Earth Biogenome Project
https://www.earthbiogenome.org/

Darwin Tree of Life Project
https://www.sanger.ac.uk/programme/tree-of-life/

This Week in Genomics
October 2,2023

1,000 Species Get Their Genomes Sequenced for the First Time

1,000 reference genomes of the highest quality have now been produced for diverse eukaryotic species across the tree of life.

Technology Network/Genomic Research
https://www.technologynetworks.com/genomics/news/1000-species-get-their-genomes-sequenced-for-the-first-time-378926

A thousand reference genomes of the highest quality have now been produced for diverse eukaryotic species across the tree of life. The feat has been achieved by the ***Wellcome Sanger Institute's Tree of Life Programme***, where a ***scientific pipeline*** has been established to ***sample organisms, extract and sequence their DNA, assemble and curate their genomic data, and openly release these for use by researchers worldwide.***

One thousand genomes is an impressive achievement: ***nearly half of all the high-quality genomes generated worldwide*** so far. Behind the 1,000 number is a ***huge amount of work*** by naturalists, taxonomists, lab specialists and computer scientists, and it ***creates a platform*** from which both Sanger Institute teams and biologists in general ***can reach for even higher goals***. We look at some other equally impressive numbers behind this thousand-species milestone.

Open science - data for all

We live in an age of ***deep concern for the biosphere*** and our place in it. If humanity is to ***weather the growing crises and preserve and enhance biodiversity*** we ***need to share our knowledge***. In this spirit, ***all the genome sequence data produced*** - ***all 1,000 genome sequences*** - have been ***released openly***, so that ***anyone, anywhere in the world can use them***. All of the data are released into the European Nucleotide Archive, and from there shared around the world. They may be ***used for monitoring species conservation status, discovering new drugs and bioproducts or for deepening understanding of the evolution and functioning of all life***.

141 taxonomic orders - diversity across the tree of life

At the heart of the programme's mission is capturing the diversity of life on Earth. The ***genomes for some species are easier to produce than others***, but the researchers have never shied away from tackling the trickier groups. As a result, the ***first 1,000 genomes span over 141 different taxonomic orders across more than 20 phyla***.

390 Lepidoptera species - comparing things that flutter

There are several reasons why so ***many of these first*** 1,000 species are ***moths and butterflies***. Most practically, nocturnal moths have a habit of ***flying towards light and into moth traps***. This was especially helpful considering the Tree of Life programme was established just prior to the Covid-19 pandemic and the team's ***early collections were largely restricted to their own back gardens***. ***Lepidoptera genomes*** are also relatively ***small***, and ***DNA extraction and genome assembly is generally more straightforward*** than for many other groups. That said, ***Lepidoptera*** are also ***important*** - as ***pests, as pollinators and as major herbivores in many ecosystems*** - and they comprise ***nearly one fifth of all described animal species.***

The ***purpose of sequencing everything*** is also to look in depth at closely related species, a ***scientific field known as comparative genomics***. Lepidoptera were selected as a group to ***showcase early on how the programme could supercharge this***. Research into Lepidoptera evolution, conservation and specific genes has already begun to emerge based on the data.

94 billion DNA base pairs - the largest genome

The ***1,000th genome released*** by the teams was also the ***largest that anyone has yet assembled***. Indeed, the ***European Mistletoe (Viscum album)*** is the species with the ***largest genome in Britain and Ireland***. At ***94 billion DNA bases long, it is 30 times the size of the human genome***.

Mistletoe also has the ***largest chromosomes*** - a continuous sequence of DNA code - that the teams have assembled and curated. The ***largest of its 10 chromosomes is almost 11 billion DNA bases*** in size. It's assembly probed the limitations of existing processes and tools.

4,011 manual interventions - the most complicated genome

The ***1,000 genomes*** have been ***achieved*** in part because of ***step changes in laboratory and computer technologies***. In the laboratory, we are able to ***generate ever more accurate and longer stretches of sequencing read data***, which make ***genome assembly easier*** - putting back together bigger puzzle pieces rather than smaller ones. While we cannot yet read each chromosome in one go, the ***accuracy of these new methods*** (about one error in 10,000 letters read) makes the ***"assembly" process much easier and more accurate.*** The assembly process uses new computer tools and hardware that ***perform at speeds and scale that were dreamed-of but not achievable just five years ago***.

Bioinformaticians use several cutting-edge computer tools - often developed by the teams themselves - to ***ensure that the DNA sequence submitted is correct***. These tools are trained to ***spot and correct a whole range of possible errors***, but can be fooled by complicated sections of the genome unique to each species, ***especially regions that are highly repetitive***. For these very difficult regions, an expert human needs to go in and complete the genome by eye, manually positioning bits of sequence correctly in the chromosomes.

The ***genome*** that required the ***most human attention*** was the ***Common Toad (Bufo bufo)*** with over 4,011 interventions needed to build a correct representation of its 5 billion letter-long genome. In contrast, ***14 of the first 1,000 genomes required no manual interventions at all!***

55 genomes with DNA sequence fully assigned to chromosomes

The Tree of Life programme strives for the highest quality genomes possible using the latest genomic technologies. These reference genomes are designed to stand the test of time and be useful to scientists for years to come.

A measure of this is how much of the DNA data generated can be placed along the chromosome structure by bioinformaticians. The ***programme aims to get as close to 100 per cent as possible*** and the completeness of their reference genomes is world-leading, nevertheless for most species ***some parts of the sequence simply won't fit***. However in ***55 species, 100 per cent of the sequence has been assigned to chromosomes*** - an astonishing feat.

90 chromosomes - the species with the most chromosomes

Ordering the DNA sequence data along a species' chromosomes is key to producing these reference genomes. But ***not all species have the same number of chromosomes***. The current record holder among the first 1,000 genomes is ***Lysandra coridon (chalkhill blue), a butterfly with 90 chromosomes***. On the opposite end, the scale ***insect Icerya purchasi has only two chromosomes.***

4 copies - the highest ploidy in a species so far

Humans are diploid, meaning we have two copies of our genetic material - one inherited from our mother and one from our father. But evolution has thrown up some wild variations on this, with species copying and losing parts of their genomes in ways that result in a much higher ploidy count. The ***species with the highest ploidy count*** in our first 1,000 genomes is the ***silverweed Potentilla anserina***, which is a ***tetraploid, meaning it has four copies of its genome in each cell***. We currently have more tetraploid and some hexaploid (six copies) genomes in the pipeline. The genome with the ***highest ploidy*** we are working with at the moment is that of the ***Seaside Arrowgrass, Triglochin maritima***, which is ***octaploid (eight copies).***

400,000 lines of code - software to support the science

The scale of Tree of Life's science requires support from some powerful platforms, many of which are built by the programme's software developers. ***400,000 lines of Python and Javascript are contained in the bespoke software products created by the Enabling Platforms team***. 70 per cent of that is in the ***Samples Tracking System***, which allows the team to keep tabs on how samples from different species are progressing through the genome-generating pipeline. This is now being adopted by teams across the Wellcome Sanger Institute.

The Enabling Platforms software team are currently working hard on a variety of websites to easily explore and track the data being generated across the programme.

Thousands more - the genomes yet to come

These first 1,000 genomes are just the beginning. The projects on which ***Tree of Life are collaborating already aim to produce thousands more genomes***, and crack the processes required for even the trickiest taxa.

A few hundred incredibly interesting marine species, including corals and their symbionts, are expected soon via the Aquatic Symbiosis Genomics project. Other international collaborations such as the Vertebrate Genomes Project and Biodiversity Genomics Europe will see thousands more species arriving at the Sanger Institute for sequencing.

All these projects are part of the Earth BioGenome Project, a global network of initiatives and institutes ultimately aiming to sequence all 1.2 million named species on the planet, transforming biology along the way.