The Upstream Transcriptional Regulation of the COPD-linked *DSP* Gene

Lucas Brown

Senior Thesis

May 2018

Abstract:

Chronic obstructive pulmonary disorder (COPD) is a disease that affects millions of people around the world. There are many factors that can lead to the development of the disease, including certain genetic risk factors. In this study, the upstream regulation of the COPD-linked gene *DSP* is investigated in an attempt to understand how the gene feeds into the pathology of the disease. A possible association with the Yap1 transcription factor induced by changes in the stiffness of the matrix was explored. 16HBE cells were grown on plastic and collagen-coated plates that mimicked the extracellular matrix, and no increase in YAP1 expression was observed. Further knockdown experiments with YAP1 did not reveal a strong relationship to *DSP* expression. The rs2076295 SNP was also investigated, having been linked to lower expression of *DSP.* A reporter plasmid containing the SNP region upstream of the *DSP* promoter controlling expression of luciferase was constructed and used to confirm that the risk allele of the SNP region resulted in lower expression levels of *DSP* in 16HBE cells. The metagenomics tool DeepSEA was used to explore and screen new transcription factor candidates for further experimentation.

Introduction:

Chronic Obstructive Pulmonary Disorder (COPD) is a chronic inflammatory lung disease that affects 11 million Americans. Symptoms of the disease include breathing difficulty, mucus production, and persistent coughing. The two main contributors to COPD are emphysema, the destruction of the alveoli, and chronic bronchitis, the inflammation of the inner bronchial tubes of the lungs.[1] COPD has many identified contributing risk factors, including long-term exposure to lung irritants, smoking and age.[2]

While environmental factors contribute the most to the development of COPD, some genetic risk factors have also been identified. The alpha-1 antitrypsin deficiency genetic disorder is one such risk factor, and accounts for around 1% of COPD cases.[3] The gene encoding a serpin peptidase inhibitor, *SERPINA1*, contains mutations that render the resulting protein, alpha-1 antitrypsin, dysfunctional. One of the functions of this protein is to protect the lungs from neutrophil elastase, an enzyme that can damage connective tissue. The loss of function of alpha-1 trypsin leads to a decrease in tissue connectivity and early onset emphysema.[3] Other gene variants have been associated with COPD as well, such as alpha1-antichymotrypsin, vitamin D-binding proteins, and variants of the cystic fibrosis transmembrane regulator gene.[4] However, the exact mechanism or importance of these links is difficult to determine. The interconnectedness of the human genome and complexity of human physiology means that the signal of serious mutations can often be lost among the thousands of variants in the human genome.

Genome Wide Association Studies (GWAS) are a tool that has arisen in the past decade as a solution to this problem, and combine advances in computational modelling, statistics, and next-generation sequencing data. These studies compare the frequency of single nucleotide polymorphisms (SNP) in the genome between healthy members of a population and those with a trait of interest, such as a disease. Alleles that are found to be more present in diseased populations are said to be "risk alleles," and indicate a possible association. This method allows the identification of candidate genes and SNPs for *in vivo* experimental verification. According to the NHGRI-EBI Catalog of published genome-wide association studies, over 3,000 GWA studies have examined over 1.800 diseases and traits, with thousands of SNP associations being identified.[5]

The *DSP* gene, encoding desmoplakin, has implicated through GWAS studies several times in connection with pulmonary fibrosis, a leading symptom of COPD.[6,7] *DSP* encodes the desmoplakin protein, a key adhesion protein in the desmosomes that make up cell to cell junctions. Variants of *DSP* have been associated with idiopathic pulmonary fibrosis (IPF), the most common of the idiopathic interstitial pneumonias (IIP).[6] An increase in *DSP* expression was found to be increased in individuals with fibrotic IIP. GWAS studies have also found that a decrease in *DSP* expression has been correlated with an increase in the risk of developing COPD. Specifically, the SNP rs2076295, located in intron 5 of the *DSP* gene, was found to affect the expression of the gene.[6.7.9] The two alleles of the SNP are the T allele, the major allele and correlated with higher/normal expression of the *DSP* gene, and the G allele, the minor allele and putative "risk" allele associated with lower expression of *DSP*.[9] It has been hypothesized that this SNP could be located in a transcriptional regulatory region of the gene.

This study focused on the investigation of the transcriptional regulation of the *DSP* gene in relation to the pathology of COPD. Understanding the mechanism of regulation will aid in the understanding of how the expression level of *DSP* features into the development of pulmonary fibrosis. The importance of the rs2076295 SNP was investigated, as well as the role of other possible transcription factors. One such transcription factor is the Yes associated protein 1 (YAP1), the effector transcription factor of the Hippo signaling pathway. The Hippo pathway is involved with development, growth, repair, and homeostasis.[10] However, the Yap1 protein has been shown to be localized to the nucleus to regulate transcription in response to changes in the stiffness of the surrounding matrix outside the cell, independent of the activation of the rest of the Hippo pathway. Interestingly, *DSP* has also been shown to increase in expression in response to matrix-induced mechanical stress.[17] This similarity in function as mechanosensors could indicate a link between the two genes.

16HBE cells, an immortalized human bronchial epithelial line, was used as the model system for this investigation. While immortalized cell lines aren't perfect models, they are easier to culture and give more consistent data than primary cells. Additionally, many previous studies of *DSP* have been done in 16HBE cells as they have excellent expression of the gene, allowing results found in this study to be verified by other research groups that study the same topic.

Materials and Methods:

Cloning of the rs2076295 SNP Reporter Plasmid

The reporter plasmid pGL3 was selected as the vector for the rs2076295 expression assay. The reporter plasmid was designed using the SnapGene software, with DNA sequences for the pGL3 vector being found on Addgene, and the sequences for the *DSP* gene and surrounding regions being taken from the GRCh38/hg38 genome found in the USCS Genome Browser. Primers were designed to amplify ~500bp of the rs2076295 SNP region, and install a kpn1 cutsite at the 5' end and a Sac1 cutsite at the 3' end. Another set of primers was designed in order to amplify ~1,000 bases of the *DSP* promoter region, installing an xho1 site at the 5' end and a Hind111 site at the 3' end. 16HBE cell genomic DNA was used as the template for the amplification of the fragments, as the cell line is heterozygous for both alleles of the rs2076295 SNP.

New England Biolab's HiFi Polymerase and restriction enzymes were used to perform the amplification of fragments and the restriction enzyme digests according to the NEB's written protocols. Fragment size was confirmed using a 1% agarose gel with eithidium bromide as the staining agent. Qiagen PCR Purification kits were used to purify each reaction. NEB T4 DNA Ligase was used to ligate the fragments together according to company protocol, and the fragments were transformed into Invitrogen DH5 alpha cells and plated onto LB agar+100mg/L ampicillin plates. Colonies were picked the next day, grown up into 5mL cultures in liquid LB media supplemented with 100mg/L ampicillin, and screened using a restriction enzyme digest

with the same restriction enzymes used in the initial cloning. Plasmids shown to have inserts were sent to Macrogen for Sanger sequencing to confirm correct sequence. The sequencing used the standard primer RVprimer3 that spans the 5' side of the multiple cloning site of the pGL3 vector. Colonies found to have the correct sequenced were grown in 100mL LB cultures and midi-prepped using the Qiagen Midi-Prep kit and protocol.

Cell Culture, Transfection, and siRNA Knockdown

16HBE cells were cultured at 37°C using EMEM media from Life Technologies supplemented with 10% Fetal Bovine Serum and 1%PenStrep from Life Technologies. Cells were cultured in 10 cm plates and split 3 times a week. Media was aspirated from the wells, which were washed with 5mL of Phosphate Buffered Saline (PBS). Cells were detached using 3 mL of 0.05% trypsin.

Plasmid transfection was performed in a 24-well plate for the luciferase reporter assay for the rs207295 SNP. 16HBE cells were split into a 24-well plate the day before the transfection so that the resulting plate was about 50-60% confluent (100,000 cells per well). The media was changed the day of transfection, and the plate was incubated during the following steps. Reporter plasmid DNA and the TK-Renella control plasmid DNA were mixed with the reagents of the Lipofection3000 kit from Invitrogen according to their company protocol. 150ng of TK-Renella and 1,500ng of reporter plasmid DNA were used per well. Mixed reagents were incubated for 15 minutes, and 40 µL of reagent/DNA mixture were added to each well drop-wise. The plate was then incubated at 37°C for 6 hours, after which the media was changed. Cells were collected for analysis 48 hours after the initial transfection.

Transfection of siRNA for the purpose of gene knock downs was performed in 6-well plates. The siRNA was designed using IDT software and ordered from IDT as a batch. 16HBE cells were split the day before the transfection so that the confluency was about 50% on the day of the transfection (500,000 cells). 4 µg of siRNA were mixed with the reagents of the RNAimax kit form Invitrogen according to protocol. Opti-mem media from Life Technologies was used as the delivery matrix. After 6 hours the media was changed.

Rs2076295 SNP Luciferase Reporter Assay

16HBE cells that had been transfected with reporter plasmids were lysed using 100 µL of passive lysis buffer from the Dual-Luciferase Reporter Assay from Promega. Cells were rocked gently for 15 minutes, and 40 µL of lysate was transferred to a white-backed plate reader plate. The lysate was analyzed using a Wallac 1420 Luminometer, with 100 µL of both the luciferase and the TK-Renilla substrate being added into each well, and a luminescence reading done.

Cell Harvesting and Real Time Quantitative PCR Analysis

16HBE cells were harvested using the RNeasy kit from Qiagen, following the company's protocols. The RNA concentration of samples was assessed using a Nanodrop spectrometer. 1 µg of the RNA was converted into cDNA using the High Capacity cDNA Reverse Transcription Kit from Thermo Fisher Scientific, following the company protocol. Quanitative PCR Analysis was carried out using TaqMan Fast Advanced Master Mix according to the company protocol. A comparative cycle threshold analysis was run, and the ppiA gene was used as the reference gene. The ppiA gene is involved in protein folding, and so is a good reference gene as it is evenly expressed in most cell types.[12] qPCR probes were designed and ordered from IDT.

Rat Tail Collagen Coated Plates

Collagen plates were created using 6-well plates. Rat tail collagen (Type 1) was diluted using cold 0.02N acetic acid to a final concentration of 50 µg/mL. 1 mL of this solution was added to each plate, and the plate was incubated at room temperature overnight. The next day, plates were washed with 1 mL of PBS, and were ready for use.

Metagenome Analysis

Metagenome analysis of the rs2076295 SNP region was done using the DeepSEA suite of online investigation tools developed by Princeton University.

Results:

Yap1 Knockdown and Collagen Matrix Effects

Yap1 is the effector protein of the Hippo signaling pathway, but has also been shown to operate independently of this pathway in epithelial cells in the regulation of the cell's response to physical stress. In order to investigate the potential connection of the Yap1 transcription factor to the regulation of the expression of desmoplakin, a knockdown experiment was designed. It was theorized that if YAP1 was a transcription factor of the *DSP* gene, then knocking down YAP1 would lead to a decrease in *DSP* expression. IDT software was used to design a pool of siRNA oligoes that would bind to Yap1 transcripts and target them for degradation. Quantitative PCR probes were order for *yap1, DSP,* and *ppiA*, a reference gene.

Additionally, it was decided to measure the effects of changes in the matrix stiffness of the cells when YAP1 was knocked down. If changes in the stiffness of the matrix surrounding the cell was linked to desmoplakin, then a change of expression level would be expected in the *DSP* gene. YAP1 levels would also be assessed, as one would expect an increase with a stiffer matrix. In order to create a softer matrix, rat tail collagen was used to coat plates. This provides a surface more akin to the extra-cellular matrix of the lung, and has less of a physical impact on the cultured cells than growing them on the standard hard plastic plates.

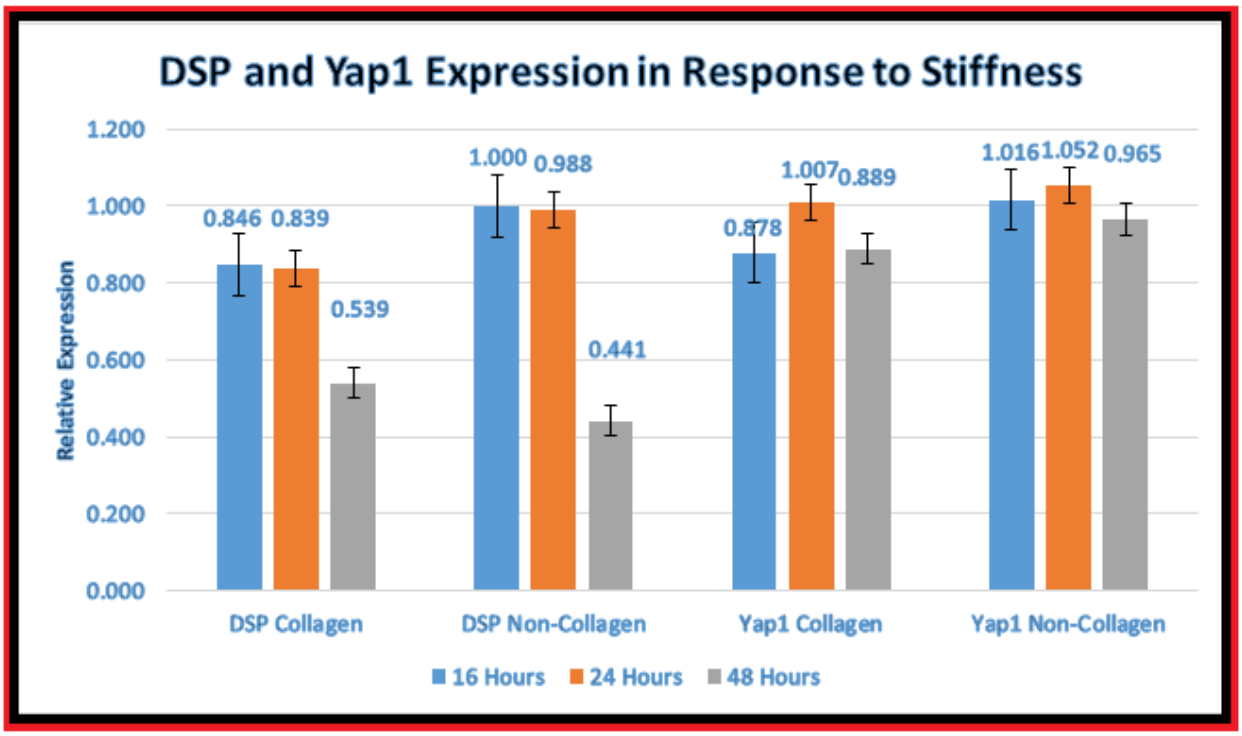


Figure 1: Expression of DSP and YAP1 in response to a change in matrix stiffness. 16HBE cells were grown on collagen plates, then split into collagen and non-collagen plates. These cells were harvested at 16, 24, and 48 hours, and analyzed using comparative CT quantitative PCR. RNA was harvested from the cells, and then converted into cDNA. All values are normalized to the 16 hour non-collagen DSP level. Error bars represent the standard deviation of repeats in the quantitative PCR. The ppiA gene was used as the reference gene.

In order to first observe the effects of the change in matrix stiffness on the expression levels of both *Yap1* and *DSP*, a growth experiment was designed (Figure 1). Cells were grown on collagen coated plates, then split into collagen and non-collagen wells. RNA was harvested from the cells at 16, 24, and 48 hours after the split, and the RNA was converted into DNA, which was analyzed using the qPCR probes. The *ppiA* gene was used as the reference gene.

There seems to be an overall higher expression of *DSP* in cells grown on collagen coated plates when compared to those grown on hard plastic. *DSP* expression increased about 20%

when grown on non-collagen plates (0.846 to 1.00). This may indicate upregulation in an attempt to provide more structure to the cell by joining it to its neighbors more tightly by forming a larger desmosome structure. This structure may not be as necessary when the cells have a harder plastic surface to grow on, resulting in lower expression of *DSP*. This same result corroborates an observation made in a previous paper, where *DSP* was upregulated in response to stiffer conditions.[17] Additionally, it was observed in both conditions that the expression of *DSP* vastly decreased at the 48 hours mark. This observation could be explained by cell crowding in the wells. While equal numbers of cells were split into each well at the beginning of the experiment, by 48 hours that confluency of the well was nearing 90%. The cells may have downregulated desmoplakin expression in an attempt to allow greater migration away from higher density clumps of cells.

There seems to be little change in the expression of YAP1 either across time points, or across matrix conditions. This is interesting considering previous works observing a connection between YAP1 and changes in stiffness. This could be because the change in matrix condition isn't drastic enough to elicit a response from the cell. However, as noted in Das *et al.*[10], YAP1 is shown to localize to the nucleus in response to mechano-transduction, not be upregulated transcriptionally. The paper goes on to find that the phosphorylation of YAP1 regulates its localization to the nucleus, which is a post-translational modification. Since the experiment only measured control of YAP1 at the transcriptional level, it is understandable that no difference in the level of transcription was observed. A western blot using an antibody that discriminates between phosphorylated and non-phosphorylated YAP1 could measure the difference between the activated and non-activated states across a number of matrix conditions.

Overall, the experiment demonstrated a link between the regulation of *DSP* and the change in matrix conditions. It was decided to continue assessing the effects of the different matrix conditions on expression levels throughout the knockdown of YAP1 mRNA.

The possibility of a link between *YAP1* and *DSP* was investigated by the use of siRNA knockdown of *YAP1* mRNA. If YAP1 acts as the transcription factor for *DSP*, there should be a significant decrease in *DSP* with a knockdown of *YAP1*. 16HBE cells were grown on collagen plates, and then transfected with either *YAP1* knockdown siRNA or control siRNA. The cells were then split into collagen and non-collagen wells and left to grow. RNA was then harvested from all the cells, and analyzed by reverse transcriptase qPCR. Given the data from Figure 1, it seemed likely that *DSP* expression would be lowered in both the knockdown and the control transfection wells, while no change in *YAP1* expression between matrix conditions was expected.

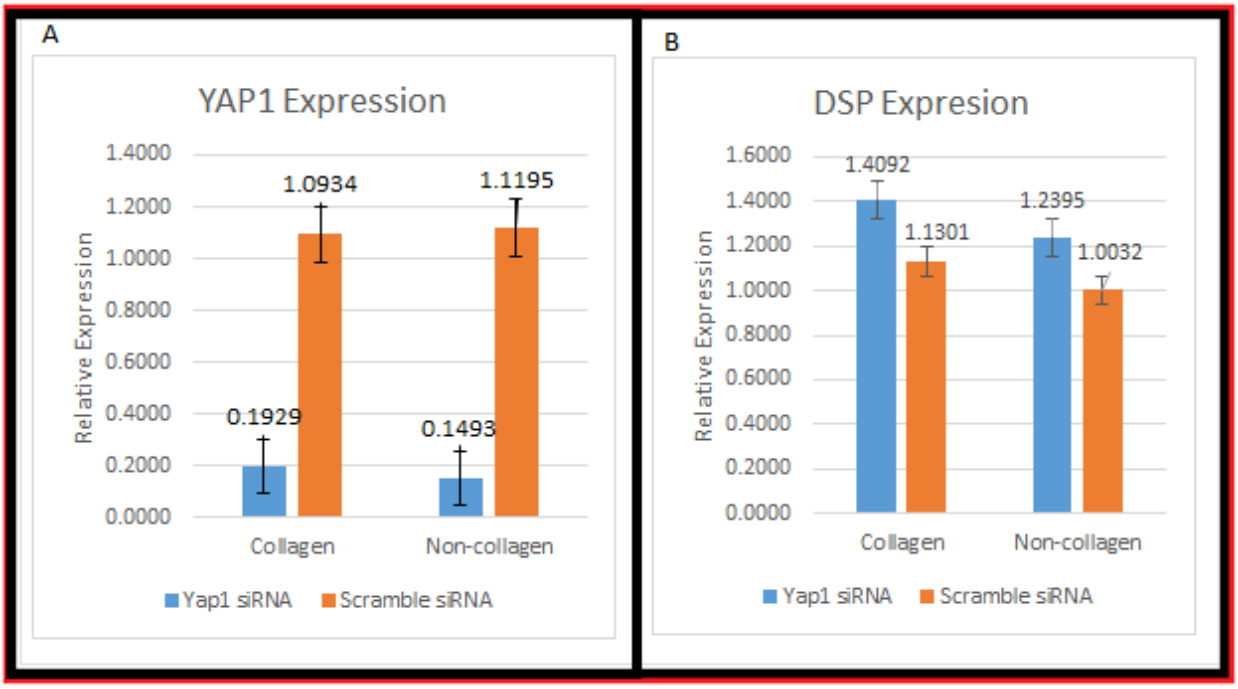


Figure 2: The expression levels of Yap1 (A) and DSP (B) mRNA in 16HBE cells. Cells were grown on collagen plates and transfected with either YAP1-targeted siRNA (KD) or control siRNA (scramble). The cells were then split into collagen and non-collagen plates, and grown for 2 days. Afterwards RNA was harvested and analyzed via RT-qPCR. Error bars represent standard deviations of three separate transfections. Levels of mRNA were normalized to the non-collagen scramble DSP mRNA level for each individual experiment, then averaged together..

The levels of *YAP1* mRNA in the cell for both the collagen and non-collagen KD wells were much lower than the scramble wells, implying an efficient and successful knockdown of the *YAP1* gene (Figure 2a). Additionally, the lower levels of desmoplakin expression see in the collagen growth experiment (Figure 1) were not observed in this experiment. The collagen *DSP* levels seemed overall a little higher than the non-collagen levels, but these were still within the error for the experiment, and so were not significantly different.

DSP expression seemed to increase with the knockdown of *YAP1*, not decrease as predicted (1.409 during the knockdown to 1.1301 for the control). The *Yap1* siRNA *DSP* levels

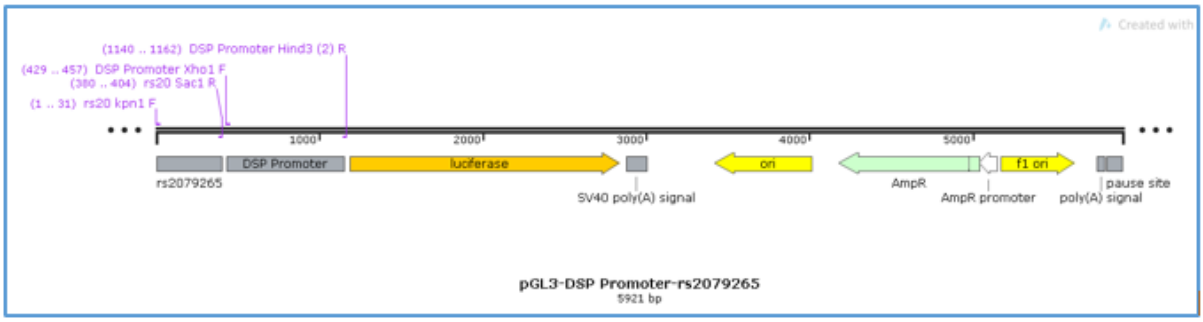
for collagen were about 27% higher than the scramble, and about 23% higher than the scramble in the non-collagen coated wells (Figure 2b). This would seemingly invalidate the hypothesis that *YAP1* is the transcription factor for the *dsp* gene, as a significant decrease in *DSP* expression would be expected from the knockdown of its transcription factor.

Given both the lack of effect on *DSP* expression that the collagen matrix seemed to have, and the lack of effect the knock down of *YAP1* seemed to have, it is likely the *YAP1* and *DSP* aren't related, at least not directly. Therefore, a new direction was needed.

Rs2076295 Cloning and Luciferase Assay

The rs2079265 SNP has been linked to the expression levels of the *DSP* gene in metagenomics analyses and GWAS studies. The T allele has been associated with higher expression of *DSP*, while the G allele is associated with lower expression. This would suggest that the rs2079265 SNP is located in the binding region of a transcription factor, or some other regulatory element of the gene. In order to investigate this phenomenon, the effect of the SNP on *DSP* expression levels needed to be validated *in vivo*. An experiment was designed using the reporter plasmid pGL3, containing firefly luciferase. Since the pGL3 vector has no promoter, ~750bp of the *DSP* promoter was cloned upstream of luciferase gene. Then about 400bp surrounding the SNP region of the rs2079265 region was cloned upstream of the *DSP* gene (Figure 3a). This amount was selected to ensure that the full sequence of any directly related elements to the SNP was included, without Primers were designed to amplify promoter and SNP fragments from 16HBE genomic DNA (Figure 3b).

A



B

Primer Name	Primer Sequence (5'-3')	Fragment Amplified (Base Pairs)
DSP Promoter Hind3 (2) R	cgt**CAAGCTTG**TCGGCGGGCAATCGG	DSP Promoter Region (727)
DSP Promoter Xho1 F	aagc**CTCGAG**GCGTGGTAGGTAGCGTTTCCAG	DSP Promoter Region (727)
Rs20 Sac1 R	cca**CCTAGG**AAAGGATTCACCAATG	Rs2079265 SNP Region (404)
Rs20 Kpn1 F	cgg**GGTACC**GCTGGCAGCTGGACAAAATCAAAGC	Rs2079265 SNP Region (404)

Figure 3: A. Schematic of the rs2079265 reporter plasmid. Cloning was designed using the SnapGene software. Ellipses indicate continuity of the construct, as it is a plasmid. Primer locations are shown on the plasmid map, and the restriction enzyme in the primer name indicates a cutsite that was engineered into the 5' end of each primer to generate fragments with the desired restriction enzyme sites. The DSP promoter was cloned into the plasmid first, as the SNP region has a HindIII site located within it. Cloning with the HindIII enzyme and a plasmid that contained the SNP region would result in partial fragment loss of the SNP region. B. Primers used for the cloning of the rs2079265 expression reporter plasmid. Numbers in the primer name indicate the iteration of that primer designed. The lower case letters in the primer sequence indicate the "runway" installed for the restriction enzyme to ensure an efficient digestion step, and the bold letters are the restriction enzyme recognition sequences.

The *DSP* promoter was cloned first as the HindIII restriction enzyme would be used, and another HindIII site existed in the SNP region. This meant that if the SNP region were cloned first into the multiple cloning site, the second digestion reaction to insert the *DSP* fragment would destroy the SNP region. The DSP Promoter Hind3 (2) R primer and the DSP Promoter Xho1 F primer were used to amplify the *DSP* promoter from genomic DNA extracted from 16HBE cells. The results of the PCR was analyzed via gel electrophoresis (Figure 4a). As can be seen from the figure, the fragment was of the expected size. This fragment was then purified. The fragment and the pGL3 vector were both digested with the HindIII and Xho1 enzymes, and the pGL3 reaction was dephosphorylated using recombinant shrimp alkaline phosphatase. The pGL3 reaction was run on a gel and gel purified to minimize false positives from empty vectors in the colonies. The two fragments were ligated together in a 10:1 insert to vector molar ratio and transformed into competent DH5α *E. coli*. Colonies were then screened for the insert using the same two enzymes as digested with, and the resulting digestion reactions analyzed by gel electrophoresis (Figure 4b). The plasmids that had an insert were confirmed by sequencing with the standard primer RVprimer3 (Figure 4c).

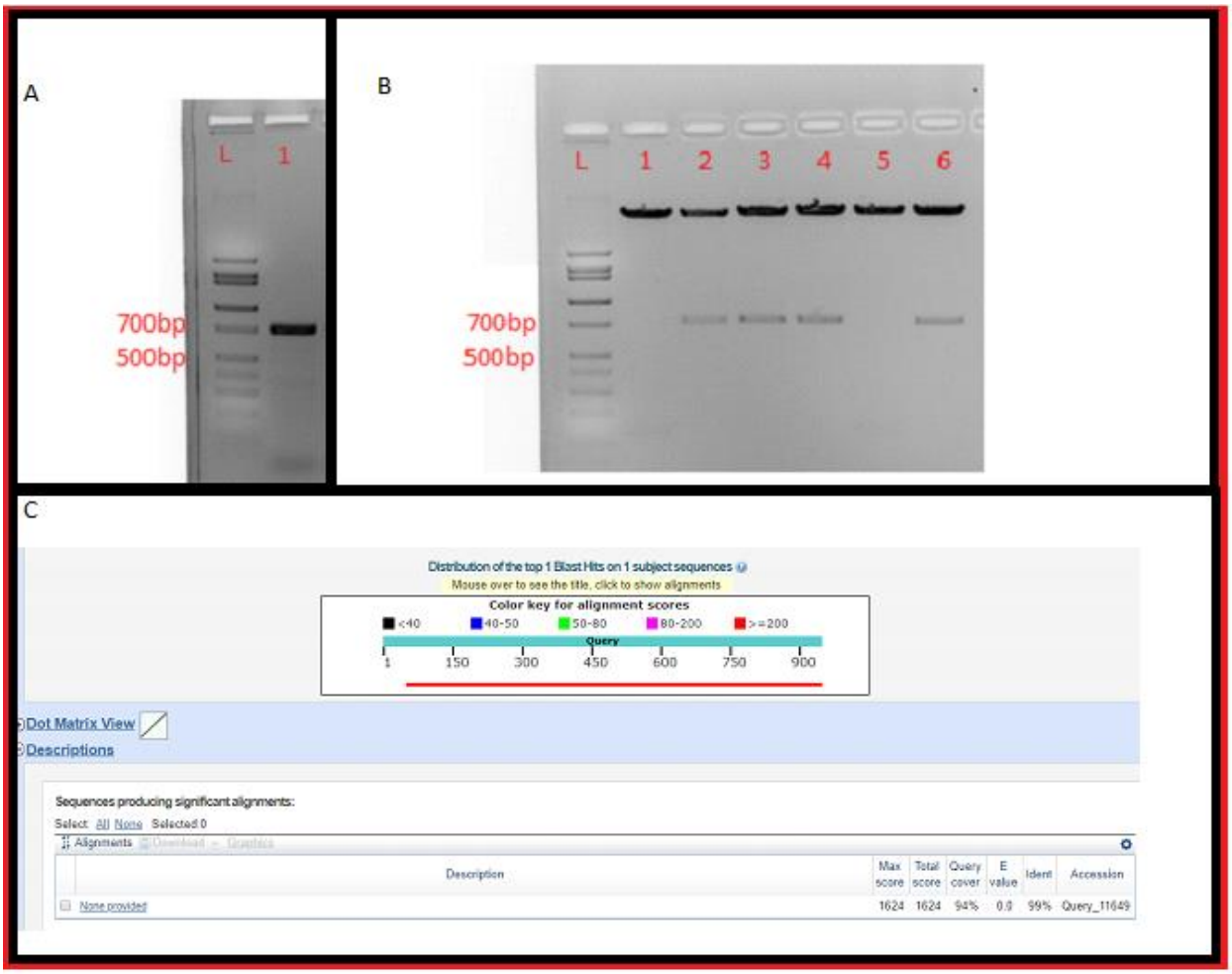


Figure 4: DSP Promtoer-pGL3 cloning results. A. DSP promoter PCR results. The primers DSP Promoter Hind3 (2) R and DSP Promoter Xho1 F were used to amplify 700bp of the promoter region by PCR, and the amplified fragment was purified and run on a gel to confirm it size (Lane 1). B. Digestion screening results. After the ligation of the DSP promoter fragment to the pGL3 vector, the colonies were screened by digestion. The miniprepped –plasmids from lanes 2, 3, 4, and 5 were sequenced to confirm insert identity. C. NCBI BLAST sequence alignment results for one cloned plasmid. The Sanger sequencing result was aligned with the DSP promoter region and the downstream luciferase gene, and was found to have sufficient homology, indicating that both sequences for the promoter and the luciferase gene were present.

The rs2076295 SNP region was cloned into the *DSP* Promoter-pGL3 construct after the first cloning. Since the 16HBE cell line is heterozygous for both alleles of the SNP, the rs20 Sac1 R and rs20 kpn1 F primers were used to amplify the SNP regions for both alleles indiscriminately. The resulting product was visualized on a gel (Figure 5a). Afterwards, the product was purified, and both the SNP region fragment and the newly created *DSP* Promoter-pGL3 construct were digested with the Sac1 and Kpn1 restriction enzymes in separate reactions. The digested vector was dephosphorylated, run on a gel, and purified. The two fragments were ligated together in a 10:1 insert to vector molar ratio and transformed into competent DH5α *E. coli*. The resulting colonies were screened as before, using digestion reactions (Figure 5b). Promising colonies were mini-prepped and sent for sequencing, where one plasmid for each allele was selected (Figure 5c).

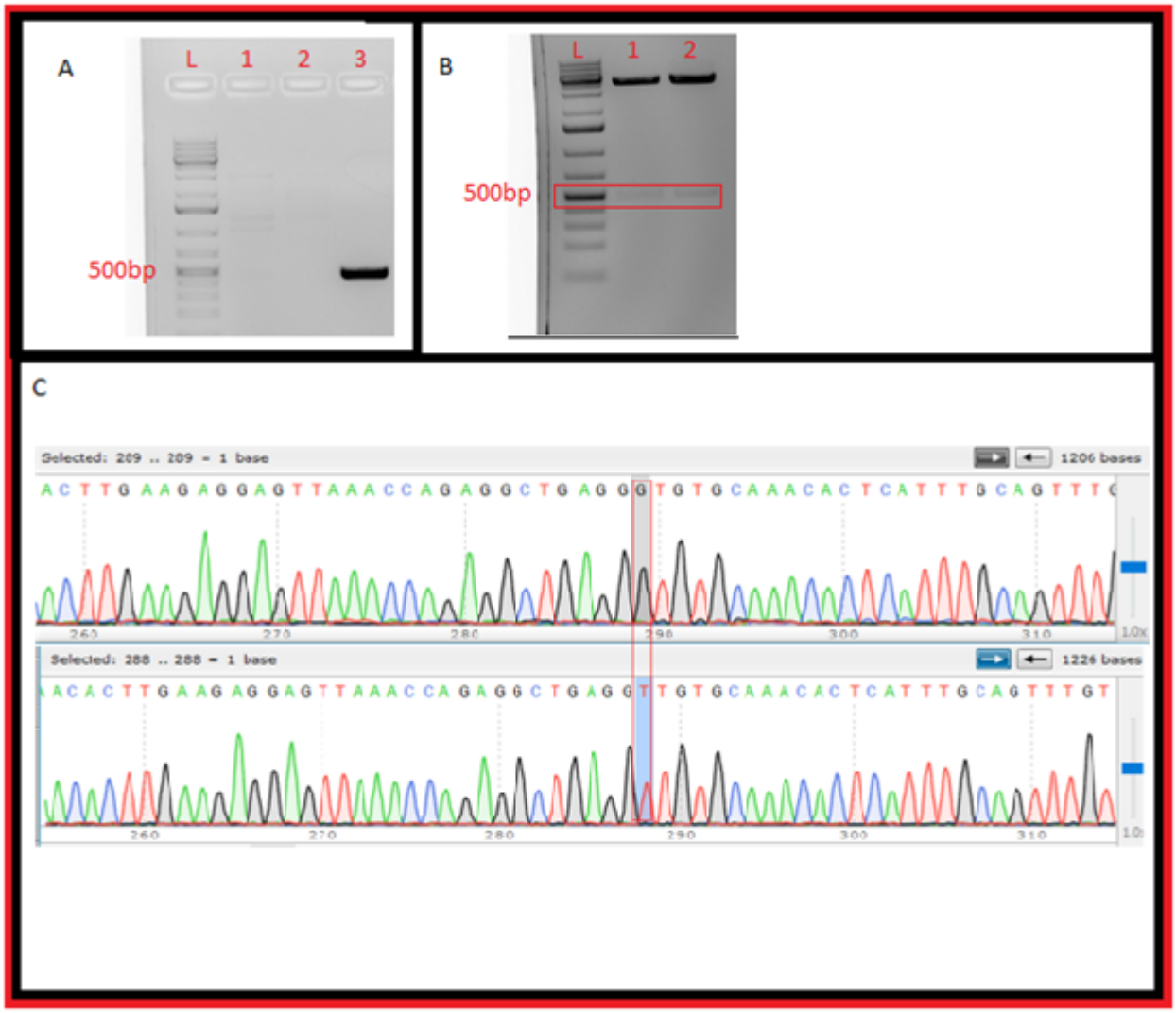


Figure 5: rs2076295 reporter plasmid cloning results. A. PCR results for the amplification of the rs2076295 SNP region. The rs20 Sac1 R and rs20 kpn1 F primers were used as different annealing temperatures to amplify the region, and the reaction run on lane 3 was the only successful amplification. B. Digestion screen results for the insertion of the rs2076295 SNP region inside the DSP Promoter-pGL3 vector. The red box indicates the bands caused by the digested insert. Comparison of two sequencing results from two different colonies. The bottom sequence contains the T allele, and the top result contains the G allele. The red box indicates the misalignment caused by the SNP.

Once the reporter constructs were successfully cloned and confirmed, an expression assay was designed using a dual-reporter structure. The luciferase encoded in the rs2076295-*DSP* Promoter-pGL3 reporter plasmid is firefly luciferase, which is a 61kDa protein isolated from the beetle *Photinus pyralis* (eastern firefly) that requires oxygen, ATP and magnesium to produce light.[13] This light produced is in the 550-570nm range. By contrast, luciferase isolated from *Renilla reniformis* (sea pansy) is 31 kDa and requires only coelenterazine and oxygen as cofactors. The light produced from this protein is found in the 480nm range, completely separate from the firefly luciferase. The difference and substrates and luminescence ranges offer a clever opportunity for a role as an internal control for TK renilla luciferase. Plasmids containing promoters of interest controlling expression of firefly luciferase are co-transfected into cell lines along with plasmids containing TK renilla under the control of a low constitutive promoter. When the cells are grown, lysed, and assayed for activity, the reagents for the firefly luciferase are added first, and a measurement of all wells is taken. Then, a mixture of inhibitor for firefly luciferase and reagents for TK renilla are added to the wells, and a second measurement is taken at 480nm. The first measurement is standardized by the second measurement, as it is assumed that the vector uptake rate is the same for both plasmids during transfection. This process allows an accurate and standardized measurement of transcription activity.

The T-allele and G-allele rs2076295 SNP reporter plasmids were co-transfected with TK renilla containing plasmids into 16HBE cells along with the original *DSP* Promoter-pGL3 plasmid with no SNP region and the empty pGL3 vector. The cells were left to grow for 48 hours and then were harvested and analyzed using the dual-reporter assay.

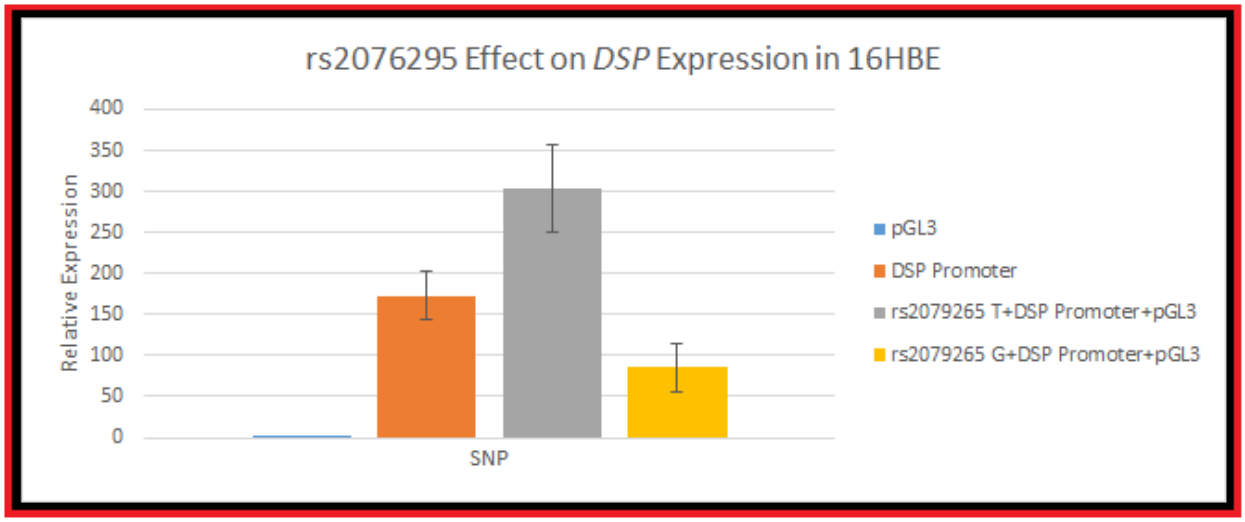


Figure 6: Dual Firefly/Tk renilla luciferase reporter assay for the rs2076295 SNP. 16HBE cells were co-transfected with a luciferase reporter plasmid and a TK renilla plasmid, and grown for 48 hours, after which they were harvested. The luminescence readings were normalized first to the TK renilla reading for each well, then to the pGL3 vector reading. The error bars represent the standard deviation from three separate transfections.

As can be seen from Figure 6, the rs2076295 SNP has a significant effect on the expression of the *DSP* promoter. The pGL3 basic vector has a comparatively low expression level, since the vector itself doesn't have a promoter, so there is no place for the RNA polymerase to initiate transcription. The addition of the rs2076295 SNP region containing the G-allele resulted in a nearly 50% decrease in expression over just the *DSP* promoter (80 to 160). The T-allele by contrast had nearly double the expression levels as the *DSP* promoter by itself (160 to 300). This change in expression levels could indicate that the SNP happens in a regulatory region of the *DSP* gene. A change of a single base pair could potentially effect the binding of a transcription factor such as an activator or repressor.

Interestingly, the T and the G alleles both resulted in higher and lower expression of luciferase, rather than any one allele having the same expression as the lone *DSP* promoter.

One would expect that if the SNP region were a binding site for an activator, the inclusion of the G-allele SNP region would result in the same levels of expression as the promoter, but that is not what is observed. Instead, it appears that the region as an activator affect with one allele, and a repressor effect with another. This observation could just be an artifact of the assay. The SNP region is located in the intronic region of the gene, so having the method of regulation so close could alter the degree of regulation in unpredictable ways. Additionally, the fold change in expression seen in this experiment probably isn't accurate to the wild-type expression of the gene, as the SNP region is looked at in isolation, and the effects of any other regulatory elements in the promoter such as methylation or additional activator sequences are lost. The experiment also doesn't take into account any post-transcriptional or translational regulation that could be happening in the cell.

Despite these limitations, it was determined that there was suitable evidence that the rs2076295 SNP region played a role in the regulation of desmoplakin expression. In lieu of testing likely transcription factors through more knock down experiments, it was decided to use a metagenomics approach to identify likely transcription factor candidates.

Metagenome Analysis

Metagenomic analysis is a revolutionary tool made possible by advances in both computing power and sequencing techniques. Here, the online tool suite from DeepSEA was used to assess transcription factor binding to the SNP region. DeepSEA was originally developed at Princeton University, and uses machine learning trained on data sets from the ENCODE

project.[14] Each transcription factor or other regulatory element in the data base has a range of sequences that have been reported to be associated with it. The algorithm can then predict the likelihood of these regulatory elements appearing in or binding to a sequence of DNA. The program can also focus on a single transcription factor and identify which base pairs in a sequence if changed would contribute positively or negatively to the probability of the factor binding. This is the saturated mutagenesis feature, and could be used to potentially identify transcription factors that are negatively impacted by changing the T to a G at the location of the rs2076295 SNP.

Two transcription factors, PU.1 and FoxA1, were identified as possible transcription factors that bind to the rs2076295 SNP. Initially, a 1,000bp region of chromosome 6 was analyzed, centered on the rs2076295 SNP. Transcription factors and other regulatory features were ranked in probability across a number of cell-line studies (Figure 7a). The transcription factor PU.1 was found to have the most probable binding of any transcription factor (0.403). This is interesting, as PU.1 has previously been mentioned in the literature as having a common binding motif in the rs2076295 region.[7] However the cell line from which the data comes from, GM12891, is a lymphoblastic cell line, which could indicate that this finding had little relevance to lung cells. Different cell lines sometimes have variable expression levels of the same gene. In order to assess the binding likelihoods of transcription factors in lung relevant cell lines, the list of chromatin features was sorted by cell line. The A549 is an alveolar basal epithelial cell line, which is much more physiologically relevant. As can be seen from the sorted probabilities (Figure 7b) the transcription factor FoxA1 was found to have the highest binding probability of

any of the lung cell lines (0.005). This isn't a high probability, but it is the highest observed for the A549 cell lines.

In order to further assess the validity of these two candidates, saturated mutagenesis was modelled using the DeepSEA program (Figure 7c). FoxA1 was calculated to have a moderate decrease in binding probability with a change from T to G, as happens in the risk allele of the rs2076296 SNP. The base pairs immediately surrounding the SNP are calculated to have similar decreases, suggesting that the area is a binding motif for the FoxA1 protein. However, no such decrease is seen with the saturated mutagenesis of PU.1, indicating that there would be no change in binding probability if any one base pair in the sequence were changed. While this is persuasive evidence to expel PU.1 as a potential transcription factor, there are some limitations to DeepSEA. DeepSEA's algorithm is based off of machine learning and observed data, not biochemistry and protein modelling, leaving it prone to overlook binding motifs just because they haven't been previously recorded.

In order to further confirm transcription factor candidacy, the expression levels of FoxA1, PU.1 and five other transcription factors with high binding probabilities were looked at in 16HBE through RNAseq data (Figure 7d). The raw data is of yet unpublished, but the sequencing data was available in lab for this analysis. The RNA transcripts of the seven transcription factors were analyzed across two wildtype 16HBE lines and three 16HBE-derived cell lines manufactured in the lab using CRISPR/Cas9: 1E6, 2B8, 1D7, 1D8, and C3. These cells lines are wildtype, wildtype, TT homozygous for the rs2076295 SNP, have insertions/deletions at the DSP location, and have a complete knockout of the DSP gene. Both FoxA1 and PU.1 were found to have minimal expression in any of these lines, while the other transcription factors

have variable levels. GR and EP300 were observed to have the highest expression levels, signaling them out as promising candidates.

Overall, this method of genomic analyses was useful in identifying and eliminating potential transcription factors for this particular SNP. However, given the degree of uncertainty surrounding the program's reliability, it was decided to use a combination of techniques to select a candidate transcription factor for knockdown experiments. ATAC-seq will be used to inform further experiments on this matter, as explained in the discussion.

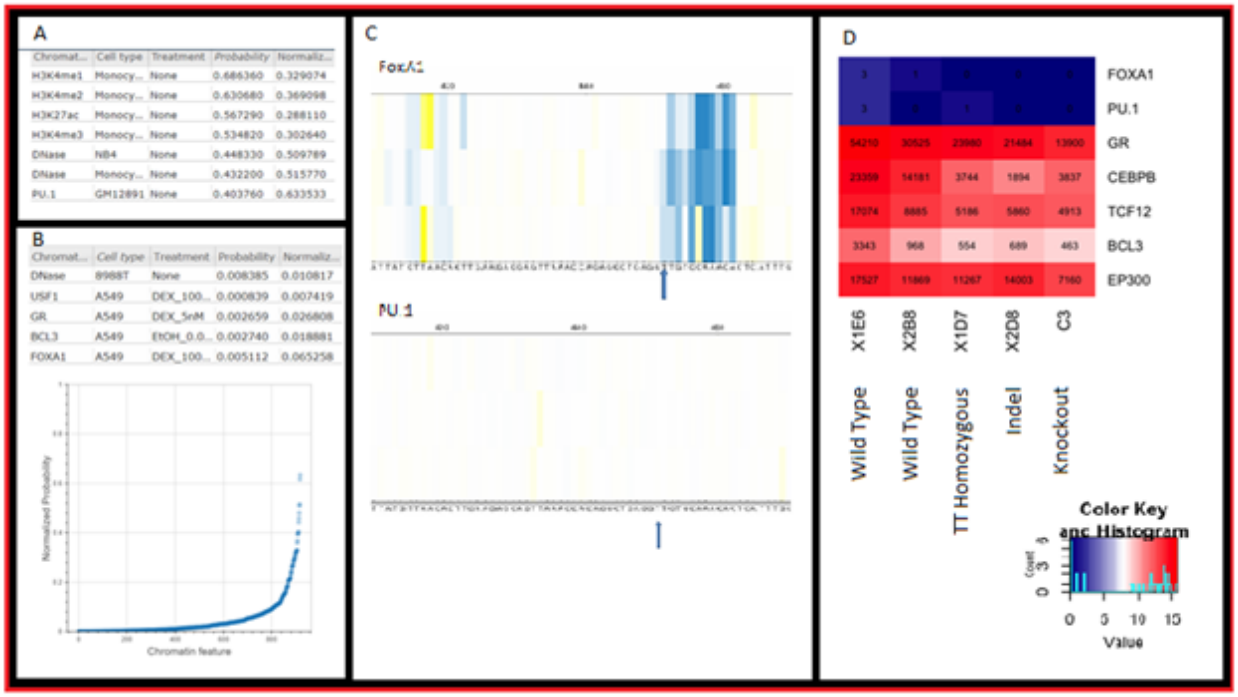


Figure 7: Metagenomic analysis of the rs2076295 SNP Region. A and B. Calculated chromatin features of 1,000 base pairs surrounding the rs2076295 SNP region sorted by probability across all cell lines (A) and sorted by lung cell lines (B). The higher the probability, the more likely that the feature is present. The treatment column refers to what stress the cell was induced with when the data set was collected, and the cell type column indicates the cell line. C. Simulated saturated mutagenesis of FoxA1 and PU.1. The rs2076295 SNP is indicated by the blue arrow. Blue indicates a more negative binding probability, while yellow indicates a more positive one. The base pair order from top to bottom is A, G, C, T, and A. If the base was a T, then the order would be A on top, followed by G and then C. D. RNA-seq analysis of expression levels of candidate transcription factors in16HBE lines. The non-wildtype lines were previously engineered using CRISPR/Cas9.

Discussion:

The investigation into the role of the Yap1 protein in the regulation of *DSP* was fruitless, but was a worthwhile area to question. The role of Yap1 in the cell's response to mechanical stress and stiffness had not yet been elucidated, and it possible that regulation of a structural protein such as desmoplakin could be one of the outputs of its localization to the nucleus. The lack of response in transcription levels of *DSP* in response to the knockdown of *Yap1* would seem to suggest that regulation of *DSP* happens independently of whatever signaling pathway Yap1 partakes in during the mechanical stress response. Additionally, it is also possible the changes in matrix stiffness were not enough to induce a response in Yap1. A concentration of 50µg/mL collagen was used in each plate as this was the standard in other experiments in the lab. It's possible that by varying concentrations, a change in expression could be seen. If work were to continue in this vein, another growth experiment could be performed in which 16HBE cells were grown on a gradient of concentrations, and collected at different time points to be analyzed for changes in *DSP* and *Yap1* transcription via qPCR. However, this line of investigation was halted in favor of more fruitful lines in the interests of having results for this study.

While the data found in GWAS studies already suggested that the risk allele of the rs2076295 SNP resulted in lower expression levels of *DSP*, but this study features the first time the SNP was validated in an *in vivo* system. The use of the reporter plasmid and the dual-reporter luciferase assay clearly shows that when the reporter plasmid features the risk allele of the SNP, expression is significantly lower than when the plasmid contains only the *DSP* promoter region, or when the plasmid contains the T allele of the SNP. It is important to note here that the actual fold difference in expression is probably not accurate. Previous

experiments in the lab on CRISPR engineered TT homozygous and GG engineered homozygous 16HBE cell lines have shown no significant increase or decrease in *DSP* expression (unpublished data). This would suggest that while the region containing the rs2076295 SNP does indeed play a role in the transcriptional regulation of *DSP*, but there are probably other regulatory elements. The dramatic change in expression observed in the reporter assay is probably an artifact of the SNP acting in isolation so much closer to the promoter than usual. Depending on the method of transcriptional regulation, the distance from the promoter could impact the degree of activation.

This could also explain the inhibitory effect that the G allele SNP region had on expression levels. One would expect a SNP that lowers the binding probability of a transcription factor to have the same activity as just the promoter, but we see here a significant decrease in *DSP* expression. It is possible that the mutation would result in a transcription inhibitor binding to the region instead. A better assay to further explore this effect might be to clone the *DSP* gene exons and intron 5 into the reporter plasmid and transfect into a *DSP* knockout CRISPR line. This way, the effect of the SNP on the actual gene can be measured in isolation from other regulatory elements via qPCR.

The metagenomics tool DeepSEA was useful in identifying potential transcription factors for the rs2076295 SNP region, but was not without its limitations. The two transcription factors initially identified, PU.1 and FoxA1, showed great promise but were shown to be virtually untranscribed in the 16HBE cell line via RNA-seq data. It should be noted that 16HBE cells are not perfect models, as there are some differences between them and alveolar epithelial cells, where the majority of the pathogenicity of COPD takes place. There is a definite possibility Even

other transcription factors that were identified as having the highest probabilities of binding and were shown to be transcribed in the 16HBE RNA-seq data had low probabilities relative to other chromatin features. As a result, it was decided that instead of trying these low probability transcription factors one by one in knockout experiments, it would be faster and less expensive to analyze the region via ATAC-seq.

ATAC-seq is a DNA foot printing technique used to assess chromatin accessibility first described in 2013.[15] While many genome-wide accessibility techniques such as MNase-seq, FAIRE-seq, and DNAse-seq already exist, ATAC-seq has two advantages. First, the technique is faster to perform, taking under 3 hours until a sample is ready for next-generation sequencing. Second, ATAC-seq can be used with as little as 50,000 cells, which is 1,000 fold less than any other technique.[16] ATAC-seq uses a hyperactive Tn5 transposase which cuts exposed DNA and ligates adapters to the strands. Primers specific to these adapters are then used to amplify the fragments before the sample is sent to next-generation sequencing. The sequences are mapped onto a reference genome, and chromatin visibility is then assessed. This technique can be used to identify high transcription areas, and can be resolve binding sites of transcription factors. Using the technique to identify transcription factor motifs binding in and around the *DSP* promoter area is cheaper and faster than attempting to test candidates one by one. The technique will also identify non-transctiption factor regulatory elements, such as methylation and DNAse hypersensitivity areas.

The scope of this study was to study the upstream transcriptional regulation of the COPD-linked *DSP* gene, and several advances were made in this direction. *DSP* expression was found to be independent of the Yap1 protein, and only slightly affected by changes in matrix

stiffness. The rs2076295 SNP was validated *in vivo* for the first time, and was shown to have an activator-like effect on *DSP* expression. This observation will provide an area to focus on when the ATAC-seq analysis is performed on 16HBE cells, and will hopefully lead to the full elucidation of the upstream regulation of *DSP* and its role in the pathology of COPD.

References:

1. "COPD - Symptoms and Causes." Mayo Clinic. Accessed March 19, 2018. http://www.mayoclinic.org/diseases-conditions/copd/symptoms-causes/syc-20353679.
2. "COPD | National Heart, Lung, and Blood Institute (NHLBI)." Accessed March 19, 2018. https://www.nhlbi.nih.gov/health-topics/copd.
3. Silverman, Edwin K., and Robert A. Sandhaus. "Alpha1-Antitrypsin Deficiency." *New England Journal of Medicine* 360, no. 26 (June 25, 2009): 2749–57. https://doi.org/10.1056/NEJMcp0900449.
4. Sandford, A. J., T. D. Weir, and P. D. Paré. "Genetic Risk Factors for Chronic Obstructive Pulmonary Disease." *The European Respiratory Journal* 10, no. 6 (June 1997): 1380–91.
5. "GWAS Catalog." Accessed March 19, 2018. http://www.ebi.ac.uk/gwas/downloads.
6. Mathai, Susan K., Brent S. Pedersen, Keith Smith, Pamela Russell, Marvin I. Schwarz, Kevin K. Brown, Mark P. Steele, et al. "Desmoplakin Variants Are Associated with Idiopathic Pulmonary Fibrosis." *American Journal of Respiratory and Critical Care Medicine* 193, no. 10 (May 15, 2016): 1151–60. https://doi.org/10.1164/rccm.201509-1863OC.
7. Hobbs, Brian D., Kim de Jong, Maxime Lamontagne, Yohan Bossé, Nick Shrine, María Soler Artigas, Louise V. Wain, et al. "Genetic Loci Associated with Chronic Obstructive Pulmonary Disease Overlap with Loci for Lung Function and Pulmonary Fibrosis." *Nature Genetics* 49, no. 3 (March 2017): 426–32. https://doi.org/10.1038/ng.3752.
8. "DSP Gene - GeneCards | DESP Protein | DESP Antibody." Accessed March 21, 2018. http://www.genecards.org/cgi-bin/carddisp.pl?gene=DSP.

9. “Rs2076295 - SNPedia.” Accessed March 21, 2018. https://www.snpedia.com/index.php/Rs2076295.
10. Chen, Wei, John M. Brehm, Ani Manichaikul, Michael H. Cho, Nadia Boutaoui, Qi Yan, Kristin M. Burkart, et al. “A Genome-Wide Association Study of Chronic Obstructive Pulmonary Disease in Hispanics.” *Annals of the American Thoracic Society* 12, no. 3 (March 2015): 340–48. https://doi.org/10.1513/AnnalsATS.201408-380OC.
11. Das, Arupratan, Robert S. Fischer, Duojia Pan, and Clare M. Waterman. “YAP Nuclear Localization in the Absence of Cell-Cell Contact Is Mediated by a Filamentous Actin-Dependent, Myosin II- and Phospho-YAP-Independent Pathway during Extracellular Matrix Mechanosensing.” *Journal of Biological Chemistry* 291, no. 12 (March 18, 2016): 6096–6110. https://doi.org/10.1074/jbc.M115.708313.
12. Obchoei S, Wongkhan S, Wongkham C, Li M, Yao Q, Chen C (Nov 2009). "Cyclophilin A: potential functions and therapeutic target for human cancer". *Medical Science Monitor*. **15** (11): RA221–32.
13. “What Are Some of the Differences between Renilla Luciferase and Firefly Luciferase?”*Promega*, Promega, www.promega.com/resources/pubhub/enotes/what-are-some-of-the-differences-between-renilla-luciferase-and-firefly-luciferase/.
14. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*. 2015;12(10):931-934. doi:10.1038/nmeth.3547.
15. Buenrostro, Jason D; Giresi, Paul G; Zaba, Lisa C; Chang, Howard Y; Greenleaf, William J (6 October 2013). "Transposition of native chromatin for fast and sensitive epigenomic

profiling of open chromatin, DNA-binding proteins and nucleosome position". *Nature Methods*. **10** (12): 1213–1218. doi:10.1038/nmeth.2688. PMC 3959825 . PMID 24097267.

16. Buenrostro, Jason D.; Wu, Beijing; Chang, Howard Y.; Greenleaf, William J. (January 2015). "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". *Current Protocols in Molecular Biology*: 21.29.1–21.29.9. doi:10.1002/0471142727.mb2129s109.
17. Qu, Jing, et al. “Desmoplakin (DSP), a GWAS-Identified Genetic Risk Allele of IPF, Is a Matrix Stiffness-Regulated Mechanosensitive Gene.” *ATS Journal*, 15 May 2016, pp. 333–342.