

# *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>

UCSC Genome Browser Home x + https://genome.ucsc.edu New Chrome available :

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

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute UCSC Genome Browser

Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us

Search genes, data, help docs and more... Search

Tools

- [Genome Browser](#) - Interactively visualize genomic data
- [BLAT](#) - Rapidly align sequences to the genome
- [In-Silico PCR](#) - Rapidly align PCR primer pairs to the genome
- [Table Browser](#) - Download and filter data from the Genome Browser
- [LiftOver](#) - Convert genome coordinates between assemblies
- [REST API](#) - Returns data requested in JSON format
- [Variant Annotation Integrator](#) - Annotate genomic variants
- [More tools...](#)

hg38 hg19 mm39

News

- Jun. 7, 2024 - [New GENCODE Versions tracks for hg19/hg38/mm39 \(V46/VM35\)](#)
- May. 22, 2024 - [New GENCODE gene tracks: V46 \(hg38\) - VM35 \(mm39\)](#)
- Apr. 25, 2024 - [New AbSplice Prediction Scores track for hg19](#)
- Mar. 26, 2024 - [New gnomAD v4 Constraint Metrics \(hg38\) and gnomAD Non-canc...](#)
- Mar. 07, 2024 - [New Prediction Scores super track and BayesDel track for hg19](#)
- Mar. 05, 2024 - [New JASPAR tracks: Human \(hg19/hg38\) - Mouse \(mm10/mm39\)](#)

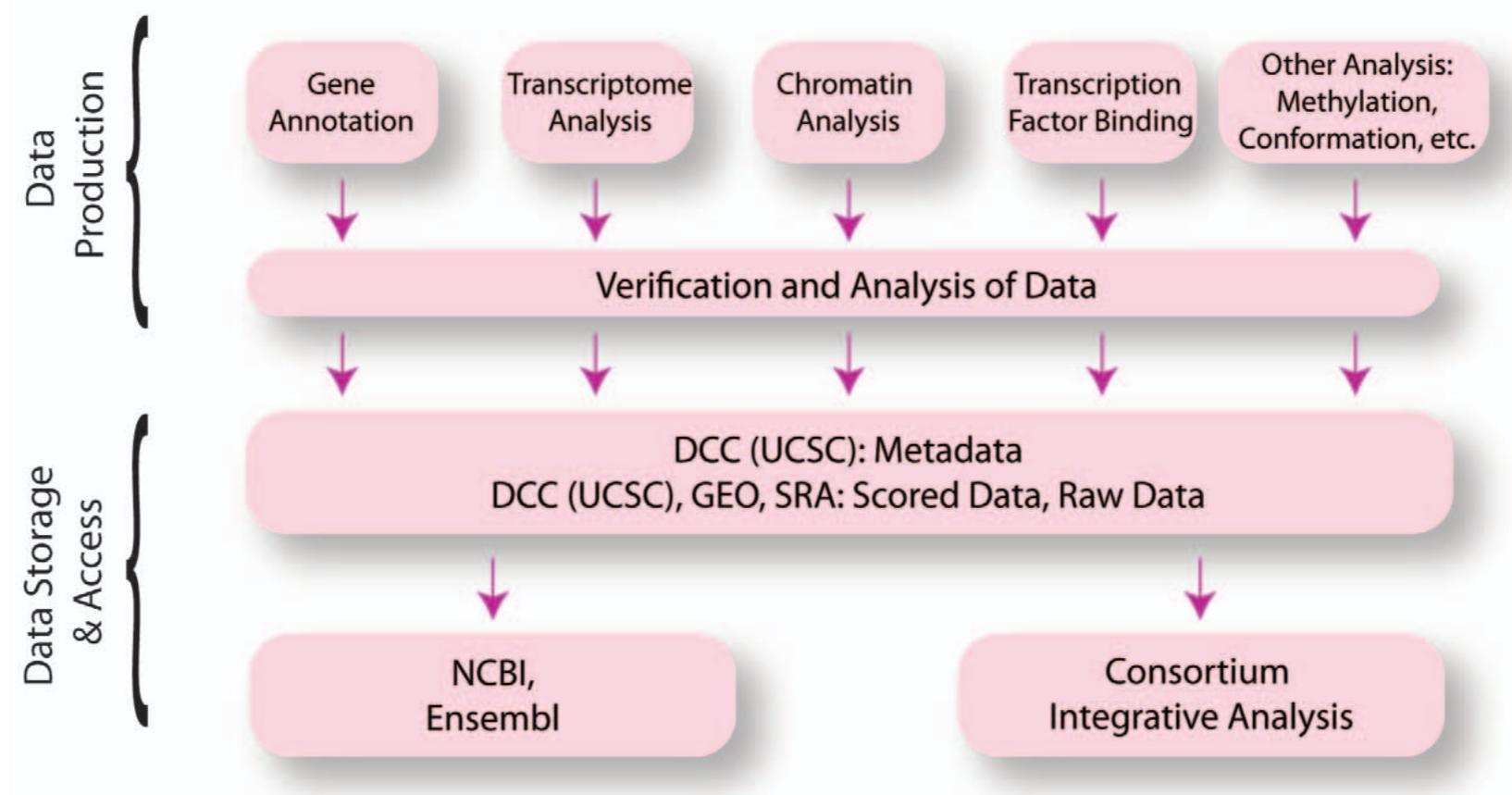
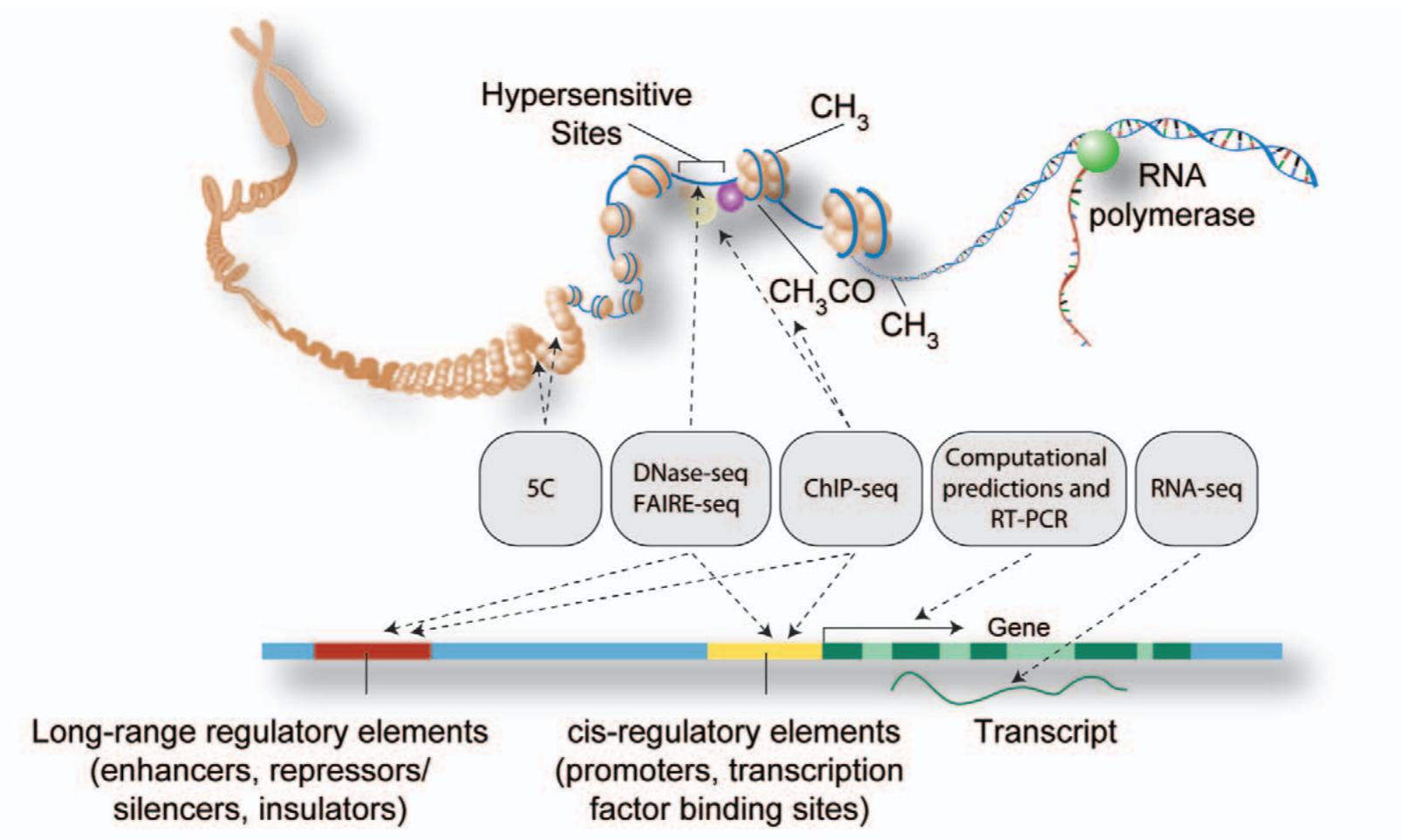
[More news...](#) [Subscribe](#)

Meetings and Workshops: Come see us in person!

- NYU Langone Center for Human Genetics and Genomics - New York, NY. June 24, 2024.
- [McKusick Summer Course in Human and Mammalian Genetics and Genomics](#) - Bar Harbor, ME. July 24, 2024.
- [Brazilian Society for Medical Genetics](#) - Cuiaba, Brazil. August 28, 2024.
- Faculty of Medicine, University of Chile, West Campus - Santiago, Chile. Sep 3-4, 2024

Feel free to [contact us](#) if you are interested in attending a workshop, or meeting someone from the team to collaborate, get help, or ask any questions at the meetings.

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



# genome browsing

---

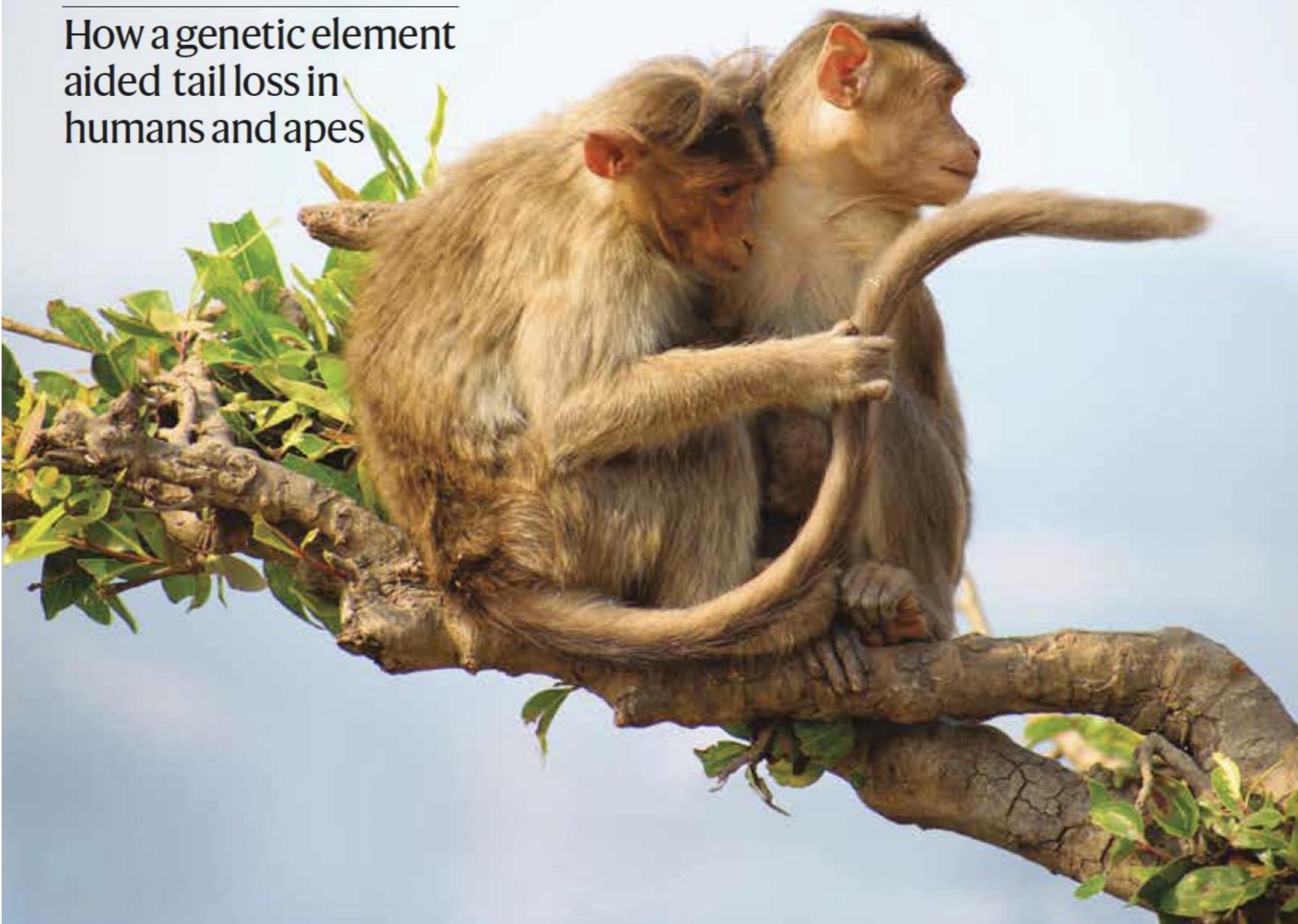
- *specific genes*
- *specific genomic regions*
- *first pass information on annotations (genes, repeats)*
- *expression across tissues*
- *conservation*
- *variation*
- *regulation*

The international journal of science / 29 February 2024

# nature

## TALE OF TAILS

How a genetic element  
aided tail loss in  
humans and apes



# 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



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

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’<sup>1–3</sup>, with a proposed role in contributing to human bipedalism<sup>4–6</sup>. 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<sup>7–9</sup>—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<sup>10</sup>. 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.

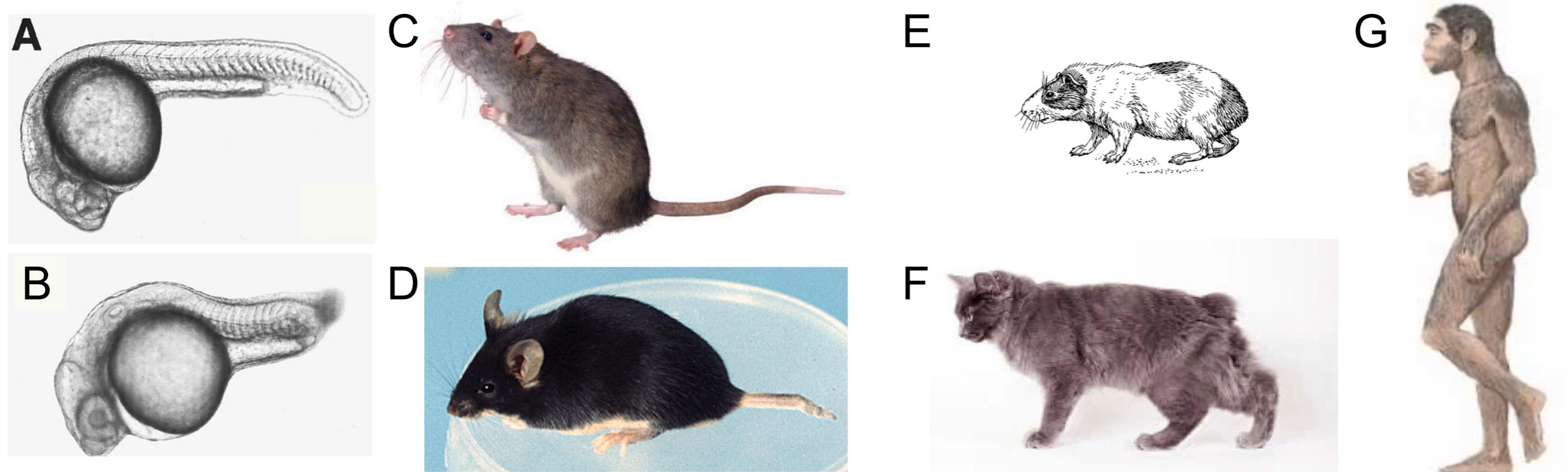
a personal story behind the science



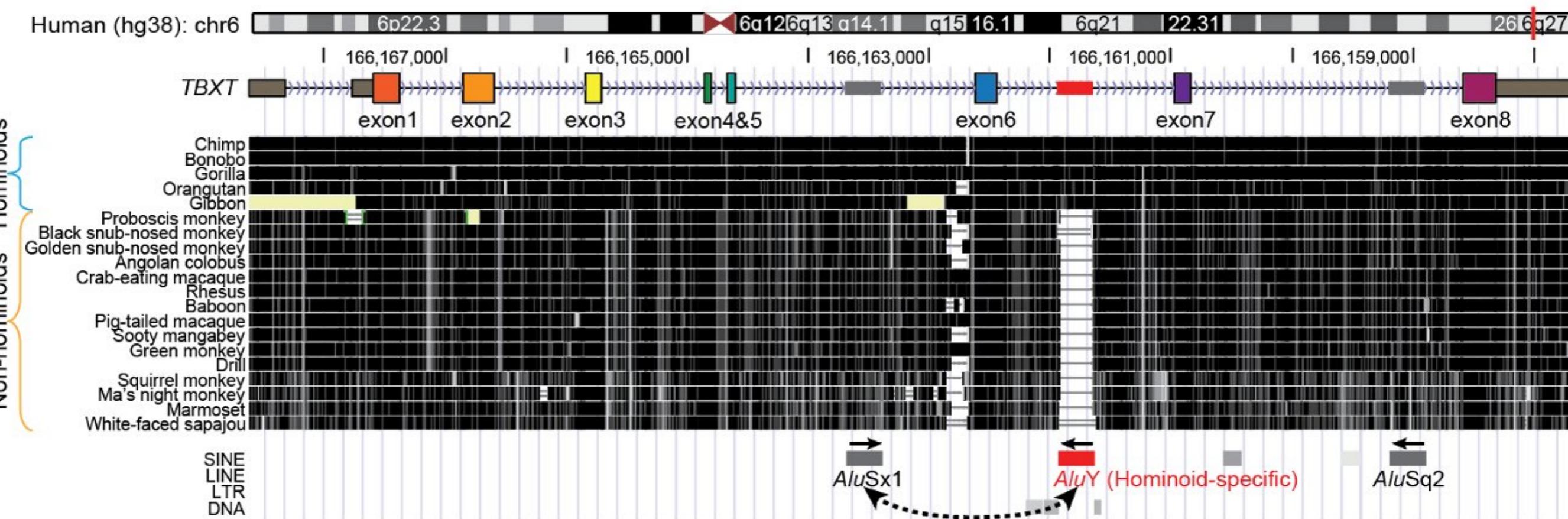
<https://podcasts.apple.com/us/podcast/bo-xia-and-a-tale-of-tails/id1563415749?i=1000647438287>

# Start with the Brachyury gene

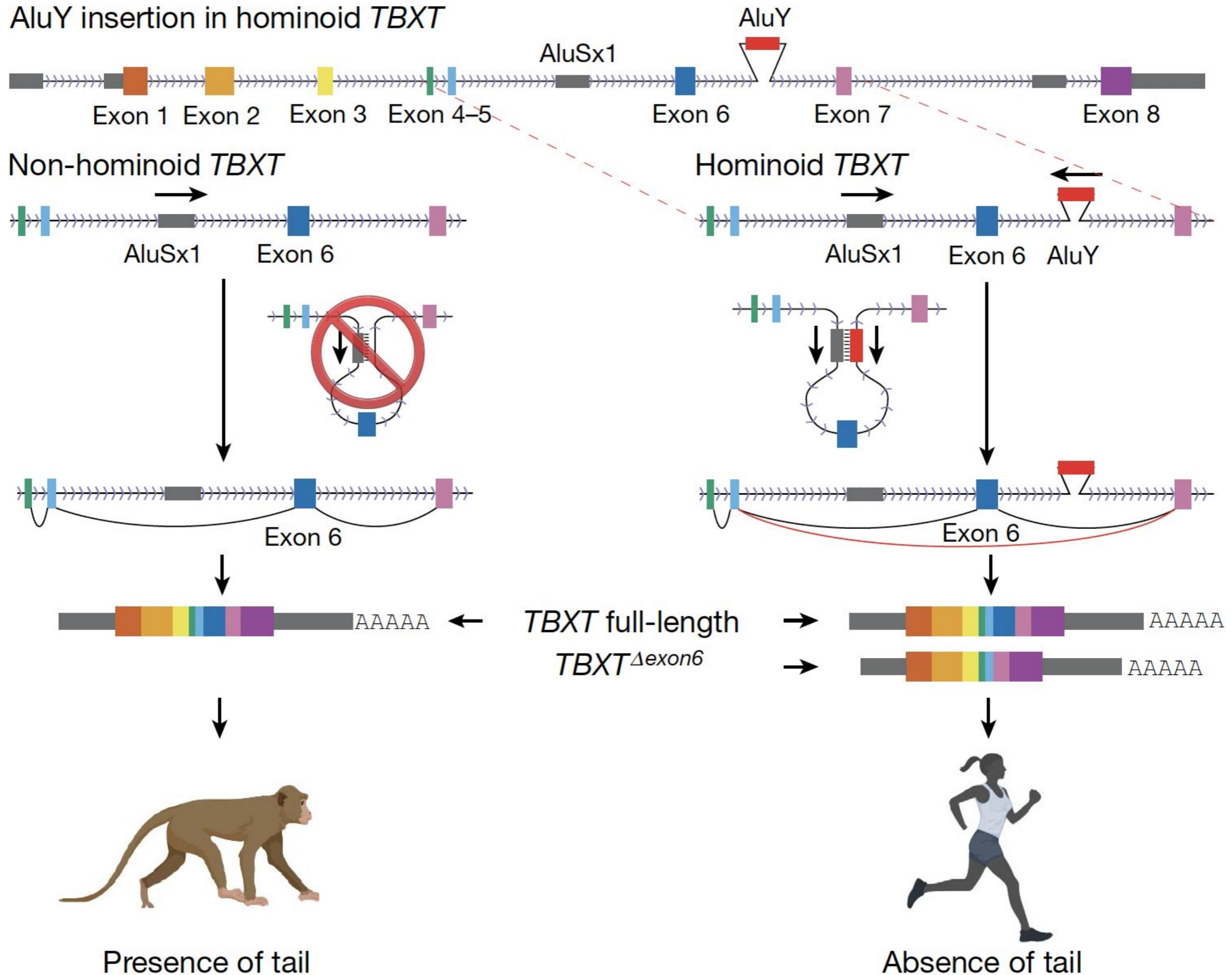
---



**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**

## AluY insertion in hominoid *TBXT*



# <https://ensembl.org>

Homo\_sapiens - Ensembl gen x + https://useast.ensembl.org/Homo\_sapiens/Info/Index New Chrome available

Bookmarks taehoonkim UTD Email Biological Science... UT Dallas Galaxy | The Universe... eLearning Google Yahoo Finance Bing Google Drive Wikipedia All Bookmarks

Login/Register

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

**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** Pax6 INS FOXP2 BRCA2 DMD ssh

**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)

**Regulation**

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

**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** ATCGAGCT ATCCAGCT ATCGAGAT

**Example phenotype**

**VeIP**

Example regulatory feature

This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our [Privacy Policy](#) and [Terms of Use](#)

I Agree

<https://epigenomegateway.wustl.edu>

WashU Epigenome Browser x +

epigenomegateway.wustl.edu/browser/ Finish update :

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

WashU Epigenome Browser Documentation Switch to the 'old' browser

CHOOSE A GENOME LOAD A SESSION

Please select a genome

 Human  
hg19  
hg38  
t2t-chm13-v1.1  
t2t-chm13-v2.0

 Chimp  
panTro6  
panTro5  
panTro4

 Gorilla  
gorGor4  
gorGor3

 Gibbon  
nomLeu3

 Baboon  
papAnu2

 Rhesus  
rheMac10  
rheMac8  
rheMac3  
rheMac2

 Crab-Eating Macaque

 Marmoset

 Cow

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

## Galaxy platform 2024

*Accessible, reproducible, and collaborative data analyses*



>11 k users  
>1 m jobs



→ ↵ >400 supported  
→ data types

Resourced with:

- 🔧 >9 k tools
- ⌚ Unlimited workflows
- 🎓 >400 tutorials

*Open source | Vibrant global community | Multidisciplinary*

# GENOME LAB

Welcome to the Galaxy Australia Genome Lab. Get quick access to tools, workflows and tutorials for genome assembly and annotation.  
[What is this page?](#)

## Data import and preparation

Tools   Workflows   Help

Common tools are listed here, or search for more in the full tool panel to the left.

- Import data to Galaxy
- FastQC - sequence quality reports
- FastP - sequence quality reports, trimming & filtering
- NanoPlot - visualize Oxford Nanopore data

# PROTEOMICS LAB

Welcome to the Galaxy Australia Proteomics Lab. Get quick access to the tools, workflows and tutorials you need to get started with proteomics on Galaxy.  
[What is this page?](#)

This page is currently under development in consultation with the [Australian Proteomics Bioinformatics community](#).

Proteomics tutorials for Galaxy

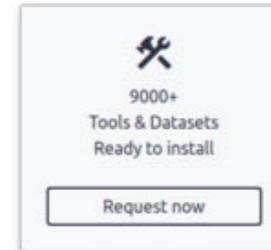
MS data file format converters

MS reference data on Galaxy



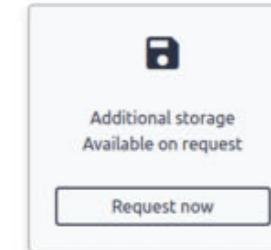
AlphaFold 2.0 on Galaxy  
Australia

Apply now



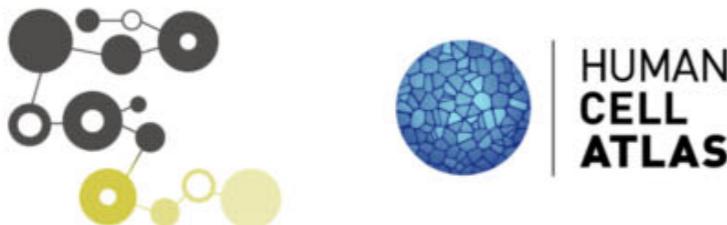
9000+  
Tools & Datasets  
Ready to install

Request now



Additional storage  
Available on request

Request now



HUMAN  
CELL  
ATLAS

## Welcome to the Single Cell Omics Galaxy Instance!

The Single Cell Omics and [The Human Cell Atlas](#) Galaxy enthusiasts have combined forces to bring you a single cell focused Galaxy instance to make your analysis even easier. This resource is based on the Galaxy framework, which guarantees simple access, easy extension, flexible adaption to personal and security needs, and sophisticated analyses independent of command-line knowledge.

This service is a joint project between different groups from the [Earlham Institute](#), the [Gene Expression Team at EMBL-EBI](#), the [Teichmann Team](#) at the [Wellcome Sanger Institute](#), [EMBL](#) the [Sorbonne University](#), [Peter MacCallum Cancer Centre](#) and the [University of Freiburg](#).

The server is part of the European Galaxy server and is maintained by the [RNA Bioinformatics Center \(RBC\)](#) as part of [de.NBI](#) and [ELIXIR](#).

## Content

1. Get started with #single-cell
2. Training
3. Workflows
4. Asking for help
5. Asking for tools
6. Join the Single Cell Community of Practice
7. Contributors

## Get started with #single-cell

Are you new to Galaxy, or returning after a long time, and looking for help to get started?

You may be interested in the following resources:

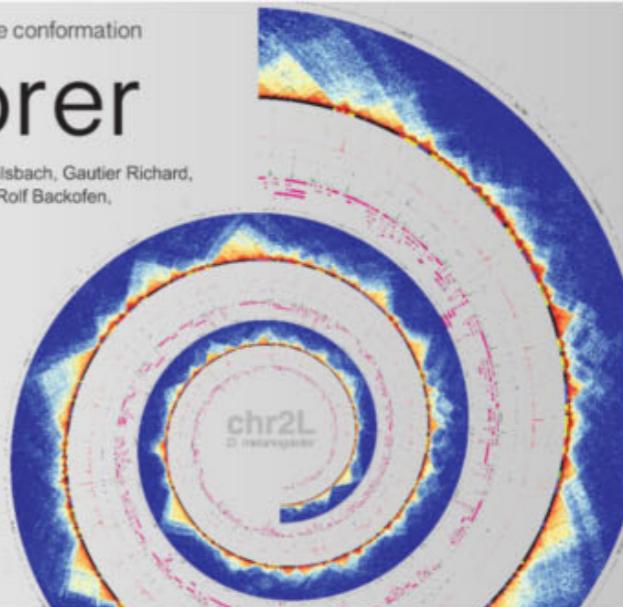
## Galaxy HiCExplorer

Welcome to the Galaxy HiCExplorer – a webserver to process, analyse and visualize Hi-C, HiChIP and single-cell Hi-C data.

Tools to process and visualize chromosome conformation  
**HiCExplorer**

Joachim Wolff, Leily Rabbani, Fidel Ramirez, Ralf Gilsbach, Gautier Richard, Vivek Bhardwaj, Stephan Nothjunge, Björn Güning, Rolf Backofen, Gina Renschler, Devon Ryan, Thomas Manke

<https://github.com/deeptools/HiCExplorer>



## Get started with Galaxy HiCExplorer

Are you new to Galaxy, or returning after a long time, and looking for help to get started? Take a [guided tour](#) through Galaxy's user interface.

Take a [guided tour](#) for an introduction to Galaxy HiCExplorer and Hi-C data analysis. This tour is guides you through the Hi-C tutorial on the [Galaxy Training Network](#) where you can analyse Hi-C data of *Drosophila melanogaster*. Follow the tutorial to understand the analysis steps better or as a help which parameters are useful.

A precomputed history of the tutorial can be viewed [here](#).

A more advanced tutorial is hosted on [readthedocs.io](#). It is designed for the shell based version of the HiCExplorer but can be easily adapted to Galaxy HiCExplorer. In this tutorial mouse stems cells from [Marks et al. \(2015\)](#) are analysed. We provided the input fastq files in our [data library](#).

**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>.

Home - Galaxy Community Hub +

galaxyproject.org

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

Galaxy COMMUNITY HUB Global Regions ▾ News Events Help ▾ Community ▾ About ▾ Applications ▾ @jtx 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

 BRNO  
— 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 Hub

galaxyproject.org

Bookmarks taehoonkim UTD Email Biological Science... UT Dallas Galaxy | The Universe eLearning Google Yahoo Finance Bing Google Drive Wikipedia My Calendar All Bookmarks Finish update

# 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.

**Vetted workflows**  
(Re)use or adapt quality workflows.

**Data**  
Pull data from popular repositories.

**FAQ**  
Common questions. Good answers.

**Videos**  
Watch to learn, including conferences.

**Blog**  
Read highlights from the experts.

**Community**  
Explore the world-wide community.

**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

Mail - Kim, Tae Hoon - Outl X Mail - Kim, Tae Hoon - Outl X SSPUSADV.pdf (SECURED) X COSMIC Actionability | Hon X About COSMIC X COSMIC.67.pdf X Human hg38 chr5:1294176 X +

cancer.sanger.ac.uk/cosmic/about

Bookmarks taehoontkim UTD Email Biological Science... UT Dallas Galaxy | The Universe eLearning Google Yahoo Finance Bing Google Drive Wikipedia My Calendar Julia Documentation All Bookmarks

**COSMIC**  
Catalogue Of Somatic Mutations In Cancer

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

## 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

```

graph TD
    NatureGenetics[nature genetics]
    Science[Science]
    Nature[nature]
    CellLines[wellcome sanger institute cell lines]
    TCGA[TCGA]
    ICgc[International Cancer Genome Consortium]
    GraduationCap[Graduation Cap]
    Database[Database]
    COSMICWebsite[COSMIC website and downloads]
    CellLinesWebsite[Cell lines website and downloads]

    Nature --> GeneFocusedScreens
    Science --> GeneFocusedScreens
    Nature --> GenomeWideScreens
    CellLines --> GeneFocusedScreens
    TCGA --> GenomeWideScreens
    ICgc --> GenomeWideScreens
    GeneFocusedScreens --> GraduationCap
    GenomeWideScreens --> GraduationCap
    GraduationCap --> Database
    Database --> COSMICWebsite
    Database --> CellLinesWebsite
  
```

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

Screenshot of the COSMIC website (<https://cancer.sanger.ac.uk/cosmic>) showing the homepage and a news section.

The top navigation bar includes links to various resources like Mail, SSPUSADV.pdf (SECURED), COSMIC Actionability, COSMIC Catalogue of Somatic Mutations In Cancer, COSMIC.67.pdf, Human hg38 chr5:1294176, and All Bookmarks.

The main header features the COSMIC logo and a search bar.

**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.

**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
- Drug Resistance** — curation of mutations conferring drug resistance

**COSMIC News**

**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 taehoontkim UTD Email Biological Science... UT Dallas 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**.

```
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>