VGP generates largest number of high-quality genomes of iconic and endangered species

August 28 2019

The Vertebrate Genomes Project (VGP) and collaborators are announcing the second data set of the largest number (101) of chromosomal-level genome assemblies of vertebrates towards completing Phase 1 of the VGP, which includes one representative species per vertebrate order or ~260 species. These 101 genomes, most finished or in their final stages of assembly, demonstrate the success of the VGP in utilizing and developing experimental and computational tools for scalability to achieve the goal of producing high-quality, near error-free, and complete chromosomal-level genome assemblies of all 70,000 extant vertebrate species on Earth. The VGP revised the number of vertebrate species upwards from 66,000 since its first data set because of updates in species identification and classifications.

These genomes will facilitate solving problems in biology, medicine, and conservation including studies of life, disease, and biodiversity such as generating a more complete and accurate family tree of vertebrates, deciphering vertebrate chromosomal genome evolution, comparing genomics of convergent traits (i.e., vocal learning, flight, loss of limbs, and aquatic/terrestrial adaptations), and reconstructing the genomes of common ancestors of all vertebrates and of key vertebrate clades (e.g., mammals, birds, reptiles, amphibians, teleost fish and tetrapods).

The VGP has now been able to scale up to about 10 genomes per month, an increase from a rate of about 1 genome per month since the first data

set of 15 genomes in September 2018, a 10-fold increase in output. The previous announcement established the strength of the G10K-VGP consortium and the capability of new sequencing technology to reliably achieve high quality, near-error free, phased reference genomes, which have since been further improved to generate even higher-quality genome assemblies.

Nearly all genomes were done in collaboration with individual scientists or other consortium projects, including the vertebrates of the Wellcome Sanger Institute's 25 Genomes for 25 Years, Bat1K genomes, and B10K bird genomes. Most of the genome data were generated at three sequencing hubs that have invested in the mission of the VGP including at the Rockefeller University Vertebrate Genome Lab (VGL) in New York, USA, the Wellcome Sanger Institute in the UK, and the Max Planck Institute (MPI) in Dresden, Germany, led by VGP Assembly Team Chair Adam Phillippy of the National Institutes of Health in Bethesda, Maryland and team members Olivier Fedrigo of VGL, Richard Durbin of Cambridge University, UK, and Gene Myers of MPI. The VGP set up their new genome assembly pipeline on DNANexus, a cloud-based computing platform for genomics.

These new assemblies include improvements they and others have made in the genome sequencing and assembly technology since the first data set such as better resolution on separating out paternal and maternal chromosome sequences that were found to cause errors in genome assemblies. Many of these species had earlier versions of their genomes assembled, but because these prior assemblies were too fragmented and did not meet the quality metrics set by the VGP, they were revisited with new long-read DNA sequencing and chromosomal assembly technologies that the VGP helped develop.

G10K Chair, Erich Jarvis, a professor at the Rockefeller University and Investigator of the Howard Hughes Medical Institute, says "This second

data set demonstrates the power of the VGP to bring together the international collective wisdom and expertise for generating the highest quality genome data possible, for the least cost possible, for the best science possible, and for the good of humanity and other species".

Of these 101 species, 100 are vertebrates and one is an invertebrate, a starfish contributed by the Sanger Institute's 25 Genomes Project as an outgroup relative. The 100 vertebrates represent 77 taxonomic orders sequenced to this completeness for the first time, which, along with 13 from the previous data set, add to a total 90 orders of the ~260 Phase 1 species. These genomes include iconic species such as the largest vertebrate—the blue whale, as well as the bottlenose dolphin, parakeet, marmoset monkey, red-bellied piranha, Great Potoo, and jawless sea lamprey, a primitive fish.

Emma Teeling, professor at the University College Dublin in Ireland and co-director of Bat1K, stated "We have completed our pilot study and sequenced the genome of six bat species from phylogenetically diverse families to chromosome level assemblies. These genomes have already revealed some unique genomic adaptations pertaining to mammalian flight, echolocation and extraordinary immunity".

For conservation efforts, these genomes will be used to help identify species most genetically at risk for extinction, preserving their genetic information for the future and helping to save them from the human-induced sixth mass extinction. This data set includes four critically-endangered species (vaquita, European eel, Bolson tortoise, and smalltooth sawfish), seven endangered species (blue whale, grey crowned-crane, green sea turtle, Atlantic halibut, ring-tailed lemur, chimpanzee, and golden aronawa), and eight vulnerable species (sterlet, thorny skate, Siamese fighting fish, Abyssinian ground hornbill, great white shark, leatherback sea turtle, Atlantic cod, and European turtle dove).

The vaquita is perhaps the most critical of this set. Through contacts made by Jacquelyn Mountcastle of the VGL at the Rockefeller University, the VGP worked with Phillip Morin of the National Oceanic and Atmospheric Administration in collaboration with Mexican researchers. Mexico's vaquita porpoise accidentally entangles and drowns in fishing nets, and the resumption of illegal fishing for an endangered fish to supply the Chinese black wildlife market accelerated the decline such that nearly half die each year. A rescue effort in 2017 involved 90 researchers from 9 countries to try to capture some of the 30 vaquitas that remain in the Gulf of California to save them from extinction. Unfortunately, one animal captured at the time died soon after of shock, but her live cells were cultured and frozen by the San Diego Frozen Zoo, which were then used to generate the high-quality reference genome sequence. This year, Leonardo Dicaprio produced a documentary, "Sea of Shadows", on the plight of the vaquita to help build public support for saving it from extinction; the documentary included the female whose genome the VGP sequenced and assembled. Her chromosomes are highly homozygous, but preliminary analyses suggest that this is due to tens of thousands of years persistence as a small population rather than recent loss of diversity that might accelerate extinction. Phillip Morin says "The vaquita genome analysis offers a strong counter to the common argument of genetic doom that has been brought up repeatedly as a red herring argument against trying to save the species."

Similarly, Lisa M. Komoroske, Assistant Professor of Conservation Genomics & Ecophysiology at the University of Massachusetts, Amherst, who led the effort to raise funds with the VGP for the Pacific leatherback turtle genome, says "The populations have declined by greater than 90% and are listed as critically endangered. This has been largely due to human activities such as direct harvest and fisheries interactions. Pacific leatherbacks are one of eight species among the most at risk of extinction in the near future protected by the United

States NOAA under the Endangered Species Act," Komoroske continues. "Often referred to as 'living dinosaurs', leatherback turtles are an ancient lineage that possess unique physiological adaptations, including those that allow them to survive in cold waters exploiting habitats far beyond many other ectotherms". Komoroske and others are using the VGP-affiliated genome to study the remaining genetic diversity in the species, particularly important as new leatherback conservation initiatives are being determined to enable genetic mixing of populations to avoid excessive inbreeding.

This data set also includes 12 trio-based assemblies, where the DNA of the parents are used to separate the DNA sequences of the child chromosomes to assemble two genomes (one each from mother and father) from one individual: zebra finch, bottlenose dolphin, common brushtail possum, common marmoset, nile rat, budgerigar, chicken, Bolson tortoise, hourglass treefrog, zebrafish SAT, sterlet, and human. Based on an assembly approach developed by Sergey Koren and Arang Rhie of the Phillippy lab at NHGRI, these trio-based assemblies are 40-60% better than the non-trio based assemblies at separating out parentally-inherited DNA. The new chicken trio assembly in progress is expected to improve agricultural and biomedical research because the chicken is the most commonly studied avian genome for these areas. The human trio assembly in progress is expected to help inform the scientific community on how to generate higher quality human genome assemblies for personalized medicine and to understand human evolution and our network of relationships between each other.

These genomes have already been used to train the next generation of scientists on how to produce high-quality, chromosomal level reference assemblies. Dr. Arang Rhie, who played a key role a developing the computational pipeline for generating the high-quality VGP assemblies, performed online training and supervision of students internationally who then assembled many of the 101 genomes. This training opportunity

provided to novice researchers is helping to democratize and scale-up the generation of high-quality reference genomes; in the future, this opportunity will enable the generation of thousands of genomes per year to meet the ambitious targets that the VGP and related projects.

The new sequences are stored and publicly available in the Genome Ark database, a new digital library of genomes generated by the G10K consortium and hosted by Amazon, and annotated and displayed in international public genome browsing and analysis databases including the National Center for Biotechnology Information (NCBI), Ensembl at the European Bioinformatics Institute, and UCSC Genome Browser —part of the UC Santa Cruz Genomics Institute—which recently launched with 24 vertebrate assemblies. Of the 101, 60 are available immediately, and the others are being deposited soon, all under the G10K data use policy to ensure equitable data use and publication.

Approximately $600 million is needed to sequence all 70,000 vertebrate species. We are currently focused on completing Phase 1, which will provide representative reference sequences for all 260 vertebrate orders, through crowdsourcing among scientists and has successfully crowdsourced $4.8 million of the $6 million thus far needed (sample and funding availability for Phase 1 species is available here).

More information: 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 made at https://giveandjoin.rockefeller.edu/vgl-donate.

Provided by Rockefeller University