The Vertebrate Genomes Project introduces a new era of genome sequencing

April 28 2021

Credit: CC0 Public Domain

The Vertebrate Genomes Project (VGP) today announces their flagship study and associated publications focused on genome assembly quality and standardization for the field of genomics. This study includes 16

diploid high-quality, near error-free, and near complete vertebrate reference genome assemblies for species across all taxa with backbones (i.e., mammals, amphibians, birds, reptiles, and fishes) from five years of piloting the first phase of the VGP project.

In a special issue of *Nature*, with companion papers simultaneously published in other scientific journals, the VGP details numerous technological improvements based on these 16 genome assemblies. In the flagship study, the VGP demonstrates the feasibility of setting and achieving high-quality reference genome quality metrics using their state-of-the-art automated approach of combining long-read and long-range chromosome scaffolding approaches with novel algorithms that put the pieces of the genome assembly puzzle together.

Growing out of the decade-old mission of Genome 10K Community of Scientists (G10K) to sequence the genomes of 10,000 vertebrate species and other comparative genomics efforts, the VGP is taking advantage of dramatic improvements in sequencing technologies in the last few years to begin production of high-quality reference genome assemblies for all ~70,000 living vertebrates. To date, the current VGP pipelines have led to the submission of 129 diploid assemblies representing the most complete and accurate versions of those species to date and is on the path to generating thousands of genome assemblies, demonstrating feasibility in not only quality standardization but also scale.

"When I was asked to take on leadership of the G10K in 2015, I emphasized the need to work with technology partners and genome assembly experts on approaches that produce the highest quality data possible, as it was taking months per gene for my students and postdocs to correct gene structure and sequences for their experiments, which was causing errors in our biological studies", said Erich Jarvis, lead of the VGP sequencing hub at The Rockefeller University, Chair of the G10K and a Howard Hughes Medical Institute Investigator. "For me this was

not only a practical mission, but a moral imperative."

Arang Rhie, first author of the flagship paper from the National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, U.S., adds, "It truly was a challenge to design a pipeline applicable to highly diverged genomes. Our largest genome, 5 Gb in size, broke almost every tool commonly used in assembly processes. The extreme level of heterozygosity or repeat contents posed a big challenge. This is just the beginning; we are continuously improving our pipeline in response to new technology improvements."

The VGP's approach combines assembly pipelines with manual curation to fix misassemblies, major gaps, and other errors, which informs the iterative development of better algorithms. For example, the VGP helped reveal high levels of false gene duplications, losses or gains, due mostly to algorithms not properly separating maternal and paternal chromosomes. One solution includes a trio binning approach of using DNA from the parents to separate out the paternal and maternal sequences in the offspring. For cases where parental data is unavailable, another solution developed by the VGP and collaborators is an algorithm called FALCON-Phase that reduces the computational complexity of phasing maternal and paternal DNA sequences at chromosome scale.

Kerstin Howe, lead of the curation team at the Wellcome Sanger Institute in the UK, says, "Our new approach to produce structurally validated, chromosome-level genome assemblies at scale will be the foundation of ground-breaking insights in comparative and evolutionary genomics."

Adam Phillippy, chair of the VGP genome assembly and informatics working group of over 100 members and head of the Genome Informatics Section of the National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, U.S., adds,

"Completing the first vertebrate reference genome, human, took over 10 years and $3 billion dollars. Thanks to continued research and investment in DNA sequencing technology over the past 20 years, we can now repeat this amazing feat multiple times per day for just a few thousand dollars per genome."

The excellent quality of these genome assemblies enables unprecedented novel discoveries which have implications for characterizing biodiversity for all life, conservation, and human health and disease. The first high-quality reference genomes of six bat species, generated with the Bat 1K consortium, revealed selection and loss of immunity-related genes that may underlie bats' unique tolerance to viral infection. This finding provides novel avenues of research to increase survivability, particularly relevant for emerging infectious diseases, such as the current COVID-19 pandemic.

Specific to conservation and in collaboration with the Māori in New Zealand and officials in Mexico, genomic analyses of the kākāp?, a flightless parrot, and the vaquita, a small porpoise and the most endangered marine mammal, respectively, suggest evolutionary and demographic histories of purging harmful mutations in the wild. The implication of these long-term small population sizes at genetic equilibrium gives hope for these species' survival.

Richard Durbin, a Professor at the University of Cambridge and lead of the VGP sequencing hub at the Wellcome Sanger Institute in the UK, says, "These studies mark the start of a new era of genome sequencing that will accelerate over the next decade to enable genomic applications across the whole tree of life, changing our scientific interactions with the living world."

Gene Myers, lead of the VGP sequencing hub at the Max Planck Institute in Dresden, Germany, elaborates, "The VGP project is at the

vanguard of the creation of a genomic catalog in analogy with Linnaeus' classification of life. I and my colleagues in Dresden are excited to be contributing superb genome reconstructions with the funding of the Max-Planck Society of Germany."

The VGP involves hundreds of international scientists working together from more than 50 institutions in 12 different countries since the VGP was initiated in 2016 and is exemplary in its scientific cooperation, extensive infrastructure, and collaborative leadership. Additionally, as the first large-scale eukaryotic genomes project to produce reference genome assemblies meeting a specific minimum quality standard, the VGP has thus become a working model for other large consortia, including the Bat 1K, Pan Human Genome Project, Earth BioGenome Project, Darwin Tree of Life, and European Reference Genome Atlas, among others.

As a next step, the VGP will continue to work collaboratively across the globe and with other consortia to complete Phase 1 of the project, approximately one representative species per 260 vertebrate orders separated by a minimum of 50 million years from a common ancestor with other species in Phase 1. The VGP intends to create comparative genomic resources with these 260 species, including reference-free whole genome alignments, that will provide a means to understand the detailed evolutionary history of these species and create consistent gene annotations. Genome data are primarily generated at three sequencing hubs that have invested in the mission of the VGP including The Rockefeller University's Vertebrate Genome Lab, New York, U.S.; Wellcome Sanger Institute, UK; and Max Planck Institute, Germany.

Phase 2 will focus on representative species from each vertebrate family and is currently in the progress of sample identification and fundraising. The VGP has an open-door policy and welcomes others to join its efforts, ranging from fundraising and sample collection to generating

genome assemblies or including their own genome assemblies that meet the VGP metrics as part of our overall mission.

The VGP collaborated with and tested many protocols from genome sequencing companies, some of whose scientists are also co-authors of the flagship study, including from Pacific Biosciences, Oxford Nanopore Technologies, Illumina, Arima Genomics, Phase Genomics, and Dovetail Genomics. The VGP also collaborated with DNAnexus and Amazon to generate a publicly available VGP assembly pipeline and host the genomic data in the Genome Ark database. The genomes, annotations and alignments are also available in international public genome browsing and analyses databases, including the National Center for Biotechnology Information Genome Data Viewer, Ensembl genome browser, and UC Santa Cruz Genomics Institute Genome Browser. All data are open source and publicly available under the G10K data use policies.

Other novel biological discoveries from the 16 genomes in the flagship paper, and 25 genomes total from over 20 papers in this first wave of publications include:

- Corrections of false gene or chromosome losses, where previous assemblies missed between 30% to 50% of GC-rich protein-coding gene regulatory regions, which were considered to belong to the 'dark matter' of the genome;
- Newly identified chromosomes in the zebra finch and platypus;
- Complete and error free mitochondrial genomes for most species, some generated in single molecule sequences without the need for assembly;
- Wild sex chromosome evolution in monotreme mammals and birds;
- Genetic variations between humans and marmosets that have implications for marmosets as an emerging non-human primate

model system for biomedical research;
- Lineage-specific changes shaping the evolution of bird and mammal genomes: duck, emu and platypus and echidna; and
- Proposal for a universal evolution-based revised nomenclature for the oxytocin and vasotocin ligand and receptor families.

More information: Towards complete and error-free genome assemblies of all vertebrate species, *Nature* (2021). DOI: 10.1038/s41586-021-03451-0

Nature collection: www.nature.com/articles/d42859-021-00001-6

Provided by Rockefeller University