applied genomics

BIOL5382 - Applied Genomics Laboratory Course

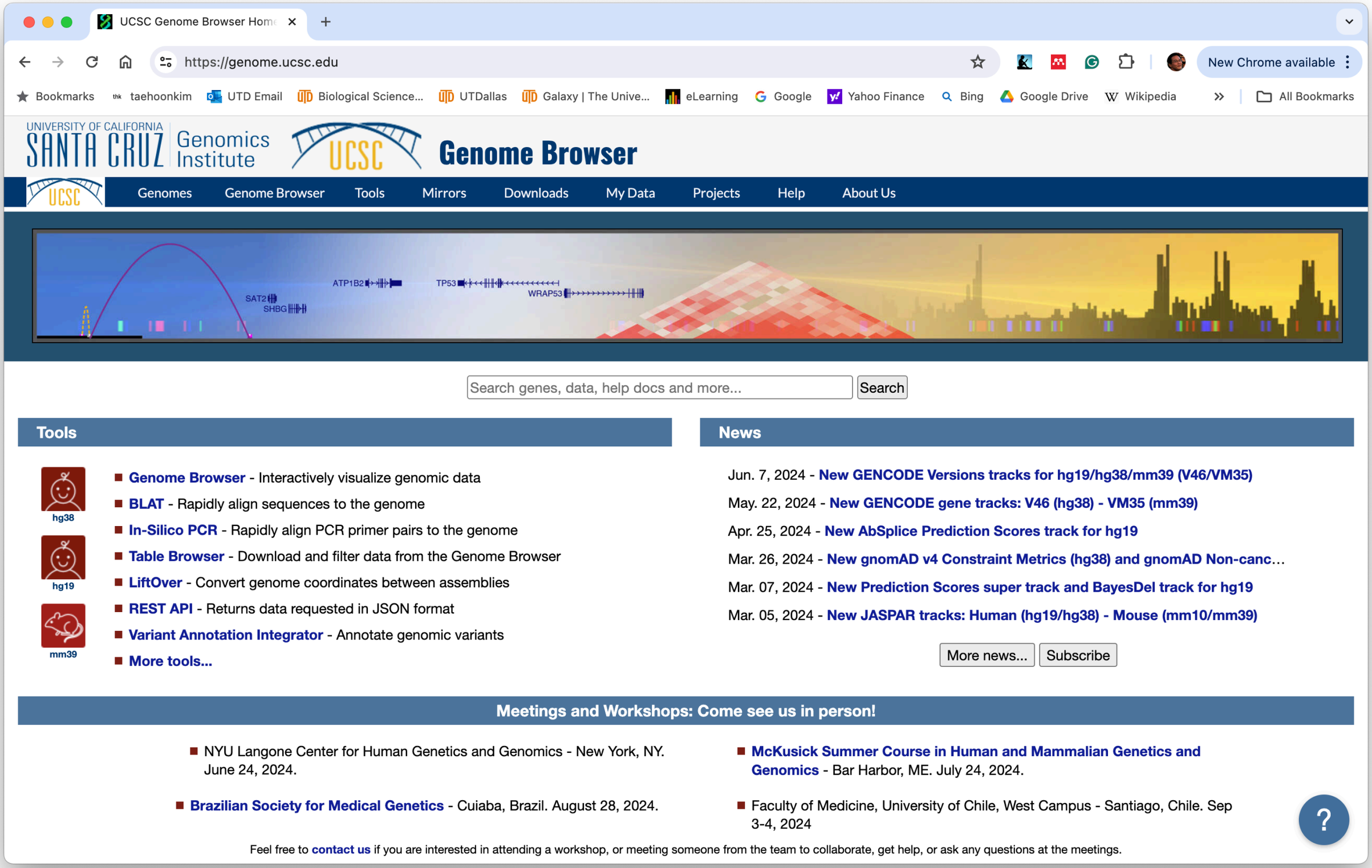
Tae Hoon Kim, Ph.D.
genome@utdallas.edu
http://taehoonkim.org

databases

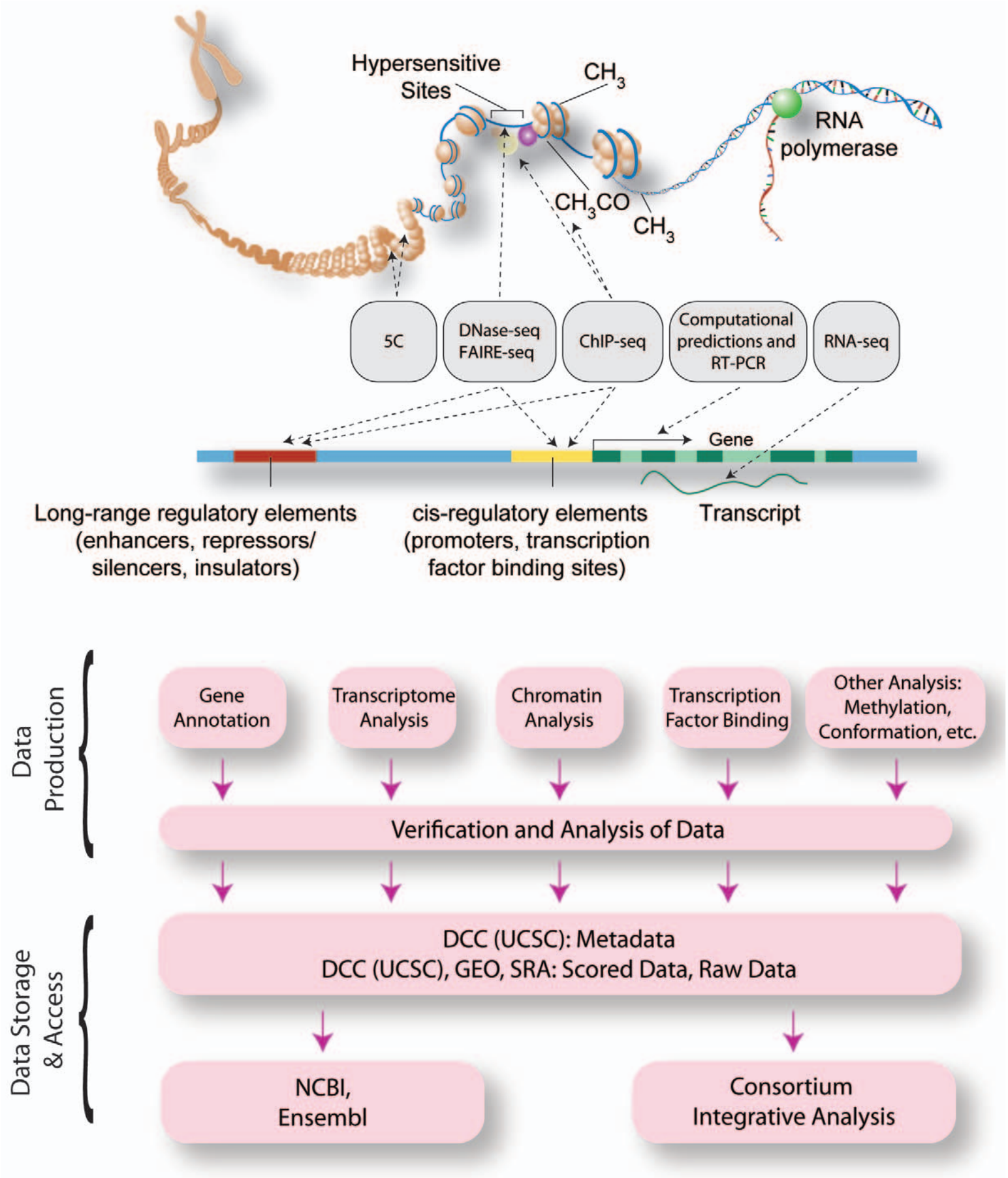
- *how to access genomic information*
- *UCSC genome browser*
  - *case story for TBXT gene evolution*

- *EnsEMBL*
- *GTEx (genome tissue expression)*

https://genome.ucsc.edu



https://news.ucsc.edu/2015/06/genome-anniversary.html


Hypersensitive Sites
CH_3
RNA polymerase
CH_3CO
CH_3
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript
Data Production
Gene Annotation
Transcriptome Analysis
Chromatin Analysis
Transcription Factor Binding
Other Analysis: Methylation, Conformation, etc.
Verification and Analysis of Data
Data Storage & Access
DCC (UCSC): Metadata
DCC (UCSC), GEO, SRA: Scored Data, Raw Data
NCBI, Ensembl
Consortium Integrative Analysis

genome browsing

- *specific genes*
- *specific genomic regions*
- *first pass information on annotations (genes, repeats)*
- *expression across tissues*
- *conservation*
- *variation*
- *regulation*

The international journal of science / 29 February 2024

nature

TALE OF TAILS

How a genetic element aided tail loss in humans and apes

Article

On the genetic basis of tail-loss evolution in humans and apes

https://doi.org/10.1038/s41586-024-07095-8

Received: 14 September 2021

Accepted: 19 January 2024

Published online: 28 February 2024

Open access

Check for updates

Bo Xia[1,2,3,4]✉, Weimin Zhang[2,10], Guisheng Zhao[1,2,10], Xinru Zhang[3,5,10], Jiangshan Bai[3], Ran Brosh[2], Aleksandra Wudzinska[2], Emily Huang[2], Hannah Ashe[2], Gwen Ellis[2], Maayan Pour[1,2], Yu Zhao[2], Camila Coelho[2], Yinan Zhu[2], Alexander Miller[6], Jeremy S. Dasen[6], Matthew T. Maurano[2,7], Sang Y. Kim[7], Jef D. Boeke[2,8,9]✉ & Itai Yanai[1,2,8]✉

The loss of the tail is among the most notable anatomical changes to have occurred along the evolutionary lineage leading to humans and to the 'anthropomorphous apes'[1–3], with a proposed role in contributing to human bipedalism[4–6]. Yet, the genetic mechanism that facilitated tail-loss evolution in hominoids remains unknown. Here we present evidence that an individual insertion of an Alu element in the genome of the hominoid ancestor may have contributed to tail-loss evolution. We demonstrate that this Alu element—inserted into an intron of the *TBXT* gene[7–9]—pairs with a neighbouring ancestral Alu element encoded in the reverse genomic orientation and leads to a hominoid-specific alternative splicing event. To study the effect of this splicing event, we generated multiple mouse models that express both full-length and exon-skipped isoforms of *Tbxt*, mimicking the expression pattern of its hominoid orthologue *TBXT*. Mice expressing both *Tbxt* isoforms exhibit a complete absence of the tail or a shortened tail depending on the relative abundance of *Tbxt* isoforms expressed at the embryonic tail bud. These results support the notion that the exon-skipped transcript is sufficient to induce a tail-loss phenotype. Moreover, mice expressing the exon-skipped *Tbxt* isoform develop neural tube defects, a condition that affects approximately 1 in 1,000 neonates in humans[10]. Thus, tail-loss evolution may have been associated with an adaptive cost of the potential for neural tube defects, which continue to affect human health today.

a personal story behind the science

https://podcasts.apple.com/us/podcast/bo-xia-and-a-tale-of-tails/id1563415749?i=1000647438287

Start with the Brachyury gene

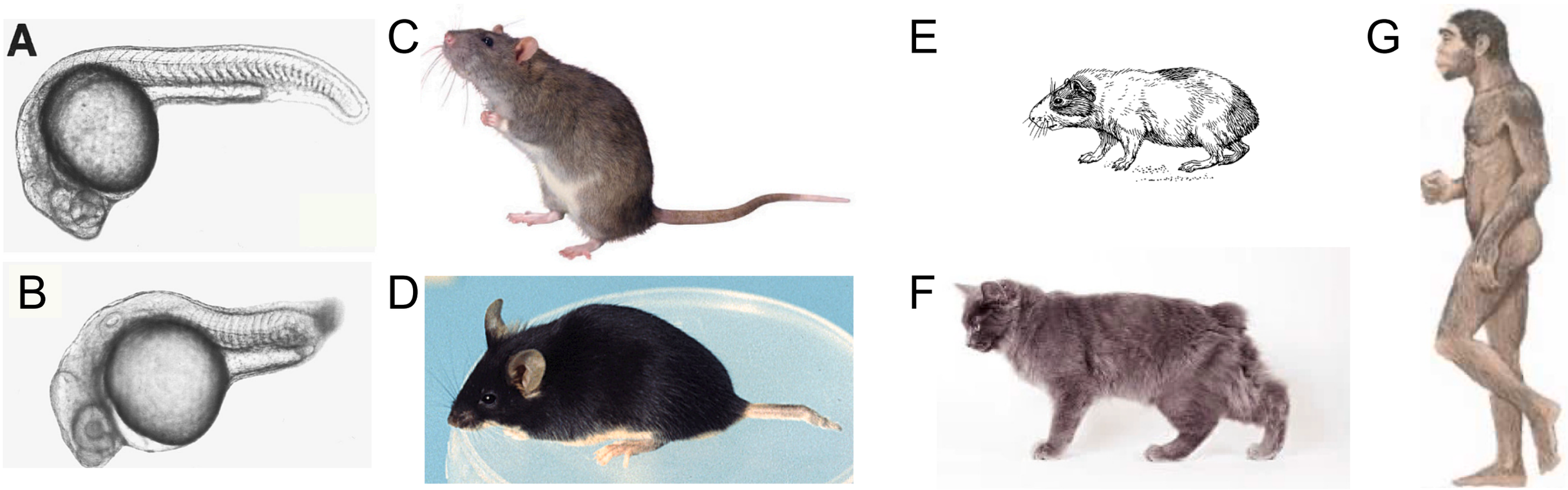
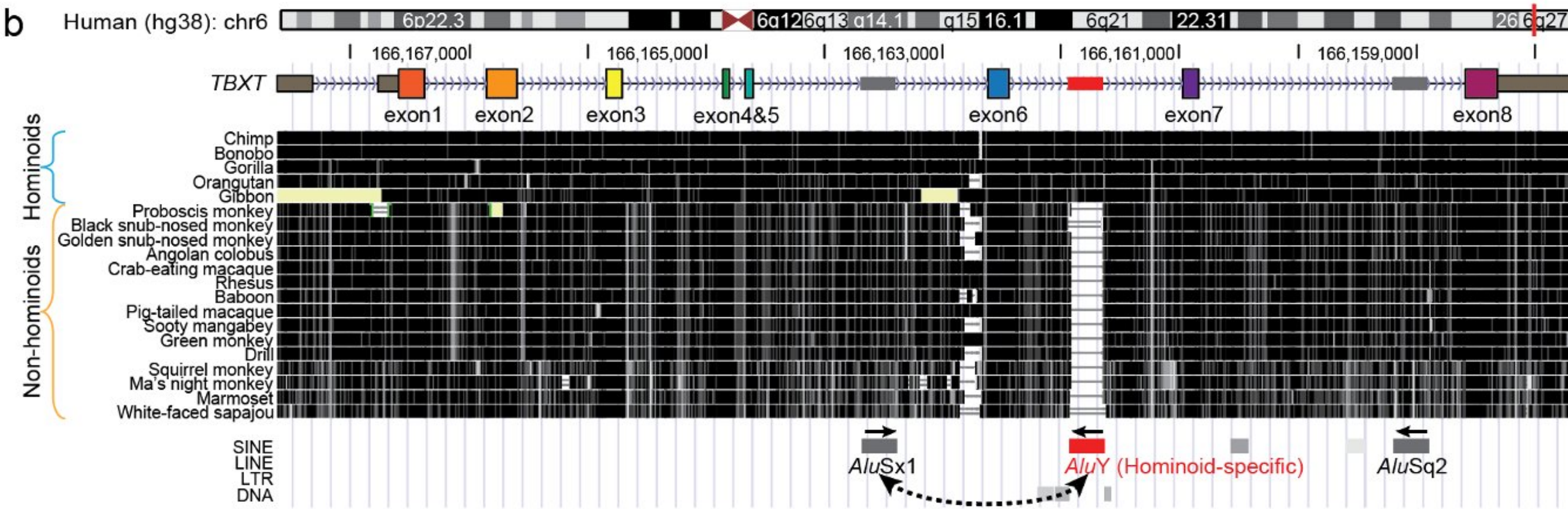
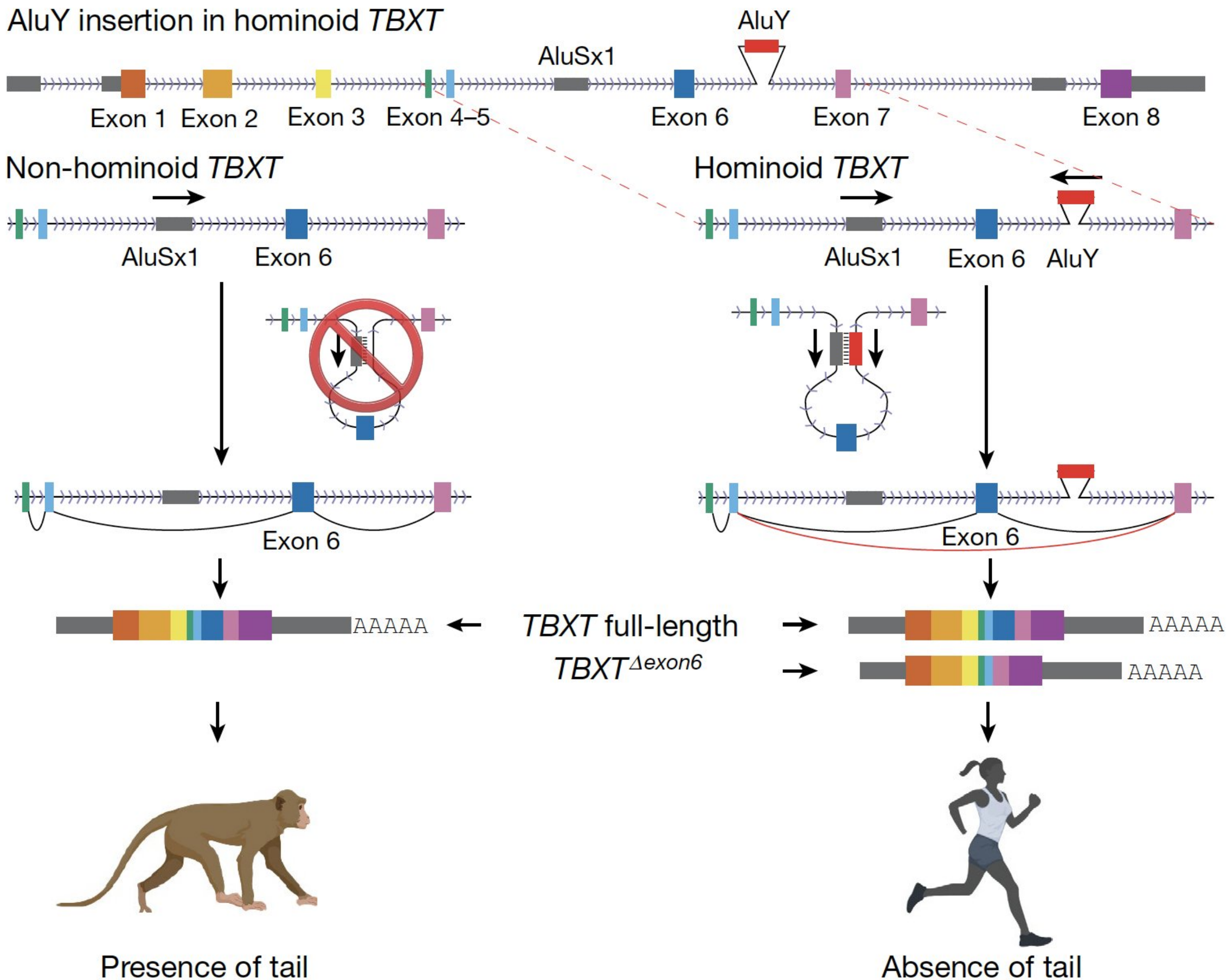


Fig. 3. The short-tailed and tailless animals. A - zebrafish, wild type; B - homozygotic *no tail* (Brachyury, T) mutant (Halpern et al., 1997); C - mice, wild type; D - mice, heterozygotic T mutant; E - guinea pig, wild type; cat, Manx, heterozygotic T mutant; G - hominid, wild type.


b
Human (hg38): chr6
6p22.3
6q12
6q13
q14.1
q15
16.1
6q21
22.31
26
6q27
166,167,000
166,165,000
166,163,000
166,161,000
166,159,000
TBXT
exon1
exon2
exon3
exon4&5
exon6
exon7
exon8
Hominoids
Non-hominoids
Chimp
Bonobo
Gorilla
Orangutan
Gibbon
Proboscis monkey
Black snub-nosed monkey
Golden snub-nosed monkey
Angolan colobus
Crab-eating macaque
Rhesus
Baboon
Pig-tailed macaque
Sooty mangabey
Green monkey
Drill
Squirrel monkey
Ma's night monkey
Marmoset
White-faced sapajou
SINE
LINE
LTR
DNA
AluSx1
AluY (Hominoid-specific)
AluSq2


AluY insertion in hominoid TBXT
AluY
AluSx1
Exon 1
Exon 2
Exon 3
Exon 4–5
Exon 6
Exon 7
Exon 8
Non-hominoid TBXT
AluSx1
Exon 6
Hominoid TBXT
AluSx1
Exon 6
AluY
Exon 6
Exon 6
AAAAA
TBXT full-length
TBXTΔexon6
AAAAA
AAAAA
Presence of tail
Absence of tail

https://ensembl.org

Homo_sapiens - Ensembl ge

https://useast.ensembl.org/Homo_sapiens/Info/Index

New Chrome available

Bookmarks | taehoonkim | UTD Email | Biological Science... | UTDallas | Galaxy | The Unive... | eLearning | Google | Yahoo Finance | Bing | Google Drive | Wikipedia | All Bookmarks

Ensembl east

BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Login/Register

Search Human...

Human (GRCh38.p14) ▾

Search Human (Homo sapiens)

Search all categories ▾ Search... Go

e.g. **BRCA2** or **17:63992802-64038237** or **rs699** or **osteoarthritis**

Genome assembly: GRCh38.p14 (GCA_000001405.29)

More information and statistics

Download DNA sequence (FASTA)

Convert your data to GRCh38 coordinates

Display your data in Ensembl

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart Go

View karyotype

Example region

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild

Download FASTA files for genes, cDNAs, ncRNA, proteins

Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins

Update your old Ensembl IDs

Example gene

Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis

Download alignments (EMF)

Example gene tree

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

More about variation in Ensembl

Download all variants (GVF)

Variant Effect Predictor

Example variant

Example phenotype

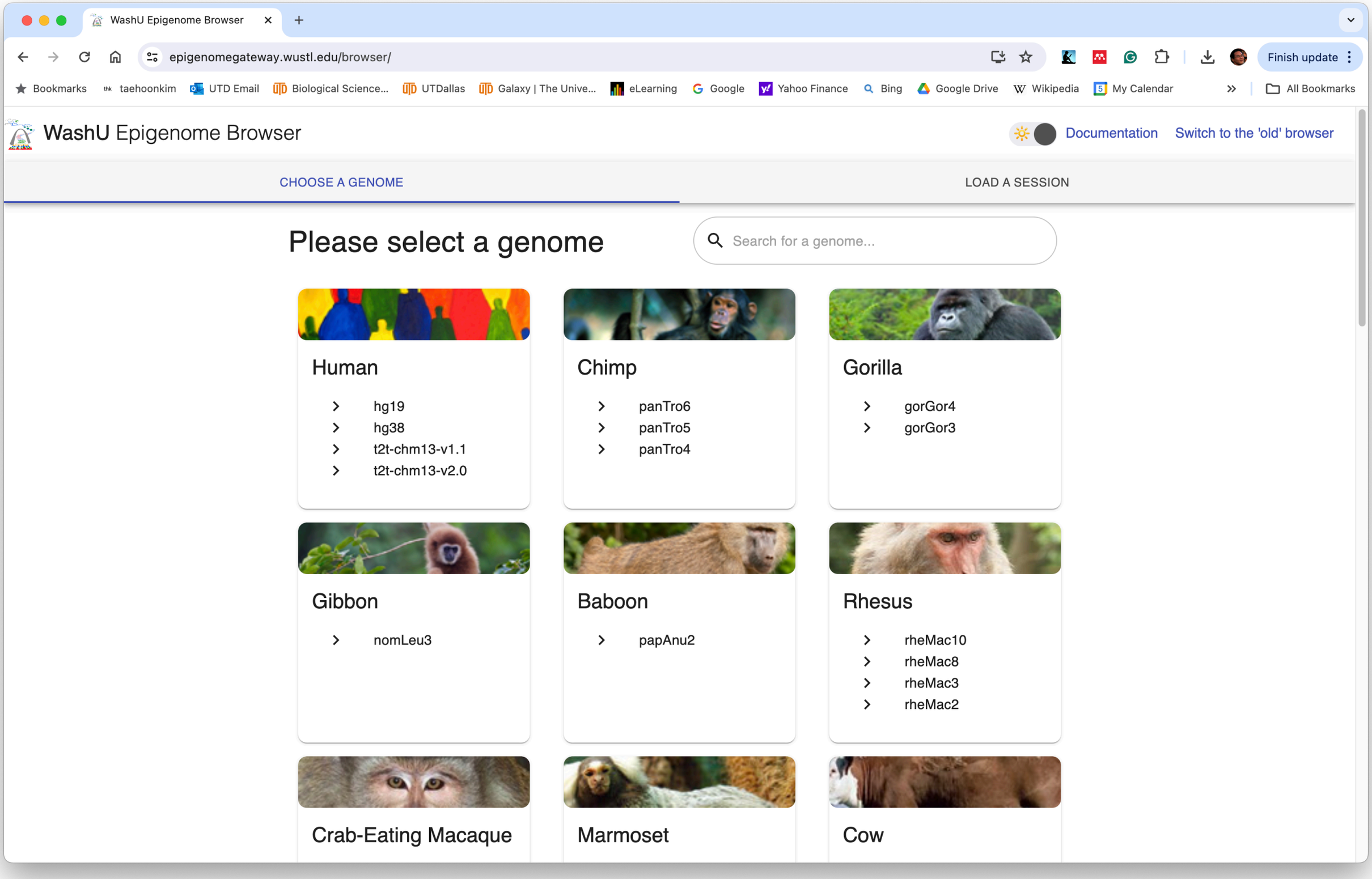
Regulation

What can I find? Regulatory features like enhancers and promoters, and regulatory activity including ATAC-seq and ChIP-seq tracks.

More about the Ensembl regulatory annotation

Example regulatory feature

https://epigenomegateway.wustl.edu



additional features of genome browser

lab exercises for genome browser

galaxy

galaxy project

- *started out as galaxy browser*
  - *beefed up UCSC genome browser with Perl and MySQL programming support*
- *now, almost 20 years later, it is a large centralized computational project that goes beyond genome browsing*

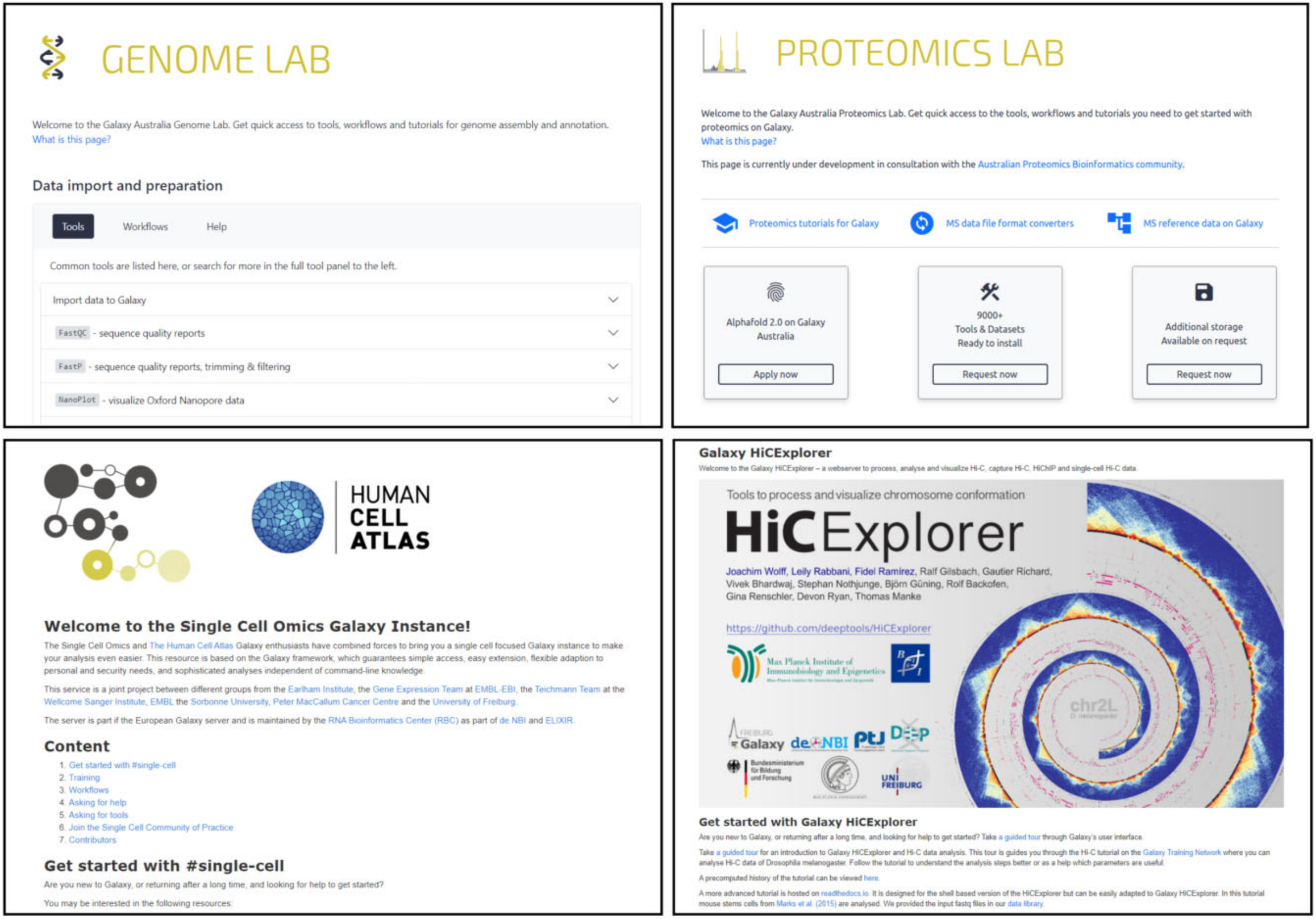


Figure 1. Examples of Galaxy Labs/subdomains. Researchers can quickly access a concentration of domain-specific tools, workflows, support, and training through Galaxy Labs or Galaxy subdomains. Top: the Genome Lab and Proteomics Lab on Galaxy Australia, https://genome.usegalaxy.org.au and https://proteomics.usegalaxy.org.au. Bottom: the Single Cell Omics subdomain on Galaxy Europe, https://singlecell.usegalaxy.eu/ and https://hicexplorer.usegalaxy.eu.

galaxyproject.org

Home - Galaxy Community Hu

galaxyproject.org

Finish update

Bookmarks | taehoonkim | UTD Email | Biological Science... | UTDallas | Galaxy | The Unive... | eLearning | Google | Yahoo Finance | Bing | Google Drive | Wikipedia | My Calendar | All Bookmarks

Galaxy COMMUNITY HUB | Global | Regions | News | Events | Help | Community | About | Applications | @jxtx | Search | Edit

Galaxy Community Conference is coming up.
Read on.

Meet Galaxy - a data analysis universe

Galaxy is a free, open-source system for analyzing data, authoring workflows, training and education, publishing tools, managing infrastructure, and more.

Use Galaxy now: US | Learn more: First Steps with Galaxy

BR NO
— GALAXY COMMUNITY CONFERENCE 2024 —

GCC is where Galaxy meets!

Join the community.

Become a GCC2024 sponsor.

Learn

Tutorials
Access topic-based trainings for free.

Videos
Watch to learn, including conferences.

Home - Galaxy Community H

galaxyproject.org

Finish update

Bookmarks | taehoonkim | UTD Email | Biological Science... | UTDallas | Galaxy | The Unive... | eLearning | Google | Yahoo Finance | Bing | Google Drive | Wikipedia | My Calendar | All Bookmarks

Learn

From curated tools and workflows to self-paced tutorials available on the Galaxy Training Network (GTN), there are plenty of materials to learn from.

Tutorials
Access topic-based trainings for free.

Videos
Watch to learn, including conferences.

Vetted workflows
(Re)use or adapt quality workflows.

Blog
Read highlights from the experts.

Data
Pull data from popular repositories.

Community
Explore the world-wide community.

FAQ
Common questions. Good answers.

Help & support
Advice is always available.

Galaxy is more than you think

Galaxy is a world-wide community.

There are over 500,000 registered Galaxy users from all over the world. Join this lively community to get help, contribute, and learn.

Galaxy has 1,000s of tools

In partnership with BioConda and BioContainers, Galaxy provides instant access to vast number of analysis tools.

https://training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-short/tutorial.html

COSMIC database

COSMIC
Catalogue Of Somatic Mutations In Cancer

What is COSMIC?

COSMIC – the Catalogue of Somatic Mutations in Cancer – is the world's largest source of expert manually curated somatic mutation information relating to human cancers. Here we outline that data in terms of structure, content and scope making it easier for you to evaluate what you will find in COSMIC, and how best to access it to fulfill your research needs.

Overview

COSMIC comprises the COSMIC database and the Cell Lines Project, two separate but related resources. This page discusses COSMIC; please see the About Cell Lines page for more information on the Cell Lines Project.

The COSMIC database combines two main types of data:

High Precision Data, Manually Curated by Experts:

- Targeted gene-screening panels
- Over 27,000 peer reviewed papers
- Metadata (environmental factors and patient history)
- Focused on known and suspected cancer genes and mutations
- Objective frequency data as a result of mutation negative samples
- Full details of the curation process and data captured

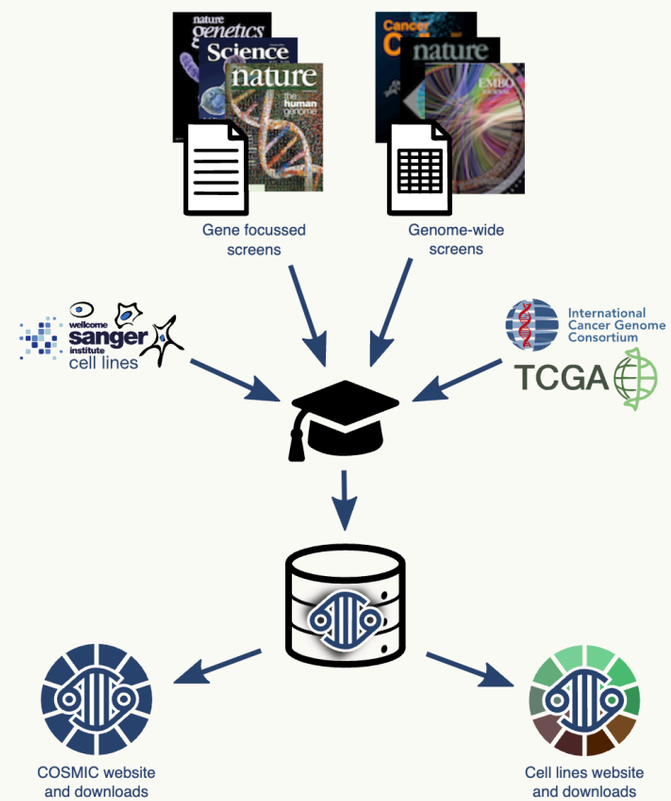
Genome-wide Screen Data:

- Over 37,000 genomes, consisting of:
  - peer reviewed large scale genome screening data
  - other databases such as TCGA and ICGC
- Provides unbiased, genome-level profiling of diseases
- Objective frequency data, by interpreting non-mutant genes across each genome
- Can be used to discover novel driver genes

Together, this compilation of data provides extensive coverage of the cancer genomic landscape from a somatic perspective. New and potentially significant data are continually captured and made available through four significant updates to COSMIC each year.

For more information on COSMIC, read more about our curation processes and the analyses that we run on mutation data, or see our answers to frequently asked questions about curation, histology, and mutation syntax.

Find out more about licensing COSMIC data



https://cancer.sanger.ac.uk/cosmic

cancer.sanger.ac.uk/cosmic

COSMIC
Catalogue Of Somatic Mutations In Cancer

Projects | Data | Tools | News | Help | About | Genome Version | Search COSMIC... SEARCH | Login

COSMIC v100, released 21-MAY-24

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below.

eg Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell SEARCH

Projects

COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:

COSMIC
The core of COSMIC, an expert-curated database of somatic mutations

Cell Lines Project
Mutation profiles of over 1,000 cell lines used in cancer research

COSMIC-3D
An interactive view of cancer mutations in the context of 3D structures

Cancer Gene Census
A catalogue of genes with mutations that are causally implicated in cancer

Cancer Mutation Census
Classification of genetic variants driving cancer

Actionability
Mutations actionable in precision oncology

Data curation

- Gene Curation — details of our manual curation process
- Gene Fusion Curation — details of our curation process for gene fusions
- Genome Annotation — information on the annotation of genomes

COSMIC News

Follow @cosmic_sanger

Largest genomic cancer resource accelerating research and drug development
Press release COSMIC has released the 100th version of its knowledgebase, containing further information on 300,000 somatic mutations linked to human cancers. More...

'You can't be what you can't see': Inspiring inclusion with Nidhi Bindal Dhir, COSMIC Head of IT
International Women's Day 2024 is focused on inspiring inclusion. We are privileged at COSMIC to work with a range of incredible women such as our Head of IT Nidhi Bindal Dhir who we caught up with to reflect on her 15 years as part of the team! More...

Curation in context: A glimpse into COSMIC v99
To celebrate the release of COSMIC v99, we delve into how focusing on expert manual curation of specific genes, resistance mutations & more emulates our dedication to being a reliable, sustainable source of genomic data on somatic mutations in cancer More...

Tools

- Cancer Browser — browse COSMIC data by tissue type and histology
- Genome Browser — browse the human genome with COSMIC annotations
- GA4GH Beacon — access COSMIC data through the GA4GH Beacon Project

Help

- Downloads — data that you can download from our SFTP site
- Documentation — view our help documentation
- FAQ — a compilation of our Frequently Asked Questions
- Release Notes — information about the latest COSMIC release

https://cancer.sanger.ac.uk/cosmic/help/tutorials

EGFR

TERT

R and bioconductor

Bioconductor - Install

bioconductor.org/install/

Bookmarks | taehoonkim | UTD Email | Biological Science... | UTDallas | Galaxy | The Unive... | eLearning | Google | Yahoo Finance | Bing | Google Drive | Wikipedia | My Calendar | Julia Documentati... | All Bookmarks

⚠ Registration Open for Bioc2024 July 24-26

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

About | Learn | Packages | Developers | Search | Get Started

Home > Install

Get started

1. Install R

The current release of Bioconductor is version 3.19; it works with R version 4.4.0. Users of older R and Bioconductor must update their installation to take advantage of new features and to access packages that have been added to Bioconductor since the last release.

The development version of Bioconductor is version 3.20; it works with R version 4.4.0. More recent 'devel' versions of R (if available) will be supported during the next Bioconductor release cycle.

Step 1
Install R

1. Download the most recent version of R. The R FAQs and the R Installation and Administration Manual contain detailed instructions for installing R on various platforms (Linux, OS X, and Windows being the main ones).
2. Start the R program; on Windows and OS X, this will usually mean double-clicking on the R application, on UNIX-like systems, type "R" at a shell prompt.
3. As a first step with R, start the R help browser by typing **help.start()** in the R command window. For help on any function, e.g. the "mean" function, type **?mean**.

2. Get the latest version of Bioconductor

Once R has been installed, get the latest version of Bioconductor by starting R and entering the following commands.

It may be possible to change the Bioconductor version of an existing installation; see the 'Changing version' section of the BiocManager vignette.

Details, including instructions to install additional packages and to update, find, and troubleshoot are provided below. A devel version of Bioconductor is available. There are good reasons for using **BiocManager::install()** for managing Bioconductor resources.

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(version = "3.19")
```

Step 2
Get Bioconductor

Step 3
Now get your packages!

https://www.bioconductor.org/packages/release/workflows/vignettes/sequencing/inst/doc/sequencing.html