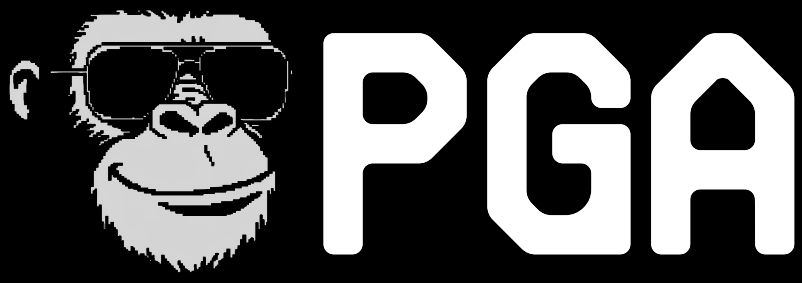

The Primate Genome-phenome Archive

Version 1.0 - July 2025

Official Documentation (v 0.9.0)

Introduction

The Primate Genome-Phenome Archive (PGA) is a comprehensive resource that maps phenotypic diversity to genetic signals across 233 primate species. It includes a collection of 263 complex traits and provides downloadable datasets that associate each trait with predicted molecular determinants inferred through comparative genomics. These predictions are based on analyses of convergent amino acid substitutions (Barteri et al., 2025) and relative evolutionary rates (RERs) (Kowalczyk et al., 2019), enabling a detailed investigation into the evolutionary basis of key traits such as brain size, lifespan, social behavior, and reproductive strategies.

The PGA builds upon the work of Valenzuela et al. (in prep., 2025), which identifies lineage-specific and convergent patterns in trait evolution throughout the primate phylogeny. The archive is openly accessible through GitHub, and all datasets are distributed in standardized, machine-readable formats suitable for integration with analytical pipelines.

The accompanying documentation is organized into clear sections to guide users through the structure and usage of the PGA. A first section outlines the distribution formats and access methods for the dataset. Subsequent sections describe the phenotypic data, detailing trait definitions, sources, and taxonomic coverage, and the genotypic tables, explaining how molecular signals were derived and linked to traits through comparative methods.

How to cite

The PGA and its associated datasets are currently part of an unpublished work by Valenzuela et al. (in preparation). A peer-reviewed publication is forthcoming. In the meantime, users are kindly asked to refrain from formal citation. For questions, collaborative opportunities, or additional information, please contact Dr. **Arcadi Navarro** at **arcadi.navarro@upf.edu**.

Support and Contact

For technical issues or to report errors, please contact **Dr. Fabio Barteri** at **fabio.barteri@upf.edu** or visit the project's forum at https://github.com/pgarchive/data/issues.

Table of Contents

Introduction – p. 2

How to Cite – p. 2

Support and Contact – p. 2

1. Distribution – p. 4

1.1 GitHub Repository – p. 4

1.2 Google Spreadsheet (phenotypic data only) – p. 4

2. Phenotypic Information – p. 5

2.1 Dataset Overview – p. 5

2.2 Trait Data Sources (trait.data.sources.tsv) – p. 6

2.3 Phenomic Dataset (nhp.phenomic.dataset.tsv) – p. 6

2.4 Quantitative Traits Summary (metadata.quantitative.traits.tsv) – p. 6

2.5 Qualitative Traits Summary (metadata.qualitative.traits.tsv) – p. 7

2.6 Trait Redundancy (traits.redundancy.tsv) – p. 7

2.7 Human Equivalent Measurements (human.equivalent.measurements.tsv) – p. 8

2.8 Discovery Groups for CAAS (caas.discovery.groups.tsv) – p. 8

3. Genetic Background Information – p. 9

3.1 RERconverge Results (ALL_RER_CONVERGE_UNIQUE_MAR23.tab) – p. 9

3.2 CAAS Results (CAAS_FINAL_MAR23.tab) – p. 10

3.3 Further Material on Request – p. 11

4. Acknowledgements – p. 11

1. Distribution

The data included in the Primate Genome-Phenome Archive (PGA) are available through multiple distribution channels to accommodate different use cases and levels of technical expertise.

1.1 GitHub repository.

The full dataset, including both phenotypic and genotypic information, is hosted on GitHub at:

https://github.com/pgarchive/data.

This repository contains standardized, machine-readable files (e.g., CSV and TSV formats) suitable for programmatic access and integration into bioinformatics workflows.

1.2 Google Spreadsheet (phenotypic data only)

A simplified version of the phenotypic dataset is also accessible via Google Sheets for easy browsing and manual inspection:

https://docs.google.com/spreadsheets/d/
1DyBKiZdaz6YW_cAj8fbupxuaTxv33SvBV8LYF2hMD0k/edit?usp=sharing

This version contains **only phenotypic traits**, excluding the genotypic data. The genetic component — which includes thousands of gene–trait associations and substitution metrics across species — is too large and complex to be effectively handled within a spreadsheet format. For full access to the genotypic data and its integration with trait information, users should refer to the GitHub repository.

Users are encouraged to refer to the GitHub repository for the most up-to-date and complete version of the archive, as well as for detailed information on data structure, file formats, and version tracking.

2. Phenotypic information

The phenotypic data included in the Primate Genome-Phenome Archive (PGA) derive from a systematic and comprehensive literature review carried out by Valenzuela et al. (in prep., 2025). This effort was especially curated by Dr. Joseph Orkin (Université de Montréal) and Dr. Borja Esteve-Altava (European Molecular Biology Laboratory – EMBL), whose contributions ensured consistency, reliability, and taxonomic breadth across all trait entries.

The dataset comprises 263 traits spanning 233 primate species, capturing a wide range of biological dimensions including behavior, ecology, morphology, physiology, and life history. Traits are provided as species-specific values, either continuous (e.g., body mass, lifespan, litter size) or categorical (e.g., nocturnality, dietary regime, arboreality). Continuous traits generally reflect average values reported in the literature, while categorical traits are assigned based on consistent descriptive observations. Each trait is associated with its original bibliographic source to ensure traceability.

2.1 Dataset overview

The phenotypic component of the PGA is organized into several structured files, available in the GitHub repository:

- **trait.data.sources.tsv** – provides the ID and name of each trait, along with the corresponding bibliographic reference.
- **nhp.phenomic.dataset.tsv** – the main phenomic matrix with species as rows and traits (by ID) as columns.
- **metadata.quantitative.traits.tsv** – summary statistics for continuous traits.
- **metadata.qualitative.traits.tsv** – summary statistics for categorical traits.
- **traits.redundancy.tsv** – number of measurements available per trait.
- **human.equivalent.measurements.tsv** – mapping of non-human traits to analogous human phenotypes in GWAS datasets.
- **caas.discovery.groups.tsv** – discovery group assignments (foreground/background) used in CAAS-based analyses.

These files serve as the backbone for comparative and evolutionary studies using the PGA. Each file is described in more detail in the following sections.

All tables are available in the GitHub repository at: https://github.com/pgarchive/data. A simplified version of the phenotypic matrix is also available for browsing as a Google Spreadsheet.

2.2 Trait data sources (trait.data.sources.tsv)

This table lists all 263 traits included in the PGA, providing for each one a unique trait ID, a human-readable name, and the corresponding bibliographic reference(s) from which the data were obtained. It serves as a master reference to trace the origin of each trait and understand the basis of its definition and inclusion.

Column 1	ID	The PGA Unique ID for the trait
Column 2	Trait	Extended trait name
Column 3	Description	Description
Column 4	Description+values	Description with files
Column 5	Source	Source (Authors, year)
Column 6	Link	Link to the publication

2.3 Non-human primates phenomic dataset (nhp.phenomic.dataset.tsv)

This is the core phenotypic matrix of the PGA, where rows correspond to primate species and columns to trait IDs. Each cell contains the value assigned to a given species for a specific trait, either as a continuous measurement or a categorical label. Traits are identified by their unique IDs, which can be cross-referenced using trait.data.sources.tsv. This table provides a standardized, species-by-trait view of phenotypic variation across the entire dataset.

Column 1	Species	The Species Name
Column 2	GroupName	Species taxon (SUPERFAMILY_family)
Columns 3 to 266	#PGA.ID	PGA ID in each column, as each column represents a specific trait.

2.4 Quantitative traits, summary statistics (metadata.quantitative.traits.tsv)

This table provides descriptive statistics for all continuous (quantitative) traits in the PGA. For each trait, it includes the number of species with available data, the mean, standard deviation, minimum, maximum, and other distributional metrics. These summaries offer a quick overview of data completeness and variability, and are useful for identifying traits with broad evolutionary ranges or limited coverage.

Column 1	ID	The PGA Unique ID for the trait

Column 2	Trait	Extended trait name
Column 3	all_covered_species	PGA ID in each column, as each column represents a specific trait.
Column 4	N_families	Number of families in the trait.
Column 5	class_traits	numeric= float, integer=integer
Column 6	mean	Mean
Column 7	sd	Standard Deviation
Column 8	min	Minimum value
Columns 9,10,11	p25, p50, p75	Quantiles
Column 12	Max	Maximum value

2.5 Qualitative traits, summary statistics (metadata.qualitative.traits.tsv)

This table summarizes all categorical (qualitative) traits in the PGA. For each trait, it lists the number of species annotated and the frequency of each category label, providing an overview of trait structure, balance across categories, and data completeness.

Column 1	ID	The PGA Unique ID for the trait
Column 2	Trait	Extended trait name
Column 3	all_covered_species	PGA ID in each column, as each column represents a specific trait.
Column 4	Type	Binary, Discrete, Multinomial*
Column 5	Categories	The distinct categories for the trait.**
Column 6		
Column 7		

* Binary = two conditions, discrete = a few numerical tags, multinomial = a few text labels.

** This columns collects all the different conditions associated with the trait. For “Seasonal Breed” for instance (PTD00013), we mention two possibilities, yes (the species breeds seasonally) or no (the species doesn’t breed seasonally). For “TrophicGuild”, which is multinomial, we have more conditions (Folivore, Folivore_frugivore, Frugivore, Gummivore, Insectivore, Omnivore).

2.6 Redundancy of the traits (traits.redundancy.tsv)

This table groups traits under broad descriptive labels (e.g., Body_size, Brain_size, Gestation) based on shared thematic scope or naming conventions. Although not organized by formal

domains, it provides a way to identify traits that may refer to related biological features or repeated measurements of similar phenomena.

Column 1	ID	The PGA Unique ID for the trait
Column 2	Ambit	Descriptive label
Column 3	Trait	Extended trait name
Column 4	nº_families	Number of families
Column 5	nº_species	Number of species

2.7 The human equivalent of Primate traits (human.equivalent.measurements.tsv)

This table links primate traits to human phenotypes. For each trait, it lists the ID, name, and the closest matching terms in causalDB and the GWAS Catalog, using standardized ontologies like UMLS and EFO, with their respective codes. It supports translational analyses and highlights potential human relevance of primate traits.

Column 1	ID	The PGA Unique ID for the trait
Column 2	Trait	Extended trait name
Column 3	Measurement	Measurement in Human
Column 4	Source	Publication mentioning the value
Column 5	Closest UMLS Human Term in causalDB	Closest UMLS Human Term in causalDB
Column 6	UMLS code	UMLS code
Column 7	Closest trait in causalDB	Closest trait in causalDB
Column 8	Closest EFO Human Term in GWAS catalog	Closest EFO Human Term in GWAS catalog
Column 9	EFO code	EFO code
Column 10	Closest UMLS Human Term in GWAS catalog	Closest UMLS Human Term in GWAS catalog
Column 11	Closest UMLS code in GWAS catalog	Closest UMLS code in GWAS catalog

2.8 Discovery groups for Convergent Amino Acid Substitutions (CAAS) analysis (caas.discovery.groups.tsv)

This table defines the discovery groups used in the detection of convergent amino acid substitutions (CAAS). Each species is assigned to either a foreground group (1) or a background

group (0) for each trait, based on phenotypic criteria. These groupings serve as input for comparative genomic analyses using tools like CAAStools (https://github.com/linudz/caastools).

Column 1	ID	The PGA Unique ID for the trait
Column 2	Species	Species name
Column 3	TOP=1/BOTTOM=0	Top/Bottom assignation (FG/BG, see caastools documentation)
Column 4	Trait	Extended trait name

3. Genetic background information

The Primate Genome-Phenome Archive (PGA) includes two large-scale datasets that describe the genetic background associated with trait evolution in primates. These tables are the result of independent computational approaches aimed at identifying gene–trait associations across the phylogeny. Both files are available in the GitHub repository under the /genetic.information/ directory. Due to their size (>100 MB), users are advised to download the files locally before consultation.

The first file, CAAS_FINAL_MAR23.tab, contains the output of an analysis performed with CAAStools, a toolkit designed to detect convergent amino acid substitutions (CAAS). For each gene–trait pair, the table reports substitution patterns consistent with convergent evolution in species sharing a given phenotype. The phenotypic groupings used for the analysis are defined in the accompanying file caas.discovery.groups.tsv, which specifies foreground and background species for each trait.

The second file consists of the raw output of RERconverge, a method that correlates relative evolutionary rates (RERs) of genes with phenotypic traits across the phylogeny. The table includes, for each gene–trait pair, the correlation coefficient, p-values, adjusted significance levels, and gene identifiers. This approach complements the CAAS analysis by identifying genes under rate shifts potentially associated with trait divergence or constraint.

Together, these resources provide a dual perspective on the molecular basis of trait evolution in primates: one focused on convergent molecular changes, and the other on evolutionary rate variation.

3.1 RERconverge results (ALL_RER_CONVERGE_UNIQUE_MAR23.tab)

This file contains the results of a genome-wide correlation analysis between relative evolutionary rates (RERs) of genes and phenotypic traits across primates. Each row represents a gene–trait association, with columns detailing statistical metrics and identifiers. These results help identify

genes whose evolutionary rates are significantly correlated with particular phenotypic traits, potentially indicating functional relevance.

Column 1	Correlation	Pearson correlation coefficient between gene RERs and the trait.
Column 2	ForegroundSpeciesCount	Number of species in the foreground group for the trait.
Column 3	PValue	Raw p-value of the correlation test.
Column 4	AdjustedPValue	Multiple testing–corrected p-value (e.g., FDR-adjusted).
Column 5	EffectSize (or SLP)	Additional statistic, possibly signed log p-value or similar effect metric.
Column 6	IsForegroundGene	Binary indicator (1/0) if gene is in foreground set (in some applications).
Column 7	GeneSymbol	HGNC symbol of the gene.
Column 8	TraitName	Trait identifier used in the analysis.

3.2 CAAS results (CAAS_FINAL_MAR23.tab)

This file contains the results of a genome-wide analysis of convergent amino acid substitutions (CAAS) performed with CAAStools. Each row represents a substitution event that occurs at a specific position in a protein alignment and is associated with a particular trait. The table includes information on the gene involved, the trait tested, the alignment position of the substitution, the evolutionary scenario identified (e.g., convergent, divergent), and the number of species in the foreground (top) and background (bottom) groups used for the analysis.

Column 1	GeneSymbol	HGNC symbol of the gene where the substitution was detected.
Column 2	TraitName	Name or identifier of the trait associated with the substitution pattern.
Column 3	Position	Amino acid position in the multiple sequence alignment.
Column 4	Scenario*	Mutational scenario detected*
Column 5	TopGroupCount	Number of species in the foreground group (trait-associated group).
Column 6	BottomGroupCount	Number of species in the background group (reference group).

* This variable indicates wether amino acid convergence is detected in both the top and bottom group (scenario1), in the top group only (scenario2) or in the bottom group only (scenario3). See CAAStools documentation, where the “scenario” is described as “Pattern”.

3.1 Further material on request

The genetic analyses included in the PGA are based on the primate genome dataset described in Kuderna et al., 2023 (Science). The multiple sequence alignments of orthologous genes used for both the CAAS analysis and the RERconverge correlations are not included in the public repository due to their size and complexity, but they are available upon request.

Researchers interested in accessing these alignments or obtaining further details about the computational methods used may contact:

- Dr. Arcadi Navarro – arcadi.navarro@upf.edu
- Dr. Fabio Barteri – fabio.barteri@upf.edu

4. Acknowledgements

This project is the result of a collaborative effort involving multiple institutions and contributors. We acknowledge the participation of:

- Arcadi Navarro (Universitat Pompeu Fabra, IBE) – Project supervisor
- David de Juan (Centro Nacional de Biotecnología, CNB) – Project supervisor
- Gerard Muntaner (Universitat Rovira i Virgili, URV) – Project supervisor
- Alejandro Valenzuela (Ace Alzheimer Center Barcelona) – First author of the companion study
- Claudia Vasallo (BarcelonaBeta Brain Research Centre, BBRC) – Human genetics integration
- Fabio Barteri (Natural Sciences Museum of Barcelona, IBE) – Co-first author, database curator, documentation author, and webmaster

We extend our sincere thanks to Dr. Joseph Orkin (University of Montreal) and Dr. Borja Esteve-Altava (EMBL) for their meticulous work in assembling and validating the phenotypic trait dataset, a cornerstone of the PGA.

We also thank Dr. Lukas Kuderna (Illumina, Inc.) for his support and for generating the foundational genomic data from 233 primate species, on which much of this work is based.

Finally, we are especially grateful to Dr. Tomàs Marquès-Bonet (Universitat Pompeu Fabra, IBE), whose support, vision, and commitment were instrumental to the success of this project — both scientifically and logistically.