

Improving AI/ML-readiness of FaceBase Research Datasets

Carl Kesselman (PI), Rob Schuler (Speaker)

University of Southern California

AI Supplements Closeout Meeting, October 31, 2022



FaceBase: Data Hub for Craniofacial and Dental Research



“Our Face is Our Identity”

Promote multidisciplinary collaboration and research in craniofacial development, molecular genetics, and genomics

Established in 2009

FaceBase
A Resource For Craniofacial Researchers

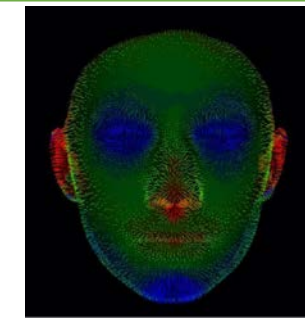
FaceBase includes...

- Datasets on normal and abnormal craniofacial and dental development in humans and animal models
- 1000+ datasets including 2D+3D imaging, sequencing, single-cell, and much more
- Worldwide userbase, 1000+ users/mo.

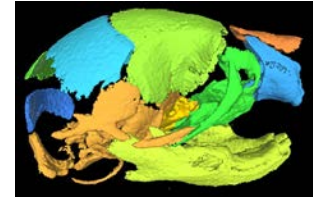
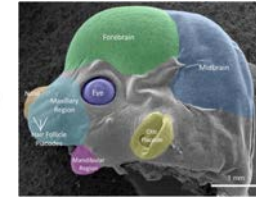
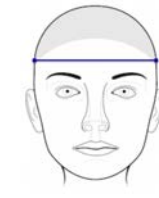
FaceBase supports craniofacial and dental researchers by...

- Providing a FAIR Data Sciences platform for the full research and analytics lifecycle: stable identifiers, versioned data, detailed metadata, standard vocabulary, and more.
- Empowering scientists with tools for viewing, annotating, and organizing data
- Promoting best practices for data sharing

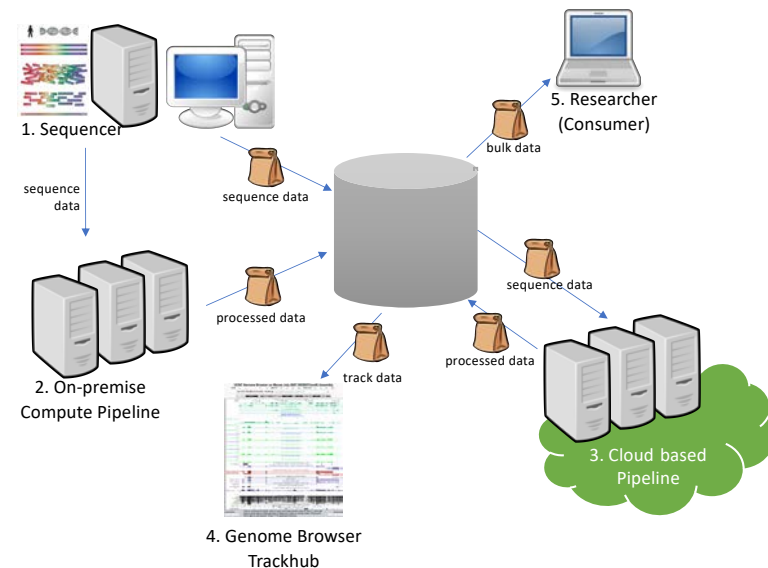
Challenge: scale curation and dissemination of data to support ML research methods



1,000s of images and genomic data on human subjects and animal models

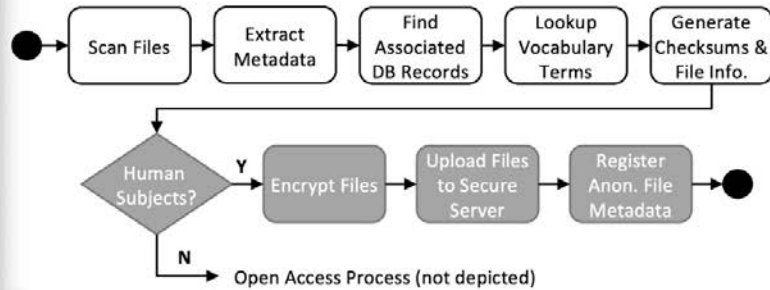
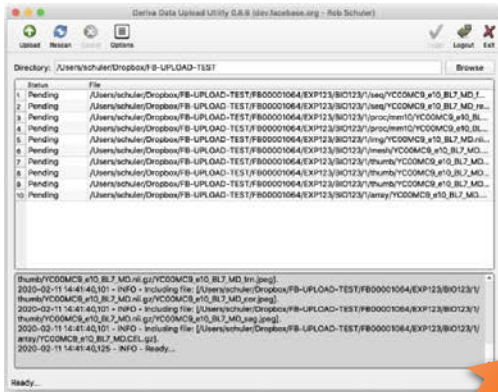


FAIR Data Sciences Platform



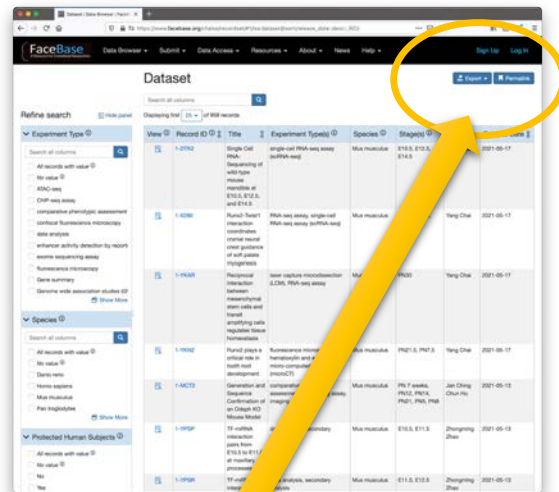
Outcomes and Future Work

Scalable Bulk Data Import and Export Pipelines, and integration with add'l community ontologies

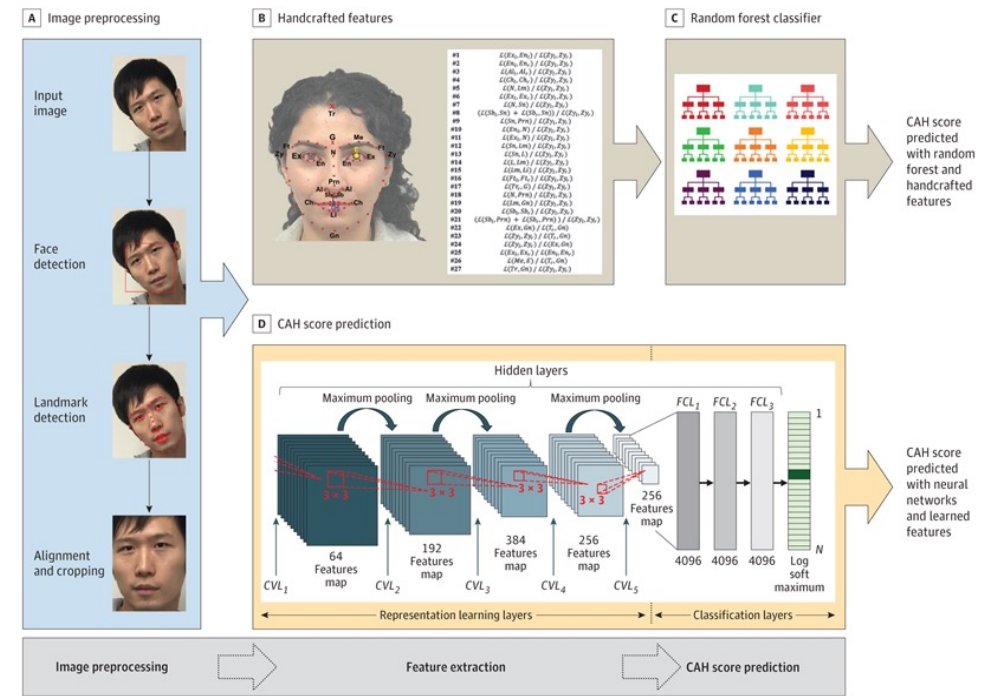


Bulk Import Pipeline Highlights

- Automated metadata extraction
- Automated vocabulary term alignment
- Automated organization and structuring of experiment and biospecimen
- Enabled automated curation of 40,000+ image dataset



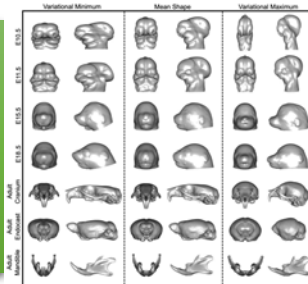
Bulk Export Pipeline integrated throughout site



Future: Evaluation of Facial Scans in ML Pipelines

Ongoing evaluation of human-subjects facial scans in a machine learning pipeline prototype based on a published ML pipeline (illustrated above) for automated phenotype prediction. (eval only)

Dataset: exemplar dataset on mouse morphology curated and released using the bulk import/export pipelines <https://doi.org/10.25550/3-HXMC>



Acknowledgments

- **FaceBase 3 Team**

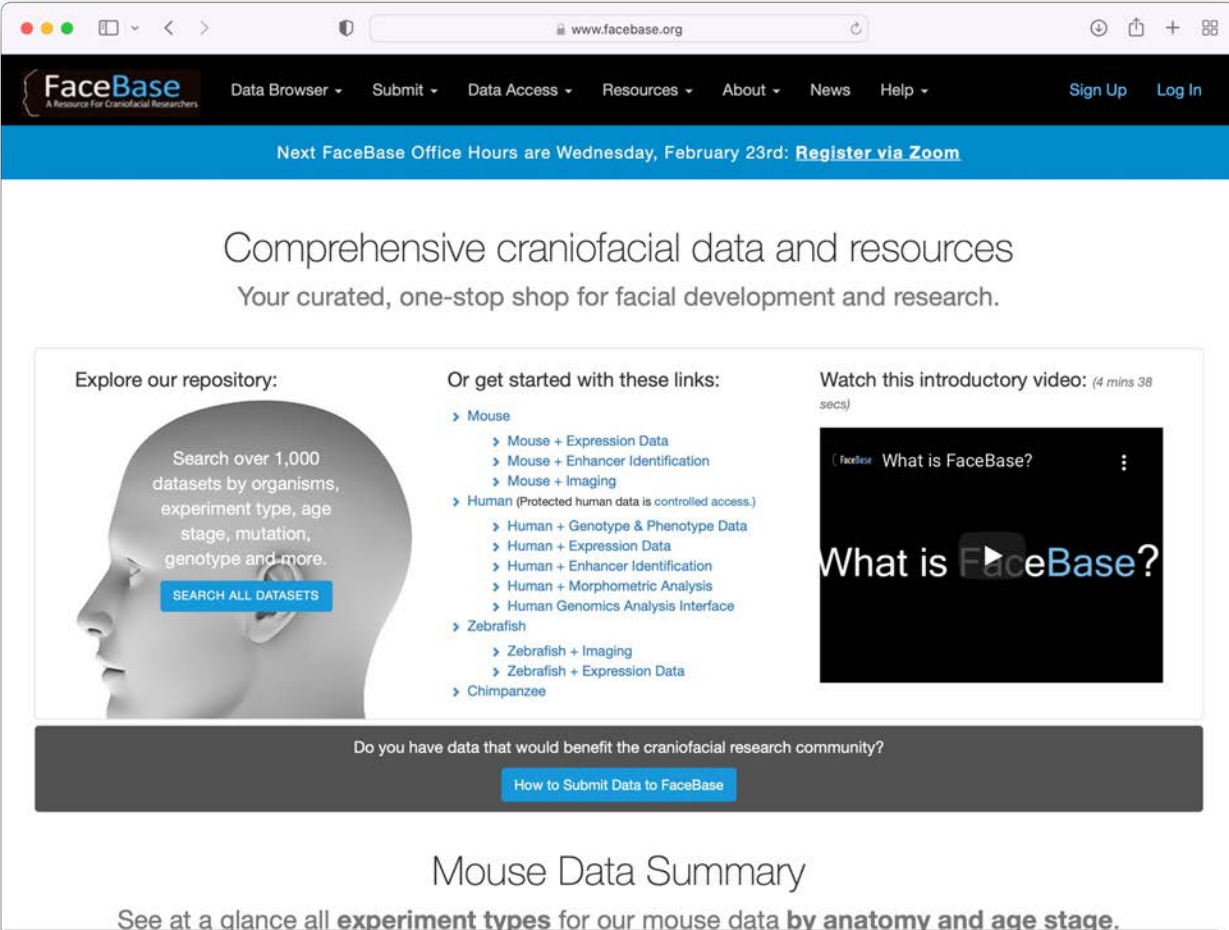
- Yang Chai (co-PI)
- Carl Kesselman (co-PI)
- Alejandro Bugacov
- Joe Hacia
- Thach Vu Ho
- Laura Pearlman
- Bridget Samuels
- Rob Schuler
- Cris Williams
- Wael Abd-Almageed (AI/ML)

- **Website:** www.facebase.org

- **Twitter:** @FaceBaseDataHub

- **Email:** help@facebase.org

- **Sponsor:** NIH / NIDCR (U01DE028729)



The screenshot shows the FaceBase website homepage. At the top, there is a navigation bar with the FaceBase logo and links for Data Browser, Submit, Data Access, Resources, About, News, Help, Sign Up, and Log In. A blue banner below the navigation bar announces the next office hours: "Next FaceBase Office Hours are Wednesday, February 23rd: Register via Zoom". The main heading reads "Comprehensive craniofacial data and resources" followed by the tagline "Your curated, one-stop shop for facial development and research." Below this, there are three main sections: "Explore our repository:" featuring a 3D head model and a "SEARCH ALL DATASETS" button; "Or get started with these links:" with a list of categories like Mouse, Human, Zebrafish, and Chimpanzee; and "Watch this introductory video: (4 mins 38 secs)" with a video player titled "What is FaceBase?". At the bottom of the main content area, there is a dark grey bar with the text "Do you have data that would benefit the craniofacial research community?" and a "How to Submit Data to FaceBase" button. Below the main content area, there is a section titled "Mouse Data Summary" with the text "See at a glance all experiment types for our mouse data by anatomy and age stage."

USC
Viterbi
Information
Sciences Institute

Center for
Craniofacial
Molecular
Biology