Assembly Data Hubs support viewing any sequence on the UCSC Genome Browser

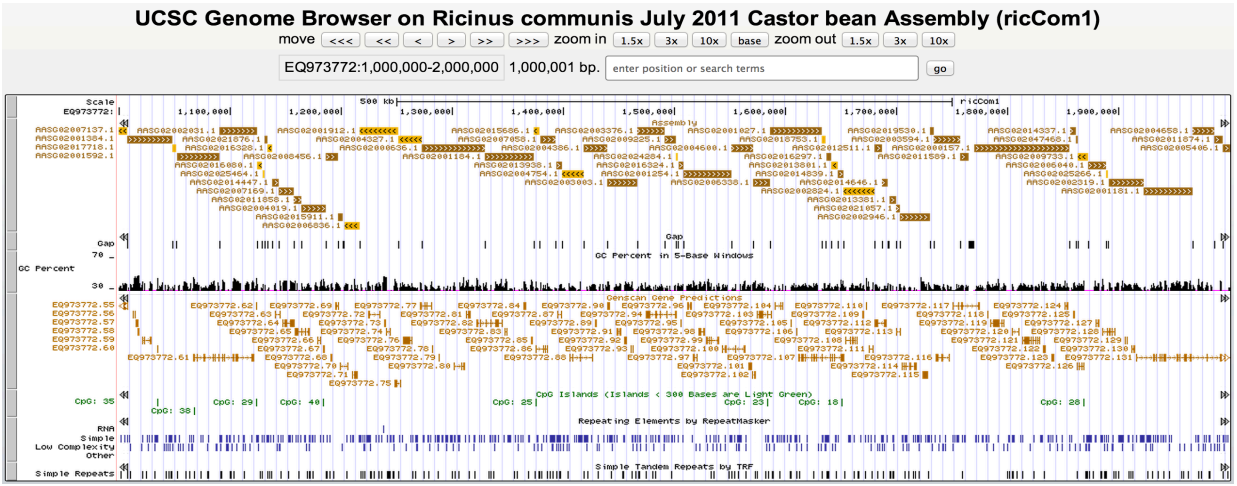
Brian J Raney, Ngan Nguyen, Tim R Dreszer , Galt P Barber, Hiram Clawson, Pauline A Fujita, Donna Karolchik, Ann S Zweig, Benedict Paten, William J Kent

University of California Santa Cruz, Center for Biomolecular Science &Engineering, Santa Cruz, CA, 95064

Easily build a UCSC Genome Browser with your sequence and annotations

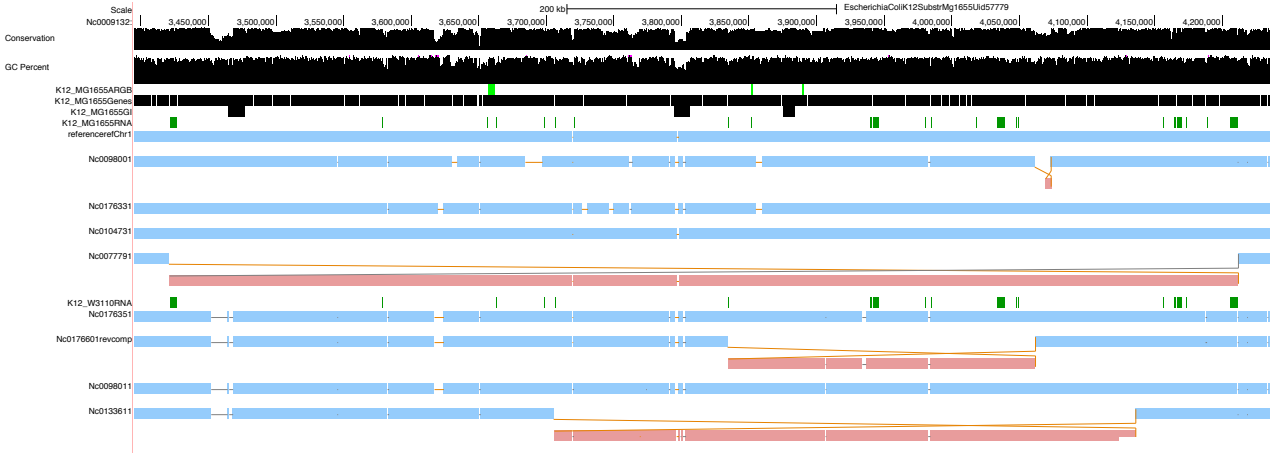
Assembly Data Hubs allow end-users to use the UCSC Genome Browser to view their own sequences with associated annotation, without the requirement that UCSC support a browser on that sequence. An Assembly Data Hub is a set of internet accessible data files that define the reference sequence to be used for a browser instance, as well as all the data files that define the annotation for that sequence. User sequences can be as complex as whole genome assemblies, or just a few scaffolds from a re-sequencing project. The end-user maintains control over the sequence and the annotations and can update them at her convenience.

Assembly Data Hubs are an extension to Track Data Hubs which allow user-level annotation on existing reference sequences. Track Data Hubs are flat files that are internet accessible. They need not exist on the same system, but can be distributed widely. Track hub annotations are stored as compressed binary indexed files in BigBed, BigWig, BAM, HAL, or VCF/tabix format. When a hub track is displayed in the Genome Browser, only the relevant data needed to support the view of the current genomic region is transmitted to UCSC, rather than the entire file. All transmitted data is cached on a UCSC server so future access is local to UCSC servers.



Build Comparative Assembly Hubs using Cactus and Snake Display

As an example of the ease of creation and the usefulness of Assembly Data Hubs we present the Cactus Alignment Pipeline, which we use to create an E. coli Reference Assembly in which 57 Escherichia coli and 9 Shigella complete genomes are aligned and a consensus reference is created. An assembly hub is created that supports the viewing of all input sequences and the consensus sequence on the UCSC Genome Browser. Each sequence is annotated with an automated pipeline that maps gene annotation from well-annotated genomes to the other genomes, and also generates useful annotations such as GC content and Mappability.



See our wiki for more information:
http://genomewiki.cse.ucsc.edu/index.php/GI2013

Acknowledgments

This work was funded by National Human Genome Research Institute award U41HG002371 to the UCSC Center for Genomic Science.

CENTER FOR BIOMOLECULAR SCIENCE & ENGINEERING
promoting discovery and invention for human health and well-being