1. Task One – Information Discovery in Databases

1.1 Human Disease and Description

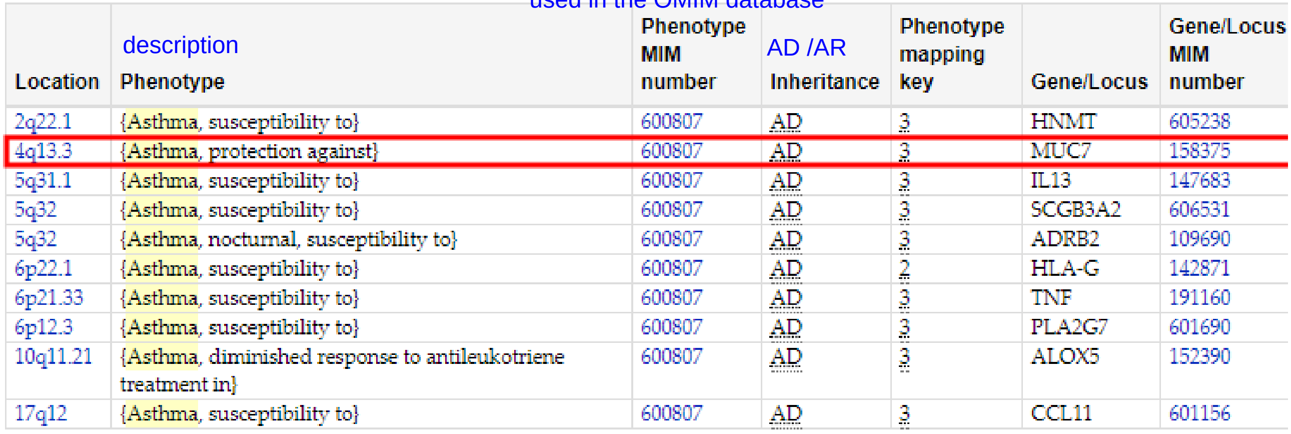
The most common chronic disease affecting children and young adults is bronchial asthma. It is a complex genetic disorder that is largely attributed to the interactions among many genes and between these genes and the environment. Asthma-related traits include clinical symptoms of asthma, such as coughing, wheezing, and dyspnea; bronchial hyperresponsiveness (BHR) as assessed by various clinical tests [1, 2, 3].

1.2 Related Gene Information for Asthma

The table below shows the known phenotype-gene relationships for Asthma. The MIM number is the numbering system used in the OMIM database. 1----- and 6----- denote autosomal loci or phenotypes. AD denotes autosomal dominant. In an AD disease, inheriting the abnormal gene from a single parent means there is a chance you can get the disease. This is the counterpart to autosomal recessive (AR) genes, where both parents must possess a copy of the gene in order for the disease to develop. Phenotype is a description of emergent characteristics or traits, it arises from the expression of genes and environmental factors. A locus is the position of a gene on a chromosome.

position of a gene
on a chromosome

used in the OMIM database



Location	description Phenotype	Phenotype MIM number	AD /AR Inheritance	Phenotype mapping key	Gene/Locus	Gene/Locus MIM number
2q22.1	{Asthma, susceptibility to}	600807	AD	3	HNMT	605238
4q13.3	{Asthma, protection against}	600807	AD	3	MUC7	158375
5q31.1	{Asthma, susceptibility to}	600807	AD	3	IL13	147683
5q32	{Asthma, susceptibility to}	600807	AD	3	SCGB3A2	606531
5q32	{Asthma, nocturnal, susceptibility to}	600807	AD	3	ADRB2	109690
6p22.1	{Asthma, susceptibility to}	600807	AD	2	HLA-G	142871
6p21.33	{Asthma, susceptibility to}	600807	AD	3	TNF	191160
6p12.3	{Asthma, susceptibility to}	600807	AD	3	PLA2G7	601690
10q11.21	{Asthma, diminished response to antileukotriene treatment in}	600807	AD	3	ALOX5	152390
17q12	{Asthma, susceptibility to}	600807	AD	3	CCL11	601156

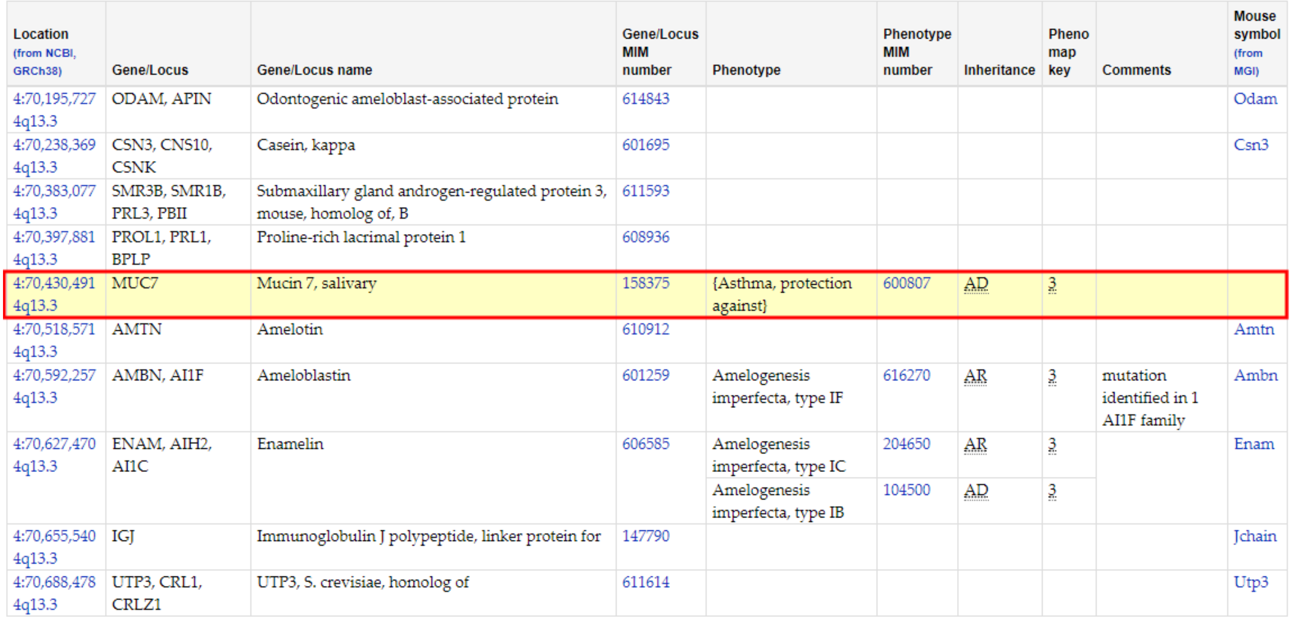
https://omim.org/entry/600807?search=asthma&highlight=asthma

Figure 1: Phenotype-gene relationships for Asthma. Highlighted in red is MUC7.

The gene I have selected for further analysis is MUC7. MUC7 is an AD gene that is potentially protective against asthma [4]. The MUC7 gene encodes a small salivary mucin, which is theorised to be protective against asthma by promoting the clearance of bacteria in the oral cavity and to aid in mastication, speech, and swallowing [5].

1.3 Genomic Context Information for MUC7

Figure 2. provides contextual information for MUC7. The locus for MUC7 is 4q13.3, which means that it is located on the 4th chromosome, it lies on the long arm (q), in region 1, band 3, and sub-band 3. None of the nearby genes in the table below have phenotypes related to Asthma, in fact, several of them (AMTN, AMBN, ENAM) are related to teeth formation. MUC7 does not have a corresponding mouse symbol. ?

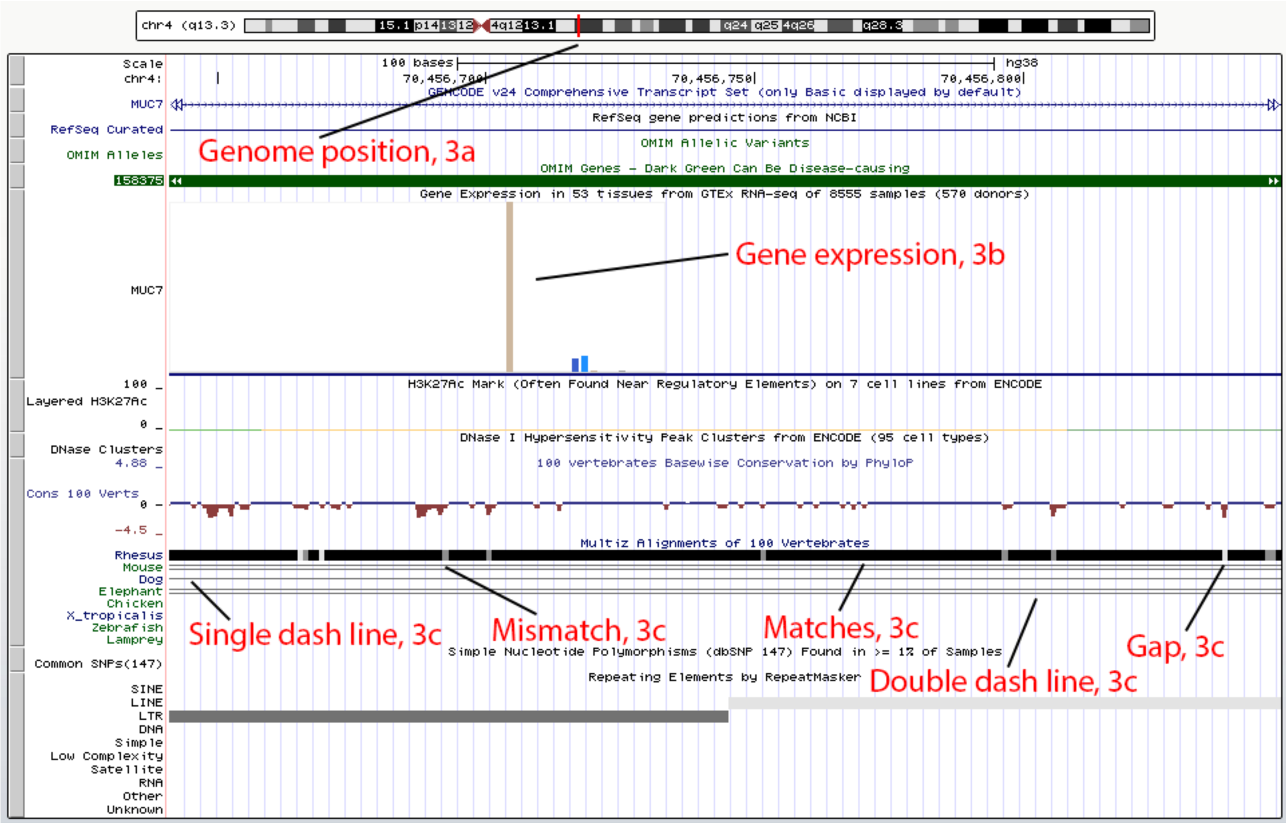


Location (from NCBI, GRCh38)	Gene/Locus	Gene/Locus name	Gene/Locus MIM number	Phenotype	Phenotype MIM number	Inheritance	Pheno map key	Comments	Mouse symbol (from MGI)
4:70,195,727 4q13.3	ODAM, APIN	Odontogenic ameloblast-associated protein	614843						Odam
4:70,238,369 4q13.3	CSN3, CNS10, CSNK	Casein, kappa	601695						Csn3
4:70,383,077 4q13.3	SMR3B, SMR1B, PRL3, PBII	Submaxillary gland androgen-regulated protein 3, mouse, homolog of, B	611593						
4:70,397,881 4q13.3	PROL1, PRL1, BPLP	Proline-rich lacrimal protein 1	608936						
4:70,430,491 4q13.3	MUC7	Mucin 7, salivary	158375	{Asthma, protection against}	600807	AD	3		
4:70,518,571 4q13.3	AMTN	Amelotin	610912						Amtn
4:70,592,257 4q13.3	AMBN, AI1F	Ameloblastin	601259	Amelogenesis imperfecta, type IF	616270	AR	3	mutation identified in 1 AI1F family	Ambn
4:70,627,470 4q13.3	ENAM, AIH2, AI1C	Enamelin	606585	Amelogenesis imperfecta, type IC Amelogenesis imperfecta, type IB	204650 104500	AR AD	3 3		Enam
4:70,655,540 4q13.3	IGJ	Immunoglobulin J polypeptide, linker protein for	147790						Jchain
4:70,688,478 4q13.3	UTP3, CRL1, CRLZ1	UTP3, S. crevisiae, homolog of	611614						Utp3

https://omim.org/geneMap/4/261?start=-3&limit=10&highlight=261

Figure 2: Local gene information for MUC7, the gene is highlighted in red.

Figure 3. shows the position (3a) of MUC7 in the genome and various annotated information. The gene expression of MUC7 across 53 tissues (3b), with the only notable expression being in minor salivary gland tissue. The Multiz alignments (3c) show alignments against various vertebrate species.



chr4 (q13.3)
15.1 p14 13 12 q12 13.1 q24 q25 4q26 q28.3
Scale
chr4:
100 bases
70,456,700
70,456,750
70,456,800
hg38
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
MUC7
RefSeq gene predictions from NCBI
RefSeq Curated
OMIM Allelic Variants
OMIM Alleles
OMIM Genes - Dark Green Can Be Disease-causing
158375
Gene Expression in 53 tissues from GTEx RNA-seq of 8555 samples (570 donors)
MUC7
100 _
Layered H3K27Ac
0 _
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
4.88 _
Cons 100 Verts
0 -
-4.5 _
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Common SNPs(147)
Simple Nucleotide Polymorphisms (dbSNP 147) Found in >= 1% of Samples
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
Genome position, 3a
Gene expression, 3b
Single dash line, 3c
Mismatch, 3c
Matches, 3c
Gap, 3c
Double dash line, 3c

Figure 3: Genomic context, expression, and alignment data for MUC7

The black regions indicate matches between the sequences, the grey bars mismatches, and the white bars gaps. Single dashed lines show no bases between the aligned species, double dashed lines show that the aligned species has one or more unalignable bases in the gap region.

https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr4%3A70430492%2D70482997&hgsid=892221727_w0sUoKFhL9ThlWxDXj6oT2ViC0aa

1.4 Protein Sequences for MUC7

FASTA is a text-based format for representing either nucleotide or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. Each letter codes for a single amino acid or nucleotide. For example, M is code for Methionine and T is code for Threonine.

1.4.1 FASTA protein sequence for Human MUC7

>sp|Q8TAX7|MUC7_HUMAN Mucin-7 OS=Homo sapiens GN=MUC7 PE=1 SV=2
MKTLPLFVCICALSACFSFSEGRERDHELRHRRHHHQSPKSHFELPHYPGLLAHQKPFIR
KSYKCLHKRCRPKLPPSPNNPPKFPNPHQPPKHPDKNSSVVNPTLVATTQIPSVTFPSAS
TKITTLPNVTFLPQNATTISSRENVNTSSSVATLAPVNSPAPQDTTAAPPTPSATTPAPP
SSSAPPETTAAPPTPSATTQAPPSSSAPPETTAAPPTPPATTPAPPSSSAPPETTAAPPT
PSATTPAPLSSSAPPETTAVPPTPSATTLDPSSASAPPETTAAPPTPSATTPAPPSSPAP
QETTAAPITTPNSSPTTLAPDTSETSAAPTHQTTTSVTTQTTTTKQPTSAPGQNKISRFL
LYMKNLLNRIIDDMVEQ

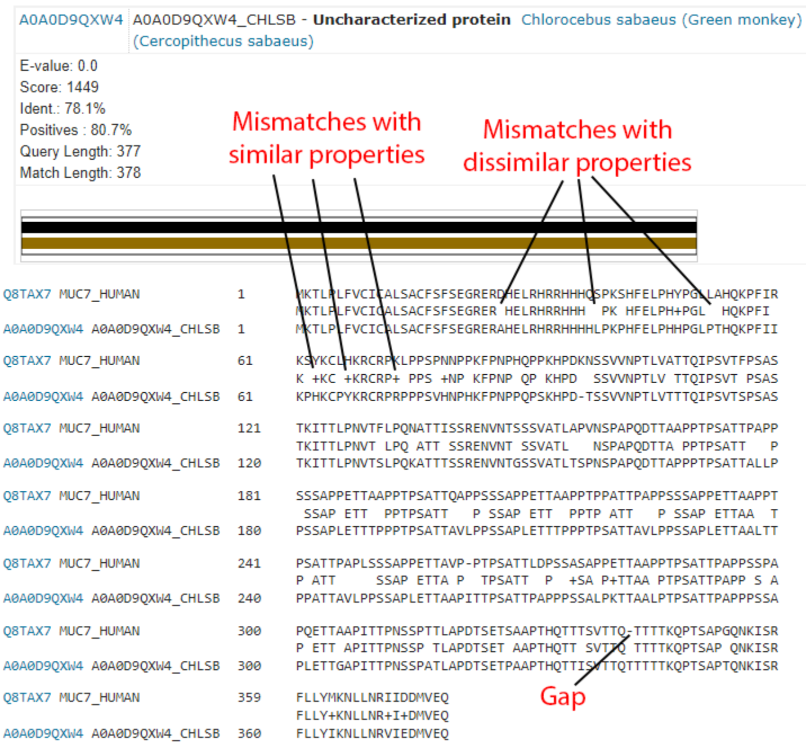
1.4.2 FASTA protein sequence for Green Monkey MUC7

>tr|A0A0D9QXW4|A0A0D9QXW4_CHLSB Uncharacterized protein OS=Chlorocebus sabaeus GN=MUC7 PE=4 SV=1
MKTLPLFVCICALSACFSFSEGRERAHELRHRRHHHHLPKPHFELPHHPGLPTHQKPFII
KPHKCPYKRCRPRPPPSVHNPHKFPNPPQPSKHPDTSSVVNPTLVTTTQIPSVTSPSAST
KITTLPNVTSLPQKATTTSSRENVNTGSSVATLTSPNSPAPQDTTAPPPTPSATTALLPP
SSAPLETTTPPPTPSATTAVLPPSSAPLETTTPPPTPSATTAVLPPSSAPLETTAALTTP
PATTAVLPPSSAPLETTAAPITTPSATTPAPPPSSALPKTTAALPTPSATTPAPPPSSAP
LETTGAPITTPNSSPATLAPDTSETPAAPTHQTTISVTTQTTTTTKQPTSAPTQNKISRF
LLYIKNLLNRVIEDMVEQ

2. Task Two – Alignment Algorithms for Protein Sequences

2.1 BLAST Sequence Alignment

BLAST (basic local alignment search tool) is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA sequences. Unlike the Smith-Waterman algorithm, BLAST will not always find optimal sequence alignments. However, BLAST still achieves a high degree of accuracy at magnitudes higher speeds. BLAST does this by essentially scanning the examined sequences for high scoring sequence pairs of relatively high complexity and using these pairs as ‘seeds’ to begin alignment between the sequences. The assumption is that these pairs are more likely to exist in high scoring alignments. Figure 4. shows the alignment between Humans and Green Monkeys for the FASTA protein sequences in 1.4 using the BLAST algorithm.



A0A0D9QXW4 A0A0D9QXW4_CHLSB - **Uncharacterized protein** Chlorocebus sabaeus (Green monkey) (Cercopithecus sabaeus)

E-value: 0.0
Score: 1449
Ident.: 78.1%
Positives : 80.7%
Query Length: 377
Match Length: 378

```
Q8TAX7 MUC7_HUMAN            1     MKTLPLFVCICALSACFSFSEGRERDHELRHRRHHHQSPKSHFELPHYPGLLAHQKPFIR
                                   MKTLPLFVCICALSACFSFSEGRER HELRHRRHHH  PK HFELPH+PGL  HQKPFI
A0A0D9QXW4 A0A0D9QXW4_CHLSB  1     MKTLPLFVCICALSACFSFSEGRERAHELRHRRHHHHLPKPHFELPHHPGLPTHQKPFII

Q8TAX7 MUC7_HUMAN            61    KSYKCLHKRCRPKLPPSPNNPPKFPNPHQPPKHPDKNSSVVNPTLVATTQIPSVTFPSAS
                                   K +KC +KRCRP+ PPS +NP KFPNP QP KHPD  SSVVNPTLV TTQIPSVT PSAS
A0A0D9QXW4 A0A0D9QXW4_CHLSB  61    KPHKCPYKRCRPRPPPSVHNPHKFPNPPQPSKHPD-TSSVVNPTLVTTTQIPSVTSPSAS

Q8TAX7 MUC7_HUMAN            121   TKITTLPNVTFLPQNATTISSRENVNTSSSVATLAPVNSPAPQDTTAAPPTPSATTPAPP
                                   TKITTLPNVT LPQ ATT SSRENVNT SSVATL   NSPAPQDTTA PPTPSATT   P
A0A0D9QXW4 A0A0D9QXW4_CHLSB  120   TKITTLPNVTSLPQKATTTSSRENVNTGSSVATLTSPNSPAPQDTTAPPPTPSATTALLP

Q8TAX7 MUC7_HUMAN            181   SSSAPPETTAAPPTPSATTQAPPSSSAPPETTAAPPTPPATTPAPPSSSAPPETTAAPPT
                                    SSAP ETT   PPTPSATT   P SSAP ETT   PPTP ATT   P SSAP ETTAA  T
A0A0D9QXW4 A0A0D9QXW4_CHLSB  180   PSSAPLETTTPPPTPSATTAVLPPSSAPLETTTPPPTPSATTAVLPPSSAPLETTAALTT

Q8TAX7 MUC7_HUMAN            241   PSATTPAPLSSSAPPETTAVP-PTPSATTLDPSSASAPPETTAAPPTPSATTPAPPSSPA
                                   P ATT     SSAP ETTA P  TPSATT  P  +SA P+TTAA PTPSATTPAPP S A
A0A0D9QXW4 A0A0D9QXW4_CHLSB  240   PPATTAVLPPSSAPLETTAAPITTPSATTPAPPPSSALPKTTAALPTPSATTPAPPPSSA

Q8TAX7 MUC7_HUMAN            300   PQETTAAPITTPNSSPTTLAPDTSETSAAPTHQTTTSVTTQ-TTTTKQPTSAPGQNKISR
                                   P ETT APITTPNSSP TLAPDTSET AAPTHQTT SVTTQ TTTTKQPTSAP QNKISR
A0A0D9QXW4 A0A0D9QXW4_CHLSB  300   PLETTGAPITTPNSSPATLAPDTSETPAAPTHQTTISVTTQTTTTTKQPTSAPTQNKISR

Q8TAX7 MUC7_HUMAN            359   FLLYMKNLLNRIIDDMVEQ
                                   FLLY+KNLLNR+I+DMVEQ
A0A0D9QXW4 A0A0D9QXW4_CHLSB  360   FLLYIKNLLNRVIEDMVEQ
```

Mismatches with similar properties
Mismatches with dissimilar properties
Gap

Figure 4: BLAST alignment between the two FASTA protein sequences from 1.4

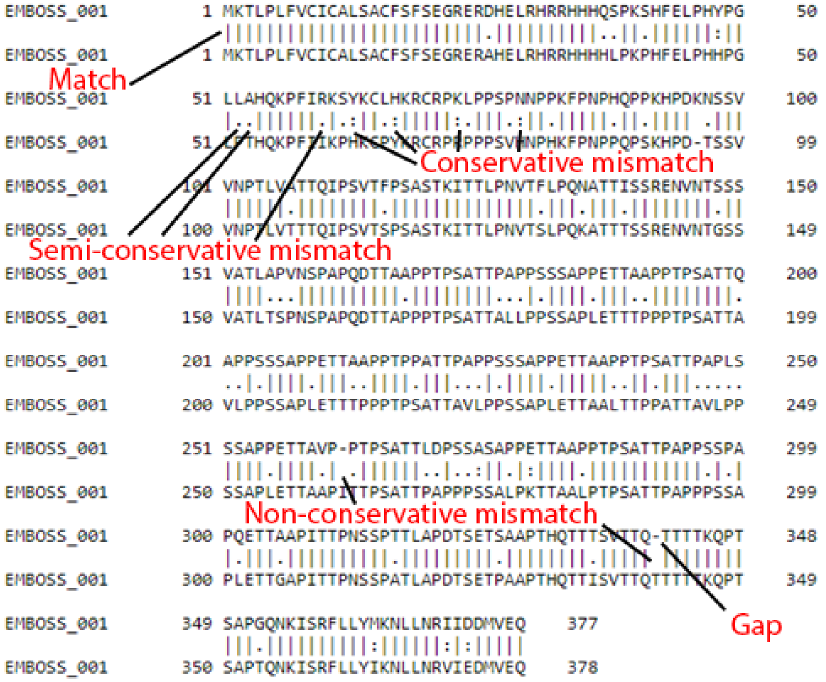
Both empty spaces and the ‘+’ symbol denote a mismatch between the query and subject sequences but, the ‘+’ symbol denotes mismatches that have chemically similar properties. This is the source of the difference in the ‘Ident’ and ‘Positives’ values. The two sequences can be said to be 78.1% identical, and 80.7% similar, as 80.7% of the bases in the two sequences are either matches or close mismatches. The ‘-‘symbol denotes gaps introduced between the two sequences. As can be seen in Figure 4., the relevant sequence for Humans is 377 bases, and the Green Monkey sequence is 378 bases. The relevant letter code is shown when the two sequences match exactly.

2.2 Needleman-Wunsch Algorithm for Sequence Alignment

The Needleman-Wunsch algorithm is used in sequence alignment. It is a dynamic programming algorithm that uses a 2D search matrix to store previous computations of edit distance. This means that it only needs O(nm) computation time instead of O((nm)^2). Figure 5. shows the 2D search matrix for two non-identical sequences. There are three main steps to the algorithm:

1. Create a scoring matrix. This can vary significantly from arbitrary integer scores for matches, mismatches, and indels (insertions or deletions), to matrices that attribute values for the letters that are involved.
2. Trace-back. When moving horizontally or vertically a gap penalty must be added. Diagonal movement is either a match or a mismatch. The value in each index is given by the highest possible score that can be achieved when travelling to that index.
3. Alignment. Walk along the highest scoring path of the matrix starting at the end. The resulting path(s) is the optimal alignment(s) for the given scoring matrix.

Figure 6. shows the alignment between the two sequences from 1.4 using the Needleman-Wunsch algorithm. Vertical dashes indicate matches between the sequences, colons indicate mismatches with similar chemical properties, dots indicate mismatches with slightly more different properties, empty spaces indicate vastly different mismatches, and horizontal dashes indicate introduced gaps between sequences.



```
EMBOSS_001         1 MKTLPLFVCICALSACFSFSEGRERDHELRHRRHHHQSPKSHFELPHYPG     50
                     |||||||||||||||||||||||||.||||||||||..||.||||||:||
EMBOSS_001         1 MKTLPLFVCICALSACFSFSEGRERAHELRHRRHHHHLPKPHFELPHHPG     50

EMBOSS_001        51 LLAHQKPFIRKSYKCLHKRCRPKLPPSPNNPPKFPNPHQPPKHPDKNSSV    100
                     |..||||||.|.:||.:|||||:.|||.:||.|||||.||.|||| .|||
EMBOSS_001        51 LPTHQKPFTIKPHRSPYKRCRPRPPPSVHNPHKFPNPPQPSKHPD-TSSV     99

EMBOSS_001       101 VNPTLVATTQIPSVTFPSASTKITTLPNVTFLPQNATTISSRENVNTSSS    150
                     ||||||.||||||||.||||||||||||||.|||.|||.||||||||.||
EMBOSS_001       100 VNPTLVTTTQIPSVTSPSASTKITTLPNVTSLPQKATTTSSRENVNTGSS    149

EMBOSS_001       151 VATLAPVNSPAPQDTTAAPPTPSATTPAPPSSSAPPETTAAPPTPSATTQ    200
                     ||||...||||||||||.||||||||...|.||||.|||..||||||||.
EMBOSS_001       150 VATLTSPNSPAPQDTTAPPPTPSATTALLPPSSAPLETTTPPPTPSATTA    199

EMBOSS_001       201 APPSSSAPPETTAAPPTPPATTPAPPSSSAPPETTAAPPTPSATTPAPLS    250
                     ..|.||||.|||..||||.|||...|.||||.|||||..||.|||.....
EMBOSS_001       200 VLPPSSAPLETTTPPPTPSATTAVLPPSSAPLETTAALTTPPATTAVLPP    249

EMBOSS_001       251 SSAPPETTAVP-PTPSATTLDPSSASAPPETTAAPPTPSATTPAPPSSPA    299
                     ||||.||||.| .||||||..|..:||.|:||||.|||||||||||.|.|
EMBOSS_001       250 SSAPLETTAAPIATPSATTPAPPPSSALPKTTAALPTPSATTPAPPPSSA    299

EMBOSS_001       300 PQETTAAPITTPNSSPTTLAPDTSETSAAPTHQTTTSVTTQ-TTTTKQPT    348
                     |.|||.||||||||||.|||||||||.||||||||.||||| ||||||||
EMBOSS_001       300 PLETTGAPITTPNSSPATLAPDTSETPAAPTHQTTISVTTQTTTTTKQPT    349

EMBOSS_001       349 SAPGQNKISRFLLYMKNLLNRIIDDMVEQ    377
                     |||.||||||||||:||||||:|:|||||
EMBOSS_001       350 SAPTQNKISRFLLYIKNLLNRVIEDMVEQ    378
```

Figure 5: Needleman-Wunsch alignment between the FASTA protein sequences in 1.4

Conservative mismatches (mutations) are mutations that change an amino acid to a different one that has similar chemical properties. Conservation mutations often have a smaller effect on function than non-conservative mutations. This distinction between conservative and non-conservative mutations is the reason there are measures for how identical and how similar two sequences are, as noted in 2.1.

2.3 Longest Common Subsequence(s)

The longest common subsequence (LCS) is the longest subsequence that can be derived from another by removing some elements without changing the order of the remaining elements. The LCS is different to the *longest common substring* in that the LCS does not have to occupy consecutive positions within the original sequences. The functional definition for the LCS of two subsequences is shown in Figure 7.

$$LCS(X_i, Y_j) = \begin{cases} \emptyset & \text{if } i = 0 \text{ or } j = 0 \\ LCS(X_{i-1}, Y_{j-1}) \frown x_i & \text{if } x_i = y_j \\ \text{longest}\left(LCS(X_i, Y_{j-1}), LCS(X_{i-1}, Y_j)\right) & \text{if } x_i \neq y_j \end{cases}$$

Figure 6: Functional notation for the LCS problem

For two sequences, X and Y, of length 1..m and 1..n indexed by i and j respectively, LCS(X_i, Y_j) represent the set of longest common subsequence of the prefixes X_i and Y_j. The values are null when the indices are 0. When the characters at the indices i and j match, the LCS of the current indices extend the previous diagonal subsequence by the match. When the characters do not match, the LCS is the longest subsequence using previous prefixes of X or Y.

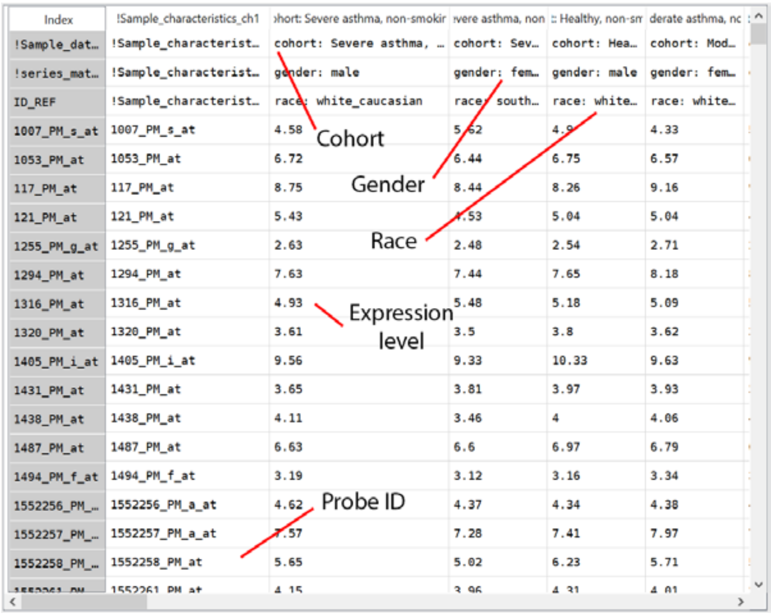
The LCS of the protein sequences from 1.4 is:

MKTLPLFVCICALSACFSFSEGRERHELRHRRHHHPKHFELPHPGLHQKPFIKKCKRCRPPPSNPKFPNPQPKHP
DSSVVNPTLVTTQIPSVTPSASTKITTLPNVTLPQATTSSRENVNTSSVATLPNSPAPQDTTAPPTPSATTAPPSSAP
ETTPPTPSATTAPPSSAPETTPPTPATTAPPSSAPETTAAPPATTAPSSAPETTAPTPSATTPSSAPTTAAPTPSATTPA
PPSSAPETTAPITTPNSSPTLAPDTSETAAPTHQTTSVTTQTTTTKQPTSAPQNKISRFLLYKNLLNRIDMVEQ

3. Task three – Knowledge discovery from microarray data

3.1 Data description

This data is part of the U-BIOPRED (Unbiased BIOmarkers in PREDiction of respiratory disease outcomes) project. The goal of the investigation was to identify transcript fingerprints in whole blood that characterize patients with severe asthma and to determine whether subgroups of severe asthmatics can be identified [6]. As each probe ID was used as a decision variable, there were 54715 decision variables in each experiment. Figure 7. shows a sample of the data with labelled cells. The data source is linked in Appendix A.



Index	!Sample_characteristics_ch1	›hort: Severe asthma, non-smokir	›vere asthma, non	:: Healthy, non-sm	derate asthma, nc
!Sample_dat…	!Sample_characterist…	cohort: Severe asthma, …	cohort: Sev…	cohort: Hea…	cohort: Mod…
!series_mat…	!Sample_characterist…	gender: male	gender: fem…	gender: male	gender: fem…
ID_REF	!Sample_characterist…	race: white_caucasian	race: south…	race: white…	race: white…
1007_PM_s_at	1007_PM_s_at	4.58	5.62	4.9	4.33
1053_PM_at	1053_PM_at	6.72	6.44	6.75	6.57
117_PM_at	117_PM_at	8.75	8.44	8.26	9.16
121_PM_at	121_PM_at	5.43	4.53	5.04	5.04
1255_PM_g_at	1255_PM_g_at	2.63	2.48	2.54	2.71
1294_PM_at	1294_PM_at	7.63	7.44	7.65	8.18
1316_PM_at	1316_PM_at	4.93	5.48	5.18	5.09
1320_PM_at	1320_PM_at	3.61	3.5	3.8	3.62
1405_PM_i_at	1405_PM_i_at	9.56	9.33	10.33	9.63
1431_PM_at	1431_PM_at	3.65	3.81	3.97	3.93
1438_PM_at	1438_PM_at	4.11	3.46	4	4.06
1487_PM_at	1487_PM_at	6.63	6.6	6.97	6.79
1494_PM_f_at	1494_PM_f_at	3.19	3.12	3.16	3.34
1552256_PM_…	1552256_PM_a_at	4.62	4.37	4.34	4.38
1552257_PM_…	1552257_PM_a_at	7.57	7.28	7.41	7.97
1552258_PM_…	1552258_PM_at	5.65	5.02	6.23	5.71
1552261_PM_…	1552261_PM_at	4.15	3.96	4.31	4.01

Figure 7: Example dataset structure for experiments one and two.

Along with cohort information, each sample contains gender and race information. Gender and race information was not used in these analyses. There were 498 samples in total, split into four cohorts:

1. Healthy, non-smoking (87 samples)
2. Moderate asthma, non-smoking (77 samples)
3. Severe asthma, non-smoking (246 samples)
4. Severe asthma, smoking (88 samples)

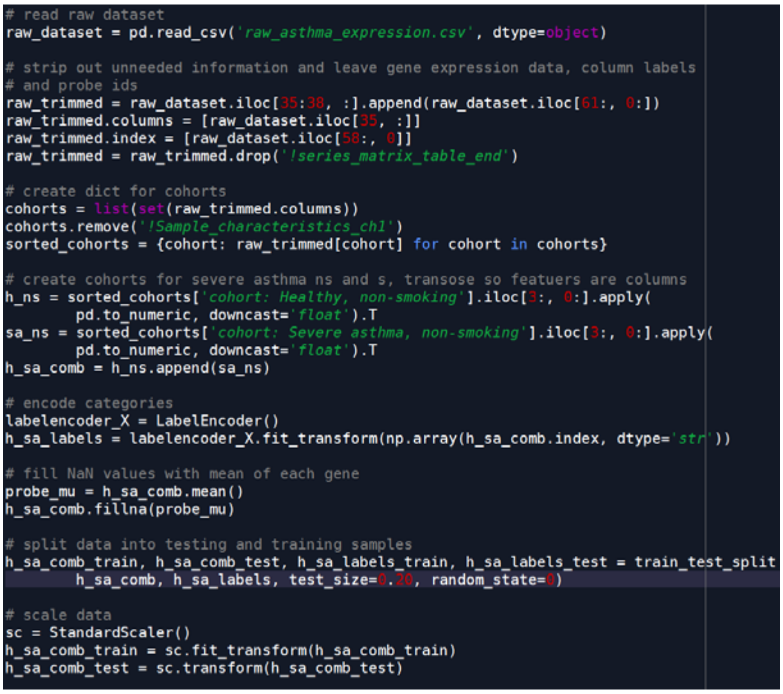
Two experiments were carried out, binary classification of non-smoking and smoking severe asthmatics, and binary classification of non-asthmatics and severe (non-smoking asthmatics). The purpose of experiment one was to determine if smoking caused significant enough effects on gene expression to allow high performance classification. The purpose of experiment two was to determine what differences in gene expression were present in severe asthmatics compared to healthy individuals. The smoking severe asthmatic cohort was excluded from the second experiment as the biomarkers associated with smoking were assumed to be significantly different to those associated with just severe asthma.

3.2 Relevance of the data

Analysis of the dataset could lead to the stated goals of the U-BIOPRED project - uncovering biomarkers that can be used in the prediction of respiratory disease. This knowledge could be used for optimising treatment of respiratory diseases by predicting patient response to medication and other treatments.

3.3 Pre-processing

Figure 8. shows pre-processing used on each of the cohorts. The data h_ns contains all the useful healthy, non-smoking information, and sa_ns contains all the useful severe asthma, non-smoking information. Homogenous and and unused information was removed from the dataset. The data was then transposed so that each probe ID was a column, the probe ID's were used as the decision variables. NaN or Null values were replaced with the mean value of each probe ID expression level. The cohort information was then encoded as either a '0' or '1'. The expression data was then split into training and testing samples at either a 75:25 or 80:20 ratio of training to testing data. Finally, the training and testing data was standardized about the mean using the training set statistics (mean and standard deviation used in standardisation calculated from the training data).



```
# read raw dataset
raw_dataset = pd.read_csv('raw_asthma_expression.csv', dtype=object)

# strip out unneeded information and leave gene expression data, column labels
# and probe ids
raw_trimmed = raw_dataset.iloc[35:38, :].append(raw_dataset.iloc[61:, 0:])
raw_trimmed.columns = [raw_dataset.iloc[35, :]]
raw_trimmed.index = [raw_dataset.iloc[58:, 0]]
raw_trimmed = raw_trimmed.drop('!series_matrix_table_end')

# create dict for cohorts
cohorts = list(set(raw_trimmed.columns))
cohorts.remove('!Sample_characteristics_ch1')
sorted_cohorts = {cohort: raw_trimmed[cohort] for cohort in cohorts}

# create cohorts for severe asthma ns and s, transose so featuers are columns
h_ns = sorted_cohorts['cohort: Healthy, non-smoking'].iloc[3:, 0:].apply(
        pd.to_numeric, downcast='float').T
sa_ns = sorted_cohorts['cohort: Severe asthma, non-smoking'].iloc[3:, 0:].apply(
        pd.to_numeric, downcast='float').T
h_sa_comb = h_ns.append(sa_ns)

# encode categories
labelencoder_X = LabelEncoder()
h_sa_labels = labelencoder_X.fit_transform(np.array(h_sa_comb.index, dtype='str'))

# fill NaN values with mean of each gene
probe_mu = h_sa_comb.mean()
h_sa_comb.fillna(probe_mu)

# split data into testing and training samples
h_sa_comb_train, h_sa_comb_test, h_sa_labels_train, h_sa_labels_test = train_test_split
        h_sa_comb, h_sa_labels, test_size=0.20, random_state=0)

# scale data
sc = StandardScaler()
h_sa_comb_train = sc.fit_transform(h_sa_comb_train)
h_sa_comb_test = sc.transform(h_sa_comb_test)
```

Figure 8: Generic pre-processing for experiments one and two.

3.4 Visualisation

Each probe ID was used as a decision variable. Given that humans struggle to visualise more than 3 dimensions, dimensionality reduction was used to make useful visualisation possible. Three techniques for visualisation used: PCA (principal component analysis), LDA (linear discriminant analysis), and T-SNE (t-distributed stochastic neighbour embedding).

PCA is used to transform a set of observations into a series of linearly uncorrelated variables called principal components. The transformation is done such that the first component has the most variance (explains the most variance from the data as possible), and each succeeding component has the most variance possible under the constraint that it is orthogonal to the preceding components. PCA does not take into account any difference in class, it is unsupervised. LDA attempts to find a linear combination of features that characterize two or more classes of objects or events. It explicitly attempts to model the difference between the classes of data, it is a supervised algorithm. T-SNE is a nonlinear dimensionality reduction technique that models each high dimensional object by a 2 or 3D point such that similar objects are modelled by nearby points. The data used in experiment one is visualised in Figure 9.

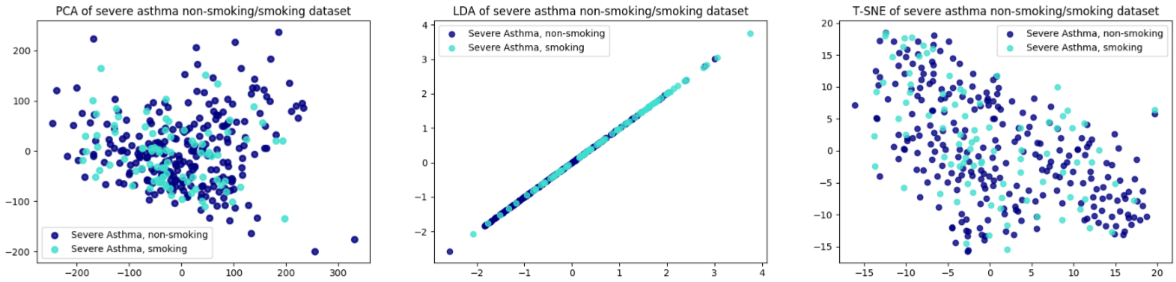


Figure 9: Experiment one data visualisation.

The separation of the data using PCA is very poor. This indicates that the relationships between the variables are non-linear and complex. LDA indicates that the distributions of the two cohorts overlap but they have different means. T-SNE indicates that there are possibly multiple complex shaped clusters that could be formed to classify the data, although overfitting could easily occur. The results of the visualisation from experiment two is shown in Figure 10.

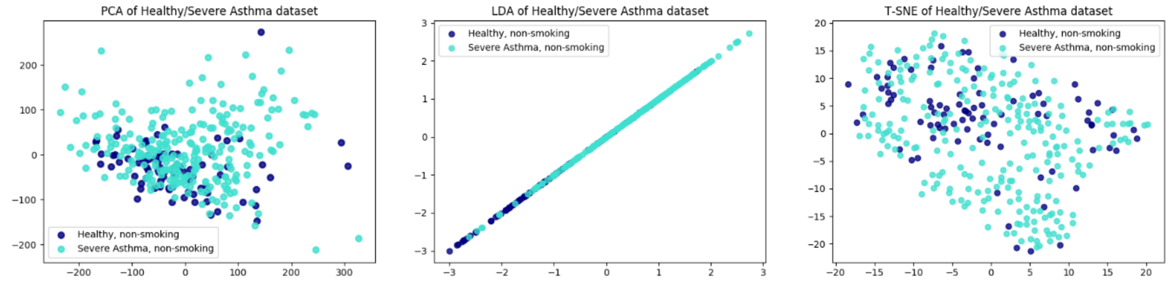


Figure 10: Experiment two data visualisation.

The separation of the data using PCA is better than experiment one, but still quite poor. The variables likely have non-linear complex relationship. LDA indicates the same as in experiment one - hat the distributions overlap but have different means. This is surprising given that the cohorts in experiment one were split only by their smoking status, and the cohorts in experiment two were split by non and severe asthma. T-SNE clearly indicates 3 or more complex clusters that could be used for classifying healthy, non-smokers.

3.5 Feature selection and extraction

Feature selection is used to reduce the number of components needed in modelling by identifying variables that have the greatest effect on outcomes and eliminating those with minimal impact. Two methods were used for feature selection in experiments one and two, PCA and SNR (signal to noise ratio). The formula shown in Figure 11. was used to extract features with a high signal to noise statistic. The threshold for SNR extraction was set to 0.275 for experiment one, and 0.45 for experiment two. These values were chosen because the two cohorts in experiment one were differentiated only by whether they did or did not smoke. It was assumed that this differentiation would be less significant than that between the cohorts in experiment two (non-asthmatic and severe asthmatic).

$$D_{sn} = \frac{(\mu_a - \mu_b)}{(\sigma_a + \sigma_b)}$$

Figure 11: Signal-to-noise statistic.

As noted previously, PCA can be used for dimensionality reduction by transforming an initial set of variables into a smaller set of linearly uncorrelated components. There are two assumptions of PCA that were likely violated in the experiments, the assumption of only linear relationships between variables, and that there are no significant outliers. PCA was used to serve as a comparison against SNR, which makes no such assumptions. The number of components used after reduction was varied from 5-1000 in steps of 5, with the best number of components being chosen based on minimizing the sum of false negatives/positives.

3.6 Classification

Four types of classification algorithms were used for each experiment: logistic regression, multi-layer perceptron, support vector machines, and K-nearest neighbours. A short explanation of the algorithms and the hyperparameters for each is shown in Appendix B, C, and D. Attempting classification without feature selection would have resulted in very long training times. PCA and SNR were used to select a varied number of features to use for classification. True/false positives/negatives were used as the main performance metric as they help identify the context in which the algorithms are performing poorly. The best results for each classification algorithm for experiment one is shown in tables 1 and 2, and for experiment two in tables 3 and 4. The best results were selected by minimising the sum of the % of false positives and false negatives. The performance metrics used are further discussed in Section 3.7. The scores with the 800 highest SNR components are indicated in brackets. In the case of experiment one, positive classification indicates severe asthma, smoking. In the case of experiment two, positive classification indicates severe asthma, non-smoking.

Performance category	Logistic regression score	Multi-layer perceptron	Support vector machine	K-nearest neighbors
True positives	46.7%	31.25%	26.7%	25%
False positives	53.3%	68.75%	73.3%	75%
True negatives	82.7%	88%	84.6%	83.8%
False negatives	17.3%	12%	15.4%	16.2%

Table 1: Experiment one classification using 260 principal components.

Performance category	Logistic regression (800)	Multi-layer perceptron	Support vector machine	K-nearest neighbors
True positives	73.3% (86.7%)	60%	26.7%	46.7%
False positives	26.7% (13.3%)	40%	77.3%	53.3%
True negatives	77% (94.2%)	96.2%	100%	98.1%
False negatives	23% (5.8%)	3.8%	0%	1.9%

Table 2: Experiment one classification using top 12 SNR selected components.

In experiment one, classification using the SNR identified components was significantly better than using PCA. As shown in Section 3.4, all of the dimensionality reduction measures experienced difficulty in separating the two cohorts. PCA also does not take into account nonlinear and complex relationships between decision variables. These two factors are likely why the PCA feature selections performed poorly even when the hyperparameters were optimised. SNR may have implicitly identified variables with complex relationships that PCA missed. Interestingly, increasing the number of components selected with SNR to 800 significantly improved the classification performance with logistic regression. The increase had negligible effects on the other classifiers.

Performance category	Logistic regression	Multi-layer perceptron	Support vector machine	K-nearest neighbors
True positives	80.9%	71.9%	92.9%	93%
False positives	19.1%	28.1%	7.1%	7%
True negatives	80%	88.9%	32%	44.4%
False negatives	20%	11.1%	68%	55.6%

Table 3: Experiment two classification using 260 principal components.

Performance category	Logistic regression (800)	Multi-layer perceptron	Support vector machine	K-nearest neighbors
True positives	97.6% (100%)	97.6%	97.6%	95.2%
False positives	2.4% (0%)	2.4%	2.4%	4.8%
True negatives	52% (72%)	52%	44%	56%
False negatives	48% (28%)	48%	56%	44%

Table 4: Experiment two classification using top 12 SNR selected components.

In experiment two, classification using SNR identified components was again significantly better than using PCA. The classifier performance was similar to experiment one, but most of the prediction errors were false negatives instead of false positives. These results are interesting, as it was assumed that classification would be much easier due to the greater separation of the cohorts seen in Section 3.4. Again, increasing the number of SNR selected components to 800 significantly improved the performance of logistic regression, but had a negligible effect on the other classifiers.

3.7 Performance metrics

The performance of classification problems needs to be assessed differently to regression problems. For example, a classification algorithm could be 90% accurate, but whether all of this inaccuracy is solely in predicting one or more classes cannot be known from overall accuracy alone. True/false positives/negative rates indicate what is not being correctly classified in a binary classification problem. These rates can be visualised in a table called a confusion matrix for quick analysis. Figure 12. shows an example confusion matrix constructed from the SNR logistic regression results of experiment one. Confusion matrices can be extended to multi-class classification problems, which can indicate areas to focus improvement efforts. Figure 13. shows how the performance of the logistic classifier varied for different numbers of PCA components in experiment two.

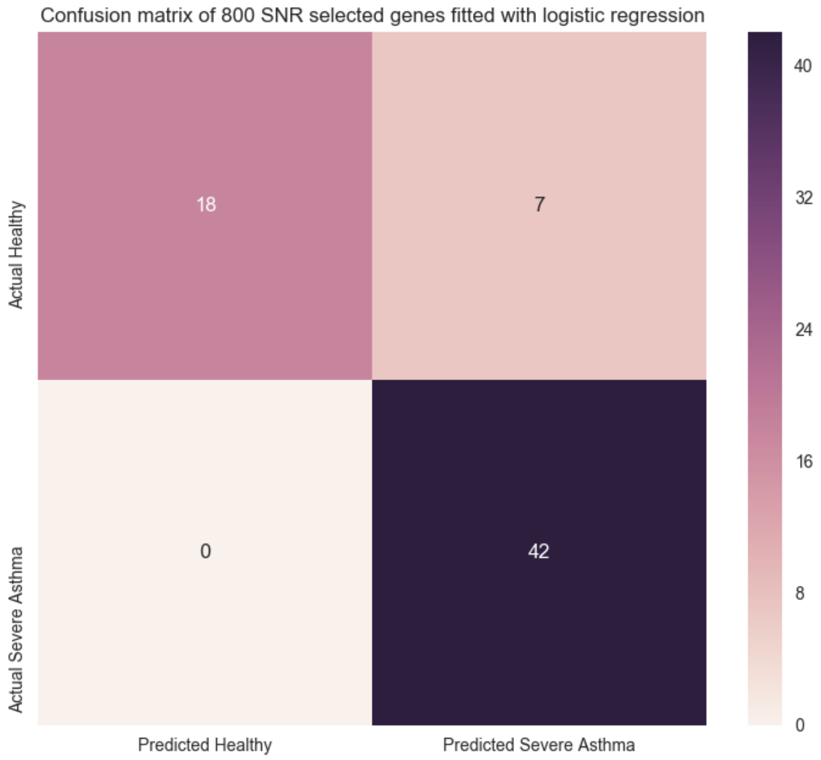


Figure 12: Confusion matrix of logistic classification using top 800 SNR selected components in experiment one.

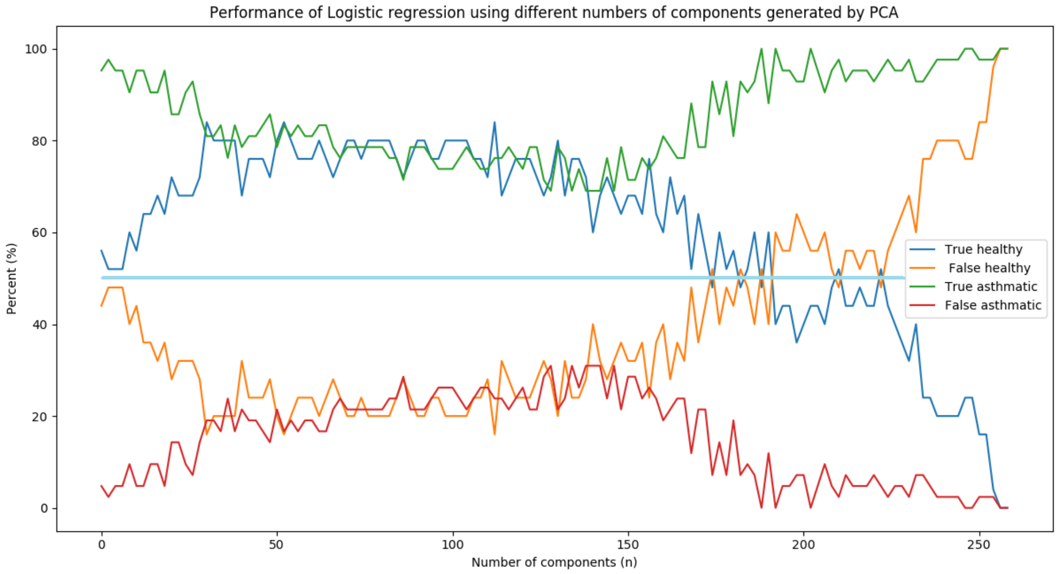


Figure 13: Logistic regression performance using varying numbers of principal components in experiment two.

3.8 Knowledge discovery

The probe ID’s with the highest SNR were mapped to their corresponding sequences. The sequences were then identified using BLAST analysis. The performance was highest in experiment 1b) and 2b) (ignoring logistic regression) using the top 12 genes from each set of cohorts. The following observations can be made about the identifiable genes from experiments one and two. Further information about the top 12 genes from each cohort can be found in Appendix E and F.

SNR rank	Gene	Information
1	GPR15	Hypomethelation of this gene is significantly greater in smokers than non-smokers [7]. Its expression is also up-regulated in rheumatoid arthritis patients [8].
3	LRRN3	Low expression of LRRN3 and high expression of MYCN is highly associated with unfavourable outcomes in patients with neuroblastoma [13]
5	CDH18	CDH18 has been implicated as a loci for leprosy susceptibility [9]. Leprosy susceptibility loci have demonstrated a high tendency to show association with autoimmune and inflammatory diseases.[9].
6	MCHR1	Polymorphisms in MCHR1 are associated with differences in body composition and interact with physiologic and energy-related lifestyle factors [11].
7	CLDND1	Implicated as a survival factor in breast cancer [10].
10	KIAA0790	May play a role in hematopoiesis [12].

Table 5: Identifiable genes from top 12 SNR experiment one.

SNR rank	Gene	Information
3	PDE7A	PDE7 inhibiting drugs are useful for regulating pro-inflammatory and immune T-cell function [17]
5	MyD88	Signalling mediated by this gene in intestinal epithelial cells is crucial for maintenance of gut homeostasis [15]
9	F5	May be involved in coagulation [16]
11	RBM17	May be involved in sickle cell anaemia [17]

Table 6: Identifiable genes from top 12 SNR in experiment two.

The selected genes are involved in a wide range of functions. None of the top 12 genes overlapped between the experiments. It should be noted that the hypomethelation of the gene with the highest SNR in experiment one, GPR15, is significantly greater in smokers than non-smokers. It is clearly an important factor in differentiating between the cohorts in experiment one. Several genes in experiment one are very highly expressed in the brain, as shown in Figure 14. The gene with the highest SNR in experiment two, PDE7A, may be involved in general inflammatory and immune function. This suggests that individuals with Asthma may experience greater inflammation in general compared to non-asthmatics. Further work is discussed in Section 3.10.

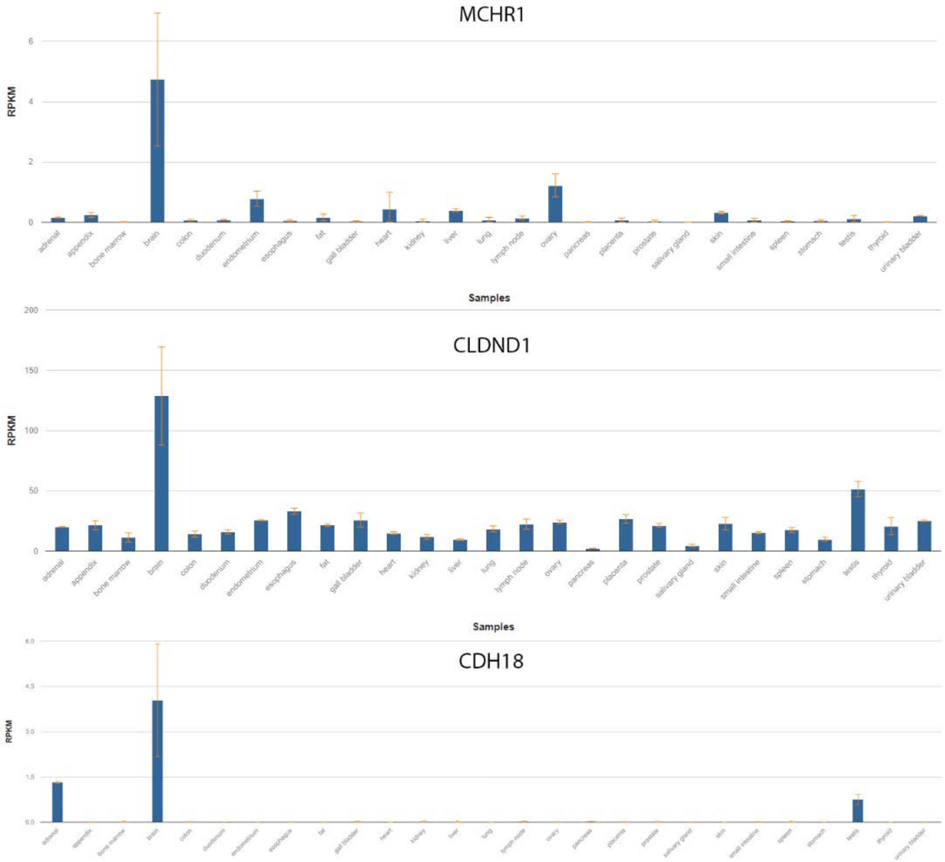


Figure 14: Expression levels of MCHR1, CLDND1, and CDH18 in different human body tissues.

3.9 Inference

Classification in this context serves two purposes:

1. Identification of novel treatment pathways
2. Predicting the severity of asthma and modulating treatment accordingly

It may be that current treatments such as the various corticosteroids and bronchodilators could be avoided entirely if targeted gene or drug therapy was used to alter the severity of asthma over a medium-long term period. Predicting the severity of asthma, even if the known symptoms do not indicate a high level of severity, could prevent fatal asthma attacks by making use of preventative (instead of reactive) treatment such as symbicort. My motivation for carrying out these experiments is entirely selfish. I have severe asthma, and I would rather that I did not.

3.10 Discussion and future work

These two experiments indicate that severe asthma and smoking have significant effect on gene expression profiles. Further analysis of this dataset and comparison with other research could help identify transcriptional 'fingerprints' for severe asthma and smokers. Possible future work directions sorted by priority and estimated information gain:

1. Split cohorts into subgroups based on gender and perform binary classification between each possible group (8C2 = 28 possible experiments)
2. Analyse the top N genes from each cohort and examine where overlaps do and do not occur.
3. Use clustering methods to determine possible subgroups that can be used for classification between different sets of cohorts.
4. Design ensemble classifiers that use weighted estimates for classification. As shown in TABLE, some methods were great at predicting one class, but no better than chance with the other. An ensemble of models could be created that uses each model to predict class membership, and then weights the prediction based on their known performance.
5. Perform feature construction using known indicators of each class. The ratio of the expression of one gene relative to another could be an important latent factor in classification.

References

[1] Laitinen, T., Daly, M. J., Rioux, J. D., Kauppi, P., Laprise, C., Petäys, T., ... & Laitinen, L. A. (2001). A susceptibility locus for asthma-related traits on chromosome 7 revealed by genome-wide scan in a founder population. *Nature genetics*, *28*(1), 87.

[2] Illig, T., & Wjst, M. (2002). Genetics of asthma and related phenotypes. *Paediatric respiratory reviews*, *3*(1), 47-51.

[3] Pillai, S. G., Chiano, M. N., White, N. J., Speer, M., Barnes, K. C., Carlsen, K., ... & Sly, P. (2006). A genome-wide search for linkage to asthma phenotypes in the genetics of asthma international network families: evidence for a major susceptibility locus on chromosome 2p. *European journal of human genetics: EJHG*, *14*(3), 307.

[4] Rousseau, K., Vinall, L. E., Butterworth, S. L., Hardy, R. J., Holloway, J., Wadsworth, M. E. J., & Swallow, D. M. (2006). MUC7 haplotype analysis: results from a longitudinal birth cohort support protective effect of the MUC7* 5 allele on respiratory function. *Annals of human genetics*, *70*(4), 417-427.

[5] Kirkbride, H. J., Bolscher, J. G., Nazmi, K., Vinall, L. E., Nash, M. W., Moss, F. M., ... & Swallow, D. M. (2001). Genetic polymorphism of MUC7: allele frequencies and association with asthma. *European journal of human genetics: EJHG*, *9*(5), 347.

[6] Bigler, J., Boedigheimer, M., Schofield, J. P., Skipp, P. J., Corfield, J., Rowe, A., ... & Roberts, G. (2017). A severe asthma disease signature from gene expression profiling of peripheral blood from U-BIOPRED cohorts. *American journal of respiratory and critical care medicine*, *195*(10), 1311-1320.

[7] Kõks, G., Uudelepp, M. L., Limbach, M., Peterson, P., Reimann, E., & Kõks, S. (2015). Smoking-induced expression of the GPR15 gene indicates its potential role in chronic inflammatory pathologies. The American journal of pathology, 185(11), 2898-2906.

[8] Cartwright, A., Schmutz, C., Askari, A., Kuiper, J. H., & Middleton, J. (2014). Orphan receptor GPR15/BOB is upregulated in rheumatoid arthritis. Cytokine, 67(2), 53-59.

[9] Liu, H., Irwanto, A., Fu, X. A., Yu, G., Yu, Y., Sun, Y., ... & Li, Y. (2015). Discovery of six new susceptibility loci and analysis of pleiotropic effects in leprosy. Nature genetics, 47(3), 267-271.

[10] Achari, C., Winslow, S., & Larsson, C. (2015). Down regulation of CLDND1 induces apoptosis in breast cancer cells. PloS one, 10(6), e0130300.

[11] Fontaine-Bisson, B., Thorburn, J., Gregory, A., Zhang, H., & Sun, G. (2014). Melanin-concentrating hormone receptor 1 polymorphisms are associated with components of energy balance in the Complex Diseases in the Newfoundland Population: Environment and Genetics (CODING) study. The American journal of clinical nutrition, 99(2), 384-391.

[12] Claudio, J. O., Zhu, Y. X., Benn, S. J., Shukla, A. H., McGlade, C. J., Falcioni, N., & Stewart, A. K. (2001). HACS1 encodes a novel SH3-SAM adaptor protein differentially expressed in normal and malignant hematopoietic cells. Oncogene, 20(38), 5373.

[13] Akter, J., Takatori, A., Hossain, M. S., Ozaki, T., Nakazawa, A., Ohira, M., ... & Nakagawara, A. (2011). Expression of NLRR3 orphan receptor gene is negatively regulated by MYCN and Miz-1, and its downregulation is associated with unfavorable outcome in neuroblastoma. Clinical Cancer Research, 17(21), 6681-6692.

[14] Castaño, T., Wang, H., Campillo, N. E., Ballester, S., González-García, C., Hernández, J., ... & Huertas, O. (2009). Synthesis, structural analysis, and biological evaluation of thioxoquinazoline derivatives as phosphodiesterase 7 inhibitors. ChemMedChem, 4(5), 866-876.

[15] Bonnert, T. P., Garka, K. E., Parnet, P., Sonoda, G., Testa, J. R., & Sims, J. E. (1997). The cloning and characterization of human MyD88: a member of an IL-1 receptor related family. FEBS letters, 402(1), 81-84.

[16] Santamaria, S., Reglińska-Matveyev, N., Gierula, M., Camire, R. M., Crawley, J. T., Lane, D. A., & Ahnström, J. (2017). Factor V has an anticoagulant cofactor activity that targets the early phase of coagulation. Journal of Biological Chemistry, 292(22), 9335-9344.

[17] Lallena, M. J., Chalmers, K. J., Llamazares, S., Lamond, A. I., & Valcárcel, J. (2002). Splicing regulation at the second catalytic step by Sex-lethal involves 3′ splice site recognition by SPF45. Cell, 109(3), 285-296.

Appendices

Appendix A – Dataset link

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69683

Appendix B - Classification algorithm explanations

Logistic regression is a regression model where the dependent variable is categorical. The output is expressed by mapping the linear simple or multinomial regression model through the logit function. In the case of a binary dependant variable, the probability of class membership can be found by taking the inverse of the logit function - the logistic function. The general form of the logistic function is shown in Figure B.1. In the context of binary classification, f(x) is the probability of class membership being ‘positive’, and x is the linear or multinomial regression equation (y = Ax + b)

$$f(x) = \frac{1}{1 + e^{-x}}$$

Figure B. 1: General form of the logistic function.

MLP is a class of feedforward artificial neural network. It consists of three or more layers of nodes with activation functions: the input layer, hidden layer(s), and the output layer. The activation function is a function that maps the weighted inputs to the node to its output. Each node in one layer connects with a weight W to every node in the following layer. Learning occurs by changing the connections weights, W, based on the error in the output compared to the desired result. Figure B.2 shows the tanh activation function, the resultant error sum, and the change in connection weights calculated via gradient descent.

$$f(x) = \tanh(x) \qquad \mathcal{E}(n) = \frac{1}{2}\sum_j e_j^2(n). \qquad \Delta w_{ji}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial v_j(n)} y_i(n)$$

Figure B. 2: Tanh activation function, sum of squared errors, and change in connection weight. In order.

SVM constructs a hyperplane (or set of hyperplanes) in an N-dimensional space that seeks to maximise the distance of two or more support vectors from the hyperplane (while keeping them equidistant). Kernel functions are usually used to map the input data to a higher dimensional space so that it becomes linearly separable. Figure B.3 shows the formulation for a soft margin SVM, which allows some values to enter the margin between the support vectors and the hyperplane.. If the value of lambda is small enough, the soft-margin SVM behaves identically to a hard-margin SVM.

$$\left[\frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i(\vec{w} \cdot \vec{x}_i - b)\right)\right] + \lambda \|\vec{w}\|^2$$

Figure B. 3: Soft margin SVM formulation.

K-NN for classification is a non-parametric method. The input consists of the k closest samples in the feature space, and the output is the class membership of the sample in question. The class with the highest membership, found by summing the k closest samples, is assigned to the unclassed sample. The k closest samples can be weighted by their distance to place preference on samples that are closer to the unclassed sample.

Appendix C - Experiment one hyperparameters

Logistic regression

```
lr = LogisticRegression(random_state=0,
                        tol=0.00001,
                        class_weight='balanced',
                        solver='lbfgs',
                        max_iter=1000,
                        multi_class='ovr',
                        verbose=True)
```

Multilayer perceptron

```
mlp = MLPClassifier(hidden_layer_sizes=10000,
                    activation='identity',
                    solver='lbfgs',
                    alpha=0.1,
                    batch_size=200,
                    learning_rate='adaptive',
                    max_iter=200,
                    tol=0.00000001,
                    verbose=True,
                    early_stopping=False,
                    validation_fraction=0.2
                    )
```

Support vector machine

```
svm_k = SVC(kernel='linear',
         tol=0.0000001,
         random_state=0,
         verbose=True)
```

K-nearest neighbors

```
nn = 13
power_param = 2
k_nn = neighbors.KNeighborsClassifier(n_neighbors=nn,
                                weights='distance',
                                p=power_param,
                                n_jobs=-1)
```

Appendix D - Experiment two hyperparameters

Logistic regression

```
lr = LogisticRegression(random_state=0,
                  tol=0.00001,
                  solver='lbfgs',
                  multi_class='multinomial',
                  verbose=True)
```

Multilayer perceptron

```
mlp = MLPClassifier(hidden_layer_sizes=10000,
                activation='identity',
                solver='lbfgs',
                alpha=0.1,
                batch_size=50,
                learning_rate='adaptive',
                max_iter=200,
                tol=0.00000001,
                verbose=True,
                early_stopping=False,
                validation_fraction=0.2
                )
```

Support vector machine

```
svm_k = SVC(kernel='linear',
         tol=0.0000001,
         random_state=0,
         verbose=True)
```

K-nearest neighbors

```
nn = 15
power_param = 2
k_nn = neighbors.KNeighborsClassifier(n_neighbors=nn,
                                weights='distance',
                                p=power_param,
                                n_jobs=-1)
```

Appendix E - Top 12 genes identified by SNR experiment one. 100% match unless stated otherwise.

https://www.ebi.ac.uk/arrayexpress/files/A-GEOD-13158/A-GEOD-13158.adf.txt

SNR rank	Probe id	GenBank entry	Nucleotide
1	208524_PM_at	NM_005290	GPR15
2	209841_PM_s_at	AL442092	DKFZp761K2424
3	209840_PM_s_at	AI221950	LRRN3
4	214153_PM_at	BE467941	BEK67_22530 (26% match)
5	206280_PM_at	NM_004934	CDH18
6	223855_PM_s_at	BC001736	MCHR1
7	1554149_PM_at	BC013610	CLDND1
8	1563346_PM_at	AY063452	Pc21g12300 (28% match)
9	239146_PM_at	AI634844	Txndc12 (42.9% match)
10	41644_PM_at	AB018333	KIAA0790
11	240577_PM_at	AI033071	Cvel_23002 (32% match)
12	207637_PM_at	NM_014653	WSCD2

Appendix F - Top 12 genes identified by SNR experiment two. 100% match unless stated otherwise.

SNR rank	Probe id	GenBank entry	Nucleotide
1	241320_PM_at	AI821449	No BLAST matches
2	216262_PM_s_at	AL050318	TGIF2
3	1552343_PM_s_at	NM_002604	PDE7A
4	212382_PM_at	BF433429	FOZG_07194 (48.5% match)
5	209124_PM_at	U70451	MyD88
6	231029_PM_at	AI740541	HMPREF0185_01232 (38.5% match)
7	232034_PM_at	AL117607	FSHD (48% match)
8	231500_PM_s_at	AV650728	H3BVE0 (99% match)
9	204714_PM_s_at	NM_000130	F5
10	229806_PM_at	AI304951	Dmoj\GI13502 (28.6% match)
11	224781_PM_s_at	AI923119	RBM17
12	213079_PM_at	AA223871	D623_10008204