Chicken Genomic Diversity consortium: large-scale genomics to unravel the origins and adaptations of chickens

Steven R Fiddaman[1*], Christophe Klopp[2], Mathieu Charles[3], Philippe Bardou[4], Ophélie Lebrasseur[5,6], Martijn Derks[7], Jens Schauer[8], Christian Reimer[8], Johannes Geibel[8,9], Almas Gheyas[10], Adrian Smith[1], Robert Schnabel[11], Maria Luisa Martin Cerezo[12], Masahide Nishibori[13], Cyrill John P. Godinez[14], John King N. Layos[15], James M. Alfieri[16,17,18], Heath Blackmon[16,17], Giridhar N. Athrey[16,18], Greger Larson[19], Ismael Ng'ang'a[20], William Muir[21], Margaret Lange[22], Dominic Wright[12], Hans Cheng[23], Henner Simianer[9], Steffen Weigend[8,9], Wesley Warren[11], Richard Crooijmans[7], Olivier Hanotte[24,25,26], Jacqueline Smith[10], Michele Tixier-Boichard[27], Laurent AF Frantz[20,28,*]

Affiliations:

1. Department of Zoology, University of Oxford, Oxford, OX1 3SZ, UK
2. INRAE, MIAT UR875, Sigenae, F-31326, Castanet Tolosan, France.
3. University Paris-Saclay, INRAE, AgroParisTech, GABI, Sigenae 78350, Jouy-en-Josas, France.
4. Université de Toulouse, INRAE, ENVT, GenPhySE, Sigenae, F-31326, Castanet Tolosan, France.
5. Centre d'Anthropobiologie et de Génomique de Toulouse (CAGT), CNRS UMR 5288, Université Toulouse III Paul Sabatier, Toulouse 31000, France
6. Instituto Nacional de Antropología y Pensamiento Latinoamericano, Ciudad Autónoma de Buenos Aires, Argentina
7. Animal Breeding and Genomics, Wageningen University and Research, Wageningen, The Netherlands
8. Institute of Farm Animal Genetics, Friedrich-Loeffler-Institut, 31535 Neustadt, Germany
9. Center for Integrated Breeding Research, University of Goettingen, 37075 Göttingen, Germany
10. The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, UK
11. Department of Animal Sciences, University of Missouri, Columbia, MO, USA
12. AVIAN Behavioural Genomics and Physiology, IFM Biology, Linköping University, Linköping 58183, Sweden
13. Laboratory of Animal Genetics, Graduate School of Integrated Sciences for Life, Hiroshima University, 1-4-4 Kagamiyama, Higashi-Hiroshima 739-8528, Japan
14. Department of Animal Science, College of Agriculture and Food Science, Visayas State University, Visca, Baybay City, Leyte 6521, Philippines
15. College of Agriculture and Forestry, Capiz State University, Burias, Mambusao, Capiz 5807, Philippines

16. Interdisciplinary Program in Ecology and Evolutionary Biology, Texas A&M University, College Station, TX, USA
17. Department of Biology, Texas A&M University, College Station, TX, USA
18. Department of Poultry Science, Texas A&M University, College Station, TX, USA
19. The Palaeogenomics & Bio-Archaeology Research Network, Research Laboratory for Archaeology and History of Art, The University of Oxford, Oxford, UK.
20. Queen Mary University of London, Bethnal Green, London, E1 4NS, UK.
21. Purdue University, Department of Animal Sciences, West Lafayette, IN, USA
22. Department of Molecular Microbiology & Immunology, University of Missouri, Columbia, MO, USA.
23. USDA, ARS, USNPRC, Avian Disease and Oncology Laboratory, East Lansing, MI, USA
24. Cells, Organisms and Molecular Genetics, School of Life Sciences, University of Nottingham, Nottingham, NG7 2RD UK
25. Centre for Tropical Livestock Genetics and Health, The Roslin Institute, Edinburgh, EH25 9RG UK
26. LiveGene, International Livestock Research Institute (ILRI), P.O. 5689, Addis Ababa, Ethiopia
27. University Paris-Saclay, INRAE, AgroParisTech, GABI , 78350 Jouy-en-Josas, France
28. Palaeogenomics Group, Department of Veterinary Sciences, LMU Munich, Germany

* corresponding authors: steven.fiddaman@biology.ox.ac.uk ; laurent.frantz@lmu.de

Author emails:

Jacqueline Smith – Jacqueline.smith@roslin.ed.ac.uk

Richard Crooijmans- richard.crooijmans@wur.nl

Martijn Derks- martijn.derks@wur.nl

Margaret Lange - langemj@missouri.edu

Michèle Tixier-Boichard: michele.tixier-boichard@inrae.fr

Mathieu Charles : mathieu.charles@inrae.fr

Christophe Klopp: christophe.klopp@inrae.fr

Philippe Bardou : philippe.bardou@inrae.fr

Jens Schauer: Jens.Schauer@fli.de; https://orcid.org/0000-0001-7303-4824

Johannes Geibel: Johannes.Geibel@fli.de; https://orcid.org/0000-0001-7172-3263

Christian Reimer: Christian.Reimer@fli.de; https://orcid.org/0000-0002-9697-2511

Steffen Weigend: Steffen.Weigend@fli.de; https://orcid.org/0000-0002-5670-2808

Ophélie Lebrasseur: ophelie.lebrasseur@univ-tlse3.fr; https://orcid.org/0000-0003-0687-8538

Dominic Wright: dominic.wright@liu.se

James M. Alfieri - jamesmalfieri@gmail.com[1,2,3]

Heath Blackmon - coleoguy@gmail.com[1,2]

Giridhar N. Athrey - giri.athrey@tamu.edu[1,3]

Consortium description

On October 25-26, 2019, a satellite meeting devoted to the preparation of a Chicken Genome Diversity Consortium was organised after the 11th European Symposium of Poultry Genetics in Prague. Researchers involved in chicken genomics from Europe, Africa and China, discussed the objectives of such a consortium with some presenting their data. However, the technical aspects of how to share and jointly analyse the data were not finalized, nor was the funding model for the cost of data storage and computation. In 2021, an opportunity arose with the call for projects of the SuperMUC computing cluster of the Leibniz-Rechenzentrum in Germany. A new consortium of scientists re-launched the discussion to establish a project with the aim to explore how the high-throughput genomics age can be harnessed to answer evolutionary questions surrounding the chicken. The FARMGENOMIC project (23826) was accepted for funding in autumn 2021, gathering around 20 members from 10 institutions in Europe, North America, and Africa. This newly formed Chicken Genomic Diversity consortium brings together members from a variety of disciplines, including genomics, palaeogenetics, animal breeding, immunology, organismal biology, evolutionary biology, and archaeology. Central to the consortium are the concepts of inclusivity and openness – all data are to be made available to all members of the consortium, and later distributed to the wider community, and collaborations between groups are fostered and actively encouraged. It is hoped this state-of-the-art resource, curated in-house by bioinformaticians, will enable the community to answer previously intractable questions in chicken evolution.

Dataset description

At the core of the consortium is a substantial genomic dataset of chickens and junglefowl. At the time of writing (September 2022), the dataset comprises 4,392 chicken and junglefowl genome sequences, of which 2,307 were derived from public databases, and the remainder provided by consortium members. In addition to domesticated chickens and red junglefowl (comprising all five subspecies: *G. g. gallus*, *G. g. bankiva*, *G. g. jabouillei*, *G. g. murghi*, *G. g. spadiceus*; total n = 291), we also included members of the congeneric *Gallus* species *G. varius* (n = 21), *G. lafayettii* (n=12), and *G. sonneratii* (n=15). Among the domesticated chickens, a wide array of geographical locations are represented (Africa, n = 1,047; East Asia, n = 856; South East Asia, n = 72; South Asia, n = 137; Middle East, n = 219; European fancy breeds, n = 462; North America, n = 835;

South America, n = 15; Oceania, n = 24) as well as commercial birds (n = 329) and experimental lines (n = 42). The wide scope of the dataset aims to capture a significant proportion of the extant genetic variation in the chicken genome. Furthermore, the addition of 15 ancient chicken genomes from Europe and the Middle East will provide supplemental time-depth, including a window into past genetic variation following the arrival of chickens into Europe from Southeast Asia.

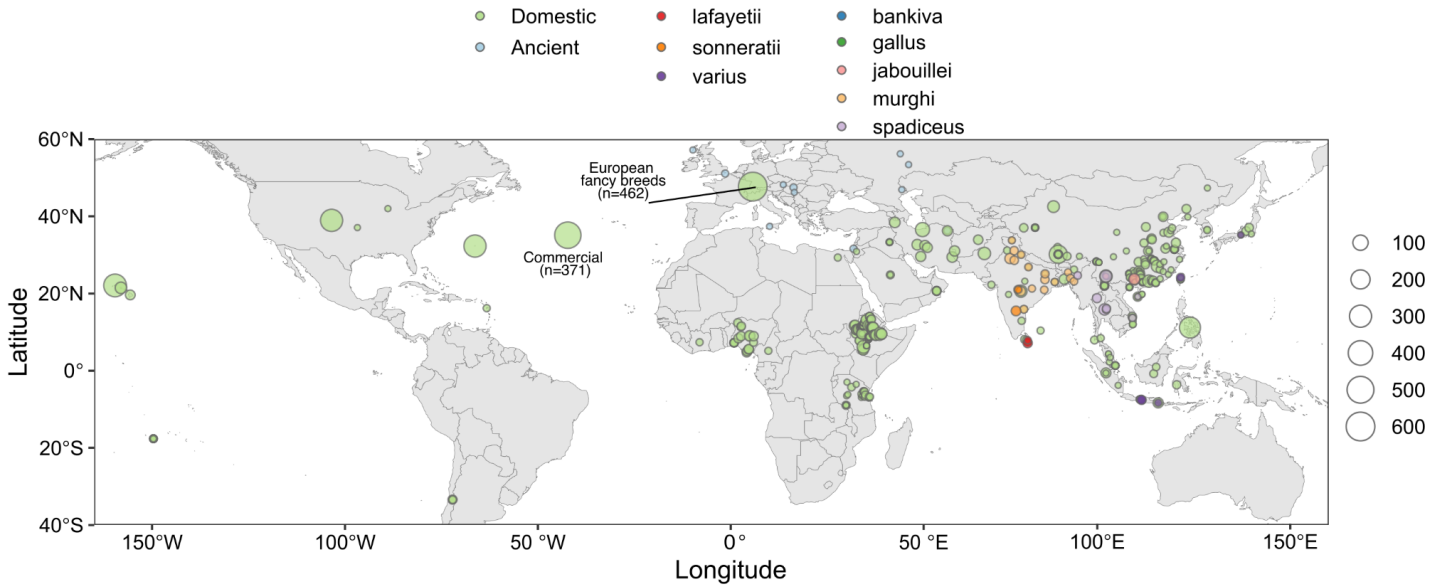

Figure 1. Sampling of global *Gallus* spp. diversity. The map shows the sampling locations for the 4,392 genomes from domestic chickens and congeneric junglefowl species. To illustrate group size, commercial birds and European fancy breeds are also included on the map, although physical sampling location is not presumed to be important for these birds.

Consortium aims

The aims of the consortium are numerous and varied, reflecting the diverse interests of the contributing groups.

Specific scientific aims include:

1. **Deleterious alleles and possible inbreeding**. Breeds with high rates of inbreeding and potential health risks will be identified based on genetic load and deleterious variants in the sequence data. We will also investigate loss-of-function variation in relation to pseudogenes and adaptation.
2. **Structural variation**. The impact of structural variants (e.g. deletions and duplications) on trait variation will be assessed. This analysis aims to produce a catalog of structural

variation and associated frequency estimates, as well as predicted functional consequences of the variants identified. We also aim to construct a graph genome from a diverse selection of breeds using a combination of long- and short-read sequencing technologies.

3. **Phenotype and trait adaptations**. We aim to identify causal gene variants that underlie adaptive traits. For instance, we are interested in covariation between genotypic variants and agro-pastoral markers to shed light on the genetic basis of adaptation to different environments. We will also investigate adaptation to phenotypic and production traits, such as feather colour and egg shell quality.
4. **Distribution of extant chicken genetic diversity**. Sequence data from such diverse geographical sources permits a detailed investigation of extant chicken diversity with respect to geographical spread. Within this investigation, finer scale analyses of diversity – particularly within the continents of Africa and Asia – are to be conducted.
5. **Evolutionary history of chickens**. The introduction of chickens into Europe (when and how many times) remains unclear. By comprehensively mapping the extant variation in chicken populations, we aim to build a high-quality reference panel for variants, which can be used to phase and impute genomes, including low-coverage ancient genomes from Europe dating to ~2000 years ago (a few centuries following the introduction of chickens in Europe). Using similar approaches, we also plan to decipher the evolutionary history of chickens in Neotropical America, in which chickens underwent a much more recent (~500 years ago) introduction.
6. **Evolutionary history of *Gallus* spp**. Combining data from domesticated chickens and congeneric junglefowl is expected to help answer questions regarding the ultimate origin of domestic chickens and the contribution of junglefowl to modern chicken ancestry.
7. **Evolution and adaptation in the immune system**. A simple prediction is that chickens have had to adapt to cope with (i) exposure to novel pathogens, and (ii) increased intensity of pathogen pressure due to increased flock size and density of rearing. This is likely to have left signatures of adaptation at immune loci of the chicken genome. Genes such as the Toll-like receptors and other pattern-recognition receptors at the front line of defense against pathogens will be investigated for signals of selection. We aim to conduct *in vitro* testing to validate bioinformatic predictions of functional change in immune receptors using methods that are well-established within the consortium.

These lines of investigation will be synthesized into several publications over the course of the consortium, led by different principal investigators depending on expertise. At the outset, the consortium has aimed to be as inclusive as possible, and as such, the studies listed above are neither exhaustive nor limited to current members of the consortium. The consortium welcomes input from any groups wishing to make the best use of this genomic resource.

Processing pipeline

In order to provide complete consistency of analysis, the entire dataset was re-processed from raw reads using a state-of-the-art mapping and processing pipeline implemented on the SuperMUC computing cluster of the Leibniz-Rechenzentrum, Bavarian Academy of Science, Germany (**Figure 2**). All reads underwent pre-processing (quality trimming, adaptor removing) with Fastp (v 0.21.0) then were mapped to the most recent version of the chicken genome (GRCg7b) (GCA_016699485.1) with BWA (v 0.7.17-r1188). The resulting BAM files from the same samples were then merged with samtools (v 1.9). For the variant calling, we generated gvcf using Elprep (v 5.1.1, compiled with go1.17; a reimplementation of GATK in GO langage). These gvcfs were integrated into a GATK genomicDB (gatk v4.2.3.0). To optimize performance, we built 48 databases corresponding to partitioning the genome into 48 intervals of equal size (~20Mb). Variant calling was performed using GATK genotypeGVCFs to obtain a VCF file by interval. We then obtained a global VCF file using GATK GatherVCFs. GATK VariantRecalibrator was then used to recalibrate variants using known SNPs.

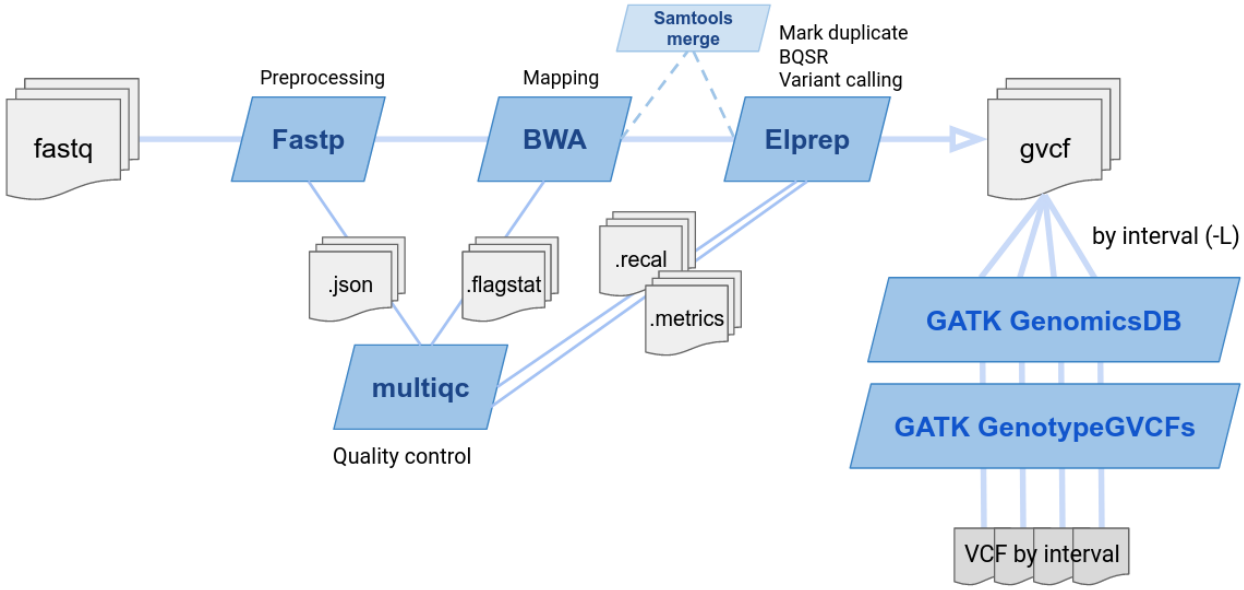

Figure 2: Processing pipeline. To ensure between-sample consistency, all samples have been re-processed from raw fastq reads. Reads underwent pre-processing and quality control before mapping to the latest version of the chicken genome (GRCg7b), variant calling, and generating a VCF.

Project timeline

The project began in 2021 and is expected to conclude (at least the first tranche of analyses) in 2023. The first phase of the project (Q3-Q4 2021) involved data gathering from both public and private sources and curation of associated metadata. In Q1 2022, the read files were quality checked to remove low quality samples and to check pre-processing from the variety of sequencing platforms included in the dataset. In Q2 2022, read mapping commenced, soon

followed by variant calling. At the time of writing (September 2022), mapping and variant calling have been completed and the VCF will shortly be made available for further analyses.

Data hosting and availability

The SuperMUC computing cluster will provide the processing power and storage capability to generate and store raw read files, alignment map files and variant call files (VCF) for the duration of the project. The final VCF will be made available in the first instance to members of the consortium, and will also be provided to the community for wider use. High quality SNPs will be made available to the community on GLOBUS, via sftp, and the European Variation Archive via the European Bioinformatics Institute.

Funding

The computational part of the project was funded by SuperMUC grant ID: 23826 awarded by The Leibniz Supercomputing Centre (LRZ), of the Bavarian Academy of Sciences and Humanities. The authors were supported by European Research Council grants (ERC-2013-StG-337574-UNDEAD and ERC-2019-StG-853272-PALAEOFARM) and by the Wellcome Trust (210119/Z/18/Z). The authors would like to acknowledge the Edinburgh Genomics Facility (Edinburgh, UK) for generation of the sequence data. This study was funded by the Bill and Melinda Gates Foundation (BMGF) and with UK aid from the UK Government's Department for International Development (Grant Agreement OPP1127286) under the auspices of the Centre for Tropical Livestock Genetics and Health (CTLGH), established jointly by the University of Edinburgh, SRUC (Scotland's Rural College), and the International Livestock Research Institute. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the BMGF nor the UK Government. We thank the CGIAR livestock program (CRP) for supporting the sampling component of the research. Sequencing data provided by INRAE teams were produced with public sources of funding, coming from INRAE, the French National Research Agency (ANR), the European Commission (SABRE project) and its HORIZON 2020 program (FEED-A-GENE, IMAGE projects). Data funded by INRAE and ANR were previously described by Tixier-Boichard et al. 2020. https://doi.org/10.20870/productions-animales.2020.33.3.4564. Sampling of data provided by FLI was funded by the German Federal Ministry of Education and Research (BMBF) via the SYNBREED project (FKZ 0315528E; www.synbreed.tum.de) and sequenced within the project IMAGE - Innovative Management of Animal Genetic Resources (www.imageh2020.eu, funded by the EU Horizon 2020 research and innovation program No 677353). Sequencing of the ancient genomes was funded by the Arts and Humanities Research Council (grant AH/L006979/1). OL is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 895107.