RNA-Seq: Differential Expression Analysis

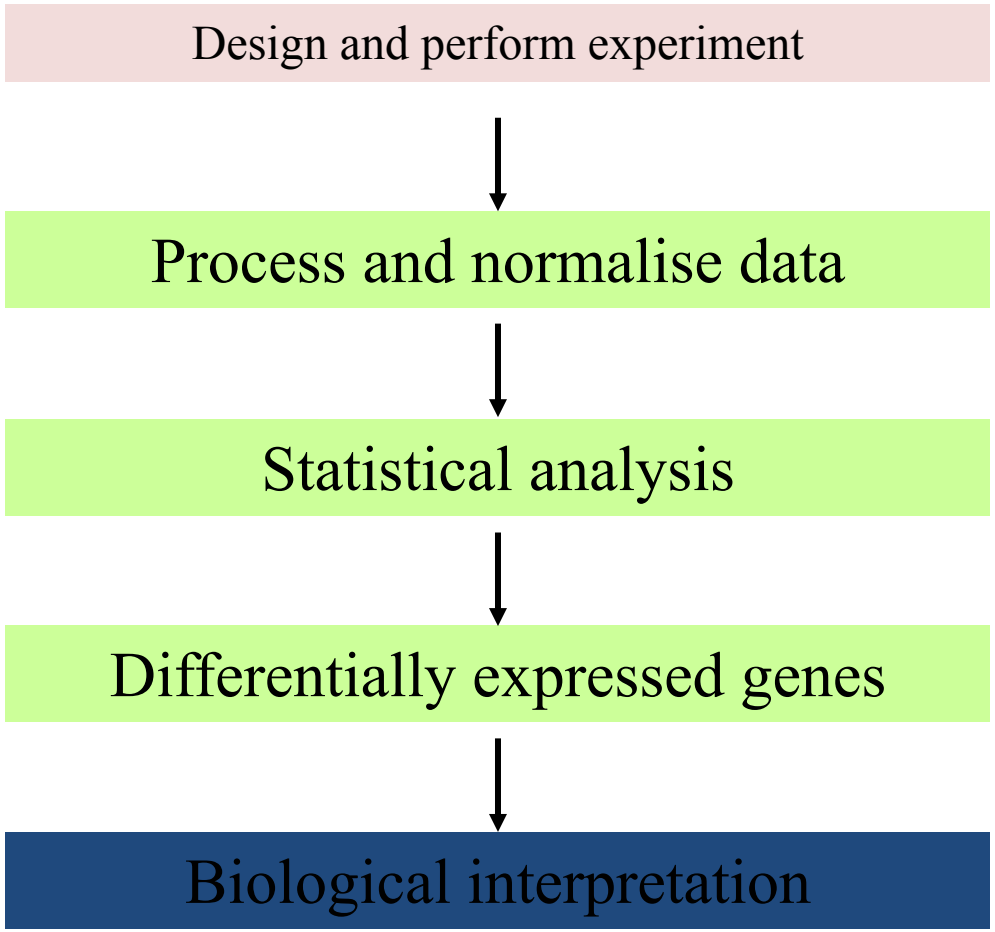
BCB 5250 Introduction to Bioinformatics II

Spring 2020

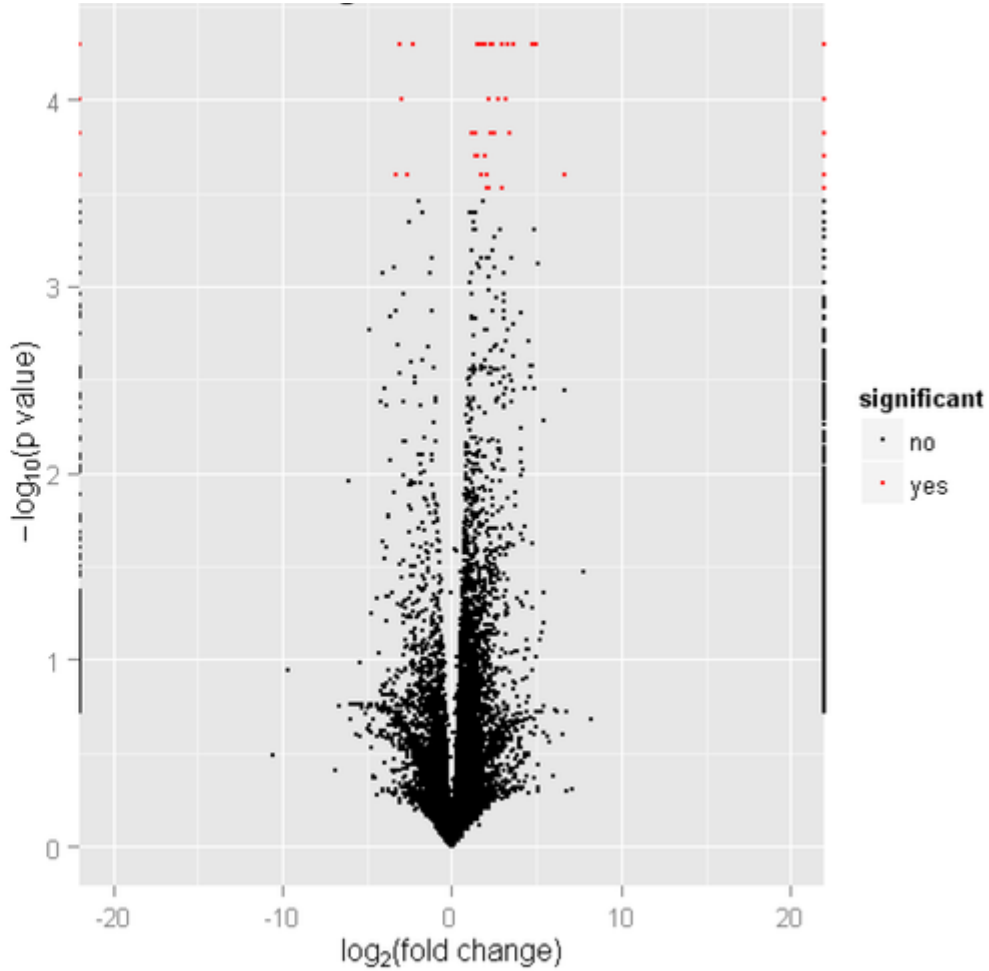
Tae-Hyuk (Ted) Ahn
Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

SAINT LOUIS
UNIVERSITY
EST. 1818

RNA-Seq Pipeline

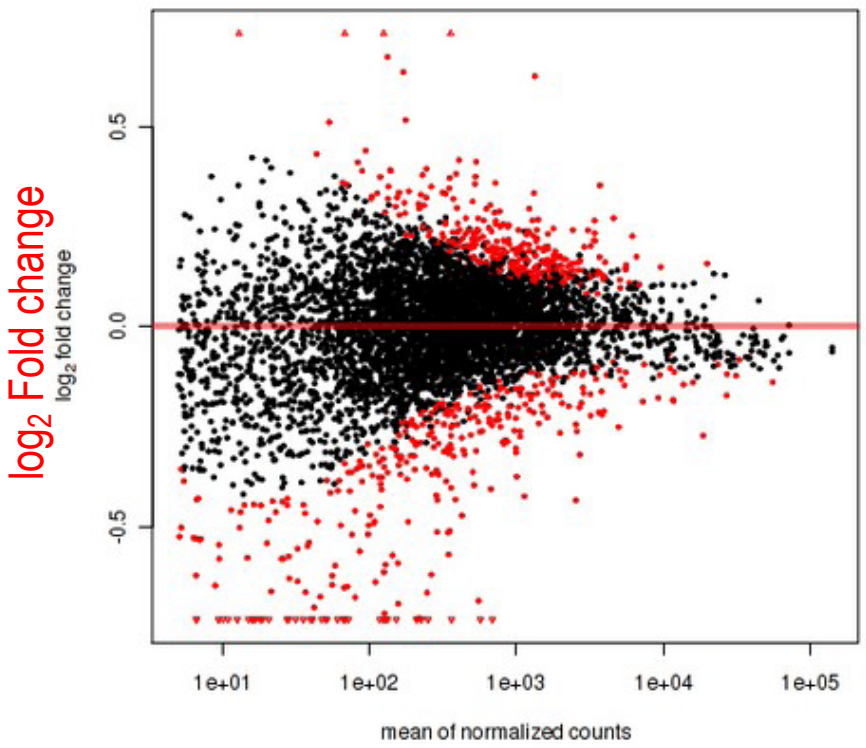


Illustrating DE results



$\log_2$ Fold change

volcano-plot: p value versus the log2 fold change between 2 conditions.



MA-plot: mean of counts versus the log2 fold change between 2 conditions.

Some popular DE tools

	DESeq2	edgeR	Cuffdiff / cummRbund
Normalization	DESeq sizeFactor/geometric	TMM/upper quartile/RLE	geometric, upper, quartile, fpkm
Read count distribution assumption	Negative binomial	Negative binomial	Negative binomial
DE Test	exact test	exact test	t test
Ref	Genome Biol 2010;11:R106.	Bioinformatics 26, 139–140 (2010)	Nature biotechnology *31*, 46-53 (2013)

Statistics are Important

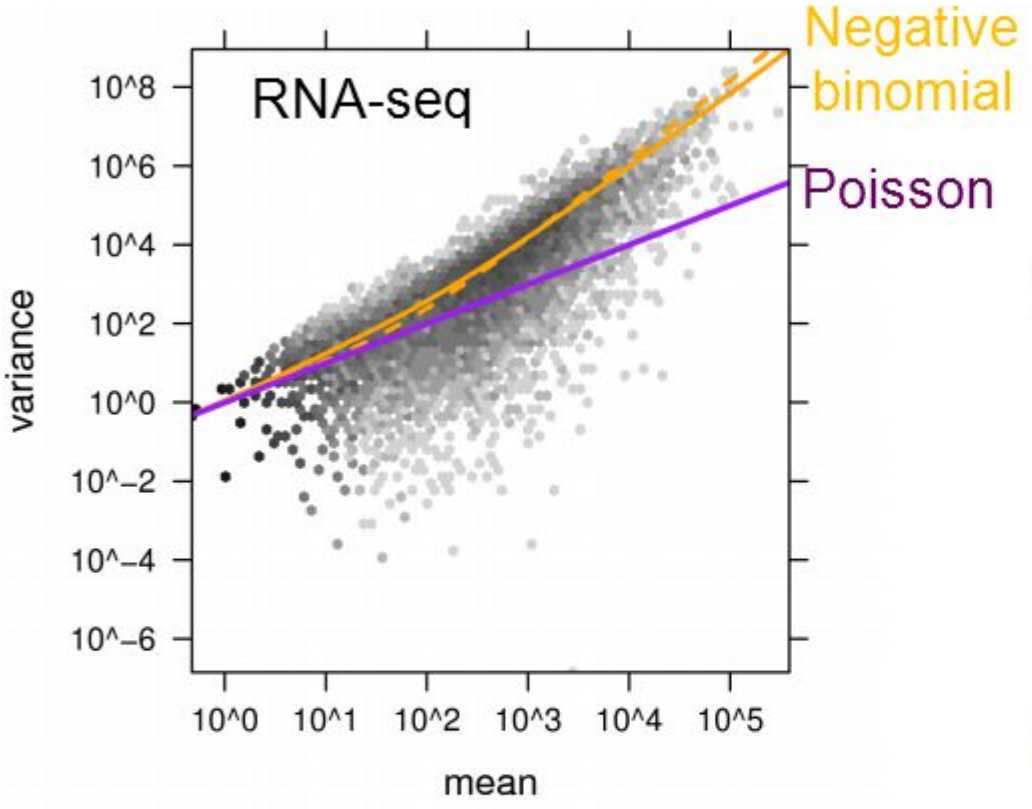
We use these tests for different reasons and under different circumstances.

1. z-test. A z-test assumes that our observations are independently drawn from a Normal distribution with unknown mean and *known variance*. A z-test is used primarily when we have quantitative data. (i.e. weights of rodents, ages of individuals, systolic blood pressure, etc.) However, z-tests can also be used when interested in proportions. (i.e. the proportion of people who get at least eight hours of sleep, etc.)
2. t-test. A t-test assumes that our observations are independently drawn from a Normal distribution with unknown mean and *unknown variance*. Note that with a t-test, we do not know the population variance. This is far more common than knowing the population variance, so a t-test is generally more appropriate than a z-test, but practically there will be little difference between the two if sample sizes are large.

With z- and t-tests, your alternative hypothesis will be that your population mean (or population proportion) of one group is either not equal, less than, or greater than the population mean (or proportion) or the other group. This will depend on the type of analysis you seek to do, but your null and alternative hypotheses directly compare the means/proportions from the two groups.

3. Chi-squared test. Whereas z- and t-tests concern quantitative data (or proportions in the case of z), chi-squared tests are appropriate for qualitative data. Again, the assumption is that observations are independent of one another. In this case, you aren't seeking a particular relationship. Your null hypothesis is that no relationship exists between variable one and variable two. Your alternative hypothesis is that a relationship does exist. This doesn't give you specifics as to how this relationship exists (i.e. In which direction does the relationship go) but it will provide evidence that a relationship does (or does not) exist between your independent variable and your groups.
4. Fisher's exact test. One drawback to the chi-squared test is that it is *asymptotic*. This means that the p-value is accurate for very large sample sizes. However, if your sample sizes are small, then the p-value may not be quite accurate. As such, Fisher's exact test allows you to exactly calculate the p-value of your data and not rely on approximations that will be poor if your sample sizes are small.

How to calculate variance



- RNA-Seq data was initially modeled as count data fitting a Poisson distribution like the microarray data.
- Issue: genes with high mean counts tend to show more variance between samples than genes with low mean counts (overdispersion)
- Solution: Negative binomial distribution (= Poisson distribution + local regression)

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2),$$

Biological Interpretation

- An obvious way to gain biological insight is to assess the differentially expressed genes in terms of their known function(s)
- Required an automated and objective (statistical) approach
- Functional profiling or pathway analysis

Why functional interpretation

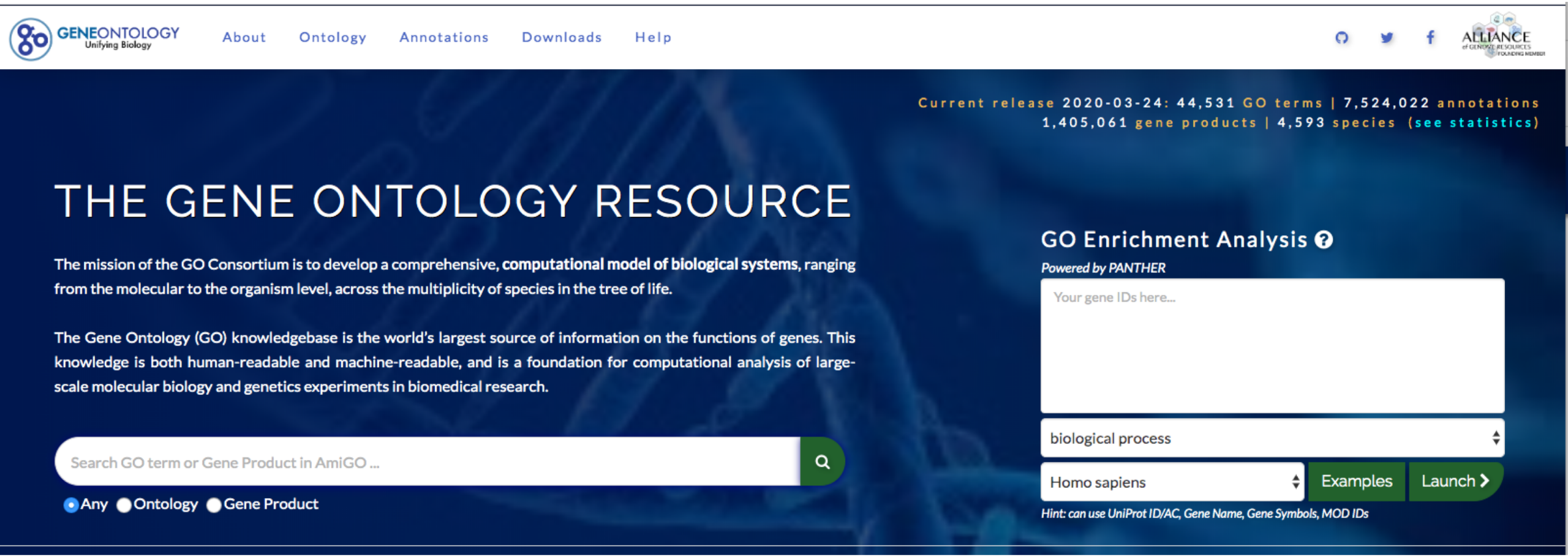
- High-throughput experiments do not produce biological findings
- Genes do not work alone, but in an intricate network of interactions
- Helps interpret the data in the context of biological processes, pathways and networks
- Global perspective on the data and problem at hand

Early functional analyses

- Manually annotate list of differentially expressed (DE) genes
- Extremely time-consuming, not systematic, user-dependent
- Group together genes with similar function
- Conclude functional categories with most DE genes important in disease/condition under study
- BUT may not be the right conclusion

Gene Ontology

- http://geneontology.org/



What is the Gene Ontology?

- Set of biological phrases (terms) which are applied to genes:
  - protein kinase
  - apoptosis
  - Membrane
- Genes are linked, or associated, with GO terms by trained curators at genome databases
  - known as ‘gene associations’ or GO annotations
- Some GO annotations created automatically

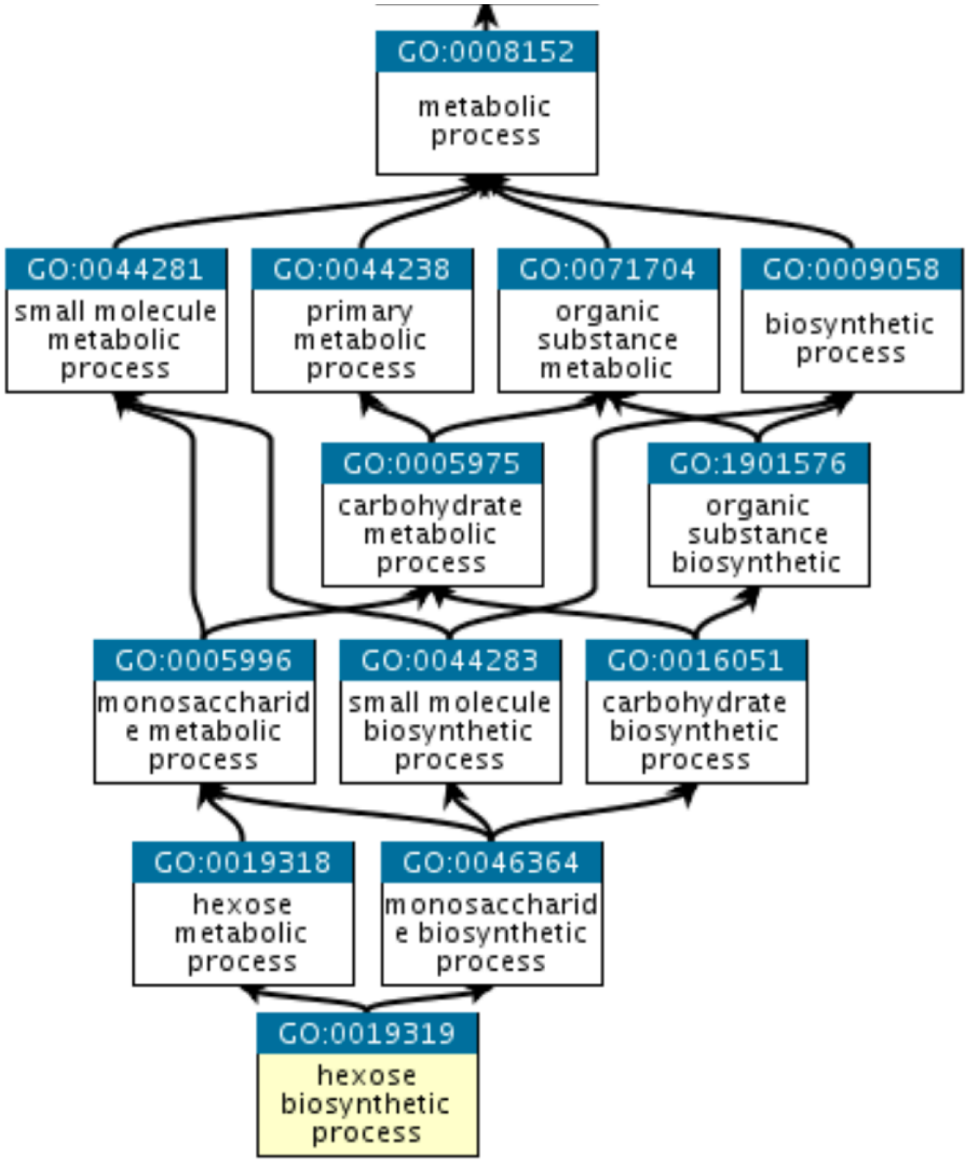
Gene Ontology overview

An ontology is a formal representation of a body of knowledge within a given domain. Ontologies usually consist of a set of classes (or terms or concepts) with relations that operate between them. The Gene Ontology (GO) describes our knowledge of the biological domain with respect to three aspects:

Molecular Function	Molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as "catalysis" or "transport". GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products (*i.e.* a protein or RNA), but some activities are performed by molecular complexes composed of multiple gene products. Examples of broad functional terms are *catalytic activity* and *transporter activity*; examples of narrower functional terms are *adenylate cyclase activity* or *Toll-like receptor binding*. To avoid confusion between gene product names and their molecular functions, GO molecular functions are often appended with the word "activity" (a *protein kinase* would have the GO molecular function *protein kinase activity*).
Cellular Component	The locations relative to cellular structures in which a gene product performs a function, either cellular compartments (*e.g.*, *mitochondrion*), or stable macromolecular complexes of which they are parts (*e.g.*, the *ribosome*). Unlike the other aspects of GO, cellular component classes refer not to processes but rather a cellular anatomy.
Biological Process	The larger processes, or 'biological programs' accomplished by multiple molecular activities. Examples of broad biological process terms are *DNA repair* or *signal transduction*. Examples of more specific terms are *pyrimidine nucleobase biosynthetic process* or *glucose transmembrane transport*. Note that a biological process is not equivalent to a pathway. At present, the GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

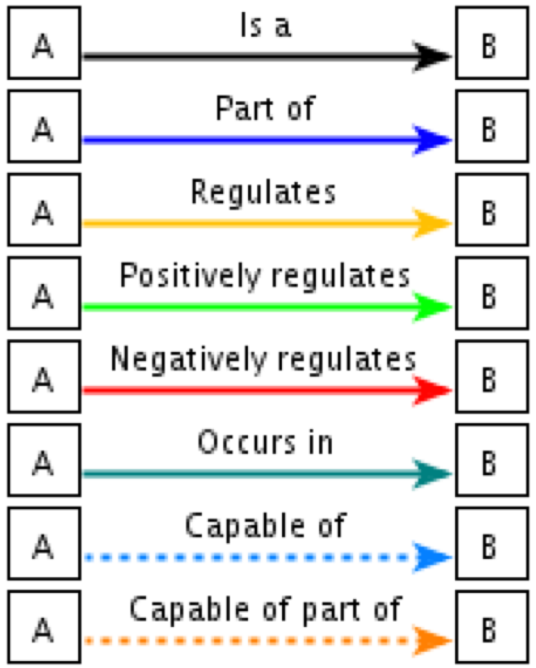
The GO graph

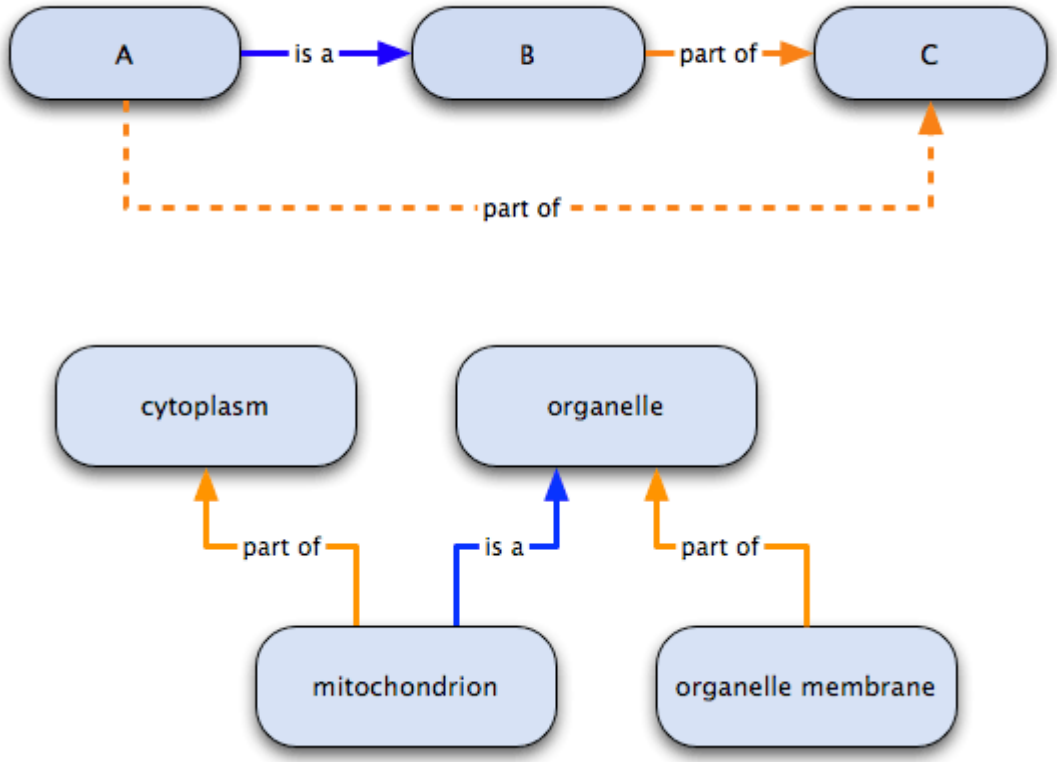
The structure of GO can be described in terms of a graph, where each GO term is a node, and the relationships between the terms are edges between the nodes. GO is loosely hierarchical, with 'child' terms being more specialized than their 'parent' terms, but unlike a strict hierarchy, a term may have more than one parent term (note that the parent/child model does not hold true for all types of relations, see the relations documentation). For example, the biological process term hexose biosynthetic process has two parents, hexose metabolic process and monosaccharide biosynthetic process. This reflect the fact that *biosynthetic process* is a subtype of *metabolic process* and a *hexose* is a subtype of *monosaccharide*.



Gene Ontology

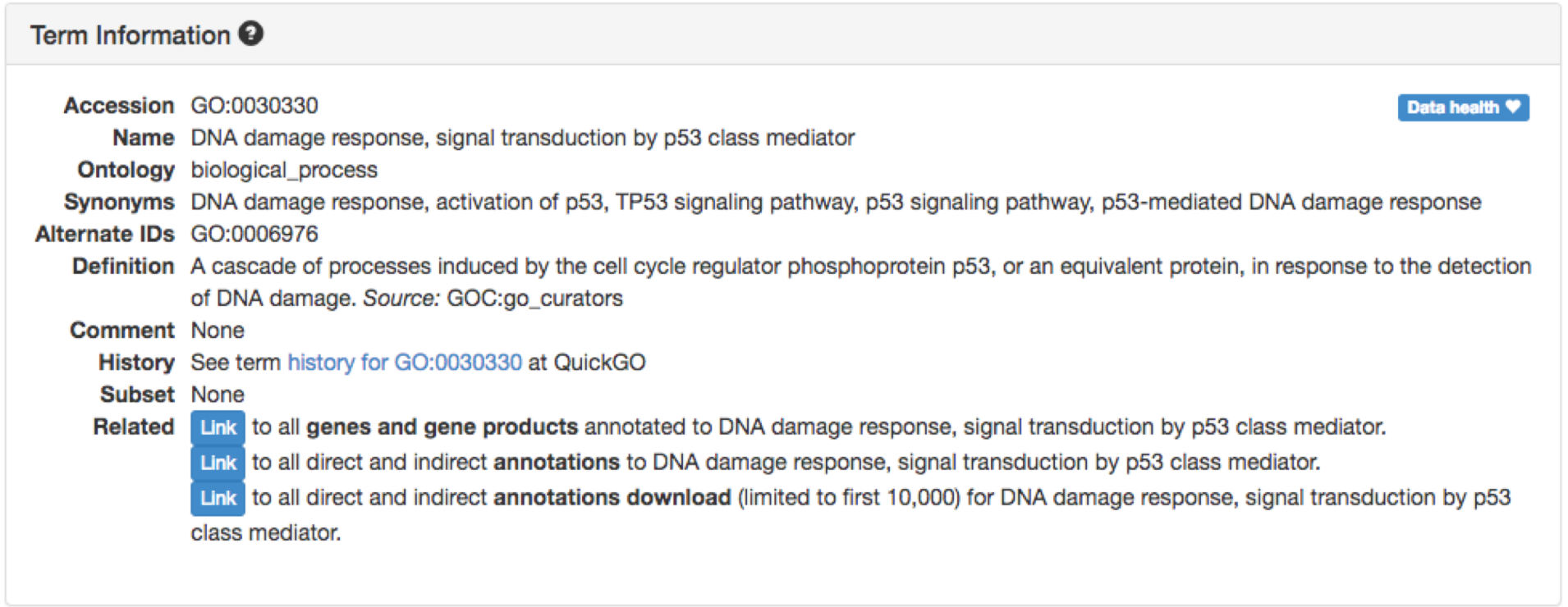
- Based on “is a” or “part of” relationship





Example: GO:0030330

- Search “GO:0030330” in google
  - QuickGO
  - AmiGO
  - Gowiki (GONUTS)



Term Information

Accession GO:0030330 Data health

Name DNA damage response, signal transduction by p53 class mediator

Ontology biological_process

Synonyms DNA damage response, activation of p53, TP53 signaling pathway, p53 signaling pathway, p53-mediated DNA damage response

Alternate IDs GO:0006976

Definition A cascade of processes induced by the cell cycle regulator phosphoprotein p53, or an equivalent protein, in response to the detection of DNA damage. *Source:* GOC:go_curators

Comment None

History See term history for GO:0030330 at QuickGO

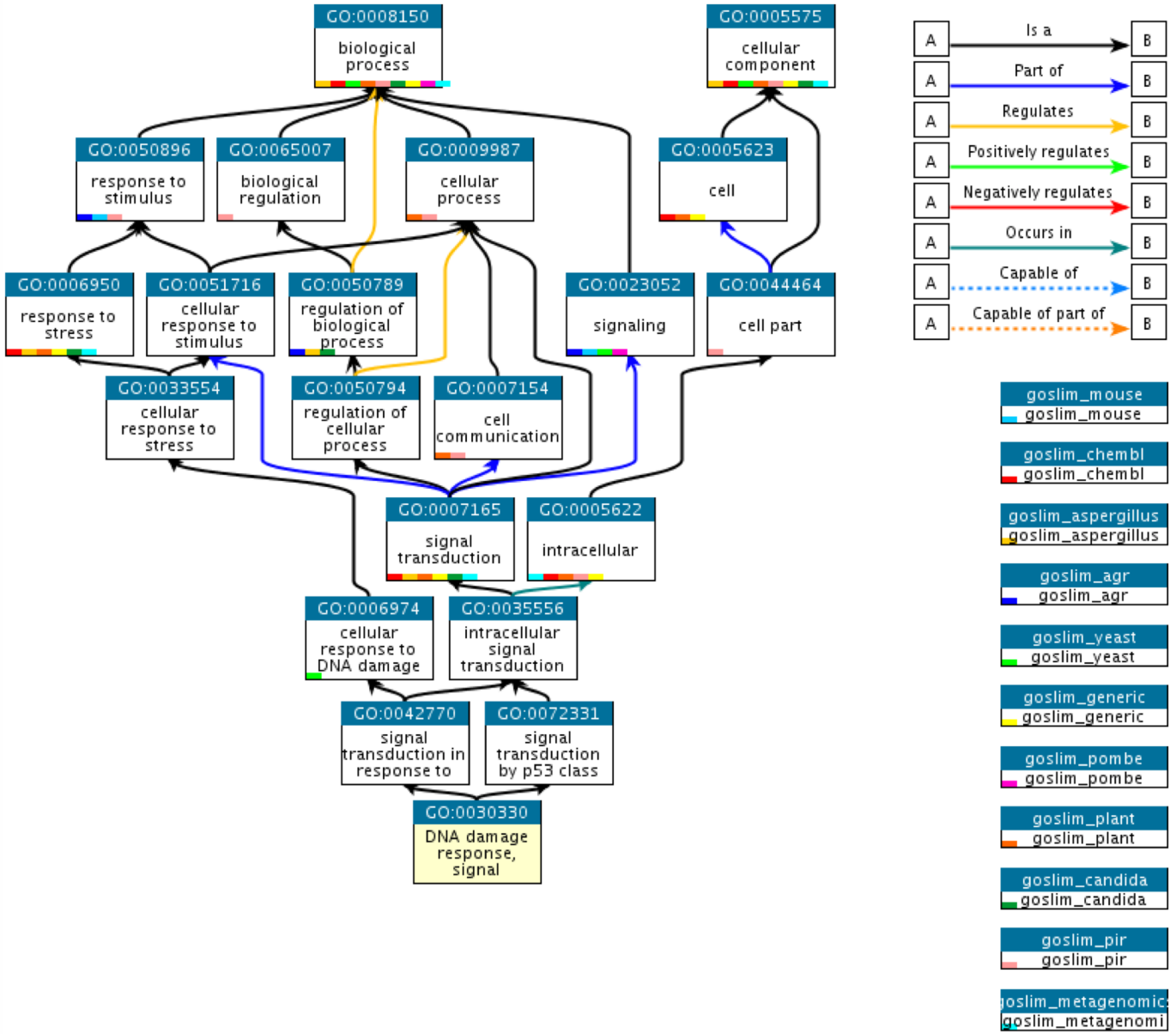
Subset None

Related Link to all **genes and gene products** annotated to DNA damage response, signal transduction by p53 class mediator.
Link to all direct and indirect **annotations** to DNA damage response, signal transduction by p53 class mediator.
Link to all direct and indirect **annotations download** (limited to first 10,000) for DNA damage response, signal transduction by p53 class mediator.

QuickGO

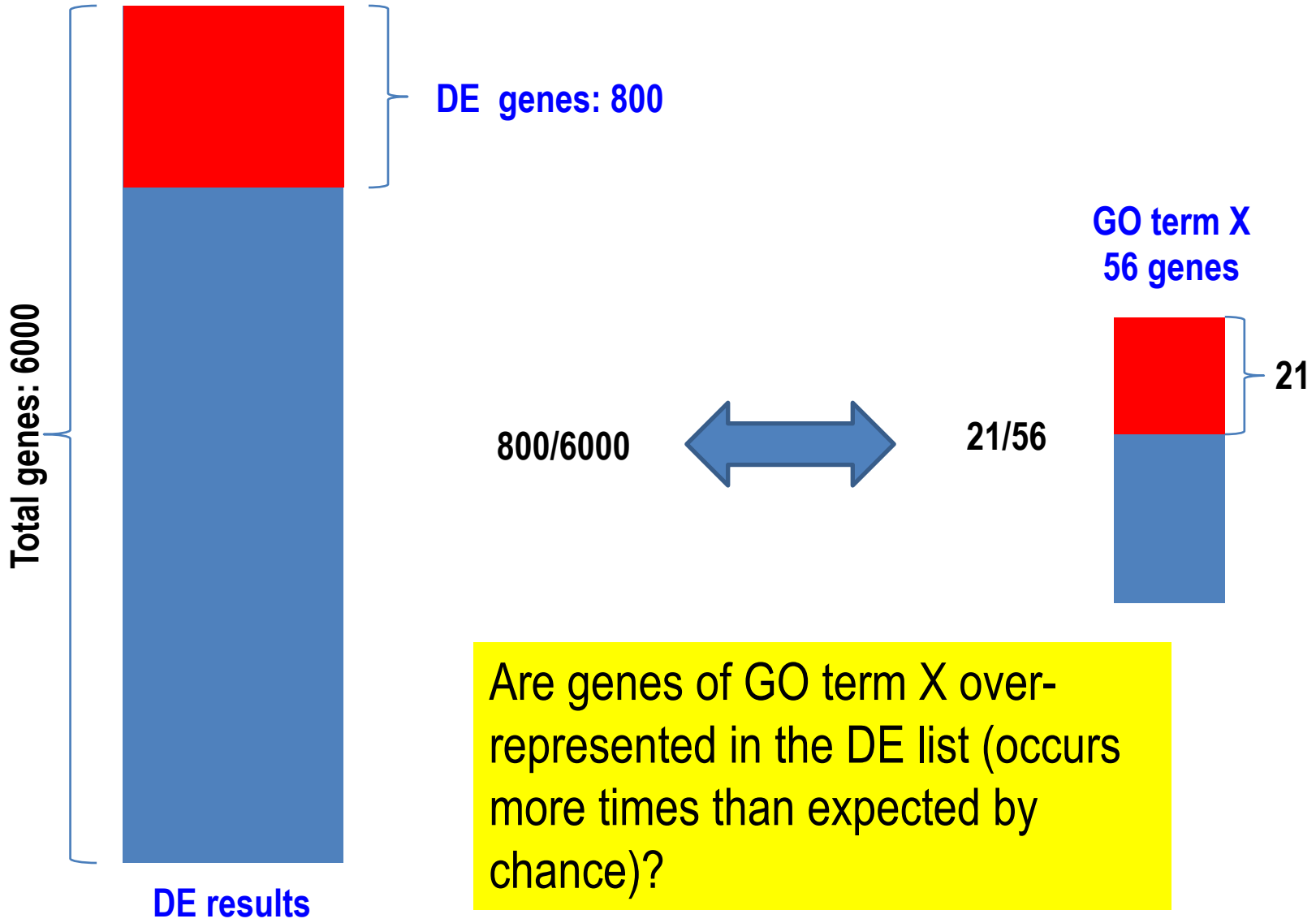
Ancestor Chart

Ancestor chart for GO:0030330



QuickGO - http://www.ebi.ac.uk/QuickGO

GO enrichment analysis



Are genes of GO term X over-represented in the DE list (occurs more times than expected by chance)?

GO Tools

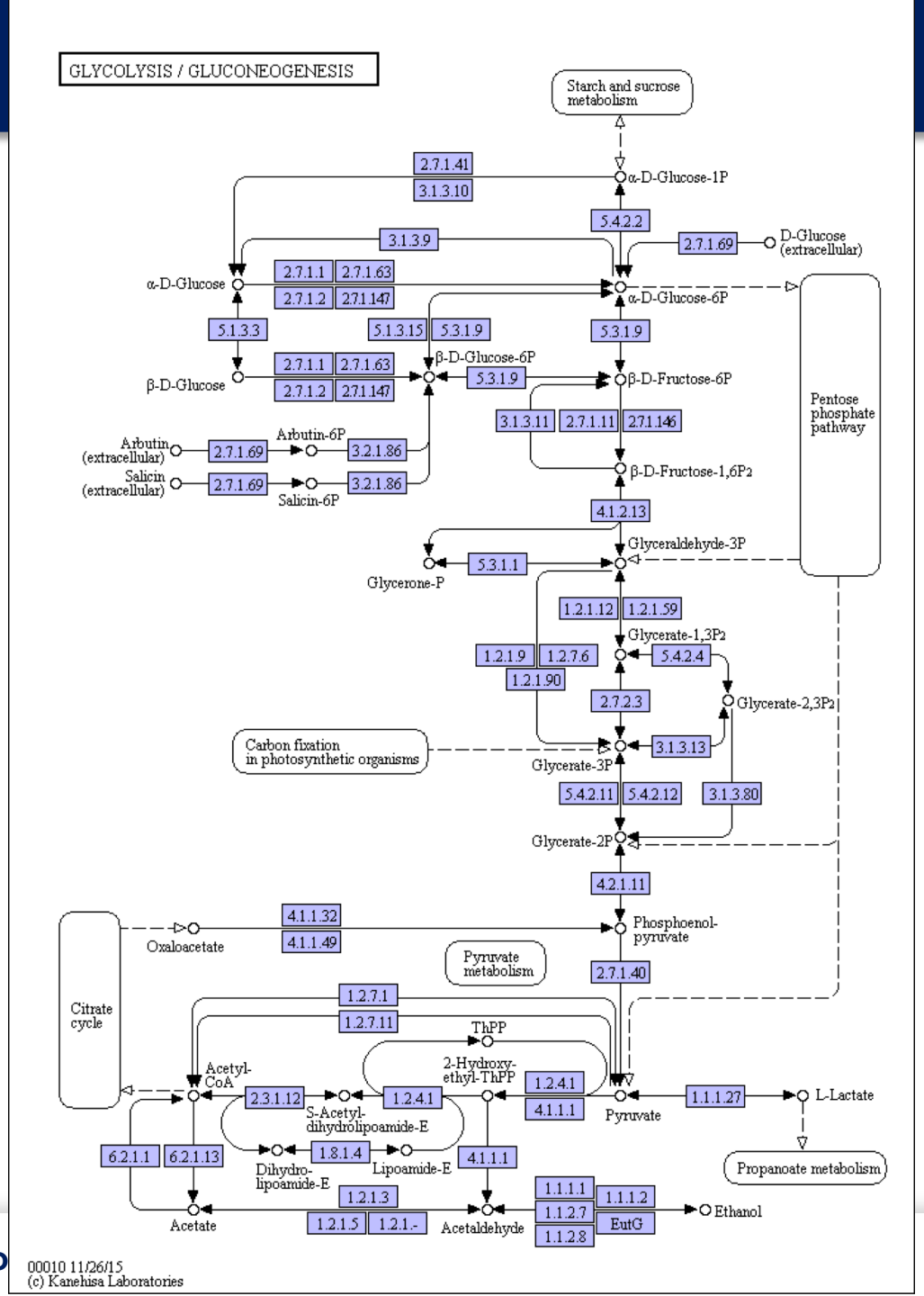
- Gene Ontology (GO) terms enrichment:
  - topGO (https://bioconductor.org/packages/release/bioc/html/topGO.html)
  - goSTAG (https://bioconductor.org/packages/release/bioc/html/goSTAG.html)
  - DAVID (https://david.ncifcrf.gov/)

Part II: pathway analysis

- Pathways focus on physical and functional interactions between genes rather than taking the gene-centered view of GO-based analyses
- Most pathway analysis tools rely on precompiled databases of pathways derived from large-scale literature analysis
- KEGG database is the most popular one

Part II: pathway analysis-KEGG

- KEGG (http://www.genome.jp/kegg/)
  - Kyoto Encyclopedia of Genes and Genomes
  - KEGG is a knowledge base for the systematic analysis of gene functions, in terms of the networks of genes and molecules
  - The major component of KEGG is the Pathway database, which consists of graphical diagrams of biochemical pathways including most of the known metabolic pathways and some of the known regulatory pathways



KEGG pathway example: Glycolysis pathway

- Glycolysis is the process of converting glucose into pyruvate and generating small amounts of ATP (energy) and NADH (reducing power).

Tools for KEGG pathway annotation and enrichment

- KOBAS (**K**EGG **O**rthology **B**ased **A**nnotation **S**ystem)
  - http://kobas.cbi.pku.edu.cn/
- Database for Annotation, Visualization, and Discovery (DAVID)
  - http://david.abcc.ncifcrf.gov/home.jsp
- KEGG Mapper – Search&Color Pathway
  - http://www.genome.jp/kegg/tool/map_pathway2.html
- Pathview

Pathway analysis:

- Pathway analysis:
  - GAGE (http://bioconductor.org/packages/release/bioc/html/gage.html)
  - Reactome (http://www.reactome.org/)
- Sample walkthrough:
  - From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline
    - https://www.bioconductor.org/help/workflows/RnaSeqGeneEdgeRQL/

Genome Annotation

- Which sequences code for proteins and structural RNAs?
- What is the function of the predicted gene products?
- Can we link genotype to phenotype? (i.e. What genes are turned on when ? Why do two strains of the same pathogen vary in their pathogenicity?)
- Can we trace the evolutionary history of an organism from its genomic sequence and genome organization? Evolutionary history of a pathway?

Genome Annotation

Genome annotation is a key process for identifying the coding and non-coding regions of a genome, gene locations and functions. Analysis of DNA sequence with genome annotation software tools allow finding and mapping genes, exons-introns, regulatory elements, repeats and mutations. Genome databases are essential to retrieve information on gene name, protein product and DNA sequence functions.

After the *de-novo* genome assembly

- NCBI RefSeq project
  - The NCBI Reference Sequence (RefSeq) project provides sequence records and related information for numerous organisms, and provides a baseline for medical, functional, and comparative studies.
  - The International Nucleotide Sequence Database Collaboration (INSDC, made up of GenBank, the European Nucleotide Archive, and the DNA Data Bank of Japan) represents an archival repository of all sequences
  - The RefSeq database is a non-redundant set of reference standards derived from the INSDC databases that includes chromosomes, complete genomic molecules (organelle genomes, viruses, plasmids), intermediate assembled genomic contigs, curated genomic regions, mRNAs, RNAs, and proteins.

RefSeq – Entrez Genomes

- This process flow provides genomic, RNA, and protein RefSeq records derived from assembled and annotated whole genome sequence data submitted to the INSDC.
- This pipeline provides all of the bacterial, viral, organelle, and plasmid RefSeq records and records for some eukaryotic genomes, including plants and fungi, as data becomes publicly available.
- Protein and transcript records are instantiated from the submitted genome sequence annotation or are predicted by NCBI's bacterial or eukaryotic computational annotation process.

RefSeq – Eukaryotic Genome Annotation Pipeline

- This process flow is an automated computational method that provides a copy of the submitted genome assembly in order to provide an annotated genome.
- RefSeq records may include chromosomes, intermediate assembled scaffolds and contigs, and transcripts and proteins.
- Depending on the species, genome annotation may reflect a mixture of transcript-based RefSeq records and computationally predicted transcripts and proteins with varying levels of support from transcript or protein alignments

The NCBI Eukaryotic Genome Annotation Pipeline

RefSeq - Curation-supported RefSeq pipeline

- NCBI staff scientists provide curation support in several ways. Staff leverage the Protein Clusters database to apply consistent nomenclature to orthologous proteins, work with collaborating groups to better represent data ranging from whole genomes to paralogous genes, and react to feedback from users reporting sequence or name improvements.
- NCBI curation staff also work closely with developer staff to provide genomic region, RNA, and protein RefSeq records for a subset of species grouped under the Bilateria node. Transcript and protein records are primarily derived from cDNA records submitted to the INSDC. This process flow is supported by a combination of bioinformatics and a significant level of manual curation.

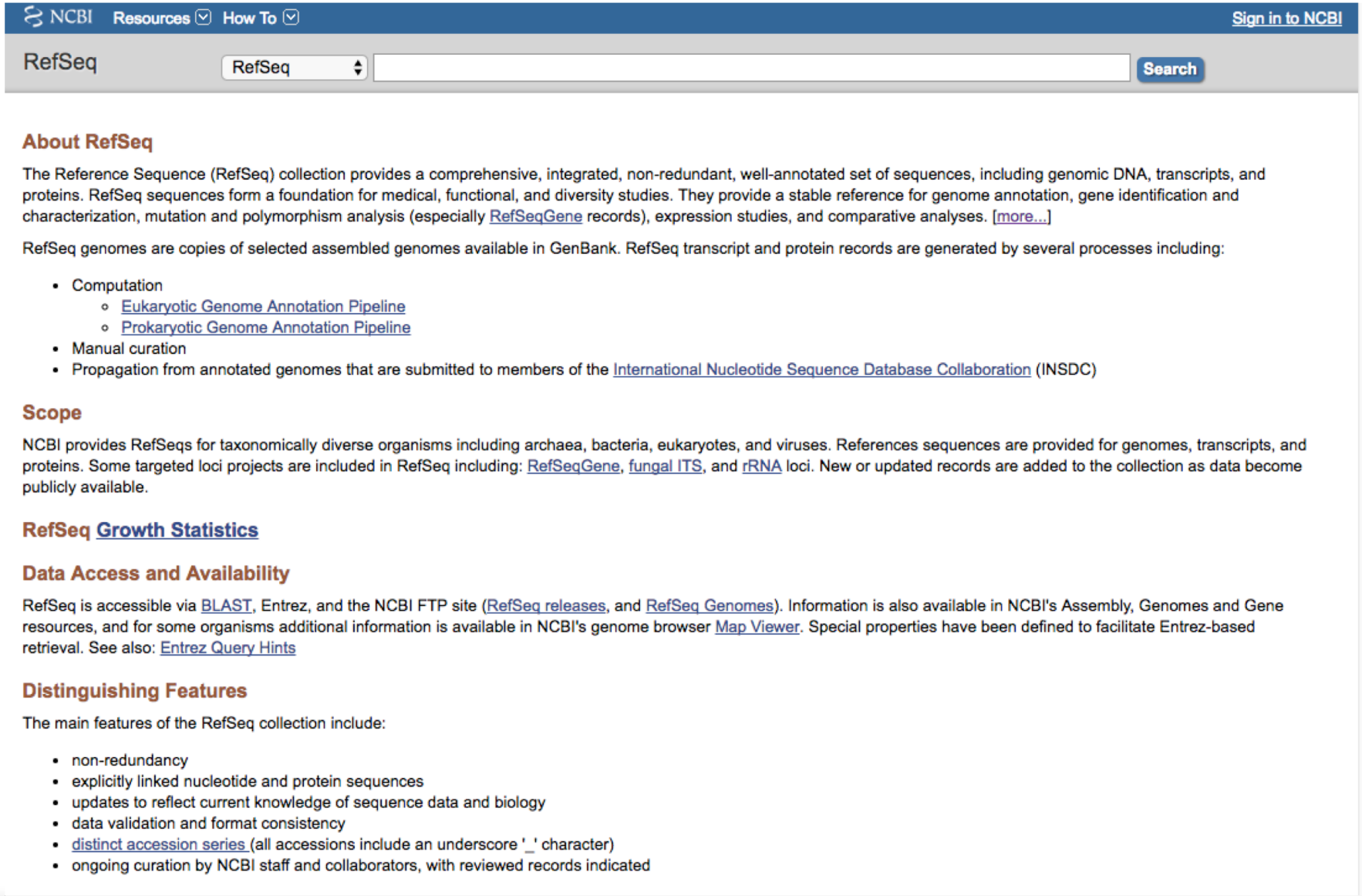
RefSeq - Collaboration

- Some RefSeq records are provided by collaborating groups. Different collaborations provide some fully annotated genomes or records for gene families or individual genes. Collaborations with official nomenclature groups, model organism databases, or other database groups also provide descriptive information, including gene symbols, names, publications, mapping data, feature annotation, database cross-references, and more.

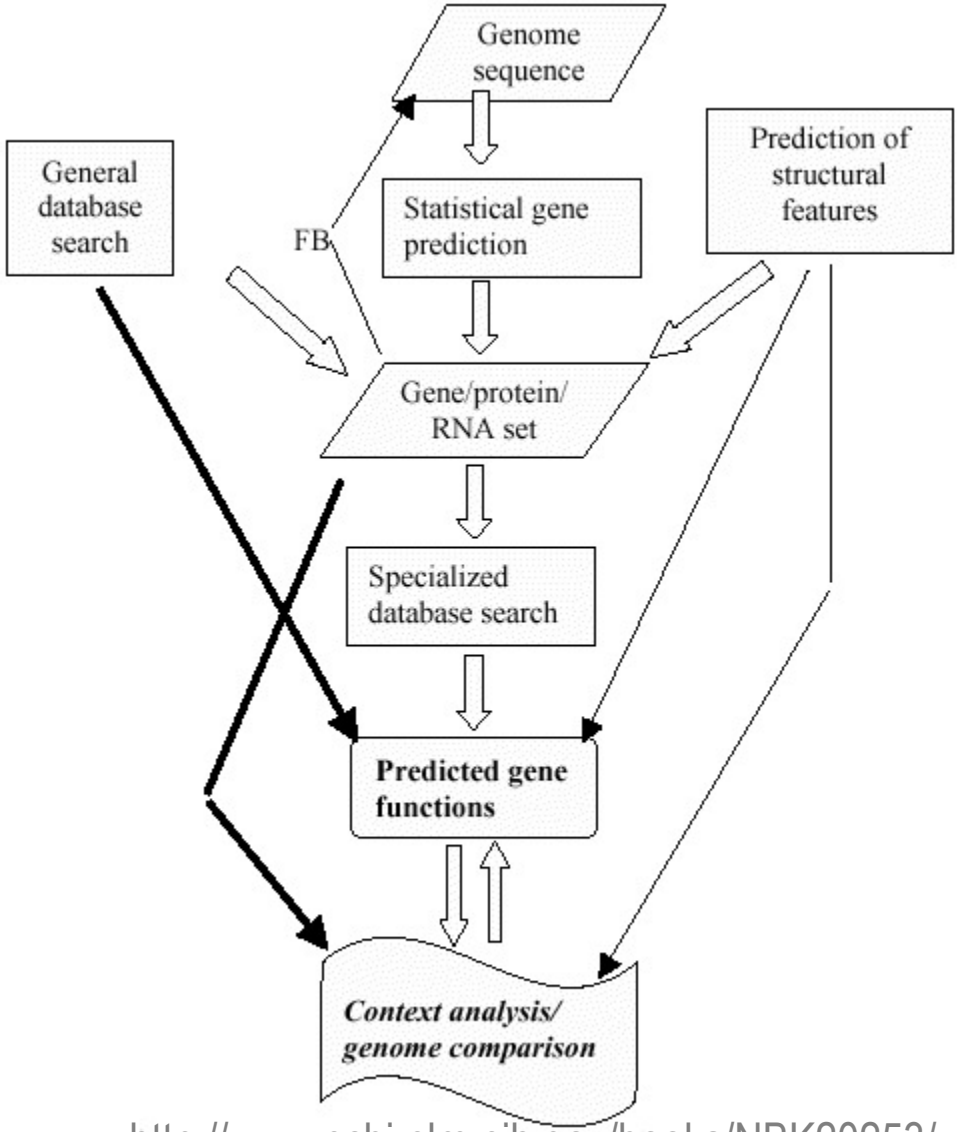
RefSeq Status Codes

Code	Description
MODEL	The RefSeq record is provided by the NCBI Genome Annotation pipeline and is not subject to individual review or revision between annotation runs.
INFERRED	The RefSeq record has been predicted by genome sequence analysis, but it is not yet supported by experimental evidence. The record may be partially supported by homology data.
PREDICTED	The RefSeq record has not yet been subject to individual review, and some aspect of the RefSeq record is predicted.
PROVISIONAL	The RefSeq record has not yet been subject to individual review. The initial sequence-to-gene association has been established by outside collaborators or NCBI staff.
REVIEWED	The RefSeq record has been reviewed by NCBI staff or by a collaborator. The NCBI review process includes assessing available sequence data and the literature. Some RefSeq records may incorporate expanded sequence and annotation information.
VALIDATED	The RefSeq record has undergone an initial review to provide the preferred sequence standard. The record has not yet been subject to final review at which time additional functional information may be provided.
WGS	The RefSeq record is provided to represent a collection of whole genome shotgun sequences. These records are not subject to individual review or revisions between genome updates.

Check the RefSeq



A generalized flow chart of genome annotation

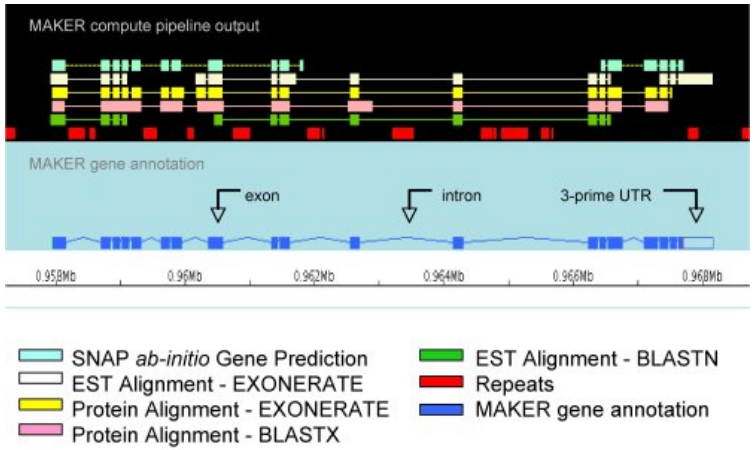


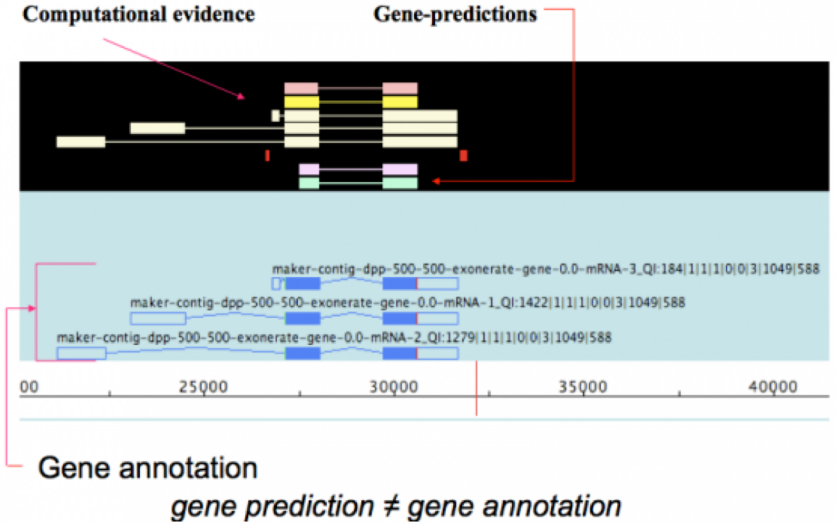
http://www.ncbi.nlm.nih.gov/books/NBK20253/

- Statistical gene prediction: use of methods like GeneMark or Glimmer to predict protein-coding genes.
- FB: feedback from gene identification for correction of sequencing errors, primarily frameshifts.
- General database search: searching sequence databases (typically, NCBI NR) for sequence similarity, usually using BLAST.
- Specialized database search: searching domain databases, such as Pfam, SMART, and CDD, for conserved domains, genome-oriented databases, such as COGs, for identification of orthologous relationship and refined functional prediction, metabolic databases, such as KEGG for metabolic pathway reconstruction, and possibly, other database searches.
- Prediction of structural features: prediction of signal peptide, transmembrane segments, coiled domain and other features in putative protein functions.

Genome Annotation Pipeline

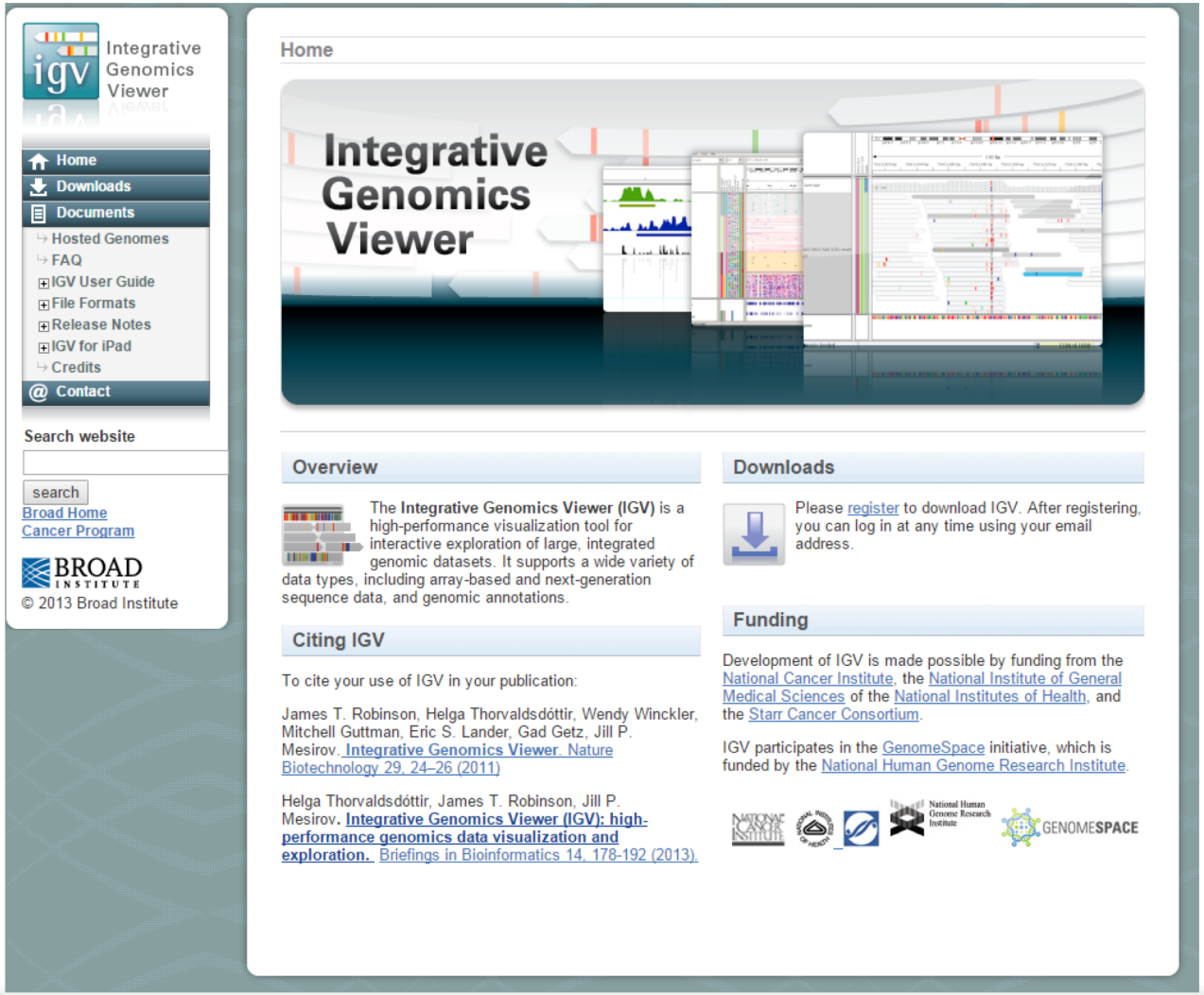
- MAKER 2 (http://www.yandell-lab.org/software/maker.html)
  - Identifies and masks out repeat elements
  - Aligns ESTs to the genome
  - Aligns proteins to the genome
  - Produces ab initio gene predictions
  - Synthesizes these data into final annotations
  - Produces evidence-based quality values for downstream annotation management





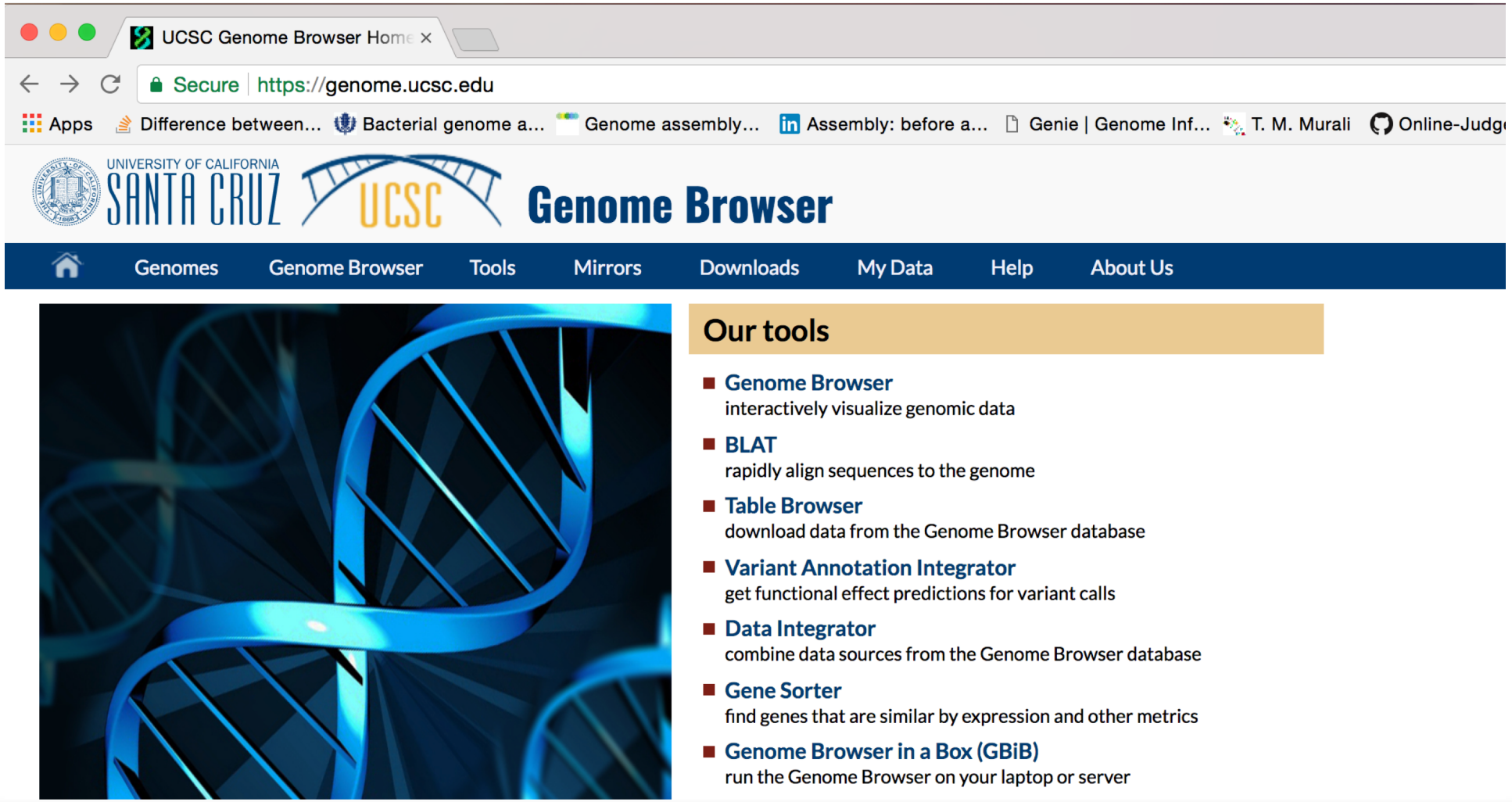
Genome Browser: IGV

- https://www.broadinstitute.org/igv/



Web: UCSC Genome Browser

- https://genome.ucsc.edu/



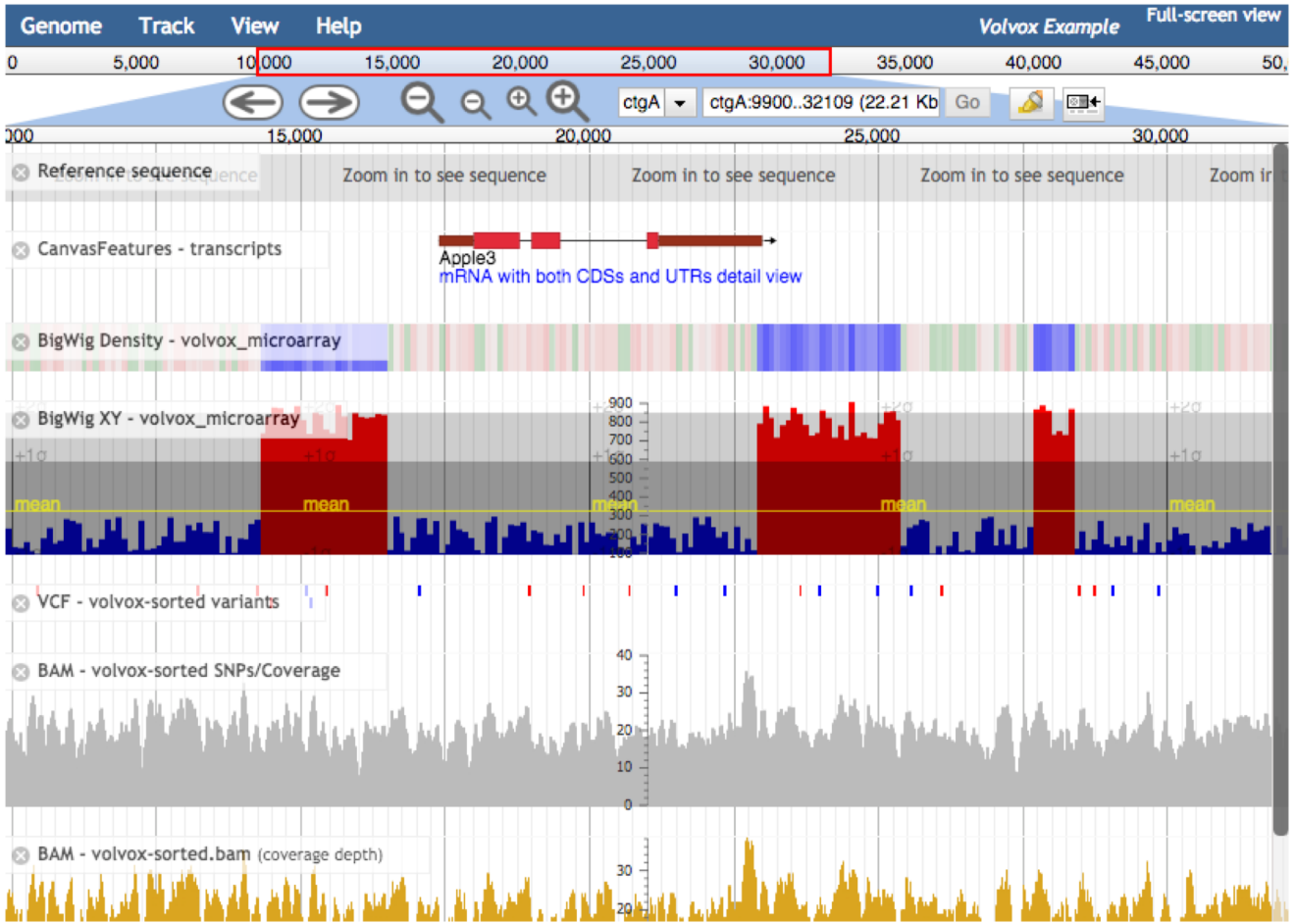
Web: JBrose

- https://jbrowse.org/

The JBrowse Genome Browser

JBrowse is a fast, scalable genome browser built completely with JavaScript and HTML5. It can run on your desktop, or be embedded in your website.

Download latest release – JBrowse 1.16.2



Example:

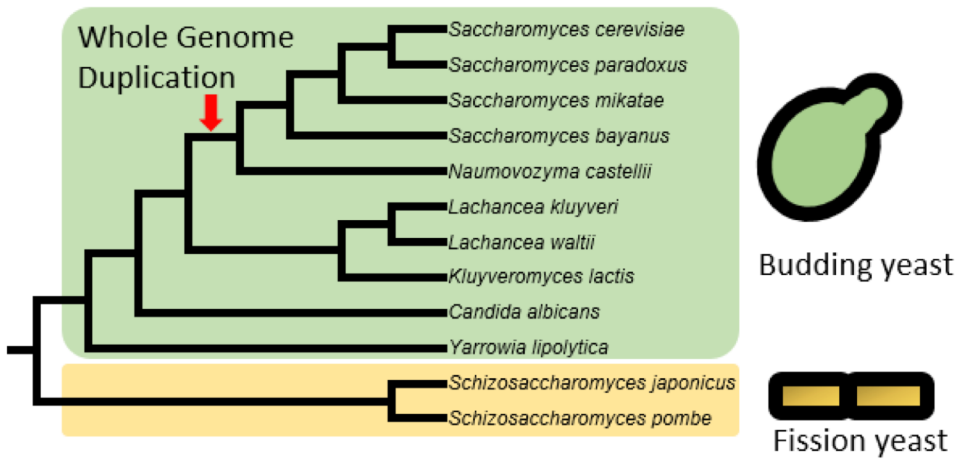
- http://www.yeastss.org/

YeasTSS

SAINT LOUIS UNIVERSITY.

Home | Genome Browser | Download | Search | Help | Contact

Welcome to the YeasTSS Atlas



Click species name to visualize its TSS maps in a new window.

The YeasTSS Atlas (Yeast Transcription Start Site Atlas) is a primary depository of yeast transcription initiation sites (TSS) data generated by Dr. Zhenguo Lin's lab at Saint Louis University. These data are valuable for precisely determining the 5'boundary and the 5'untranslated region (5'UTR) of protein coding genes, improving genome annotation quality, and predictions of novel genes, core promoter elements, transcription factor binding sites, and other motifs associated with transcription and inferring gene regulatory network.