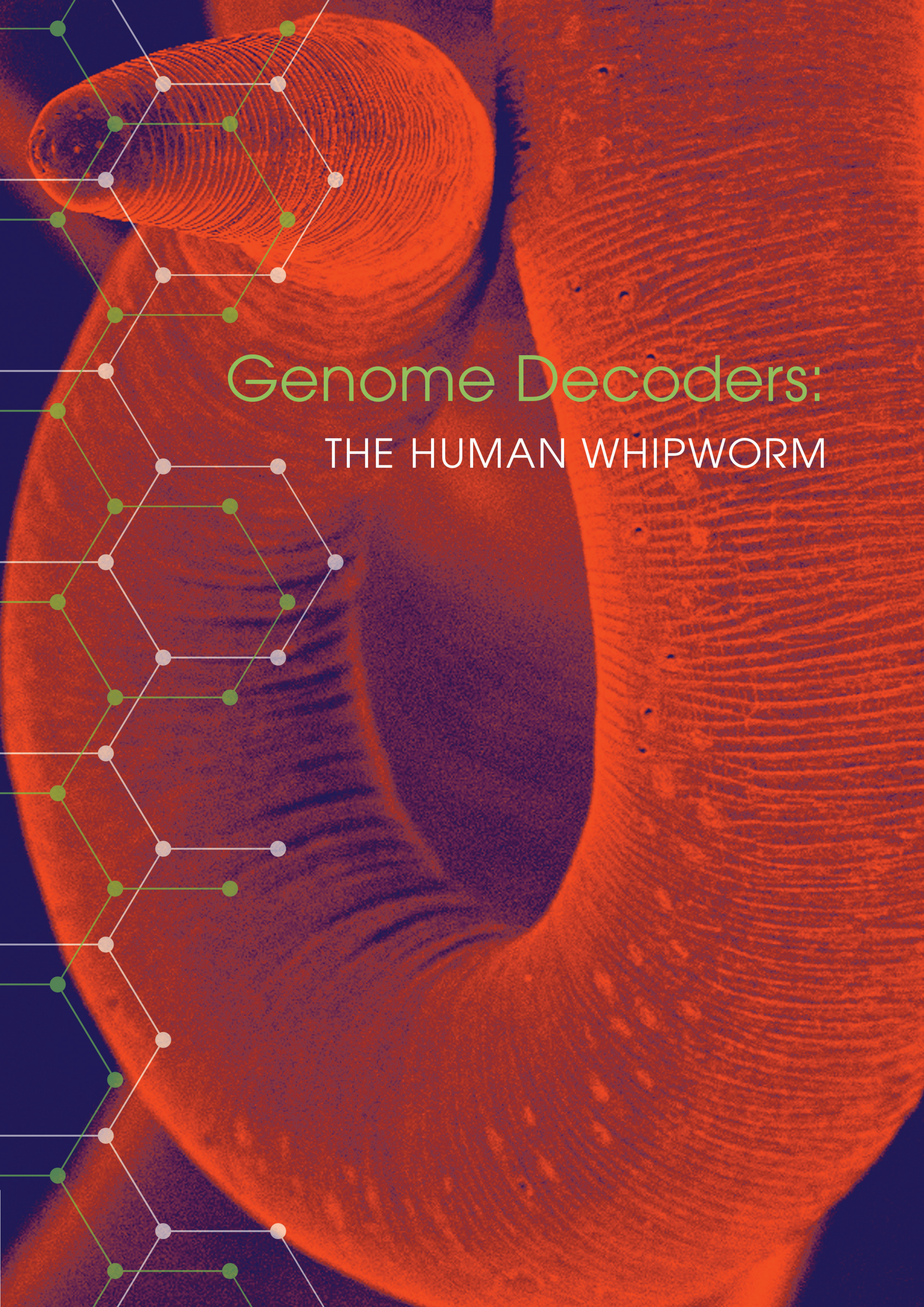
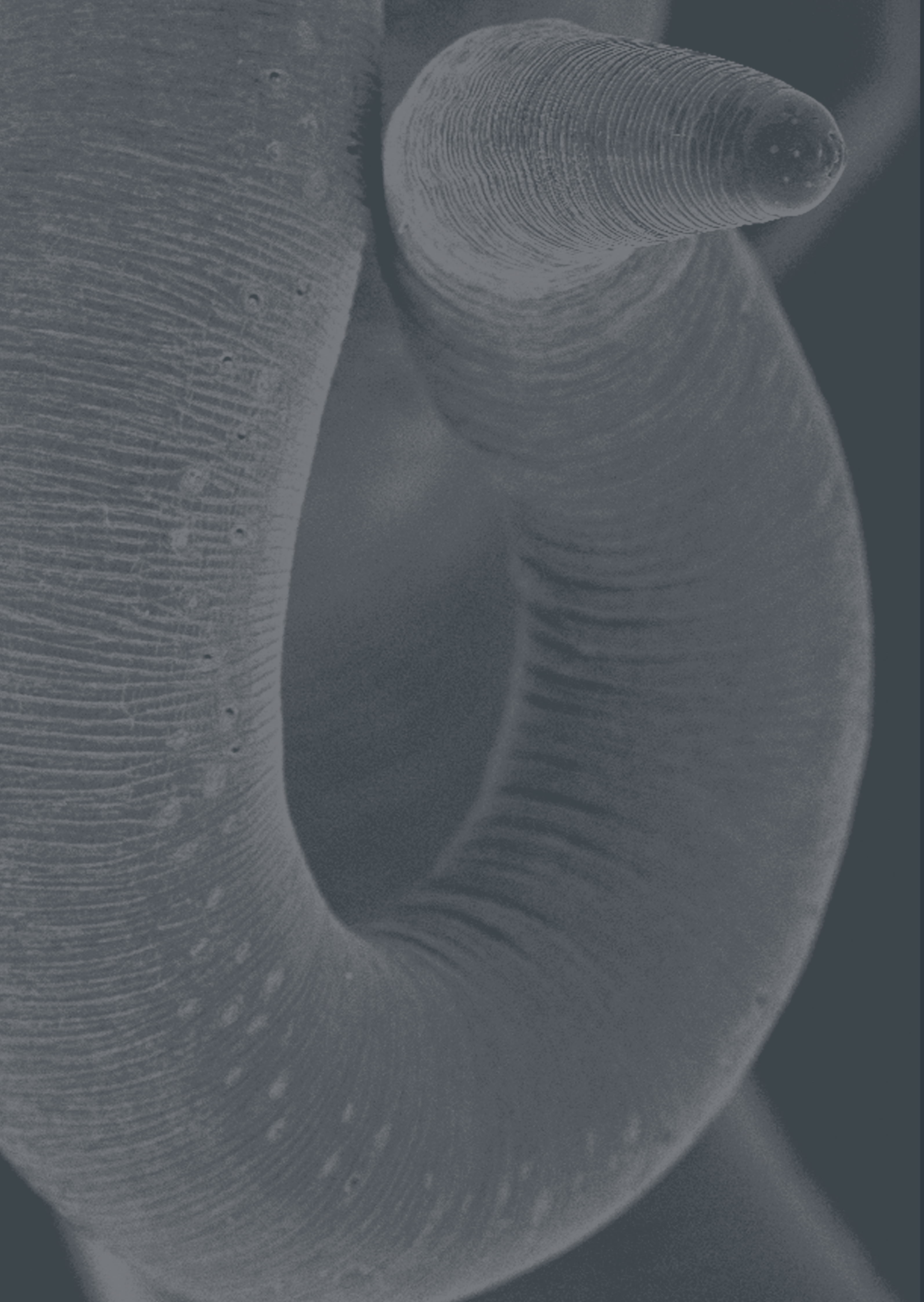


Genome Decoders:

THE HUMAN WHIPWORM



Contents

WELLCOME GENOME CAMPUS PUBLIC ENGAGEMENT

Introduction

The Institute for Research in Schools (IRIS) offers teachers and students the opportunity to contribute to original scientific research. We are absolutely delighted to be partnering with the Wellcome Sanger Institute and European Bioinformatics Institute for this genomics project.

The project scientists will work with schools to annotate the genome of a whipworm. Not only does this support the students' learning but will contribute to improving the understanding of the whipworm and potentially open up avenues for drugs and vaccines that target it.

We can't wait to see the results of the students' work.

Becky Parker

Professor Becky Parker
Director,
Institute for Research in Schools

Hundreds of millions of children and adults around the world have to put up with worm infestations for much of their lives. For many, this means living with miserable health problems, struggling to attend school or being unable to work.

Dealing with worms is a big problem. First, there are no vaccines and very few drugs. This could be a disaster waiting to happen because over-using individual drugs drives drug-resistance. Second, many of the drugs that we do have work well on some species but not others.

To hunt for new ways to control worms, we really need to understand them. And in modern biology, the best way to know your enemy is to understand its genome. Clues to every facet of an organism's biology, every future drug target or vaccine candidate are encoded somewhere in its genome.

Genomes are "sequenced" from random fragments, so assembling the 80 million-base genome of the human whipworm, in the correct order, was a jigsaw that took several years. But that work is largely done. Now we need to figure out where, along this long string of bases, are the genes, and what do they do? That's where you can really help.

Gene structures in animals can be complicated – there are all kind of things like exons and introns, untranslated regions and alternative splicing to deal with. The good news is that, when viewed across the landscape of a genome, this abstract sounding stuff will really start to make a lot of sense. Of course there are thousands of genes to find but computer programs do a good job of finding the rough positions of genes. What they are less good at, and people are great at, is finding the precise boundaries of all the bits that make up a gene. This requires sifting through lots of different predictions and examining visual evidence. It requires getting a sense for what is normal and then spotting the unusual. With lots of volunteers looking at the genome, we should be able to find every gene and find plenty of new clues to what makes human whipworms different to other animals, and maybe some new vulnerabilities that can be attacked.

I've always found it exciting to be the first person to see the basic building blocks of an organism. For the human whipworm, we haven't yet looked, so you'll be the first. Can we really work as a team to analyse every gene? I'm really excited to see what you'll uncover.

Dr Matt Berriman
Senior Group Leader
Wellcome Sanger Institute

Project overview

Trichuriasis is a major Neglected Tropical Disease (NTD) affecting millions of children in Asia, Africa and South America. The disease is the result of infection by the parasitic worm *Trichuris trichiura,* also known as the human whipworm. High parasite burden has been linked to malnutrition as well as physical and cognitive developmental problems, which can have chronic socio-economic consequences.

Genome sequencing of parasitic worms, such as the whipworm, can help in the fight against NTDs. Annotating the genome enables the identification of the genes that code for proteins important during infection and disease. This understanding can aid the development of new treatments and vaccines.

Working with scientists from the Wellcome Sanger Institute and EMBL-European Bioinformatics Institute, students will have the opportunity to contribute to the annotation of the human whipworm genome. They will learn the fundamentals of genome annotation and curation and will work together to annotate different regions of the whipworm genome, identify genes.

Curriculum links

This project can link to various aspects of key stage 5 and senior phase biology curriculum, including the following areas:

- DNA and the genome.
- Proteins.
- Parasitism.

Helminths

What are Helminths?

Helminths are parasitic worms that survive by feeding on a living host to gain nourishment and protection, sometimes resulting in illness of the host. There are a variety of different helminths from the very large to the microscopic and they are classified in two major groups:

1. Platyhelminths (also known as flatworms)
2. Nematodes (also known as roundworms)

The image above shows *Schistosoma mansoni*, a type of platyhelminth, which causes the disease Schistosomiasis.
Image: David Goulding, Wellcome Sanger Institute

Helminths infect a range of hosts such as mammals (including our pets and livestock), amphibians, reptiles, fish, birds, invertebrates and humans.

Their effects inside their host vary, causing a wide spectrum of diseases from mild to potentially deadly. Helminths are one of the leading causes of morbidity in the developing world, with over two billion people affected – almost a third of the world's population.

Neglected Tropical Diseases

The diseases caused by helminths are catalogued as Neglected Tropical Diseases (NTDs). They are designated as 'neglected' because they generally receive less attention, treatment and research funding than other infectious diseases such as HIV/AIDS, malaria and tuberculosis.

Helminth infections are generally characterised by long term chronic illness, although acute complications can be deadly. Infections can retard mental and physical development and impede cognitive development and education.

By impairing agricultural and economic productivity, helminth infections have a detrimental impact on socio-economic development. This then completes a vicious cycle, whereby poverty and poor sanitation produce ideal conditions for the transmission of the disease.

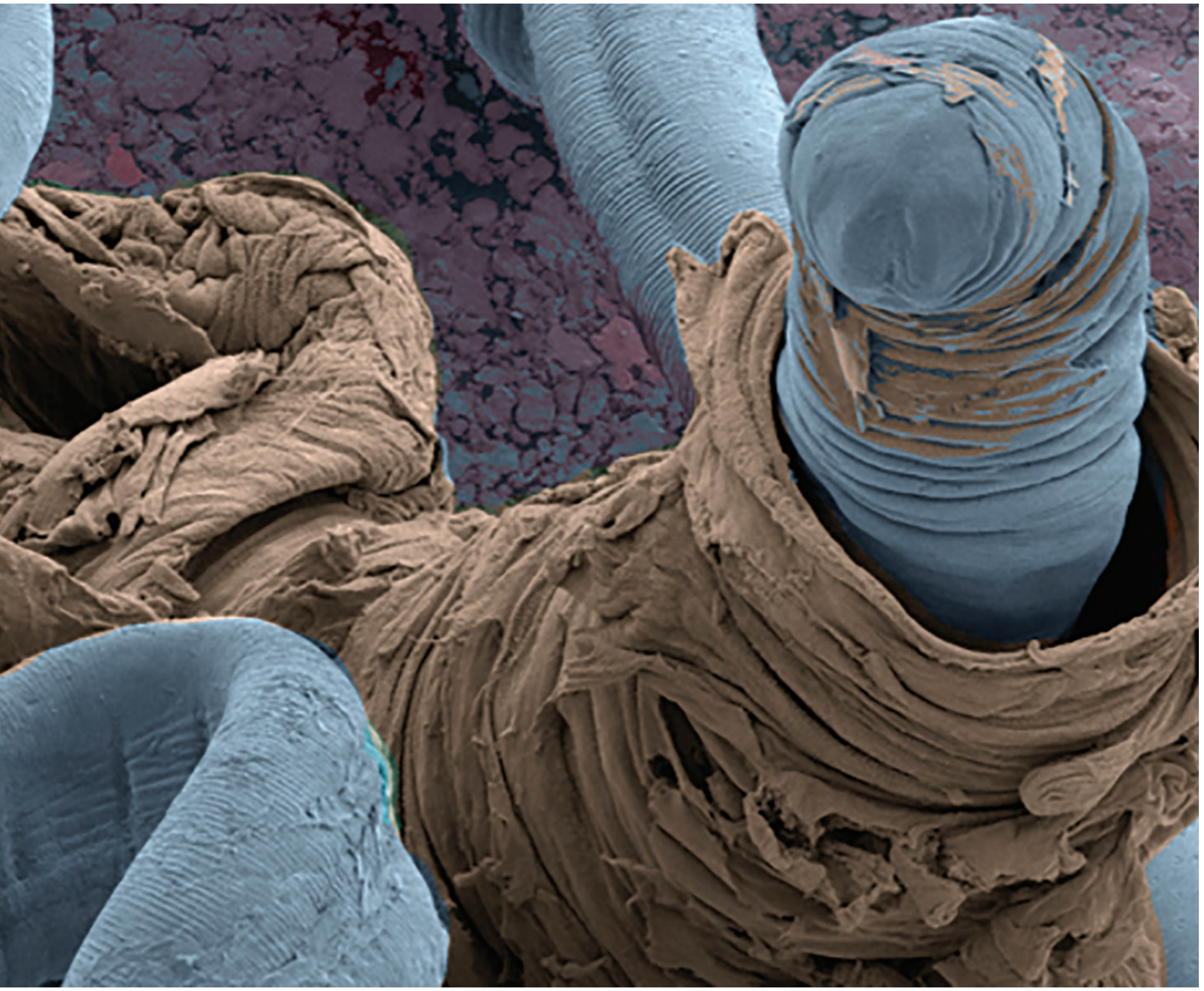
The most common NTDs worldwide are the infections caused by intestinal parasitic nematodes, such as whipworms, roundworms and hookworms. Infections with intestinal nematodes are widely distributed, occurring mainly in tropical and subtropical regions. These diseases are characterised by abdominal pain, gut inflammation, diarrhoea and anaemia. They are linked to physical, nutritional and cognitive impairment in young, developing children.

What is a Whipworm?

Whipworms are a type of intestinal parasitic nematode. This group also includes threadworms and hookworms.

There are several species of whipworm, each having evolved to infect a specific host. The human whipworm, *Trichuris trichuria*, infects humans and causes the disease trichuriasis.

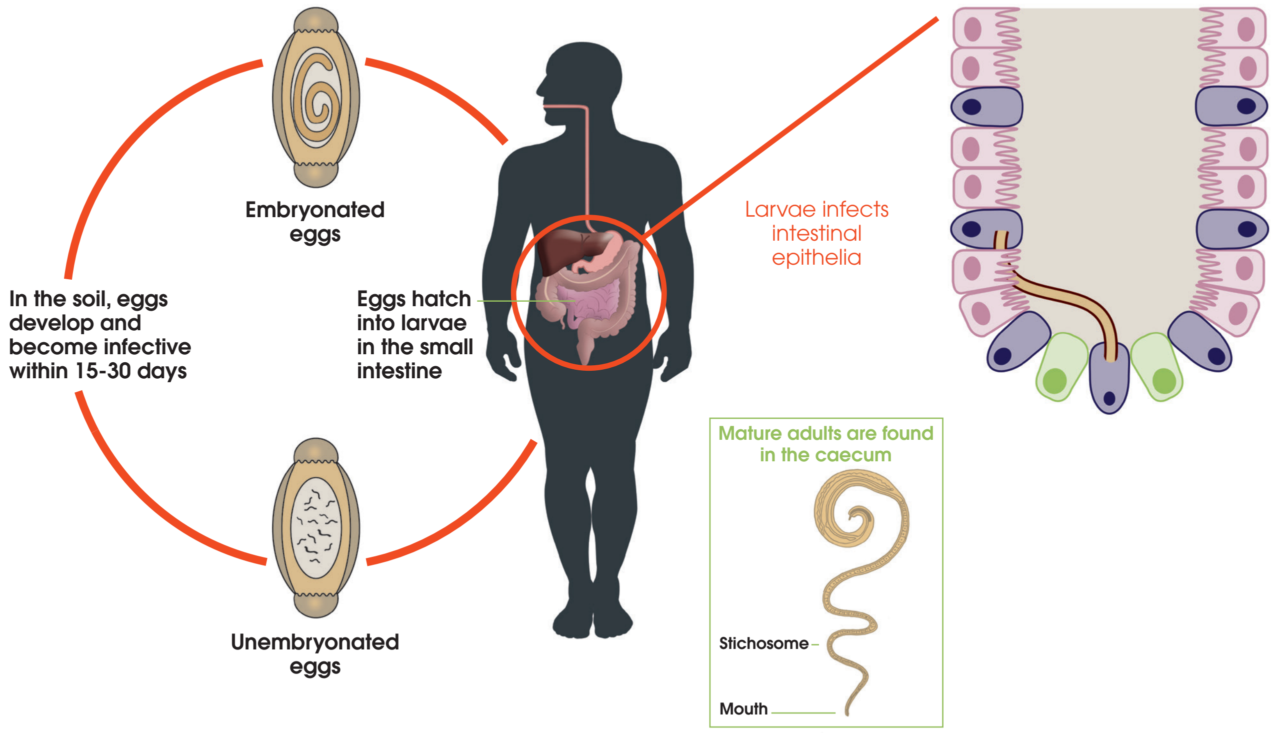
Image: David Goulding, Wellcome Sanger Institute



Whipworm lifecycle

- Adult whipworms live in the caecum at the top of the large intestine where they produce thousands of microscopic un-embryonated eggs each day. At this stage the eggs are not infective.
- The eggs are passed in the faeces of infected people. In areas where there is a lack of adequate toilet sanitation the eggs pass into the soil contaminating it and the surrounding environment.
- After 15-30 days in the environment, a tiny parasitic larvae develops inside the egg, causing it to become infective (embryonated).
- Infection with whipworms occurs upon ingestion of whipworm eggs present in food or water.
- Upon arrival to the caecum the eggs hatch and the liberated larvae burrow through the intestinal epithelia.
- In the epithelia the larvae develop until they become adults and the cycle begins again as these adults produce eggs.

Trichuriasis: Life cycle of *Trichuris trichuria*


Embryonated eggs
In the soil, eggs develop and become infective within 15-30 days
Eggs hatch into larvae in the small intestine
Unembryonated eggs
Larvae infects intestinal epithelia
Mature adults are found in the caecum
Stichosome
Mouth

How do we study whipworms?

Studying human infections is not feasible because the human whipworm cannot be maintained in the lab. To study the process of infection and the disease caused by the parasite in the host, a mouse model of infection is used: the mouse whipworm, *Trichuris muris*.

By studying the human whipworm genome it is possible to learn about its biology with a view to finding ways to tackle the disease. Knowing the genetic makeup of the parasite will enable researchers to search for chinks in its armour. These vulnerabilities may make it susceptible to drugs or potential vaccine candidates.

The genetic makeup can be determined by sequencing and annotating the genome. Genomics is explained in-depth in the following section, but some key points include:

- A genome contains the DNA instructions used to make an organism.
- The sequence of the genome is the order of the DNA nucleotides: adenine, thymine, cytosine and guanine (A, T, C, G).
- The sequence is very long indeed - millions of nucleotides - making this joint effort to help with its analysis all the more vital.
- Annotation of the genome involves searching this sequence data to identify genes, the regions of the genome that encode for proteins.

A draft genome of the human whipworm was produced in 2014, however it was fairly fragmented. Due to this fragmentation research scientists have had to rely on the mouse whipworm genome as a reference because it is of higher quality than the human whipworm.

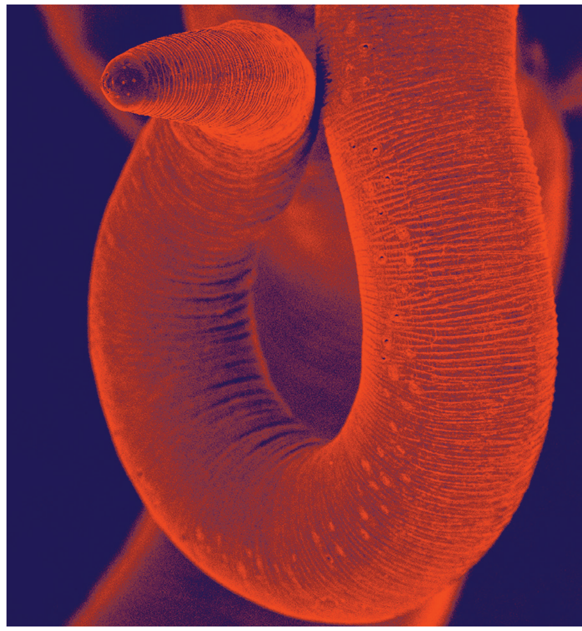
Thanks to the acquisition of new samples, high quality genome data is now available for the human whipworm but this needs to be curated.

Worm therapy

There is currently research underway looking into the role helminths, in particular the pig whipworm (*Trichuris suis*), can play as a therapy for auto-immune diseases such as inflammatory bowel disease, psoriasis and asthma.

Auto-immune diseases have become more common in industrialised nations over recent decades and it is proposed that in increasingly sterile lives, the lack of contact with micro-organisms and parasites could be driving this trend, leading to more unbalanced immune systems.

Deliberate infection with helminths is being investigated as a possible therapy. A patient is given a dose of parasite eggs, which hatch and colonise the gut. Clinical trials are underway in some countries where it has been associated with reducing inflammation caused by auto-immune diseases.



Where did the samples of the human whipworm come from?

Human whipworm research is not possible without a supply of good quality, human adult whipworms. Thanks to a bold endeavour by researcher Peter Nejsum from Aarhus University in Denmark, we now have access to really high quality material. Peter infected himself with whipworms in order to treat his psoriasis. He ingested six hundred eggs, so he now has a living supply inside him. This means when new adult worms are needed, he has access to them.

Genomics

What is a genome?

A genome is an organism's complete set of genetic material. It is organised into chromosomes. Each genome contains all of the information needed to build the organism and allow it to develop and function. All living things have a unique genome. Closely related organisms have similar genomes.

How are genomes studied?

In order to study an organism's genome, scientists try to reconstruct its sequence and structure as completely and accurately as possible. To do this they first need to get a good quality sample of the organism's DNA. Some organisms, such as parasitic worms, are difficult to isolate so this can be quite challenging.

- The chromosomes first have to be fragmented into sections short enough to be sequenced. The chemistry of sequencing is constantly improving, currently the most cutting edge technology is able to sequence a stretch of around ten thousand base pairs.

- Each sequence that comes out of the sequencing machine is referred to as the sequencing read. Given the size of the *Trichuris trichiura* genome is around eighty million base pairs, putting these sequences into the correct order is quite a computational challenge.

- To enable reconstruction of the genome from the sequencing reads, each region of the genome is covered by multiple reads.

- The number of times a region of the genome is sequenced is known as the sequencing depth. This means there are many short overlapping sequences. This allows the order they should go in to be determined, a process known as assembly.

A genome assembly represents the best efforts to reconstruct the genome as it appears in the organism's cells. The *Trichuris trichiura* genome has three chromosomes. In the current genome assembly, these are represented by 113 stretches of sequence (known as scaffolds).

What is a gene?

A gene is a small section of DNA containing the instructions for making a specific molecule, usually a protein. A human has approximately twenty thousand protein-coding genes. The whipworm is currently estimated to have between eleven and fifteen thousand. The process of finding the genes in the genome and predicting which proteins they encode is known as annotation.

Proteins

The central dogma of molecular biology tells us genes are transcribed into RNA and RNA is translated into proteins. Proteins are essential to the functioning of our cells. The structure and function of the human body depends on proteins and the regulation of the body's cells, tissues and organs cannot exist without them. Proteins perform essential functions, for example:

- Enzymes act like "factories" for chemical reactions.
- Signalling proteins tell the cells where they are and what's happening in their environment.
- Structural proteins provide cellular scaffolds in tissues such as cartilage, hair and nails.
- Mechanical proteins, such as actin and myosin, help to make muscles move.

When a researcher is studying a biological process within an organism this often comes down to studying how a protein is functioning, which other proteins it is interacting with, how its expression is regulated and so on.

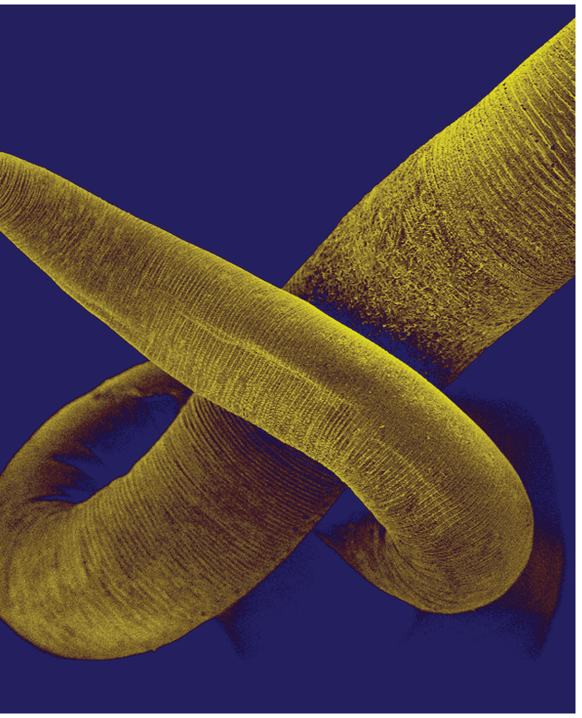
An essential starting point to understanding how an organism works is a complete and accurate catalogue of its proteins. Proteins themselves are difficult to study in a high throughput manner for many reasons, including:

- They are not all expressed at the same time in the same cell types
- All proteins have different properties so they cannot be easily isolated
- They can't be amplified, which makes analysing proteins from limited tissue samples very challenging.

Therefore the best way to define a full catalogue of an organism's proteins is to sequence and annotate its genome.

Comparative Genomics

One very powerful approach to extracting information about an organism's biology from its genome is comparative genomics. This is based on the understanding that every species within the tree of life evolved from a common ancestor.



When two species first branch from one another their genomes will essentially have the same set of genes. As the two species diverge, they will each adapt to their different environments and be under different selective pressures. In the case of the whipworm, *Trichuris trichiura* adapted to infect humans, whilst *Trichuris muris* adapted to infect mice. It is possible to detect signatures of these different lifestyles in the genomes of the two species.

Genes can have very similar sequences. Gene sequences found in two species are presumed to have been present in their common ancestor and to have been maintained in both genomes after they speciated (evolved into different species). Such genes are said to be orthologues of one another. Orthologues are maintained because they encode proteins involved in biological processes or structures important for both species. Identifying *Trichuris trichiura* genes that have an orthologue in T*richuris muris* will be useful, as studying these genes in the mouse model of Trichuriasis will likely give results also applicable to the human disease.

Comparative genomics allows us to identify differences between the genomes of two species. Genes present in *Trichuris trichiura* but not in *Trichuris muris* might have evolved specifically in the human-infecting species because the proteins they encode are involved in helping the worm infect and survive within the human body. Identifying and studying these proteins could help find biological processes and structures to be targeted to kill the worm in the human body or to prevent it from establishing an infection.

How are genomes annotated?

The process of annotating a genome has two main steps:

1. Gene prediction algorithms are used to predict the position and structures of the genes in the genome (an algorithm is a computer program used to solve a problem).
2. Some or all of the gene models predicted by the algorithms are refined by humans in a process of manual curation.

Gene prediction algorithms use two main methods to find genes and predict their structures:

1. The genome sequence itself must support the presence of a gene. The algorithm will scan the genome for the presence of certain sequence motifs, such as a start codon, a stop codon, splice donor/acceptor pairs, etc (see gene structure section for more information on these). Regions where these features appear in the correct order, with appropriate spacing between them, indicate the presence of a gene.
2. In addition, most algorithms incorporate evidence into their predictions. In most cases this takes the form of RNA sequencing (RNAseq) data. RNAseq evidence helps in the resolution of more complex gene structures. For example, in regions where numerous potential splice sites are present in the genome sequence, RNAseq data will guide the algorithm towards the functional splice site in the organism.

Algorithms provide a good starting point for producing a complete gene set. However, in some cases they still cannot do as good a job as a human. This is especially true for complex regions, for example where the genome is very repetitive or the gene has a lot of alternative transcripts.

Alternative transcripts allow for increased complexity within an organism's genome as this means multiple proteins can be coded for from a single gene. Curation is the process of manually altering the gene model to make it the best it can be.

The human genome and a handful of model organism genomes such as the mouse and zebrafish have been curated intensively over several years by many researchers, to the extent that almost all of the genes have been found and reviewed. This is not often the case for genomes of parasitic worms. A well curated *Trichuris trichiura* gene set will be a great asset to researchers working on this parasite.

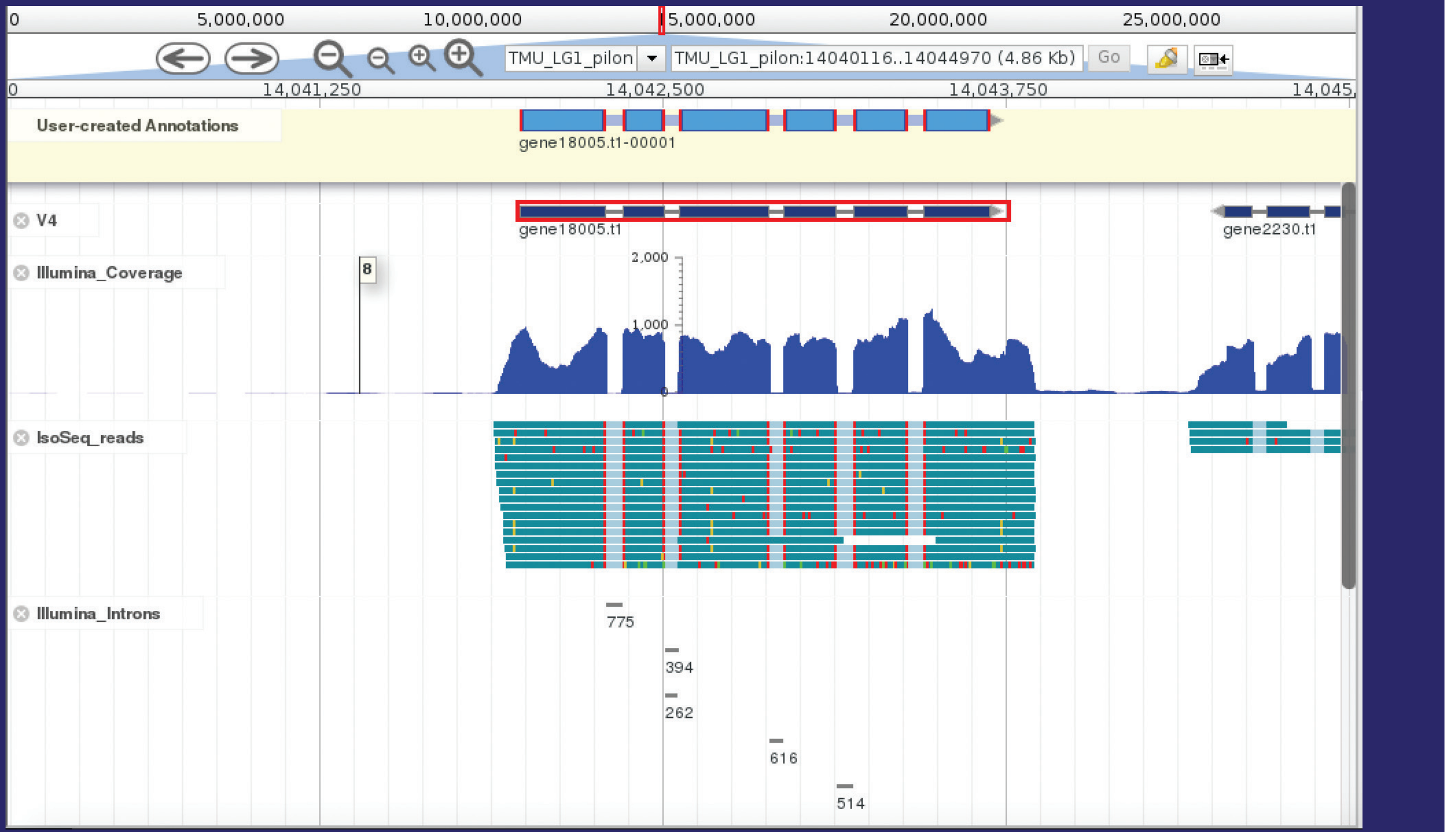
APOLLO:
gene curation software

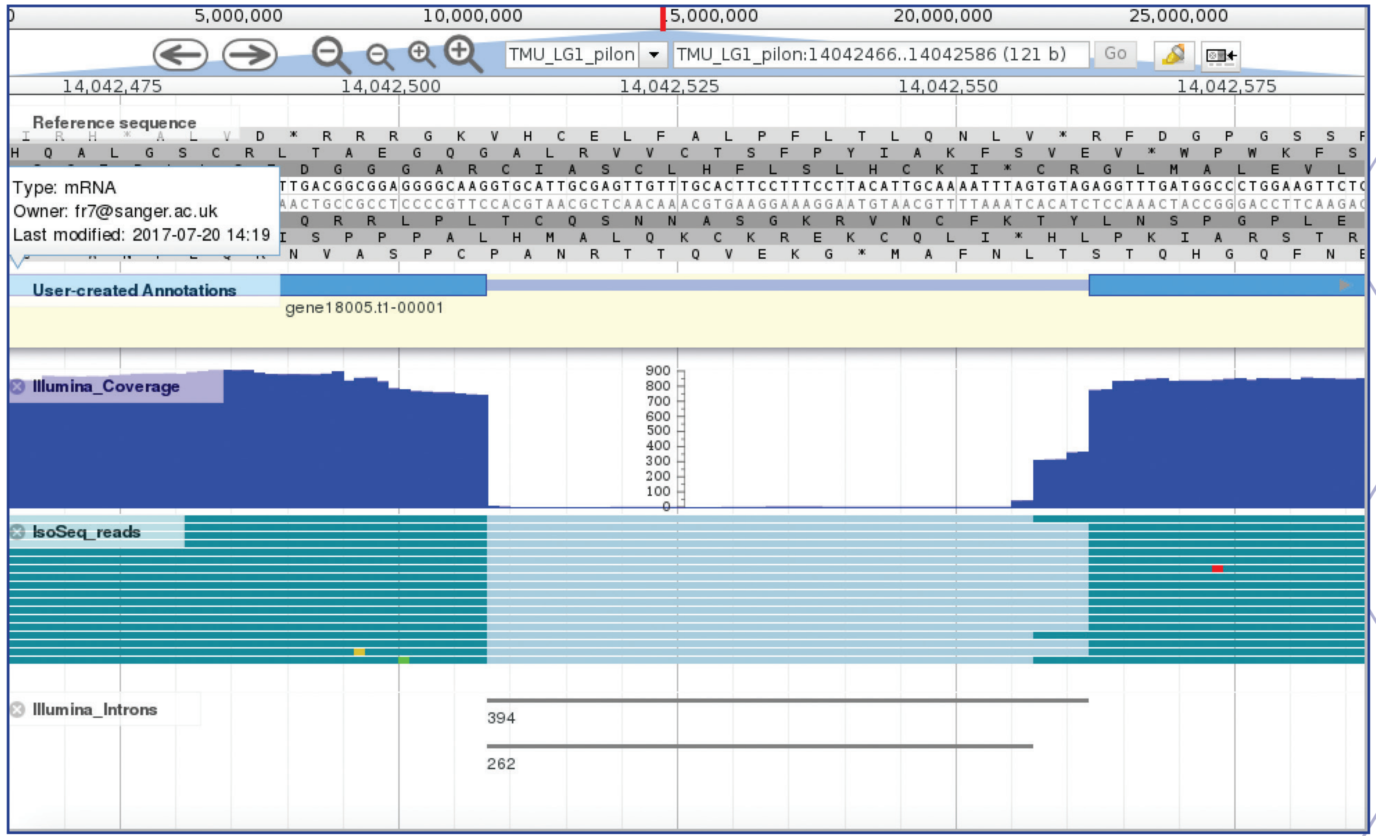
The manual curation is carried out using a piece of software called Apollo. This is accessible from a web browser. The details on how to access this will be provided separately to the participating schools by IRIS. A more detailed users guide for the software will also be supplied.

Apollo is based on a genome browser, which is a piece of software used by biologists to visualise genomes. Using the browser you can select scaffolds belonging to the genome assembly and scroll along them, observing features that have been annotated (in the case of the whipworm, these will be genes).

Evidence tracks can also be used: these display data from *Trichuris trichiura* or related organisms to help the curator refine the gene model. For an explanation of each of the different types of data potentially used as an evidence track, see the section on Evidence tracks on page 17.

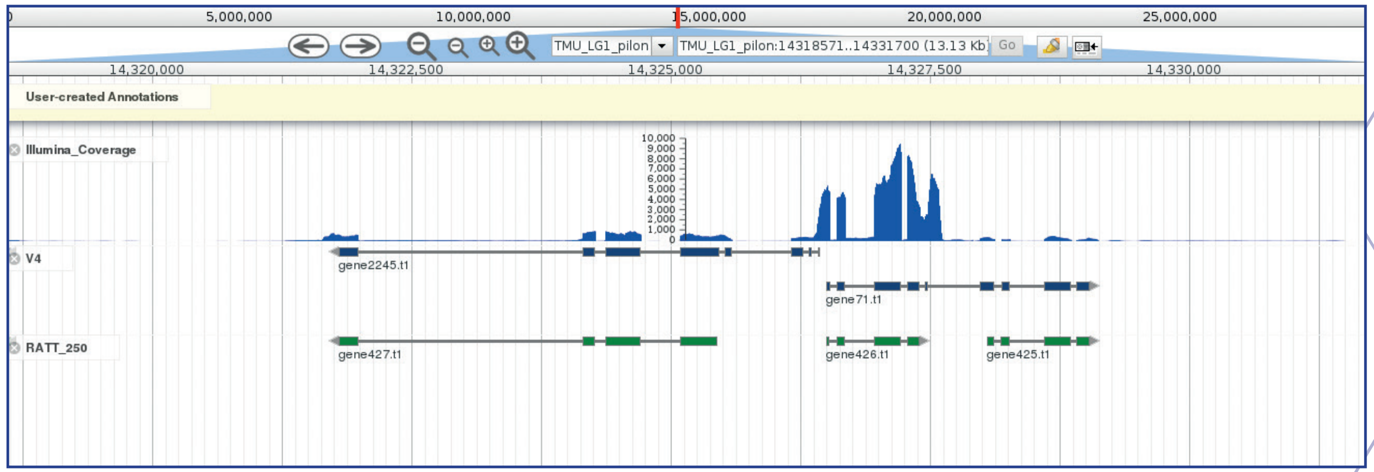
Apollo software allows you to compare gene prediction models with evidence tracks to determine if the prediction is correct. In this example the predicted gene (gene 18005) shown in dark blue matches the RNA sequencing data (Illumina coverage, Isoseq reads and intron data) quite well, indicating this is a good gene model. Dragging the gene to the yellow scratch area at the top of the screen selects this gene for further review.





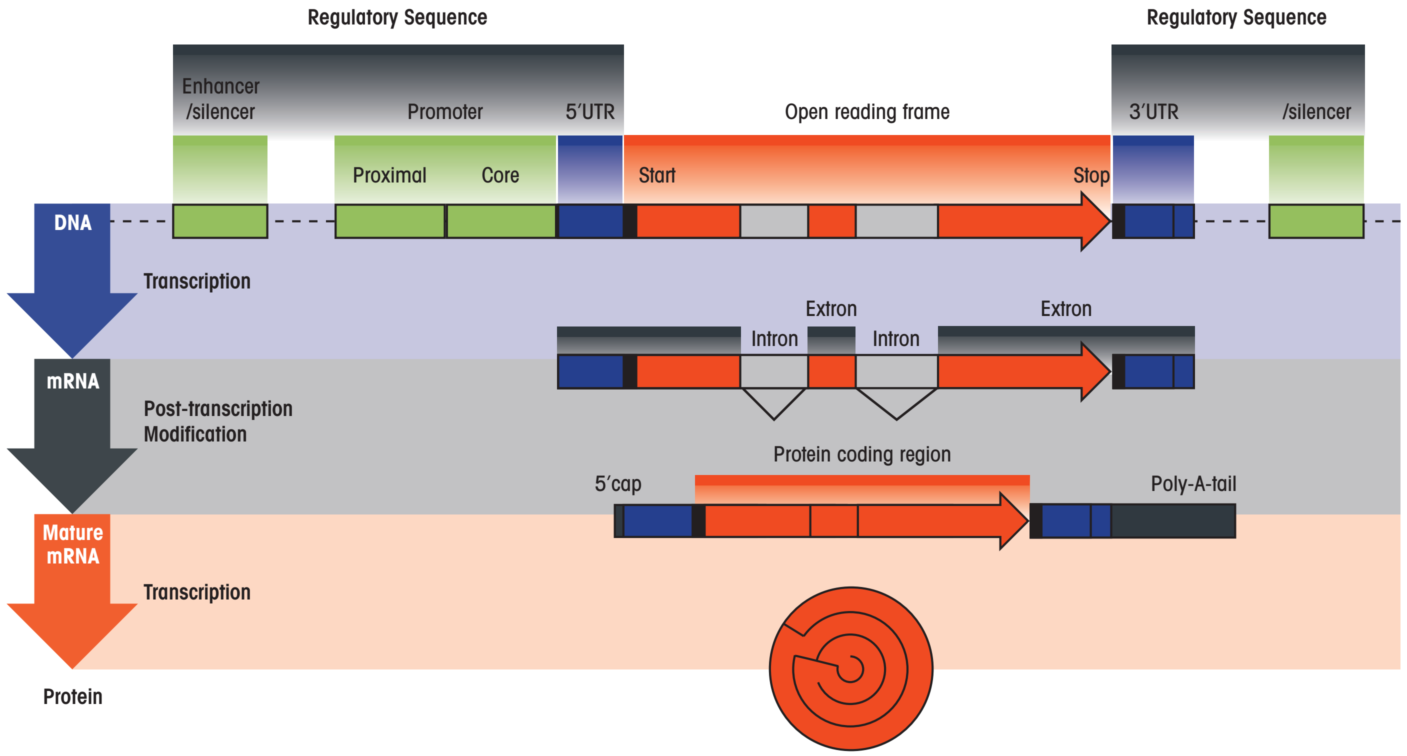
Type: mRNA
Owner: fr7@sanger.ac.uk
Last modified: 2017-07-20 14:19

Apollo allows you to zoom in to see the sequence data of the whipworm genome. In this example we have zoomed in to investigate whether there is support for an alternative splicing event.



Apollo allows you to view multiple gene prediction models and compare them to sequence evidence to choose the best model. The data may suggest an alternative model that has not been predicted by any of the algorithms: Apollo allows you to create and edit new gene models in these cases.

Gene structure



It is important to be familiar with the following features of genes in order to start curating gene models. Terms are listed alphabetically.

- **5' and 3': pronounced "5 prime" and "3 prime".** Nucleic acid sequences have directionality: DNA is transcribed from the 5′ end to the 3′ end. The translation of mRNA transcripts begins from the 5′ end.
- **Alternative transcripts:** thanks to splicing, a single gene can code for more than one protein, by splicing together different combinations of exons from the primary transcript. This increases the complexity and provides an additional challenge for annotation.
- **CDS:** CoDing Sequence. It is the region of the mRNA to be translated into protein.
- **Exon:** Pre-mRNAs are comprised of alternating exons and introns. Exons are the regions that are spliced together to form the mature mRNA. They contribute to both the protein coding sequence (CDS) and upstream and downstream untranslated regions (UTRs).

- **Introns:** introns are the regions of DNA to be removed and not included in the mature mRNA. They do not contribute to the protein.
- **mRNA:** messenger RNA. A generic term for mRNA which can refer to either pre-mRNA or mature mRNA.
- **Mature mRNA:** this is fully processed and ready to be translated. It has had its introns removed by alternative splicing. It will have a PolyA tail at its 3′ end. It consists of the CDS plus 5′ and 3′ UTRs.
- **PolyA tail:** a string of adenosine residues ligated to the 3′ end of the transcript.
- **Pre-mRNA:** the mRNA molecule produced directly upon transcription. It consists of alternating exons and introns, the introns not yet having been removed by alternative splicing.
- **Promoter region:** a sequence within the DNA near the transcription start site that instructs the cell to start transcribing. A very common promoter sequence motif in eukaryotes is TATAAA referred to as the "TATA box".
- **Splice acceptors and donors:** these sequences demarcate intron/extron boundaries. The splice donor site is at the 5′ most end of the intron, whilst the splice acceptor site is at the 3′ most end. The most common splice donor/acceptor sequences are GU/AG (or GT/AG in the corresponding DNA): these are observed in approximately 99 per cent of cases and are described as "canonical" splice sites. Other sequences can act as splice donors/acceptors: these are described as "non-canonical" splice sites.
- **Start codon:** the codon at which translation starts. It is usually AUG (ATG in DNA), which encodes methionine (represented as an M in protein sequences).
- **Stop codon:** the codon at which translation stops. This can be either UAG, UAA or UGA (or in DNA: TAG, TAA, or TGA). In protein sequences a stop codon is represented as a *.
- **Transcription start site:** the site at which a gene starts to be transcribed.
- **Transcription termination site:** the site at which a gene stops being transcribed.
- **UTR:** untranslated region. These can be at the 5′ or 3′ end of the mature mRNA. They do not contribute to the protein product. They play a role in the regulation of translation and transcript stability.

APOLLO

Evidence tracks

Evidence tracks are used by curators to assess how well the gene model is supported by evidence. Evidence tracks might be experimental data generated from the species being annotated or evidence from other related species.

- RNAseq tracks: mature mRNA can be extracted from parasite samples and sequenced. The RNA sequences can then be aligned back to the genome. This is achieved by using algorithms which are "splice aware", meaning the RNA reads are aligned with gaps between the exons such that we can infer where the introns are. RNAseq data aligning to a region where a gene is predicted is very good evidence the gene is real. Even more powerfully, RNAseq evidence can be used to understand the alternative transcripts a gene produces.

- Protein tracks: many genes encode proteins fulfilling core functions common to the organisms across the eukaryotes. For example, proteins involved in core metabolic pathways have very similar sequences in all organisms studied. One form of evidence is to take the protein products of genes from well curated species and align them to the genome you are trying to annotate. This could indicate where well conserved genes are located in the new genome.

GLOSSARY

ANNOTATION:
The process of marking and labelling interesting regions in the genomic DNA sequence. These will usually be genes (and the features that define the structure of genes).

COMPARATIVE GENOMICS:
The comparison of genomic features between different organisms. The genomic features may include the DNA sequence, genes, gene order, regulatory sequences, and other genomic structural landmarks.

CURATION:
The process whereby evidence from experimental data and computational prediction methods is scrutinised by a person in order to create new annotations or refine and improve existing ones.

GENOME BROWSER:
After a genome has been sequenced, assembled and annotated it needs to be visualised in a human understandable way. This can be done via software known as a genome browser.

ORTHOLOGUES:
Genes in different species that evolved from a common ancestral gene. They encode proteins with similar sequences that often have similar functions.

TRANSCRIPTION:
The first step during protein synthesis when the DNA in a gene is copied to produce an RNA transcript called messenger RNA (mRNA).

TRANSLATION:
The second step during protein synthesis where the message from DNA for making a protein has been taken to the ribosome and a protein is constructed with the help of tRNA.

FURTHER SOURCES

The yourgenome.org website has been produced by the Public Engagement team and scientists from the Wellcome Genome Campus. It contains a range of information relating to DNA and sequencing. Below are some specific links relevant to this project.

Gene expression and protein synthesis

http://www.yourgenome.org/facts/what-does-dna-do

http://www.yourgenome.org/facts/what-is-gene-expression

http://www.yourgenome.org/video/from-dna-to-protein

From DNA sample to useful DNA data

http://www.yourgenome.org/facts/what-happens-to-dna-sequence-when-it-comes-off-a-sequencing-machine

http://www.yourgenome.org/facts/how-do-you-put-a-genome-back-together-after-sequencing

http://www.yourgenome.org/facts/how-do-you-identify-the-genes-in-a-genome

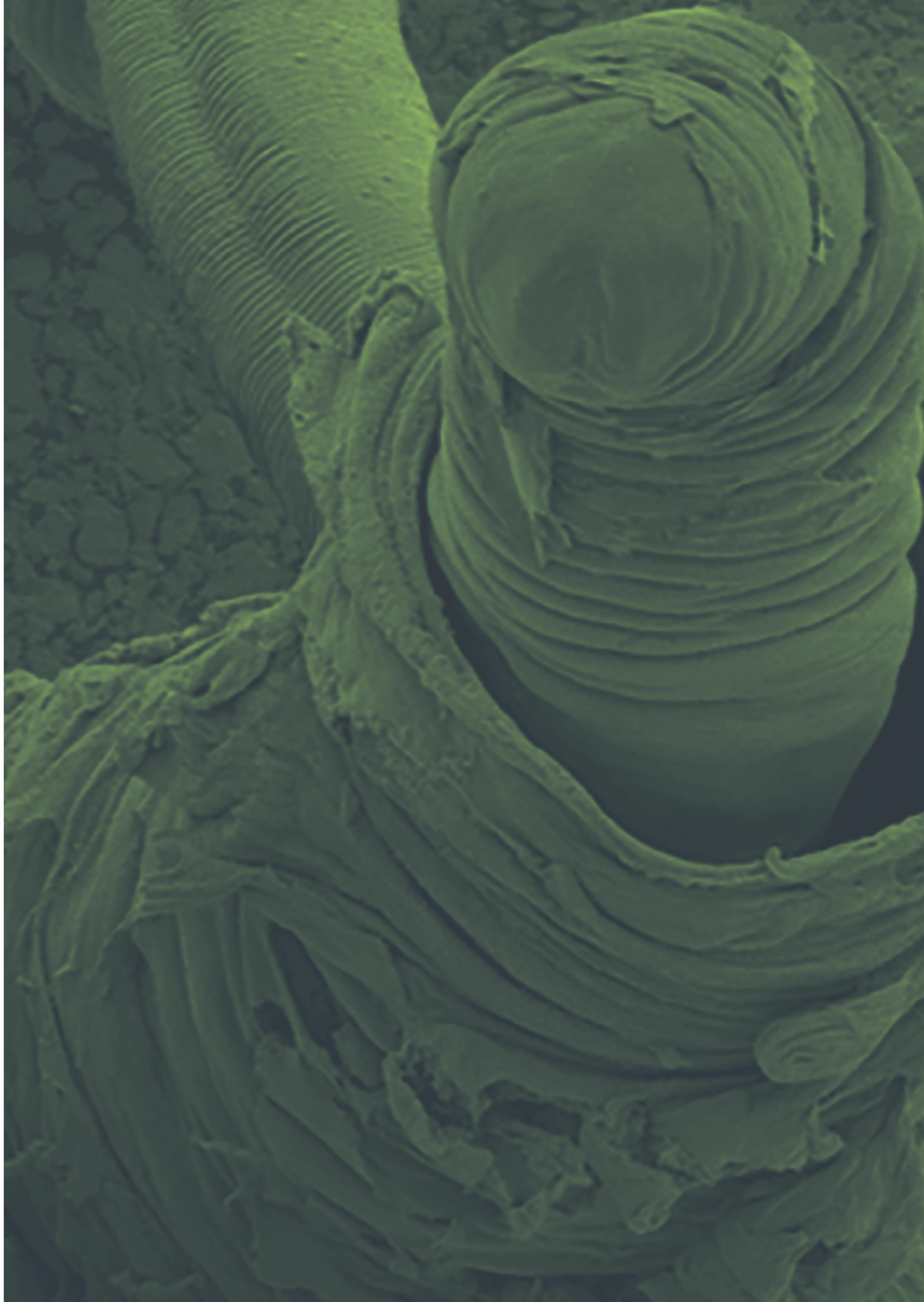
http://www.yourgenome.org/facts/how-do-you-find-out-the-significance-of-a-genome-after-sequencing

http://www.yourgenome.org/facts/how-are-sequenced-genomes-stored-and-shared

Activity: Function finders BLAST

Decode DNA sequences and discover the proteins they code for using online scientific databases.

http://www.yourgenome.org/activities/function-finders-blast



Institute for Research in Schools

Wellcome Wolfson Building,
165 Queen’s Gate,
London,
SW7 5HD

www.researchinschools.org

THE INSTITUTE for RESEARCH in Schools
wellcome trust sanger institute
EMBL-EBI
WELLCOME GENOME CAMPUS PUBLIC ENGAGEMENT
Marie Skłodowska-Curie Actions
CONNECTING SCIENCE