Differentially expressed genes associated with resistance to a parasite in the freshwater snail *Biomphalaria glabrata*

Evan Lavelle

**Abstract**

Motivation: *Biomphalaria glabrata* is a species of freshwater snail that acts as an intermediate host for *Scistosoma mansoni*, a trematode parasite responsible for over 200 million cases of schistosomaiasis in humans. We hypothesize that the resistance to infection in some strains of *B. glabrata* is due in part to expression of lectin genes involved in their immune system.

Results: The expressional changes between two strains of snails upon exposure to *S. mansoni* are compared to determine if there is evidence for this hypothesis by differential gene expression analysis. We found that 43 genes are significantly differentially expressed in response to the parasite challenge between the two strains. Notably, the list includes a lectin gene down-regulated in the resistant strain.

Conclusions: Our findings provide further insight into what gene regulatory mechanisms are responsible for resistance to parasite infection in fresh water snails, not conclusively reported in the original study that generated the data. This, in turn, could produce a means by which to ultimately prevent transmission of the infection to humans.

**Introduction**

*Biomphalaria glabrata* is a species of freshwater snail that acts as an intermediate host for *Schistosoma mansoni*, a trematode parasite responsible for several million cases of schistosomaiasis in humans. We hypothesize that the resistance to infection in some strains of *B. glabrata* is due in part to expression of lectin genes involved in their immune system. We incorporated previously acquired RNA-seq data for a differential gene expression analysis between parasite susceptible/resistant strains and exposed/unexposed groups of *B. glabrata*. We examined the data with regard to presence of lectin genes in the *B. glabrata* genome to determine if there is significantly changed transcription upon exposure to *S. mansoni* in the resistant group compared to the susceptible group.

The data used is from *B. glabrata* snail samples descendent from a population collected from Guadelope (Tennessen *et al*., 2015). To maintain a control population, one group of snails were indiscriminately bred, whereas others were selectively bred based on resistance to *S. mansoni*. The observed infection rate upon exposure at the conclusion of this practice was 52% and 10%, respective to the susceptible and resistant bred strains. RNA-seq data were generated to identify genes in the GRC and perform differential gene expression between the two phenotypic groups. Although the researchers distinguished between two susceptible genotypes in their analyses (S1 and S2), no significant difference was determined in resistance of the snails representing them. Therefore, they are considered synonymous in this project.

Each single RNA-seq sample consists of pooled materials from three individual snails, all homozygous for their GRC locus. The researchers sequenced 18 such samples for both phenotypes unexposed to the parasite *S. mansoni*, and 6 both for those exposed to the parasite. The exposed group samples were evenly divided between those exposed for two and six hours. However, the researchers found no significant difference in the results between the two exposure time subgroups. Thus, they are considered as one group in the project.

Lectin proteins are characterized by a conserved carbohydrate recognition domain (Drickamer, 1989). C-type lectins in particular are known to possess varied immunological functions (Brown *et al*., 2018). To date, the agents responsible for resistance in some strains of *B*. *glabrata* to *S*. *mansoni* have not been conclusively identified. While elucidation of the mechanisms governing this interaction is beyond the scope of this project, we report several genes with significantly different expressional changes to exposure. Among them is a C-type lectin, lending evidence to our hypothesis that it holds an influential role in the immune response of *B*. *glabrata* to *S. mansoni*. Furthermore, for every DEG, we present on whether the resistant phenotype’s expressional changes are positive or negative in relation to the susceptible phenotype’s.

**Data and code availability**

Table 1 summarizes the 48 samples used in this analysis. There are 12 resistant strain, single-end read samples in the control group and 6 in the experimental group. Likewise, there are 12 susceptible, single-end read samples in the control group and 6 in the experimental group. In addition to these 36 single-read samples, there are 6 susceptible strain and 6 resistant strain paired-end samples, both in the control group. RNA-sequencing was conducted at Oregon State University on an Illumina HiSeq 2000 machine. Six of the 18 unexposed samples for both phenotypic groups used paired-end reads, while all others are single-end reads. Raw RNA-seq reads are available from the NCBI Sequence Read Archive (SRA) using Bioproject Accession PRJNA264063.

The shell script files on the Discovery HPC cluster for transcriptome assembly and gene abundance estimation are titled “assembly\_script.sh” and “estimate\_script.sh”, respectively. Source code for the differentially expressed gene (DEG) analysis is present in “glabrata\_DGE.R” with a script to search the genome annotation file in “annotation\_search.R”.

**Methods**

*Transcriptome assembly*

Transcriptome assembly was performed on NMSU’s high performance computing cluster, Discovery. From within the project space in Discovery, a script calling the wget command was used with sample data links present in the .csv file acquired via the study’s accession number to download .srr files to the server. Following this, the .srr files were converted to fastq files with NCBI’s SRA toolkit directly from the command line of the head node. For the 12 paired-end read samples, the necessary --splitfiles option was used. The genome used for the subsequent steps was submitted by The Genome Institute at the Washington University School of Medicine to NCBI’s genome database, where it is available via accession numbers PRJNA290623 and PRJNA12879. It (genome assembly ASM45736v1) represents the sequence of a susceptible *B*. *glabrata* line BB02, and was indexed with the hisat2-build command. The concomitant alignment file is present as a .gff that was checked for suitability with gffcompare. The multi-fastq files resultant from the previous step were input to hisat2 (Version 2.1.0; Kim et al., 2015; Daehwan et al., 2015) with the --dta option for mapping to the chromosome, with the output being a collection of .sam files. These were converted to .bam files and sorted by a single command with samtools v1.3.1. Transcript assembly was then performed with stringtie v1.2.2, with the results compiled into a single .gtf transcriptome by the stringtie --merge function option. Finally, the stringtie -e and - B options were used in conjunction to produce a gene counts file in .csv format.

*Differential gene expression analysis*

We applied generalized linear models with a negative binomial noise model using the R package DESeq2 (Version 1.24.0; Love et al, 2014) to prepare the count data for differential gene expression analysis. For quality control, 5,654 genes with a total count under 50 are omitted from the analysis. Being that the goal of the analysis is to identify genes belonging to the resistant strain that are either up or down regulated to an extent significantly distinguished from the response of the susceptible strain, interaction between these two factors is the final layer of the model design and shown below:

dataSet <- DESeqDataSetFromMatrix(countData = cleanCounts, colData = samples,

design = ~Sequencing\_Type + Phenotype + Treatment + Phenotype:Treatment, tidy = FALSE)

dds <- DESeq(dataset, minReplicatesForReplace = 6)

*Outlier removal in analysis of each gene*

The software defaults to a value of 7 for the minimum number of replicates in a sample group required for replacement of outlier values. Because several such groups in the sample (e.g. “control”/“susceptible”/“paired”) are represented by a total of 6 replicates (as can be observed in Table 1), this value was specified to allow for filtering via the Cook’s cutoff value. This statistic is attached to each individual count, and is a measure of how much the given value influences analysis of the group in which it is present. An outlier count among several more consistent samples in a group, for example, will yield a higher cook’s Cutoff value than it would among fewer, where it is more difficult to categorize the outlier as anomalous. Applying this standard to the results, we effectively eliminated genes which would be significantly differentially expressed due to single outliers within a sample group.

ddsResults <- results(dds, list(“Phenotyperesistant.Treatmentchallenge”), cooksCutoff=.05)

summary(ddsResults, alpha=0.05)

out of 59245 with nonzero total read count

adjusted p-value < 0.05

LFC > 0 (up) : 17, 0.029%

LFC < 0 (down) : 26, 0.044%

The intersection of DEGs that meet this criterion and the DEGs that meet the p-value is 43, or 0.073% of all input genes. For contrast, calling the results of the single factor comparison between the experimental group and the control group for the baseline susceptible strain yields 6,164 genes. The resistant strain involves a somewhat more complex extraction, and has 5,760 DEGs between the treatment groups, as shown by the following summary. The intersection between these results is 4,730 genes.

sus <- results(dds, contrast = c(“Treatment”,“challenge”,“control”),

cooksCutoff=.05, alpha = .05)

summary(sus)

out of 59245 with nonzero total read count

adjusted p-value < 0.05

LFC > 0 (up) : 4655, 7.9%

LFC < 0 (down) : 1509, 2.5%

res <- results(dds,

contrast=list(c("Treatment\_challenge\_vs\_control",

"Phenotyperesistant.Treatmentchallenge")), cooksCutoff = 0.05, alpha = .05)

summary(res)

out of 59245 with nonzero total read count

adjusted p-value < 0.05

LFC > 0 (up) : 4356, 7.4%

LFC < 0 (down) : 1404, 2.4%

To examine the model’s mechanisms in more detail, a series of PCA plots made with the ggplot2 package demonstrates the outcome of library normalization and sequencing type effect adjustment. Figure 1 is produced from the code below and shows log values of the raw input counts after a pseudocount of 1. The expected exposure effect is noticeable along with an undesired paired-end read effect.

lognormcounts <- DESeqTransform(dds)

assay(lognormcounts) <- log2(1 + counts(dds, normalized = FALSE))

Figure 2 introduces library size normalization to the data, as evidenced in the code below. It can be seen that this operation largely mitigates what was a glaring segregation of paired-end sequenced samples. It can be concluded from this that the paired-end sequenced samples are essentially reliable; they are distinct mostly in that technique correlates with library size.

lognormcounts <- DESeqTransform(dds)

assay(lognormcounts) <- log2(1 + counts(dds, normalized = TRUE))

However, it is still possible to isolate and correct for paired-end read effect apart from performing the normalization. To accomplish this, a vector is extracted from the model results with values reflecting sequencing type effect calculated post-normalization. Then, it is vector-wise subtracted from the previously taken log values of normalized values after pseudocount for samples sequenced via paired-end reads. The code below demonstrates this process on an excerpt of the data it is functioning on:

> coef(dds)[1:5,1:2]

Intercept Sequencing\_Type\_paired\_vs\_single

MSTRG.35554|LOC106071988 4.591944 0.4721993815

MSTRG.52545|LOC106050692 9.243902 0.0001842889

MSTRG.54602 1.343604 0.3465093957

MSTRG.54601 3.376364 -1.0608916384

MSTRG.54600 4.055727 0.3733778408

adjustVector <- coef(dds)[,2]

adjustVector <- unname(adjustVector)

removed <- lognormcounts

for (index in which(samples$Sequencing\_Type == "paired")){

assay(removed)[,index] <- assay(lognormcounts)[,index] - adjustVector

}

assay(lognormcounts)[1:5,13:17]

X1617536 X1617537 X1617538 X1617539 X1617540

MSTRG.35554|LOC106071988 5.321338 4.771742 5.756312 3.316063 5.733260

MSTRG.52545|LOC106050692 8.889877 9.764902 8.545270 8.438101 8.657352

MSTRG.54602 0.000000 2.696610 0.000000 4.372307 2.569857

MSTRG.54601 2.033490 3.766634 0.000000 0.000000 0.000000

MSTRG.54600 4.420887 3.513144 4.426123 4.801069 5.033032

assay(removed)[1:5,13:17]

X1617536 X1617537 X1617538 X1617539 X1617540

MSTRG.35554|LOC106071988 4.8491385 4.299543 5.2841125 2.843863 5.261060

MSTRG.52545|LOC106050692 8.8896931 9.764718 8.5450858 8.437916 8.657168

MSTRG.54602 -0.3465094 2.350100 -0.3465094 4.025798 2.223348

MSTRG.54601 3.0943815 4.827525 1.0608916 1.060892 1.060892

MSTRG.54600 4.0475088 3.139766 4.0527448 4.427692 4.659654

The result of plotting “removed” object as a PCA graph generates Figure 3. DEGs of the final model can be extracted with the following commands:

# **selects for genes with log fold change in regulation by chosen false discovery rate**

ddsSig <- ddsResults[which(ddsResults$padj < 0.05),]

**# reorders by pvalue**

ddsSigOrdered <- ddsSig[order(ddsSig$padj),]

> nrow(ddsSig)

[1] 43

Another program, “annotation\_search.R” scans the genome annotation file for characterizations of all genes with LOC identification. The rtracklayer package is used to load the .gff annotation file for this purpose.

exonIndeces <- which(annotation$type == "exon")

**# matches to smaller gene vector, because transcript rows(first encounters), aren't named**

falseIndeces <- match(SigNames, annotation$gene[exonIndeces])

productIndeces <- exonIndeces[falseIndeces]

productNames <- annotation$product[productIndeces]

**Results**

The experimental design is summarized in Table 1 for the 48 samples of the dataset. Of the 59,245 of the 64,899 unique transcripts identified by StringTie with total counts of at least 50 in the dataset, DESeq2 determines 43 to exhibit a significant difference in expression between the two phenotypes upon *S. mansoni* exposure with an adjusted P-value no more than 0.05. Of the 43 DEGs, 18 are annotated by transcripts in the ASM45736v1 assembly, with the remaining 25 being de novo transcripts predicted by the StringTie assembly. For comparison, 30,069 of the constructed assembly’s total transcripts of 64,899 are shared by the ASM45736v1 assembly, 6,303 of which are confidently identified. No fully identified genes are present in the DEGs. Taking advantage of the annotation file makes it possible to view the results of analysis with suspected gene names incorporated (Table 2). DEG #41 (row-wise on Table 2) contains two names, for which StringTie built a transcript bridging these two zones. It can also be seen that DEG #34 is one of 77 lectin-characterized transcripts in the genome annotation.

Figure 1 lends insight into factors that contribute variability to the data. Principal component 1, representing the linear combination of genes with the largest variance in the sample data, largely corresponds to the effect of parasite exposure. The positive values agree with biological response to infection; while some genes may certainly be downregulated in a complex gene regulatory network, the net protein production can be expected to rise as the organism’s defenses are fortified. Normalization for the sequencing library size, as seen in Figure 2, neatly merges distributions of paired-end and single-end control group samples. More noteworthy however, is the greater range of PC2 values in the control group than in the experimental one that remains. In part, this could be the consequence of less replicates for the latter, but it also suggests a dominating exposure effect. This idea is substantiated by the comparatively huge number of genes that significantly respond to the parasite, as shown in this single factor comparison in the Methods section. Paired-end sequencing effect adjustment (Figure 3), unlike library-size normalization, slightly reduces PC1 variance, though it only marginally impacts the placement of plotted points.

Figure 4 shows a DEG #34 (annotated as C-type lection domain family 7 member A-like) by adjusted counts for this gene. While exposed samples of the susceptible strain are in accordance with the general upregulation observed across genes (as seen in Figures 1—3), those of the resistant strain exhibit less copies when compared to their own control group samples. It is reasonable that a resistant phenotype would display a less drastic response to infection, given it may have a baseline expression of immune-related genes at higher abundance as a preemptive measure. It is curious, however, that this gene appears actually downregulated in response to exposure in the resistant strains. The implications of this possibility are considered in Discussion.

Annotations are present for 14 of the 43 genes, though they are not confirmed to be *B. glabrata* orthologs for the genes named. We summarize their functions based on information from genecard.org. Androgen-induced gene 1, the 4th most DEG, produces a protein involved in long-chain fatty acid catabolism. The 5th most DEG, PARP14, is an anti-apoptotic gene hypothesized to regulate glycolysis. PRRT1, the 9th most DEG, is a transmembrane component of the AMPAR pore complex. The 12th most DEG, MRPS2, is involved in forming the structure of mitochondrial ribosomes. The 15th most DEG is HSPA1, a protein folding chaperone that also assists in transport and marking proteins for degradation. The 16th, MAP3K3, is a part of protein kinase signal transduction. The 21st most DEG, TNRC6C, is involved in gene silencing via miRNAs. The 26th most DEG is a component in cytoskeletal structure, TUBA3. The 24th is rho GTPase-activating protein gacU. The aforementioned 34th most DEG is CLEC7A, a type 2 membrane receptor cited to recognize plant and fungi glucans, and stimulates T-cell manufacture. The 35th is a general characterization for STK, which can act in a variety of capacities depending on the family, and the 36th is a transcript variant of HSPA1 (the 15th most DEG). The 37th is MIPEP, which processes proteins related to oxidative phosphorylation by cleaving down to mature size upon their induction into the mitochondrion. The last gene with annotation is the 41st most DEG, the endonuclease RBBP8, an endonuclease that associates with the protein BRCA1, and is thought to assist in DNA repair, mitotic control, and transcriptional regulation. Of the DEGs, 17 are more positively expressed in the resistant strain upon exposure and the remaining 26 are more positively expressed in the susceptible strain upon exposure.

**Discussion**

Authors of the study from which the published RNAseq was used (Tennessen *et al*., 2015) performed a fundamentally different analysis in that it is focused solely on the region they initially identified to exhibit high allelic frequency divergence correlated to the resistant/susceptible phenotypes. It was only for the 27 genes in this region that differential expression analysis was performed, and only for independent comparisons between the susceptible and resistant strains in the control group, and the two strains in the challenge group. Only two DEGs were found, both in the comparison between strains in the control group, and both non-coding genes. The authors discussed the underwhelming evidence resultant from their DGE analysis and suggested an alternate cause behind *S*. *mansoni* resistance: “Although we cannot rule out a role for subtle expression differences, including in noncoding transcripts, the simplest mechanism of dominant resistance via gene expression would be over expression of the resistance allele in a coding gene, and such a pattern is never observed. In contrast, the strikingly high amino acid sequence divergence is a much stronger candidate for the functional mechanism” (Tennessen *et al*., 2015 p. 11).

The researchers performed restriction site associated (RAD) genotyping on samples of the groups to reveal a genomic region (referred to as the Guadeloupe Resistance Complex, or GRC) closely correlated to the susceptible/resistant phenotypes. Because the 27 genes in this set exhibit near-perfect linkage disequilibrium, the term “allele” was used to describe expression of the entire haplotype, with the hypothesis that the dominant gene set (R) confers resistance, and the two recessive ones (S) confer susceptibility.

In our analysis, none of the genes of the Guadeloupe Resistance Complex were present as annotated genes in the results of our multi-factor DGE model. While the possibility cannot be discounted that some of the DEGs we found serve some function in a resistance mechanism, it is more likely that most are passively affected by the overall strain-specific reaction of *B. glabrata* to *S. mansoni*.

One gene that may plausibly help endow resistance is heat shock protein 70A. Notably, there are two variants of HSP70A in the list of DEGs, one with a more positive change in the resistant strain and the other with a more positive change in the susceptible strain. While this protein was once thought to be relegated to intracellular maintenance, they are now known to also be exported outside the membrane (De Maio, 2014). Further, they have been found to have active immunological roles, though only those pertaining to autoimmunity have been detailed (Binder, 2014).

As the only DEG coding for a membrane receptor, CLEC7Aa type has unique opportunity for involvement in resistance to infection. Interestingly, four genes in the GRC that Tennessen *et al*. hypothesize responsible for resistance are also single-pass transmembrane proteins with the same orientation: “Specifically, we hypothesize that their extra cellular (N-terminal) domains recognize foreign substances such as parasite PAMPs (pathogen-associated molecular patterns), while their intracellular (C-terminal) domains transmit this signal to other cellular components, leading to a physiological response” (Tennessen *et al*., 2015 p. 11).

C-type lectin receptors are calcium ion-dependent proteins containing a conserved carbohydrate recognition domain (Kerrigan and Gordon, 2010). Since their ability to activate natural killer (NK) cells was discovered (Lanier, 2008), the role of CLEC7A (a.k.a. Dectin-1) in immune response to viral, fungal, and bacterial infection has become well established (Drummond and Brown, 2011). Having passed a strict means of outlier detection without any counts being replaced and possessing plausible biological potential to combat infection, CLEC7A stands out as a candidate benefactor of the resistant *B*. *glabrata* strain. However, Figure 4 indicates an association of the upregulation of this gene with the susceptible strain after exposure. At least two C-type lectins have been shown to behave as a liability when challenged with particular viruses by helping the particles propagate (Pohlmann *et al*., 2003). Although this precedent may not apply for the interaction with *S*. *mansoni*, it is not unfathomable that the parasite could be taking advantage of the protein by binding to it while evading the immune response.

Unfortunately, very little information is available for C-type lectins on which base more detailed speculation, particularly in relationship to animal parasites. In regard to their functions apart from those related to autoimmunity and microbe defense, Brown *et al*. remarks: “Despite their importance, many of the C-type lectins that are involved in these activities are poorly characterized, and our understanding of their contributions and underlying mechanisms remains incomplete. A logical next step to confirm CLEC7A’s involvement in reaction would be quantitative PCR analysis on various tissues from the same conditions of *B*. *glabrata* samples, where more stark differences in expression might be observed. Gene expression network inference with time-series exposure data would allow for a clearer picture of the *B*. *glabrata*’s response, and a knock-down experiment targeting CLEC7A could confirm or refute a causative role of the gene’s activity in immunity to *S*. *mansoni* in some freshwater snails.

**References**

Binder, R.J. (2014) Functions of Heat Shock Proteins in Pathways of the Innate and Adaptive Immune System. Journal of Immunology 193(12): 5765-5771.

Brown G.D., Willment J.A., Whitehead L. (2018) C-type lectins in immunity and homeostasis. Nature Review Immunology 18(6): 374-389.

Daehwan K., Langmead B., Paggi J., Pertea G., Salzberg S. (2015). HISAT2: graph-based alignment of next generation sequencing reads to a population of genomes (Version 2.1.0) [Software]. Available from: <http://ccb.jhu.edu/software/hisat2/index.shtml>

De Maio A. (2014) Extracellular Hsp70: export and function. Current Protein Peptide Science 15(3):225-31.

Drickamer K., (1989) Demonstration of carbohydrate-recognition activity in diverse proteins which share a common primary structure motif. Biochem Soc Trans (17): 13-15

Drummond R.A., and Brown G.D. (2011). The role of Dectin-1 in the host defence against fungal infections. Current Opinion in Microbiology (14): 392-399.

Kim D., Langmead B., Salzberg S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. Nature Methods (12): 357-360.

Kerrigan A.M., and Brown G.D. (2010) Syk-coupled C-type lectin receptors that mediate cellular activation via single tyrosine based activation motifs. Immunological Reviews (234): 335-352.

Lanier L.L., 2008. Up on the tightrope: natural killer cell activation and inhibition. Nature Immunology 9(5): 495-502.

Love M.I., Anders S., Huber W. (2014) DESeq2: Differential gene expression analysis based on the negative binomial distribution (Version 1.24.0) [Software]. Available from: <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>

Love M.I., Anders S., Huber W. (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology (15): 550.

Tennessen J.A., Théron A., Marine M., Yeh J-Y, Rognon A., Blouin M.S. (2015) Hyperdiverse Gene Cluster in Snail Host Conveys Resistance to Human Schistosome Parasites. PLoS Genet 11(3)