Molecular Cell

CellPress

Voices

# Voices on technology: The molecular biologists' ever-expanding toy box

With the focus on technology for this issue of *Molecular Cell*, a group of scientists working in different areas of molecular biology provide their perspective on the most recent important technological advance in their field, where the field is lacking, and their wish list for future technology development.

Rachel Patton McCord
Assistant Professor of Biochemistry & Cellular and Molecular Biology, University of Tennessee, Knoxville

## Choose the right tool for the question

Like kids at Christmas, we as scientists are often eager to hear about the exciting new toys that we can apply in our research. Like parents shopping on Black Friday, we can get swept away in the hype of new technologies. But, if we aren't careful, we can also end up like children who no longer want to play with those toys by January 15th. In the scientific world, this can look like a wave of publications all jumping on the bandwagon to use a new technique, followed by the realization that the biological insights gained effectively just recapitulate what we already knew from more established methods. How can we make the most of new technologies while avoiding these pitfalls? From my observations of progress in my own field, I have seen two factors that propel technologies from ‘‘cool in theory’’ to actual major impact: crosstalk and integration across different technologies and a deep understanding of biological questions.

In my field, where we seek to understand the principles and functions of 3D chromosome structure, it is not a single technology, but instead the integration of molecular genomic data with single-molecule imaging techniques that has been revolutionary in the past several years. The 3C/Hi-C (chromosome conformation capture) family of techniques combined proximity ligation with the power of high-throughput sequencing to provide a completely new view of the 3D structure of chromosomes over the past decade. Among other observations, this technology revealed that 3D genome contact domains (called TADs) existed at the gene regulatory scale and were bounded by the proteins CTCF and cohesin. But, even though this sequencing-based technology was amazing, it reached a limit of what it could do by itself. The new technology of auxin-inducible degrons was critical to enable the field to test the role of candidate architecture proteins in the formation of these TADs. The genomic data, combined with computational simulations, gave rise to the idea that the cohesin complex could help form TADs by extruding loops of DNA, but critical tests of this idea needed the power of single-molecule experiments and high-speed atomic force microscopy. Suddenly, we could actually watch a DNA loop being formed by condensin or cohesin in real time! Understanding what TADs mean in living cells has also required new single-molecule imaging technologies that allow us to trace and track the locations of chromosomes and gene expression in single cells. None of these technologies by themselves could have led in such a short time to the understanding we now have of chromosome folding mechanisms. If a single-molecule imaging lab had observed that cohesin extrudes DNA loops, they wouldn't have known how to interpret this result without the perspective of the Hi-C data. And if the ‘‘genomics people’’ had stayed in a separate bubble away from ‘‘the microscopists’’ and the ‘‘single-molecule people,’’ our perspective on the 3D genome would still be stuck at the stage of potentially interesting correlations but no mechanistic understanding.

It is also key that people who are at the forefront of developing new technologies work closely with those who have a deep understanding of the biological questions. It can be tempting, with an exciting new technology, to want to apply it indiscriminately to any question that arises. But the biggest impacts are rarely made by a tool looking for a problem. For example, the auxin-inducible degron system mentioned above has been absolutely amazing as a tool to rapidly deplete certain proteins in a cell. But, this tool has had the most impact in cases where biologists had deeply studied a question, and therefore knew exactly where they needed this capability to answer their question. Though my field is full of exciting new technologies, I caution my students not to phrase their research interests as ‘‘I want to use CRISPR to test something,’’ but instead to first identify the key question. Sometimes, we also need the humility to admit that the very best technology to answer our biological question may be an ‘‘old and boring’’ one rather than a shiny new one.

**Mikko Taipale**
Associate Professor, Donnelly Center, University of Toronto

## Predict, design, synthesize

I think the most significant advance in the last several years has been the sudden emergence of highly accurate protein structure prediction. It has only been six months since AlphaFold2 and RoseTTAfold were released, but they have already changed the way we think about biological questions and conduct experiments. Everyone can now access remarkably accurate predictions of hundreds of thousands of proteins, and soon we can count these in the hundreds of millions. We also have the first drafts of structurally resolved yeast and human protein interaction networks. In the coming years, these algorithms will help us understand how coding variation impacts protein function and rewires interaction networks, how deadly pathogens exploit host proteins to their own advantage, and how evolution has sculpted diverse proteomes from *E. coli* to elephants. They will also help us design completely new proteins for therapeutic, diagnostic, and industrial purposes.

While designing novel proteins is now much easier, testing them in the lab is still hampered by our limited capacity in DNA synthesis. That’s one reason why I can’t wait for the moment when writing DNA is as cheap as reading DNA. Imagine ordering and receiving 100,000 full-length genes in a week for a hundred dollars. Imagine synthesizing entire chromosomes and genomes for a few thousand dollars. Oh, what we could do with that! We would screen millions of synthetic proteins to develop next-generation therapeutics. We would dive deep into the world of uncharacterized genes in the most obscure organisms. We would recode, reshuffle, recombine, regenerate, and degenerate genomes to understand the forces that have shaped them for 4 billion years. And at the very least, we would marvel at scientists of the bygone era, those who had to culture bacteria for plasmid DNA instead of just ordering it from their local supplier. I don’t know when this revolution will happen, because like nuclear fusion, it’s always just a few years away. But I do know that when it comes, we should all be prepared!

Finally, I have one request for all technology developers out there: Can you please come up with a way to freeze and thaw mammalian cells as easily as yeast and bacterial glycerol stocks?

Sarah Teichmann
Head of Cellular Genetics, Wellcome Sanger Institute
Director of Research, Cavendish Laboratory, University Cambridge

### Location, location, location—generating cell maps

It is now over a decade since the first report using next-generation technology for single-cell mRNA sequencing, and the pace of technological advances in the intervening years has been breath-taking. As a result, high-throughput scRNA-seq analyses of tissues and organs in development, physiology, and disease are now commonplace. In tandem, techniques to measure epigenomic and proteomic characteristics of single cells have blossomed, continuing the resolution revolution of single-cell genomics to give a much richer characterization of cell phenotypes.

But while these methods provide incredibly useful cellular parts lists, their reliance on tissue dissociation loses the vital spatial information we need to make representative cell maps. Thankfully, we have recently benefitted from another ground-breaking advance: spatial transcriptomics, which encompasses several methods in which spatial information is retained in whole-transcriptome studies. When used in tandem with single-cell multi-omic studies, spatial transcriptomics allows us to pinpoint cell types in tissues and provides an integrated understanding of tissue anatomy. Spatial methods are becoming ever more widely available to the community, as well as being more robust and scalable. They are crucial to the various projects of the Human Cell Atlas, the international consortium aiming to map each cell type of the human body.

A big challenge now is how to effectively integrate data from spatial and single-cell modalities. One recent advance is Cell2Location, a computational tool using a probabilistic framework that can map the spatial location of up to hundreds of reference cell types. It's an exciting tool as it helps define not only new cell types but also tissue zones where modules of cell types work together.

For the future, we will need a better integration of imaging and next-generation sequencing techniques. We need to be able to capture characteristics such as cell morphology and link this to the molecular details. Again, one of the core problems here is computational, managing to find a way to integrate imaging and transcriptomics, modalities that are currently entirely separate. Finally, I would really like to be able to analyze cell types in large volumes of tissue, not just the single planes we have access to in slices. Tissues are complex, three-dimensional structures with many microenvironments, and a way to expand spatial transcriptomics to larger volumes, especially in the Z axis, would be a game changer.

Rebecca Voorhees
Assistant Professor of Biology and Biological Engineering
Investigator, Heritage Medical Research Institute
California Institute of Technology

### A switch in structure: Generating, predicting, interpreting

In essence the ultimate goal of all structural biology is to "watch" a molecule carry out its function at the atomic level. Historically, each technological advance in the field has led to a major leap in our understanding of the chemistry that underpins cellular life.

In the 1950s, the first structures of proteins determined using X-ray crystallography gave us the earliest glimpses into proteins function at the molecular level. However, not all complexes are easily amenable to X-ray crystallography because of the large quantities of extremely pure and homogeneous sample required to generate crystals.

Almost ten years ago, technological advances in single-particle cryoelectron microscopy (cryo-EM) expanded the types of molecules we could visualize at atomic or near-atomic resolution. Without the need to form crystals, the sample quantities and homogeneity required for cryo-EM are far less stringent. As a result, cryo-EM has enabled structure determination of proteins purified directly from patient tissues and has proven particularly powerful for studying membrane proteins.

Further, because cryo-EM relies on imaging of individual particles that are not constrained by a crystal lattice, the conformational heterogeneity and dynamics of a sample are captured in each image. Over the past year, the first computational tools to deconvolute the continuous motion of particles from single-particle cryo-EM datasets paved the way toward directly visualizing molecular movement.

Nevertheless, both X-ray crystallography and single-particle cryo-EM require purifying a protein away from its cellular context. The frontier of structural biology will undoubtedly center around strategies to image biological molecules *in situ*, including cryoelectron tomography (cryo-ET) coupled with correlative fluorescence microscopy. These methods will be enabled by increasingly powerful structural prediction algorithms like AlphaFold and Rosetta. Computationally predicted structures will facilitate interpretation of the somewhat lower resolution EM density maps more typical of cryo-ET.

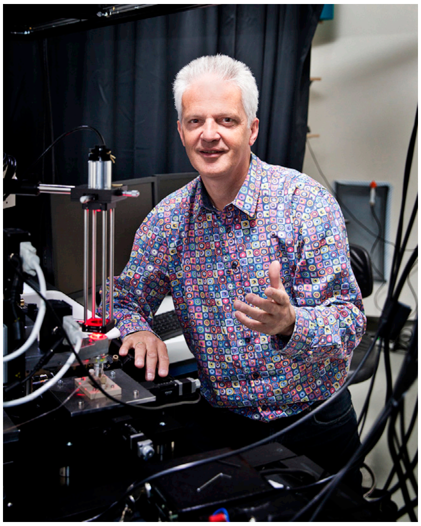
More broadly, however, computational protein prediction is the technology that will most transform the field of structural biology. As software algorithms continue to improve for prediction of multisubunit complexes and nucleic acids, structural biology will center less and less on generating structures and more on their interpretation. Combining computational and experimental methodologies will thus bring us ever closer to directly visualizing the molecular details of a dynamic complex in its native cellular environment.

**Yue Wan**
Associate Director, Genome Institute of Singapore Agency for Science, Technology And Research (A*STAR)

## From studying RNA structures to understanding RNA function

While studying how a region of RNA folds traditionally involves using radioactivity and running of large sequencing gels, we have seen an explosion of high-throughput experimental and computational methods that enable us to decipher RNA structures in a massively parallel way in the past few years. These methods have evolved from the initial sequencing strategies that allow us to know which bases are paired and unpaired to enabling us to determine the RNA interaction partners and even higher-order tertiary interactions in different systems. Recently, developments in third-generation high-throughput sequencing has further enabled us to identify structure information along long RNA sequences, allowing us to identify isoform-specific structures and their impact in isoform gene regulation. Additionally, an RNA can form multiple conformations, making the problem of studying RNA structures even more complex. Advances in single-molecule RNA analysis has enabled clustering of structurally distinct RNA structures from the same sequence, providing insights into the extent of RNA structural heterogeneity and their biology.

All of these technological advances have dramatically changed our ability to understand RNA structure and their functions in the cell. In the coming years, the continued development of new structural probes and the combinatorial mapping of different aspects of RNA structures, combined with measurements of cellular processes such as translation and decay, in the same cells will provide deeper and more direct insights into structure and regulation. Additionally, technology breakthroughs that allow interrogation of structural diversity in single cells and single molecules will broaden our understanding of the ability of RNA structures to determine cell fates. Another exciting direction that the field is taking is the application of AI to learn and predict RNA structures. The development of AlphaFold has changed the way we can obtain protein structures. Similarly, the application of AI to RNA structure is also likely to revolutionize our ability to predict RNA structures accurately from primary sequences, making it easy for anyone to obtain accurate RNA structures. Collectively, the new technologies and the biology that we learn using them will enable us to better understand the diversity of RNA structures and their functions in different systems and accelerate the promise of RNA molecules as drug targets in the coming years.



**Cees Dekker**
Distinguished University Professor, Delft University of Technology

## Nanopores: From DNA sequencing to proteomics

Cells feature a myriad of small pores that transport ions, metabolites, proteins, and nucleic acids across membranes. In the past decades, such nanopores have been at the heart of developing a variety of biotech applications. The most prominent example is nanopore-based single-molecule DNA sequencing, where a helicase slowly traverses a single DNA molecule across a nanopore while an ion current that is flowing through the pore exhibits slight changes as bases move along the pore constriction, yielding sequence information. As this nanopore technology is commercialized into pocket-sized sequencers, it is progressively making inroads into the genomics market, worth more than ten billion dollars.

Yet nanopores can be used in many different ways, as single-molecule biophysics tools, selective filters, sensors for metabolities, nanoreactors for chemistry in confinement, mimics for complex protein pores such as the nuclear pore complex, etc. Let me here specifically single out one very exciting novel direction: nanopores as tool for *protein* identification and sequencing. While DNA sequencing was the major driver of the nanopore field since the 1990s, attention has been redirected from DNA to proteins in the past years. While genomes obviously are a key source of basic information, it has become clear that splicing, transcriptional variants, and post-translational modifications lead to an enormous diversity of proteins, and neither the DNA genotype nor the RNA transcriptome can fully describe the protein phenotype. Hence, mapping the enormous and dynamic variations in the proteome is urgently needed, for which nanopores can be employed. Early experiments showed the fast (∼microsecond) translocation of folded proteins through solid-state nanopores. Proteins could also be unfolded using strong denaturants like urea or SDS to allow linear translocation through the pore. Recently a technique was developed to trap a folded protein against a DNA-origami sphere docked onto a nanopore, and this so-called NEOtrap was shown able to hold and study a single protein for very long times (hours), allowing researchers to study intrinsic conformational dynamics of individual proteins.

An ultimate goal in nanopore protein research is to sequence the amino acids along the peptide backbone of a protein. The challenges associated with this holy grail of single-protein sequencing are humongous, since proteins are folded, amino acid residues are hydrophobic or hydrophilic as well as positively or negatively charged, and 20 different amino acids need to be distinguished. Speed control of moving a peptide only slowly through the nanopore is yet another challenge. While early work employed a ClpX protease motor to unfold and translocate a protein through a nanopore, very recent work provided a breakthrough in obtaining exquisite control by using DNA-peptide hybrids, where a peptide was drawn through a nanopore in single amino acid steps by a helicase walking on a lead DNA strand. This allowed researchers to discriminate even single amino acid substitutions in single reads of a peptide, while moreover the same molecule could be re-read hundreds of times, which drove the read accuracy to basically 100%. While much more needs to be done to develop these first proof-of-principle data on protein identification into a *de novo* nanopore single-molecule protein sequencer, this approach provides an exciting step forward. Combined with other recent breakthroughs such as the AlphaFold AI platform that predicts protein folding, it is clear that new techniques are currently providing an enormous boost to proteomics.

Neville E. Sanjana
Core Faculty Member, New York Genome Center
Assistant Professor, Department of Biology, NYU

## CRISPR: Gene editing and beyond

Although the human genome was first sequenced in 2003, biologists have spent much of the last 2 decades trying to understand the function of those 3 billion As, Cs, Ts, and Gs. There has been a particular emphasis on which regions of the genome contribute to specific diseases. In some cases, like serious monogenic disorders, this has been straightforward, but for other cases, such as polygenic, common diseases, it has been much more challenging to pinpoint where in the genome matters most.

As with the race to sequence the first genome, new technologies have played an oversized role in helping us understand the function of genes and noncoding regions of the genome. In particular, the rapidly expanding genome-engineering toolbox—driven by the development of CRISPR programmable nucleases and related tools—has enabled us to move beyond correlational studies to assign causal roles to genome elements.

With these tools, we can now edit, silence, or activate genes and then measure changes in disease-associated molecular or cellular phenotypes. But this is not limited to just genes! For noncoding regions of the genome identified through large-scale genome-wide association studies (GWASs), we often cannot distinguish which noncoding variants drive disease phenotypes and which ones are mere passengers, typically in linkage with the causal variant. Discovery of these noncoding regulators can have a tremendous impact on human health: one of the first gene editing therapies to enter the clinic (for hemoglobin disorders) targets a noncoding region—a binding site of the transcription factor GATA1—to restore expression of fetal hemoglobin.

These breakthroughs are enabled by the incredible programmability of CRISPR-based tools: A short RNA guides them to a specific genome location. This easy programmability enables massively parallel experiments, where a single scientist can investigate how loss of each of the 20,000 genes in the human genome modulates drug resistance in cancer or infection with pathogens like SARS-CoV-2.

Finally, the expanding CRISPR toolbox is now moving beyond the genome. New CRISPR enzymes like Cas13 are similarly easy-to-program but target RNA instead of DNA. This opens the door to a new world of transcriptome engineering with several applications that are only possible on the transcript level. The last decade has brought a dizzying array of new programmable technologies to engineer the genomes and transcriptomes of human cells. In the next decade, these technologies will enable us to improve human health through a deeper understanding of our own genomic code.