Effects of Preserved and Preservative-Free Glaucoma Drugs on Proteomic Expression Levels in Corneal and Conjunctival Epithelial Cells in vitro

Effects of Preserved and Preservative-Free Glaucoma Drugs on Proteomic Expression Levels in Corneal and Conjunctival Epithelial Cells *in vitro*

By Janika Nättinen

Master's Degree Programme in Bioinformatics

SILK, Department of Ophthalmology, University of Tampere, Finland

BioMediTech, Tampereen yliopisto (University of Tampere)

Tampere

28.08.2015

Acknowledgements

Master's Thesis

Abstract

Benzalkonium chloride (BAC), is the most commonly used preservative world-wide in cosmetics and eye drops; however, it is toxic to epithelial cells and the pathways in this toxicity are largely unknown. In this study, the effects of common ophthalmic drugs for glaucoma were uncovered by exposing human corneal and conjunctival epithelial cells (HCE and IOBA-NHC respectively) to treatments with and without BAC. It was hypothesized that these drugs may be linked to inflammatory mechanisms and cell death. The aim of this study was to explore proteomic data in order to discover some potential biomarkers relating to BAC-induced effects, which could be beneficial in further studies.

HCE and IOBA-NHC cells were exposed to a preservative-free topical medication tafluprost, a similar drug latanoprost which contains BAC or BAC by itself for 24 hours and then the proteomic profiles of treated and untreated cells were analysed with NanoLC-TOF-MS using SWATH technique. Central tendency normalization was applied to the $\log_2$-transformed proteomic data once the quality of the data was initially ensured with descriptive statistics including correlation and clustering methods. Mixed-effects ANOVA model was implemented to data to uncover differentially

expressed protein levels and Benjamini-Hochberg procedure was used for multiple testing correction. All these procedures were performed using R software.

Statistical analysis identified 29 differentially expressed proteins for IOBA-NHC cells (fold change>1.25 or <0.8, q-value<0.25) and 28 for HCE cells (fold change>1.5 or <0.67, q-value<0.25). Based on the significance estimations, enrichment analyses were performed using several online tools, including GOrilla and DAVID tools. After examining both individual statistically and biologically significant proteins and the enrichment analyses results for both cell lines, it appeared that changes in mitochondrion functions are affected by exposure to BAC. This was supported by the enrichment analyses and in addition NDFUA5 and NDUFS3, proteins associated to the mitochondrial membrane respiratory chain NADH dehydrogenase, were under-expressed for BAC-treated samples in IOBA-NHC cell line. Furthermore, in both cell lines the cholesterol production and therefore the plasma membrane permeability and structure could be altered due to reduced abundance of HMGCS1, which is an essential catalyst in this process. In addition, actin cytoskeleton contractions were at least in the HCE cell line increased, which then in turn could affect the permeability of the cell junctions. This was initially noted due to the over-expression of MYH9, MYL12A and MYL6 in HCE cell line samples treated with BAC. These potential novel proteomic biomarkers will be further analysed in ongoing clinical studies of glaucoma patients.

Tiivistelmä

Bentsalkoniumkloridi (BAC), on maailmanlaajuisesti yleisimmin käytetty säilöntäaine kosmetiikassa ja silmätipoissa; se on kuitenkin myrkyllistä epiteelisoluille ja tähän myrkyllisyyteen johtavat solunsisäiset reitit ovat suurelta osin edelleen tuntemattomia. Tässä tutkimuksessa yleisten glaukoomalääkkeiden vaikutuksia tutkittiin altistamalla ihmisen sarveiskalvon ja sidekalvon epiteelisoluja (HCE ja IOBA-NHC) käsittelyihin, joista osa sisälsi BAC-säilöntäainetta. Hypoteesina oli, että nämä säilöntäaineita sisältävät lääkkeet voisivat aiheuttaa soluissa muutoksia, jotka liittyvät tulehduksellisiin mekanismeihin ja solukuolemaan. Tavoitteena oli identifioida proteomiikka datan avulla potentiaalisia biomarkkereita, jotka ilmaisisivat BAC-säilöntäaineen vaikutuksia soluissa.

HCE ja IOBA-NHC solut altistettiin joko säilöntäaineettomalle tafluprostille, samankaltaiselle BAC-säilöntäainetta sisältävälle latanoprostille tai pelkästään BAC-säilöntäaineelle 24 tunnin ajan ja näytteet analysointiin NanoLC-TOF-MS-laitteella SWATH-tekniikkaa käyttäen. Log_2-transformoitu data normalisoitiin mediaani-normalisointia käyttäen sen jälkeen, kun näytteiden laatu oli ensin vahvistettu alustavien kuvaajien ja metodien, mm. korrelaation ja klusteroinnin, kautta. Tilastollisessa analyysissa käytettiin sekamallia ja Benjamini-Hochberg-menetelmää sovellettiin p-arvojen väärien positiivisten löydösten kontrolloimiseen. Kaikki yllämainitut menetelmät tehtiin käyttäen R-tietokoneohjelmaa.

Tilastollinen analyysi identifioi 29 tilastollisesti mielenkiintoista proteiinia IOBA-NHC solujen näytteistä ja 28 vastaavasti HCE solujen näytteistä. Tilastollisten tulosten perusteella tehtiin rikastusanalyyseja useilla verkko-ohjelmilla, joihin kuuluivat mm. GOrilla ja DAVID. Kun sekä yksittäiset, mielenkiintoiset proteiinit että rikastusanalyysien tulokset käytiin läpi, tulokset viittasivat siihen, että BAC vaikuttaa mitokondrioon ja siihen liittyviin mekanismeihin soluissa, soluhengitykseen erityisesti. Tätä tukevat sekä rikastusanalyysien tulokset sekä se, että NDFUA5 ja NDUFS3, jotka liittyvät mitokondrion prosesseihin, olivat aliekspressoituneita IOBA-NHC näytteissä, jotka oli altistettu BAC-säilöntäaineelle. Lisäksi, molemmissa solulinjoissa kolesterolin tuotanto ja sitä kautta solukalvon läpäisevyys saattaa muuttua alentuneen HMGCS1:n johdosta. Lisäksi ainakin HCE-solujen tuloksissa oli viitteitä nousseeseen aktiinitukirakenteiden supistusten määrään, sillä MYH9, MYL12A ja MYL6 olivat yliekspressoituja näytteissä, jotka oli altistettu BAC-säilöntäaineelle. Tämä voi osaltaan vaikuttaa solujen liitosten läpäisevyyteen. Nämä potentiaaliset biomarkkerit tullaan analysoimaan tarkemmin tulevien kliinisten tutkimusten avulla.

Abbreviations

Abbreviation	Meaning
ACG	Angle closure glaucoma
ANOVA	Analysis of variance
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
BAC	Benzalkonium chloride
BP	Biological process
CAI	Carbonic anhydrase inhibitors
CC	Cellular component
DAVID	Database for Annotation, Visualization and Integrated Discovery
DIA	Data-independent acquisition
ES	Enrichment score
FDR	False discovery rate
GO	Gene ontology
GOrilla	Gene Ontology enRIchment anaLysis and visuaLization tool
GSEA	Gene set enrichment analysis
HCE	Human corneal epithelial cells
HK2	Hexokinase-2
HMGCS1	Hydroxymethylglutaryl-CoA synthase, cytoplasmic
IOBA-NHC	Immortalized normal human conjunctival epithelial cells
IOP	Intraocular pressure
IPI	International protein index
iTRAQ	Isobaric tag for relative and absolute quantitation
KEGG	Kyoto encyclopaedia of genes and genomes
LC	Liquid chromatography
MF	Molecular function
MI	Mutual information
MLC	Myosin light chain
MRM	Multiple reaction monitoring
MS	Mass spectrometry
MYH9	Myosin-9
MYL12A	Myosin regulatory light chain 12A
MYL6	Myosin light polypeptide 6
NDUFA5	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 5
NDUFS3	NADH dehydrogenase [ubiquinone] iron sulphur protein 3, mitochondrial
OAG	Open angle glaucoma
OSD	Ocular surface disease
ROS	Reactive oxygen species
SRM	Selected reaction monitoring
SWATH	Sequential window acquisition of all theoretical spectra
WST-1	Water-soluble tetrazolium salt

Table of Contents

1 Introduction

Glaucoma is a progressive optic neuropathy characterized by structural and functional changes in ganglion cell axons of the optic nerve (Salim, 2012). These changes can eventually result in a loss of vision if the condition goes untreated. The underlying causes of glaucoma are still largely unknown, but it is known that high intraocular pressure (IOP) is one of the main risk factors of glaucoma (Lang, 2007). Hence, glaucoma is usually treated by using mainly topical medication, laser therapy or surgery, which aim to decrease IOP. Currently approximately 60 million people are affected by glaucoma and this figure is expected to rise considerably in the future (Quigley & Broman, 2006).

Topical treatment of chronic glaucoma can in many cases have consequences that reduce the efficacy and safety of the medical therapy. Such consequences include ocular surface changes (Pisella et al., 2002), cataract development (Chandrasekaran et al., 2006) and topical anti-glaucoma treatments are also connected to increased risk of failure in other types of therapies such as surgery (Broadway et al., 1994). Furthermore, as also described by Pisella et al. (2002) some patients are known to be more sensitive to the drugs and their adjunctive agents such as preservatives, which are present in the topical treatments. In addition, the use of preservative containing topical medications has been associated to the development and worsening of ocular surface diseases (OSD) (Zhou & Beuerman, 2012). Especially benzalkonium chloride (BAC), which is the most common preservative used in eye drops, has been linked to adverse effects in some patients as it is known to not only protect the ophthalmic solution from microbial contamination but also to be cytotoxic to cells of the surface of the eye (Badouin et al., 2010). Several studies have demonstrated the apoptotic effects of BAC on various cell lines (Epstein et al., 2009; Furrer et al., 2002), yet the pathways leading to this effect are not fully known.

The aim of this study was to examine the proteomic profiles of epithelial cell lines, which were exposed to either BAC-containing or preservative-free treatments and to identify individual proteins which play a role in the cellular reactions caused by preserved glaucoma medication in the cell lines. More specifically, corneal and conjunctival epithelial cells were exposed to preserved latanoprost, unpreserved tafluprost and preservative BAC. By identifying some statistically significant protein expression level changes between the treatments, it was hoped that some further information could be gained from the underlying pathways which cause cell apoptosis and other adverse effects in cells which are exposed to preservatives. In addition, enrichment analysis was applied once differentially expressed proteins were identified to further broaden the understanding of biological aspects. The results obtained from this study could be later on validated in further studies.

2 Review of literature

Many patients suffering from an ophthalmic condition are likely to be subjected to various topical medications during the time of their treatment, which for example in the case of glaucoma, can last for several years or even decades (Freeman & Kahook, 2009). Many topical treatments, most importantly multiple-dose eye drops, contain some form of preservatives or a mixture of them. The benefits of added preservatives include limiting the microbial proliferation and preventing any unwanted alterations in the formula during the time of use (Baudouin et al., 2010). However, preservatives are often also cytotoxic for the epithelial cells of the ocular surface and due to this, patients who are subjected to long-term medication containing preservatives, may also experience unwanted side effects (Pisella et al., 2002). These adverse effects are often allergic or inflammatory symptoms varying from redness, stinging, burning, irritation and eye dryness to occasionally conjunctivitis or corneal damage (Baudouin et al., 2010). Not only is the duration of exposure a factor, but it has further been explained that the unwanted symptoms appear to be proportional to the concentration of the preservative, i.e. a larger proportion of a preservative often means a greater reaction (Uusitalo et al., 2010). Recent studies have shown that by switching from a preserved multiple-dose eye drops to preservative-free formulations, patients who suffered from adverse effects during the use of glaucoma medication containing preservatives, have experienced reduction in their allergic and inflammatory signs and symptoms after switching to preservative-free medication, whilst still maintaining a lowered level of IOP (Uusitalo et al., 2010; Pisella et al., 2002).

2.1 Glaucoma

Glaucoma is a term describing a group of neurodegenerative diseases causing progressive optic neuropathy due to loss of retinal ganglion cells and it is the second leading cause of blindness worldwide (Quigley & Broman, 2006). In a majority of glaucoma cases, the aqueous humor formed by the ciliary body is not successfully drained from the eye through trabecular outflow pathways as it should, resulting in unusually high IOP levels in the eye (ocular hypertension) which causes optic neuropathy (Wiggs, 2007). Most commonly, glaucoma takes place later in life as the risks to obtain glaucoma are significantly increased after the age of 40 and early-onset glaucoma is less common (Lang, 2007).

The two main types of glaucoma, i.e. open angle glaucoma (OAG), which is the most predominant one of all glaucoma groups, and angle closure glaucoma (ACG) are affecting over 60 million people

worldwide, and it is estimated that the number of affected individuals will increase to 79.6 million by 2020 (Quigley & Broman, 2006). OAG and ACG are differentiated by the status of the angle between iris and the cornea and in OAG this angle is open but the aqueous humor outflow through trabecular meshwork has gotten blocked slowly overtime and the normal outflow of humor is this way prevented (Lang, 2007). The slow development of this blockage is the reason why many patients only discover noticeable symptoms later on in the process of the condition when it has already developed further and some level of vision loss has already occurred (Kroese & Burton, 2003). Figure 2.1.1 visualizes the effects of the blockage in the eye.
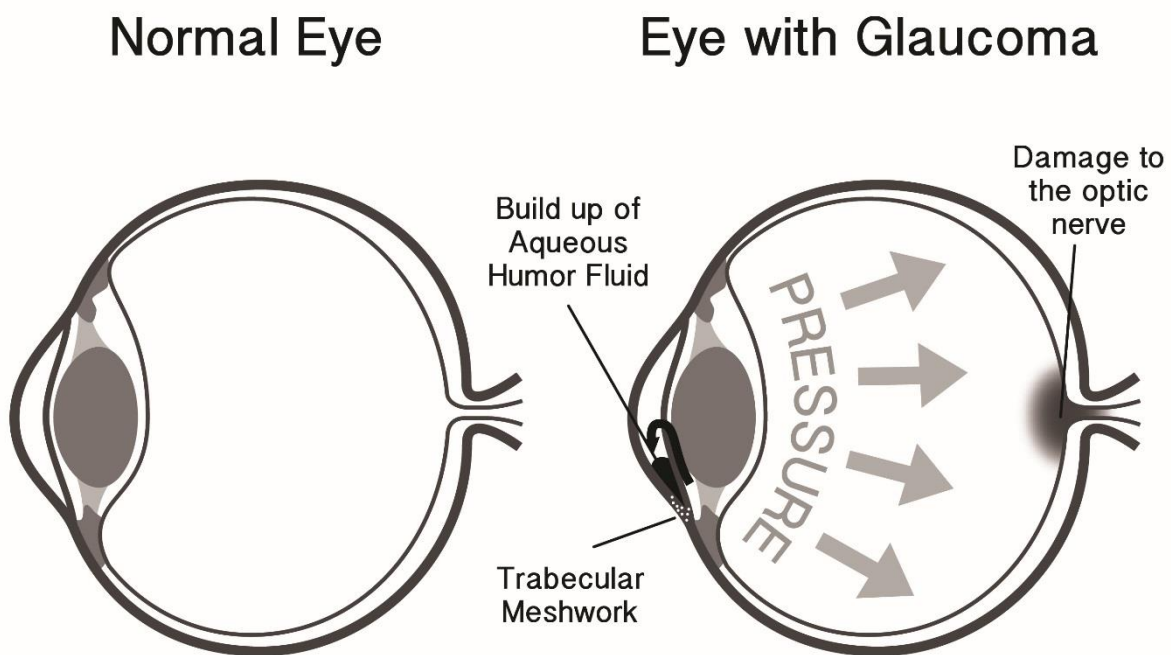
Figure 2.1.1 Comparisons of normal eye and eye with glaucoma (modified from a picture by Thomas Bond & Partners, 2013). In the eye with glaucoma, the trabecular meshwork has been blocked preventing the aqueous humor flow, which results in pressure in the eye and hence damage to the optic nerve.

In ACG the blockage is more sudden and it is caused by a quick increase in the IOP which is causing the iris to occlude the trabecular meshwork causing blurred vision, severe eye and head ache, nausea and sudden sight loss are often quickly noticed by the patient (Lang, 2007). This type of glaucoma requires immediate treatment. The discussion of glaucoma in this study will refer usually to primary OAG, which instead of ACG requires long-term topical treatment of the condition, even though some aspects may be applicable to other types of glaucoma as well.

The word "primary" or "secondary" added in the front of the glaucoma type refers to the cause of glaucoma. Primary refers to glaucoma that is not caused by any other ocular disorders and secondary indicates that the glaucoma may be a result of another ocular disorder or a side effect from another treatment (Lang, 2007; Kroese & Burton, 2003).

One of the main and most well-known risk factors of glaucoma is elevated IOP, though it is no longer considered the only distinctive factor as it has been discovered that the presence or absence of intraocular hypertension does not have a direct causal relationship with glaucoma and it is also possible to develop glaucoma where the IOP is considered normal (Noecker, 2006; Lang, 2007). In fact it is thought that 25%-50% of POAG patients have what is considered a normal IOP (Kroese & Burton, 2003). Normal IOP in adults is approximately 15 mmHg and the threshold for intraocular hypertension is considered to be at 21 mmHg.

Other risk factors connected to glaucoma include old age, African origin, myopia and family history of glaucoma (Kroese & Burton, 2003). Additionally certain medical conditions, e.g. diabetes, high blood pressure are considered risk factors and furthermore, vascular dysregulation is considered to be linked to glaucoma (Lang, 2007). Despite the knowledge of risk factors, the underlying causes of glaucoma are still largely unknown and therefore instead of trying to identify the general risk factors, the research now concentrates on finding particular genes and proteins, which could be responsible for causing the condition. The identification of these genes and proteins could then help to develop methods of treatment and even prevention of glaucoma.

Recent studies have identified some of the genes which are thought to be associated with glaucoma and early-onset glaucoma can be inherited as a mendelian autosomal-dominant or autosomal-recessive trait through these genes (Wiggs, 2007). However, the adult-onset glaucoma does not often exhibit mendelian inheritance patterns but instead, the condition is a result of interactions between multiple genetic factors and the environment also plays a role in the development of the disease (Wiggs, 2007). Some further distinctions can be made between the genes associated to POAG. For example, mutations in myocilin (MYOC), WD-repeat domain 36 (WDR36) and optineurin (OPTN), can cause POAG alone without any further influence from other genes or risk factors (Fingert, 2011). Other genes associated with POAG can instead be considered risk alleles, which in the case of mutations can promote the development of POAG but these genes are not solely responsible for its development.

2.2 Treatments of glaucoma

To this date glaucoma is not curable, but through early detection and correct treatment, most often topical medication, the development of glaucoma progression towards blindness can be halted or delayed (Noecker, 2006). The main aim of topical glaucoma medication is to lower the IOP and stop any further glaucomatous damage from occurring in the optic nerve.

There are different ways the reduction in IOP can be achieved and the main functions are listed in an article by Noecker (2006). First approach is that the production of aqueous humor fluid is inhibited and the ocular hypotensive agents achieving this include beta-blockers and carbonic anhydrase inhibitors (CAIs). Alternatively, with prostamides, prostaglandin analogs or parasympathomimetic drugs the trabecular or uveoscleral outflow can be increased. The trabecular outflow is the conventional method for the eye to remove the excess aqueous humor into the canal of Schlemm, whilst uveoscleral outflow, which is not as effective (Lang, 2007), happens through the ciliary body. Furthermore, some agents like α-adrenergic agonists work by combining both of the aforementioned methods of reducing IOP, i.e. reduction of aqueous humor production and increase in outflow. It is worth noting that occasionally it may be beneficial for the patient to tackle the increased IOP by combining multiple medications with different effects (Noecker, 2006).

2.3 Preservatives

Preservatives can cause inflammatory and allergic reactions with some patients. These effects are often caused by damage to the cells first in contact with the eye drops, i.e. corneal and conjunctival epithelium cells, due to the cytotoxic nature of the preservatives. Despite some unwanted side effects, the preservatives provide essential protection for the formulation from any microbial contamination, which may occur via the patient's hands or by other surface areas of the patient whilst applying the medication. Further on, the preservatives may help the drug to maintain its potency and prevent any possible biodegradation. However, this type of protection is considered to be the task of stabilizing agents and not preservatives. (Furrer et al., 2002)

Preservatives can be divided into two main groups: detergents and oxidizing preservatives (Noecker, 2001), however many studies increase the number of ophthalmic preservative classes up to four (Epstein et al., 2009): detergents, oxidants, chelating agents, and metabolic inhibitors. There are

additional division methods based on the preservatives' chemical classes (Furrer et al., 2002). Only the first two major groups, detergents and oxidants, are explained in some further detail.

The detergents work by altering the lipid component of cell membranes of the affected (microbial) cells causing membrane instability (Noecker, 2001). This is different from oxidizing preservatives, which instead enter the cell and alter the lipids, proteins and DNA elements inside (Noecker, 2001). Both of these methods promote lysis of plasma membrane, inhibition of cellular metabolism, oxidization or coagulation of cellular constituents or promotion of hydrolysis (Noecker, 2001). The method of action with oxidizing preservatives is considered less drastic compared to detergents, however, with sufficiently large doses, both types of preservatives are capable of causing cytotoxic effects in eukaryotic cells, leading to inflammation. Since the evidence of harmful effects of preservatives, and especially detergent preservatives, has piled up, many new approaches have been developed that attempt to tackle the issues with cytotoxic effects, e.g. preservative-free and sustained-release medications (Kaur et al., 2009). Another new approach is the sofZia (Alcon) preservative system, which contains chemical substances, which are not cytotoxic to ocular surface cells but still maintain antimicrobial environment in the solution (Kaur et al., 2009).

2.3.1 Benzalkonium chloride

Benzalkonium chloride (BAK or BAC) is one of the most commonly used preservative in topical ophthalmic medications. It is classified as a quaternary ammonium compound composed of a mixture of alkylbenzyl-dimethylammonium chloride homologues (Epstein et al., 2009). Based on the division discussed in the earlier section, BAC could be classified as a detergent-like substance and it is considered highly effective due to its ability to efficiently prevent microbial contamination by protein denaturation and lysis of cytoplasmic membranes (Noecker, 2001), whilst also affecting the cell membrane permeability by allowing the ingredients in the medication to enter the anterior chamber by breaking cell-cell junctions in the epithelium (Kaur et al., 2009).

However, BAC is also known to be interrupting the metabolic processes of the cell, causing lysis of the cell contents and allowing vital substances to escape the cell (Epstein et al., 2009). Further on, there is evidence showing that BAC induces necrosis and apoptosis in bacterial cells, in concentrations of 0.05-0.1% and 0.01% respectively, by disturbing their plasma membrane, as desired. Unfortunately these effects can be very similar in human ocular surface cells (Kaur et al., 2009; Baudouin et al., 2010). BAC also has a tendency to interrupt cell mitosis (Guo et al., 2007) and

reduce the tear film breakup time which reduces the tear film stability. This instability can heighten the risk for adverse effects in particular with patients suffering from dry eye syndrome, and this is particularly prevalent with patients also suffering from glaucoma as they have a decreased rate of basal tear turnover (Kaur et al., 2009). Therefore, patients suffering from both of the conditions mentioned above, i.e. dry eye and glaucoma, are highly susceptible to encounter the adverse effects of BAC.

In a study by Epstein et al. (2009) cytokines in BAC-treated cells were quantified via enzyme linked immunosorbant assays cells and it was shown that cells treated with BAC contain significantly increased quantities of two well-known inflammation biomarkers; interleukin (IL-) 1 and tumor necrosis factor (TNFα). In addition, some other inflammation-related markers increased moderately. Further on, some *ex vivo* observation studies on rabbits have shown that even after a short period of exposure, some corneal and conjunctival damage starts occurring and these effects are more severe when higher concentrations of BAC are applied (Furrer et al., 2002). The epithelial cells in the eye can be damaged by the cytotoxicity of BAC and this effect can be notable since the epithelial cells form the protective barrier in the surface of the eye (Guo et al., 2007).

2.4 Epithelial conjunctival and corneal cell lines

2.4.1 Conjunctival epithelial cell line (IOBA-NHC)

The conjunctiva is a smooth, continuous membrane which lines not only the inside of the eyelid (palpebral conjunctiva) but it also covers the sclera around the cornea (bulbar conjunctiva). This can be seen illustrated in Figure 2.4.1.1. The conjunctival epithelium helps to maintain healthy ocular surface and physiological changes in it are thought to be connected to inflammatory diseases of the ocular surface (Brasnu et al., 2008a). The process of collecting human biopsies is one of the methods of obtaining conjunctival epithelium samples. However, these samples are often not ideal for research purposes (Brasnu et al., 2008b) and instead, immortalized cell lines are often very popular in research studies due to their ease of access and quick cell growth. The reproducibility is also considered better than with primary cultures (Brasnu et al., 2008b), which could be considered another alternative to cell lines.

The IOBA-NHC cell line was characterized by Diebold et al. (2003). This spontaneously immortalized cell line is commonly used in studies examining and comparing toxicity profiles of varying topical medications (Diebold, 2003; Brasnu et al., 2008b). In studies examining the effects

of BAC in epithelial conjunctival cell it was noted that cells showed signs of caspase-dependent and -independent apoptosis, oxidative stress, increase of cell membrane permeability and cell shrinkage and blebbing amongst other typical "symptoms" (Clouzeau et al., 2012; Buron et al., 2006). According to a study by Pellinen et al. (2012), IOBA-NHC cells appear to be more sensitive to the effects of BAC in comparison to HCE cells.
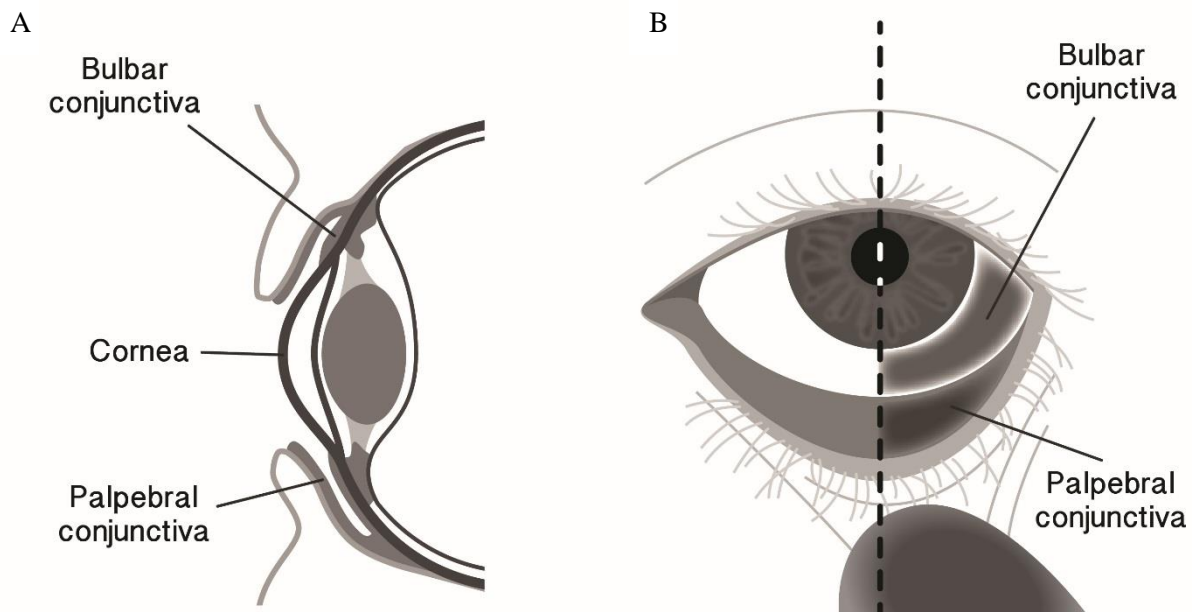
Figure 2.4.1.1 The structure of the eye from the side (A) and front (B). The conjunctiva layers the sclera around the cornea and the inside of the eyelid. Cornea covers the iris and lens. (Modified from a picture by Azari & Barney, 2013)

2.4.2 Corneal epithelial cell line (HCE)

The cornea, is an area covering the iris and bordering the sclera. The cornea is built up from several parts and the most anterior layer is the corneal epithelium (see Figure 4.2.1.1). The purpose of the corneal epithelial cells is the protection of cornea and hence this is another interesting cell line to study in relation to the cytotoxic effects of BAC. As discussed in a study by Guo et al. (2007), BAC has a tendency to accumulate in the corneal epithelium as it does not penetrate well through it and hence BAC has various effects on this barrier. These effects include the disruption of the barrier function, reduced wound healing and interruption of mitosis along with the usual BAC induced effects already described earlier. The same study evaluated the effects of BAC in relation to myosin light chain, which controls the barrier integrity, adhesion and migration.

The epithelial corneal cell line HCE was established and immortalized by Araki-Sasaki et al. (1995) by infecting primary cultured human corneal epithelial cells with a recombinant sv-40-adenovirus vector (Huhtala et al., 2002). HCE cell line retains the properties (e.g. well-developed desmosomes and abundance of microvilli) of normal corneal epithelial cells. Furthermore, a study comparing the cytotoxic effects between HCE and rabbit primary corneal epithelial cell line reached similar conclusions and it stated that the HCE "is a better model for studies of the corneal toxicity of drugs" (Huhtala et al., 2002).

2.5 Mass spectrometry and proteomics

Genomics has developed into cheaper and faster research topic than it was just a decade ago and it continues to develop at a rapid pace and due to this, it is a popular research area. However, where genomics can be used to investigate the DNA structure and expression, and this way explain phenotypes that present themselves, proteomics does this using proteins and the peptides forming them. It could be argued, that in medicine development and in personalized medicine in particular, the proteins play an even larger part than the genomic information available since the proteins are essentially the end product of the genes and these are the elements in biological systems that the drugs are essentially used to target. Many things may occur between the gene and finished gene product, a protein, and only the end product is what actually causes effects to take or not to take place. Hence, though genomics is naturally still an important aspect of "omics" research, proteomics could also be expected to become growingly interesting and popular starting point for research, especially in medical research. (Schmidt et al., 2014; Noble & MacCoss, 2012; Kumar & Mann, 2009)

In proteomics, proteins in given samples can be identified and quantified and there are many methods that can be applied to the samples depending on the information requirements and interests. For example, it could be of interest to just identify which proteins are present in a given sample, or it could be more useful to also obtain the expression levels of the proteins identified in a sample. As the quantification processes have evolved quickly in the past years, these days more and more often the output from proteomic experiments is more than just a list of proteins. (Kumar & Mann, 2009)

The main core of the experiments often include liquid chromatography (LC) coupled with mass spectrometry (MS). The general workflow of a shotgun MS experiment, can be considered to consist of three major parts as explained by Noble & MacCoss (2012). In the first part, the proteins are isolated from a mixture, i.e. sample, and they are digested into peptides using a protease. Next, in

order to reduce the complexity, liquid chromatography is applied, which separates the peptides based on their chemical properties. Finally, third step includes tandem mass spectrometry and the "tandem" here refers to the two rounds of mass spectrometry which are applied at this step. At this stage mass spectrometer first selects several peptides for fragmentation from the liquid chromatography based on an initial analysis of distinction, and this is more specifically referred to as LC-MS. These chosen peptides are then processed individually so that fragmentation spectra, referred to as daughter ions, of the subpeptides are gathered. These spectra are characterized by mass-to-charge ratio (m/z), retention times and intensity (Noble & MacCoss, 2012) and each individual spectrum helps then to identify the original peptides, or parent ions, which were fragmented.

The peptide identification can be achieved with a variety of methods: database search, de novo spectrum identification, tag-based methods or library search and of these, the database search is the most commonly used (Noble & MacCoss, 2012). Next, and in some cases finally, the protein identification takes place. Again, there are programs and different methods available for the execution of this part and it should be noted that one main complication in this stage is that some peptides are so-called degenerate peptides, which means that they may be present in several proteins and naturally this can complicate process down the line if not accounted for (Noble & MacCoss, 2012).

In order to produce different types of proteomics data, several different methods have been developed to meet the needs. For example initially, when no prior information is available of a given sample or proteomic profile, shotgun proteomics can provide a good starting point as this takes a "discovery"-approach to proteomics and the output consists of high-throughput data. In this process, a very large number of proteins is essentially identified from complex mixtures and the deepest possible coverage of the proteome can be achieved. (Schmidt et al., 2014)

Another one of the main techniques in proteomics is the targeted MS which is essentially the other traditional approach next to shotgun proteomics. In some cases the researcher may already have a fairly good idea of the proteins which are of interest in a particular study. Targeted MS methods include selected and multiple reaction monitoring (SRM and MRM respectively). Having prior information of the proteins of interest can be taken advantage and by focusing on these proteins and peptides alone, the sensitivity and reproducibility of the method can be increased significantly. It should be noted that targeted and shotgun proteomics can naturally be both performed in the same study in order to obtain maximal amount of information of given samples. (Schmidt et al., 2014)

In shotgun proteomics proteins obtained from a sample of interest are first fragmented into smaller and smaller subparts and the final goal is then to solve what these subparts were in the beginning of

the process. The reason for this process is that currently, at least with most MS machinery, complete proteins are simply too large to be processed as they are without very expensive tools and hence the fragmenting and defragmenting is necessary (Noble & MacCoss, 2012). The whole process of MS techniques is further visualized in Figure 2.5.1.
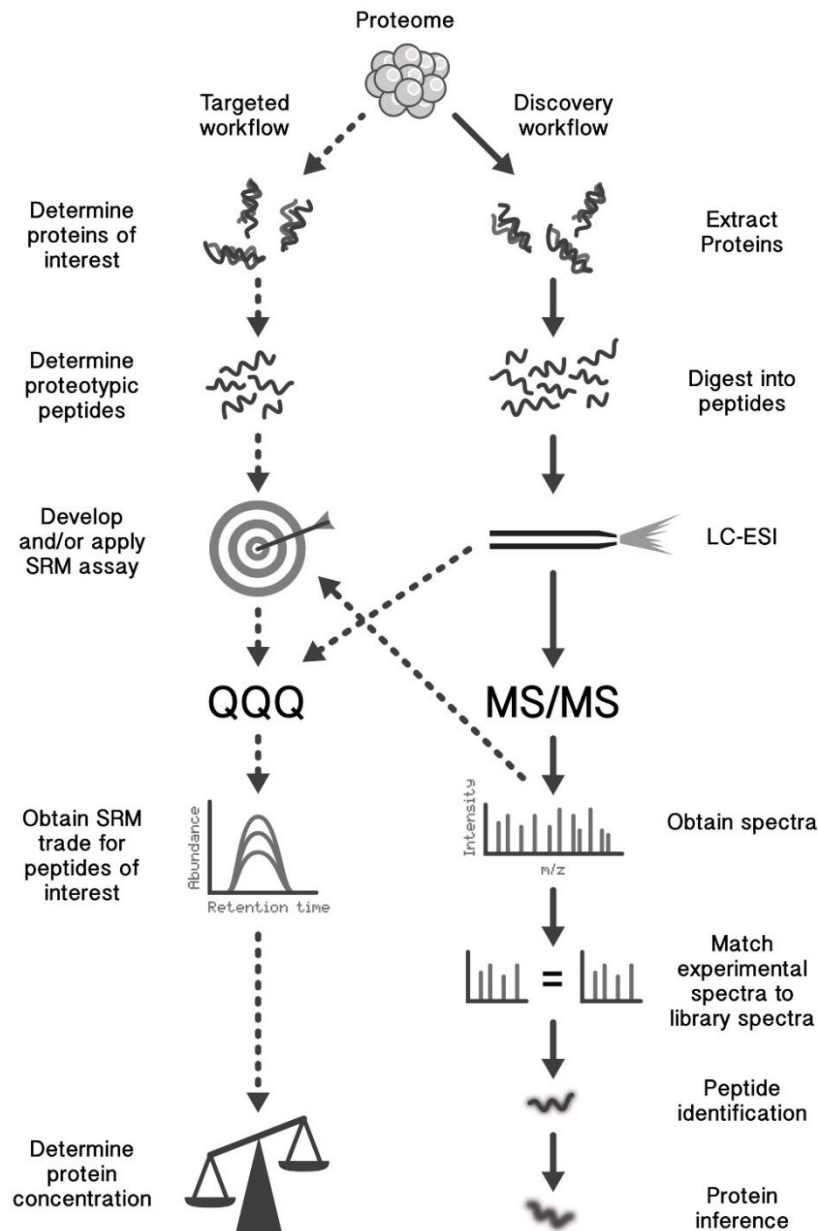
Figure 2.5.1 Shotgun and targeted proteomics workflow (modified from a picture by Perez-Riverol, 2014). The track on the left illustrates the targeted workflow where the proteins of interest are first identified and, after SRM assay or similar has been developed, the concentration of these proteins can be determined. On the right, the discovery workflow shows how all proteins are extracted and digested and a spectra is obtained after MS/MS. Once peptides have been identified, e.g. via library search, the proteins can be quantified.

One of the drawbacks of shotgun proteomics is sensitivity and reproducibility, or the lack of it (Schmidt et al., 2014) and this can be overcome by implementing e.g. MRM or other targeted method instead. Additionally, the method described above, i.e. shotgun proteomics only identifies the existence of proteins although this alone can already provide some indication of the abundance levels (Kumar & Mann, 2009). Techniques have been developed which allow quantification in shotgun as well as targeted proteomics. These techniques are a collection of approaches including stable isotope labelling, spectral counting and peptide chromatographic peak intensity methods. As an example, in stable isotope labelling, e.g. isobaric tag for relative and absolute quantitation (iTRAQ) labelling, a heavy isotope label is added into some of the samples and this enables absolute rather than relative quantification. Semi-quantitative approaches are not widely used as label-free quantification methods are becoming more and more popular nowadays (Kumar & Mann, 2009). Label-free methods are not considered as precise as isotope labelling, which however is still complicated, expensive and limit the sample number (Huang et al., 2015).

Recently, new MS methods have been produced and the one method of interest in particular is SWATH acquisition, which is a data-independent acquisition (DIA) strategy using MS/MS data (Perez-Riverol, 2014). SWATH essentially combines the high throughput aspect of shotgun proteomics and the SRM's ability to produce accurate, complete and reproducible data (Collins et al., 2013). In SWATH, the peptides are continuously fragmented in fixed windows and the results are matched to a spectral library, which has been previously produced using shotgun proteomics (Huang et al., 2015). The resulting output, as already mentioned, is a high throughput data with high accuracy and it depends on the prior information (library) obtained in an earlier stage.

2.6 Statistical methods with quantified proteomics data

Once quantified proteomics data have been produced, the next step is the statistical analysis. Following subsections will discuss different statistical methods, which can be implemented to this type of data.

2.6.1 Reduction of bias

All measurement data are subject to degree of bias and noise. This can be caused by several systematic and random measurement errors and hence, before proceeding to any differential expression analysis,

these aspects should be evaluated and, if necessary, adjusted. In addition to bias, data can have extreme measurements present, which can skew the results if not accounted for correctly. (Callister et al., 2006; Karpievitch, Dabney, & Smith, 2012)

One way to reduce and smoothen the effects of extreme measurements is data transformations. For example $\log_2$-transformation, where each observation is converted to $\log_2$-scale, has several advantages. As explained in an article by Callister et al. (2006), "it converts the distribution of ratios of abundance values of peptides into a more symmetric, almost normal distribution". Furthermore, the article goes on to describe that this type of transformation also allows the use of many robust normalization techniques developed for this type of data, as it reduces the leverage of a low number of highly abundant species on the regression analysis used by these robust techniques. For these reasons, this technique is also very commonly used with microarray data.

Central tendency normalization, also known as global adjustment, is a method that is implemented in order to reduce systematic bias between biological replicates and it is quite common in proteomics data analysis. This method essentially subtracts a chosen measure, e.g. mean or median, from each observed value (1), which in other words means that "distribution of the log intensity values to center around a constant such as mean, median or some fixed value for each sample" (Karpievitch, Dabney, & Smith, 2012). In this particular analysis, median of each biological replicate is subtracted from each ($\log_2$-transformed) individual observation of that biological replicate. This then results in the assumption that a large amount of the protein abundances remain unchanged. The formula below shows the process:

$$x'_{i,j} = x_{i,j} - \mu_j, \tag{1}$$

where $x_{i,j}$ is the ith protein abundance value in the jth sample, μ_j is the median of the jth sample and $x'_{i,j}$ is the resulting normalized value. This method further has benefits when it comes to the differential expression analysis as the observed protein abundances are now centered around zero as a result.

Ideally, this method further enables the differentially expressed proteins to become more identifiable, whilst differences between random fluctuations are "smoothened". However, it should be noted that unfortunately all normalization will at the same time also result in loss of some information (Hu & He, 2007). The aim is to minimize the bias, whilst also making sure that the loss of actual information is kept as small as possible. It should be in addition noted, that this sort of normalization does not

eliminate systematic trends which sometimes occur in the data (Karpievitch et al., 2012) and therefore MA-plots can provide importantly further information of any possible underlying biases.

2.6.2 Technical replicates

Technical replicates are more specifically replicates, which have been obtained from the same biological replicate, i.e. sample, and hence they cannot be though to be separate observations and are not considered fully independent of each other. Ideally, a data should consist of several biological replicates, which could be considered independent, as the scope of conclusions can become very limited when approaching data with only technical replicates as mentioned in the review article by Cui and Churchill (2003). This aspect should be accounted for in the pre-processing and in choice of statistical methods in particular.

It should be noted that there are different approaches that can be taken with technical replicates. More specifically, it is possible to keep the technical replicates separate or alternatively take a chosen statistic, often arithmetic or geometric mean of the technical replicates, given that there are no large differences between them. The variability between technical replicates can be evaluated for example by using MA-plots or correlation measures. Whatever the decision is with the pre-processing of technical replicates, it should be noted that these cannot be treated as independent observations of each other similar to biological replicates. Hence, as done in this work, when the technical replicates are kept separate, the statistical methods used should be capable of taking the non-independence into account as well by applying more complex methods, such as mixed-model analysis of variance (ANOVA). (Cui & Churchill, 2003)

2.6.3 Differential expression

After establishing the initial quality of the data and performing the necessary transformations and normalizations, the next step is to establish, if any statistically and biologically significant differences occur between the cellular states (Kumar & Mann, 2009). Several statistical tools can be implemented here depending on the structure of the data and research question of the study.

Mixed-effects model is a statistical model, which could be considered a more evolved version of the repeated measures ANOVA in the sense that in addition to the ability to take into account repeated or connected measures, it can also account for various other aspects such as a nested structure. As

described by Cui and Churchill (2003) mixed models are one option in the case where data is constructed hierarchically or there are non-independent replicates involved. Mixed models treat some of the factors in an experimental design as random samples from a population and these factors are modelled as sources of variance.

2.6.4 Multiple testing correction

One issue, which should be taken into account before identifying the differentially expressed proteins, is the multiple testing issue. For example, when around 2,000 proteins are quantified, the testing for differential expression will be also done approximately 2,000 times since each protein is tested individually. This means that we could expect to observe approximately 100 p-values below 0.05 just by chance. These would be considered false positives if no true underlying difference was indeed present and the p-value had occurred just by chance. Hence, as the number of tests increases, so does the number of false positives. These issues, more specifically multiple comparison issues, should be controlled carefully. (Gutstein et al., 2008)

One option to account for the multiple testing is to adjust the obtained p-values using one of the common multiple testing corrections, e.g. Bonferroni correction and Benjamini-Hochberg correction. The two methods mentioned above are perhaps the most commonly used correction methods and the Bonferroni correction is often considered far too conservative for this type of data, resulting in an increased number of false negatives (Gutstein et al., 2008). Benjamini-Hochberg, which allows to control the FDR, will produce corrected p-values, also known as q-values, which will tell the proportion of false positives observed in the data when a given threshold is chosen. For example, with a q-value threshold of 0.1, 10% or fewer are expected to be false positives (Gutstein et al. 2008). The classical threshold for statistical significance is often 0.01 or 0.05.

2.6.5 Thresholds

Once the coefficient estimates or other measures relating to biological differences and corresponding p-values, with necessary adjustments, are obtained for all of the quantified proteins, it is time to evaluate, which one of them could be considered interesting, i.e. statistically significant.

Two aspects should be taken into account: fold change and the adjusted p-value. The first, fold change, tells about the effect size of the difference, i.e. how far apart two mean values are from each

other. However, when the variability between individual points in a group is large, this may skew the fold change estimates and produce extremely large fold changes even when no true statistically significant difference exists between groups. Hence, the p-values, or q-values, provide further valuable information at the same time. Yet, just relying on p-values alone is not desirable as the differences detected this way could be so small that no biological significance could be derived from the results (McCarthy & Smyth, 2009). Traditionally, fold changes above 1.5 or 2 are considered interesting. A fold change of 2 means that the quantity changes two-fold, i.e. in layman's terms it doubles.

2.6.6 Network reconstruction

Network reconstruction can be used to establish theoretical connections between the proteins of interest. Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) is one such algorithm, which can be used for the process of network reconstruction (Margolin et al. 2006).

ARACNE uses expression profiles and the original article by Margolin et al. (2006) used microarray expression profiles to demonstrate its functions. As explained in the article, it works by first establishing mutual information (MI) measures for pairs of proteins. Mutual information measure tells how much information one variable, here a protein, contains about another variable, i.e. another protein (Cover & Thomas, 1991). The MIs are then filtered based on a chosen threshold computed for a specific p-value (Margolin et al. 2006). Next, the algorithm removes indirect candidate interactions and this step bases heavily on the data processing inequality (DPI). This step enables that the number of false positive interactions is reduced, which could otherwise arise due to the co-regulation of genes, or proteins in this case. The resulting networks can then be visualized using for example Cytoscape or other similar tool.

2.6.7 Enrichment analysis and tools

Once the differentially expressed proteins are identified using a chosen method and thresholds, the next question is: what do these interesting proteins have in common and are they related to any specific pathway etc. in the cell, which could explain the changes we are for example clinically observing? Enrichment analyses essentially provide information about any chosen enriched GO terms or pathways. More specifically, the algorithm identifies, by performing hypergeometric-based tests,

terms which are containing so many of the interesting proteins, that it can be considered statistically unlikely that this could occur by chance (Falcon & Gentleman, 2007). Hence, it could then be expected that the changes observed in the cellular systems and such are somehow connected to a given enriched GO term or for example KEGG term (Kyoto Encyclopedia of Genes and Genomes, a biological pathway database) term.

There are few notable further aspects in the GO terms. First, they are ordered hierarchically and hence there can be highly similar terms showing up as enriched. In these cases it is up to the user to define, which is the most descriptive term which could be used and in addition the number of genes and enrichment score (ES) should be kept in mind. Secondly, the GO annotations are divided into three main groups: biological process (BP), cellular component (CC) and molecular function (MF) (Huang et al., 2009 (1)). All of these will be checked in the analyses further on.

There are different approaches to performing enrichment analysis. One alternative is that a list of interesting proteins, often referred to as candidate proteins, and a background set, also known as "protein universe", are required. This approach is more commonly referred to as singular enrichment analysis (SEA) (Huang et al., 2009 (1)). Here the list of interesting annotations therefore includes annotations obtained from a differential analysis, which have satisfied chosen thresholds. Further on, the background, is usually the list of all proteins in a given data, i.e. both the candidate proteins and the proteins which are not considered differentially expressed. However, the background could also be considered to be the total species-specific genome or proteome. In any case it should be noted that the choice of the background also has an effect on the results (Falcon & Gentleman, 2007).

Another approach to enrichment analysis is that the user first orders the full list of quantified proteins based on some obtained measure, this method is referred to when discussing gene set enrichment analysis (GSEA) (Subramanian et al., 2005; Mootha et al., 2003). This measure could for example be a correlation coefficient or indeed $\log_2$ fold change as is available in this analysis. Here it should be noted that the limits between interesting and not interesting proteins are not as clear as in the earlier approach and results are likely to be different. Whichever approach is chosen, the multiple comparison issue is encountered again and this is why most tools provide several p-value measures for the enriched terms and it is up to the user's discretion which adjustment and threshold is used (Huang et al., 2009 (1)).

3 Aims of the study

This study had more than one aim. Firstly, it was performed in order to merely test if, and how well, this type of study could be performed successfully. More specifically, it was of interest to examine how immortalized cell lines would respond to these treatments with these specific concentrations and how the resulting proteomics would turn out. This type of study has not previously been performed using MS SWATH method.

The aim of this study was not to tell that all preservatives are bad and should not be used in any topical treatments of glaucoma. Instead, it should merely be acknowledged that there is a subgroup of patients who are currently not benefitting from the medication provided for them and other alternatives could be more favorable for them. In addition, it could be expected that in the future, new and hopefully less harmful preservatives are developed to replace the current ones at least partially, whilst also the popularity of preservative-free treatments could also be expected to rise. By examining the pathways and effects of BAC, perhaps some new approaches to the topical medications and their ingredients could be revealed.

The second aim of this study, which is what this paper is mostly concentrating on is, what proteins are showing different abundance levels between the treatments and what could explain these differences in a larger, biological scale. It was hoped, that once these proteins, and their place in the vast network of biological processes was established, a few chosen ones could be further verified in following studies. Once a confirmation of the effects of BAC in a proteomic level can be achieved, it can assist in the development of new medication or in fact new preservatives, which could be beneficial
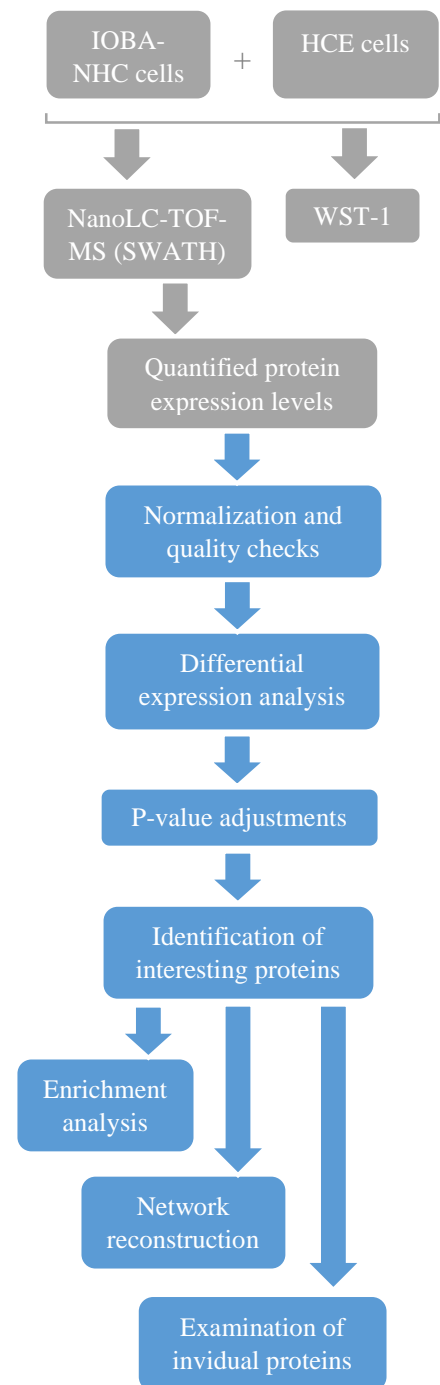
Figure 3.1 Flowchart of the steps taken in this study. The tasks in the grey boxes were done prior to the start of this thesis and the blue boxes show the tasks described in more detail in this study. Some tasks are excluded from the start.

26

particularly to those patients needing long-term ophthalmic topical treatment who are sensitive to this very common preservative.

The steps taken to achieve the goals of this study are visualized in Figure 3.1, where it can be seen in more detail, which steps/tasks were carried out prior (grey boxes) and during (blue boxes) this study. By completing these tasks, it was hoped that the results could then be used to assess and answer the research questions outlined here.

4 Materials and methods

This section describes the steps taken in the research from sample preparation and processing to the actual analysis of the obtained data. For clarity, there are two main subsections; the first one concentrates on the sample treatments, their preparation and the mass spectrometry processing carried out with the samples and the second subsection discusses the various statistical analysis tools used in further processing of the obtained data.

4.1 Study material

As discussed previously, HCE and IOBA-NHC were used in this study to examine the cytotoxic effects of BAC. One reason for collecting this type of data was to see if producing reasonable data from cell lines was in fact possible as studies exactly like this have not been performed previously. Other alternatives to this type of data collection exist, e.g. tear samples from patients, but the interest here is cell lines in particular.

Sample preparation consisted of exposing the HCE and IOBA-NHC cells to either preservative-free prostaglandin tafluprost (Talfotan® 15 µg/ml, Santen), preserved latanoprost (Xalatan® 50 µg/ml, Pfizer) or preservative BAC for 24 hours. Xalatan contains high concentrations of BAC (0.02%) unlike Talfotan, which is a preservative free medication but both a contain prostaglandin as the effective agent. In addition, samples with no additional treatment were included as controls. Hence, a total of four different sample groups, with varying number of biological replicates for HCE and IOBA-NHC cell lines, were processed and analysed for both cell lines.

The water-soluble tetrazolium salt (WST-1) assay (Roche) was used to evaluate the cytotoxicity of treatments and it is based on functions of mitochondrial dehydrogenase enzymes as an indication of cellular growth and viability; hence, it enables the approximation of cytotoxicity via loss of cells. Figure 4.1.1 indicates the results of the WST-1 cytotoxicity test after treating the cells with different glaucoma drugs. When the dilution rate is increased from 1:300 towards 1:20, the cell survival for BAC- and latanoprost-treated cells drops dramatically to only few percent. However, tafluprost-treated cells have a very good survival rate throughout, on average over 90% in all dilutions for both cell lines. For the purposes of this study, a dilution of 1:300, with 24 hour exposure, was chosen for

A IOBA-NHC 20 000c/w exposure 24 h -FBS WST-1

B HCE 20 000c/w exposure 24 h -FBS WST-1
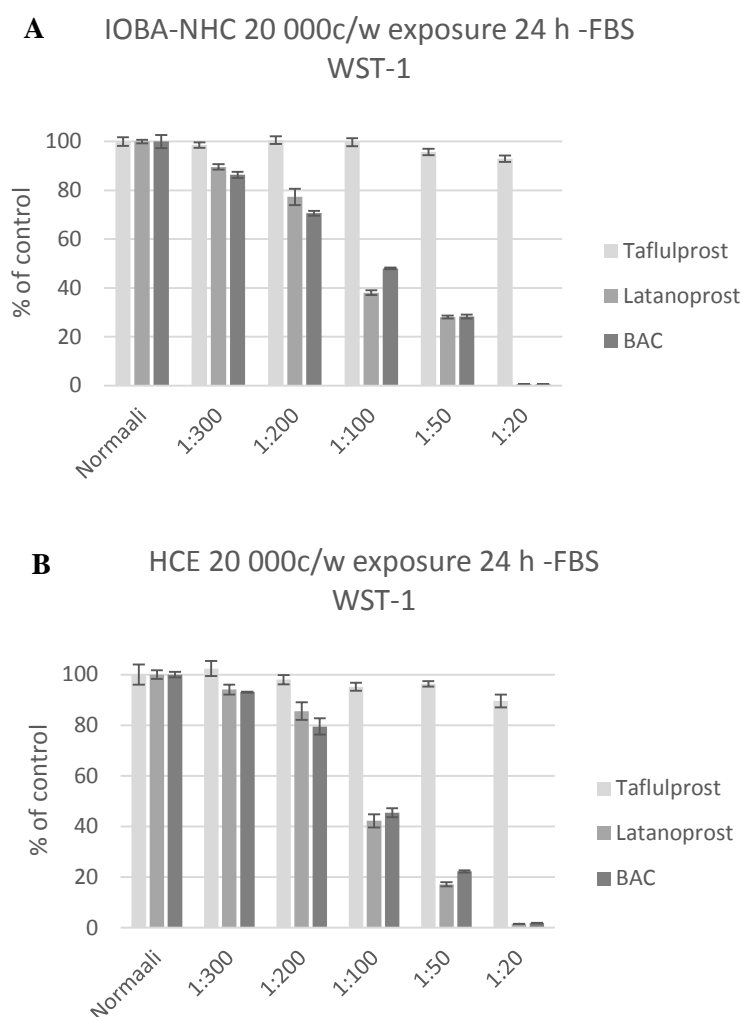
Figure 4.1.1 The cytotoxicity of different treatments was evaluated with water-soluble tetrazolium salt (WST-1) assay (Roche) in order to identify an ideal exposure concentration to be used further on in the study. The results obtained with untreated control cells were set as 100% and the BAC concentration in BAC-treated sample is kept equal to the corresponding exposure of BAC in preserved latanoprost. As the figures indicate, the cell survival is over 80% for all treatment groups in both A) IOBA-NHC and B) HCE cell lines when the dilution is 1:300 for tafluprost and latanoprost.

latanoprost and tafluprost and an equivalent dilution of 0.000067% was used for BAC. This way it can be assumed that whilst the cell survival is still relatively high, the effects of BAC could be observed in the proteomic profiles of these samples. Hence, all following processing methods and analysis relate to these parameters.

For each cell line, three biological replicates were then produced of treated and untreated cells and their proteomic profiles were analysed with NanoLC-TOF-MS using SWATH^TM. The structure for each individual cell line can be seen in Table 4.1.2. One of the sample groups for the HCE cell line was damaged during sample processing and could not be included in the analysis and therefore HCE data only has two biological replicates for each treatment. Otherwise the structure of the data remains the same as shown in the table below. The SWATH library for >2700 proteins was created from the samples and 2299 and 1920 proteins using ProteinPilot and PeakView and Marker Viewer were used to match and relatively quantify the results respectively for IOBA-NHC and HCE cells. False discovery rate (FDR) of 1% was applied to the analysis.

Table 4.1.2 Two technical replicates were produced of each biological replicate and the data structure of IOBA-NHC cell line is shown here as an example. HCE cell line data structure is highly similar though the number of biological replicates for each treatment is only two.

Preservative	Preservative						Preservative-free					
Treatment	BAC			Latanoprost			Tafluprost			Control		
Biological replicate	1	2	3	1	2	3	1	2	3	1	2	3
Technical replicates	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2

One additional important aspect of the data, which should be noted here is that in a sense the data is organized in an ordinal manner. More specifically, the data for each cell line is structured so that the proteins in the top rows of the matrix, have a better quantification quality when compared against the proteins in the bottom end, which could be considered less reliable in their abundance. This is important to take into account in later parts of the analysis and especially when interpreting the differential expression results.

4.2 Statistical methods

Once the processed and quantified protein abundance data were obtained, the next step is to apply statistical methods to it in order to identify interesting proteins or any possible patterns between the treatments. In the following subsections, the statistical methods applied to IOBA-NHC and HCE data are described. Where applicable, some further descriptions about both of the datasets are included but majority of the actual analysis and results are explained and illustrated then in the following sections (see sections 5 and 6).

4.2.1 Data processing – from raw data to quantified protein measures

Altogether 2299 and 1920 proteins, all with a unique International protein index (IPI) and associated, not necessarily unique, gene symbol accession, were relatively quantified for the IOBA-NHC and HCE cell line samples respectively. The relative quantification step used a SWATH library for >2700 proteins as a reference and this library was created from the samples.

4.2.2 Reduction of bias and checking for data quality

First, $\log_2$-transformation was applied to the data and as the number of quantified proteins was relatively high for proteomic data, and some variability between biological replicates and their medians were observed for both cell lines, a central tendency normalization was applied to the data after $\log_2$-transformation. Other normalization methods were also tested, e.g. loess normalization, which also accounts for the aforementioned trend bias, and quantile normalization. However, these alternative methods did not result in any considerable improvements in the data.

Quality of the data, in the case of technical replicates in particular, was in addition checked prior and after the transformation and normalization steps. This was performed by producing MA-plots, correlations coefficients for technical replicates (Spearman's rank correlation coefficient) and visualizations applying hierarchical clustering.

4.2.3 Differential expression analysis

Once it had been established that the data was adequately preprocessed, the next step involved differential expression analysis. More specifically it was checked, if any statistically significant differences will arise between the preservative-free and preservative-treated samples. As mentioned, the differential expression analysis, similar to all of the earlier steps, was done separately for the IOBA-NHC and HCE cell line data.

The presence of technical replicates in the data means that the chosen statistical model should be able to account for the multiple "layers" in the data. Note, that not only should the non-independence of the technical replicates be accounted for, but also the nested structure of the preservative-free (tafluprost and control) and preservative-treated (latanoprost and BAC) samples needs to be included in the model.

By including both latanoprost- and BAC-treated samples under the preservative-treated factor, the full effects of BAC-treatment, and in particular the potential proteins affecting the development of BAC-induced adverse effects in topical treatments could be better discovered. Same applies to the preservative-free group, which naturally includes both control and tafluprost-treated samples. The model was implemented using *lme4* package in R and *lmer* function in it in particular. The resulting model, in terms of R code, is shown below:

$$protein\ abundance \sim preservative + (1|treatment:sample) + (1|sample:run) \quad (3)$$

In the above function, the protein abundance is a continuous variable whilst the independent variables are all nominal. Here, the last two terms starting with "(1|" show the random effects and the ":" symbols indicate, that the factors are nested within each other. For example, as is known, there are two runs (technical replicates), which relate to the same sample (biological replicate). This model is not stating the additional i.i.d. residual error but it should be considered to be part of the model. In addition, below are the statistical hypothesis relating to this statistical model:

H_0: There is no difference between the protein abundance levels of the preservative-free and preservative-treated samples.

H_1: There is a difference between the protein abundance levels of the preservative-free and preservative-treated samples.

It should be noted that the *lmer* function does not supply p-values for the coefficients and hence these were separately produced using varying methods, which were compiled in a text by Mirman (2014). As Mirman explains, the most conservative of the methods listed uses Kenward-Roger approximation to obtain approximate degrees of freedom and the t-distribution to get p-values. The results based on the less-conservative methods could have produced p-values with much higher initial FDR and hence Kenward-Roger approximation was implemented and more specifically the p-value derivation is implemented in the *pbkrtest* package in R.

4.2.4 Establishing thresholds – fold change and p-value adjustments

Benjamini-Hochberg, which allows to control the FDR, was implemented here to obtain the corrected p-values, i.e. q-values. In order to establish, which proteins have an interesting difference in abundance between preservative-free and preservative-treated samples, the groups of interesting proteins for IOBA-NHC and HCE cell line results were evaluated based on both their statistical significance (q-values) and biological effect size ($\log_2$ fold changes in this case).

As the data in this study is now in $\log_2$ scale, the threshold for proteins of biological effect needs to be transformed to $\log_2$ scale as well, hence for example $\log_2(2)$ for 2-fold changes. Both of the thresholds are applied to the results at the same time, which should ensure that the chosen proteins of interest have both statistical and biological significance. Because of this, and due to the initially conservative p-values, the traditional thresholds were slightly relaxed as described in the results.

4.2.5 Network reconstruction

Next some networks were produced by reconstructing them using reverse engineering. More specifically, the algorithm used to achieve this is called ARACNE, Algorithm for the Reconstruction of Accurate Cellular Networks, which is available to use in R via several packages. In this study, *minet* package was used to pre-process the data and run the ARACNE algorithm (Meyer et al., 2008). Cytoscape (version 3.2.0) was then used to visualize the resulting potential connections (Shannon et al., 2003).

4.2.6 Enrichment analysis

In this study, several tools were tested for enrichments and these tests implemented both KEGG and GO terms. The tools used include freely available online tools The Database for Annotation, Visualization and Integrated Discovery, DAVID, (Huang et al., 2009 (1); Huang et al., 2009 (2)), Gene Ontology enRIchment anaLysis and visuaLizAtion tool, GOrilla, (Eden et al., 2009; Eden et al., 2007) and GSEA software provided by the Broad Institute (Subramanian et al., 2005; Mootha et al., 2003).

5 Results

This section describes in further detail the results together with the analyses performed with the IOBA-NHC and HCE cell lines. In the following subsections, the results from both cell line analyses are shown, often side-by-side though some parts are kept separate, and the cell line specific results are then later on discussed in relation to each other together in more detail. In addition, any differences between the cell lines are examined.

5.1 Descriptive statistics

First aspect of interest was the number of proteins quantified from the samples; 2299 and 1920 proteins were relatively quantified for IOBA-NHC and HCE cells respectively. Venn diagram in Figure 5.1.1 shows how many of these quantified proteins were found in both cell lines. It should be noted, that there are several proteins which are very similar in their function and annotation and only due to small variation they are not considered to be the same protein in the two cell lines and hence, with more relaxed matching, the figure in the overlapping section could be expected to be much higher.
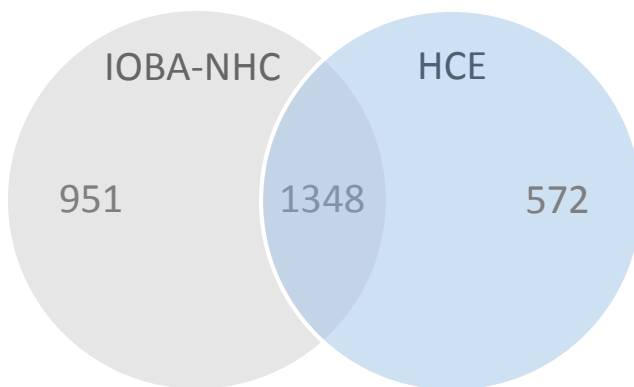
Figure 5.1.1 Venn diagram of the numbers of quantified proteins between the two cell lines, i.e. IOBA-NHC (grey) and HCE (blue).

As discussed earlier, the data, after $\log_2$ transformation, should be further normalized if seen necessary, and since there were some differences between the technical replicate distributions and more specifically their medians and variance (see Appendix A), this was deemed as an appropriate next step. The normalization was performed using the central tendency normalization by subtracting the sample-wise median from each sample-specific observation.
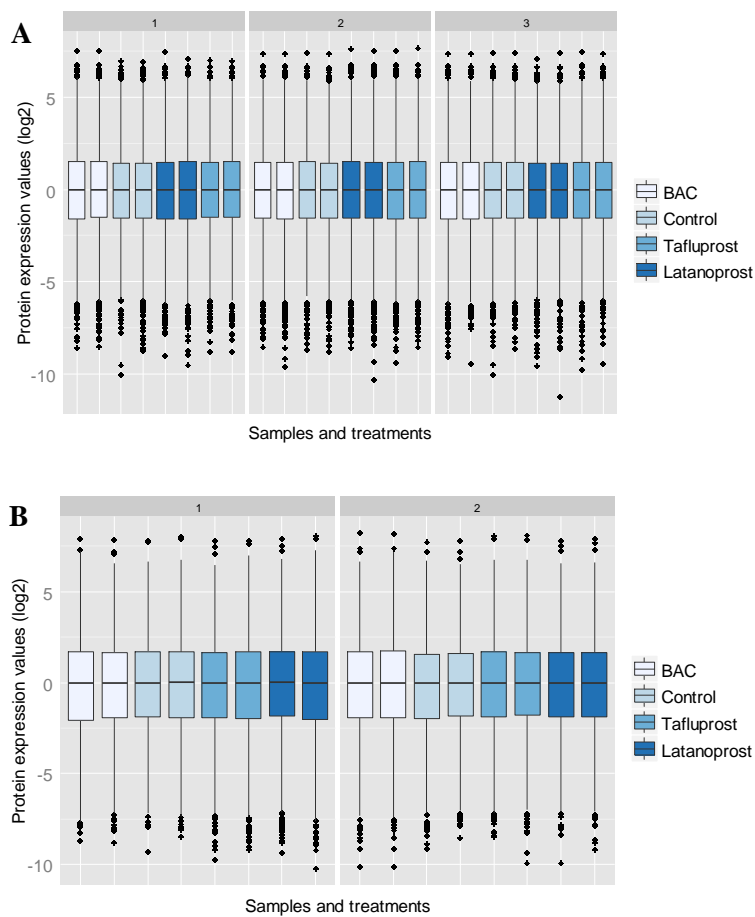
Figure 5.1.2 Boxplots of protein expression levels post normalization for (A) IOBA-NHC cell line and (B) HCE cell line. Normalization was performed on log$_2$-transformed data using central tendency normalization.

In Figure 5.1.2, the protein expression level distributions for individual technical replicates can be seen post normalization. This normalization appears to have succeeded well for both cell lines as it can now be assumed that the protein expression levels in all biological samples are roughly similar to each other based on medians and variance (see Appendix A for comparison). As can further be seen from Figure 5.1.2, there are three biological replicates present in the IOBA-NHC cell line (A) and two biological replicates for the HCE cell line (B) and each biological replicate has two technical replicates.

Figure 5.1.3 shows the discretized protein abundance levels in heat maps for both cell lines separately. Discretization, i.e. binning, into 10 categories was applied to the data due to the large homogeneity of the datasets and the few higher outliers, which otherwise would have prevented efficient visualization of differences between the samples. The dendrogram in the top shows the hierarchical clustering of the samples and this is similarly done to proteins in the rows, though the names of the proteins are omitted due to their large number. The clustering here was performed using hierarchical clustering with complete linkage method and Euclidean distance as the similarity measure.

As Figure 5.1.3 also suggests, the technical replicates are not highly dissimilar as is evident from their clustering close together for both cell lines. Further on, the control and tafluprost-treated samples are more often clustered together similar to samples treated with BAC and latanoprost. This is particularly true for the IOBA-NHC samples (A), and the smaller number of biological replicates in HCE data (B) could be causing the clustering becoming more ambiguous. Yet, some clearer differences between preservative-free and preservative-treated samples appear to be evident for the HCE cell line. The

clustering is not perfect for either of the cell lines and this would not be expected either as all identified proteins are included in this plot.
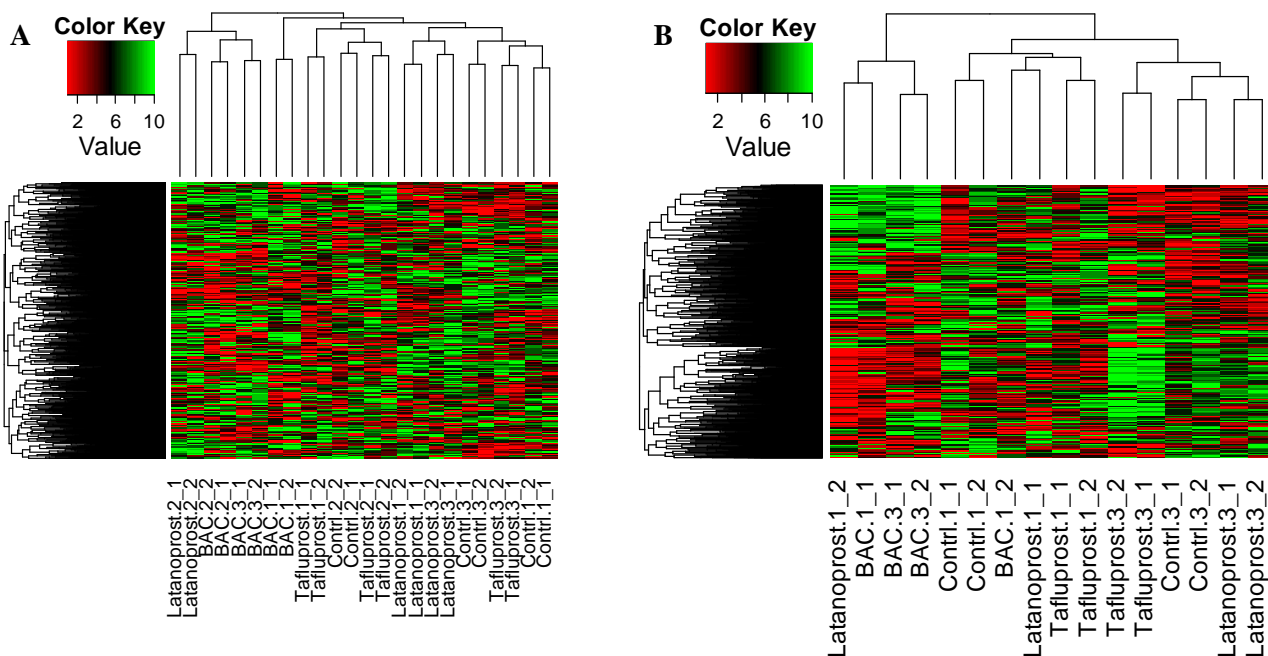
Figure 5.1.3 Heat maps of the discretized expression levels between all samples for (A) IOBA-NHC and (B) HCE cell lines. Hierarchical clustering using Euclidean distance was applied in the clustering and generation of dendrograms. The number following the treatment name tells the sample number and the final figure tells the technical replicate number. As described by the colour keys, green signifies highly expressed proteins and red indicates that the expression level is relatively low in comparison to other proteins on average. BAC- and latanoprost-treated samples cluster frequently close to each other in both cell lines similar to control and tafluprost-treated samples.

The quality of technical replicates was further evaluated by calculating the Spearman's correlation coefficient and the connected p-value for all technical replicates. For all technical replicates, the correlation coefficient was well above 0.9 and in fact the minimum coefficient was 0.9489 and largest p-value obtained was 2.2e-16, which is still considerably small. In addition, the MA-plots (Appendix B) did not show any worryingly large trends between the technical replicates. Hence, the technical replicates appear to be considerably similar and no exclusions are necessary. The preliminary descriptive visualizations and statistics described in this subsection do suggest that the data are of sufficient quality.

As mentioned, other normalization methods, e.g. quantile normalization, were tested but these options were discarded as they did not appear to improve the quality of data. In fact, it appears that these "stricter" methods cause a larger loss information which is necessarily no longer due to bias. This type of approaches can be more relevant with data containing more quantified proteins or for example

mRNA. Indeed quantile normalization is often used to eliminate bias from mRNA or similar data which can include hundreds of thousands of different mRNAs.

5.2 Differential expression

This subsection will have the IOBA-NHC and HCE cell lines separated, which will enable more detailed analysis of the results for both cell lines and hopefully to avoid confusion between the results and their corresponding interpretations.

5.2.1 Differential expression analysis for IOBA-NHC

The differential expression analysis was performed using mixed-effects model as described previously and the multiple correction issue was accounted for by adjusting the p-values based on Benjamini-Hochberg correction, which results in q-values. Figure 5.2.1.1 displays the results from the differential expression analysis. The chosen thresholds for this cell line, once taking into account the conservative nature of the original p-values and the fact that two different thresholds are applied to the results, were chosen to be >1.25 and <0.8 for the fold change and <0.25 for the q-values. Effectively, values satisfying the fold change and the q-value thresholds are considered interesting and potentially significant both statistically and biologically. In Figure 5.2.1.1 these proteins are colored red and they are also further named (gene symbol ID) in the volcano plot.

A few proteins pop out from the plot. In general, interesting proteins include the ones, which are mapped very high in relation to the y-axis, i.e. they have a very low q-value, and hence even lower original p-value. In the case of IOBA-NHC, proteins satisfying this criteria include at least hydroxymethylglutaryl-CoA synthase, cytoplasmic (HMGCS1), NADH dehydrogenase [ubiquinone] iron-sulfur protein 3, mitochondrial (NDUFS3) and NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 5 (NDUFA5). These proteins are all under-expressed in the BAC-treated samples.

It should be noted that General transcription factor 3C polypeptide 4 (GTF3C4), is likely to have obtained its significance due to very large variability which becomes evident when plotting the observed values of this protein and it is unlikely, that there would be a truly significant difference between the preservative-free and preservative-treated samples in the case of this protein.

Figure 5.2.1.1 Volcano plot of IOBA-NHC data analysis results. The x-axis represents the biological effect size, fold change (in log$_2$ scale) so that proteins at the far left in this scale can be considered to be under-expressed in preservative-treated samples. The proteins in the far right are then over-expressed in the preservative treated samples. The y-axis displays the statistical significance scale, the least statistically significant proteins being at the bottom close to zero. Proteins, which have a q-value below 0.25 and fold change above 1.25 or below 0.8 are named and labelled red.

As already mentioned previously, the order of the proteins is an aspect which should be kept in mind since the proteins higher in the list can be considered to be more reliable. Table 5.2.1.2 lists the proteins of interest, i.e. proteins which satisfy the chosen thresholds, together with their corresponding log$_2$ fold changes and p-values (both non-adjusted and adjusted). More specifically, as a result the statistical analysis 29 differentially expressed proteins were identified for IOBA-NHC cells (fold change>1.25 or <0.8, q-value<0.25). This table shows the proteins in the order they were discovered in the loop, hence the proteins, their expression levels and hence, their results in the top rows can be considered to be of higher quality than the ones in the bottom rows.

Table 5.2.1.2 Values for interesting proteins, i.e. proteins which have a fold change > 1.25 or <0.8 and q-value < 0.25, in IOBA-NHC cell line in the order of quality of quantification.

Gene symbol	IPI	Log_2 fold change	P-value	Adj. p-value
PYGL	IPI00943894.1	0.434	0.005	0.190
HADHA	IPI00031522.2	-0.518	0.004	0.175
HNRNPA1	IPI01022801.1	-0.349	0.008	0.221
HMGCS1	IPI00963899.3	-0.580	0.000	0.004
PYGB	IPI00004358.4	0.324	0.002	0.112
DDX21	IPI00015953.3	-0.562	0.001	0.091
HADHB	IPI01018954.1	-0.362	0.003	0.142
HMGB2	IPI00219097.4	0.416	0.009	0.222
AIP	IPI00953925.1	0.330	0.009	0.222
AP3D1	IPI00411453.3	0.460	0.009	0.222
RTCD1	IPI00011726.1	-0.392	0.002	0.112
GTF3C4	IPI00016725.2	3.015	0.006	0.191
HK2	IPI00917193.1	-0.759	0.005	0.190
MCTS1	IPI00179026.2	-0.498	0.008	0.221
GALK1	IPI00940264.1	0.363	0.001	0.085
KIAA0101	IPI00014147.3	-0.651	0.001	0.109
CYP51A1	IPI01013163.2	-0.979	0.002	0.112
MRTO4	IPI00106491.3	0.553	0.004	0.164
PSMB8	IPI01017933.1	1.333	0.008	0.220
NDUFA5	IPI00412545.4	-0.410	0.000	0.023
MACROD1	IPI00155601.1	-0.684	0.007	0.203
UBE2Z	IPI00829749.2	0.763	0.009	0.222
NDUFS3	IPI00025796.3	-0.395	0.000	0.004
FMR1	IPI00872761.2	1.260	0.010	0.227
NUP54	IPI00386702.1	0.380	0.012	0.237
INHA	IPI00007080.1	0.534	0.000	0.058
TPP1	IPI00554617.2	-0.387	0.010	0.225
B4GALT5	IPI00011656.1	-0.418	0.001	0.091
MRPS35	IPI00073779.1	-0.410	0.010	0.227
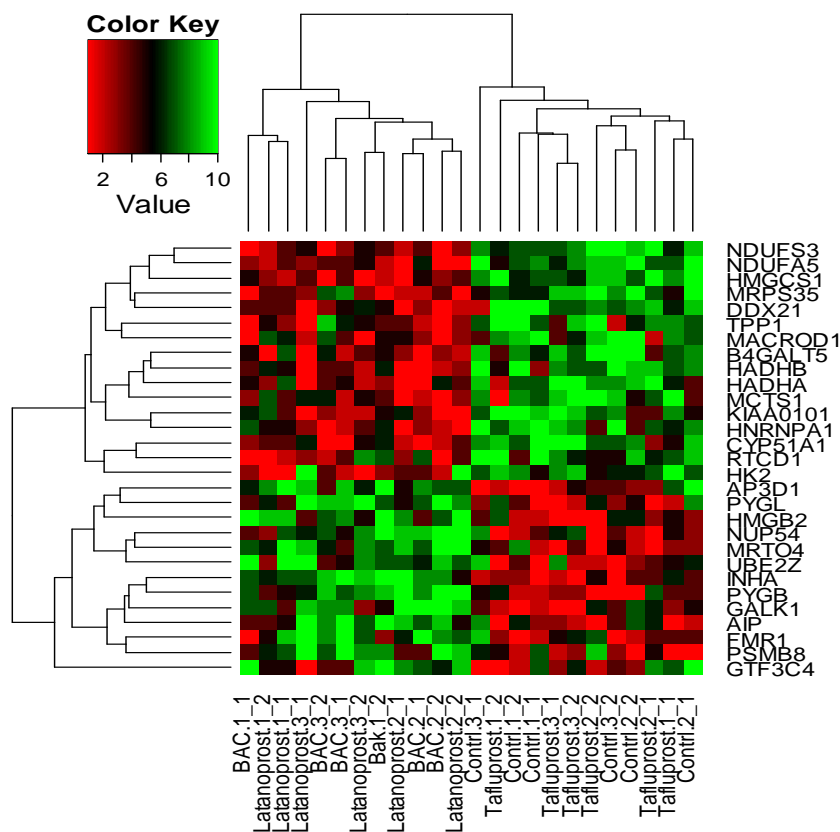
Figure 5.2.1.3 Heat map of proteins, which displayed statistically significant (fold change > 1.25 or < 0.8 and q-value < 0.25) expression levels in IOBA-NHC data. The red colour indicates lower expression and green alternatively tells of higher expression levels.

Figure 5.2.1.3 visualizes the protein expression levels for the statistically significant proteins which were listed above. This time, the expression levels are again discretized so that large values do not overshadow the smaller ones. As it can be seen from the figure, there are approximately as many over- and under-expressed proteins for this cell line when comparing the preservative-free and preservative-treated samples. The next question is, how these proteins are connected and how they can explain the results we see through observing the cells.

5.2.2 Differential expression analysis for HCE

Similar analysis, as was done for the IOBA-NHC cell line data in the previous subsection, was carried out for the HCE cell line as well. Hence, again this part of the analysis was performed using mixed-effects model and the multiple correction issue is accounted for by adjusting the p-values based on Benjamini-Hochberg correction, which results in q-values.

Figure 5.2.2.1 shows the relative biological and statistical significance of the proteins based on the differential expression analysis results. Some of the interesting proteins based on the figure above include myosin light polypeptide 6 (MYL6), myosin-9 (MYH9) and myosin regulatory light chain 12A (MYL12A), which are all highly statistically significant and the biological significance is also consistently high for all of these proteins. In addition, HMGCS1 is again under-expressed in the BAC-treated samples almost exactly in the same biological and statistical significance level as in IOBA-NHC cell line. It would appear that this protein is playing a role in both cell lines possibly in similar

40

aspects. However, as the biological effect threshold is higher for HCE cell line since it has in relation much higher number of statistically significant proteins, the values corresponding to HMGCS1 are not shown in the following table.
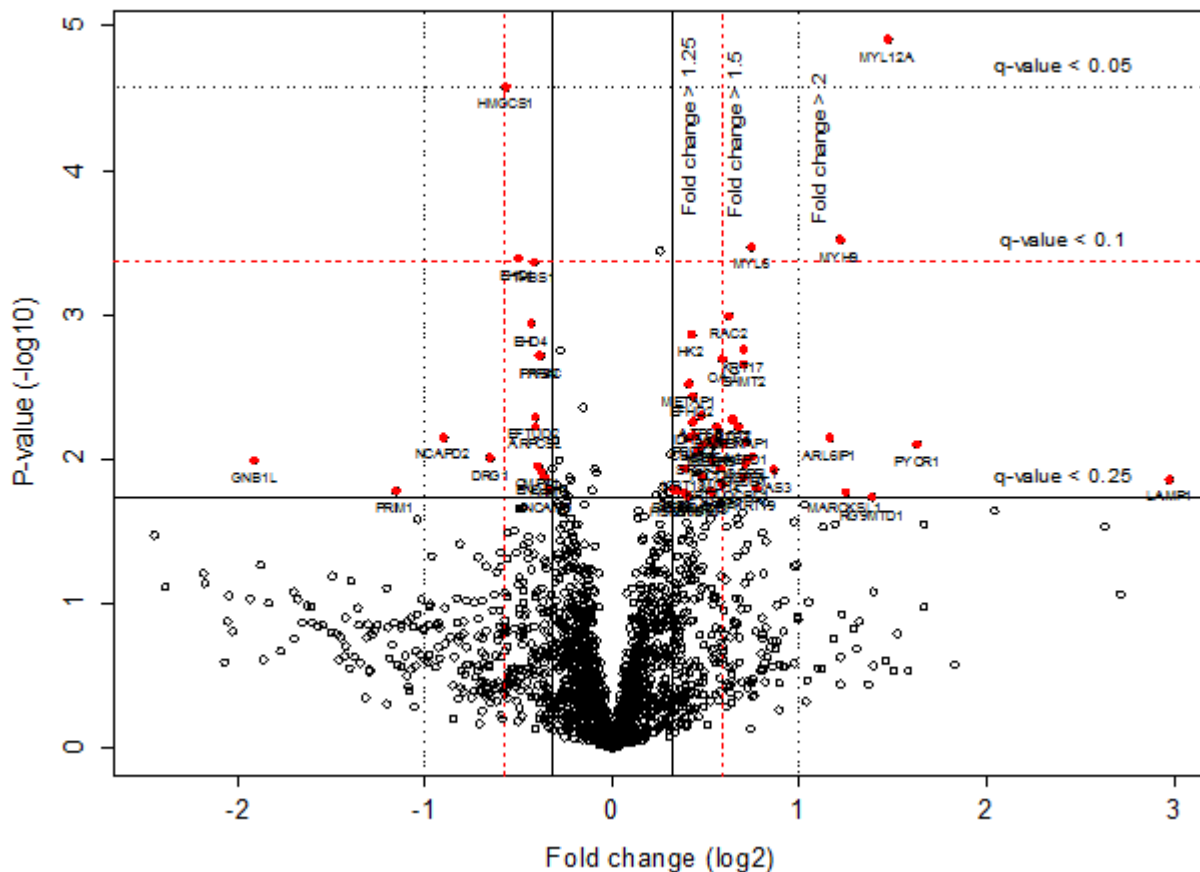
Figure 5.2.2.1 Volcano plot of HCE data analysis results. The x-axis represents the biological effect size, fold change (in log₂ scale) so that proteins at the far left in this scale can be considered to be under-expressed in preservative-treated samples. The proteins in the far right are then over-expressed in the preservative treated samples. The y-axis displays the statistical significance scale, the least statistically significant proteins being at the bottom close to zero. Proteins, which have a q-value below 0.25 and fold change above 1.25 or below 0.8 are named and labelled red.

In the volcano plot, the thresholds used were the same as with the IOBA-NHC but in Table 5.2.2.2 the fold change threshold has been increased to 1.5 as the number of interesting proteins here was already relatively high as already mentioned previously. By increasing the threshold for the biological significance, the number of proteins of interest was now roughly similar for both cell lines. Hence, as a result 28 proteins were identified as statistically and biologically significant for HCE cells (fold change>1.5 or <0.67, q-value<0.25). In addition, it should be noted that a fold change threshold of

1.5 is more traditional and hence, it can be assumed that these results were more reliable than some of the ones obtained with the IOBA-NHC cell line data.

Table 5.2.2.2 Values for interesting proteins, i.e. proteins which have a fold change > 1.5 or <0.67 and q-value < 0.25, in HCE cell line in the order of quality of quantification.

| Gene symbol	IPI	Log$_2$ fold change	P-value	Adj. p-value
MYH9	IPI00019502.3	1.219	0.000	0.069
HSPD1	IPI00784154.1	0.714	0.008	0.239
KRT19	IPI00479145.3	0.772	0.016	0.244
SHMT2	IPI00789370.3	0.698	0.002	0.157
NCAPD2	IPI00299524.2	-0.900	0.007	0.239
GOT2	IPI00018206.4	0.641	0.005	0.239
OAT	IPI00022334.1	0.586	0.002	0.155
HSPE1	IPI00220362.5	0.714	0.010	0.244
GLUD1	IPI01014382.2	0.704	0.011	0.244
MYL12A	IPI00220573.4	1.475	0.000	0.014
MYL6	IPI00796366.2	0.741	0.000	0.069
DRG1	IPI00031836.3	-0.651	0.010	0.244
UQCRC1	IPI00013847.4	0.697	0.013	0.244
NIPSNAP1	IPI00894205.2	0.672	0.006	0.239
OAS3	IPI01015393.1	0.862	0.012	0.244
SSBP1	IPI00029744.1	0.589	0.015	0.244
RAC2	IPI00010270.1	0.622	0.001	0.145
PYCR1	IPI00941557.1	1.631	0.008	0.239
GRPEL1	IPI00909181.1	0.753	0.010	0.244
ARL6IP1	IPI01010487.1	1.163	0.007	0.239
PRIM1	IPI00027704.5	-1.159	0.017	0.244
CTSZ	IPI00002745.1	0.646	0.005	0.239
ERH	IPI00029631.1	0.597	0.008	0.239
LAMP1	IPI00884105.2	2.972	0.014	0.244
KRT17	IPI00450768.7	0.704	0.002	0.155
MARCKSL1	IPI00641181.5	1.247	0.017	0.244
RG9MTD1	IPI00099996.2	1.387	0.018	0.246
GNB1L	IPI00107339.4	-1.916	0.010	0.244
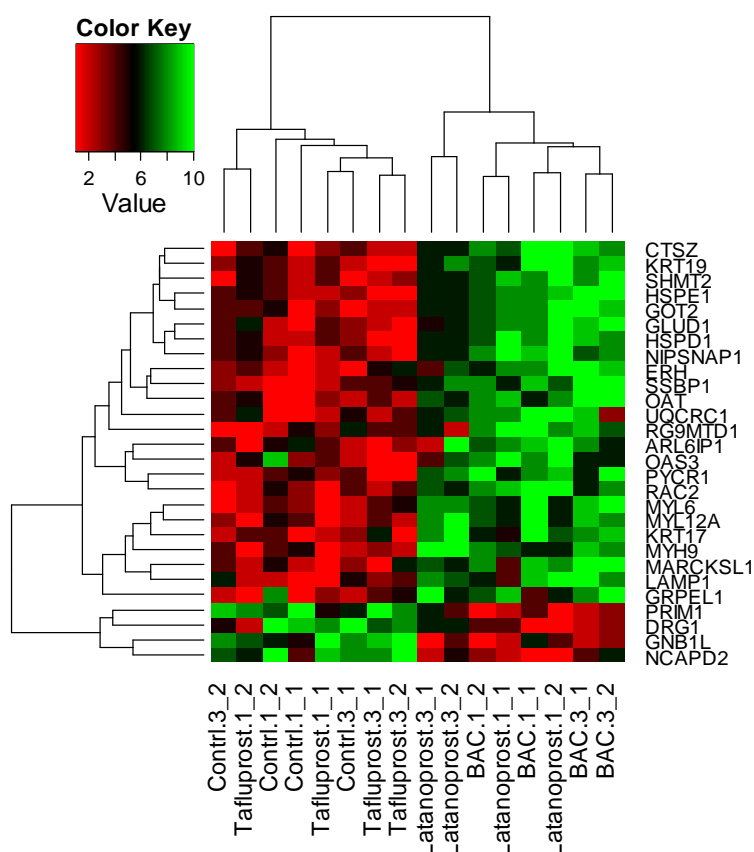
Figure 5.2.2.3 Heat map of proteins, which displayed statistically significant (fold change > 1.5 or < 0.67 and q-value < 0.25) expression levels in HCE data. The red colour indicates lower protein abundance level and green alternatively tells of higher abundance.

The interesting proteins identified in the differential analysis shown in the previous table, are again clustered and visualized in a heat map in Figure 5.2.2.3. The protein abundance is again discretized for better visualization. There is a considerably higher number of significant proteins, which are over-expressed in BAC-treated samples in comparison to the preservative-free samples which was also evident from the volcano plot seen previously. Recall, that with IOBA-NHC this was approximately evenly halved for the "top" 29 proteins.

5.2.3 Visualization of individual proteins of interest

When keeping the thresholds same as previously described, i.e. 0.25 q-value and 1.25/0.8 or 1.5/0.67 fold change for IOBA-NHC and HCE cell lines respectively, there are no statistically and biologically significant proteins, which would be found from both cell lines. This is unfortunate, as there was notably a large number of proteins, which were quantified for both cell lines. However, it would seem more reasonable to do this types of comparisons with exactly the same thresholds and hence, whilst the q-value threshold was kept the same (0.25), the fold change threshold was set to look at proteins with fold change above 1.25 or below 0.8 for both cell lines. This way two proteins were identified to be present in both cell lines: HMGCS1 and HK2. The results for these proteins are shown next together with other interesting proteins, which were deemed statistically and biologically significant in only one cell line but are nevertheless in some way interesting and worthy of further investigation.

HMGCS1 showed signs of under-expression with BAC treatment for both cell lines similarly (IOBA-NHC: -0.58±0.07, p-value < 0.001, HCE: -0.57±0.05, p-value < 0.001). Its ranking was 239 and 211

in IOBA-NHC and HCE data respectively, which is relatively high in both cases and this suggests good accuracy of results. Figure 5.2.3.1 shows the expression levels for different treatments and this further clearly shows that the expression is much lower for cells treated with BAC, in both cell lines.
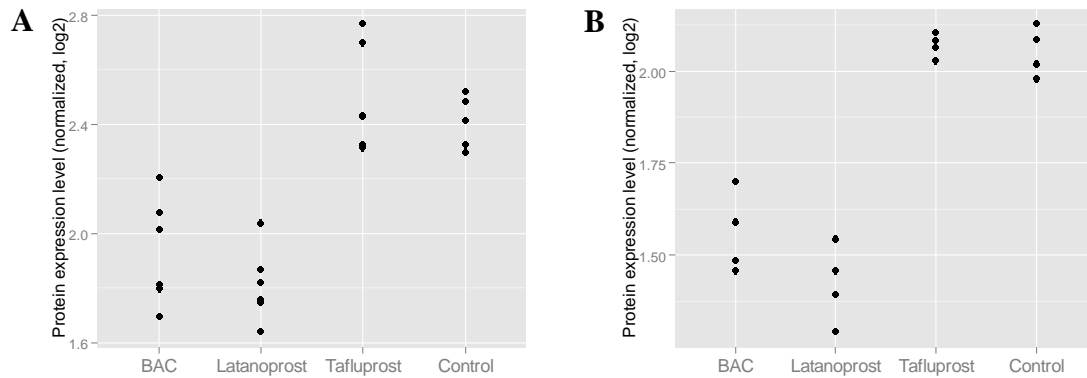
Figure 5.2.3.1 The expression levels of HMGCS1 between different treatments for IOBA-NHC (A) and HCE (B) cell lines. It can be noted, that the relative expression levels are consistently lower for preservative-treated samples for both cell lines.

HK2 expression levels in BAC-treated cells were under-expressed for IOBA-NHC (-0.76 ± 0.21, p-value 0.005) and over-expressed HCE (0.42 ± 0.07, p-value = 0.001). The true significance of this protein is in this point difficult to determine as the ranking of this particular protein was 1176 for IOBA-NHC and 1169 for HCE and these larger values suggest, that the quality of measurements for this particular protein are not high. Hence, there should be more caution with this protein. Figure 5.2.3.2 shows the expression levels for the cell lines and it demonstrates how the expression of control and Tafluprost samples is "flipped" between the cell lines.
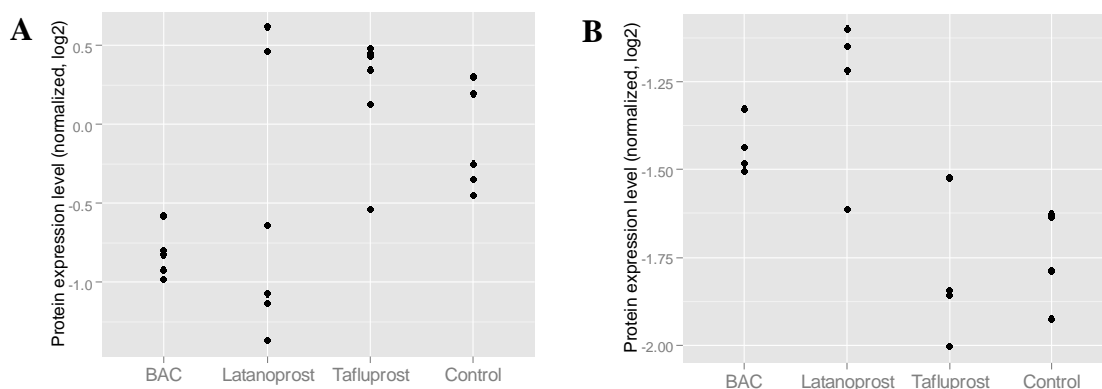
Figure 5.2.3.2 The expression levels for HK2 between different treatments for IOBA-NHC (A) and HCE (B) cell lines. The differences between the preservative-free and preservative-treated samples are not as clear as hoped. It appears that there is a large amount variability between the measurements.

Myosin-related proteins were over-expressed in the BAC-treated samples, but only in the HCE cell line. These clearly stood out from the rest of the proteins as highly statistically significant. Figure 5.2.3.3 shows the expression levels for these proteins in HCE samples and it can be clearly seen that the three proteins, more specifically MYH9 (1.22±0.16, p-value < 0.001), MYL12A (1.48±0.11, p-value < 0.001) and MYL6 (0.74±0.1, p-value < 0.001), had much higher expression levels for the BAC- and latanoprost-treated samples in all cases in comparison to the preservative-free treatments.

Figure 5.2.3.4 visualizes differences between the treatments in IOBA-NHC cells. As it can be seen from the figure, both NDUFA5 (-0.41±0.06, p-value < 0.001) and NDUFS3 (-0.39±0.04, p-value < 0.001) had higher expression levels in tafluprost-treated and control samples in comparison to the two other treatments. The differences were not very large, but it could be thought that even small deviations from the norm could potentially cause changes in the cells.
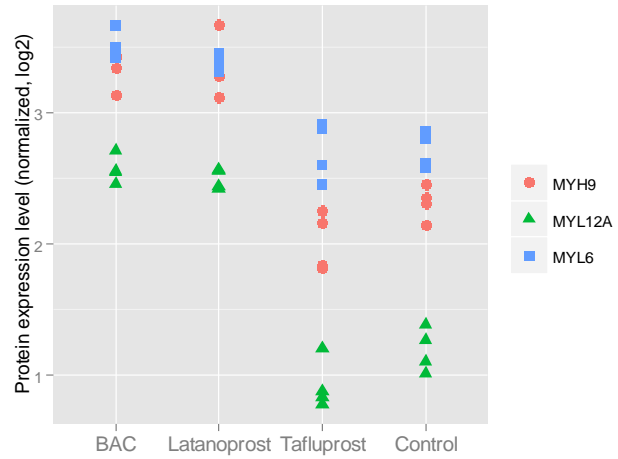
Figure 5.2.3.3 The expression levels between samples in HCE cell line for MYH9, MYL12A and MYL6. All three proteins had much higher expression levels for the BAC- and latanoprost-treated samples in all cases in comparison to the preservative-free treatments.
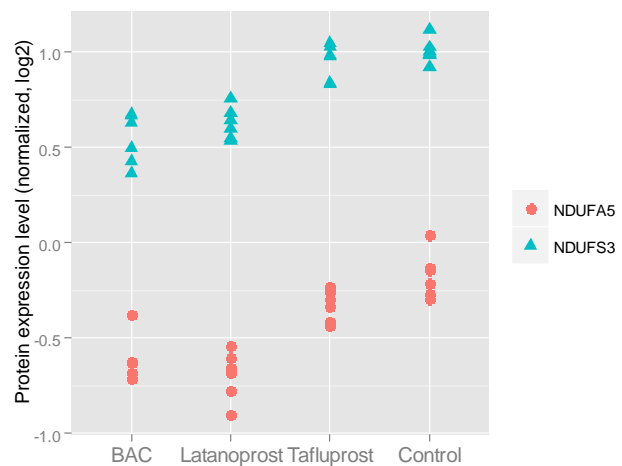
Figure 5.2.3.4 The expression levels between samples in IOBA-NHC cell line for NDUFA5 and NDUFS3. The relative differences between preservative-free and preservative-treated samples were not as large as for the previous proteins but it would still appear that the preservative-free samples have higher relative expression levels.

5.2.4 Network construction

The next step in the analysis process was the network construction, which was performed by applying ARACNE algorithm based on the MI scores available for the protein pairs, as already described earlier. This was performed using the *minet*-package in R. Cytoscape was then used to visualize the resulting potential connections, which are visualized in Figures 5.2.4.1 and 5.2.4.2. Note, that the resulting graphs are mainly just additional visualizations of the results but the potential connections could also further give indication of the similar expression levels between proteins.

Figure 5.2.4.1 Connections between the proteins deemed significant, based on both their biological and statistical effect size, for the IOBA-NHC cell line. The colour of the circle tells whether the protein is under- or over-expressed, green and red respectively, in the BAC-treated samples in comparison to the preservative-free samples. The size further indicates how statistically significant the protein is in terms of the differences. More specifically, the larger the circle, the more statistically significant the protein can be thought to be.

Figure 5.2.4.2 Connections between the proteins deemed significant, based on both their biological and statistical effect size, for the HCE cell line. The size and color of the circle are again descriptive of the statistical significance and expression level (see Figure 5.2.3.1 description).

The visualizations were created with Cytoscape and they display connections deemed significant based on ARACNE algorithm. Any connections seen here could be deemed to be potential indicators, that the connected proteins display similar patterns in their abundance based on the data. From Figure 5.2.4.1 it can be noted that HMGCS1, NDUFS3 and NDUFA5 are all connected based on the mutual information. These three proteins were in addition, the three perhaps most interesting proteins arising from the IOBA-NHC cell line results. Most notably, in Figure 5.2.4.2, the myosins (MYL6, MYL12A and MYH9) are all connected in the graph. Since the connections suggest that the data is "behaving" similarly with these proteins, this further support the findings that they all appear to be affecting the same functions in the cell.

5.3 Enrichment analyses

In the next two subsection, the enrichment analysis results for GO and pathways are displayed. The analyses were carried out for both cell lines and these results are based on the earlier differential expression analysis results.

5.3.1 GO enrichments

The GO enrichments were tested with two different online tools: DAVID and GOrilla. The analyses were conducted differently; the enrichment analysis in DAVID was applied using a protein set, i.e. proteins of interest as specified in previous subsection, and a background of all quantified proteins. The enrichment analysis carried out with GOrilla, uses a list of all of the proteins, which were ordered based on the $\log_2$ fold change. One further difference between these two tools is that DAVID further on clusters the enriched GO terms together unlike GOrilla.

DAVID

DAVID tool clusters the enrichment results into groups that contain similar terms and gives an ES for the overall cluster. In Figure 5.3.1.1, the clusters are visualized for the three GO term groups, i.e. biological process, cellular component and molecular function, as indicated by their titles.

Based on Figure 5.3.1.1 (A) and (B), these are no consistent enrichments between the two cell lines but when moved into the cellular components in Figure 5.3.1.1 (C) and (D), it can be noted that with both cell lines, the interesting proteins are highly concentrated on mitochondrion and other components, which could be closely associated to mitochondrion. In addition, HCE cell line also displays enrichments in myosin complex. Both cell lines also have enrichments in nucleotide binding in molecular functions (Figure 5.3.1.3 (E) and (F)). Note, that only the mitochondrion-related terms are statistically significant and these can be seen in more detail in Appendix C.
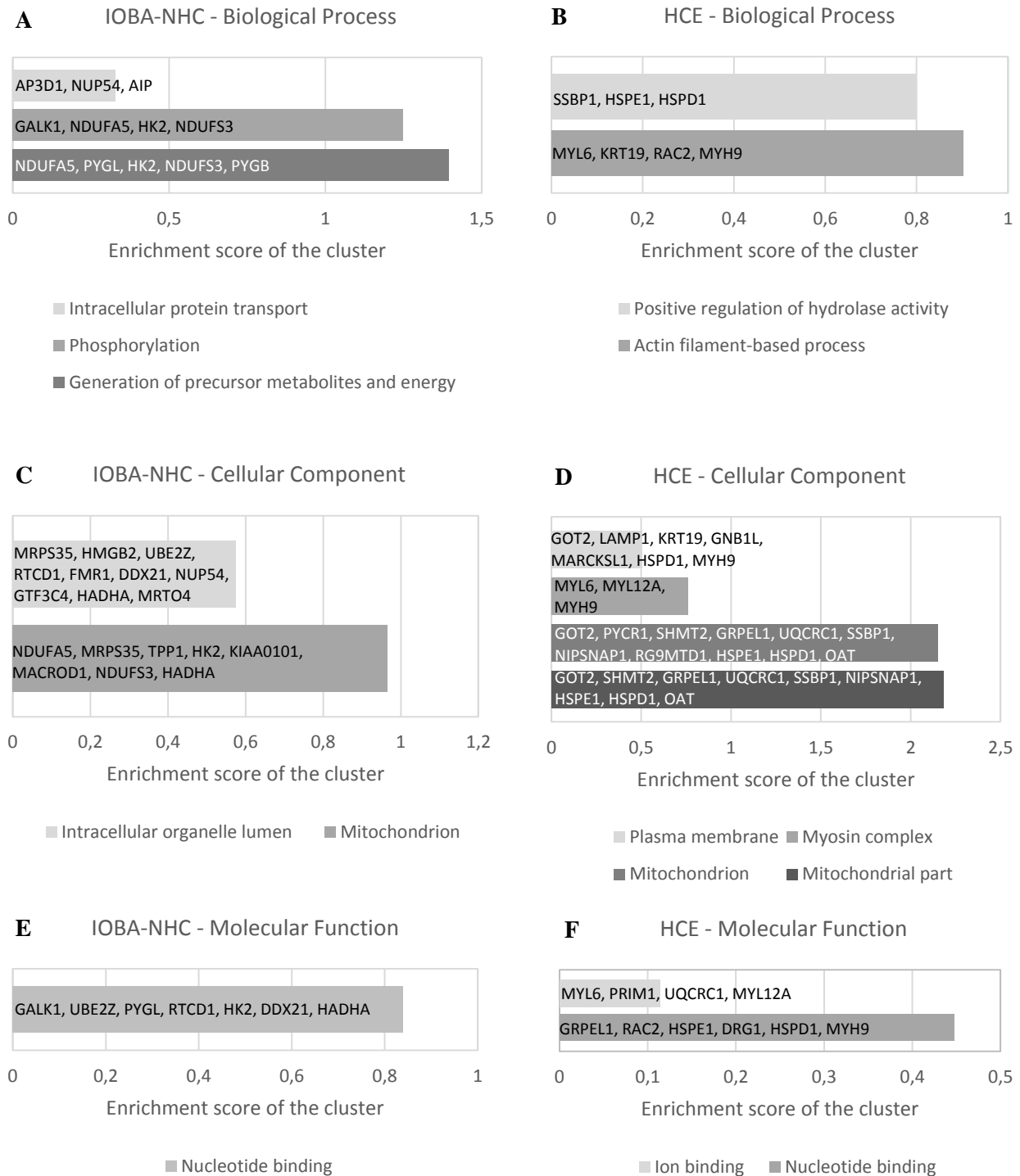
Figure 5.3.1.1 Clusters of enriched GO terms for IOBA-NHC and HCE cell lines. Biological process (A, B), cellular component (C, D) and molecular function (E, F) domains are all shown separately for the IOBA-NHC and HCE cell lines respectively. The horizontal axis shows the ES of the cluster and in general, the higher the ES, the more significant the cluster can be thought to be. All of the proteins associated to a given cluster, are shown inside the bars and the cluster name is given below the plot in the legend. The cluster naming was done so that it best describes the nature of all terms inside the cluster, whilst also taking into account the names which had the most statistically significant GO terms within the given cluster. See Appendix C for further information of the specific terms.

<u>GOrilla</u>

As already discussed, ranking all of the proteins based on some significance measure, e.g. fold change, is another method of performing an enrichment analysis and this was essentially performed with GOrilla. In GOrilla tool, there are no annotation clustering possibilities and hence in the results, only the top 3 enriched GO terms for each cell line and GO term subgroup, when available, are listed. The ranking of enriched terms was done based on their q-values, i.e. p-values which have been corrected in order to account for the multiple testing, which also is present in the enrichment analysis results. No thresholds as such are applied here but the q-values should be kept in mind when making assumptions based on these results.

It can be seen from Table 5.3.1.2 that the proteins, which are over-expressed in the presence of preservatives include many GO terms related to the actin filaments and cell adhesion for IOBA-NHC cell line and mitochondrion and translation for HCE cell line. The mitochondrion terms have a particularly low q-value. Individual proteins are not listed in the following results, as due to the nature of the test, a very large number of proteins is often included in one group.

Table 5.3.1.2 Top 3 enrichment terms, where available, for both cell lines and for all three GO domains. The enrichment analysis is carried out ordering the $\log_2$ fold changes from positive to negative, i.e. the proteins over-expressed for preservative-treated samples are at the top and given most importance. The FDR q-value tells about the statistical significance of a given term.

Cell line	GO domain	GO Term	Description	FDR q-value
IOBA-NHC	BP	GO:0007614	short-term memory	6.86E-01
		GO:1903385	regulation of homophilic cell adhesion	7.17E-01
		GO:0030029	actin filament-based process	7.60E-01
	CC	GO:0042995	cell projection	4.23E-05
		GO:0070062	extracellular exosome	1.48E-04
		GO:1903561	extracellular vesicle	1.85E-04
	MF	GO:0008092	cytoskeletal protein binding	7.75E-03
		GO:0003779	actin binding	9.90E-03
HCE	BP	GO:0043624	cellular protein complex disassembly	2.62E-03
		GO:0009056	catabolic process	2.75E-03
		GO:0006415	translational termination	2.91E-03
	CC	GO:0044429	mitochondrial part	7.97E-22
		GO:0005759	mitochondrial matrix	9.70E-18
		GO:0070013	intracellular organelle lumen	6.27E-12
	MF	GO:0003735	structural constituent of ribosome	1.75E-04
		GO:0003954	NADH dehydrogenase activity	1.36E-01
		GO:0051537	2 iron, 2 sulfur cluster binding	1.54E-01

On the other hand, when the proteins were ranked so that the highest ranking proteins corresponded to proteins, which were very highly expressed in preservative-free samples in relation to preservative-treated samples, the results are different and they can be seen from Table 5.3.1.3. Now the IOBA-NHC cell line displays enrichments in terms related to mitochondrion and translation as well as transcription and HCE cell line has various terms. All terms in HCE cell line have relatively high q-values and this is expected as there were a very small amount of statistically significant proteins, which were over-expressed in the preservative-free samples. Overall, it would appear that the mitochondrion and its related structures are significantly enriched based on both online tools and hence both methods, i.e. protein set and ranked proteins.

Table 5.3.1.3 Top 3 enrichment terms, where available, for both cell lines and for all three GO domains. The enrichment analysis is carried out ordering the $\log_2$ fold changes from negative to positive, i.e. the proteins under-expressed for preservative-treated samples are at the top and given the most importance. The FDR q-value tells about the statistical significance of a given term.

Cell line	GO domain	GO Term	Description	FDR q-value
IOBA-NHC	BP	GO:0000375	RNA splicing, via transesterification reactions	8.24E-05
		GO:0070124	mitochondrial translational initiation	8.72E-05
		GO:0000398	mRNA splicing, via spliceosome	1.23E-04
	CC	GO:0044429	mitochondrial part	7.91E-08
		GO:0019866	organelle inner membrane	7.84E-07
		GO:0005743	mitochondrial inner membrane	2.74E-06
	MF	GO:0001103	RNA polymerase II repressing transcription factor binding	5.75E-02
		GO:0016655	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	1.06E-01
		GO:0016651	oxidoreductase activity, acting on NAD(P)H	1.12E-01
HCE	BP	GO:1903008	organelle disassembly	3.27E-01
		GO:0018212	peptidyl-tyrosine modification	3.69E-01
		GO:0016126	sterol biosynthetic process	4.82E-01
	CC	GO:0005665	DNA-directed RNA polymerase II, core complex	8.08E-01
	MF	GO:0004713	protein tyrosine kinase activity	5.95E-02
		GO:0003899	DNA-directed RNA polymerase activity	2.86E-01
		GO:0031267	small GTPase binding	2.95E-01

5.3.2 Pathway enrichments

The pathway enrichment analysis was performed using a tool provided by Broad Institute. Here the list of proteins was again ranked from the highest to lowest $\log_2$ fold change and the default software

settings were used to run the tests. Further on, all curated annotations were searched through all latest (version 5.0) curated KEGG annotations. Other annotation databases were also tested, but KEGG database appeared to produce more general and applicable information and when multiple databases were combined, the amount of information became overwhelming. KEGG enrichments were also tested using DAVID annotation tool but no further information was obtained this way.

As it can be seen from Tables 5.3.2.1 and 5.3.2.2, the results are not highly significant for many of the pathways after adjusting for the multiple comparison. The two tables display only the KEGG terms, which had an unadjusted p-value below 0.3. Keeping the threshold high enables us to get a wider picture of the possible underlying pathways being affected, though the significance of the enrichment should also be kept in mind. Another aspect is that these enrichment results are merely guiding the discussion and any arising hypotheses should be further tested before any conclusions can be made. This naturally also applies to the other results shown in this subsection.

Table 5.3.2.1 The most statistically significant results obtained from a pathway enrichment analysis for the IOBA-NHC cell line. The order tells which way the $\log_2$ fold changes were ordered, i.e. "neg" refers to enrichments in the proteins, which were under-expressed for the BAC-treated samples and vice versa for "pos". Size shows how many proteins from the list were found to associate to any particular KEGG annotation. ES and NES refer to non-normalized and normalized ESs respectively and the final three columns refer to unadjusted (NOM p-val) and adjusted significance measurements.

Order	KEGG Annotation	Size	ES	NES	NOM p-val	FDR q-val	FWER p-val
neg	Spliceosome	81	-0.431	-1.549	0.008	0.796	0.563
	Butanoate metabolism	16	-0.578	-1.462	0.061	0.748	0.789
	Apoptosis	17	-0.538	-1.367	0.112	0.878	0.936
	Lysine degradation	15	-0.528	-1.288	0.178	0.997	0.986
	Valine, leucine and isoleucine degradation	25	-0.442	-1.253	0.172	0.952	0.993
pos	Pentose phosphate pathway	18	0.514	1.337	0.096	1.000	0.974
	Vascular smooth muscle contraction	18	0.520	1.333	0.137	1.000	0.976
	Lysosome	29	0.399	1.173	0.233	1.000	1.000
	Focal adhesion	39	0.368	1.133	0.261	1.000	1.000

Table 5.3.2.2 The most statistically significant results obtained from a pathway enrichment analysis for the HCE cell line. The same descriptions apply as with Table 5.3.2.1.

Order	KEGG Annotation	Size	ES	NES	NOM p-val	FDR q-val	FWER p-val
neg	Pyrimidine metabolism	31	-0.617	-1.774	0.002	0.073	0.081
	Purine metabolism	39	-0.546	-1.644	0.010	0.155	0.306
	Cell cycle	31	-0.510	-1.449	0.046	0.512	0.829
	Prostate cancer	16	-0.527	-1.307	0.153	0.921	0.981
	Pathways in cancer	47	-0.410	-1.249	0.141	0.989	0.997
	Chronic myelois leukemia	18	-0.488	-1.227	0.192	0.910	0.997
	MAPK signaling pathway	33	-0.424	-1.211	0.186	0.842	0.999
	Pentose phosphate pathway	18	-0.483	-1.196	0.229	0.788	1.000
pos	Arginine and proline metabolism	19	0.707	1.877	0.003	0.027	0.024
	Citrate cycle TCA cycle	19	0.663	1.786	0.000	0.047	0.081
	Alzheimers disease	23	0.590	1.667	0.015	0.089	0.207
	Fatty acid metabolism	15	0.618	1.535	0.033	0.196	0.514
	Ribosome	70	0.412	1.458	0.015	0.259	0.698
	Valine, leucine and isoleucine degradation	19	0.514	1.389	0.098	0.337	0.842
	Lysosome	28	0.411	1.247	0.166	0.590	0.982
	Adherens junction	20	0.440	1.237	0.182	0.543	0.985
	Parkinsons disease	15	0.473	1.218	0.236	0.526	0.987
	Tight junction	27	0.404	1.194	0.203	0.526	0.991

The tables do not present any major results, which would be consistent between the two cell lines. However, there are individual terms relating to cell death in both cell lines when looking at the proteins which are under-expressed in BAC-treated samples, i.e. "Apoptosis" (IOBA-NHC) and "MAPK signalling pathway" (HCE). Furthermore, in the "pos"-designated rows, it can be noted that IOBA-NHC has "Focal adhesion" and HCE has "Adherens junction" and "Tight junction" and these can be thought to refer to BAC's ability to break junctions between cells and this way enhance the entrance of the drug as well. Several terms refer to different metabolic pathways and citric acid cycle, mainly in HCE cell line.

5.4 Summary

The results of this study are shown in the subsections above and the results may include potential biomarkers for the adverse effects of BAC and the proteins identified as important or interesting can then be verified in further studies.

Statistical analysis identified 29 differentially expressed proteins for IOBA-NHC cells (absolute fold change > 1.25 or < 0.8, q-value < 0.25) and 28 for HCE cells (absolute fold change > 1.5 or < 0.67, q-value < 0.25). Many of the proteins identified are connected to mitochondrion as shown by the enrichment analysis results.

Several results observed in this study strongly suggest that changes in mitochondrion functions are affected by exposure to BAC Furthermore, in both cell lines HMGCS1 had a similar protein expression level profile, where this protein's expression levels were more specifically under-expressed in the preservative-treated samples. In addition, in HCE cell line results, MYH9, MYL12A and MYL6 were over-expressed in samples treated with BAC and in IOBA-NHC cell line, NDFUA5 and NDUFS3 were under-expressed for BAC-treated samples. These individual proteins and their functions, together with the other results seen in this section are discussed in further detail in the next section.

6 Discussion

In this section the results from the previous section will be discoursed in particular with respect to the known effects of BAC cytotoxicity and other related biological and proteomic aspects. The two cell lines used in this study and their corresponding results are discussed in some parts separately but mainly in comparison to each other in order to understand the underlying mechanisms more in depth.

6.1 Differential analysis

Once the analysis was performed for all proteins in both cell line data, it was noted that the number of proteins with a relatively high fold change ratio and low adjusted p-value, i.e. q-value, was quite low for both cell lines. However, the relative number of proteins, with high statistical and biological significance, was higher for HCE cell line than it was for the IOBA-NHC cell line. This could be explained by the lower number of independent biological replicates for the HCE data, which could then lead to a higher proportion of false positives. However, the quality on the other hand is much higher for the HCE cell line, at least based on some of the initial visualizations and in fact normalization of this data would not have been necessary (see Appendix A) and it was performed just to be consistent with the other similar analyses. Hence, the reason for a higher number of interesting proteins is yet to be established. This does raise questions about the research settings and whether more samples should have been included in the study.

It could be argued that the thresholds should be kept exactly the same for both cell lines but as there are naturally differences between the cell lines and the aim was to identify the "top" interesting proteins, this type of threshold approach seemed reasonable. This is particularly true, since this way it was possible to achieve groups of interest for both cell lines which had approximately the same number of proteins. With some parts, it could however make more sense to keep the thresholds consistent and one such is described in the next paragraph.

Next subsections will discuss some of the individual proteins of interest, which were looked at in the previous section as well in terms of their results and visualizations. The proteins and their known functional properties in cells are explained and in the light of this information, further hypotheses are made of their possible connections to BAC-induced effects.

6.1.1 HMGCS1

HMGCS1 is a cytosolic, soluble HMG-CoA synthase which, according to National Center for Biotechnology Information (NCBI) database, is connected to lipid metabolic process, response to cholesterol and drug as well as brain development. Perhaps one of the most important roles of this protein lies in the cholesterol biosynthesis, where HMG-CoA synthase plays a part as a catalyst (King, 2015).

Cholesterol most importantly contributes to the physical properties of the plasma membrane of cells. More specifically, it affects the thickness, permeability, fluidity and phase behavior (Crockett, 1998) as well as cell signaling (Sheng et al., 2012). In our results HMGCS1 was under-expressed for BAC-treated samples, which could imply that less cholesterol would become available for the use of the cells. This could then in turn result in less stable plasma membranes in BAC-treated cells. It could be hypothesized that this in its part is affecting the cytotoxicity of the preservative since it is further already known that BAC does disturb the plasma membranes of both prokaryotic and eukaryotic cells.

Note further, that interestingly lanosterol 14-alpha demethylase (CYP51A1), which is a cytochrome P450 enzyme, is also found in the list of interesting proteins in IOBA-NHC cell line (-0.98 ± 0.24, p-value = 0.002) and it happens to be an enzyme which is also involved in de novo cholesterol biosynthesis (King, 2015).

6.1.2 HK2

Hexokinase-2, i.e. HK2 is an enzyme involved in the glucose metabolism pathways according to information in NCBI. It is commonly located in the outer membrane of mitochondria and once dissociated from it HK2 triggers apoptosis via mitochondrial membrane permeabilization and release of apoptogenic proteins (Chiara et al., 2008).

It should be recalled that the true significance of this protein is in this point difficult to determine as further the ranking of this particular protein was 1176 for IOBA-NHC and 1169 for HCE and this is not as high as one could have hoped for and the results and visualizations suggested that the actual underlying difference between samples with regards to this protein are not as clear as desired since the expression levels between preservative-free and preservative treated samples were "flipped" between the cell lines. Hence, there should be more caution about the actual significance of this protein with regards the BAC-induced effects.

6.1.3 Myosin enzymes

Overall, there are not many groups of proteins that initially stand out from the list of proteins as similarly expressed but one group worth noting are the myosin-related proteins. There are already some previous studies which have identified that the myosins do play a role in BAC-induced inflammation (Guo et al., 2007; Droy-Lefaix et al., 2013). Myosins and their related myosin light chains (MLC), and more specifically the phosphorylation of MLC is needed to produce the contractions of the actin cytoskeleton and this event in turn regulates the barrier integrity, adhesion and migration. Furthermore, the study by Guo noted that exposure to BAC leads to a reduced phosphorylation of MLC, which in turns affects the contractility of the actin cytoskeleton and the maintenance of epithelial barrier suffers from these changes as the cell migration ability is reduced. Combined with reduced levels of proliferation and adhesion, adverse effects are likely to take place.

MLC kinase phosphorylates the MLC whilst MLC phosphatase dephosphorylates it. The combination of these two events results in cell migration. As the cytoskeleton contracts, it can be further thought that the paracellular permeability is increased as the intercellular tight junctions are opened. A study by Droy-Lefaix et al. (2013) based their starting point on this information and evaluated how the inhibition of MLC phosphorylation affected the BAC-induced effects in the eyes of rats. According to their results, inhibition of cytoskeleton contractions via MLC kinase inhibitor reduces the inflammatory effects of BAC.

It should be noted, that articles by Droy-Lefaix and Guo claim somewhat different things, but the main point remains the same; BAC is affecting the actin cytoskeleton contractions via MLC. Whether the inflammatory effects are caused by phosphorylation or dephosphorylation does not appear to be clear yet and more research is needed. Based on the results obtained from this study, it appears that the over-expression of the three myosins, or myosin light chains, mentioned does suggest increased mobility of the cells and it could be expected that the adhesion is this way also increased as described by Droy-Lefaix. Based on these results, it is not difficult to believe, that inhibition of MLC could reduce the inflammation in eyes, or at least reduce some of the adverse effects.

6.1.4 Mitochondrial membrane respiratory chain NADH dehydrogenase (Complex I)

It has already been discussed previously that BAC is inducing oxidative stress in human conjunctival cells and the under-expression of both NDUFA5 and NDUFS3 in the IOBA-NHC cell line results

appears to further confirm these observations. These two proteins are subunits of the mitochondrial membrane respiratory chain NADH dehydrogenase (Complex I), which will be referred to simply as Complex I here. When Complex I becomes inhibited, as has happened with the preservative-treated samples in IOBA-NHC cells in this study, the production of reactive oxygen species (ROS) is increased (Raha & Robinson, 2000; Fato et al., 2009). Similar results were shown in an article by Nianyu et al. (2003), where rotenone, which inhibits Complex I, was shown to induce apoptosis by affecting the ROS production. Based on this, it could be hypothesised that the two proteins discussed here, and their under-expression is causing apoptosis of IOBA-NHC cells via ROS induced oxidative stress.

6.2 Enrichment analyses

Next the results obtained from the enrichment analyses are discussed. These results are largely complementing many aspects conversed in the previous subsections when interesting individual proteins and their potential effects were discussed.

The enrichment results from DAVID had the highest ESs for "Generation of precursor metabolites and energy" for IOBA-NHC cell line and "Actin filament-based process" for HCE cell line in the biological process domain. These results support the evidence provided by some of the statistically significant proteins discussed earlier. More specifically, this could further strengthen the arguments about conjunctival cells experiencing oxidative stress due to defects in the electron transport chain as a result of exposure to BAC. Additional proteins affecting the enrichment of this term, in addition to NDUFA5 and NDUFS3, were glycogen phosphorylases, liver form (PYGL) and brain form (PYGB) and HK2. Further on, the actin filament-based movement was also discussed earlier in relation to myosin enzymes and as expected, MYL6 and MYH9 are associated to this GO term together with keratin, type I cytoskeletal 19 (KRT19) and Ras-related C3 botulinum toxin substrate 2 (RAC2).

For both cell lines, the highest scoring terms in the cellular component domain were connected to mitochondrion and its parts. This further suggests, that BAC is largely affecting the mitochondrion and more studies are needed to investigate, what exactly is happening. One likely explanation is that BAC is causing instability in the mitochondrion membranes and, together with disrupting the respiratory chain functions, it could even cause lysis of this organelle structure. The destruction of mitochondrion, would result in apoptosis as this event would initiate apoptosis in cells.

The enrichment results achieved from GOrilla tool were slightly different, though the mitochondria and its parts were again highly enriched for the HCE in relation to proteins over-expressed in BAC-treated samples and for the IOBA-NHC in relation to proteins under-expressed in BAC-treated samples. The terms related to the mitochondrion were the most statistically significant terms in these results. It should be noted that the tables describing the results do contain some terms which have a very high, i.e. non-significant, q-values. These are nevertheless included in order to describe the top terms regardless of their individual statistical significance.

Furthermore, IOBA-NHC cell line results relating to the proteins over-expressed in the BAC-treated samples suggested that there were enrichment in relation to extracellular exosomes and actin binding. The actin binding again suggests increase in cell mobility and adhesion, but this time in IOBA-NHC cell line. The extracellular exosomes can be related to several cell functions but it is generally accepted, that they are playing a part in cell signaling and they can be connected to inflammation and immune system. In addition, interestingly, exosomes are thought be produced by cells to avoid accumulation of drugs or waste (van der Pol et al., 2012). Hence, potentially these extracellular vesicles are formed in conjunctiva cells in attempt to remove BAC-associated waste from them.

Enrichments with proteins, which were under-expressed in the BAC-treated samples did not result in any statistically significant results for HCE cell line but, as mentioned, IOBA-NHC cell line includes statistically significant terms relating to mitochondrion in cellular component domain and furthermore to mitochondrial translation initiation and mRNA splicing in biological process domain. This would suggest, that translation of mitochondrial genome is reduced and since the genome located in the mitochondria encode proteins related to the respiratory chain complex, this could further in its part explain how this pathway is interrupted with BAC (Kuzmenko et al., 2014).

Final part of the enrichment analysis consisted of the pathway enrichments, which unfortunately did not result in any further consistent and statistically significant results for IOBA-NHC cell line, although terms such as "apoptosis" and "focal adhesion" were mentioned. These terms had still such high p- and q-values that it is perhaps unnecessary to start forming specific hypotheses based on them. However, it could be argued that BAC is somehow affecting the energy production in cells, possibly via mitochondrion, which is then resulting in unwanted and apoptotic effects in the affected cells.

For HCE cell line only a few interesting terms appeared when results were ordered in such a manner that the proteins, which were over-expressed in the BAC-treated samples were on the top. According to the pathway enrichments, when the HCE cells are treated with BAC, this may affect the cells' citrate cycle and arginine and proline metabolism. However, based on my personal experience with

this tool, the latter term appears to be very frequently in the top enrichment results. Hence, the main interest should be on the first term mentioned, which also refers that the main BAC-induced effects are affecting the mitochondrion and its related functions.

6.3 Data assessment

This subsection will discuss some of the issues relating to the sample processing and data analysis. A critical approach is taken to evaluate the successfulness of the research overall and the ability to answer the proposed research questions. Some of the topics discussed here are not necessarily issues but merely thoughts about the data and methods used, which arose in the process of writing the study.

It should be noted, that there are no ethical issues relating to this study, as cell lines were used to perform this study. Since cell lines were used, the results should be treated carefully as the situation in real-life could differ from the results in some aspects. However, this should not be a major concern as the cell lines in this study have been studied extensively in similar studies and they have been estimated to represent the actual human epithelial conjunctival and corneal cells well (as mentioned previously). For the HCE cell line, another commercially available untransfected cell line is the Wong-Kilbourne derivative of Chang conjunctival cell line. Unfortunately, this cell line is reported to suffer from contamination with HeLa cells. However, it has also been noted that IOBA-NHC cells do show some differences in phenotypic and karyotypic patterns in comparison to normal conjunctival epithelium (Brasnu et al., 2008b). Brasnu et al. (2008b) also identified further small differences between these two cell lines. Chang cells could be more sensitive towards the effects of BAC despite the reported contamination. However, the study concluded that both cell lines can be useful in toxicological *in vitro* studies.

In addition to the SWATH data, the data was also produced using iTRAQ-labeling method. However, the protein quantification steps prior to any further data analysis were not successful and hence, this data was discarded. This does not affect the results here, though it would have been interesting and informative to compare the results obtained in these two different methods. Perhaps some further confirmation of the significance of some of the proteins could have been confirmed. However, this can be similarly achieved in further studies which are currently being carried out for more treatments, including the ones mentioned in this study.

The data annotation was performed with IPI annotations as previously mentioned. However, IPI database was discontinued in September 2011 and unfortunately for this data, the switch to mapping

with UniProt Knowledgebase was not yet completed and as a result the data were mapped using potentially outdated IPI accessions. However, as the gene symbols were additionally readily available in the data, majority of the proteins could be quite easily mapped to known genes or proteins. Yet, in an ideal case, the proteins accessions would have been according to the new recommendations (Griss et al., 2011).

Another aspect, again this cannot be considered as an issue per se, but more as an aspect to keep in mind, is that the cells' exposure to any given treatment is 24 hours. One of the topics of this study discusses how the preservatives are affecting sensitive patients, who are exposed to these treatments often for extensive periods of time, i.e. years or decades. Hence, a cell line's exposure to the treatment for only 24 hours is not fully comparable to this, and it is not attempted either. This study merely tries to enlighten some of the possible early steps occurring in the corneal and conjunctival cells and the reality in patients could naturally be very different from this, and from each other.

Furthermore, it would have been interesting, if the concentrations of BAC, and naturally other treatments, had been varied in this study, which could further verify the obtained results. Incidentally, this is being done in the next related study, which is being processed currently.

One of the main issues affecting the data analysis quality of the study in my opinion was the number of biological samples as this was rather small for HCE cell line in particular. Ideally, the number should have been higher than two or three, which was not the case here. Small number of biological replicates can result in reduced quality of results as the number of false positives and negatives can rise. However, this problem becomes more significant when the quality of measurement techniques is not good to begin with. In this particular study, as SWATH method was implemented in the protein quantification, I would argue that the issue is not as severe as it could be with some other methods. Furthermore, it can be thought that the sensitivity was slightly increased by implementing the mixed effects model instead of merely taking the means of the technical replicates. Yet, in further studies, if the resources allowed, I would be inclined to wish for larger numbers of biological replicates per treatment.

7 Conclusion

This study had more than one aim. First, it was performed in order to merely test if, and how well, this type of study would work out. More specifically, we were interested to examine how immortalized cell lines would respond to these treatments and how the resulting proteomics would turn out. This type of study has not previously been performed using MS SWATH method. As visualized in the results, we demonstrated that evidently this type of research is possible and it does produce some results, which are, based on technical replicates, of high quality, whilst still showing some interesting differences between the varying treatments.

The aim of this study is not to tell that all preservatives are bad and should not be used in any topical treatments of glaucoma. Instead, it should merely be acknowledged that there is a subgroup of patients who are currently not benefitting from the medication provided for them. By examining the pathways and effects of BAC, perhaps some new approaches to the topical medications and their ingredients could be developed for these patients in particular.

The second aim of this study, which is what this paper is mostly concentrating on is, what proteins are showing different abundance levels between the treatments and what could explain these differences in a larger, biological scale. It was hoped, that once these proteins, and their place in the vast network of biological processes was established, a few chosen ones could be further verified in following studies. Once a confirmation of the effects of BAC in a proteomic level can be achieved, it can hopefully help us in development of new medication or in fact new preservatives, which could be beneficial particularly to those patients needing long-term ophthalmic topical treatment who are sensitive to this very common preservative.

Several results observed in this study strongly suggest that changes in mitochondrion functions are affected by exposure to BAC. This became evident not only from the enrichment analyses but it was also noted that in IOBA-NHC cell line, NDFUA5 and NDUFS3 were under-expressed for BAC-treated samples. These proteins are associated to the mitochondrial membrane respiratory chain NADH dehydrogenase and once this process is inhibited, it is noted to increase ROS production, which could then in turn lead to oxidative stress. Furthermore, in both cell lines the cholesterol production and therefore the plasma membrane permeability and structure could be altered due to reduced abundance of HMGCS1, which is an essential catalyst in this process. It would further appear, with both individual proteins and enrichment results, that actin cytoskeleton contractions are at least in the HCE cell line increased, which then in turn affects the permeability of the cell junctions.

This was initially noted due to the over-expression of MYH9, MYL12A and MYL6 in HCE cell line samples treated with BAC.

Together these results show not only supporting evidence of the previously conducted studies but also identify individual proteins, which could potentially play a key role in the effects induced in cornea and conjunctiva cells as a result of exposure to BAC.

As already mentioned a few times, currently a new, very similar study is being carried out with more treatments. More specifically only IOBA-NHC cell are exposed to the treatments in this study and at least two new treatments are included in the study. These new added treatments are unfavorably affecting some patients but they do not contain BAC and hence it will be of interest not only to confirm some of the results seen in study but to also add more possible comparisons in the equation. Furthermore, the concentrations of treatments are also varied in this new study, which will further complicate the structure of statistical analysis but it is hoped to at the same time to provide more reliable answers to the current and future research questions. The processing and analysis of the samples will again be carried out using MS SWATH method.

8 References

Araki-Sasaki, K., Ohashi, Y., Sasabe, T., Hayashi, K., Watanabe, H., Tano, Y., & Handa, H. (1995). An SV40-immortalized human corneal epithelial cell line and its characterization. Investigative Ophthalmology & Visual Science, 36(3), 614-621.

Azari, A. A., & Barney, N. P. (2013). Conjunctivitis

A systematic review of diagnosis and treatment. The Journal of the American Medical Association, 310(16), 1721-1730. doi:10.1001/jama.2013.280318

Baudouin, C., Labbé, A., Liang, H., Pauly, A., & Brignole-Baudouin, F. (2010). Preservatives in eyedrops: The good, the bad and the ugly. Progress in Retinal and Eye Research, 29(4), 312-334. doi:10.1016/j.preteyeres.2010.03.001

Brasnu, E., Brignole-Baudouin, F., Riancho, L., Guenoun, J. M., Warnet, J. M., & Baudouin, C. (2008). In vitro effects of preservative-free tafluprost and preserved latanoprost, travoprost, and bimatoprost in a conjunctival epithelial cell line. Current Eye Research, 33, 303-312. doi:10.1080/02713680801971857

Brasnu, E., Brignole-Baudouin, F., Riancho, L., Warnet, J. M., & Baudouin, C. (2008). Comparative study on the cytotoxic effects of benzalkonium chloride on the wong-kilbourne derivative of chang conjunctival and IOBA-NHC cell lines . Molecular Vision, 14, 394-402.

Broadway, D. C., Grierson, I., O'Brien, C., & Hitchings, R. A. (1994). Adverse effects of topical antiglaucoma medication: II. the outcome of filtration surgery. Archives of Ophthalmology, 112(11), 1446-1454. doi:10.1001/archopht.1994.01090230060021

Buron, N., Micheau, O., Cathelin, S., Lafontaine, P. O., Creuzot-Garcher, C., & Solary, E. (2006). Differential mechanisms of conjunctival cell death induction by ultraviolet irradiation and benzalkonium chloride. Biochemistry and Molecular Biology, 47, 4221-4230. doi:10.1167/iovs.05-1460

Callister, S. J., Barry, R. C., Adkins, J. N., Johnson, E. T., Qian, W., Webb-Robertson, B. J., & ...Lipton, M. S. (2006). Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. Journal of Proteome Research, 5(2), 277-286. doi:10.1021/pr050300l

Chandrasekaran, S., Cumming, R. G., Rochtchina, E., & Mitchell, P. (2006). Associations between elevated intraocular pressure and glaucoma, use of glaucoma medications, and 5-year incident cataract: The blue mountains eye study. Ophthalmology, 113(3), 417-424.

Chiara, F., Castellaro, D., Marin, O., Petronilli, V., Brusilow, W. S., Juhaszova, M., & ... Rasola, A. (2008). Hexokinase II detachment from mitochondria triggers apoptosis through the permeability transition pore independent of voltage-dependent anion channels. Plos One, 3(3), e1852. doi:10.1371/journal.pone.0001852

Clouzeau, C., Godefroy, D., Riancho, L., Rostène, W., Baudouin, C., & Brignole-Baudouin, F. (2012). Hyperosmolarity potentiates toxic effects of benzalkonium chloride on conjunctival epithelial cells in vitro. Molecular Vision, 18, 851-863.

Collins, B. C., Gillet, L. C., Rosenberger, G., Röst, H. L., Vichalkovski, A., Gstaiger, M., & Aebersold, R. (2013). Quantifying protein interaction dynamics by SWATH mass spectrometry: Application to the 14-3-3 system. Nature Methods, 10, 1246-1253. doi:10.1038/nmeth.2703

Cover, T. M., & Thomas, J. A. (1991). Entropy, relative entropy and mutual information. Elements of information theory (pp. 12) John Wiley & Sons, Inc.

Crockett, E. L. (1998). Cholesterol function in plasma membranes from ectotherms: Membrane-specific roles in adaptation to temperature. American Zoologist, 38, 291-304.

Cui, X., & Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. Genome Biology, 4(4) doi:10.1186/gb-2003-4-4-210

Diebold, Y., Calonge, M., Enriquez de Salamanca, A., Callejo, S., Corrales, R. M., Sáez, V., & ...Stern, M. E. (2003). Characterization of a spontaneously immortalized cell line (IOBA-NHC) from normal human conjunctiva. Investigative Ophthalmology & Visual Science, 44(10), 4263-4274. doi:10.1167/iovs.03-0560

Droy-Lefaix, M. T., Bueno, L., Caron, P., Belot, E., & Roche, O. (2013). Ocular inflammation and corneal permeability alterationby benzalkonium chloride in rats: A protective effect of a myosin light chain kinase inhibitor. Investigative Ophthalmology & Visual Science, 54, 2705-2710. doi:10.1167/iovs.12-10193

Eden, E., Lipson, D., Yogev, S., & Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. PLoS Computational Biology, 3(3) doi:10.1371/journal.pcbi.0030039

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics, 10(48) doi:10.1186/1471-2105-10-48

Epstein, S. P., Chen, D., & Asbell, P. A. (2009). Evaluation of biomarkers of inflammation in response to benzalkonium chloride on corneal and conjunctival epithelial cells. Journal of Ocular Pharmacology and Therapeutics, 25(5), 415-424. doi:10.1089/jop.2008.0140

Falcon, S., & Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. Bioinformatics, 23(2), 257-258. doi:10.1093/bioinformatics/btl567

Fato, R., Bergamini, C., Bortolus, M., Maniero, A. L., Leoni, S., Ohnishi, T., & Lenaz, G. (2009). Differential effects of mitochondrial complex I inhibitors on production of reactive oxygen species. Biochimica Et Biophysica Acta, 1787(5), 384-392. doi:10.1016/j.bbabio.2008.11.003

Fingert, J. H. (2011). Primary open-angle glaucoma genes. Eye, 25, 587-595. doi:10.1038/eye.2011.97

Freeman, P. D., & Kahook, M. Y. (2009). Preservatives in topical ophthalmic medications: Historical and clinical perspectives. Expert Review of Ophthalmology, 4(1), 59-64.

Furrer, P., Mayer, J. M., & Gurny, R. (2002). Ocular tolerance of preservatives and alternatives. European Journal of Pharmaceutics and Biopharmaceutics, 53, 263-280.

Griss, J., Martín, M., O'Donovan, C., Apweiler, R., Hermjakob, H., & Vizcaíno, J. A. (2011). Consequences of the discontinuation of the international protein index (IPI) database and its substitution by the UniProtKB ''complete proteome'' sets. Proteomics, 11, 4434-4438.

Guo, Y., Satpathy, M., Wilson, G., & Srinivas, S. P. (2007). Benzalkonium chloride induces dephosphorylation of myosin light chain in cultured corneal epithelial cells. Investigative Ophthalmology & Visual Science, 48(5), 2001-2008.

Gutstein, H. B., Morris, J. S., Annangudi, S. P., & Sweedler, J. V. (2009). Microproteomics: Analysis of protein diversity in small samples. Mass Spectrometry Reviews, 27(4), 316-330.

Hu, J., & He, X. (2007). Enhanced quantile normalization of microarray data to reduce loss of information in gene expression profiles. Biometrics, 63(1), 50-59. doi:10.1111/j.1541-0420.2006.00670.x

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research, 37(1), 1-13. doi:10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols, 4(1), 44-57. doi:10.1038/nprot.2008.211

Huang, Q., Yang, L., Luo, J., Guo, L., Wang, Z., Yang, X., & ... Zhang, Y. (2015). SWATH enables precise label-free quantification on proteome scale. Proteomics, 15, 1215-1223. doi:10.1002/pmic.201400270

Huhtala, A., Mannerström, M., Alajuuma, P., Nurmi, S., Toimela, T., Tähti, H., & ... Uusitalo, H. (2002). Comparison of an immortalized human corneal epithelial cell line and rabbit corneal epithelial cell culture in cytotoxicity testing. Journal of Ocular Pharmacology and Therapeutics, 18(2), 163-175.

Karpievitch, Y. V., Dabney, A. R., & Smith, R. D. (2012). Normalization and missing value imputation for label-free LC-MS analysis. BMC Bioinformatics, 13(Suppl 16), S5. doi:10.1186/1471-2105-13-S16-S5

Kaur, I. P., Lal, S., Rana, C., Kakkar, S., & Singh, H. (2009). Ocular preservatives: Associated risks and newer options. Cutaneous and Ocular Toxicology, 28(3), 93-103.

King, M. W. (2015). Cholesterol. Retrieved from http://themedicalbiochemistrypage.org/cholesterol.php

Kroese, M., & Burton, H. (2003). Primary open angle glaucoma. the need for a consensus case definition. Journal of Epidemiology and Community Health, 57(9), 752-754.

Kumar, C., & Mann, M. (2009). Bioinformatics analysis of mass spectrometry-based proteomics data sets. FEBS Letters, 583, 1703-1712. doi:10.1016/j.febslet.2009.03.035

Kuzmenkoa, A., Atkinsona, G. C., Levitskiib, S., Zenkinc, N., Tensona, T., Hauryliuka, V., & Kamenskib, P. (2014). Mitochondrial translation initiation machinery: Conservation and diversification. Biochimie, 100, 132-140. doi:10.1016/j.biochi.2013.07.024

Lang, G. K. (2007). Glaucoma. Ophthalmology : A pocket textbook atlas (2nd edition ed., pp. 239). Germany: Georg Thieme Verlang.

Li, N., Ragheb, K., Lawler, G., Sturgis, J., Rajwa, B., Melendez, J. A., & Robinson, J. P. (2003). Mitochondrial complex I inhibitor rotenone induces apoptosis through enhancing mitochondrial reactive oxygen species production. The Journal of Biological Chemistry, 278, 8516-8525.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics, 7(Suppl 1), S7. doi:10.1186/1471-2105-7-S1-S7

McCarthy, D. J., & Gordon, K. S. (2009). Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics, 25(6), 765-771. doi:10.1093/bioinformatics/btp053

Meyer, P. E., Lafitte, F., & Bontempi, G. (2008). Minet: An open source R/Bioconductor package for mutual information based network inference. BMC Bioinformatics, 9:461 doi:10.1186/1471-2105-9-461

Mirman, D. (2014, February 11). Three ways to get parameter-specific p-values from lmer. Retrieved from http://mindingthebrain.blogspot.fi/2014/02/three-ways-to-get-parameter-specific-p.html

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., & ... Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics, 34, 267-273. doi:10.1038/ng1180

Noble, W. S., & MacCoss, M. J. (2012). Computational and statistical analysis of protein mass spectrometry data. PLoS Computational Biology, 8(1) doi:10.1371/ journal.pcbi.1002296

Noecker, R. (2001). Effects of common ophthalmic preservatives on ocular health. Advances in Therapy, 18(5), 205-215.

Noecker, R. J. (2006). The management of glaucoma and intraocular hypertension: Current approaches and recent advances. Therapeutics and Clinical Risk Management, 2(2), 193-206.

Pellinen, P., Huhtala, A., Tolonen, A., Lokkila, J., Mäenpää, J., & Uusitalo, H. (2012). The cytotoxic effects of preserved and preservative-free prostaglandin analogs on human corneal and conjunctival epithelium in vitro and the distribution of benzalkonium chloride homologs in ocular surface tissues in vivo. Current Eye Research, 37(2), 145-154. doi:10.3109/02713683.2011.626909

Perez-Riverol, Y. (2014,). SWATH-MS and next-generation targeted proteomics.

Pisella, P. J., Pouliquen, P., & Baudouin, C. (2002). Prevalence of ocular symptoms and signs with preserved and preservative free glaucoma medication. British Journal of Ophthalmogy, 86(4), 418-423.

Quigley, H. A., & Broman, A. T. (2006). The number of people with glaucoma worldwide in 2010 and 2020. British Journal of Ophthalmogy, 90(3), 262-267. doi:10.1136/bjo.2005.081224

Raha, S., & Robinson, B. H. (2000). Mitochondria, oxygen free radicals, disease and ageing. Trends in Biochemical Sciences, 25(10), 502-508. doi:10.1016/S0968-0004(00)01674-1

Salim, S. (2012). The Role of Anterior Segment Optical Coherence Tomography in Glaucoma. Journal of Ophthalmology, 2012. doi:10.1155/2012/476801

Schmidt, A., Forne, I., & Imhof, A. (2014). Bioinformatic analysis of proteomics data. BMC Systems Biology, 8(Suppl 2)

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., & … Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research, 13(11), 2498-2504. doi:10.1101/gr.1239303

Sheng, R., Chen, Y., Gee, H. Y., Stec, E., Melowic, H. R., Blatner, N. R., & ... Cho, W. (2012). Cholesterol modulates cell signaling and protein networking by specifically interacting with PDZ domain-containing scaffold proteins. Nature Communications, 3, 1249. doi:10.1038/ncomms2221

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., & ...Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545–15550. doi:10.1073/pnas.0506580102

Thomas Bond & Partners. (2013). Glaucoma. Retrieved from http://www.thomasbond.co.uk/eye-health/glaucoma.html

Uusitalo, H., Chen, E., Pfeiffer, N., Brignole-Baudouin, F., Kaarniranta, K., Leino, M., & ...Ropo, A. (2010). Switching from a preserved to a preservative-free prostaglandin preparation in topical glaucoma medication. Acta Ophthalmologica, 88, 329-336. doi:10.1111/j.1755-3768.2010.01907.x

van der Pol, E., Böing, A. N., Harrison, P., Sturk, A., & Nieuwland, R. (2012). Classification, functions, and clinical relevance of extracellular vesicles. Pharmacological Reviews, 64(3), 676-705.

Wiggs, J. L. (2007). Genetic etiologies of glaucoma. Archives of Ophthalmology, 125(1), 30-37. doi:10.1001/archopht.125.1.30

Zhou, L., & Beuerman R. W. (2012). Tear analysis in ocular surface diseases. Progress in Retinal and Eye Research, 31(6), 527-550. doi:10.1016/j.preteyeres.2012.06.002

Appendix A – pre-normalized data
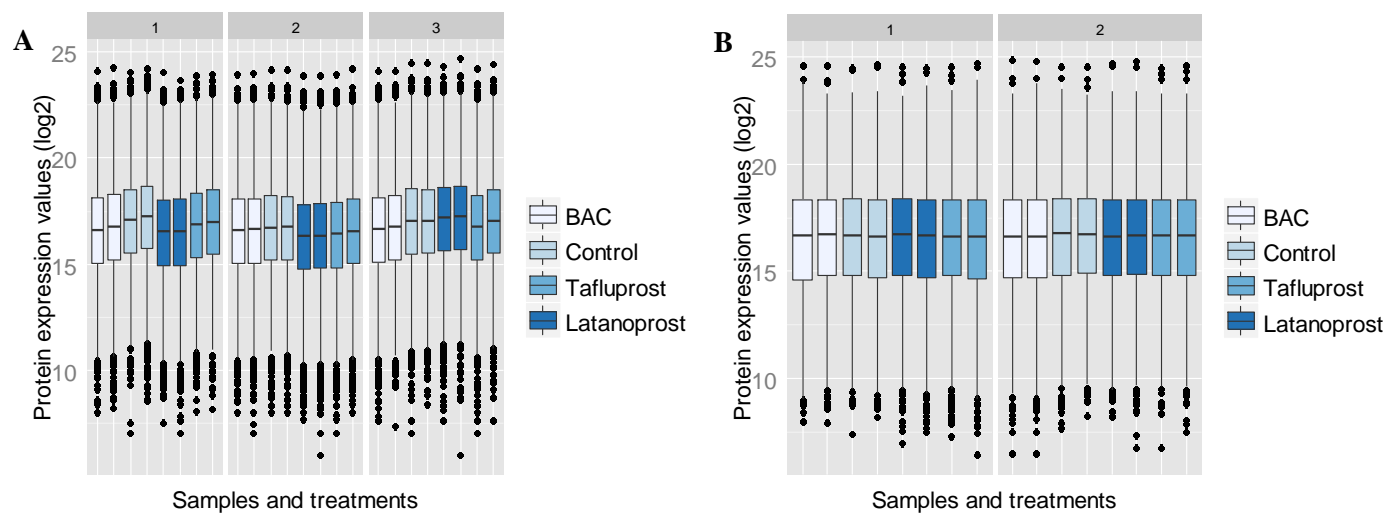
Figure A-1 Boxplots of the non-normalized data. As it can be seen, the IOBA-NHC cell line data (A) suffers from more variation than the HCE data (B). The HCE data was in fact nearly perfect in terms of median and variance similarity. However, in order to be consistent with the data analysis, normalization was applied to both even though the effects on HCE data are likely to be very minor.

Appendix B – MA-plots of technical replicates
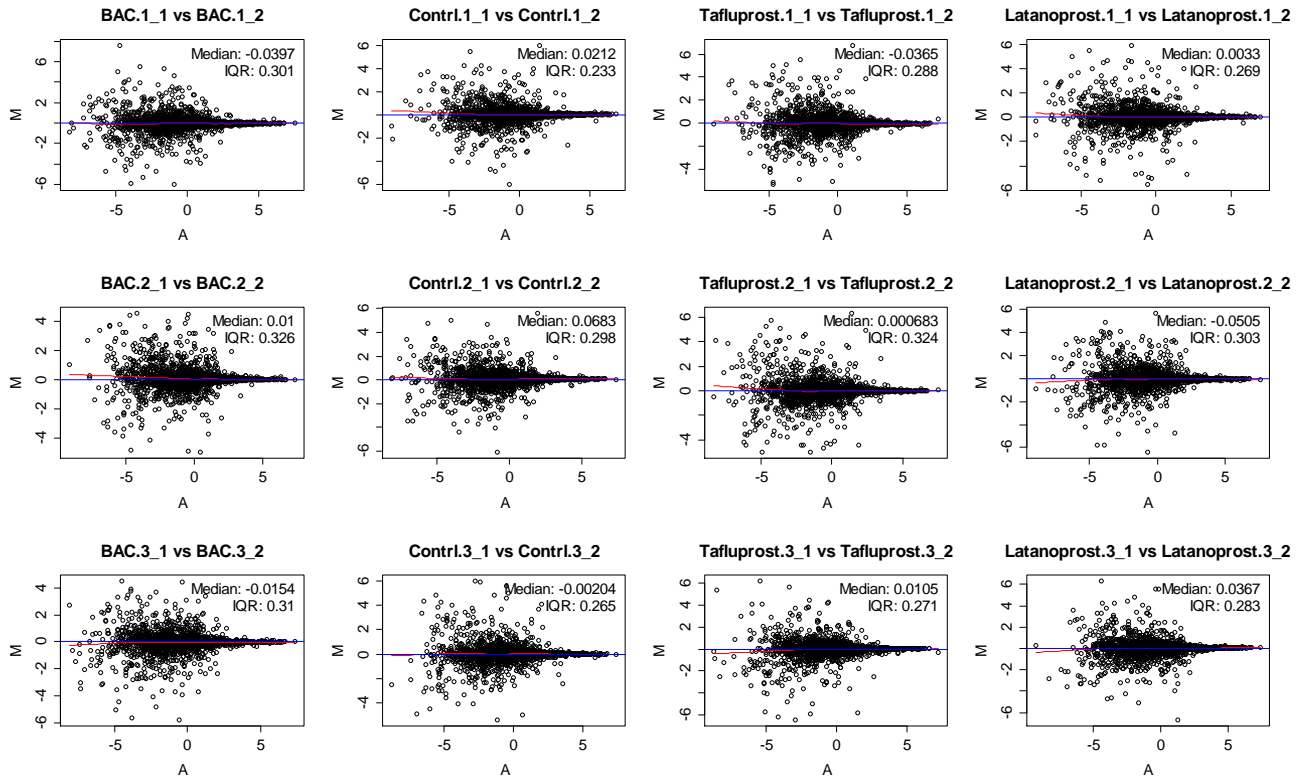
Figure B-1 MA-plots of technical replicates for IOBA-NHC cell line data. These plots are based on the normalized data and the aim is to establish if there was any need for any other forms of normalization as well. It appears that there are no specific trends here as the red line (trend) is very flat, horizontal and close to the blue line which is a line specifying zero on y-axis. Also, the median in each plot is relatively close to zero, which further confirms, that the data is sufficiently normalized between the technical replicates and that they are of good quality.
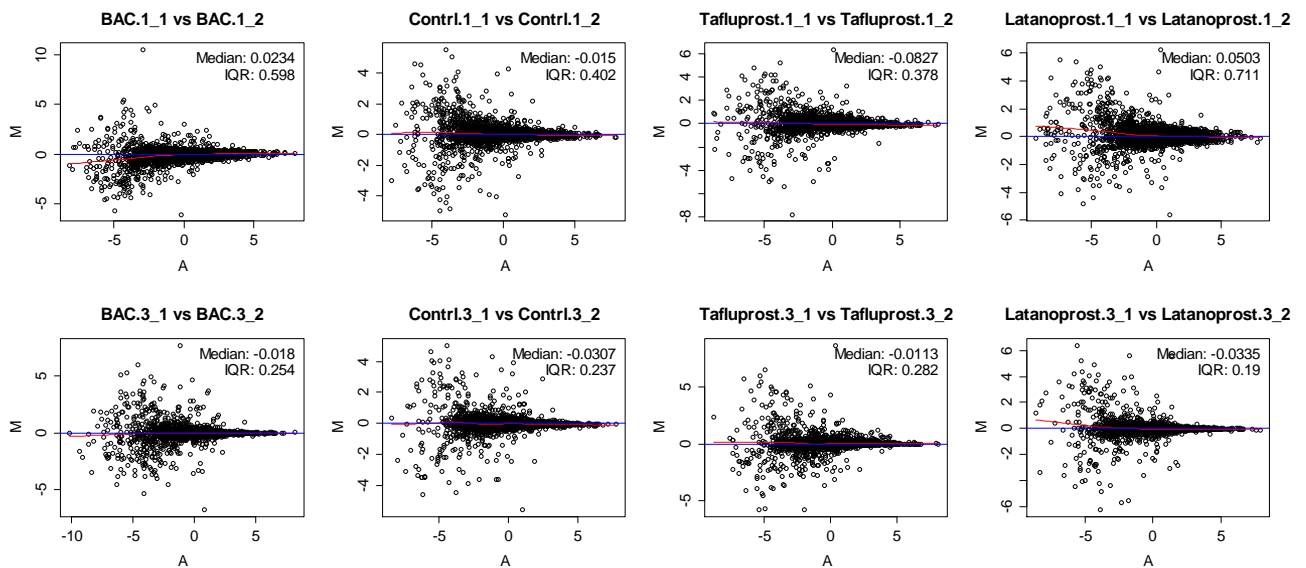
Figure B-2 MA-plots of technical replicates for HCE cell line data. These plots are based on the normalized data and similar to Figure B-1, the aim is to establish if there was any need for any other forms of normalization as well. Again, there are no worryingly large trends visible and the medians are close to zero. It can be concluded that the technical replicates are of good quality for this cell line as well.

Appendix C – More detailed tables of the GO term enrichment results from DAVID

Table C-1 Clustered enrichment terms under biological process domain for IOBA-NHC cell line results.

| Domain	Cluster/ES	Term	Count	PValue	Gene symbol	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	Cluster 1 - Enrichment Score: 1.40	GO:0006091~generation of precursor metabolites and energy	5	0.01	NDUFA5, PYGL, HK2, NDUFS3, PYGB	5.19	0.97	0.84	13.87
		GO:0019318~hexose metabolic process	4	0.03	GALK1, PYGL, HK2, PYGB	5.66	1.00	0.95	30.82
		GO:0005996~monosaccharide metabolic process	4	0.04	GALK1, PYGL, HK2, PYGB	5.10	1.00	0.95	38.55
		GO:0044275~cellular carbohydrate catabolic process	3	0.06	PYGL, HK2, PYGB	7.23	1.00	0.98	54.87
		GO:0016052~carbohydrate catabolic process	3	0.06	PYGL, HK2, PYGB	6.86	1.00	0.97	58.39
		GO:0006006~glucose metabolic process	3	0.11	PYGL, HK2, PYGB	4.87	1.00	0.98	80.51
	Cluster 2 - Enrichment Score: 1.25	GO:0006091~generation of precursor metabolites and energy	5	0.01	NDUFA5, PYGL, HK2, NDUFS3, PYGB	5.19	0.97	0.84	13.87
		GO:0016310~phosphorylation	4	0.07	GALK1, NDUFA5, HK2, NDUFS3	3.84	1.00	0.97	63.56
		GO:0006793~phosphorus metabolic process	4	0.11	GALK1, NDUFA5, HK2, NDUFS3	3.16	1.00	0.99	80.43
		GO:0006796~phosphate metabolic process	4	0.11	GALK1, NDUFA5, HK2, NDUFS3	3.16	1.00	0.99	80.43
	Cluster 3 - Enrichment Score: 0.33	GO:0006886~intracellular protein transport	3	0.32	AP3D1, NUP54, AIP	2.48	1.00	1.00	99.46
		GO:0034613~cellular protein localization	3	0.34	AP3D1, NUP54, AIP	2.35	1.00	1.00	99.67
		GO:0070727~cellular macromolecule localization	3	0.35	AP3D1, NUP54, AIP	2.33	1.00	1.00	99.70
		GO:0046907~intracellular transport	3	0.57	AP3D1, NUP54, AIP	1.52	1.00	1.00	100.00
		GO:0015031~protein transport	3	0.60	AP3D1, NUP54, AIP	1.45	1.00	1.00	100.00
		GO:0045184~establishment of protein localization	3	0.60	AP3D1, NUP54, AIP	1.45	1.00	1.00	100.00
		GO:0008104~protein localization	3	0.64	AP3D1, NUP54, AIP	1.34	1.00	1.00	100.00

Table C-2 Clustered enrichment terms under biological process domain for HCE cell line results.

Domain	Cluster/ES	Term	Count	PValue	Gene symbol	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	Cluster 1 - Enrichment Score: 0.90	GO:0030029~actin filament-based process	4	0.04	MYL6, KRT19, RAC2, MYH9	5.15	1.00	1.00	38.97
		GO:0030036~actin cytoskeleton organization	3	0.16	KRT19, RAC2, MYH9	3.94	1.00	1.00	91.31
		GO:0007010~cytoskeleton organization	3	0.34	KRT19, RAC2, MYH9	2.36	1.00	1.00	99.70
	Cluster 2 - Enrichment Score: 0.80	GO:0051345~positive regulation of hydrolase activity	3	0.04	SSBP1, HSPE1, HSPD1	8.60	1.00	1.00	44.66
		GO:0051336~regulation of hydrolase activity	3	0.11	SSBP1, HSPE1, HSPD1	5.11	1.00	1.00	78.59
		GO:0043085~positive regulation of catalytic activity	3	0.36	SSBP1, HSPE1, HSPD1	2.31	1.00	1.00	99.76
		GO:0044093~positive regulation of molecular function	3	0.41	SSBP1, HSPE1, HSPD1	2.08	1.00	1.00	99.92

Table C-3 Clustered enrichment terms under cellular component domain for IOBA-NHC cell line results.

Domain	Cluster/ES	Term	Count	PValue	Gene symbol	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_CC_FAT	Cluster 1 - Enrichment Score: 0.97	GO:0031966~mitochondrial membrane	4	0.05	NDUFA5, HK2, NDUFS3, HADHA	4.44	0.99	0.99	43.34
		GO:0005739~mitochondrion	8	0.05	NDUFA5, MRPS35, TPP1, HK2, KIAA0101, MACROD1, NDUFS3, HADHA	2.12	0.99	0.91	44.86
		GO:0005740~mitochondrial envelope	4	0.07	NDUFA5, HK2, NDUFS3, HADHA	3.94	1.00	0.87	53.84
		GO:0044429~mitochondrial part	5	0.10	NDUFA5, MRPS35, HK2, NDUFS3, HADHA	2.67	1.00	0.89	66.54
		GO:0031967~organelle envelope	5	0.11	NDUFA5, HK2, NUP54, NDUFS3, HADHA	2.57	1.00	0.86	70.70
		GO:0031975~envelope	5	0.11	NDUFA5, HK2, NUP54, NDUFS3, HADHA	2.57	1.00	0.86	70.70
		GO:0005743~mitochondrial inner membrane	3	0.15	NDUFA5, NDUFS3, HADHA	4.13	1.00	0.83	82.82
		GO:0019866~organelle inner membrane	3	0.19	NDUFA5, NDUFS3, HADHA	3.62	1.00	0.77	89.06
		GO:0031090~organelle membrane	4	0.34	NDUFA5, HK2, NDUFS3, HADHA	1.85	1.00	0.91	98.90
	Cluster 2 - Enrichment Score: 0.58	GO:0070013~intracellular organelle lumen	10	0.14	MRPS35, HMGB2, UBE2Z, RTCD1, FMR1, DDX21, NUP54, GTF3C4, HADHA, MRTO4	1.53	1.00	0.88	79.71
		GO:0043233~organelle lumen	10	0.15	MRPS35, HMGB2, UBE2Z, RTCD1, FMR1, DDX21, NUP54, GTF3C4, HADHA, MRTO4	1.52	1.00	0.85	81.59
		GO:0031981~nuclear lumen	8	0.16	HMGB2, UBE2Z, RTCD1, FMR1, DDX21, NUP54, GTF3C4, MRTO4	1.65	1.00	0.81	84.70
		GO:0031974~membrane-enclosed lumen	10	0.16	MRPS35, HMGB2, UBE2Z, RTCD1, FMR1, DDX21, NUP54, GTF3C4, HADHA, MRTO4	1.49	1.00	0.78	84.72
		GO:0005730~nucleolus	5	0.23	HMGB2, UBE2Z, FMR1, DDX21, MRTO4	1.91	1.00	0.83	94.28
		GO:0005654~nucleoplasm	5	0.28	HMGB2, RTCD1, FMR1, NUP54, GTF3C4	1.76	1.00	0.87	97.28
		GO:0043232~intracellular non-membrane-bounded organelle	7	0.86	MRPS35, HMGB2, UBE2Z, FMR1, DDX21, HADHA, MRTO4	0.85	1.00	1.00	100.00
		GO:0043228~non-membrane-bounded organelle	7	0.86	MRPS35, HMGB2, UBE2Z, FMR1, DDX21, HADHA, MRTO4	0.85	1.00	1.00	100.00

Table C-4 Clustered enrichment terms under cellular component domain for HCE cell line results.

Domain	Cluster/ES	Term	Count	PValue	Gene symbol	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_CC_FAT	Cluster 1 - Enrichment Score: 2.18	GO:0044429~mitochondrial part	9	0.00	GOT2, SHMT2, GRPEL1, UQCRC1, SSBP1, NIPSNAP1, HSPE1, HSPD1, OAT	5.46	0.01	0.01	0.08
		GO:0005743~mitochondrial inner membrane	5	0.00	GOT2, SHMT2, UQCRC1, NIPSNAP1, HSPD1	7.80	0.26	0.06	2.93
		GO:0019866~organelle inner membrane	5	0.00	GOT2, SHMT2, UQCRC1, NIPSNAP1, HSPD1	6.65	0.42	0.09	5.23
		GO:0031966~mitochondrial membrane	5	0.01	GOT2, SHMT2, UQCRC1, NIPSNAP1, HSPD1	6.50	0.45	0.08	5.70
		GO:0005740~mitochondrial envelope	5	0.01	GOT2, SHMT2, UQCRC1, NIPSNAP1, HSPD1	5.35	0.70	0.14	11.15
		GO:0031967~organelle envelope	6	0.02	GOT2, SHMT2, UQCRC1, RAC2, NIPSNAP1, HSPD1	3.31	0.95	0.28	24.83
		GO:0031975~envelope	6	0.02	GOT2, SHMT2, UQCRC1, RAC2, NIPSNAP1, HSPD1	3.31	0.95	0.28	24.83
		GO:0031090~organelle membrane	5	0.13	GOT2, SHMT2, UQCRC1, NIPSNAP1, HSPD1	2.44	1.00	0.76	78.44
	Cluster 2 - Enrichment Score: 2.15	GO:0044429~mitochondrial part	9	0.00	GOT2, SHMT2, GRPEL1, UQCRC1, SSBP1, NIPSNAP1, HSPE1, HSPD1, OAT	5.46	0.01	0.01	0.08
		GO:0005739~mitochondrion	11	0.00	GOT2, PYCR1, SHMT2, GRPEL1, UQCRC1, SSBP1, NIPSNAP1, RG9MTD1, HSPE1, HSPD1, OAT	3.32	0.05	0.02	0.47
		GO:0031980~mitochondrial lumen	7	0.00	GOT2, SHMT2, GRPEL1, SSBP1, HSPE1, HSPD1, OAT	6.06	0.06	0.02	0.57
		GO:0005759~mitochondrial matrix	7	0.00	GOT2, SHMT2, GRPEL1, SSBP1, HSPE1, HSPD1, OAT	6.06	0.06	0.02	0.57
		GO:0070013~intracellular organelle lumen	8	0.47	GOT2, PRIM1, SHMT2, GRPEL1, SSBP1, HSPE1, HSPD1, OAT	1.19	1.00	1.00	99.93
		GO:0043233~organelle lumen	8	0.49	GOT2, PRIM1, SHMT2, GRPEL1, SSBP1, HSPE1, HSPD1, OAT	1.18	1.00	0.99	99.95
		GO:0031974~membrane-enclosed lumen	8	0.52	GOT2, PRIM1, SHMT2, GRPEL1, SSBP1, HSPE1, HSPD1, OAT	1.15	1.00	0.99	99.98
	Cluster 3 - Enrichment Score: 0.76	GO:0016459~myosin complex	3	0.00	MYL6, MYL12A, MYH9	54.57	0.10	0.03	0.99
		GO:0015629~actin cytoskeleton	3	0.26	MYL6, MYL12A, MYH9	2.87	1.00	0.95	96.97
		GO:0044430~cytoskeletal part	4	0.50	MYL6, KRT19, MYL12A, MYH9	1.46	1.00	0.99	99.96
		GO:0005856~cytoskeleton	5	0.59	MYL6, KRT19, DRG1, MYL12A, MYH9	1.20	1.00	0.99	100.00
		GO:0043228~non-membrane-bounded organelle	9	0.65	MYL6, PRIM1, KRT19, SHMT2, SSBP1, DRG1, MYL12A, MYH9, NCAPD2	1.03	1.00	1.00	100.00
		GO:0043232~intracellular non-membrane-bounded organelle	9	0.65	MYL6, PRIM1, KRT19, SHMT2, SSBP1, DRG1, MYL12A, MYH9, NCAPD2	1.03	1.00	1.00	100.00
	Cluster 4 - Enrichment Score: 0.50	GO:0005886~plasma membrane	7	0.12	GOT2, LAMP1, KRT19, GNB1L, MARCKSL1, HSPD1, MYH9	1.93	1.00	0.77	76.41
		GO:0044459~plasma membrane part	4	0.33	LAMP1, GNB1L, HSPD1, MYH9	1.90	1.00	0.97	98.92
		GO:0016021~integral to membrane	3	0.48	LAMP1, HSPD1, MYH9	1.82	1.00	0.99	99.94
		GO:0031224~intrinsic to membrane	3	0.52	LAMP1, HSPD1, MYH9	1.67	1.00	0.99	99.98

Table C-5 Clustered enrichment terms under molecular function domain for IOBA-NHC cell line results.

Domain	Cluster/ ES	Term	Count	PValue	Gene symbol	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_MF_FAT	Cluster 1 - Enrichment Score: 0.84	GO:0048029~monosaccharide binding	3	0.01	GALK1, PYGL, HK2	17.46	0.71	0.71	11.59
		GO:0005529~sugar binding	3	0.02	GALK1, PYGL, HK2	11.91	0.93	0.73	22.95
		GO:0030246~carbohydrate binding	3	0.05	GALK1, PYGL, HK2	7.94	1.00	0.85	43.04
		GO:0005524~ATP binding	6	0.18	GALK1, UBE2Z, PYGL, RTCD1, HK2, DDX21	1.84	1.00	1.00	90.08
		GO:0032559~adenyl ribonucleotide binding	6	0.19	GALK1, UBE2Z, PYGL, RTCD1, HK2, DDX21	1.81	1.00	0.99	91.36
		GO:0030554~adenyl nucleotide binding	6	0.22	GALK1, UBE2Z, PYGL, RTCD1, HK2, DDX21	1.73	1.00	0.98	93.92
		GO:0001883~purine nucleoside binding	6	0.23	GALK1, UBE2Z, PYGL, RTCD1, HK2, DDX21	1.71	1.00	0.98	94.63
		GO:0001882~nucleoside binding	6	0.23	GALK1, UBE2Z, PYGL, RTCD1, HK2, DDX21	1.69	1.00	0.97	95.27
		GO:0032553~ribonucleotide binding	6	0.34	GALK1, UBE2Z, PYGL, RTCD1, HK2, DDX21	1.47	1.00	0.99	99.12
		GO:0032555~purine ribonucleotide binding	6	0.34	GALK1, UBE2Z, PYGL, RTCD1, HK2, DDX21	1.47	1.00	0.99	99.12
		GO:0017076~purine nucleotide binding	6	0.37	GALK1, UBE2Z, PYGL, RTCD1, HK2, DDX21	1.42	1.00	0.99	99.47
		GO:0000166~nucleotide binding	7	0.41	GALK1, UBE2Z, PYGL, RTCD1, HK2, DDX21, HADHA	1.29	1.00	0.99	99.78

Table C-6 Clustered enrichment terms under molecular function domain for HCE cell line results.

Domain	Cluster/ES	Term	Count	PValue	Gene symbol	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_MF_FAT	Cluster 1 - Enrichment Score: 1.42	GO:0070279~vitamin B6 binding	3	0.02	GOT2, SHMT2, OAT	14.41	0.72	0.47	15.34
		GO:0030170~pyridoxal phosphate binding	3	0.02	GOT2, SHMT2, OAT	14.41	0.72	0.47	15.34
		GO:0019842~vitamin binding	3	0.04	GOT2, SHMT2, OAT	8.65	0.97	0.57	35.97
		GO:0048037~cofactor binding	3	0.22	GOT2, SHMT2, OAT	3.23	1.00	0.94	92.82
	Cluster 2 - Enrichment Score: 0.45	GO:0051087~chaperone binding	3	0.01	GRPEL1, HSPE1, HSPD1	24.02	0.36	0.36	5.78
		GO:0051082~unfolded protein binding	3	0.12	GRPEL1, HSPE1, HSPD1	4.80	1.00	0.86	73.09
		GO:0017076~purine nucleotide binding	6	0.38	GRPEL1, RAC2, HSPE1, DRG1, HSPD1, MYH9	1.40	1.00	0.99	99.40
		GO:0032553~ribonucleotide binding	5	0.58	RAC2, HSPE1, DRG1, HSPD1, MYH9	1.20	1.00	1.00	99.99
		GO:0032555~purine ribonucleotide binding	5	0.58	RAC2, HSPE1, DRG1, HSPD1, MYH9	1.20	1.00	1.00	99.99
		GO:0000166~nucleotide binding	6	0.66	GRPEL1, RAC2, HSPE1, DRG1, HSPD1, MYH9	1.06	1.00	1.00	100.00
		GO:0030554~adenyl nucleotide binding	4	0.67	GRPEL1, HSPE1, HSPD1, MYH9	1.15	1.00	1.00	100.00
		GO:0001883~purine nucleoside binding	4	0.68	GRPEL1, HSPE1, HSPD1, MYH9	1.13	1.00	1.00	100.00
		GO:0001882~nucleoside binding	4	0.69	GRPEL1, HSPE1, HSPD1, MYH9	1.12	1.00	1.00	100.00
		GO:0005524~ATP binding	3	0.85	HSPE1, HSPD1, MYH9	0.91	1.00	1.00	100.00
		GO:0032559~adenyl ribonucleotide binding	3	0.85	HSPE1, HSPD1, MYH9	0.90	1.00	1.00	100.00
	Cluster 3 - Enrichment Score: 0.11	GO:0046872~metal ion binding	4	0.76	MYL6, PRIM1, UQCRC1, MYL12A	1.01	1.00	1.00	100.00
		GO:0043169~cation binding	4	0.77	MYL6, PRIM1, UQCRC1, MYL12A	0.99	1.00	1.00	100.00
		GO:0043167~ion binding	4	0.78	MYL6, PRIM1, UQCRC1, MYL12A	0.98	1.00	1.00	100.00