International Vertebrate Genomes Project releases first 15 high-quality reference genomes
Publicly available data will impact studies on life, disease, and conservation efforts

The Genome 10K (G10K) announces the official launch of a new project, the international Vertebrate Genomes Project (VGP), and its first release of 15 new, high-quality reference genomes for 14 species representing all five vertebrate classes – mammals, birds, reptiles, amphibians, and fishes. The mission of the VGP is to provide high-quality, near error-free, and complete genome assemblies of all 66,000 vertebrate species on Earth to address fundamental questions in biology, disease, and conservation.

The new sequences are stored and publicly available in the Genome Ark database, a new digital open-access library of genomes generated by the G10K-VGP consortium and hosted by Amazon, and will soon be processed for gene identifications in international public genome browsing and analyses databases, including the National Center for Biotechnology Information (NCBI), Ensembl, and University of California, Santa Cruz (UCSC) genome browser. The G10K-VGP consortium has convened more than 150 experts from academia, industry, and government, from over 50 institutions in 12 countries, to develop high-resolution sequencing and genome assembly methods that reduce cost and eliminate errors that plague current reference genomes. The new VGP genomes eliminate many of these errors. For conservation efforts, these VGP genomes will be used to identify species most genetically at risk for extinction, preserving their genetic information for the future and helping to save them from extinction.

One of the species included in the first release is the kakapo, a flightless parrot found only in New Zealand that is on the brink of extinction, with less than 150 alive. In partnership with the Kakapo Genetic Rescue Project, G10K Chair Erich Jarvis, professor at Rockefeller University and Howard Hughes Medical Institute Investigator, and his group helped sequenced samples from a bird named Jane to create a high-quality assembly that will now become the reference genome for her species. Jane unfortunately died on May 17, 2018, just before the completion of her genome. This first data release of species is being dedicated to Jane and to conservation efforts all over the world to preserve Earth's biodiversity.

The 15 genomes created through the VGP are a proof of principle demonstrating the strength of the G10K-VGP consortium and the new sequencing technology's dependability and scalability to sequence all vertebrate genomes. These genomes are currently the most complete versions of their species to date:

- **Mammals (4 species)**
  - Two bat species, Greater horseshoe bat (*Rhinolophus ferrumequinum*) and Pale spear-nose bat (*Phyllostomus discolor*), used as models for longevity and vocal learning
  - The Canada lynx (*Lynx canadensis*), once nearly extinct in the United States and now recovering
  - The duck-billed platypus (*Ornithorhynchus anatinus*), an egg-laying mammal with reptilian traits

- **Reptiles (1 species)**
  - A newly discovered turtle species from Mexico, Goode's Thornscrub Tortoise (*Gopherus evgoodei*)

- **Amphibians (1 species)**
  - Two-lined caecilian (*Rhinatrema bivittatum*), a limbless amphibian that resembles a snake

- **Birds (3 species. 4 genomes)**

- In addition to the kakapo (*Strigops habroptilus*), the VGP re-sequenced species from two other bird orders to represent the only three vocal learning birds among more than 40 avian orders
- A male and female zebra finch (*Taeniopygia guttata*), the most commonly studied vocal learner
- Anna's hummingbird (*Calypte anna*), belonging to the smallest group of birds

- **Fish (5 species)**
  These species represent a large diversity of traits and are used to study species evolution and adaptation:
  - Flier Cichlid (*Archocentrus centrarchus*), native to Central America
  - Eastern happy (*Astatotilapia calliptera*), also a cichlid fish Native to Lake Malawi, Africa
  - Climbing perch (*Anabas testudineus*), native to inland waters of Southeast Asia
  - Tire track eel (*Mastacembelus armatus*), native to rivers of Southeast Asia
  - Blunt-snouted clingfish (Gouania willdenowi), native to north Mediterranean coast, Syria to Spain

Over the last three years, the G10K-VGP consortium worked behind the scenes to compare all the major sequencing and analysis technologies on just a few animals to help advance and develop the needed technologies to create higher quality, "platinum-level" genomes. They found, as some others have, that sequencing technologies with long reads always gave higher-quality results than with short reads and that technologies that measure long-range genome interactions are necessary to "assemble" these DNA reads into whole chromosomes. Further, they found that the common practice of merging the paternal and maternal chromosomes (haplotypes) into one genome was causing numerous errors. Therefore, they are now assembling the paternal and maternal DNA of an individual separately (called phasing).

Dr. Jarvis says "I got tired of having my students spend months to a year or more, and more money, re-cloning and re-sequencing genes because the current draft genome assemblies were not good enough for our studies of genetics of vocal learning and spoken language in songbirds and humans. So, when I was asked and voted in as G10K Chair, I decided to make it a mission to help generate high-quality genome assemblies for studies using any vertebrate species. The bird genomes are also being generated as part of an associated Bird 10,000 (B10K) genomes project."

Gene Myers, a director of the Max-Planck Society in Dresden and well-known bioinformatician, G10K Council member and lead of one of the sequencing hubs, says, "The advances in long-read sequencing and long-range scaffolding technologies is revolutionizing *de novo* DNA sequencing. After a 10-year hiatus, this trend inspired me to return to genome assembly as I believe we will ultimately be able to produce near-perfect, telomere-to-telomere genome reconstructions, and if current cost trends continue, for less than $1,000 on average per vertebrate species, thus dramatically altering the landscape of genomics."

The current Phase 1 genomes are being built with Pacific Biosciences long reads to generate an initial assembly of pieces of chromosomes (called contigs), 10X Genomics linked reads to join them together in bigger pieces (called scaffolds), Bionano Genomics optical DNA maps to link them at a larger scale and correct structural errors in the sequence assembly, Arima Genomics (also Dovetail Genomics and Phase Genomics) Hi-C proximity-ligation data to bring larger pieces together into whole chromosomes, and G10K-VGP genome assembly computer algorithms, which were specifically developed by this consortium and will become useful for all species.

Adam Phillippy, Chair of the VGP Assembly Working Group and head of the Genome Informatics Section at the National Human Genome Research Institute, says, "Until recently, sequencing the complete genome of a single animal required millions of dollars and years of effort. New sequencing technologies have dramatically reduced the cost and made it possible to reconstruct near-perfect genomes for the first time. Despite these advances, the computational challenges of assembling and analyzing thousands of genomes remain. To tackle these remarkable challenges, we have assembled an all-star team of bioinformaticians and are recruiting help from around the world. In addition, our corporate informatics partners at DNAnexus and Amazon Web Services have been instrumental in getting this project off the ground."

The G10K-VGP consortium plans to complete the VGP in taxonomic hierarchy from Phase 1 representing all 260 orders of living vertebrates, to Phase II representing 1,045 families, Phase III representing 9,478 genera, and finally Phase IV, representing approximately all 66,000 species of vertebrates. Additionally, the VGP will sequence the heterogametic sex where it exists, so that both sex chromosomes can be recovered for each species. The species in Phase 1 are based on a proposed new definition of orders based on species that diverged from each other soon after the last mass extinction event that killed off the dinosaurs 66 million year ago. Studying these ordinal-level species will help scientists determine what type of species survived that mass extinction and inform efforts on how to help species survive the current anthropogenic 6th mass extinction event.

Richard Durbin, of the University of Cambridge and the Wellcome Sanger Institute, G10K Council member and lead of the sequencing hubs, says, "The last 20 years have proven the value of openly available high-quality reference genome sequences to scientific research, but until now, these have mostly been available just for humans and other key organisms. We are entering an era in which we will obtain reference genome sequences for all species across the Tree of Life. This announcement and data release are key steps towards this goal, for vertebrates, the phylum of animals that we belong to."

Prof. Emma Teeling, University College Dublin, Ireland and Director of the associated Bat 1K Project, said, "Today represents a monumental example of what is possible when determined people imagine the future. Working together we have sequenced 15 exquisite genomes from across deep evolutionary time, unique in their quality and perfection, enabling us for the first time to uncover the genetic basis of vertebrate life. Now that we started producing exquisite genomes of all living vertebrate orders at high-quality, imagine doing so for all life. Why not?".

Jenny Graves, one of the pioneers of comparative genomics and sex chromosome evolution who was not involved in recent sequencing projects, exulted, "This is a real tour-de-force. We could not have imagined, twenty years ago, that we would ever have genome sequences of more than a handful of animals. Now we have real prospects of solving evolutionary mysteries and charting population health in endangered (even extinct) animals."

The G10K-VGP leadership consists of a 15-member council, a board of trustees, and 16 subgroups that perform the daily operations of the VGP, including obtaining tissue sample permits, executing DNA extractions, sequencing genomes, performing genome alignments and annotation, and managing the project within and across institutions and countries. The genome sequencing hubs are currently based at the Rockefeller University in New York led by Olivier Fedrigo and Erich Jarvis, the Sanger Institute in the United Kingdom led by Richard Durbin and his team including Shane McCarthy and Kerstin Howe, and the Max Planck Institute of Molecular Cell Biology and Genetics in Dresden, Germany led by Gene Myers and his team including Martin Pippel and Sylke Winkler. The assembly team to which they all belong is led by Adam Phillippy, along with his team members Arang Rhie and Sergey Koren at the NIH. Building on her previous experience assembling and phasing human and animal genomes, Dr. Rhie made a massive effort to help develop a standard assembly process for the VGP. Harris Lewin and his postdoc Joana Damas at UC Davis and others played essential roles in evaluation of assemblies and other stages of the project. The VGP hubs are currently working with major sequencing and assembly companies to further test, improve, and generate new approaches for producing the most complete and error-free reference genomes possible. The G10K-VGP has an open-door policy for any scientist and others that want to join, so long as they follow the G10K policies.

Approximately $600 million is needed to complete all VGP phases. The G10K-VGP is currently focused on completing Phase 1 through crowdsourcing among scientists, having raised $2.5 million of the $6 million thus far needed for this phase. For those in the public that wish to help support the project, or even sponsor a species, more information is available at *https://vertebrategenomesproject.org/ways-to-help-1/*. Financial gifts to the G10K-VGP can be donated at *https://giveandjoin.rockefeller.edu/vgl-donate*.

Contact:
Dr. Sadye Paez, G10K-VGP Program Director
212-327-8206
spaez@rockefeller.edu

The Rockefeller University

The Rockefeller University is the world's leading biomedical research university and is dedicated to conducting innovative, high-quality research to improve the understanding of life for the benefit of humanity. Our 82 laboratories conduct research in neuroscience, immunology, biochemistry, genomics, and many other areas, and a community of 1,800 faculty, students, postdocs, technicians, clinicians, and administrative personnel work on our 14-acre Manhattan campus. Our unique approach to science has led to some of the world's most revolutionary and transformative contributions to biology and medicine. During Rockefeller's 117-year history, 25 of our scientists have won Nobel Prizes, 23 have won Albert Lasker Medical Research Awards, and 20 have garnered the National Medal of Science, the highest science award given by the United States.

The Vertebrate Genome Laboratory at the Rockefeller University

The Vertebrate Genome Laboratory (VGL) at the Rockefeller University is a Resource Center specializing in ultra-High-Molecular Weight DNA (uHMW DNA) and long-read genomic technologies. The primary objective of the VGL is to generate at least one high-quality, phased, chromosome-level, annotated, reference genome assembly of all approximately 66,000 vertebrate species for the Vertebrate Genomes Project (G10K-VGP). The VGL consists of ~1,500 sq. ft. located on the 7th floor of the Weiss Research Building. The team is composed of four members, including a Director, two Research Support Specialists/Associates, and a Research Assistant. The VGL is equipped with four state of the art Pacific Biosciences Sequel™ sequencers, one Bionano Genomics Saphyr™ optical mapper, one 10x Genomics Chromium™ microfluidics platform, and all the necessary ancillary instruments for preparing uHMW DNA.

The Wellcome Sanger Genome Institute

The Wellcome Sanger Institute is one of the world's leading genome centres. Through its ability to conduct research at scale, it is able to engage in bold and long-term exploratory projects that are designed to influence and empower medical science globally. Institute research findings, generated through its own research programmes and through its leading role in international consortia, are being used to develop new diagnostics and treatments for human disease. To celebrate its 25th year in 2018, the Institute is sequencing 25 new genomes of species in the UK. Find out more at www.sanger.ac.uk or follow @sangerinstitute

The MPI-CBG

The Max Planck Institute of Molecular Cell Biology and Genetics (MPI-CBG) is one of 84 institutes of the Max Planck Society, an independent, non-profit organization in Germany. 500 curiosity-driven scientists from over 50 countries ask: How do cells form tissues? The basic research programs of the MPI-CBG span multiple scales of magnitude, from molecular assemblies to organelles, cells, tissues, organs, and organisms.

#   #   #