Intro to the Postgenomic Biotechnology course

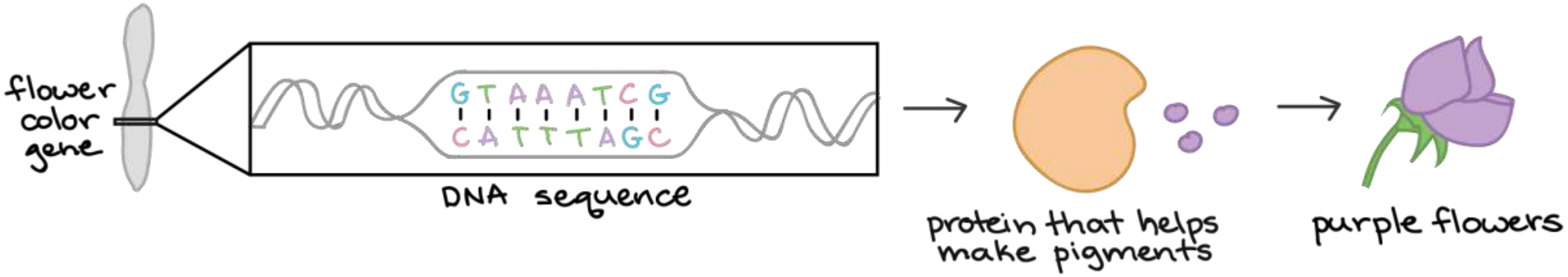
Candidate of bio. Sci. **Meldebekova Aliya A.**

Agenda

- **Genes. Basic dogma of molecular biology**
- **Postgenomics:**
  - **Genome Size, the C-value Paradox, G-value paradox and Junk DNA**

Genes.
Basic dogma of molecular biology

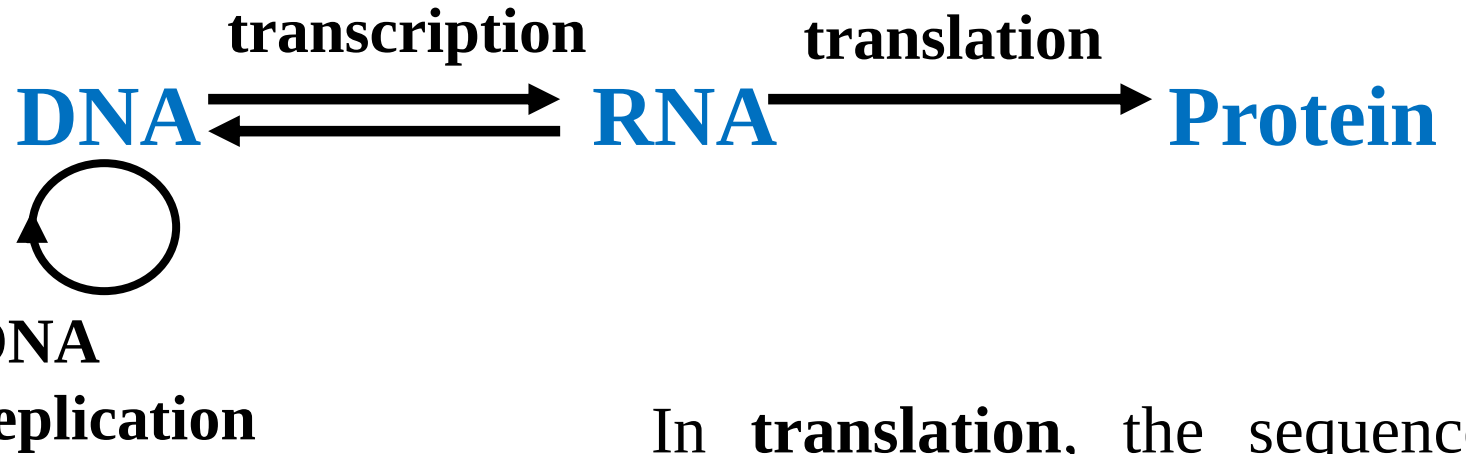
- A DNA molecule isn't just a long, boring string of nucleotides. Instead, it's divided up into functional units called **genes**.
- Each gene provides instructions for a functional product, that is, a molecule needed to perform a job in the cell.
- In many cases, the functional product of a gene is a protein. For example, Mendel's flower color gene provides instructions for a protein that helps make colored molecules (pigments) in flower petals.



- The functional products of most known genes are proteins, or, more accurately, polypeptides.
- **Polypeptide** is just another word for a chain of amino acids.
- Although many proteins consist of a single polypeptide, some are made up of multiple polypeptides.
- Genes that specify polypeptides are called **protein-coding** genes.

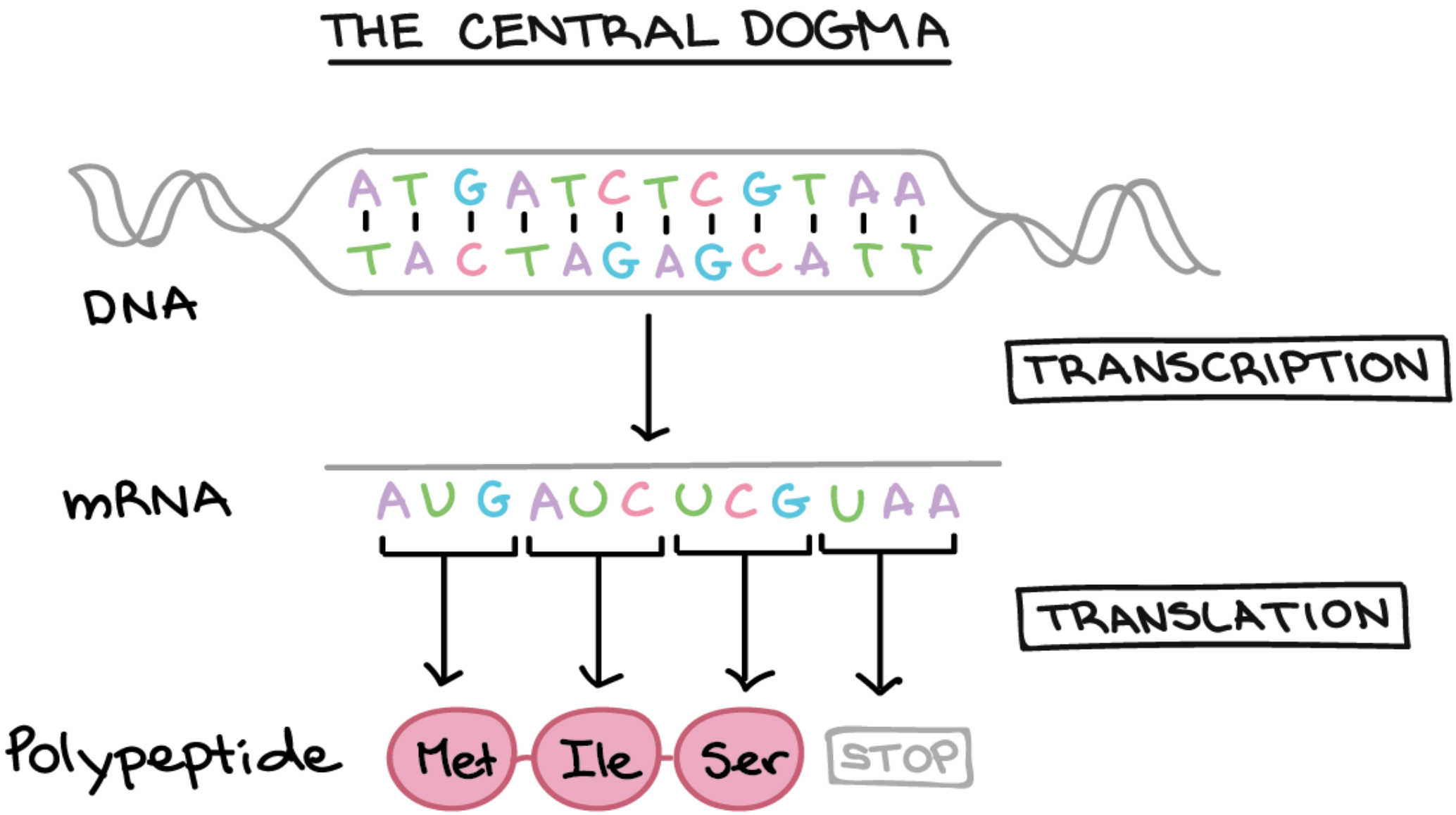
Not all genes specify polypeptides.
Instead, some provide instructions to build functional RNA molecules, such as the **transfer RNAs** and **ribosomal RNAs** that play roles in translation.

Central Dogma of Biology



In **transcription**, the DNA sequence of a gene is copied to make an RNA molecule.

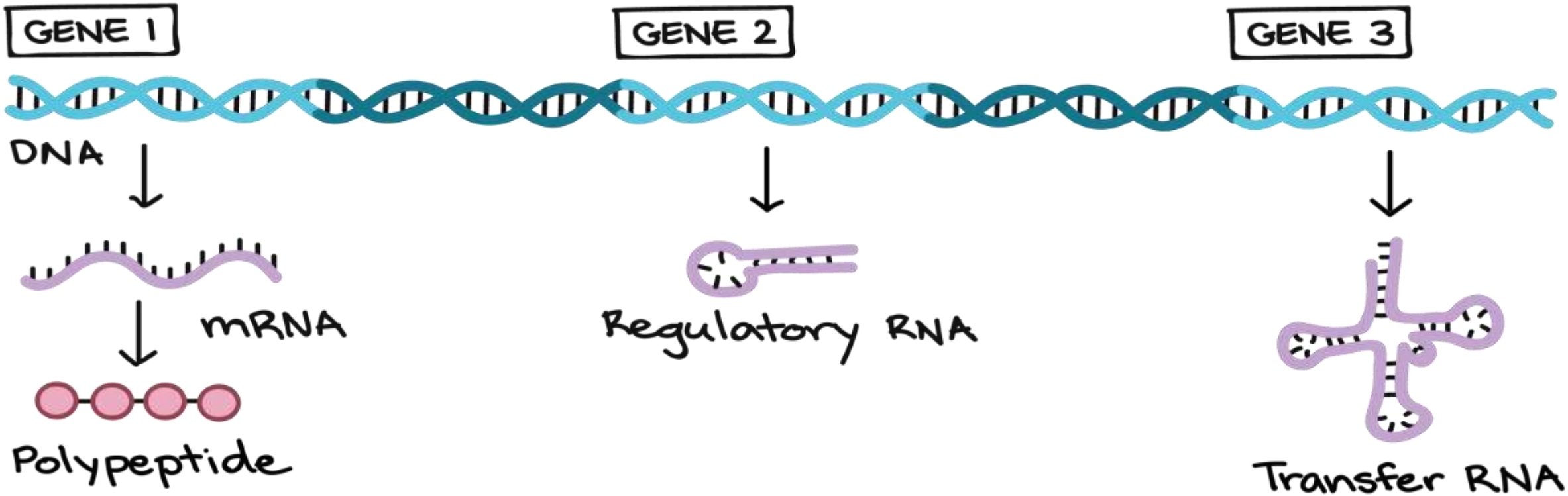
In **translation**, the sequence of the mRNA is decoded to specify the amino acid sequence of a polypeptide.
The name *translation* reflects that the nucleotide sequence of the mRNA sequence must be translated into the "language" of amino acids.



https://www.youtube.com/watch?v=vWBxqCe3iBA

As mentioned above, an organism's DNA can be divided into functional units called **genes**. Each gene consists of a sequence of DNA, and that sequence provides instructions to build a product needed by the cell.

Some products are polypeptides, while others are functional RNAs.



- The idea that genes encode polypeptides has been around for many years (experiments by **Beadle** and **Tatum** in the 1940s).

- Certain types of functional RNAs (such as **transfer RNAs** and **ribosomal RNAs**) have been known for many years.
- However, scientists have only recently discovered **many other genes** that encode **regulatory RNAs, non-protein-coding RNAs** that change the expression of other genes.
- How these RNAs work is an active area of research.

Gene expression

Thus, during expression of a **protein-coding gene**, information flows from **DNA → RNA → protein**

- This directional flow of information is known as the **central dogma** of molecular biology.
- **Non-protein-coding genes** (genes that specify functional RNAs) are still transcribed to produce an RNA, but this RNA is not translated into a polypeptide.
- For both type of gene, th**e process of going from DNA to a functional product** is known as **gene expression**.

Summary

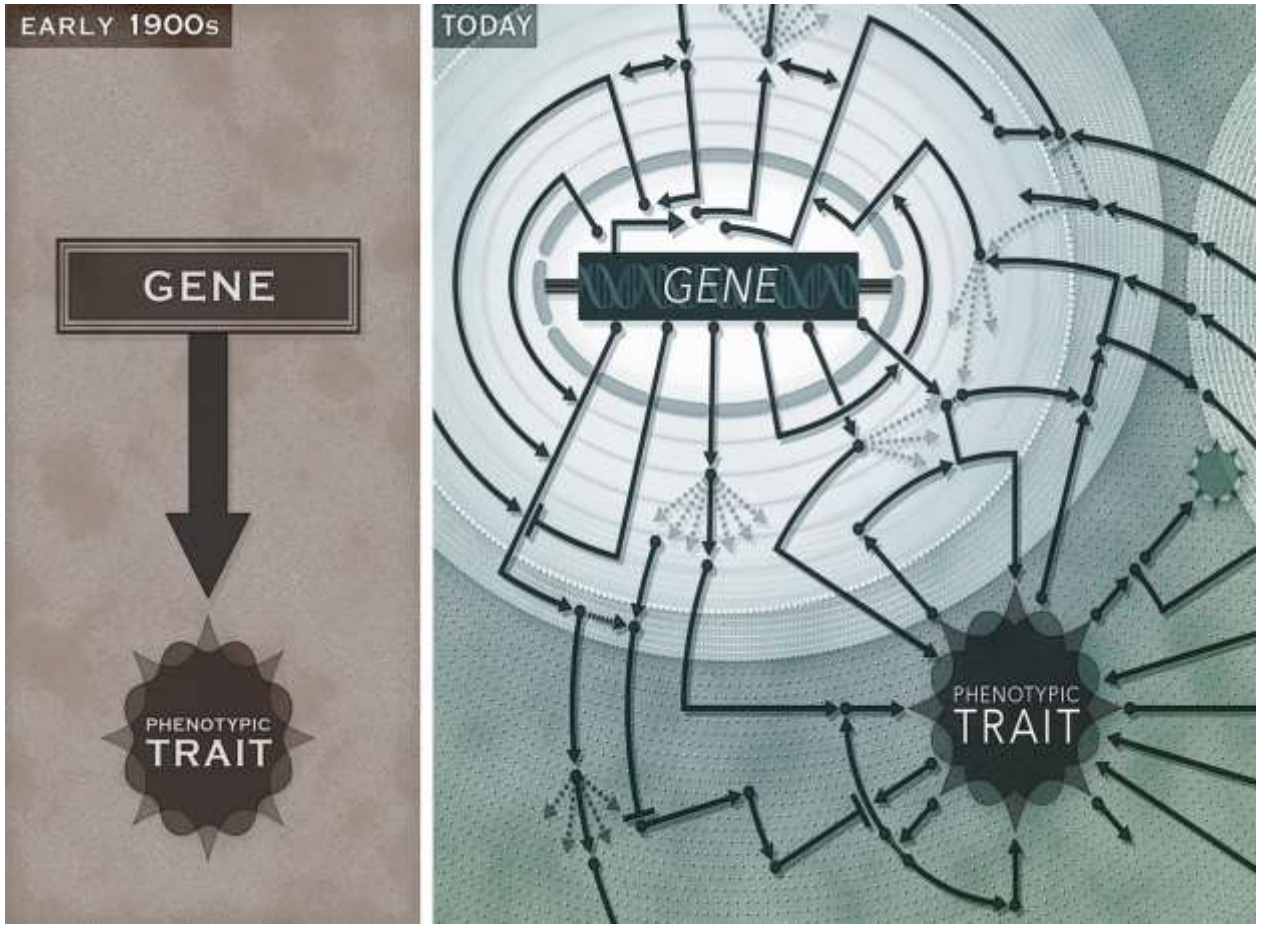
- **DNA is divided up** into functional **units** called **genes**, which may specify polypeptides (proteins and protein subunits) or functional RNAs (such as tRNAs and rRNAs).
- Information from a gene is used to build a functional product in a process called **gene expression**.
- A gene that encodes a polypeptide is expressed in two steps. In this process, information flows from DNA → RNA → protein, a directional relationship known as the **central dogma** of molecular biology.

Postgenomics

Introduction

Years after the Human Genome Project's completion, the life sciences stand in a moment of uncertainty, transition, and contestation.
The "postgenomic era" has seen rapid shifts in research methodology, funding, scientific labor, and disciplinary structures.
Postgenomics is transforming our understanding of disease and health, our environment, and the categories of race, class, and gender.
At the same time, "the gene" retains its centrality and power in biological and popular discourse.
The contributors to Postgenomics analyze these ruptures and continuities and place them in historical, social, and political context.
Postgenomics, they argue, forces a rethinking of the genome itself, and opens new territory for conversations between the social sciences, humanities, and life sciences.

- The gene was first defined as a Mendelian trait, and then, it became a locus and then linked to a protein and, finally, to elements of proteins (amino acids). This development reflects the desire to focus the gene's definitions on a unique entity : a phenotype, a mutation, a protein, an enzyme, a nucleotide sequence, and so on.
- Yet, despite the progress made in the molecular understanding of genes, functionalist expressions - “genes for” – have never stopped multiplying**: the gene “for” cancer, or schizophrenia, diabetes, intelligence, homosexuality, crime depression**, and so on.



According to the postgenomic concept, the gene is no longer interpreted as a unique functional or molecular entity.

- In the 1990s, it became possible to sequence whole genomes, which gave rise to genomics.
- Rather than analyzing the sequence and function of individual genes, **genomics tries to identify and understand both individual genes and their interaction with each other**. Together with transcriptomics, proteomics, metabolomics, and bioinformatics, genomics spawned systems biology, which aims to understand and model the interaction of many components in the cell, not just genes, in an effort to explain how genetic information translates into phenotypic traits.

Post-genomics is the term that describes the group of **omics sciences** that emerged following the sequencing of the human genome, including **nutrigenomics, metabolomics, transcriptomics**, and others including **proteomics**.

The term **proteomics** for the large scale study of proteins, was only coined in the mid-1990s.

Although protein biochemistry has a long history, new high-throughput technologies using **mass-spectrometers** allowed the identification of many proteins at once, in contrast to older methods that focused on one protein at a time. Proteomics' historical foundation in biochemistry, recent technological innovations, and the complex scientific challenges specific to protein studies that are not present in the study of genetic DNA makes it a valuable case to compare with genomics.

About 30 years ago researchers and other stakeholders started setting up the first genomics initiative, the **Human Genome Project (HGP**).

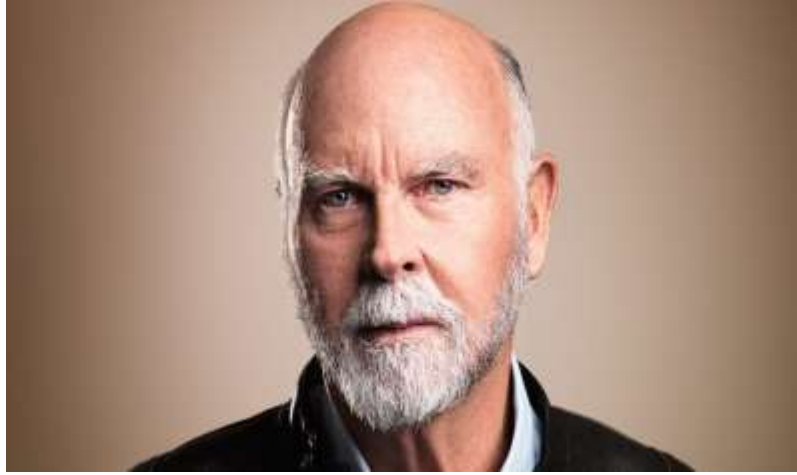
A multi-centre, international program has started in 1990 and was brought to a conclusion in 2003.

More than a decade later **genomics is still big in business** (and big business):

- the Obama administration announced in January 2015 that they intend **to sequence one million human genomes** (search for Precision Medicine Initiative).

- **Craig Venter**, the commercially minded nemesis of the publicly-funded HGP is also in the mix again, this time involved in a privately-funded collaboration that aims to sequence two million genomes over the course of the next ten years.

-And equally important, we see not only the same players clash again but also the same promises being made, with talk of "groundbreaking health benefits" and "new medical breakthroughs" appearing once again in press releases and other announcements.



J. Craig Venter, PhD

J. Craig Venter, PhD, is regarded as one of the leading scientists of the 21st century for his numerous invaluable contributions to genomic research. Dr. Venter is Founder and CEO of the J. Craig Venter Institute (JCVI), a not-for-profit, research organization **dedicated t**o human, microbial, plant, synthetic and environmental **genomic research**, and the exploration of social and ethical issues in genomics.

- Dr. Venter also is a co-founder of **Synthetic Genomics, Inc. (SGI**) and **Human Longevity, Inc. (HLI).**

SGI is a privately held company developing products and solutions including sustainable bio-fuels, vaccines, biotherapeutics and transplantable organs. HLI is a genomic-based, health intelligence company empowering proactive healthcare.

In 1992, Dr. Venter founded The Institute for Genomic Research (TIGR, now part of JCVI), a not-for-profit research institute, where in 1995 he and his team decoded the genome of the first free-living organism, the bacterium *Haemophilus influenzae*, using his new whole genome shotgun technique.
In 1998, Dr. Venter **founded Celera Genomics** to sequence the human genome using new tools and techniques he and his team developed. This research culminated with the February 2001 publication of the human genome in the journal, Science. He and his team at Celera also sequenced the fruit fly, mouse and rat genomes.
Dr. Venter and his team at JCVI continue to blaze new trails in genomics. They have sequenced and analyzed hundreds of genomes, and have published numerous important papers covering such areas as environmental genomics, the first complete diploid human genome, and the groundbreaking advance in creating the first self-replicating bacterial cell constructed entirely with synthetic DNA.

But many things are also different now.

- For instance**, China has emerged as a major player in the genomics** field, with the **BGI (formerly the Beijing Genomics Institute**) already announcing in 2011 the aim **to sequence one million genomes**.

Moreover, **DNA sequencing is no longer the only goal** of these large-scale initiatives:

- the new genomics is of course still a genome-based effort, but it is a transformed enterprise that also focuses on **data about proteins, DNA methylation patterns** or the **physiology** and **the environment of the people** studied;

- DNA sequence data now forms only part of a much larger picture in the push for what is called '**precision**' or '**personalised' medicine**.

- Developments such as these have led many to refer to the present as a '**postgenomic' age**.

Terminology and Definitions

1. Gene – Genome – Genomics

- The term '**genomics**' derives from the term '**genome**', which itself derives (in part) from the term '**gene**'. The meaning(s) of and the relationships between these different terms is by no means simple.
- The term '**gene**' was introduced in 1909 by the Danish biologist **Wilhelm Johannsen**, who used it to refer to the (then uncharacterised) elements that specify the inherited characteristics of an organism.

The term '**genome**' was introduced in 1920 by the German botanist **Hans Winkler** (1877–1945) in his publication "Prevalence and Cause of Parthenogenesis in the Plant and Animal Kingdom". Winkler defined the term as follows:

"I propose to use the expression 'genome' for the haploid set of chromosomes that, in conjunction with the associated protoplasm, represents the material foundation of the systematic unit [often translated as "species".

- The term ‘**genomics**’, finally, was invented in 1986 at a meeting of several scientists who were brainstorming to come up with a name for a new journal that Frank Ruddle (Yale University) and Victor McKusick (Johns Hopkins University) were setting up.
- The aim of this journal was to publish data on the sequencing, mapping and comparison of genomes.
- To capture these different activities Thomas Roderick (Jackson Laboratory) proposed the term ‘genomics’.

This was a significant moment in the history of the life sciences, as it is here that the **-omics** suffix appears for the first time.

Reading the Genome

- The first genome to be sequenced was that of a virus, namely bacteriophage ΦX174, sequenced by Frederick Sanger in 1977 (Sanger et al. 1977).
- Up to about 1985, work on several other viruses was initiated in different laboratories across the world and even the sequencing of model organisms such as the bacterium ***Escherichia coli*** or the roundworm ***Caenorhabditis elegans*** was being tackled.
- Among all the different sequencing efforts at the time **the human genome project** (**HGP**) of course stands out. Not only is the human genome relatively large (roughly 3.2 billion base pairs (bps)) and of key interest to us as human beings, but the HGP itself was envisioned as a diverse large-scale research project with various strands and aims.
- Getting the sequence out of this project was the one goal that got the most attention in the wider media, but surely many would agree that other findings and practices developed within the HGP were of equal or even greater importance.

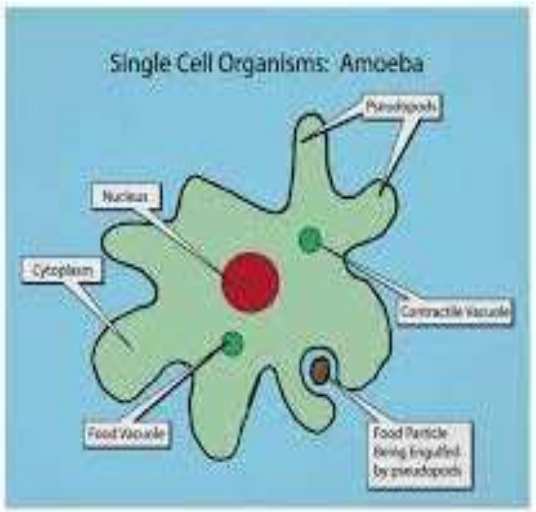
Genome Size, the C-value Paradox and Junk DNA

- It has been known since the 1950s that genome size varies greatly between different organisms, but from the very beginning it was also clear that this diversity has some surprising features.
- One of these features is the **absence of correlation** between **the complexity of an organism** and **the size of its genome**.

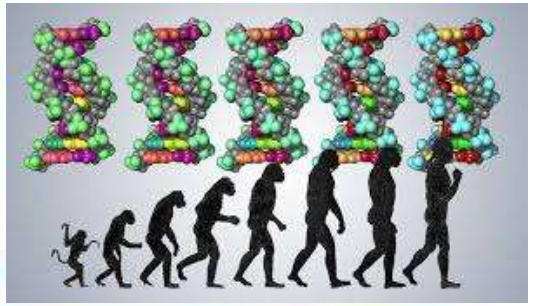
The C-value Paradox

Assuming an informational account of the genome one would expect that the more complex an organism is, the more DNA its genome should contain (this is in fact what many biologists assumed at least until about the 1960s).
How to define and assess the complexity of an organism is a tricky issue, but intuitively it seems reasonable to assume that a single-celled amoeba is less complex than an onion, which in turn is less complex than a large metazoan such as a human being, both in terms of the complexity of the workings and the structure of the organism.

- The expectation was that the DNA content of human cells should be much larger than that of onions or amoebae.
- As it turns out, however, **both the onion and the amoeba have much larger genomes than human beings**.
- The onion, for instance, has a genome of about 16 billion base pairs, meaning it is about five times the size of the human genome (Gregory 2007). The same lack of correlation between genome size and complexity can be found in many other instances.



Amoeba proteus, has a mere **290 billion base pairs**, making it 100 times larger than the human genome



Onion has a genome of **about** 16 billion base pairs

- It was also found early on that very similar species in the same genus show large variation in genome size, despite having similar phenotypes and karyotypes (i.e., number and shape of chromosomes in a genome).
- Within the family of buttercups, for instance, DNA content varied up to 80-fold (Rothfels et al. 1966).
- Also, Holm-Hansen (1969) showed that species of unicellular algae display a 2000-fold difference in DNA content despite all being of similar developmental complexity.
- It was findings such as these that gave a real urgency to addressing this discrepancy that was now labelled the **C-value paradox**.
- The term '**C-value'** refers to the **constant ('C') amount** ('value') of **haploid DNA per nucleus** and is **measured in picograms of DNA per nucleus**.
- The C-value is a measure of the amount of DNA each genome contains.

Junk DNA

- These discussions of genome sizes were closely related to concerns about **gene numbers**. And this consideration of **genome size vs. gene numbers** is what originally gave rise to the concept of '**junk DNA'** (Ohno 1972).
- The reasoning behind this concept was the following:

if one assumes:

- a) that **more complex organisms will have more DNA** than less complex organisms
- b) that **gene numbers** **increase in proportion with** **genome size**, then the genome of the more complex organism should have more genes than the less complex one.
- Human cells, for instance, contain about 750x more DNA than *E. coli*, meaning that they should turn out to have in the range of 3.7 million genes, as *E. coli* has about 5000 genes. This is clearly not the case; even in the 1970s it was generally supposed that the human genome might contain no more than 150,000 genes (Crollius et al. 2000).
- This discrepancy leads to the conclusion that the vast majority of the DNA in our genome cannot be genes and is therefore what **Susumu Ohno** referred to as '**junk**' (In 1972 the late geneticist Susumu Ohno coined the term "junk DNA" to describe **all noncoding sections of a genome, most of which consist of repeated segments scattered randomly throughout the genome)**.

G-value paradox

- The problem that the junk DNA discussion brings up has also been referred to as the 'G-value paradox' ('G' stands for 'gene'), which directly concerns the discrepancy between the number of genes in an organism and its complexity (Hahn & Wray 2002).
- This paradox has been reinforced by the findings of the HGP.
- As Gregory (2005) and other commentators have pointed out, the finding that the human genome contains many fewer genes than expected was one of the most surprising outcomes of the HGP.
- Initial estimates from before the project were in the range of 50,000 to 150,000. These were reduced to about 30,000-35,000 after the publication of the first sequence draft in 2001 and have now been further revised to the order of 20,000 (Gregory 2001).

- Some researchers assumed that the C-value paradox was fully resolved by the recognition that there is non-coding DNA in genomes (Gregory 2001).
- Larger genome size in ‘simpler’ organisms merely means that they have large quantities of non-coding DNA.
- But as Gregory points out, the fact that the majority of DNA in our genomes is non-coding might make the C-value discrepancies less of a paradox, but it gives rise to a whole range of further puzzles (Where does this extra DNA come from? What is its function? Etc.),

The publication of the draft genome sequence in 2001 and the conclusion of the HGP in 2003 **did not give researchers all the tools and insights they needed to tackle these long-standing problems.**

But after the HGP, building on the initial sequencing effort, researchers could start to go beyond the mere sequence and gain a deeper understanding of the workings of the genome.

This put them in a position to tackle issues such as the significance of junk DNA and the C-value paradox more directly (or at least from a different angle).

- Through alternative splicing, most genes can code for many different proteins; moreover, the messenger RNA and the transcribed protein can be further modified to increase the variety of protein products.
- The regulation of transcription and translation is also controlled by a plethora of proteins, RNA molecules, DNA segments, and chemical modification of DNA.
- In eukaryotes, many regulatory regions may be located far away from the coding sequence and may also be involved in the regulation of more than one gene; mechanisms that control transcription therefore involve the three-dimensional structure of DNA.
- The result of this bewildering complexity is that the number of “functional units” is much greater than the number of coding sequences.