# MakeHub: Creating Individual Assembly Hubs for Display with the UCSC Genome Browser

**Katharina J. Hoff**[1,2]

1) Institute for Mathematics and Computer Science, University of Greifswald, Greifswald, GERMANY
2) Center for Functional Genomics of Microbes, University of Greifswald, Greifswald, GERMANY

Contact: katharina.hoff@uni-greifswald.de

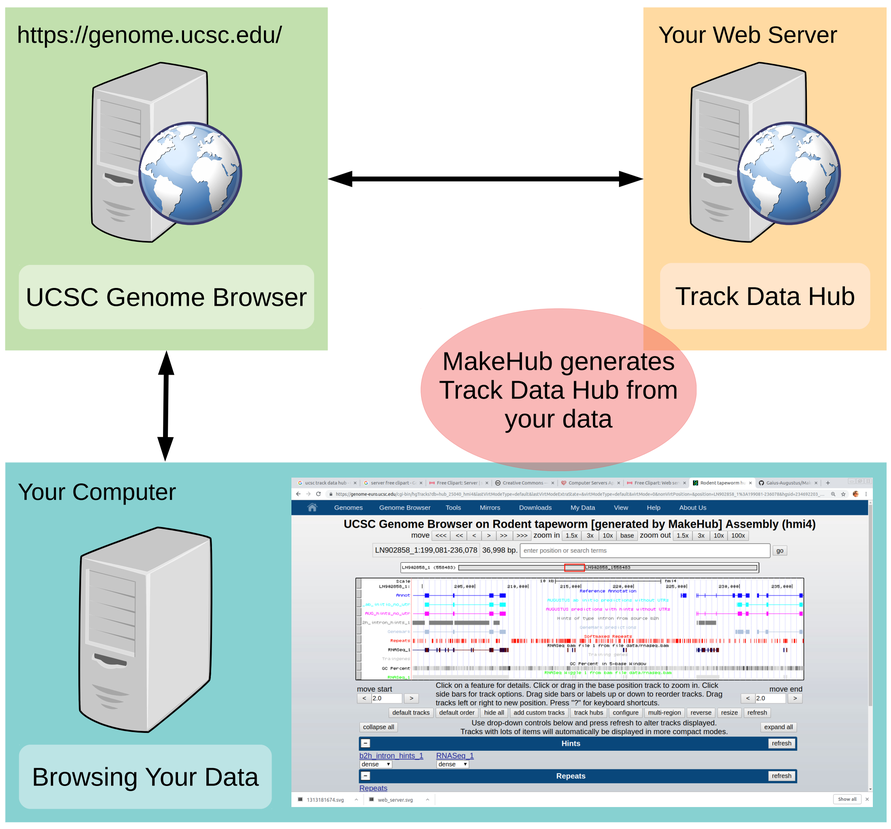
UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456

## Abstract

The UCSC Genome Browser [2] is one of the most powerful and most convenient tools for visualization of genomes with their annotation. Track data hubs [1] allow the display of externally hosted genomic data via publicly available UCSC Genome Browser instances. Creating your own assembly hub for a novel genome, however, is often a tedious task that involves many steps that are in part difficult for scientists without programming background.
MakeHub [3] has the goal to enable scientists to quickly and automatically generate assembly hubs of novel genomes, their annotation and informative RNA-Seq read alignments. Producing a complete assembly hub is a one-step process with MakeHub.
Implemented in Python, MakeHub utilizes tools provided by the UCSC Genome Browser group, SAMtools [4], and components of the gene Prediction tool AUGUSTUS [5]. MakeHub is integrated in the BRAKER [6,7] pipeline for fully automated and unsupervised RNA-Seq and/or protein based structural genome annotation. It is further compatible with the outputs of MAKER [8], GlimmerHMM[9], SNAP [10] and GeMoMa [11].

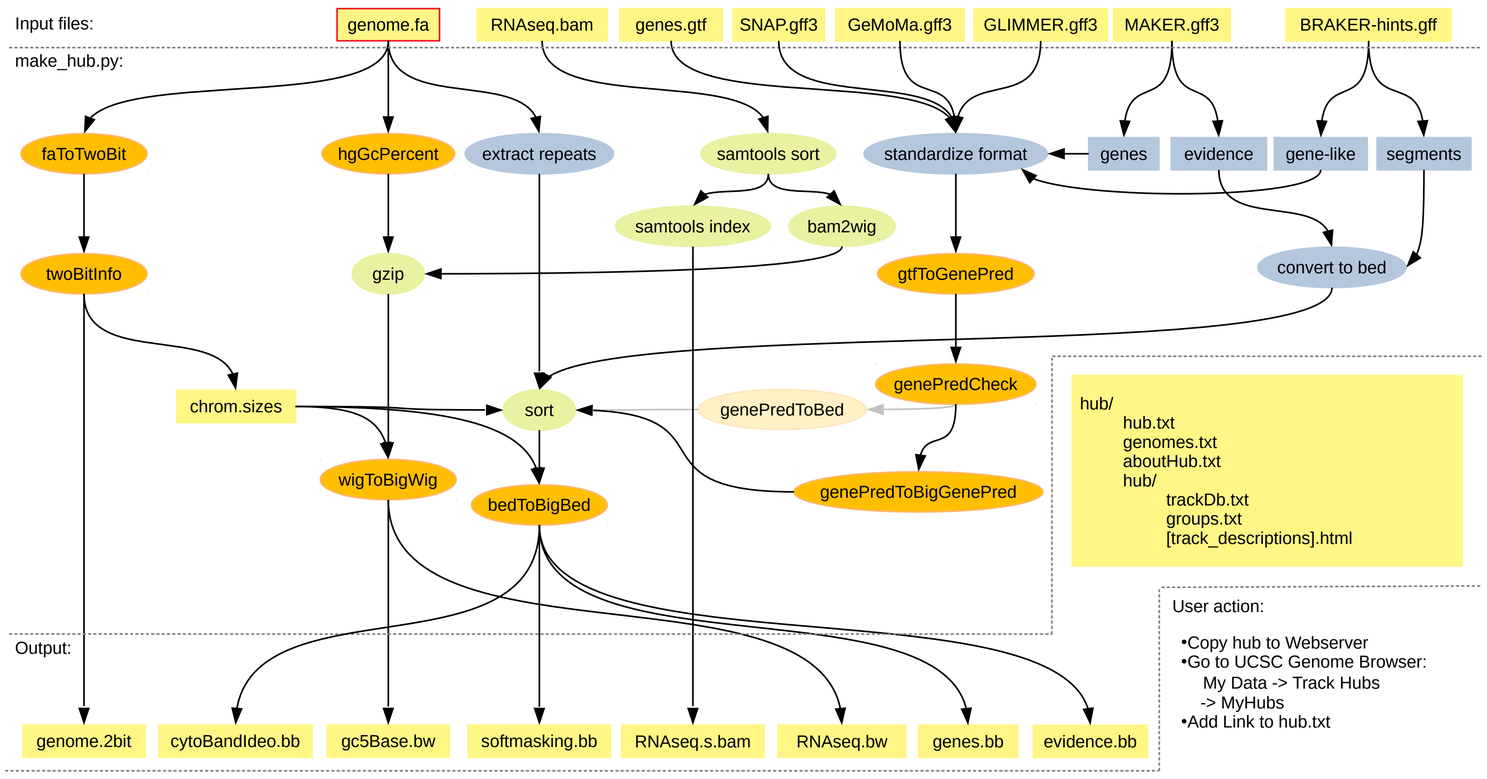
MakeHub is freely available at **https://github.com/Gaius-Augustus/MakeHub**.

## The Track Data Hub Principle




## MakeHub Creates Track Data Hubs

A single command invokes the MakeHub pipeline:




An example call that generates a hub with the following tracks:

- Repeats,
- Intron hints from RNA-Seq data,
- RNA-Seq coverage data,
- a gene model reference annotation,
- GeneMark-ET/EP [12, 13] predictions,
- training genes used to train AUGUSTUS,
- AUGUSTUS ab initio gene predictions,
- AUGUSTUS predictions with hints,
- base position,
- GC-content,
- restriction enzymes,
- perfect match to short sequence.

```
make_hub.py --short_label hmi4 --long_label "Rodent tapeworm" \
  --genome data/genome.fa --email katharina.hoff@uni-greifswald.de \
  --annot data/annot.gtf --bam data/rnaseq.bam \
  --braker_out_dir data --latin_name "Hymenolepsis microstoma" \
  --assembly_version GCA_000469805.2
```

## Software Dependencies

- Unix, e.g. Ubuntu Linux, including sort & gzip
- Python3
- Biopython
- UCSC Tools*:
  bedToBigBed, genePredCheck, faToTwoBit, gtfToGenePred, hgGcPercent, ixIxx, twoBitInfo, wigToBigWig, genePredToBed, genePredToBigGenePred
- SAMtools

*) will be downloaded automatically if missing

## Connecting Your Track Data Hub


## References

[1] Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, Kent WJ. 2014. "Track Data Hubs". Bioinformatics 30(7):1003-1005
[2] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. "UCSC Genome Browser". Genome Research 12(6):996-1006
[3] Hoff KJ. 2019. "MakeHub: Fully automated generation of UCSC Genome Browser Assembly Hubs". Genomics, Proteomics and Bioinformatics, in press; preprint at https://www.biorxiv.org/content/10.1101/550145v2
[4] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 2009. "The sequence alignment/map format and SAMtools". Bioinformatics 26(16):2078-2079
[5] Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. "Using native and syntenically mapped cDNA alignments to improve de novo gene finding". Bioinformatics 24(5):637-644
[6] Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2015. "BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS." Bioinformatics 32(5):767-769
[7] Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. "Whole-Genome Annotation with BRAKER". Methods of Molecular Biology 1962:65-95
[8] Holt C, Yandell M. 2011. "MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects". BMC Bioinformatics 12(1):491
[9] Majoros WH, Salzberg SL. 2004. "TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders". Bioinformatics 20(16):2878-2879
[10] Korf I. 2004. "Gene finding in novel genomes". BMC Bioinformatics 5:59
[11] Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. 2018. "Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi". BMC Bioinformatics 19(1):189.
[12] Lomsadze A, Burns PD, Borodovsky M. 2014. "Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm". Nucleic Acids Research 42(15):e119.
[13] Bruna T, Lomsadze A, Borodovsky M. 2020. "GeneMark-EP and -EP+: automatic eukaryotic gene prediction supported by spliced alignments:. Preprint at https://www.biorxiv.org/content/10.1101/2019.12.31.891218v1

## Acknowledgements

The international collaboration between the groups of Mark Borodovsky and Mario Stanke, supported by US National Institute of Health grant HG000783, gave rise to the development of MakeHub.