect the development of similar tools tailored to the unique features of living systems. Sequencing is

well placed to take on this challenge, especially when combined with complementary analysis methods such as LC-MS [63]. Given these growing capabilities and falling costs, the application of sequencing to synthetic biology is likely to grow rapidly over the next decade, becoming a crucial tool to verify that engineered living systems work precisely as we expect before being used in real-world settings.

Acknowledgements

This work was supported by BrisSynBio, a BBSRC/EPSRC Synthetic Biology Research Centre grant BB/L01386X/1 (T.E.G.), the EPSRC/BBSRC Centre for Doctoral Training in Synthetic Biology grant EP/L016494/1 (P.-A.G.), and a Royal Society University Research Fellowship grant UF160357 (T.E.G.)

Author Contributions

All authors helped to write the manuscript.

Declaration of Interest

None.

References

- of special interest
- •• of outstanding interest

1. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH: **DNA sequencing at 40: past, present and future**. *Nature* 2017, **550**:345–353.

2. Stark R, Grzelak M, Hadfield J: **RNA sequencing: the teenage years**. *Nat Rev Genet* 2019, **20**:631–656.

3. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al.: **Highly parallel direct RNA sequencing on an array of nanopores**. *Nat Methods* 2018, **15**:201.

4. Nivala J, Marks DB, Akeson M: **Unfoldase-mediated protein translocation through an α-hemolysin nanopore**. *Nat Biotechnol* 2013, **31**:247–250.

5. Greco FV, Tarnowski MJ, Gorochowski TE: **Living computers powered by biochemistry**. *The Biochemist* 2019, **41**:14–18.

6. Danino T, Prindle A, Kwong GA, Skalak M, Li H, Allen K, Hasty J, Bhatia SN: **Programmable probiotics for detection of cancer in urine**. *Sci Transl Med* 2015, **7**:289ra84-289ra84.

7. Ceroni F, Boo A, Furini S, Gorochowski TE, Borkowski O, Ladak YN, Awan AR, Gilbert C, Stan G-B, Ellis T: **Burden-driven feedback control of gene expression**. *Nat Methods* 2018, **15**:387–393.

8. González LM, Mukhitov N, Voigt CA: **Resilient living materials built by printing bacterial spores**. *Nat Chem Biol* 2020, **16**:126–133.

9. Brophy JAN, Voigt CA: **Principles of genetic circuit design**. *Nat Methods* 2014, **11**:508.

10. Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP, Endy D, Church GM: **Composability of regulatory sequences controlling transcription and translation in *Escherichia coli***. *Proc Natl Acad Sci* 2013, **110**:14024.
    •• One of the first studies demonstrating the power of using chip-based oligo synthesis and Flow-seq to unravel the contextual effects between promoters and ribosome binding sites. Although most combinations lead to predictable expression levels of a

reporter, a minority deviate significantly. The authors suggest that screening large libraries of circuits may be more productive than standardization.

11. Gorochowski TE, Avcilar-Kucukgoze I, Bovenberg RAL, Roubos JA, Ignatova Z: **A Minimal Model of Ribosome Allocation Dynamics Captures Trade-offs in Expression between Endogenous and Synthetic Genes**. *ACS Synth Biol* 2016, **5**:710–720.

12. Gyorgy A, Jiménez JI, Yazbek J, Huang H-H, Chung H, Weiss R, Del Vecchio D: **Isocost Lines Describe the Cellular Economy of Genetic Circuits**. *Biophys J* 2015, **109**:639–646.

13. Weiße AY, Oyarzún DA, Danos V, Swain PS: **Mechanistic links between cellular trade-offs, gene expression, and growth**. *Proc Natl Acad Sci* 2015, **112**:E1038.

14. 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.
    •• The authors use an adapted form of Ribo-seq that employs synthetic RNA spike-in standards to characterize the function of transcriptional and translational parts in absolute units. They further demonstrate how this information can be used to more precisely measure the burden synthetic genetic constructs place on a host cell.

15. Gorochowski TE, Espah Borujeni A, Park Y, Nielsen AA, Zhang J, Der BS, Gordon DB, Voigt CA: **Genetic circuit characterization and debugging using RNA-seq**. *Mol Syst Biol* 2017, **13**:952.
    •• The first application of RNA-seq to a large synthetic genetic circuit. It is shown how RNA-seq data can be processed to generate transcription profiles and use these to infer the performance of basic regulatory parts, response dynamics of genetic logic gates, and host response to the burden imposed by the circuit.

16. 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.
    •• This paper uses MPRAs to measure the relative contribution of several factors potentially involved in translation efficiency in bacteria. The sheer breadth of biological features measured is a testament to current sequencing capabilities and experimental techniques available to bioengineers.

17. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N: **Shotgun metagenomics, from sampling to analysis**. *Nat Biotechnol* 2017, **35**:833–844.

18. Bharanikumar R, Premkumar KAR, Palaniappan A: **PromoterPredict: sequence-based modelling of Escherichia coli σ(70) promoter strength yields logarithmic dependence between promoter strength and sequence**. *PeerJ* 2018, **6**:e5862–e5862.

19. Salis HM, Mirsky EA, Voigt CA: **Automated design of synthetic ribosome binding sites to control protein expression**. *Nat Biotechnol* 2009, **27**:946–950.

20. Liu Q, Schumacher J, Wan X, Lou C, Wang B: **Orthogonality and Burdens of Heterologous AND Gate Gene Circuits in E. coli**. *ACS Synth Biol* 2018, **7**:553–564.

21. Yim SS, Johns NI, Park J, Gomes AL, McBee RM, Richardson M, Ronda C, Chen SP, Garenne D, Noireaux V, et al.: **Multiplex transcriptional characterizations across diverse bacterial species using cell-free systems**. *Mol Syst Biol* 2019, **15**:e8875.
    •• A rapid in vitro approach called DRAFTS is introduced allowing for cell-free extracts of diverse bacteria coupled and RNA-seq to measure the performance of large libraries of regulatory elements *en masse*. This method opens up ways to quickly develop part libraries for non-model organisms and generate valuable characterization data.

22. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al.: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing**. *Nat Methods* 2007, **4**:651–657.

23. Warren CL, Kratochvil NCS, Hauschild KE, Foister S, Brezinski ML, Dervan PB, Phillips GN, Ansari AZ: **Defining the sequence-recognition profile of DNA-binding molecules**. *Proc Natl Acad Sci U S A* 2006, **103**:867.

24. Stanton BC, Nielsen AAK, Tamsir A, Clancy K, Peterson T, Voigt CA: **Genomic mining of prokaryotic repressors for orthogonal logic gates**. *Nat Chem Biol* 2014, **10**:99–105.
    •• The authors make use of randomly synthesized DNA and ChIP-seq to generate the binding motifs of 16 TetR repressor homologs from genome databases. Once binding motifs are found, these are used to build synthetic promoters that are regulated by the repressors and these are finally used to create a set of orthogonal NOT and NOR logic gates.

25. Seetin MG, Mathews DH: **RNA Structure Prediction: An Overview of Methods**. In *Bacterial Regulatory RNA: Methods and Protocols*. Edited by Keiler KC. Humana Press; 2012:99–122.

26. Green AA, Kim J, Ma D, Silver PA, Collins JJ, Yin P: **Complex cellular logic computation using ribocomputing devices**. *Nature* 2017, **548**:117.

27. Watters KE, Abbott TR, Lucks JB: **Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq**. *Nucleic Acids Res* 2015, **44**:e12–e12.
    • Adaptation of the SHAPE-seq protocol to work *in vivo*. Important verification of the technique to ensure that the chemical treatments do not cause cellular responses that would affect the accuracy of the RNA structures obtained.

28. Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, Schroth GP, Pachter L, Doudna JA, Arkin AP: **Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)**. *Proc Natl Acad Sci* 2011, **108**:11063.

29. Takahashi MK, Watters KE, Gasper PM, Abbott TR, Carlson PD, Chen AA, Lucks JB: **Using in-cell SHAPE-Seq and simulations to probe structure–function design principles of RNA transcriptional regulators**. *RNA* 2016, **22**:920–933.

30. Watters KE, Strobel EJ, Yu AM, Lis JT, Lucks JB: **Cotranscriptional folding of a riboswitch at nucleotide resolution**. *Nat Struct Mol Biol* 2016, **23**:1124–1131.
    •• The SHAPE-seq protocol is extended to allow for cotranscriptional folding to be observed adding a temporal component to the RNA structures generated. This applied to a fluoride riboswitch and the precise point where the structural dynamics split depredating upon the presence of fluoride is pinpointed.

31. 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.

32. Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E: **Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters**. *Nat Biotechnol* 2012, **30**:521–530.

33. de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A: **Deciphering eukaryotic gene-regulatory logic with 100 million random promoters**. *Nat Biotechnol* 2019, doi:10.1038/s41587-019-0315-8.
    - The largest MPRA preformed to date in which >100 million random synthetic yeast promoters are tested. Machine learning is used to understand the features influencing expression levels based on known transcription factor binding sites. Weak regulatory interactions, often not considered when deigning sequences, are found to play a key role.

34. Cuperus JT, Groves B, Kuchina A, Rosenberg AB, Jojic N, Fields S, Seelig G: **Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences**. *Genome Res* 2017, **27**:2015–2024.

35. Movva R, Greenside P, Marinov GK, Nair S, Shrikumar A, Kundaje A: **Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays**. *PLOS ONE* 2019, **14**:e0218073.

36. LeCun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**:436–444.

37. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A: **A primer on deep learning in genomics**. *Nat Genet* 2019, **51**:12–18.

38. Ghorbani A, Abid A, Zou J: **Interpretation of Neural Networks Is Fragile**. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*. 2019:3681–3688.
    - This paper shows how slight modifications to an input image can substantially change a trained neural networks interpretation based on saliency maps. Coining the term “adversarial perturbations”, the authors warn against blind faith in trusting that the interpretations of deep neural networks fully capture what has been learnt.

39. Weiss R, Homsy GE, Knight TF: **Toward *in vivo* Digital Circuits**. In *Evolution as Computation*. Edited by Landweber LF, Winfree E. Springer Berlin Heidelberg; 2002:275–295.

40. Canton B, Labno A, Endy D: **Refinement and standardization of synthetic biological parts and devices**. *Nat Biotechnol* 2008, **26**:787.

41. Ingolia NT: **Ribosome profiling: new views of translation, from single codons to genome scale**. *Nat Rev Genet* 2014, **15**:205.

42. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B: **Synthetic spike-in standards for RNA-seq experiments**. *Genome Res* 2011, **21**:1543–1551.

43. Beal J, Haddock-Angelli T, Baldwin G, Gershater M, Dwijayanti A, Storch M, de Mora K, Lizarazo M, Rettberg R, with the iGEM Interlab Study Contributors: **Quantification of bacterial fluorescence using independent calibrants**. *PLOS ONE* 2018, **13**:e0199432.

44. Cabal GG, Genovesio A, Rodriguez-Navarro S, Zimmer C, Gadal O, Lesne A, Buc H, Feuerbach-Fournier F, Olivo-Marin J-C, Hurt EC, et al.: **SAGA interacting factors confine sub-diffusion of transcribed genes to the nuclear envelope**. *Nature* 2006, **441**:770–773.

45. Brickner DG, Cajigas I, Fondufe-Mittendorf Y, Ahmed S, Lee P-C, Widom J, Brickner JH: **H2A.Z-Mediated Localization of Genes at the Nuclear Periphery Confers Epigenetic Memory of Previous Transcriptional State**. *PLOS Biol* 2007, **5**:e81.

46. Stoof R, Wood A, Goñi-Moreno Á: **A Model for the Spatiotemporal Design of Gene Regulatory Circuits**. *ACS Synth Biol* 2019, **8**:2007–2016.

47. Richardson SM, Mitchell LA, Stracquadanio G, Yang K, Dymond JS, DiCarlo JE, Lee D, Huang CLV, Chandrasegaran S, Cai Y, et al.: **Design of a synthetic yeast genome**. *Science* 2017, **355**:1040.

48. Dymond J, Boeke J: **The *Saccharomyces cerevisiae* SCRaMbLE system and genome minimization**. *Bioeng Bugs* 2012, **3**:168–171.

49. Blount BA, Gowers G-OF, Ho JCH, Ledesma-Amaro R, Jovicevic D, McKiernan RM, Xie ZX, Li BZ, Yuan YJ, Ellis T: **Rapid host strain improvement by *in vivo* rearrangement of a synthetic yeast chromosome**. *Nat Commun* 2018, **9**:1932.

50. Liu W, Luo Z, Wang Y, Pham NT, Tuck L, Pérez-Pi I, Liu L, Shen Y, French C, Auer M, et al.: **Rapid pathway prototyping and engineering using *in vitro* and *in vivo* synthetic genome SCRaMbLE-in methods**. *Nat Commun* 2018, **9**:1936.

51. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al.: **Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome**. *Science* 2009, **326**:289.

52. Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J: **Hi–C: A comprehensive technique to capture the conformation of genomes**. *3D Chromatin Archit* 2012, **58**:268–276.

53. Kempfer R, Pombo A: **Methods for mapping 3D chromosome architecture**. *Nat Rev Genet* 2019, doi:10.1038/s41576-019-0195-2.

54. Mercy G, Mozziconacci J, Scolari VF, Yang K, Zhao G, Thierry A, Luo Y, Mitchell LA, Shen M, Shen Y, et al.: **3D organization of synthetic and scrambled chromosomes**. *Science* 2017, **355**:eaaf4597.
    •• Application of Hi-C to native, synthetic and SCRaMbLE'd chromosomes. This is the first study to investigate the effect of large-scale chromosomal rearrangements, finding that overall chromosome structure is fairly robust.

55. Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, Parkes D, Freeman C, Dhalla F, Patel SY, et al.: **Sequencing of human genomes with nanopore technology**. *Nat Commun* 2019, **10**:1869.

56. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W: **Detecting DNA cytosine methylation using nanopore sequencing**. *Nat Methods* 2017, **14**:407–410.

57. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, Paten B: **Mapping DNA methylation with high-throughput nanopore sequencing**. *Nat Methods* 2017, **14**:411–413.

58. McDonald JI, Celik H, Rois LE, Fishberger G, Fowler T, Rees R, Kramer A, Martens A, Edwards JR, Challen GA: **Reprogrammable CRISPR/Cas9-based system for inducing site-specific DNA methylation**. *Biol Open* 2016, **5**:866.

59. Müller CA, Boemo MA, Spingardi P, Kessler BM, Kriaucionis S, Simpson JT, Nieduszynski CA: **Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads**. *Nat Methods* 2019, **16**:429–436.

60. Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, Lönnerberg P, Linnarsson S: **Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq**. *Genome Res* 2011, **21**:1160–1167.

61. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, et al.: **Three-dimensional intact-tissue sequencing of single-cell transcriptional states**. *Science* 2018, **361**:eaat5691.

62. Riglar DT, Giessen TW, Baym M, Kerns SJ, Niederhuber MJ, Bronson RT, Kotula JW, Gerber GK, Way JC, Silver PA: **Engineered bacteria can function in the mammalian gut long-term as live diagnostics of inflammation**. *Nat Biotechnol* 2017, **35**:653–658.

63. Gowers G-OF, Chee SM, Bell D, Suckling L, Kern M, Tew D, McClymont DW, Ellis T: **Improved betulinic acid biosynthesis using synthetic yeast chromosome recombination and semi-automated rapid LC-MS screening**. *Nat Commun* 2020, **11**:868.

64. Fernandez-Rodriguez J, Yang L, Gorochowski TE, Gordon DB, Voigt CA: **Memory and Combinatorial Logic Based on DNA Inversions: Dynamics and Evolutionary Stability**. *ACS Synth Biol* 2015, **4**:1361–1372.

65. Carstens S, Nilges M, Habeck M: **Inferential Structure Determination of Chromosomes from Single-Cell Hi-C Data**. *PLOS Comput Biol* 2016, **12**:e1005292.

Tables

Table 1: Sequencing methods and the biological design applications they can support.

Method	Type of Data	Applications	Refs.
DNA-seq	DNA sequence	• Verification of synthetic DNA constructs • Genome sequencing • Genetic part mining	[1,55]
	DNA structural variation	• Monitoring DNA rearrangements (*e.g.* SCRaMbLE) • Characterization of recombinases and integrases • Monitoring of mobile genetic elements • Measurement of recombination rates during evolution	[49,63,64]
	DNA base modifications	• Engineering synthetic epigenetics • Novel DNA memory storage mechanisms	[56,57]
	Species identification	• Metagenomics • Microbiome engineering • Design of stable multi species consortia	[17]
RNA-seq	Transcript concentrations	• Monitoring steady-state mRNA concentrations • Calculating differential gene expression • Analysis of stress responses • Riboswitch design	[2,3,21,31]
	RNA polymerase flux	• Calculating initiation rate of promoters • Identifying transcription termination sites • Measuring termination efficiency of terminators • Calculating RNAP processivity	[15,20]
	Transcript isoforms	• Engineering controlled gene splicing • Discovery of protein variants • Refactoring gene sequences (intron removal)	[3]
	RNA base modifications	• Developing new mechanisms for gene regulation • Control of translation (tRNA modifications)	[3]
Ribo-seq	Ribosome occupancy	• Relative codon occupancy times • Identification of ribosome pausing sites • Optimization of protein translation efficiency	[14,41]
	Ribosome flux	• Measurement of stop codon termination efficiency • Measurement of translation initiation rates	[14]
ChIP-seq	Protein-DNA binding sites	• Identification of transcription start sites • Discovery of transcription factor binding motifs	[22–24]
SHAPE-seq	RNA secondary structure	• Design of RNA sensors and regulators • Monitoring of co-transcriptional RNA folding • Optimization of gene expression	[27–30]
Flow-seq	Gene expression (fluorescence)	• Construction of genotype-phenotype maps • Training machine learning models	[10,16,32–35]
Hi-C	Chromosome structure	• Control of chromosome structure • Engineering long-range genetic interactions	[44,45,51–54,65]

Figures and captions

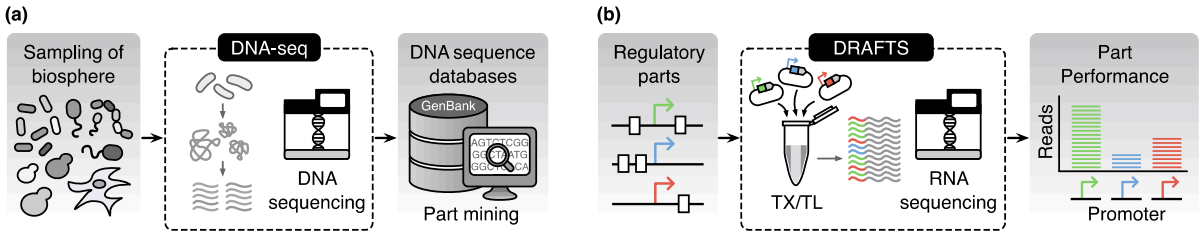

Figure 1: Using sequencing to search for and characterize genetic parts. (a) DNA-seq has been extensively used to sample genomic DNA from across the biosphere [1]. Much of this data is available in centralized databases like GenBank offering the chance to search/mine for sequences encoding potentially useful genetic parts. **(b)** The DRAFTS method allows for the performance of many transcriptional regulatory parts (*e.g.* promoters) to be rapidly measured using cell-free transcription and translation (TX/TL) followed by RNA sequencing [21]. Each part is first uniquely barcoded (denoted by colored regions in the figure) and inserted into a standardized expression plasmid. These are then pooled and expressed in a single pot reaction using cell-free TX/TL. RNA is extracted and sequenced, and then reads are demultiplexed and assigned to an individual construct using the unique barcodes. Read counts directly correspond to the relative performance of the part. By creating cell-free TX/TL expression systems using different organisms the functionality of genetic parts across cellular contexts can be rapidly tested.

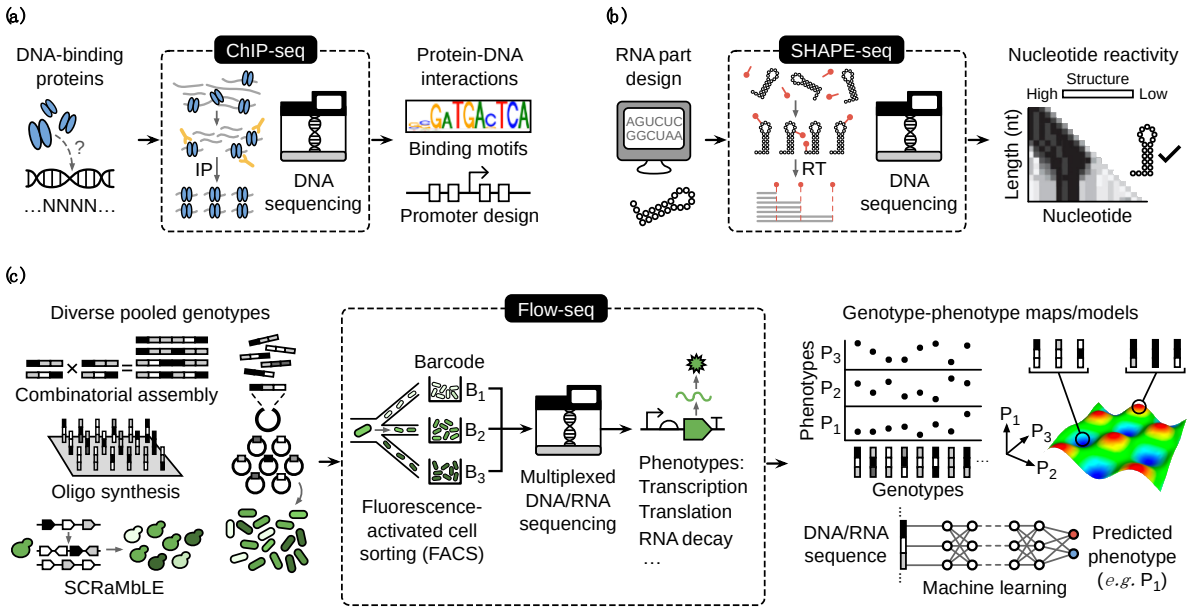

Figure 2: Sequencing methods to aid the design and optimization of genetic and molecular parts. (a) ChIP-seq can be used to study protein-DNA interactions allowing for the binding motifs of DNA-binding proteins to be measured [22]. These operator sequences can then be used to construct synthetic promoters [24]. **(b)** SHAPE-seq can probe the structural state of RNAs at a nucleotide resolution [27,28]. It relies on the use of SHAPE chemistry that selectively modifies unstructured nucleotides of RNA (red markers). These modified nucleotides cause reverse transcription (RT) to halt, providing the location of the unstructured nucleotide within the RNA. The data generated from sequencing the DNA produced can then be used to infer the structure for the entire RNA. Nucleotide reactivity graph is adapted from data in [30] and is not related to the structure shown. **(c)** Flow-seq allows for a genotype-phenotype (GP) map to be generated for diverse pool of genetic constructs where phenotype is indicated by the expression of a fluorescent reporter [32]. A diverse set of genotypes can be generated in numerous ways including chip synthesis of oligos, combinatorial DNA assembly, or by employing systems like SCRaMbLE. Measuring large numbers of cellular phenotypes and their associated genotypes is made possible by fluorescence active cell sorting (FACS), barcoding of cells in each bin with different fluorescence (B_1–B_3), and then multiplexed pooled DNA or RNA sequencing of these. Depending on how fluorescence is linked to a particular phenotype of each genetic design, this approach can measure many different aspects of a phenotype related to transcriptional, translational and post-translational processes (P_1–P_3). Each genotype links to a particular point in phenotypic space and it is possible for multiple genotypes to exhibit very similar phenotypes and therefore become clustered at particular points in this space. The schematic for machine learning shows how a genotype (on the left) can feed into a deep neural network and connecting to a particular phenotype (on the right). The network will generally connect a genotype to a single phenotype

and multiple networks can be used to predict different aspects of a phenotype (*e.g.* P_1–P_3). Initially, genotypes and known phenotypes (*e.g.* from an MPRA study) are provided and a learning algorithm used to update the parameters (weights) of the network. Once sufficient samples have been viewed, the network is able to predict the phenotype of an unseen genotype.

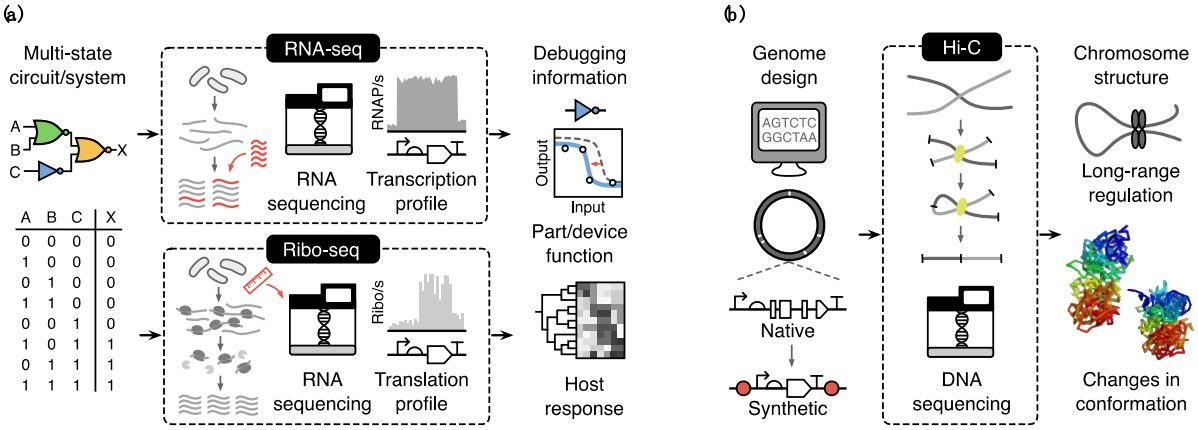

Figure 3: Applications of sequencing to large circuits and cellular systems. (a) Large multi-state genetic circuits can be debugged through a combination of RNA-seq and Ribo-Seq [14,15] to provide detailed information about internal circuit states, the *in situ* function of parts and devices, and the response of the host cell to the additional burden. RNA-seq can be supplemented with synthetic RNA spike-ins and Ribo-seq with measurements of growth rate and cellular mass (red features in dashed boxed) to convert read counts into absolute units (*i.e.* RNAP/s and Ribosomes/s for transcriptional and translational signals, respectively) [14]. **(b)** Hi-C allows for chromosome structure to be measured [51,52]. This is valuable as synthetic biology moves towards genome design [47] and begins to exploit long-range interactions to regulate synthetic genetic circuits. Chromosome confirmations are adapted from [65].