August 2014
Version 2.0

Tutorial 1 – Overview of CHO Genome Resources

As of August 2014, The Chinese hamster genome database (Hammond et al. 2012) contains information from the CHO-K1 genome (Xu et al. 2011), the Chinese hamster mitochondrial genome (Partridge et al. 2007), and the Chinese hamster nuclear genome (Lewis et al. 2013; Brinkrolf et al. 2013), as described below.

The genome databases at CHOgenome.org are updated shortly after new or updated *Cricetulus griseus* information is published or released. Moving forward, upon the addition of updated genome databases to CHOgenome.org, the outdated databases will be reclassified as legacy databases. Only one legacy database for each unique *Cricetulus griseus* cell type or cell line will be maintained at a time; upon the reclassification of an outdated genome database to a legacy database, the previous legacy database (should there be one) will be retired. This process will promote the most recent information while maintaining outdated information in a manner that benefits all CHO genome community users.

Similarly, when significant upgrades are made to the website, the previous version of the website will be temporarily maintained in parallel, permitting users a transition period. Please feel free to contact us via email with any issues, comments, or suggestions.

1. CHO-K1 genome assembly and annotation

The CHO-K1 genome (CriGri_1.0) was assembled from Illumina sequencing data by a whole genome shotgun (WGS) approach in which the CHO genome is sheared into smaller pieces, sequenced, and then assembled into a draft genome sequence.

A) CHO-K1 GenBank

The CHO-K1 genomic information is available at NCBI GenBank using the accession codes described below.

Overlapping reads from short-insert (< 500 bp) paired-end libraries are aligned and assembled into contigs. Contigs are continuous sequences containing no gaps.

The CHO-K1 WGS contigs have accession numbers of AFTD01000001-AFTD01265786.

Reads from all paired-end libraries, in increasing order of insert size, are used to order contigs and assemble scaffolds. Scaffolds may contain gaps (size can be estimated from library insert size).

The CHO-K1 WGS scaffolds have accession numbers of JH000001-JH109151.

The WGS scaffolds are annotated by prediction of coding sequences (CDS) and protein sequences.

All genes derived from this genome sequencing project have been assigned the locus tag prefix I79.

Gene symbols were assigned to CHO protein-coding genes in the Chinese hamster genome database based on annotation of homologous proteins.

B) CHO-K1 RefSeq (2012)

Using the same CHO-K1 GenBank assembly, NCBI annotated the CHO-K1 genome using a different annotation pipeline and created the CHO-K1 RefSeq (2012) database. Additional information regarding the NCBI RefSeq initiative and annotation process can be found at http://www.ncbi.nlm.nih.gov/refseq/about/, while details regarding the differences between RefSeq and GenBank can be found at http://www.ncbi.nlm.nih.gov/projects/RefSeq/GenBankvsRefSeq.pdf.

The scaffold sequences were unaltered, but the scaffolds were assigned different accession numbers. The annotated genes were assigned new gene ID numbers.

The CHO-K1 RefSeq scaffolds have accession numbers of NW_003613580.1-NW_003722730.1 for the nuclear genome sequences and NC_007936.1 for the mitochondrial genome sequences.

The CHO-K1 RefSeq (2012) genes have ID numbers of 100750331-100775103.

As the updated RefSeq annotations, i.e. RefSeq (2014), are released, many of the NCBI transcript and protein IDs in the RefSeq (2012) assembly have been discontinued and replaced with new IDs. The discontinued IDs are no longer accessible on the NCBI RefSeq database and may or may not contain links to the new IDs. We have archived the RefSeq (2012) database to enable users who were working with this database to track the outdated information (IDs) from the previous genome database entries.

C) CHO-K1 RefSeq (2014)

In May 2014, NCBI reannotated the CHO-K1 RefSeq assembly using an updated annotation pipeline. The RefSeq (2014) database scaffolds and contigs were not altered and have the same accession numbers as the RefSeq (2012) assembly. The RefSeq (2014) assembly does contain many new and updated transcript variants and protein isoforms.

CHO-K1 RefSeq (2014) is the most recent CHO-K1 reference genome assembly and database.

Within the RefSeq FTP site, the CHO-K1 RefSeq (2014) files are labeled as species reference (ref_CriGri).

2. Chinese hamster (CH) genome assembly and annotation

A) CH RefSeq (2014)

Lewis et al. (2013) reported the Chinese hamster genome sequences (C_griseus_v1.0). The CH genome was assembled from Illumina sequencing data by SOAPdenovo2.2 assembler. The CHO-K1 and CH genome sequences used the same NCBI annotation process to generate the CHO-K1 RefSeq (2014) and CH RefSeq (2014) databases, respectively.

Within the NCBI RefSeq FTP site, the CH RefSeq (2014) files are labeled alternative RefSeq (alt_C_griseus).

While the CHO-K1 RefSeq files are species representative, as evidenced by the different labels within the NCBI RefSeq FTP files, homologous genes between the CHO-K1 and CH genomes were assigned the same gene IDs.

The CH scaffolds have accession numbers of NW_006834731.1-NW_006887440.1.

The CH RefSeq gene IDs range between 100682525 and 103163833.

CH RefSeq (2014) is the most recent CH reference genome assembly and database.

B) CH-17A/GY Genome Chromosome Sequences

In 2013, Brinkrolf et al. sequenced the physically isolated Chinese hamster chromosomes (Cgr1.0). The genome sequences were assembled at a scaffold level and the scaffolds are organized according to the chromosome they are affiliated with.

This genome draft was not selected for the RefSeq annotation process and as a result, the assembly is not hosted on the CHO genome database (genome search or genome browser pages). However, the scaffold and protein sequence databases are available on the CHO BLAST server.

Genome Sequence Publications

Brinkrolf K, Rupp O, Laux H, Kollin F, Ernst W, Linke B, Kofler R, Romand S, Hesse F, Budach WE, Galosy S, Müller D, Noll T, Wienberg J, Jostock T, Leonard M, Grillari J, Tauch A, Goesmann A, Helk B, Mott JE, Pühler A, Borth N (2013) Chinese hamster genome sequenced from sorted chromosomes. Nature Biotechnology 31(8):694-695.

Hammond S, Kaplarevic M, Borth N, Betenbaugh MJ, Lee KH (2012) Chinese hamster genome database: an online resource for the CHO community at www.CHOgenome.org. Biotechnology Bioengineering 109(6):1353-1356.

Lewis NE, Liu X, Li Y, Nagarajan H, Yerganian G, O'Brien E, Bordbar A, Roth AM, Rosenbloom J, Bian C, Xie M, Chen W, Li N, Baycin-Hizal D, Latif H, Forster J, Betenbaugh MJ, Famili I, Xu X, Wang J, Palsson BO (2013) Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome. Nature Biotechnology 31(8):759-765.

Partridge MA, Davidson MM, Hei TK (2007) The complete nucleotide sequence of Chinese hamster (Cricetulus griseus) mitochondrial DNA. DNA Sequence 18(5): 341-346.

Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, Chen W, Xie M, Wang W, Hammond S, Andersen MR, Neff N, Passarelli B, Koh W, Fan HC, Wang J, Gui Y, Lee KH, Betenbaugh MJ, Quake SR, Famili I, Palsson BO, Wang J (2011) The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nature Biotechnology 29(8): 735-741.