Studying fermentation via RNA-seq

Ignat Sonets, Kamilla Faizullina

Abstract. We analyze RNA-seq data from yeast to study gene expression during fermentation. We make alignment and analyze expressions level using utilities. Using The Saccharomyces Genome Database, we propose the gene which could be important for fermentation process.

1 Introduction

Hello! Today we perform differential expression analysis for RNA-seq data of S.cerevisiae to estimate changes in gene expression levels while baking the bread. But before we start, I want to briefly introduce some data about yeasts and fermentation process. As we all know, for making bread we use yeasts. Yeasts eat sugars and put dough to rise. But what processes occur? First, it is an anaerobic process. You might be surprised, but yeasts undergo ethanol fermentation during dough rising (and if you make your dough with more water and leave your mixture near a heat source for about 2-3 days, you will recreate the ancient beer recipe, as it was done by ancient Egyptians.). Fermentation of sugars in flour is the core of bread making. How fermentation works? Ethanol fermentation transforms one mole of glucose into two moles of ethanol and two moles of carbon dioxide (which is the most important component for bread making), producing two moles of ATP (unfortunately, it won't make bread a Red Bull) in the process. Details about given process are taken from [1]. The overall chemical formula for alcoholic fermentation is:

$C6H12O6 \rightarrow 2C2H5OH + 2CO2$

Sucrose is a sugar composed of a glucose linked to a fructose. In the first step of alcoholic fermentation, the enzyme invertase cleaves the glycosidic linkage between the glucose and fructose molecules.(NB: almost the same process can be made with saccharose, which is also dimer consist of glucose and fructose, but with different glycoside bond).

$C12H22O11 + H2O + \text{invertase} \;\; 2\ C6H12O6$

Next, each glucose molecule is broken down into two pyruvate molecules in a process known as glycolysis [2]. Glycolysis is summarized by the equation:

C6H12O6 + 2 ADP + 2 Pi + 2 NAD+ → 2 CH3COCOO + 2 ATP + 2 NADH + 2 H2O + 2 H+

CH3COCOO is pyruvate, and Pi is inorganic phosphate. Finally, pyruvate is converted to ethanol and CO2 in two steps, regenerating oxidized NAD+ needed for glycolysis:

1. CH3COCOO + H+ → CH3CHO + CO2

catalyzed by pyruvate decarboxylase

2. CH3CHO + NADH + H+ → C2H5OH + NAD+

This reaction is catalyzed by alcohol dehydrogenase. So, we think, that yeasts would behave differently between 2 states: before the start of the fermentation and during the fermentation, because cells will(and should) respond to its changing environment. To provide necessary enzymes for fermentation, the yeast cell should start their synthesis (i.e. translation). To start their synthesis, transcription (DNA to RNA information transfer) should begin. So your gene expression(i.e. the process by which information from a gene is used in the synthesis of a functional gene product that enables it to produce protein as the end product[3]) changes. But how to e s t i m a t e this? Differential expression analysis at your service. To do this, we need RNA-seq data, reference genome of S.cerevisiae, annotation file (to find which gene changed its expression levels) and skills. By measuring differences in gene expression we can not only confirm our suggestions, but also discover new data that could potentially be useful in biotechnology(i.e. modifying enzymes, maybe adding new enzymes or even massive genomic rearrangements) and even on your kitchen by, for example, tweaking flour/water proportions, changing flour type, adding more sugar etc. This could be done with tries and errors and many repeats, but if we can obtain some evidence-based discoveries, why not? Let's get started.

2 Data

We analyze transcriptome data of Saccharomyces cerevisiae obtained using IconIllumina HiSeq 2000. In order to study RNA expression levels, we use RNA-seq data from yeast obtained before and during fermentation [4].

We also use Saccharomyces cerevisiae assembly R64 strain S288C from NCBI [5] as a reference genome data. The data were obtained via Illumina HiSeq 2000.

3 Methods

RNA-seq differential gene expression analysis allows to measure quantitative changes in the expression levels between the experiments. We analyze yeast data before and

during fermentation as we would like to compare expression of different genes between fermentation process and normal growth.

First, we should make alignment. We use HISAT2 which allows us make alignment for RNA sequencing reads [6]. HISAT2 alignment is based on extended the Burrows-Wheeler transform with indexing sequences.

The utility DESeq2 is used to perform differential gene expression analysis [7]. This method is based on the assumption negative binomial distribution in the model. We use command featureCounts from the Subread package [8] for counting reads and command gffread [9] to prepare results for DESeq2.

After using DESeq2, we have output results. To visualize RNA-Seq results, we can use heatmap plots and volcano plots [10].

To make interpretation of results, we use Saccharomyces Genome Database by Stanford [11]. The Saccharomyces Genome Database provides information about genes functions from the yeast genome.

4 Results

We successfully aligned our data to S288C reference genome using HISAT. Results are in Table 1.

Reads	SRR941816	SRR941817	SRR941818	SRR941819
Assigned	7291723	7987001	1402166	4975466
Unassigned-Unmapped	520138	511726	66371	234531

Table 1: HISAT alignment results short summary

Later we filtered and sorted our results using SAMtools. After coverting GFF annotation file to GTF, we used FeatureCounts to count reads for various genes and features. DESEQ2 package outputs gave us normalized counts for each feature, and heatmap(Fig. 1) shows us 1)clustering of two states (0 min/30 min) and 2)difference of gene expression between 2 states (the more red point is, the bigger the value and so the bigger is expression).To put in simplified way, when observing the heatmap we can see that across all genes, proprotion of expressed genes of 0 min/30 min are approx. 50/50, and genes which are expressed in 0 min state are not expressed in 30 min state.

Finally, GO Slim mapper was used to "attach" top 50 genes with lowest adjusted p-value(so the most important observations were selected) for corresponding metabolics pathway(See Suppl.Table 1 for more info) To visulaize up/downregulation of the genes according to organism state(before/strat fermentaion), we made a volcano plot showing all of 50 genes.

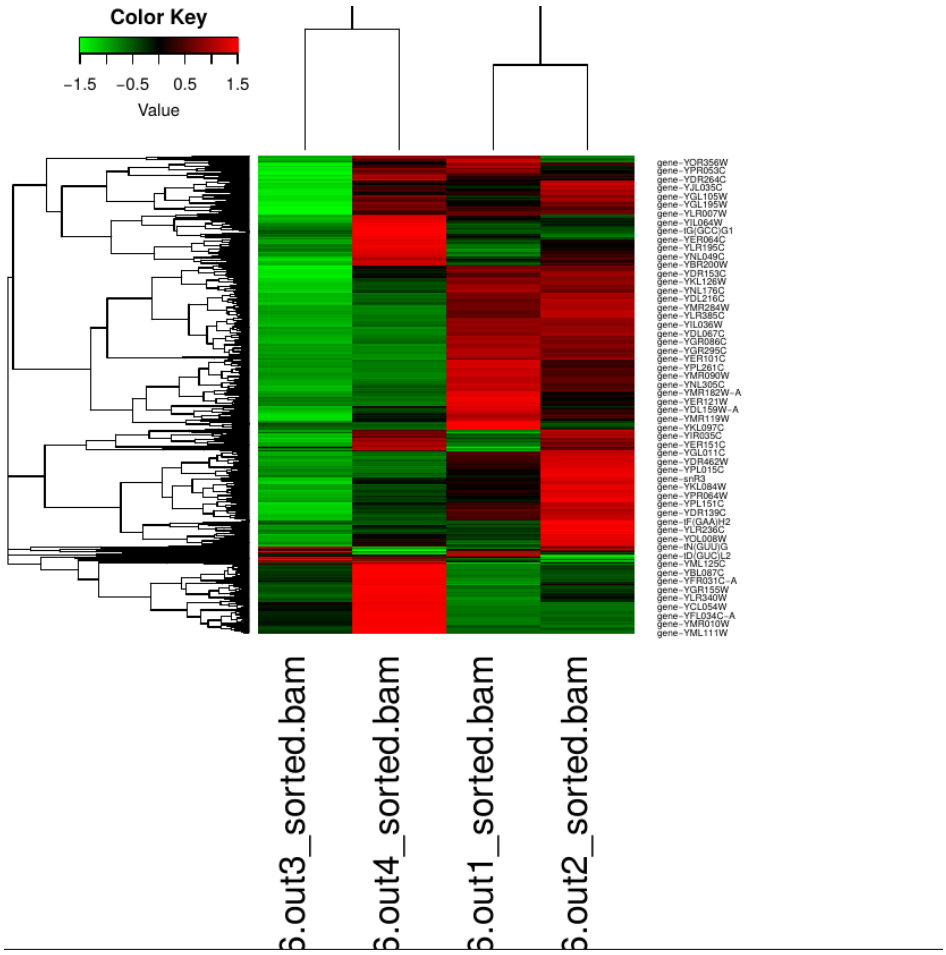

Figure 1: Heatmap of differentially expressed genes made with DESEQ2 results.

In total(according p-adj. value) 3450 gene changed its expression before starting the fermentation and after 30 min of fermentation. But describing each of the changes would take an eternity, so we narrowed down the focus to top 50 genes. We found 39 GO terms describing these pathways (see Suppl.Table 1 for more details.), and all of them are upregulated. According to the volcano plot, 48 out of 50 genes are upregulated (see Fig.2), except for 2 genes.

The main reason for such a result is a need for fast adaptation for new conditions, main limitaion of them(these new conditions) is lack of O2, and to survive, yeast cell should assemble ribosomes, create a lot of new RNAs of all types, assemble new enzymes, transport new sugars and start converting them to ATP as fast as possible. And we can clearly see, that is happening exactly as we predicted. Let's investigate GO terms and genes more thoroughly.

We start with downregulrated genes, because there are only 2 of them. First, YKR097W (PCK1)[12] . This gene belongs to carbohydrate metabolic process term (GO:0005975). This gene encodes phosphoenolpyruvate carboxykinase; it is the key

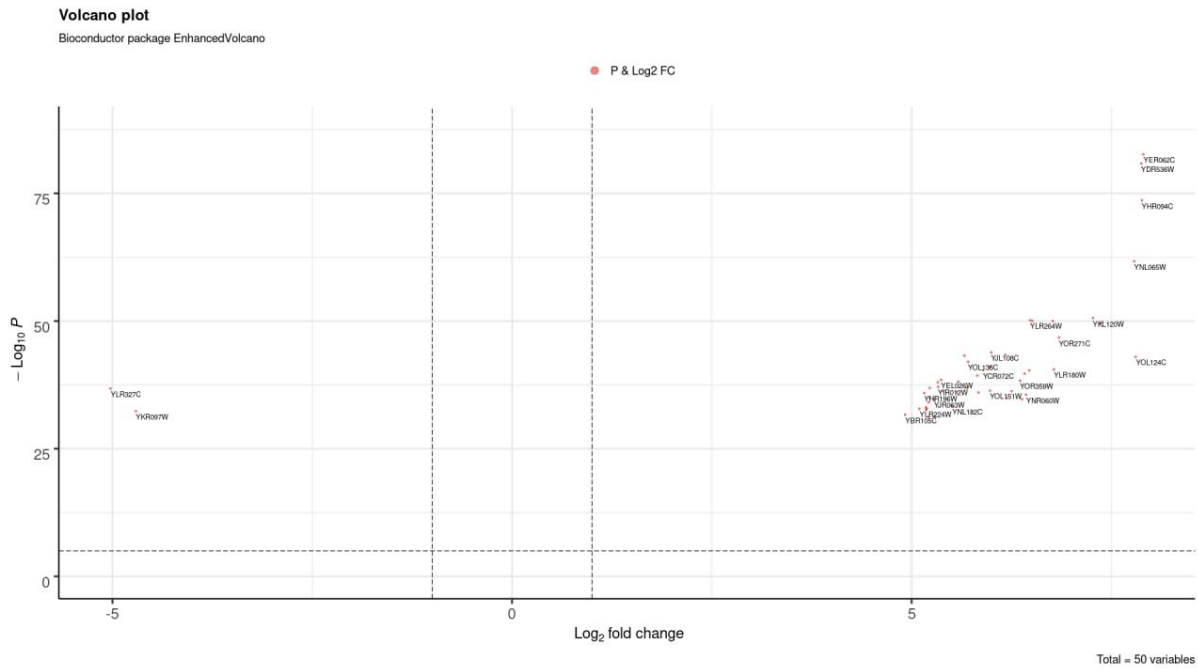

Figure 2: Volcano plot showing top-50 up/downregulated genes.

enzyme in gluconeogenesis, catalyzes early reaction in carbohydrate biosynthesis, located in the cytosol. By switching metabolism to sugars degradation in lack of O2, gluconoegenesis is not needed, so no surprise it is downregulated. One can assume that gene is essential for gluconeogenesis which is crucial when growing on sugars-free media, but flour is rich for carbohydrates, so no senese in this gene and its product during bread making.

The second gene–YLR327C(TMA10)[13] – is not allocated for any of GO terms. It encodes protein of unknown function that associates with ribosomes; it is known that protein abundance increases in response to DNA replication stress. Here things got more speculative. At first glance, ribosomes number should rise (as bread:)), and so this protein should also produce more active, but we don't see this. So I can't give the exact answer that could reveal role of this protein. The only thing I can assume that this protein forms dimeric complexes with ATP-synthase (Dienhart M, et al. (2002)[14]). As we all know, fermentation is much less energy-efficient than aerobic metabolism,so cell doesn't nedd so much ATP-synthases units,and numbers of TMA10 and TMA10 RNA decreases.

Other genes are upregulating, and because we can't discuss all of 48 genes and their roles in pathways, we will choose only 1 gene and try to hypothesize how it might be a part of actual changes in yeast metabolism. For example, SYO1 / YDL063C [15], part of the ribosomal large subunit biogenesis (GO:0042273). It is SYnchronized impOrt or SYmpOrtin ; it is assembly chaperone that co-translationally associates with nascent

Rpl5p, preventing aggregation; facilitates synchronized nuclear coimport of two 5S-rRNA binding proteins, Rpl5p and Rpl11p, mediated by import receptor Kap104p; required for biogenesis of the large ribosomal subunit. By helping with an assembly of LSU (5S rRNA is a structural component of LSU), thus accelerating ribosome assembly as a functional complex, it provides the synthesis of necessary enzymes and maintaining fermentation speed and keeping the cell alive. Thank you for your attention!

References

[1] https://en.wikipedia.org/wiki/Ethanol_fermentationBiochemical_process_of_fermentatio

[2] Stryer, Lubert (1975). Biochemistry. W. H. Freeman and Company. ISBN 978-0-7167-0174-3.

[3] https://en.wikipedia.org/wiki/Gene_expression

[4] ftp.sra.ebi.ac.uk/vol1/fastq/SRR941/SRR941816/SRR941816.fastq.gz
ftp.sra.ebi.ac.uk/vol1/fastq/SRR941/SRR941817/SRR941817.fastq.gz
ftp.sra.ebi.ac.uk/vol1/fastq/SRR941/SRR941818/SRR941818.fastq.gz
ftp.sra.ebi.ac.uk/vol1/fastq/SRR941/SRR941819/SRR941819.fastq.gz (282 Mb)

[5] Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, Connor R, Fiorini N, Funk K, Hefferon T, Holmes JB, Kim S, Kimchi A, Kitts PA, Lathrop S, Lu Z, Madden TL, Marchler-Bauer A, Phan L, Schneider VA, Schoch CL, Pruitt KD, Ostell J. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2019 Jan 8;47(D1):D23-D28. doi: 10.1093/nar/gky1069. PubMed PMID: 30395293; PubMed Central PMCID: PMC6323993. 2: Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2019 Jan 8;47(D1):D94-D99. doi: 10.1093/nar/gky989. PubMed PMID: 30365038; PubMed Central PMCID: PMC6323954.

[6] Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature Methods 2015

[7] Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8.

[8] Liao Y, Smyth GK and Shi W (2014). featureCounts: an efficient general pur-pose program for assigning sequence reads to genomic features.Bioinformatics,30(7):923-30.http://www.ncbi.nlm.nih.gov/pubmed/24227677

[9] How to cite this article Pertea G and Pertea M. GFF Utilities: GffRead and GffCompare [version 1; peer review: 3 approved]. F1000Research 2020, 9:304 (https://doi.org/10.12688/f1000research.23297.1)

[10] Maria Doyle, 2021 Visualization of RNA-Seq results with Volcano Plot (Galaxy Training Materials). /training-material/topics/transcriptomics/tutorials/rna-seq-viz-with-volcanoplot/tutorial.html Online; accessed Fri Apr 02 2021

[11] Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. Jan;40(Database issue):D700-5. [PMID: 22110037]

[12] https://www.yeastgenome.org/locus/YKR097W

[13] https://www.yeastgenome.org/locus/S000004319

[14] Dienhart M, et al. (2002) Formation of the yeast F1F0-ATP synthase dimeric complex does not require the ATPase inhibitor protein, Inh1. J Biol Chem 277(42):39289-95

[15] https://www.yeastgenome.org/locus/YDL063C

These 8 identifier(s) represent valid gene names that either could not be mapped to terms in the current GO slim set or are currently annotated to the root node for the slim set being used: YOR360C YBL028C YLR264C-A YML018C YLR327C YGR079W YJL107C YJL108C

GO Terms from the biological process Ontology

GO Term (GO ID)	Genes Annotated to the GO Term	GO Term Usage in Gene List	Genome Frequency of Use
rRNA processing (GO:0006364)	YDR449C, YEL026W, YER127W, YGR159C, YHR066W, YHR196W, YJL069C, YLR264W, YMR093W, YNL112W, YNL182C, YOL041C, YOL080C	13 of 49 genes, 26.53%	366 of 6443 annotated genes, 5.68%
ribosomal large subunit biogenesis (GO:0042273)	YCR072C, YDL063C, YEL026W, YHR066W, YIR012W, YJL122W, YNL182C, YOL041C, YOL080C	9 of 49 genes, 18.37%	130 of 6443 annotated genes, 2.02%
ribosomal small subunit biogenesis (GO:0042274)	YDR449C, YEL026W, YER127W, YGR159C, YHR196W, YJL069C, YLR264W, YMR093W	8 of 49 genes, 16.33%	146 of 6443 annotated genes, 2.27%
ribosome assembly (GO:0042255)	YCR072C, YGR159C, YHR066W, YIR012W, YLR264W, YNL182C, YOL080C	7 of 49 genes, 14.29%	79 of 6443 annotated genes, 1.23%
transcription by RNA polymerase I (GO:0006360)	YHR196W, YJL148W, YJR063W, YML043C, YMR093W, YNL248C	6 of 49 genes, 12.24%	71 of 6443 annotated genes, 1.10%
ion transport (GO:0006811)	YDR536W, YHR094C, YKL120W, YNL065W, YNR060W, YOR271C	6 of 49 genes, 12.24%	340 of 6443 annotated genes, 5.28%
transmembrane transport (GO:0055085)	YDR536W, YHR094C, YKL120W, YNL065W, YOR271C	5 of 49 genes, 10.20%	468 of 6443 annotated genes, 7.26%
nucleobase-containing small molecule metabolic process (GO:0055086)	YBL039C, YMR300C, YNL141W, YOL136C	4 of 49 genes, 8.16%	220 of 6443 annotated genes, 3.41%
carbohydrate metabolic process (GO:0005975)	YBR105C, YER062C, YKR097W, YOL136C	4 of 49 genes, 8.16%	253 of 6443 annotated genes, 3.93%
RNA catabolic process (GO:0006401)	YLR264W, YNL112W, YOR359W	3 of 49 genes, 6.12%	166 of 6443 annotated genes, 2.58%

cellular amino acid metabolic process (GO:0006520)	YBL039C, YLR180W, YMR300C	3 of 49 genes, 6.12%	218 of 6443 annotated genes, 3.38%
regulation of translation (GO:0006417)	YLR264W, YNL112W, YOR359W	3 of 49 genes, 6.12%	234 of 6443 annotated genes, 3.63%
nucleobase-containing compound transport (GO:0015931)	YHR196W, YLR264W	2 of 49 genes, 4.08%	183 of 6443 annotated genes, 2.84%
proteolysis involved in cellular protein catabolic process (GO:0051603)	YBR105C, YLR224W	2 of 49 genes, 4.08%	265 of 6443 annotated genes, 4.11%
RNA modification (GO:0009451)	YOL124C, YPL212C	2 of 49 genes, 4.08%	186 of 6443 annotated genes, 2.89%
DNA-templated transcription, termination (GO:0006353)	YJR063W, YNL112W	2 of 49 genes, 4.08%	42 of 6443 annotated genes, 0.65%
tRNA processing (GO:0008033)	YOL124C, YPL212C	2 of 49 genes, 4.08%	134 of 6443 annotated genes, 2.08%
response to chemical (GO:0042221)	YLR224W, YNL065W	2 of 49 genes, 4.08%	530 of 6443 annotated genes, 8.23%
transcription by RNA polymerase II (GO:0006366)	YJR063W, YNL112W	2 of 49 genes, 4.08%	556 of 6443 annotated genes, 8.63%
DNA-templated transcription, elongation (GO:0006354)	YJL148W, YNL248C	2 of 49 genes, 4.08%	109 of 6443 annotated genes, 1.69%
lipid metabolic process (GO:0006629)	YBL039C, YOL151W	2 of 49 genes, 4.08%	348 of 6443 annotated genes, 5.40%
amino acid transport (GO:0006865)	YNL065W, YOR271C	2 of 49 genes, 4.08%	56 of 6443 annotated genes, 0.87%
regulation of DNA metabolic process (GO:0051052)	YNL182C, YOR359W	2 of 49 genes, 4.08%	97 of 6443 annotated genes, 1.51%
DNA-templated transcription, initiation (GO:0006352)	YML043C, YNL248C	2 of 49 genes, 4.08%	83 of 6443 annotated genes, 1.29%
carbohydrate transport (GO:0008643)	YDR536W, YHR094C	2 of 49 genes, 4.08%	46 of 6443 annotated genes, 0.71%
organelle assembly (GO:0070925)	YLR180W	1 of 49 genes, 2.04%	125 of 6443 annotated genes, 1.94%
cellular ion homeostasis (GO:0006873)	YNR060W	1 of 49 genes, 2.04%	162 of 6443 annotated genes, 2.51%
protein modification by small protein conjugation or removal (GO:0070647)	YLR224W	1 of 49 genes, 2.04%	223 of 6443 annotated genes, 3.46%
mRNA processing (GO:0006397)	YEL026W	1 of 49 genes, 2.04%	220 of 6443 annotated genes, 3.41%
regulation of organelle organization (GO:0033043)	YLR180W	1 of 49 genes, 2.04%	279 of 6443 annotated genes, 4.33%
RNA splicing (GO:0008380)	YEL026W	1 of 49 genes, 2.04%	153 of 6443 annotated genes, 2.37%
generation of precursor metabolites and energy (GO:0006091)	YOL136C	1 of 49 genes, 2.04%	113 of 6443 annotated genes, 1.75%
response to osmotic stress (GO:0006970)	YER062C	1 of 49 genes, 2.04%	73 of 6443 annotated genes, 1.13%
monocarboxylic acid metabolic process (GO:0032787)	YOL136C	1 of 49 genes, 2.04%	164 of 6443 annotated genes, 2.55%
cytoplasmic translation (GO:0002181)	YLR264W	1 of 49 genes, 2.04%	205 of 6443 annotated genes, 3.18%
tRNA aminoacylation for protein translation (GO:0006418)	YDR037W	1 of 49 genes, 2.04%	37 of 6443 annotated genes, 0.57%
protein targeting (GO:0006605)	YBR105C	1 of 49 genes, 2.04%	256 of 6443 annotated genes, 3.97%
DNA replication (GO:0006260)	YNL182C	1 of 49 genes, 2.04%	151 of 6443 annotated genes, 2.34%
DNA recombination (GO:0006310)	YGR159C	1 of 49 genes, 2.04%	255 of 6443 annotated genes, 3.96%