Functional Analysis

BCB 5200 Introduction Bioinformatics I

Fall 2017

Tae-Hyuk (Ted) Ahn

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

SAINT LOUIS UNIVERSITY™

EST. 1818

RNA-Seq HW and Final Exam Announcement

- RNA-Seq HW will be posted until 9am Wed, Dec 6th (Due: 4pm Monday, Dec 11th)
- Take-Home Exam
- Two Parts:
  - Part 1: Multiple questions (70%)
  - Part 2: Coding problem (Python) (30%)
- I will post the exam at 12:45pm Thursday, Dec 7th on Blackboard
- Part 1 due is 4:00pm Fri, Dec 8th (Turn-in by BB or my office)
- Part 2 due is 4:00pm Mon, Dec 11th

Download NGS data using SRA Toolkit

- Data are saved as SRA format at NCBI SRA database
- NCBI SRA Toolkit is used to download SRA files from the NCBI SRA site and convert them into FASTQ files
  - https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/
  - Available for: Linux, Mac, Windows
  - for CentOS Linux 64 bit architecture

```
$wget https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-centos_linux64.tar.gz
$ tar -zxvf sratoolkit.current-centos_linux64.tar.gz
# add SRA toolkit path to PATH
$ source ~/.bashrc
```

Download and Converting SRA format data into FASTQ

Usage:

```
$ fastq-dump [options] <path/file> [<path/file> ...]
$ fastq-dump [options] <accession>
```

For example, download a paired-end SRA data SRR2567784 :

```
$ fastq-dump --split-files SRR2567784
```

After completion, it generates two fastq files

```
SRR2567784_1.fastq
SRR2567784_2.fastq
```

If you want to download many SRA files

If run accession # are continuous integers, i.e. SRR1265943- SRR1265952

```
$ COUNTER=1265943
$ while [  $COUNTER -lt 1265953]; do
> echo "fastq-dump --split-files SRR$COUNTER" >> sra_download_command.list
> let COUNTER=COUNTER+1
> done
```

It generates a file sra_download_command.list

```
fastq-dump --split-files SRR1265943
fastq-dump --split-files SRR1265944
fastq-dump --split-files SRR1265945
fastq-dump --split-files SRR1265946
fastq-dump --split-files SRR1265947
fastq-dump --split-files SRR1265948
fastq-dump --split-files SRR1265949
fastq-dump --split-files SRR1265950
fastq-dump --split-files SRR1265951
fastq-dump --split-files SRR1265952
```

```
$ sh sra_download_command.list
```

If you want to download many SRA files

If run accession # are NOT continuous integers, i.e. SRR1264528, SRR1265109, SRR1265122, SRR1265901, SRR1265914

```
$ echo "fastq-dump --split-files SRR1264528" >> sra_download_command.list
$ echo "fastq-dump --split-files SRR1265109" >> sra_download_command.list
$ echo "fastq-dump --split-files SRR1265122" >> sra_download_command.list
$ echo "fastq-dump --split-files SRR1265122" >> sra_download_command.list
$ echo "fastq-dump --split-files SRR1265914" >> sra_download_command.list
```

It generates a sra_download_command.list file

```
fastq-dump --split-files SRR1264528
fastq-dump --split-files SRR1265109
fastq-dump --split-files SRR1265122
fastq-dump --split-files SRR1265122
fastq-dump --split-files SRR1265914
```

```
$ sh sra_download_command.list
```

Microarray and RNA-Seq Pipeline

Design and perform experiment

↓

Process and normalise data

↓

Statistical analysis

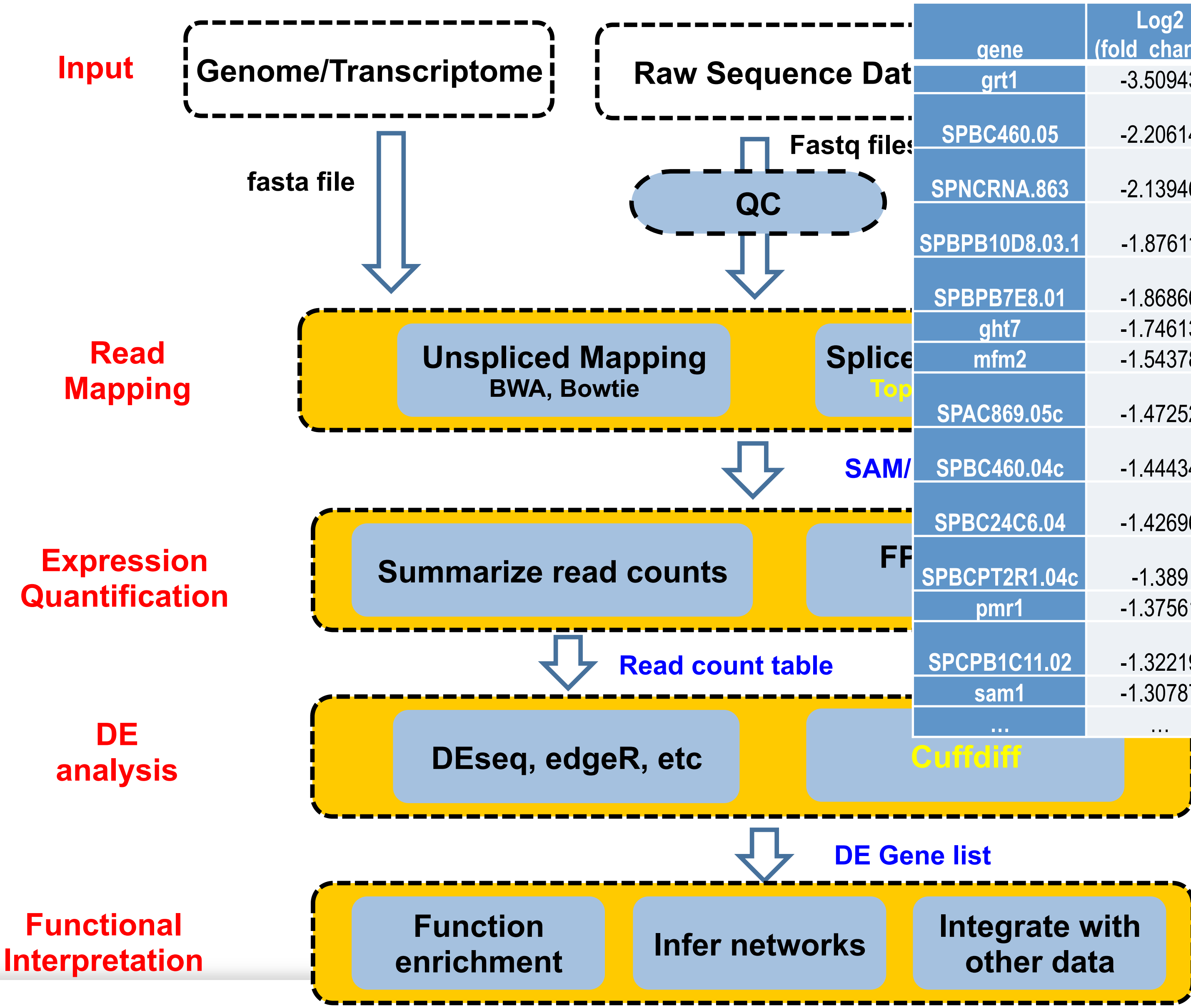
↓

Differentially expressed genes

↓

Biological interpretation

RNA-Seq



gene	Log2 (fold_change)	p_value	q_value
grt1	-3.50943	5.00E-05	0.001384
SPBC460.05	-2.20614	5.00E-05	0.001384
SPNCRNA.863	-2.13946	5.00E-05	0.001384
SPBPB10D8.03.1	-1.87611	0.00065	0.011576
SPBPB7E8.01	-1.86866	5.00E-05	0.001384
ght7	-1.74613	0.00255	0.033104
mfm2	-1.54378	5.00E-05	0.001384
SPAC869.05c	-1.47252	5.00E-05	0.001384
SPBC460.04c	-1.44434	0.0003	0.006203
SPBC24C6.04	-1.42696	5.00E-05	0.001384
SPBCPT2R1.04c	-1.389	0.00085	0.014346
pmr1	-1.37561	5.00E-05	0.001384
SPCPB1C11.02	-1.32219	5.00E-05	0.001384
sam1	-1.30787	5.00E-05	0.001384
…	…	…	…

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 (GO)

- http://geneontology.org/
- The Gene Ontology (GO) provides consistent descriptions of gene products across databases
  - Describe the genes or gene products
    - Location, function, process
  - Genes have relationships to others
  - Gene product has multiple features

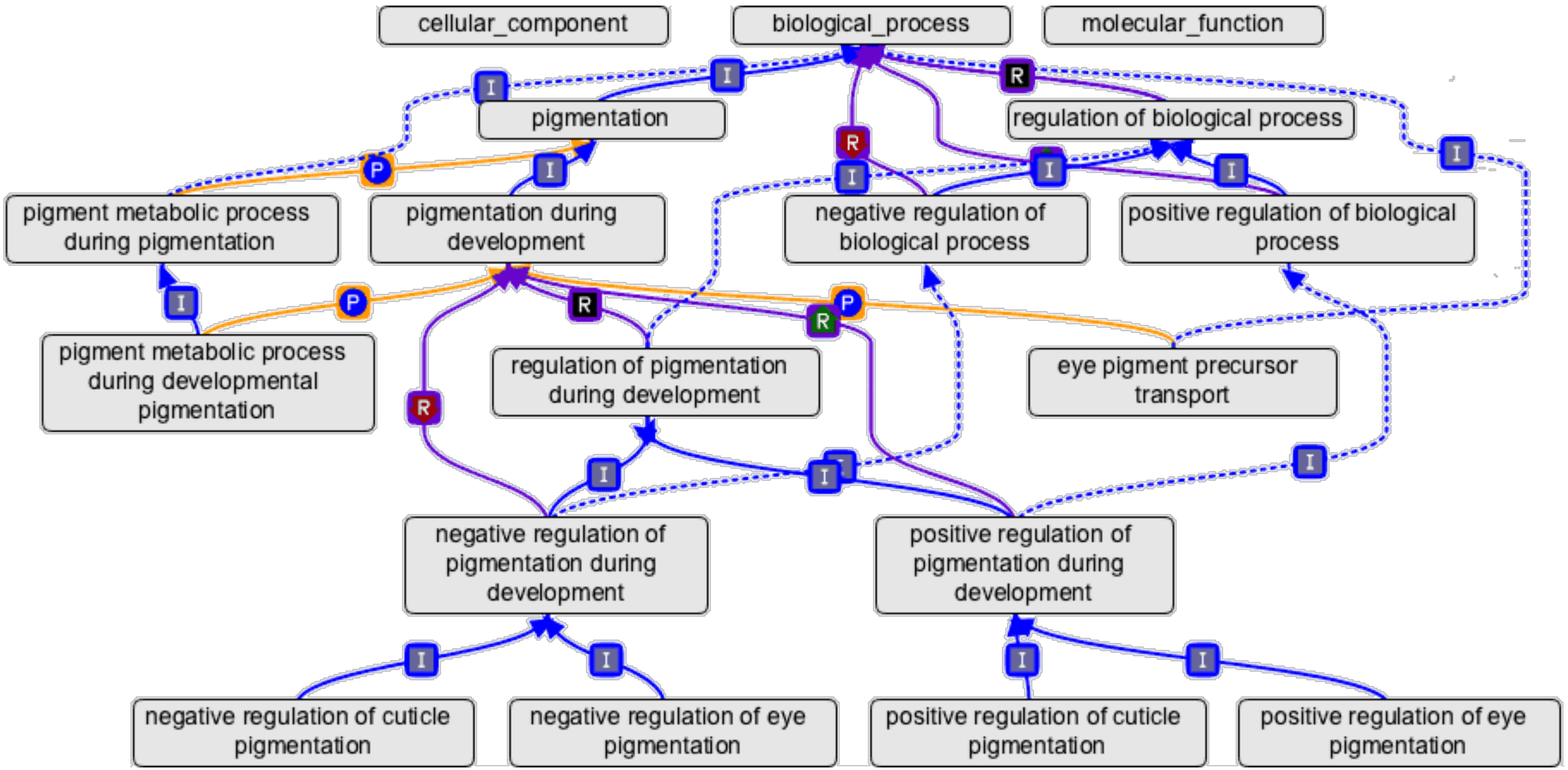
What is the Gene Ontology?

- 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

Three domains of ontology

- Molecular function: activities of a gene product at the molecular level
  - e.g. catalytic activity, calcium ion binding
- Biological process: broad objective or goal
  - e.g. signal transduction, immune response
- Cellular component: location or complex
  - e.g. nucleus, mitochondrion

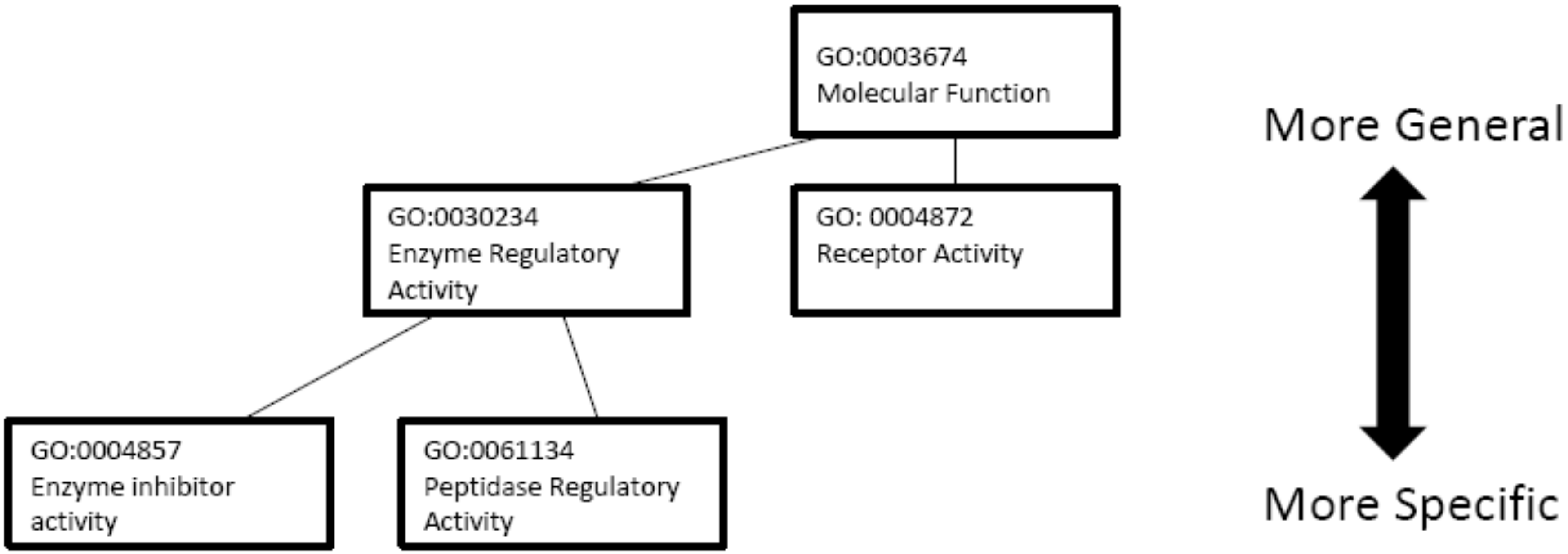
Hierarchy structure of ontology



http://geneontology.org/page/ontology-structure

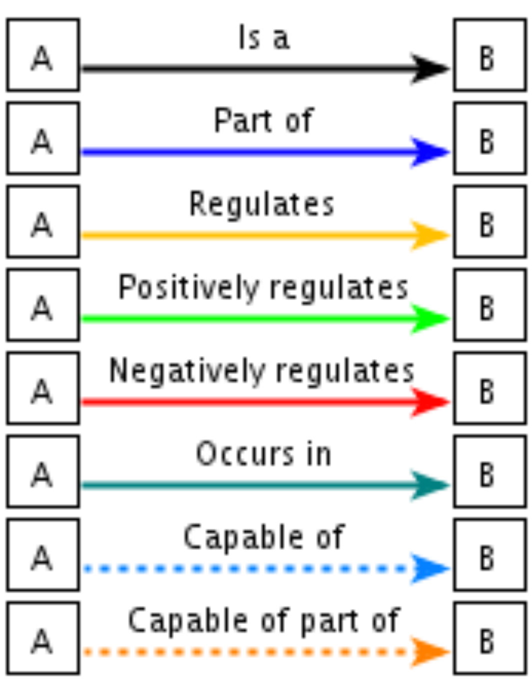
Gene Ontology

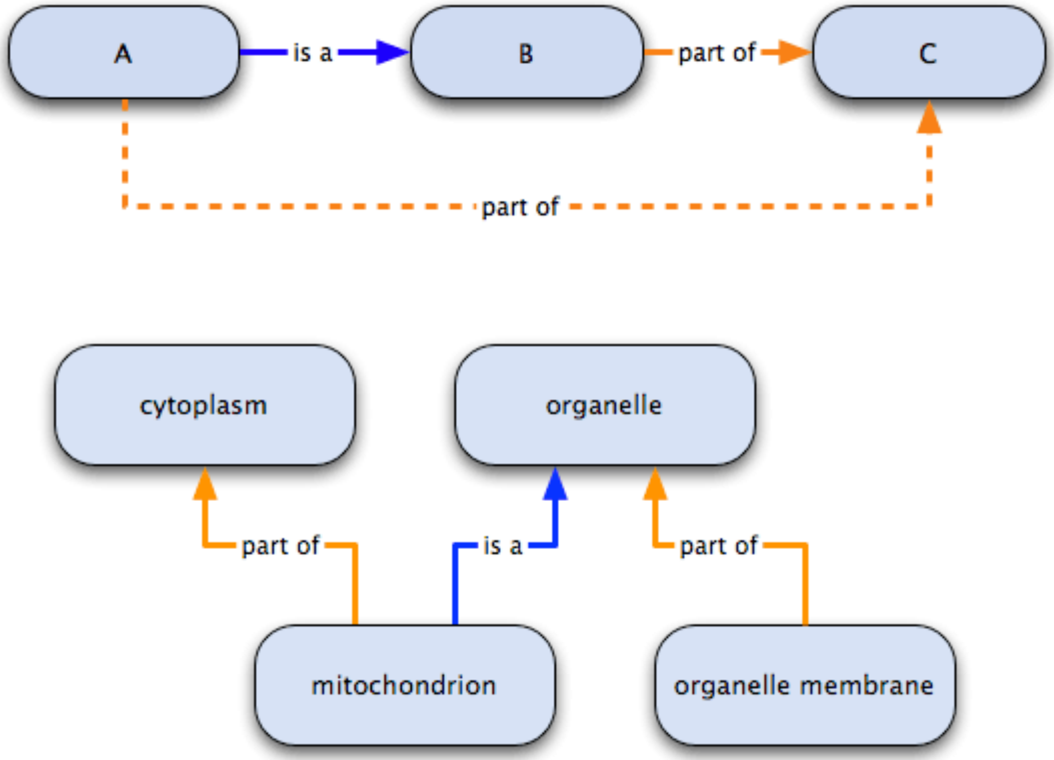
- Hierarchical relationship



Gene Ontology

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





Example: GO:0030330

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

Term Information

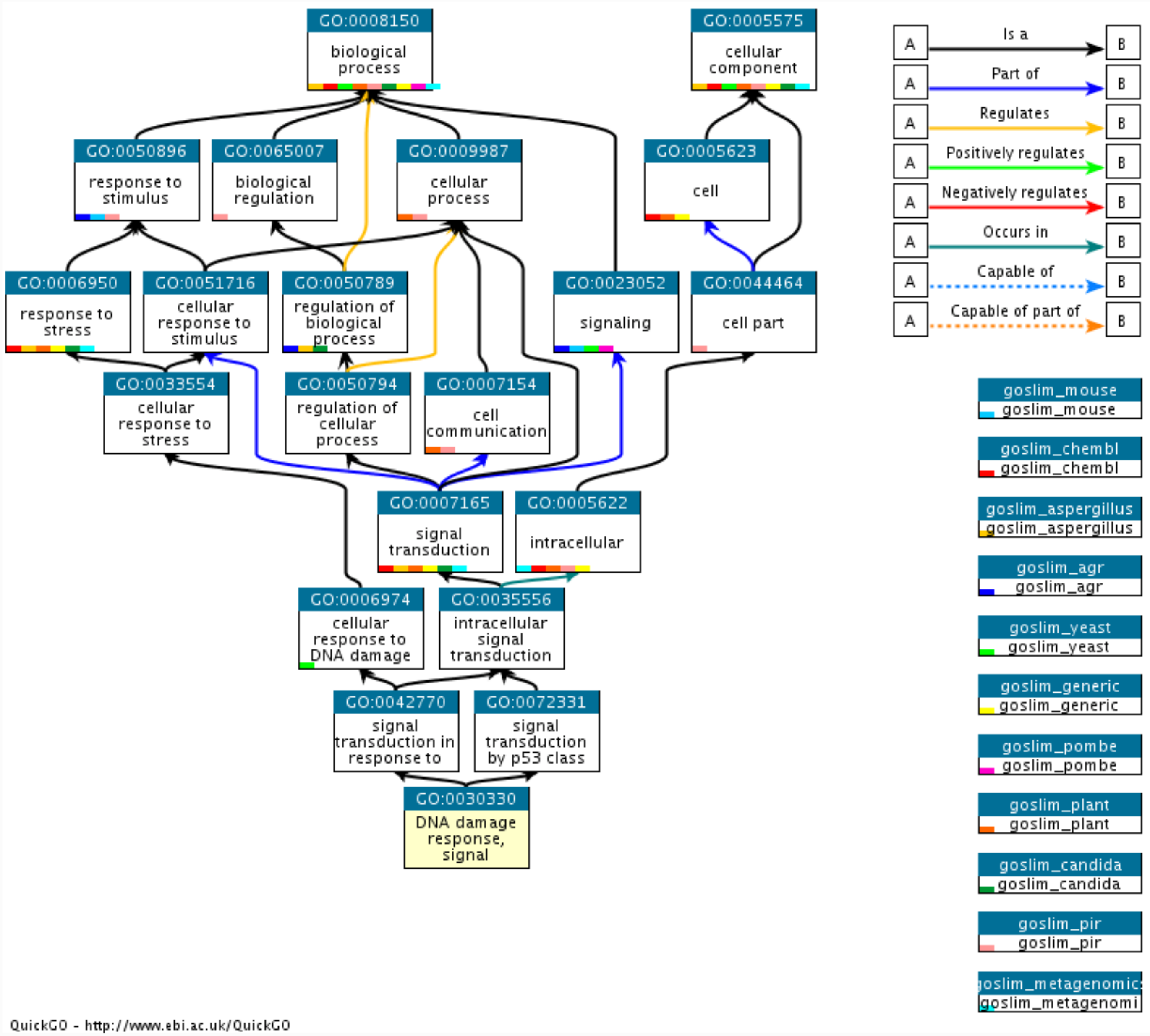
Data health

Accession	GO:0030330
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.

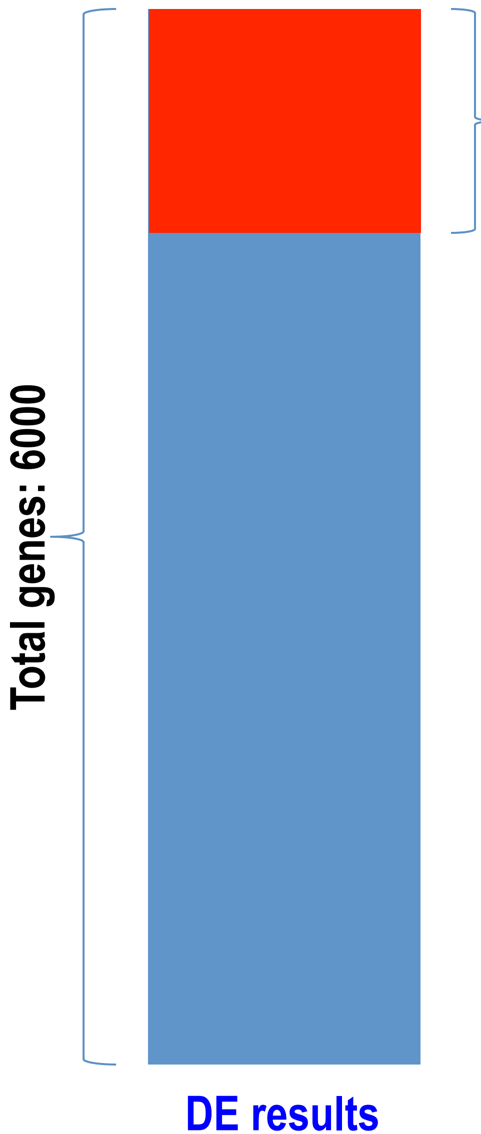
QuickGO

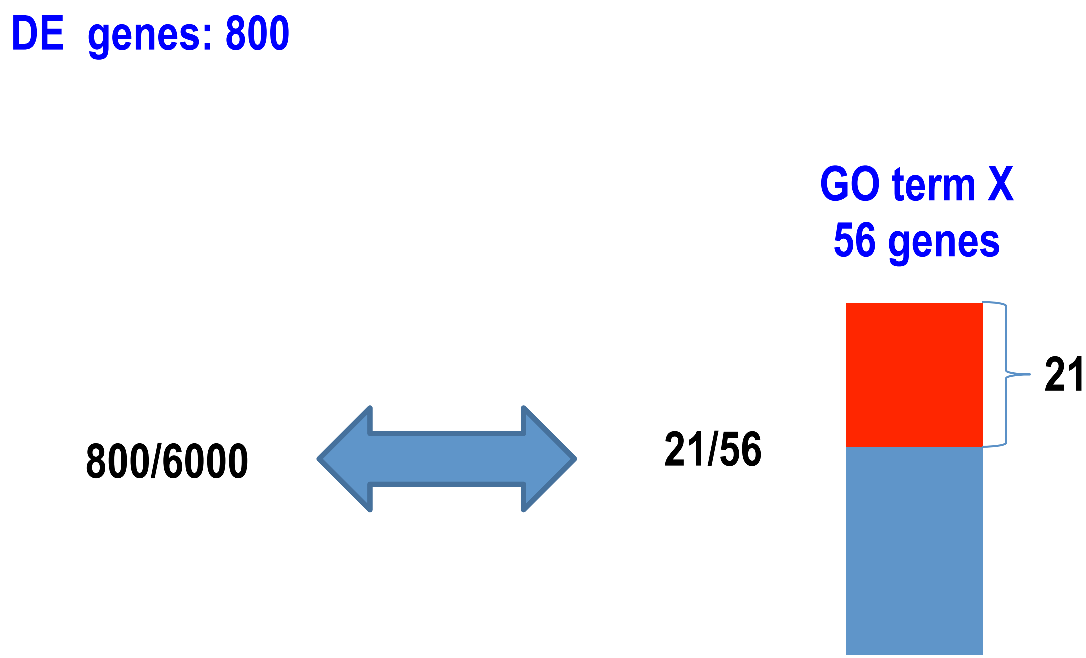
Ancestor Chart

Ancestor chart for GO:0030330



GO enrichment analysis



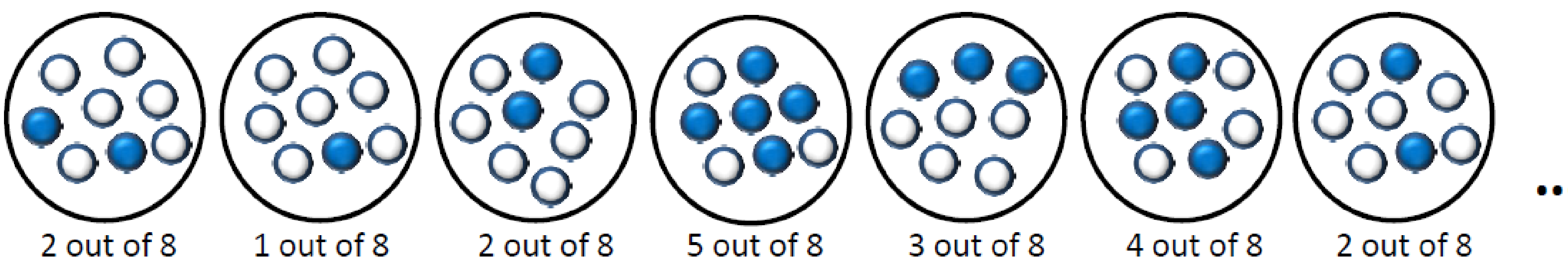


Are genes of GO term X over-represented in the DE list (occurs more times than expected by chance)?

A quick review: Modified Fisher's exact test

Do I have a surprisingly high number of blue genes?

Null model: the 8 genes/balls are selected randomly



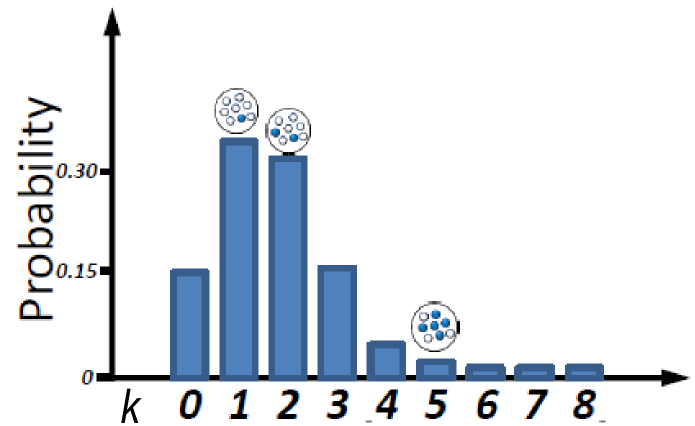
So, if you have 50 balls, 10 of them are blue, and you pick 8 balls randomly, what is the probability that k of them are blue?

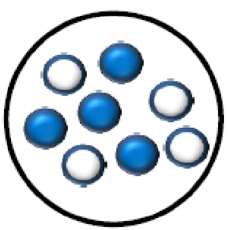
A quick review: Modified Fisher's exact test

Hypergeometric distribution

$$\mathbb{P}(\sigma_t = k) = \frac{\binom{m_t}{k}\binom{m-m_t}{n-k}}{\binom{m}{n}}$$

$m=50$, $m_t=10$, $n=8$





So … do I have a surprisingly high number of blue genes?

What is the probability of getting at least 4 blue genes in the null model?

$P(\sigma t >=4)$

GO enrichment analysis tools

- Gene Ontology Consortium
  - http://www.geneontology.org/page/go-enrichment-analysis
- Database for Annotation, Visualization, and Discovery (DAVID)
  - http://david.abcc.ncifcrf.gov/home.jsp
- GENERIC GENE ONTOLOGY (GO) TERM FINDER
  - http://go.princeton.edu/cgi-bin/GOTermFinder
- Gene Set Enrichment Analysis (GSEA)
- GOrilla(Gene Ontology enRIchment anaLysis and visuaLizAtion tool)
- WebGestalt (WEB-based GEne SeT AnaLysis Toolkit)
- Blast2GO
  - https://www.blast2go.com

[4] RNA-Sequencing Analysis for the yellow fever mosquito *Aedes aegypti*

Research Objectives

Illumina Analyzer

```
>1
CTCGGTGATCGATTGGAAAAAAAAATGGTTCGCACAGATC
>1
CGCCAAGCAGTGTTCGTTGTTGGTGGGGCCTGCTGCTGA
>1
CTGTCAACTGCCATCAGCAGCGTGCTCACTGGGCGATCG
>1
TCGATTTCGGCGCTGCAGTGGCTGGGCAAAGCCGATAAA
>2
GCAGTTCTGGGAGAAACAGGAAGCCTCTCGCATCGAAGA
>1
AGCAACTCCAGTTGGAGCTGTCGCTTTCTCCGTTGCACT
>1
CCGAAAATCTACGACGTAATCTACGAACTGAACGTCATC
>1
ATGGGACCACGTTGGAGGCGGTGAATACTTTTAGTGGCC
>1
TGTCACCCTCGGAGAGCGGTCGTCTGCGTACTCCCCAGT
>1
GACAATGCCGAACTGGGCTTGATCTTCCGAGGTGGACCA
>1
TAGCACCGGATGCGCAGGGATTGTTCATGTTCTTCTGAC
>2
TATAGCCTCACTGTTAACGGTAAAGTGTACTTGGGGGAA
>1
ATCAATTCCTGGCCCTGCTCGATACCTTGCATGAGGAAG
```

12 *Aedes aegypti* transcriptome

	Description	BP
Sample 01	Embryo 0-2 Hours	33 bp
Sample 02	Embryo 2-4 Hours	33 bp
Sample 03	Embryo 4-8 Hours	33 bp
Sample 04	Embryo 8-12 Hours	33 bp
Sample 05	Larvae	39 bp
Sample 06	Pupae	41 bp
Sample 07	1-5D Male	38 bp
Sample 08	0-1D Ovary	39 bp
Sample 09	72H PBMOvary	40 bp
Sample 10	Embryo 0-1 Hour	83 bp
Sample 11	24H PBM Ovary	83 bp
Sample 12	24H PBM Carcass	83 bp

- Bioinformatics
- Wet Experiments

Gene Annotation Comparison

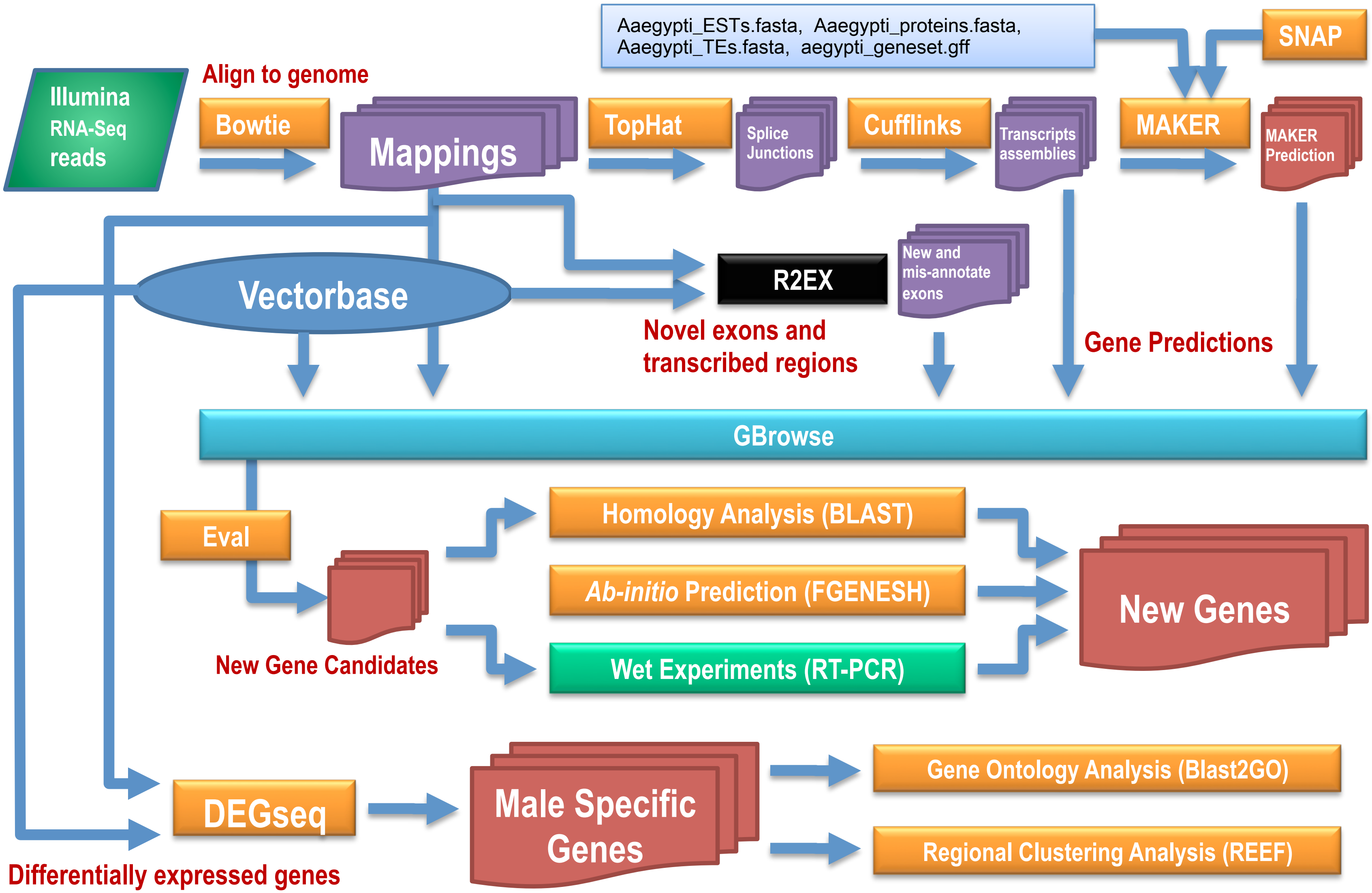
- Find New Transcripts
- Find Mis-Annotated Transcripts

Gene Expression Analysis

- Gene Expression Profile
- Male vs. Female Transcripts

[4] RNA-Sequencing Analysis for the yellow fever mosquito *Aedes aegypti*

Analysis Workflow



[4] RNA-Sequencing Analysis for the yellow fever mosquito *Aedes aegypti*

R2EX Program

R2EX

- **R2EX-finding**

  `R2EX-finding[options] bowtie-map gene-annotations`

  R2EX-finding is a tool to report possible new and misannotated exons by user defined criteria.

- **R2EX-express**

  `R2EX-express[options] bowtie-map gene-annotations`

  R2EX-express reports normalized expression abundance of exons.

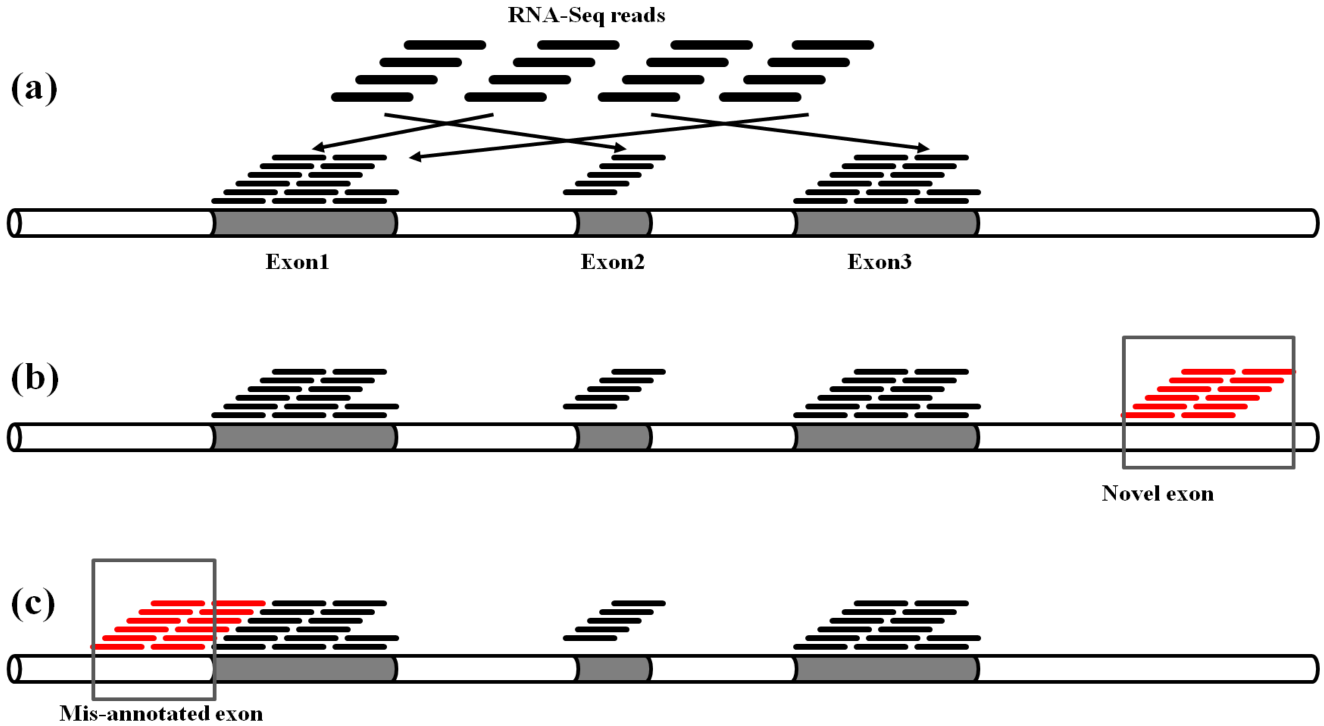
- **R2EX-splicing**

  `R2EX-splicing[options] R2EX-express annotation`

  R2EX-splicing is a tool to report mRNA and exon expression abundance comparison to find alternative splicing of input samples.

[4] RNA-Sequencing Analysis for the yellow fever mosquito *Aedes aegypti*

R2EX Program



(a) Reads are aligned to the genome using a mapping program such as Bowtie.

(b) R2EX detects new aligned regions where the mapping density of the region is higher than a threshold and satisfies several user input options such as minimum length.

(c) Previous mis-annotated regions also can be detected by R2EX program.

[4] RNA-Sequencing Analysis for the yellow fever mosquito *Aedes aegypti*

Results – Web browser

Web Link

Genome Browser for *Aedes Aegypti* RNA-Seq

Tae-Hyuk (Ted) Ahn and Zhijian (Jake) Tu
The Departments of Computer Science and Biochemistry
Virginia Tech

NGS RNA-Seq Analysis for *Aedes aegypti*

We analyze Next Generation Sequencing (NGS) RNA-Seq data of *Aedes aegypti* mosquito, the yellow fever mosquito that can spread degue fever, Chikungunya and yellow fever viruses, and other diseases. 12 samples of RNA-Seq data has been used for this browser. Vectorbase gene annotation set and MAKER gene predictions are included.

- *Aedes aegypti* whole genome RNA-Seq analysis

In this browser, we have tried to incorporate all software results for supercont1.1, the largest supercontig of *aedes aegypti* .

- *Aedes aegypti* supercont1.1 RNA-Seq analysis

Learning More about GBrowse

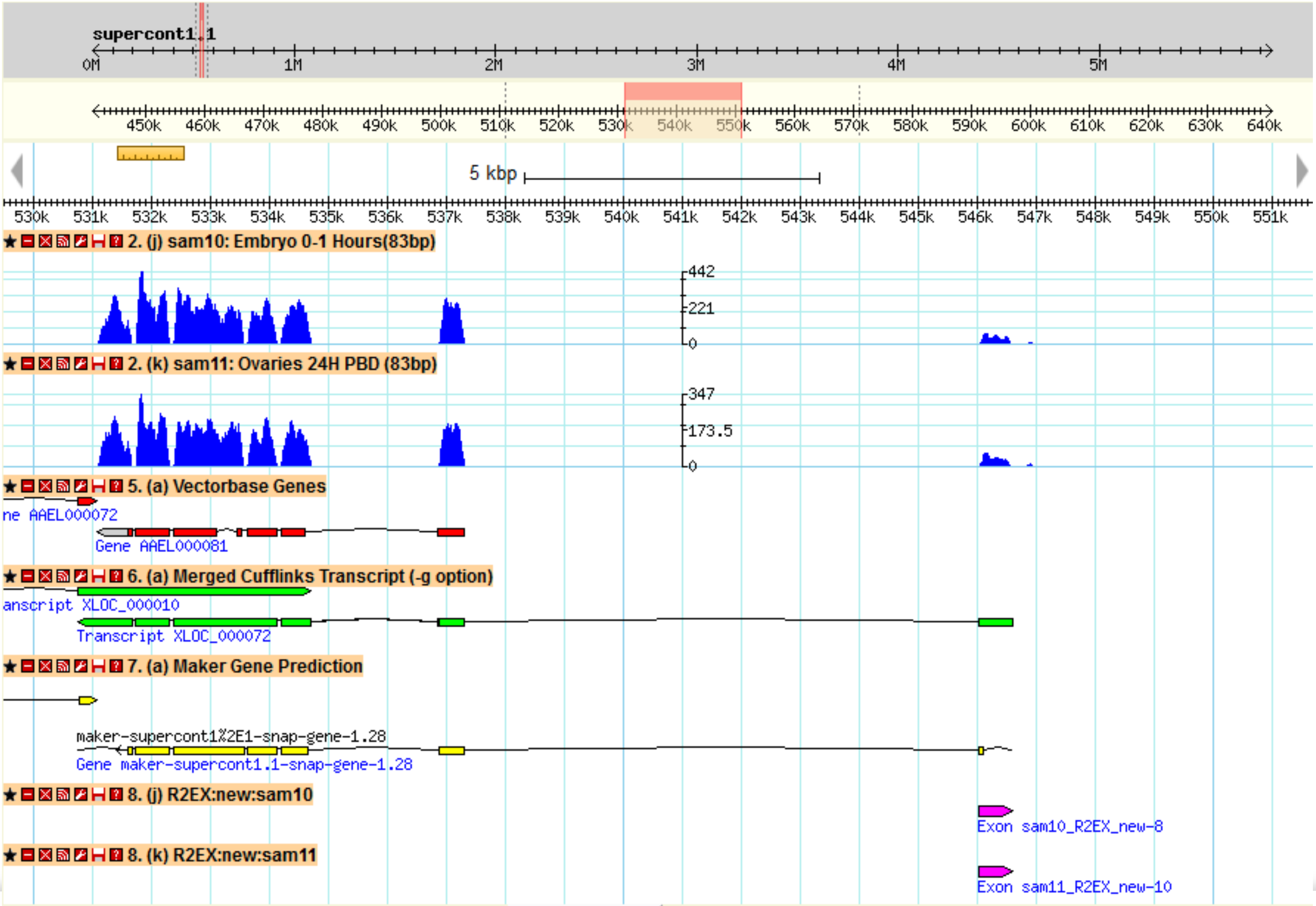
To find out more about Gbrowse, try the:

- GBrowse 2.0 HOWTO
- GBrowse Tutorial

Tae-Hyuk (Ted) Ahn, thahn@cs.vt.edu

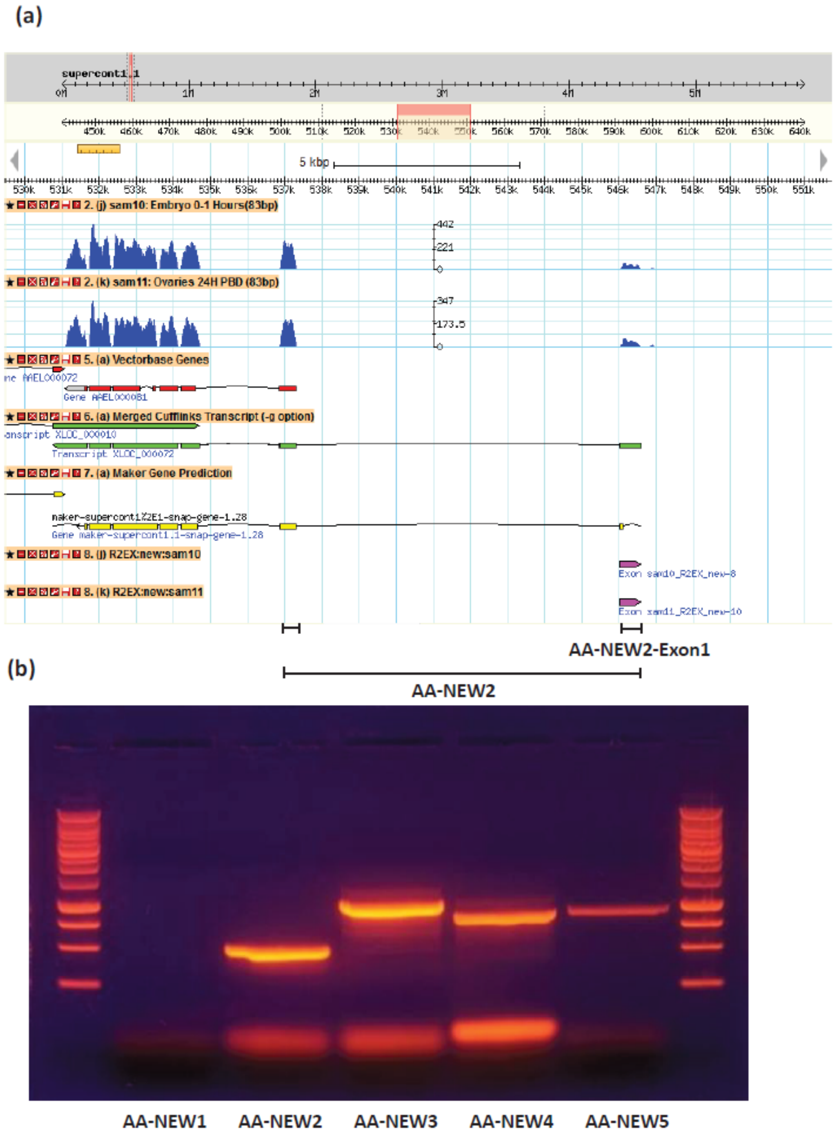
[4] RNA-Sequencing Analysis for the yellow fever mosquito *Aedes aegypti*

Results – Web Browser



[4] RNA-Sequencing Analysis for the yellow fever mosquito *Aedes aegypti*

Results: New gene transcripts



- We selected several strong new gene candidates.
- Homology-based analysis and ab inito gene validations confirm the gene predictions.
- Several possible genes were validated by RT-PCR (wet-lab experiments).

[4] RNA-Sequencing Analysis for the yellow fever mosquito *Aedes aegypti*

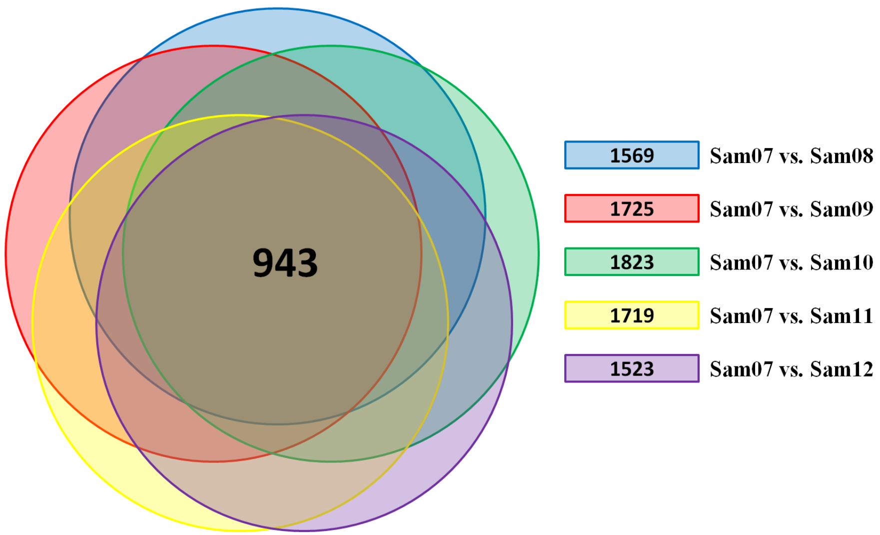
Results: How many possible new genes?

Sample	Cufflinks # of new transcripts (# of new exons)	Cufflinks # of new transcripts have one exon	Cufflinks # of new transcripts have two more exons	Cufflinks # of annotated transcripts required revision	R2EX # of new exons	R2EX # of mis-annotated exons
sam01	1,069 (1140)	1,003	66	4822	10,458	3,188
sam02	1,377 (1660)	1,149	228	5479	12,136	4,979
sam03	1,583 (1874)	1,320	263	6065	17,271	5,075
sam04	2,413 (2632	2,206	207	6124	22,689	4,244
sam05	2,549 (2975)	2,250	299	6037	21,467	6,403
sam06	4,800 (5481)	4,232	568	7502	32,879	7,841
sam07	5,817 (6663)	5,147	670	8587	31,010	8,388
sam08	2,101 (2545)	1,783	318	6956	10,902	5,353
sam09	1,422 (1876)	1,109	313	5606	6,912	5,162
sam10	2,109 (3790)	1,116	993	6257	5,235	6,391
sam11	3,279 (4882)	2,346	933	7046	6,126	7,903
sam12	3,174 (3995)	2,699	475	7239	14,826	7,712

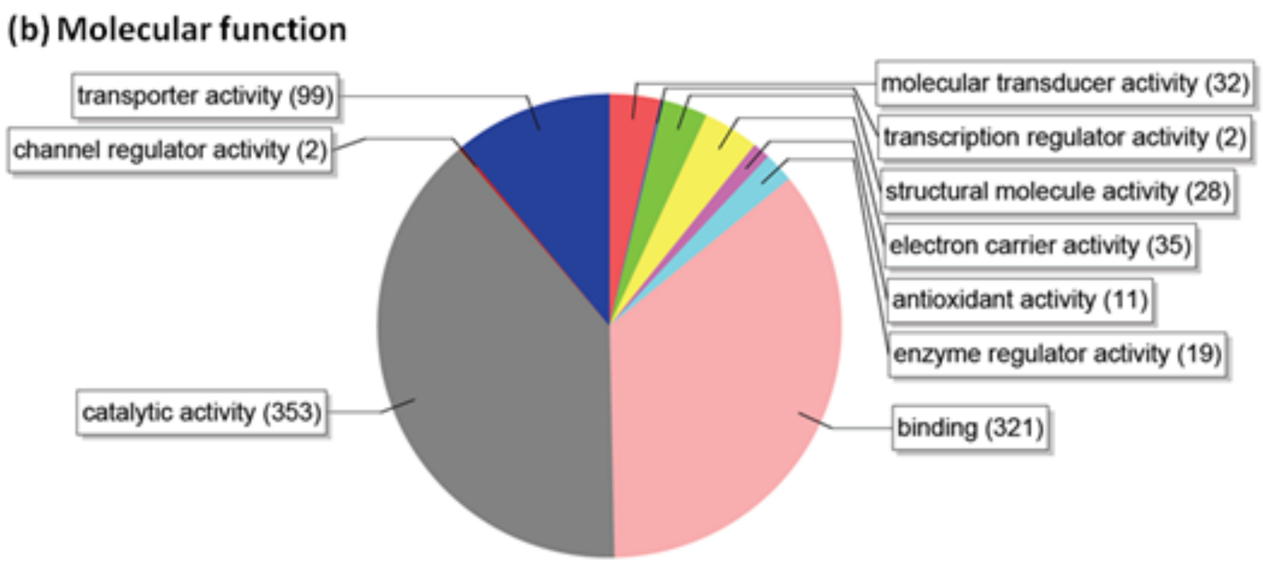
- Current annotation: 17,402 transcripts
- Cuffmerge transcripts prediction: 37,634 (12,466 possible new transcripts)
- MAKER gene prediction: 21,091 (110 possible new transcripts)

[4] RNA-Sequencing Analysis for the yellow fever mosquito *Aedes aegypti*

Results: Male-biased transcripts



- Another contribution of this study is to analyze differential transcript expression of RNA-Seq data especially on male-biased transcripts.
- Using gene expression profile program (DEGseq), we identified 943 male-biased gene transcripts.

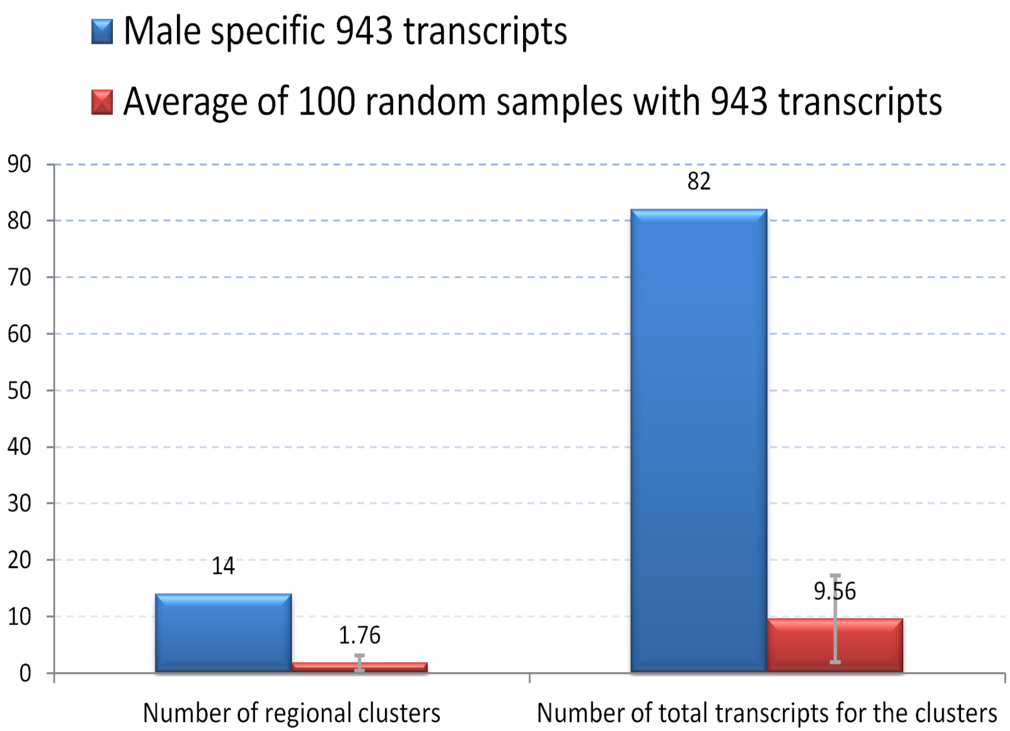


- Gene ontology (GO) analysis of the 943 male-biased transcripts was performed for functional annotation. The software Blast2GO.

[4] RNA-Sequencing Analysis for the yellow fever mosquito *Aedes aegypti*

Results – Male-biased genes are regionally clustered?

- As genes are non-randomly distributed in the genome, sex-biased genes are non-randomly distributed in many organisms.
- 82 of 943 male-biased transcripts are regionally clustered.
- Random sampling test indicates that a significant portion of the predicted male-biased transcripts are located very close to each other.



- However, previous study of *Anopheles gambia* (malaria mosquito) regarded the male-biased genes as random chromosomal distributions due to fail to detect any significant deviation in the distribution of male-biased genes.

[4] RNA-Sequencing Analysis for the yellow fever mosquito *Aedes aegypti*

Conclusions

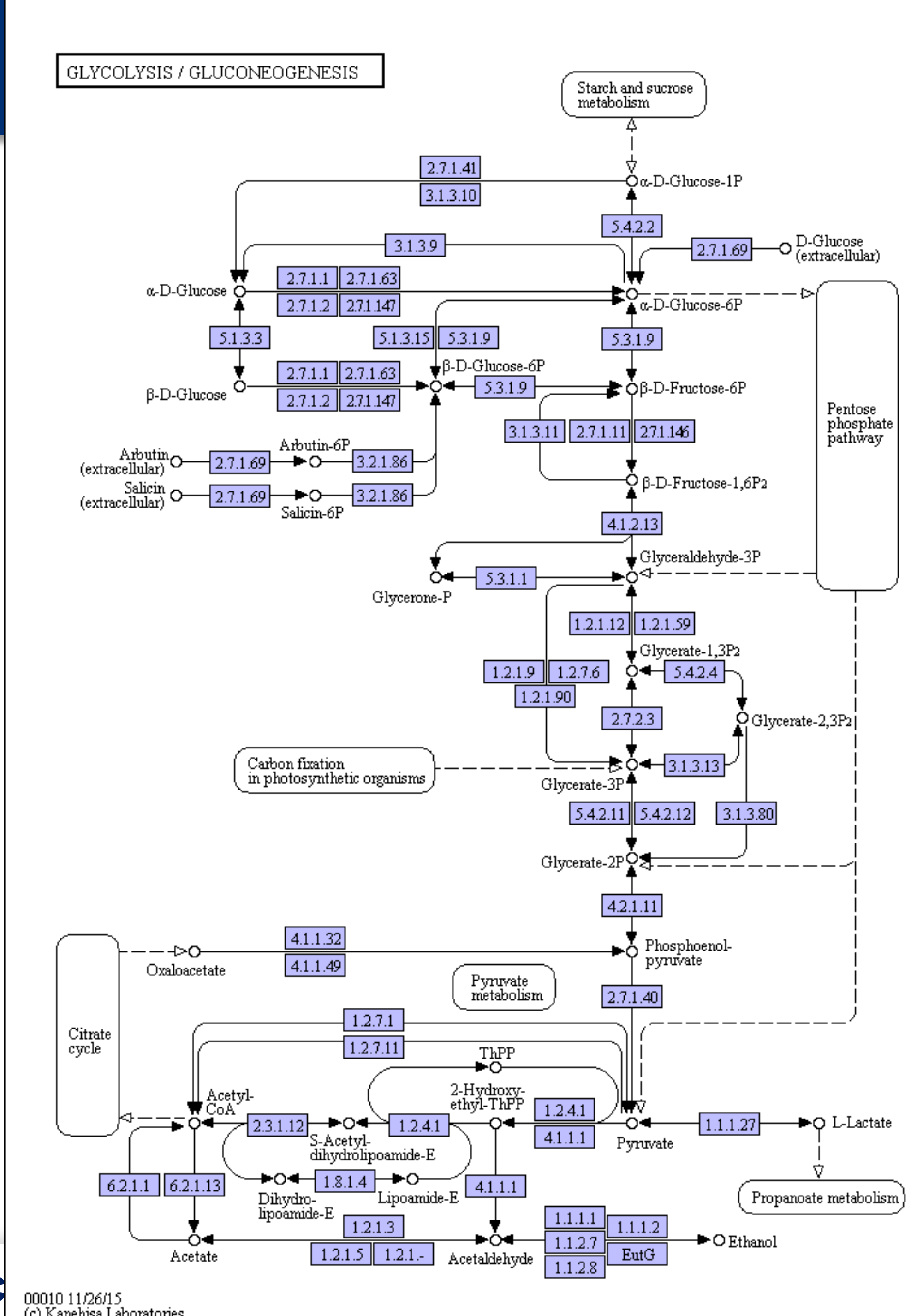
- We provided a ***detailed analysis to find un-annotated and mis-annotated transcripts and sex-biased transcripts*** using *Aedes aegypti* RNA-Seq data and bioinformatics tools.
- The volume and complexity of the RNA-Seq data require scalable algorithms on the high-performance computing machines.
- We implemented a new tool R2EX that efficiently finds new highly expressed regions (not annotated in previous, and can be a new exon or gene) and overlapped with previous exons (possible mis-annotated exons).
- Another contribution of this work is to analyze differential transcript expression of RNA-Seq data especially on sex-biased transcripts.
- We found ***male-biased transcripts of Ae. aegypti using RNA-Seq and these genes are non-randomly distributed and regionally enriched***.

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