Probing the dark matter of the human genome with big DNA

Jon M. Laurent, Sudarshan Pinglay, Leslie Mitchell and Ran Brosh (New York University Langone Health, USA and Yale University, USA)

Less than 2% of our genome is protein-coding DNA. The vast expanses of non-coding DNA make up the genome's "dark matter", where introns, repetitive and regulatory elements reside. Variation between individuals in non-coding regulatory DNA is emerging as a major factor in the genetics of numerous diseases and traits, yet very little is known about how such variations contribute to disease risk. Studying the genetics of regulatory variation is technically challenging as regulatory elements can affect genes located tens of thousands of base pairs away, and often, multiple distal regulatory variations, each with a very small effect, combine in an unknown way to significantly modulate the expression of genes. At the Center for Synthetic Regulatory Genomics (SyRGe) we directly tackle these problems in order to systematically elucidate the mechanisms of regulatory variation underlying human disease.

The sequencing of the human genome, completed in 2003, brought with it a torrent of new questions about the function of our genetic blueprint. It was as if we were given a book explaining the basis of our existence but hadn't yet learned the meaning of more than a few words. The subsequent years have ushered in a new era of understanding for human biology, as we've begun to learn how to read this blueprint and assign function to its various parts.

One of the earliest observations made by the nascent field of human genomics was the vast disparity in the proportion of genomic sequence devoted to protein-coding genes as compared with the rest, the so-called 'non-coding' DNA. While proteins represent the major functional unit of any cell, the 22,000 or so human proteins are encoded by less than 2% of the approximately 3.3 billion bp of DNA in the human genome. The role of the remaining 98% has puzzled researchers for decades. In the past, it has even been disparaged as simply 'junk DNA', having little-to-no functional consequence. However, similar to the invisible 'dark matter' created after the Big Bang having very measurable effects on the physics of the universe, we prefer to think of the non-coding genome as the 'dark matter of the genome', with its many secrets waiting to be unveiled.

There has been a renaissance of sorts in recent years as researchers have developed ever more advanced technologies for investigating the function of this genomic dark matter. Some functional elements of the non-coding genome are already known, many of which have been identified by similarity to well-characterized elements in better-studied species. These sequences include, among others, elements that regulate the expression of coding genes in space and time (promoters, enhancers) or elements that direct the 3D organization of the genome. The development of systems-level studies has brought about the ability to identify these types of functional elements with greater throughput and sensitivity.

Researchers have been applying high-throughput techniques to connect specific genomic sequences to disease onset and progression. One of the most common type of such efforts, called genome-wide association studies, or GWAS, have led to the identification of thousands of sites in our genomes where person-to-person sequence variations have been associated with various phenotypes or diseases. Variants found in coding sequences can often be easily rationalized mechanistically as disease causing because they often alter protein function. However, more than 90% of variants identified by GWAS studies lie in non-coding regions of the genome, and it is thus unclear what role they play in influencing disease risk. These variants are often located at or near potential regulatory sites and hence, many of them likely exert their effect by altering nearby genome structure or changing expression of nearby genes. Identified sites are also typically clustered together as dozens of potentially causative variants spanning a large continuous region of the genome in the order of hundreds of thousands of base pairs. Thus, pinpointing functional sites and elucidating the mechanism by which they affect nearby genes has proven to be extremely challenging.

We, a diverse group of scientists at the Center for Synthetic Regulatory Genomics (SyRGe) at NYU Langone Health, are tackling this problem using a 'bottom-up' approach, writing entire segments of the human genome from scratch rather than the more painstaking 'top-down' methods that involve stepwise editing of individual variants. In our approach, which we call 'synthetic regulatory genomics', we are leveraging the expertise of the Boeke lab to synthesize large chunks of DNA and subsequently deliver those synthetic genomic regions...
to a relevant cell line where they can be interrogated for function. Importantly, our specific methodology allows straightforward parallel assembly of dozens of variant versions of any given genome region, giving us the ability to precisely and simultaneously test multiple variants and their combinations.

**How we do it**

The synthetic regulatory genomics approach (Figure 1) starts by identifying a region of the human genome that has been linked to a disease, for example through GWAS hits. We call this region a ‘locus’; it is typically about 100,000 bp in length and contains multiple sequence variants that could be important for the disease.

The second step of the workflow is to assemble, in parallel, many versions of the locus from the bottom-up. To do this, we start with small fragments of synthetic DNA, usually ~3,000 bp in length, that tile across the entire length of the locus. Wherever there is a sequence variant we are interested in studying, we produce multiple fragments covering the sequence variation of that region. We then stitch the fragments together into full-length sequences, substituting in the fragments containing sequence variants where appropriate. When complete, one full-length version will match the sequence that is present in healthy individuals and will provide a reference point, or control, for our later experiments in cells. The rest of the assemblies will carry combinations of the disease-relevant sequence variants. The assembly step is carried out with the help of the budding yeast *Saccharomyces cerevisiae*, which is highly proficient at precisely stitching together pieces of DNA to produce larger DNAs.

The third and final step is to transfer the library of assemblies into mammalian cells growing in culture in the lab. Generally speaking, this step is carried out by replacing the resident version of the locus inside the cells with the new versions that we built. The end product is a set of new cell lines, one carrying the control locus, and all others carrying a specific set of variant segments of known sequence. We then perform experiments to measure differences in all of these cell lines, usually related to gene and protein expression. Together, the data we produce allows us to begin to understand, at base pair resolution, the role of sequence variants with respect to disease.

**Future implications of synthetic genomics**

The bottom-up approach is a powerful strategy to infer the rules governing biological systems. To this end, scientists are currently applying this strategy to ‘write’ designer genomes from the bottom up (e.g. the Synthetic Yeast Genome Project, aka Sc2.0 and the Genome Project -Write, aka GP-Write). With increased understanding of biological systems, we can leverage the bottom-up approach to construct biological systems for the benefit of society. Synthetic biologists are currently pursuing such engineering at all levels – individual molecules, cells and whole organisms powered by designer DNA. For example, designer organisms can be endowed with abilities that their natural counterparts do not possess, such as degradation of plastics or the production of useful chemicals. As Sydney Brenner put it: “Progress in science depends on new techniques, new discoveries and new ideas, probably in that order.” The ability to manipulate synthetic DNA specifically on the scale discussed in this article is sure to engender avenues of thought that are currently unimaginable.

Understandably, the rise of such powerful new capabilities to manipulate cells and organisms at the genetic level sparks debate around legal, societal and ethical issues. We foresee at least three major lines of concern:

First, is the application of genome-writing technology to introduce modifications into the human germline. At present we do not have the requisite understanding of biological systems required to make these modifications in a reliable and risk-free manner. Further, the ethical considerations of such modifications are far from being properly discussed by the scientific community or the public. It is our position that for the foreseeable future genome writing in human cells be limited to experiments in culture.

Second, is the concern about the release and possible adverse impact of synthetic organisms on the environment or human health. There are many suitable approaches available to address this problem. For instance, synthetic organisms have been engineered to carry genetic ‘kill switches’, making the genetically modified organisms incapable of living outside the lab. Beyond this, scientists must develop more rigorous testing methodologies to...
study the possible impact of these organisms on the environment. It is unlikely that we will ever be completely sure that synthetic organisms will pose no risk to the environment. However, if the proper risk-mitigation approaches can be identified and undertaken, we submit that the benefits will far outweigh the possible costs.

Third, are issues of ‘dual-use’—that the technologies developed by the synthetic biology community could be used malevolently by nefarious agents. While this concern is legitimate, it suggests only that we must better regulate the technology rather than halt scientific progress.

Other concerns have nothing to do with the real-world impact of synthetic biology. Such arguments have been made throughout history at every great scientific or technological breakthrough. Benjamin Franklin was accused of having usurped the power of God when he put lightning rods on buildings in order to prevent them from being struck by lightning. Might there have been a caveman who said the power of fire was best left alone? Was it only yesterday that newspapers screamed ‘test-tube baby born,’ and now 1–2% of the US population is conceived in vitro? Such concerns are invariably left behind as humanity moves on. It is more important to understand the risks and reap the benefits than to leave such powerful technology unutilized.

We believe the promise of big DNA will help solve issues in human health and society and should inspire researchers and the public alike. The loci we are currently working on, plus those that we plan to include, are associated with macular degeneration, lupus and schizophrenia, among others. Tackling these projects with our synthetic regulatory genomics approach has the potential to help thousands of people maintain a healthy and productive life. We are also investigating loci that will greatly enrich our understanding of broad principles of genetic regulation. We invite you to participate in our effort by nominating a locus on the SyRGE Center website that may fit with this vision. We fundamentally believe that the work ongoing at the SyRGE Center, the other applications discussed in this article, and the many other applications not discussed here and yet to be realized, will overall be beneficial to humanity as a whole.

Further reading

- Mitchell, L.A., McCulloch, L.H., Pinglay, S., et al. bioRxiv 423426; https://doi.org/10.1101/423426
- Laurent, J.M., Fu, X., German, S., Maurano, M.T., Zhang, K. and Boeke, J.D. (2019) Big DNA as a tool to dissect an age-related macular degeneration-associated haplotype. Precision Clinical Medicine 2, 1–7; Synthetic Yeast Genome Project, Sc.2.0 www.syntheticyeast.org
- Boeke lab website: https://med.nyu.edu/research/boeke-lab/
- Center for Synthetic Regulatory Genomics (SyRGe) website: https://www.thedarkmatterproject.org
- National Human Genome Research Institute website: https://www.genome.gov

Sudarshan Pinglay is a PhD student at NYU Langone Health currently working towards engineering biological systems in order to understand their function. He is from Bangalore, India and completed undergraduate degrees in molecular biology and philosophy at Johns Hopkins University. Sudarshan is passionate about soccer, heavy metal guitar, food and exposing everyone to the joys of a scientific worldview. Email: sudarshan.pinglay@nyulangone.org

Ran Brosh is currently a senior research scientist in the Institute for Systems Genetics at NYU Langone Health, where he leads the stem cell effort of the Center for Synthetic Regulatory Genomics (SyRGe). Ran earned his PhD from the Weizmann Institute, where he studied the tumor suppressor p53. For his post-doc training, Ran joined the lab of Dr Ihor Lemischka at Mount Sinai School of Medicine, where he developed a new method to induce protein degradation in mammalian cells. Ran is passionate about science, travel, photography and sports-watching. Email: Ran.Brosh@nyulangone.org

Leslie Mitchell is a postdoctoral fellow in the Institute for Systems Genetics at NYU Langone Health. She has worked extensively on chromosome and genome engineering in both yeast and mammalian systems and helped lead the international Synthetic Yeast Genome Project. Leslie is also a co-founder of the biotech start-up Neochromosome, Inc., focused on pioneering new human cell therapies using Big-DNA technology. Leslie is passionate about traveling, gardening, wine tasting and sports-playing. Email: Leslie.Mitchell@nyulangone.org

Jon M. Laurent is a postdoctoral fellow in the Boeke lab at NYU Langone Health who is interested in the using power of synthetic genomic approaches to dissect genome function. He originally hails from Montana, where he did his undergraduate studies in molecular biology. His PhD work was at the University of Texas at Austin using synthetic biology techniques to probe the evolution of protein function. Email: Jon.Laurent@nyulangone.org