DNA BARCODING

EVERYTHING YOU NEED TO KNOW

WHAT IS IT?

HOW DOES IT WORK?

WHY IS IT IMPORTANT?

DNA BARCODING

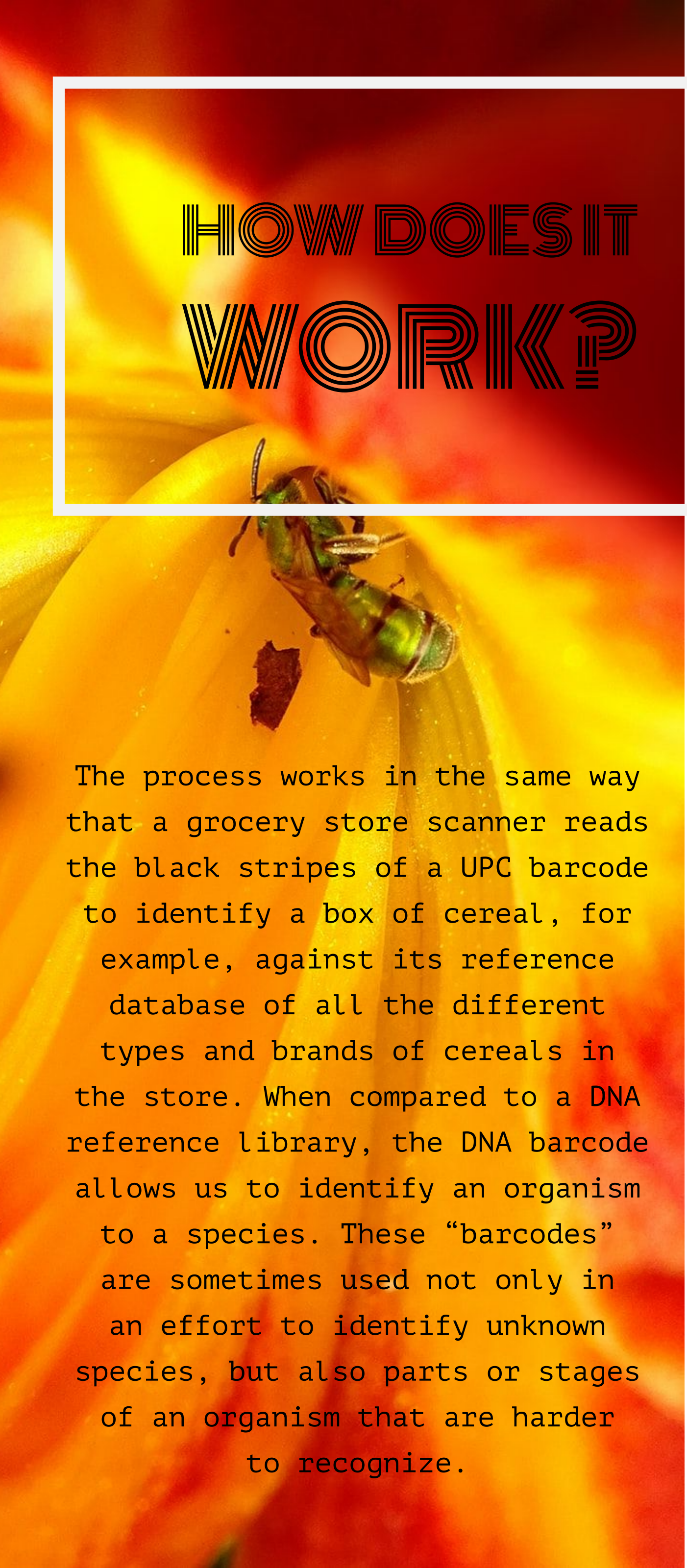
IN THE 250 YEARS THAT SCIENTISTS HAVE BEEN DESCRIBING LIFE ON EARTH, TWO MILLION SPECIES HAVE BEEN REVEALED. YET, IT IS ESTIMATED THAT OUR SPECIES IS ONE AMONG 20 MILLION. MOST SPECIES STILL AWAIT DISCOVERY!

Introduction

Species are a vital part of biodiversity, the variation of life on Earth. They provide food, medicines, and clean our water systems. Identifying new species isn't always easy - especially when examining smaller groups like insects. One way scientists have been able to identify species is by reading DNA. DNA is the molecule key to life, and it carries the instructions for its functioning, growth and reproduction. All this information is written in sequences of four 'letters' called nucleotides - **A,C,T,G.** It is the huge number of possible combinations of these four letters that is at the foundation of all the biodiversity of this planet.

WHATISIT?

DNA barcoding is a method used to identify species using a short, specific section of DNA known as the DNA barcode region.
For example, if you caught a butterfly in your butterfly net, you could find out if it is a monarch butterfly or a viceroy or a painted lady by using DNA barcoding.



DNA BARCODING APPLICATIONS

INCLUDE IDENTIFYING PLANTS EVEN WHEN FLOWERS OR FRUITS ARE NOT AVAILABLE, IDENTIFYING THE POLLEN OF THE BODIES OF POLLINATING ANIMALS, IDENTIFYING INSECT LARVAE THAT HAVE FEWER RECOGNIZABLE CHARACTERS THAN ADULTS, OR INVESTIGATING THE DIET OF AN ANIMAL USING ITS STOMACH CONTENT OR FECES.

The idea for DNA barcodes was conceived by a Canadian evolutionary biologist and professor, Paul Hebert. Once he and his colleagues at the University of Guelph developed the DNA barcoding method, they then began work to create an entire system dedicated to cataloguing life.

BARCODING BASICS

The workflow begins with the collection of individual specimens in the environment, then the sorting of these individuals based on defining physically characteristics in the lab, before extracting DNA without destroying the organism. From this DNA, the barcode region can be targeted using a machine about the size of a rice cooker. This machine “cooks” the DNA, increasing and decreasing its temperature in a process called Polymerase chain reaction (PCR) which eventually make many copies of the specific target region. Then the PCR product - the multiple copies of the barcode region - is sequenced, meaning the unique position of each nucleotide that makes up the DNA barcode is identified. Then the barcode is compared against a reference database. When each unique position along the barcode region matches to the unique position along the barcode region of an identified species in the library, we have an ID for the unknown organism!

The match doesn’t need to be perfect. The barcode region does have some variation amongst the individuals of a particular species. But it works because the particular gene regions chosen as the barcode has less difference within species than between species. This is referred to as the barcode gap.

BARCODINGBASICS

TECHNOLOGICAL TRANSFORMATION

DNA barcoding began by examining single specimens to obtain a sequence record for each individual. As technology advanced, DNA metabarcoding was born, enabling us to examine samples in bulk - thousands of specimens together in a single sample - and resulting in a set of sequences linked to a particular sample. While DNA barcoding is ideal for the construction of DNA barcode reference library, it is more expensive per specimen than DNA metabarcoding which is now being used to cheaply detect whether a species can be found in a particular area, crucial information for bio-monitoring efforts.

BARCODINGBASICS

The long & short of DNA barcoding

DNA sequencing instruments use a variety of read lengths - the number of base pairs (bp) sequenced from a DNA fragment. Current platforms fall into two categories: short-read and long-read instruments.

The short-read platforms give us sequences with less than 500 bp each, but they can produce up to 500 million reads per run. These are best for DNA metabarcoding because they produce such a high read count.

The long-read platforms give us only 3 million reads per run, but each read can be greater than 50,000 bp in length. As the barcode region is quite short, this might seem unimportant for DNA barcoding. But by taking advantage of a process called Circular Consensus Sequencing, we get reads with incredible reliability.

HOW?

Each PCR product is circularized to allow the sequencer to read the barcode region multiple times, correcting errors each time. This process is fast, reliable and very inexpensive per specimen. The SEQUEL 1 machine can run 9,216 specimens at a time while the new SEQUEL II platform can analyze 73,000 specimens in a run. If used for 250 days a year, it could process 40 million georeferenced, time-stamped specimen records, all assigned to a species each year at 1 CAD dollar apiece

NEW SEQUENCING TECHNOLOGY IS GIVING US A LOT OF DATA! AND SPECIALIZED INFORMATICS SUPPORT IS ESSENTIAL TO CURATE AND ANALYZE RESULTS. CUE, THE BOLD AND MBRAVE PLATFORMS.

BARCODINGBASICS

DATA, DATA, DATA!

BOLD stores barcode records and also has tools for data validation and analysis. Among them, the Barcode Index Number system is the most important. It is based on the barcode gap - the fact that members of a species usually have little variation in the barcode region while those of different species have a lot more. The BIN system uses algorithms to separate sequence groups and then assigns a unique identification tag to each group thus automating species discovery and naming.

There are now 656,000+ BINs on BOLD, each assembling information on the distribution, morphology, and taxonomy of a species. This type of a system is truly revolutionary - it provides a fast species count for any taxonomic group in any setting (simply collect specimens, sequence them, and BOLD assigns them to a BIN), and it can tell you the species shared among sites.

WANT TO FIND OUT HOW MUCH YOUR CITY HAS IN COMMON WITH COSTA RICA? YOU CAN GO TO BOLD RIGHT NOW AND GET AN ANSWER WITH ONLY A FEW CLICKS OF YOUR MOUSE.

MBRAVE IS THE NEW PLATFORM WHICH WAS DEVELOPED TO AID THE ANALYSIS OF DATA STREAMS GENERATED BY HIGH-THROUGHPUT SEQUENCING PLATFORMS. IT IS A CLOUD-BASED PLATFORM TO DEPOSIT, ANALYZE, AND VISUALIZE METABARCODING DATA.

There is a special urgency to our work that will never be experienced by other disciplines – the subjects of our science are disappearing.

Paul Hebert

WHY IS IT IMPORTANT?

The International Barcode of Life Consortium involves research organizations in 31 nations that have joined forces to complete three tasks before mid-century. Register all species, document their interactions, and track their population trends.

HUMAN ACTIVITIES ARE THREATENING THE PERSISTENCE OF MANY SPECIES.

UNTIL NOW, BIODIVERSITY SCIENCE WAS ILL-PREPARED TO ADDRESS THIS CHALLENGE. MOST SPECIES ARE UNKNOWN, THEIR INTERACTIONS ARE UNRECOGNIZED, AND OUR KNOWLEDGE OF THEIR DISTRIBUTION AND ABUNDANCE IS VERY INCOMPLETE.

The International Barcode of Life Consortium is helping to lay the foundation for an earth observing system for species through large-scale DNA analyses. It has just launched its second major research program, BIOSCAN, to progress this work.