The Functional Annotation of Animal Genomes

Progress and challenges for our continued global effort

Peter Harrison
Genome Analysis Team Leader
EMBL-European Bioinformatics Institute

peter@ebi.ac.uk
A coordinated international action to accelerate genome to phenome research

- Coordinated international effort to provide high quality functional annotation of animal genomes it has a focus on the livestock and aquaculture community, but extends to other animals.
- Core aims of data openness, reusability, rich metadata and standardisation.
- Supporting the community and bringing researchers from different funded projects to create a harmonised rich genome to phenome resource.
A global collaboration

- Global membership, with significant focus in Europe and North America.
- Governed by a steering committee and scientific advisory board.
- Workings groups, open to all members, focussing on sampling, metadata, data analysis and communication.
- A lot of the discussion agreements focus upon the standardisation of data acquisition, analysis, presentation and data sharing.
A coordination of numerous individual projects

- Coordination and agreement between scientific groups pursuing separate (but often coordinated) grants on particular species or research areas.
- Why standardisation is so important to ensure comparability and coordination on data analysis and infrastructure.
- US and Europe groups have coordinated their on separate grants to be able to sequence more tissues and timepoints.
- Projects can build upon and compliment each other for maximal benefit.
EuroFAANG is a coordinated effort to harmonize FAANG research in Europe to create free resources that link genotype to phenotype to improve animal production and welfare.

Document genome function to understand the basis for trait variation and disease resistance in farmed fish
www.aqua-faang.eu

Identify genome features for phenotypic diversity in cattle
www.bovreg.eu

Identify genome features whose activity, during development and when facing environmental challenges, determines complex traits in chicken and pigs
www.gene-switch.eu

These projects have received funding from the European Union’s Horizon 2020 Research and Innovation Programme under the Grant agreements no 817923, 8156668 and 817998.
A particularly close relationship formed between three European FAANG Horizon funded projects to focus on:

- Collaborative development of analysis pipelines under nf-core NextFlow platform
- Hold shared training sessions to be open to more participants.
- Form comparative analysis and Pipeline development working groups
- Coordinate outreach and dissemination.
- Develop shared infrastructure and data resources

Further European projects now commenced from European Horizon funding.

Remain focussed on position as part of global FAANG
Global context

• FAANG remains global with new projects continually funded under the FAANG banner, with further new members and projects welcome to join.

Creating a Shared Vision Across Crop and Livestock Communities. FAANG is involved in this initiative that brings together crop and livestock sciences, genetics, genomics, computational and data sciences, and engineering.

The Farm Animal Genotype-Tissue Expression (FarmGTEx) Consortium aims to provide an atlas of tissue-specific gene expression and regulation in livestock species. Aligns well with the aims of FAANG, with many groups data feeding into current and future work.
Standardised assays

- Particularly important that global FAANG projects focus on same core assays and histone marks.
- Further these efforts to standardise protocols.
- Evolve with new technology, ATAC-Seq, single cell sequencing.
The Data Coordination Challenge

- Having hundreds of researchers, across multiple funded projects and from many institutions spread all over the world.

- Key to ensure reproducible and comparable research within and across different projects and species.

- G2P requires rich, validated and standardised metadata.

- High quality datasets require standardisation on data generated and analyses performed.
EMBL-EBI FAANG Data Coordination Centre

• For five years EMBL-EBI have been providing data coordination solutions and infrastructure to solve these issues:
  • Ensuring data is richly described, available and searchable.
  • Ensuring data is consistently reported and presented.
  • Facilitating standardisation of protocols and analysis methods.
  • Facilitating data openness, reusability and cross-project analysis.
What makes the FAANG data collection special

- Rich, consistent and validated metadata descriptions.
- Standardised set of core assays from contributing projects.
- Mandatory sampling and experiment protocols connected to each dataset and available with the datasets for download.
- Now moving towards standardised analysis pipelines attached to each dataset.
- A data platform that ensures data is open and FAIR.

To accelerate genome to phenome research
A full metadata solution for FAANG

- FAANG Metadata and Data Sharing working group identified the sampling, experiment and analysis questions for FAANG.
  - Requires >200 different metadata questions for different studies.
  - Constantly evolving with the community.
  - Recent changes include aquaculture, single cell sequencing and focus on developmental timepoints.

https://data.faang.org/ruleset/samples
FAANG Metadata standard aims to

- be easily understandable, highly repeatable and reusable.
- use consistent terminology
- be well structured, rich and specific.
- Terminology controlled through standardised ontologies to make downstream search and analysis more powerful.
- Enhancing reproducibility, accelerating research and enabling cross-depositor analyses to be performed.

https://data.faang.org/ruleset/samples
Not all metadata is equal

- FAANG has adopted three levels of requirement for metadata.
  - **Mandatory**: Fields are always required, validation will fail if not provided
  - **Recommended**: Should always be provided, validation will fail unless a specific ‘missing’ term is supplied and warn even if it is:
    - 'not applicable'
    - 'not collected' (i.e. will always be missing)
    - 'not provided' (i.e. may be added later)
    - 'restricted access' (i.e. it isn't missing, we just can't include it in a public document)
  - **Optional**: Useful but not required to pass validation.

- FAANG also has legacy standards for importing community data of importance to the project that does not meet its rich standards.
The FAANG data model

• Separately register animals, samples and experiments and analyses.

Organism
SAMEA6265168

BioSamples
http://phylopic.org/
The FAANG data model

- Separately register animals, samples and experiments and analyses.

http://phylopic.org/
The FAANG data model

- Separately register animals, samples and experiments and analyses

http://phylopic.org/
The FAANG data model

- Separately register animals, samples and experiments and analyses.
Detailed protocols mandatory with every submission

- Enhances reproducibility and comparative study possibilities.
- Useful reference of methodologies for future studies.
- Encourages standardisation on common protocols.
- Sampling performed more similarly across projects.

https://data.faang.org/protocol/samples
Validation and brokered submission

• Rich metadata rulesets are only useful if they are met by all submissions.

• All FAANG data goes through pre-submission validation, that blocks submission till compliant.

• Users upload metadata excel templates.

• Validation service not only highlights errors, it warns on suggested improvements such as being more specific in ontologies.

• Brokered submission to underlying archives.

• Mandatory protocol PDFs linked to every submission.

• Active helpdesk to support users with issues.

https://data.faang.org/validation/samples
Ontologies

• In biology we use ontologies to unambiguously describe a biological entity.
• Each ontology entity has a unique identifier and groups synonyms under it.
• Ontologies have hierarchy allowing for intuitive search at different levels of specificity.

endocrine system disease

Synonyms: Endocrinopathy, ENDOCRINE DISORDER NOS, Diseases, Endocrine System, Hormone disturbance NOS, Endocrine gland disease NOS, Endocrine System Diseases, ENDOCRINE DIS, Disease, Endocrine System, Hormone disorders, Diseases of Endocrine System, Disease of endocrine gland, Endocrine disturbance, Disorder of endocrine system (disorder), Endocrine Diseases and Manifestations, System Disease, Endocrine, Endocrine disorder NOS (disorder), Disease, Endocrine, Disorder of endocrine gland, Hormone disturbance,

EFO_0001379
FAANG Ontology improvement service

- Frequently ontologies are not appropriate for use in agriculture and G2P, as are medical or model organism based.

- We have released a FAANG Ontology improvement service that will community source improvements to ontologies of importance in animal agriculture.

- Allows users to list ontologies of importance, flag ontologies that need improvement or are missing.

- Users submit improvements that are forwarded to underlying ontologies for update.

https://data.faang.org/
FAANG Standardised analyses – two streams

• Centralised standardised annotation platforms:
  • Ensembl Genome Browser annotation – species processed by a standardised universal pipeline, presented uniformly and with comparative resources built in.

• Standardising individual projects/researchers analyses:
  • Standardised distributable pipelines for reproducible and comparative research. Such as the Nf-core pipelines being adopted by many FAANG projects.
  • Encouraging pipeline reuse, recoding of detailed protocols and parameters.
  • Ensuring all pipelines are publicly available with permissive licensing.
FAANG pipeline vision

- Shared development of a complete set of open pipelines across the entire set of FAANG (and associated) projects for all FAANG sequencing technologies and versions compatible with all species.
- Development based on the principles of open science, open source code and reproducible workflows and environments.
- Researchers improve reproducibility and prevent duplicated development effort by reusing or improving a common set of FAANG pipelines.
- EuroFAANG projects are exemplifying this shared development approach.
- Will never be one pipeline that fits all so it is key we capture information on pipeline used and its parameters.
Technology for standardisation widely available

Ensuring usage of identical pipelines and maximum reproducibility.

• **Workflow managers**, such as Nextflow, enable easy pipeline construction from existing components and manage data flow scalability.

• **Containerisation technologies** such as Docker, ensure an identical software environment can be distributed for the pipeline wherever it is installed.

• **Cloud technologies** that support the above ensure reproducibility as multiple projects use the same platform, accelerate research speed and lower cost.
FAANG community pipelines: GitHub and nf-core

- Standardise analysis use within and across FAANG projects to accelerate research and avoid reinvention.
- EuroFAANG projects agreed to use and contribute to nf-core where possible.
- Ensure each dataset has a link to the pipeline and config file available.

https://github.com/FAANG

https://nf-co.re/
GENE-SWitCH: Open science with modern cloud analysis

- Sample metadata
- Raw data immediately submitted
- Colocalised/synced raw data for analysis

Images: EMBL-EBI, Illumina

- Data to archives as rapidly as possible.
- Cloud analysis gets data from FAANG.
- Analyses linked to data, all under project identifier.
FAANGs standardised datasets

Rich, validated metadata

Detailed sampling and analysis protocols

FAANG dataset

Legacy contextual datasets

Publications

Analysis pipeline and parameters

Secondary analyses and annotations
FAANG Data Portal

• Single access point to all FAANG metadata and data.
• Automated identification of dataset use in publications, that links publications to datasets.
• Access to detailed protocols.
• Direct access to download all data from underlying archives.
• Sub project pages to access all of their data, e.g. EuroFAANG.

https://data.faang.org/
FAANG Data Portal

- Rich validated sampling and material preparation metadata.
- Sampling and preparation protocols.
- Primary analysis metadata, datasets and protocols.
- Analysis datasets and protocols.
- Links to pipelines and configuration used to generate each analysis.
- Automated links to publications.
- Optional track hub links to genome browsers

https://data.faang.org/
Ensembl Genome Browser FAANG annotation

• Ensembl is genomics platform for enabling and accelerating downstream science

• Produces and updates annotations for all FAANG species.

• Connects gene annotation, comparative genomics, variation and regulatory data.

• Genome Browser supports customized FAANG track hubs.

https://www.ensembl.org/
The FAANG collection

• 11811 samples from 1031 organisms across 11 species (4 more Salmonid fish species currently in submission process).

• 66 publicly available richly described datasets indexed thus far comprising 26,102 downloadable files.

• 19 FAANG dataset papers published.

http://phylopic.org/
FAANG 1.0
Establishing functional maps

Much of FAANG work to date, adding the genetic/epigenetic link to complex phenotypes
FAANG 1.0
Establishing functional maps

Much of FAANG work to date, adding the genetic/epigenetic link to complex phenotypes

FAANG 2.0
Will layer the functional maps into different genotypes in informative populations

Using powerful tools for predictive biology
To understand the biological processes driving phenotypes

Taking into account current scenarios of animal production
Including societal, ethical and environmental constraints on breeding and husbandry

New paradigms for applied genomics to tackle challenges in animal production – based on new ‘drivers’ for breeding and management

Slide adapted from Chris Tuggle
So what is next for FAANG?

From FAANG to fork: application of highly annotated genomes to improve farmed animal production

Emily L. Clark, Alan L. Archibald, Hans D. Darby2,3, Martien A. M. Groenen, Peter W. Harrison, Ross D. Houston, Christa Köhr, Sigbjørn Lien, Daniel J. Macqueen, James M. Rees, Diego Robledo, Mick Watson, Christopher K. Tuggle and Elisabeta Giulia

Introduction

The Food and Agriculture Organisation of the United Nations (FAO) reports that by the year 2050 the global human population is likely to reach 9.7 billion, rising to 11.2 billion by 2100. This population growth poses several challenges to the global food system, which will need to produce more healthy food using fewer natural resources, reducing the environmental impact, conserving biodiversity and flexibly adjusting to changing societal expectations. Meeting this demand requires environmentally sustainable improvements to farmed animal health and welfare, and of efficiency and diversi-
Challenge: Centralising and expanding infrastructure

• Scaling to more species and handling exponential growth in data.

• Requirement to develop next generation of infrastructure to support:
  • single cell atlases
  • Pangeneomes
  • biorepositories

• Improved presentation layers and web accessible genome browsers and tracks

• Automated presentation graphs.
Challenge: Multiple references

• Huge challenge for FAANG to manage multiple references, need to reanalyse past data and for researchers to agree and switch in coordination.

• Ensembl is increasing multiple breed support, such as its 11 pig breeds, and now also multiple chicken assemblies to support.

• Future advances of graph representations and scalability of genome browsers.

• Currently looking at improving support for community annotation
Challenge: Representing Community Annotation

- Supports the community to advertise Track Hubs to make it easier for researchers to utilise them within Ensembl and UCSC genome browsers.
- New version due for release in coming months.
- Need for greater support and uptake of community annotation tools such as WebApollo, that feeds community annotation into EMBL-EBI resources.
Challenge: Verification and Collation of community annotation?

- The Ensembl Automated annotation of FAANG species is powerful, but it is a generalist high throughput approach for vertebrates.
- A growing need to perform high quality manual annotation and to capture community manual/automated annotation efforts into an accessible resource.
- EMBL-EBI is currently looking for partners for a trial to take community annotation improvements to FAANG species to accept, verify and incorporate annotation.
- Expand upon EMBL-EBI Human and Vertebrate Analysis and Annotation (HAVANA) teams model organism work to evaluate FAANG species annotations to improve key areas of animal genomes.
Challenge: further harmonizing data analysis at a global level

- An indexed supported set of containerised, documented and benchmarked pipelines. E.g. FAANG GitHub and nf-core.
- Must capture pipeline used and parameters with each dataset for reproducibility.
- Potential need for cloud based preconfigured analyses to ease entry point and further standardise.
Concluding Remarks

• FAANG is producing high quality datasets supported by rich metadata, protocols and pipelines. Making them high reproducible and reusable.

• First phase of FAANG has produced functional maps of key species and continues to broaden its scope.

• A number of key challenges for the community to address as it moves to its next phase of both layering more in depth genetic and phenotypic information as well as the applicability in industry for selection, editing and improvement of animal production and welfare.
Acknowledgments

peter@ebi.ac.uk
@peterwharrison

• Alexey Sokolov (FAANG DCC Project Leader)
• Akshatha Nayak (FAANG DCC Bioinformatician)
• Raheela Aslam (DCC Web developer)
• Paul Flicek (Associate Director and Head of Ensembl)
• Entire Ensembl team

• H2020 EuroFAANG collaborators from
  • AQUA-FAANG
  • BovReg
  • GENE-SWitCH
• Members of FAANG steering committee and working groups
• Fantastic FAANG community for their continued data submission, feedback and support

https://data.faang.org/