Embargo Data Use Policy for the International Reference Vertebrate Genomes Project (VGP)

by the G10K and Affiliated groups

**Last updated January 9th, 2018**

The goal of the Vertebrate Genomes Project ([VGP](https://www.genomeark.org/)) is to generate at least one high-quality, error-free, near gapless, chromosome-level, haplotype phased, and annotated reference genome assembly for all extant vertebrate species, and to utilize those genomes to advance the mission of the VGP, which include addressing fundamental questions in biology, disease, and conservation. The VGP is a project of the [G10K organization](https://genome10k.soe.ucsc.edu/), with affiliates from the Bird 10,000 genomes ([B10K](https://b10k.genomics.cn/)), [Bat1K](http://bat1k.ucd.ie/), and others in the future that will join our effort. The VGP will release raw and assembled genome and transcriptome data publicly before publishing on those genomes. To support fair and productive use of this data, the [G10K Council](https://genome10k.soe.ucsc.edu/leadership) has developed the following data use policy, which is consistent with those of genome annotation centers we are working with ([Ensembl](http://www.ensembl.org/index.html), [NCBI](https://www.ncbi.nlm.nih.gov/home/about/policies/), and [UCSC](https://ucscgenomics.soe.ucsc.edu/)) and will be enforced by journal editors (e.g. [Nature](http://www.nature.com/nature/about/editors/), [Science](http://www.sciencemag.org/about/editors-and-editorial-boards)). We ask all users to respect and follow this embargo data use policy. Our policy follows standards in genomics, and the text here was adopted from the [Sanger Institute Data Use Policy](http://www.sanger.ac.uk/about/who-we-are/policies/open-access-science).

**VGP Data Use Policy**

Before publishing on them, the G10K VGP releases raw reads, assembled genomes, transcriptome sequence data, and annotations as a service to the research community. These data releases occur through the public archives, such as [GeneBank](https://www.ncbi.nlm.nih.gov/genbank/) and [Gene Expression Omnibus](https://www.ncbi.nlm.nih.gov/geo/) (GEO) at NCBI, [European Nucleotide Archive](https://www.ebi.ac.uk/ena/about) (ENA) and [ArrayExpress](https://www.ebi.ac.uk/arrayexpress/about.html) at EMBL-EBI, the [DNA Data Bank of Japan](http://www.ddbj.nig.ac.jp/intro-e.html) (DDBJ), as well as a partnership with [DNAnexus](https://www.dnanexus.com/) and Amazon Web Services ([AWS](https://aws.amazon.com/)) in a public S3 Bucket specifically dedicated to the VGP. We encourage others to use these data, but expect them to respect our right to first presentation (including journal publications, pre-prints such as in bioRxiv, public conference talks, and press releases) of a genome-wide analysis of the data we generate, including the use of genome-wide data for phylogenetic and evolutionary analysis, on behalf of ourselves as data producers, the sample providers and collaborators. Therefore, please respect the embargo on the presentation of analyses using pre-publication data that we release via this website and the relevant archives. Exceptions to the policy are for analyses of either a single locus, or a single gene family in a species, or a maximum of 5 gene loci across multiple species, or for use as a reference for mapping reads from independent studies. Individual genomes and datasets will be considered released from this embargo when they are expressly published by members of the VGP or if released by a G10K announcement as described below. For any queries about using the data, referencing/publishing analyses based on pre-publication VGP data from this project, or joining the VGP, contact the G10K Chair, currently Erich D. Jarvis, [ejarvis@rockefeller.edu](mailto:ejarvis@rockefeller.edu), copying when appropriate the named individuals responsible within VGP for the genome(s) or questions of interest.

**Timeline of Embargo**

It is the intent of the VGP members to publish the genomes in phases, with Phase 1 being an ordinal project of species representing all vertebrate orders and divergences before and soon after the last mass extinction event 66 million years ago. We have not set a time limit on publications, as it is difficult to predict how long a phase or subproject takes to be completed. However, our intended timeline for specific phases are on the order of two years from the time of public data release. If work being conducted by the VGP project members ceases or is considered by G10K Council to take unreasonably long beyond several years according to scientific standards, then G10K Council by majority vote may release the embargo on a genome or a set of genomes. Members of the G10K Council can be found at <https://genome10k.soe.ucsc.edu/leadership>. At the time that the embargo is released due to either publication or formal G10K announcement, the relevant releases will be updated on the G10K, VGP, and annotation Center websites.

**VGP Data Use Statement**

The VGP consortium intends to use the genomic data that it produces for multiple studies. A list of studies planned for the Phase 1 ordinal VGP include:

1. Genome-scale family tree of vertebrates
2. Comparative genomics of specialized traits in each vertebrate lineage
3. Comparative genomics of convergent traits (e.g. vocal learning, flight, loss of limbs, and aquatic / terrestrial adaptations)
4. Developing universal vertebrate gene orthology and nomenclature
5. Deciphering vertebrate chromosomal genome evolution
6. Reconstruction of the common ancestor genomes of all vertebrates and of key vertebrate clades (e.g. mammals, birds, reptiles, amphibians, teleost, bony vertebrates, jawed vertebrates, and tetrapods)
7. Evolution of nucleotides to chromosomes of the human genome
8. Genetics of why some lineages are more disease resistant than others
9. Conservation genomics of endangered species sequenced
10. The genomes of all remaining Kakapo parrots on the planet
11. Genetic signatures of domestication across vertebrates
12. Genetics of sex determination and sex chromosome evolution among vertebrates
13. Brain cell type evolution and homologies using genomics and transcriptomics
14. 3-Dimensional genome structure across vertebrates
15. Consequences of the evolutionary battle between transposons and host factors

We ask that if persons are interested in using the VGP data for these and other studies to contact the G10K Chair and generators of the unpublished data for permissions. This list will be periodically updated as data and publications from the VGP are generated.

**Data Sharing by VGP members**

Data generated by VGP members requires that researchers share their data as widely and effectively as possible with the following considerations:

* **Access** – Primary data and final assemblies should be deposited in accessible repositories. Dataset accession numbers must be included in any publications.
* **Rights of Data Providers** - Researchers should be appropriately credited for their contribution to data generation and analyses.

**G10K-VGP affiliated projects**

Although the above policy is focused on the VGP, it will be applied to affiliated projects unless otherwise noted. These currently include: Bird 10,000 ([B10K](https://b10k.genomics.cn/)), [Bat1K](http://bat1k.ucd.ie/), and the Kakapo projects.