Asthma and Allergic Disease: Sequence Comparison and an Exploration of Mitotic Kinetics

John Komoll

Rice University, *Department of Biochemistry and Cell Biology*, 6100 Main St., Houston TX 77005

Introduction

Asthma and allergy are conditions characterized by immunological responses caused by a combination of both genetic and environmental factors. They share some immune response pathways with a few genes common to both diseases. However, the biological mechanisms for the development of each condition are not well understood, and neither condition has a cure (1).

Recent literature has suggested a link between the synchronicity of mitosis between cells and the level of asthmatic inflammation of these cells (2). This suggests some undiscovered connection between the kinetics of mitosis and immunological response from asthma and/or allergy.

In this paper, I examine four significant genes related to the immunological responses of both allergy and asthma with the hopes of finding categorical information or patterns about these conditions to contribute to the search for a more complete understanding of each of them. These four genes are: interleukin-33 (IL-33) and thymic stromal lymphopoietin (TSLP), which have central roles in the innate immune response pathways of both conditions (1), and Orosomucoid like 3 (ORMDL3) and Gasdermin-B (GSDML), which are associated with the development of childhood onset asthma (1).

First, I examine the nucleotide sequences of each gene in an attempt to elucidate sequence similarities that may be telling of some common genetic motif of these cytokine proteins. Second, I gather data available on the growth of cells over time with each of the genes of interest silenced during growth to try to find an effect on the kinetics of mitosis from each gene.

Methods

Acquiring Sequence Data for Genes of Interest

Accession numbers for each gene were obtained from the NCBI UniGene database (3). These accession numbers were then passed to MATLAB's "getgenbank" function to obtain the base pair sequences for each gene of interest from the NCBI GenBank database (4), as well as other relevant information for the gene.

Comparing Sequences with Smith-Waterman Sequence Alignment

Once gene sequences were obtained for each gene, they were each then compared to every other gene sequence using the Smith-Waterman algorithm. This algorithm pairs two sequences against a scoring matrix. The highest scoring alignment of sequences in the matrix defines the alignment path. The final alignment is the total path that yields the largest score based on the scoring matrix. This algorithm was implemented using MATLAB's "swalign" function between each pair of two gene sequences, using the standard "nuc44" scoring matrix. A score space representing the scores of each possible path was displayed by setting the "SHOWSCORE" parameter to True.

Comparing Sequences with BLAST

To ensure that no optimal alignment sequences were missed in analyzing the Smith-Waterman optimal alignment, and to examine the possibility of multiple aligning

sequence sections, the BLAST algorithm was used to compare each gene sequence with all other gene sequences studied. The BLAST algorithm makes a list of words (or short sequence subsets) from a given sequence and compares each word to a matrix of word scores. High scoring words in the target sequence are aligned with the word of the first sequence, and the local alignment is scored with the same word scoring matrix. All aligned sequences above a threshold score are returned as significant alignments. This was performed on the NCBI BLAST webpage (5), comparing two gene sequences when given their previously found accession numbers. Each gene sequence was also compared against the entire NCBI nucleotide database (6) using MATLAB's "blastncbi" function to make a request of the BLAST service, and MATLAB's "getblast" function to access the results of this request.

Collecting Cell Growth Data For Processing in MATLAB

Videos of cell growth with knockdown of a wide variety of genes are available from the Mitocheck database (7). Five such videos of cell growth were downloaded for each silenced gene of interest (ORMDL3, GSDML, IL-33, TSLP), as well as five videos of a control cell growth. The silenced gene selected for the control group was RHO, the gene encoding Rhodopsin. Rhodopsin is a light-sensitive protein involved in vision with no known link to mitosis (8). These videos were then each converted into a VideoReader object in MATLAB. The "readFrame" method of the VideoReader object was used to move to the next frame for iteration of each video's set of frames.

Counting Cells Across Video Files

A function named "cell_counter" was defined in MATLAB to count the number of cells in each frame of a video file following the format of the video data on the Mitocheck

Database. Within “cell_counter,” a vector “num_cells” is created to store the number of cells in each frame of the inputted video. Then, each iterated frame is smoothed with a Gaussian image filter (radius = 5 pixels, sigma = 5 pixels). The background of the image is generated with MATLAB’s “imopen” function, using a disk structuring element of radius 100, and is subsequently subtracted from the smoothed frame. Then, a mask is applied to the iterated frame, showing all pixel values above 0.08 as a logical 1. MATLAB’s “regionprops” function is then used to obtain the areas of each continuous white object in the mask, and the number of areas is stored as the number of cells in the frame in “num_cells.” After all frames are iterated through, the function outputs the “num_cells” vector, giving a cell count across the input video.

Fitting Data to Cell Growth Models

Once cell counts across growth videos are obtained, they must be fit to a cell growth model in order to obtain growth rate constants to compare. Two possible growth models are the following:

Exponential Model: $\frac{dx}{dt} = c \cdot x$ $\qquad x = a \cdot e^{c \cdot t}$

Logistic Model: $\frac{dx}{dt} = c \cdot x \cdot (1 - \frac{x}{b})$ $\qquad x = \frac{b}{1 + a \cdot e^{-c \cdot t}}$

The exponential growth model assumes that the growth of cells in the video is never restricted by the surrounding density of cells or availability of growth substrate. The logistic growth model denies this assumption, and assumes the cell count is approaching some upper asymptote.

The cell count data from the first video of each gene’s silencing was fitted to both models using MATLAB’s “fit” function, and the model that better characterized the observed data for all samples was chosen. Then, this model was applied to all sample

video data with MATLAB's "fit" function, and the growth rate constant *c* was obtained from each fit for comparison.

ANOVA Statistical Analysis of Cell Growth Rates

Once rate constants *c* for each of the five sample videos for each of the silenced genes was obtained, they were grouped based on which gene was silenced, and a 1-way ANOVA test was performed to compare the average means of each group. This was performed with MATLAB's "anova1" function, which also generates a box plot representing the different groups of rate constant data. The resulting p-value was used to determine whether any means between the silenced genes and the control group were significantly different.

Results

Smith-Waterman score readouts (Figure 1) show many alignment paths with similar low scores, indicating no good alignments between any combination of two gene sequences of interest.

BLAST between all combinations of two gene sequences of interest yielded no significant similarities in gene sequences.

BLAST of each sequence against the NCBI nucleotide database reveals more similarities between species of the same gene than between genes in the same species. None of the first 30 hits

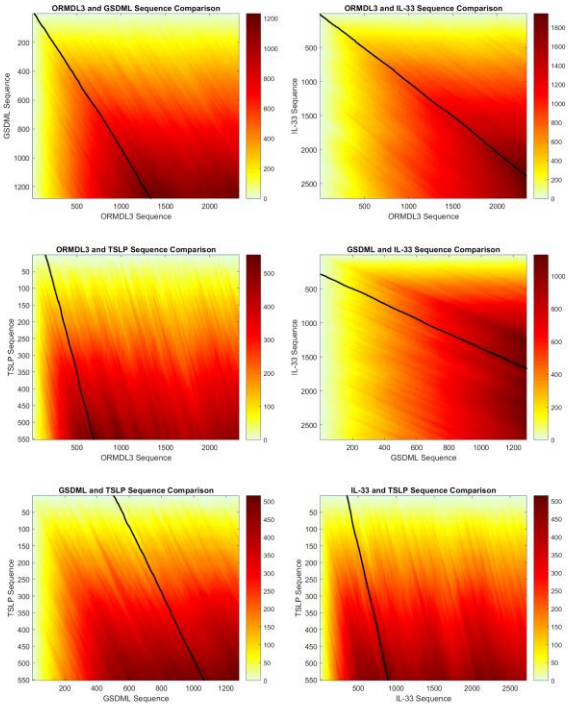

Figure 1. Smith-Waterman alignments of gene sequences. Scale bars represent score magnitudes.

for any of the genes in question include any of the other genes related to asthma and allergy.

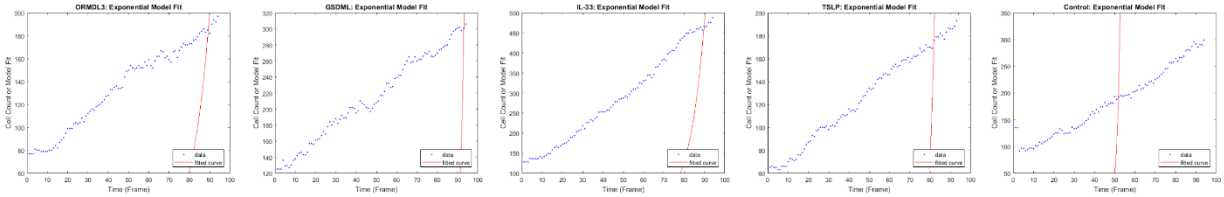

Figure 2. Exponential model fit plotted against experimental data for cell counts. The fit is visibly poor for all samples tested.

The logistic growth model fits cell growth data much better than the exponential growth model. Fitting the exponential model to one sample video from each testing group (Figure 2) reveals a poor characterization of cell count data. Fitting the logistic model to the same sample videos from each testing group (Figure 3) shows that this model accurately describes the change in cell counts over time. The average R^2 value for the logistic model fit was 0.988, across all 25 samples examined.

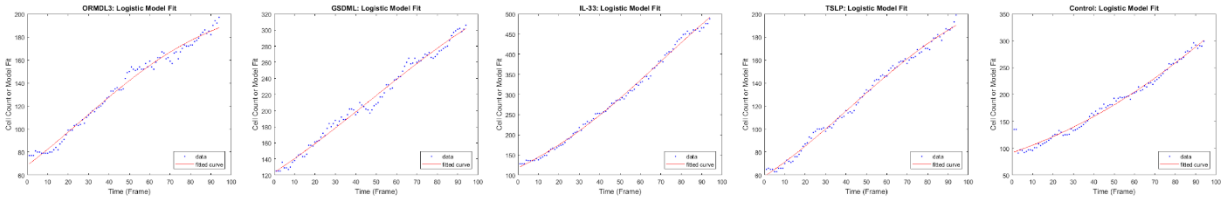

Figure 3. Logistic model fit plotted against experimental data for cell counts. The fit is visibly appropriate for all samples tested.

1-way ANOVA analysis of rate constants (n = 25, 5 categories) from the logistic regression model show no significant differences in growth rate between cells with any given silenced gene and a control group of cells in the same conditions (p = 0.845). Figures 4 shows the box plot displayed from MATLAB's "anova1" analysis of the rate constants.

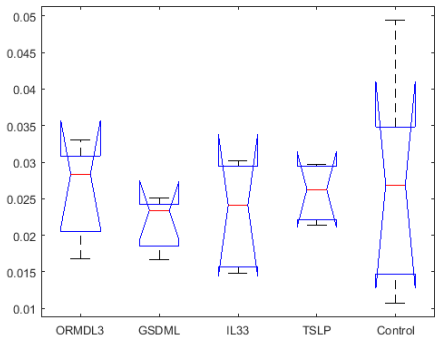

Figure 4. Box plot of logistic model rate constant data for all testing groups (ORMDL3, GSDML, IL33, TSLP, Control).

Discussion

In this study, four major genes encoding cytokines involved with the immunological pathways of asthma and allergy were examined. These genes were ORMDL3, GSDML, IL-33, and TSLP. Neither Smith-Waterman nor BLAST algorithms could find any sequence alignments between the four genes of interest. This indicates that the genes involved in asthmatic and allergic immune responses are varied in structure, and exhibit a high degree of biological complexity. Ultimately, I was unable to find any common genetic motifs between proteins involved in these conditions.

Analyzing BLAST results for each individual gene compared to the NCBI nucleotide database revealed that each gene had more alignment hits with the same genes of different species than with different human genes involved with allergy and asthma. This indicates that the individual biological pathways of these genes are highly conserved through evolution, and are common between many species.

The logistic model of cell growth was ultimately more successful in describing the cell counts computed from the Mitocheck video data of cell growth. In both the control and the experimental groups, cells were not able to grow in an uninhibited manner, but rather were limited to an upper bound. This indicates that the assumptions made in the exponential growth model were invalid. The environments of the cells, including their access to growth substrate and proximity to other cells, imposed a logarithmic behavior to their growth once their populations reached a relatively high density.

The p-value obtained from the 1-way ANOVA comparison of the five rate constants (1 from the control group, and 1 from each of the silenced genes examined in this paper) was much higher than 0.05, indicating that the computed rate constants were not even close to being statistically different from one another. This indicates that none of the genes related to asthma and allergy examined in this paper has a noticeable effect on the rate of mitosis when silenced. Taking this result into consideration, I fail to find any desired link between mitotic kinetics and the immunological responses of the allergic and asthmatic conditions.

In the future, it would be interesting to compare the ORMDL3, GSDML, IL-33, and TSLP gene sequences of patients with asthma or allergies to those without these conditions. Such a comparison could indicate if there are any key differences in the features of these proteins between healthy and affected patients that contribute to the existence or absence of this condition. A lack of difference in these gene sequences could indicate that the conditions are more dependent on the expression of the encoded proteins.

In addition, a future experiment in which the genes of interest are overexpressed, rather than silenced, would provide more insight into the effect of ORMDL3, GSDML, IL-33, and TSLP on mitotic kinetics. It is conceivable that one of these proteins might alter the rate of mitosis when present in high quantities, but the methods explored here would not reveal such a connection.

Reference

1. Ober, Carole, and Tsung-Chieh Yao. “The Genetics of Asthma and Allergic Disease: A 21st Century Perspective.” *Immunological reviews* 242.1 (2011): 10–30. *PMC*. Web. 7 Dec. 2017.

2. Benton, Angela S., et al. “Synchronizing Mitosis Reduces Intrinsic Inflammation In Asthmatic Airway Epithelium.” American Journal of Respiratory and Critical Care Medicine, vol. 183, 2011, doi:10.1164/ajrccm-conference.2011.183.1_meetingabstracts.a2814.

3. “Home - UniGene - NCBI.” National Center for Biotechnology Information, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/unigene/.

4. “GenBank Database.” National Center for Biotechnology Information, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/genbank/.

5. “Nucleotide BLAST: Align Two or More Sequences Using BLAST.” National Center for Biotechnology Information, U.S. National Library of Medicine, blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&LINK_LOC=align2seq.

6. “Nucleotide BLAST: Search Nucleotide Databases Using a Nucleotide Query.” National Center for Biotechnology Information, U.S. National Library of Medicine, blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome.

7. Neumann, Beate et al. “Phenotypic Profiling of the Human Genome by Time-Lapse Microscopy Reveals Cell Division Genes.” Nature 464.7289 (2010): 721–727. PMC. Web. 7 Dec. 2017.

8. Rowena, Matthews G, et al. “Tautomeric Forms of Metarhodopsin.” The Journal of General Physiology, vol. 47, 1963, pp. 215–240., www.ncbi.nlm.nih.gov/pmc/articles/PMC3108885/.