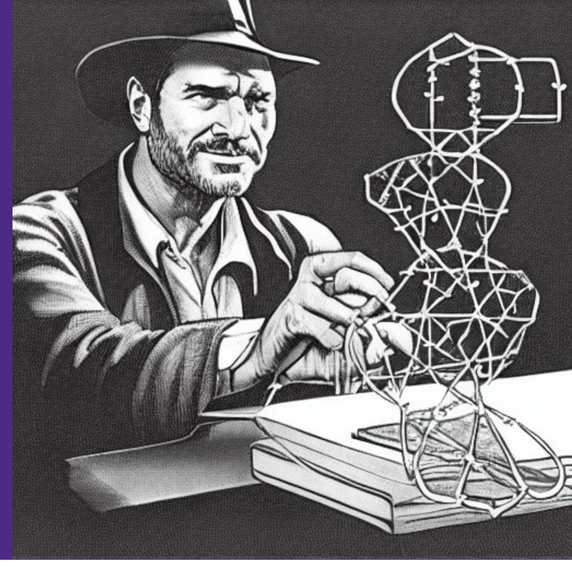
GeneHTracker: Improving reproducibility and reusability of datasets based on gene identifiers

Hugo A. Guillen-Ramirez[1,2,3,4], Daniel Sanchez-Taltavull[5], Rory Johnson[1,2,3,4]

1 Department of Medical Oncology, Inselspital, Bern University Hospital, University of Bern, 3010 Bern, Switzerland.
2 Department for BioMedical Research, University of Bern, 3008 Bern, Switzerland.
3 School of Biology and Environmental Science, University College Dublin, Dublin D04 V1W8, Ireland.
4 Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Dublin D04 V1W8, Ireland.
5 Department of Visceral Surgery and Medicine, Inselspital, Bern University Hospital, University of Bern, 3010 Bern, Switzerland



ENSEMBL/GENCODE gene IDs can be retired or deprecated, have their gene type changed, or even a new ID can be assigned to a given locus. These problems limit the reproducibility and reusability of published gene lists. Here, we present GeneHTracker, a **Python package** and **index** that allows to:

1) retrieve the latest coordinates for a list of gene IDs;
2) find the complete (GENCODE) annotation history of a gene ID or gene symbol;
3) generate putative mappings for deprecated/retired IDs based on overlapping annotations curated since Gencode version 5/Ensembl version 60, representing almost 12 years of annotations.

Currently GeneHTracker is generated for human genes. **Do you need to index for other organism? Do you want to create an index for other annotation set?** Let's have a chat!

Where is ENSG00000259484?

You find an interesting gene...

https://lncipedia.org › transcript
Transcript: lnc-ZNF280D-3:4 - LNCipedia
Alternative gene names: ENSG00000248500; **ENSG00000259484**.1; RP11-323F24.1; OTTHUMG00000172613.1; ENSG00000285331.1; AC010999.3. RNA sequence: Structure: ...

https://www.ncbi.nlm.nih.gov › articles › PMC3946172
Long Non-Coding RNA Expression Profiles in Hereditary ...
by PM Tørring · 2014 · Cited by 21 — As long non-coding RNAs (lncRNAs) are increasingly recognized as key regulators of gene expression and constitute a sizable fraction of the ...

http://www.enhanceratlas.org › browseenhancer
EnhancerAtlas 2.0: an updated resource with typical enhancer ...
RP11, **ENSG00000259484**. ZNF280D, ENSG00000137871. snoU13, ENSG00000239035 · dbSUPER. Number of super-enhancer constituents: 3. ID, Coordinate, Tissue/cell ...

...is not longer active in Ensembl...

Only searching Human | ENSG00000259484
1 results match **ENSG00000259484** when restricted to species: Human

ENSG00000259484 (Human Gene)
ENSG00000259484
Ensembl gene ENSG00000259484 is no longer in the database but it has been mapped to 1 deprecated identifier.

https://www.ensembl.org/

...there are no associated genes...

Gene: ENSG00000259484

This identifier is not in the current EnsEMBL database

ID History

Stable ID	ENSG00000259484.1
Status	Retired (see below for possible successors)
Latest Version	ENSG00000259484.1 Release: 72 Assembly: GRCh37 Database: homo_sapiens_core_72_37

Associated archived IDs for this stable ID version

No associated IDs found

...or even mapped IDs!

ID History Converter results

Job details

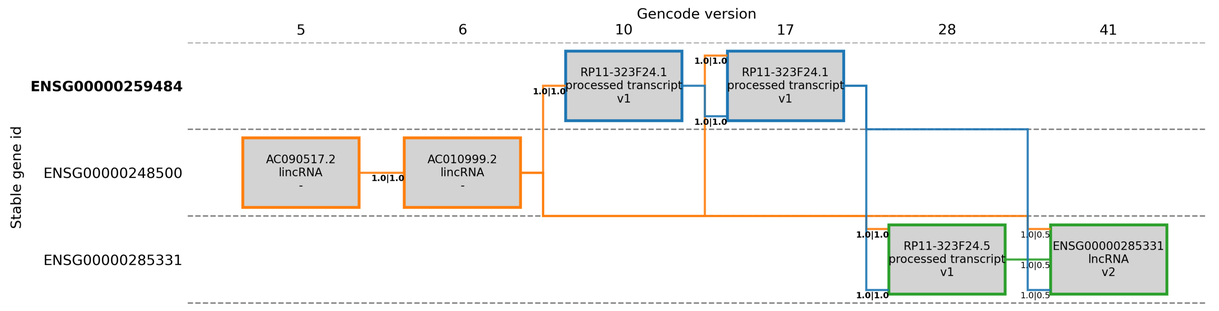
Download results file | New job

No results

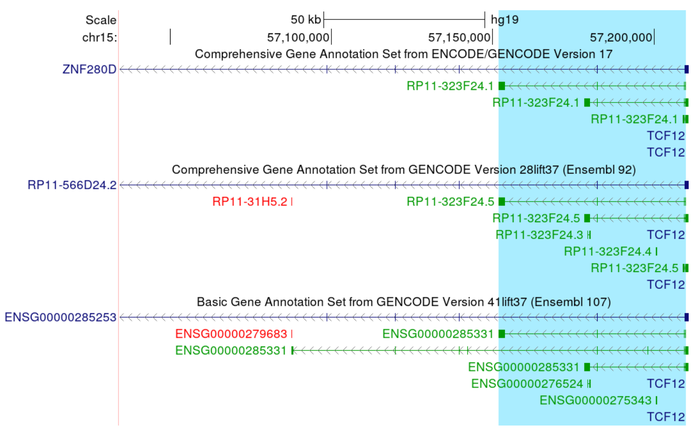
No stable IDs mapped to the given IDs

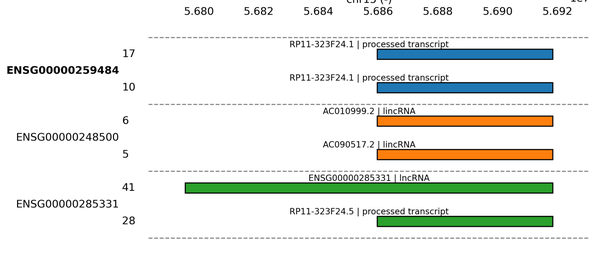
Finding ENSG00000259484

Gene history diagram



Genome browser view



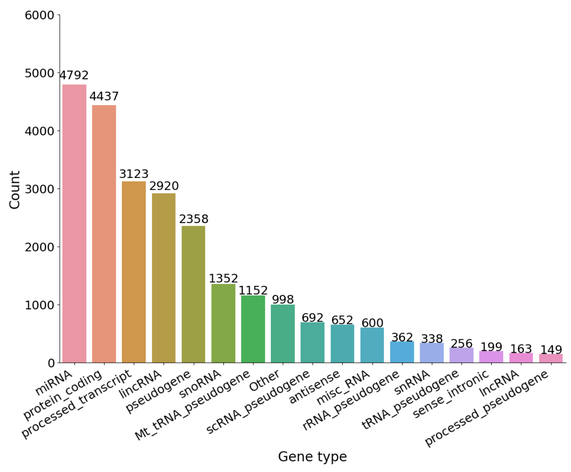


Default GeneHTracker browser

Custom session at UCSC

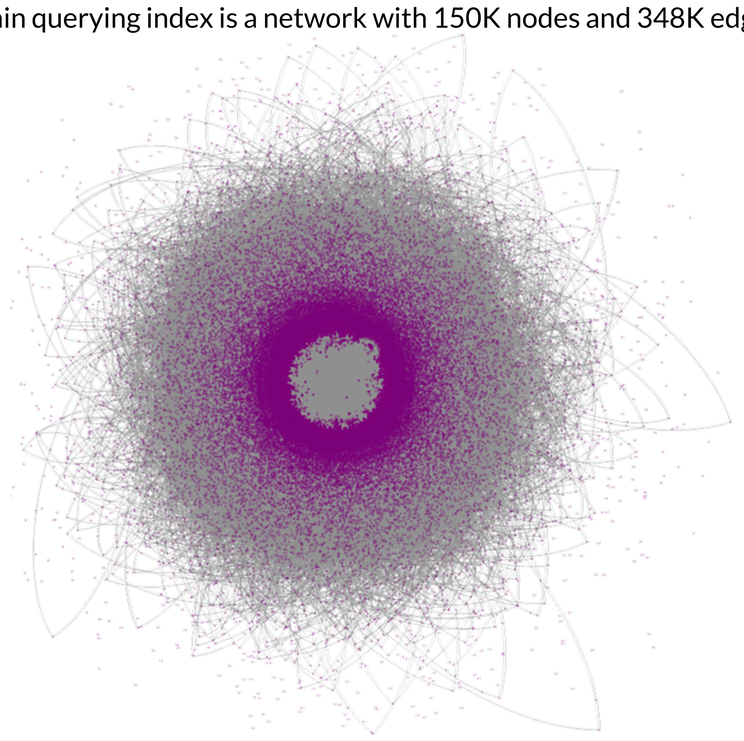
Our solution: GeneHTracker

GeneHTracker indexes:
37 GENCODE versions
259,726 total gene identifiers
76,052 stable gene identifiers
14,200 deprecated gene identifiers



Deprecated identifiers per gene type

The main querying index is a network with 150K nodes and 348K edges.



Automatise your gene finding jobs

Using a gene list as input, GeneHTracker will provide the following information:

- gene_shortid
- latest_gencode
- latest_coord
- latest_gene_id
- latest_gene_name
- latest_gene_type
- latest_hgnc_id
- latest_deprecated
- latest_ensembl
- latest_assembly
- latest_date
- mapped_gene_shortid
- mapped_overlap
- mapped_gencode
- mapped_coord
- mapped_gene_id
- mapped_gene_name
- mapped_gene_type
- mapped_hgnc_id
- mapped_deprecated
- mapped_ensembl
- mapped_assembly
- mapped_date

Conclusions

GeneHTracker enhances reproducibility of gene based datasets by proposing putative mappings and providing tools to manually inspect them.

References
- https://github.com/HugoGuillen/genehtracker
- Frankish, Adam, et al. "GENCODE 2021." Nucleic acids research 49.D1 (2021): D916-D923.
- Banner image generated with stable-diffusion v1.0: Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

@HugoGuillen
GitHub
GLDLab
INSELSPITAL UNIVERSITÄTSSPITAL BERN HÔPITAL UNIVERSITAIRE DE BERNE
UCD DUBLIN
CRG Centre for Genomic Regulation
NCCR RNA & Disease
Swiss National Science Foundation
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No **884178, RNADOMAIN**.