FACEBASE WORKSHOP:
FAIR DATA FOR CRANIOFACIAL RESEARCH

Robert Schuler
43rd Annual SCGDB Meeting
October 20, 2020
Comprehensive resource on dental and craniofacial research -- fostering data sharing and curated resources.

Transitioned from a “hub and spoke” consortium to a community-wide effort in 2019.

Comprised of a core team of craniofacial and computer science expertise w/ advisory and experts groups for oversight.
Growing, Multi-faceted Digital Repository

942 Datasets and growing
- 820 mouse, 82 human, 39 zebrafish, 3 other
- 634 imaging, 143 sequencing, and other experiment types
- 800+ detailed experiment records
- 3,500~ detailed biosample records
- 5,000+ images, 3,700+ sequencing, 2,270+ tracks, and more (~10 TB)

12 Month Usage
- 16,000+ visitors, 63,900~ views
- 2,000+ downloads
- 7,400+ image views
- 77,000+ track reads

13 new projects contributing data from FB3:Y1

Serves a worldwide community: 50% US + 50% International (traffic)
Mission to Improve the Stewardship of Research Data

Science is dependent on generating and analyzing data, but...

- 80% of time spent on accessing, cleaning, integrating data
- Scarcity of data sharing
- 10% reproducibility of data
- Recent high-profile retractions in COVID-19 research, for example.
Findable Accessible Interoperable Reusable Data

- **Organization** of data files to understand what and how they were produced
- **Terminology** from community standards with broad acceptance
- **Description** of the experimental methods and biological materials
- **Protocols** that are enumerated or referenced for precise details
Organization of Data on FaceBase

- **Projects** contribute **Datasets**

- **Datasets** organize **Experiments**, **Biosamples**, and **Data**

- **Protocols** give additional details on the materials and procedures conducted by Experiments

FaceBase data model (high-level view)
Organization of Data on FaceBase

- **Projects** contribute **Datasets**
- **Datasets** organize **Experiments**, **Biosamples**, and **Data**
- **Protocols** give additional details on the materials and procedures conducted by Experiments
- **Search generally begins at the Dataset**
Search and Refine

- Refinements by experiment type, species, gene, age, anatomy, etc.
Search and Refine

- Refinements by experiment type, species, gene, age, anatomy, etc.
- Summary view with key properties listed.
Summary of Mouse Datasets

Mouse Data Summary
Click a cell or label to see the related datasets.

Headings and Cells link directly to the Search interface
Explore By Anatomical Region of Interest

Linked to standard terminology

Surface mesh viewer with each structure coded to the terminology

Links above drive search results of related datasets

“Explore” to go deeper into the search and refine actions
Overview of Dataset Display

- Summary
- Visualization
  - Genome Browser
  - Surface Models
  - Orthoslice Views
  - Images
- Experiments & Biosamples
- Data Files
  - Sequencing, Processed, Imaging, Array, Supplementary, other...
**Dataset:** ChIP-seq of multiple histone marks and RNA-seq from embryonic face subregions

<table>
<thead>
<tr>
<th>Identifiers</th>
<th>Record ID: TMJ</th>
<th>Accession: FB0000806.2</th>
<th>DOI: 10.25559/TMJ/</th>
<th>(Released: 2017-03-30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>ChIP-seq, RNA-seq and transgenic assays to identify non-coding regulatory elements (enhancers) active during craniofacial development</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project</td>
<td>Genomic and Transgenic Resources for Craniofacial Enhancer Studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contributors</td>
<td>Yang Chai, David FitzPatrick, Steven Liigo, Len Perucchio, Edward Rubin, Axel Visel, Yoko Yuzawa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keywords</td>
<td>(The following keywords are associated with this dataset.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment Type</td>
<td>ChIP-seq assay, RNA-seq assay</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Species</td>
<td>Mus musculus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage</td>
<td>E11.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary of the Dataset Contents

**Dataset:** ChIP-seq of multiple histone marks and RNA-seq from embryonic face subregions

**Identifiers:**
- **Record ID:** TMJ
- **Accession:** FB000000806.2
- **DOI:** 10.25550/TMJ
  - (Released: 2017-03-30)

**Description:**
ChIP-seq, RNA-seq and transgenic assays to identify non-coding regulatory elements (enhancers) active during craniofacial development.

**Project:**
Genomic and Transgenic Resources for Craniofacial Enhancer Studies

**Contributor(s):**
Yang Chai, David FitzPatrick, Steven Liago, Len Permacchio, Edward Rubin, Axel Visel, Yoko Yuzawa

**Keywords:**
- (The following keywords are associated with this dataset.)
- ChIP-seq assay, RNA-seq assay

**Experiment Type:**
ChIP-seq assay, RNA-seq assay

**Species:**
Mus musculus

**Stage:**
E11.5
### Dataset: ChIP-seq of multiple histone marks and RNA-seq from early cranium and face subregions

<table>
<thead>
<tr>
<th>Identifiers</th>
<th>Record ID: TMJ</th>
<th>Accession: FB000000806.2</th>
<th>DOI: 10.25550/TMJ/13609</th>
<th>[Released: 2017-03-30]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>ChIP-seq, RNA-seq, and transgenic assays to identify non-coding regulatory elements (enhancers) active during craniofacial development</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project</td>
<td>Genomic and Transgenic Resources for Craniofacial Enhancer Studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contributor(s)</td>
<td>Yang Chai, David Fitzpatrick, Steven Liogo, Len Permacchio, Edward Rubin, Axel Visel, Yoko Yuzawa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keywords</td>
<td>(The following keywords are associated with this dataset.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experiment Type</td>
<td>ChIP-seq assay, RNA-seq assay</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Species</td>
<td>Mus musculus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage</td>
<td>E11.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Persistent Identifiers (e.g., DOI)

- **Record ID**: TMJ
- **Accession**: FB000000806.2
- **DOI**: 10.25550/TMJ/13609

#### Attribution of contributors

- Yang Chai
- David Fitzpatrick
- Steven Liogo
- Len Permacchio
- Edward Rubin
- Axel Visel
- Yoko Yuzawa
Summary of the Dataset Contents

**Dataset:** ChIP-seq of multiple histone marks and RNA-seq from embryo face subregions

- **Identifiers:** Record ID: TMJ, Accession: FB0000806.2, DOI: 10.25550/TMJ (Released: 2017-03-30)
- **Description:** ChIP-seq, RNA-seq and transgenic assays to identify non-coding regulatory elements (enhancers) active during craniofacial development
- **Project:** Genomic and Transgenic Resources for Craniofacial Enhancer Studies
- **Contributor(s):** Yang Chai, David FitzPatrick, Steven Liago, Len Permacchio, Edward Rubin, Axel Visel, Yoko Yuzawa
- **Keywords:** The following keywords are associated with this dataset.
  - Experiment Type: ChIP-seq assay, RNA-seq assay
  - Species: Mus musculus
  - Stage: E11.5

**Persistent Identifiers (e.g., DOI)**

**Keywords from standardized terminology**

**Attribution of contributors**
Embedded Visualization
Integrated Genome Browser Track Hub

FaceBase disseminated on Genome Browser

Tracks linked back to FaceBase
Orthogonal Slice Viewer with Volume Rendering

Orthoslice View

Web-based Viewer

3D Volume View
Highly Detailed Surface Models

- Surface models may contain any number of anatomical structures
- Linked to underlying (meta)data
- Linked to standard vocabulary (ZFA, MA, UBERON,...)

60+ anatomical structures linked to standard terminology in Zebrafish atlas (pictured)
Upcoming Visualization Features

High-resolution Microscopy Images with Annotations

Single Cell Browser through Integration with UCSC
Details on Experiments, Biosamples, etc.

Further drill-down to related biosamples, files, and other related data.
Export (i.e., Download) Dataset

Dataset: ChIP-seq of multiple histone marks and RNA-seq from embryonic face subregions

- Identifiers: Record ID: TMJ
- Accession: FB00000806.2
- DOI: 10.25559/TMJ
- (Released: 2017-03-30)
- Description: ChIP-seq, RNA-seq and transgenic assays to identify non-coding regulatory elements (eQTLs) active during craniofacial development
- Project: Genomic and Transgenic Resources for Craniofacial Enhancer Studies
- Contributor(s): Yang Chai, David FitzPatrick, Steven Liago, Len Pennacchio, Edward Rubin, Axel Visel, Yoko Yuzawa
- Keywords: (The following keywords are associated with this dataset)
- Experiment Type: ChIP-seq assay, RNA-seq assay
- Species: Mus musculus
- Stage: E11.5

Exporting BAG

Your request is being processed...
You will be prompted to download the file when it is ready.
“Big Data Bag” Format for Large Datasets

- Spreadsheets (CSV)
- File Manifest (links + checksums)
- Client tool for reliable download
Detailed gene summaries
Interactive plots of GWAS data
Facial norms measures

More at the Resources Hub...

https://www.facebase.org/resources/
Why Share Your Data?

• Dedicated focus on craniofacial and dental
• Increase the visibility and impact of your research
• Cross-reference with publication
• Data are “published” like first-class academic works
• Satisfy data sharing requirements for grants and publications
Data Citation

- **Persistent actionable identifiers for all database entities**
- **Explore any historical version of entire database**

Follows leading publisher’s recommendations

Import into reference managers

Persistent, actionable identifiers

Versioned or “live”

Digital Object Identifiers
Empower Contributors to Submit Valuable Data

1. Online forms with multi-record entry and edit
2. Graphical desktop applications
3. Command-line, Python and Web services interfaces for custom data submission workflows (not pictured)

Researchers have submitted their own datasets with 100s to 1000s of files, usually in a few days.
Quality Control Dashboards

• Automated Quality Control Rules to check each dataset

• QC issues are displayed on project pages (visible only to contributors)
Streamlined Process for Data Submission

Timeline (approximate)
- T0: Form submitted
- T+2 weeks: Review decision
- T+3 weeks: Project setup
- T+5 weeks: Submit data*
- T+6 weeks: QC review

* Based on user averages

IRB Certification Process
- Individual level data classified as human subjects
- Requires USC certification of your IRB decision
- Tracks are not considered restricted data
- Timeline Varies

User Activity
Hub Activity
A Test of the Reproducibility of FaceBase Data

**Evaluation:** 3 labs, 13 datasets, on 192 RNA-Seq or ChIP-Seq experiments (1000+ data files)

a) Uniform processing pipeline
b) Output strongly correlated with researcher’s results

**Diagram:**
- 1. Sequence
- 2. On-premise Compute Pipeline
- 3. Cloud-based Pipeline
- 4. Genome Browser Trackhub
- 5. Researcher (Consumer)

**Comparison of RNA-seq log2 counts from FaceBase vs JGI pipeline**

**Comparison of ChIP-seq peaks from FaceBase vs LBNL pipeline**

**Table:**

<table>
<thead>
<tr>
<th></th>
<th>H3K27ac</th>
<th>H3K27me3</th>
<th>H3K4me1</th>
</tr>
</thead>
<tbody>
<tr>
<td>lnps</td>
<td>np</td>
<td>dp</td>
<td>mx</td>
</tr>
<tr>
<td>md</td>
<td>1</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>mnp</td>
<td>1</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>mx</td>
<td>1</td>
<td>12</td>
<td>16</td>
</tr>
</tbody>
</table>

**Log2 counts of T6IF1FR RNA-seq**

Pearson correlation = 0.99

TPM JGI

TPM FaceBase

**Graph:**
- RNA-seq log2 counts
- ChIP-seq peaks (overlap top-2000)
Thank You

- **Co-PIs:** Carl Kesselman; Yang Chai
- **Core Team:** Rob Schuler (CS & Technical); Bridget Samuels (Biocuration); Alejandro Bugacov (Data science); Cris Williams (Communications); Joe Hacia (Bioinformatics); Thach Vu Ho (Data curation)

- **Website:** [www.facebase.org](http://www.facebase.org)
- **Sponsor:** NIH / NIDCR (U01DE028729)