Introduction to 'Omics and Bioinformatics

Chris Overall

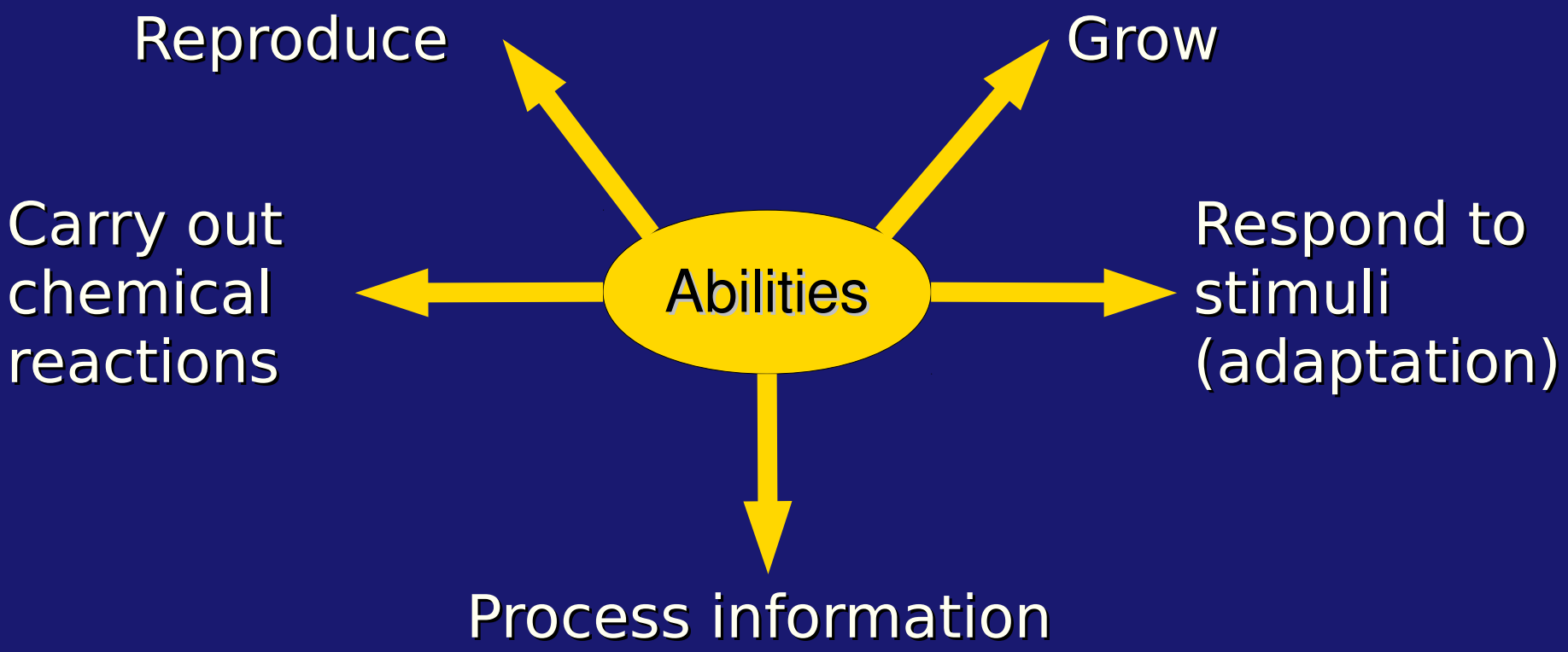
Department of Bioinformatics and Genomics

University of North Carolina – Charlotte

Acquire → Store → Analyze → Visualize

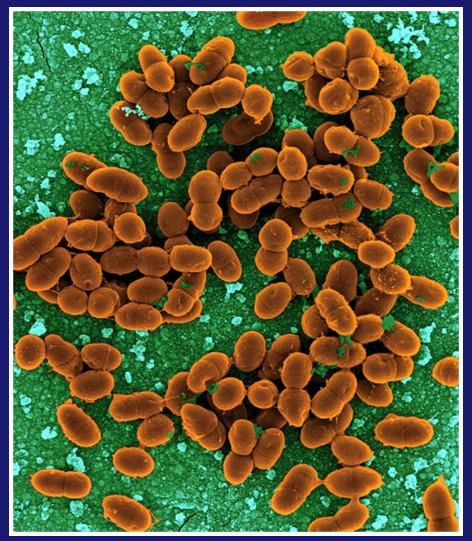
Bioinformatics makes many current biological and biomedical studies possible

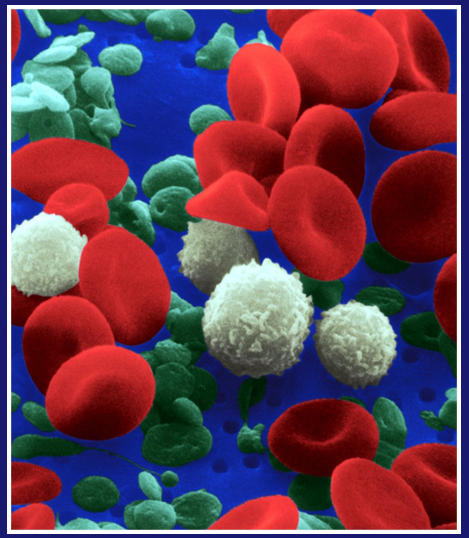
Cells are the fundamental units of life, with many common abilities

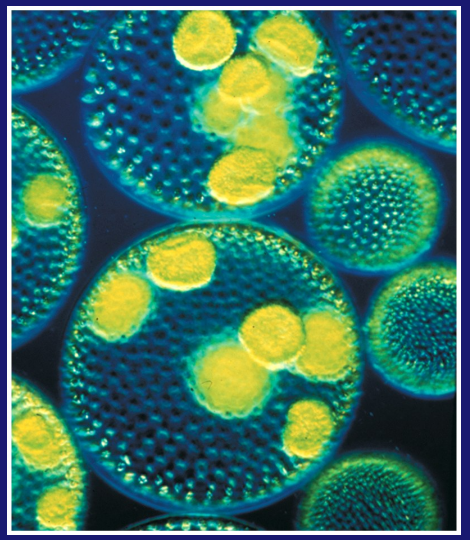
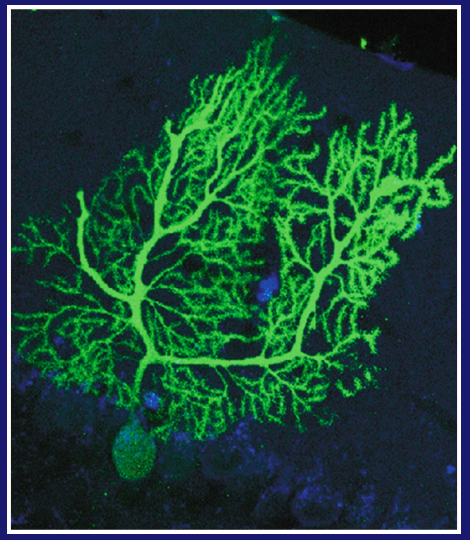
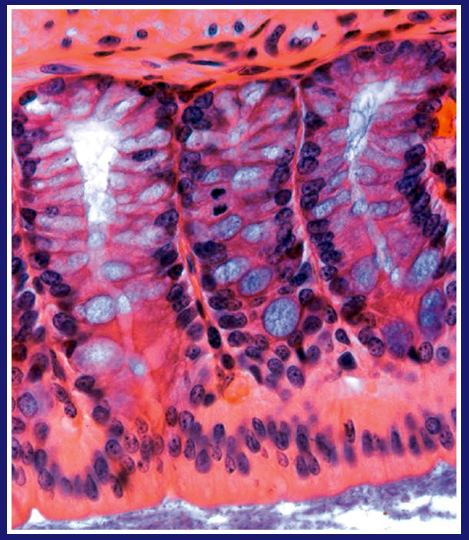


All cells share certain structural features and carry out complicated processes in basically the same way

Despite the similarity of all cells, they vary greatly in size, shape and function

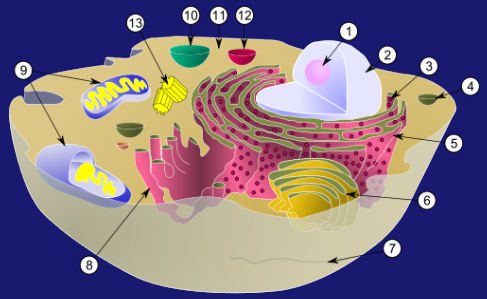


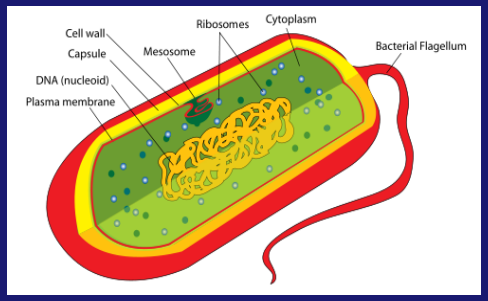



What makes them different?

There are two basic types of cells: prokaryotes and eukaryotes





Eukaryote		Prokaryote
10-100 µm	⟷	1–2 µm
Nucleus	⟷	Nucleoid
Membrane-bound organelles	⟷	No membrane-bound organelles
Multicellular or unicellular	⟷	Mostly unicellular

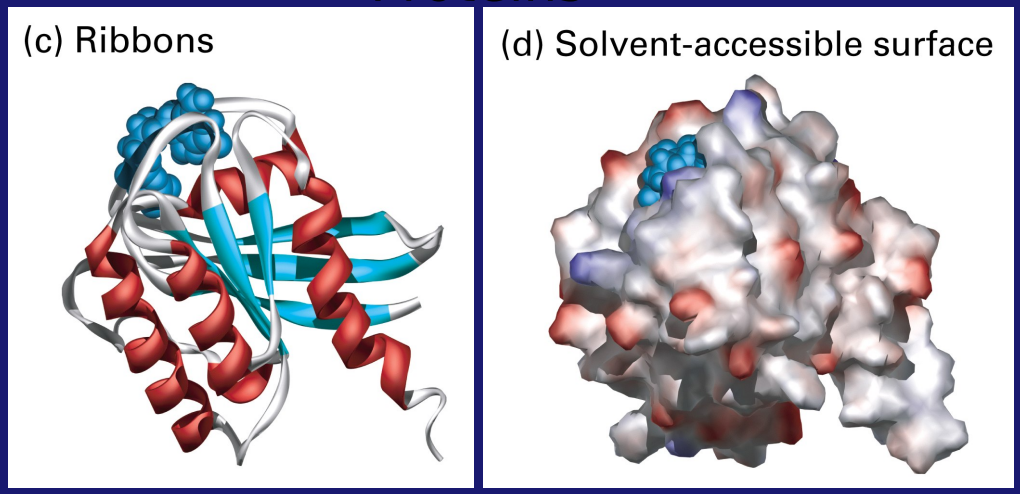
Humans start as a single eukaryotic cell and end up with 100 trillion cells

100 trillion = $100 * 10^{12}$ = 100,000,000,000,000

If you lined up all of your cells in a row, they would stretch across the continental U.S. over 1 million times

All cells use the same four types of macromolecules

Proteins



Nucleic acids

Lipids

Image

needed

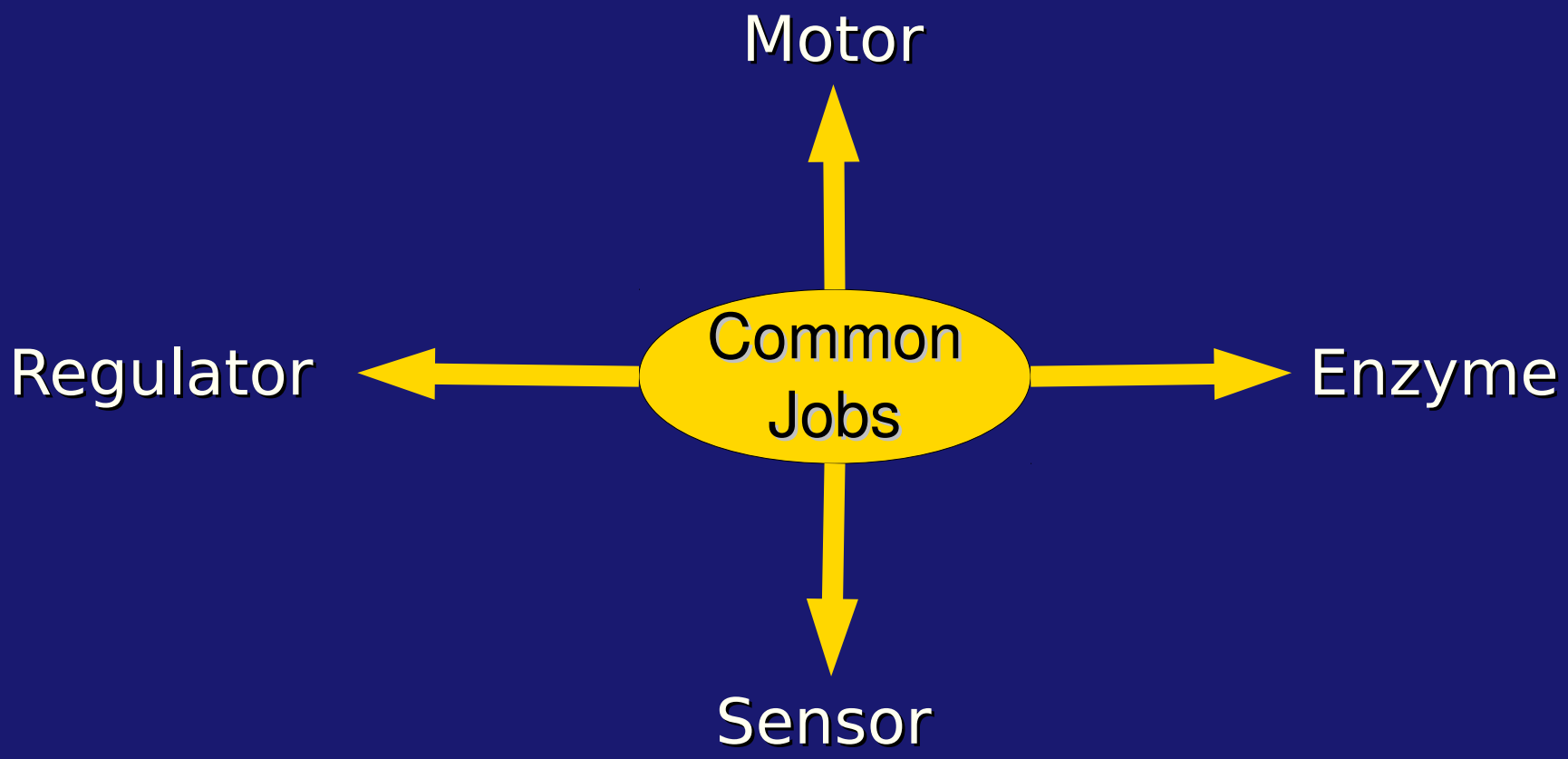
Carbohydrates

Image

needed

We will focus on proteins and nucleic acids

Proteins give a cell structure and perform most cellular tasks



A protein is a linear sequence of amino acids, which uniquely defines it

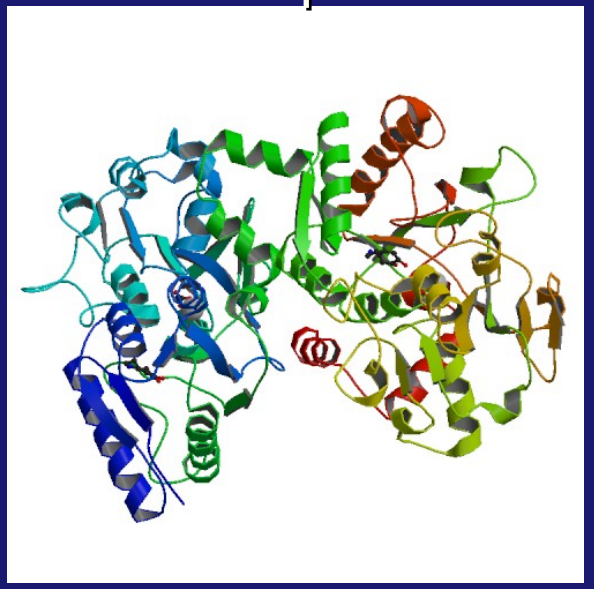
Amino acid sequence

LGLCLAAPRKSVRWCTISPAEAAKCAKFQRNMKKVRGPSVSCIRKTSSFECIQAIAANKA
DAVTLDGGLVYEAGLHPYKLRPVAAEVYQTRGKPQTRYYAVAVVKKGSGFQLNQLQGVKS
CHTGLGRSAGWNIPIGTLRPYLNWTGPPEPLQKAVANFFSASCVPCADGKQYPNLCRLCA
GTEADKCACSSQEPYFGYSGAFKCLENGAGDVAFVKDSTVFENLPDEADRDKYELLCPDN
TRKPVDAFKECHLARVPSHAVVARSVDGREDLIWRLLHRAQEEFGRNKSSAFQLFKSTPE
NKDLLFKDSALGFVRIPSQIDSGLYLGANYLTATQNLRETAAEVAARRERVVWCAVGPEE
ERKCKQWSDVSNRKVACASASTTEECIALVLKGEADALNLDGGFIYVAGKCGLVPVLAEN
QKSQNSNAPDCVHRPPEGYLAVAVVRKSDADLTWNSLSGKKSCHTGVGRTAAWNIPMGLL
FNQTGSCKFDKFFSQSCAPGADPQSSLCALCVGNNENENKCMPNSEERYYGYTGAFRCLA
EKAGDVAFVKDVTVLQNTDGKNSEPWAKDLKQEDFELLCLDGTRKPVAEAESCHLARAPN
HAVVSQSDRAQHLKKVLFLQQDQFGGNGPDCPGKFCLFKSETKNLLFNDNTECLAELQGK

TTYEQYLGSEYVTSITNLRRCSSSPLLEACAFLRA

Folding

Folded protein

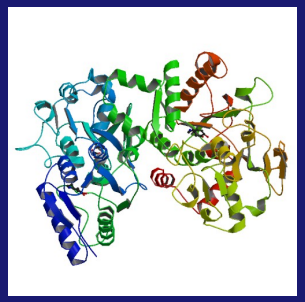
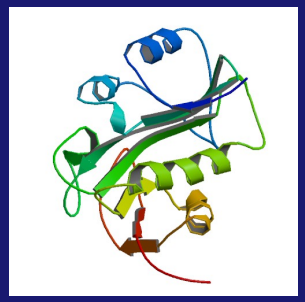
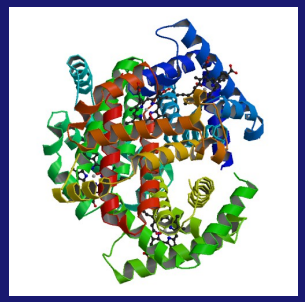


3D structure = function

All organisms use basically the same 20 letter alphabet

Typical length: 100 – 1,000 amino acids

An organism's proteome is the entire set of proteins that it can produce

proteome =  +  +  + … N

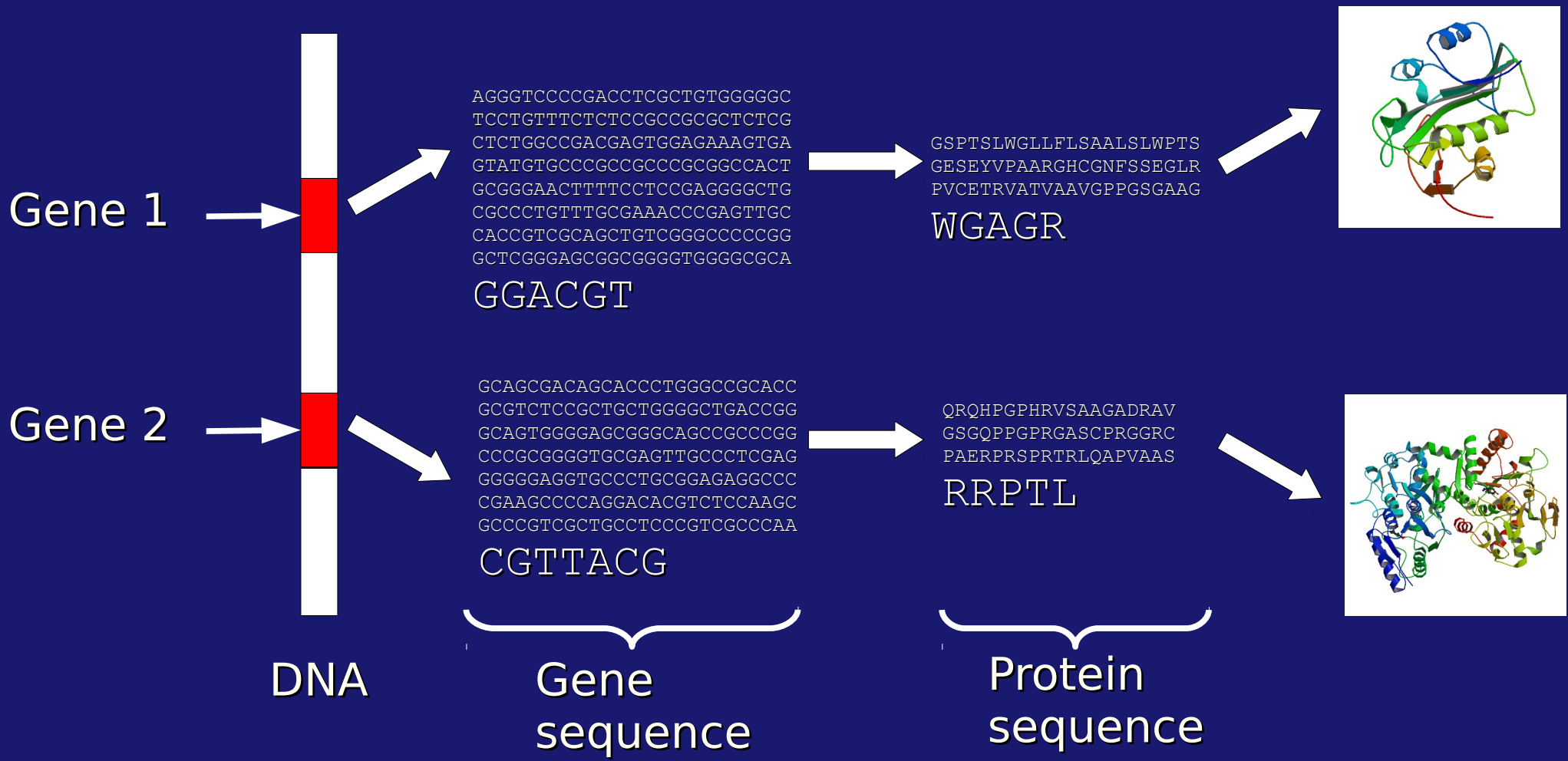
What is the estimated size of the human proteome?

Phrased differently, how many unique proteins can the human body create?

Answer: 2 million

The genes encoded in DNA determine which proteins can be created by a cell

DNA alphabet = {A, C, T, G}



Change this slide!!! DNA → RNA → Protein

An organism's genome is the entirety of its hereditary (genetic) information

Genome = genes + non-coding regions

How many genes do humans have?

Answer: approximately 20,000

The genome of every cell in the human body is basically identical: they have the same genes

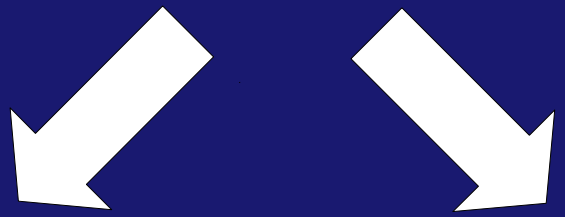
You have 10 times more bacterial cells than human cells in your body

You have 1 quadrillion bacterial cells

Approximately 3 pounds of your weight is bacteria

Most of them won't make you sick

The beneficial bacteria is mostly in your gut

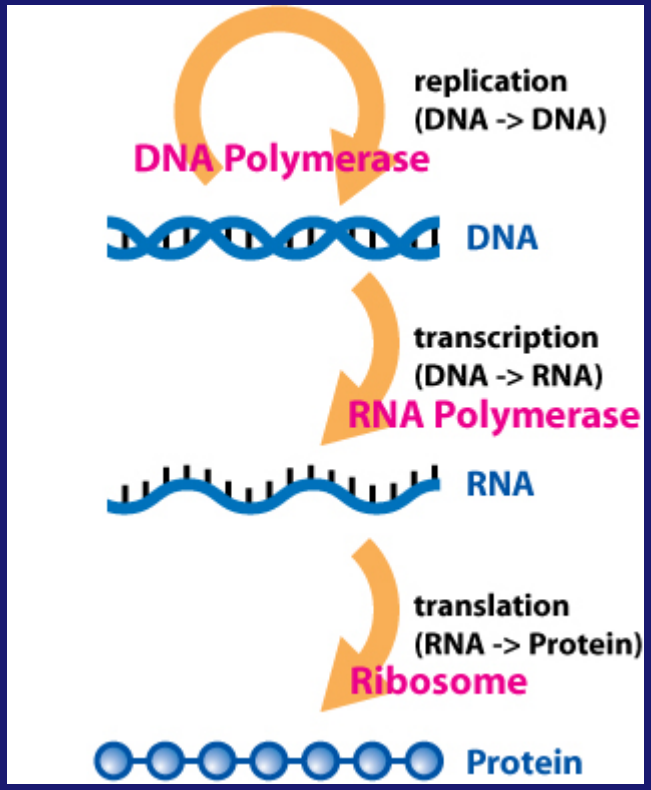


You need them to digest food

E. coli is the most common one

Your metagenome is the entirety of the hereditary (genetic) information of the bacteria in your body

Questions



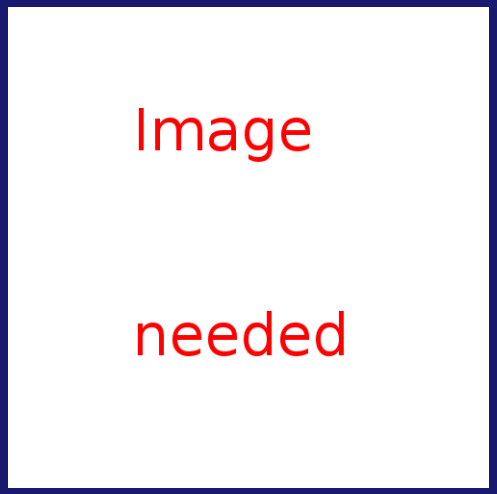
(1) What is a genome?

(2) What is a transcriptome?

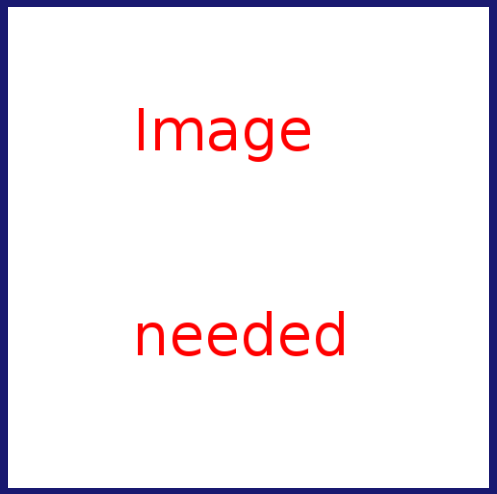
(3) What is a proteome?

(4) What is a metagenome?

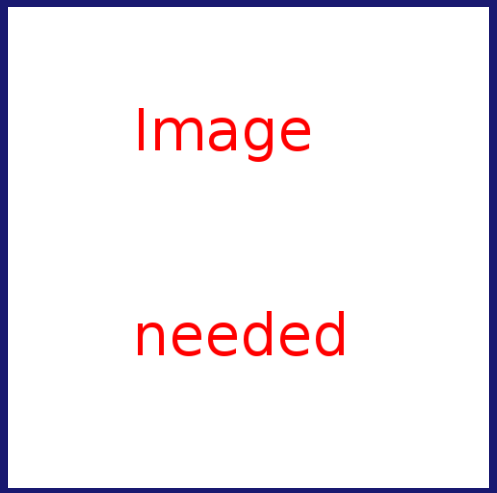
How do we get different types of human cells if their genomes are the same?



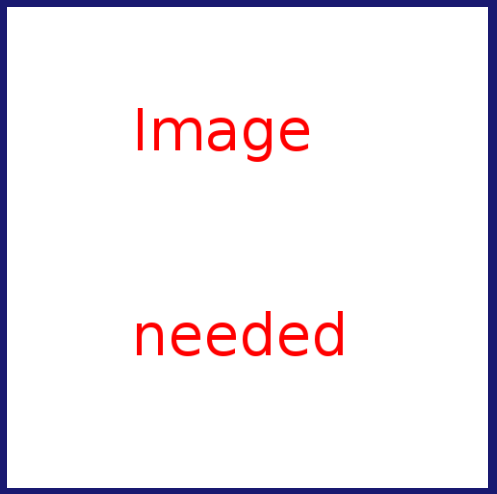
Typical cell



Cancer cell



Neuron

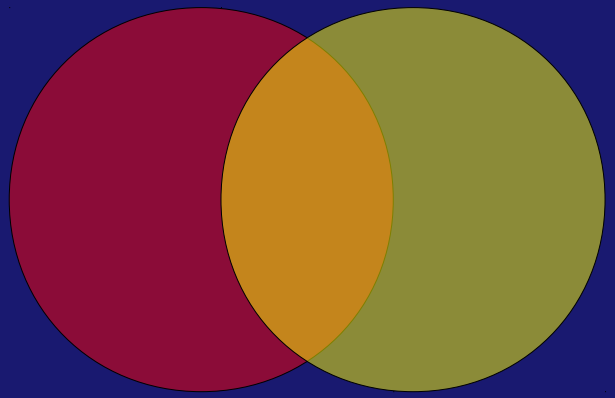


Macrophage

The genome also specifies how, when and where to produce each kind of protein

Specialization

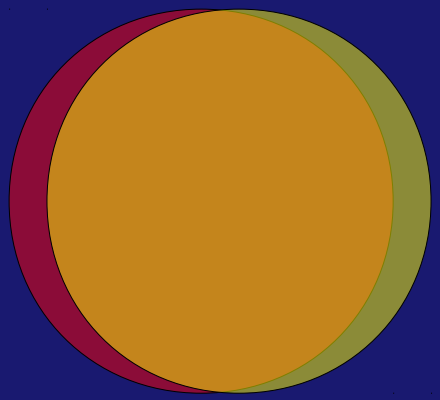
Different human cell types, such as those found in different organs, have different proteomes



Liver vs. neuron

Adaptation

A cell responds to changes in the environment by slightly changing its proteome



Liver 1 vs. liver 2

This is known as gene regulation.

End of Act 1: Questions

(1) What is a cell?

(2) What is a protein?

(3) What is a gene?

(4) How is a protein created from a gene?

(5) What is a genome?

(6) What is a a proteome?

(7) What are the letters (nucleotides) in the DNA alphabet?

(8) How many letters (amino acids) are in the protein alphabet?

'Omics is large-scale, high-throughput molecular biology

*omics

- Studies the entire set of a certain type of molecule in an organism
- Studies when, where and how the molecule interacts, is created and is destroyed

Genomics →

- (1) Study of all the genes and non-coding regions in an organism
- (2) Study of an organism's genome

Functional genomics → Study of what, when, where and how the genome produces proteins or other biological products

There are four core 'omics discriplines, and many other subdisciplines

Genome	→	Genomics
Metagenome	→	Metagenomics
Proteome	→	Proteomics
Transcriptome	→	Transcriptomics

Since the year 2000, there has been an explosion of biological data

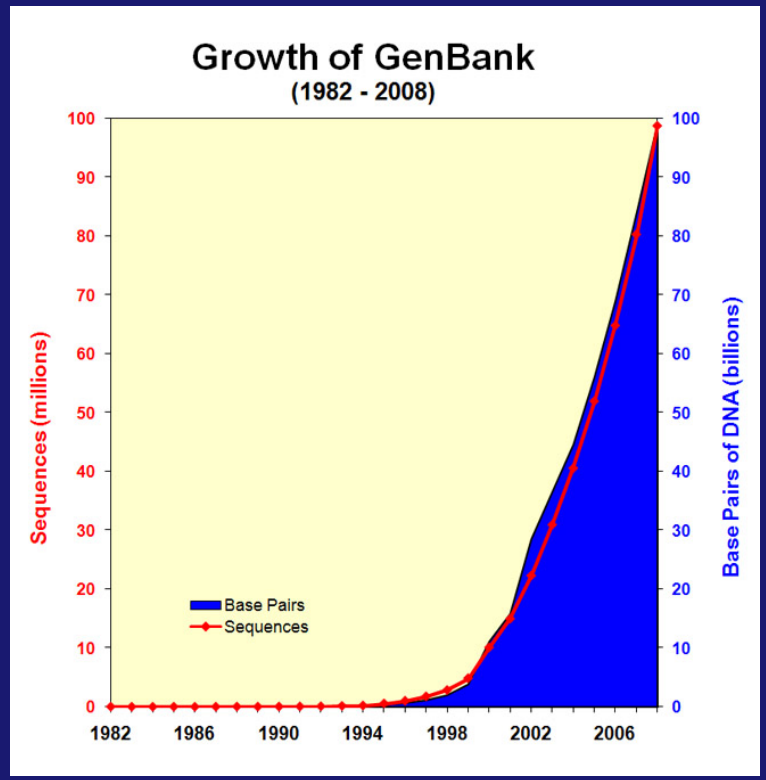
Sequenced genomes = 5,062

Base pairs = 99,116,431,942
= 396 GB

Sequences = 98,868,465

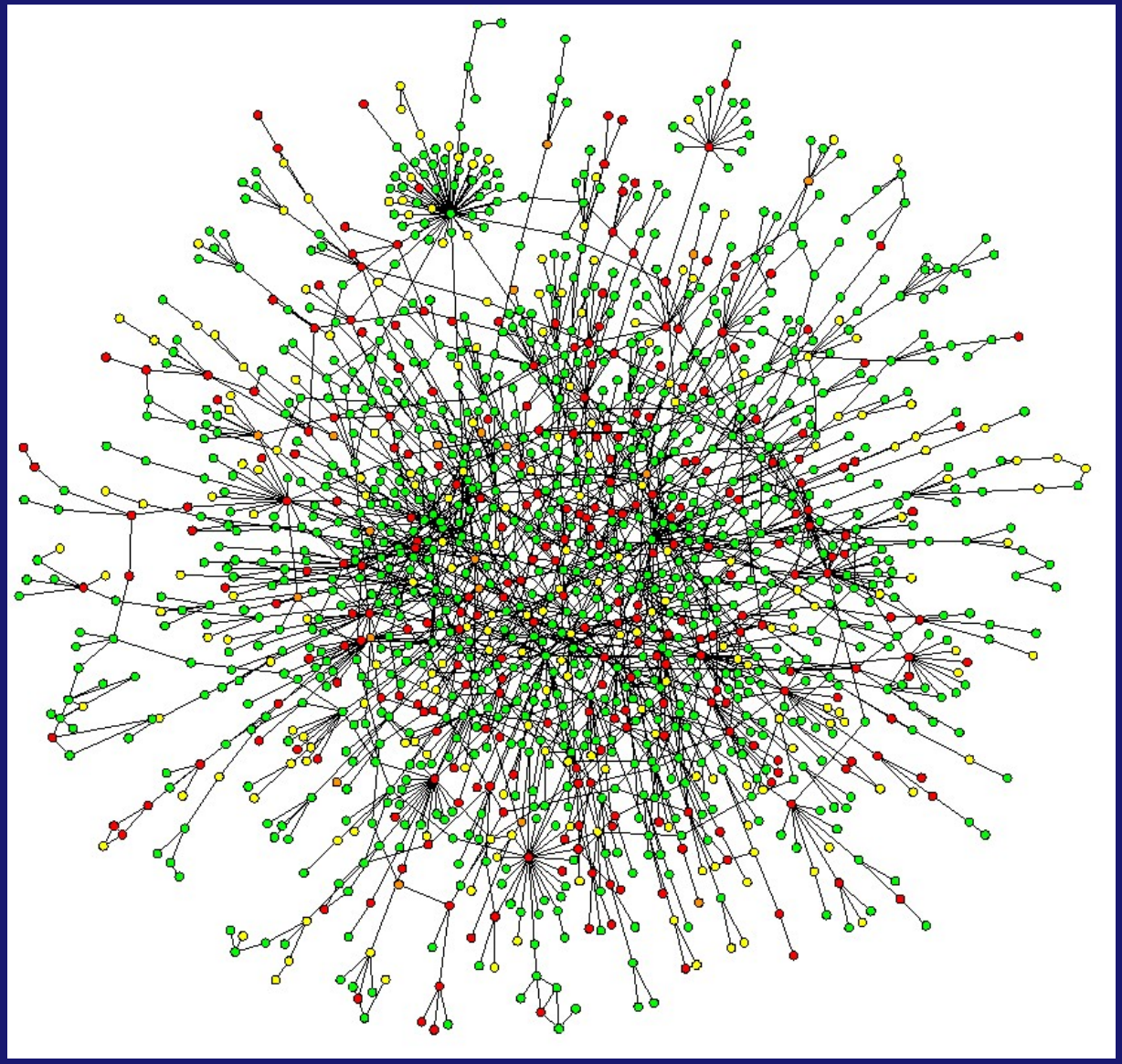
Known genes = 7,095,197

Known proteins = 523,151

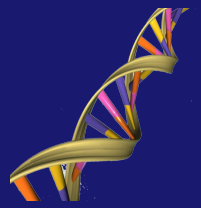


This is only a small portion of 'omics data that has been generated in the past 10 years

For even a single organism, there are thousands of protein interactions



We need Bioinformatics to make sense of the bewildering amount of 'omics data

Bioinformatics =  + 

Computers 'Omics

Acquire → **Store** → **Analyze** → **Visualize**

Experimental techniques for studying generating 'omics data

Computational techniques for storing, organizing and sharing 'omics data

Computational techniques for finding patterns in and making sense of 'omics data

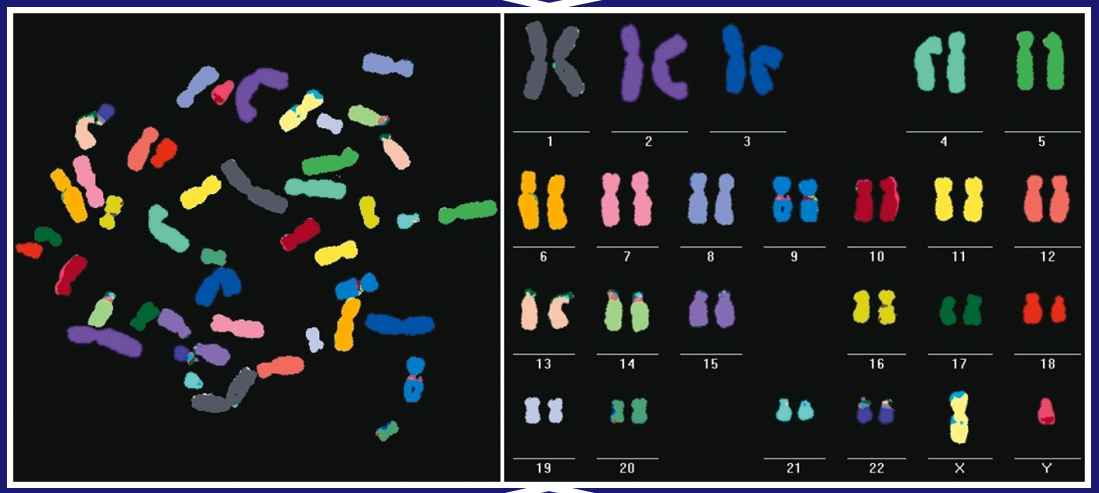
Computational techniques for visualizing 'omics data; pictures are easier to understand than numbers

End of Act 2: Questions

The human genome is fairly large, but the mouse genome is even larger

The human genome is stored in 23 pairs of chromosomes

There are an estimated 20 – 25,000 protein-coding genes in our genome



3 billion base pairs in our DNA; 3 gigabytes of storage space

There are approximately 2 million proteins in our proteome

It would take a person a 100 years to recite the human genome, saying one nucleotide per second, 24 hours a day

Human Genome Project: sequence all of our DNA and identify all of our genes

Aside from answering basic biological questions, why is this important?

Answer: every disease has a genetic component

- Inherited
- Response to environment

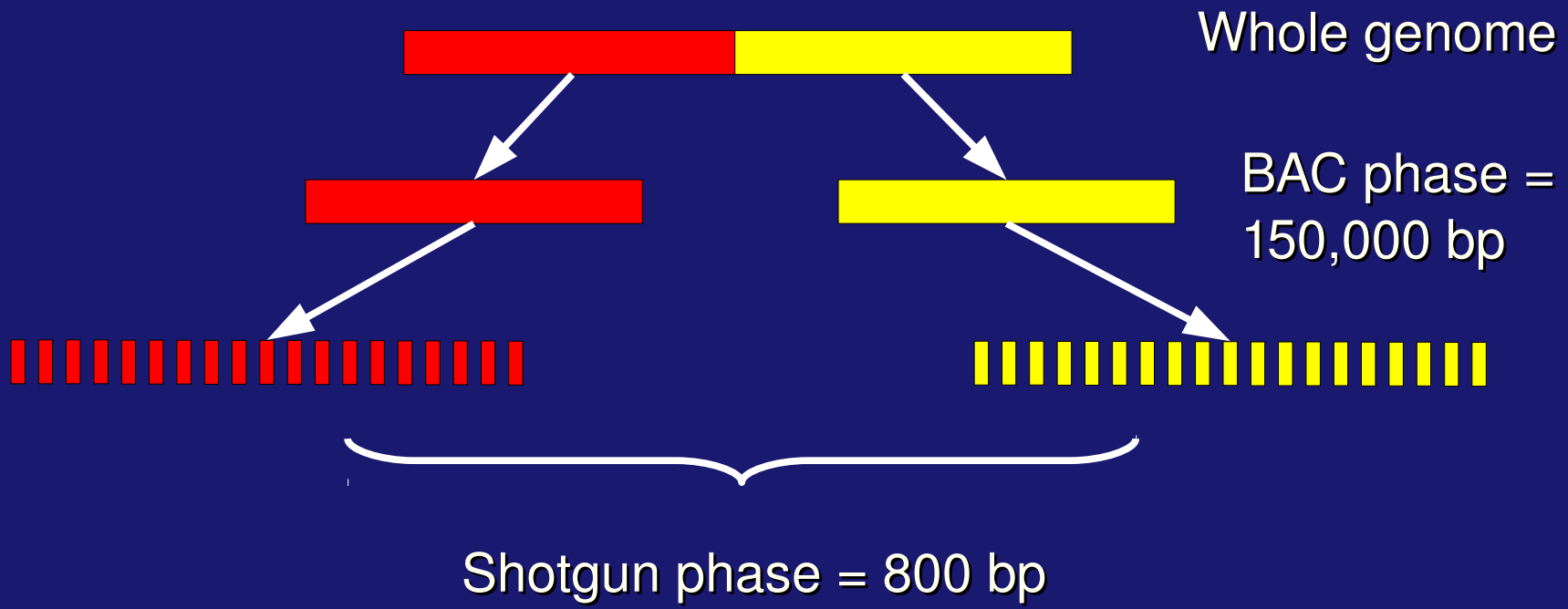
The draft sequence of the human genome was published in 2000; the complete sequence was published in 2003

Bioinformatics was, and continues to be, an essential part of the Human Genome Project. How did we do it?

Acquire: determine the sequence of the human genome

It wasn't possible to sequence the whole genome at one time

We had to break it up, sequence smaller pieces and then put put it back together



Bioinformatics was necessary to assemble the genome sequence

Bioinformatics challenge → paste together the small shotgun sequences to get the entire human genome sequence

ACCTTGGCCTAGGCT

GCTGGAATCCAGTGCC

TAGGCTACTGGCTGGA

TAACTAGCTTAATCCG

GTGCCCGGGTTAACTA

Required the use of sophisticated assembly algorithms that ran on supercomputers

Sequence assembly game

Store: organize and share the sequence of the human genome

Problem: we have too much data to store in a notebook and to search manually

Bioinformatics challenge	Solution
Develop efficient methods to store, organize and access the 'omics data	Created databases; The major data repositories are GenBank, EMBL and DDBJ
Develop easy ways for experimental biologists to access the data	Created web-based applications to access the data

A substantial portion of biological data is freely available to anyone with a web browser

Analyze: Search the genome sequence for biological features

Bioinformatics challenge	Solution
Identify genes	Developed gene finding algorithms (i.e. spot the gene)
Identify genes that are similar to an unknown one	Developed sequence alignment algorithms, such as BLAST
Find regulatory sequences	Developed pattern identification algorithms to find out how those genes are regulated
Store, organize and share these feature annotations	Developed databases to store information about genes

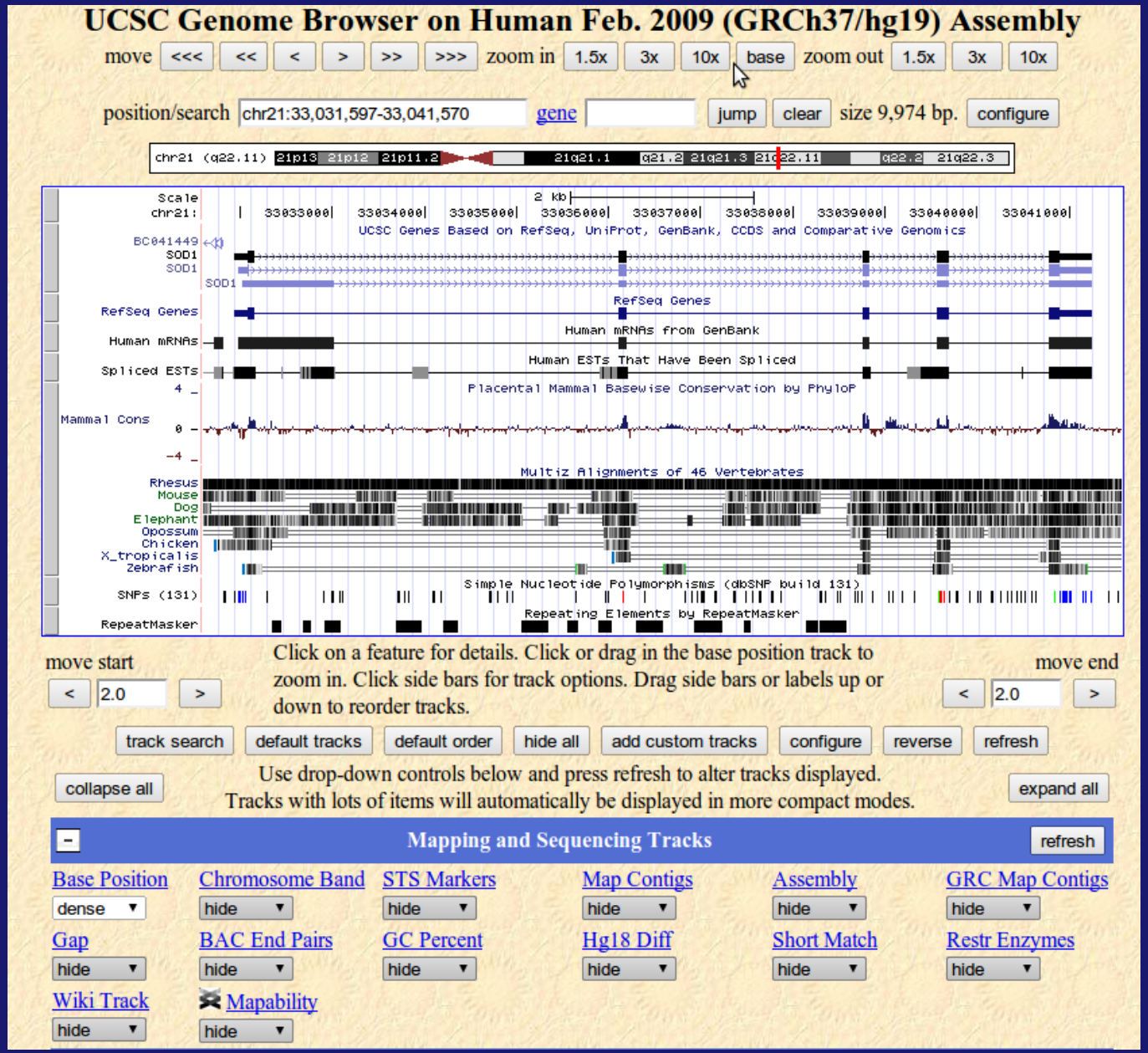
Analyze: OK, we know the genes now. What do they make?

The draft sequence of the human genome was published in 2000; the complete sequence was published in 2003

Analyze: When, how and where are the genes expressed?

The draft sequence of the human genome was published in 2000; the complete sequence was published in 2003

Visualize: We need tools to visualize the information in our genome.



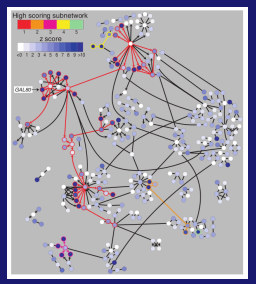
Network screening can uncover novel, context-dependent interactions

Developed a method to score (rate) a subnetwork based upon gene differential expression

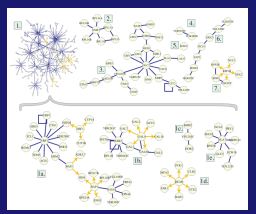
NC-17

Developed a method to find high-scoring subnetworks

Identified active subnetworks in a subset of the yeast interaction network when GAL80 was deleted



Identified active subnetworks in the full yeast interaction network when perturbed twenty different ways.



Questions?