FAANGMine: Tools for Exploring Functional Annotation of Animal Genomes

Christine G. Elsik[1], Deborah A. Triant[1], Amy T. Walsh[1],
Darren E. Hagen[2]

[1]University of Missouri

[2]Oklahoma State University

Welcome to FAANG

Announcement January 6 2023: F ble. We are still in the process of adding new RNAseq and FAANG sequence annotation tracks to the JBrowse browsers, and the end of January 2023.

About FAANGMine

The FAANG (Functional Annotation o m is "a coordinated international action to accelerate genome to phenome" and aims to generate comprehensive maps of fu of domesticated animals. Data generated by the FAANG Consortium will enhance the use of animal models in the biologica will hasten discoveries in fundamental biology, as well as those that impact human health, animal well-being and agricu e FAANG effort to have substantial impacts on science and society, the data must be accessible to scientists and students e use of the FAANG data will require data mining tools that enable easy search and retrieval.

FAANGMine is a project supported b ation to address the need for a high performance data mining resource that enables fine-grained querying and integrating the vith existing information, such as functions of known genes and research datasets. The specific aims of the project are to 1) develop FAANGMine - a high-performance data mining system that integrates genome assemblies and currently available annotation data for FAANG species, 2) extend FAANGMine by integrating new data generated by the FAANG Consortium, 3) create a FAANGMine user community that consists of students and scientists working on genetics of domesticated animal species. FAANGMine will empower animal researchers, with or without bioinformatics programming skills, to leverage the FAANG data in their research, thereby accelerating discoveries that elucidate the genetic basis of phenotypic variation.

FAANGMine v1.3 (January 2023) is available. See the FAANGMine Release Update page for details. This release includes published FAANG sequence annotation data (e.g. histone marks, open chromatin regions, CTCF binding sites, chromatin states and TADs) with associated metadata for bovine, chicken, goat, horse and pig. See the FAANGMine Data Source page for a complete data list and cited publications. FAANGMine also integrates the reference genomes with a variety of external data sources, including genes (NCBI, Ensembl), proteins (UniProt), protein families and domains (InterPro), http://FAANGMine.org

NSF

Objective: Integrate genome annotation data to enable researchers to create customized annotation datasets merged with their own data for use in downstream analysis

NCBI

FAANG
Functional Annotation of Animal Genomes

IntAct

Ensembl

BioGRID

OMIM
Human Genetics Knowledge
for the World

OrthoDB

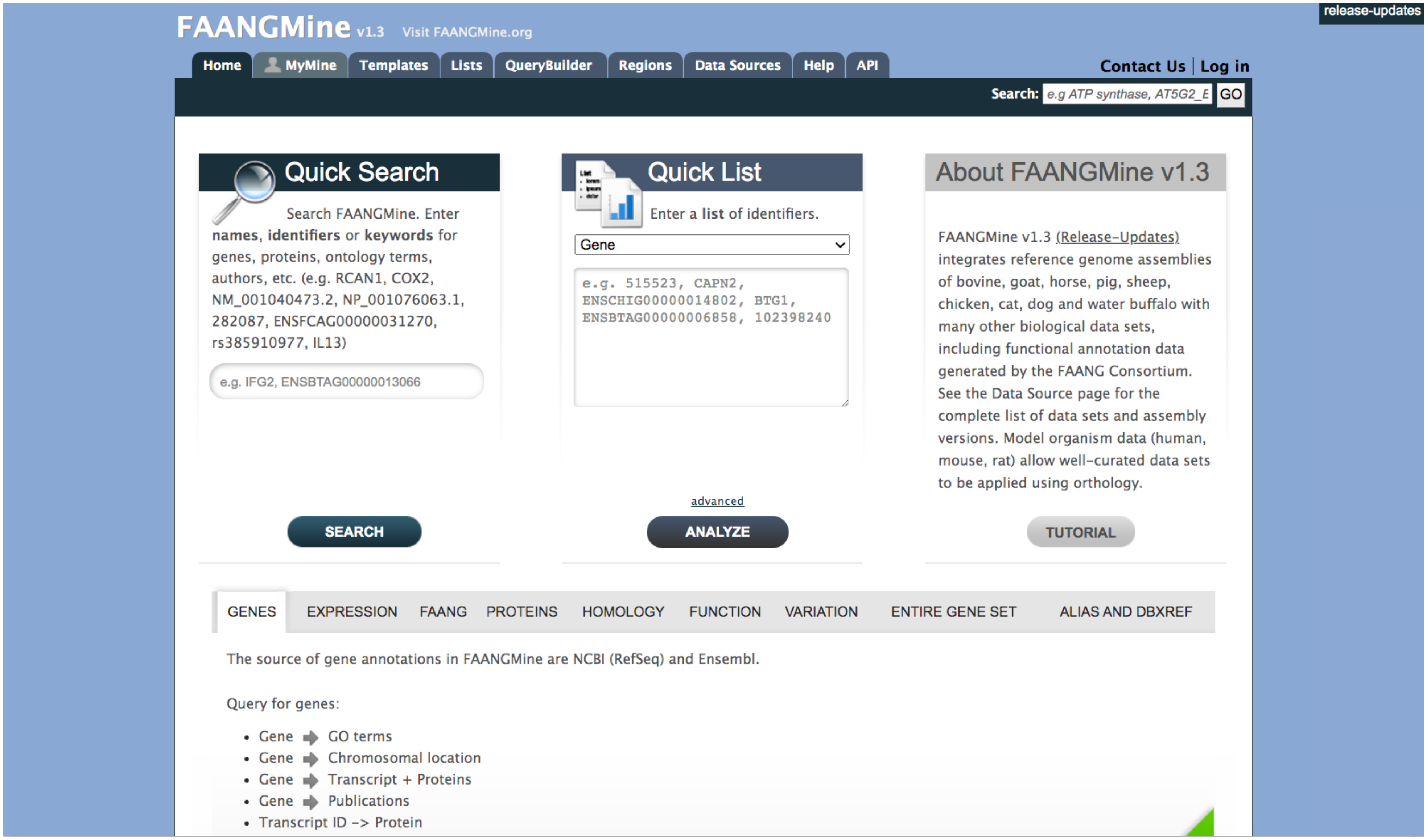
interpro
Protein sequence analysis & classification

KEGG
Kyoto Encyclopedia of Genes and Genomes

REACTOME
A CURATED PATHWAY DATABASE

UniProt

Animal QTLdb



Search FAANGMine. Enter **names**, **identifiers** or **keywords** for genes, proteins, ontology terms, authors, etc. (e.g. RCAN1, COX2, NM_001040473.2, NP_001076063.1, 282087, ENSFCAG00000031270, rs385910977, IL13)

Enter a **list** of identifiers.

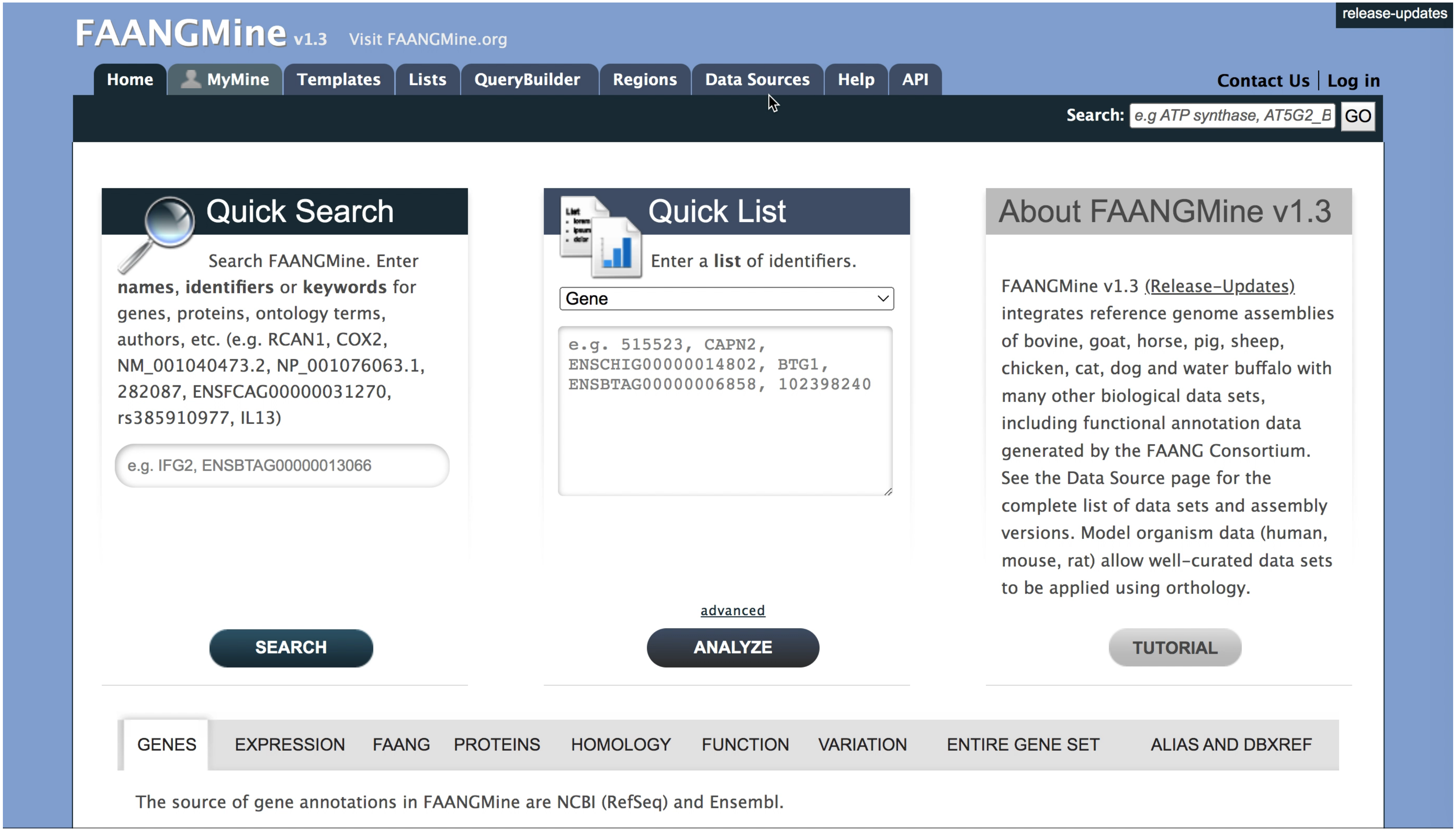
e.g. 515523, CAPN2, ENSCHIG00000014802, BTG1, ENSBTAG00000006858, 102398240

FAANGMine v1.3 (Release-Updates) integrates reference genome assemblies of bovine, goat, horse, pig, sheep, chicken, cat, dog and water buffalo with many other biological data sets, including functional annotation data generated by the FAANG Consortium. See the Data Source page for the complete list of data sets and assembly versions. Model organism data (human, mouse, rat) allow well-curated data sets to be applied using orthology.

The source of gene annotations in FAANGMine are NCBI (RefSeq) and Ensembl.

Query for genes:

- Gene ➡ GO terms
- Gene ➡ Chromosomal location
- Gene ➡ Transcript + Proteins
- Gene ➡ Publications
- Transcript ID -> Protein



Quick Search

Search FAANGMine. Enter **names**, **identifiers** or **keywords** for genes, proteins, ontology terms, authors, etc. (e.g. RCAN1, COX2, NM_001040473.2, NP_001076063.1, 282087, ENSFCAG00000031270, rs385910977, IL13)

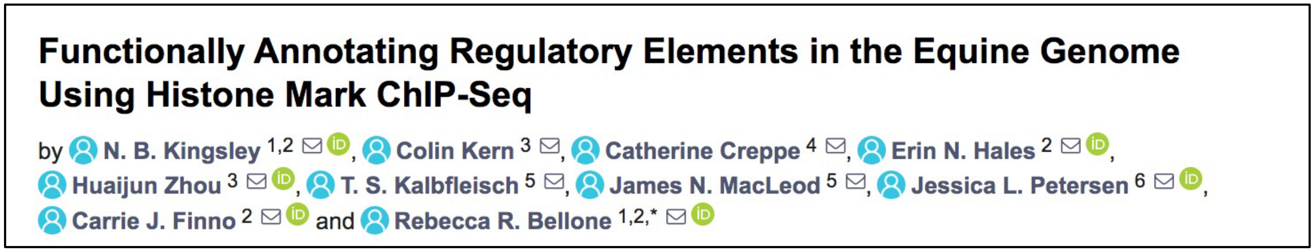
Quick List

Enter a **list** of identifiers.

About FAANGMine v1.3

FAANGMine v1.3 (Release-Updates) integrates reference genome assemblies of bovine, goat, horse, pig, sheep, chicken, cat, dog and water buffalo with many other biological data sets, including functional annotation data generated by the FAANG Consortium. See the Data Source page for the complete list of data sets and assembly versions. Model organism data (human, mouse, rat) allow well-curated data sets to be applied using orthology.

The source of gene annotations in FAANGMine are NCBI (RefSeq) and Ensembl.



Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq

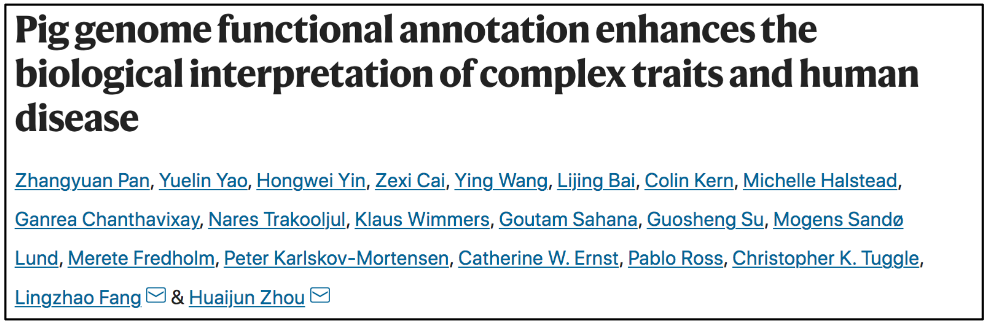
by N. B. Kingsley [1,2], Colin Kern [3], Catherine Creppe [4], Erin N. Hales [2], Huaijun Zhou [3], T. S. Kalbfleisch [5], James N. MacLeod [5], Jessica L. Petersen [6], Carrie J. Finno [2] and Rebecca R. Bellone [1,2,*]

“Adopt-a-Tissue” Initiative Advances Efforts to Identify Tissue-Specific Histone Marks in the Mare

N. B. Kingsley[1,2], Natasha A. Hamilton[3], Gabriella Lindgren[4,5], Ludovic Orlando[6], Ernie Bailey[7], Samantha Brooks[8], Molly McCue[9], T. S. Kalbfleisch[7], James N. MacLeod[7], Jessica L. Petersen[10], Carrie J. Finno[2] and Rebecca R. Bellone[1,2*]

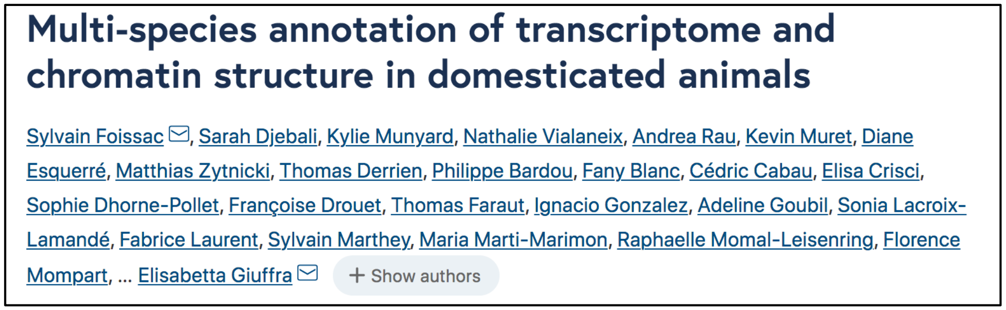
Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research

Colin Kern, Ying Wang, Xiaoqin Xu, Zhangyuan Pan, Michelle Halstead, Ganrea Chanthavixay, Perot Saelao, Susan Waters, Ruidong Xiang, Amanda Chamberlain, Ian Korf, Mary E. Delany, Hans H. Cheng, Juan F. Medrano, Alison L. Van Eenennaam, Chris K. Tuggle, Catherine Ernst, Paul Flicek, Gerald Quon, Pablo Ross & Huaijun Zhou



Pig genome functional annotation enhances the biological interpretation of complex traits and human disease

Zhangyuan Pan, Yuelin Yao, Hongwei Yin, Zexi Cai, Ying Wang, Lijing Bai, Colin Kern, Michelle Halstead, Ganrea Chanthavixay, Nares Trakooljul, Klaus Wimmers, Goutam Sahana, Guosheng Su, Mogens Sandø Lund, Merete Fredholm, Peter Karlskov-Mortensen, Catherine W. Ernst, Pablo Ross, Christopher K. Tuggle, Lingzhao Fang & Huaijun Zhou



Multi-species annotation of transcriptome and chromatin structure in domesticated animals

Sylvain Foissac, Sarah Djebali, Kylie Munyard, Nathalie Vialaneix, Andrea Rau, Kevin Muret, Diane Esquerré, Matthias Zytnicki, Thomas Derrien, Philippe Bardou, Fany Blanc, Cédric Cabau, Elisa Crisci, Sophie Dhorne-Pollet, Françoise Drouet, Thomas Faraut, Ignacio Gonzalez, Adeline Goubil, Sonia Lacroix-Lamandé, Fabrice Laurent, Sylvain Marthey, Maria Marti-Marimon, Raphaelle Momal-Leisenring, Florence Mompart, ... Elisabetta Giuffra

Goals in developing FAANG data model for InterMine

- Individual functional annotation elements (e.g. histone marks, open chromatin regions, regulatory regions) are data objects with unique identifiers and attributes
- Each element can be linked to bioproject, biosample, experiment and analysis metadata
- Annotation elements from the same sample (individual/tissue) identified in different experiments can be associated with each other and with gene expression levels in that sample

FAANG Data Curation Approach

- Biosample metadata
  - Primary source: EBI BioSamples - includes data submitted via the FAANG Data Portal which is very detailed and strict standardization
  - Secondary source: NCBI Biosamples – less detailed, standards not enforced
- Experiment metadata – NCBI SRA RunInfo files
- Analysis metadata – FAANG Data Portal
- Functional annotation data (peak files in bed format) – FAANG Data Portal, sources provided in publications
- Individual elements in peak files were assigned unique ids that contained the experiment accession

Data Curation Challenges and Solutions

Some biosamples were not submitted via the FAANG data portal

- Metadata information is incomplete
- Ontology terms must be manually curated

Challenge in tying information together for individual/tissue samples

- Individual/tissue sample used in different experiments submitted to NCBI Biosamples separately for each experiment (e.g. different biosample id for ATAC-seq, different histone marks, CTCF and RNA-seq) - does not allow tying this information together

- Solution: identify connected samples based on tissue and individual id within the NCBI Biosample report, and assign a single new identifier to connected samples called “Specimen tag”.

Challenges due to analyses not being submitted to FAANG Data Portal

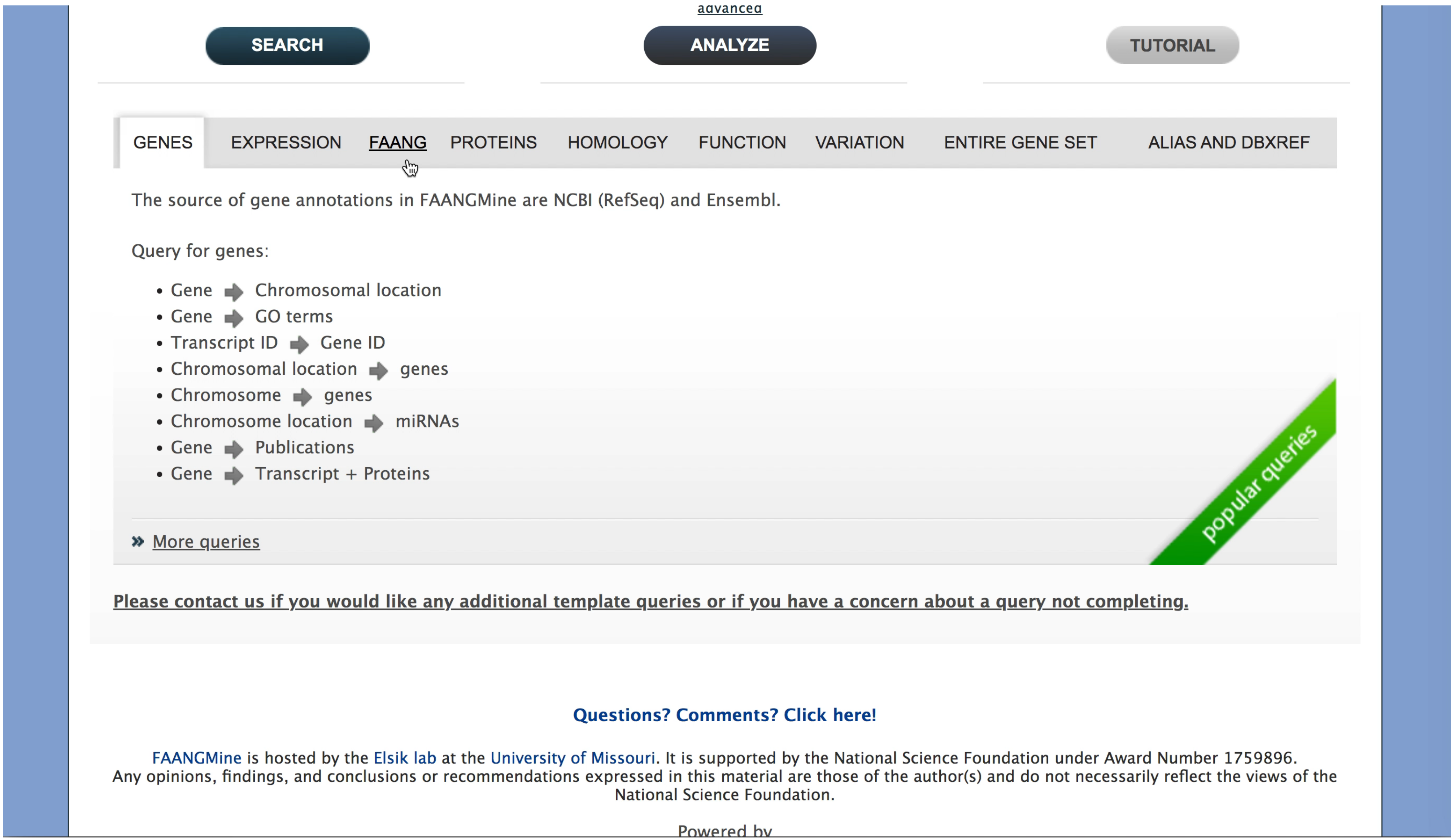
- Challenge in finding analysis data
  - Solution: Data locations were found in publications (ftp site or website)
- Analysis data did not have analysis accessions.
  - Solution: We assign a unique identifier ("Analysis tag") to each analysis.
- Individual analysis files are not connected to repository experiment or biosample accessions.
  - Solution: We manually curate experiment and biosample accessions based on file names containing tissue and individual id, or by contacting the research group

The location of ChIP-seq antibody information varies

- The ChIP-seq antibody may be provided in the NCBI Biosample report, as part of the library name in the SRA RunInfo file, or with the analysis report of the FAANG Data Portal
- Solution: The information is findable but requires extra effort

Insufficient controlled vocabulary terms

- Different publications use different terminology for chromatin states
  - Solution: We use the exact terminology provided in the publication as the “Name” of each element in FAANGMine
- All genome annotation features in InterMine must be tied to Sequence Ontology (SO) terms, but some element types lack specific SO terms
  - Solution: we assign the closest term
    - A/B compartment -> “biological region”
    - all chromatin states -> “regulatory region”
    - histone modification mark -> “histone binding site”



The source of gene annotations in FAANGMine are NCBI (RefSeq) and Ensembl.

Query for genes:

- Gene ➡ Chromosomal location
- Gene ➡ GO terms
- Transcript ID ➡ Gene ID
- Chromosomal location ➡ genes
- Chromosome ➡ genes
- Chromosome location ➡ miRNAs
- Gene ➡ Publications
- Gene ➡ Transcript + Proteins

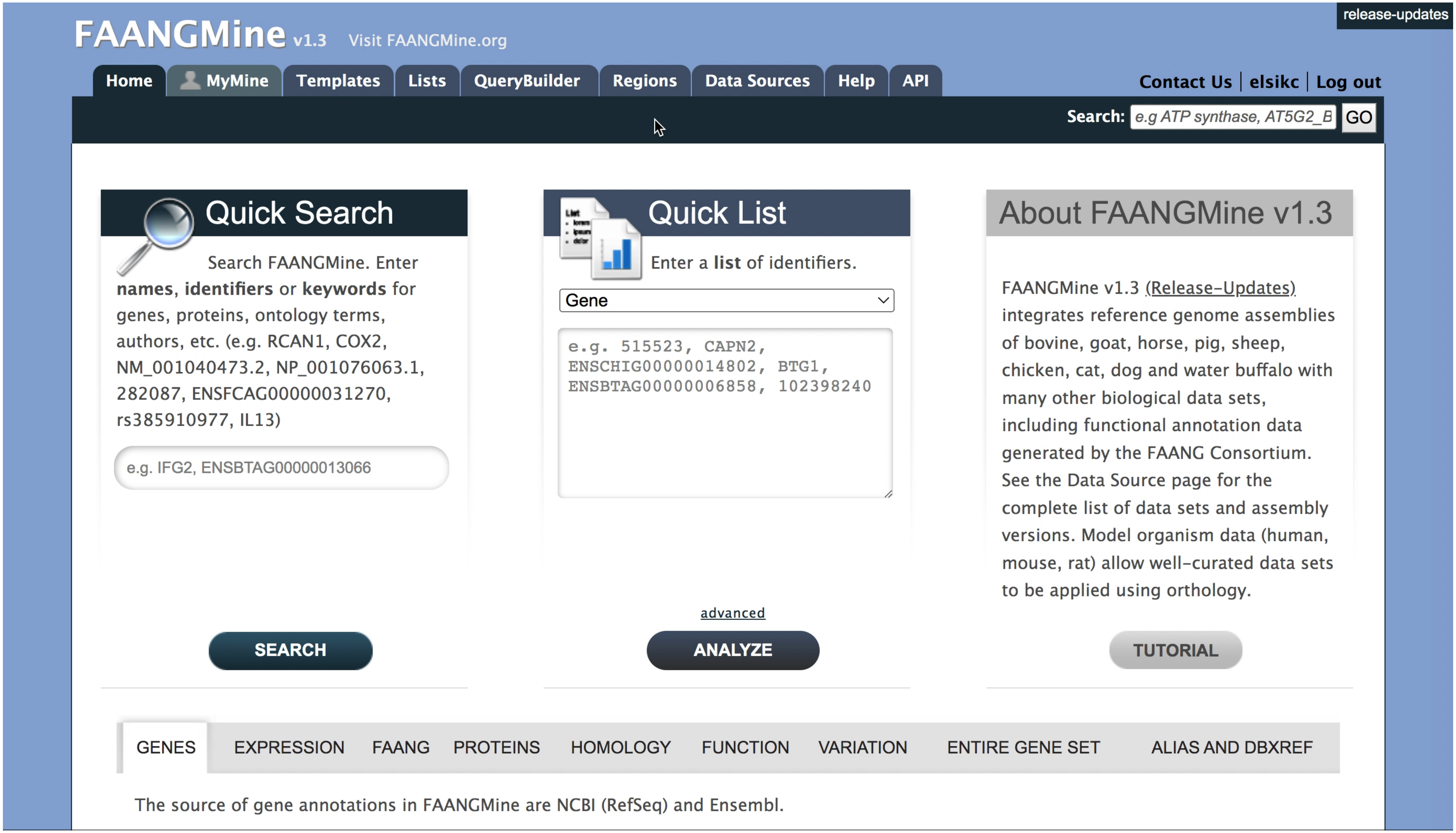
» More queries

Please contact us if you would like any additional template queries or if you have a concern about a query not completing.

Questions? Comments? Click here!

FAANGMine is hosted by the Elsik lab at the University of Missouri. It is supported by the National Science Foundation under Award Number 1759896.

Powered by



Search FAANGMine. Enter **names**, **identifiers** or **keywords** for genes, proteins, ontology terms, authors, etc. (e.g. RCAN1, COX2, NM_001040473.2, NP_001076063.1, 282087, ENSFCAG00000031270, rs385910977, IL13)

Enter a **list** of identifiers.

FAANGMine v1.3 (Release-Updates) integrates reference genome assemblies of bovine, goat, horse, pig, sheep, chicken, cat, dog and water buffalo with many other biological data sets, including functional annotation data generated by the FAANG Consortium. See the Data Source page for the complete list of data sets and assembly versions. Model organism data (human, mouse, rat) allow well-curated data sets to be applied using orthology.

The source of gene annotations in FAANGMine are NCBI (RefSeq) and Ensembl.

Acknowledgements

Developers:
Deb Triant and Amy Walsh

FAANGMine CO-PIs:
Tom Spencer, MU
Darren Hagen, Oklahoma State

FAANGMine Advisors:
Noelle Cockette, Utah State
Cathy Ernst, Michigan State
Carrie Fino, UC-Davis
James Koltes, Iowa State
Hannes Lohi, University of Helsinki
Leslie Lyons, MU
Fiona McCarthy, University of Arizona
Ben Rosen, ARS-Beltsville
Huaijun Zhou, UC-Davis

- This material is based upon work supported by the National Science Foundation under Award Number 0054449. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.