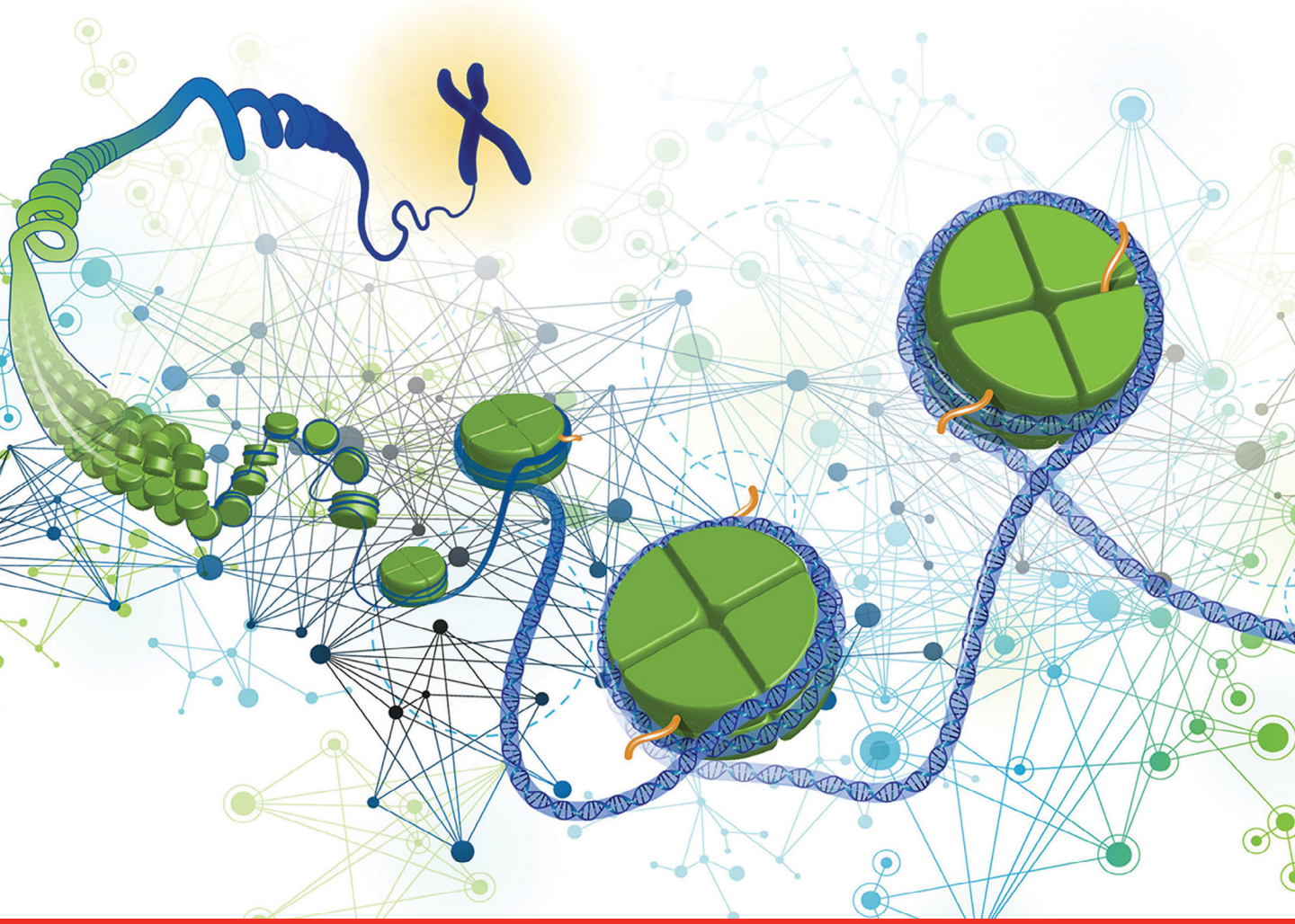


# CIFAR ROUNDTABLES ON CANCER AND THE DARK GENOME

## Summary Report

Johnny Kung, PhD

SEPTEMBER 2022

CIFAR 40 YEARS ANS

# TABLE OF CONTENTS



## CIFAR

CIFAR is a global research organization that convenes extraordinary minds to address the most important questions facing science and humanity.

We are supported by the governments of Canada, Alberta, and Quebec, as well as foundations, individuals, corporations, and Canadian and international partner organizations.

## LAND ACKNOWLEDGEMENT

CIFAR’s office is located on the traditional territory of many nations including the Mississaugas of the Credit, the Anishnabeg, the Chippewa, the Haudenosaunee, and the Wendat peoples. This land is now home to many diverse First Nations, Inuit, and Métis peoples. It is covered by Treaty 13 with the Mississaugas of the Credit.

**COVER IMAGE:** Darryl Leja, National Human Genome Research Institute (NHGRI)

# INTRODUCTION

**In the past two decades, advances in genomic sequencing have led to significant leaps forward in our understanding of how genetic changes contribute to the development of cancer. Major publicly-funded collaborative projects, such as The Cancer Genome Atlas (TCGA), have generated valuable resources for the entire cancer research community. These resources have allowed researchers to develop new insights into how alterations in genetic pathways could lead to cancer.**

Much of these existing resources and research are focused on "coding genes", which encode for the proteins that constitute many of the functional and structural components of our cells. However, more than 98% of our genome is "noncoding", and this "dark matter" of the genome has been shown to have important biological functions. Many regulatory elements reside within the noncoding genome, controlling how genes are turned on or off, and much of the noncoding genome is also transcribed into noncoding RNAs that play important roles in gene regulation. Research such as the Pan-Cancer Analysis of Whole Genomes (PCAWG) project has identified many cancer-associated genetic changes that lie within the noncoding genome. What is currently lacking is a comprehensive repository of all the different types of noncoding elements relevant to cancer, and more importantly, a systematic effort to understand the functional significance of genetic changes in the noncoding genome for cancer development.

On November 3, 2021, CIFAR organized a roundtable of international experts in the biology of the noncoding genome and cancer biology, as well as leaders of "Big Biology" programs. A broad consensus emerged that the field would be considerably advanced by the development of shared resources such as databases and new research tools and techniques. On April 26, 2022, a second roundtable was convened to discuss the potential scientific priorities and governance structure for a project to build these shared resources.

This report provides a brief introduction to the biology of the dark genome and highlights key insights that emerged from these two roundtables, including promising next steps for progress in this area.

This series of roundtables was made possible by the generous support of the MacMillan Family Foundation.

# WHAT IS THE "DARK GENOME", AND WHY IS IT IMPORTANT IN THE CONTEXT OF CANCER?

The misregulation of genes plays key roles in the development of cancer, e.g., when there is decreased activity of genes that prevent the over-proliferation of cells, or increased activity of genes that drive cell proliferation. Genetic changes, or *mutations* — including the addition, removal or substitution of individual DNA letters (or *bases*), or large-scale *structural variants* such as duplication, deletion or rearrangement of whole segments of DNA — can affect gene function by directly altering the DNA sequence of the genes themselves, or by changing whether, when, and how strongly genes are *expressed* (turned on or shut off).

Within the human genome, only less than 2% of its DNA sequence encodes for proteins. The cellular machinery *transcribes*, or copies, this sequence into messenger RNA (mRNA) *transcripts*, the information in which is then used as templates to assemble amino acids into proteins in the process of *translation*. Of the remaining 98% of the genome — the noncoding DNA — a portion is transcribed as part of the same mRNA transcripts as protein-coding genes, but are either *spliced* out (segments known as *introns*) from the mature transcripts, or remain in the mature transcripts but do not encode for the amino acid sequence of the protein product (*untranslated regions* at the two ends of transcripts). Much of the rest of the noncoding genome is also transcribed independently of protein-coding genes, and some of these *noncoding RNAs* (ncRNAs) have been found to play important biological functions in normal development and in disease, including cancer. ncRNAs have long been known to be central components of the cellular machineries for the transcript splicing and the mRNA-to-protein translation processes, but in recent years, researchers have found multiple additional roles for ncRNAs, include recruiting (or interfering with) gene regulatory factors, orchestrating the 3D architecture of the genome, or interfering with translation. Much about these functions, and their mechanisms, remains to be investigated.

At the same time, many parts of the noncoding genome also contain important *regulatory elements*, sites that control gene expression. These include *promoters*, from where transcription is initiated; *enhancers* and *silencers*, which boost or suppress gene expression, respectively; and *insulators* (or *boundary elements*), which block the effects of other elements by, e.g., interacting with other insulators to demarcate distinct domains in the genome. These elements may regulate their target genes by directly recruiting relevant protein factors to assemble, activate or repress the transcription machinery; or, by modifying how loosely or tightly the DNA is packaged (into structures known as "open" or "closed" *chromatin*), which affects the accessibility of that DNA to the transcription machinery. Other than promoters, many of these regulatory elements may be located far away from the genes that they regulate (in terms of the linear DNA sequence; in reality, DNA in our genome forms loops in 3D space that bring the elements and their target genes physically close together). As such, the gene targets of regulatory elements are not always obvious without detailed experiments.

Much progress has been made over the years in understanding how genetic changes within, or near, protein-coding genes contribute to cancer. However, there is still a significant gap in fully understanding the effects of changes to the noncoding parts of the genome, whether in genes that encode for ncRNAs, or in regulatory elements that are distant from their target genes.

# WHAT ARE THE MAJOR UNANSWERED QUESTIONS, AND THE KEY TECHNICAL CHALLENGES TO ANSWERING THEM?

The expert roundtable participants posed a range of unanswered questions about the role of the dark genome in cancer, and identified challenges to answering some of them:

## What are all the aspects of the noncoding genome relevant to cancer?

- There is still a need to map the location and work out the function of many **enhancers, boundary sites** and other regulatory elements. These are not all known in normal cells, let alone how they are mutated or misregulated in cancer.
- There is also a need to map out all the variations in **chromatin state** throughout the genome - which can provide information about whether and how different parts of the genome are expressed, under different conditions (including in cancerous cell states) — and correlate that information with cells' biological characteristics (or *phenotypes*)
- The **3D architecture** of the genome determines how genes and their regulatory elements are brought together to control gene expression. A number of experimental techniques have been developed in the last few years to investigate this architecture, including Hi-C (high-throughput chromosome conformation capture), high-throughput FISH (fluorescent in situ hybridization), live-cell imaging techniques, and *in situ* sequencing technologies. However, many of these techniques still need to be scaled up or optimized.
- Studying **ncRNAs** poses additional challenges. Not only is the inventory of ncRNAs expressed in different cell types not all worked out and often poorly annotated, but ncRNA function also depends on how the RNA molecules are folded into 3D shapes and how certain RNA bases are chemically modified. Moreover, some ncRNAs, while not encoding for full proteins, have been found to contain sections that are translated into short chains of amino acids (or *peptides*) which in some cases may have distinct biological functions. New analysis techniques have been developed to study RNA structure, and a new generation of genetic sequencing techniques (known as "long-read" sequencing) can help in creating a better catalogue of ncRNAs.
    - Importantly, a full accounting of ncRNAs may not be a finite problem — the cellular transcription machinery is imperfect, and the more comprehensively scientists sequence the RNA population in cells, the more they may find RNA variants expressed at very low levels that may or may not be biologically functional.

- Well over half of the human genome consists of ***repetitive elements*** — stretches of DNA, several hundreds or thousands of bases long, that are repeated thousands or even millions of times throughout the genome, a portion of which is transcribed into ncRNAs. Many of these are remnants of viruses and transposons (genetic parasites sometimes called "jumping genes") that entered the genome during evolutionary history, but some have since acquired potentially important biological roles both in normal development and in disease. Because of their repetitive nature, they have been very difficult to study with existing techniques, but newer tools like long-read sequencing could help.

## How do genetic changes alter the function of noncoding elements?

- Associating noncoding genetic variants with their phenotypes remains a key challenge, and effort should be dedicated to **causally attributing biological consequences to specific mutations**, through experiments to perturb the function of noncoding elements throughout the genome in a high-throughput manner. Recent advances in experimental techniques useful in this regard include:
    - Genome editing - e.g., the CRISPR/Cas system, which can introduce targeted genetic changes to the genome, as well as variations of the technique for targeted activation or inhibition of gene expression without making genetic changes (known as CRISPRa and CRISPRi, respectively)
    - Synthetic biology - which allows scientists to create long stretches of DNA with multiple genetic changes, and then introduce that DNA into cells

  However, the phenotypic readouts that should be assayed after functional perturbation need to be thoughtfully considered, i.e., experiments should not simply look at whether or not a genetic change leads to cell death, but also other meaningful biological effects.

## How should we study the role of noncoding elements in cancer?

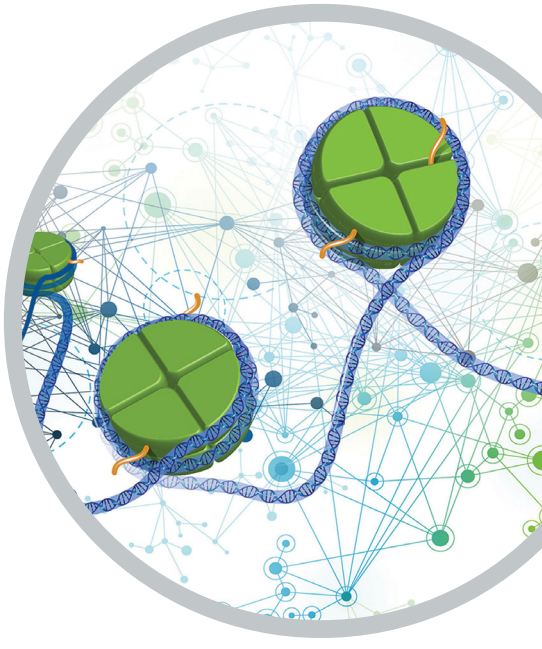
- A proper understanding of the dark genome's role in cancer will require studying it in **appropriate model systems**. Many current studies use "cell lines" — cells that have been cultured to grow relatively more easily in dishes in a lab, and which have been much studied and so have lots of available experimental data. However, cell lines are quite artificial systems that are not necessarily the most biologically representative. A "best case" model that can allow researchers to fully study the context and dynamics of cancer might be a longitudinal study of the entire **developmental hierarchy of cancer** — from precancerous tissues, through primary tumours, to metastases, along with the surrounding tissue microenvironment and immune cells. However, obtaining all these clinical samples would be logistically (and even ethically) difficult, and these more biologically realistic samples are also harder to perturb functionally. New techniques for culturing cells/tissues, such as organoids (3D, miniature versions of "organs" in a dish), which better reflect biological context but are still relatively easy to grow and study, could be useful compromises.
- Because tumours are **highly heterogeneous** structures in which individual cells may have different biological functions or behaviours, experimental tools such as single-cell techniques (which can isolate individual cells for analysis) and spatial genomics/transcriptomics (which can profile the DNA or RNA content of individual cells within their tissue context) will also need to be optimized and used for any such studies.

# WHAT KIND OF SHARED RESOURCES CAN DRIVE PROGRESS IN THE FIELD?

Overall, the roundtable participants agreed that there is a need for a multi-prong effort to map, perturb, and computationally analyze and model the role of cancer-relevant noncoding elements. In order to benefit the broader genomics and cancer biology communities, data and functional models generated by such an effort should be openly shared and easily accessible through an interactive online portal. This portal should have built-in analysis tools, so that all interested researchers can not only analyze the generated data, but could also input their own data to model and make predictions using standardized tools and computational pipelines. Standardized protocols for performing similar mapping and functional studies should also be openly shared with the broader community so that they might generate compatible data to add to the database.

In building this database and portal, computational biologists need to be engaged early on to build the necessary tools, which can help narrow down which noncoding elements to prioritize for study. There should be plans to sunset older techniques while still allowing for ways to compare data generated by competing methodologies, as well as the capacity for backward compatibility that allows for integration and comparison with existing cancer genomics datasets.

A database and portal as envisioned require long-term commitment, with an institutional home and dedicated staff. Future-proofing measures should be put in place, e.g., making sure that donor consents are obtained in such a way that there could be a possibility of re-accessing samples or re-contacting donors, in case of new technologies coming online that could allow for new experiments or analyses. To build sustained interest and support, at least part of the data generation and analysis should be done with an eye to possible future development of clinical prediction tools and therapeutics — where the identification of certain changes in the noncoding genome could be useful for predicting patient outcomes, and where such changes could potentially be targeted to treat cancer. As such, attention should be paid early on to any intellectual property implications of, e.g., the techniques and tools used in the effort.



# WHAT WOULD A PROJECT TO DEVELOP SUCH SHARED RESOURCES LOOK LIKE?

Finally, the roundtable participants discussed the outline of a pilot project to generate the shared resources (database and portal, new experimental and computational tools, standardized protocols) described in the previous section, identifying some of the key elements or decision points for the project.

## Cataloguing noncoding elements, or performing functional analyses?

An important early decision is the balance that such a project should strike between, on one hand, further mapping and cataloguing noncoding elements, and on the other, carrying out functional assays to study the biological effects of genetic changes to these elements. Mapping of elements may be needed for specific cancer subtypes of interest that are not well covered in existing databases such as PCAWG. Noncoding elements like ncRNA and boundary sites will in particular need further mapping and annotation, and perhaps preliminary functional analyses to model what elements are biologically relevant for data collection efforts to focus on.

## What should the project study?

A pilot project should focus on at most a handful of cancer types, informed by a gap analysis on what is available and missing in existing cancer genomics data repositories. New data should be generated from samples that allow for clinically interesting comparisons (which could be different for different cancer types), such as primary tumours and treatment-resistant tumours, as well as some well-studied cell lines and model systems that could more easily permit perturbation studies downstream.

## How should the project be structured?

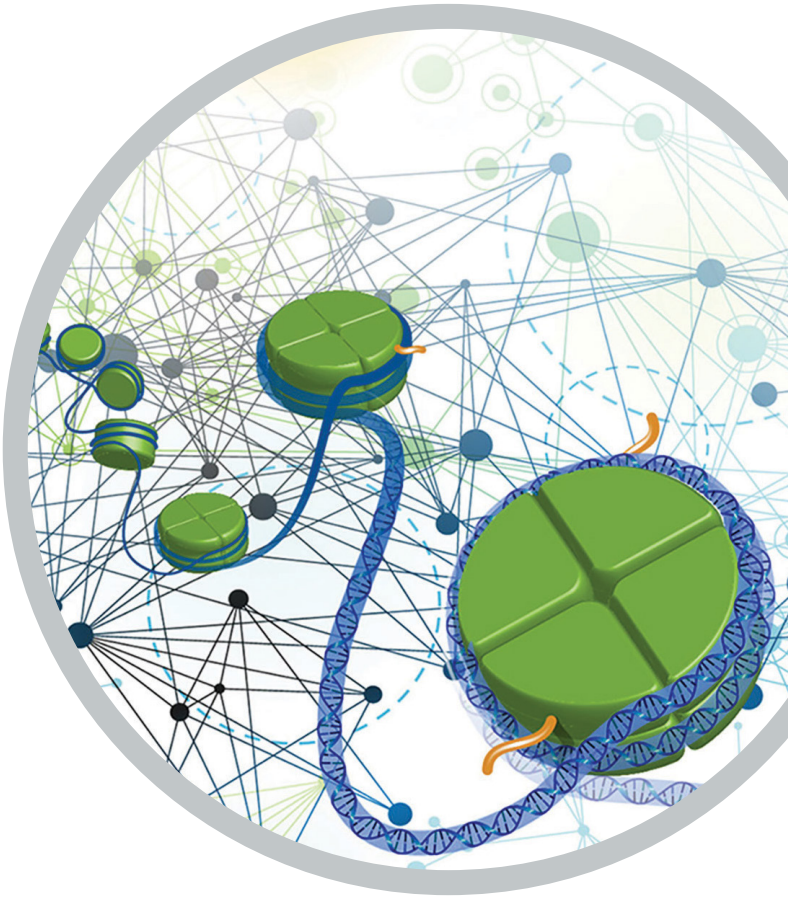
The pilot could potentially establish a “full stack” pipeline that consists of curating and annotating a catalogue of noncoding elements, performing functional analyses, and building biological models for generalizing the functional insights to other elements and cancer types. Researchers should work backwards by first deciding on the biological questions of interest, to determine what functional studies are needed to answer those questions, and finally identify what samples need to be collected for such experiments.

## How could the pilot project most broadly impact the community?

The pilot could serve as an example, and build a set of standardized protocols and data standards, for the rest of the research community to replicate for studying other cancers. The project (and the resources it creates) should be dynamic and responsive to community input as well as evolving biological/clinical questions. It also needs to build in mechanisms to take in data from the broader community, including incentives

for the community to contribute data. Partnerships should be built with organizations and institutions for sourcing samples and performing analyses, and ones that can host the database and user portal for the long run. To alleviate potential privacy concerns related to using patient samples, the project may need to consider measures such as data federation, or to build in the capacity to analyze encrypted data.

With a consortium of motivated researchers, a mechanism for seeking community input and determining scientific priorities, and a thoughtfully built governance structure, a pilot project could begin to generate shared resources that will significantly advance our understanding of the role of the majority of our genome in shaping the development of cancer, with the potential to make important contributions to improving human health.



# ROUNDTABLE PARTICIPANTS

## CONVENORS

▲● **Alan Bernstein**, President and CEO, CIFAR

▲● **Shirley Tilghman**, President and Professor Emerita, Princeton University

▲● **Harold Varmus**, Lewis Thomas University Professor of Medicine, Weill Cornell Medicine; Senior Associate Member, New York Genome Center

---

▲● **Bradley Bernstein**, Professor, Harvard Medical School; Chair of Cancer Biology, Dana-Farber Cancer Institute; Institute Member, Director of the Epigenomics Program and Co-Director of the Gene Regulation Observatory, Broad Institute

▲● **Jason Buenrostro**, Assistant Professor, Harvard University; Associate Member and Co-Director of the Gene Regulation Observatory, Broad Institute

▲ **Howard Chang**, Professor, Stanford University

● **Jonah Cool**, Science Program Officer, Chan Zuckerberg Initiative

▲ **Titia de Lange**, Leon Hess Professor and Director of the Anderson Center for Cancer Research, Rockefeller University

● **Stacey Edwards**, Associate Professor and Laboratory Group Leader, QIMR Berghofer

▲● **Juliet French**, Associate Professor, Laboratory Group Leader and Deputy Department Coordinator, QIMR Berghofer

▲ **Eileen Furlong**, Head of Genome Biology Unit, EMBL

● **Todd Golub**, Director, Broad Institute; Professor, Harvard Medical School

▲ **Marcin Imieliński**, Associate Professor, Weill Cornell Medicine; Core Member, New York Genome Center

● **Rory Johnson**, Associate Professor, University College Dublin

▲● **Ekta Khurana**, Associate Professor, Weill Cornell Medicine; Co-Leader of the Cancer Genetics and Epigenetics Program, Meyer Cancer Center

▲● **Jeannie Lee**, Professor, Harvard Medical School; Vice-Chair of Molecular Biology, Massachusetts General Hospital

▲● **Yang Liu**, Associate Professor, University of Pittsburgh

▲● **Mathieu Lupien**, Senior Scientist, Princess Margaret Cancer Centre; Professor, University of Toronto; Investigator, Ontario Institute for Cancer Research

▲ **Michael McManus**, Professor, University of California, San Francisco; Member, Innovative Genomics Institute

● **Matthew Meyerson**, Professor, Harvard Medical School; Director of the Center for Cancer Genome Discovery, Dana-Farber Cancer Institute; Institute Member and Director of Cancer Genomics, Broad Institute

▲● **Steve Murray**, Associate Professor and Director of Knockout Mouse Project (KOMP) Model Development, The Jackson Laboratory

▲ **Stephen Quake**, Professor, Stanford University; Co-President, Chan Zuckerberg Biohub

▲● **Neville Sanjana**, Assistant Professor, New York University; Core Member, New York Genome Center

● **Martin Smith**, Assistant Professor, Université de Montréal

▲● **Sarah Teichmann**, Senior Group Leader and Head of Cellular Genetics Programme, Wellcome Sanger Institute; Co-Founder and Co-Chair of Organizing Committee, Human Cell Atlas

▲● **Igor Ulitsky**, Associate Professor, Weizmann Institute of Science

▲ **Karen Vousden**, Principal Group Leader, Francis Crick Institute; Chief Scientist, Cancer Research UK

▲ first roundtable / ● second roundtable


MaRS Centre, West Tower
661 University Ave., Suite 505
Toronto, ON, M5G 1M1 Canada
cifar.ca

**STAY CONNECTED:**

@CIFAR_News /CIFARVideo /CIFAR

Charitable Registration Number: 11921 9251 RR0001