 DOI: 10.24425/academiaPAS.2024.151184

Big Data in Every Cell

The emergence of a new field of science in recent years – *bioinformatics* – has given us a more precise understanding of the workings of human DNA and RNA.

Beata Małachowska, MD, PhD specializes in medicine, biostatistics, bioinformatics, and computational biology. She works at Albert Einstein College of Medicine in New York City as a postdoctoral fellow and data analyst in the Department of Radiation Oncology. Her research focuses on the analysis of single-cell sequencing data, with particular emphasis on areas such as radiation-induced intestinal injury and the role of thrombopoietin mimetics in the treatment of radiation-induced syndromes. She has received numerous scientific awards, including the ISPAD-JDRF Fellowship Award, the START Scholarship from the Foundation for Polish Science, the Golden Otis Award, and the L'Oreal-UNESCO Award for Women in Science.
b.e.malachowska@gmail.com

Beata Małachowska

Department of Radiation Oncology, Albert Einstein College of Medicine, New York, USA

When the Human Genome Project kicked off back in 1990, scientists estimated it would take 15 years to decipher the entire sequence. The human genome consists of approximately 3 billion base pairs of DNA, and the assumption is that this information remains constant and uniform across all the cells of an organism. This monumental project brought together 20 major laboratories across six countries.

Fast forward over 30 years, and we can now not only decode the complete DNA sequence of an individual but also analyze the dynamic changes in RNA production across their different cells based on this DNA. Moreover, this can be done within several dozen hours in a single laboratory, simultaneously examining thousands of cells. With the cost of sequencing a single genome reduced to roughly $1,000, obtaining this information or funding such projects is no longer the primary challenge. Instead, the focus has shifted to how to effectively process the vast amounts of data generated from individual biological experiments in a manner that is comprehensible and useful to people.

Within each living human cell lies a nucleus containing 46 DNA molecules. If unraveled and stretched out end to end, these molecules would extend approximately two meters in length. To accommodate this immense length within the confines of a tiny cell, the DNA must be tightly coiled and carefully arranged. This organization ensures that the cell has access to the necessary DNA fragments when required. The tightly coiled portion in any given cell, termed *heterochromatin*, is inaccessible for RNA production. Conversely, the remaining portion, or *euchromatin*, is accessible, serving as the site for RNA production. Through transcription, some of our DNA (genes) are converted into RNA molecules, which act as templates for protein synthesis. The quantity of proteins produced depends, in part, on the number of RNA templates generated from DNA.

Differing functions

While every living cell within an organism shares the same DNA, the shapes and functions of those cells vary significantly. The cells of the immune system, for instance, have the ability to recognize bacteria within our bodies utilizing special receptors (proteins) on their surfaces to identify bacterial fragments as foreign. Neurons, the cells of the nervous system, on the other hand, are able to rapidly transmit signals from one end of our body to the other using electrical signals, facilitated by specialized channels (proteins) on their surfaces. These examples highlight how a cell's identity and function are determined by the proteins it can produce, which are, in turn, depend on the available RNA (or more specifically mRNA, the type of RNA used as an informational template).

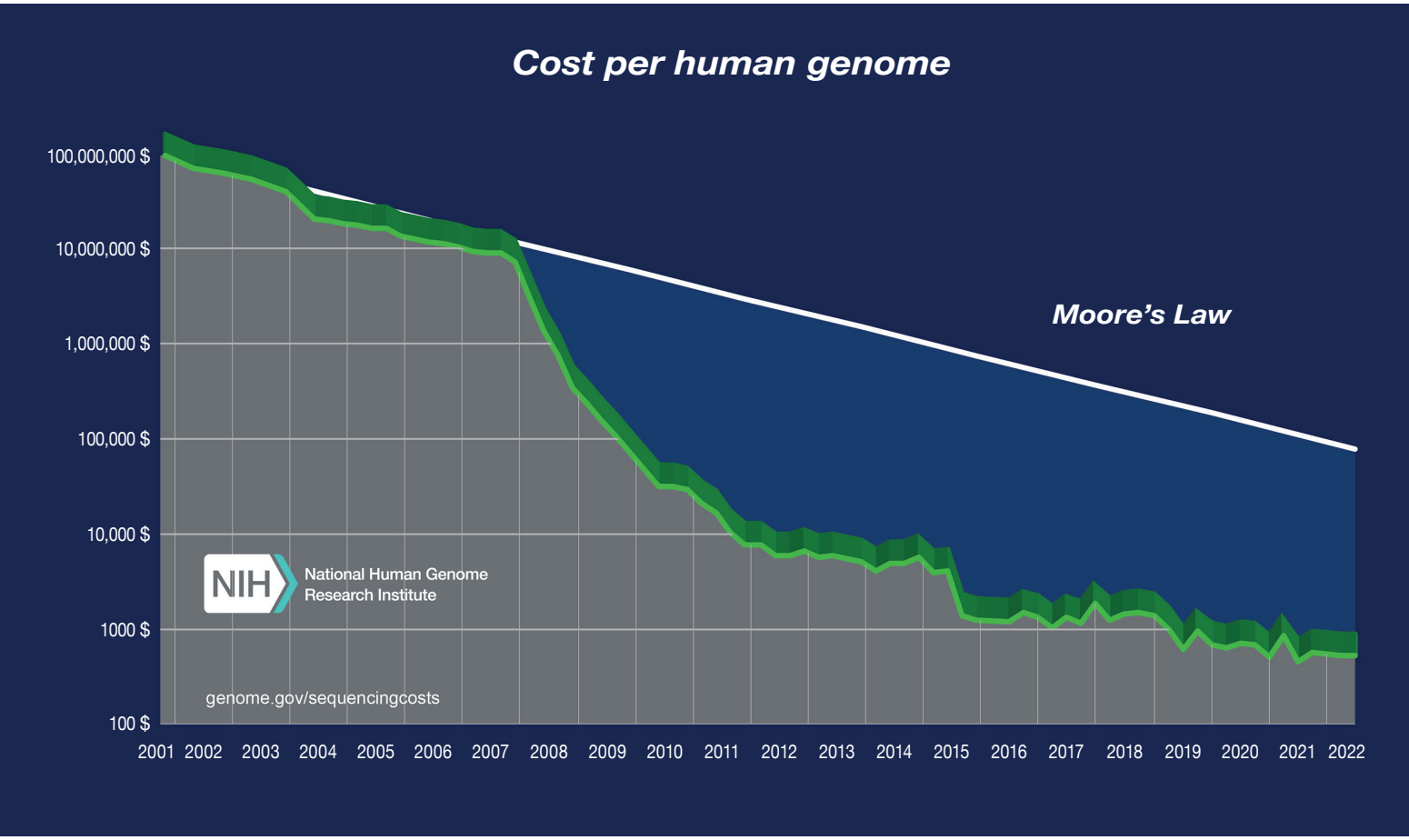
Thanks to next-generation sequencing techniques (commercially available since 2005), retriev-

ing information from DNA/RNA on a massive scale has become simpler and progressively more affordable over time. The Human Genome Project ended up lasting 13 years, two less than planned, primarily due to the introduction of *shotgun sequencing*. This method involves breaking long DNA molecules into smaller fragments, deciphering their sequences, and then reassembling the information using computers. Given the vast amount of information carried by the DNA of a single human (3.4 GB), manually comparing sequences and aligning them to form a coherent linear ensemble would have been a tremendously time-consuming task. Computerized data-processing therefore became indispensable in the field of biology, giving rise to the field of *bioinformatics*. "Dry labs" emerged, focusing solely on the analysis of data obtained from biological experiments, as distinct from the "wet labs" traditionally used in biology. Dry labs employ individuals with a range of diverse backgrounds, including bioinformaticians, biologists, biotechnologists, physicians, who became interested in bioinformatics for their research needs, as well as software engineers and computer scientists specializing in bioinformatic analyses. This amalgam of people and types of experience works as a crucible for continuously developing new data analysis algorithms – a particularly appealing workplace for young individuals who grew up in the computer era.

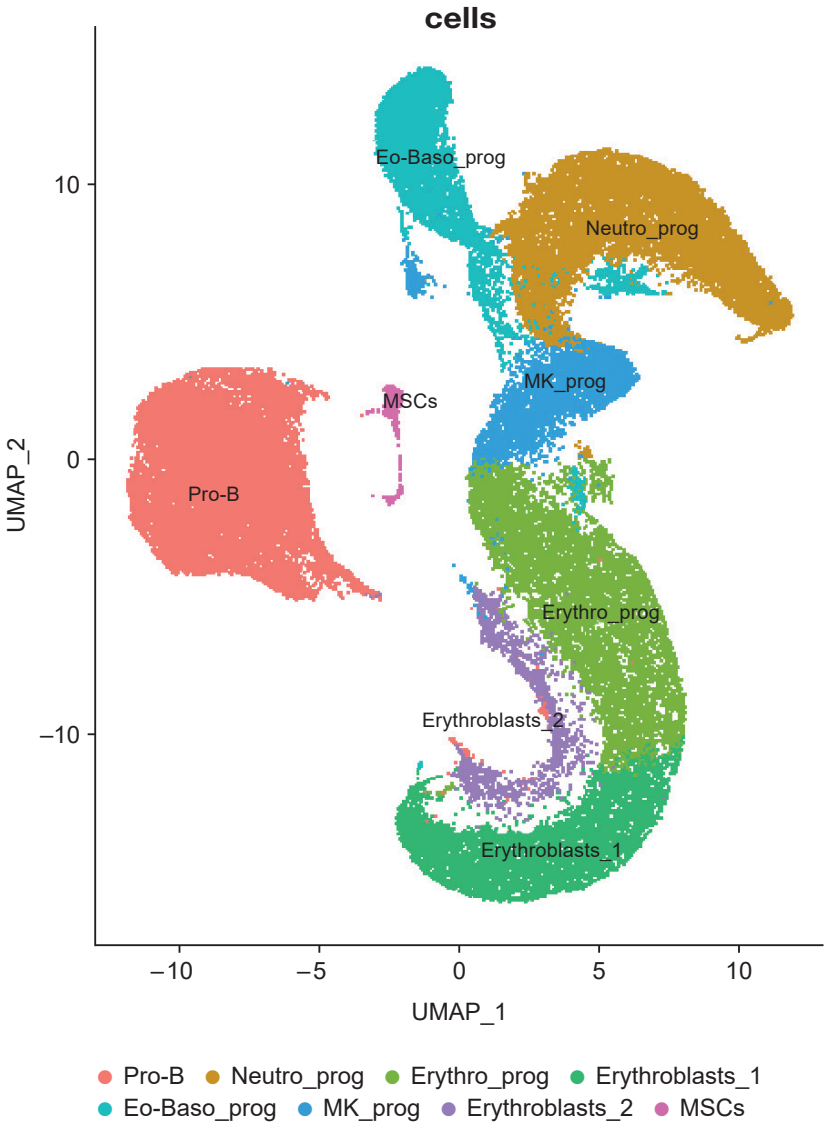
Now that the Human Genome Project has been completed, we have unlocked the foundational blueprint of human DNA. Interestingly, DNA sequencing techniques can also be applied to decipher RNA, as it can be transcribed directly into complementary DNA (akin to a mirror image). However, as previously mentioned, RNA is not uniform across all cells. Its type and quantity can vary depending on factors such as function, shape, disease state, or drug influence. Some

The DNA record

Mutations can give rise to variations in the DNA composition across different cells of a person's body. Such discrepancies are notably pronounced in cancer cells, where DNA alterations involve not only changes in individual nucleotides but also amplifications of specific segments, deletions of others, and structural modifications like fragment inversions.



The declining cost of human genome sequencing over the decades



A graphical presentation of results from single-cell sequencing analysis. Each point represents an individual cell, and the distance between points measures the similarity of RNA profiles between cells

genes may reduce their expression or completely shut it down, while others may activate or heighten their expression levels. Presently, it is estimated that the human genome contains 20,000 to 25,000 genes, with their diverse expressions serving as the basis for biological diversity among cells, tissues, and even entire organisms. The human body encompasses around 200 different cell types, each with its own distinct expression profile, which can fluctuate due to various factors like sex, disease state, time of day, diet, or environmental conditions, thus resulting in unique expression profiles for every cell within our body.

With further advancements in sequencing techniques, we now possess the capability to acquire insights into the type and quantity of RNA produced by individual cells (scRNAseq – single-cell RNA sequencing). It is estimated that a single cell harbors approximately 360,000 mRNA molecules, of which 12,000 exhibit a unique sequence (these are *single transcripts*, a subtype of mRNA), while the remainder are duplicates. In a typical experiment of this nature, about 10,000–20,000 cells from one sample are sequenced, with the possibility of multiple samples in a single experiment. However, the challenge lies in analyzing such vast volumes of data.

Whence the difficulties?

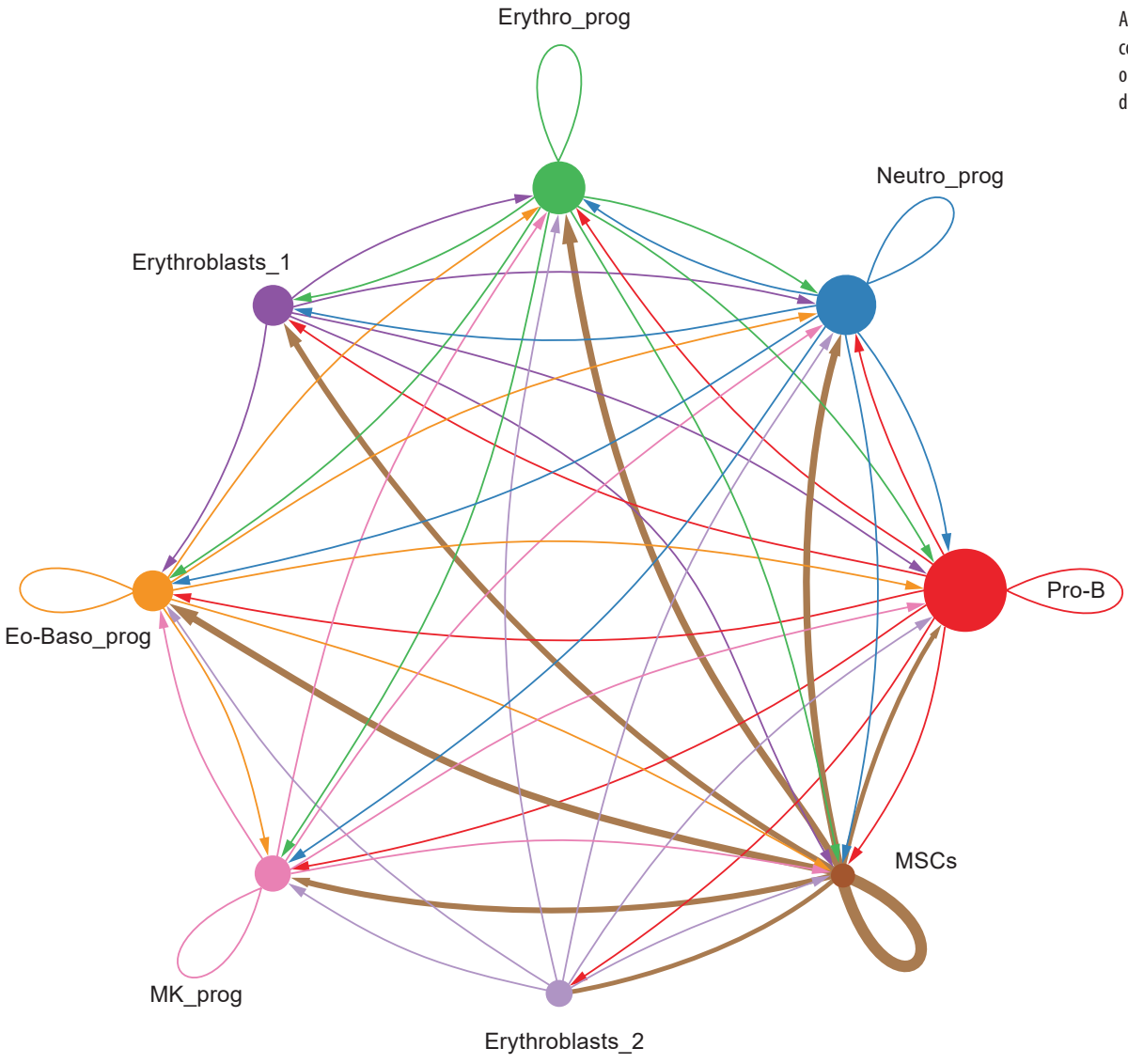
First and foremost, we must grasp how data is generated by sequencing a single cell. Initially, individual cells from a given sample are separated and placed into droplets containing RNA markers and appropriate reagents. These droplets serve as miniature reaction test-tubes where cell-specific reactions take place. RNA from each cell is fragmented, and distinct markers are attached to these fragments. To ensure that each cell resides in a separate droplet with its own marker, most of the droplets do not contain cells. Subsequently, we sequence approximately one million droplets to glean information from 10,000–20,000 cells.

The initial stage of analysis involves discerning which droplets contain genuine cells, and which ones contain only contaminants. Next, RNA fragments are cross-referenced with the genome to piece together fragments originating from a single gene, thereby determining which genes are active in each cell and their expression levels. Given the vast number of droplets and RNA fragments to compare (e.g. 450 million for 20,000 cells), this process demands substantial computational power and takes several days. Once we have identified the data stemming from genuine cells, we proceed to the proper data analysis. First, we have to refine and manipulate the data to eliminate any technical artifacts. Subsequently, drawing upon our biological knowledge, we classify the cell types by clustering them and identifying characteristic genes for each group.

Next, a specific analysis is conducted, tailored to the objectives of the particular research project. Currently, there are thousands of algorithms available for analyzing this type of data, but selecting the right one for a given project is difficult and using it correctly is even more challenging. Thanks to data on the RNA profile of a single cell, we can determine how different types of cells will react to a treatment,

Non-coding DNA

Not all genes – DNA segments transcribed into RNA – lead to the formation of proteins. Some of the resulting RNA functions by regulating the expression of other genes, stabilizing the genome, or assisting in the process of building a particular protein, but not as a template recording its sequence.



Analyzing intercellular communication based on single-cell sequencing data

or explain the mechanism by which new drugs act at the molecular level, allowing us to predict optimal conditions for their use and potential side effects. We can also describe how cells communicate with each other about certain dangers (infection, damage, the need for repair). With the knowledge so gained, we can simulate signals emitted by individual cells, using appropriate chemical substances to induce cells to work for the benefit of the whole organism (e.g. by repairing damaged tissues). Research underway at the Department of Radiation Oncology at Albert Einstein College of Medicine is seeking new drugs that stimulate the regeneration of specific tissues after exposure to ionizing radiation (as in the case of radiation incidents).

In today's era of modern medicine and biology, bioinformatic analysis plays a crucial role. It enables a precise understanding of gene activity at the single-cell level, which is invaluable for uncovering subtle differences in cell functioning in various states of health and disease. Through scRNAseq, we can identify specific cell subtypes, understand their functions, and interactions in a way that was previously impossible.

These analyses provide valuable insights that contribute to the development of *personalized medicine*, enabling precise diagnoses and tailored therapies. In diseases such as cancer, knowledge of disturbed gene expression profiles in individual cells can lead to the development of more effective treatment methods. Moreover, understanding RNA dynamics in cells is crucial for research into various biological processes, from embryonic development to immune responses.

Therefore, in an era where we have vast amounts of biological data, bioinformatic analysis is becoming an essential tool for translating this data into useful knowledge, thereby accelerating progress in science and medicine. ■

Further reading:

Loewe L., Hill W.G., The population genetics of mutations: good, bad and indifferent, *Philos Trans R Soc Lond B Biol Sci*. 2010, https://doi.org/10.1098/rstb.2009.0317

Vercellino J., Małachowska B., Kulkarni S. et al., Thrombopoietin mimetic stimulates bone marrow vascular and stromal niches to mitigate acute radiation syndrome, *Stem Cell Res Ther* 2024, https://doi.org/10.1186/s13287-024-03734-z