

RNA-Seq: GTF/GFF and Splice Aligner

BCB 5250 Introduction to Bioinformatics II

Spring 2020

Tae-Hyuk (Ted) Ahn

Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University



SAINT LOUIS
UNIVERSITY™

— EST. 1818 —

RNA-Seq Lab

- <https://www.yeastgenome.org/dataset/GSE53720>

Dataset: RNA-seq Experiments from *Saccharomyces cerevisiae*

GEO ID:	GSE53720
Reference:	Dang W, et al. (2014)
Channels:	1
Conditions:	4
Description:	RNA-seq Experiments from Calorie Restricted and Non-Restricted WT Yeast
Categories:	nutrient utilization, stress

NCBI GEO Overview

- <https://www.ncbi.nlm.nih.gov/geo/info/overview.html>

General overview

GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community.

The three main goals of GEO are to:

1. Provide a robust, versatile database in which to efficiently store high-throughput functional genomic data (see [Data organization](#))
2. Offer simple submission procedures and formats that support complete and well-annotated data deposits from the research community (see [Submission guide](#))
3. Provide user-friendly mechanisms that allow users to query, locate, review and download studies and gene expression profiles of interest (see [Query and analysis](#))

Please see the [GEO Documentation](#) listings to find more information about various aspects of GEO.

RNA-Seq Lab

- <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53720>

Platforms (1) [GPL13821](#) Illumina HiSeq 2000 (*Saccharomyces cerevisiae*)

Samples (4) [GSM1299413](#) WT_NR_A

[Less...](#) [GSM1299414](#) WT_NR_B

[GSM1299415](#) WT_CR_A

[GSM1299416](#) WT_CR_B

This SubSeries is part of SuperSeries:

[GSE53721](#) Calorie Restricted and Non-Restricted Yeast

Relations

BioProject [PRJNA232717](#)

SRA [SRP034844](#)

NCBI BioProject

- <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA232717>

Saccharomyces cerevisiae (baker's yeast)

RNA-seq Experiments from Saccharomyces cerevisiae

RNA-seq Experiments from Calorie Restricted and Non-Restricted WT Yeast Overall design: We used RNA-seq to study transcriptome changes under Calorie Restricted and Non-restricted *Saccharomyces cerevisiae*

Accession	PRJNA232717; GEO: GSE53720
Data Type	Transcriptome or Gene expression
Scope	Multiisolate
Organism	Saccharomyces cerevisiae [Taxonomy ID: 4932] Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces; <i>Saccharomyces cerevisiae</i>
Publications	Dang W <i>et al.</i> , "Inactivation of yeast Isw2 chromatin remodeling enzyme mimics longevity effect of calorie restriction via induction of genotoxic stress response.", <i>Cell Metab</i> , 2014 Jun 3;19(6):952-66
Submission	Registration date: 30-Dec-2013 Departments of Cell & Developmental Biology; Biology; Genetics, University of Pennsylvania
Relevance	Model Organism

Accession: PRJNA232717 ID: 232717

See Genome Information for *Saccharomyces cerevisiae*

NAVIGATE UP

This project is a component of the Calorie Restricted and Non-Restricted Yeast

NAVIGATE ACROSS

1 additional project is a component of the Calorie Restricted and Non-Restricted Yeast.

3576 additional projects are related

NCBI SRA

- https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=232717

- [GSM1299416: WT_CR_B; *Saccharomyces cerevisiae*; RNA-Seq](#)
 - 1. 1 ILLUMINA (Illumina HiSeq 2000) run: 101.2M spots, 5.1G bases, 3Gb downloads
Accession: SRX403445
- [GSM1299415: WT_CR_A; *Saccharomyces cerevisiae*; RNA-Seq](#)
 - 2. 1 ILLUMINA (Illumina HiSeq 2000) run: 93.1M spots, 4.7G bases, 2.7Gb downloads
Accession: SRX403444
- [GSM1299414: WT_NR_B; *Saccharomyces cerevisiae*; RNA-Seq](#)
 - 3. 1 ILLUMINA (Illumina HiSeq 2000) run: 110.9M spots, 5.5G bases, 3.4Gb downloads
Accession: SRX403443
- [GSM1299413: WT_NR_A; *Saccharomyces cerevisiae*; RNA-Seq](#)
 - 4. 1 ILLUMINA (Illumina HiSeq 2000) run: 83.8M spots, 4.2G bases, 2.4Gb downloads
Accession: SRX403442

Summary ▾

SRA Toolkit

- \$ fastq-dump SRR1066660
- <https://github.com/ncbi/sra-tools>
- <https://github.com/ncbi/sra-tools/wiki/HowTo:-fasterq-dump>

fasterq-dump

a faster *fastq-dump*

The fasterq-dump tool uses temporary files and multi-threading to speed up the extraction of FASTQ from SRA-accessions.

If a minimal commandline is given:

```
$fasterq-dump SRR0000001
```

Getting started with HISAT, StringTie, and Ballgown

<https://www.nature.com/articles/nprot.2016.095#an1>



Protocol | Published: 11 August 2016

Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek & Steven L Salzberg 

Nature Protocols **11**, 1650–1667 (2016) | Download Citation 

Getting started with HISAT, StringTie, and Ballgown

- wget ftp://ftp.ccb.jhu.edu/pub/RNAseq_protocol/chrX_data.tar.gz
 - You can copy the data from
hopper.slu.edu:/public/ahnt/courses/bcb5250/rna_seq_lab/chrX_data.tar.gz
 - How to recursively wget the files in the directory?
 - wget -c -r ftp://ftp.ccb.jhu.edu/pub/RNAseq_protocol

Build Star Index

- Download STAR aligner
- Check the STAR manual
- Map all reads to the reference
 - Build the index
 - Recommendation:
 - \$ mkdir star_indexes
 - \$ STAR --runThreadN 12 --runMode genomeGenerate --genomeDir star_indexes --genomeFastaFiles genome/chrX.fa --sjdbGTFfile genes/chrX.gtf --sjdbOverhang 74 --genomeSAindexNbases 12)
 - Create a shell script to align all reads to once
 - Recommendation:
 - \$ STAR --genomeDir star_indexes --runThreadN 6 --readFilesIn XXX.fq --outFileNamePrefix star_align/XXX --outSAMtype BAM SortedByCoordinate --outSAMunmapped Within --outSAMattributes Standard

Base Shell Example

```
#!/bin/bash
dir="chrX_data"
for entry in $dir/samples/*1.fastq.gz
do
    name=$(echo $entry | egrep -o "ERR[^_]+")
    hisat2 -p 8 --dta -x "$dir"/indexes/chrX_tran -1 "$dir"/samples/"$name"_chrX_1.fastq.gz -
2 "$dir"/samples/"$name"_chrX_2.fastq.gz -S "$dir"/map/"$name"_chrX.sam
done
```