Accelerated cell line development and improved characterisation of lentiviral vector production through application of MALDI-ToF mass spectrometry and multivariate data analysis.

Volume 1 of 1

By Jakub Krakowiak

Submitted for Engineering Doctorate (EngD)

Research conducted at the School of Chemical Engineering and Advanced Materials at Newcastle University

Submitted 1st October 2017

Abstract

Several cell and gene therapies will be commercially launched within the next few years using lentiviral vectors as the gene delivery vehicle. Oxford BioMedica's Lentivector® platform is an advanced lentiviral-based gene delivery system designed for improved safety and efficacy. The growing interest in these vectors has created a strong demand for large scale production of lentiviral vectors as well as for development of packaging and producer cell lines. This EngD project used a combination of matrix assisted laser desorption ionisation time of flight mass spectrometry (MALDI-ToF MS) and multivariate data analysis (MVDA) to analyse cell and lentiviral vector samples. A comparison between mass spectra of samples produced across small and large scale in adherent and suspension culture was used to identify what aspects of the manufacturing process had the biggest impact on cell and vector variation. Principal component analysis was applied to compare different lentiviral vector production methods, assess data structure of the process parameters and examine whole cell and vector mass spectrometry data. This approach led to improved characterisation of lentiviral vectors and HEK293T cells. It demonstrated the capability to differentiate between adherent and suspension cells as well as cell lines of different levels of performance as defined by lentiviral vector infectious titre. Partial least squares discriminant analysis (PLS-DA) was used to calibrate and validate a predictive model of cell line performance based on mass spectrometry and viral vector titre data obtained from multiple HEK293T cell lines. PLS-DA model validation resulted in 87.5% accuracy in classification of cell lines as high or low producers based on a discrimination threshold determined by viral vector titre. The results of PLS-DA modelling indicated that this method can be used for accurate cell line performance prediction, accelerating cell line development by several weeks, improving cell selection and reducing campaign timelines.

Acknowledgements

This EngD thesis is a result of collaboration between Newcastle University, the University of Kent and Oxford BioMedica funded by EPSRC (grant number EP/G037620/1).

I would like to give my warm and sincere thanks to all the people who have helped me along the way and contributed to the successful completion of this EngD project:

My academic and industrial supervisors Dr Chris O'Malley (Newcastle University), Prof Gary Montague (Teesside University), Prof Mark Smales (University of Kent) and Peter Jones (Head of Manufacturing Development, Oxford BioMedica) for committing a large amount of their time towards guidance and support throughout all stages of the project;

Dr Carol Knevelman (Head of Process R&D, Oxford BioMedica) for providing technical advice and support as a co-supervisor at Oxford BioMedica;

Dr Jane Povey (University of Kent) for the MALDI-ToF mass spectrometry training, organising and performing mass spectrometry analysis and for advice throughout the project;

I would also like to thank the team within Oxford BioMedica, especially in Process R&D, Manufacturing, Sciences and Technology and the Cell Engineering Group for the training and technical assistance in the laboratories throughout the project and provision of materials, including the parental cell lines used for cell line development in the project;

I would also like to thank my EngD colleagues especially Kirstie Pemberton for sharing her experience at Oxford BioMedica and helping me generate some early-stage materials that were used in the initial MALDI-ToF viability study; all of my peers in cohort 5 for the scientific and technical discussions and for sharing this long EngD journey;

My girlfriend and colleague Reka Nagy for her emotional support, sharing the difficult experience of pursuing a postgraduate degree and also for tutoring me in R scripting;

Finally, all members of my family who always supported my decision to pursue higher education overseas; I would especially like to honour the memory of my late grandfather, Mieczysław Krakowiak (aged 75) who passed away before I could finish my degree. He always wanted me to become a doctor (a medical doctor but I think he would settle for an engineering one) and I regret I couldn't fulfil his wish earlier.

Table of Contents

List of tables

List of figures

Nomenclature

α-cyano – α-cyano-4-hydroxycinnamic acid
AAV – adeno-associated virus
ACSS - automated cell screening system
ADA-SCID – adenosine deaminase deficiency severe combined immunodeficiency
AEC – anion exchange chromatography
ATMP – advanced therapy medicinal products
AV – adenovirus
BMR(s) - batch manufacturing record(s)
CaPO – calcium phosphate
CF (2/10) – cell factory (2-stack/10-stack)
CLC – cholesterol lipid concentrate
CMV – cytomegalovirus
CV – coefficient of variation
(k)Da – (kilo)Dalton
DHB – 2,5-dihydroxybenzoic acid
DMEM – Dulbecco's modified eagle medium
DNA – deoxyribonucleic acid
ds – double stranded
EIAV – equine infectious anaemia virus
EngD – engineering doctorate
FACS – flow assisted cell sorting
FBS – foetal bovine serum
GFP – green fluorescent protein
GMP – good manufacturing practice
HEK293(T/FE) – human embryonic kidney 293 (T/FE)
HIV – human immunodeficiency virus
(HP)LC – (high-performance) liquid chromatography
HSV – herpes simplex virus
IMAC – immobilised metal ion affinity chromatography
iPSC(s) – induced pluripotent stem cell(s)
LTR – long terminal repeat
LV(s) – latent variable(s)
LVV(s) – lentiviral vector(s)
MALDI-ToF (MS) – matrix assisted laser desorption ionisation time of flight (mass spectrometry)
MLV – murine leukaemia virus
MVDA – multivariate data analysis
m/z – mass to charge ratio
NaBu – sodium butyrate
OXB – Oxford BioMedica
PaCL - packaging cell line
PC – principal component
PCA – principal component analysis
PCR – polymerase chain reaction
PrCL - producer cell line
P/V – power per volume
QbD – quality by design
RNA – ribonucleic acid
RCL – replication competent lentivirus

RRE – rev regulatory element
(γ-)RV – (γ-)retrovirus
SA – sinapinic acid
SEC – size exclusion chromatography
SG – Savitzky-Golay
SIN – self inactivating
SOP – standard operating procedure
ss – single stranded
STR – stirred tank reactor
SV40 – simian virus 40
TFA – trifluoroacetic acid
TSSM – Tri, sodium chloride, sucrose and mannitol (buffer)
TU/ml – transducing units per millilitre
VSV-G – vesicular stomatitis virus G
X-SCID – X-linked severe combined immunodeficiency

Introduction

This Engineering Doctorate (EngD) research project is closely tied to biopharmaceutical industry through collaboration with Oxford BioMedica, a world leading gene and cell therapy company involved in process development and manufacturing of lentiviral vectors (LVVs). Gene therapy is an emerging field of medicine which was initially met with concerns over safety and ethics but through the development of advanced methods of gene delivery has gathered increased support and interest of investors. Gene therapy offers unique benefits of delivering long term treatment and addressing the unmet needs of patients in areas not covered by currently available treatment options (Naldini, 2015). Viral vectors are one of the major platforms used in gene therapy with multiple applications and different systems available for a variety of therapeutical needs (Keeler et al., 2017). LVVs can be used for long-term transduction of non-dividing cells which positions them as a versatile platform with therapeutic potential in multiple *in vivo* and *ex vivo* applications (Escors & Breckpot, 2010). Oxford BioMedica specialises in LVV development and manufacturing and this EngD project focuses on characterisation and optimisation of the LVV production process.

Overview

The novelty of the gene therapy field creates a need for the development of efficient and robust production processes. LVV production is a demanding process with several unique challenges which are not present in other areas of biopharmaceutical production. Manufacturing of viral vectors is covered in an extensive review of upstream (Merten et al., 2014a) and downstream processing (Merten et al., 2014b). The authors also specifically cover production of LVVs and its specific challenges (Merten et al., 2016). Increased demand for the development of new products and transitioning them through clinical trials and into commercial stage drives the growing need for industrialisation of the current laboratory scale processes to enable market supply. This leads to a need for improved process understanding and establishment of high-throughput process development methods as well as capability to scale up manufacturing processes to meet the demands of the market. These needs are addressed through transition from adherent cell culture to serum-free suspension cell culture which offers significant benefits in the area of process scalability, control and consistency (van der Loo & Wright, 2016). This transition not only extends the potential capacity for high volume production of LVVs but also provides additional data about the process through advanced process analytical technologies. Another major trend is the push for the development of stably transfected packaging and producer cell lines in an effort to reduce

cost of goods supplied and improve process consistency (Broussau et al., 2008; Kafri et al., 1999; Stewart et al., 2009).

The significance of the improved process control and monitoring becomes more prominent in the context of continuous process improvement as described in the quality by design (QbD) industry guideline Q8 R2 (ICH, 2009). The trend of data-driven process development showcases the principle of quality built into the process while advanced monitoring and control ensures that the process remains within the design space to ensure consistent product quality, safety and efficacy. The abundance of process data requires an approach to analysis capable of extracting the underlying information which can be used to inform process development decisions and ensure optimal production. The multivariate character of the data warrants the use of advanced analytical methods characteristic for the area of chemometrics. Multivariate data analysis (MVDA) allows examination of large data sets characterised by multiple variables as is the case in LVV production process monitoring where multiple process parameters are recorded throughout the process. Reducing the dimensionality of data and extracting the information about the most important parts of the process can help drive improvements in the process.

Methodology

The major MVDA method used in this project is principal component analysis (PCA). It is a method of transforming a data matrix into linearly uncorrelated variables - principal components (PCs). The PCs capture the highest possible amount of data variance while each subsequent PC remains orthogonal to the previous ones. It is a well-established technique originating in the early 20th century (Hotelling, 1933; Pearson, 1901). It is capable of reducing data dimensionality and providing a platform for graphical representation of complex data and it has been applied in multiple areas of academia and industry with several more recent articles summarising the underlying method and its applications (Abdi & Williams, 2010; Bro et al., 2014; S. Wold et al., 1987). Throughout this EngD project, PCA has been used for exploratory data analysis of process parameters and spectrometry data from cell and viral vector samples and is described in detail in Chapter 2 and Chapter 3.

Mass spectrometry (MS) is being increasingly used in analysis of biological processes and it can be used to address the need for improved understanding of LVV production. The general principle of this technique is based on ionisation of biological samples and estimation of their mass to charge ratio based on their behaviour in an electric field (Glish & Vachet,

2003). This allows identification of peptides, proteins and complex entities such as cells. It is a powerful technique potentially capable of accurate identification of individual molecules as well as mixtures where the entire mass spectrum of multiple molecules provides a characteristic fingerprint that can be used as a basis for further analysis. This approach opens up the possibility of examining cells and their structures and therefore can provide more insight into the LVV production process.

The main MS method used in this EngD project was matrix assisted laser desorption-ionisation time of flight mass spectrometry (MALDI-ToF MS) which allows ionisation of large biological molecules. This powerful technique was a significant advancement in the MS field, recognised by a shared Nobel award for its inventor (Tanaka et al., 1988). Karas & Krüger (2003) provide a description of the ionisation method while further studies describe sample preparation (Veloo et al., 2014). MALDI-ToF MS is routinely used in analysis of biological samples (Caprioli et al., 1997), bacteriology (Seng et al., 2009) and in clinical studies (Clark et al., 2013). In this project it has been applied to the analysis of HEK293T cells and LVV samples generated at Oxford BioMedica as part of the project and analysed with help of Dr Jane Povey at the University of Kent.

A major challenge in handling the MS data obtained for cell and vector samples is the complexity and structure of the data. These samples are composed of multiple proteins, the resulting mass spectrum is characterised by thousands of variables, each representing signal intensity at certain mass to charge ratio. Interpreting this data can be challenging but this problem is addressed by applying MVDA techniques to the mass spectra to simplify the analysis and extract the important information about cell and vector samples. This EngD project aims to combine the MS methodology with MVDA to improve process understanding of LVV production and develop methods for improved process characterisation. This approach is extended by developing a predictive model that can be used in cell line development to debottleneck the selection of a high LVV producer cell line. Modelling has been increasingly important for advanced process development and the use of MVDA combined with MS presents an attractive avenue for such an approach.

Using MS data presents a range of challenges. Sample preparation, instrument calibration and signal generation can all introduce variation and noise in the data. To reduce the impact of this variability the data can be processed through mathematical algorithms which aim to reduce the noise and improve visibility of important trends. Typically, the data

is corrected for signal background, normalised and smoothened before final analysis. Details of pre-processing applied in the context of this project are described in Chapter 3.

Modelling of the MS data was performed using partial least squares discriminant analysis (PLS-DA). It is a variant of partial least squares (PLS) regression, originally used in econometrics (Wold, 1966) and then widely applied in chemometrics (Wold et al., 2001). The method is based on transformation of two matrices: X matrix of predictors and Y matrix of responses. The model projects the data into latent variables (LVs) in a way that maximises covariance between matrices, therefore creating a model that explains the variation in response data in relation to the variation in predictor data. This mathematical method can be used to generate predictive models and is particularly suitable for data with multiple correlated variables. Through this method it is possible to predict the response values of unknown data based on input of predictor values. PLS-DA is a variant of this method using categorical rather than continuous data which provides advantages in classification of data into distinct populations based on cateogircal properties or threshold of critical parameters (Barker & Rayens, 2003). This method, combined with MALDI-ToF MS has been used in classification of CHO cells based on their performance in monoclonal antibody production (Povey et al., 2014) and shows potential to be applied in LVV production and cell line development. PLS and PLS-DA methodology and application in this project is described in detail in Chapter 4.

Development of a working model should be an iterative process where data quality forms the basis of the model development. Acquisition of experimental data, model calibration and validation can be a balancing act limited by time and resources. To ensure model quality the data should be processed to ensure it is applicable to the model and consistent. Selection of data sets, outlier detection and application of modelling method suitable for the features and limits of the data are all important for model development. Finally, model validation is an essential part of development where the performance is tested within the initial calibration data set (cross-validation) and independent data set obtained after model development (validation). Model development and validation in this project is described in detail in Chapter 4.

Current gene therapy challenges

When reviewing the literature, it becomes apparent that while LVV production shares many common traits with other biopharmaceutical manufacturing processes, it also faces

unique challenges. There is an unmet need for improved characterisation at the molecular level as well as process optimisation through the use of advanced process analytical technologies and data-driven analysis. MS and MVDA methods were applied to mammalian cell based production, especially in the lucrative area of monoclonal antibody production in CHO cells (Povey et al., 2014; Schwamb & Wiedemann, 2015). However, viral vector production has received little attention in this area of research. Cell transfection, production of viral proteins and viral vector assembly all have profound effects on the host cell. This presents both a challenge in process characterisation as well as an opportunity to explore the application of MS and MVDA in this novel field.

MS has been used in the research of mammalian cells and viruses but its application in process characterisation has been limited. In particular, the analysis of viral vectors and the effect of scale, production methodology and processing are lacking, presenting an opportunity for improved characterisation of viral vectors. Subsequently, the trend to shift LVV production modality from adherent to serum-free suspension cell culture provides a unique opportunity for characterisation of the cells and viral vectors generated from these two different production methods. MS is a robust technique which has been demonstrated to be effective in protein characterisation and identification. In the case of this project, whole cell and vector analysis can be applied to multiple aspects of process characterisation and offers an unique opportunity for exploring the approach covering the entire mass spectrum rather than focusing on individual proteins. Combined with MVDA techniques it can provide a wealth of data applicable to improved process development and product characterisation.

Currently the development of stably transfected packaging and producer cell lines for LVV production has been a major focus of the industry but remains challenging and time consuming (Merten et al., 2016). At the same time few attempts have been made to implement statistical modelling as part of the development process. The PLS-DA based method was demonstrated to be effective in cell line development for antibody production (Povey et al., 2014) and could be applied to packaging cell line development with significant benefits. It is an opportunity for exploring a new application for a technique traditionally used in chemometrics for transforming complex measurements and multi-instrument readings into data capable of prediction and classification of process outputs. Viral vector production could greatly benefit from advanced analytical methods both in the area of process control and viral vector quality control.

EngD objectives

The justification for this EngD project stems from the need for debottlenecking of the cell line development process and for better understanding and characterisation of LVV production. Gene therapy is a rapidly developing field of medicine and it could greatly benefit from application of advanced techniques which have been demonstrated to benefit other areas of biopharmaceutical production. There has been a significant interest in improving the LVV production process and commercialisation of gene therapy products but there is still room for improvement in the area of understanding the underlying reactions. Exploration of cells and viral vectors through a combination of MS and other analytical techniques is an opportunity to bridge the gap in knowledge in this area while providing practical benefits for process development and LVV manufacturing through optimised process control and development of modelling tools for improved decision making. The application of MS and MVDA in the novel field of gene therapy expands the current knowledge of these techniques and provides tangible benefits to the emerging gene therapy industry. Following the ICH guidelines for QbD (ICH, 2009), techniques used in this project demonstrate feasibility of improving process understanding and consistency through analysis of batch and product data.

The need for improved understanding of LVV production through cell and vector characterisation and implementation of MS and MVDA into the process leads to the research questions forming the basis of this EngD project. Are HEK293T cells and LVVs suitable for MS analysis and what information can be obtained from comparison of different samples? Establishing an optimal method of cell and vector analysis will be necessary and the methodology will need to be well characterised to demonstrate its reliability and robustness. The study aims to establish which traits of cell and vector samples can be characterised using the combination of MS and MVDA as well as how that knowledge can be applied to improve LVV production. Simultaneously, MVDA will be investigated as a way to improve process monitoring and control through analysis of critical process parameters.

Additionally this projects aims to determine whether the method established by Povey et al., (2014) can be applied in a new setting of development of packaging and producer cell lines for LVV production. Manufacturing of LVVs is significantly different than the process used for monoclonal antibodies. The analytical methods for estimation of cell productivity are also different and can lead to new challenges. How accurate is the PLS-DA based method when applied to different type of cells and is it able to reliably predict cell line performance?

Moreover, can this methodology be transitioned to industrial application and assist future cell line development at Oxford BioMedica? There is a significant value in addressing these questions to improve the understanding of LVV production process and to establish novel analytical and modelling methods to facilitate progress of manufacturing in the novel field of gene therapy.

The cell line development (described in general terms in Chapter 1 and in context of the project in Chapter 4) is a time consuming process which requires generation of a large amount of initial clones which are progressively characterised and the pool is narrowed down to select the best producers. The classical approach to cell line development (Figure 1) relies on manual generation of clones through limited dilution cloning in antibiotic media followed by characterisation of individual clones at increasingly larger scale of production.

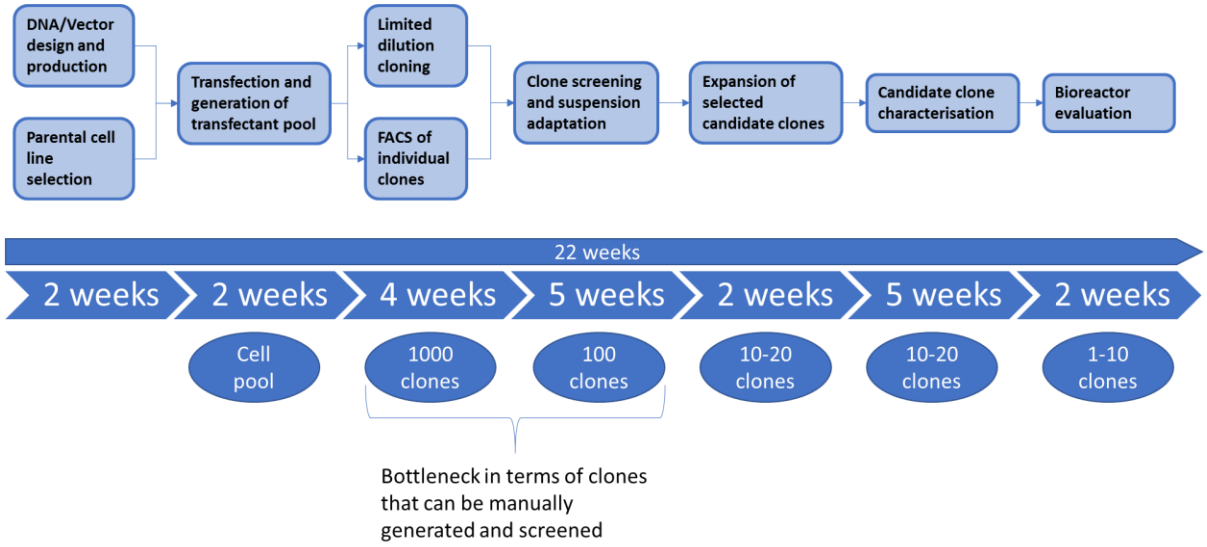


Figure 1: Diagram and timelines of manual cell line development process

Oxford BioMedica has developed an automated cell screening system (ACSS) which addresses one of the limitation of the manual process by automating the clone generation process and allowing generation of thousands of clones (Figure 2). However, the limitation in terms of cell lines that can be characterised is still a bottleneck further along the process.

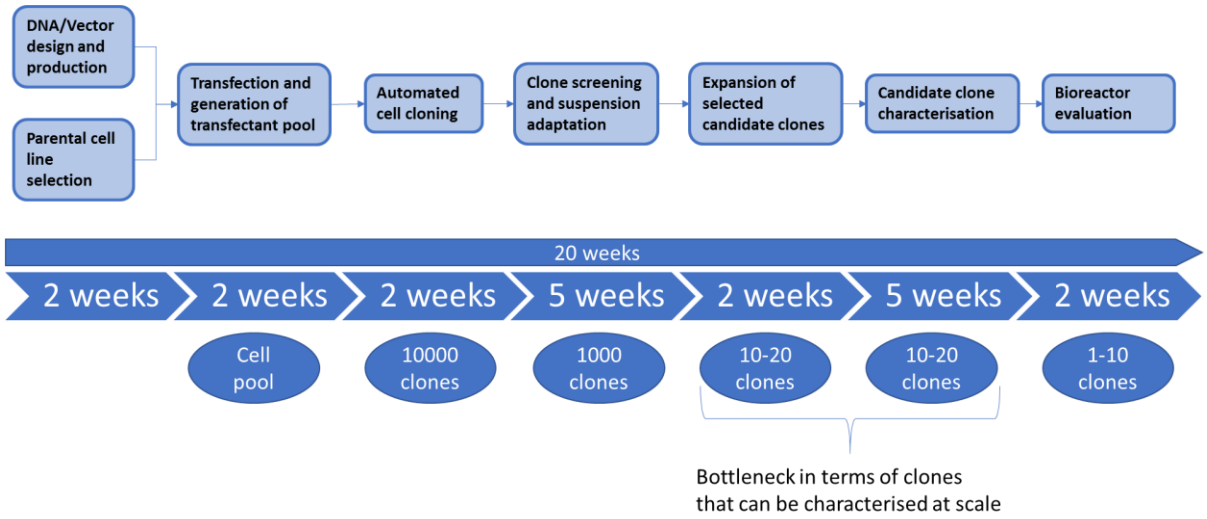


Figure 2: Diagram and timelines of Oxford BioMedica's automated cell line development process

The PLS-DA model aims to address the limitation of cell line development and increase the number of clones which can be successfully characterised and categorised as low or high producer at an early stage of development process. (Figure 3). By using the predictive model described in detail in Chapter 4 the cell line development process can generate a higher number of clones which can be characterised earlier and with improved accuracy which reduced the overall timeline.

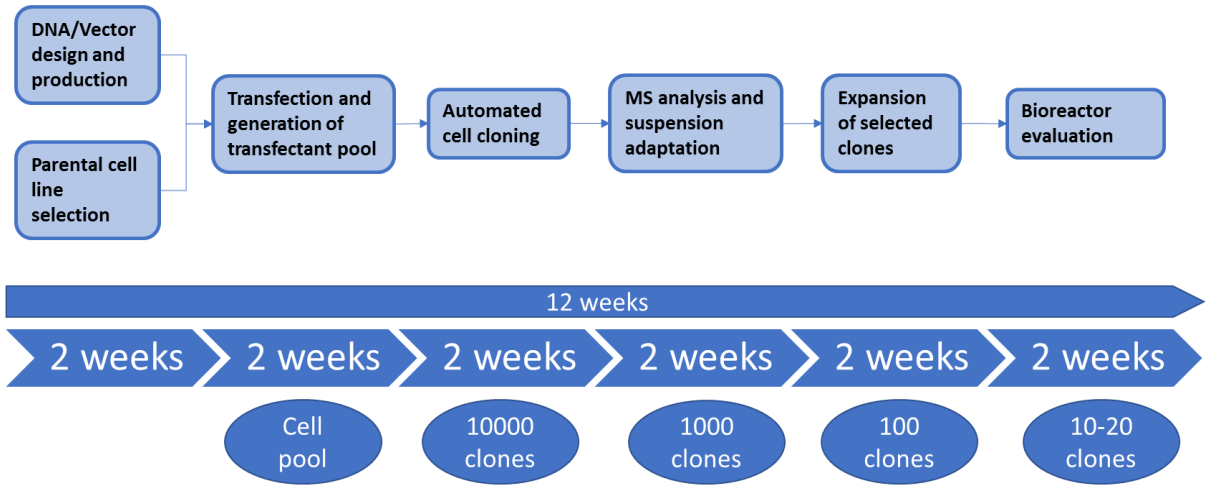


Figure 3: Diagram and timelines of automated PLS-DA assisted cell line development process.

Thesis outline

This EngD thesis is organised into four chapters, each outlining different part of the research project conducted at Oxford BioMedica. The first chapter provides a background of gene therapy and viral vector production in a form of literature review. It covers the history

of this novel field of medicine, different types of vectors used in the field, advancements in the development of LVVs and the details of different approaches to upstream and downstream processing of LVVs. The second chapter covers process characterisation through PCA of batch manufacturing records for production of LVVs in adherent cell cultures. Additionally, the analysis of laboratory-scale batches for the suspension cell culture process is discussed to provide a contrast between the two different modalities and examine the applicability of MVDA in process monitoring and characterisation. The third chapter describes the development of MS methodology for use with HEK293T cells and LVVs as well as results of analysis of multiple cell and vector samples. Different approaches to sample generation and processing are discussed along with pre-processing of the MS signal. Results of PCA are presented for cell and vector samples produced at different scales and using a variety of production methods. The final chapter covers the development and implementation of the predictive model of cell line performance intended for use in packaging cell line development. The process of cell line development is described along with characterisation of cell lines used in development and validation of the model. Refinements and optimisation of the methods are discussed for multiple versions of the model. The results of model calibration and validation are discussed highlighting the benefits and potential challenges in implementation of the method within an industrial setting.

Chapter 1 Lentiviral vector production literature background

1.1. Introduction

Advanced therapy medicinal products (ATMP) comprise a group of promising therapeutics which have gained interest and increasing investment in the pharmaceutical sector. They offer new approaches to currently unmet therapeutic needs as well as alternatives to currently available treatments. Gene therapy is a prominent subgroup of ATMP based on the principle of modifying patient genetic material. Development and manufacturing of gene therapy medicines is a major focus of Oxford BioMedica (OXB) who specialise in lentiviral vector (LVV) technology. Lentiviruses are capable of integrative transduction of non-dividing cells, ensuring a long-term expression of therapeutic genes, which is highly advantageous for some of the gene therapy applications. Currently the production process remains the biggest challenge, especially in terms of the process scale-up and maintaining the product consistency and safety. This engineering doctorate (EngD) project, sponsored by OXB, aims to improve process understanding and timelines across the various stages of LVV production through application of MVDA, MS and predictive modelling. This objective aligns with the Quality by Design approach which emphasises process understanding and continuous improvement. Ultimately this leads to better control and efficiency of the process which benefits both the manufacturer and the patient through improved quality and consistency of the product. This chapter will outline the literature background of viral vector-based gene therapy, provide an overview of LVV production, and discuss the approach and methodologies applied throughout the project.

1.2. History and advancement of gene therapy

Gene therapy is one of the major groups of ATMP and it provides promising and innovative solutions to unmet needs in medicine. To fully understand the current methods used in development and manufacturing of viral vectors, it is important to review the past advances in the field which became the foundation of the contemporary platforms. This work highlights the development of gene therapy from its early days to the recent regulatory approval of viral vector-based therapies.

1.2.1. Early days of gene therapy

Interest in human gene therapy was initially motivated by the opportunity to introduce new approaches to deal with severe genetic disorders. Scientists were examining the therapeutic potential of viral vectors in early 1970s, when Rogers *et al*. (1973) conducted the first trial of viral gene therapy in humans. The study aim was to treat urea cycle disorder using wild-type *Shope papilloma* virus but it was unsuccessful. The first gene therapy study using recombinant DNA in humans was performed by Martin Cline in 1980 as reviewed by the author of the study (Cline, 1985). He introduced a β-globulin-encoding gene along with herpes simplex virus (HSV) thymidine kinase gene as a marker to the bone marrow stem cells extracted from two patients. The modified cells were then re-introduced to patients following irradiation of the native bone marrow. The treatment proved ineffective (Cline, 1985) and was performed while the permission from University of California Human Subject Protection Committee was still under review. Eventually, the permission was not granted. Cline's study was deemed premature and ethics behind it were questioned (Beutler, 2001).

Despite these initial setbacks, increasing advancements throughout 1970s and 1980s in basic science especially molecular genetics pushed gene therapy towards therapeutic use at a rapid pace (Selkirk, 2004). The first gene therapy clinical trial was performed in 1988 where two patients with adenosine deaminase deficiency severe combined immunodeficiency (ADA-SCID), a variant of immunodeficiency disorder caused by an autosomal recessive dysfunction of adenosine deaminase gene, were treated. They had their white blood cells extracted and genetically modified to express the correct variant of adenosine deaminase gene. The cells were then introduced back to the patients. In theory this should rescue the disease phenotype by producing the appropriate protein. However, only a transient response was observed in one of the patients while the other did not respond to the treatment at all (Blaese et al., 1995).

1.2.2. Gene therapy advancement and improvement

Further trials followed with more examples of ADA-SCID treatment (Bordignon et al., 1995) and the first effective in vivo gene transfer (Puumalainen et al., 1998). The advancement of gene therapy was set back in 1999 when an 18 year old ornithine transcarbamylase deficient patient suffered a fatal multiple organ failure as a result of a severe immune response to adenoviral gene therapy (Stolberg, 1999). This incident elucidated the risks associated with gene therapy and the requirement for a better understanding of the viral vectors used in gene therapy. Despite the setback, further studies were conducted leading to more X-linked severe combined immunodeficiency (X-SCID) and ADA-SCID trials using γ-retrovirus (γ-RV) that resulted in good efficacy of the treatment. However, follow-up examination revealed that some of the patients developed leukaemia which was associated with insertional mutagenesis caused by the viral vector (Fischer et al., 2005). This demonstrated that while there is a clear benefit to the therapy, it requires improved techniques and understanding. Later trials using RVs for ADA-SCID treatment (Aiuti et al., 2009) and X-SCID (Hacein-Bey-Abina et al., 2010) resulted in therapeutic effect without development of leukaemia or other adverse side effects, therefore giving hope for safe and successful treatment

In 2003 the first gene therapy product was approved for clinical use in China. Genidicine, an adenoviral vector encoding human p53 gene was used for treatment of head and neck squamous carcinoma (Peng, 2005; Wilson, 2005). A year later EU granted a GMP certificate to Ark Therapeutics for manufacturing of Cerepro, an adenoviral gene therapy vector encoding HSV thymidine kinase for treatment of malignant brain tumours. In 2008 phase 3 clinical trial was completed for Cerepro (Wirth et al., 2009). The first viral vector based gene therapy medicine approved in Europe was Glybera, an adeno-associated virus (AAV)-based *in vivo* gene therapy for the treatment of lipoprotein lipase deficiency (Scott, 2015). The latest approved viral vector therapeutic include Strimvelis, a γ-RV-based *ex vivo* gene-modified cell therapy for ADA-SCID (CHMP, 2016) and Kymriah, a LV-modified T-cell treatment for B-cell acute lymphoblastic leukaemia (CHMP, 2018). The number of gene therapy products going through clinical trials is increasing and the available products are constantly being improved, resulting in enhanced efficacy and safety. This includes several products developed by or in collaboration with OXB.

1.3. Gene therapy viral vectors overview

Viral vectors are one of the major platforms used in gene therapy. There are currently multiple types of viruses used as vectors, each with distinct advantages and challenges (Table 1 for summary). While OXB specialises in LVV-based therapies, it is important to understand the properties of other similar viral vectors.

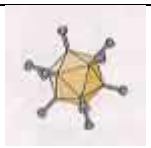
	AV	AAV	SV40	HSV	γ-RV	LV
Family	*Adenoviridae*	*Parvoviridae*	*Polyomaviridae*	*Herpesviridae*	*Retroviridae*	*Retroviridae*
Image						
Genome	dsDNA linear	ssDNA linear	dsDNA circular	dsDNA linear	ssRNA	ssRNA
Coating	naked	naked	naked	enveloped	enveloped	enveloped
Major capsid proteins	Hexon, penton, proteins IIIa, VI, VIII, IX	VP1, VP2, VP3	VP1, VP2, VP3	VP5, VP19C, CP23	Group-specific antigen proteins (gag)	HIV: p24 CA, SP1, SP2
Diameter	70-90 nm	18-26 nm	45-55 nm	150-200 nm	80-130 nm	80-130 nm
Genome size	36-38 kb	5 kb	5-8 kb	120-200 kb	3.5-9 kb	3.5-9 kb
Packaging capacity	7.5-7.9kb	4.5kb	5 kb	>30 kb	8 kb	8kb
Infectivity	Non-dividing /dividing cells	Non-dividing /dividing cells	Non-dividing /dividing cells	Non-dividing /dividing cells	Dividing cells	Non-dividing /dividing cells
Integrative	NO	YES/NO	YES	YES/NO	YES	YES
Gene expression	Transient	Transient/ Long term	Long term	Limited long term	Long term	Long term
Immune response	High	Low	Low	High	Low	Low
Main advantage	Efficient and easy to produce	Safe, broad infectivity range	Stable, effective transduction	High packaging capacity	Long-term gene expression	Long-term gene expression
Main limitation	Highly immunogenic	Low packaging capacity	Low packaging capacity	Difficult application, immunogenic	Insertional mutagenesis, transduces only dividing cells	Difficult to obtain high titres

Table 1: Properties, advantages and limitations of the viral vectors commonly used in cell and gene therapy. References: Deyle & Russell, 2009; Fink, & Glorioso, 2007; Goverdhana et al., 2005; Jacobs, Breakefield, & Fraefel, 1999; Strayer, 2000; Volpers & Kochanek, 2004. Images adapted from http://www.nlv.ch/Virologytutorials/Classification.htm

1.3.1. Adenoviruses

Adenovirus (AV) is one of the first virus classes that were adopted as a vector for gene therapy. It is still commonly used in certain types of gene therapy and has become a well characterised and advanced method of gene delivery. AVs are double stranded DNA

viruses with linear DNA at their core enclosed in an icosahedral capsid. Their linear DNA can be modified using currently available molecular genetics techniques. As a result of the extensive toolset available for DNA modification AVs are easier to work with than RNA-based viruses and are therefore especially suited for applications requiring extensively modified recombinant DNA.

AV is a non-integrating virus, resulting in transient infection that is generally safer compared to genome-integrating vectors but limited to short-term therapy. The main advantages of AVs are their good characterisation and wide range of infectivity profile. Extensive knowledge of the virus' mode of action, metabolism and toxicity profile resulted in development of efficient and safe vector that can be produced at high titres (Volpers & Kochanek, 2004).

The main issue of AVs that must be considered in therapeutic use is their immunogenicity profile. AVs are usually met with quick immune response from the patient, resulting in quick clearing of the vector from the body. In case of high dose injection, AV vector can lead to severe immune response that would be dangerous for the patient and could even result in death (Stolberg, 1999). This issue can limit some of AVs applications as a systemic injection is unlikely to be safe or effective. A possible approach to solve this issue could involve the use of immunosuppressants but such radical measure bears risks and problems of its own. Instead, AV vectors are best used for localised short term therapy. They are easy to modify and efficient and terefore suitable for use in cancer therapy, where a localised injection resulting in short-term therapy can reduce the tumour size (Rein et al., 2006).

1.3.2. Adeno-associated virus

Another class of commonly used DNA viruses is AAV. It is a class of small non-pathogenic single stranded DNA viruses with an icosahedral capsid. Unlike AVs, AVVs are characterised by low immunogenicity, making them safe for human systemic applications. They are capable of infecting non-dividing cells, making them capable of delivering their genetic payload to multiple types of tissues. These properties make it an attractive vector for use in gene therapy as it is considered safe, effective and flexible. AAV is a genome-integrating virus that preferentially integrates into a defined locus of chromosome 19 where it forms a provirus (Deyle & Russell, 2009). However, during the development of AAV as a gene therapy vector this functionality has been removed by deleting the *rep* gene needed for

chromosomal integration. Viruses with *rep* deletion persist in the cell with low rate of random integration that is regarded as safe for the patient (Deyle & Russell, 2009). The property of AAV to persist in the cell results in long term expression of the therapeutic gene, making this type of vector suitable for use in treatment of wide variety of diseases. It can be especially effective to treat single gene hereditary diseases such as cystic fibrosis (Moss et al., 2004). AAV vectors have been tested with multiple methods of delivery, including local and systemic injections as well as aerosol inhalers for treatment of respiratory tract disorders. The main disadvantage of AAVs is their low size gene packaging capacity. The vector is able to deliver only genes up to 4.8kb long which limits some of the potential applications as many proteins targeted by gene therapy exceed that size limit. Nevertheless, AAV vectors are regarded as a safe and effective gene therapy delivery method for long-term expression of small size genes.

1.3.3. Simian virus 40

An alternative vector for long-term therapy is simian virus 40 (SV40). It is a circular dsDNA virus with an icosahedral non-enveloped capsid. The vector is produced using packaging cell lines that contain essential structural genes encoding capsid proteins and the T-antigen. The cells are transfected with a minimal plasmid containing the transgene and a promoter. The cells are then capable of producing non-replicative gene therapy vectors. The vectors prepared in this way are capable of high-efficiency transduction in non-dividing and dividing cells resulting in integrative long-term expression of the transgene, making SV40 an attractive platform for multiple applications (Strayer, 2000).

SV40 was found to lack immunogenicity and result in stable long-term expression of the transgene. However, it is limited by its packaging capacity; similar to AAV it can only contain genes of limited size. Moreover, the vector results in integration of DNA into host genome which may result in insertional mutagenesis, an event where gene integration activates host genes that promote oncogenesis. Moreover, wild-type SV40 is capable of inducing tumours through suppression of the *p53* gene which poses an additional risk when used in integrative gene therapy. These issues suggest that while effective, SV40 has certain risks associated with its use as a gene therapy vector and therefore has not become popular.

1.3.4. Herpes simplex virus

The last major group of DNA viruses to discuss are HSVs. They are dsDNA viruses with an enveloped icosahedral capsid. They are characterised by a good transduction rate and a high affinity for neurons. The wild-type virus can alternate between lytic and latent stage

which presents both an opportunity and a challenge in vector design. There are several approaches of accommodating the virus for the gene therapy application. One of them is production of amplicons which consist of the transgene, HSV promoter and a packaging signal sequence which is used in producing the vector in a packaging cell line expressing structural viral genes. This results in production of a non-replicative minimal viral vector capable of long-term non-integrative expression specific for neuron cells (Jacobs et al., 1999), however the promoter can be substituted to allow gene expression in other cell types.

An alternative approach involves production of non-replicative virus capable of inducing latent-like state upon infection in neurons and other cell types (Fink & Glorioso, 2007). However, in this case gene expression is problematic as most promoters get silenced. This issue creates the need for extensive vector engineering to achieve gene expression in peripheral or central nervous system, which limits the use of HSV. Additionally, HSV can be toxic for cells and therefore multiple deletions are required to ensure its safety for the cell and persistence of the therapeutic effect. As such HSV presents a unique opportunity for gene therapy but remains problematic in application.

1.3.5. γ-retroviruses

RVs are ssRNA enveloped viruses capable of replication thanks to the process of reverse transcription of RNA to DNA. The major group of RVs are γ-RVs represented by viruses such as murine leukaemia virus (MLV). γ-RV genome comprises of the core gene groups *gag*, encoding structural capsid proteins, *pol* that encodes integration and replication enzymes such as reverse transcriptase and integrase, and *env* that encodes envelope proteins. The viral genome sequence also involves other accessory genes and is flanked by long terminal repeats (LTRs) which facilitate virus integration into the host genome. γ-RVs are capable of efficient, long-term genome integration in a biased-random location, providing a persistent gene transduction method that can be retained throughout multiple cell generations.

The main drawback of γ-RVs is the fact that they can only infect dividing cells, combined with their ability to provide persistent gene integration γ-RVs are therefore well suited for genetic modification of expanding cells such as stem cells e.g. HSC of bone marrow. Current production methods using replication-defective vectors and packaging cell lines provide vector titres sufficient for gene therapy applications and γ-RVs were some of the first viruses applied in in gene therapy clinical trials. However, the trials elucidated the risks of insertional mutagenesis associated with the use or γ-RVs, especially MLV which

integrates at semi-random places in the genome but has a preference for regulatory sequence of the genes (Wu et al., 2003). Therefore, it is more likely to cause a mis-regulation of an important oncogene compared to other integrating viruses. The potential risk requires extensive viral vector characterisation and engineering to limit the possibility of insertional mutagenesis. An alternative approach involves the use of Lentiviruses, a group of RVs with properties similar to γ-RVs but offering distinct advantages.

1.4. Lentiviral vectors overview

LVVs are one of the prominent groups of vectors used in gene therapy and the focus of this EngD project. As part of the *retroviridae* family, Lentiviruses share many of their properties while offering several added benefits (Table 1). The lentivirus-based vectors have been extensively engineered to improve their safety and efficacy, providing a robust platform for gene delivery and therefore an attractive cell and gene therapy vector (Escors and Breckpot, 2010).

1.4.1. Lentiviral vectors development

The use of LVVs is the focus of the gene delivery platform technologies developed by OXB. As such it is important to realise the improvements in the LVV platform since it was first used in research. The first application of LVVs was treatment of human immunodeficiency virus (HIV) infection. This approach helped to overcome the initial safety issues as the vector used to deliver therapeutic gene was derived from HIV (which is a lentivirus itself). Taking advantage of this opportunity, the first clinical trial was initiated in 2003 where patients' cells were extracted, transduced *ex vivo* and then reintroduced to the patients with a later follow-up study (McGarrity et al., 2013). In order to apply LVVs in the treatment of more diseases, improved safety and production efficiency was required. This was achieved by engineering systems for LVV production based on 2nd and 3rd generation LVVs as described in detail in the section 1.4.5 - 1.4.6. The first therapeutic in vivo use of LVVs was performed by OXB using ProSavin®, resulting in encouraging results of the phase I/IIa clinical trial where, reporting no serious adverse events related to administration of ProSavin®. Lasting improvement in patients' motor function over 6 and 12 months period was measured by increased UPDRS scores (6 months, mean score 38 with ProSavin® vs 26 without ProSavin®, p=0.0001; 12 months: 38 vs 27, p=0.0001) (Palfi et al., 2014). Improvement in LVV platform, along with the access to enhanced production methods led to use of LVVs in treatment of a variety of diseases, with multiple products being currently tested in phase I and II clinical trials, including the ones developed by OXB.

1.4.2. Lentiviral vectors applications

While OXB offers a variety of ophthalmological and neurological gene therapy products, LVVs have been widely applied in other areas of science and medicine as well. An extensively studied application includes transduction of HSC/progenitor cells with the aim of treatment of a variety of diseases, most notably several different variants of SCID. γ-RV vectors demonstrated both efficacy and risks associated with gene therapy of HSCs.

Addressing the risks of gene therapy, LVVs are intrinsically less likely to cause insertional mutagenesis and resulting leukaemia, development of self-inactivating (SIN) vectors with insulator sequences further reduces the risks. The efficacy of LVV-mediated HSC transduction has been demonstrated in several mouse models (Adjali et al., 2005; Miyoshi et al., 1999) as well as in human cell cultures (Case et al., 1999; Zielske et al., 2003). Due to the ability of lentiviruses to infect non-dividing cells, *in situ* therapy offers an attractive alternative protocol compared to *ex vivo* transduction usually used with γ-RVs. Using a similar approach of LVV transduction of HSCs, proof-of-concept studies have been performed for treatment of β-thalassemia (May et al., 2000) and haemophilia (Sadelain et al., 2009).

The first clinical application of LVVs was the therapeutic treatment of a HIV infection through inhibition of wild type HIV replication. Since the first clinical trial (Lu et al., 2004) several improvements have been achieved including use of optimised anti-sense RNA together with a conditionally replicating LVV active in the presence of wild-type HIV packaging proteins, resulting in both HIV inhibition and competition for viral proteins, reducing virus packaging ability (Levine et al., 2006). An alternative treatment method has been demonstrated in mice models where CD34+ HSCs were transduced with a combination vector encoding TRIM5α (which inhibits HIV uncoating in the cell), a CCR5 shRNA (which prevents virus entry by preventing cells from displaying CCR5 receptor used by HIV), and a TAR decoy (which inhibits proviral transcription by binding the TAT protein.) (Anderson et al., 2009; Walker et al., 2012). Another example of this approach involves a development of polycistronic vector expressing several antiviral small RNAs (Chung et al., 2014). Overall, both early and current variants of LVV-based gene therapy show promising results in HIV therapy, limiting wild type HIV load and improving immune function in patients.

A different application of gene therapy can be found in generation of induced pluripotent stem cells (iPSCs). iPSCs are generated by transducing somatic cells with *Oct4*, *Klf4*, *Sox2*, and *cMyc* genes. Initially this was performed using separate retroviral vectors (Takahashi & Yamanaka, 2006) but improved methods using a gene cassette with 4 reprogramming genes have been developed (Sommer et al., 2009). Moreover, the problem of the residual expression of iPSC transcription factors in differentiated cells has been addressed through development of a “hit and run” system where the *loxP* site is integrated in one of the LTRs. When expression of Cre recombinase is activated the LVV sequence is excised from the cells, deleting the iPSC transcription factors in differentiated cells (Chang et al., 2009).

The development of LVV-based iPSC generation systems offers a useful tool for stem cell research improving consistency and safety of the transformation, potentially bringing iPSCs closer to a clinical application.

1.4.3. Wild type lentivirus

Wild type lentiviruses consist of ssRNA enclosed in an enveloped capsid. Like other RVs, their genome contains *gag, pol* and *env* sequences encoding structural capsid proteins, integration and replication enzymes and envelope glycoproteins respectively. Lentiviruses also encode *tat* (trans-acting activator that enhances transcription process) and *rev* (regulating mRNA splicing and transport after transcription) and accessory genes such as *nef, vif, vpr* and *vpu*, depending on the type of the virus. The genome also contains cis acting LTRs that facilitate viral genome integration and polypurine tract that serves as an initiation site for complementary DNA synthesis during reverse transcription.

Lentiviruses are integrative viruses capable of infecting both dividing and non-dividing cells that can deliver a large amount of genetic information. Unlike RVs, they prefer inserting into non-regulatory regions of host DNA and therefore are less likely to inactivate or misregulate genes. These properties make lentiviruses a promising gene delivery vector. However, efficient production and safe application required extensive engineering of the virus to avoid risks associated with insertional mutagenesis and formation of replication competent lentiviruses (RCLs). At the same time, high titres of the viral vector are required to achieve a therapeutic effect. These issues have been addressed through the development of progressively more advanced LVVs constructs (see vector engineering in Figure 4).

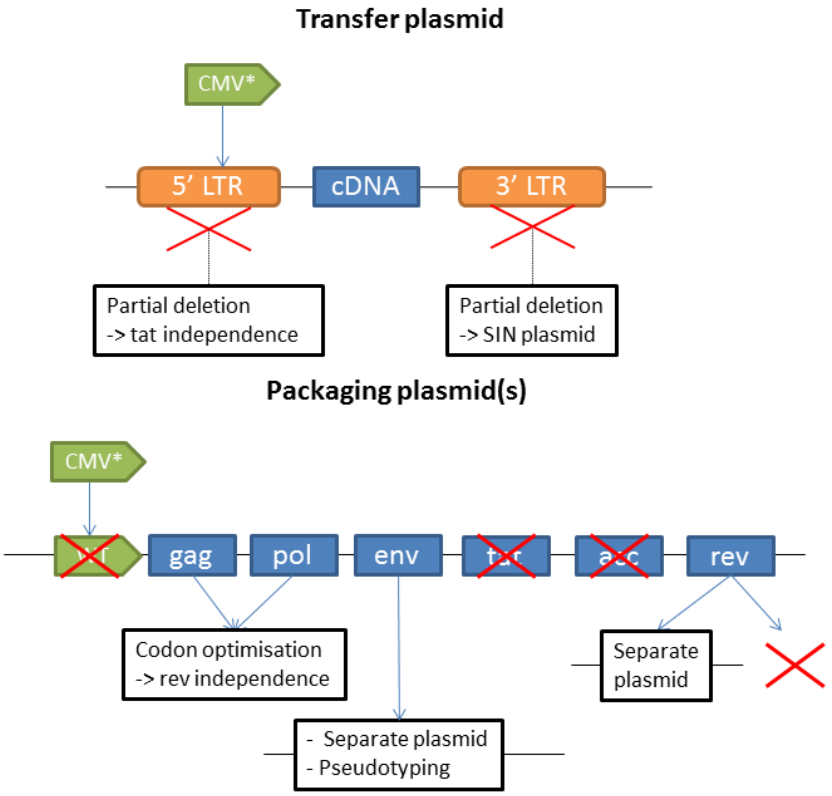


Figure 4: A summary of plasmid engineering introduced in second and third generation LVV. The presented changes have been implemented across several generations of LVVs. Red cross marks a deletion of a sequence; WT – wild type promoter; acc - accessory genes; * - CMV enhancer/promoter is common but other promoters can be used.

1.4.4. First generation LVV

The first step in developing LVVs was to separate trans-acting genome elements from the main viral genome bearing the therapeutic transgene. To achieve this goal, the packaging genes (gag, pol as well as accessory and regulatory elements) were placed in a separate plasmid while all cis-acting elements along with the transgene were placed in a transfer plasmid. Additionally, the env sequence was placed in a separate envelope plasmid, resulting in 3 plasmid expression system (Amado & Chen, 1999). This also allowed production of viruses where the original HIV env genes in the envelope plasmid were replaced with an alternative envelope gene from other virus (a process called pseudotyping). The most common choice is vesicular stomatitis virus G (VSV-G) envelope that is characterised by affinity for multiple cell types and an efficient transduction profile. Pseudotyping reduces homology between different vector elements and improves safety while enhancing the flexibility of the system in terms of possible applications.

In order to produce LVVs, animal cells such as human embryonic kidney cells 293 (HEK293) are transfected with the packaging plasmids. This results in replication of the transfer vector and production of packaging proteins that form the capsid and envelope. The resulting virus particles contain genes from the transfer plasmids in form of RNA enclosed in the capsid and envelope produced by the packaging genes. This way there are no infection and replication genes in the viral RNA, making it impossible to replicate upon infection and therefore limiting the risk of production of creating RCLs.

1.4.5. Second generation LVV

To facilitate further improvement of vector safety as well as to eliminate unnecessary gene elements and optimise production, second generation LVVs were designed. While the details of design may vary, the core changes revolve around removing unnecessary genes and limiting sequence overlap between plasmids decreasing the chance of recombination between vectors. Such recombination event could result in creation of transfer plasmid with some or all of the packaging genes and production of RCLs.

One major change involved a deletion of the accessory genes associated with virus virulence (*nef, vif, vpr, vpu* for HIV based vectors; Kim, Mitrophanous, Kingsman, & Kingsman, 1998) in the packaging vector that in second generation includes *gag, pol, rev* and usually *tat* genes along with a cytomegalovirus (CMV) enhancer/promoter sequence and rev regulatory element (RRE). However, the *tat* sequence can be eliminated without an impact on vector titres by using a chimeric 5' LTR instead of the wild-type sequence. The chimeric sequence involves a deletion which is replaced with a heterologous sequence such as CMV enhancer/promoter. This change makes the gene expression *tat*-independent and allows for deletion of this element from the packaging plasmid (Kim et al., 1998; Miyoshi et al., 1998). Moreover, a deletion in 3' LTR results in the SIN vector that is incapable of gene transcription after genome integration, preventing transcription of any undesirable genes obtained through unintended recombination (Miyoshi et al., 1998; Zufferey et al., 1998). Another important modification involves creating vectors based on non-primate lentiviruses such as equine infectious anaemia virus (EIAV) which is safer than HIV and can be generated using a smaller number of genes. This alternative vector was also found to be equally effective and especially useful in transducing post-mitotic neurons (Mitrophanous et al., 1999).

1.4.6. Third generation LVV

The further improvement was achieved after development of third generation LVVs that offer additional safeguards and improved efficiency. First improvement involves splitting the packaging plasmid into two elements, one with *gag* and *pol* sequences along with CMV enhancer/promoter and RRE and the other with *rev* sequence under control of CMV enhancer/promoter. Splitting the packaging plasmid improves safety by making it less likely that a single recombination event could resolve in production of RCLs (Dull et al., 1998). Other improvements involve codon optimisation of the *gag-pol* sequence and the introduction of an additional open reading frame in the vector genome which makes the vector expression partially *rev*-independent. Due to the optimisation, *gag-pol* expression is *rev*-independent, however *rev* still plays a role in transfer gene expression and is required in currently used HIV-based therapeutic vectors (Kotsopoulou et al., 2000). Codon optimisation also improves the vector expression profile and decreases the chance of homologous recombination with any wild-type RV that could be present in the cell culture or patient's body.

Overall, advancement of the LVVs has greatly improved their safety and allowed the technology to be applied in new fields, resulting in the development of a variety of LVV-based therapeutics that are currently undergoing clinical trials. At the same time the vectors are still being improved, involving a design of two vector system (offering higher titres but at a cost of increased risk of recombination and production of RCLs), five vectors system (with further reduced risk of RCL generation but increased complexity of transfection and possible lower vector production) or replacing the constitutive promoters used in most vectors with inducible promoters for improved control over viral protein expression (Pan et al., 2008). Improved production methods are also in demand, as clinical application requires high titres of the vector and increasingly high volumes to treat more patients in further phases of the clinical trials.

1.5. Lentiviral vectors production (Upstream processing)

LVVs have characteristics such as large therapeutic payload (up to 9 kb), permanent modification of dividing and non-dividing cells and no pre-existing immunity that makes them attractive for developing clinical and commercial applications of gene and cell therapies. However, the requirement for high titres puts pressure on the development of highly effective production methods. The first stage of production includes cell culture and transfection which comprises the upstream processing part of the production.

1.5.1. Cell line development and selection

Selection of the cell line is an important step before LVV production, once the cell line is developed or selected for the process it is banked and individual vials are revived ath the beginning of production cycle. Among multiple available options HEK293 cells are the most popular choice. It is a cell line characterised by high transfection rate allowing for effective expression of the plasmid genes used in production of LVVs. They are easy to culture and readily grow using adherent culture in foetal bovine serum (FBS)-supplemented media. Currently used protocols use HEK293 cells adapted to adherent culture to produce high titres of LVVs for the use in clinical trials. HEK293T cell line is a variant of HEK293 cells which have undergone additional selection and contain the SV40 large T-antigen. HEK293T cell line is characterised by enhanced growth, transfection rate and ability to produce viral vectors, making it the preferred cell line for LVV production. These cells can also be grown in a serum-free suspension culture system bearing in mind that growth medium composition required in that case is different to that used in adherent cell culture. HEK293T cells are therefore well suited for large scale production of viral vectors which is a substantial improvement over other cell lines (Merten et al., 2011). Another variant of HEK293 cells is the HEK293E cell line which includes EBV nuclear antigen which is used for expression of EBV origin of replication. This promotes episomal persistence of plasmids and higher expression of proteins which can be advantageous in production of LVVs (Tom et al., 2008), however HEK293E cell line is dependent on serum for growth.

While different variants of HEK293 cells are commonly used for viral vector production there are also several alternative cell lines available for both research and commercial use. Most of them offer no clear advantage and result in lower or comparable final vector titres. However, COS-1 cells were found to perform better in standard conditions used for growth (adherent culture in FBS-supplemented medium), resulting in 3-4 times higher titre production (Smith & Shioda, 2009). Due to the animal source of COS-1 cells (As

opposed to HEK293 derived from human kidney tissue) and associated risk of animal-derived adventitious agents, they are less popular than HEK293. They are also dependent on serum in adherent growth and therefore offer no advantage in suspension culture, therefore they are more suitable for small scale research application and do not offer improvement for clinical and commercial scale production.

1.5.2. Transient transfection and stable packaging/producer cell lines

In order to produce the viral vector, cells need to be transfected with the plasmids containing viral proteins and the therapeutic gene. This can be achieved either by a transient transfection of the cells in each production cycle or through development of stable packaging/producer cell lines that can be maintained and subcultured to produce the vector as described in more detail later (section 3.1.4). Currently LVV production relies on transient transfection. In this process the essential plasmids (the exact number and genetic structure may vary as detailed earlier in sections 1.4.3 - 1.4.6) may differ for the different vector generation used in the process) are introduced into the cell and start producing proteins and RNA encoded by the plasmids. The expression plasmids include *gag*, *pol* and *env* (and possibly *tat* and *rev*) sequences as well as the transfer plasmid with transgene and cis-acting elements. There is a variety of protocols available for the process that can achieve high transfection rate and subsequently high vector titres. The critical parameter is the delivery of appropriate plasmid DNA concentration at an appropriate plasmid ratio. One of the commonly used transfection agents is calcium phosphate (CaPO) which forms a CaPO/DNA complex used for transfection of adherent cell monolayers by forming a precipitate on the surface of the cells and facilitating endocytosis of the DNA (Kingston et al., 2003). However, there are multiple issues associated with use of CaPO as it is toxic to the cells and highly pH dependent, it also requires use of serum and medium change as part of the transfection protocol. As such CaPO is an inexpensive but highly variable and sensitive transfection agent which is used mainly in research setting (Schweizer & Merten, 2010). A popular alternative is polyethyleneimine (PEI) and its derivatives such as PEI Pro®, which allow for serum-free transfection and are less toxic for the cells making it an effective method of cell transfection for both the adherent and suspension cell culture. It requires use of correct PEI/DNA ratio for effective transfection but in optimised conditions can result in high transfection rates and vector production (Pham et al., 2006). An alternative transfection reagent is Lipofectamine™, a lipid-based transfection agent that transfers plasmids into the cell through liposomes. It is highly effective but expensive and therefore used predominantly in small scale production

(Pham et al., 2006). Another method includes electroporation which uses electric field to temporarily increase the cell membrane permeability. It can generally be used only with small volumes, but flow electroporation is a semi-continuous process where the cells are transfected as they pass through the electric field. It results in production of high titre LVVs and involves no reagent and no cell toxicity. It is therefore a potentially useful method; however its scalability is limited (Witting et al., 2012). An interesting transfection method involves the use of baculovirus which can introduce plasmids into the cells (Lesch et al., 2008). However, each plasmid has to be introduced by a separate virus and therefore the method is less reliable for systems with higher number of plasmids i.e. 3rd generation LVVs which are regarded as superior to earlier generations.

Transient transfection is commonly used for production of LVVs in both small and large scale for research, clinical trials and small-scale commercial purposes. In this production system, high titres of the vector can be achieved but it requires an additional, high-cost transfection step before production and the results can vary significantly between batches. This limits its use in large scale industrial production of LVVs as biopharmaceuticals. Therefore, development of effective inducible stable packaging and producer cell lines is considered critical to produce sufficient amount of vector required to support therapeutic applications. A stable packaging cell line encodes all trans-acting viral genes and therefore it only requires a transfection of the transfer plasmid in a single efficient step. Alternatively, inducible stable producer cell lines contain all required plasmids (including the transfer plasmid) allowing to skip the transfection step. However, development of stable packaging/producer cell lines faces several challenges. First of all, some of the vector proteins are toxic to the cell when expressed at a high level such as VSV-G envelope glycoproteins and some of the enzymes encoded by *gag-pol* sequence (Karacostas et al., 1993; Rohll et al., 2002). Therefore, the preferred inducible system can accumulate cell biomass before producing the vector. An early approach involved using the tet-off system where tetracycline would inhibit production of vector proteins (Kafri et al., 1999). However, this system lacks flexibility and requires a change of medium (for the medium without tetracycline) as a method of induction which can be both problematic and expensive.

As an alternative, a tet-on/cumate system was developed by Broussau et al. (2008) where protein production is induced by addition of tetracycline into the medium. It was possible to generate final titres of 10^7 TU/ml using this system which were more viable for both clinical and commercial applications compared to alternative systems. Stewart et al.

(2011) have developed two stable producer cell lines capable of LVV production in adherent culture resulting in high titres (comparable to transient transfection), good vector quality, tight control over induction (no viral proteins were detected in non-induced samples) and genetic stability of the cell lines over up to 111 days. Another approach of improving stable cell lines is to use a different non-toxic envelope protein such as RD114-TR, used for constitutive expression of viral genes resulting in vector production titres of 10^6 TU/ml (Stornaiuolo et al., 2013). This could potentially allow for constitutive expression of viral proteins with no toxicity. However, this non-toxic envelope has limited use as it is specific for HSC. An inducible stable producer cell line would be well suited for use in a suspension culture bioreactor where cells could accumulate biomass and then be induced to produce the viral vector. Perfusion process could potentially be used where the media is exchanged through filtration to provide optimal cell growth conditions and viral vector production is either continuous or induced once sufficient biomass concentration is achieved. Producer cell line-based system has a good potential for scale-up and industrial production of LVVs. Overall, the inducible stable cell lines show promising results but it is difficult to generate cell lines which retain both high growth rate and are capable of producing high titre viral vectors. Therefore, transient transfection is still a preferred method for LVV production at least until a reliable stable producer cell lines with high production efficiency can be developed.

1.5.3. Growth medium

Another important factor in viral vector production is selection of the growth medium composition. Different variants of Dulbecco's Modified Eagle Medium (DMEM) are commonly used for cell culture. The medium is supplemented with a variety of additives, the one with highest impact is FBS. Historically FBS has been used as an additive for culturing mammalian cells and it has been used in production of LVVs as it contains many elements that help cell growth and make the production process easier and more effective. However, FBS has multiple drawbacks. It comes from an animal source and therefore leads to risk of contamination with animal viruses. They are difficult to eliminate from the final formulation as no virus inactivation step can be used in downstream processing due to the nature of the product. FBS also makes downstream processing overall more complicated and less effective. Moreover, FBS has a complex composition that can vary between batches and therefore a more defined alternative would be desirable to decrease the process variability. FBS is also difficult to source due to limited supply and increases the risk to supply chain continuity

which can be especially problematic in time sensitive manufacturing processes such as in the case of viral vectors used to modify patient-derived cells (e.g. CAR-T therapy). While FBS has its drawbacks, it is used in cell culture because it improves growth of cells and adherence and therefore it is difficult to substitute. One method to limit problems of using FBS is a protocol that uses a serum-containing medium at first to generate biomass which is then followed by a medium exchange with serum-free medium (Schambach et al., 2009). This approach reduces the contamination level for downstream processing; however, it does not eliminate risks of viral contamination and therefore does not improve biosafety. Overall, FBS is difficult to substitute for adherent culture, however elimination of FBS inhibits cell adhesion and therefore a serum-free medium is much better suited for the suspension culture. It has been demonstrated that HEK293 cells can grow in a serum-free suspension medium and produce LVVs; therefore a serum-free medium is likely the optimal choice for suspension culture (Merten et al., 2014b). Media provided for suspension culture are chemically defined and multiple solutions are available from manufacturers (e.g. Thermo Fisher Scientific, Fujifilm Irvine Scientific).

Apart from FBS, there are several other media supplements that can affect LVV production. Cholesterol can be added to improve infectivity of the vector which was attributed to change of membrane composition of either the producer cells or the vector itself (Chen et al., 2009). Sodium butyrate is another common additive; it is a histone deacetylase inhibitor which leads to hyperacetylation of histones, chromatin decondensation and therefore higher transcription and expression rate of viral proteins (Davie, 2003). However, it was reported that sodium butyrate action may be envelope protein-dependent and in commonly used VSV-G enveloped vectors it has little to no effect on final titres and therefore its application is limited (Sena-Esteves et al., 2004). Nevertheless, the use of sodium butyrate is cited in multiple protocols for boosting LVV production. Chloroquinone is an additive that increases pH of endosomes and lysosomes which in turn inhibits DNA degradation. However, it is toxic to the cells and therefore its use is limited. Moreover its action was found to be transfection agent-dependent and it had no effect when used with commonly utilised PEI (Kuroda et al., 2009). There is also a single example of use of caffeine which resulted in improved titres, however the mechanism of this reaction is unclear (Ellis et al., 2011).

1.5.4. Adherent and suspension culture

As discussed above, different media have been developed for use in adherent or suspension culture systems. The most common approach until recently was to scale out using

2D adherent culture for vector production using cell lines such as HEK293. It is commonly used at the laboratory scale for production of LVVs due to the stable growth and high titres that can be achieved. However, protocols for culturing the cells can be problematic during the cell expansion phase as whenever the cell culture vessel has to be switched to a larger surface area the cells have to be detached and then reattached to the new surface. This procedure usually involves a change in culture conditions (temperature, aeration, volumes used) and therefore can affect cell stability. Detachment is commonly achieved through use of purified cell-dissociation enzymes (such as TrpLE) followed by medium exchange. Alternatively, cells can be kept in a single cell culture vessel with periodically exchanged medium, however this limits the area available for growth of the cells. The cells require a surface for growth and therefore process scaling is limited and mostly involves increasing the number of vessels, i.e. linear scale-out which offers limited efficiency (Merten et al., 2014a). Another issue is the limited control of the process as well as variability in some of the parameters such as the temperature or gas distribution when using multiple vessels. An alternative involves use of microcarriers in suspension culture where cells can attach to small beads which are then freely floating in the medium. This approach can be adapted using existing bioreactor technology which can be freely scale up while providing benefits of a well-controlled environment of a bioreactor. Maintaining cell attachment to the microcarriers can be difficult and require dedicated cell culture medium. Fixed bed bioreactors such as iCELLis® can also be utilised, where cells are attached to a high surface fibrous membrane and media is recirculated to provide uniform nutrient and gas distribution.

For improved scale-up potential, cells can be adapted to suspension culture using methods such as stirred tank reactor (STR) or WAVE bioreactor. Using serum-free medium promotes cell detachment and HEK293 and HEK293T cells can be adapted to suspension culture. This method results in better mass transfer of gases and nutrients and is generally characterised by better control methods, especially when used in an STR where multiple standardised probes can be easily used to measure temperature, pH and oxygen level as well as other parameters. These reasons make suspension culture much better suited for large scale production. However, despite the advantages of suspension culture, cell growth rate is generally lower compared to adherent cell culture (Merten et al., 2014a). This issue results in continued use of the adherent cell culture in production of small to medium scale LVVs used for clinical trials. However, transition to optimised serum-free suspension culture can offer multiple benefits.

1.5.5. Process scale-up

Difference between adherent and suspension cell culture is especially evident when considering the potential for scale-up. At the small scale there are reliable methods available for both kinds of cultures, however adherent cells are predominant because they are easier to grow. Commonly adherent cells are grown in flat bottom T-flasks kept in incubators where temperature and atmosphere can be controlled. For suspension cells shake flasks are commonly place on a rocking platform in an incubator with controlled temperature, humidity and gas composition. Small scale production is usually performed for research purposes or to subculture the cells. Flasks can also be used to prepare a seed culture for bigger vessels.

Large scale production methods for adherent cells involve cell factories which consist of vessels with multiple layers stacked on top of each other. They offer large surface for cell growth and can be scaled up linearly by increasing the number of chambers in a stack and the number of stacks. Adherent cell production using this technology is relatively simple and cost-effective, however gas exchange in the flasks can become a problem, especially when using high number of chambers per stack. This issue is addressed by using a permeable film at the bottom of each chamber as in the HyperFlask® design (Kutner et al., 2009). Another design utilises a concept of roller bottles where rotation distributes media across the entire surface of a large bottle. This design provides a large surface for growth and can be partially automated. In both cases the scaling is linear and therefore problematic for large scale production. An alternative involves use of microcarriers and packed bed bioreactors which can utilise large scale reactors for growth of adherent cells either on beads suspended in the media or on a matrix fixed in the reactor. However, in either case the hydrodynamics of the reactor associated with stirring and aeration can cause cell detachment and death which is a major issue.

For the suspension cell culture, low volume STRs (5-10 L) are often used for research and test runs as well as for growth of seed culture for larger bioreactors. They offer good control of process parameters and can be partially automated, involving temperature and pH control, aeration and stirring rate. They can be further scaled, however increasing the size of the vessel requires scaling the impeller and maintaining the same power per volume and stirring rate increases tip speed which in turn can exhibit more shear stress on the cells and eventually lead to decreased cell viability and reduced viral vector production. Another major issue is the need to keep STRs sterile between runs as it becomes more problematic as the size of the vessel increases. To address this issue a variety of disposable single-use

bioreactors are available. An alternative to STRs for suspension culture are WAVE bioreactors which can be used with bags up to 500L. This technology utilises rocking motion to mix medium in a sterile bag which addresses the issue of shear stress and is suitable for suspension culture or adherent culture using microcarriers. However, while the agitation method used in WAVE generates less bubbles and shear when compared with STR, it also limits mass transfer of oxygen and therefore larger scale application may offer limited ability to sustain cell metabolism required to produce high titre viral vectors (Merten et al., 2014b).

1.6. Lentiviral vectors purification (Downstream processing)

Upstream processing is followed by a series of purification and concentration steps required to remove impurities and reach a therapeutically active concentration of the viral particles. High purity is especially important to ensure patient safety. Downstream processing follows the initial production and ensures sufficient product quality. Common downstream processing steps are summarised in Figure 5

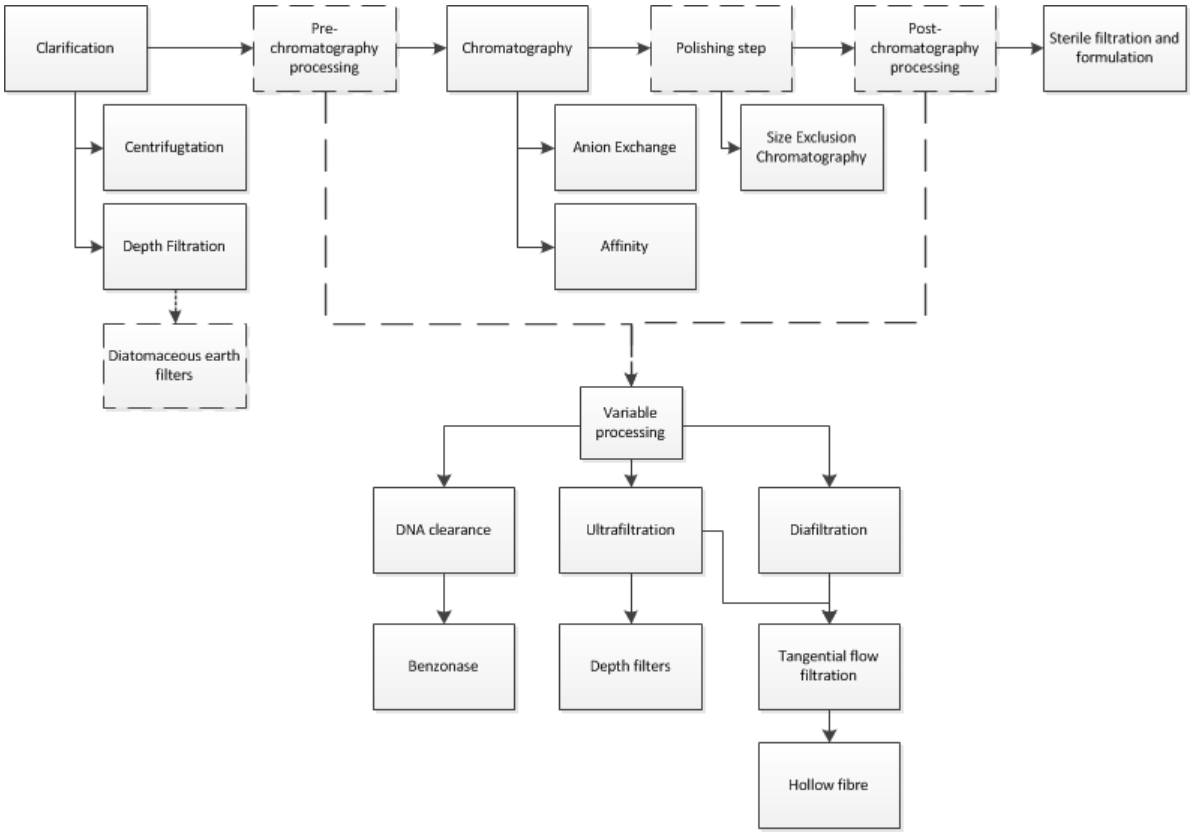


Figure 5: Summary of common downstream processing steps used in purification of lentiviral vectors.

1.6.1. Initial steps

Once the viral particles is produced it is necessary to perform a series of capture and purification steps to remove cells and cellular debris, reduce volume and purify the viral particles to a level of quality required for a clinical application and to get the viral particles in to the correct formulation buffer before vialing. The common impurities can be product or process related. The first group involves inactive, incomplete, and aggregated viral particles which result from a variety of virus assembly issues. These kinds of impurities are difficult to avoid and remove due to their similarity to a properly assembled particle. However, due to this similarity they tend to pose smaller risk to the product quality compared to other impurities. Process related impurities can originate from the cells, medium or transfection

reagents used in the process, including cell debris, host cell and transfection plasmid DNA, cellular proteins as well as any serum proteins if FBS is used in the process. Any reagents and enzymes used in cell processing and purification (e.g. Benzonase endonuclease) must be removed as well. Moreover, DNA from host cell and plasmids used for transfection can end up contaminating the sample. Finally, the bulk volume of water and media has to be removed to concentrate the viral particles to achieve a therapeutically effective dose.

The first step in purification cycle is clarification used to remove the cells and cell debris that can be easily separated from LVVs. A common method involves centrifugation and ultracentrifugation of the sample where impurities are pelleted out and viral particles remains in the supernatant and can be recovered. It is a useful method for small scale purification; however, it is not scalable. For larger volumes long cycles of lower speed centrifugation can be used but such approach tends to be less effective and can disrupt the virus and result in low step yield. Instead, large scale production relies on dead end filtration. Using the fact that viral vectors are small particles, many bulky impurities can be retained on the membrane and quickly purified out. However, this leads to membrane fouling which can be a major problem addressed by using several filter membranes in succession, using progressively smaller pore sizes. This way less material is retained at each membrane and fouling is less of a problem. The clarification process can also be supplemented by use of diatomaceous earth which can help to separate cell debris and small molecules (including LVVs).

The next step involves purification of DNA contamination using enzymes such as Benzonase®. It is a purified endonuclease that cleaves cellular and plasmid DNA but is unable to access viral RNA enclosed in the capsid (Sastry et al., 2004). This step involves seemingly straightforward incubation and can be performed early or late during the purification; however, the incubation can have a significant effect on process timing and therefore viral vector stability. Additionally, Benzonase® is a process related impurity itself and needs to be removed. Therefore it is commonly applied before the chromatography step where it can be removed in the column flow-through (Zufferey, 2002). However, this approach requires a larger quantity of Benzonase® due to higher initial volume. Alternative protocols involve Benzonase® treatment after chromatography and sample concentration where a smaller enzyme quantity can be used. Higher DNA removal rate has been recorded for this approach, however Benzonase® has to be purified in a separate step, usually through filtration (Bandeira et al., 2012). This results in lowering the cost associated with Benzonase®

treatment but may result in loss of viral vector recovery due to additional filtering step. It is also possible to perform two Benzonase® treatment steps to maximise the removal of contaminating DNA. Overall, the decision about positioning of Benzonase® purification depends on the required cost to recovery ratio and the overall protocol used.

The final step which is sometimes applied before or after chromatography is a hollow fibre ultrafiltration which is a concentration and purification step where small pore size membrane (usually 100-500 kDA cut-off) is used to retain the viral vector and get rid of a bulk volume of the media as well as small impurities that pass through the membrane. As in other membrane based methods, membrane fouling is a major problem which can be significantly reduced by use of tangential flow filtration (Reiser, 2000). This method is used in some protocols for sample concentration before, after or instead of chromatography in conjunction with more filtration/diafiltration steps. As with benzonase purification, the protocols can be flexible and allow for variation where it is difficult to find an optimal solution.

1.6.2. Chromatography

Chromatography is a purification process widely applied in the biopharmaceutical industry due to flexibility offered by the large number of available methods and protocols as well as high the purification quality that can be achieved. In case of LVVs the most common method is anion exchange chromatography (AEC) which uses the fact that viral particles are negatively charged and can be bound by a positively charged column. This method can be scaled for industrial use and results in 20-30 times concentration of the sample and up to 65% step recovery yield (Slepushkin et al., 2003). The main challenge involves handling the vector particle size which is not compatible with commonly available columns designed for protein purification which results in low binding as well as shear stress that can disrupt the vector (Maria de las Mercedes Segura et al., 2013). Therefore, alternative methods including monolithic columns such as CIM DEAE or membranes such as Mustang® Q or Sartobind® Q have been utilised to develop a method more suited for purification of viral particles. Another issue is the fact that elution buffer containing high salt concentration is detrimental to virus activity and therefore AEC requires quick desalination after elution to improve the recovery yield (Segura, Kamen, Trudel, & Garnier, 2005) .

Another method that can be used for LVVs recovery is affinity chromatography. A common approach involves a heparin packed column which shows affinity for the vector

particles. It does not interact with the envelope protein and therefore can be used for differently pseudotyped vectors. This method results in 33-53% step recovery yield (Segura, Kamen, Lavoie, & Garnier, 2007). Alternatively, other commonly used affinity columns have been adopted for viral vectors purification, including immobilised metal ion affinity chromatography (IMAC) packed with Ni ions. This application requires a specifically engineered viral envelope protein with hexa-histidine tag attached in order for the virus to bind to the column. While effective, this particular method faces the problem of vector inactivation caused by use of high concentration imidazole for elution as well as the potential metal-mediated oxidation (Cheeks et al., 2009). The tag can also cause an immune reaction in a patient and its use is discouraged by FDA for clinical applications unless the tag can be removed. Approach similar to the principles used in IMAC can be applied to avidin-biotin interaction as an alternative system but it has similar limitations (Nesbeth et al., 2006). For purification of VSV-G pseudotyped LVV it could also be possible to use VSV-G binding antibodies in immobilised phase. Overall, adaptation of affinity chromatography can be considered as an alternative to AEC but it makes it more difficult to meet regulatory requirements for viral particle safety and immunogenicity profile.

Following the early purification steps, size exclusion chromatography (SEC) can be used as a polishing step. It uses separation based on size of the particles passing through the columns and it offers a step yield of 70-85% (Slepushkin et al., 2003). However, it has low loading capacity and requires initial concentration of the sample; the column can also trap the viral particles, depending on the size of the pores and structure of the column. Therefore, SEC is used only in some of the protocols; it may be particularly useful if a Benzonase® step is used after AEC as at that point the sample is concentrated and most other impurities are eliminated. It is also possible to utilise multi-modal approaches such as Capto Core resin which combines size exclusion and binding capacity

Finally, following the main purification step of chromatography and any filtration/diafiltration steps that may follow, the sample needs to be sterile filtered and prepared for storage and delivery to the patient. The sterile filtration using a 0.22 µm dead end membrane filter ensures the sample is free of infectious agents and it is usually a requirement for the sample to be used in clinical application unless sample sterility can be assured through validated aseptic process. The LVVs then need to be stored at -80^{o}C as there is no formulation that would completely counteract very short half-life of the vector preparation reaching 1-2 days at room temperature and 8 days at 4^{o}C. A study addressing use

of lipoproteins and recombinant serum albumin proteins has shown improved stability of the preparation (Carmo et al., 2009). However, until a more effective formulation is available storage and transport of the vector remain problematic.

1.7. Conclusions

Despite the initial setbacks, gene therapy has seen significant advancement in recent decades and become a promising area of medicine capable of delivering solutions to previously unmet needs. By learning from the past problems, it is possible to identify and overcome the major challenges of viral vector based gene therapy such as immunogenicity, insertional mutagenesis and difficulties in manufacturing. With a variety of available platforms based on different virus types, it is possible to develop varying therapies designed for specific applications. The improvements introduced in second and third generation LVVs led to improved safety and efficacy which led to development of effective therapies. OXB's expertise and product catalogue is strongly based in the LVV platform which makes it an important focus of this EngD project in terms of process characterisation and optimisation.

One of the main factors limiting LVV application is the challenge to provide high titres and develop large scale manufacturing. Over the last two decades significant progress has been achieved in improving vector production methods. Improved control and efficiency of transfection along with development of serum-free suspension process provides a base for production of consistent product. Downstream processing consisting of a combination of filtration and chromatography steps ensures high product concentration, safety and quality. Based on the need for improved process understanding of LVVs production, this EngD project aim is to apply a combination of experimental, statistical and computational methods to process characterisation, optimisation and modelling.

The overview of gene therapy area presented in this chapter, in particular viral vector production process, presents the context of research described in this thesis. In Chapter 2, LV manufacturing and process development data is analysed using several statistical techniques. This includes multivariate data analysis methods introduced to analyse data trends and to assess feasibility of using these methods as part of manufacturing and process development workflow.

Chapter 2 Lentiviral vector production process characterisation and analysis

2.1. Introduction

As established in the previous chapter, lentiviral vectors (LVVs) have multiple applications in gene therapy and offer unique benefits compared to other related platforms. However, large scale production remains challenging and requires a consistent and efficient process to meet the production goals and satisfy regulatory bodies. Oxford BioMedica (OXB) is actively involved in manufacturing LVVs in serum-dependent adherent cell culture as well as development of improved production process based on serum-free suspension cell culture. LVVs must be produced according to Good Manufacturing Practice (GMP) regulations to ensure product quality and patient safety. This is achieved by providing detailed information about the process and maintaining batch records. The process description is covered by standard operating procedures (SOPs) which detail each step to ensure that each batch is produced in a standard manner, thereby reducing process variation and ensuring product consistency. This approach requires identification and good understanding of the process parameters and quality attributes at each unit operation as well as their interactions and combined effects. These parameters are captured in batch manufacturing records (BMRs) to facilitate process control and monitoring as part of quality control. BMRs along with other records (data acquisition systems and engineering configuration records which capture the details of equipment operation) ensure that individual batches are produced according to the same procedure, taking note of any deviations and unexpected variations in process parameters which may lead to a difference in the final product quality. The adherent cell culture-based LVV manufacturing process is complex, and it is closely monitored by the operators. The abundance of data present in the executed BMRs could potentially be useful for determining the impact of process parameters and identifying those steps in the process which introduce the most variability or impact product quality. Demonstrating the understanding of process design space and control needed to operate within it is an important principle of QbD that ensures product consistency throughout process lifetime (ICH, 2009) and the analysis evaluated in this chapter would demonstrate compliance with the guidelines on continuous improvement.

LVV production can be performed using either adherent or serum-free suspension cell culture which impacts multiple parts of the process, including critical process parameters,

control regime as well as the overall performance (Merten et al., 2014b). The main parameters monitored throughout the process focus on cell performance reflected by cell count, viability and confluence. In suspension process there are additional parameters which can be directly controlled and affect the LVV production including oxygen levels, pH and temperature as demonstrated in similar systems used for mammalian cell culture (F. Li et al., 2010). Process performance is assessed based on the final viral vectors titre measured through several assays including polymerase chain reaction (PCR) and flow assisted cell sorting (FACS) to ensure product quality and consistency (Geraerts et al., 2006). Additional information is collected when using new cell lines, especially when developing packaging and producer cell lines where cell performance and productivity are assessed. Overall LVV production and process development monitoring outputs a large amount of data which describes process flow and product properties. It can be used to examine trends present throughout the process and inform optimisation steps. It is therefore useful but challenging to analyse the wealth of generated data to improve product and process understanding as well as process control.

There are a number of statistical methods available to analyse the process data documented in BMRs and process developments reports. Selecting the best method to apply to the data set is important for obtaining meaningful results. These methods can range from simple analysis to complex multivariate methods. Assessing correlation between different variables is a simple approach but can highlight some trends and point to significant interactions. They can be further assessed through use of multivariate data analysis (MVDA) which helps to identify and visualise combined effects of multiple variables in the process (Glassey, 2012). Data structure is an important consideration for MVDA – in BMRs the data is characterised by variables at multiple timepoints and batches. These three dimensions of data would typically be converted into a two-dimensional data matrix before analysis (Wold et al., 2009). This data unfolding can be performed as batch-wise unfolding where each batch (as a single row in data matrix) is described by variable values and time (each variable at different time points is a separate column in data matrix) to focus on batch-to-batch variation. In contrast, observation-wise unfolding describes each variable (presented as individual columns in data matrix) using batch and time (each batch at different time points is a row in data matrix). Data unfolding applied in OXB's BMRs is further described in section 2.2.1.

Principal component analysis (PCA) is a powerful exploratory method which can be used to reduce data dimensionality, cluster data points, assess the effect of individual

variables in the multivariate trends and serve as a basis for quality monitoring model (S. Wold et al., 1987). After analysing a set of reference batches, PCA and other multivariate methods can be applied to assess future batches and compare them to the reference standard ("golden batch") to detect any variation (Nomikos & MacGregor, 1994). MVDA can be further utilised to introduce additional methods of process analysis based on chemometrics. Near infrared and Raman spectroscopy have seen increasingly more application in process monitoring and control (Rowland-Jones et al., 2017) while MS can be used for process and product analysis and development (Povey et al., 2014). Overall a combination of different statistical methods can be used throughout the process to inform decision making and help extend process understanding and maintain product quality in accordance with QbD guidelines. This chapter present results of process analysis using a selection of statistical methods which was performed to improve process understanding and assess applicability of multivariate data analysis approaches to process analysis and manufacturing batch quality monitoring.

2.2. Methods

A vital part of GMP is maintaining process consistency and limiting risks to product quality. OXB BMRs contain information about process parameters and quality attributes and were analysed to provide insight into LVV production process using a combination of traditional statistics and MVDA. Details of methods used to obtain and analyse the data are provided below.

2.2.1. Batch manufacturing records for LVV production in adherent cells

Executed BMRs of the OXB LVV production process are available for multiple products and contain detailed information about several process parameters at different stages of the process. To improve process understanding and analyse interaction between these parameters, the executed BMRs were analysed using PCA. OXB BMRs describe the process of lentiviral vector production in HEK293T cells cultured adherently in 10-layer cell factories (CF-10, Nunc™ EasyFill™ Cell Factory™ Systems, Thermo Fisher Scientific) with intermediate expansion in T-flasks (Thermo Fisher Scientific) and 2-layer (CF2) cell factories (Thermo Fisher Scientific). Cells were first revived from a vial (1-1.5 ml, -150°C storage) and cultured in Dulbecco's modified eagle medium with phenol red (DMEM, GE Healthcare Biosciences) with 10% foetal bovine serum (FBS, Life technologies or Gibco) in T225 T-flasks (Thermo Fisher scientific). They were then expanded to vessels with increasing volume and number of vessels with steps involving CF 2 cell factory, CF10 cell factory, 11 CF10s and finally 25 CF10s (Thermo Fisher Scientific). In the final 25 CF10s 24 hours after inoculation cells were transfected with a set of EIAV based plasmids proprietary to OXB using Lipofectamine® 2000CD Transfection Reagent (Thermo Fisher Scientific). 18 hours after transfection the cells were induced with sodium butyrate (NaBu, 10mM, Sigma Aldrich). The vector product was harvested 8 hours after induction (harvest 1, followed by media replacement) and 22 hours after induction along with the cell samples (harvest 2). The bulk harvest was then pooled and purified using a combination of normal flow filtration, AEC and ultrafiltration by hollow fibre filtration. The process is summarised in Figure 6 and Figure 7. The data from downstream processing steps was not assessed in this analysis as records were not available at the time of writing which would further reduce the number of batches suitable for analysis.

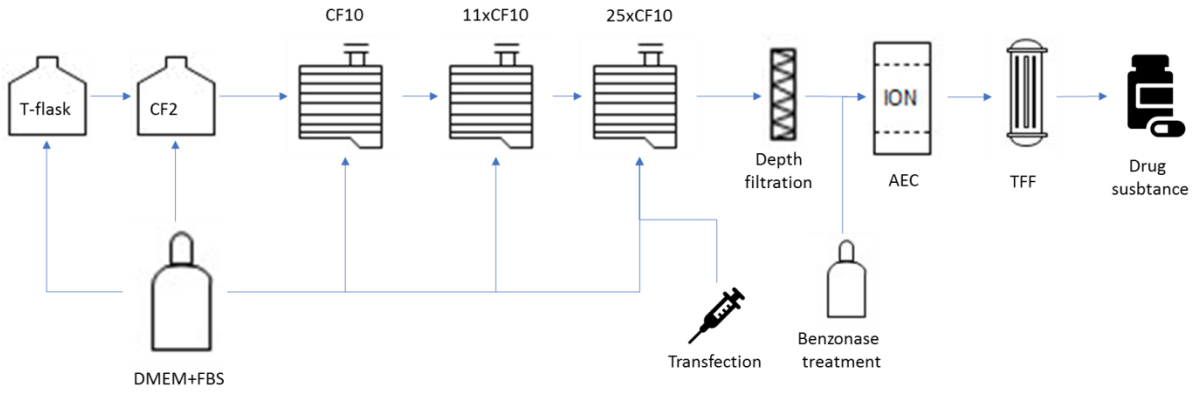


Figure 6: Process diagram of LVV production process as captured in the OXB's batch manufacturing records.

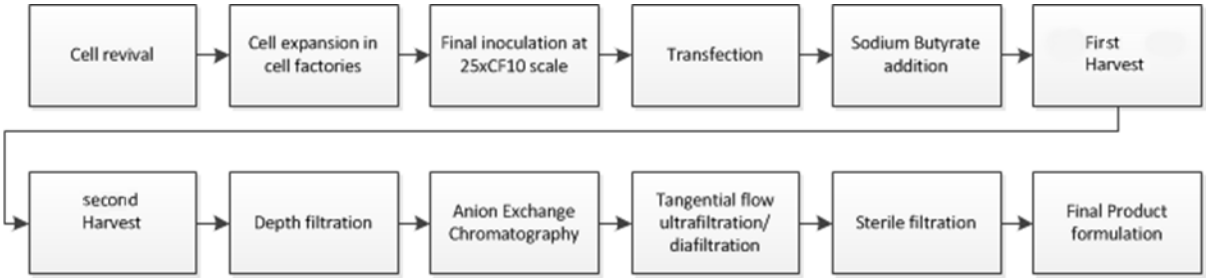


Figure 7: Summary of LVV production process captured in the batch manufacturing records.

The BMRs examined in this analysis were provided for total of 34 batches of adherently produced LVV spread between 4 different EIAV-based products. In total 24 quantifiable variables were identified and measured at different time points during the 14-day manufacturing cycle. Due to the large amount of data available, including pilot runs, many of the batch records were incomplete and therefore difficult to include in the analysis. As a result, a total of 15 batches were identified and 20 variables with the most complete records were chosen for the statistical analysis. All 15 batches were manufactured according to the current platform process and included data for four different gene therapy products, produced using the same adherent serum-dependent HEK293 cell line. The 20 variables used in the analysis included viable cell count at 4 different time points, cell viability at 5 time points, cell confluency at 9 time points, and 2 measurements of the final viral vector titre based on RNA copy number or a transduction assay based on FACS (Table 2). The variables were measured following OXB's GMP-compliant quality management system, including rigorous titre analysis resulting in coefficient of variation (CV) <15% across biological repeats. BMRs include certain variables at multiple time points which means that the data was unfolded batch-wise in the records. Each batch was characterised by a single row of variables at different time points are presented as separate columns (Wold et al., 2009). Batch-wise unfolding was the preferred approach over observation-wise unfolding to focus on variability

between batches and their evolution over time rather than trending the variables. Therefore, no changes were applied to data folding. The results were collated into a single matrix in Microsoft Excel and imported to MATLAB version R2013a (Mathworks). Prior to PCA, the data was normalised by scaling each variable to units of standard deviation. The data was then mean centred using the PLS toolbox version 8.0.1 (Eigenvector) pre-processing to compensate for the variable scale of the different variables and to maximise the captured information. PCA was performed using the Eigenvector PLS toolbox with venetian blinds cross-validation settings. Classes were assigned to samples based on the viral vector product of each batch. PC scores and loadings plot were examined for 2 major datasets: 15x20 matrix including all selected batches and variables; 15x18 matrix including all batches and all variables except the final titre data (RNA and FACS assays).

Variable	Type	Description	Variable	Type	Description
1	Offline	Day 1 VCN	11	Offline	Day 11 Confluency
2	Offline	Day 1 Viability	12	Offline	Day 11 Viability
3	Offline	Day 4 Confluency	13	Offline	Day 11 VCN
4	Offline	Day 4 Viability	14	Offline	Day 12 Confluency
5	Offline	Day 5 Confluency	15	Offline	Day 13 control confluency
6	Offline	Day 6 Viability	16	Offline	Confluency pre-induction
7	Offline	Day 6 VCN	17	Offline	Confluency post-induction
8	Offline	Day 8 Confluency	18	Offline	Day 14 confluency
9	Offline	Day 8 Viability	19	Analysis	Vector titre (RNA assay)
10	Offline	Day 8 VCN	20	Analysis	Vector titre (FACS assay)

Table 2: List of variables used in statistical analysis of BMRs documenting EIAV vector production at OXB

2.2.2. Suspension process development data

The executed BMRs describe the monitoring of GMP compliant manufacturing of LVVs in adherent cell culture. Serum-free suspension process has been developed and is being implemented in manufacturing at OXB. There is abundant process data for the suspension cell culture development which can be used similarly to the BMRs for the purpose of MVDA. There is a significantly higher degree of sample variability in the data set as LVVs were produced at different scales using varying process parameters. Vessels used in the recorded experiments included 0.5L MiniBio reactors (Applikon biotechnology), 7L EZ Bioreactors (Applikon biotechnology) and 50L BIOSTAT® CultiBag® stirred tank reactor (Sartorius Stedim Biotech). Cells used for the experiments were producer cell lines encoding one of the OXB products as well as baseline HEK293T cells used for transient transfection.

All bioreactor experiments shared a common protocol with a degree of variation based on the nature of the experiment. Cells were revived from a vial (1-1.5ml, -150°C

storage) and cultured in FreeStyle™ 293 media (Thermo Fisher Scientific) supplemented with cholesterol lipid concentrate (0.1% v/v using 250x Cholesterol Lipid Concentrate, Thermo Fisher scientific) in shake flasks (250 or 500 ml, Corning) for at least a week prior to inoculation. Bioreactor vessels were assembled and autoclaved prior to the experimental use (except for pre-sterilised single use bioreactors). Bioreactors were charged with FreeStyle™ 293 media and inoculated with previously cultured cells (cell concentration varied between experiments). Cells were induced with doxycycline and NaBu (10mM) at varying points after inoculation. Throughout the process stirring rate, oxygen level, pH (0.05 deadband) and temperature were controlled at set points varying between experiments according to a design of experiments method or standard values used in the current version of the process. pH, oxygen, CO_2 and metabolite levels (glucose, lactate, glutamine, glutamate, measured in YSI Biochemistry analyser) were all measured offline for some or all of the experiments. Vector samples were harvested at varying time points according to experimental protocol and used to calculate the viral vector titre.

The preliminary analysis (descriptive statistic of sample size, range and percentage of complete records) was performed on a data set of 125 batches characterised by 57 variables to assess quality of the data set (Appendix 1, Table 11). A second data set was formed by omitting batches and variables with a significant amount of missing data (over 70%), resulting in a data set of 123 batches characterised by 26 variables (Table 3)

Some MVDA techniques can be used with incomplete data and it is possible to extrapolate data e.g. through linear and non-linear regression. In this case variables with over 70% missing data were excluded to avoid extrapolating data where it could result in false prediction and to focus the analysis on better characterised and more complete variables. In general terms the variables used for final analysis (described in section 2.3.2) were process setpoints, measurements taken at the time of inoculation, transfection and harvest as well as final titre analysis. The data sets were imported into MiniTab® statistical software and analysed using the correlation analysis tool as well as into MATLAB (R2013a) where PCA was performed using the PLS Toolbox (8.0.1).

Variable	No	Type	Details
Working volume	S1	setpoint	Bioreactor volume
DO2 setpoint	S2	setpoint	Process dissolved oxygen setpoint
Temp setpoint	S3	setpoint	Process temperature setpoint
pH setpoint	S4	setpoint	Process pH setpoint
agitation [rpm]	S5	setpoint	Process stirring speed setpoint
Tip speed	S6	derived	Calculated from impeller diameter and stirring speed (agitation)
P/V	S7	derived	Calculated from impeller power number, impeller diameter, stirring speed (agitation) and fluid density
pH post INOC	S8	online	pH measured at the time of inoculation
pH post TFX	S9	online	pH measured at the time of transfection
pH at HRV	S10	online	pH measured at the time of harvest
pCO2 post INOC	S11	offline	CO2 concentration measured at the time of inoculation
pCO2 post TFX	S12	offline	CO2 concentration measured at the time of transfection
pCO2 at HRV	S13	offline	CO2 concentration measured at the time of harvest
pO2 post INOC	S14	offline	O2 concentration measured at the time of inoculation
pO2 post TFX	S15	offline	O2 concentration measured at the time of transfection
pO2 at HRV	S16	offline	O2 concentration measured at the time of harvest
VCN post INOC	S17	offline	Viable cell number measured at the time of inoculation
VCN post TFX	S18	offline	Viable cell number measured at the time of transfection
VCN post HRV	S19	offline	Viable cell number measured at the time of harvest
Viability post INOC	S20	offline	Cell viability measured at the time of inoculation
Viability post TFX	S21	offline	Cell viability number measured at the time of transfection
Viability post HRV	S22	offline	Cell viability number measured at the time of harvest
FACS at HRV	S23	analysis	Infectious titre assay from first harvest sample
FACS at HRV48	S24	analysis	Infectious titre assay from second harvest sample
RNA at HRV	S25	analysis	Viral genome titre assay from first harvest sample
RNA at HRV48	S26	analysis	Viral genome titre assay from second harvest sample

Table 3: List of variables used in final statistical analysis of process development for suspension-based LVV production.
Time points refer to: INOC – Final inoculation; TFX – Transfection; IND – Sodium Butyrate addition; HRV – first harvest; HRV48 – second harvest as in the process flow diagram (Figure 7);
Types refer to: setpoint – process setpoint, determined by operator; derived – value calculated from other parameters; online – parameter measured as part of process monitoring, using bioreactor sensors; offline – parameters measured after sampling, using standalone equipment; analysis – variable obtained from an analytical assay after process is finished

2.3. Results

In order to gain a better understanding of the manufacturing platform processes developed by OXB, a statistical analysis was performed. It is important to realise there are significant differences in the way data is recorded for these two different processes. Batch manufacturing records provide a more structured and consistent format while many of the experimental data records were either incomplete or additional parameters were added on a batch-to-batch basis. The analysis approaches taken for these two different data sets were therefore challenging. The results highlighting the most prominent trends are presented in the sections below examining the effect of PCA and other statistical method based on correlation analysis.

2.3.1. Batch manufacturing records analysis

The executed BMRs from the adherent EIAV GMP production process were analysed using PCA. The main aim of the analysis was to check whether PCA could be applied to historical batch data to improve process knowledge and understanding of the process through the multivariate analysis of process parameters. The data was accumulated from multiple batches of multiple products and as such the data recorded is often inconsistent or missing multiple data points, however the LVV production process was consistent between products, as outlined in the methods section. Approximately half of the samples and a third of the variables were omitted from the analysis compared to the original dataset. A single missing confluency data point was estimated based on the historical data. The main quantitative parameters described in the records are the final vector titres and also the cell viability, confluency and total cell counts at different stages of cell expansion upstream process and during transient transfection. The records also include qualitative and descriptive records which are often related to in-process controls and in-process monitoring used to ensure batch-to-batch consistency. However, the majority of the qualitative and descriptive records had a uniform value across all batches as a negative value would be associated with deviation from expected results. As such there were multiple records which were unsuitable for multivariate analysis and were omitted in the final analysis. Overall, the final analysis was performed on a matrix of 15 batches, each characterised by 20 variables (Table 2). The data was normalised to account for numerical differences in the scale of the variables (percentages for confluency and cell viability, millions for cell count and titre) and allow for direct comparison of variables based on their standard deviation within variable. The resulting dataset is presented in Figure 8.

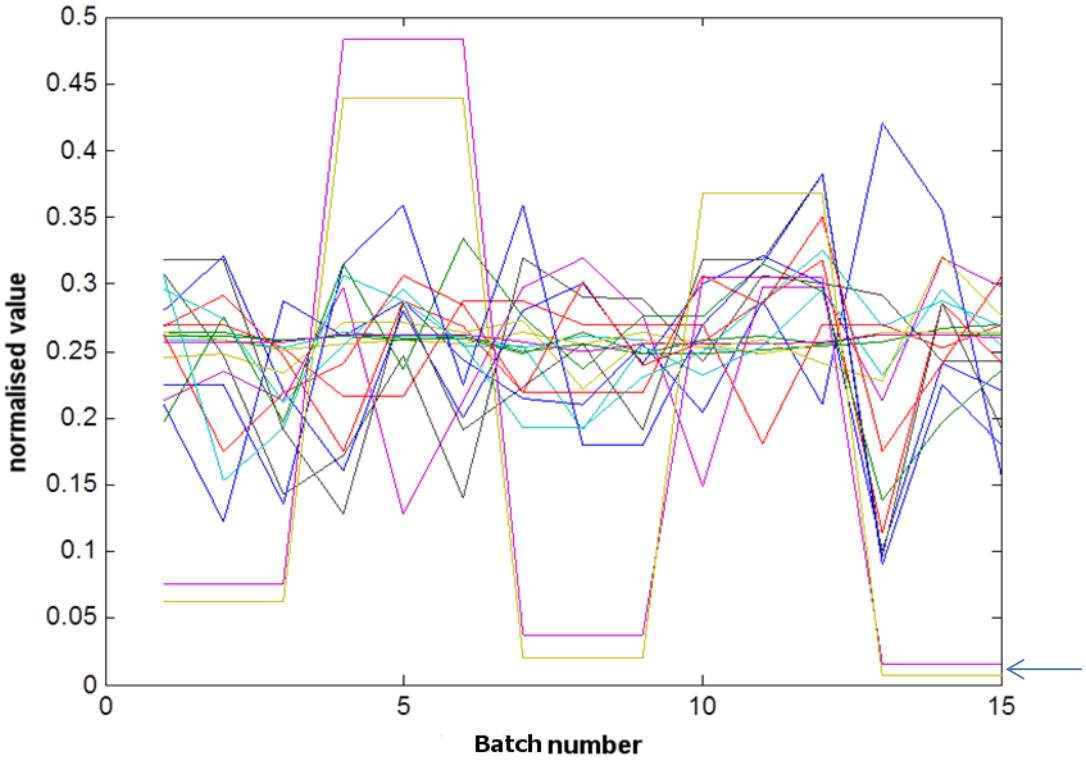


Figure 8: Diagram of 20 variables (normalised values) from 15 manufacturing batches.
X axis represents 15 manufacturing batches; y axis shows normalised values of all variables. Arrow indicates variables 19 (RNA copy number) and 20 (FACS infectious titre)

Initially, the data was mean centred and then followed by PCA which resulted in 67.81% of the variability captured in PC1 and 14.83% of variability captured in PC2. Examination of the loadings plot revealed that PC1 score is mostly impacted by variables 19 and 20 i.e. the final titre measured by FACS and RNA copy number assays (Figure 9). This indicates that the titre variation has a major impact on the overall data structure which would be expected given its high variation as observed in Figure 8.

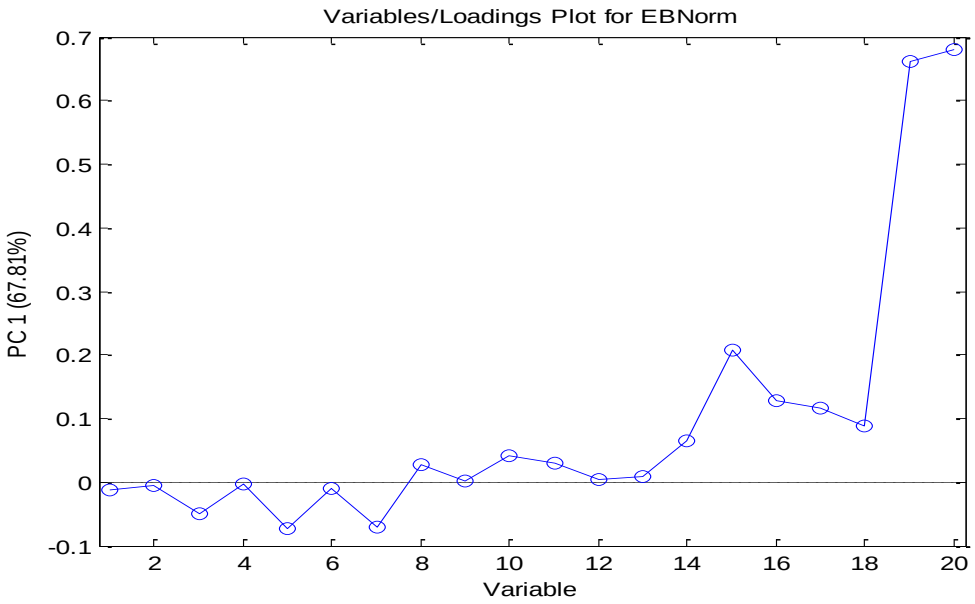


Figure 9: PCA loadings plot of PC1 for the BMR data comprised of 20 variables and 15 batches.

PC2 is impacted by multiple different variables (Figure 10)

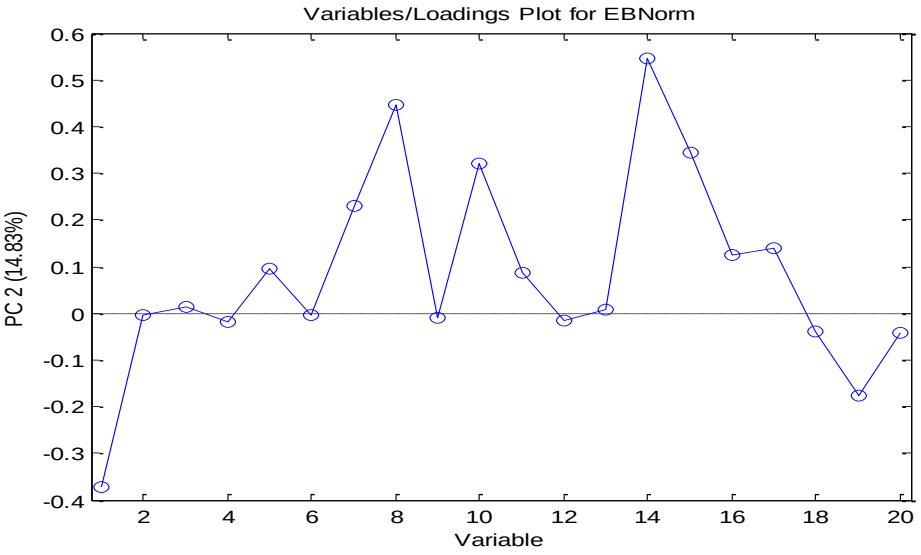


Figure 10: PCA loadings plot for PC2 for the BMR data comprised of 20 variables and 15 batches.

Inspection of the scores plot (Figure 11) showed that data points with high PC1 scores relate to high final titre achieved by all the batches of product 2 (samples 4-6 and 10-12) while the other reached lower titres and PC1 scores. The factors impacting PC2 are less uniform, however the loading plot suggests that variables 1,8,14 and 15 have the highest impact (Respectively: Total viable cell count on day 1 and cell confluency on days 8, 12 and 13). Samples 3, 4, 6 and especially 13 are all characterised by negative PC2 score, high Day 1 cell count and mostly low confluency on days 8, 12 and 13, forming a group of batches which could be described as having lower than average performance.

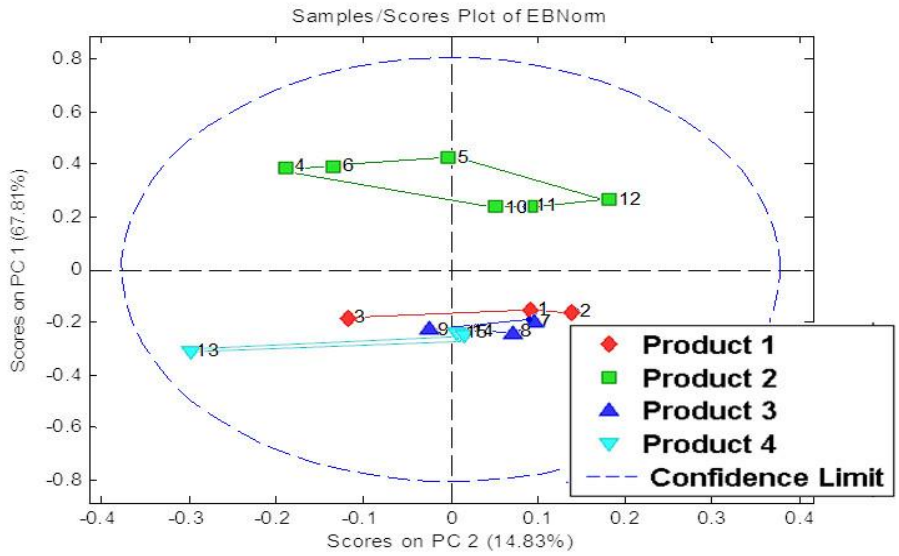


Figure 11: PCA scores plot for PC1 and PC2 of the BMR data comprised of 20 variables and 15 batches.

The major impact of vector titre value may be masking the effect of other variables and therefore a second round of analysis was performed without titre data, using 15 batches characterised by 18 variables (variables 19 and 20 i.e. RNA and FACS titre were excluded, see Table 2). The contribution of individual parameters was more significant and varied compared to the initial analysis dominated by the vector titre values. This approach revealed data structure where PC 1 and PC2 score values correlated with the product type, indicating that the cells behave differently when transfected with different transgenes (Figure 12). However, most of the samples remain close to the centre of the plot except for sample 13 which was already identified as a potential outlier in the previous analysis. Samples from product 1 and 3 significantly overlap, indicating higher correlation between batches using these two products compared to product 2 and 4.

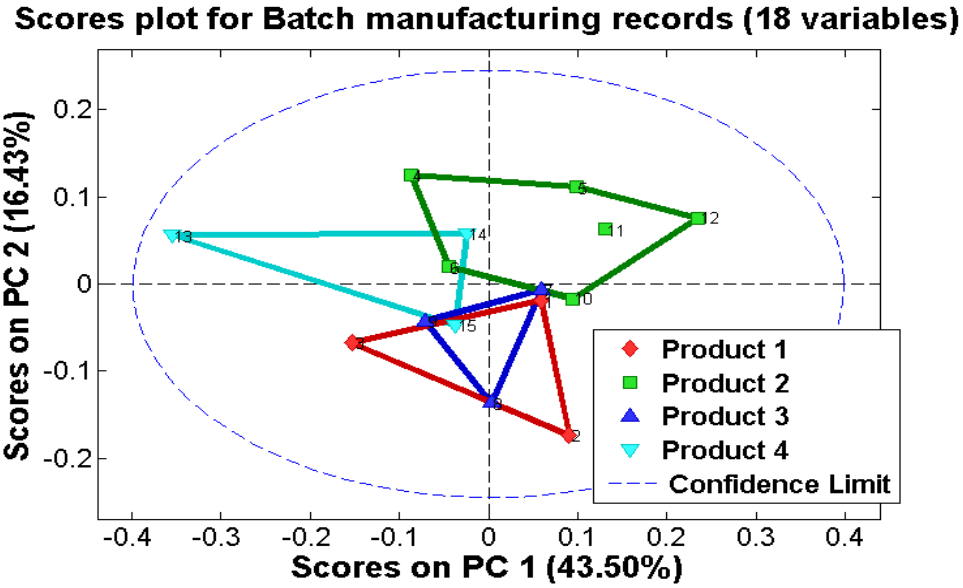


Figure 12: PCA scores plot (PC1 and PC2) of the trimmed BMR data using 18 variables from 15 batches.

Overall, the results have shown that PCA can be used to visualise batch-to-batch variation. However, the major impact of the final product titre on the analysis and the fact that the titre is affected by the product type suggest that in order to reliably assess batch-to-batch variation only batches of a single product should be assessed in future with a separate model established for each product. LVV titration assay variation can often be a consideration when analysing such data, however in case of BMRs the assays were performed to GMP standard set by OXB. The observed CV between biological replicates was below 15%, a significantly lower variation than observed between batches of different products and due to process variation.

The PCA method described above could be used to establish desired conditions (golden batch approach) and monitor subsequent batches for deviation in multivariate space. The combination of process variables recorded in the current BMRs can be used to identify potentially less optimal batches, mostly based on cell confluency and viability values which were associated with low PC2 score in Figure 11 and lower titre compared to other batches of the same product. However, it is unlikely that PCA can be used to differentiate between batches of different products using the current set of variables maintained in the BMRs unless the product has a major impact on the final titre (as in the case of product 2). The main benefit may come from identifying batches with sub-optimal cell properties with either low viability or confluency throughout the process as well as showing variation of multiple variables from standard values. The main concern would be that the correlation between final titre and cell count, confluency and viability is often inconsistent and therefore while it may be possible to identify outlier batches in terms of cell performance, how it relates to vector production may not be so easy to distinguish.

A further drawback in this analytical approach is the fact that the variables recorded in the BMRs are in-process measurements such as cell viability and confluency which can be monitored to ensure process consistency but cannot be directly controlled by an operator. This issue is inherent to adherent process where process parameters are controlled indirectly (e.g. pH is controlled by CO_2 concentration in the incubator and medium buffer composition). PCA can still be used as a visualisation tool of batch performance (based on critical quality attributes) and to demonstrate compliance to continuous improvement requirements based on in-process measurements. However, to gain improved process understanding, further process inputs (not captured in current BMRs) should also be included in the analysis. To improve process knowledge the inputs could be subjected to PCA and process critical quality

attributes could be used to categorise the data, providing further information on interaction of different control parameters and their combined effect on titre. Suspension or perfusion-based processes would be better suited for such an approach due to more direct control and enhanced monitoring offered by these systems.

2.3.2. Suspension

The data available for suspension batches at the time of writing was only available for development batches where records vary significantly between multiple experiments due to changing experimental set-up. There are parameters which have been recorded more diligently while others are underrepresented, limiting the ability to analyse some of the process parameters and performance indicators. The most prevalent parameters recorded throughout suspension process development were cell count, cell viability and pH at different time points. The levels of these parameters over time were inspected for all the batches (Figure 13-Figure 15).

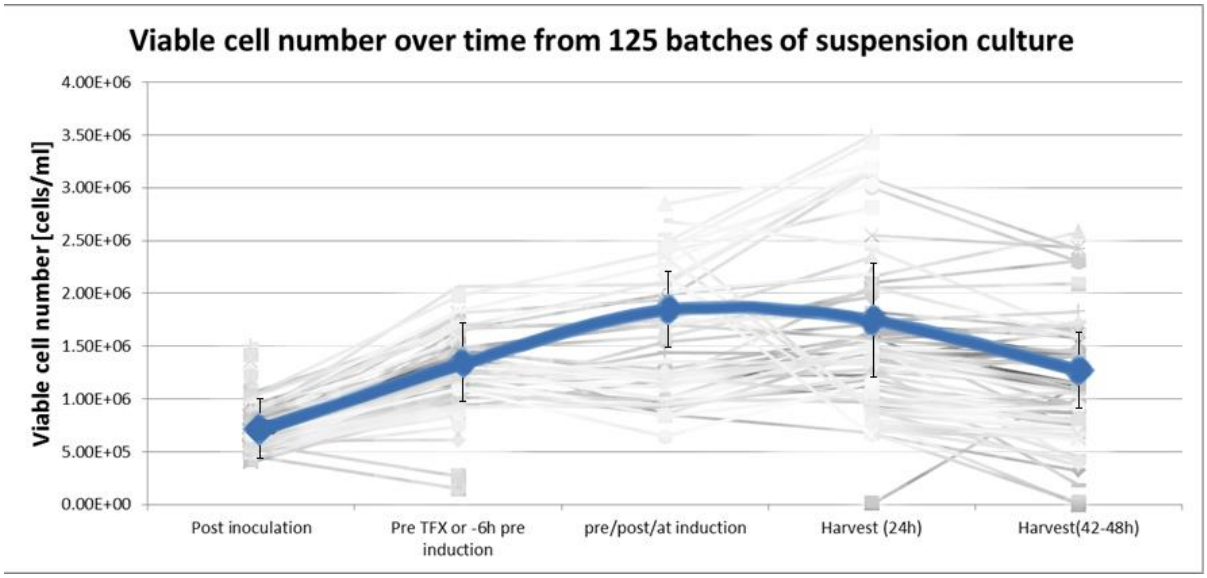


Figure 13: Viable cell number over time from 125 batches of suspension-based LVV production process. The average viable cell number at each time point is drawn as a thick blue line.

This simple analysis has shown that on average the viable cell number initially increases over the first 48 hours of the process until induction. For the next 24 to 48 hours the cell count starts to decrease. There are however exceptions from this trend where individual batches show a decrease in cell count after transfection or a continued increase in cell number up until harvest.

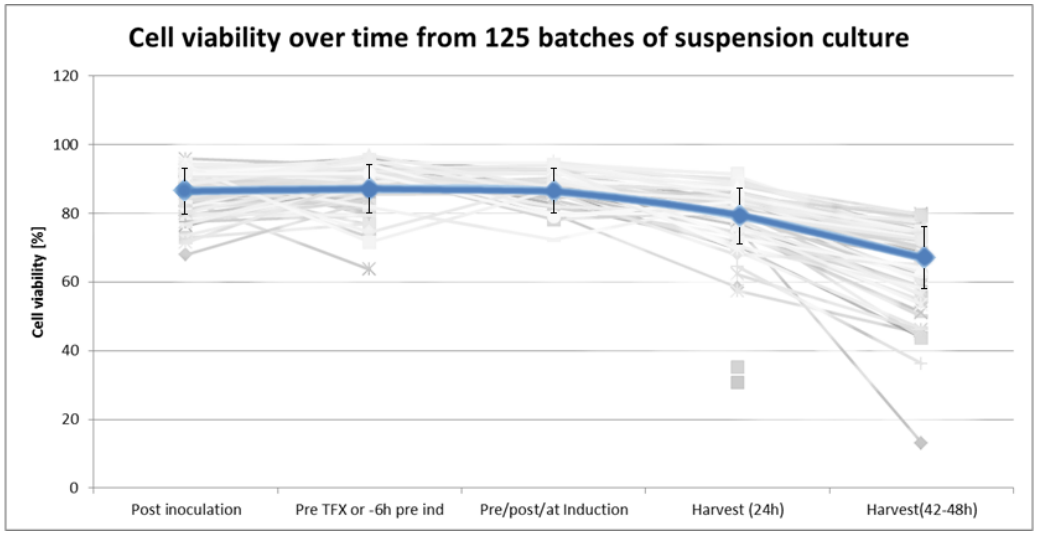


Figure 14: Cell viability over time from 125 batches of suspension-based LVV production process. The average viability at each time point is drawn as a thick blue line.

The changes in viable cell number are related to the cell viability in culture outlined in Figure 14. Initially the viability remains stable at a high level of about 80-90% viability with individual batches showing an increase or decrease in viability over time with no significant change in the average viability. However after induction (about 48h after inoculation) cell viability decreases and reaches a significantly lower level at harvest and especially the later harvest (some of the batches were harvested earlier or later than others during process development) where many batches reach cell viabilities below 60%.

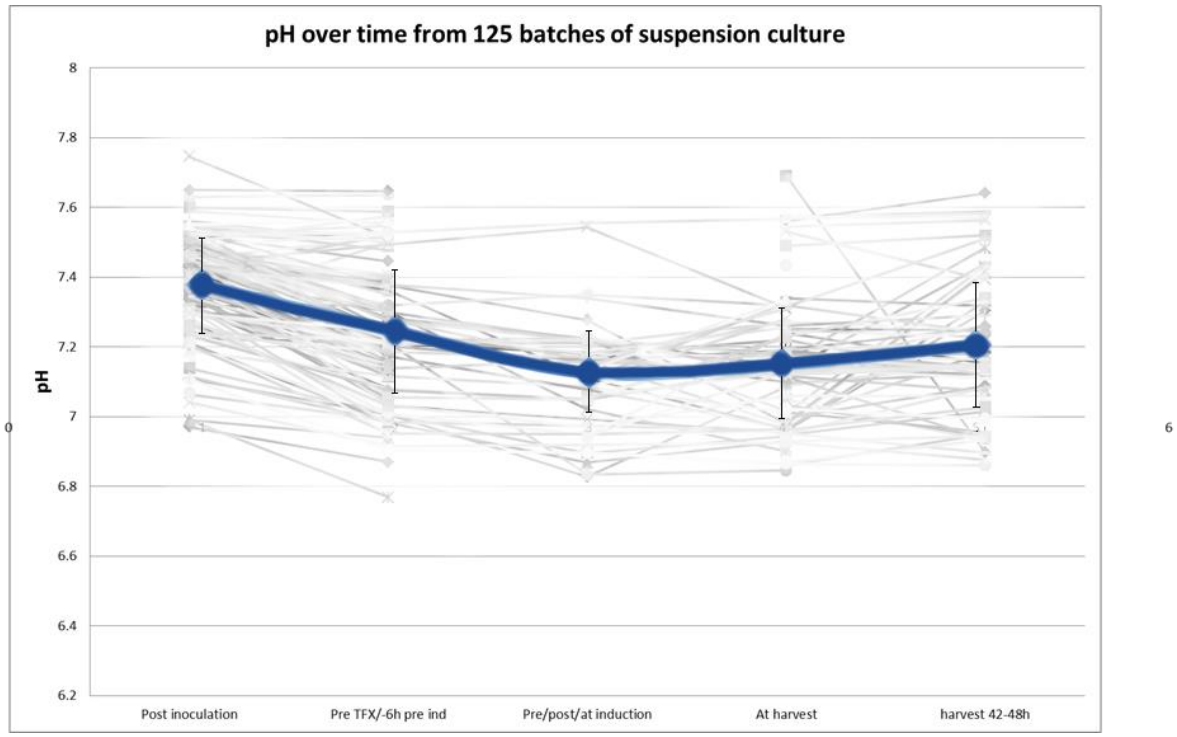


Figure 15: pH over time from 125 batches of suspension-based LVV production process. The average pH at each time point is drawn as a thick blue line.

As shown on Figure 15, the overall pH level is buffered and controlled above 7 due to lentiviral vector instability at low pH. pH level has shown moderate variation in the rate of change with the majority of the samples either remaining stable or decreasing for the first 24h. After 48 hours there is a significant decrease in pH which is followed by a reverse trend of pH increasing for the next 24 and 48 hours until harvest. pH level range between experiments varies between 6.8 to 7.6 due to changing set points and large controller dead band to avoid excessive addition of NaOH and CO_2

A further analysis was undertaken to eliminate any missing entries in the data set. This included two batches with multiple missing records and 29 variables with records missing from over 70% of the batches. These underrepresented variables included the metabolites data (glucose, lactate, glutamine, glutamate) and multiple parameters at induction (48h) and second or late harvest (48h after induction or 96h in the process). These records were missing because appropriate analysis was not performed on all the data or a particular step was omitted in the process as in the case of second harvest which was not performed for the majority of experiments due to low vector recoveries compared to the initial harvest.

A 26-parameter correlation table has been constructed to inspect the relationship between different process parameters and the trends in the process. A number of easily explainable correlations was found (Figure 16): a relationship between working volume, agitation rate, tip speed and P/V which are all related to culture volume and aeration rate; both pH and dO_2 had a correlation between their set points and levels of a corresponding parameter at different time points which is expected as a consequence of control of these parameters throughout the upstream process. PCA was performed and a loadings plot generated for the first two PCs to assess the relationships between variables and their impact on the data structure (Figure 16). Overall variability captured was relatively low (29.9%) indicating that the iterative analysis with different set of batches (e.g. from single product or single scale) and variables (omitting the least significant ones) could improve the quality of the results. Besides PCA a correlation matrix was used to inspect correlation coefficients of individual pairs of parameters which were identified as potentially significant across the PCA.

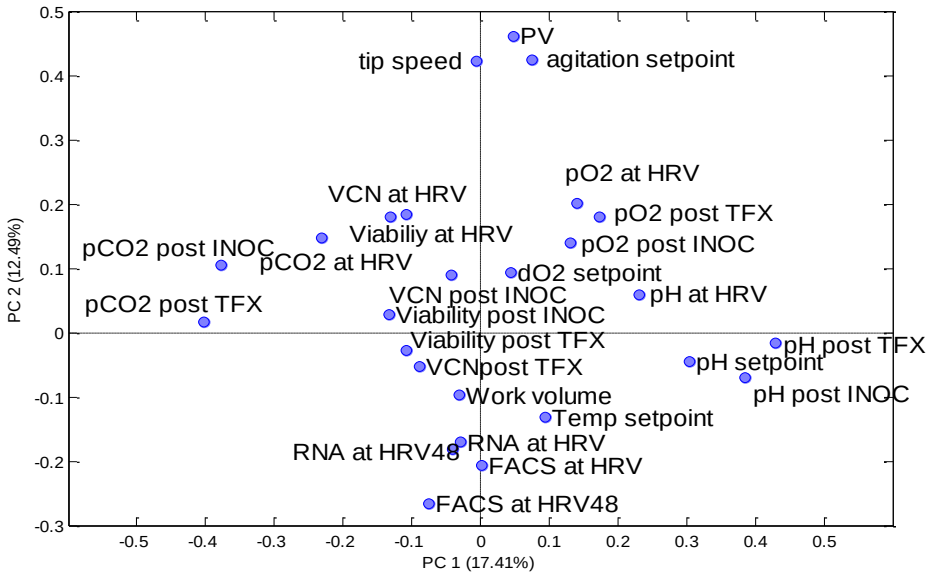


Figure 16: PCA loadings plot (PC1 and PC2) for suspension-based LVV production process parameters. Variables at the opposing ends of the scale have an inverse effect in determining the PCA score of analysed samples suggesting a negative correlation.

As shown in Figure 16, the PCA loadings plot can be divided into four major subsections, two for each PC. PC1 (capturing 17.41% of overall variance) is focused around pCO_2 and pH as well as several parameters which are affected by these parameters. The inverse relationship between pCO_2 and pH is easily explained by the fact that pH is controlled by CO_2 sparging and therefore lower pH is associated with a higher concentration of the gas in solution. Similarly, the positive correlation between pH and aeration where pH and pO_2 is characterised by a positive loading value can be explained by air sparging displacing CO_2 which affects pH. The correlation analysis of these parameters has shown similar trends where a significant ($p<0.05$) positive correlation between pH and pO_2 is observed especially after transfection and at harvest. A negative correlation was observed between pO_2 and pCO_2 after transfection.

PC2 (12.49% variance captured) is dominated by the inverse relationship between stirring rate related parameters (agitation rate set point, tip speed, power per unit volume (P/V)) and overall process performance as indicated by the quality attributes RNA copy number and FACS titre. This is further supported by a highly significant ($p<0.01$) negative correlation between all stirring parameters and FACS titre. This could indicate that a high P/V may reduce effective viral vector titre due to shear forces within the bioreactor causing

damage to either the cells (reducing their ability to produce vector) or the vector itself (reducing the effective viral vector titre).

Analysis of the correlation matrix has highlighted several other interesting trends. For example, pH varies throughout the process as indicated in the initial analysis (Figure 15). It correlates with cell count and viability ($p<0.05$) at several time points in the process. In this analysis lower pH after transfection and at harvest was correlated only to FACS vector titre obtained from late or second harvest ($p<0.05$). However, there was no significant correlation with early harvest titres or RNA copy number. Whilst changes in pH may be attributed to overall cell culture performance, there are also indications that it may also be directly used to increase productivity.

The effect of cell count and cell viability on the final titre is not straightforward. A positive correlation at inoculation and transfection but negative after induction was observed, suggesting that initial high cell density may improve productivity but high density towards the end of the process could be detrimental. These results align with the initial analysis which showed an initial increase in viable cell number followed by a decrease after induction.

2.4. Discussion

Manufacturing of lentiviral vectors is a complex and costly process which justifies extensive monitoring and control as well as continuous improvements to the production methods. Statistical analysis of the process can be used to gain a better process understanding, highlight trends in the process data and focus areas for further improvements. The results of PCA and correlation analysis were used to identify key properties of the current LVV production processes in manufacturing and process development settings. The discussion below summarises the results and outlines its potential benefits and future direction.

2.4.1. Improvements in BMR analysis

Analysis of the current BMRs has demonstrated that MVDA can be used to assess the batch-to-batch variability based on all the available process data. Through PCA it is possible to identify outlier batches which can be further inspected and analysed to maintain consistency in the process. However, in this report the examined data set was dominated by in process measurements such as cell confluency and cell viability as well as viral vector titre (Table 2). These variables are measured to monitor the process and cannot be directly controlled; therefore, performing an analysis based on these data sets is of limited value. Analysis of parameters used in process control such as temperature, pH, concentration of dissolved gases and nutrients could provide more relevant and reliable information on vector production. Knowing the behaviour of the production system over time and the effect of individual process parameters on product quality and yield would greatly improve the benefits of PCA performed on BMRs (Figure 11 and Figure 12).

The use of PCA for the adherent process is hampered by the limited process parameters which are directly controlled. The BMRs contain information about variables that can be used to describe overall conditions of the cells but do not record parameters such as incubator temperature and CO_2 levels. The problem is further complicated by the fact that monitoring incubator conditions does not directly reflect the cell culture conditions which are difficult to measure due to the limited ability to implement on-line sensors in adherent cell factories. Consequently, even the parameters such as incubator CO_2 concentration, humidity or temperature that can be controlled, do not directly reflect the state of the cell culture in individual vessels. Moreover, the data presented in the executed BMRs can be highly variable as shown in the data for cell confluency, viability and cell number which can vary between operators. Furthermore, measurements are based on subjective operator observations from a

single layer of the cell factories which may not be representative of the entire culture vessel. Additional monitoring and records of potential cell performance indicators such as metabolite profile over time, pH and more detailed cell density and viability data could improve future MVDA by providing mode detailed and relevant data sets. However, this would require changes to the established production process carried out by OXB.

The other issue when applying MVDA to BMR analysis and monitoring is the dominant effect of product type and viral vector titre. Titre is an important indicator of process productivity; however, it can overshadow the over variables as it can be highly variable for different products. Therefore, depending on the goal of analysis, it may be beneficial to omit titre data in the analysis to focus more on other indicators of process performance. Understanding the source of titre variability is also an important factor: while the assay variation of the data used in this work was relatively low (CV <15% due to high rigour of the analysis as described in methods section 2.2.1), often high assay variation of LVV titration could affect the data. Ideally, only batches of the same product should be compared to limit the effect of genome on the process performance. Alternatively, titre data for each product could be normalised to the highest titre value to reduce the impact of titre variation between products. MVDA could be applied to either monitoring process output, in which case titre data may be an important variable, or analysis of process parameters and their effect on cell performance, in which case titre data may dominate other parameters and limit the insight gained from the analysis. A further step in analysis could be use of partial least squares analysis (described in more detail in Chapter 4) to model the correlation between process parameters (X-block) and LVV titre (Y-block)

Despite the limited ability to directly monitor and record adherent cell culture parameters, PCA can be used as a tool for assessing batch-to-batch product quality based on in process measurements. As demonstrated in Figure 12 the PC scores plot can be used to visualise and cluster data to identify potential outliers. Combined with examination of the loadings plot it is certainly possible to use this type of analysis to help in investigating deviations from the standard operations which could affect product quality, thereby finding root cause to improve understanding and reduce batch-to-batch variation. This approach could greatly benefit from establishing a data base of reference batches which could be analysed and used as a comparison for future manufacturing runs. Application of PCA would demonstrate ongoing improvement to process monitoring and understanding, satisfying QbD guidelines

2.4.2. Suspension

Adherent cell culture is limited in terms of process monitoring and control which presents an opportunity to examine the suspension-based process which is easier to control and provides more information. The robust monitoring system allows recording and easy access to experimental data generated by process research and development. Currently there are no BMRs available for the suspension process due to ongoing use of the adherent process in manufacturing. However, there is an abundance of records for the suspension process from development batches obtained from multiple varying experiments. While the research data is less consistent due to different conditions used in the experiments, it can still be useful for analysis of cell and vector properties due to the large amount of data.

The combination of PCA and correlation analysis highlighted several trends in the data obtained from research batches. Increasing viral vector titre is an important objective of the process development studies as well as this statistical analysis. There are several correlations which could guide process development decisions such as the effect of pH and P/V on the final titre which show benefits of increased pH control and conservative stirring rate. However, before committing to process changes it would be beneficial to perform studies dedicated to examination of these particular interactions. The data set used in the statistical analysis has limited reliability due to missing data points and the varying goals of multiple studies which were pooled together for the analysis.

The data quality used in the PCA analysis in this chapter is limited due to inconsistency of the process itself. The research data set consists of multiple variants of the process where setpoints, control schemes as well as monitoring and recording of data varies significantly between experiments. The initial analysis of the entire data set points to correlation between certain parameters (e.g. agitation rate and titre) but also highlights limits of the analysis due to high variation between process types used in research. Dividing the data sets further e.g. based on working volume, process duration or other fundamental ways of categorising process variants could improve the quality of analysis. In cases where there are not enough experiments to support splitting data sets, it could be beneficial to perform additional experiments to fill the gaps. In the current form the variation in process categories used in the analysis could be limiting the quality of the model. Nevertheless, the analysis points to several correlations and interactions between process parameters that provide insight into viral vector production

Understanding the underlying mechanisms of viral vector production in cells is an important part of process development. The statistical analysis highlights the behaviour of cell culture throughout the process with the initial cell expansion indicated by stable cell viability, increasing viable cell number and a decrease in pH. The pH decrease could be attributed to increased cell growth and metabolic rate leading to reduction of glucose to lactate and related by-products. This in turn may result in lowering the pH due to the acidic character of these metabolites. However, it has been suggested that maintaining lower pH during cell culture can increase viral vector infectivity and effective titre (Holic et al., 2014). Increased control and change in pH over the duration of the process could be beneficial but a dedicated study to determine optimal pH range and extent of control deadband would be recommended.

Viral vector production begins after transfection of cells which hinders cell growth and leads to lowering of cell viability and cell count due to shift in cell metabolism and the cytotoxic effect of the viral vector. The major cause of the cytotoxicity is the VSV-G protein as well as HIV protease and Vpr protein which were described as one of the major obstacles in development of stable packaging cell lines (Kafri et al., 1999). Higher viral vector production can lead to increased cell toxicity and death and change in cell count/viability over time may be a useful indicator of process performance. However, cell death due to undesirable process conditions (loss of control, nutrient depletion, contamination) could also lead to a decrease in cell viability and number, making this method potentially unreliable as a process performance predictor. The viral vectors are harvested at pre-determined time points after transfection and a certain level of viral vector related cell death is expected. It can be used as an indicator of process performance as long as other possible causes of cell death are accounted for.

The research data analysis can prove useful when the suspension process is introduced to manufacturing. Being able to correlate cell performance and vector quality with the critical process parameters profiles from research runs would lead to improved operation of the manufacturing process. Better understanding of the process would lead to an expanded design space and tighter control allowing reduced variability between batches and more consistent production cycle. More consistency means that a desired total titre can be reached with less raw materials or in shorter time, therefore improving the overall process efficiency in accordance with the QbD principles. As the lentiviral vector-based therapies progress into clinical application there will be an increased need for demonstrable understanding and

control of the manufacturing process and use of PCA and other statistical methods could prove essential.

2.5. Conclusions

Oxford BioMedica has the capability for bioprocess development and scale-up for clinical and commercial supply. The statistical analysis reported was able to compare data from both pilot and production scale batches. The BMRs hold consistent records of variables monitored during production runs, although some entries were not available at the time of writing. They are dominated by in process measurements which limit the potential to improve process understanding through MVDA but they present an opportunity for use in monitoring and analysis of the process to identify deviations and outliers. The suspension process development records offered a more varied insight into several process parameters and therefore they were used to gain further insight into the process under development. These two data sets required different approaches to extract useful information about LVV production but they both benefited from statistical analysis. MVDA has not been previously performed on OXB records and the early results presented in this chapter demonstrate that useful information can be extracted both for process monitoring and development. However, the analysis also highlights shortcomings of the data and analysis which can be improved upon in the future.

One of the major objectives of this study was to examine results obtained through PCA and assess usefulness of this method in process characterisation and batch-to-batch variation monitoring. PCA presents an opportunity to reduce data complexity and examine relationships between variables explained by data structure. PC scores can be used for clustering of data points such as individual batches. It can be used as a graphical representation of data structure which can make it easier to examine a process characterised by multiple variables. Combined with BMRs this method can be used to monitor batch consistency (golden batch approach). PC loadings can be used to examine the effect of individual parameters on the analysis as well as their relationship through the effect each variable has in determining the scores of each PC. This can provide insight into combinatorial effect of multiple variables which can be used in improving process understanding and guiding process development as observed in the suspension data.

Overall, the statistical analysis applied to the manufacturing and process development batches demonstrated that this methodology can be used to examine trends in the data and identify relationships of process parameters which can then be used to guide better understanding and highlight areas for improvement. PCA could be applied as a routine batch monitoring tool and with a carefully selected set of reference batches could be used to track

process variation between batches. With additional comparison and expansion of currently monitored parameters, product quality could benefit from the introduction of this type of PCA-based analysis especially in suspension-based process.

The current analysis could be refined by modelling smaller data subsets, grouped by product type, cell line and production vessel scale. PCA highlighted that the data forms clusters based on product type. It is feasible to inspect a simpler system on a product-by-product basis in more detail, which may further help to characterise the manufacturing process. Similarly, for suspension samples, aligning data points produced in the same type and size of vessel would eliminate some of the process variation in the data and would reduce complexity of the analysed system. The adjustments can be further investigated but were outside of the scope of this project where focus was placed on assessing the feasibility of applying broad MVDA methods to process data.

Chapter 2 used process data to assess feasibility of applying MVDA in manufacturing and process development setting. In particular, PCA was used to analyse data trends of process parameters captured in BMRs and process development documents. While several areas were identified as potentially relevant for further research (data clustering based on product type, correlation of process parameters) the diversity of samples and processes used resulted in PCA models which would require further refinement, especially if these method were to be introduced as part of manufacturing or process development workflow. The next chapter further utilises MVDA methods to inspect data trends in cell and viral vectors samples generated specifically for this EngD project and analysed using mass spectroscopy.

Chapter 3 The use of MALDI-ToF mass spectrometry in analysis of HEK293T cells and lentiviral vectors

3.1. Introduction

One of the major goals of this EngD project with OXB was to improve process understanding of LVV production. This chapter outlines how this goal is achieved through characterisation of HEK293T cells and LVVs. The main methodology employed here was a combination of MALDI-ToF MS and PCA. MALDI-ToF MS was optimised both for HEK293T cells and LVV samples. PCA was used as the main method of transforming mass spectra and enabling graphical representation of the data. The analysis was performed on a variety of cell and vector samples generated in a range of conditions. This included varying cell culture type (adherent or suspension adapted), vessel size, process conditions, viral vector type (EIAV or HIV) and downstream processing of the viral vector. The data is used to differentiate between cell and vector types, assess the impact of varying process conditions and establish a method for analysis of process and product consistency.

3.1.1. Mass spectrometry overview

Mass spectrometry (MS) is an analytical method based on generation and measurement of charged molecules which are differentiated based on their mass-to-charge ratio, typically visualised as a graph of signal intensity over a spectrum of mass-to-charge ratio values. MS can focus on different aspects of analysed molecule to identify individual elements of a mixture, generate a unique mass spectrum fingerprint to differentiate between molecules or mixtures, or to identify specific properties of a molecule (e.g. purity, interaction with other molecules). Two most common MS methods used in the field of biotechnology are MALDI-ToF MS (utilised in this EngD project and described in detail in the following section 3.1.2) and LC/MS. LC/MS (also used in tandem setup as LC/MS-MS) is the main alternative to MALDI-ToF used for biological molecules. The molecules (or mixture) undergo initial separation of molecules through a HPLC column. Different columns can be used based on the mixture and desired separation effect. The analyte is then ionised, typically through electrospray ionisation (ESI) where molecules are dispersed, charged and transitioned into gas phase before entering mass spectrometer. Both in single and tandem setup, quadrupoles (four metal rods set in parallel, charged with variable voltage) are used to modulate the travel of charged molecules and enable analysis of molecules of specific mass to charge ratios. LC/MS based methods allow high sensitivity of analysis and are well suited

for a range of chemical and biological molecules, especially for analysis of individual molecules (Pitt, 2009; Korfmacher, 2007).

The primary application of MS in the field of viral vector research and production was proteomic analysis. Segura et al (2008) analysed retroviral vector associated host cell proteins using LC/MS. Viral vectors were extensively purified using a combination of ultracentrifugation and chromatography methods. The viral vectors were then subjected to subtilisin digestion and individual proteins were separated using SDS-PAGE method. The individual proteins were further fragmented through tryptic digestion. The resulting peptides were analysed through LC/MS method (using combination of HPLC and electrospray quadrupole MS).

Similar approach was employed for quantitative analysis of LVV proteins (Denard et al, 2009). LVV were purified through a combination of TFF chromatography and ultracentrifugation methods. Viral proteins were extracted using Proteinase K which was followed by separation and densitometry-based quantitative analysis on 2D gel (based on protein size and isoelectric point). The individual proteins were fragmented through tryptic digestion. The resulting peptides were analysed using MALDI-ToF followed by bioinformatic analysis to identify individual proteins. The authors identified 10 co-purified protein, 18 protein incorporated into virion and 6 viral proteins. Another study identified even larger range of LVV-associated proteins using LC/MS method (Wheeler et al, 2007)

The MS analysis of viral vectors has primarily focused on proteomic analysis of highly purified vectors where individual proteins were separated and fragmented before analysis. The analysis focuses on protein identity and properties, using mass spectrometry as one of the analytical tools rather than the subject of study (e.g. no mass spectra are presented in publications described above, instead focusing on protein gel images and protein sequence). A wide range of identified proteins highlights complexity of viral vectors and the effect of methods used in their production, processing, and analysis. With a variety of available methods, MALDI-ToF was selected based on its suitability for whole cell and intact vector analysis which was the focus of the research described in Chapter 3 and modelling described in Chapter 4.

3.1.2. MALDI-ToF mass spectrometry

MALDI-ToF MS has been used in multiple fields of science for analysis of a variety of molecules. One of the major applications includes protein analysis using either fragmented or intact proteins. Proteins can be identified by applying a combination of gel or chromatography separation, tryptic digestion and peptide mass fingerprinting (Fenselau, 1997). Additionally whole protein mass spectra can be used for protein analysis and proteomic studies (Liu & Schey, 2005a). MALDI is a soft ionisation method, which means it can be used for more fragile and difficult to manage molecules compared to other MS methods such as electron ionisation (M. Karas & Bahr, 1990; Knochenmuss, 2006). The principle of action (Figure 17) is that the sample suspended in an ionisation matrix and spotted on a plate is irradiated with a laser which leads to desorption of the matrix and the sample, transfer of electrons and ionisation of the sample which is required for a successful MS measurement. The charged sample molecule is vaporised and accelerated in an electric field and its time of flight is measured to estimate the sample's mass to charge ratio (m/z) based on its travel time (Vestal, 2009). Most ions generated by MALDI are single charged and therefore m/z can be used to estimate the mass of the molecule. It is one of the advantages of MALDI when compared to other common MS methods such as electrospray ionisation which are often used to generate multiply charged ions (Loo et al., 1992) which complicate the analysis of the resulting spectrum compared to singly charged ions generated through MALDI-ToF MS. Several parameters of MALDI-ToF can be adjusted to achieve high quality mass spectrum. Sample preparation has a large impact on the spectrum quality and reproducibility. Varying the sample concentration, matrix type and matrix composition will affect the final spectrum and therefore needs to be optimised and controlled for each assay (AlMasoud et al., 2014; Liu & Schey, 2005b; Schaiberger & Moss, 2008). The laser intensity and voltage should be adjusted based on matrix and sample composition to minimise generation of multiply charged ions (Williams et al., 2003). A common improvement over typical MALDI-ToF MS is the use of a reflectron. This device generates an electrical field capable of reversing the path of an ion. This technique extends the distance travelled by the ion and reduces the variability of time of flight for molecules with the same m/z, therefore increasing the resolution of the resulting mass spectrum (Cornish & Cotter, 1993). MALDI-ToF MS can be used to analyse protein complexes such as antibodies (Bodnar et al., 2015) or whole cells (Koubek et al., 2012; Povey et al., 2014) where the sample is ionised and its individual components are detected. This results in a complex mass spectrum which can be used as a fingerprint of the sample. Depending on the complexity of

the sample, individual proteins can be separated using methods such as gel electrophoresis and analysed individually through MALDI-ToF MS or the whole mass spectra can be used for fingerprint analysis. By using a tandem setting with high energy fragmentation of amino acids MALDI-ToF/ToF MS can also be used for protein sequencing (Gogichaeva et al., 2007).

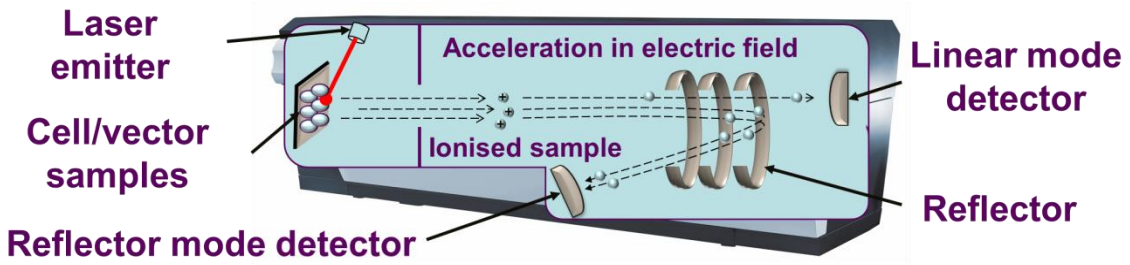


Figure 17: Schematic of the MALDI-ToF MS mode of action.
Sample suspended in the matrix is ionised by a laser, accelerated in an electric field and subsequently detected. By measuring its time of flight it is possible to estimate the mass to charge ratio (m/z) of the molecules.

The methodology used in this project was whole cell and vector analysis which utilises cell or viral vector samples suspended in ionisation matrix. Incubation in the matrix solution lyses the cells and upon laser desorption/ionisation the proteins are ionised (with varying efficiency depending on individual protein properties, which is one of the factors affecting signal intensity). The resulting mass spectrum includes signals from multiple proteins present in the cell. While such a spectrum cannot be used to identify individual proteins, the fingerprint can be subjected to MVDA and compared against other samples or a data base. This technique has been demonstrated in bacterial (Seng et al., 2009) and mammalian (Feng et al., 2010) cells and used in process characterisation (Momo et al., 2013), cell line development (Povey et al., 2014) and clinical diagnosis (Tudó et al., 2015), demonstrating how powerful and versatile this technique can be.

3.1.3. Principal component analysis

The study discussed in this report covers MALDI-ToF MS analysis of multiple cell and vector samples using PCA. It is an exploratory MVDA method that transforms high dimensional data into a smaller number of principal components (PCs) characterised by scores and loadings values (Abdi & Williams, 2010). This results in a graphical representation of data structure and allows analysis of complex multivariate interactions between the variables. For each PC, loadings values determine weight of each original variable that is used to generate the PC score value. In case of MS the variables are values of mass to charge ratio (m/z). For each sample, the PC scores represent the value corresponding to a transformed spectrum. In the case of mass spectra, PCA transforms the whole mass

spectrum of each sample into a single data point in principal component space which can be placed on a PC scores plot to compare it to other samples and assess its characteristics.

3.1.4. Cells for LVV production

Cells used in the EngD project are HEK293T cells, which are robust in their ability to maintain a healthy population in adherent and (once adapted) suspension culture. They are easy to transfect and often used for viral vector production. They also contain a SV40 Large T-antigen which differentiates them from the original HEK293 cells (Gama-Norton et al., 2011). The experiments described in this work made use of multiple HEK293T cell lines, developed at Oxford BioMedica or as part of this project. While HEK293T cells are typically grown in adherent culture, many cell lines derived from HEK293 and HEK293T were adapted to suspension growth in serum-free medium through progressive media exchange and cell selection. Use of serum-free medium improves consistency of the process as batch-to-batch variation of animal-derived serum is eliminated. Serum can have a significant effect on cell analysis. While the serum proteins are washed away during sample preparation, their varying composition can affect cell growth as well as the transfection process during LVV production, resulting in a lasting variation within cell protein composition.

Transiently transfected cells initially do not contain any of the viral genes and all plasmids are introduced into the cell as a part of vector production once a high enough cell density has been achieved. This can result in high variability of the LVV production process and it involves a transfection step which adds complexity to the process and can be expensive depending on the reagents used. Transient transfection has been demonstrated to achieve high titres and is commonly used in LVV production. However, transient transfection leads to increased variability in vector titre and quality and higher process costs due to costs of goods involved in transfection, specifically DNA and transfection reagents (Merten et al., 2014b). To improve process consistency and eliminate the need for transfection, stable producer cell lines (PrCLs) are transfected with the required viral vector and genome DNA (such as GFP reporter if used for research or a therapeutic transgene for clinical application) during cell line development. The cells then follow the development cycle outlined above which means that once a high performance PrCL is selected it can be used for vector production without the need for transfection which should reduce process variability (Stewart et al., 2011). However, producer cell lines require an induction step to activate expression of genes needed for viral particle production. Moreover, this means that any single PrCL can only be used for a single product because the genome is fixed. An intermediate solution between transient and

producer cell lines is the use of packaging cell lines (PaCLs). For PaCLs, structural genes of the virus are integrated into the cell during cell line development and their expression is induced during production. The genome is added during production in a single plasmid transfection step (Stewart et al., 2009). This reduces the complexity of the transfection step and allows flexibility in terms of the genome encoded by the vector, but it retains most of the costs of transient transfection process. A summary of different transfection approaches is provided in Figure 18.

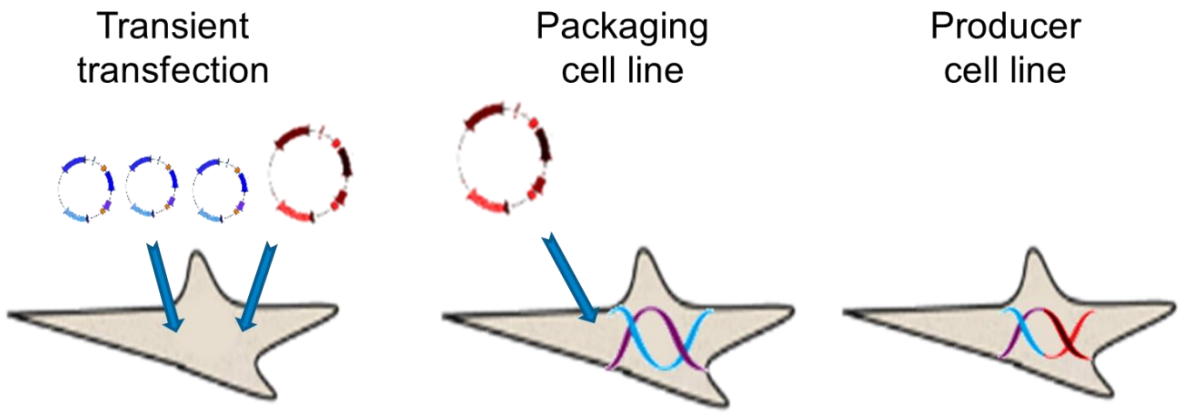


Figure 18: Comparison of cell lines and cell transfection methods used to induce LVV production in cells. Blue DNA/plasmids indicate viral genes; red DNA/plasmid indicates genome. Blue arrows indicate a transfection step required to introduce plasmid DNA into cells.

3.1.5. Lentiviral vector

Lentiviruses are part of the *Retroviridae* family. They are single stranded RNA viruses, characterised by their ability to stably transduce dividing and non-dividing cells. They can encode a large amount of RNA, up to 18 kilobases which is a significant advantage over other types of vectors in terms of large gene or multi-gene delivery (Kumar et al., 2001). They are capable of stable transduction of a variety of host cells, resulting in long-term expression of the target protein and a long-lasting therapeutic effect with a single dose administration and no toxicity or immunogenicity effects (Palfi et al., 2014; Quinonez & Sutton, 2002). LentiVector® is OXB's proprietary gene delivery platform that can be used for both *in vivo* and *ex vivo* LVV products. The benefits of a defined large scale LVV platform include high productivity and safety of the vectors compared to early generation vectors (see Chapter 1). OXB's platform performance has been exemplified in OXB-102 treatment for Parkinson's disease (currently in clinical trials) and FDA-approved Kymriah CAR-T therapy developed in partnership with Novartis. An established viral vector platform provides benefits in terms of shorter development times, improved supply chain management and overall reduction in costs and time required for the product to reach the market. LVVs are

used in development of multiple gene and cell therapy products, designed for a long-lasting treatment in the areas of ophthalmology, neurology and oncology. A downside to LVVs is the fact that producing high titres required for therapeutic effect has proven challenging and balancing process optimisation is a complex task (Merten et al., 2016; Schweizer & Merten, 2010). Vector can be produced in cells cultured adherently or in suspension with the industry trend moving towards serum-free suspension culture due to its lower variability, improved safety and better potential of scale up. The cells are transfected with plasmids encoding modified viral protein genes which are then expressed in the cells leading to production, assembly and secretion of the viral vectors which typically takes up to a week. The vector is then purified using a series of clarification, filtration, Benzonase® treatment and chromatography steps (Segura et al., 2013). All these steps can affect the final vector titre and quality and are a subject to optimisation. However, to optimise the production process without compromising the vector quality there is a need for improved understanding of the impact of changes in the process on the cells and vector (Cockrell & Kafri, 2007).

Efficient production of LVV heavily relies on the cell line performance. As such development of cell lines, especially PaCLs and PrCLs are a major focus within the cell and gene therapy industry, including OXB's in-house cell line development programme. Use of MALDI-ToF and PCA provides a range of tools for in-depth characterisation of cell samples in terms of process consistency (e.g. variation between batches, between scales and different processes). It can also be used to assess viral vector composition to improve product quality monitoring, especially in terms of purification. Finally, as described in more details in Chapter 4, these methods can be applied to cell line development to provide additional information at the critical early stage of cell line selection and therefore reduce campaign timeliness and increase the amount of clones that can be characterised in more detail without the need for scale up into bioreactors.

3.2. Methods

The experiments in this chapter required generation of a large volume of cell and viral vector samples produced at different scales and using different methods. The samples were subjected to different downstream processing regimens. The resulting data was used in PCA to inspect the data structure and identify underlying information about cell and viral vector properties. The total number of biological cell samples of 61 was examined using a variety of methods as described throughout this chapter, in particular section

3.2.1. Material generation

In the experiments, several HEK293T cell lines were cultured in different conditions to obtain a variety of cell samples and produce viral vectors which were sampled as well. The type of cell, transfection method, cell culture method and downstream processing are listed below and summarised in Table 4. A total of 61 different biological cell samples were used to obtain mass spectra as described through sections 3.3.1-3.3.5 and summarised together in the robustness study described in section 3.3.6. For viral vectors, 36 biological samples from different sources were used to obtain mass spectra as described in sections 3.3.7-3.3.9.

Data set 1 and 2 (DS1, DS2) include cell and LVV samples from adherent production used throughout the following analysis

- Sample preparation (Chapter 3.3.1)
- MS Pre-processing (Ch 3.3.2)
- Adherent cell analysis (Ch 3.3.4)
- Overall robustness study (Ch 3.3.6)
- HIV and EIAV viral vector analysis (Ch 3.3.7)
- Downstream processing (Ch 3.3.9)

DS3-6 samples originated from 4 different cell lines and were primarily used in the following analysis

- Cell line impact on MS (Ch 3.3.5)
- Overall robustness study (Ch 3.3.6)

DS 7 and 8 are samples from suspension production used in the following analysis

- Sample preparation (Ch 3.3.1)
- MS Pre-processing (Ch 3.3.2)
- Suspension cell analysis (Ch 3.3.3)

- Overall robustness study (Ch 3.3.6)
- HIV and EIAV viral vector analysis (Ch 3.3.7)
- Downstream processing (Ch 3.3.9)

DS 9 and 10 are samples differentiated by downstream processing and used throughout viral vector analysis

- HIV and EIAV viral vector analysis (Ch 3.3.7)
- Viral vector concentration (Ch 3.3.8)
- Downstream processing (Ch 3.3.9)

DS 11 are sample from large scale production and were used in assessment of DSP effect on viral vector MS (Ch 3.3.9)

DS 12-15 are samples obtained from 2 different packaging cell lines at ambr15 and MiniBio scale. These samples were primarily used to support work described in Chapter 4 but also in the MS robustness study (Ch 3.3.6)

ID	Cell line	Transfection	Cell culture method	Downstream processing	Additional comments
DS1	1	Transient (EIAV)	Adherent, CF2	Two-step centrifugation	Material prepared with help from Kirstie Pemberton, an EngD student at OXB
DS2	1	Transient (HIV)	Adherent, CF2	Two-step centrifugation	
DS3	1	None	Adherent, culture plate	None	
DS4	2	None	Adherent, culture plate	None	
DS5	3	None	Adherent, culture plate	None	
DS6	4	None	Adherent, culture plate	None	
DS7	5	Transient (EIAV)	Suspension, MiniBio	Two-step centrifugation	
DS8	5	Transient (HIV)	Suspension, MiniBio	Two-step centrifugation	
DS9	5	Transient (HIV)	Suspension, EZ	Two-step centrifugation	
DS 10	5	Transient (HIV)	Suspension, EZ	Filtration/ chromatography	
DS 11	5	Transient (HIV)	Suspension, Pilot SUB	Filtration/ chromatography	LVV Samples provided by OXB PR&D team
DS 12	6	Packaging cell line (HIV)	Suspension, ambr®15	Simple filtration	
DS 13	6	Packaging cell line (HIV)	Suspension, MiniBio	Simple filtration	
DS 14	7	Packaging cell line (HIV)	Suspension, ambr®15	Simple filtration	
DS 15	7	Packaging cell line (HIV)	Suspension, MiniBio	Simple filtration	

Table 4: Summary of data sets used in MALDI-ToF MS analysis of LVV and cell samples.
Each data set was generated from a unique set of samples obtained using a different combination of cell lines and production process. Two step centrifugation refers to initial clarification centrifugation followed by overnight ultracentrifugation. Chromatography refers to anion exchange chromatography used to purify viral particles.

3.2.2. Adherent cell culture

Cells grown adherently were cultured in cell culture plates (10 cm^2, Thermo Fisher Scientific), T-flasks (T75, T150 or T225, Thermo Fisher Scientific) or 2-tray layer cell factories (CF2, Nunc™ EasyFill™ Cell Factory™ Systems, Thermo Fisher Scientific), sharing a common protocol adjusted for scale and purpose of cell culture. Cells were revived from a vial (1-1.5 ml, -150°C storage) and cultured in Dulbecco's modified eagle medium with phenol red (DMEM, GE Healthcare Biosciences) with 10% foetal bovine serum (FBS, Life technologies or Gibco). Cells were initially cultured for at least a week in T150 flasks after initial revival. For 10 cm^2- cell culture plates cells were detached from T-flasks and seeded on the plates in DMEM media with 10% FBS. Cells remained untransfected and were harvested after 4 days of culture. CF2 cells were seeded in a similar fashion in DMEM media with 10% FBS in a volume of 200 ml per layer (400ml total volume). 24 hours after inoculation cells were transfected with a set of HIV or EIAV based plasmids (proprietary to OXB) and a green fluorescent protein (GFP)-encoding genome plasmid using Lipofectamine® 2000CD Transfection Reagent (Thermo Fisher Scientific). 20 hours after transfection cells were induced with sodium butyrate (NaBu, 10mM, Sigma Aldrich) vector samples were harvested 6.5 hours after induction (harvest 1, followed by media top-up with DMEM with 10% FBS) and 24 hours after induction along with the cell samples (harvest 2). Details of cell harvest and downstream processing for both adherent and suspension cultured cells are provided in section 3.2.4 and 3.2.5.

3.2.3. Suspension cell culture

Cells grown in suspension were cultured in ambr®15 microbioreactors (Sartorius Stedim Biotech), MiniBio reactors (500ml, Applikon biotechnology) or EZ bioreactors (7L, Applikon biotechnology). All bioreactor protocols shared common elements and followed the same overall procedure with control parameters adjusted for scale according to OXB's proprietary process scale up settings. Cells were revived from a vial (1-1.5ml, -150°C storage) and cultured in FreeStyle™ 293 media (Thermo Fisher Scientific) supplemented with cholesterol lipid concentrate (0.1% v/v using 250x Cholesterol Lipid Concentrate, Thermo Fisher scientific) shake flasks (250 or 500 ml, Corning) for at least a week prior to inoculation. Bioreactor vessels were assembled, tubed and autoclaved prior to the experiment. Ambr®15 is a single use micro bioreactor system, where the head plate is autoclaved prior to the experiment while the vessels are provided pre-sterilised and ready to use. All following steps are programmed and automated when using ambr®15 with reagents handled by a robotic dispenser. MiniBio reactors were filled with FreeStyle™ 293 media to a working

volume of 400 ml, EZ reactors were filled to a working volume of 5L. The bioreactors were inoculated with previously cultured cells. After 24 hours the bioreactors were transfected with viral vector plasmids. For transient transfection the cells were transfected with a set of HIV or EIAV based packaging plasmids (proprietary to OXB) and a GFP-encoding genome plasmid using Lipofectamine® Transfection Reagent. For PaCLs only GFP-encoding genome plasmid was transfected using the same method; for PaCLs, viral protein gene expression was induced during transfection step using doxycycline addition. 20 hours after transfection all cells were induced using NaBu (10mM). Cells and vector were harvested 24 hours after induction by pouring or pumping the culture fluid under aseptic conditions (microbial safety cabinet). Additionally, small volume samples were collected with a syringe. Pilot bioreactor scale HIV-GFP vector samples were provided by OXB downstream processing group (produced in a single use 50L BIOSTAT® CultiBag® stirred tank reactor, Sartorius Stedim Biotech). Details of downstream processing are outlined in a further section.

3.2.4. Cell samples harvest

Cell samples were collected from all experiments. For adherent culture, cells were detached by incubation with TrypLE™ Select (Thermo Fisher Scientific) for 5 minutes at 37 °C which was subsequently neutralised by addition of DMEM media with 10% FBS. For suspension culture cells were harvested directly from the culture vessels. Cell concentrations were measured using Cedex XS Analyser System (Roche Life Sciences) or NucleoCounter® NC-200 (Chemometec). Cells were aliquoted and centrifuged (3000 rpm/1000g, 5 minutes, room temperature) to match the total viable cell count of $7x10^4$, $1x10^5$ or $1.5x10^5$ cells per sample. Supernatant was carefully discarded, and pellets were washed with phosphate buffered saline (Thermofisher scientific) and frozen at -20°C for all cell samples and -80°C for all viral vector samples following OXB's protocol.

3.2.5. Viral vector harvest and downstream processing

Vector samples were harvested once for the suspension and adherent EIAV process and twice for the adherent HIV process (second harvest after 24h). The summary of downstream processing actions is presented in Table 4. Simple filtration refers to collecting small volumes of vector samples for analysis and viral vector titre measurement. Media are harvested (5 to 10 ml) and centrifuged (3000 rpm, 5 minutes). Supernatant is filtered (0.22 μm, Fisherbrand™ syringe filter, Fisher Scientific) and aliquoted (1.5ml or 2ml cryotubes). The remaining terms apply to harvesting of the entire bulk material and purification through different methods. Two step centrifugation refers to the following steps: initial clarification

centrifugation (3000 rpm for 5 minutes , followed by 0.22 µm filtration with Stericup® filter, Merck Millipore), overnight centrifugation (6000rpm/4000g, 4 °C, supernatant is discarded and pellet re-suspended in 13 ml PBS), ultra-centrifugation (20 000 rpm/100 000g, 4 °C, 1.5 hours, re-suspended in 100 µl (adherent samples) formulation buffer TSSM (20mM Tris, 100mM NaCl, 1% w/v Sucrose, 1% w/v Mannitol, pH 7.3) or 150 µl TSSM (suspension)). Vector samples reached the final volumetric concentration of 2000x and were subsequently frozen (-80 °C). Crude vector samples were collected and frozen down as well. Filtration/chromatography refers to a full downstream processing procedure: Samples were clarified using a depth filtration filter (Sartoclear® P MaxiCap®, Sartorius; retention rate: 1,5 µm; diameter: 100 mm, cartridge height: 365 mm). Clarified harvest was frozen at -80°C. The material was thawed, supplemented with Benzonase® endonuclease (Merck Millipore) with addition of magnesium chloride (2mM final concentration) and incubated for one hour at 37°C. The sample was then purified using AEC Äkta purifier, using Sartobind®Q SingleSep membrane adsorber (Sartorius AG) using 1M NaCl elution conditions). The resulting purified sample was then concentrated using ultrafiltration/diafiltration using hollow fibre filtration and spin filtration to the final x1000 volumetric concentration.

Vector samples were used to calculate the vector titre using a live cell transduction assay utilising fluorescence-activated cell sorting flow cytometry (FACSVerse™, BD BioScienses). One day 1 HEK293T cells were seeded in a 96 flat bottom well plate and incubated for 24 hours in 150 µl DMEM media with 10% FBS (Gibco) and polybrene (1 in 400 dilution, Sigma Aldrich). 24 hours after seeding, on day 2, viral vector dilutions are prepared by mixing them with DMEM media (1 in 100 dilution). 4 µl of diluted vector is added to each well of the originally seeded plate to transduce the cells. 3 hours after vector addition, 250 µl of DMEM supplemented with 10% FBS and polybrene (1 in 400 dilution) are added to each well to reach total of 400 µl volume. The plates are incubated for 72 hours. On day 5 the media are removed and 100 µl TrypLE™ is added to detach the cells. After 5 mins incubation at 37°C TrypLE™ is neutralised with 150 µl DMEM media with 10% FBS and cells are re-suspended and transferred to a round bottom 96 wells plate in identical order as on original plate. Transduced cells are analysed using FACSVerse™ to estimate the viral vector titre by gating live cells and analysing histogram of GFP positive cells. The assay followed the same protocol as for data described in Chapter 2 but the assay variation was higher with CV of 10-50% between biological replicates. However, analysis performed in this chapter was less focused on LVV titre which was only used as reference.

3.2.6. MALDI-ToF MS

Several parameters and processing steps of MALDI-ToF MS were varied and optimised to achieve consistent quality of mass spectrum, (Figure 19).

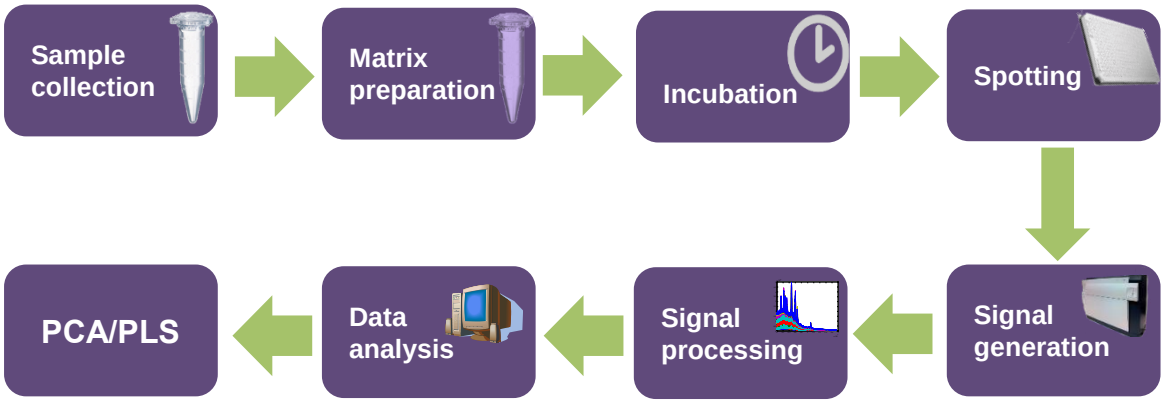


Figure 19: Flow diagram of MALDI-ToF MS sample generation and analysis process.

The parameters optimised during method development included the chemical composition of the matrix, sample incubation time, spotting technique and number of spots on a plate per sample, laser settings and signal processing (discussed in a separate section). Sample preparation has a large impact on spectrum quality and reproducibility. Varying the sample concentration, matrix type and composition will affect the final spectrum and therefore needs to be optimised and controlled for each assay. The laser intensity and voltage should be adjusted based on matrix and sample composition to minimise generation of multiply charged ions (Williams et al., 2003). These parameters were optimised with guidance from Dr Jane Povey and Prof Mark Smales from the University of Kent. Design of Experiment approach could be considered for further optimisation of matrix composition and samples preparation, however the quality of the method established in the initial investigation was sufficient to perform MS analysis and further optimisation was outside of the scope of this EngD project. In the following experiment MALDI-ToF MS was used for characterisation of HEK293T cells (sections 3.3.3 and 3.3.4) as well as EIAV and HIV based LVVs (sections 3.3.7 - 3.3.9).

MALDI-ToF MS matrices were prepared by first preparing a solution of HPLC grade water with 40% HPLC grade acetonitrile and either 0.06% or 0.15% trifluoroacetic acid (TFA, 99% purity, Across chemicals). Appropriate chemicals (α-cyano-4-hydroxycinnamic acid (α-cyano), sinapinic acid (SA) or 2,5-dihydroxybenzoic acid (DHB)) were added to the solution to reach concentration of 10mg/ml , mixed thoroughly and sonicated for 15 minutes

in a sonicating water bath. The solutions were then spun down in a microcentrifuge at 13 000 rpm (14 000g) for 5 minutes.

Frozen cell samples were thawed (room temperature, 15-20 °C), re-suspended in 50 µl of each matrix and incubated at 4 °C for 1,2 or 3 hours. 1 µl of the sample/matrix mix was spotted on 384 well ground steel MALDI-ToF plate (Bruker) several times (ranging from 3 to 10 repeats per sample) and left to dry at room temperature (15-20°C). For vector samples, they remained suspended in TSSM (no centrifugation step was performed after initial sample concentration) and 0.5 µl of vector sample was spotted, immediately followed by 0.5 µl of matrix. The plate was loaded to MALDI-ToF MS and analysed through automated protocol with settings recommended by Dr Jane Povery, University of Kent (Bruker Ultraflex; laser intensity 62% (cells) or 75% (vector); laser frequency 500 Hz; polarity: positive; ions sources:1. 24.93 kV, 2. 23.08 kV, lens 7.5 kV; Pulsed ion extraction 400 nS; Suppress at 4 kDa; spectra collected in the range of 4-60 kDa; sample rate 0.13 Gs/s, 3600 ionisation laser shots summed and saved per sample). The resulting data files were collated using R script and imported to MATLAB (R2013)

3.2.7. Principal component analysis of mass spectrometry data

For the analysis, the mass spectra were either inspected directly or pre-processed to eliminate the noise and intensity variation resulting from sample handling. For some types of analysis, the technical replicates of individual samples were averaged while other types of analysis used individual mass spectra, based on the desired analysis outcome. Several signal pre-processing steps were applied using both MATLAB® bioinformatics toolbox and Eigenvector PLS toolbox. The pre-processing optimisation was an iterative process, initially following method described by Povey et al. (2014) which was adapted over time along with optimisation of sample preparation as described in section 3.2.6. The pre-processing was adjusted to best match the cell and viral vector samples used throughout the experiment by adjusting the settings of individual commands in MATLAB as well as the order and total number of processing steps. This optimisation process was guided by Dr Chris O'Malley and Prof. Gary Montague whose expertise allowed to make efficient adjustments to the method. An alternative development method could include a Design of Experiments approach using different pre-processing steps and their individual settings as factors with the goal of minimising data noise and variation between identical samples. This approach was not taken in this project thanks to the expertise of the supervisors supporting the project. The final pre-processing steps were applied to all processed spectra and are following:

1. Repeats of individual samples were averaged. Non-averaged samples from individual repeats were also analysed in section 3.3.2 to examine assay variability
2. Baseline correction using MATLAB® bioinformatics toolbox function '*msbackadj*' to reduce the variation introduced by variable background reading caused by small differences in matrix chemical composition and sample desorption process. The command estimates the baseline and adjusts the intensities by subtracting the baseline value. The correction was done with average spectrum used as intensity matrix and command used window size 200 and step size 200.
3. Data normalisation using MATLAB® bioinformatics toolbox function '*msnorm*' to adjust the peak intensities by standardising the area under the curve to the median values for the group.
4. Curve smoothening using Savitzky-Golay (SG) algorithm through MATLAB® bioinformatics toolbox function '*mssgolay*' to further reduce noise. Parameters for SG smoothening: window width 20, polynomial order 1st and no derivative. Different sets of parameters, including 1st and 2nd order derivative treatment to some of the samples was examined as a part of the potential pre-treatment but abandoned for the final analysis. Some of the early results are presented in the results and discussion sections.
5. Peak alignment using MATLAB® bioinformatics toolbox function '*msalign*' to correct minor differences in overall spectra position.
6. During initial pre-processing optimisation, mean centring was initially applied to the samples using Eigenvector PLS toolbox. This method was not used in the final analysis of MS data due to effect it had on the data structure. Some of the early results with mean centring are presented in results and discussion sections to examine effects of mean centring on data structure.

Following signal pre-processing the samples were subjected to PCA. It is an exploratory multivariate data analysis method that transforms complex data into a smaller number of principal components (PCs), which reduces the dimensionality of the data. This allows easier analysis and graphical representation of data structure as well as analysis of complex multivariate interactions between variables (Abdi & Williams, 2010). For each PC, loadings determine contribution of each variable to the PC score value. In case of MS the variables are values of mass to charge ratio (m/z). For each sample, the PC score represents the value corresponding to a transformed spectrum. In the case of mass spectra, PCA transforms the whole mass spectrum of each sample into a single data point which can be placed on a PC scores plot to compare it to other samples and assess its characteristics. For

the cell analysis all pre-processed mass spectra from cell samples described in Table 4 were arranged in 6 data sets based on cell culture type, sample concentration, matrix composition, incubation time and time between sample spotting and analysis. This was done to examine the effect of different properties on the variation in mass spectrum.

PCA was performed on 6 major data sets:

1. Adherent and suspension cell samples transfected with either EIAV or HIV, incubated with 6 different buffers (α-cyano, SA or DHB, each with 0.06% or 0.15% TFA) for different amount of time (1, 2 or 3 hours). The analysis focused on the effect of incubation time and matrix composition

2. Suspension cell samples transfected with either EIAV or HIV incubated with 2 different buffers (α-cyano + 0.15% TFA or SA + 0.15% TFA) for 1 hour and 15 minutes at 3 different cell concentrations ($7x10^4$, $1x10^5$, $1.5\ x10^5$). The analysis focused on the effect of cell concentration as well as a comparison between adherent and suspension cell samples

3. Adherent cell samples transfected with either EIAV or HIV incubated with α-cyano + 0.15% TFA for 1 hour and 15 minutes at 3 different cell concentrations ($7x10^4$, $1x10^5$, and $1.5\ x10^5$). The analysis was focused on the effect of cell concentration as well as a comparison between adherent and suspension cell samples

4. The same adherent data set analysed 24h after spotting cells on the MALDI plate (samples were spotted separately from data set 2). The analysis focused on the effect of time between spotting and analysis

5. Concentrated viral vector samples based on HIV or EIAV produced in HEK293T cells grown adherently or in suspension, incubated with either α-cyano + 0.15% TFA or SA + 0.15% TFA and analysed immediately after spotting or 24h later. The analysis focused on characterisation of viral vector samples

6. Sets 2, 3 and 4 combined together for a comprehensive comparison between adherent and suspension cells.

The data sets were pre-processed as described above and subjected to PCA using Eigenvector PLS toolbox using default cross validation settings. Labels were assigned to data points based on the sample type and the PC scores and loadings plots were examined.

Finally, all cell and vector samples were collated into a cell matrix and a vector matrix (Pilot and EZ samples were used only for vector analysis) and subjected to PCA to assess the impact of cell culture method, scale and, in case of the viral vector samples, downstream processing. Q residuals and Hotelling T^2 values were examined to assess the quality of the PCA model as well as identify outliers or samples where variability is not explained by the model (Bro et al., 2014). For the vector analysis the high concentration of vector required to obtain high intensity mass spectra was a limiting factor. Only the 5 samples purified with two-step centrifugation or filtration/chromatography (as summarised in Table 4) were used in the analysis.

3.3. Results

MALDI-ToF MS was examined as a method for cell and viral vector characterisation. The method was first optimised to achieve optimal signal intensity and consistency which was followed by a robustness study, examining the performance of MS using a variety of samples. PCA was used throughout the process to examine data structure and underlying trends.

3.3.1. Effects of samples preparation and matrix composition

MALDI-ToF MS matrix composition has a significant impact on the quality and consistency of mass spectra as defined by signal-to-noise ratio and variation between repeats from the same sample. Different chemicals are suitable for use with different types of samples where optimal matrix composition should be determined experimentally for each new application. The data below is part of DS1-2, DS7-8 (See Table 4), generated from adherent and suspension cell pellets (at $1.5x10^5$ cells per sample) suspended in 6 different matrices and incubated for 1 to 3 hours (3 time points). The data can be used to compare differences in incubation time, matrix composition and sample preparation. All samples were measured in triplicates.

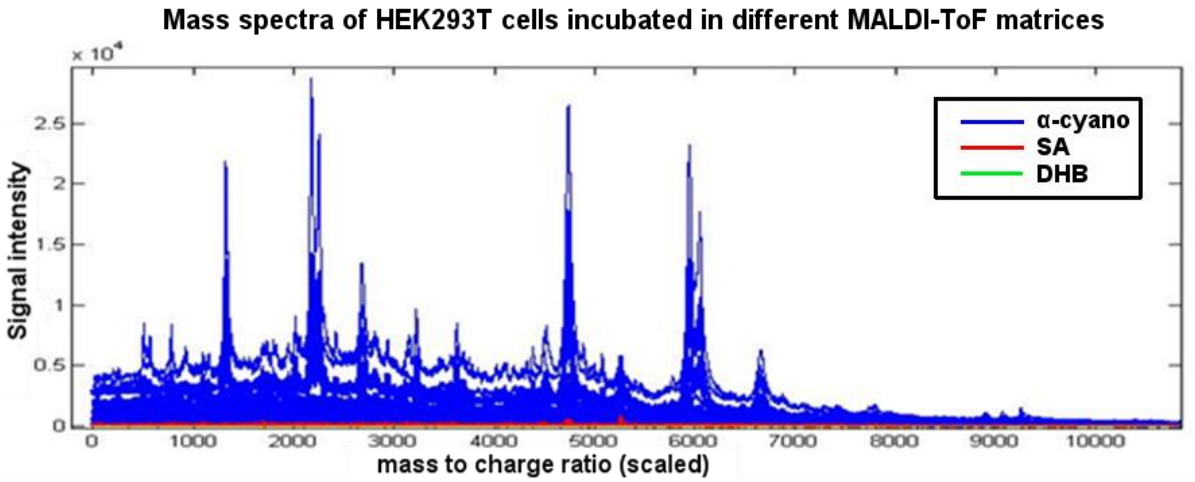


Figure 20: Line plot of mass spectra generated from 108 HEK293 cell samples prepared using three different matrices.
Presented mass spectra from all samples (DS1-2, DS7-8) used in the analysis are unprocessed. Colour based on the matrix used: blue – α-cyano, red – SA (low visibility due to low signal level), green – DHB (not visible due to very low signal level).

The first point to address is the choice of buffer/matrix used for sample ionisation. From a first glance α-cyano results give stronger signal (blue in Figure 20-Figure 22), SA gives weaker signal where individual peaks are still visible (red in Figure 20-Figure 22) while DHB resulted in a signal with no detectable peaks (green in Figure 20-Figure 22). α-cyano

was determined as the optimal choice for the matrix while SA and DHB result in much poorer mass spectra; however, SA was still used in some of the experiments for comparison and to confirm the initially observed trend.

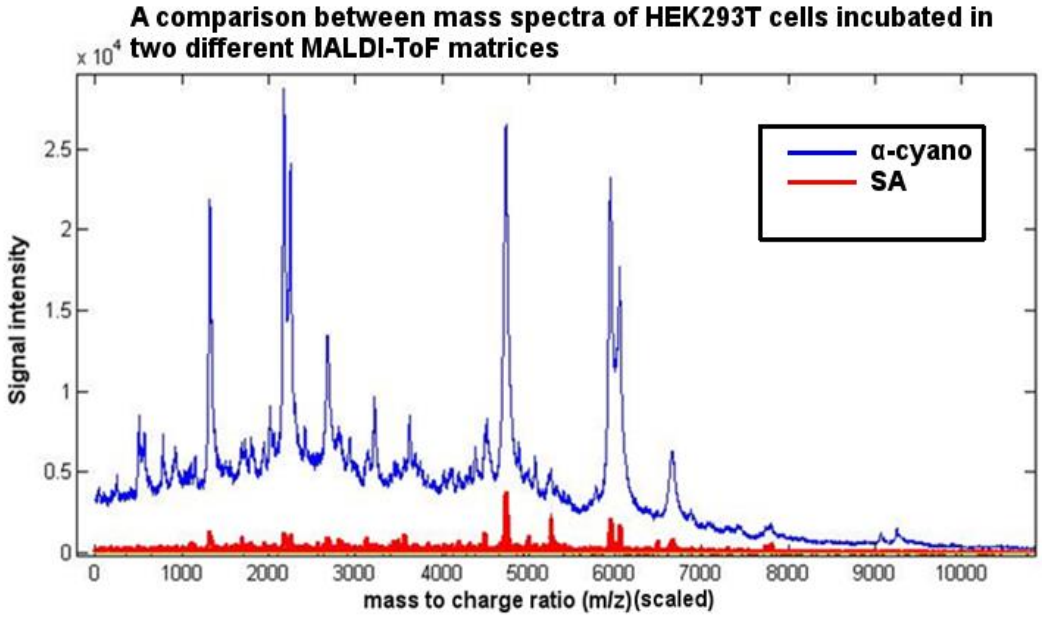


Figure 21: Line plot of mass spectra generated using SA matrix.
Unprocessed mass spectra from samples (DS1-2, DS7-8) incubated with SA (red) for 1, 2 or 3 hours. A single spectrum from a randomly selected sample incubated with α-cyano is displayed as a reference (blue).

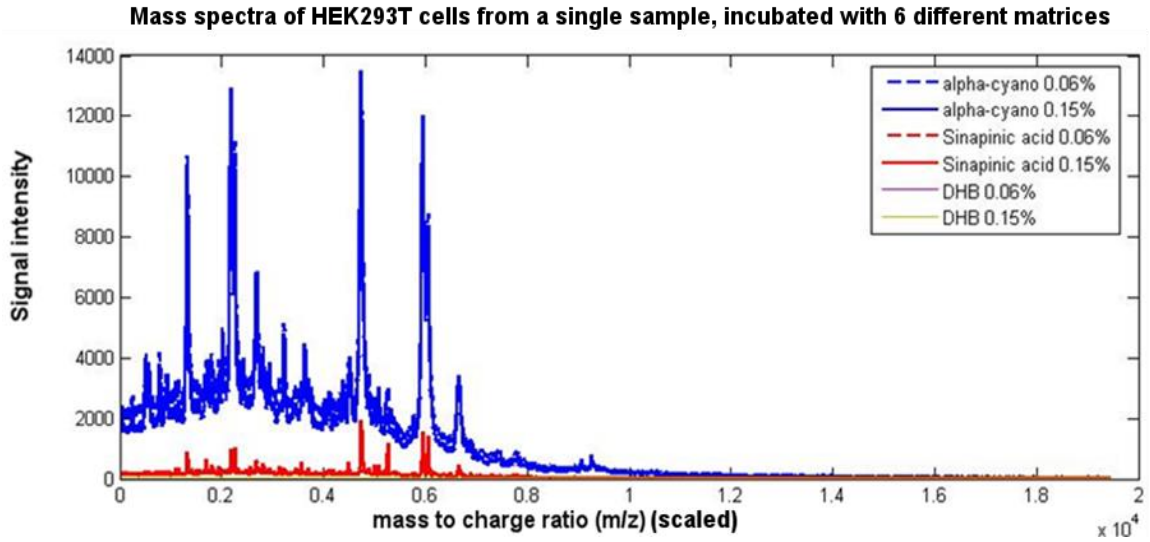


Figure 22: Line plot of mass spectra of a single cell sample, generated using 3 different matrices and 2 TFA concentrations.
Unprocessed mass spectra of a single sample (adherent cells incubated for 1 hour, DS1) were incubated with different matrices. DHB signal is not visible due to very low signal intensity

The second buffer optimisation step was the choice of TFA concentration to use with the matrix. For α-cyano it seems to have little effect with 0.06% having minimally higher signal strength but 0.15% resulting in sharper peaks and lower baseline intensity. Overall both TFA concentrations give comparable results that allow distinguishing individual peaks (Figure 23). For SA, there is some difference in intensity as well as shape of the spectrum

between 0.06% and 0.15% TFA. Some of the peaks are visible for either 0.06% or 0.15% TFA but overall the spectrum obtained with SA and 0.15% TFA has more distinct peaks and overall higher intensity compared to SA with 0.06% TFA (Figure 24).

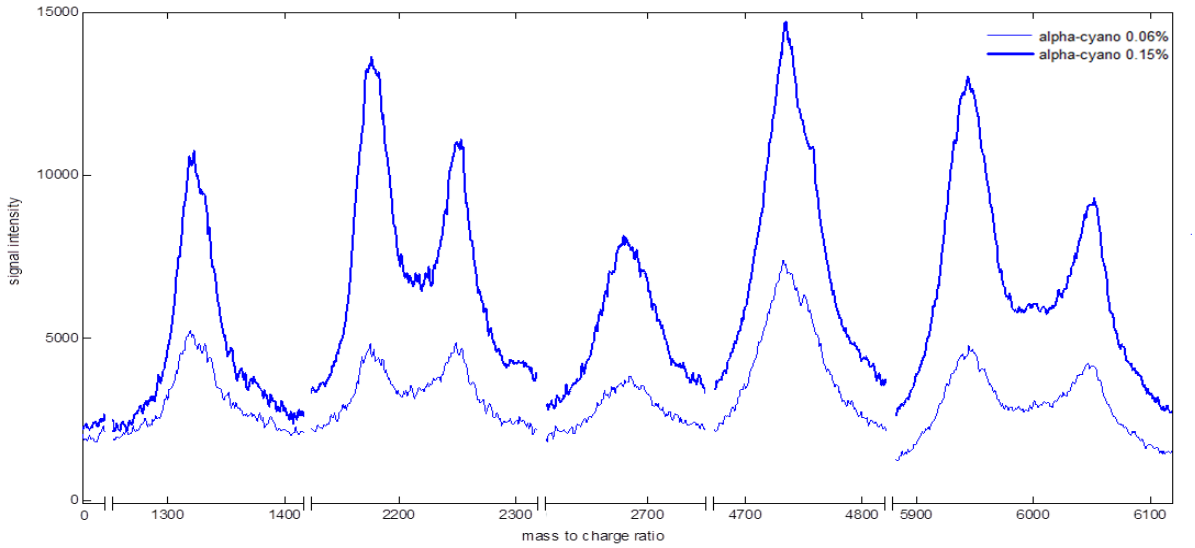


Figure 23: Line plot of selected peaks of mass spectra generated using α-cyano matrix at 2 different TFA concentrations.
Spectra obtained from a single adherent cell sample incubated with α-cyano and either 0.06% or 0.15% TFA for 3 hours (DS1).

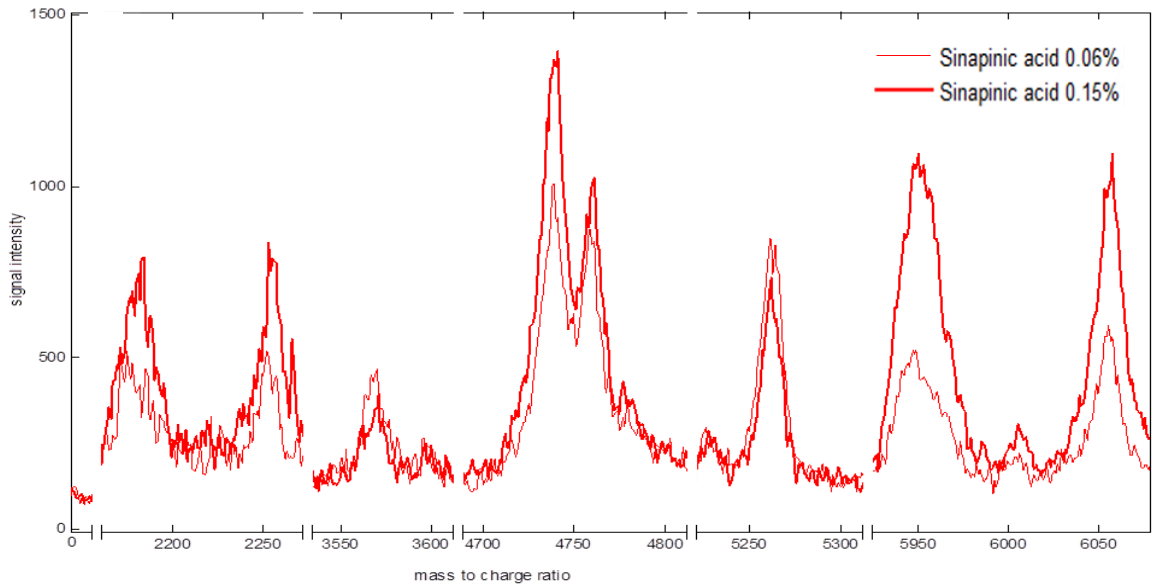


Figure 24: Line plot of selected peaks of mass spectra generated using SA matrix at 2 different TFA concentrations, Spectra obtained from a single adherent cell sample incubated with SA and either 0.06% or 0.15% TFA for 3 hours (DS1).

After visual analysis of raw mass spectra, PCA was performed to reduce data complexity and compare sources of variation captured by different PCs. Both PCA analysis (Figure 25) and simple plots indicate that α-cyano is a superior matrix for use with HEK293

cells, resulting in higher signal intensity as well as increased number and quality of signal peaks (peaks are smoother, with less overlap and bigger difference from baseline spectrum) compared to other matrices used in the experiment, therefore providing more information. In the PCA the majority of sample variability (95%) is contained in the first principal component. PC1 score values correlate with the matrix used for sample preparation as well as sample incubation time, with highest results for α-cyano and 1 hour incubation (Figure 25). For PCA using only α-cyano samples, PC1 score correlates with incubation time. Samples 1-6 and 19-24 (T1) have the highest scores values (Figure 26)

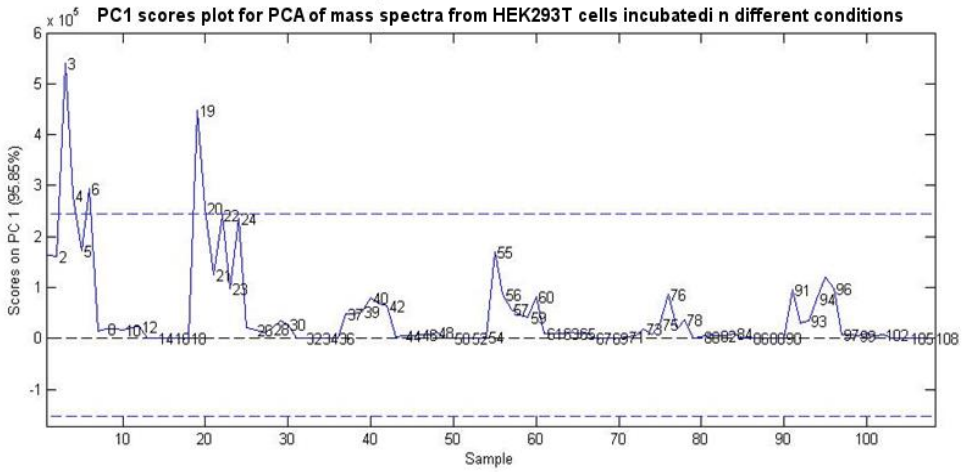


Figure 25: PCA scores plot (PC1) for mass spectra of 108 HEK293 cell samples generated using 3 different matrices. PCA was applied to raw mass spectra (no pre-processing). Legend for sample numbers (DS1-2, DS7-8) in Table 5 below. Highest scores are all for samples incubated with α-cyano.

Sample No	Description	Sample No	Description	Sample No	Description
1-6	α-cyano Adh 1 hour	7-12	SA Adh 1 hour	13-18	DHB Adh 1 hour
19-24	α-cyano Susp 1 hour	25-30	SA Susp 1 hour	31-36	DHB Susp 1 hour
37-42	α-cyano Adh 2 hour	43-48	SA Adh 2 hour	49-54	DHB Adh 2 hour
55-60	α-cyano Susp 2 hour	61-66	SA Susp 2 hour	67-72	DHB Susp 2 hour
73-78	α-cyano Adh 3 hour	89-84	SA Adh 3 hour	85-90	DHB Adh 3 hour
91-96	α-cyano Susp 3 hour	97-102	SA Susp 3 hour	103-108	DHB Susp 3 hour

Table 5: List of samples used in the MALDI-ToF MS matrix analysis.
Each description column determines the type of matrix used for incubation, type of cells (grown adherently (DS1-2) or in suspension (DS7-8)) and the duration of incubation. The samples within box highlighted in bold were used in the following section to examine the effect of spectra pre-processing. Each set of 6 samples represents a single biological sample with 6 mass spectrometry analytical replicates.

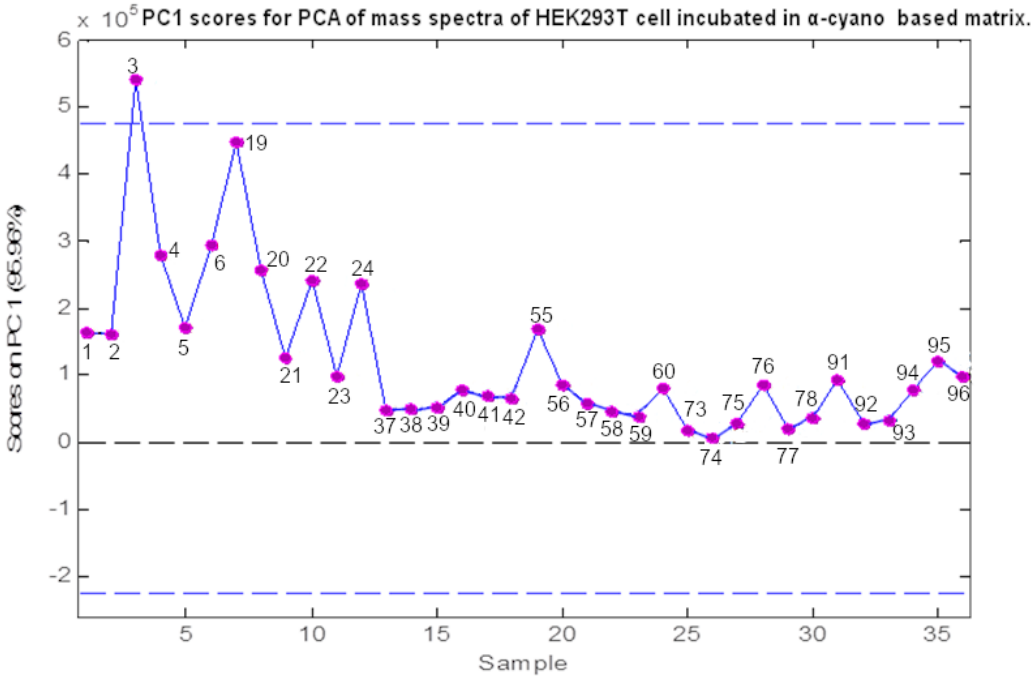


Figure 26: PCA scores plot (PC1) for 36 cell samples (adherent and suspension, Table 5) incubated with α-cyano. PCA was applied to mass spectra from the cell samples (DS1-2 and DS7-8) using raw spectra (no-pre-processing).

3.3.2. Effects of spectra pre-processing

MS data is complex and often affected by assay variation resulting in noisy data. This can be partially alleviated by performing data pre-processing which reduces the level of noise and eliminates some of the assay variation. Multiple data pre-processing approaches were applied to MS data obtained in the experiment. Pre-processing of cell and vector data was identical and the first step applied to the raw spectra (Figure 27) was baseline correction (Figure 28). Spectroscopic data is characterised by an inherent variability in baseline intensity of the signal which is caused by variation in sample desorption and ionisation process. Baseline correction is a method which brings all spectra to a common starting level and reduces the noise caused by the matrix composition and ionisation process. The next step involves data normalisation (Figure 29) which scales the y axis values based on the area under the curve and further reduces noise from variation in spectrum intensity caused by the varying amount of ionised material and the degree of material desorption.

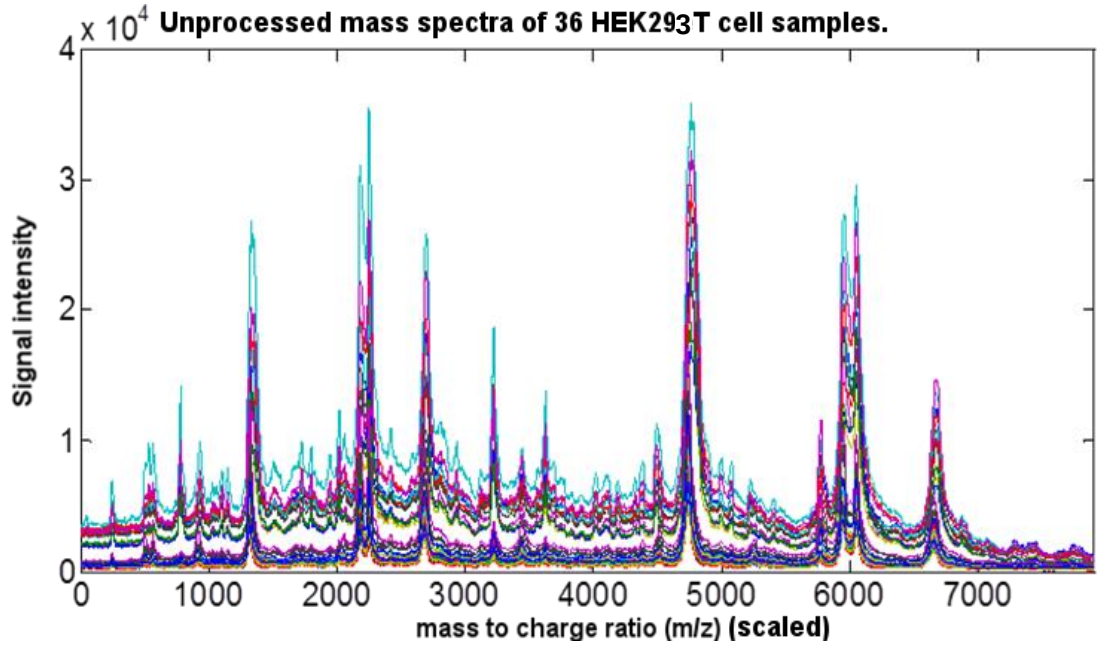


Figure 27: Unprocessed mass spectra of 36 cell samples (as in Figure 26).
All samples (DS1-2, DS7-8) were incubated with α-cyano matrix to obtain the mass spectra

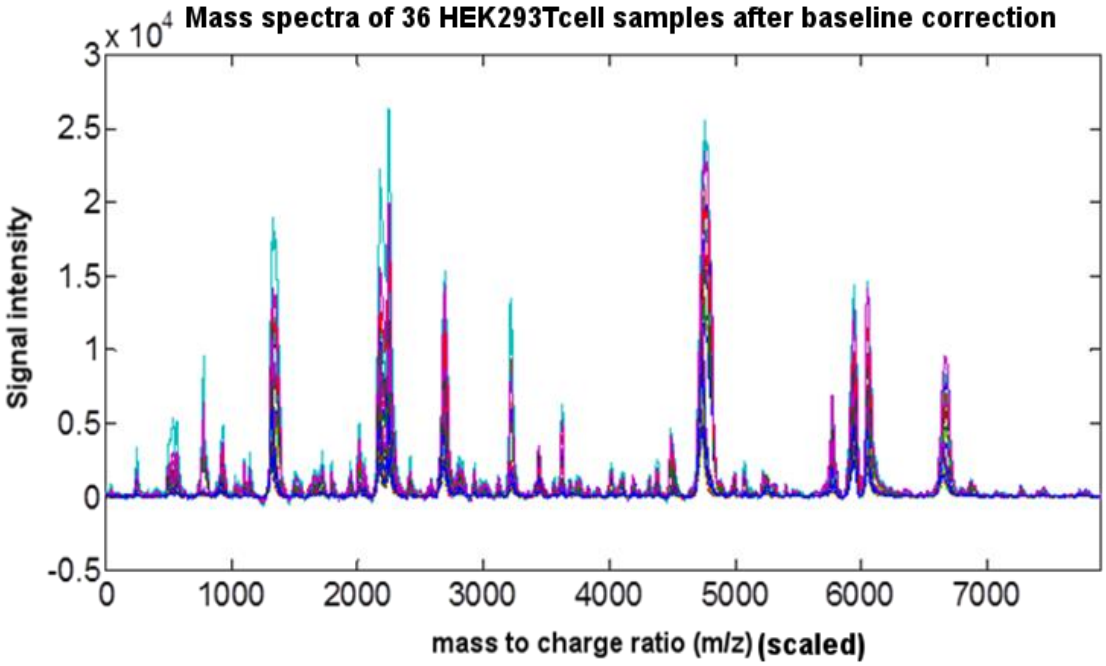


Figure 28: Mass spectra of 36 cell samples (as in Figure 26) after baseline correction step.
All samples (DS1-2, DS7-8) were incubated with α-cyano matrix to obtain the mass spectra

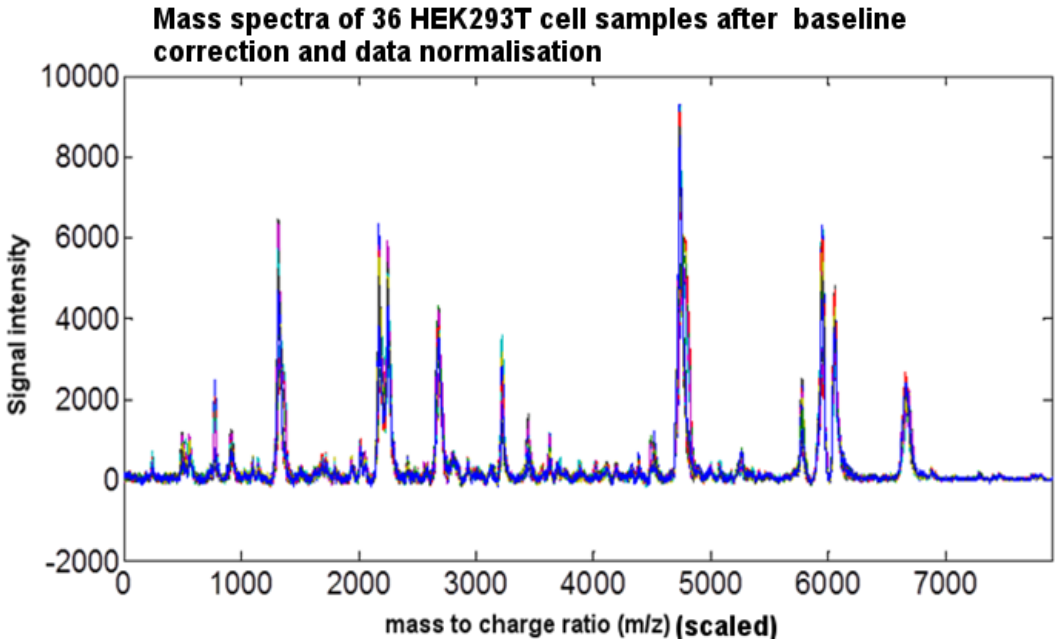


Figure 29: Mass spectra of 36 cell samples (as in Figure 26) after applying baseline correction and data normalisation steps
All samples (DS1-2, DS7-8) were incubated with α-cyano matrix to obtain the mass spectra

The final pre-processing step uses SG smoothening algorithm to filter the data and smoothen highly variable peaks which should further enhance the analysis. SG smoothening allows taking derivative of the original data as part of the algorithm. Analysis was initially performed using both data with no derivative (Figure 30) as well as 1st order derivative (Figure 31). Addition of 1st derivative has a dramatic effect on data structure which is reflected in the PCA analysis. The model using SG with no derivative captures the total of 99.13% variability in the first 2 PCs while the model using 1st order derivative captures 87.09% of variability and has significantly different results. This is caused by the fact that 1st order derivative transforms the data based on its primary structure. MS data analysis heavily relies on the signal intensity, positioning of the peaks relative to each other and overall structure of the mass spectrum. Therefore, after the initial analysis SG smoothening was performed using no derivatisation. Smoothening window width (span) was assessed in range between 10-30 with no significant effect on outcome and was maintained at 20 for all experiments as described in section 3.2.7.

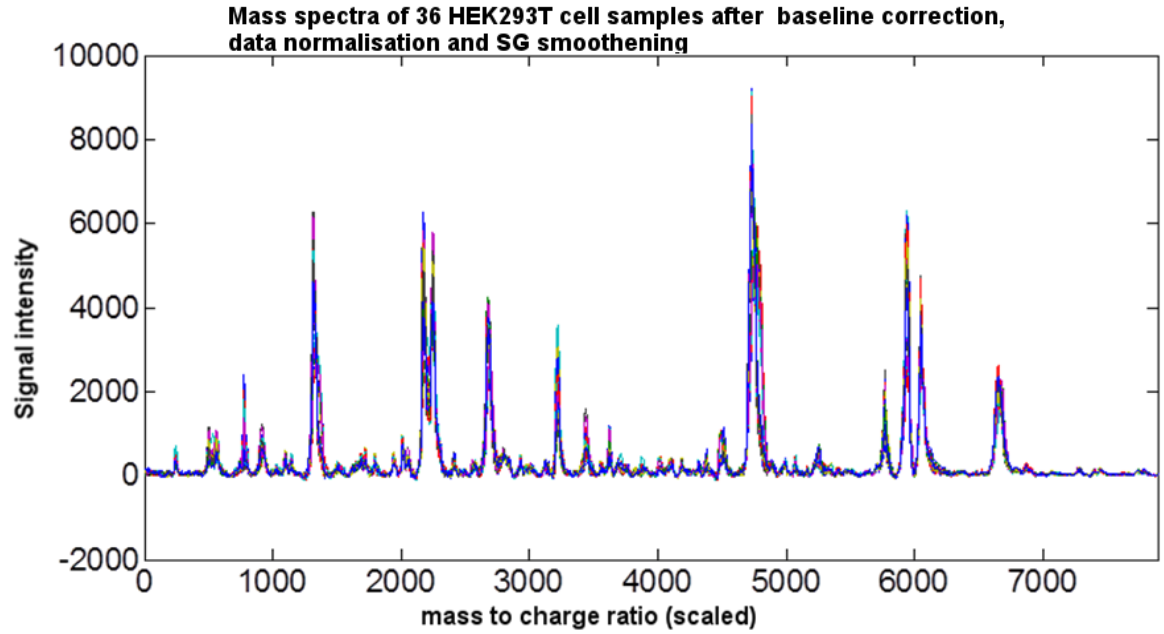


Figure 30: Mass spectra of 36 cell samples (as in Figure 26) after applying baseline correction, data normalisation and SG smoothening.
All samples (DS1-2, DS7-8) were incubated with α-cyano matrix to obtain the mass spectra

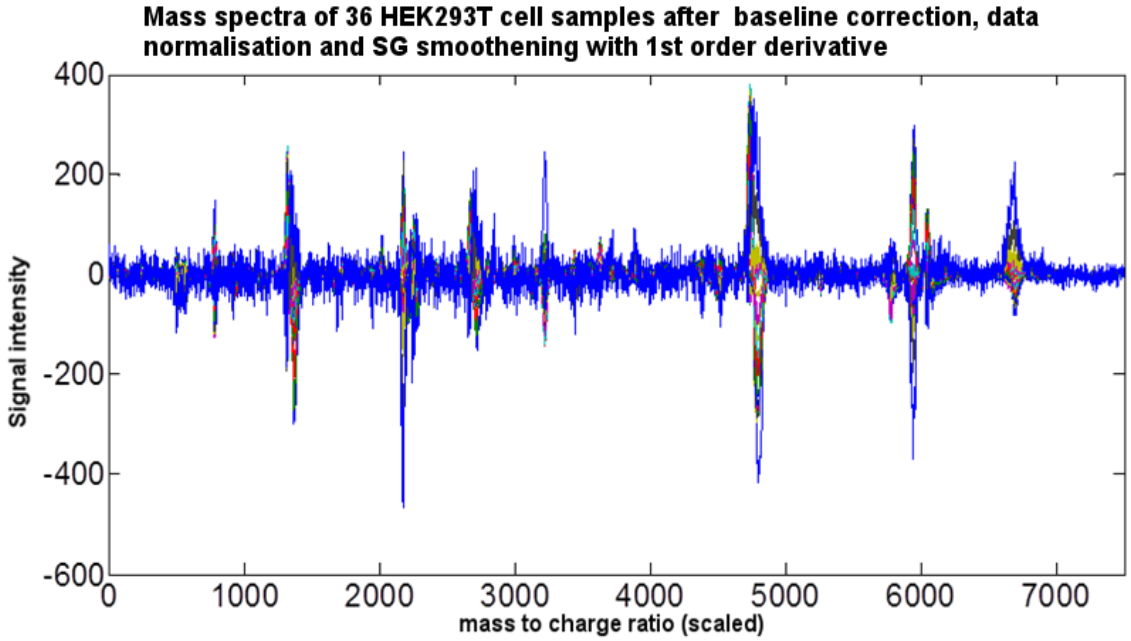


Figure 31: Mass spectra of 36 cell samples (as in Figure 26) after applying baseline correction, data normalisation and SG smoothening (1st order derivative).
All samples (DS1-2, DS7-8) were incubated with α-cyano matrix to obtain the mass spectra

Finally, another pre-treatment method considered for MS data was mean centring. It is a method where the average spectrum is subtracted from each sample. It has a major impact on the data structure (Figure 32). Similar to taking a derivative of the data, mean centring changes the structure of the data. The focus of this analysis is on the absolute values of the peak intensities and their relative position and shapes. Therefore, mean centring was not used

in the final round of pre-treatment. Another method trialled in the initial approach was re-sampling and trimming of the data. These methods were not applied in the final analysis to avoid bias from selection of specific values or regions of unknown spectra. The only modification was trimming of the high m/z values (above 7500 m/z) which did not show any high intensity signals.

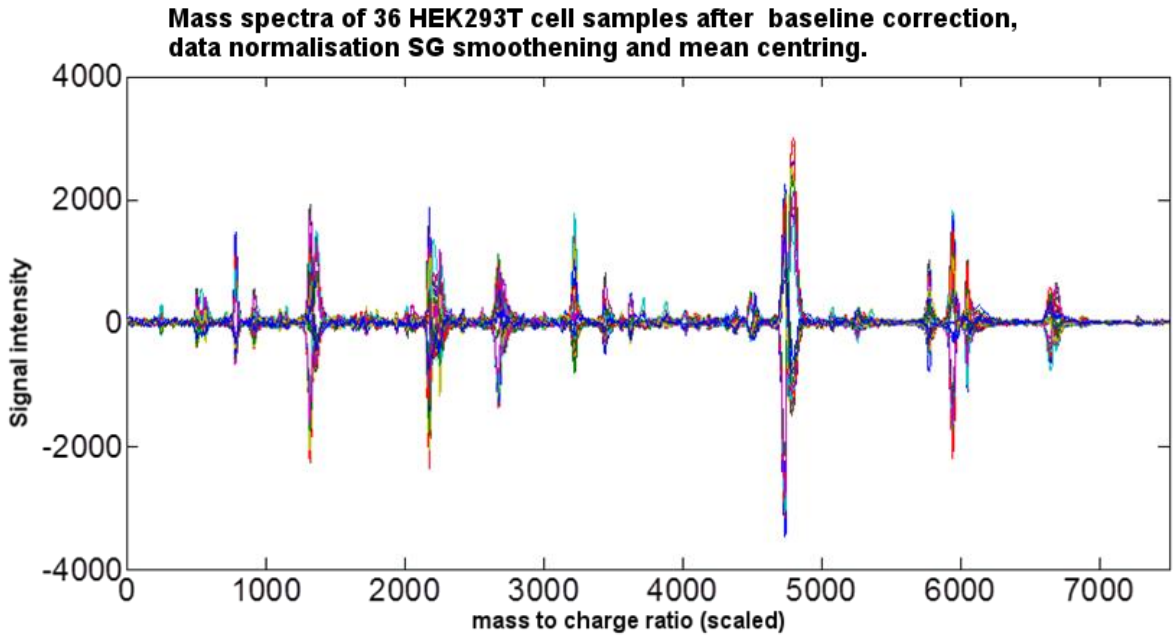


Figure 32: Mass spectra of 36 cell samples (as in Figure 26) after applying baseline correction, data normalisation, SG smoothening and mean centring.
All samples (DS1-2, DS7-8) were incubated with α-cyano matrix to obtain the mass spectra

Throughout the pre-treatment procedure samples were subjected to PCA to assess the effect of filtering on the data structure. For the 36 cell samples (See first column of Table 5), data structure of PC1 scores remained similar to the original data (see Figure 26). Interestingly, for PC2 scores (Figure 33), adherent samples (1-6; 13-18; 25-30) generally have higher PC2 scores value than suspension samples incubated for the same amount of time (7-12; 19-24; 31-36). The results are clearer when using averaged spectra where each data point represents 6 averaged mass spectra from adherent or suspension cell samples, incubated for 1,2 or 3 hours. (Figure 34-Figure 35). Figure 36 highlights the separation between adherent and suspension cell sample PC scores. This variation between cells is further investigated in the cell analysis section 3.3.3 and 3.3.4.

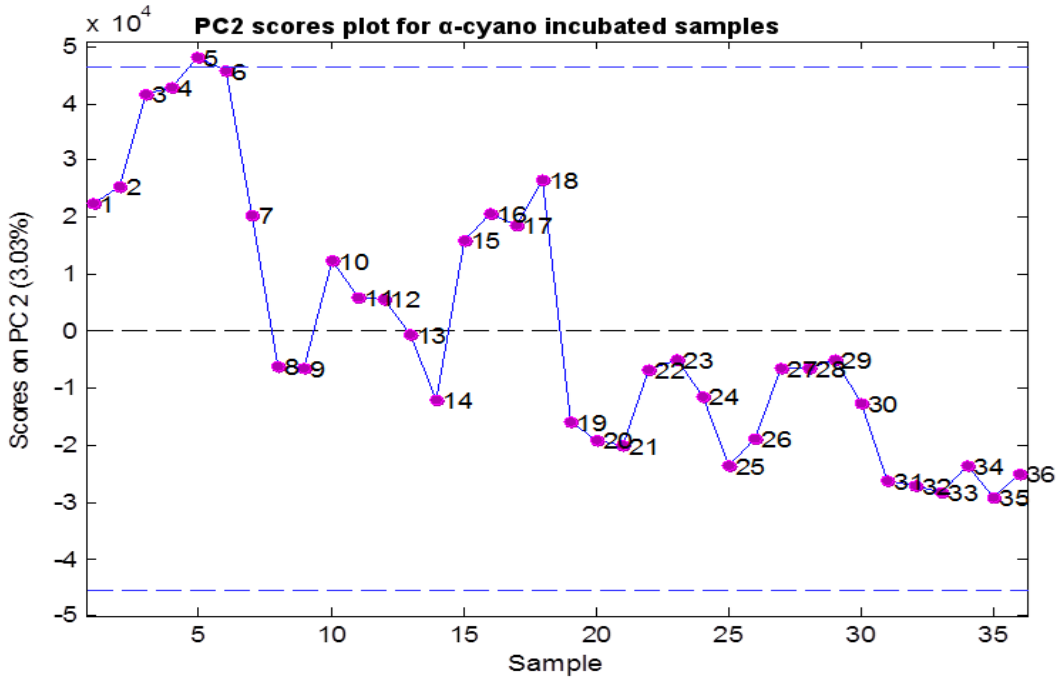


Figure 33: PCA scores plot (PC2) from mass spectra of 36 cell samples (as in Figure 26).
Cell samples were incubated with α-cyano and mass spectra were pre-treated using baseline correction and data normalisation.
All samples (DS1-2, DS7-8) were incubated with α-cyano matrix to obtain the mass spectra

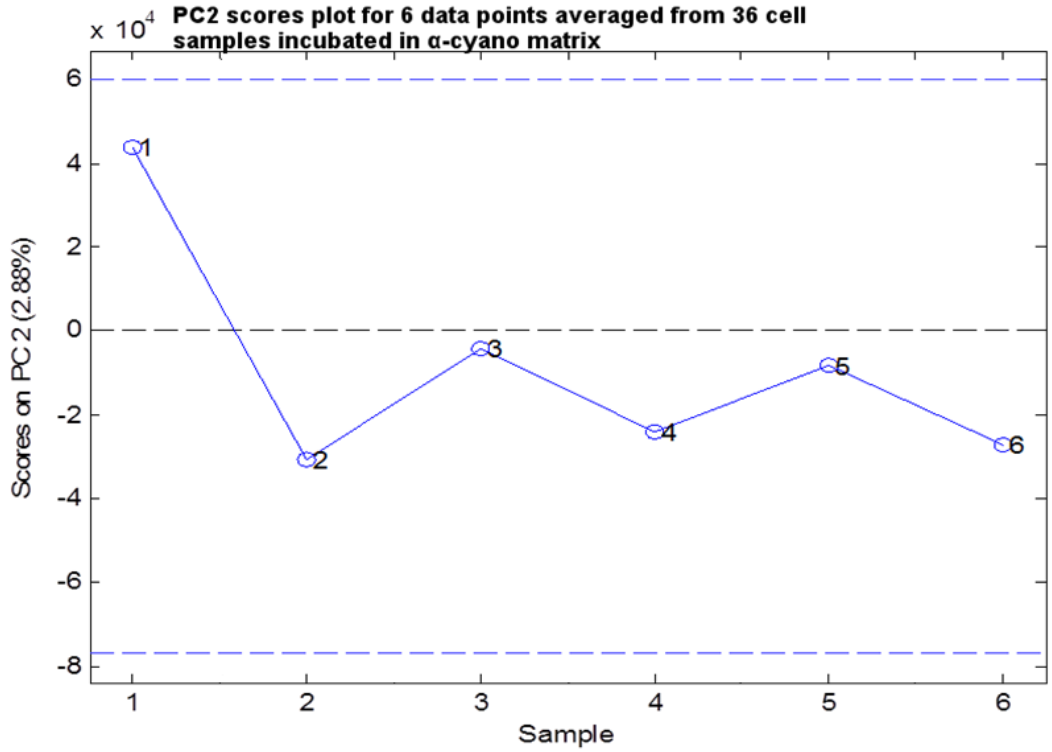


Figure 34: PCA scores plot (PC2) of mass spectra from 36 cell samples (averaged).
All samples (DS1-2, DS7-8) were incubated with α-cyano matrix to obtain the mass spectra. Mass spectra were averaged and no further pre-processing was applied. 1 – adherent cell, α-cyano matrix, 1 hour incubation; 2- suspension cell, α-cyano matrix, 1 hour incubation; 3 - adherent cell, α-cyano matrix, 2 hour incubation; 4- suspension cell, α-cyano matrix, 2 hour incubation; 5- adherent cell, α-cyano matrix, 3 hour incubation; 6 – suspension cell, α-cyano matrix, 3 hour incubation

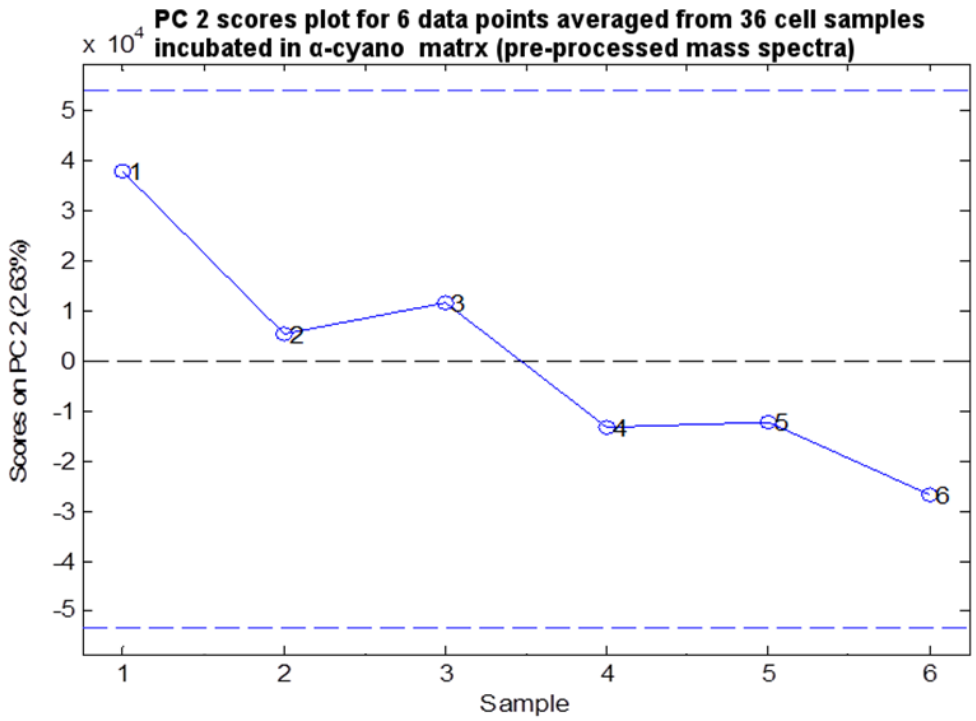


Figure 35: PCA scores plot (PC2) of mass spectra from 36 cell samples (averaged and pre-processed).
Mass spectra (DS1-2, DS7-8) were processed as in Figure 34 above with the difference of additional pre-processing using baseline correction and data normalisation as described in the methods section 3.2.7.

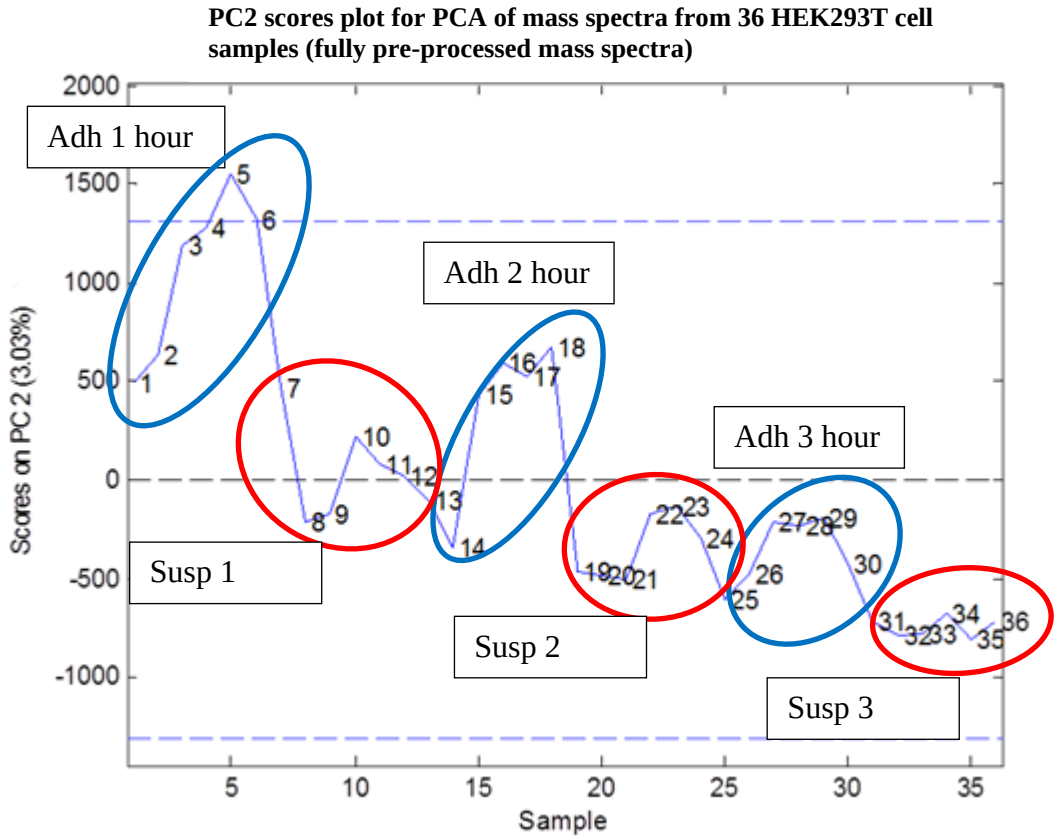


Figure 36: PCA scores plot (PC2) with highlighted adherent and suspension 36 cell samples (as in Figure 26). All samples (DS1-2, DS7-8) were incubated with α-cyano matrix to obtain the mass spectra
Mass spectra from adherent cell samples are highlighted by blue circles, suspension cell samples are highlighted in red; each circle represents a set of samples from a single batch.

3.3.3. Suspension cell analysis

The next step involved additional analysis of exclusively suspension cultured cell samples (DS7-8, see Table 4) to determine the impact of cell concentration, initial number of spots on MALDI plate per sample and the variability of mass spectra of cells cultured in suspension. This was a separate analysis using 2 types of matrices to confirm the effect of α-cyano and SA on the mass spectrum signal. The first analysis inspected MS spectra for HEK293T cells transfected with either HIV or EIAV), grown in suspension in 2 batches for each combination (Figure 37). The measurements are done using two buffers (α-cyano and SA) and 3 different cell concentrations ($7x10^4$, $1x10^5$, $1.5\ x10^5$ cells per sample). Each sample is measured in quadruplicate. Figure 37 shows spectra from all samples in the first data set.

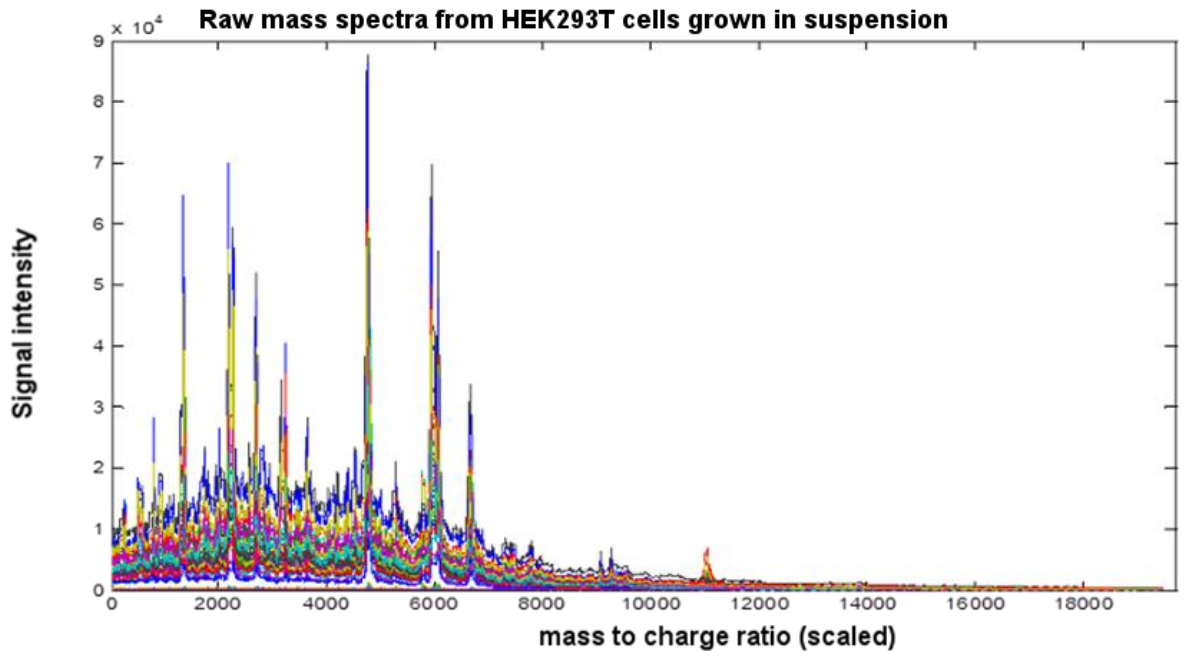


Figure 37: Line plot of raw mass spectra from 104 suspension cell samples.
HEK293T cells grown in suspension (DS7-8), transfected with EIAV or HIV or untransfected as a control, each sample was collected and analysed at 3 different cell concentrations and with 2 different buffers (α-cyano or SA)

PCA was used as the main method of comparing the different sample sets. The analysis (initially performed without signal pre-processing) confirms that buffer composition has a major impact on the spectrum as illustrated by significant difference as observed through PCA scores (Figure 38). PC1 scores were clearly separated according to matrix used to incubate the samples (positive PC1 scores for α-cyano samples, negative PC1 score for SA

incubated samples). This stands in line with the results from previously described buffer/matrix analysis.

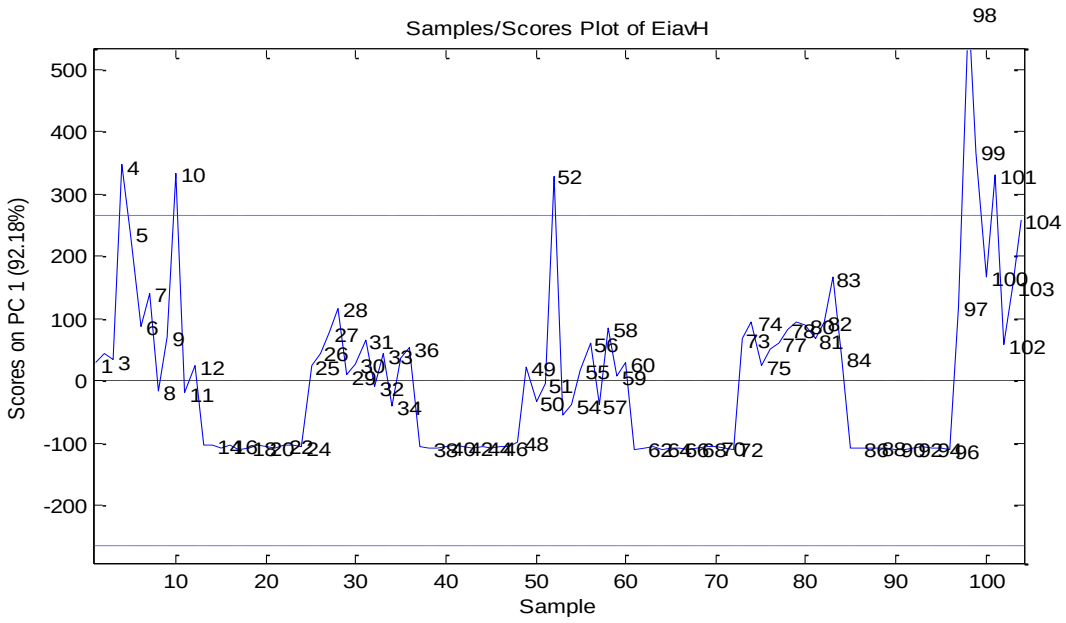


Figure 38: PCA scores plot (PC1) of mass spectra from 104 suspension cell samples (DS7-8).
Samples 13-24, 37-48, 61-72 and 85-96 form near flat lines around value of -100, all were incubated with SA matrix, the rest of the samples were incubated with α-cyano matrix. No pre-processing was applied to data.

Subsequently, the analysis focused only on α-cyano samples (Table 6) and it highlighted several potential outliers with unexpected PC score values (e.g. samples 52). All of them are single technical replicates from different samples (DS7-8), indicating a degree of variation in sample spotting and mass spectrum acquisition (less than 5% rate) which can be compensated for with enough additional technical replicates (Figure 39-Figure 41).

Sample No (averaged)	Description	Sample No (averaged)	Description
1-4 (1)	EIAV batch 1 7x10^4 cells	25-28(7)	HIV batch 1 7x10^4 cells
5-8(2)	EIAV batch 1 1x10^5 cells	29-32(8)	HIV batch 1 1x10^5 cells
9-12(3)	EIAV batch 1 1.5 x10^5 cells	33-36(9)	HIV batch 1 1.5 x10^5 cells
13-16(4)	EIAV batch 2 7x10^4 cells	37-40(10)	HIV batch 2 7x10^4 cells
17-20(5)	EIAV batch 2 1x10^5 cells	41-44(11)	HIV batch 2 1x10^5 cells
21-24(6)	EIAV batch 2 1.5 x10^5 cells	45-48(12)	HIV batch 2 1.5 x10^5 cells

Table 6: List of suspension cell samples incubated with α-cyano matrix and used in PCA.
Samples (DS7-8) were analysed in quadruplicates or as averages from 4 repeats (numeration in brackets).

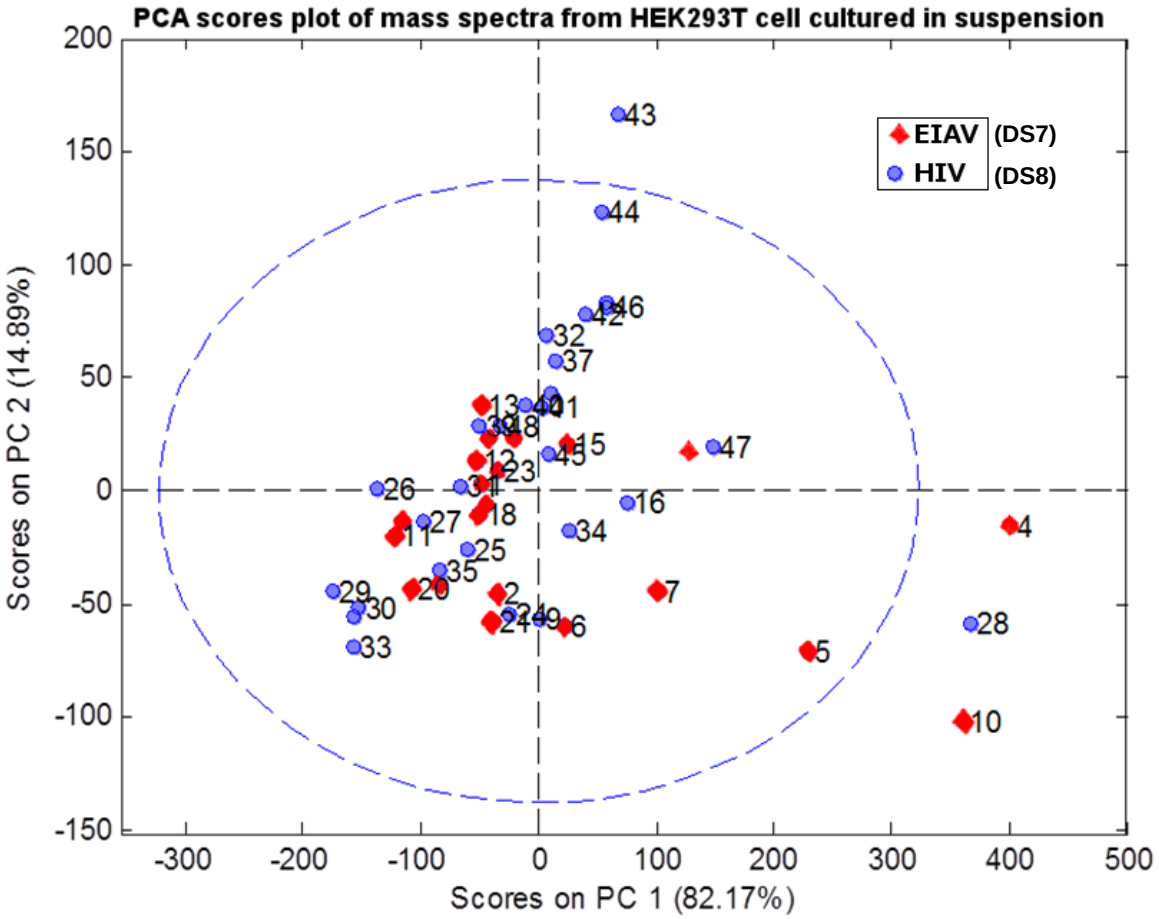


Figure 39: PCA scores plot (PC1 and PC2) of mass spectra from 48 suspension cell samples (Table 6). All samples (DS7-8) were incubated with α-cyano. Samples 4, 10, 28 and 43 are located outside of the confidence limit and are potentially significantly different when compared to the majority of MS samples. No pre-processing was applied to data.

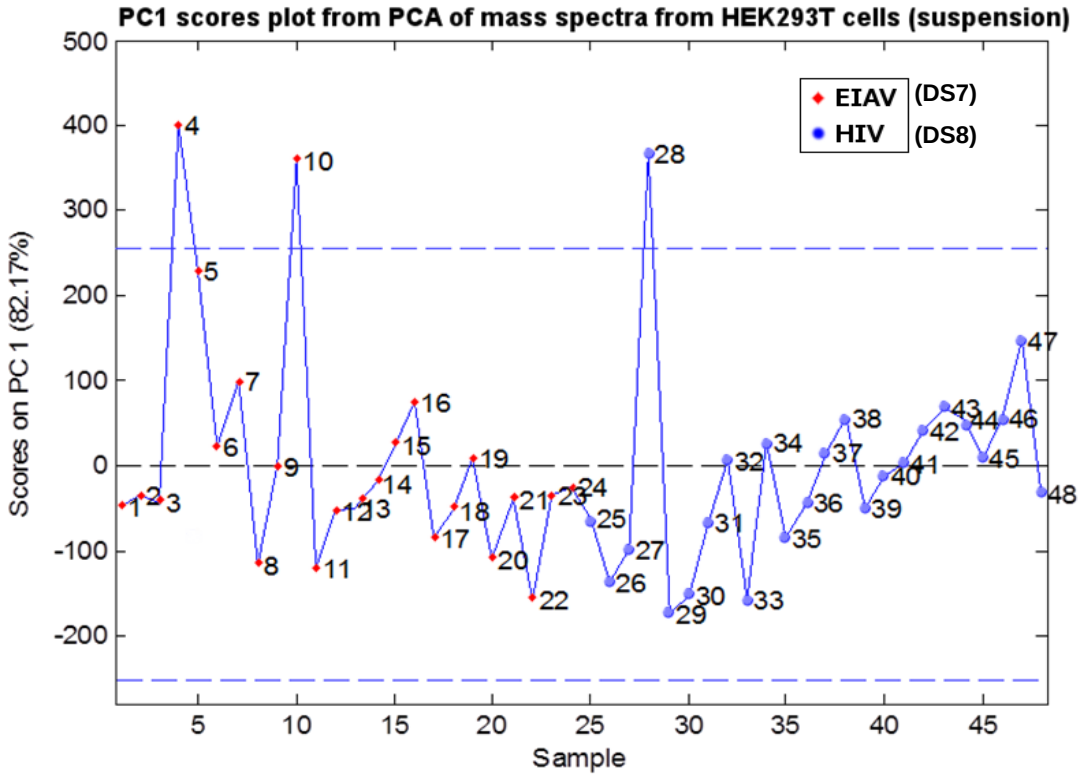


Figure 40: PCA scores plot (PC1) of mass spectra from 48 suspension cell samples (Table 6).
All samples (DS7 and DS8) were incubated with α-cyano. Samples 4, 10, and 28 are located outside of the confidence limit and are potential outliers. No pre-processing was applied to data.

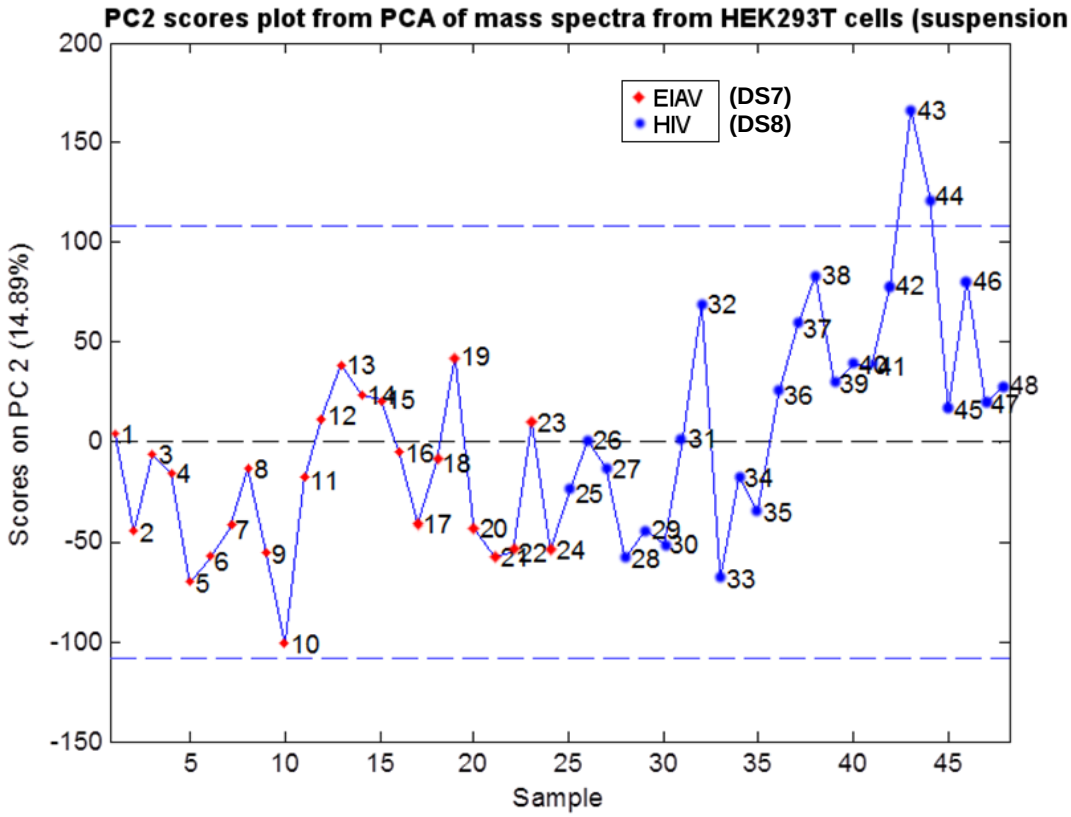


Figure 41: PCA scores plot (PC2) of mass spectra from 48 suspension cell samples (Table 6).
All samples (DS7-8) were incubated with α-cyano. Sample 43 is located outside of the confidence limit and is a potential outlier. No pre-processing was applied to data.

Effects of MS signal pre-processing were described in section 3.3.2. After initial experiments described above, MS data from suspension cell samples was fully processed (as described in section 3.2.7). The PC scores become less variable with PC1 scores remaining well within the analysis 95% confidence interval for all samples and capturing the majority of variability (Figure 42), indicating that while some samples may have unusually high or variable signal strength (outliers in raw data) they still bear representable information which can be assessed after data pre-processing.

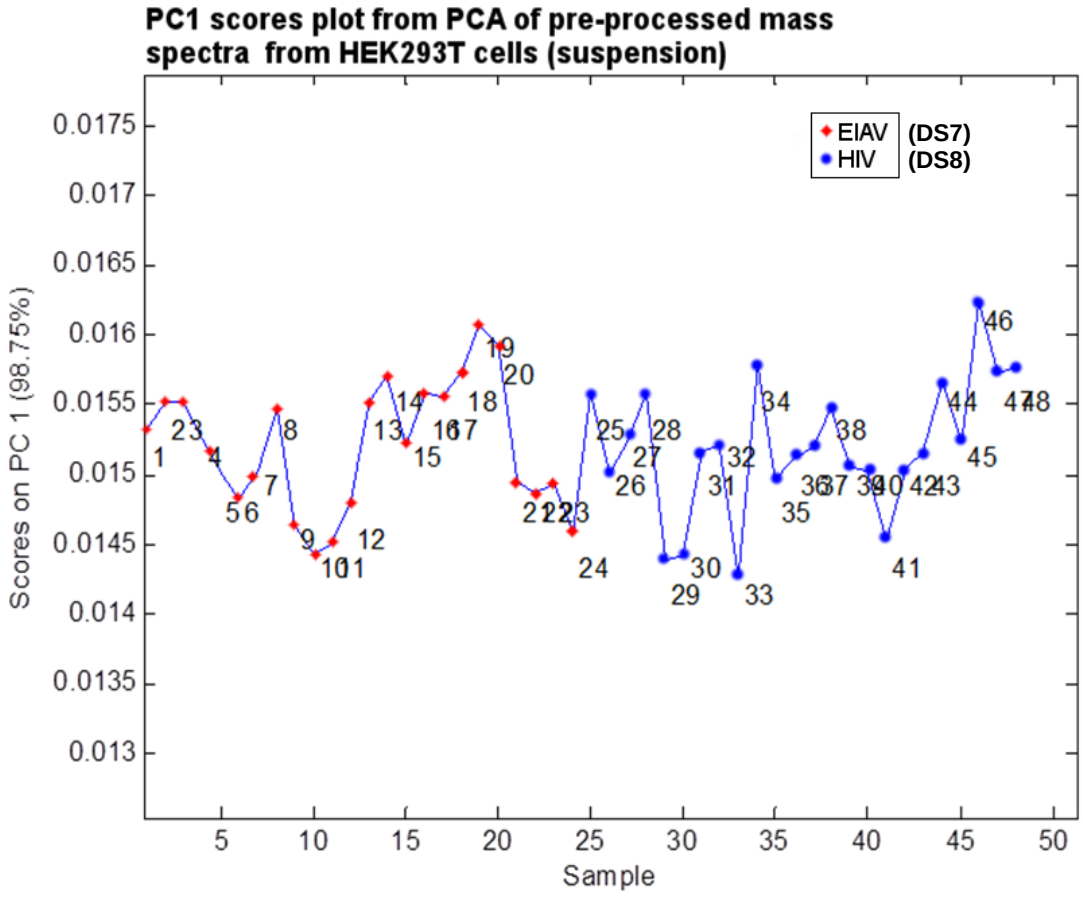


Figure 42: PCA scores plot (PC1) of pre-processed mass spectra from 48 suspension cell samples (Table 6). All samples (DS7 and DS8) were incubated with α-cyano and pre-processed using baseline correction and data normalisation. All samples are within confidence limit. Y axis scale was set to illustrate differences between samples, 95% confidence interval limit is outside Y axis scale (0.027 PC1 score)

Given that samples 1-12 come from 1st batch of EIAV transfected cells, 13-24 come from second batch of EIAV transfected cells, 25-36 come from 1st batch of HIV transfected cells and 37-48 from 2nd batch HIV transfected cells, there is no particular trend in data that would suggest a difference either between batches or between HIV and EIAV transfected cells (Figure 42-Figure 43). Additionally, samples from each batch were sampled at 3 different cell densities (Table 6). Given that there are no major differences between subsets of samples, the analysis is not significantly affected by the cell concentration within range tested

in the experiment. The majority of variance is captured in PC1 (98.75%, Figure 42) while the further PCs capture only a small percentage of variance (0.36% for PC2, Figure 43), indicating that overall the mass spectra are similar to each other and after accounting for variation in average signal intensity (captured by PC1) there is little information contained in the higher number PCs.

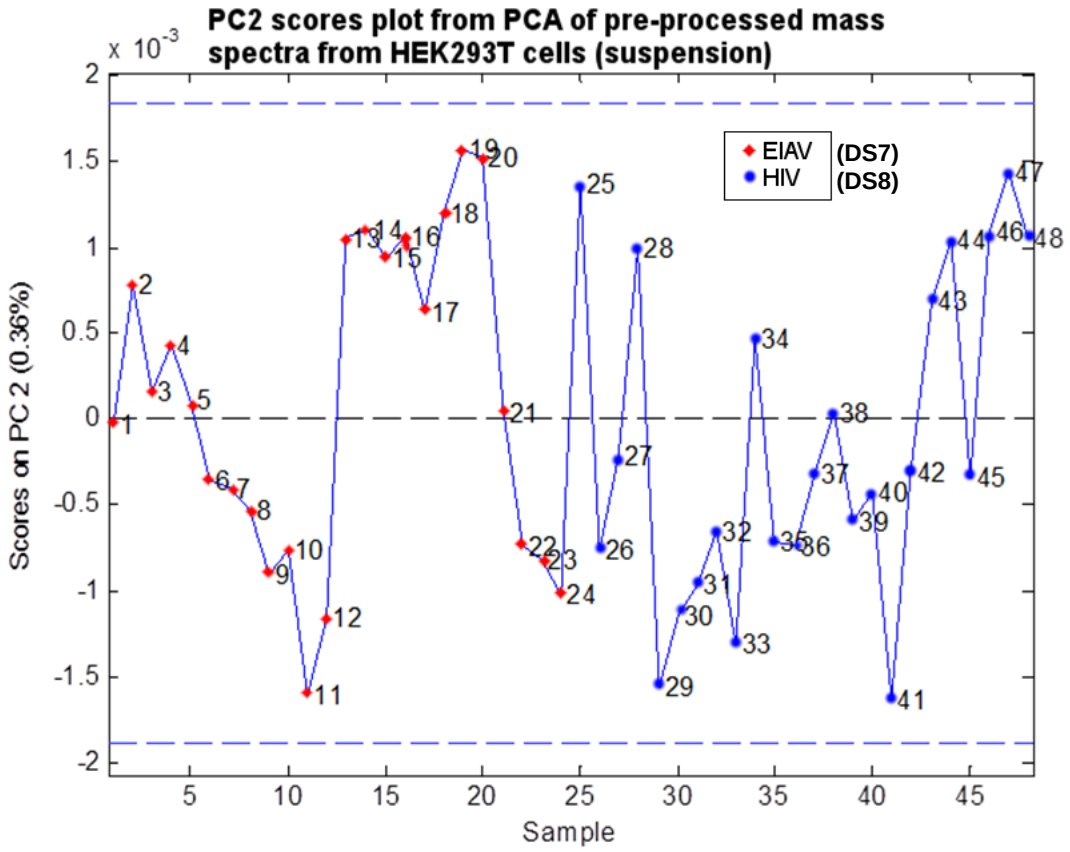


Figure 43: PCA scores plot (PC2) of pre-processed mass spectra from 48 suspension cell samples (Table 6). All samples (DS7 and DS8) were incubated with α-cyano and pre-processed using baseline correction and data normalisation. All samples are within confidence limit.

PCA analysis of averaged data (Table 6, sample numbers in brackets) shows that there is no clear distinction between EIAV or HIV transfected cells and different batches of cells transfected with the same vector as all these groups are overlapping (Figure 44). This most likely indicates consistent mass spectra signal across batches and very little difference between the two transfected cell types which remains consistent with the PCA results of individual repeats (Figure 42-Figure 43).

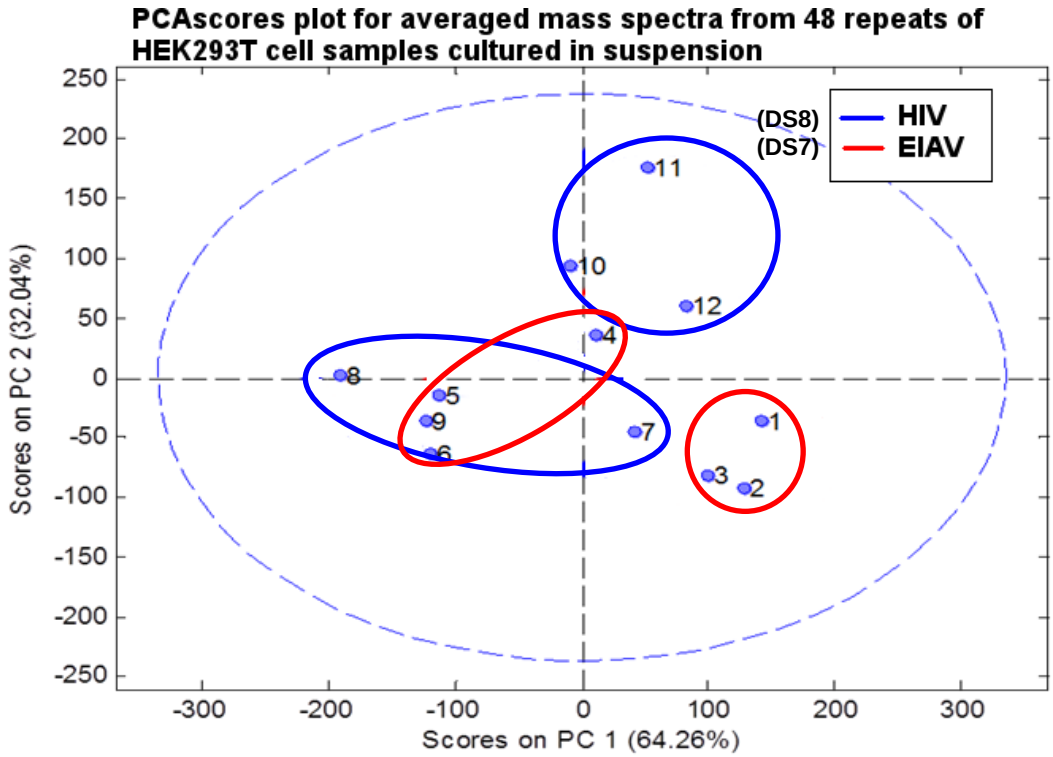


Figure 44: PCA scores plot (PC1 and PC2) of 12 averaged mass spectra from 48 suspension cell samples (Table 6). Spectra were averaged and no further pre-processing was applied. Circles indicate batches, red – transfected with EIAV, blue – transfected with HIV. Dataset used was DS7-8.

When using processed data (baseline correction and normalisation, no SG smoothening, Figure 45) the difference becomes even smaller with PC1 capturing over 99% of data variability and showing very little difference between individual samples. This data also shows that within the range used in the experiment cell concentration has little effect on the overall spectrum. PC1 score increases slightly with cell concentration for EIAV samples but the trend is not consistent for HIV samples.

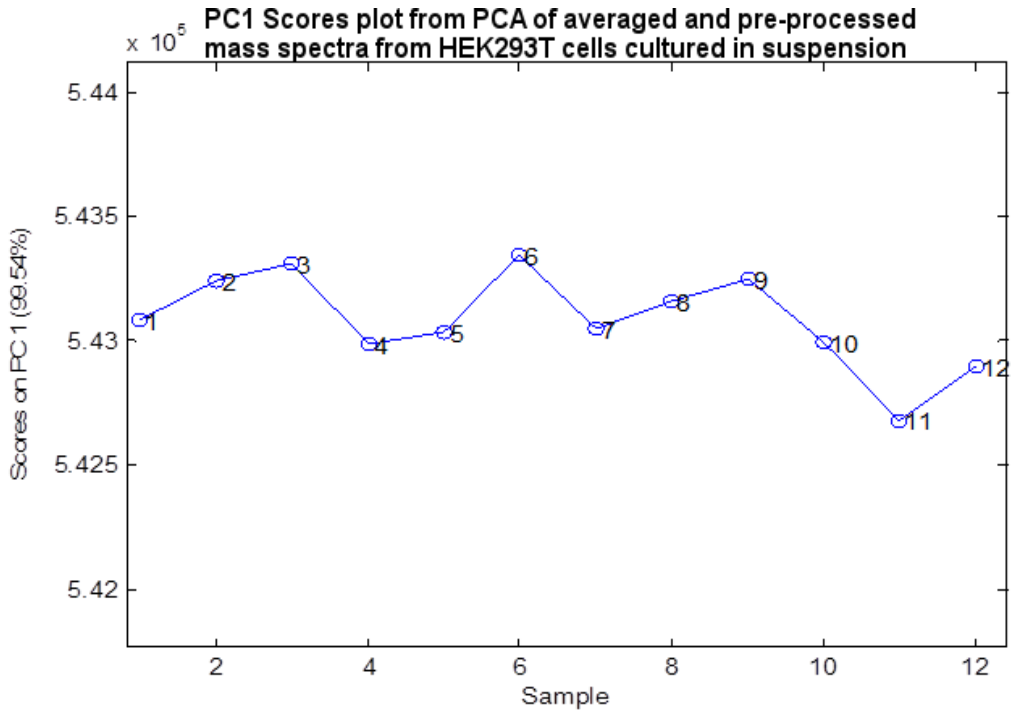


Figure 45: PCA scores plot (PC1) of 12 averaged and pre-processed mass spectra from 48 suspension cell samples (Table 6).
Mass spectra were pre-processed using baseline correction and data normalisation. Dataset used was DS7-8.

More extensive pre-processing (baseline correction, normalisation, SG smoothening with 1st order derivative, 2nd order polynomial, mean centring, Figure 46) was applied to this data set to examine any trends which are not detectable in the primary signal. This resulted in much less variability captured in the first two PCs; the results show separation of the individual batches but leave a significant amount of data variation unexplained by the first two PCs. Because of derivative and mean centring the results are difficult to interpret but the overall structure is similar to the previous analysis: there is little difference between HIV and EIAV transfected cells and individual batches. Based on this outcome and as discussed in section 3.3.2 SG derivative and mean centring was not used in further analysis. While the individual data sets form more distinct clusters, there is no consistent trend of either EIAV-transfected of HIV-transfected cells obtaining high or low scores on either PC. Based on PC loading values, PC1 scores are mainly driven by m/z signal in the ranges of 1200, 2300, 2600, 3300, 4800, 6000 while PC2 scores are driven by similar regions of the spectra but in a broader range. The loadings values and their effect on PCA are discussed further using a larger data set as part of MALDI-ToF MS robustness study in section 3.3.6.

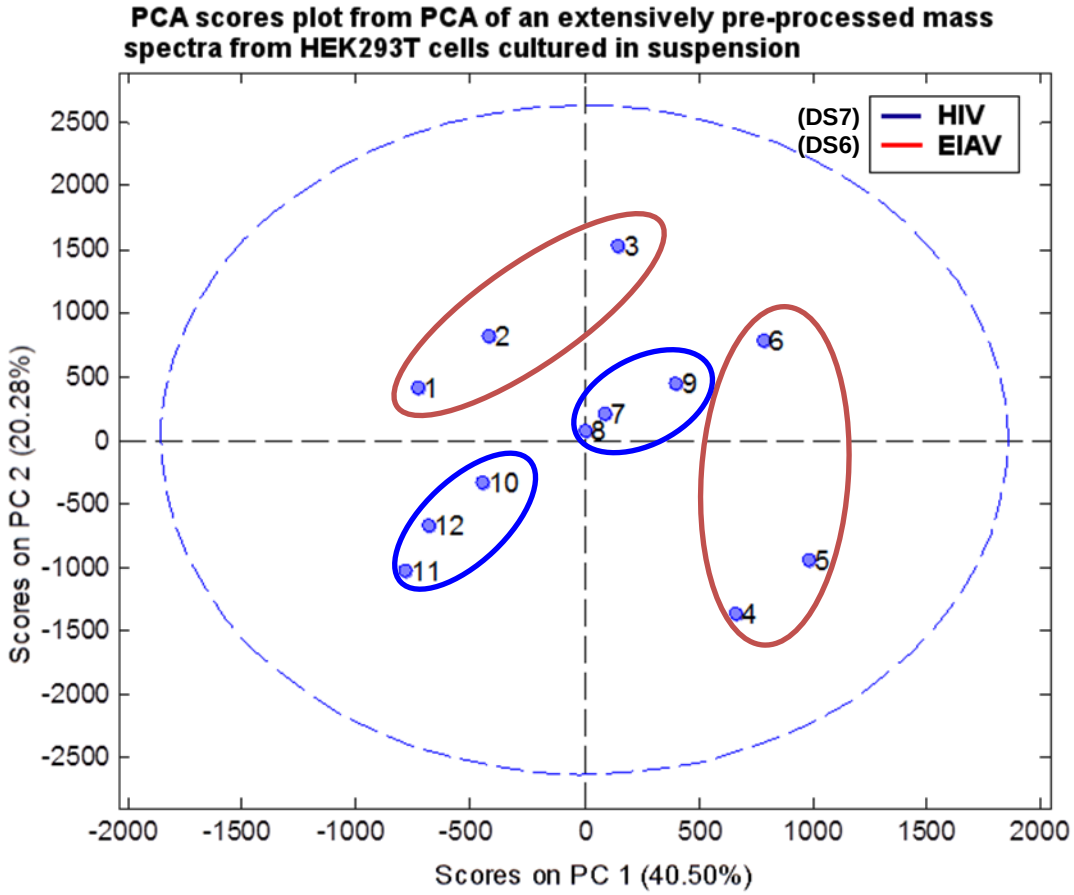


Figure 46: PCA scores plot (PC1 and PC2), of 12 averaged and extensively pre-processed mass spectra from 48 suspension cell samples (Table 6)
Mass spectra were pre-processed using baseline correction, data normalisation, SG smoothening (including 1st order derivative), mean centring and averaged data. Circles indicate batches, red – EIAV, blue – HIV. Dataset used was DS7 and DS8.

3.3.4. Adherent cells analysis

The next analysis step used cells grown adherently (DS1-2, samples 1-96, Table 7) to perform a similar analysis as was applied to suspension cells focusing on the variability of adherent cells, cell concentration effect and effect of delays in measurement. Adherent set contains 96 samples from 2 adherent batches of HIV-transfected cells and 2 adherent batches of EIAV-transfected cells each measured in 3 cell concentrations ($7x10^4$, $1x10^5$, $1.5\ x10^5$) with 4 replicates across 2 days (preparation described in section 3.2.2.).

Sample No (averaged)	Data set (Table 4)	Cell line (Table 4)	Virus produced	Cell concentration	Time before analysis
1-8 (1-2)	DS2	1	HIV	$7x10^4$	1h
9-16(3-4)	DS2	1	HIV	$1x10^5$	1h
17-24(5-6)	DS2	1	HIV	$1.5x10^5$	1h
25-32 (7-8)	DS1	1	EIAV	$7x10^4$	1h
33-40 (9-10)	DS1	1	EIAV	$1x10^5$	1h
41-48 (11-12)	DS1	1	EIAV	$1.5x10^5$	1h
49-56 (13-14)	DS2	1	HIV	$7x10^4$	24h
57-64 (15-16)	DS2	1	HIV	$1x10^5$	24h
65-72 (17-18)	DS2	1	HIV	$1.5x10^5$	24h
73-80 (19-20)	DS1	1	EIAV	$7x10^4$	24h
81-88 (21-22)	DS1	1	EIAV	$1x10^5$	24h
89-96 (23-24)	DS1	1	EIAV	$1.5x10^5$	24h

Table 7: List of adherent cell samples used in PCA analysis.
Samples were analysed in quadruplicate or averaged (values in brackets). For Figure 54 to Figure 56 additional 12 suspension samples are used for reference, shifting the numbering of adherent samples upwards by 12 (13-36 instead of 1-24).

Similar to the cell analysis from suspension data set, PCA of adherent cells shows little difference between cells transfected with HIV or EIAV based vector using raw MS spectra (Figure 47, red diamonds and blue triangles for HIV transfected cells, green squares and cyan inverted triangles for EIAV transfected cells). Likewise, there is no consistent trend based on cell concentration used for spotting on MS plate. This initial analysis was likely influenced by noise introduced by instrument and matrix variation. Pre-processing was applied to further analysis to reduce the noise as discussed previously for suspension samples. The main difference observed between the 4 groups of adherent samples is the time between sample spotting and MALDI-ToF MS analysis. As described in the methods section 3.2.2 and 3.2.6 as well as Table 7, samples 1-12 were analysed immediately after spotting while samples 13-24 were analysed 24 hours after spotting. In Figure 48 the 2 data sets have slightly different position in the PC scores plot, where samples from day 1 on average have higher score on PC1. However, the difference is small and the groups are still overlapping. The major trend (no difference between HIV and EIAV transfected cells) remains the same on both days.

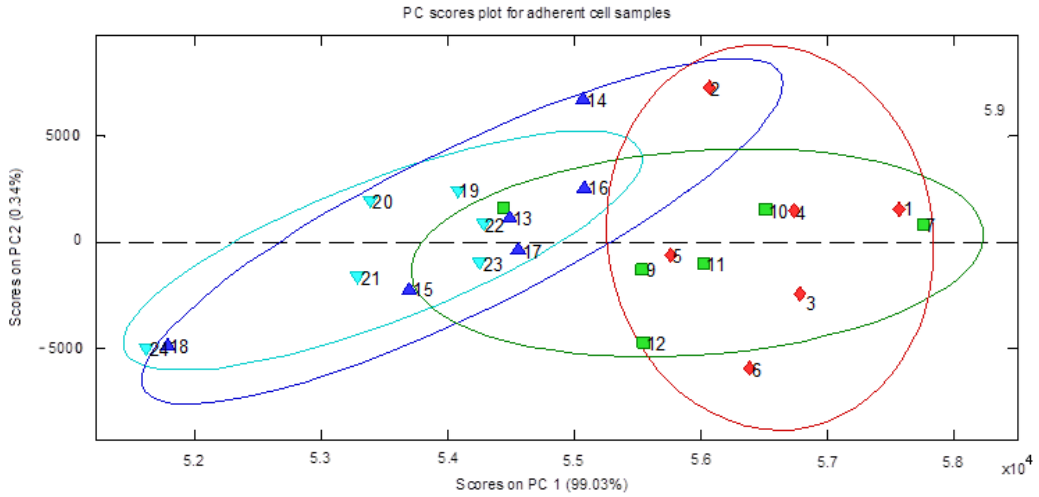


Figure 47: PCA scores plot (PC1 and PC2) for unprocessed mass spectra of 24 adherent cell samples incubated with α-cyano.
1-6: adherent cells transfected with HIV vector, analysed immediately after spotting; 7-12: as 1-6 but transfected with EIAV vector; 13-18: as 1-6 but analysed 24h after spotting; 19-24: as 7-12 but analysed 24h after spotting. Dataset used was DS1-2.

In order to examine the samples from 2 days in more detail, different pre-processing was applied. The data measured at two time points originated from the same biological source, therefore more extensive pre-processing was required to further reduce noise and focus on underlying data trends and differences between individual samples rather than absolute values of MS intensities. This included SG smoothening with 1st order derivative and mean centring of the data. As expected this resulted in changed data structure and less variability being captured (49.31% in the first 2 PCs for analysis of individual repeats (Figure 48) and 68.53% for data with averaged repeats (Figure 49). There is some difference in samples inspected on different days, on average the scores on PC1 are negative for first day data and positive for second day data indicating a clear separation (Figure 48 for data with individual repeats and Figure 49 for data with averaged replicates). This can also be illustrated as clustering on Figure 50 where the samples from two days are clearly separated except for the 2 outliers. This implies that a delay in mass spectrum acqusition has a significant effect on the mass spectra, either through changes in the matrix composition or the matrix effect on the samples over time once spotted on the plate. Longer exposure to the matrix solution has a stronger effect on the sample resulting in difference in MS spectra.

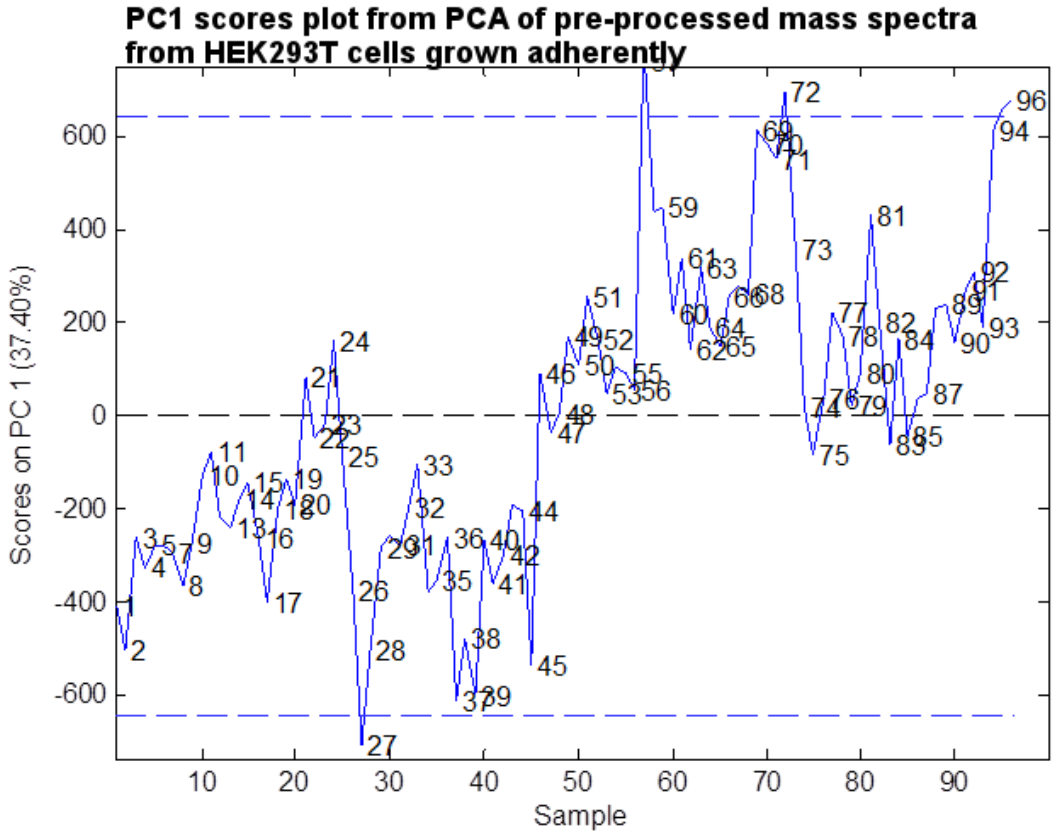


Figure 48: PCA scores plot (PC1) for pre-processed mass spectra of 48 adherent cell samples. Pre-processing involves baseline correction, data normalisation, SG smoothening (including 1st order derivative) and mean centring; samples (DS1-2) 1-48: analysed immediately after spotting, 49-96: analysed 24h after spotting.

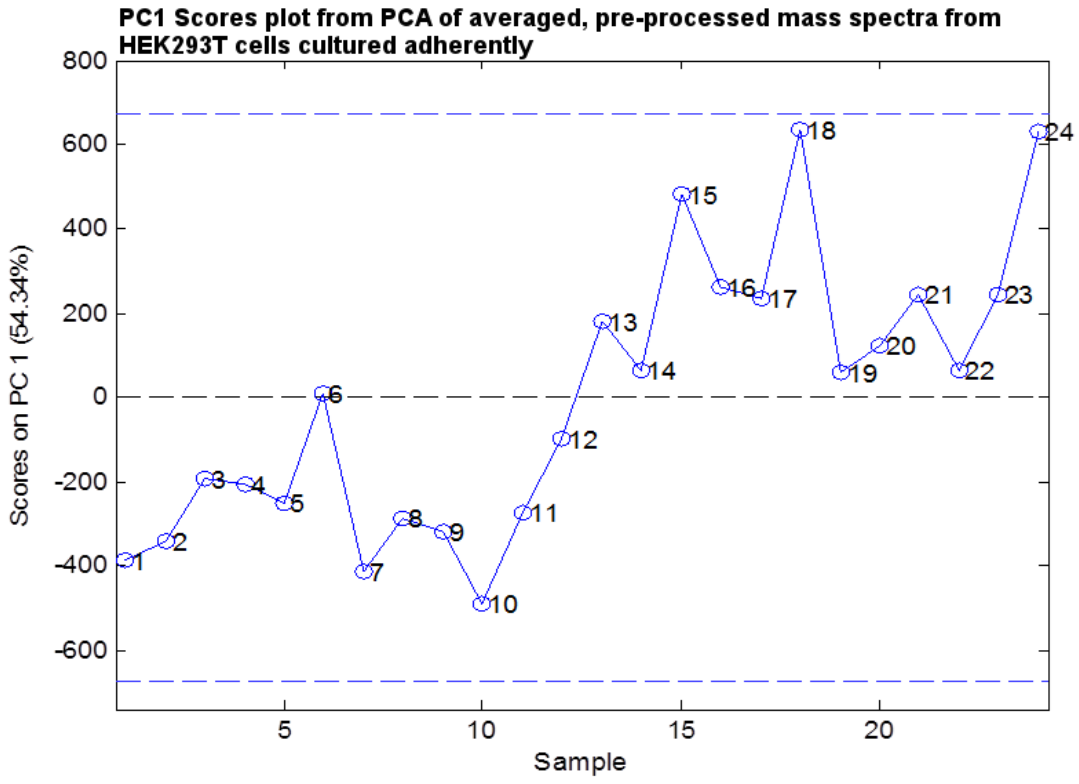
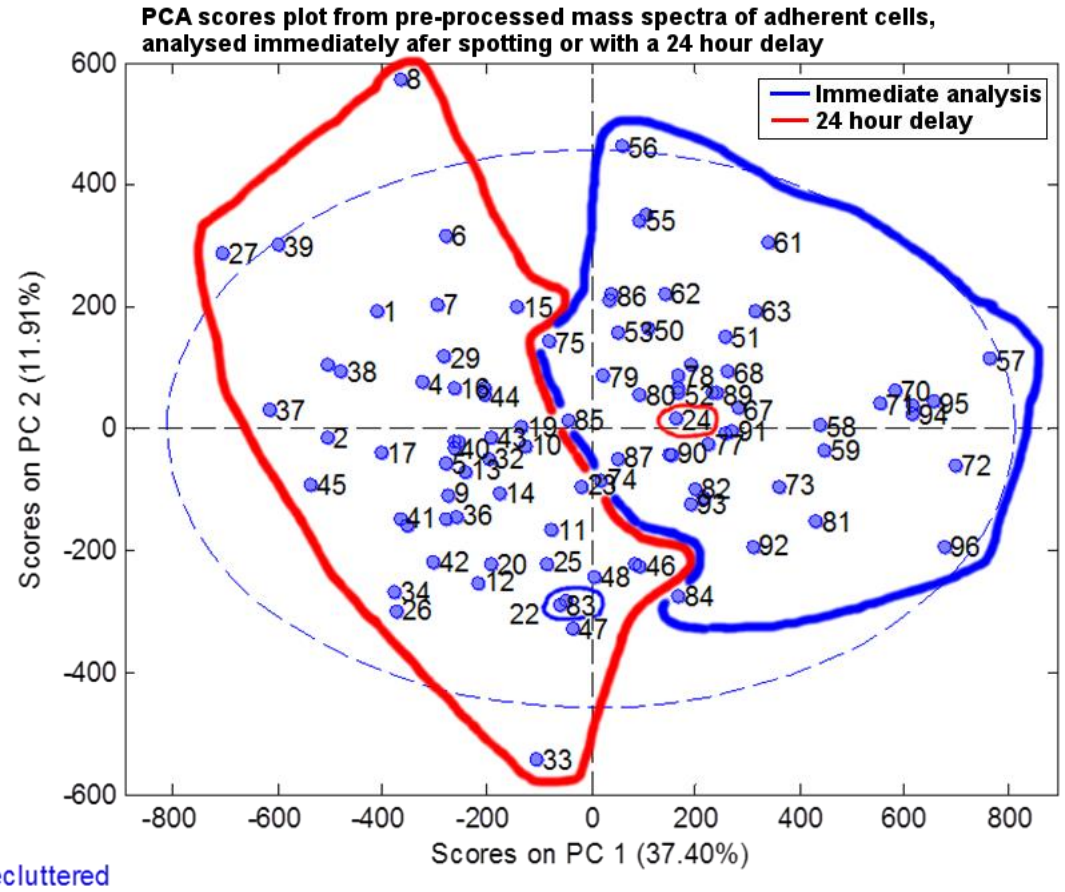


Figure 49: PCA scores plot (PC1) for averaged and pre-processed mass spectra of 48 adherent cells. Pre-processing involved baseline correction, data normalisation, SG smoothening (including 1st order derivative) and mean centring; samples (DS1-2) 1-12: analysed immediately after spotting, 13-24: analysed 24h after spotting.



Decluttered

Figure 50: PCA scores plot (PC1 and PC2) for pre-processed mass spectra of 48 adherent cells with highlighted clusters of samples analysed on day 1 and day 2.
Pre-processing included baseline correction, data normalisation, SG smoothening (including 1st order derivative) and mean centring; samples (DS1-2) 1-48 highlighted in red: analysed immediately after spotting; 49-96 highlighted in blue: analysed 24h after spotting (Table 7).

Further analysis of averaged data highlights the difference between the two days even more strongly (Figure 51). At the same time data from both days retains a similar shape on PC2 score plot (Figure 52). To illustrate this similarity, samples 13-24 were interposed over samples 1-12 in a separate graph (Figure 53). This similarity is reflected by clusters on Figure 51 retaining their overall shape while shifting their position. Interestingly the samples taken using higher cell count are always in the bottom-right part of their cluster, suggesting that the cell count affects the spectrum, however the effect is small and it is difficult to determine what concentration is optimal.

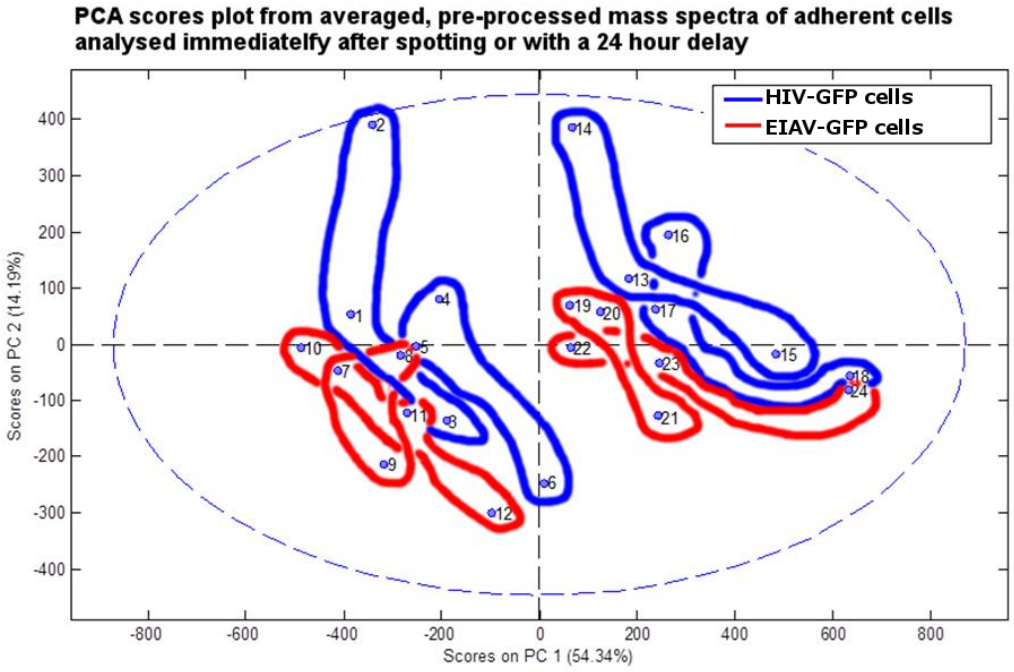


Figure 51:PCA scores plot (PC1 and PC2) for averaged and extensively pre-processed mass spectra of 48 adherent cell samples with highlighted clusters of samples analysed on day 1 and day 2.
Pre-processing included baseline correction, normalisation, SG smoothening (including 1st order derivative) and mean centring. Blue - HIV transfected cell batches, red - EIAV-transfected batches; immediately analysed data clustered on the left, data analysed 24h after spotting is clustered on the right. Dataset used was DS1-2.

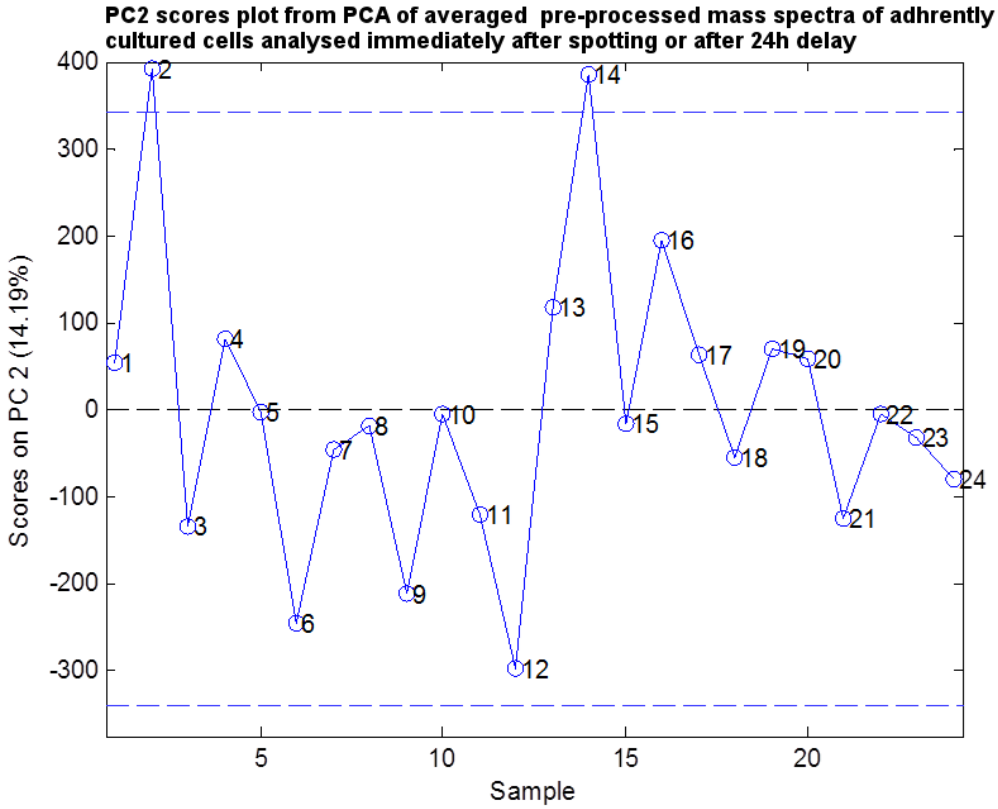


Figure 52: PCA scores plot (PC2) for averaged and extensively pre-processed mass spectra from 48 adherent cell samples.
Pre-processing included baseline correction, data normalisation, SG smoothening (including 1st order derivative) and mean centring. Data analysed immediately after spotting (1-12) and analysed 24h after spotting (13-24). Dataset used was DS1-2.

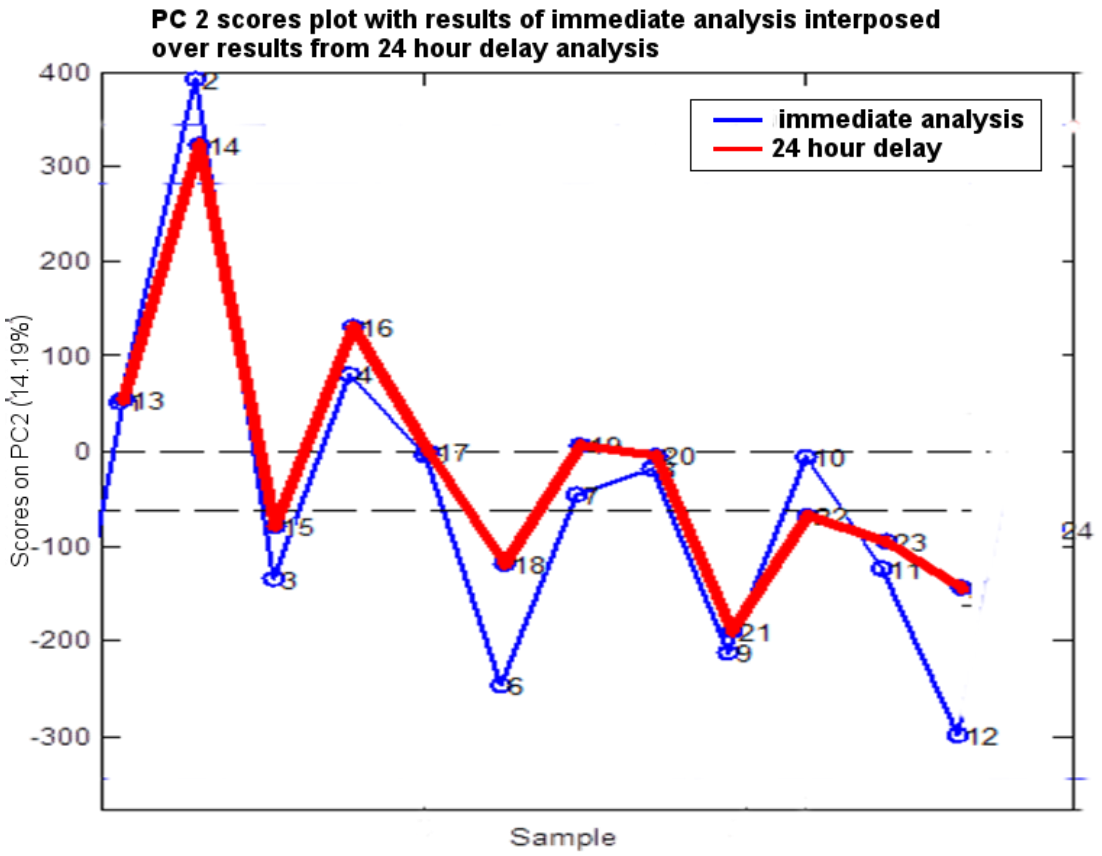


Figure 53: Interposed scores plot (PC2) for adherent cells analysed immediately or after 24h. Mass spectra were averaged and extensively pre-processed (Baseline correction, normalisation, SG smoothening (including 1st order derivative) and mean centring). Data was analysed immediately after spotting (1-12, blue) and 24h after spotting (13-24, red). Dataset used was DS1-2.

Overall, the results suggest that the time spent between sample spotting and measurement can affect the data but at the same time the data retains its overall structure over the 1 day waiting period. This suggests that performing MS analysis can be done a day after spotting the samples and retain the underlying information. However, the time between spotting and analysis should remain consistent between experiments in order to compare them. Moreover, this effect of time is mostly prominent in data pre-processed with 1st order derivative and mean centring which was used specifically to investigate this difference and was not the pre-processing approach applied to the majority of data (see section 3.2.7).

Finally, analysis of the entire data set for cell samples (i.e. combined adherent and suspension cell sample data) compares all variables with the focus on differences between adherent and suspension cells. Addition of suspension samples introduces more diversity to the MS data set and therefore further analysis follows the standard approach to pre-processing as described in section 3.2.7 (no derivative was used during SG smoothening). Based on PC1 (93.94% variance captured, Figure 54) there is a significant difference between

suspension cells data and adherent cells data from either day one or 2. This trend is clearer on PC scores plot (Figure 55) where the suspension dataset is clearly separated from adherent dataset. As described before, the two adherent cells datasets (analysed immediately after spotting or 24h after spotting) can be distinguished, however they overlap each other and they are both clearly separated from suspension cells (Figure 54 and Figure 55).

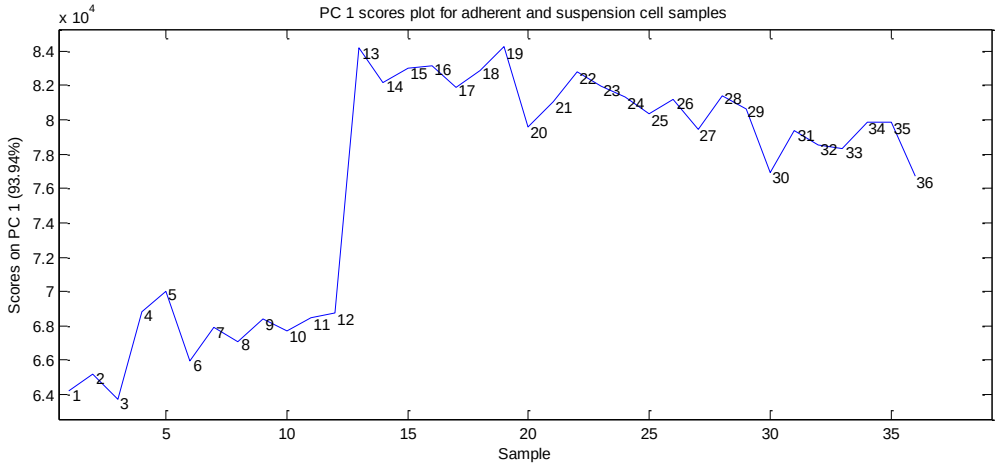


Figure 54: PCA scores plot (PC1) for 36 mass spectra of adherent and suspension cell samples (averaged, pre-processed).
Pre-processing included baseline correction, data normalisation and SG smoothening (no derivative). Samples 1-12: suspension cells (DS7-8). Samples (DS1-2) 13-24: adherent cells (day 1); Samples 25-36: adherent cells (day 2).

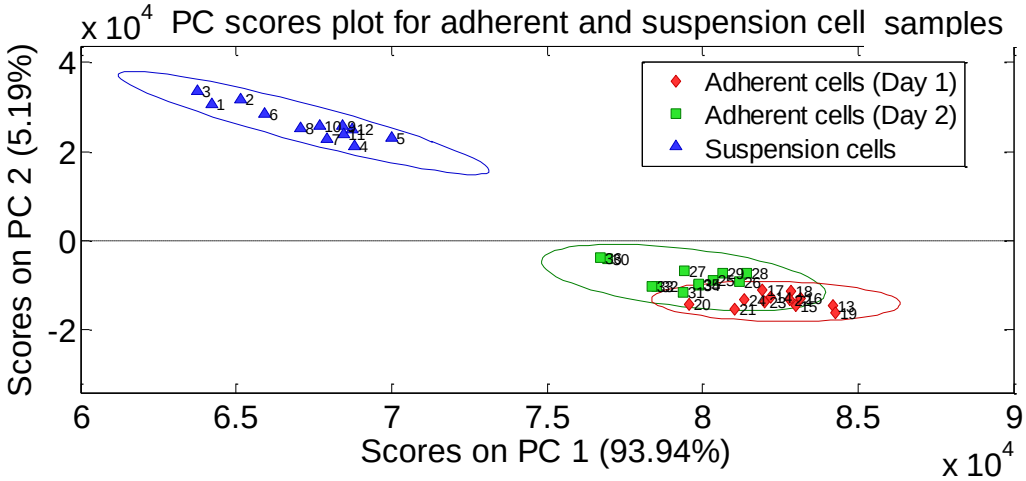


Figure 55: PCA scores plot (PC1 and PC2) for 36 mass spectra of adherent and suspension cell samples (averaged, pre-processed).
Pre-processing included baseline correction, normalisation and SG smoothening (no derivative). Samples 1-12: suspension cells (DS7-8). Samples (DS1-2) 13-24: adherent cells (day 1); Samples 25-36: adherent cells (day 2).

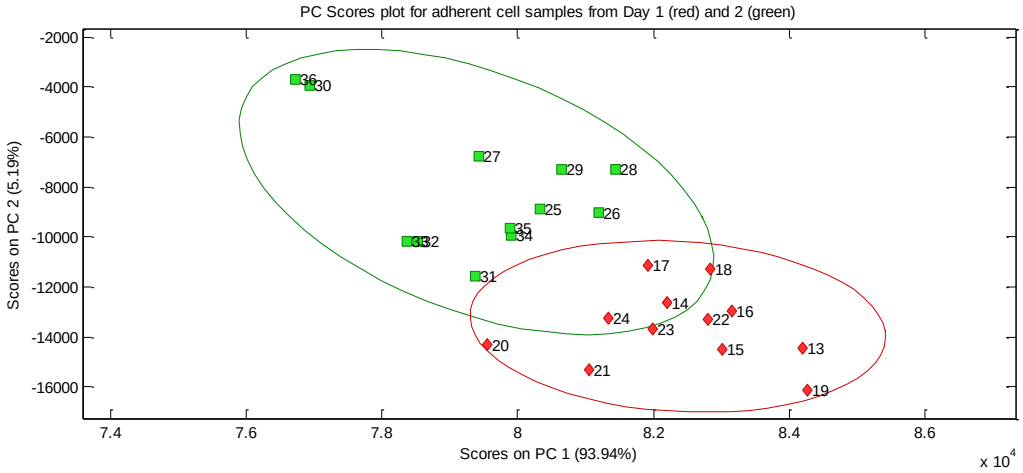


Figure 56: PCA Scores plot from Figure 55 zoomed in on data points from adherent cell samples. Day 1 samples (13-24) are slightly shifted compared to day 2 samples (25-36). Dataset used was DS1-2.

3.3.5. Effect of cell line variation

The next goal of the MALDI-ToF MS study was to assess the effect of higher number of repeats per sample and to measure variation between samples from different cell lines. Samples obtained for this experiment are from 4 different cell lines (DS3, DS4, DS5 and DS6, Table 4), cultured adherently. For each cell line 3 samples were taken and each sample was spotted on the plate 10 times. PCA was used to examine the variation within cell line (Figure 57-Figure 60) as well as between all the cell lines (Figure 61 for all repeats Figure 62 for averaged spectra). The results for cell samples are similar for all 4 cell lines: the individual repeats are concentrated around a central point while several points are more distant, representing repeats of higher variation and outliers. The sample to sample variation is low in cell lines 1 and 2 (samples 1-10 represent first sample, 11-20 second and 21-30 third) as individual repeats from the different samples are overlapping randomly. The situation is slightly different for cell lines 3 and 4 where the repeats from different samples seem to form broad and overlapping groups, in both cases lower repeat number samples have slightly lower values on both axes (PC1 and PC2 scores) while the higher number repeats have higher score values. This could be associated with slight differences in incubation time and spotting of the different samples, however the impact of this variable should be minimal and the difference is more likely attributed to sampling variation.

Sample No	Data set (Table 4)	Cell line (Table 4)	Virus produced	Cell concentration	Time before analysis
1.1-1.30	DS3	1	None	$1.5x10^5$	24h
2.1-2.30	DS4	2	None	$1.5x10^5$	24h
3.1-3.30	DS5	3	None	$1.5x10^5$	24h
4.1-4.30	DS6	4	None	$1.5x10^5$	24h

Table 8: List of adherent samples from 4 different cell lines used in PCA analysis of different cell line samples. Each biological sample was analysed using 10 technical replicates for a total of 30 spots per cell line.

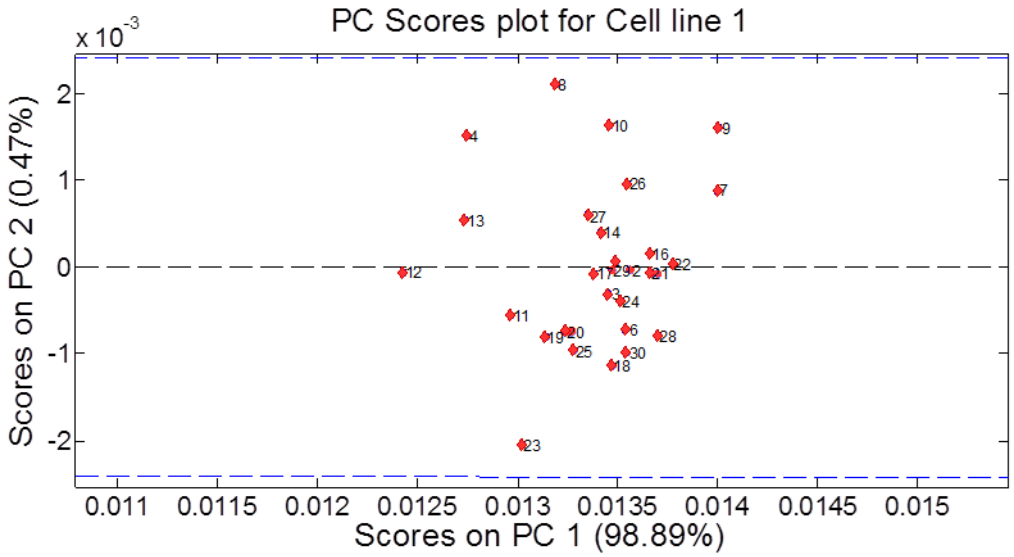


Figure 57: PCA scores plot (PC1 and PC2) for mass spectra of cell line 1.
The cells were grown adherently (DS3, see Table 4); the figure shows all individual repeats. Repeats 1-10 are for 1st sample, 11-20 for 2nd sample and 21-30 for 3rd sample. Spectra were pre-processed (baseline correction, normalisation, SG smoothening).

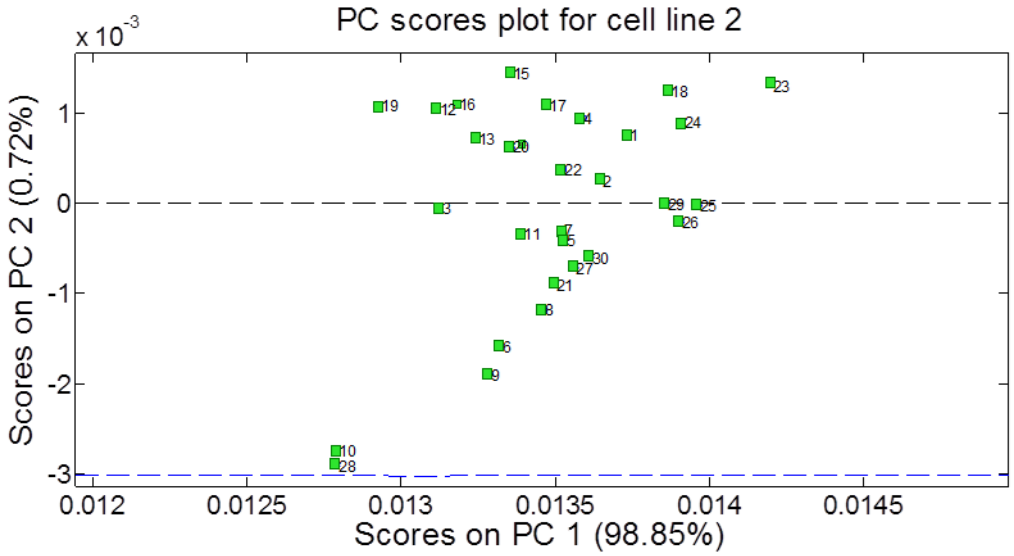


Figure 58: PCA scores plot (PC1 and PC2) for mass spectra of cell line 2.
The cells were grown adherently (DS4, see Table 4); the figure shows all individual repeats. Repeats 1-10 are for 1st sample, 11-20 for 2nd sample and 21-30 for 3rd sample. Spectra were pre-processed (baseline correction, normalisation, SG smoothening).

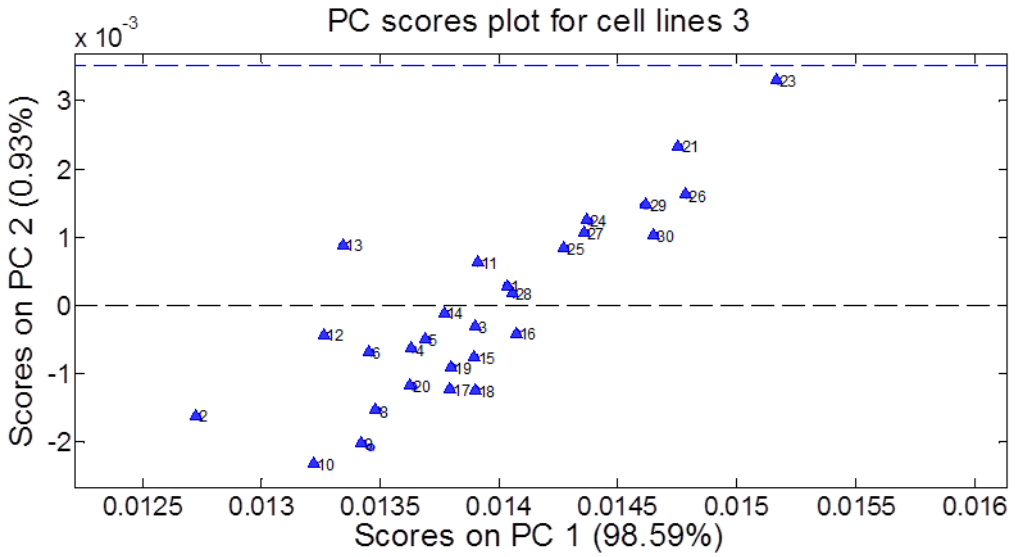


Figure 59: PCA scores plot (PC1 and PC2) for mass spectra of cell line 3.
The cells were grown adherently (DS5, see Table 4); the figure shows all individual repeats. Repeats 1-10 are for 1st sample, 11-20 for 2nd sample and 21-30 for 3rd sample. Spectra were pre-processed (baseline correction, normalisation, SG smoothening).

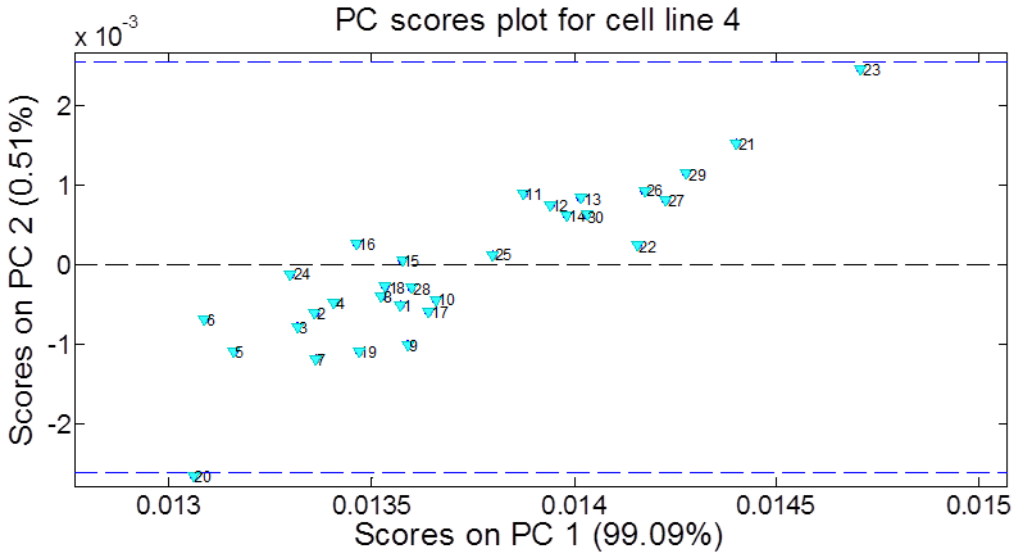


Figure 60: PCA scores plot (PC1 and PC2) for mass spectra of cell line 4.
The cells were grown adherently (DS6, see Table 4); the figure shows all individual repeats. Repeats 1-10 are for 1st sample, 11-20 for 2nd sample and 21-30 for 3rd sample. Spectra were pre-processed (baseline correction, normalisation, SG smoothening).

The analysis of this data set shows a trend where individual technical replicates from all cell lines are overlapping and it is difficult to separate individual clusters. At this point the variation between individual repeats is too high to distinguish individual cell line samples (Figure 61). However when the repeats are averaged to give a single spectra for each sample the samples from each cell lines form small groups which allow to clearly distinguish the cell lines (Figure 62). None of the 4 cell line groups are overlapping and they have distinctly different PC score values while the 3 samples for each cell line remain close to each other (except for cell line 4 which is more variable)

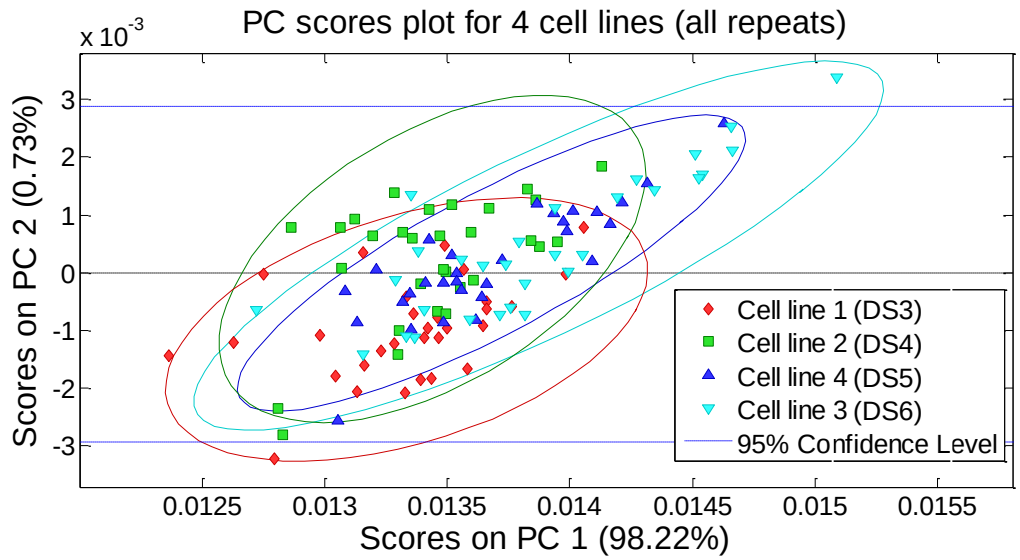


Figure 61: PCA scores plot (PC1 and PC2) for mass spectra of 4 adherent cell lines. Datasets used are DS3-6, see Table 4. Spectra were pre-processed (baseline correction, normalisation, SG smoothening).

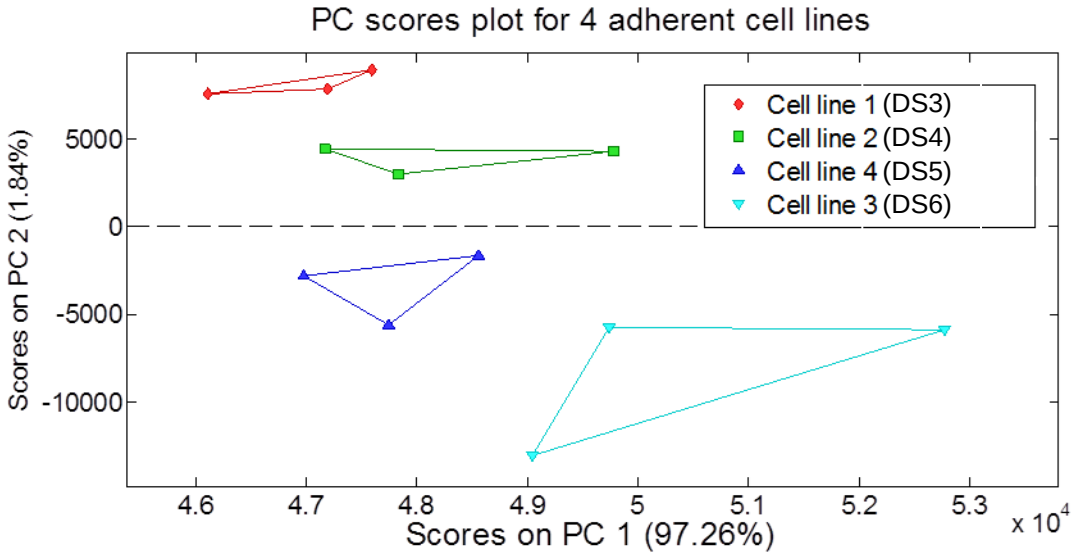


Figure 62: PCA scores plot (PC1 and PC2) for mass spectra of 4 adherent cell lines (averaged). Datasets used are DS3-6, see Table 4. Spectra were averaged for each sample (10 repeats per sample) and pre-processed (baseline correction, normalisation, SG smoothening). All samples are within 95% confidence limit (not visible due to zoom level).

3.3.6. Overall MALDI-ToF MS robustness study

The final step in the MALDI-ToF MS robustness study was a comparison analysis between all available cell data sets (DS1-8, DS12 and DS14), focused on assessing the variability within and between suspension and adherent cells as well as individual cell lines. The aim of analysing a wide array of cell and vector samples with MALDI-ToF MS was to establish a robust method capable of generating reproducible results which can be used to characterise and model the cells and vectors. The optimal MALDI-ToF matrix composition (0.15% α-cyano with 0.15% TFA, see section 3.2.6), sample preparation (10 spots per sample, matrix incubation for 1h before spotting, see section 3.2.6), laser settings (see section

3.2.6) and signal pre-processing (averaging of technical repeats, baseline correction, data normalisation and SG smoothening, see section 3.2.7) were determined as described in the previous sections 3.3.1 - 3.3.5. The robustness study presented below outlines the results of PCA of all cell data sets available to-date, produced in suspension (ambr®(DS12 and DS14), MiniBio (DS7-8)) or adherently (culture plates (DS3-6), cell factories (DS1-2)) ranging from small to large scale production with a total of 61 samples (Figure 63-Figure 64). A close examination of some of the major peaks (Figure 65a-b) shows that there is a difference between suspension and adherent cell samples, where suspension cells result in a broader peak with overall lower intensity. This trend is observed for both ambr®15 and Applikon® MiniBio reactor samples while all adherent samples are characterised by sharper, higher intensity peaks in theses variable regions of the spectra. PCA was performed to thoroughly examine the differences between multiple cell samples.

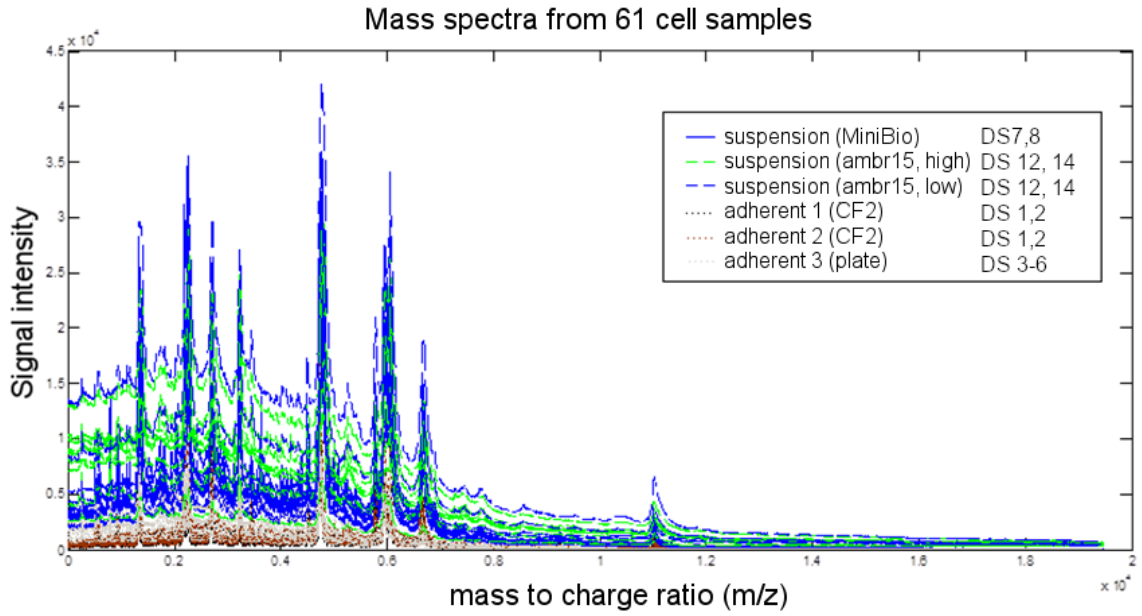


Figure 63: Mass spectra of 61 cell samples taken across different scales.
Samples were obtained from adherent (DS1-6) and suspension cell cultures (DS7-8, 12 and 14 as summarised in Table 4).

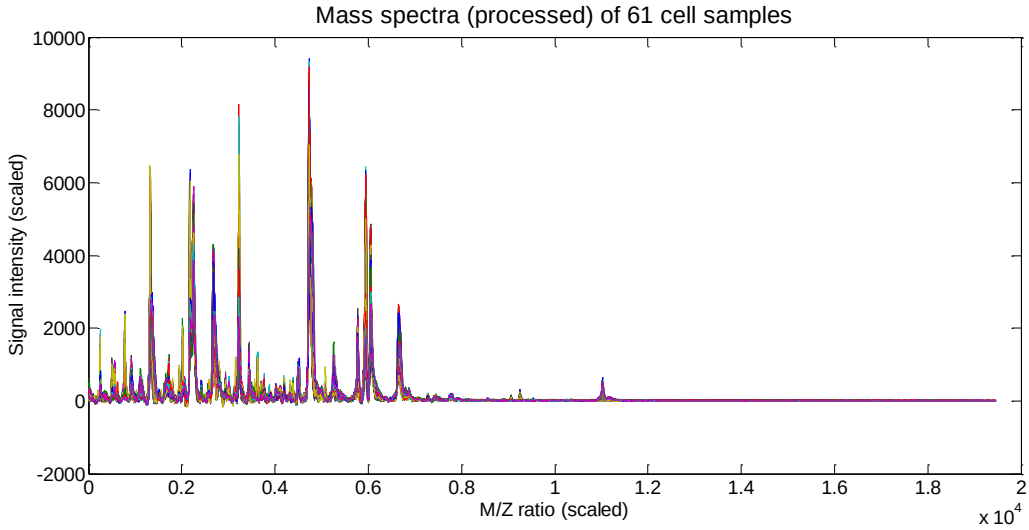


Figure 64: Mass spectra of 61 cell samples taken across different scales (pre-processed).
Processing included baseline correction, signal normalisation, Savitzky-Golay smoothening and peak alignment.
Samples were obtained from adherent(DS1-6) and suspension cell cultures ((DS7-8, 12 and 14 as summarised in Table 4).

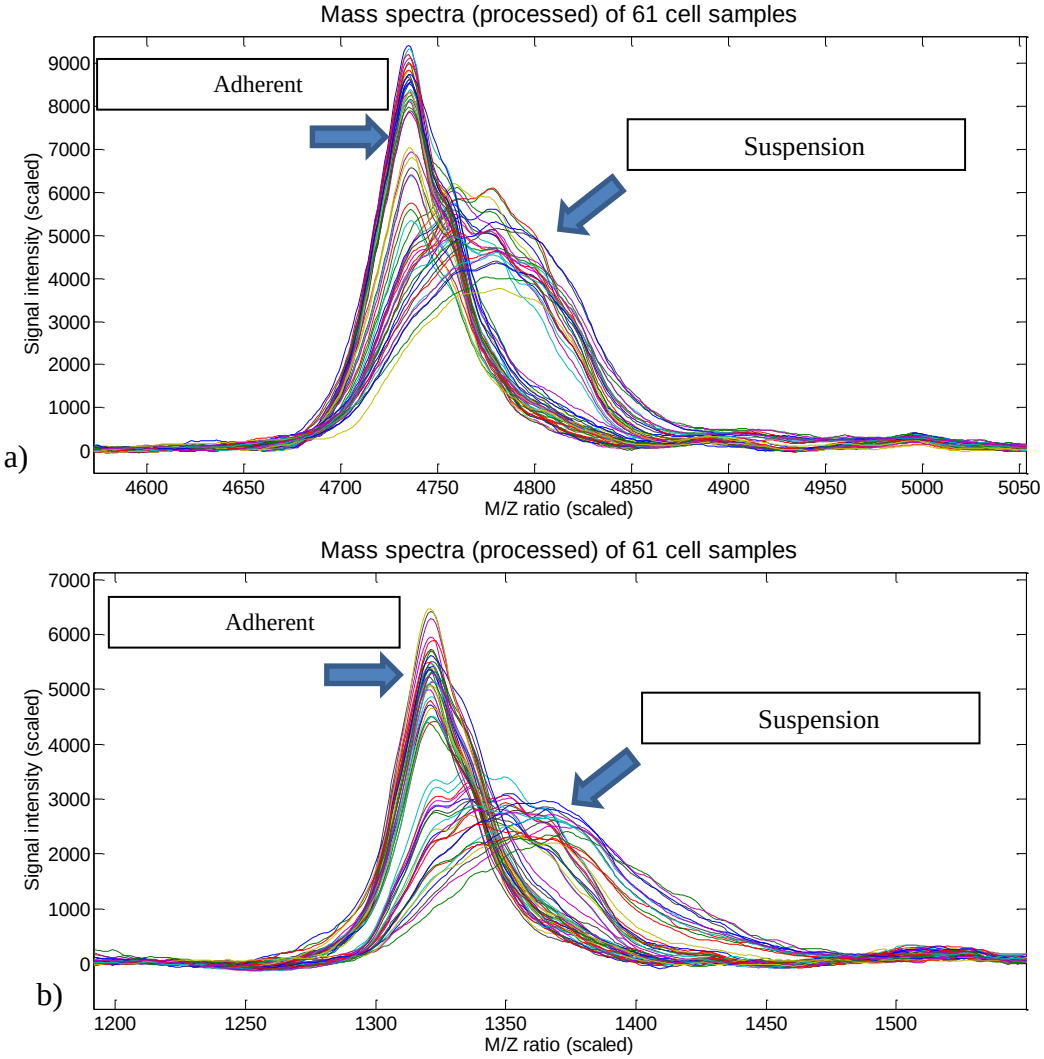


Figure 65a-b: Processed mass spectra of 61 cell samples zoomed on two regions with significant variation between adherent and suspension spectra.
Samples were obtained from adherent (DS1-6) and suspension cell cultures (DS7-8, 12 and 14 as summarised in Table 4).

The overall results of PCA are presented in Figure 66. It can be clearly observed that all suspension samples are separated from adherent samples. This is mainly observed due to the difference of scores on PC2 while PC1 scores of some of the adherent and suspension samples are similar. The variance captured by PC1 is mostly associated with the overall shape and intensity of the spectra as indicated by the loadings plot of PC1 and the average mass spectra where PC1 scores are driven primarily by several major peaks around m/z values of 1200, 2300, 2600, 3300, 4800 and 6000 (Figure 67). PC1 scores indicate that there are differences in spectra intensity and shape but this information alone is not sufficient to analyse the cell samples and it does not capture some of the important variance between the samples. PC2 scores contribute to the distinction between suspension and adherent cells as already demonstrated in Figure 66. The loadings plot for PC2 (Figure 68) shows the regions of the mass spectra which contribute to the separation of suspension and adherent cells on the PC1 vs PC2 scores plot (Figure 66). For several peaks a very sharp change from negative to positive value can be observed indicating variation in these peaks captured in PC2. This indicates that there are regions of the spectra where a population of samples is characterised by low signal intensity relative to the remaining samples followed by an increase in signal intensity relative to the other samples. This behaviour reflects the relationship between adherent and suspension cells' signals (Figure 65). This could be caused by a shift in the spectra position for some of the samples which would result in peak misalignment. However, the spectra were pre-processed and for both types of the cells there are regions in the spectra which are completely aligned. The more likely explanation is that the difference is caused by the flattened peak of suspension cell samples discussed before (Figure 65). This would explain an initially low signal for the suspension cells which then increases relative to the sharp peak of the adherent cells.

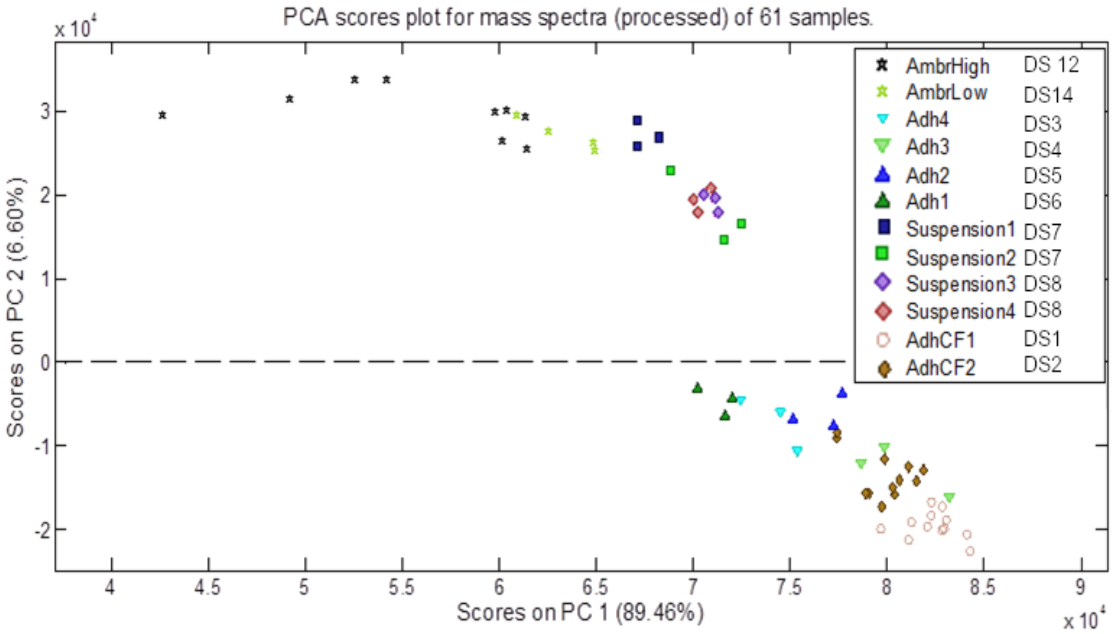


Figure 66: PCA scores plot (PC1 and PC2) for processed mass spectra of 61 cell samples.
Samples were obtained from adherent (DS1-6) and suspension cell cultures using cell lines (DS7-8, 12 and 14 as summarised in Table 4).

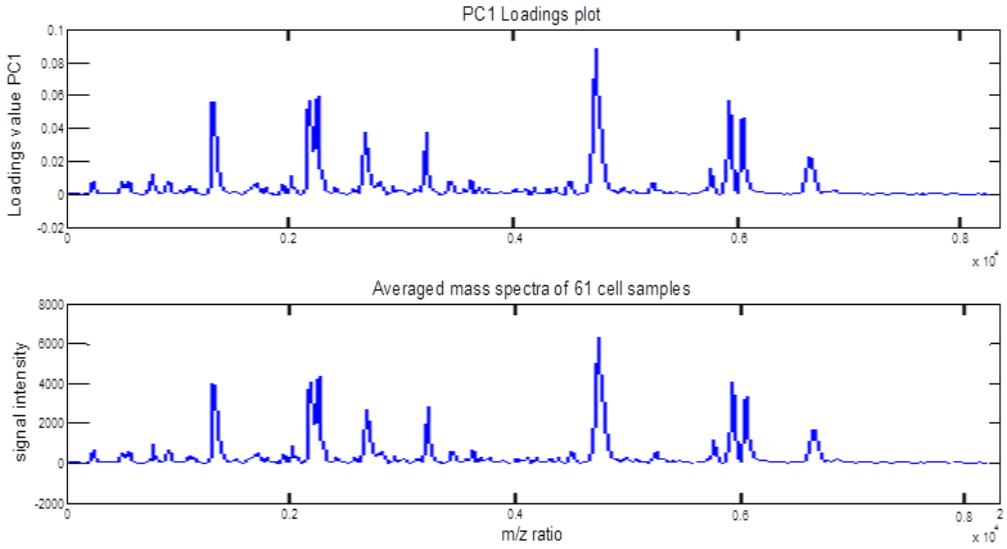


Figure 67: Comparison of PC1 loadings value and signal intensity of an averaged spectrum.
Top: PC1 loadings plot of the mass spectra of 61 cell samples. Bottom: Averaged mass spectra of the 61 cell samples.
Samples were obtained from adherent (DS1-6) and suspension cell cultures using cell lines (DS7-8, 12 and 14 as summarised in Table 4).

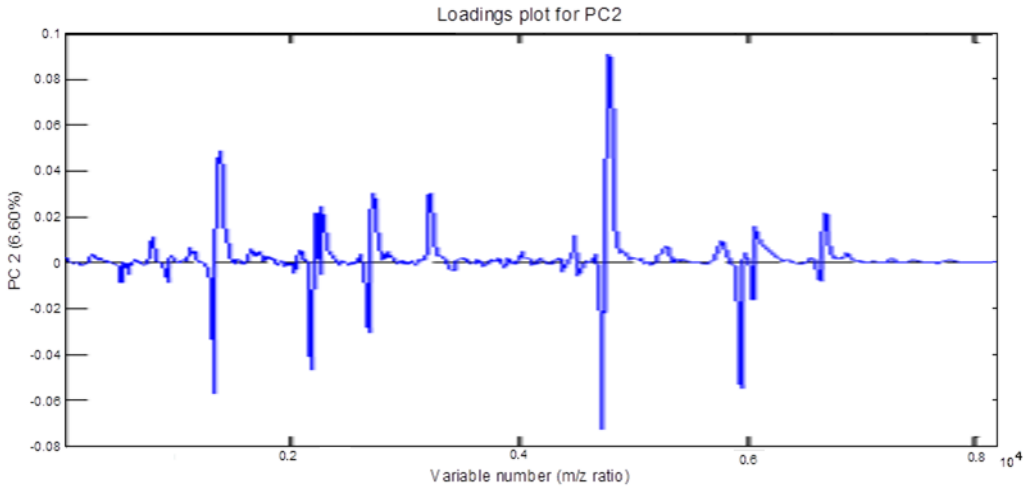


Figure 68: PCA Loadings plot (PC2) for the mass spectra of 61 cell samples.
Samples were obtained from adherent (DS1-6) and suspension cell cultures using cell lines (DS7-8, 12 and 14 as summarised in Table 4).

The last PC used in the model development is PC3 (Figure 69). Scores on this PC have high values for the 4 adherent cell lines cultured in 10 cm^2 plates along with several of the ambr®15 samples. The overall variance captured in PC3 is quite low (2.41%) and loadings plot (Figure 70) shows there is a significant impact of a single peak leading to positive values in this PC. Due to the low variance captured and only several samples showing positive score while the others are relatively uniform in value, it is difficult to interpret these results.

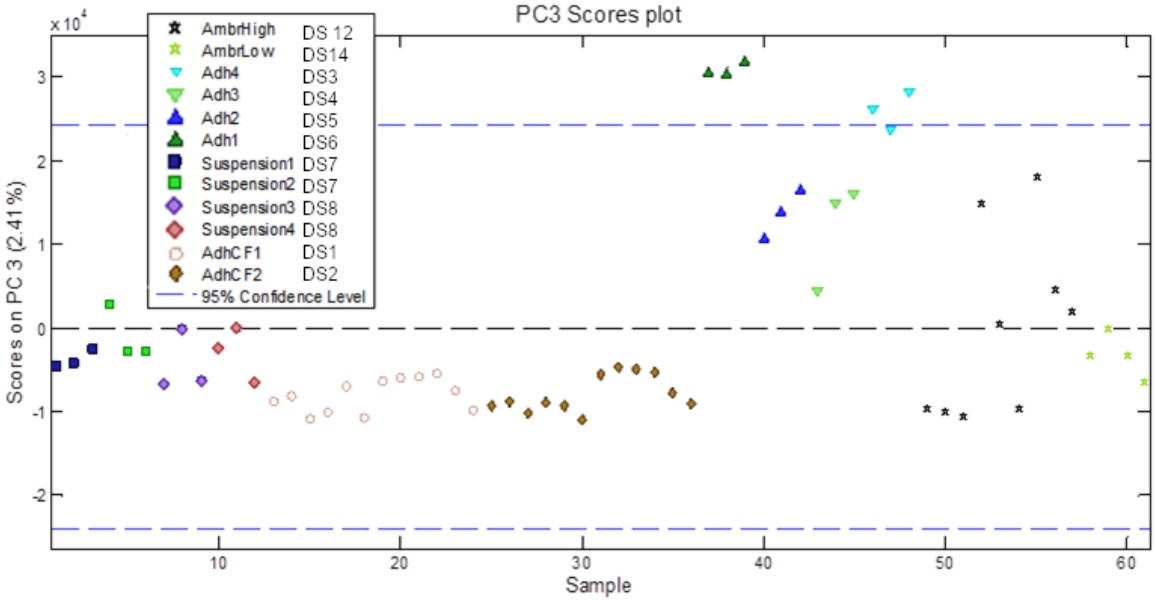


Figure 69: PCA Scores plot (PC3) for the mass spectra of 61 cell samples.
Samples were obtained from adherent (DS1-6) and suspension cell cultures using cell lines (DS7-8, 12 and 14 as summarised in Table 4).

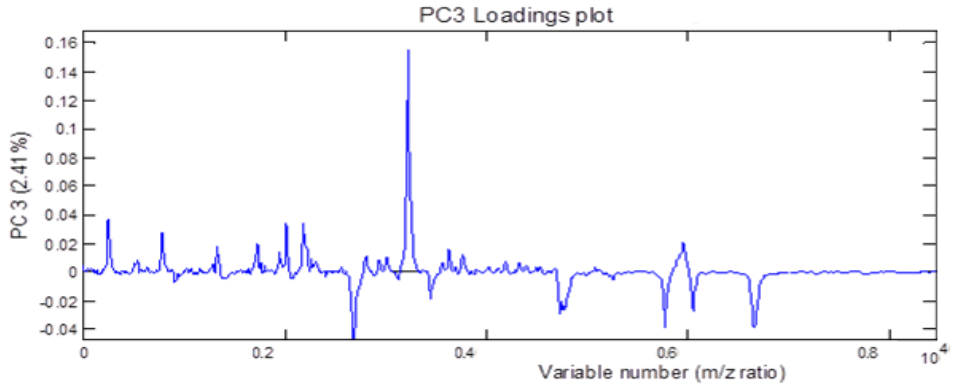


Figure 70: PCA loadings plot (PC3) for the mass spectra of 61 cell samples.
Samples were obtained from adherent (DS1-6) and suspension cell cultures using cell lines (DS7-8, 12 and 14 as summarised in Table 4).

Following the inspection of the PCs, PC1 and PC2 have been chosen to be used for visual inspection of the results with 96.06% of variance captured in these first two PCs. As outlined on Figure 71, the individual data sets can be clustered together. There is a significant amount of overlap between some of the data sets indicating a high degree of similarity between the particular mass spectra. The two ambr® samples clusters are close to each other similar to the 4 datasets obtained from Applikon® MiniBio reactors. A similar situation is observed for adherent cells where cell factory samples are overlapping with each other while the culture plate samples obtained from 4 separate cell lines are spaced slightly further from each other. Interestingly, one of the culture plate samples is overlapping with the cell factory samples, most likely due to the shared cell line (cell line 1, refer to Table 4). This indicated that both cell line and scale of culture have an impact of the protein profile measured by MS; however, the difference between adherent and suspension culture remains the main factor separating the clusters where there is no overlap between adherent and suspension cell samples.

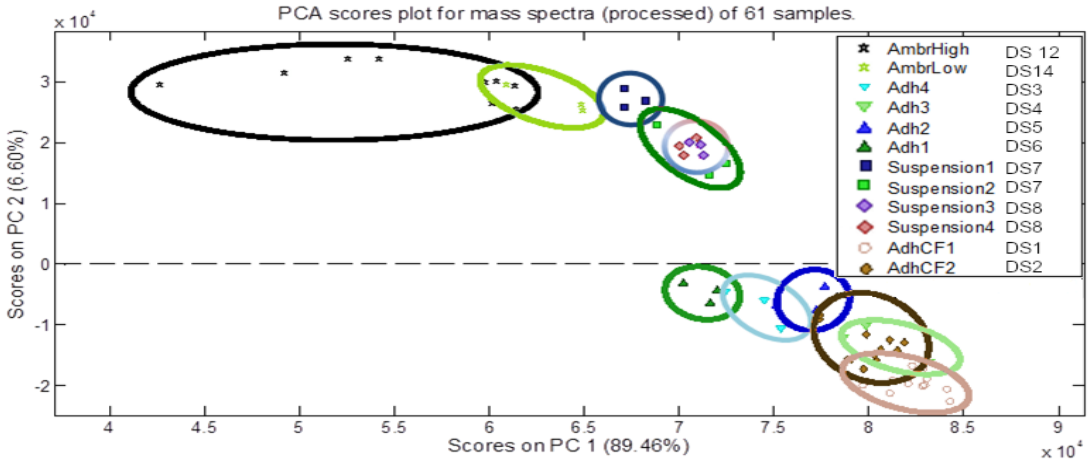


Figure 71: PCA scores plot (PC1 and PC2) for the pre-processed mass spectra of 61 cell samples with highlighted clusters.
Samples were obtained from adherent (DS1-6) and suspension cell cultures using cell lines (DS7-8, 12 and 14 as summarised in Table 4).

In order to assess the model quality and therefore the confidence in the presented results, values of Q residuals and Hoteling's T^2 were examined in an influence plot (Figure 72). Overall, 3 samples were identified with Q residual value above 1, indicating potential outliers which have higher amount of variability left unexplained by the model's PCs 1-3. There were 6 data points with Hoteling's T^2 between 0.5 and 1 which is higher than the remaining results. These 6 data points are the same adherent cell samples which were characterised by a high PC3 value suggesting that there is a degree of difference between these data points when compared to the rest of the data.

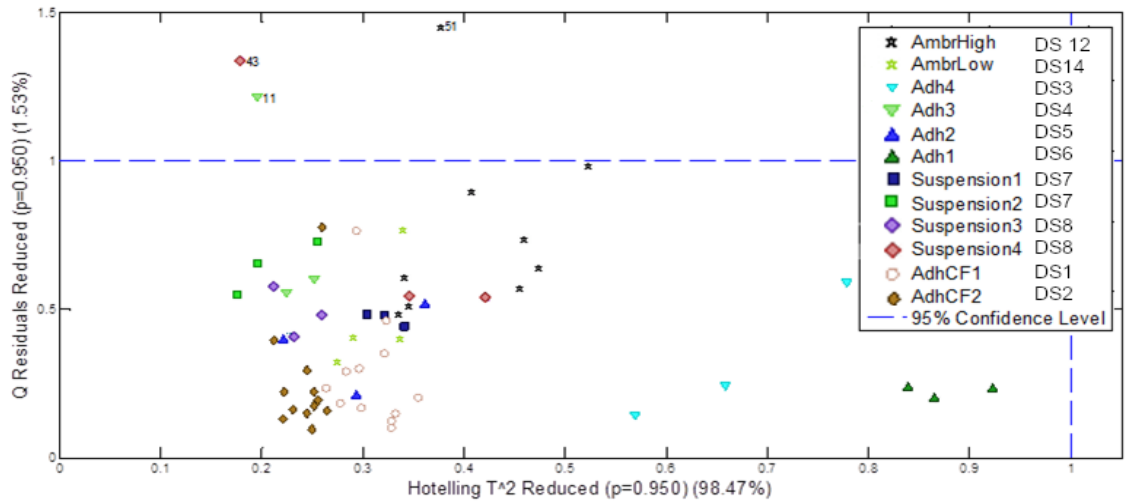


Figure 72: Hoteling's T^2 and Q residuals plot for the PCA of the mass spectra of 61 cell samples. Samples were obtained from adherent (DS1-6) and suspension cell cultures using cell lines (DS7-8, 12 and 14 as summarised in Table 4).

3.3.7. Viral vector samples analysis

For the vector analysis the important factors to compare are the virus type (EIAV/HIV), growth mode of cells used to produce it (adherent, suspension) use of buffer (α-cyano (buffer B) and SA (buffer D), both with 0.15% TFA), effect of the time difference between experiments (MALDI analysis immediately after spotting or 24h after spotting) and viral vector concentration. The viral samples were produced using cells and process from DS1-2 and, DS7-8 (See Table 4).

The comparison between different samples revealed that EIAV based vector (DS1, DS6) tends to give more variable results, where the use of different buffers had a major impact on positioning on the PC scores plot (Figure 73). HIV based vector samples (DS2 and DS7) were more consistent where neither difference in buffer nor the cells used for vector production (adherent or suspension) led to a significant difference in data structure. Additionally, for EIAV samples incubated with SA (buffer D) there were major differences

between vector produced in adherent or suspension cells. These changes were absent from samples incubated with α-cyano (buffer B).

A second model using only α-cyano (buffer B, Figure 74) again shows that EIAV samples are more variable where vector produced in suspension cells has a different position than vector produced in adherent cells. Interestingly for HIV there is no difference between samples analysed on different days which indicates that these vector samples are more robust than cell samples after spotting.

Overall there is a clear difference between EIAV and HIV vector spectra as evidenced by clustering in PC scores plot (Figure 73-Figure 74). Interestingly, suspension and adherent grown vectors show more differences in EIAV while they are mostly uniform in HIV (Figure 73-Figure 74). In all cases it is important to note a smaller sample size of vector data compared to cell data which was caused by high vector concentration used in the analysis (all samples in this analysis were concentrated by ultracentrifugation as described in section 3.2.5)

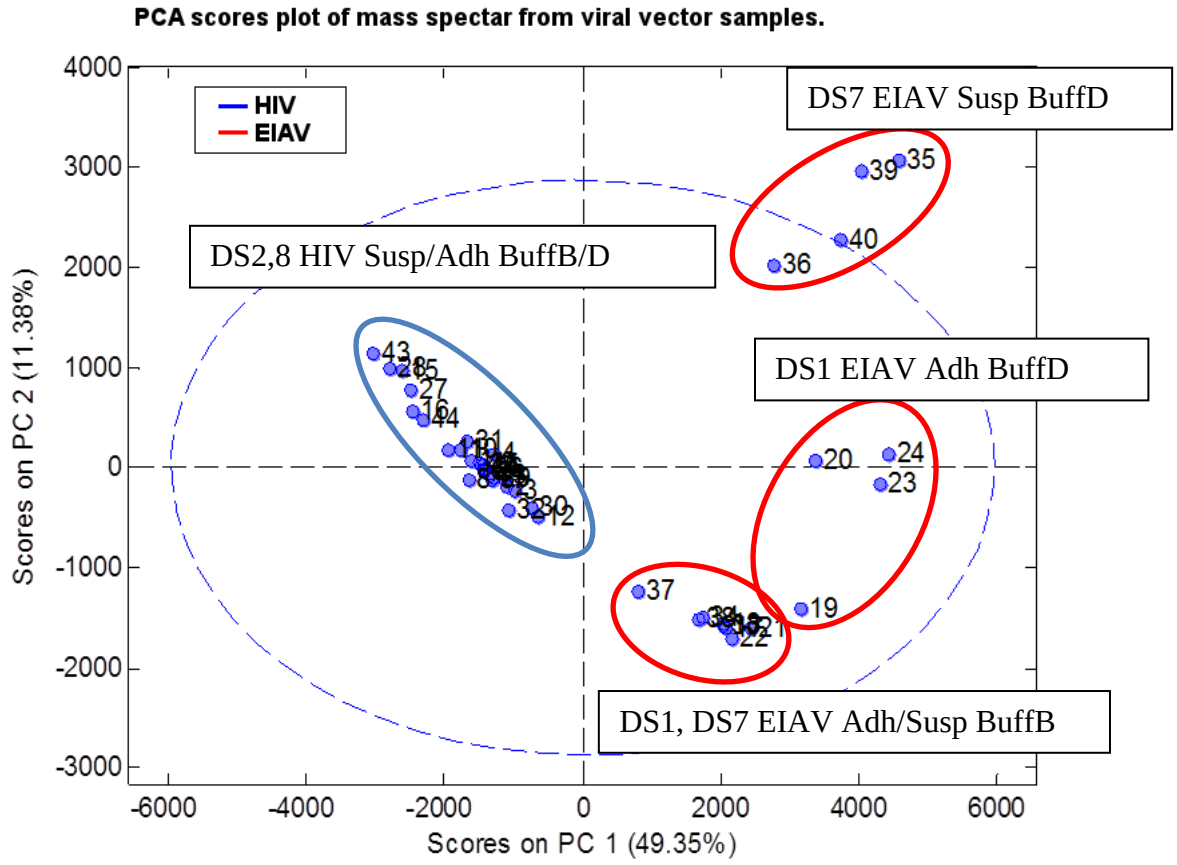


Figure 73: PCA scores plot (PC1 and PC2) for mass spectra of viral vector data.
Red circles indicate clusters of EIAV vector data points, blue circle indicates cluster of HIV vector data points. Viral vectors were produced in adherent (DS1-2) or suspension (DS7-8) systems with mass spectra generated in buffer B (α-cyano) or D (SA).

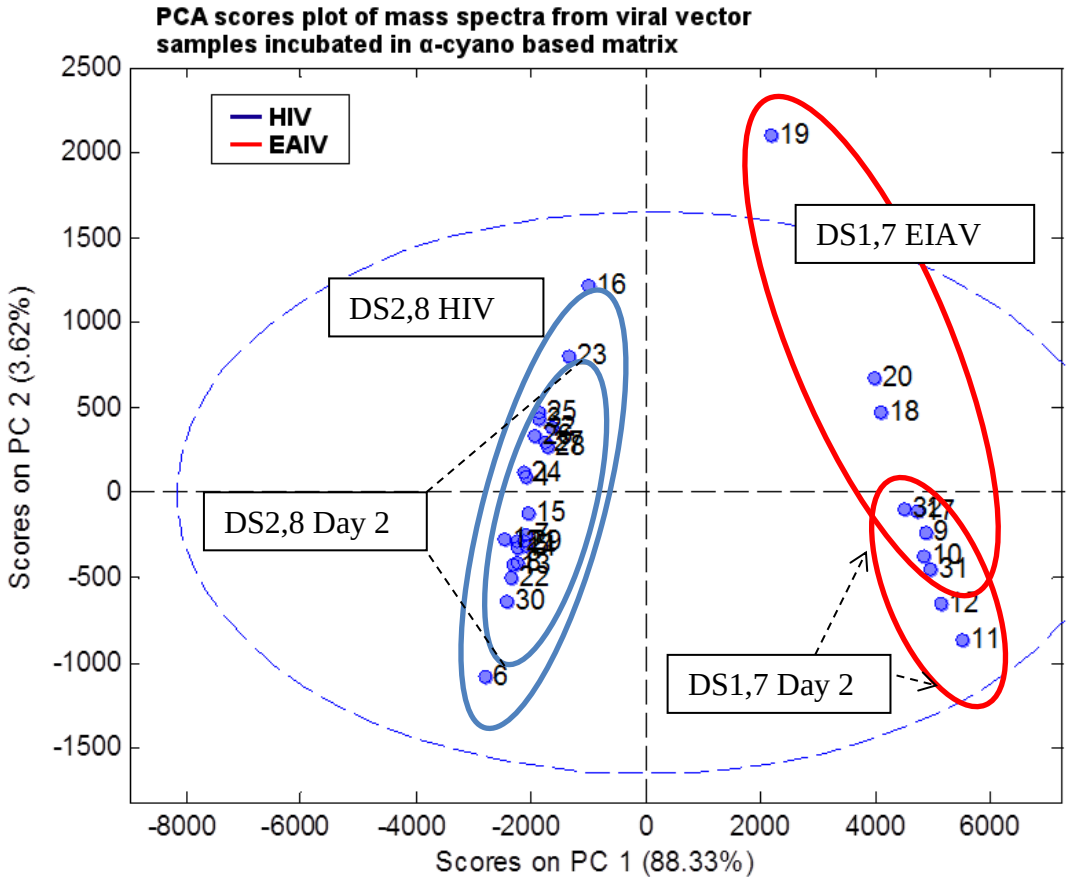


Figure 74: PCA scores plot (PC1 and PC2) for the vector samples incubated with buffer B (α-cyano). Viral vector samples were produced in adherent (DS1-2) and suspension (DS7-8) systems and mass spectra were obtained either 1 hour after spotting (default) or 24h after spotting (Day 2).

3.3.8. Effect of vector concentration

To assess the effect of vector concentration on MS results as well as the effect of downstream processing, a lower concentration viral vector samples were analysed along with the previous high concentration samples. These additional HIV vector samples were concentrated up to 60-fold volumetric concentration using a combination of filtration, ion exchange chromatography, diafiltration and spin filtration (DS10). Vector samples were analysed as a concentrated sample (60-fold volumetric concentration) or diluted with water or TSSM (final 30 or 15-fold volumetric concentration). The mass spectra of this data set are presented in Figure 75. Some of the spectra show distinct peaks of high intensity while other only results in irregular flat line of low or high intensity. There is a certain level of background baseline shift caused by the fact that the vector samples have been suspended in TSSM buffer. Combined with a low concentration of the vector in some of the samples, the poor results in some of the samples are to be expected.

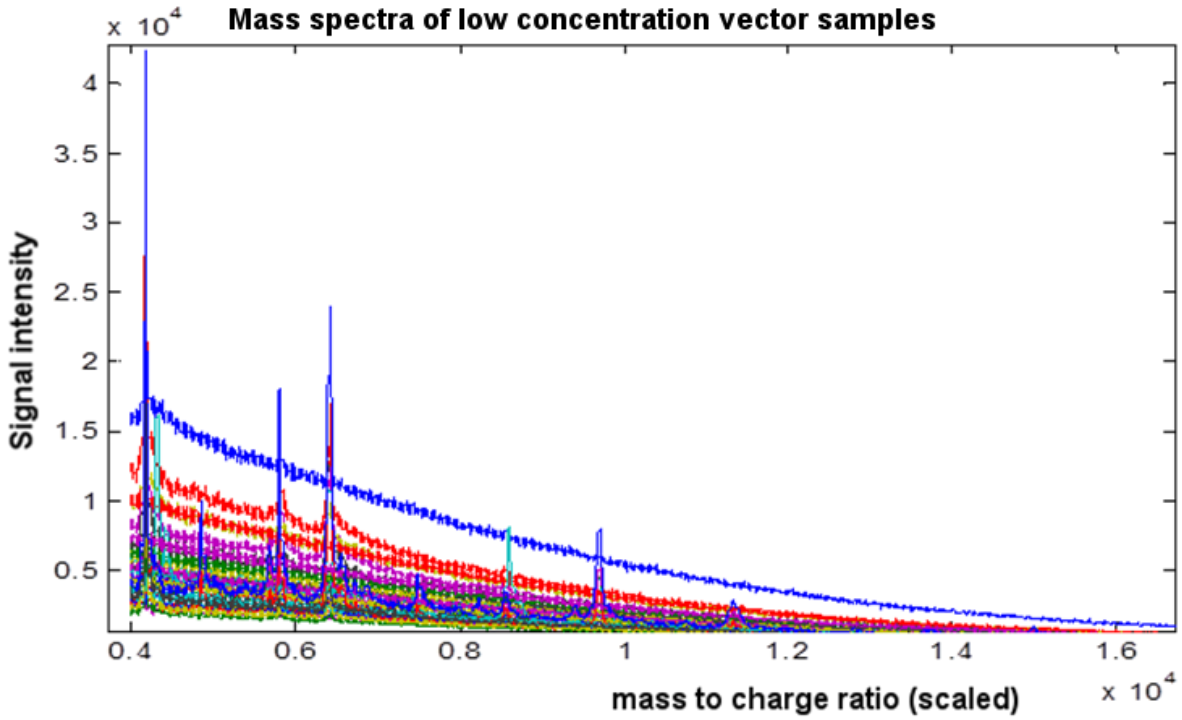


Figure 75: Mass spectra of low concentration LV samples.
Concentration ranging from 15 to 60x volumetric concentration compared to harvest samples. Notably, some samples have distinct peaks while other spectra have none. Dataset used was DS10

Overall, the only spectra which gave clear results are the highly concentrated sample (Figure 76). However within that sample there is a significant degree of variation where 4 of the repeats (Figure 76a) show a different structure than the remaining 6 repeats (Figure 76b). Some of the peaks present in the more detailed repeats are not visible in the other repeats. There are approximately 5 major peaks with several lower intensity peaks which corresponds with the number of viral proteins reported in other LVV MS analysis focusing on proteomics (Denard et al., 2009). Mass spectra and protein composition of samples is discussed further in section 3.4.3. Processing and filtering of the spectra followed by PCA (Figure 77) supports the initial assessment: There are two distinct subsets of mass spectra for the vector samples which can be clearly distinguished.

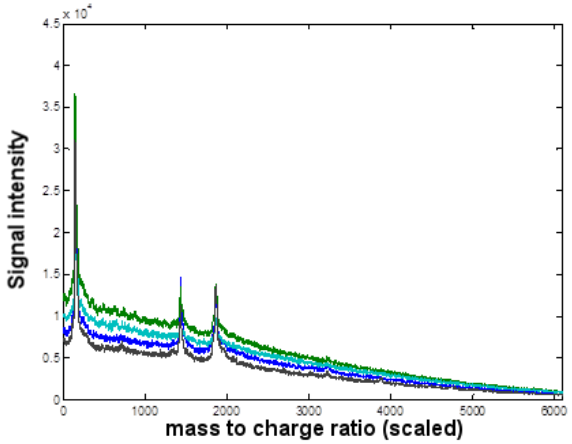


Figure 76a: Mass spectra of vector sample (60x concentration) repeats 1,2,4,7 (DS10)

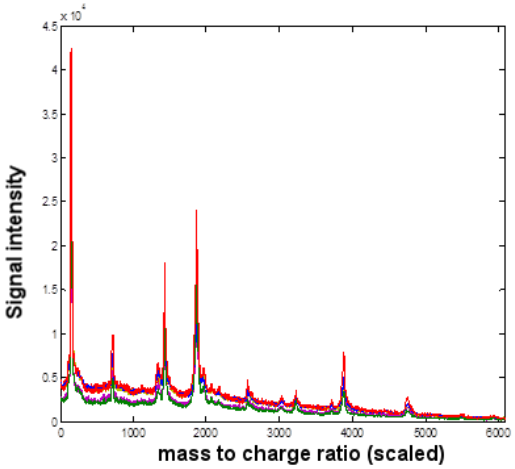


Figure 76b: Mass spectra of vector sample (60x concentration) repeats 3,5,6,8,9,10 (DS10)

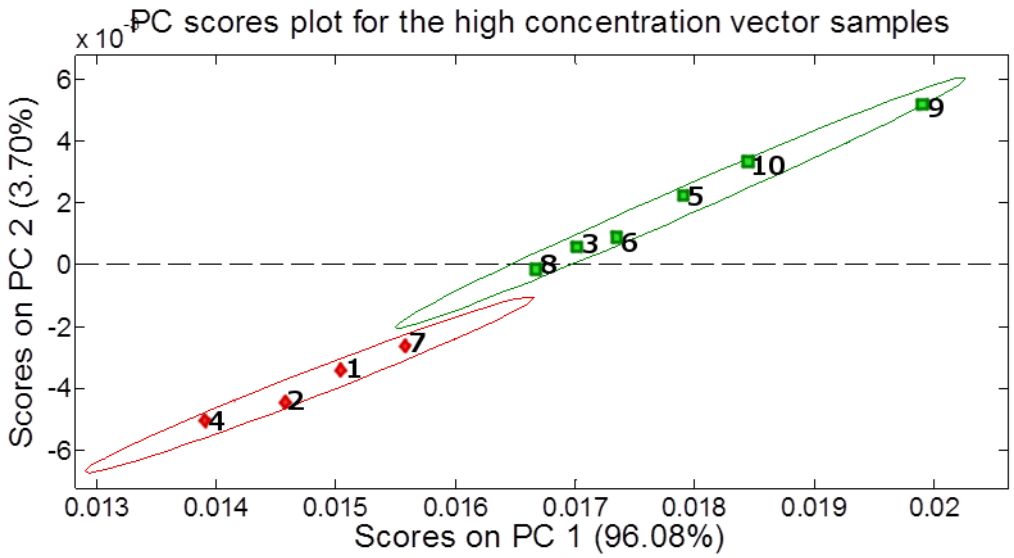


Figure 77: PC scores plot mass spectra of concentrated vector samples.
60x concentration, repeats 1-10 (as in Figure 76). Sample number is corresponding to the running order of MALDI-ToF MS. Dataset used was DS10

From the remaining samples, only 2-fold dilution with water resulted in several mass spectra with distinctive signals (Figure 78). For some of the repeats there are distinctive high intensity peaks present in the spectrum. However, for other repeats there are no peaks at all where the spectrum consists only of the background signal which can be attributed to the matrix and TSSM forming a baseline of the spectra. Overall, the results between separate repeats are highly inconsistent. The remaining samples of 4-fold dilution and 2-fold dilution with TSSM showed no vector associated signal and the only visible signal was associated with the matrix and TSSM background signal.

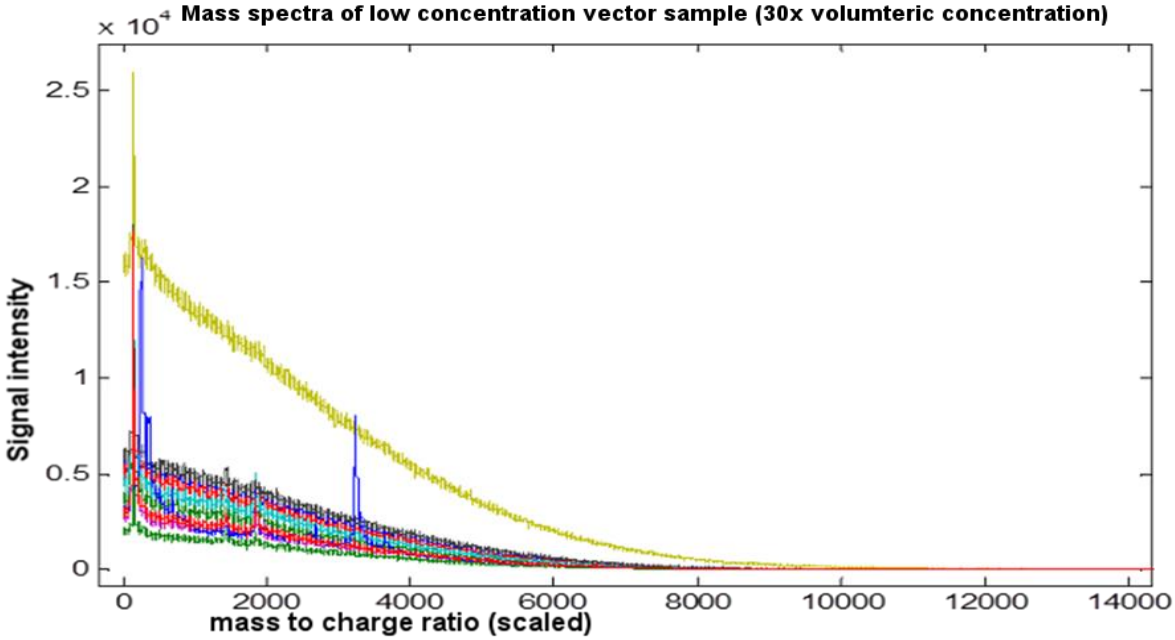


Figure 78: Mass spectra of medium concentration viral vector samples.
30x concentration, diluted with water from 60x concentrated vector. Dataset used was DS10

3.3.9. Effect of downstream processing

Additional vector samples were analysed to determine the effect of downstream processing on mass spectra. The samples were collected from a variety of sources to examine the variation caused by different modes and scales of vector production and the method of downstream processing. All samples were HIV based viral vectors encoding GFP to eliminate the impact of difference between EIAV and HIV vector described in the previous section.

The four analysed data sets included the original HIV-GFP data set used in analysis of EIAV/HIV difference (DS2,8). The samples were used as a single data set as there was no significant difference in HIV-GFP samples produced adherently or in suspension (Applikon® MiniBio reactors and CF2, Figure 73). The remaining data sets are samples from 50L pilot bioreactor run (DS11) and 7L Applikon® EZ bioreactor study (DS9-10) which used two different methods of vector concentration and purification (Two step centrifugation for DS9 and chromatography for DS10). Overall the mass spectra of the vector samples (Figure 79) are less complex than the mass spectra obtained from cell samples. Vector data is mostly comprised of several peaks which show a degree of variation between data sets. Compared to the cell line samples it is potentially easier to identify individual proteins, however this remains difficult without separating individual proteins from the whole vector sample. In the mass spectra, there are three major peaks which can be observed in all data sets but there is a degree of variation in their intensity. There are also several peaks which are present or absent

only in some data sets but not in the others. This indicates that while there are dominant traits in protein profiles of all vector samples, they are not identical. The most noticeable differences can be observed in the two data sets of vector purified using a combination of filtration and chromatography methods (blue and black on Figure 79). Both samples show peaks which are not present in any other data sets. The fourth data set (MiniBio/Cell factory, red in Figure 79) also shows peaks not present in other samples in the high m/z region of the spectra, however these peaks are overall low intensity compared to the other data.

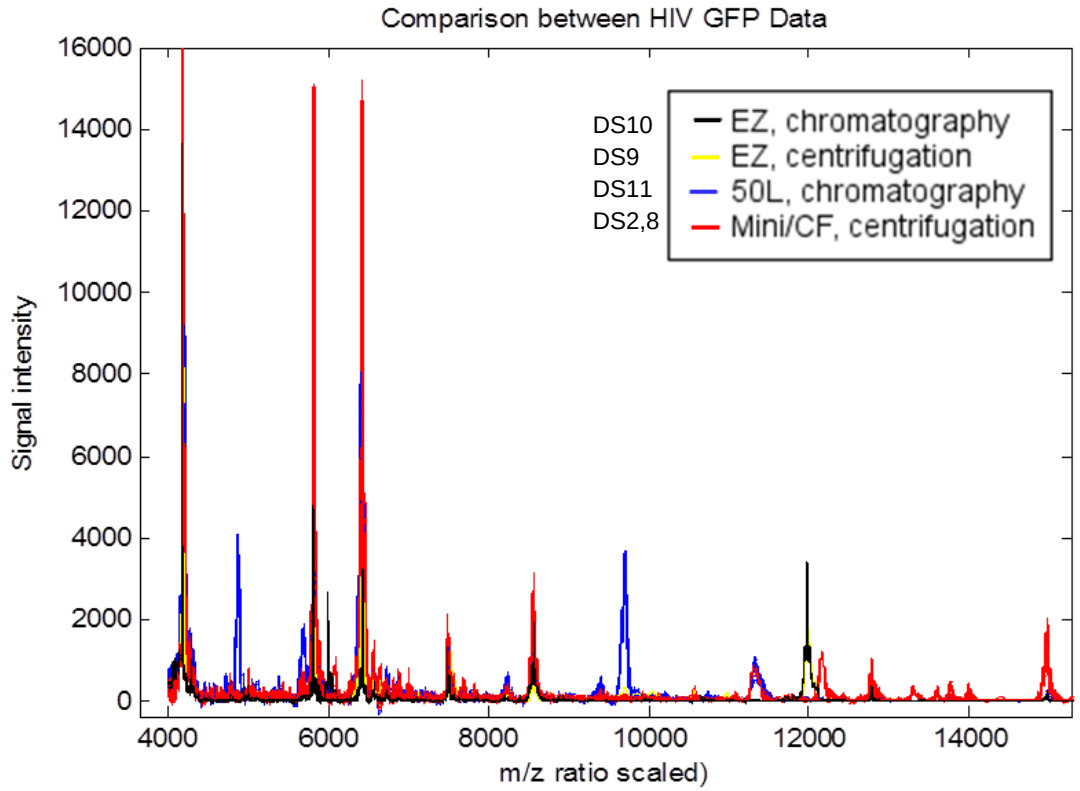


Figure 79: Raw mass spectra of 4 data sets of HIV-GFP viral vectors purified using different methods. The mass spectra were averaged for data sets consisting of viral vector samples produced and purified using different scale and downstream processing (DS2 and 8-11).

Based on the results discussed above, the effect of downstream processing was examined more closely. The vector samples from 7L Applikon® EZ bioreactors were collected together and clarified, followed by either concentration through centrifugation (DS9) or purification through a multi-step downstream processing involving filtration, Benzonase® treatment, AEC and ultra/diafiltration (DS10). This way the material was identical for all samples and the differences resulted from downstream processing which was the focus of the study. The resulting mass spectra (Figure 80) show a clear difference between samples where the vector material originated from the same source and the only variable was downstream processing post-clarification. As observed before, the purified

samples (DSP, red in Figure 80) show 2 peaks which are not present in the vector concentrated through centrifugation.

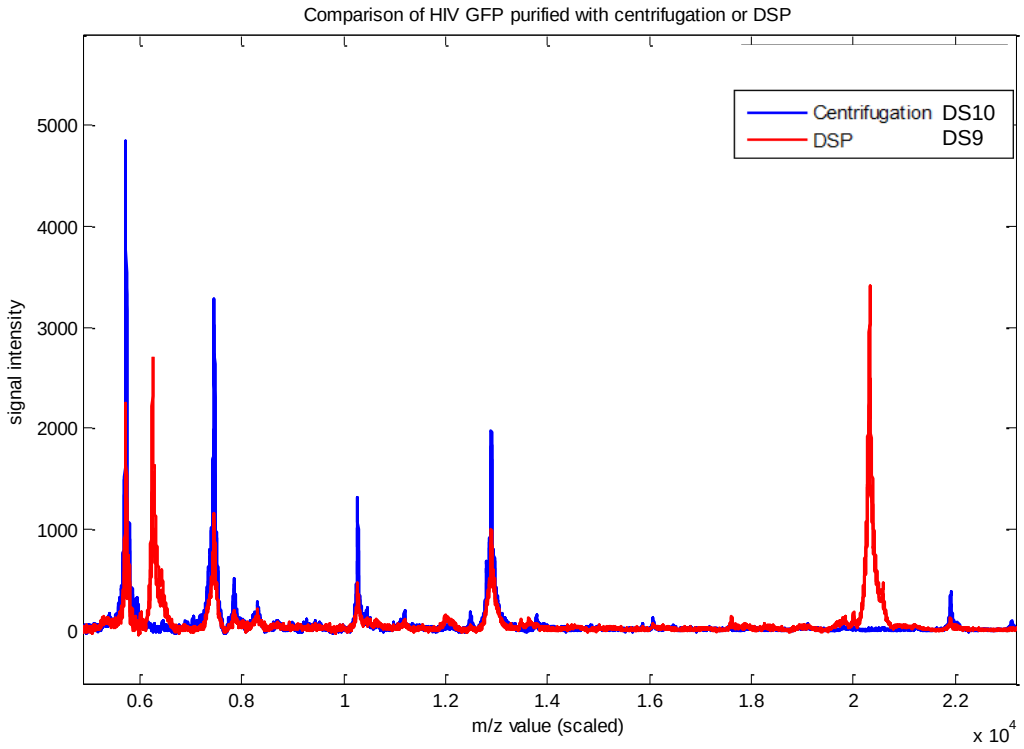


Figure 80: Mas spectra comparison between viral vectors purified with ultracentrifugation or chromatography. Raw mass spectra of HIV GFP vector samples were produced in Applikon® 7L EZ bioreactor and concentrated using centrifugation (DS9) or purified using filtration/chromatography (DSP, DS10).

The mass spectra discussed above were processed and subjected to PCA (Figure 81). Similarly to the cell samples, the majority of variation is captured in the first PC which is related to the average spectrum of all samples. The samples show only small differences in scores on this PC. For PC2 there is a higher degree of separation. Samples from the different data sets show a range of scores with negative and positive values. When both PCs are considered the samples from the same data sets form distinct clusters. However, the overall distance between clusters is small. The two data sets of vector produced in 7L Applikon® EZ bioreactor (DS9-10, red and purple in Figure 81) are clearly separated, indicating a difference between the two as expected given the two peaks observed in one of these data sets but not the other (Figure 80). The two data sets (7L Applikon® EZ reactor DS10 and 50L pilot-scale bioreactor DS11) from the vector purified through filtration/chromatography are placed close to each other on the PCA scores plot while the last cluster of centrifuged vector samples (DS2 and 8, pink in Figure 81) is more spread than the rest (which may be associated with a bigger sample number from several batches). However, the distribution of samples within the cluster was mostly random i.e. not associated with the

vessel used for their production (Applikon® MiniBio reactor or CF2). Overall, the PC2 scores are aligned with the scale of production: larger batches produced at 7L and 50L scale have positive PC2 scores values while small scale (MiniBio and CF2) batches are characterised by low or negative PC2 scores values.

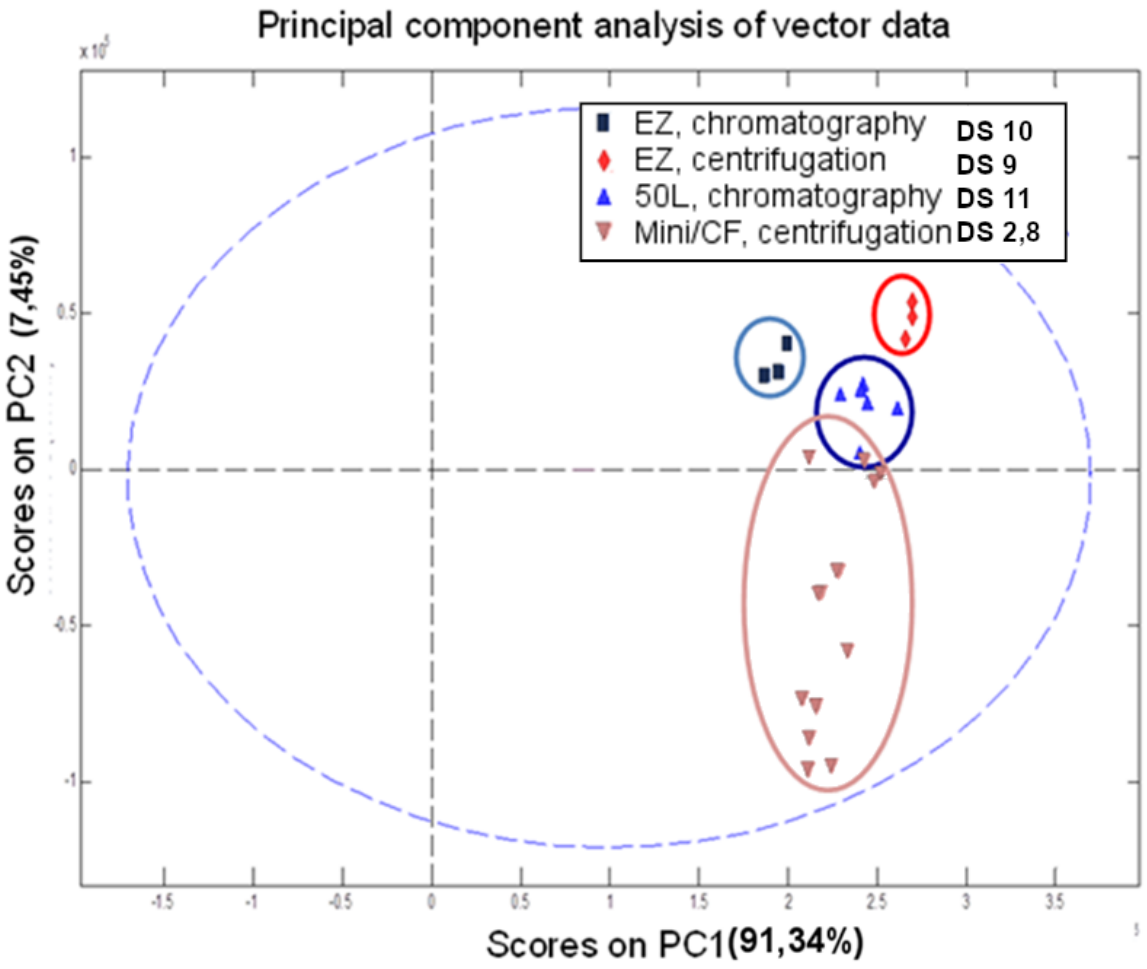


Figure 81: PCA scores plot (PC1 and 2) for the mass spectra of 4 data sets of HIV-GFP viral vectors purified using different methods.
Datasets used are DS2 and 8-11.

3.4. Discussion

Development of a functional MVDA model based on mass spectrometry needs to address the optimisation of signal to noise ratio. Obtaining spectral data is a complex process that can be affected by the sample preparation protocol, matrix-sample interaction, equipment settings and operator influence. All these factors can introduce noise into the data, reducing its quality. To improve signal to noise ratio sample preparation and the effect of different matrices was examined in section 3.3.1. To reduce the variability of samples preparation and operator variability each sample was prepared in multiple replicates. These steps were taken to develop a robust mass spectrometry method and improve signal-to-noise ratio. Additionally, the mass spectrometry signal was pre-processed to further reduce noise and emphasise the variation in data caused by variability between different cell lines, vector types and processes used in production.

MALDI-ToF MS robustness study was performed to validate the assay as a reliable method of cell and vector assessment and to form a baseline for future experiments. The study assessed different types of assay variability, addressing the impact of matrix composition, required number of repeats for each sample (variability within sample), differences between measurements for individual samples and minimal number of samples required to achieve consistent results (between sample/within batch variability), differences between separate batches of the same process and product (between batches variability). Finally, variability between different products, growth modes and cell lines was assessed in more detail as well.

With enough information about assay variability and the impact of different process factors on final results it would be possible to study more complex interactions such as the effect of process scale up on cell and vector qualities. There is a significant difference in the way that e.g. a shake flask and a bioreactor operate which is likely to impact the cell behaviour and vector structure. Having the ability to analyse them using MS could yield useful data and improved process understanding.

3.4.1. MALDI-ToF MS Robustness study

The effect of matrix composition was clear throughout all samples, with α-cyano offering superior signal intensity and peak identity for both cell and vector samples. Use of SA based matrix resulted in a lower intensity spectrum, suggesting SA could be used as an alternative to α-cyano. While it results in significantly lower signal intensity, it still retains

distinctive peaks which allow identifying key cell and vector characteristics. However, lower signal intensity means that to acquire meaningful data, higher concentration of material or more assay repeats may be required to obtain consistent data. This is especially important in case of the vector samples, where amount of the vector and its concentration is a major limiting factor. This makes SA much more problematic to use with the viral vector samples. Finally, DHB showed no results for either cell or vector samples and is therefore unsuitable for this application. Overall α-cyano based matrix provided superior results. It was used for all subsequent experiments (except for several initial studies performed to confirm the results) and should be used in preference to other possible matrix components when working with HEK293T cells as well as HIV and EIAV based LVVs. In the initial comparison of matrices multiple regions of the spectra showed significant variation depending on the matrix used and their concentration (Figure 23, Figure 24), highlighting signal variation introduced by use of different matrices. The early-stage development aimed to minimise noise in the data and determine the optimal sample preparation protocol.

Another important issue is the time spent between sample spotting and MALDI-ToF analysis. It would be beneficial to have flexibility in sample preparation time because of the logistics of the project and potential future industrial application. A study in bacteria demonstrated that an extended incubation time and exposure to oxygen had no significant effect on the mass spectrum quality (Veloo et al., 2014). However, the effect of incubation time and the period between sample spotting and MALDI-ToF analysis is uncharacterised for the mammalian cells and viral vector samples. It is therefore crucial to assess the impact of time on the sample characteristics. In the case of viral vectors samples, time between spotting and analysis had no impact on the final measurement (Figure 74). Similarly, several rounds of freezing and thawing of the sample did not change the MS results. The conclusion is that vector samples are robust and allow for flexible handling and scheduling of MS analysis. However, for cell samples 24 hour waiting period led to small changes in the data structure when performing PCA. The shift in positioning on the PC scores plot is consistent for most samples, suggesting that while cell samples are affected by variable time between spotting and analysis, the relative data structure is maintained throughout time. Therefore, the samples are comparable as long as the incubation time on the plate is consistent between experiments. It is therefore critical that the sample plate is prepared the same way every time to ensure consistent results. Small deviations are unlikely to affect the overall analysis as the change in data is minimal but extended incubation or delay before analysis may affect assay

consistency. Consequently, a standard protocol used for most experiments was established, with 1 hour 30 min sample incubation in the matrix at 4°C and MALDI-ToF analysis 24 hours after spotting.

The results show that there is a degree of variation between mass spectra generated by MALDI-ToF from a single sample, as evidenced by occasional outlier results observed in PCA scores plot as well as in the raw mass spectra. The sampling and spotting process is important for maintaining consistency of the mass spectrum, but the matrix ionisation process and detection of ionised material plays a role in signal variation as well. In the experiment comparing 4 adherent cell lines (Figure 57-Figure 62), for half of the cell lines (cell lines 1 and 2) the spread of individual repeats was very uniform while for the other half (cell lines 3 and 4) repeats for individual samples tended to group together. Analysis of the whole data set showed that averaging the repeats to form a single spectrum allows distinguishing between individual cell lines which indicates that 10 repeats per sample is a sufficient number of plate spots to limit the effect of assay variability. The difference between individual samples (especially apparent for cell line 4) indicates that a higher number of samples could be beneficial, however both the amount of material and space on the MALDI plate may become problematic if more samples are to be analysed. Overall, the results indicate that with a high enough number of repeats (10 per sample) MALDI-ToF MS can become a precise assay capable of detecting subtle differences in cell protein composition. Further studies have demonstrated spectrum consistency with 6 plate spots per sample, indicating it as a viable minimal number of repeats while a higher number can be used when enough material and plate space is available. This allows a single cell sample to be analysed in a few minutes depending on the exact amount of repeats; when accounting for sample preparation the entire analysis can take several hours depending on the number of samples.

As discussed in the previous sections, the mass spectra are complex and difficult to interpret in their raw form. Furthermore, there is a significant amount of noise in the data which further complicates the analysis. Signal pre-processing addresses these issues by reducing data noise and enabling the use of MVDA. One of the main sources of variation between samples is the different level of base intensity of the signal (e.g. Figure 63). The variation in matrix chemical composition, sample incubation and spotting as well as the laboratory conditions and ionisation events result in higher or lower level of background reading of the signal (Liu & Schey, 2005b). This is adjusted through the baseline correction method which calculates the baseline of the spectra and adjusts peak intensity by subtracting

the baseline values from the overall spectrum. Subsequently, the spectra intensities are brought to a common level which makes it more feasible to compare them. The normalisation step is used to adjust the peak heights relative to the other spectra. Finally, polynomial smoothening of the signal using the Savitzky-Golay algorithm is applied to further reduce the noise in the data while maintaining the overall peak shapes by adjusting individual wavelength intensities based on neighbouring values. This helps in the signal analysis while ensuring that no important information is overlooked. Finally, the peak alignment is performed to ensure that there is no shift in the spectra due to variations in equipment settings or any of the data processing steps. This is especially important if the samples are taken using different ranges of m/z measured by the mass spectrometer.

3.4.2. Cell samples analysis

The MS PCA results show clear differences between some of the cell and vector types. This translates to differences in proteome composition which can be used for further characterisation. For the cell samples the main factor affecting data structure was the adaptation of the cells to grow in suspension. The cells used in the initial experiment came from the same HEK293T cell line (see sections 3.3.1 and 3.3.4); however some of the cells were developed into a new line by adapting them to growth in serum-free suspension media (see section 3.3.3). For mammalian cells which usually rely on adherence to a solid surface it is a drastic change in growth mode which affects their gene expression and therefore the overall protein profile. This change is reflected in the values of PC scores of the different cell lines, which in turn demonstrates that MALDI-ToF MS is a technique sensitive enough to pick up the differences between cells adapted to different growth conditions (see section 3.3.5). These results are further supported by the experiment where 4 adherent cell lines were inspected to assess the variability between samples and repeats. The primary goal of the study was to look at the data consistency and distribution of individual repeats and samples. However, the pattern emerging from the PCA analysis clearly shows that with enough samples and repeats it is possible to identify subtle differences in protein levels between similar but not identical adherent cell line samples (see section 3.3.6). Moreover, except for a single outlier (which could be attributed to sampling variation), the results are consistent and distribution of samples from the same cell line is very close. This further supports the claim that MALDI-ToF MS is a technique well suited for cell line analysis and comparison as it is capable of differentiation between individual cell lines.

In the cell analysis study, the main finding was that the difference in peak shapes at certain m/z values also had a major effect on PCA through the values of PC2 scores. The flattened shape of the peaks for suspension cells when compared to adherent cell samples suggests a substantial difference in protein composition at this position which is significant enough to distinguish between the two cell types (adherent and suspension). Whole cell MALDI-ToF MS measures the spectra based on the protein profile of the inspected cells, and therefore the difference in peak shape would suggest a potential difference in protein structure, composition, modification or interaction which is shown to be present in all suspension-adapted HEK293T cells (see Figure 65). However, it is difficult to pinpoint the exact effect as the difference in spectra can only be observed for several peaks and some of them have complex shapes. This suggests that the effect is caused by multiple proteins, protein-protein interaction, other cellular elements (e.g. lipids) or a combination of these factors. It is difficult to say whether the root cause of the difference is genetic i.e. the selective pressure of adaptation to suspension culture is leading to changes in protein structure or it is environmental i.e. cells cultured in suspension have a different protein profile due to the growth conditions affecting protein expression. The adherent and suspension culture conditions have significant differences, including the shear stress applied to cells, media composition, presence or lack of FBS and handling of the cell culture itself during the process. Adaptation of the cells from adherent growth to suspension is a lengthy process including gradual changes in media composition and selection of cells which are responsive to the changes. While the current data is not sufficient to make a conclusion about the cause of the difference it is an interesting area to investigate and it highlights the difference between adherent and suspension cell culture.

The difference in the MALDI-ToF mass spectra of adherent and suspension adapted cells was large enough to completely separate the samples from these two data sets on the PCA scores plot (Figure 66). However, the mass spectra variation between individual cell samples cultured under the same culture conditions (adherent or suspension) was much smaller. The samples clustered together quite closely, suggesting a similarity between the cells. For the suspension cells, the two distinct clusters were formed by cell lines 6 and 7 samples, both cultured in ambr®15 micro bioreactor. The remaining clusters which are overlapping with each other while also being located close to ambr® samples are cell line 5 samples cultured in four 0.5L Applikon® MiniBio reactors. They are biological repeats using the same cell lines and culture conditions, which explains the close positioning of the samples

on the PCA scores plot. These results suggest that while suspension cells give quite similar spectra, there are enough differences to distinguish between samples taken from different cell lines. For the adherent cell samples the situation is similar, however the clusters are closer to each other than in the case of suspension cell samples and there is more overlap. This can be explained by the fact that all adherent samples have been cultured in similar conditions and at similar scale while the difference between 0.5L Applikon® MiniBio reactor and 15 ml ambr®15 system is more significant. Adherent cells require a surface for attachment and scaling up the culture involves increasing the surface area available for the cells. In the case of suspension culture, the scale up of the process involves changes in vessel geometry (rectangular for ambr®15 and cylindrical for Applikon® MiniBio reactor and larger vessels), aeration and agitation rates (and therefore mass and energy transfer). The only different data sets are the 4 cell lines cultured in 10 cm^2 dishes which are characterised by significantly higher score on PC3. While PC3 only explains 2.41% variability in the data, it is an interesting indicator which may be associated with the scale of the culture. Overall, the adherent samples are all clustered close to each other on the main PCA plot and only 4 of the subsets show a slightly different profile when looking at PC3 scores.

It is worth noting that mass spectrum region between 1300-1400 m/z and 4700-4800 m/z was highlighted as hypervariable when comparing adherent and suspension cells (Figure 65). This high degree of variation was also present in the loadings plot (Figure 68). These regions of high variability with high contribution to PCA scores can be key in differentiating between cell lines of different properties. While the difference between adherent and suspension adapted cells is clear, the differences between low and high producer cell lines are more subtle and more difficult to spot without in-depth MVDA as discussed later in Chapter 4.

As described in the methods section, most of the cell samples were transfected with HIV-based vector encoding GFP. However, there is a subset of samples transfected with either EIAV-based vector or left untransfected. Cells from two of the Applikon® MiniBio reactors were transfected with EIAV-based vector (Suspension 1 and suspension 2 on Figure 66) which did not cause major differences when compared to the remaining 2 MiniBio samples transfected with HIV (Suspension 3 and suspension 4 on Figure 66). For adherent cells analysis, AdhCF1 was transfected with EIAV while AdhCF2 with HIV and there were no major differences between the clusters. This suggests that the vector type used for

transfection has no impact on the PCA scores of the samples when compared to the baseline of cells transfected with HIV-GFP based on the cells' mass spectra.

Based on the results discussed above, MALDI-ToF MS has been assessed as a robust and reproducible method of generating mass spectra of different types of cells. The protocol has been optimised and standardised to ensure continuity and comparability between experiments. The spectra of suspension and adherent cells are clearly separated when using PCA while individual samples cluster with samples from the same data set or separate data sets of the same origin (e.g. samples from 4 separate Applikon® MiniBio reactors which form overlapping clusters). Differences in the cell lines used for the culture also impact the positioning where the two data sets from ambr®15 micro bioreactors are different enough to be identified as separate clusters, separate from the other suspension cells.

3.4.3. Viral vector analysis

In the viral vector analysis, there are several peaks present in only some of the samples. Viral vectors are complex entities that consist of multiple proteins and lipids which interact with host cells. This means that the production process has a significant impact on the final vector composition. All analysed samples were characterised by three major, high-intensity peaks which may correspond to some of the viral structural proteins encoded by *gag*, *pol* and *env* polycistronic genes. Additional two to five peaks were presented in several samples, with the exact number varied between vectors of different concentration and processing. The number of MS peaks corresponds to number of proteins previously identified in LVV samples (Denard et al., 2009) including VSV-G, p66 reverse transcriptase, p31 integrase, Pr55Gag polyprotein and its subcomponents p17 matrix protein and p24 capsid protein. Both viral envelope and capsid are important for viral vector functionality and the structural proteins which form these two elements are expected to be highly abundant in the whole vector sample. However, it is difficult to determine what proteins are contributing to the mass spectra peaks without a high degree of protein separation and purification prior to MS analysis.

Trends observed for the viral vector samples are similar to those from cell analysis, but they are based on viral vector type rather than cell growth mode. EIAV and HIV based vectors are clearly separated on PCA scores plot, the difference is significant and consistent. While the two types of vectors are different in terms of protein composition, they also share some of the features which make the results particularly interesting. The overall function of

HIV and EIAV *gag* and *pol* sequences is the same; however, the MS suggests that the proteins are different for each of the vectors, showing different peak positions and intensities. Moreover, there are proteins such as Rev which are present in one of the vectors but not in the other. Interestingly HIV based vector shows no difference based on the type of cells used for production (suspension or adherent) while EIAV based vector is less consistent. This could suggest that EIAV vector is more closely associated with the cells used for production and the difference could be associated with residual cellular proteins which are present on the vector's surface or are retained within the vector particle. However, the sample size for the vector sample was relatively low due to limited material availability and therefore this experiment should not be used as a sole indicator of a difference between EIAV and HIV production mechanism.

The analysis of viral vector results shows that the mass spectra of viral vectors can be highly variable depending on the sample concentration and buffer composition. As determined in a preliminary study, TSSM leaves a distinctive background signal when examined with MALDI-ToF. Therefore, low concentration samples face a major problem: the protein signal is not always strong enough to overcome the background noise and therefore some or all of the repeats for the low concentration vector sample result in poor quality mass spectrum. While downstream processing does not seem to affect the vector detectability, the low final concentration is often insufficient for reliable detection. As demonstrated in Figure 75 and Figure 76 even 60-fold concentrated vector forms 2 different patterns for the same sample, making these results unreliable. It is therefore likely that a higher vector concentration is necessary for a reliable and repeatable viral vector analysis.

When looking at viral vectors produced and purified using different scale and downstream processing, the results from 7L Applikon® EZ bioreactor study show that there is a difference between samples concentrated using centrifugation or purified using multi-step downstream processing including filtration and chromatography. Both samples were produced in the same conditions and clarified before processing. The two peaks present in the mass spectra of samples purified with chromatography suggest that there are either proteins which are lost during the long duration centrifugation process due to stability issues or proteins which are acquired during the complex downstream processing (filtration/chromatography). One of the downstream processing steps is Benzonase® treatment which is used for degrading host cell DNA (Sastry et al., 2004). The vector is incubated with the enzyme at 37°C early during the process. The samples are then purified using AEC and

ultra/diafiltration which are designed to eliminate any impurities present in the sample and to concentrate the vector. However, it is possible that either Benzonase® treatment or one of the other steps introduces a molecule detectable through MS, resulting in additional peaks present in the spectra. The vector samples from 50L pilot were also purified using a combination of filtration, AEC and ultra/diafiltration with the adjustments made for larger scale of the vessel. Again, mass spectra show distinct peaks which are not present in other data sets, suggesting a presence of a unique protein or protein complex. However, in this case the amount of samples was limited and only one type of downstream processing method was investigated. It is therefore possible that the additional peaks are associated with upstream processing in the large scale vessel and it is not possible to confirm that the extra peaks are caused by downstream processing without looking at more samples from 50L reactor which is not possible at this time.

The PCA analysis of HIV-GFP vector samples showed that while there are several peaks in the spectra which are present only in a subset of the samples, overall the samples are positioned close to each other on the PCA plot. This indicates that despite the difference in protein profile between data sets the individual clusters are not separated as clearly as e.g. suspension and adherent data sets discussed before. There are still trends which can be observed i.e. a separation of clusters of vector samples from 7L Applikon® EZ bioreactor concentrated through centrifugation or purified through filtration/chromatography. This indicates that the method of downstream processing has a significant impact on the viral vector mass spectrum. The distribution of samples on PCA scores plot is difficult to attribute to a specific trait of the vector as there are multiple peaks which cause the differences between samples and there is no single trend of either vessel scale or downstream processing method clearly affecting the PC scores. The samples concentrated using centrifugation occupy opposite ends of the PC2 scores range, with vector samples produced in 7L Applikon® EZ bioreactor having positive values while vector produced at smaller scale 0.5L Applikon® MiniBios and cell factories have negative scores. At the same time first two PCs capture most of the data variation, leaving other PCs with less significance.

The high amount of variation between the data sets suggests that vector production and purification introduces a degree of variation in vector composition. It is possible to separate samples processed in a different way but it is difficult to reliably identify the cause of the difference as neither production scale or downstream processing has a consistent

impact on the PC scores that could be recognised as a trend in the data. This makes interpreting the vector analysis results difficult.

There is also an issue of vector concentration which was demonstrated to affect the mass spectrum quality. The vector samples must be concentrated before analysis which is done based on volumetric concentration with optimal results obtained at about 1000-fold concentration compared to crude vector. This limits the number of samples that can be collected and limits the analysis of vector produced at small scale. These results put together indicate that vector samples are more problematic to analyse than cell samples. The variation within HIV-GFP vector samples suggests that MALDI-ToF MS may not be a sufficient method for modelling of viral vector properties. However, it is still able to identify small differences in the spectra and when combined with PCA it may be a useful tool for monitoring of viral vector purity profile if a standardised reference spectrum can be determined. However, the high concentration of vector required per sample may limit the use of this method to large scale application.

MALDI-ToF MS is a powerful tool for whole cell and vector analysis, capable of measuring the mass spectra of complex samples and generating results which can be used in fingerprinting individual samples based on their protein composition. However its ability to identify individual proteins is limited as the mass spectrum is affected by the ionisation process and protein behaviour in the electromagnetic field of the mass spectrometer (Albrethsen, 2007). As such whole cell and vector MALDI-ToF can be used to assess the differences between samples, to identify samples based on their mass spectrum fingerprint and to correlate sample properties with their mass spectra which can be used for modelling. However, it is not a suitable method for identification of individual proteins. The first step in protein characterisation would require a high degree of protein separation using methods such as simple or two-dimensional polyacrylamide gel electrophoresis. This would allow obtaining the spectrum of individual protein and with enough reference data it would be possible to measure the mass of individual proteins. However, other MS methods may be more appropriate for this kind of analysis. The main recommendation would be to use electrospray ionisation which can also be coupled with liquid chromatography to separate and analyse individual proteins. The proteins can be further characterised by fragmentation into simple peptides, which can be used for peptide sequencing to identify a protein based on its sequence (Stone et al., 1998).

3.5. Conclusions

Recent developments in the gene therapy field instil increasing amount of confidence in gene and cell therapy-based therapeutics. This rapidly developing field of medicine offers great advancements in some of the areas traditionally difficult to innovate. The novelty of this field means that there is room for use of equally novel approaches to process development and characterisation. As part of this EngD project LVV production was assessed using MALDI-ToF MS and MVDA. The results of the project show that a data-oriented approach to process analysis has a great potential in improving the understanding of the underlying reactions and maintaining continuous improvement in line with the QbD guidelines. As OXB is involved in good manufacturing practice (GMP) production of LVVs as well as development of new products and implementation of new processes it is critical that the process and product understanding are continuously improved to ensure high quality of the product delivered to the patients.

MALDI-ToF MS has been shown to be a potentially robust and reproducible method of obtaining mass spectra from cell samples. A standard protocol for MALDI-ToF MS was established, using α-cyano based matrix incubated with cell samples for 1,5h or spotted with the viral vector directly on a MALDI plate. The MS signal data was pre-processed and subjected to PCA. This methodology was used to characterise samples from seven different cell lines, cultured adherently or in suspension, in vessels ranging from 15 ml (ambr®15 micro bioreactor) to 7L (Applikon® EZ bioreactor). The effects of scale, method of cell culture and cell lines were assessed to examine a diverse set of samples. A significant difference in several mass spectra peaks was observed between adherent and suspension cell samples which was also reflected by cluster separation on the PCA scores plot. Analysis of several cell lines showed a small but significant and consistent difference in the spectra between the cell lines, providing a method for identifying different cell lines based on MS analysis. Viral vector was characterised using the same method and samples sourced from vessels ranging from 0.5L to 50L and processed through different purification methods to achieve concentration required for MALDI-ToF MS analysis. Vector concentration and formulation buffer composition plays a major role in obtaining a high-quality spectrum, where extensive concentration is required, with MS signal obtainable with 60-fold concentration (measured by volumetric concentration from raw material) but optimal conditions require 1000-fold concentrated sample. There was a significant difference between EIAV and HIV based vectors but the type of the cells used for production had little to no

impact on the mass spectrum. The results of HIV-GFP analysis were less clear due to variation in the vector data sets and the impact of vector production and processing methods on the final sample composition and mass spectra. This led to several peaks in the spectra which were present only in single data sets.

The methods described above form an extensive toolset which can be applied in the industrial setting. The main obstacle is the cost of purchasing or outsourcing a mass spectrometer which would require either a significant up-front investment or persistent cost and establishment of procedures for generating mass spectra off-site. MALDI-ToF MS is a time sensitive method where sample incubation time, spotting and the time between spotting and analysis can all have an impact on the mass spectra. Providing enough material for viral vector analysis would also be problematic as it requires a highly concentrated sample and multiple repeats. Therefore, a reliable arrangement would have to be established, allowing consistent collection of mass spectra with minimal variation in sample processing and transport. MALDI-ToF MS is a useful technique but requires a significant time and resource investment for consistent implementation.

Visual analysis of the mass spectra as well as PCA loadings and contribution plots highlights regions of the spectrum which can be identified as either highly variable between different types of samples or with particularly high effect on the PCA scores. These regions of the spectra will become the most prominent element of the data structure. However, both cells and viral particles are complex structure comprised of multiple proteins. While visible contribution of particular spectrum m/z values will have a high impact on overall interpretation of the spectrum and the PCA, the combined effect of multiple lower intensity peaks could be significant for capturing the variability between cell lines or different viral particle types. As such the analysis focused on capturing and processing whole spectra for these complex entities in order to maximise the information captured in the data and to limit introducing bias

Use of PCA for characterisation of LVVs and HEK293T cells was demonstrated to be feasible; however, the identification of individual proteins which contribute to the mass spectrum is a difficult task which would require a significant amount of sample processing protein separation, digestion and bioinformatic processing. PCA is suitable for analysis of multivariate problems such as evaluation of process parameters, monitoring of cell productivity and analysing vector quality. Characterisation of cell samples was successfully

demonstrated in this report while vector samples proved to be more variable and problematic to assess. However, the combination of MALDI-ToF MS and PCA is still viable as a method to represent the variation observed in inspected samples and could be used for analysis of vector purity and quality profile across multiple production cycles.

Overall, this chapter outlines how the EngD project managed to improve process understanding of LVV production through extensive analysis of cell and vector samples using a combination of MALDI-ToF MS and PCA. This approach proved to be an effective and robust method for assessing a large number of mass spectra from a diverse set of samples and has a potential to be applied for process analysis and quality monitoring in an industrial setting (after further refinement). A robust methodology was established for cell and vector analysis with optimised matrix composition, sample preparation, signal acquisition, processing and analysis. This lays down groundwork for a standardised approach to cell and viral vector analysis usable in process characterisation and monitoring. MALDI-ToF methods and some of the MVDA principles described in Chapter 3 are further investigated in Chapter 4. Mass spectra and LV titre data from a selection of packaging cell lines along with MVDA-based classification algorithm are used to develop a predictive model for accelerated cell line development system.

Chapter 4 Partial least squares discriminant analysis model of cell line development for lentiviral vector production

4.1. Introduction

The application of MALDI-ToF MS described in Chapter 3 in the context of process characterisation can be extended further to improve other aspects of viral vector production. The data used in cell and vector characterisation can be used to develop a statistical model capable of distinguishing between significantly different populations of cells and vectors. This concept was applied to design a predictive model of cell lines' lentiviral vector production performance based on partial least squares discriminant analysis (PLS-DA) algorithm and used to accelerate cell line development process.

4.1.1. Background

Mass spectrum obtained from cells forms a complex pattern which varies between cells cultured in different conditions as well as between individual cell lines. While the expressed proteins are similar, their levels as detected by MS can vary from cell line to cell line (Geiger et al., 2012). MS fingerprint is based on detection of ionised particles of varying mass to charge ratios (m/z) which for whole-cell analysis are mostly cell proteins (Zhang et al., 2006). Through this approach it is possible to differentiate between cells of different protein compositions and therefore different properties including adaptation to growth in specific conditions and productivity. With enough variation between cells, it is possible to identify elements of the spectrum which affect their fitness and production performance (Feng et al., 2011; Povey et al., 2014). Multivariate data analysis (MVDA) methods can be applied to design a model capable of exploring the data structure (principle component analysis as in Chapter 3) or identifying underlying data trends which can be used for classification of samples and prediction of their properties (through PLS). This chapter discusses the application of MS to characterise HEK293T packaging cell lines cultured in suspension culture through predictive modelling of cell line productivity at bioreactor scale.

4.1.2. Cell line development

Cell line development for viral vector production (Figure 82) is a lengthy process starting with selection of a parental cell line which is then stably transfected to express some (packaging cell lines, PaCL) or all (producer cell lines, PrCL) of the lentiviral vector (LVV) components in a process controlled by an inducible promoter (Kafri et al., 1999; Stewart et al., 2009). This is followed by limited dilution cloning or flow cytometry assisted sorting of

individual cell lines, characterisation of their growth, screening based on viral vector productivity, followed by bioreactor and stability studies. Manual limited dilution cloning in antibiotic selective media used for developing new clones is labour intensive and takes a long time. Culturing a large number of individual clones creates a bottleneck in cell line development process and extends the overall development time and cost.

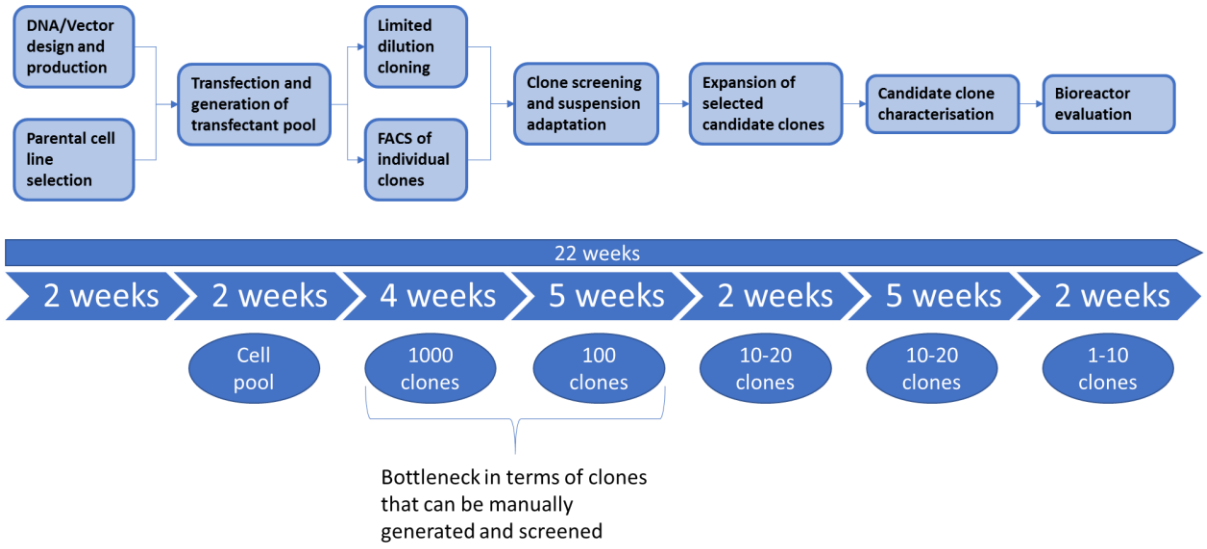


Figure 82: Manual approach to cell line development summary flow chart. A bottleneck in number of generated clones is highlighted.

Oxford BioMedica (OXB) has moved away from manual cell line development and transitioned to an automated cell screening system (ACSS) capable of high throughput screening of clones to find best producers within a shorter time frame (Figure 83). ACSS allows generation of thousands of clones. This high throughput method addresses the need for debottlenecking the cell line development process through automation. However, increased number of clones means that it becomes increasingly challenging to select the clones correctly and efficiently with a high productivity potential. While it is possible to screen cell productivity at small scale this still does not address the issue of assessing cell line performance at larger bioreactor scale which is required for successful transition of a cell line into manufacturing. It has been demonstrated that small scale studies often do not reflect large scale performance and some potentially high producers are discarded due to poor small-scale performance (Porter *et al.*, 2010). Therefore, a method for large scale evaluation is required to confirm the results of small-scale studies and help in triaging the clones potentially capable of achieving high LVV titres.

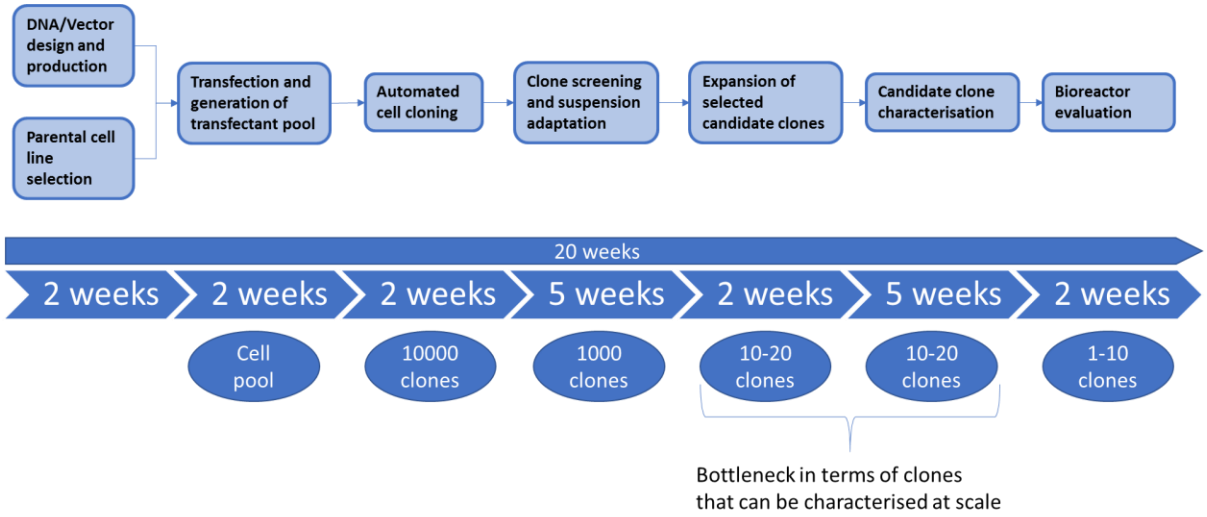


Figure 83: Automated approach to cell line development summary flow chart.

4.1.3. Partial least squares discriminant analysis

Partial least squares (PLS) regression is a statistical method commonly applied in multiple areas of industry and academia, especially chemometrics (S. Wold et al., 2001). It models the covariance structure between two matrices where X block is a predictor matrix and Y block is a response matrix. The data is projected to latent variables (LVs) determined by scores and loadings for both data sets in a way that the predictor (X) data set explains the most variation of the response (Y) data through maximisation of covariance between the two matrices (S. Wold et al., 2001). Through this process the model can be applied to uncharacterised set of predictors (X) to estimate their associated response (Y) values which is achieved by comparing new samples' scores to the calibration data set. Partial least squares discriminant analysis (PLS-DA) is a variant of the algorithm where the Y block data is categorical i.e. it belongs to defined classes such as a high or low producer rather than being characterised by a continuous variable such as viral vector titre produced by a cell (Barker & Rayens, 2003). The functional LVV titre is determined through fluorescence assisted cell sorting (FACS) method which is a highly variable measurement (FACS results and variation is discussed in more detail in sections 4.3.1 and 4.4.1). Therefore the ability to accurately predict the LVV titre through modelling is limited and the use of PLS-DA as a categorical classification algorithm is more justified as discussed further in section 4.3.2. This approach allows greater control over sample classification through adjustment of classification threshold and can be used to develop a more focused and practically oriented model which can be beneficial when dealing with highly variable measurements.

PLS and its variants have been applied to monitor and guide multiple aspects of process development and control. This includes process monitoring and control using Raman spectroscopy in ambr15® system (Rowland-Jones et al., 2017) where PLS was applied to spectroscopy data to measure metabolite concentration at-line to improve daily monitoring capabilities. PLS-DA was used along with MALDI-ToF MS to develop a predictive model of cell line performance in Chinese hamster ovary cells used for monoclonal antibody production (Povey et al., 2014). In this case the cell performance (as determined by monoclonal antibody titre produced by the cells) data was used to classify cells as either high or low producers with 4000 mg/L titre as selection threshold. The mass spectra of the cell samples were pre-processed using MATLAB Bioinformatics Toolbox (in order of application: resampling, baseline correction, curve smoothening, peak alignment, outlier detection, and normalisation. See section 3.2.7 for detailed description of MS pre-processing) and used to first calibrate the model and then predict the productivity of new cell lines using PLS-DA. The method reduced cell line development campaign timeline by up to 7 weeks through improved and faster clone selection.

This chapter discusses an application of PLS-DA and MALDI-ToF MS in human embryonic kidney cells 293T (HEK293T) to predict cell line performance in LVV production which can be used to accelerate cell line development process. Methods of cell line cloning, mass spectra generation and PLS-DA modelling are described, followed by discussion of model performance and its application within industry.

4.2. Methods

Multiple cell lines have been developed and used for LVV production to develop a PLS-DA model. Cell samples were collected for model calibration and validation and used to obtain mass spectra. MS and LVV titre data were used to develop the PLS-DA model which was subsequently validated to assess its performance.

4.2.1. Cell line development

Cell samples used in the experiments were generated from multiple packaging cell lines. The parental cell lines were engineered to conditionally express *gag*, *pol* and *env* viral vector genes but not the gene encoding the therapeutic protein. These cells were taken through a series of cloning and screening steps including automated cell screening system (ACSS) operated by the OXB cell engineering group. Several cells were selected as a representative sample of packaging cells developed at the company. These cells were adapted for suspension culture. The cells were cultured in T75 flasks (Thermo Fisher Scientific) in DMEM (GE Healthcare Biosciences) supplemented with 10% foetal bovine serum (FBS, Life technologies or Gibco). The cells were adapted to suspension growth through gradual transition from DMEM with FBS to serum-free FreeStyle™ 293 media (Thermo Fisher Scientific) by supplementing an increasing proportion of FreeStyle™ 293 media at each subsequent cell passage. During this process, cell counts and viability were monitored and seeding densities and rate of FreeStyle™ 293 addition was adjusted based on these results. At the final stage of suspension adaptation the cells were transferred to 125 ml conical shake flasks (Corning) and cultured in FreeStyle™ 293 media supplemented with 0.1% cholesterol lipid concentrate (CLC; 0.1% v/v using 250x Cholesterol Lipid Concentrate, Thermo Fisher Scientific). Cell line development was performed in two batches, one for calibration and one for validation, resulting in generation of 10 calibration cell lines and 8 validation cell lines, developed independently from each other using the same process. The cell culture was scaled up and used as a seed culture for cell and vector sample production in bioreactors.

Cells were cultured in FreeStyle™ 293 media supplemented with 0.1% CLC in conical shake flasks (250 or 500 ml, Corning) for a week prior to inoculation. Bioreactor vessels were assembled, tubed and autoclaved prior to the experiment. MiniBio reactors (500ml, Applikon biotechnology) were filled with FreeStyle™ 293 media to a working volume of 350 ml. The bioreactors were inoculated with previously cultured cells. After 24 hours the bioreactors were transfected with a HIV-based genome plasmid encoding green fluorescent protein (GFP) using Lipofectamine® Transfection Reagent (Thermo Fisher

Scientific) in FreeStyle™ 293 media. At the same time, viral protein gene expression was induced using doxycycline addition. 20 hours after transfection all cells were induced using sodium butyrate (NaBu, 10mM). Cells and vector samples were harvested 24 hours after induction in aseptic conditions. For ambr15® micro bioreactors (Sartorius Stedim Biotech) the abovementioned steps were programmed, automated and performed at smaller scale (12 ml working volume). All bioreactors were sampled daily to monitor process parameters.

Cell concentrations were measured using NucleoCounter® NC-200 (Chemometec). Cells were aliquoted and centrifuged (3000 rpm, 5 minutes) to match the total viable cell count of $1.5x10^5$ cells per sample. Supernatant was carefully discarded and pellets were washed with phosphate buffered saline (Thermofisher scientific) and frozen. For the vector samples, cells were separated through centrifugation (3000 rpm, 5 minutes) and the supernatant was filtered (0.22 µm, Fisherbrand™ syringe filter, Fisher Scientific), aliquoted (1.5ml cryotubes) and frozen at -80^{o}C.

4.2.2. Sample preparation and analysis overview

Frozen cell samples were thawed and re-suspended in 50 µl of MS matrix consisting of HPLC grade water (Sigma-Aldrich) with 40% HPLC grade acetonitrile (Sigma-Aldrich), 0.15% trifluoroacetic acid (99% purity, Across chemicals) and 10 mg/ml α-cyano-4-hydroxycinnamic acid (Sigma-Aldrich). Cells were incubated in the matrix at 4 °C for 1 hour 15 minutes. 1 µl of the sample/matrix mix was spotted on 384 well ground steel MALDI-ToF plate (Bruker) 6 times per each samples and left to dry at room temperature. The sample plate was left at room temperature for 24 hours before the analysis to allow for transport. The plate was loaded to MALDI-ToF MS and analysed through automated protocol (Bruker Ultraflex; laser intensity 62%; laser frequency 500 Hz; polarity: positive; ions sources:1. 24.93 kV, 2. 23.08 kV, lens 7.5 kV; Pulsed ion extraction 400 nS; Suppress at 4 kDa; spectra collected in the range of 4-60 kDa; sample rate 0.13 Gs/s, 3600 ionisation laser shots summed and saved per sample). The resulting data files were collated using a script in R version 3.2.4 and imported to MATLAB® version R2013 (Mathworks).

Viral vector samples were used to calculate the viral vector titre using fluorescence-activated cell sorting (FACS) flow cytometry-based transduction assay (FACSVerse™, BD BioScienses). HEK293T cells were seeded in a 96 well plate and incubated for 24 hours in 150 µl DMEM with 10% FBS media with polybrene (1 in 400 dilution, Sigma Aldrich). In a separate plate, 24 hours after seeding, viral vector was diluted 1 in 100 in DMEM and mixed

thoroughly. Media was removed from the original plate and 50 µl of diluted vector was added to each well to transduce the cells. 3 hours after vector addition 100 µl of DMEM supplemented with polybrene were added to each well and incubated for 72 hours. Afterwards the media were removed and 50 µl TrypLE™ (Thermo Fisher Scientific) was added to detach the cells. After 5 mins incubation at 37°C TrypLE™ was neutralised with 150 µl DMEM with 10% FBS and cells were re-suspended. 100 µl of transduced cells were transferred to a round bottom 96 wells plate with 150 µl in each well for a total volume of 250 µl per well. Transduced cells were analysed using FACSVerse™ to estimate the viral vector titre based on the number of cells which express GFP. The assay followed the same protocol as for data described in Chapter 2 but the assay variation was higher with CV of 10-50% between biological replicates which was accounted for by secondary FACS analysis to confirm original titres.

4.2.3. Partial least squares discriminant analysis modelling

PLS-DA was used to model performance of proprietary OXB packaging cell lines. 18 cell lines have been characterised and used to develop and validate the PLS-DA model. For model calibration, 10 cell lines were grown in ambr15® and MiniBio reactors, each in three vessels; for each vessel three harvest samples were collected and analysed with MALDI-ToF and FACS where each individual sample was incubated with the matrix and spotted on the plate six times (Figure 84). The total number of calibration data points was 540. 38 data points were rejected due to poor MS signal resulting in 502 samples used in model calibration.

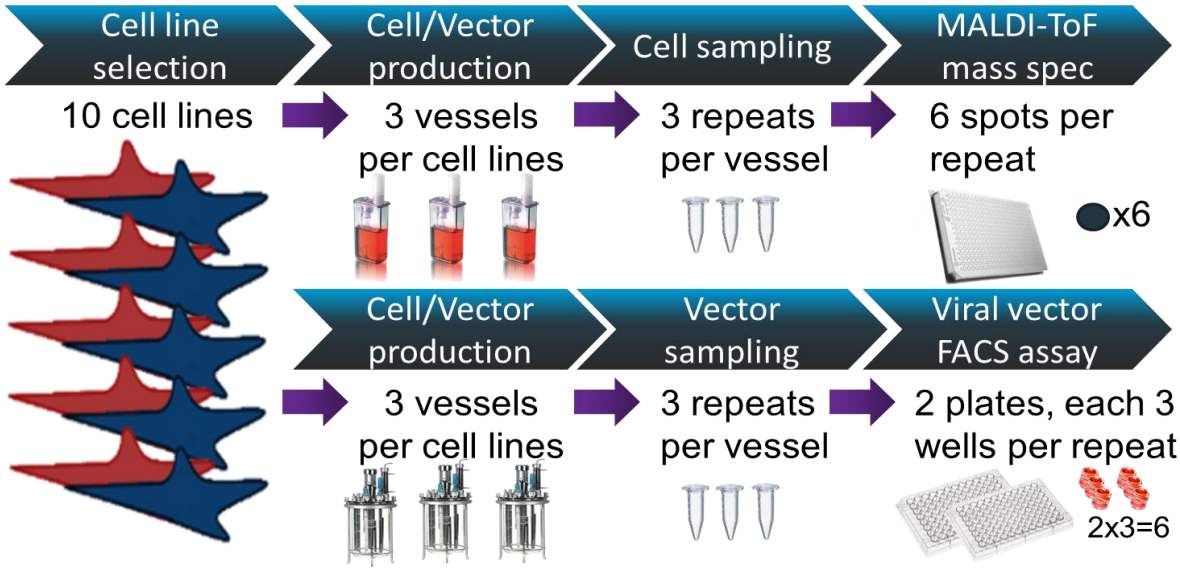


Figure 84: PLS-DA model calibration data generation summary flow chart.
Each of the 10 cell lines was cultured in 3 bioreactors (3 ambr15® and 3 MiniBio systems); 3 cell and vector samples were collected from each bioreactor; 6 spots were placed per each cell sample; each vector sample was assessed in 6 replicates.

Mass spectra obtained from ambr®15 were pre-processed as described for principal component analysis (PCA) using baseline correction, signal normalisation and Savitzky-Golay smoothening (see Chapter 3 section 3.2.7) and used as X-block for the PLS-DA model. LVV titre from MiniBio reactors measured by FACS assay was used to assign a class of low or high producer to each of the cell lines samples. Classification threshold of a high producer was chosen as $2x10^7$ TU/ml with other classification thresholds examined as possible alternatives. High/low producer classification was used as Y block data represented by a number: 1 (class 1, low producer) or 2 (class 2, high producer). The model was developed using PLS Toolbox version 8.0.1 (Eigenvector) in MATLAB® version R2013 (Mathworks). The model was cross-validated using several custom methods where data was organised into blocks based on either cell line, production vessel or biological repeat. The results were inspected to select the optimal method. The final cross-validation was performed by organising the data into 10 blocks corresponding to the 10 cell lines and leaving out one of the cell lines as a validation set in multiple iterations of cross-validation. Through this method, data from one of the cell lines was used as cross-validation data against the remaining 9 cell lines and the process was repeated for the next cell line until all cell lines were used for cross-validation. The number of LVs used in the model was selected based on the variability of both X and Y block captured by the model. The target variability captured was between 80-90% variability in order to account for complexity of the spectra but also to prevent overfitting the model to data which would occur with too high number of LVs. This resulted in a model based on 11 LVs. Other models based on different number of LVs were examined to ensure best model performance.

For model validation, 8 new cell lines were used to generate MS and viral vector titre data as described for the calibration. However, only one culture vessel was used per cell line (Figure 85). The total number of validation data points was 144. The samples were applied to the model and their classification scores and overall classification were examined to assess model performance. To ensure understanding of the classification process, contribution plots for each LV were examined by multiplying the loadings matrix by the average spectrum.

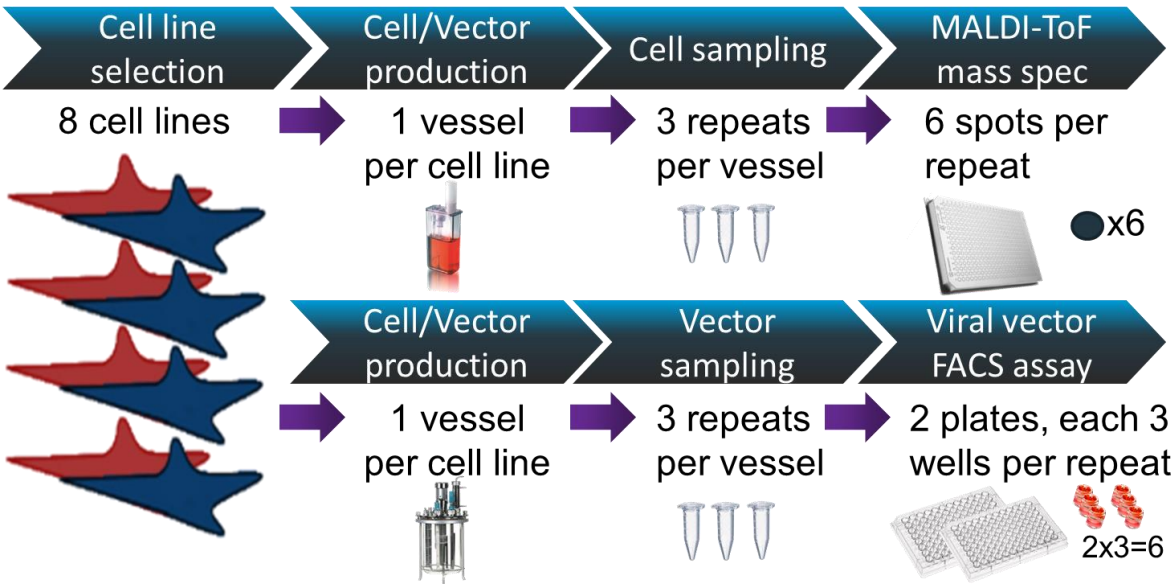


Figure 85: PLS-DA model validation data generation summary flow chart.
Each of the 8 cell lines was cultured in 1 bioreactor (1 ambr15® and 1 MiniBio systems); 3 cell and vector samples were collected from each bioreactor; 6 spots were placed per each cell sample; each vector sample was assessed in 6 replicates.

4.3. Results

Development of the PLS-DA model spans a large body of work, starting with development of packaging cell lines used for calibration and validation of the model using MS data from the cell samples and viral vector titre data as measured by FACS. Model development itself was an iterative process that required optimisation of model parameters. Multiple variants of the model were developed to identify a combination of parameters needed for accurate prediction of cell line performance.

4.3.1. Cell line development

The first set of cells has been selected from packaging cell lines developed at Oxford BioMedica. The cells were adapted from adherent culture to suspension culture. Selection was performed based on viral vector productivity of the cells at T-flask scale adherent culture as determined by FACS assay. The selection was aimed to obtain a representative distribution of high and low producers (Figure 86). However, because the initial selection was focused on obtaining a sufficient number of high and low producers, the medium producers (with viral vector titre range between $1x10^7$ and $2x10^7$ TU/ml) were underrepresented in the calibration data. While this was not an intentional decision, the variation in cell line performance, differences between adherent and suspension culture performance and lack of information to guide the initial selection process made it difficult to select cells with a continuous performance distribution.

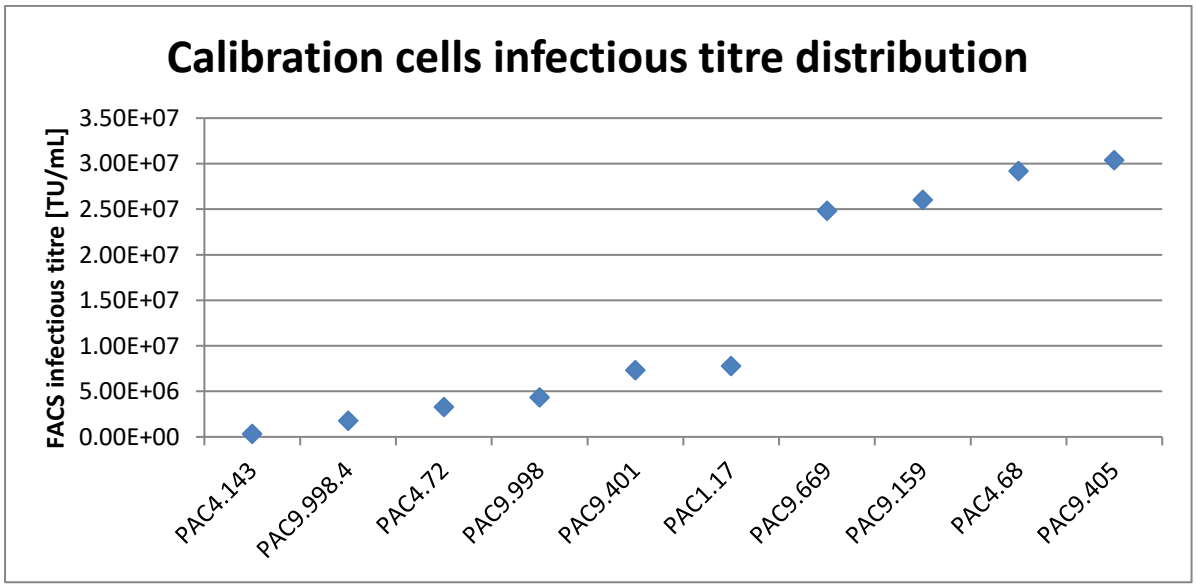


Figure 86: Distribution of LVV infectious titre obtained from each cell line in the calibration set.

The cells used in model calibration were obtained from OXB cell engineering group shortly after their suspension adaptation and were directly used for seed culture of the

bioreactors. For the validation, several different cell lines were selected at small scale adherent culture stage and then adapted to suspension through gradual media exchange. Cell viability (Figure 87) and count (Figure 88) were monitored throughout the process. Most cells maintained high viability (above 90%) throughout the adaptation process with individual episodes of temporary drop in viability which was compensated for by slowing the rate of media change between cultures. At the end of the adaptation process all cells achieved viability of over 90%. The viable cell number is characterised by alternating high and low values which is caused by bi-weekly cell splitting regimen which caused the cells to be transferred to new media every 3 or 4 days which results in higher cell counts every second count when cell were cultured for an additional day compared to the previous cell count. The cells maintained a high cell count (above $1x10^6$ cells/ml) following the initial decrease after revival. Some cells showed a drop in cell count at the final stage of adaptation when cells were completely transferred to FreeStyle™ 293 media which is often observed during the final stage of suspension adaptation. 4 of the 12 cell lines suffered from decreased viability and slow cell growth after adaptation and during preparation of the seed culture. These cells were not suitable for use with bioreactors due to insufficient number of viable cells. PAC9.333, PAC9.691, PAC9.1133 and PAC 9.1358 were discarded.

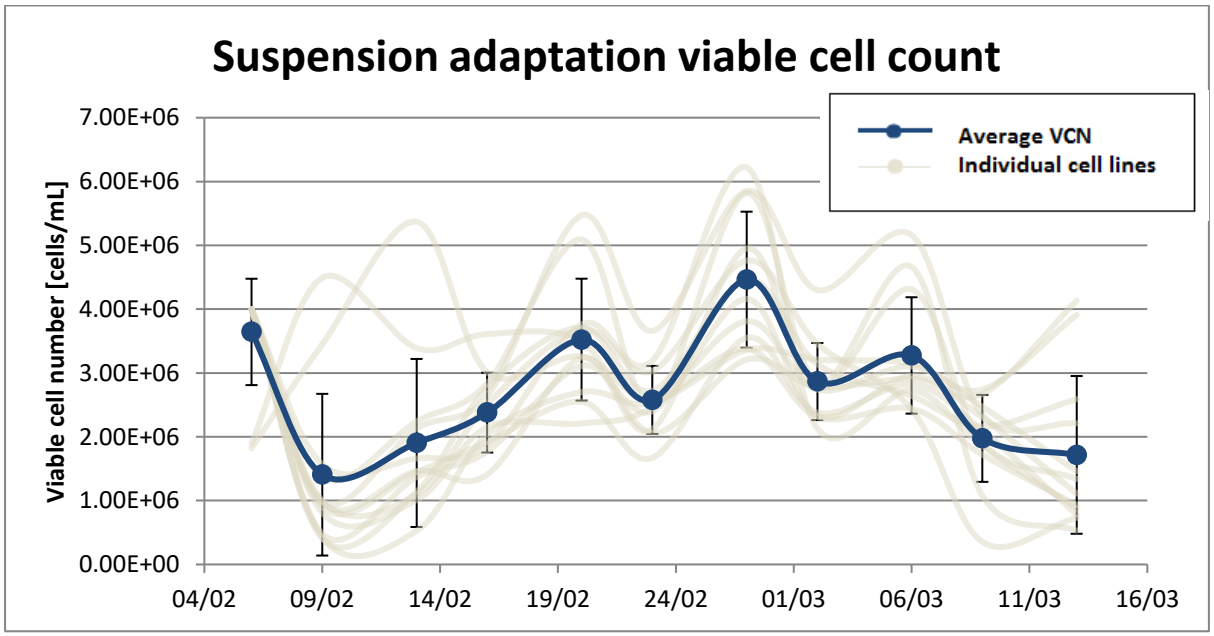


Figure 87: Cell viability of 12 cell lines (validation set) throughout the suspension adaptation process. The error bars represent standard deviation of all cell lines.

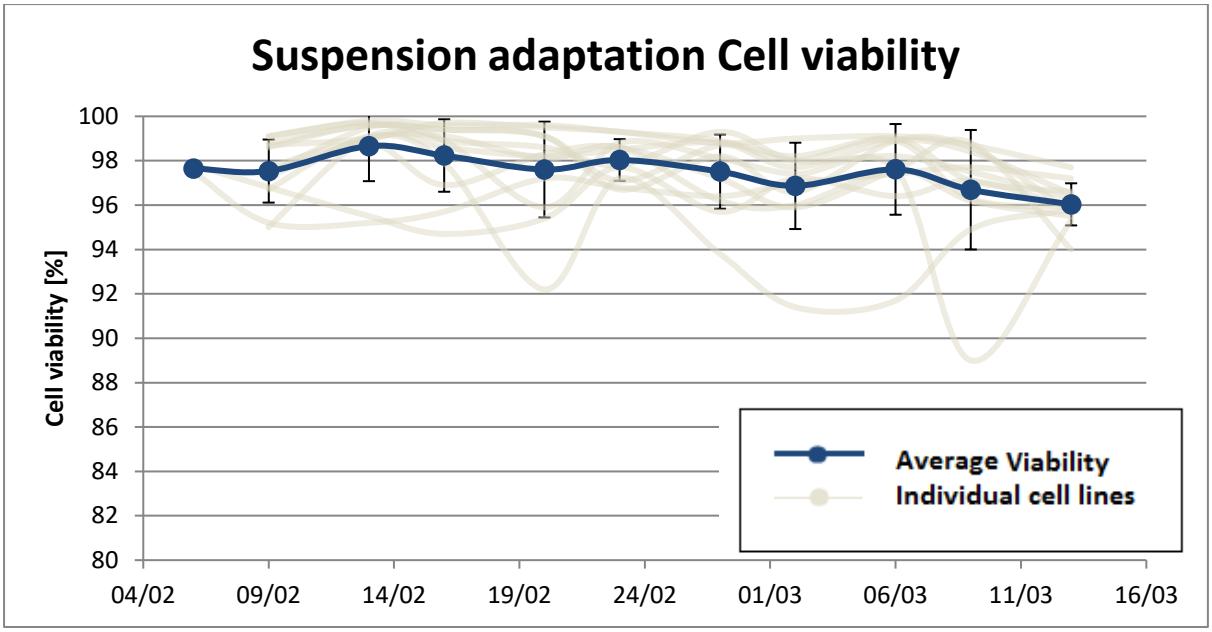


Figure 88: Viable cell number of 12 cell lines (validation set) throughout the suspension adaptation process. The error bars represent standard deviation of all cell lines.

The remaining cells were cultured in ambr15® and MiniBio reactors and their fitness has been monitored throughout the process (Figure 89-Figure 92). For MiniBio reactors all cell lines with the exception of PAC4.143 showed a similar trend of initially high viability (90%+) which drops throughout the process due to viral vector production (down to 70%-80% at the point of harvest). A similar trend was observed for ambr15 with overall slightly lower cell viability throughout the process with some cell lines reaching less than 60% viability at harvest. There is a degree of difference between cell lines rate of growth which was already observed during the seed train culture and it results in the final cell counts at the point of harvest varying between $1.5x10^6$ to $3x10^6$ cells/ml. There is a complex relationship between LVV production and the total cell number in culture. Viral vector production reduces cell viability and therefore reduces the total number of cells, suggesting that cell lines characterised by a high final viable cell count at the point of harvest may have produced less viral vectors and therefore their growth was not inhibited. At the same time a low cell number may indicate overall poor cell fitness and insufficient biomass to produce viral vectors in high concentration. Therefore, while the viable cell count and cell viability are good indicators of the overall cell culture fitness throughout the process, they should not be used as an indicator for viral vector productivity of the cell lines. The cell numbers used for MS sample preparation were kept constant at $1.5x10^5$ viable cells per sample.

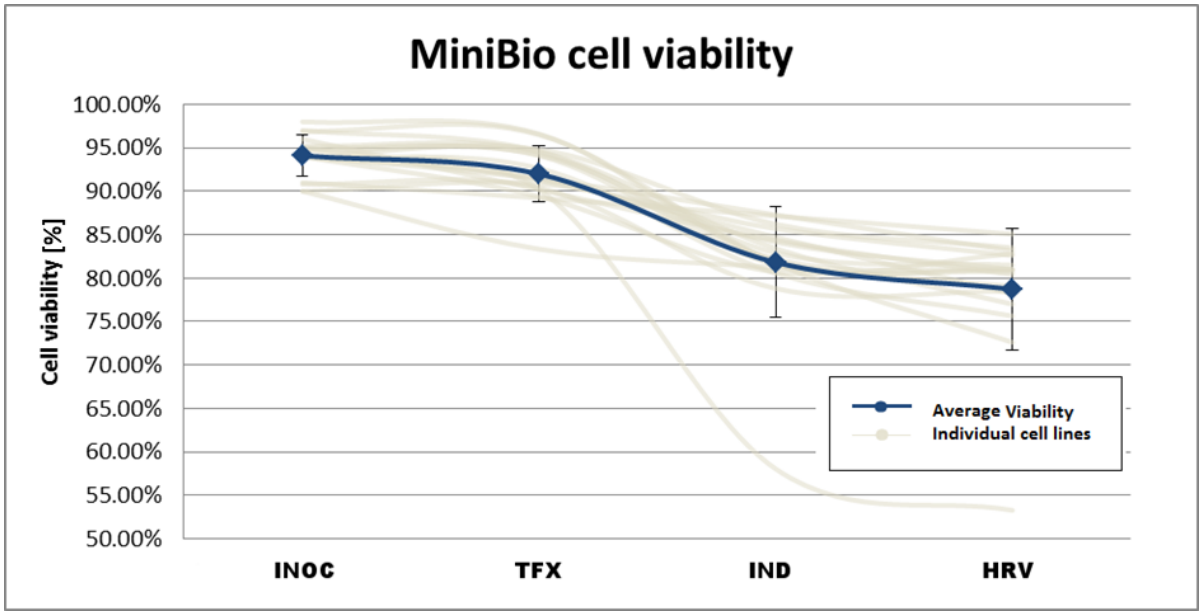


Figure 89: Cell viability of 18 cell lines (calibration and validation sets) throughout production in MiniBio reactors. INOC – Inoculation; TFX – Transfection; IND – Induction; HRV – Harvest; The error bars represent standard deviation of all cell lines.

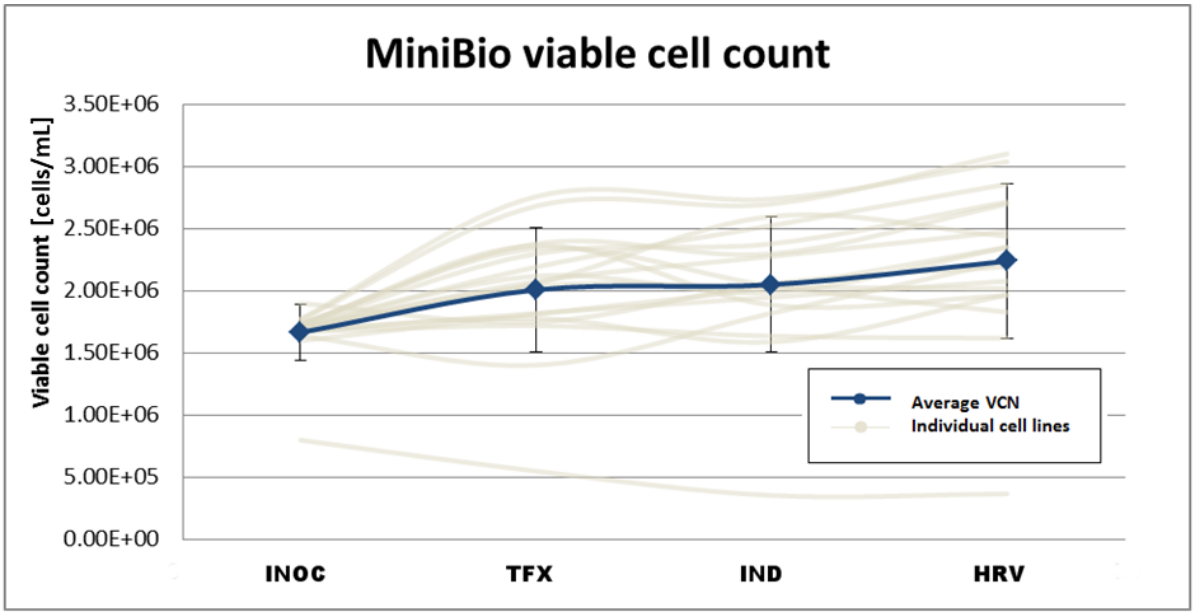


Figure 90: Viable cell number of 18 cell lines (calibration and validation sets) throughout production in MiniBio reactors.
INOC – Inoculation; TFX – Transfection; IND – Induction; HRV – Harvest; The error bars represent standard deviation of all cell lines.

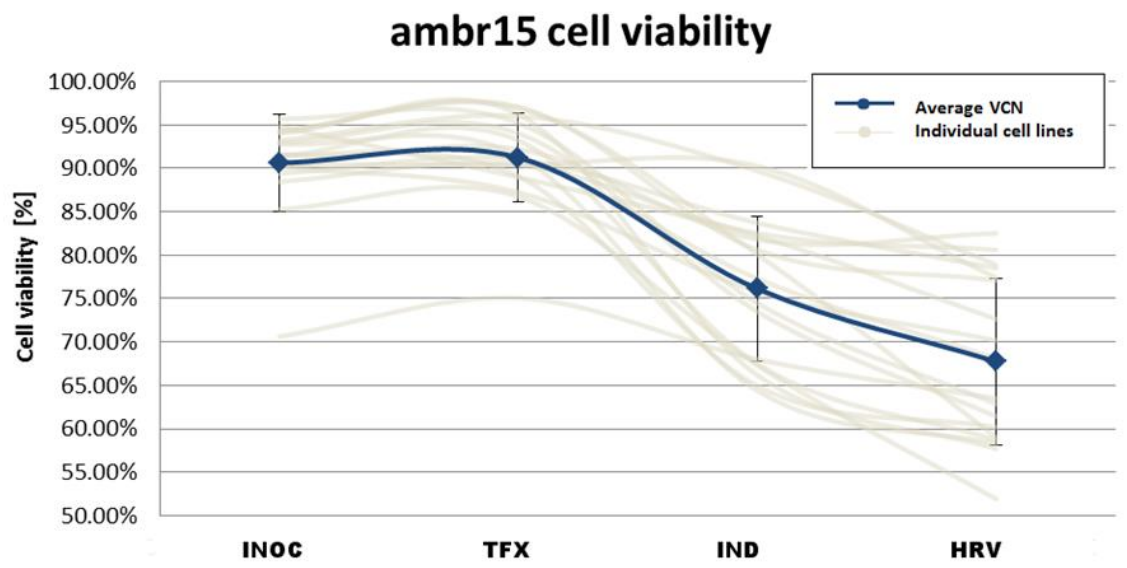


Figure 91: Cell viability of 18 cell lines (calibration and validation sets) throughout production in ambr15®. INOC – Inoculation; TFX – Transfection; IND – Induction; HRV – Harvest; The error bars represent standard deviation of all cell lines.

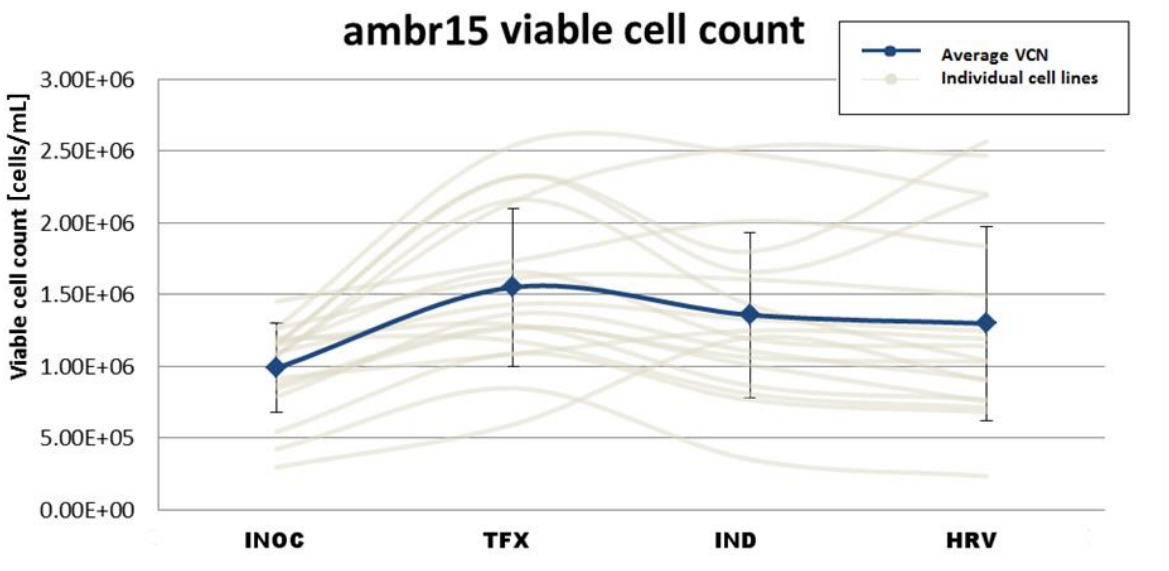


Figure 92: Viable cell number of 18 cell lines (calibration and validation sets) throughout production in ambr15®. INOC – Inoculation; TFX – Transfection; IND – Induction; HRV – Harvest; The error bars represent standard deviation of all cell lines.

Cell performance was assessed after vector production using FACS transduction assay. It is an inherently variable method influenced by the raw materials, cell culture and stochastic nature of the transduction process (Geraerts et al., 2006). Each sample was tested in triplicate and the assay was performed twice at separate time points to compensate for assay variability and to ensure accuracy of the viral vector titre data. Viral vector titre data was obtained for two different scales of bioreactor system. The first was the Sartorius Stedim Biotech ambr15® fermentation automated microscale bioreactor system that uses 10-15mL (working volume) single-use vessels to mimic the characteristic of a classical lab scale

bioreactor. Secondly, the Applikon® MiniBio bioreactors (0.5L total volume) are a true scale down of large bioreactors. Only the larger scale data (from MiniBio reactors) was used to classify the cells as high or low producers to reflect the significance of large-scale performance over small scale performance.

Figure 93 shows the results of FACS transduction assay analysis of the 10 cell lines used for calibration of the model along with a negative and positive (HIV-GFP standard) control. The results were averaged from 54 assays per cell line. Overall, there were four cell lines which produced above $1x10^7$ TU/ml and three of these consistently produced above $2x10^7$ TU/ml. Several iterations of the model used different titre level as classification threshold. Samples from PAC9.669 achieved an average titre of $2.48x10^7$ TU/ml but with higher than average variation and multiple observations below $2x10^7$ TU/ml. As such PAC9.669 was used as either high or low producer depending on the iteration of the model and the exact value of the classification threshold used. Cell lines PAC 4.68, PAC9.159 and PAC9.405 achieved above $2x10^7$ TU/ml and were classified as high producer. The remaining cell lines were classified as low producers.

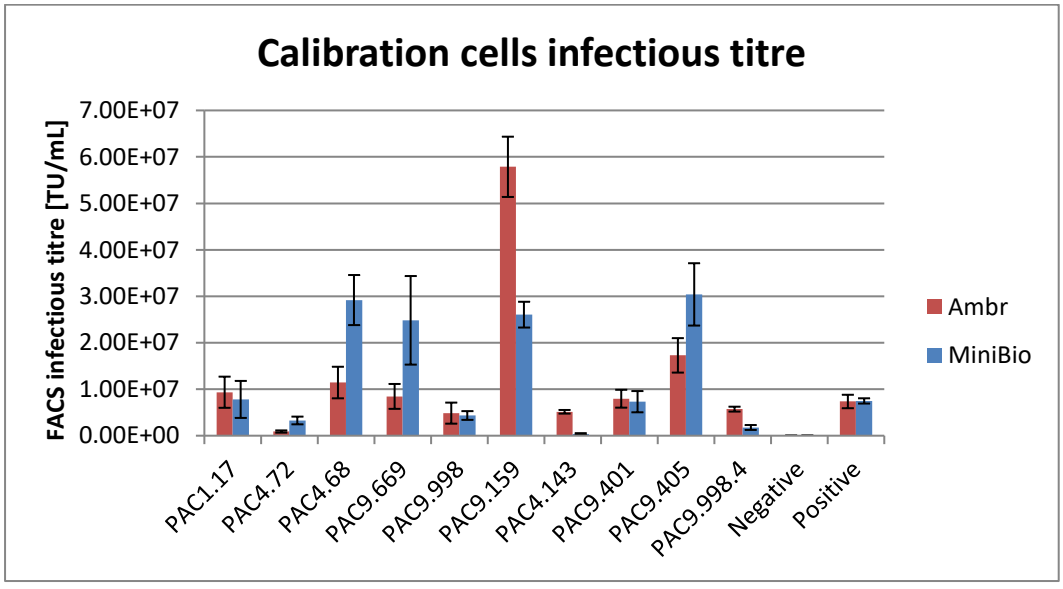


Figure 93: Viral vector infectious titre (FACS assay) for calibration set cell lines as measured. Error bars represent standard deviation of analytical replicates (3 levels of dilutions, 3 repeats per dilution, total 9 replicates).

The second set of cells was obtained by selecting 12 cell lines at the early stage of the development process. The cells were adapted to suspension culture following the protocol developed by OXB cell line engineering group and originally used for the first set of cell lines as described above. During the adaptation process cell viability was monitored to adjust the rate of media transition from DMEM with FBS to serum-free FreeStyle™ 293 medium.

After suspension adaptation process 8 cell lines with the highest viability were selected for bioreactor characterisation in ambr15® and MiniBio systems (a distribution of infectious titre reached in each cell line is presented in Figure 94).

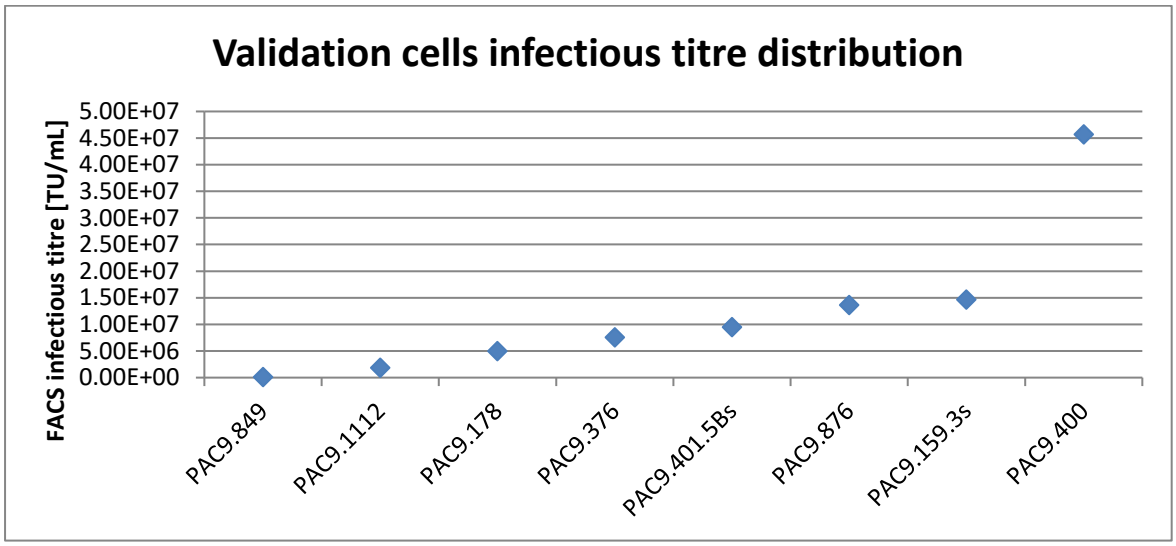


Figure 94: Distribution of LVV infectious titre obtained from each cell line in the validation set.

As with previous data set, the bioreactor viral vector production process was monitored and viral vector titre was assessed through FACS transduction assay with results averaged from 18 assays per cell line. Among the 8 selected cell lines, 3 achieved infectious titre above $1x10^7$ TU/ml while only a single cell line produced more than $2x10^7$ TU/ml (Figure 95). PAC9.400 was classified as a high producer; PAC9.159.3s and PAC9.876 were classified as either low or high producer depending on the classification threshold used in the model and the remaining cell lines were classified as low producers.

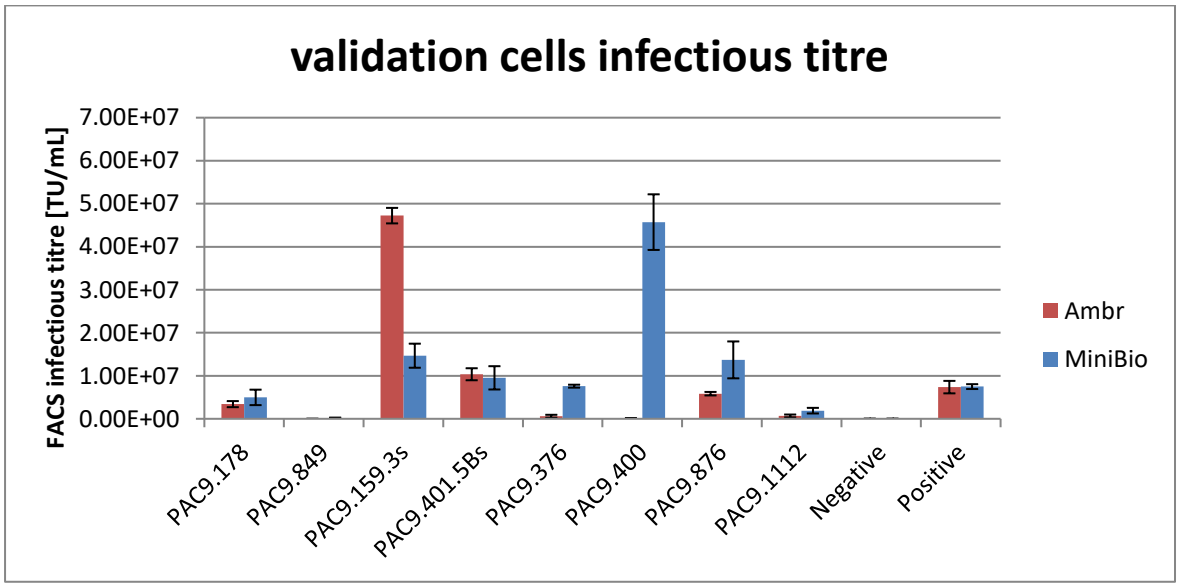


Figure 95: Viral vector infectious titre (FACS assay) for validation set cell lines.
Error bars represent standard deviation of analytical replicates (3 levels of dilutions, 3 repeats per dilution, total 9 replicates).

Altogether the cell line development resulted in characterisation of 18 cell lines used in calibration and validation of the PLS-DA model (Figure 96) The range of viral vector titre produced by the cells was from $1.46x10^5$ to $4.57x10^7$ TU/ml with 11/18 cells characterised by titre below $1x10^7$ TU/ml. 4 cell lines were classified as high producer based on $2x10^7$ TU/ml classification threshold. 3 cell lines were classified as medium range producers based on the titre between $1x10^7$ and $2x10^7$ TU/ml. The remaining cells were classified as low producers.

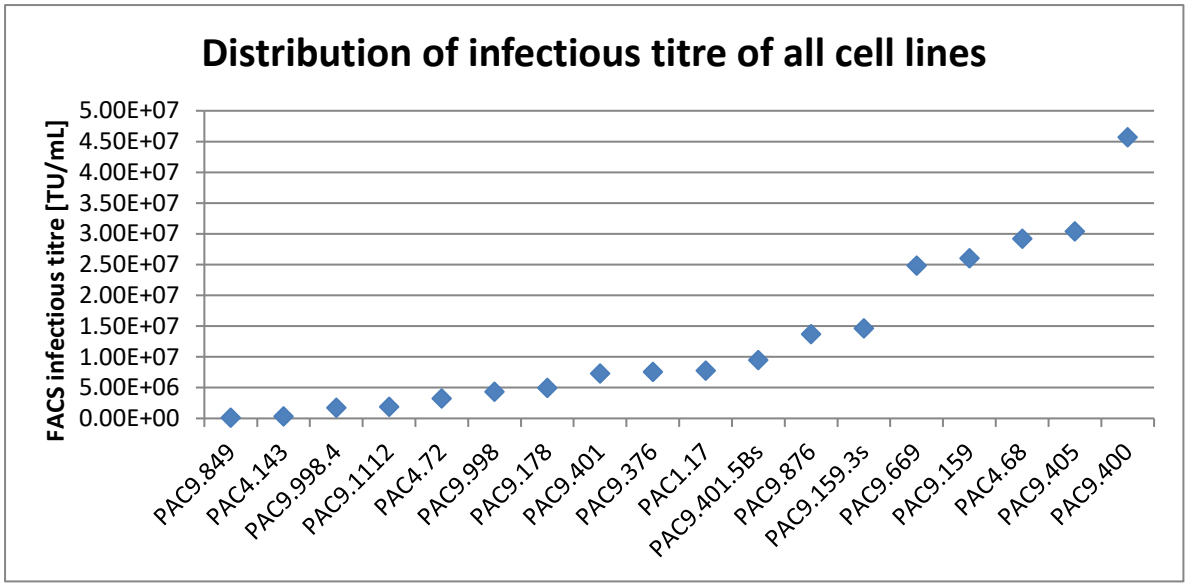


Figure 96: Distribution of LVV infectious titre obtained from all cell lines (calibration and validation set).

When looking at the cell lines used in calibration and validation data sets together it becomes apparent that there is a degree of variation in production performance between ambr15® and MiniBio scale (Figure 97). A significant difference in titre across scales was observed in about half of the cell samples. This is illustrated in Figure 98 where top producers at ambr scale are characterised by medium to high infectious titre in MiniBios while at MiniBio scale the top producers show high variability in the ambr scale infectious titre. This highlights that small scale performance can be indicative of large-scale potential but it is not always the case and there is a need for characterisation of production at larger scale than ambr15®. Without large scale characterisation, potentially high producers could be discarded due to poor performance at early stage of development (e.g. PAC9.400).

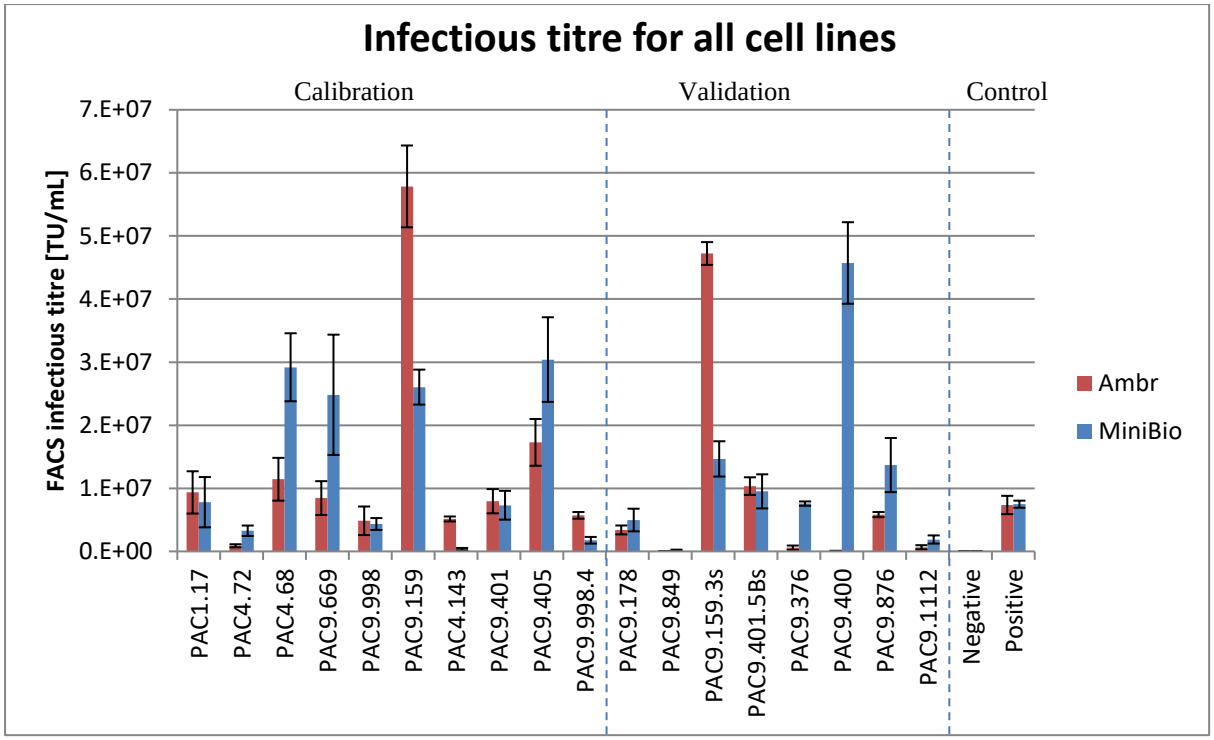


Figure 97: Viral vector infectious titre (FACS assay) for all cell lines (calibration and validation sets). Error bars represent standard deviation of analytical replicates (3 levels of dilutions, 3 repeats per dilution, total 9 replicates).

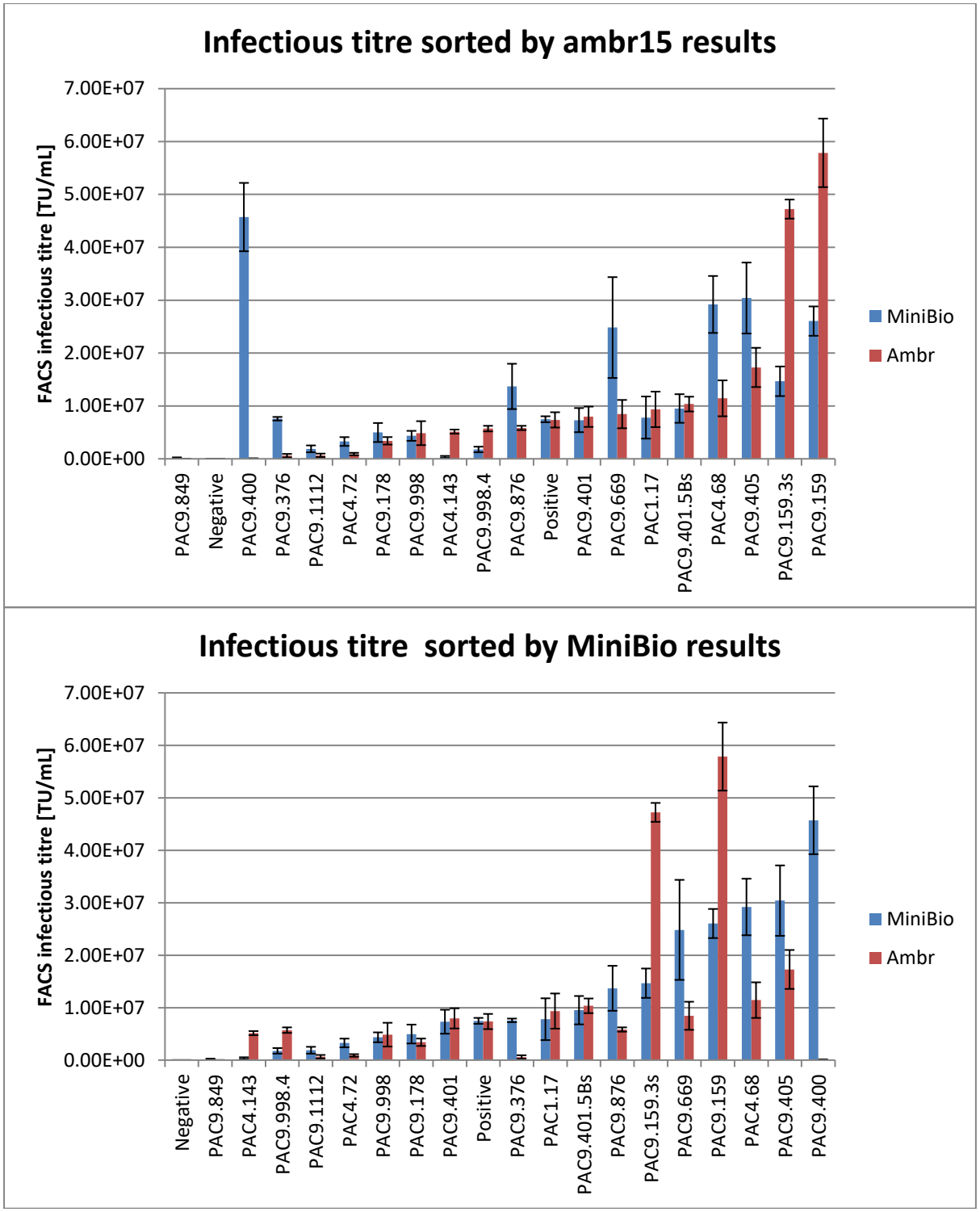


Figure 98: Viral vector infectious titre (FACS assay) sorted by ambr results (top) or MiniBio results (bottom). Error bars represent standard deviation of analytical replicates (3 levels of dilutions, 3 repeats per dilution, total 9 replicates).

4.3.2. Partial least squares discriminant analysis model development

MALDI-ToF MS methodology was applied in the area of cell line development to develop a PLS-DA model of large-scale cell line performance. In this experiment MALDI-ToF MS is used to obtain the mass spectra from cells cultured in ambr®15 micro bioreactor while viral vector titre is assessed at 0.5L Applikon® MiniBio reactor scale using the same

cell lines. These two types of data are combined to calibrate the PLS-DA model. The model's aim is to provide a platform for characterisation of clones generated by OXB's ACSS.

The initial MS data was screened by examining the intensity of individual samples. Several samples from cell line PAC4.72 were characterised by unusual spectral profiles compared to the other mass spectra. Direct inspection of the unusual samples revealed low intensity of the spectra where they showed no peaks. The overall low intensity contributed to poor quality of 38 samples. These samples were discarded and not used in the final analysis to improve overall model performance. The remaining samples were inspected using PCA (Figure 99) to assess the data structure and guide further selection of data for PLS-DA model calibration.

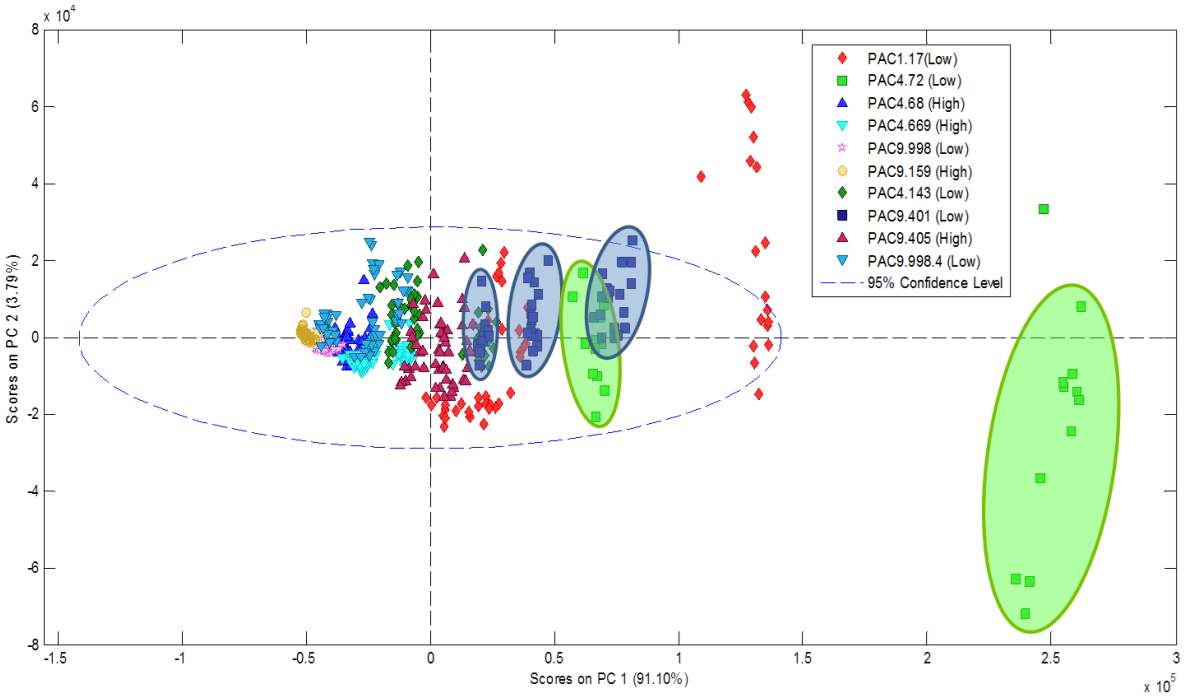


Figure 99: PCA scores plot (PC1 and 2) for the mass spectra of cell samples used in PLS-DA calibration. Blue and green circles indicate PAC9.401 and PAC4.72 clusters of samples originating from different bioreactors.

The analysis shows a degree of data clustering based on several properties. Principal component (PC) 1 captures the majority of variation (91.1%) and the data points are distributed between negative and positive PC1 scores. The distribution aligns with cell line origin of the data points; PAC1.17, PAC4.72 PAC 9.401 and some of the PAC 9.405 have positive PC1 scores while the rest of the samples were characterised by negative values. For PC2 (3.79% variability captured) the distribution of negative and positive values does not align with cell line origin and samples from multiple cell lines are characterised by both low and high scores. Looking at the PC1 and 2 scores plots, there is no clear separation of high and low producer cells. An important observation is that for some of the cell lines

(e.g.PAC4.72 or PAC9.401) there are 2 or 3 clusters of data points which are corresponding to samples that were obtained from separate bioreactors (blue and green circles on Figure 99). Moreover, there is a higher degree of variation in PC scores for cell lines PAC1.17 and PAC4.72 compared to the other cell lines. Despite the higher level of variation in the samples from these cell lines they were included in the model due to low number of available data points for PAC4.72 due to previously mentioned low intensity of some of the spectra.

PLS-DA Model calibration was performed using 10 cell lines. 502 MS data points were used (54 per cell line except PAC4.72) to maintain a high number of repeats for each cell line. All data was pre-processed using baseline correction, signal normalisation and Savitzky-Golay smoothening. The MS data was used as the X block data while the classification of cells as high or low producer based on viral vector titre was used as the Y block data. Several cross-validation methods have been considered to assess the quality of the model (Figure 100). The default method using venetian blinds selection to split samples into calibration and cross-validation sets was replaced by custom methods taking into account the organisation of data into blocks of either cell lines (54 data points per block, except PAC4.72), bioreactors (18 data points per block except PAC4.72) or biological samples (6 data points per block). While these methods were similar to the contiguous blocks method (second example in Figure 100), the custom methods account for variation in the number of samples per cell line and bioreactor caused by skipping some of the poor quality samples as explained at the beginning of section 4.3.2. Each method was tested to determine the optimal approach to cross-validation.

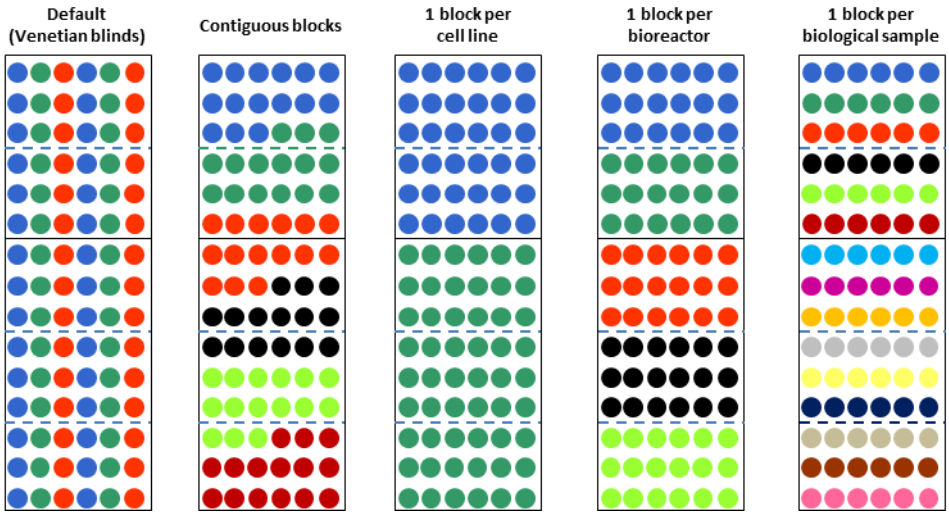


Figure 100: Summary diagram of PLS-DA model cross-validation methodology. An example using data from 2 cell lines; each dot represents a single mass spectrum corresponding to a single spot on the MALDI plate; each row represents 6 spots per biological sample; samples from separate bioreactors are separated by a dashed line; Samples from separate cell lines are separated by a straight line. Different colours indicate the split of samples into data sets used in calibration or left out for cross-validation.

The initial cross-validation using 6 data points per block resulted in cross-validation error ranging from 0.126 to 0.369 depending on the number of LVs chosen for the model (Figure 101). The method using combined samples from each bioreactor in the block achieved similar results with cross-validation error ranging from 0.186 to 0.374 (Figure 101). In both cases the error tends to decrease as the number of LVs included in the model increases. The final cross-validation method included all samples from each cell lines as a single block to represent the intended application of the model where it is used to classify samples from new cell lines. In this case the cross-validation error was higher, ranging from 0.495 to 0.708 (Figure 101). The error values varied with the number of included LVs with a minimal error achieved at 6 LVs and increasing towards both a higher and lower number of LVs.

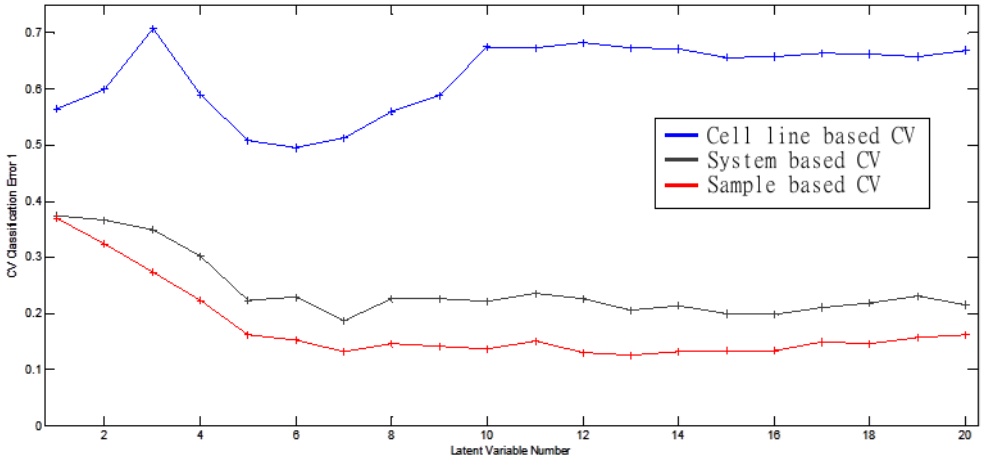


Figure 101: Cross-validation error for different cross-validation methods.
Cross-validation error was calculated for cross-validation using data block of all cell lines, individual bioreactor systems and individual samples.

In order to better understand the optimal number of LVs to be included in the model, contribution plots were inspected. Figure 102 provides an overview of contribution plots based on loadings values of individual LVs and averaged spectra from each cell line used in the model. There is a high number of peaks which are significant in the different LVs with some of them being more common than others. The peak in the 4700-4900 region achieves high positive or negative intensity in 10/11 LVs, the peak in 2100-2300 achieves a high or medium intensity in 9/11 LVs and there are several other regions which are shared by multiple LVs. At the same time there are regions of low to medium intensity which are only present in small number of LVs. This diversity demonstrates the complexity of factors contributing to the variation in the mass spectra between different samples while there are also regions which dominate the contribution plots.

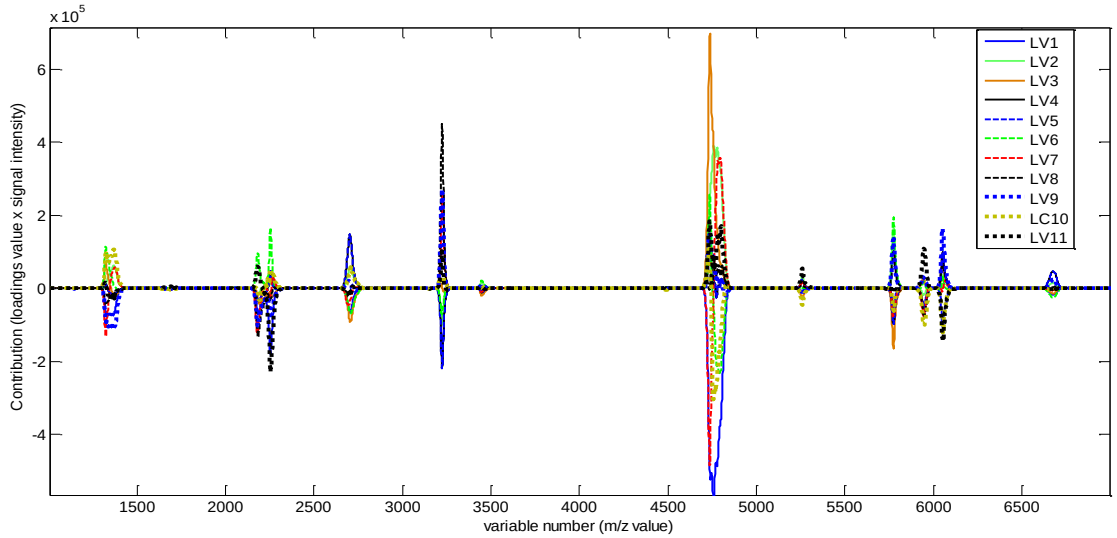


Figure 102: Overview of PLS-DA model contribution plots (11 LVs).
Each line represents a contribution plot of one of the 11 LVs based on LV's loadings value and signal intensity.

Another interesting observation from the contribution plots, specifically some of the higher numbered ones (LV 6,7,10 and 11), is a trend for some of the peaks to have a sharp change from positive to negative contribution or vice versa in one or more regions, e.g. the region 4700-4900 in LVs 6,7 and 10 (Figure 103) or the region 2100-2300 in LVs 10 and 11 (Figure 104).This sharp change reflects a difference in the shape of a peak which is specific to one population of samples but not present in others. This highlights that while the higher number LVs capture less variability per LV compared to the earlier LVs, they can be used to explain minor differences between samples which can be significant when aiming to assess the productivity of the cell lines based on differences in their mass spectra.

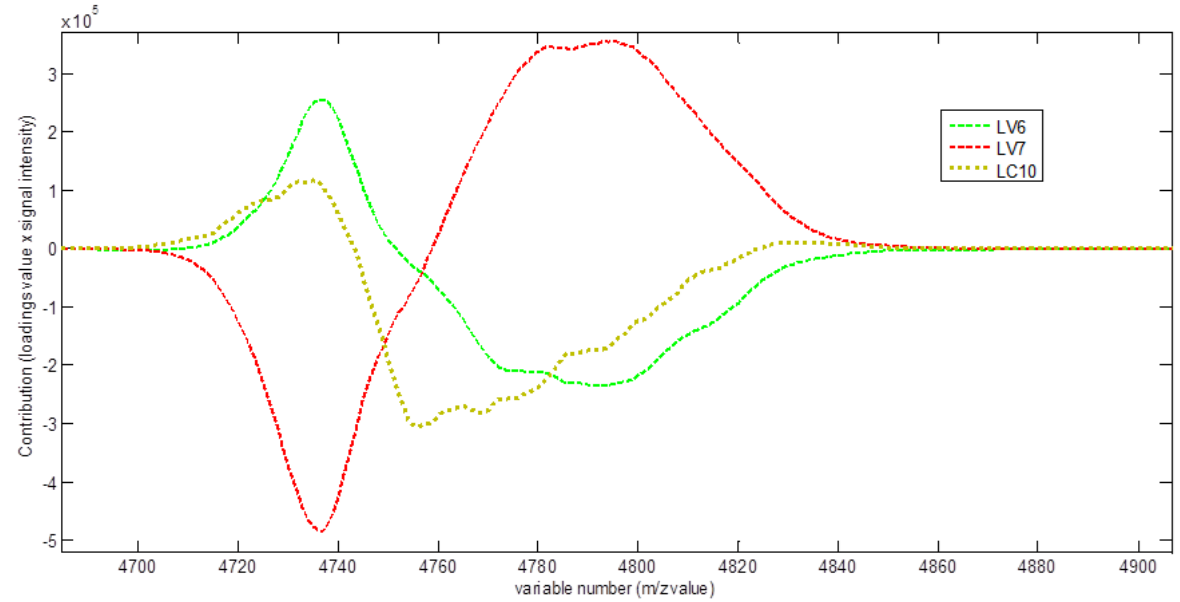


Figure 103: Contribution plot for LVs 6,7 and 10 in 4700-4900 region.

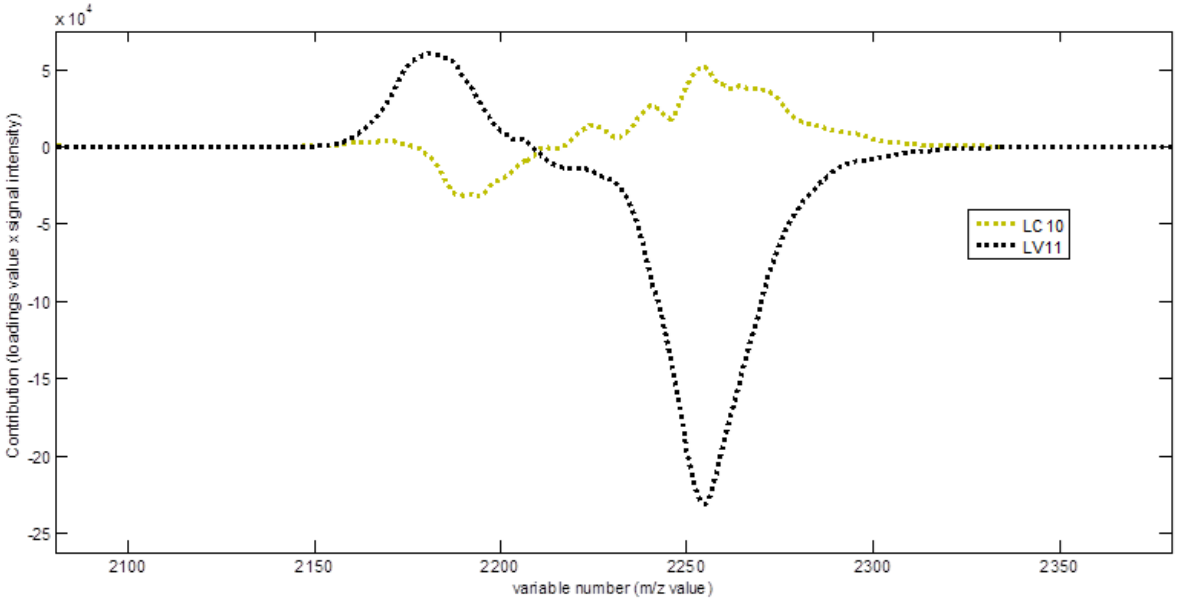


Figure 104: Contribution plot for LVs 10 and 11 in 2100-2300 region.

While minimising the cross-validation error is an important step in designing the model, a major objective was to capture a representative amount of variance in the Y block. In order to design a model that explains the high amount of sample variation the aim was to capture at least 80% of the variation of Y block. This was achieved by using 11 LVs for model calibration which captures 98.52% variation of the X block and 83.46% variation of the Y block, fulfilling the initial requirements (Figure 105).

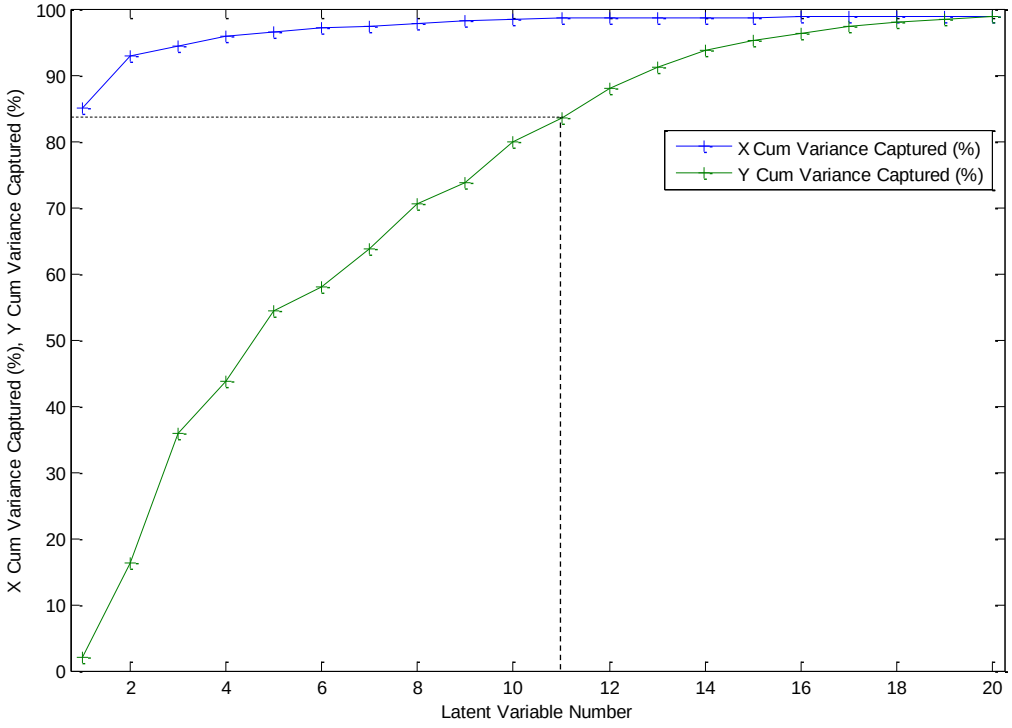


Figure 105: Plot of cumulative variance captured per number of LVs used in the PLS-DA model. Variance captured is provided for X and Y block.

The results of model cross-validation were assessed by looking at the classification probabilities for low and high producer classes and by counting the misclassification events. Using a model with 11 LVs, most samples were classified correctly during cross-validation. The predicted Y values were found above the discrimination threshold for the classification of the corresponding classes for both class 1 (low producer) and class 2 (high producer) samples (Figure 106). For class 1, the majority of samples from the low producer cell lines (below $2x10^7$ TU/ml threshold) achieved Y1 scores above the low producer classification threshold. For the remaining 3 high producer cell lines the results were reversed, with Y2 values predicted above high producer classification threshold. Overall, the predicted Y1 and Y2 values were almost opposite to each other, indicating a high model confidence in classification of samples from the two different classes. Only a few individual misclassification events were observed, resulting in over 95% prediction accuracy. Most samples were predicted to belong to their corresponding class with a high probability (Figure 107).

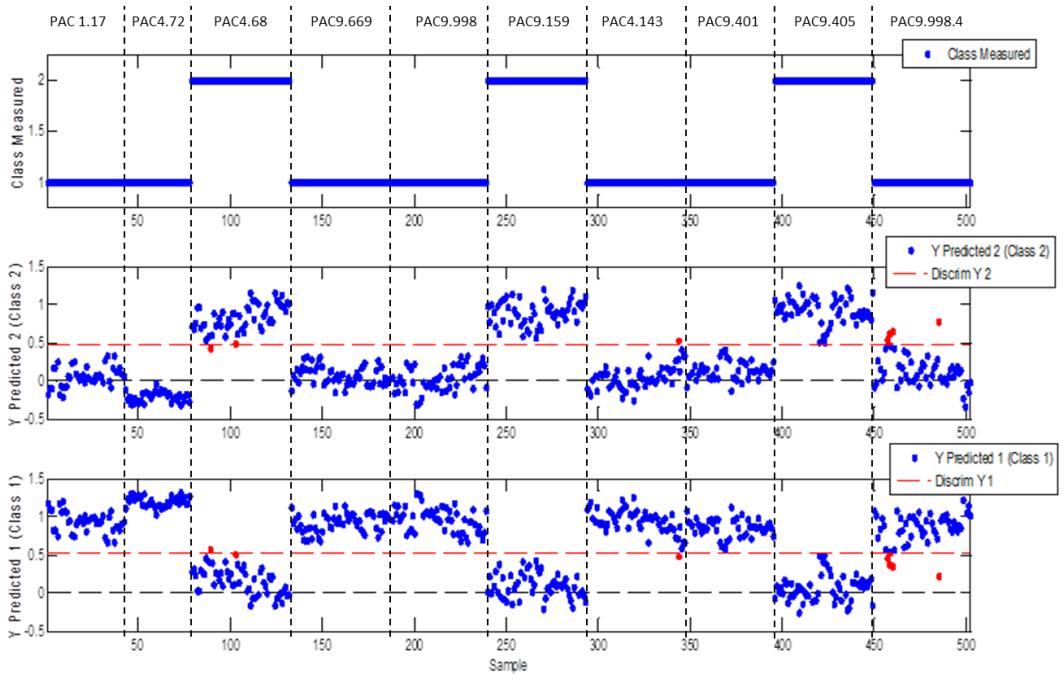


Figure 106: Cross-validation results of the model based on 10 cell lines, using 11 LVs. For each cell line, the measured class along with classification threshold, predicted class 1 and predicted class 2 scores are presented. Class 1 – low producer (below $2x10^7$ TU/ml); class 2 – high producer (above $2x10^7$ TU/ml)

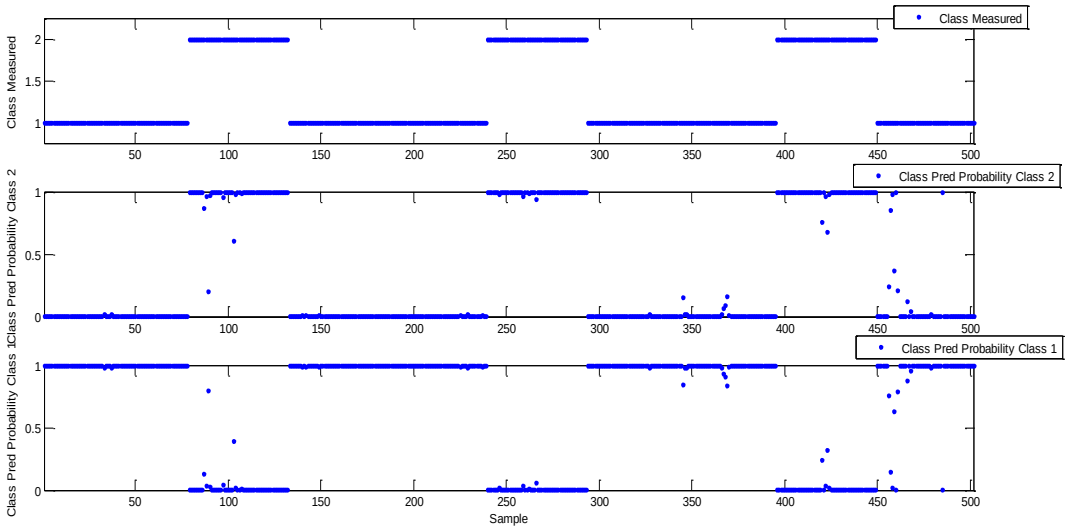


Figure 107: Probability of correct class prediction for the 10 cell line model.
Class 1 – low producer (below $2x10^7$ TU/ml); class 2 – high producer (above $2x10^7$ TU/ml)

A different version of the model was tested, using a lower number of LVs. A model using 6 LVs captured 97.1% of X block variance and 57.9% of Y block variance. This change resulted in a higher degree of misclassification leading to accuracy of prediction below 90% (Figure 108). This version of the model achieves lower accuracy of its predictions in the cross-validation and it does not meet the goal of over 80% of Y block variance captured. Therefore the 11 LVs model was used in the further studies and as a base and point of reference for testing of other variants of the model.

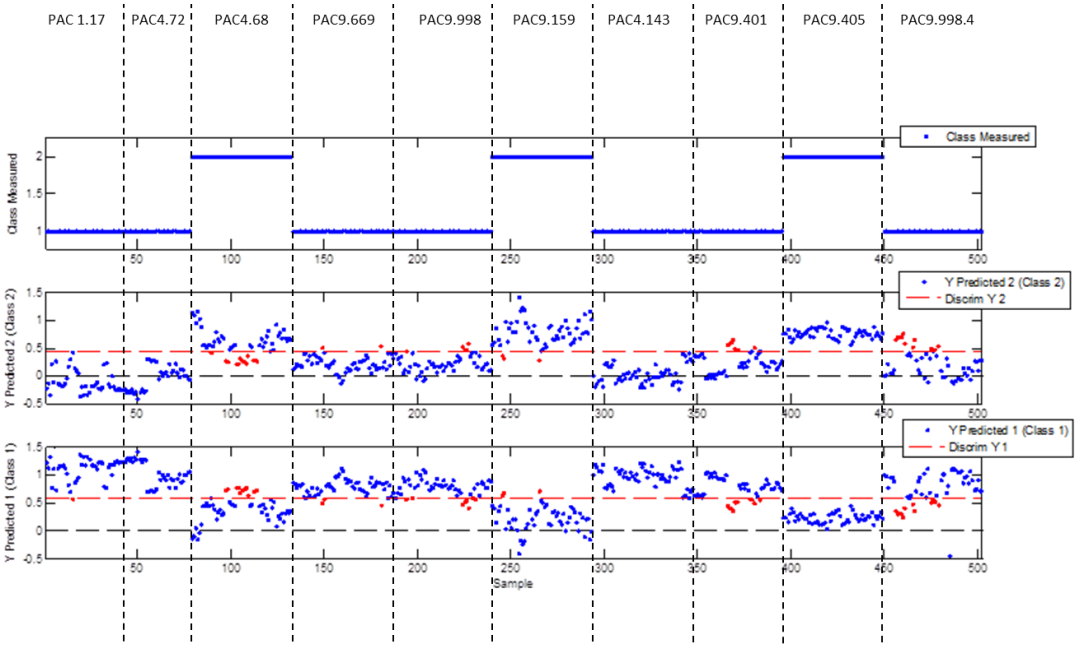


Figure 108: Cross-validation results of the model based on 10 cell lines, using 6 LVs.
For each cell line the measured class along with classification threshold, predicted class 1 and predicted class 2 scores are presented. Class 1 – low producer (below $2x10^7$ TU/ml); class 2 – high producer (above $2x10^7$ TU/ml)

After model calibration and cross-validation using PLS-DA, alternative approaches to the modelling were considered to explore different modelling methods and the initial assumptions about the data being best suited for categorical classification. Using the data from 18 cell lines, PLS-DA was compared to PLS, a method capable of predicting continuous data rather than categorical data used in PLS-DA. The averaged viral vector titre of each cell line was used as the Y block data while the same mass spectra used in PLS-DA model were used as the X block data. The cross-validation was performed using the same method as before where samples from each cell lines were used as a single block to iteratively validate the model calibrated using the remaining samples. The resulting PLS model had a similar structure to the PLS-DA model where to capture over 80% of Y block variability at least 11 LVs had to be included in the model. The final version of the model used 11 LVs, capturing 98.49% of X block data variability and 84.48% of Y block data variability. The measured and predicted titre values formed a close fit with the R^2 value of 0.84. However, due to variation in the mass spectra and the titre measurements, there was a significant degree of prediction variation between individual samples (Figure 109). Moreover, because there were no calibration samples with measured viral vector titre between 1-2x10^7 TU/ml, the prediction of a continuous variable is more difficult compared to discrimination between low and high producers. Overall, the available data was deemed to be better suited for PLS-DA rather than PLS and further model validation and optimisation was performed only for that method.

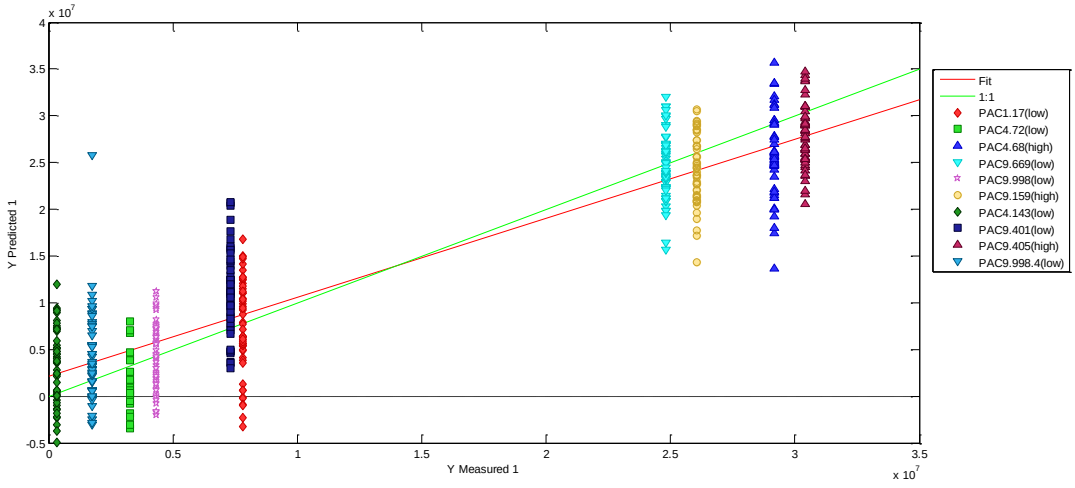


Figure 109: Comparison of measured and PLS-DA model-predicted infectious titre (FACS) values for 10 cell lines.

The PLS-DA model was further tested by varying certain parameters. Several combinations of different classification thresholds (based on different viral vector titre values) and different number of total LVs used in the model were inspected. The threshold was set as 2x10^7 TU/ml, 1x10^7 TU/ml or 5x10^6 TU/ml. The calibration and validation data

sets classes were adjusted according to the different thresholds. Each version of the model was then assessed by looking at variance captured and cross-validation error value (Figure 110). Based on these values several variants of the model were developed to test classification accuracy as a part of the model validation explained in detail in section 4.3.3.

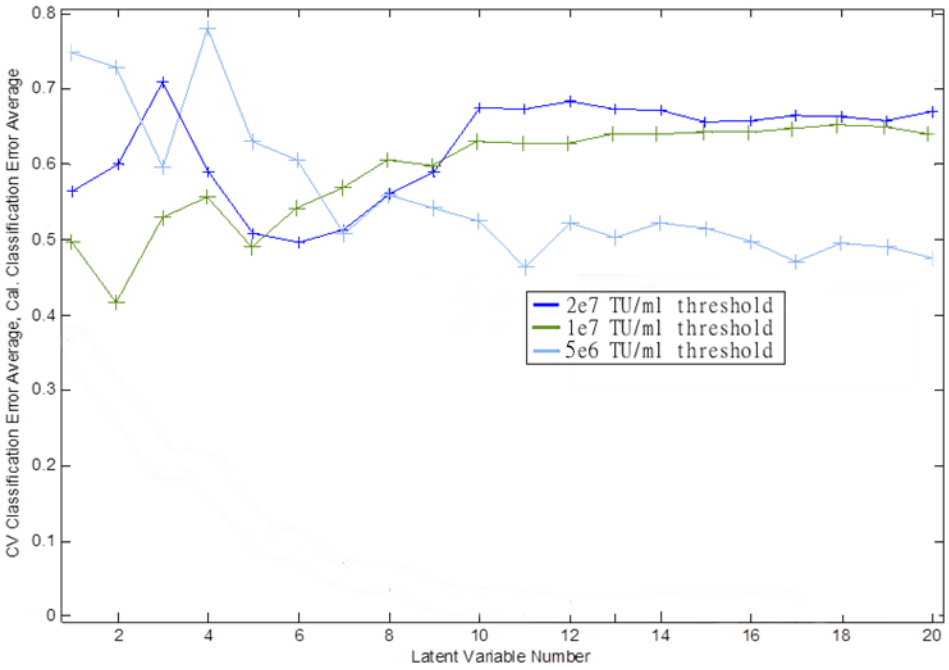


Figure 110: Cross-validation error for PLS-DA models with different classification threshold.

4.3.3. Partial least squares discriminant analysis model validation

The PLS-DA model was validated using a set of cell lines developed independently from the calibration cell lines but using the same methods. The mass spectra of new cell lines were assessed visually and using PCA (Figure 111). By looking at the PCA scores plot it is evident that there are sample populations which are clustered away from the remaining data points, namely PAC9.400 which is a high producer and is characterised by a high PC2 score compared to low and negative values of the remaining cell samples. Regardless of the position on the PC scores plots, the samples from individual cell lines are clustered close to each other. This indicates a good reproducibility of mass spectra between the samples and therefore all 144 data points were used in model validation.

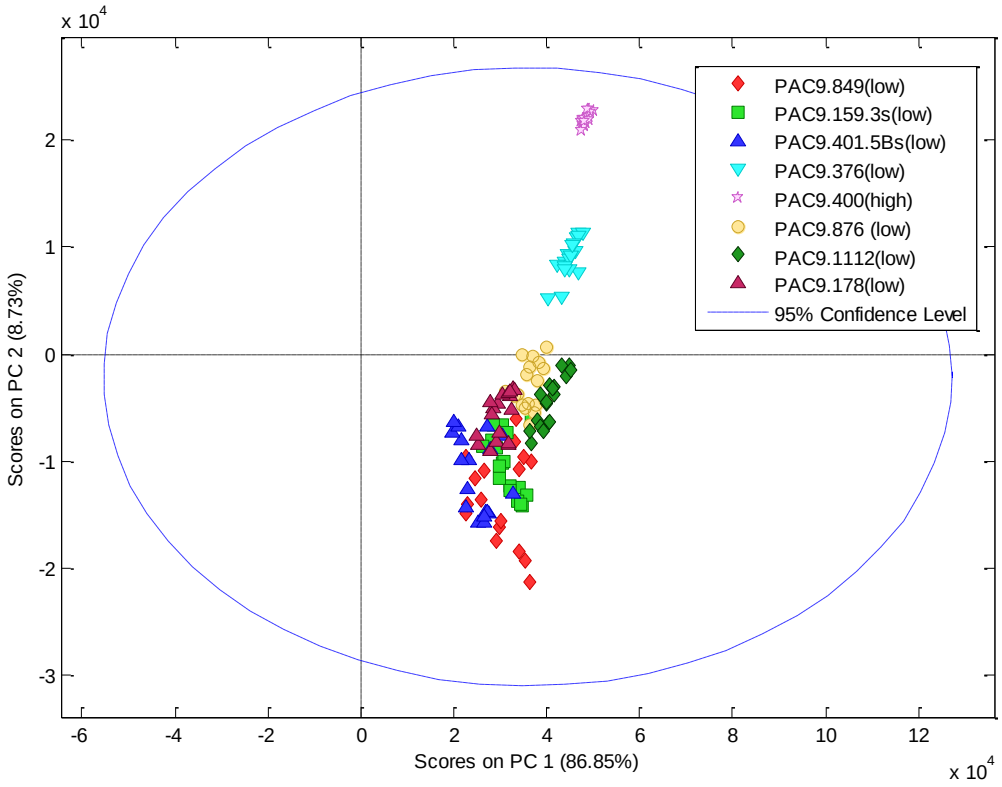


Figure 111: PCA scores plot (PC1 and PC2) for mass spectra from 8 validation cell line samples.

The validation data set was first applied to the default model generated using 502 data points from 10 cell lines and including 11 LVs. 126 out of 144 data points were classified correctly, resulting in 87.5% accuracy of the prediction by the PLS-DA model (Figure 112). The misclassification events were observed for individual samples in several cell lines. All 18 samples from PAC9.400 (high producer) were classified correctly with the predicted Y2 value significantly above the discrimination threshold. In order to further improve model performance and to compensate for the variation between mass spectra samples, the same validation method was applied to averaged mass spectra, using a single mean for each cell line (Figure 113). This approach resulted in 100% correct classification of each cell line.

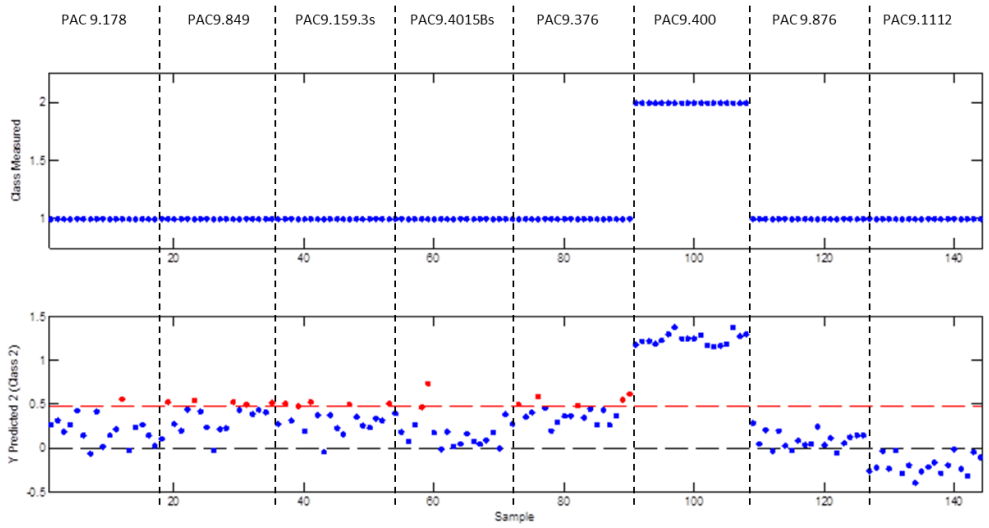


Figure 112: Model validation results for 11 LVs model with 2e7 TU/ml classification threshold.
For each cell lines measured class and predicted class 2 scores are displayed. Class 2 – high producer (above $2x10^7$ TU/ml)

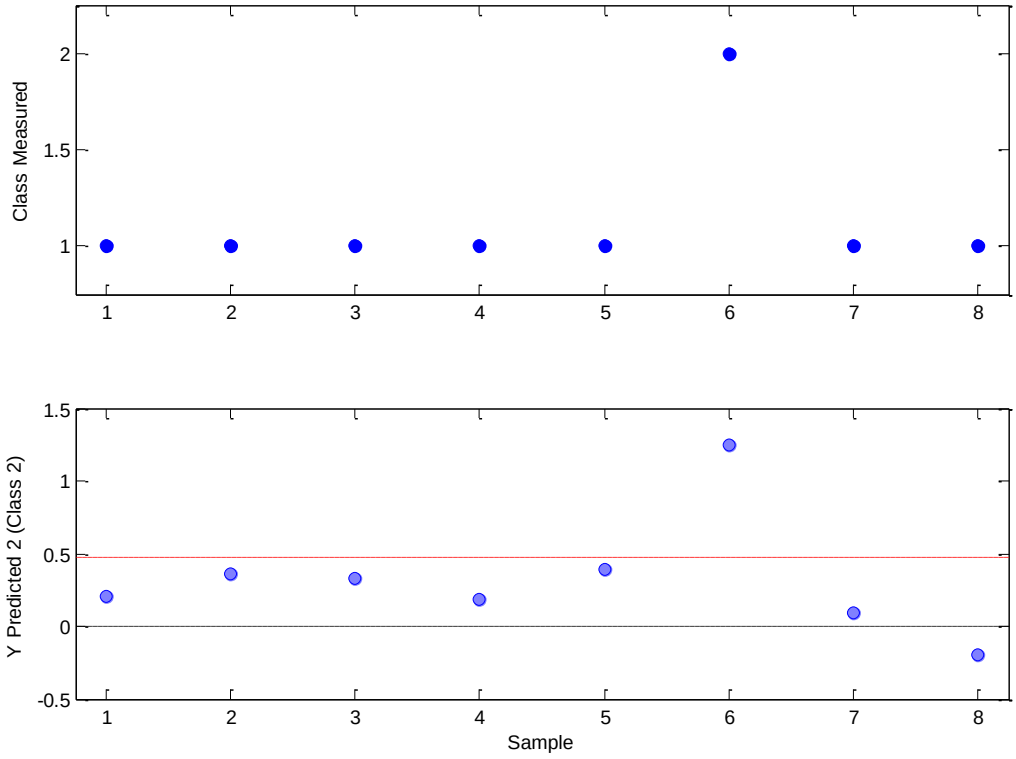


Figure 113: Model validation results for 11 LVs model with $2x10^7$ TU/ml classification threshold using averaged mass spectra.
For each cell lines measured class and predicted class 2 scores are displayed. Class 2 – high producer (above $2x10^7$ TU/ml)

Several variants of the model were developed and validated to test the robustness of the PLS-DA method for cell line development (Table 9). An approach using different viral vector titre thresholds for model calibration and validation was used along with testing of

lower and higher numbers of LVs included in the model to examine the effect on validation data classification.

Model	LVs used	Viral vector titre threshold	X block variability captured	Y block variability captured	Prediction accuracy	Figure reference
1	11	$2x10^7$ TU/ml	98.52%	83.46%	87.5%	Figure 112
2	8	$2x10^7$ TU/ml	97.72%	70.6%	86%	Figure 114
3	6	$2x10^7$ TU/ml	97.1%	57.9%	83%	Figure 115
4	10	$1x10^7$ TU/ml	98.33%	80.17%	65%	Figure 118
5	5	$1x10^7$ TU/ml	96.45%	53.62%	53%	Figure 119
6	10	$5x10^6$ TU/ml	97.32%	82.21%	37.5%	Figure 122

Table 9: Summary of PLS-DA model variants used in model validation.

First of all, the model using $2x10^7$ TU/ml threshold and a lower number of LVs was inspected. One version used 8 LVs to reach at least 70% of Y block variation captured (Figure 105) while the second variant used 6 LVs to achieve the lowest cross-validation error value (Figure 110). The first model captured 97.72% of X block variability and 70.6% of Y block variability and resulted in correct classification of the high producer cell lines but also an increased amount of misclassification events, especially for cell line PAC9.159.3s (Figure 114). Overall prediction accuracy reached 86% accuracy which is a result close to the original model. Most of the misclassification events occurred for PAC9.159.3s which had the 2nd highest titre among the validation cell lines. As such this version of the model performance is close to the original and could be considered an improvement when looking for medium producers.

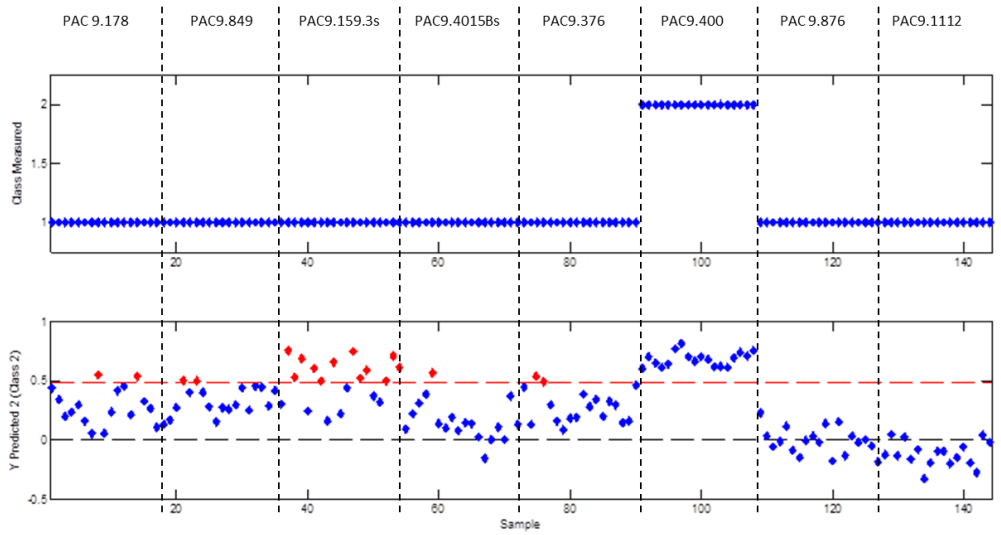


Figure 114: Model validation results for 8 LVs model with $2x10^7$ TU/ml classification threshold. For each cell lines measured class and predicted class 2 scores are displayed. Class 2 – high producer (above $2x10^7$ TU/ml)

The second variant of the model included 6 LVs, capturing 97.1% of X block variability and 57.9% of Y block variability with the cross-validation error value of 0.496. In this case the prediction accuracy was lower, reaching 83%. However, the misclassification events were concentrated in PAC9.376 where all samples from this cell line were falsely identified as high producers which would result in a complete misclassification of an unknown sample when using this version of the model (Figure 115).

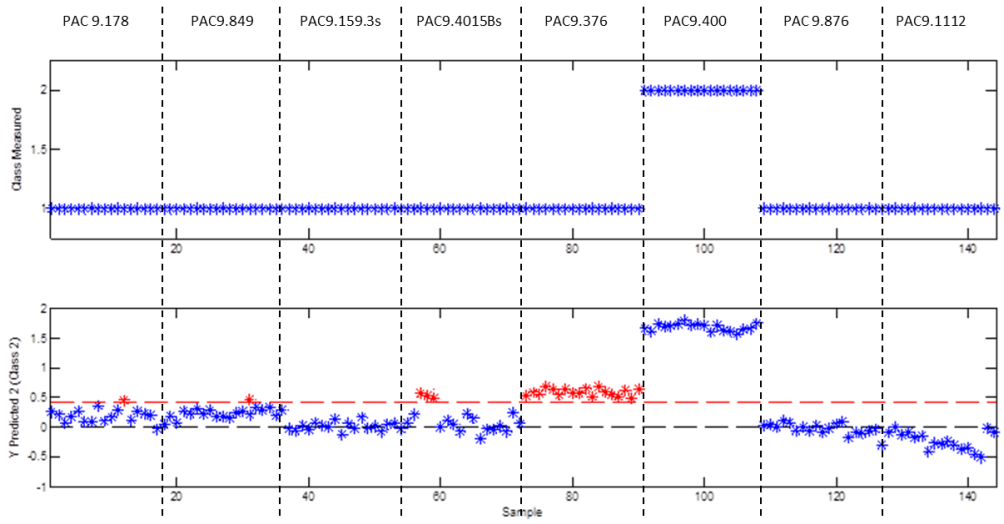


Figure 115: Model validation results for 6 LVs model with $2x10^7$ TU/ml classification threshold. For each cell lines measured class and predicted class 2 scores are displayed. Class 2 – high producer (above $2x10^7$ TU/ml)

Another version of the model was designed using a lower discrimination threshold, set at $1x10^7$ TU/ml. This change resulted in a calibration data set with 4 cell lines defined as high producers and 6 low producers. In the validation data set, there were 2 cell lines with titre consistently above $1x10^7$ TU/ml and 6 below this threshold. Two variants of the model were examined, the first including a higher number of LVs to reach 80% captured Y block variability (Figure 116) and the second with the aim to minimise cross-validation error (Figure 117).

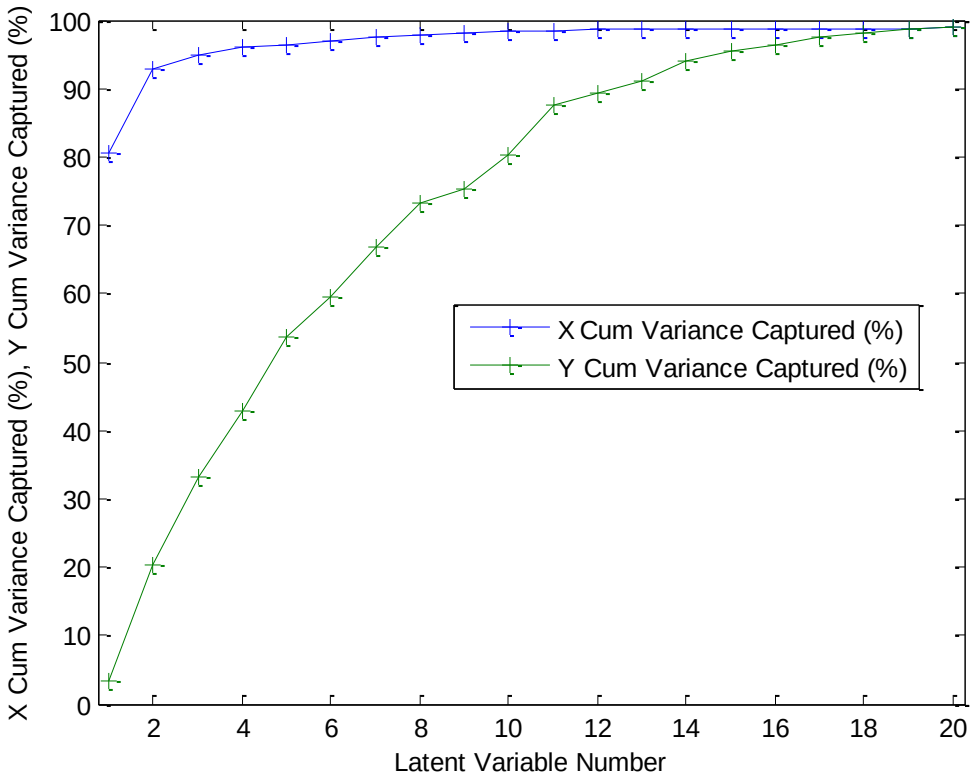


Figure 116: Cumulative variance captured per number of LVs used for PLS-DA model using 1x10^7 TU/ml classification threshold.
Variance captured is provided for X and Y block.

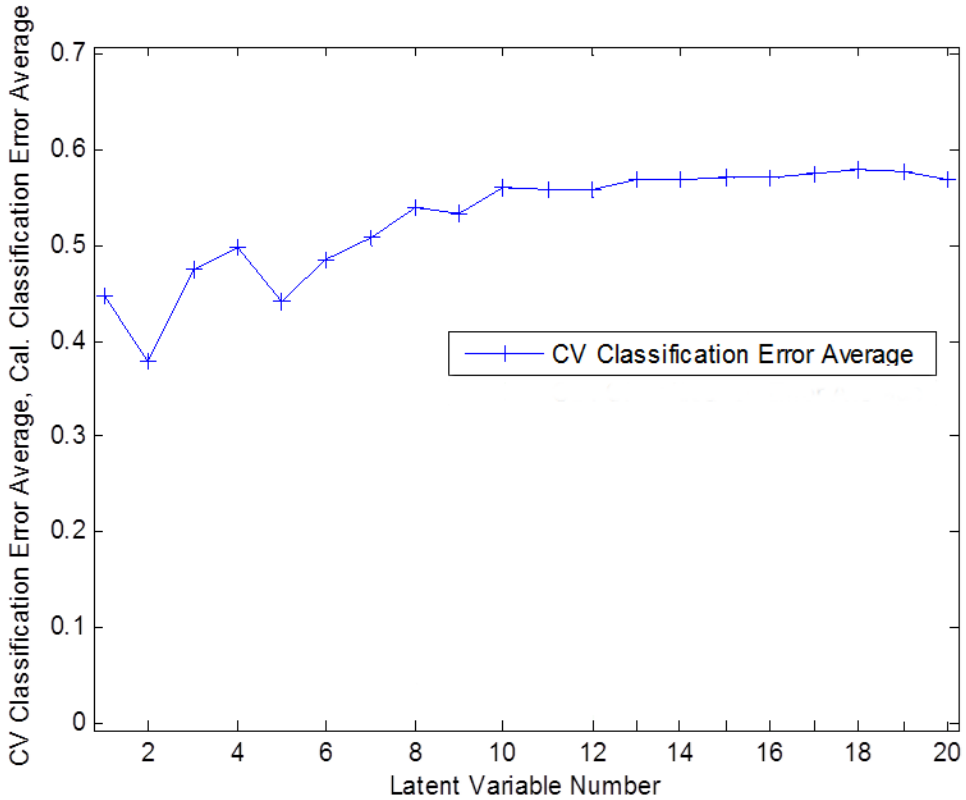


Figure 117: Cross-validation error per number of LVs used for PLS-DA model using 1x10^7 TU/ml classification threshold.

The first model included 10 LVs, captured 98.33% of X block variability and 80.17% of Y block variability with a cross-validation error of 0.55. While the samples from PAC9.400 (a high producer) were classified correctly, there was a high degree of misclassification, where majority of samples from PAC9.159.3s (the second highest producer) were classified as low producers while PAC9.376 was falsely predicted to be a high producer (Figure 118). Overall, the prediction accuracy was 65% with 2 misclassified cell lines, resulting in poor model performance.

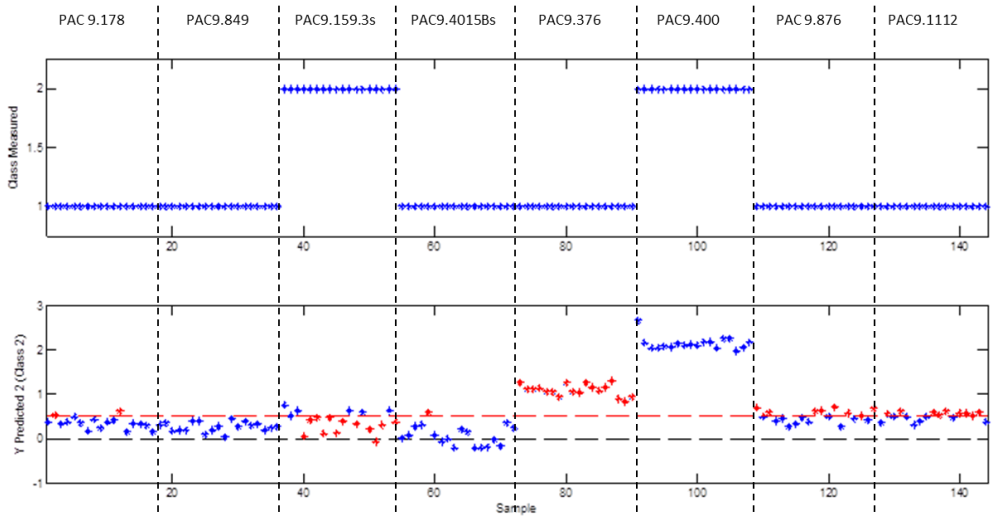


Figure 118: Model validation results for 10 LVs model with 1x10^7 TU/ml classification threshold. For each cell lines measured class and predicted class 2 scores are displayed. Class 2 – high producer (above 1x10^7 TU/ml)

The second variant of the model used 5 LVs, captured 96.45% of X block variability, 53.62% of Y block variability and achieved cross-validation error of 0.44. Similar to the previous iteration of the model, PAC9.159.s and PAC9.376 samples were misclassified while PAC9.400 was correctly predicted as a high producer. However, a significant amount of samples from PAC9.178 and PAC9.876 were misclassified as well (Figure 119). The overall accuracy of the model predictions was 53% with multiple misclassifications and decrease in model performance compared to the previous version.

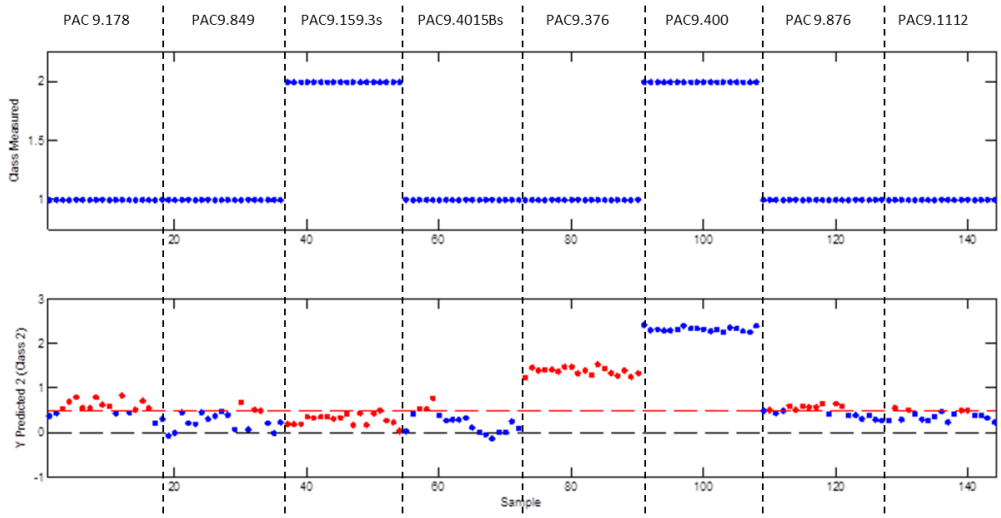


Figure 119: Model validation results for 5 LVs model with 1x10^7 TU/ml classification threshold.
For each cell lines measured class and predicted class 2 scores are displayed. Class 2 – high producer (above 1x10^7 TU/ml)

The final model variant was designed using a classification threshold of 5x10^6 TU/ml which would be considered a lower limit of acceptable infectious titre reach bye cell line during cell line development. With this threshold 5 of the 10 cells used for calibration were classified as a high producer while the remaining 5 were low. In the validation data set 3 of the 8 cell lines consistently produced over 5x10^6 TU/ml. A model using 10 LVs was inspected. It captured 97.32% of X block variability, 82.21% of Y block variability (Figure 120) and achieved cross-validation error of 0.53 (Figure 121).

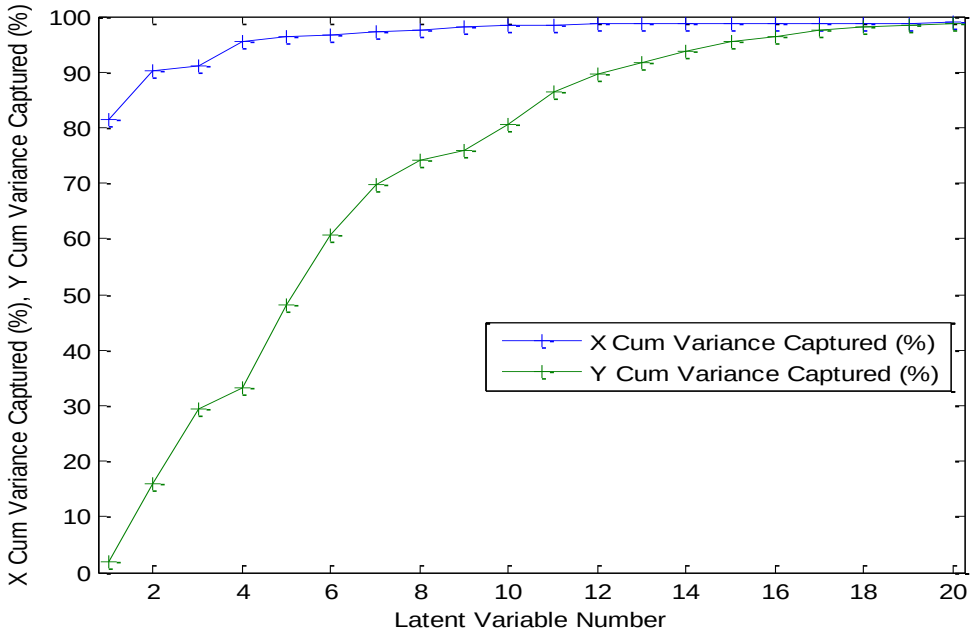


Figure 120: Cumulative variance captured per number of LVs used for PLS-DA model using 5x10^6 TU/ml classification threshold.
Variance captured is provided for X and Y block.

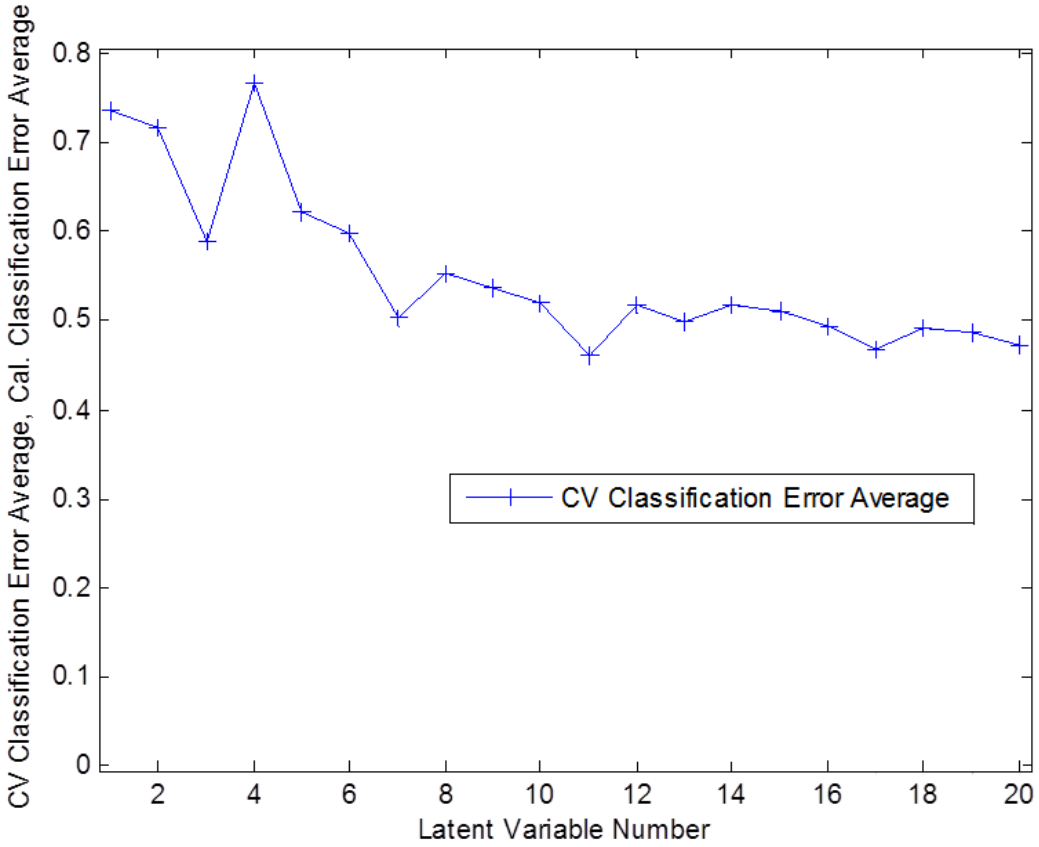


Figure 121: Cross-validation error per number of LVs used for PLS-DA model using $5x10^6$ TU/ml classification threshold.

The low threshold model correctly predicted 2 of the cell lines classified as high (PAC9.159.3s and Pac9.401.5Bs); however, there were multiple misclassifications for all the other cell lines. Both the remaining high producer (PAC9.400) and all the low producers were misclassified, resulting in 37.5% accuracy of the model which makes this version of the model highly unreliable compared to the other ones (Figure 122).

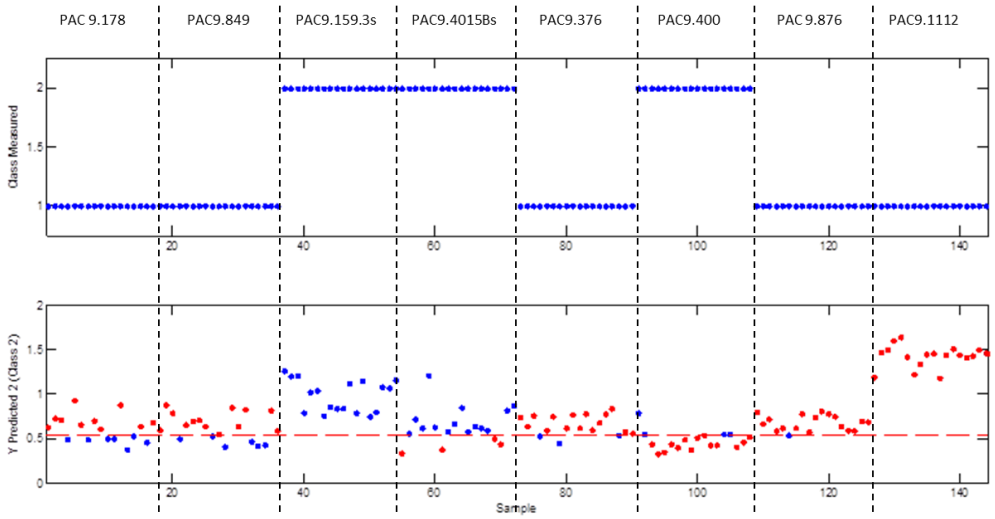


Figure 122: Model validation results for 10 LVs model with $5x10^6$ TU/ml classification threshold. For each cell lines measured class and predicted class 2 scores are displayed. Class 2 – high producer (above $5x10^6$ TU/ml)

4.4. Discussion

The PLS-DA model was applied in cell line development for HEK293T PaCLs stably transfected with viral vector genes. They were used in production of HIV-GFP which was used as a model system to assess LVV titre reached by cell lines as a basis for selecting clones as either high or low producers. These data along with mass spectra of the cells was used in design of multiple PLS and PLS-DA predictive models to determine the suitability of this approach for cell line development at OXB. These results as well as the further refinement of the method are discussed below.

4.4.1. Cell line development

18 PaCLs were developed within the Cell Engineering Group at OXB and selected for the use in the PLS-DA model development. The initial 10 cell lines used for model calibration were selected based on their productivity in small-scale adherent culture, aiming to obtain multiple high and low producers. Whilst performance at this scale can be used as a guideline, it does not always directly correlate with large scale productivity (Porter *et al.*, 2010). The initial selection bias and lack of the initial cell characterisation caused the cells of medium ($1x10^7$-$2x10^7$ TU/ml) productivity to be under-represented in the calibration data set (Figure 86). This cell performance distribution profile had a significant impact on model performance and will be discussed as part of the model validation assessment.

A different approach was taken for the cells used in model validation where 12 random cell lines were selected from a set of packaging cell lines banked at the late stage of selection in adherent culture. These cells were adapted to suspension which resulted in 8 out of 12 cell lines achieving a high (above 80%) viability. Following cell expansion and characterisation it was discovered that only a single cell line produced over $2x10^7$ TU/ml of the viral vector. However, it is an expected outcome as high producer cells are rare and in a random selection pool it is more likely to find low producers. As such the validation data set can be considered representative of an average cell line development result, with multiple low producers and a small subset of high producers. Medium producers ($1\text{-}2x10^7$ TU/ml) were also found among the developed cells.

Cell line adaptation from adherent to suspension culture is a gradual process. In the case of the validation cell lines the cells maintained high viability throughout the process with small exception where the viability decreased but it was still maintained at a satisfactory level above 85%. Of the PaCLs investigated in this study, the viability of four clones decreased

after the suspension adaptation process during seed culture scale up. Decreased viability of some of the cell cultures is an expected side-effect of suspension adaptation process due to variation in cell properties (Tsao et al., 2001). The eight remaining high viability cell lines were used for viral vector production in the 15 mL microscale bioreactor ambr15® and larger 0.5L MiniBio reactors and used as a basis for validation of the model.

The infectious titre reached by cells was assessed using a FACS transduction assay following in-house method developed and validated by OXB. The assay is affected by the cell culture used for transduction, the transduction process itself as well as varying levels of GFP expression and detection and the results can be variable. FACS assay was performed multiple times and always in triplicate to compensate for assay variability and to ensure a high number of repeats which were then used to average the results and estimate the standard deviation of the results. While individual assays CV ranged from 10-50%, by performing multiple assays this value was reduced to more reliable value of 10-20%. FACS infectious titre is routinely used at OXB and with a high number of repeats it is representative of the results used to assess the cell line development process.

Differences in viral vector titre across scales have been observed using the ambr15® and 0,5L MiniBio (Figure 97) confirming that whilst the microscale mimic is sufficient to assess cell line potential, there are examples where it does not accurately represent the performance at the commercial scale. In the most extreme case of clone PAC9.400 there is more than 10-fold difference between small- and large-scale bioreactors, highlighting a need for better characterisation tool in order to identify all potential high producers. During a traditional cell line development cycle a cell line with low productivity in adherent or small scale suspension culture could be discarded while it was demonstrated that these cells may still have a potential for high productivity at larger scale due to differences in the process conditions.

Throughout the process of cell line development, suspension adaptation, vector production and performance characterisation there were several events which introduce variability and outlier results into the process (e.g. due to bioreactor or mass spectrometer technical or programming fault). To ensure data quality and consistency all experiments followed the same production and analysis protocol, where records were reviewed to identify potential technical and operator errors that could affect the process. It is important to evaluate whether variation in the data was caused by unexpected events or inherent biological

variability of the different cell lines. Reviewing the process records can help to justify whether problematic data points should be excluded from the model or incorporated into it. Having a diverse set of samples for model calibration increases its robustness and helps to compensate for unexpected results once it is applied to real data. However, a high degree of unexplained variation in the model can decrease its performance. As such, several mass spectra from PAC4.72 were omitted due to low intensity. In a real application of the model, low intensity mass spectra would not be used as they are a result of samples preparation errors rather than poor performance of the cell lines themselves. These kind of data points would be eliminated during a pre-screening step as was the case for the calibration data. However, to ensure accurate classification of cells, both MS and FACS readings were repeated multiple times to compensate for assay variability and provide reliable and consistent data.

4.4.2. Partial least squares discriminant model performance

MS data analysis shows that there are multiple sources of signal variation. As demonstrated by low signal intensity in some of the samples from PAC4.72, sample preparation and the process of obtaining the spectra can cause a large variation to the point where spectra are not usable for analysis or modelling. Therefore, quality control of the raw spectra should be performed shortly after analysis to ensure that the mass spectra acquisition has been successful. A large number of repeats per sample help to compensate for any potential problems as well. As observed during PCA there is also a degree of variation between samples obtained from different bioreactors (e.g. for PAC9.401). At the same time there are cell lines where separate biological repeats result in consistent and similar mass spectra. This highlights the need for process monitoring and control to ensure consistency between vessels but also demonstrates that in a well-controlled system a high degree of consistency can be achieved. While sample variation can be disruptive to analysis and classification of unknown samples, including minor variation in the model calibration data helps to compensate for imperfection of data that will be applied to the model. Exclusion of all variable samples would greatly reduce the number of data points available for model training and would likely result in over-representation of certain type of cells. Including some of the less consistent data points can help improve model robustness as long as any major outliers such as low intensity spectra are accounted for.

In order to assess PLS-DA model performance and compare different model variants examined in the experiment, it is important to establish what criteria are used to determine

model's success. Ultimately, the method is needed for prediction of cell line performance but to achieve it in a reliable and reproducible way, the model needs to capture the variance within the data. In PLS and PLS-DA the data is projected to LVs so that the measured data (mass spectra) explain the classification data (infectious titre reached by cell lines). This is achieved by including a number of LVs sufficient to capture enough of variation from both data blocks to reliably model the relationship between them. Both X and Y block variance captured should exceed a number (80% in this experiment) based on the expected model performance and relationship between the two data sets. However, using too many LVs and capturing too much variability could result in overfitting the model to the calibration data which would negatively affect its ability to work with unknown data. As a point of balance, the model was expected to capture minimum 80% of X and Y block variance. This way the major sources of variation in the MS data should be captured and explain the variation in class data, resulting in a model capable of accurate prediction without overfitting the data. This required inclusion of 11 LVs in the model and was justified by the positive results of model validation (high prediction accuracy). Based on validation performance, the alternative models using a lower number of LVs (and therefore capturing less of data variance) or lower viral vector discrimination thresholds led to decrease in prediction performance.

Another argument in support of inclusion of a higher number of LVs is based on the contribution plots. As observed for LVs 6-11, there are regions of the spectra where contribution plot values drastically shift from positive to negative values or vice versa (Figure 103-Figure 104). This indicated a significant difference in the shape of an individual peak between two or more populations of cells and can be used as a basis for differentiation of samples. While these LVs capture proportionally smaller amount of variability, these unique interactions can be useful to distinguish between similar samples. In case of the cell line analysis, all samples share the same basic protein profile needed to support vector production. Capturing small differences between cell populations could be essential for differentiation between high and low producers. The trend of a sharp change from positive to negative loading is similar to the one observed for PCA of adherent and suspension cells. The adherent and suspension samples substantially differ in the shape of several peaks (Figure 123) which leads to a sharp change in loadings values (Figure 124). In fact, the major peak displaying this behaviour is the same for both PLS model and PCA in the region of 4600-4800 m/z. This indicates that the characteristics which differentiate between adherent and suspension cells may be influencing cell productivity in bioreactors. Given that the cell line development

includes adaptation of cells from adherent culture to suspension, it is possible that cells which display characteristics closer to suspension adapted cells are capable of better viral vector production in bioreactors. A more detailed study of this phenomenon could greatly improve the understanding of cell line development process as well as cellular factors which contribute to viral vector productivity in suspension culture.

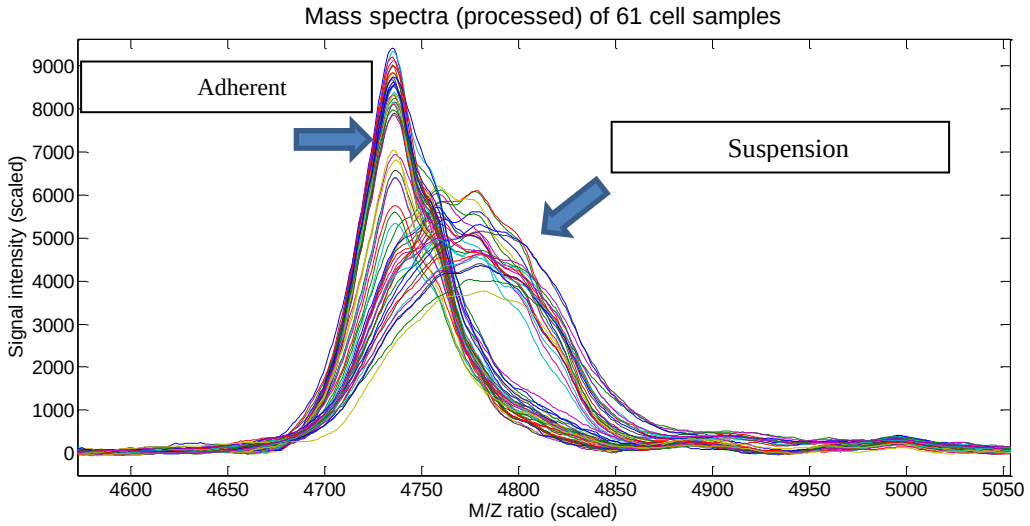


Figure 123: Highlight of the difference in mass spectra of adherent and suspension cell samples in the 4600-4800 region.
Mass spectra obtained from 61 cell samples from suspension and adherent culture described in detail in Chapter 3. The 4600-4800 region highlights the difference between suspension and adherent cell mass spectra.

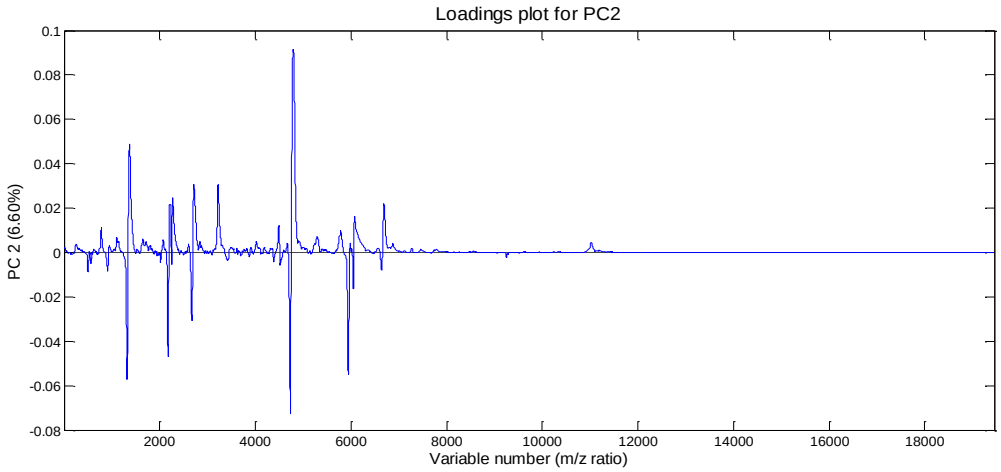


Figure 124: Highlight of PCA loadings plot (PC2) for mass spectra of adherent and suspension cell samples.
Loadings plot from PCA of mass spectra of 61 cell samples from suspension and adherent culture described in detail in Chapter 3. Sharp shift from negative to positive values indicates a difference between suspension and adherent cells.

Cross-validation of the model is an important step in the initial development as well as a method to assess its performance and guide selection of LVs. Due to exclusion of some of the mass spectra, individual cell lines were characterised by varying amount of data points. At the same time these data points shared a large degree of similarity between cell lines, bioreactor runs of the same cell lines and especially repeat samples from the same bioreactor.

These similarities mean that a default cross-validation based on random selection of samples would result in over-estimation of model performance because it would be using samples with direct relationship with those remaining in calibration data set. A custom cross-validation based on block selection was applied instead to compensate for sample distribution (Figure 101). While using biological repeats and bioreactor samples as a base for block separation, the cross-validation error is significantly lower compared to the method using cell lines as blocks. This is caused by the abovementioned overfitting where blocks of samples are taken out and the model is validated using samples similar to the ones included in the model. This inevitably leads to a perceived improved model performance, but it is caused by bias rather than actual improvement in performance. Therefore, using all samples from individual cell lines as blocks for cross-validation is a more realistic method of assessing the model quality. However, because of a large number of samples used in each block, the model performance can be underestimated during cross-validation because when an entire cell line is taken out from calibration data for purposes of validation, the available data is significantly decreased.

Model performance relies on the quality of data used for calibration and validation. The cells used in the current experiment were mostly low producers with a smaller number of high producers. The cells available for model calibration provided a solid basis for the model. However, among the 10 cell lines used in model calibration, there were no cells which could be described as medium producers (Figure 86), i.e. below the discrimination threshold of $2x10^7$ TU/ml but above $1x10^7$ TU/ml which still indicates good productivity. This underrepresentation of a population of cells is reflected in modelling results. By changing the discrimination thresholds, classification accuracy is decreased. The accuracy is assessed based on the percentage of correctly predicted samples in the validation data set. The calibration data lacks reference data for the range of medium producers and therefore decreasing the threshold leads to medium producers ($1\text{-}1.5x10^7$ TU/ml) from validation to be misclassified. This indicates that the model is prone to underperformance in prediction of infectious titre in the range close to the threshold because of the gap in calibration data set titre distribution. For this reason, the discrimination threshold was kept at $2x10^7$ TU/ml to improve prediction accuracy for the highest producers. While this approach can reduce detectability of medium producers which have potential to produce commercially relevant amount of viral vectors, it reduces chances of discarding a high producer. For one of the model variant using 8 LVs the majority of misclassification events were caused by a medium

producer cell line which was partially classified as a high producer (Figure 114). However, the high producer samples were characterised by lower Y2 scores, placing them closer to the classification threshold compared to 11 LVs model. This would indicate that while the current model has a high prediction accuracy of high producers, there are adjustments which can improve prediction of medium producers at a cost of increased risk of misclassification of high producers.

Overall performance of the model was satisfying with the version using 11 LVs and $2x10^7$ TU/ml threshold correctly predicted 87.5% of the validation samples which was increased to 100% prediction accuracy when using averaged mass spectra. This indicates that despite limited project timeframe and the variation in the data it is possible to apply PLS-DA to predict cell performance and therefore apply the method to triage cell lines during development. There are limitations to the model caused by the amount of cell lines available to calibrate and validate it but it can be used as a tool to guide decision making and can be improved over time through addition of more cell line data and model recalibration.

4.4.3. Application and implementation of the model

The PLS-DA model was designed with an intention to be integrated into OXB cell line development process. The main benefit of using the PLS-DA model is time and resource savings as well as improved characterisation of cell lines at the early stage of development. The manual process of cell line development is lengthy and only a small amount of cells can be fully characterised (Figure 125). Through process automation and use of predictive modelling the development timeline can be significantly reduced while also generating more clones and providing more information at an early stage, further reducing the need for lengthy bioreactor studies (Figure 126). By implementing these changes, the overall time required to develop new cell lines can be reduced by about 10 weeks while providing more information for each clone. This is achieved through limited dilution cloning in suspension (reducing the time needed for suspension adaptation), high-throughput screening and use of the PLS-DA model for large scale performance prediction at an early stage of the development. This approach also saves the cost of materials needed for large scale studies which is especially significant in the case of packaging cell lines which still require expensive reagents for transfection of the genome plasmid. In case of the potential application of the model in PrCLs the material savings are smaller but still significant.

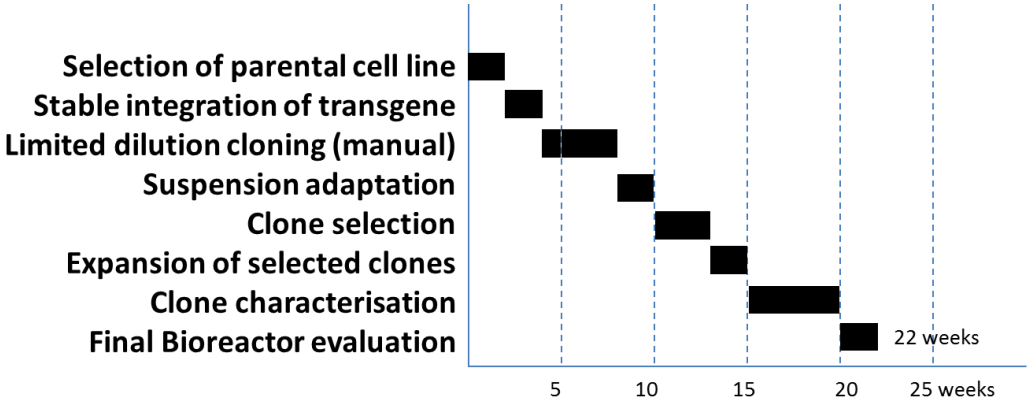


Figure 125: Packaging cell line development timeline using manual method.

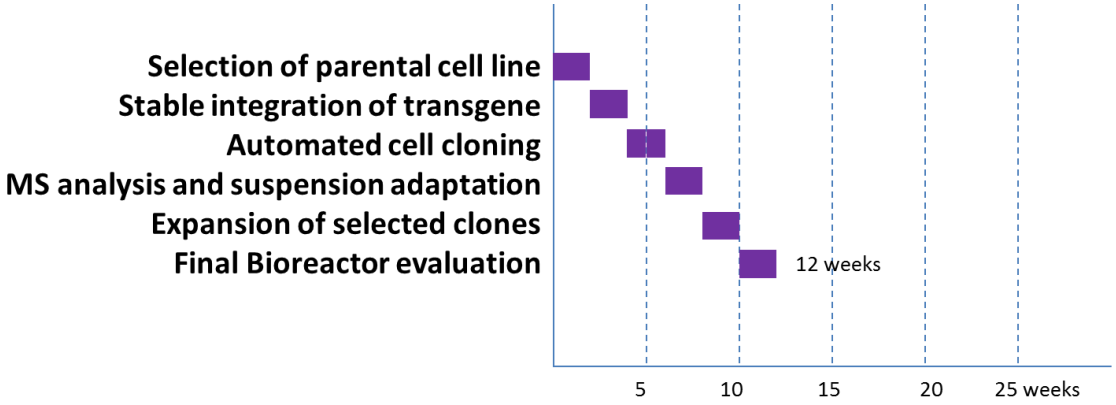


Figure 126: Reduced packaging cell line development timeline through automated method with support of the PLS-DA model.

Implementation of this PLS-DA supported method for cell line development would require following an accelerated and automated process (Table 10). Some of the materials and equipment required are already implemented in OXB workflow: ACSS has been successfully used for cell line development and could be applied to the new method as well; ambr15® is used for early cell line characterisation for selected candidates, it was used to generate samples for the model and could be accommodated for generation of samples for MALDI-ToF MS. The main limitation is obtaining MALDI-ToF MS results and implementation of the PLS-DA model itself.

Process step	Requirement for implementation
Automated high throughput cell line generation	ACSS has been used as part of new cell line development campaigns; implementation should not require any additional adjustments
Initial cell line development and selection	Remains the same as in the current cell line development process
Small scale cell sample generation in ambr15® bioreactor	Already used as part of new cell lines' characterisation, sample acquisition requires a small adjustment in harvest procedure
Generation of MALDI-ToF MS data	Requires development of in-house MS capacity or establishment of outsourcing through external contractor
PLS-DA analysis and prediction of cell line performance	Requires software installation, establishment of standard protocol and staff training; Model recalibration may be required for new types of cells

Table 10: Requirements for practical implementation of PLS-DA based accelerated cell line development process.

Preparation of samples for MALDI-ToF MS was examined and validated as a robust method for cell and viral vector analysis (as described in Chapter 3). An established and well characterised protocol is available for sample characterisation. The main obstacle in use of MALDI-ToF MS is the lack of in-house equipment at OXB. For the EngD project all mass spectra were obtained in collaboration with the University of Kent. For a commercial application an alternative approach would be required and it would involve either purchase of new equipment at OXB or outsourcing of MS sample generation. In the first case the benefits of accelerated cell line development may not justify a purchase of expensive and specialised equipment. Moreover, operation of a mass spectrometer would require development of in-house expertise which may further delay implementation of the PLS-DA model. As an alternative, sample analysis can be outsourced. In such a case the logistics of sample transport and analysis need to be established to ensure that the methodology remains aligned with the one used in calibration and validation of the model. In both cases a comparison study would be required to ensure that the sample analysis step using a different arrangement provides results which would fit the data used during model development.

Concerning the PLS-DA model, while it was calibrated and validated with independent data sets demonstrating its viability, its implementation may require some adjustments. Current version of the model uses a combination of MATLAB® and PLS Toolbox software and requires a degree of expertise to use. MS data requires pre-processing which has been simplified by scripting. An R script was used to import data generated from MALDI-ToF MS. MATLAB® bioinformatics toolbox commands were arranged sequentially to facilitate spectra pre-processing. Finally, the PLS-DA method was applied through PLS

Toolbox graphic user interface. These steps can be recreated and new data can be applied to the model following the same protocol with relative ease and little training. Currently the process is simplified but requires a significant input from the user to properly arrange the data and execute the scripting. The analysis could be simplified and semi-automated by improved scripting and development of a custom GUI. However, while using the model can be streamlined, troubleshooting and maintenance of the model and associated data processing can become an issue in the future and would require a high level of understanding of the methodology to ensure continuous support for the method.

4.4.4. Future work

In this EngD thesis MALDI-ToF MS and PLS-DA demonstrated the potential of these methods for improving and accelerating cell line development for LVV production. As described above, there is a potential for improvement in terms of method implementation and longevity. MS is a method capable of generating a wealth of data which can be explored beyond what is covered by this EngD project. As described in Chapter 3, cell and vector samples could be pre-processed to separate individual proteins and obtain proteomic data which could be used as a reference for the MS fingerprints. While protein and peptide analysis is possible to perform using MALDI-ToF MS (L. Li et al., 2000), there are methods such as liquid chromatography coupled with MS (e.g. electrospray ionisation MS) which are better suited for proteomic analysis and would be recommended for further experiments (Geiger et al., 2012). This approach would improve the understanding of how the cell lines are influenced by the process and along with analysis of PCA and PLS loadings and contribution plots it would open a wealth of data to further guide cell line development. Identifying the individual proteins that contribute to clustering of the different cell and vector populations in PCA, or to classification of cell line productivity in PLS would have a beneficial impact on both cell line development and vector design. Proteomic analysis of samples generated throughout this EngD project combined with MVDA could lead to an improved process understanding and productivity. The trade-off is the additional time needed and the expertise to develop new methods and guide the proteomic analysis project.

The PLS-DA has shown to be able to predict LVV infectious titre reached by cells based on mass spectra from cells cultured on the ambr15® scale and based on viral vector titre data. This approach was chosen as appropriate within the practical constraints of this EngD project while providing a significant improvement in the cell line development process. The benefits could be extended further by using cell samples generated during earlier stages of

development as well as by using vector titre data from larger scale production bioreactors. Cell samples could be obtained from shake flasks which would accelerate the process and eliminate the need for use of ambr15® thus reducing material costs. It would also be possible to use cell samples from early adherent culture. While this would eliminate the need for suspension adaptation, it was demonstrated that cells grown adherently or in suspension have significantly different MS fingerprints which could interfere with the large-scale productivity prediction. Adaptation of adherent cell lines to serum-free suspension culture can be a lengthy process which can cause significant change in cell's protein profile and quality attributes. An ability to predict cell performance before suspension adaptation (i.e. significantly earlier than performed in this project) could be valuable, however currently it is not possible to guarantee a connection between cell characteristics in early adherent culture and large scale suspension culture productivity. Investigating the exact relationship between small scale adherent cell culture and large scale suspension culture and the cell protein profile could provide data that could improve how cells are adapted to serum-free conditions and form a basis for improved predictive models. These results could also inform the decision on whether it is viable to decrease the scale of cell culture used for MS data generation. The main benefit of using larger scale bioreactors for assessment of viral vector titre would be better comparison to the commercial scale conditions. While 0.5L MiniBio reactors are used for process development at OXB, the larger 7L EZ bioreactors are routinely used as a direct scale-down model of the 50 and 200L production scale bioreactors.

While scaling down the sampling has its own benefits, scaling up the process used for viral vector titre assessment can greatly improve the practical application of the model. The current version of the model uses 0.5L MiniBio reactors which are used in OXB for early stage process development and optimisation studies, often in a design of experiment configuration to assess multiple parameters. While this was not practical for the purpose of this EngD project, studying performance at larger scale would help to improve accuracy of the predictive model towards the commercial process. In an ideal scenario each cell line would be tested in a commercial scale bioreactor which is unrealistic due to time and cost constraints. However, using 7L EZ bioreactors or a similar system could result in improved viral vector titre data and therefore better predictive model performance. Additional studies would be required to ensure that the titre prediction was applicable across all bioreactor scales.

The main issue with such alternative approaches is the risk of decreased model performance especially using smaller scale cell samples and the increased time and material costs when using larger scale bioreactors. The difference in process conditions between a bioreactor and a shake flask can be significant (Humphrey, 1998) and its effect on mass spectra is not well characterised, especially in the area of LVV production. The current version of the model was designed based on the concept of obtaining significant improvement in cell line development whilst maximising the similarity in process conditions and cell composition to ensure the success of the project. Using smaller scale suspension culture could provide improvements to the model but there is a risk that the change in protein profile would make it more difficult to correctly classify cell productivity across the different scales. This problem could be magnified when using adherent cell samples as it was demonstrated that there is a significant change in mass spectra when transitioning from adherent to suspension culture, as described in Chapter 3. Nevertheless, further study of the observed differences in the protein profiles between adherent and suspension culture cells as well as at different scales of production is recommended. The potential benefits of this approach could help further streamline cell line development timeliness if the samples for the PLS-DA model could be obtained earlier in the process.

The improvements suggested above are unlikely to be justifiable for the current version of the model. However, a continuous drive for development of stable PrCLs for LVV production could justify further improvements to the PLS-DA model. First of all, it is worth considering whether the current data could be used to assist with PrCL development. While the majority of cell metabolism remains similar, the vector production process is different for PaCLs and PrCLs in the fact that the latter have fixed genome and therefore do not require a transfection step. Before the current model can be fully applied to work with PrCLs, a trial run should be performed to assess the effects of transgene on the cell performance prediction. The vector used in the EngD project had a transgene that encoded GFP which would not be present in the PrCLs when used to generate therapeutic products. Any detrimental effect of the genome could be minimised through the use of the transgene repression in production system (Maunder et al., 2017). It is still advisable to thoroughly test the model robustness before applying it to PrCLs. It is possible that the model would need to be recalibrated and re-validated for use with any new PrCLs. It is difficult to determine the exact difference between the two types of cell lines due to their novelty and the lack of available samples that could be used in this EngD project. Whilst the current model may or may not be applicable to PrCLs,

the method itself was demonstrated to predict cell performance and should be applicable to new cell types. If recalibration is deemed necessary for use with PrCLs, the model could benefit from extending the scale difference between cell samples used for MS and the scale used for viral vector productivity characterisation, as described above.

Regardless of the future extension of the PLS-DA model functionality, it can be applied in development of high titre LVV PaCLs. For the best results, this application should be treated as a continuous improvement process. The model can be further refined by including more samples through recalibration and the prediction performance can be monitored and compared to production scale as new cells are used. This way after each campaign the model can be updated with new cell line data to improve its predictive power. This approach would require a significant resources and time investment to ensure that the new data is compatible with the previous one but the benefits of continuous improvement would be significant. While the PLS-DA model can be used in its current form to improve decision making in cell line development, the method could benefit from increased numbers of samples. This requirement could be met over time along with more characterisation data to further improve the robustness of the predictive model.

4.5. Conclusions

MALDI-ToF MS was used to develop a PLS-DA model capable of predicting large scale performance of cell lines without the need for bioreactor studies and FACS analysis. The model was calibrated using data from ten HEK293T PaCLs which were high and low producers. The mass spectra data from microscale culture (ambr®15) was used alongside LVV titre data from larger bioreactor scale (0.5L Applikon® MiniBio) to design the model. It was then validated by classifying eight independently developed packaging cell lines with high prediction accuracy. The PLS-DA model was demonstrated to be capable of predicting cell lines performance based on MS data.

The methods described in this chapter form an extensive toolset which can be applied in the industrial setting to the development of both stable PaCLs and PrCLs. The main application is improving cell line development process through prediction of cell lines performance, therefore enabling a greater number of clones to be screened and improved accuracy of performance prediction. The PLS-DA model could be implemented in the high throughput process to triage thousands of clones allowing further evaluation studies to focus on small number of the most promising high producing clones.

The main obstacle to implementing such an approach is the cost of either purchasing a suitable mass spectrometer which would require a significant up-front investment, or identifying a suitable contract research organisation that would be able to run these analyses routinely for OXB. MALDI-ToF MS is a time sensitive method where sample incubation time, spotting and the time between spotting and analysis can all have an impact on the mass spectra itself. Therefore, a reliable arrangement would have to be established by OXB, allowing consistent collection of mass spectra with minimal variation in sample processing and transport. A similar limitation exists for the other potential applications of MALDI-ToF MS at OXB.

Overall, this EngD project set out to explore and develop MVDA-based methodology to improve LVV production process. This included assessing feasibility of using MVDA in monitoring and development of LVV production process which was achieved by collecting process data from OXB's GMP manufacturing records of LVV. In Chapter 2 PCA was determined as a viable method to monitor batch-to-batch variation, however the method requires further refinement by establishing a golden batch for comparison and using better classification of data based on titre and product. Further PCA using research data from

suspension process development highlighted correlations between process parameters and major trends of cell performance and pH throughout the process as well as the effect of power per volume on LVV production. Analysis of the BMRs and research data provided further process understanding by identifying key areas of improvement: in manufacturing the effect of product type on titre and PCA which impacts future approach to application of MVDA in batch monitoring, along with recommendation to establish golden batch for each product; in process development the effect of power per volume and trends of pH and cell viability throughout the process were identified with lower pH setpoint and power per volume as conditions recommended for further investigation to increase the titre.

A major goal of the project was also to establish MS methodology for analysis of viral vectors and cells used in their production and to assess LVV production process and cell lines using this methodology. Chapter 3 described how combination of MALDI-ToF MS and MVDA were developed and used to improve the process understanding of LVV production through analysis of available cell and viral vector samples. The use of these tools has been shown to be an effective and robust approach for assessing a large number of mass spectra from a diverse set of samples. Characterisation of cell samples was successfully used to identify major areas of variation between cells: adherent and suspension culture as well as difference of mass spectra between individual cell lines. Viral vector samples proved to be more variable and problematic to assess. However, the combination of MALDI-ToF MS and PCA is still viable as a method to present the variation observed in viral samples and could be used for analysis of LVV purity and quality profile across multiple production cycles. However, the identification of individual proteins which contribute to the mass spectrum would require a different approach using LC-MS/MS technology and protein profiling.

Finally, the methods developed to assess cells and viral vector were applied to develop a predictive model of cell line performance designed to accelerate cell line development process. PLS-DA was used to design a model based on MS data from multiple packaging cell lines as described in Chapter 4. The model was demonstrated to successfully predict high performing cell lines within the validation data set. A road map was established to further develop and deploy this method within OXB and accelerate cell line development process. The future development of this modelling approach could involve continued support and increased sample size for model calibration through characterisation of more cell lines. The model application could also be expanded to include cell samples from earlier, smaller scale development and viral vector titre data from larger scale production to extend the scope

of the model and enhance its commercial potential. The PLS-DA model combines the MS methods and process understanding developed throughout this EngD thesis and brings it to a conclusion through practical application.

References

Abdi, H., & Williams, L. J. (2010). **Principal component analysis**. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4):433–459.

Adjali, O., Marodon, G., Steinberg, M., Mongellaz, C., Thomas-Vaslin, V., Jacquet, C., Taylor, N., & Klatzmann, D. (2005). **In vivo correction of ZAP-70 immunodeficiency by intrathymic gene transfer**. *Journal of Clinical Investigation*, *115*(8):2287–2295.

Aiuti, A., Cattaneo, F., Galimberti, S., Benninghoff, U., Cassani, B., Callegaro, L., Scaramuzza, S., Andolfi, G., Mirolo, M., Brigida, I., Tabucchi, A., Carlucci, F., Eibl, M., Aker, M., Slavin, S., Al-Mousa, H., Al Ghonaium, A., Ferster, A., Duppenthaler, A., Notarangelo, L., Wintergerst, U., Buckley, R. H., Bregni, M., Marktel, S., Valsecchi, M. G., Rossi, P., Ciceri, F., Miniero, R., Bordignon, C., & Roncarolo, M.-G. (2009). **Gene Therapy for Immunodeficiency Due to Adenosine Deaminase Deficiency**. *New England Journal of Medicine*, *360*(5):447–458.

Albrethsen, J. (2007). **Reproducibility in Protein Profiling by MALDI-TOF Mass Spectrometry**. *Clinical Chemistry*, *53*(5):852–858.

AlMasoud, N., Xu, Y., Nicolaou, N., & Goodacre, R. (2014). **Optimization of matrix assisted desorption/ionization time of flight mass spectrometry (MALDI-TOF-MS) for the characterization of Bacillus and Brevibacillus species**. *Analytica Chimica Acta*, *840*:49–57.

Amado, R. G., & Chen, I. S. (1999). **Lentiviral vectors--the promise of gene therapy within reach?** *Science (New York, N.Y.)*, *285*(5428):674–6.

Anderson, J. S., Javien, J., Nolta, J. A., & Bauer, G. (2009). **Preintegration HIV-1 Inhibition by a Combination Lentiviral Vector Containing a Chimeric TRIM5α Protein, a CCR5 shRNA, and a TAR Decoy**. *Molecular Therapy*, *17*(12):2103–2114.

Bandeira, V., Peixoto, C., Rodrigues, A. F., Cruz, P. E., Alves, P. M., Coroadinha, A. S., & Carrondo, M. J. T. (2012). **Downstream Processing of Lentiviral Vectors: Releasing Bottlenecks**. *Human Gene Therapy Methods*, *23*(4):255–263.

Barker, M., & Rayens, W. (2003). **Partial least squares for discrimination**. *Journal of Chemometrics*, *17*(3):166–173.

Beutler, E. (2001). **The Cline affair**. *Molecular Therapy*, 4(5):396-397.

Blaese, R. M., Culver, K. W., Miller, A. D., Carter, C. S., Fleisher, T., Clerici, M., Shearer, G., Chang, L., Chiang, Y., Tolstoshev, P., Greenblatt, J. J., Rosenberg, S. A., Klein, H., Berger, M., Mullen, C. A., Ramsey, W. J., Muul, L., Morgan, R. A., & Anderson, W. F. (1995). **T lymphocyte-directed gene therapy for ADA- SCID: initial trial results after 4 years.** *Science (New York, N.Y.)*, *270*(5235):475–80.

Bodnar, E., Ferreira Nascimento, T., Girard, L., Komatsu, E., Lopez, P., Gomes De Oliveira, A. G., Roy, R., Smythe, T., Zogbi, Y., Spearman, M., Tayi, V. S., Butler, M., & Perreault, H. (2015). **An integrated approach to analyze EG2-hFc monoclonal antibody N-glycosylation by MALDI-MS**. *Canadian Journal of Chemistry*, *93*(7):754–763.

Bordignon, C., Notarangelo, L. D., Nobili, N., Ferrari, G., Casorati, G., Panina, P., Mazzolari, E., Maggioni, D., Rossi, C., Servida, P., Ugazio, A. G., & Mavilio, F. (1995). **Gene therapy in peripheral blood lymphocytes and bone marrow for ADA-immunodeficient patients.** *Science (New York, N.Y.)*, *270*(5235):470–5.

Bro, R., Smilde, A. K., Smilde, A. K., Hubert, M., Song, X., Yu, R., Holmberg, M., & Lundstrom, I. (2014). **Principal component analysis**. *Analytical Methods*, *6*(9):2812.

Broussau, S., Jabbour, N., Lachapelle, G., Durocher, Y., Tom, R., Transfiguracion, J., Gilbert, R., & Massie, B. (2008). **Inducible Packaging Cells for Large-scale Production of Lentiviral Vectors in Serum-free Suspension Culture**. *Molecular Therapy*, *16*(3):500–507.

Caprioli, R. M., Farmer, T. B., & Jocelyn, G. (1997). **Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS**.

Carmo, M., Alves, A., Rodrigues, A. F., Coroadinha, A. S., Carrondo, M. J. T., Alves, P. M., & Cruz, P. E. (2009). **Stabilization of gammaretroviral and lentiviral vectors: from production to gene transfer**. *The Journal of Gene Medicine*, *11*(8):670–678.

Case, S. S., Price, M. A., Jordan, C. T., Yu, X. J., Wang, L., Bauer, G., Haas, D. L., Xu, D., Stripecke, R., Naldini, L., Kohn, D. B., & Crooks, G. M. (1999). **Stable transduction of quiescent CD34(+)CD38(-) human hematopoietic cells by HIV-1-based lentiviral vectors.** *Proceedings of the National Academy of Sciences of the United States of*

America, *96*(6):2988–93.

Chang, C.-W., Lai, Y.-S., Pawlik, K. M., Liu, K., Sun, C.-W., Li, C., Schoeb, T. R., & Townes, T. M. (2009). **Polycistronic Lentiviral Vector for "Hit and Run" Reprogramming of Adult Skin Fibroblasts to Induced Pluripotent Stem Cells**. *STEM CELLS*, *27*(5):1042–1049.

Cheeks, M. C., Kamal, N., Sorrell, A., Darling, D., Farzaneh, F., & Slater, N. K. H. (2009). **Immobilized metal affinity chromatography of histidine-tagged lentiviral vectors using monolithic adsorbents**. *Journal of Chromatography A*, *1216*(13):2705–2711.

Chen, Y., Ott, C. J., Townsend, K., Subbaiah, P., Aiyar, A., & Miller, W. M. (2009). **Cholesterol supplementation during production increases the infectivity of retroviral and lentiviral vectors pseudotyped with the vesicular stomatitis virus glycoprotein (VSV-G)**. *Biochemical Engineering Journal*, *44*(2–3):199–207.

CHMP (2016). **Strimvelis, autologous CD34+ enriched cell fraction that contains CD34+ cells transduced with retroviral vector that encodes for the human ADA cDNA sequence.** Accessed online on 12 July 2017 at: www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/003854/humn_med_001985.jsp

CHMP (2018). **CHMP summary of positive opinion for Kymriah**. Accessed online on 10 Aug 2018 at: www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/004090/smops/Positive/human_smop_001319.jsp

Chung, J., Scherer, L. J., Gu, A., Gardner, A. M., Torres-Coronado, M., Epps, E. W., DiGiusto, D. L., & Rossi, J. J. (2014). **Optimized Lentiviral Vectors for HIV Gene Therapy: Multiplexed Expression of Small RNAs and Inclusion of MGMTP140K Drug Resistance Gene**. *Molecular Therapy*, *22*(5):952–963.

Clark, A. E., Kaleta, E. J., Arora, A., & Wolk, D. M. (2013). **Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry: a Fundamental Shift in the Routine Practice of Clinical Microbiology**. *Clinical Microbiology Reviews*, *26*(3):547–603.

Cline, M. J. (1985). **Perspectives for gene therapy: Inserting new genetic information**

into mammalian cells by physical techniques and viral vectors. *Pharmacology and Therapeutics*, *29*(1):69–92.

Cockrell, A. S., & Kafri, T. (2007). **Gene delivery by lentivirus vectors.** *Molecular Biotechnology*, *36*(3):184–204.

Cornish, T. J., & Cotter, R. J. (1993). **A curved-field reflectron for improved energy focusing of product ions in time-of-flight mass spectrometry**. *Rapid Communications in Mass Spectrometry*, *7*(11):1037–1040.

Davie, J. R. (2003). **Inhibition of histone deacetylase activity by butyrate.** *The Journal of Nutrition*, *133*(7 Suppl):2485S–2493S.

Denard, J., Rundwasser, S., Laroudie, N., Gonnet, F., Naldini, L., Radrizzani, M., Galy, A., Merten, O.W., Danos, O. & Svinartchouk, F. (2009). **Quantitative proteomic analysis of lentiviral vectors using 2-DE**. *Proteomics* 9(14):3666-3676

Deyle, D. R., & Russell, D. W. (2009). **Adeno-associated virus vector integration.** *Current Opinion in Molecular Therapeutics*, *11*(4):442–7.

Dull, T., Zufferey, R., Kelly, M., Mandel, R. J., Nguyen, M., Trono, D., & Naldini, L. (1998). **A third-generation lentivirus vector with a conditional packaging system.** *Journal of Virology*, *72*(11):8463–71.

Ellis, B. L., Potts, P. R., & Porteus, M. H. (2011). **Creating Higher Titer Lentivirus with Caffeine**. *Human Gene Therapy*, *22*(1):93–100.

Escors, D., & Breckpot, K. (2010). **Lentiviral vectors in gene therapy: their current status and future potential.** *Archivum Immunologiae et Therapiae Experimentalis*, *58*(2):107–19.

Feng, H., Sim, L. C., Wan, C., Wong, N. S. C., & Yang, Y. (2011). **Rapid characterization of protein productivity and production stability of CHO cells by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry**. *Rapid Communications in Mass Spectrometry*, *25*(10):1407–1412.

Feng, H., Wong, N. S. C., Sim, L. C., Wati, L., Ho, Y., & Lee, M. M. (2010). **Rapid characterization of high/low producer CHO cells using matrix-assisted laser desorption/ionization time-of-flight**. *Rapid Communications in Mass Spectrometry*,

24(9):1226–1230.

Fenselau, C. (1997). **MALDI MS and strategies for protein analysis.** *Analytical Chemistry*, *69*(21):661A–665A.

Fink, D. J., & Glorioso, J. C. (2007). **Herpes Simplex Viral Vectors in Gene Therapy**. In *Encyclopedia of Life Sciences*. Chichester: John Wiley & Sons, Ltd.

Fischer, A., Hacein-Bey-Abina, S., Lagresle, C., Garrigue, A., & Cavazana-Calvo, M. (2005). **[Gene therapy of severe combined immunodeficiency disease: proof of principle of efficiency and safety issues. Gene therapy, primary immunodeficiencies, retrovirus, lentivirus, genome].** *Bulletin de l'Academie Nationale de Medecine*, *189*(5):779-85–8.

Gama-Norton, L., Botezatu, L., Herrmann, S., Schweizer, M., Alves, P. M., Hauser, H., & Wirth, D. (2011). **Lentivirus Production Is Influenced by SV40 Large T-Antigen and Chromosomal Integration of the Vector in HEK293 Cells**. *Human Gene Therapy*, *22*(10):1269–1279.

Geiger, T., Wehner, A., Schaab, C., Cox, J., & Mann, M. (2012). **Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins.** *Molecular & Cellular Proteomics : MCP*, *11*(3):M111.014050.

Geraerts, M., Willems, S., Baekelandt, V., Debyser, Z., & Gijsbers, R. (2006). **Comparison of lentiviral vector titration methods.** *BMC Biotechnology*, *6*:34.

Glassey, J. (2012). **Multivariate Data Analysis for Advancing the Interpretation of Bioprocess Measurement and Monitoring Data**. In *Advances in biochemical engineering/biotechnology* (Vol. 132, pp. 167–191).

Glish, G. L., & Vachet, R. W. (2003). **The basics of mass spectrometry in the twenty-first century**. *Nature Reviews Drug Discovery*, *2*(2):140–150.

Gogichaeva, N. V., Williams, T., & Alterman, M. A. (2007). **MALDI TOF/TOF tandem mass spectrometry as a new tool for amino acid analysis**. *Journal of the American Society for Mass Spectrometry*, *18*(2):279–284.

Goverdhana, S., Puntel, M., Xiong, W., Zirger, J. M., Barcia, C., Curtin, J. F., Soffer, E. B., Mondkar, S., King, G. D., Hu, J., Sciascia, S. A., Candolfi, M., Greengold, D. S., Lowenstein, P. R., & Castro, M. G. (2005). **Regulatable gene expression systems for**

gene therapy applications: progress and future challenges. *Molecular Therapy : The Journal of the American Society of Gene Therapy*, *12*(2):189–211.

Hacein-Bey-Abina, S., Hauer, J., Lim, A., Picard, C., Wang, G. P., Berry, C. C., Martinache, C., Rieux-Laucat, F., Latour, S., Belohradsky, B. H., Leiva, L., Sorensen, R., Debré, M., Casanova, J. L., Blanche, S., Durandy, A., Bushman, F. D., Fischer, A., & Cavazzana-Calvo, M. (2010). **Efficacy of Gene Therapy for X-Linked Severe Combined Immunodeficiency**. *New England Journal of Medicine*, *363*(4):355–364.

Holic, N., Seye, A. K., Majdoul, S., Martin, S., Merten, O. W., Galy, A., & Fenard, D. (2014). **Influence of Mildly Acidic pH Conditions on the Production of Lentiviral and Retroviral Vectors**. *Human Gene Therapy Clinical Development*, *25*(3):178–185.

Hotelling, H. (1933). **Analysis of a complex of statistical variables into principal components.** *Journal of Educational Psychology*, *24*(6):417–441.

Humphrey, A. (1998). **Shake Flask to Fermentor: What Have We Learned?** *Biotechnology Progress*, *14*(1):3–7.

ICH. (2009). **ICH HARMONISED TRIPARTITE GUIDELINE PHARMACEUTICAL DEVELOPMENT Q8(R2)**.

Jacobs, A., Breakefield, X. O., & Fraefel, C. (1999). **HSV-1-based vectors for gene therapy of neurological diseases and brain tumors: part II. Vector systems and applications.** *Neoplasia (New York, N.Y.)*, *1*(5):402–16.

Kafri, T., van Praag, H., Ouyang, L., Gage, F. H., & Verma, I. M. (1999). **A packaging cell line for lentivirus vectors.** *Journal of Virology*, *73*(1):576–84.

Karacostas, V., Wolffe, E. J., Nagashima, K., Gonda, M. A., & Moss, B. (1993). **Overexpression of the HIV-1 Gag-Pol Polyprotein Results in Intracellular Activation of HIV-1 Protease and Inhibition of Assembly and Budding of Virus-like Particles**. *Virology*, *193*(2):661–671.

Karas, M., & Bahr, U. (1990). **Laser desorption ionization mass spectrometry of large biomolecules**. *TrAC Trends in Analytical Chemistry*, *9*(10):321–325.

Karas, M., & Krüger, R. (2003). **Ion Formation in MALDI: The Cluster Ionization Mechanism**.

Keeler, A., ElMallah, M., & Flotte, T. (2017). **Gene Therapy 2017: Progress and Future Directions**. *Clinical and Translational Science*, *10*(4):242–248.

Kim, V. N., Mitrophanous, K., Kingsman, S. M., & Kingsman, A. J. (1998). **Minimal requirement for a lentivirus vector based on human immunodeficiency virus type 1.** *Journal of Virology*, *72*(1):811–6.

Kingston, R. E., Chen, C. A., & Okayama, H. (2003). **Calcium phosphate transfection.** *Current Protocols in Cell Biology*, *Chapter 20*:Unit 20.3.

Knochenmuss, R. (2006). **Ion formation mechanisms in UV-MALDI**. *The Analyst*, *131*(9):966.

Korfmacher, W.A. (2007). **Foundation review: Principles and applications of LC-MS in new drug discovery**. *Drug Discovery Today* 10(20):1357-1367

Kotsopoulou, E., Kim, V. N., Kingsman, A. J., Kingsman, S. M., & Mitrophanous, K. A. (2000). **A Rev-independent human immunodeficiency virus type 1 (HIV-1)-based vector that exploits a codon-optimized HIV-1 gag-pol gene.** *Journal of Virology*, *74*(10):4839–52.

Koubek, J., Uhlik, O., Jecna, K., Junkova, P., Vrkoslavova, J., Lipov, J., Kurzawova, V., Macek, T., & Mackova, M. (2012). **Whole-cell MALDI-TOF: Rapid screening method in environmental microbiology**. *International Biodeterioration & Biodegradation*, *69*:82–86.

Kumar, M., Keller, B., Makalou, N., & Sutton, R. E. (2001). **Systematic Determination of the Packaging Limit of Lentiviral Vectors**. *Human Gene Therapy*, *12*(15):1893–1905.

Kuroda, H., Kutner, R. H., Bazan, N. G., & Reiser, J. (2009). **Simplified lentivirus vector production in protein-free media using polyethylenimine-mediated transfection**. *Journal of Virological Methods*, *157*(2):113–121.

Kutner, R. H., Puthli, S., Marino, M. P., & Reiser, J. (2009). **Simplified production and concentration of HIV-1-based lentiviral vectors using HYPERFlask vessels and anion exchange membrane chromatography**. *BMC Biotechnology*, *9*(1):10.

Lesch, H. P., Turpeinen, S., Niskanen, E. A., Mähönen, A. J., Airenne, K. J., & Ylä-Herttuala, S. (2008). **Generation of lentivirus vectors using recombinant**

baculoviruses. *Gene Therapy*, *15*(18):1280–1286.

Levine, B. L., Humeau, L. M., Boyer, J., MacGregor, R.-R., Rebello, T., Lu, X., Binder, G. K., Slepushkin, V., Lemiale, F., Mascola, J. R., Bushman, F. D., Dropulic, B., & June, C. H. (2006). **Gene transfer in humans using a conditionally replicating lentiviral vector**. *Proceedings of the National Academy of Sciences*, *103*(46):17372–17377.

Li, F., Vijayasankaran, N., Shen, A. Y., Kiss, R., & Amanullah, A. (2010). **Cell culture processes for monoclonal antibody production.** *mAbs*, *2*(5):466–79.

Li, L., Garden, R. W., & Sweedler, J. V. (2000). **Single-cell MALDI: a new tool for direct peptide profiling**. *Trends in Biotechnology*, *18*(4):151–160.

Liu, Z., & Schey, K. L. (2005a). **Optimization of a MALDI TOF-TOF mass spectrometer for intact protein analysis**. *Journal of the American Society for Mass Spectrometry*, *16*(4):482–490.

Liu, Z., & Schey, K. L. (2005b). **Optimization of a MALDI TOF-TOF mass spectrometer for intact protein analysis**. *Journal of the American Society for Mass Spectrometry*, *16*(4):482–490.

Loo, J. A., Loo, R. R. O., Light, K. J., Edmonds, C. G., & Smith, R. D. (1992). **Multiply charged negative ions by electrospray ionization of polypeptides and proteins**. *Analytical Chemistry*, *64*(1):81–88.

Lu, X., Humeau, L., Slepushkin, V., Binder, G., Yu, Q., Slepushkina, T., Chen, Z., Merling, R., Davis, B., Chang, Y.-N., & Dropulic, B. (2004). **Safe two-plasmid production for the first clinical lentivirus vector that achieves >99% transduction in primary cells using a one-step protocol**. *The Journal of Gene Medicine*, *6*(9):963–973.

Maunder, H. E., Wright, J., Kolli, B. R., Vieira, C. R., Mkandawire, T. T., Tatoris, S., Kennedy, V., Iqball, S., Devarajan, G., Ellis, S., Lad, Y., Clarkson, N. G., Mitrophanous, K. A., & Farley, D. C. (2017). **Enhancing titres of therapeutic viral vectors using the transgene repression in vector production (TRiP) system**. *Nature Communications*, *8*:14834.

May, C., Rivella, S., Callegari, J., Heller, G., Gaensler, K. M. L., Luzzatto, L., & Sadelain, M. (2000). **Therapeutic haemoglobin synthesis in beta-thalassaemic mice expressing**

lentivirus-encoded human beta-globin. *Nature*, *406*(6791):82–86.

McGarrity, G. J., Hoyah, G., Winemiller, A., Andre, K., Stein, D., Blick, G., Greenberg, R. N., Kinder, C., Zolopa, A., Binder-Scholl, G., Tebas, P., June, C. H., Humeau, L. M., & Rebello, T. (2013). **Patient monitoring and follow-up in lentiviral clinical trials**. *The Journal of Gene Medicine*, *15*(2):78–82.

Merten, O.-W., Charrier, S., Laroudie, N., Fauchille, S., Dugué, C., Jenny, C., Audit, M., Zanta-Boussif, M.-A., Chautard, H., Radrizzani, M., Vallanti, G., Naldini, L., Noguiez-Hellin, P., & Galy, A. (2011). **Large-Scale Manufacture and Characterization of a Lentiviral Vector Produced for Clinical *Ex Vivo* Gene Therapy Application**. *Human Gene Therapy*, *22*(3):343–356.

Merten, O.-W., Hebben, M., & Bovolenta, C. (2016). **Production of lentiviral vectors.** *Molecular Therapy. Methods & Clinical Development*, *3*:16017.

Merten, O.-W., Schweizer, M., Chahal, P., & Kamen, A. A. (2014a). **Manufacturing of viral vectors: part II. Downstream processing and safety aspects**. *Pharm. Bioprocess*, *2*(3):237–251.

Merten, O.-W., Schweizer, M., Chahal, P., & Kamen, A. A. (2014b). **Manufacturing of viral vectors for gene therapy: part I. Upstream processing**. *Pharmaceutical Bioprocessing*, *2*(2):183–203.

Mitrophanous, K. A., Yoon, S., Rohll, J. B., Patil, D., Wilkes, F. J., Kim, V. N., Kingsman, S. M., Kingsman, A. J., & Mazarakis, N. D. (1999). **Stable gene transfer to the nervous system using a non-primate lentiviral vector**. *Gene Therapy*, *6*(11):1808–1818.

Miyoshi, H., Blömer, U., Takahashi, M., Gage, F. H., & Verma, I. M. (1998). **Development of a self-inactivating lentivirus vector.** *Journal of Virology*, *72*(10):8150–7.

Miyoshi, H., Smith, K. A., Mosier, D. E., Verma, I. M., & Torbett, B. E. (1999). **Transduction of human CD34+ cells that mediate long-term engraftment of NOD/SCID mice by HIV vectors.** *Science (New York, N.Y.)*, *283*(5402):682–6.

Momo, R. A., Povey, J. F., Smales, C. M., O'Malley, C. J., Montague, G. A., & Martin, E. B. (2013). **MALDI-ToF mass spectrometry coupled with multivariate pattern recognition analysis for the rapid biomarker profiling of Escherichia coli in**

different growth phases. *Analytical and Bioanalytical Chemistry*, *405*(25):8251–8265.

Moss, R. B., Rodman, D., Spencer, L. T., Aitken, M. L., Zeitlin, P. L., Waltz, D., Milla, C., Brody, A. S., Clancy, J. P., Ramsey, B., Hamblett, N., & Heald, A. E. (2004). **Repeated adeno-associated virus serotype 2 aerosol-mediated cystic fibrosis transmembrane regulator gene transfer to the lungs of patients with cystic fibrosis: a multicenter, double-blind, placebo-controlled trial.** *Chest*, *125*(2):509–21.

Naldini, L. (2015). **Gene therapy returns to centre stage**. *Nature*, *526*(7573):351–360.

Nesbeth, D., Williams, S. L., Chan, L., Brain, T., Slater, N. K. H., Farzaneh, F., & Darling, D. (2006). **Metabolic Biotinylation of Lentiviral Pseudotypes for Scalable Paramagnetic Microparticle-Dependent Manipulation**. *Molecular Therapy*, *13*(4):814–822.

Nomikos, P., & MacGregor, J. F. (1994). **Monitoring batch processes using multiway principal component analysis**. *AIChE Journal*, *40*(8):1361–1375.

Palfi, S., Gurruchaga, J. M., Ralph, G. S., Lepetit, H., Lavisse, S., Buttery, P. C., Watts, C., Miskin, J., Kelleher, M., Deeley, S., Iwamuro, H., Lefaucheur, J. P., Thiriez, C., Fenelon, G., Lucas, C., Brugières, P., Gabriel, I., Abhay, K., Drouot, X., Tani, N., Kas, A., Ghaleh, B., Le Corvoisier, P., Dolphin, P., Breen, D. P., Mason, S., Guzman, N. V., Mazarakis, N. D., Radcliffe, P. A., Harrop, R., Kingsman, S. M., Rascol, O., Naylor, S., Barker, R. A., Hantraye, P., Remy, P., Cesaro, P., & Mitrophanous, K. A. (2014). **Long-term safety and tolerability of ProSavin, a lentiviral vector-based gene therapy for Parkinson's disease: a dose escalation, open-label, phase 1/2 trial**. *The Lancet*, *383*(9923):1138–1146.

Pan, H., Mostoslavsky, G., Eruslanov, E., Kotton, D. N., & Kramnik, I. (2008). **Dual-promoter lentiviral system allows inducible expression of noxious proteins in macrophages**. *Journal of Immunological Methods*, *329*(1–2):31–44.

Pearson, K. (1901). **On lines and planes of closest fit to systems of points in space**. *Philosophical Magazine Series 6*, *2*(11):559–572.

Peng, Z. (2005). **Current Status of Gendicine in China: Recombinant Human Ad-p53 Agent for Treatment of Cancers**. *Human Gene Therapy*, *16*(9):1016–1027.

Pham, P. L., Kamen, A., & Durocher, Y. (2006). **Large-Scale Transfection of Mammalian Cells for the Fast Production of Recombinant Protein**. *Molecular Biotechnology*, *34*(2):225–238.

Pitt, J.J. (2009). **Principles and Applications of Liquid Chromatography-Mass Spectrometry in Clinical Biochemistry**. *Clinical Biochemistry Reviews* 30(1): 19–34 (2009).

Porter, A.J., Dickson, A.J., Racher and A.J. (2010). **Strategies for selecting recombinant CHO cell lines for cGMP manufacturing: realizing the potential in bioreactors.** *Biotechnology progress*, *26*(5):1446-54.

Povey, J. F., O'Malley, C. J., Root, T., Martin, E. B., Montague, G. A., Feary, M., Trim, C., Lang, D. A., Alldread, R., Racher, A. J., & Smales, C. M. (2014). **Rapid high-throughput characterisation, classification and selection of recombinant mammalian cell line phenotypes using intact cell MALDI-ToF mass spectrometry fingerprinting and PLS-DA modelling**. *Journal of Biotechnology*, *184*:84–93.

Puumalainen, A.-M., Vapalahti, M., Agrawal, R. S., Kossila, M., Laukkanen, J., Lehtolainen, P., Viita, H., Paljärvi, L., Vanninen, R., & Ylä-Herttuala, S. (1998). ***β*** **-Galactosidase Gene Transfer to Human Malignant Glioma *In Vivo* Using Replication-Deficient Retroviruses and Adenoviruses**. *Human Gene Therapy*, *9*(12):1769–1774.

Quinonez, R., & Sutton, R. E. (2002). **Lentiviral Vectors for Gene Delivery into Cells**. *DNA and Cell Biology*, *21*(12):937–951.

Rein, D. T., Breidenbach, M., & Curiel, D. T. (2006). **Current developments in adenovirus-based cancer gene therapy.** *Future Oncology (London, England)*, *2*(1):137–43.

Reiser, J. (2000). **Production and concentration of pseudotyped HIV-1-based gene transfer vectors**. *Gene Therapy*, *7*(11):910–913.

Rogers, S., Lowenthal, A., Terheggen, H. G., & Columbo, J. P. (1973). **Induction of arginase activity with the Shope papilloma virus in tissue culture cells from an argininemic patient.** *The Journal of Experimental Medicine*, *137*(4):1091–6.

Rohll, J. B., Mitrophanous, K. A., Martin-Rendon, E., Ellard, F. M., Radcliffe, P. A.,

Mazarakis, N. D., & Kingsman, S. M. (2002). **Design, production, safety, evaluation, and clinical applications of nonprimate lentiviral vectors.** *Methods in Enzymology*, *346*:466–500.

Rowland-Jones, R. C., van den Berg, F., Racher, A. J., Martin, E. B., & Jaques, C. (2017). **Comparison of spectroscopy technologies for improved monitoring of cell culture processes in miniature bioreactors**. *Biotechnology Progress*, *33*(2):337–346.

Sadelain, M., Chang, A., & Lisowski, L. (2009). **Supplying Clotting Factors From Hematopoietic Stem Cell–derived Erythroid and Megakaryocytic Lineage Cells**. *Molecular Therapy*, *17*(12):1994–1999.

Sastry, L., Xu, Y., Cooper, R., Pollok, K., & Cornetta, K. (2004). **Evaluation of Plasmid DNA Removal from Lentiviral Vectors by Benzonase Treatment**. *Human Gene Therapy*, *15*(2):221–226.

Schaiberger, A. M., & Moss, J. A. (2008). **Optimized sample preparation for MALDI mass spectrometry analysis of protected synthetic peptides**. *Journal of the American Society for Mass Spectrometry*, *19*(4):614–619.

Schambach, A., Swaney, W. P., & Loo, J. C. M. van der. (2009). **Design and Production of Retro- and Lentiviral Vectors for Gene Expression in Hematopoietic Cells**. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 506, pp. 191–205).

Schwamb, S., & Wiedemann, P. (2015). **MALDI-TOF mass spectrometry biotyping: At line monitoring of recombinant CHO cell cultures**. *BMC Proceedings*, *9*(Suppl 9):P53.

Schweizer, M., & Merten, O.-W. (2010). **Large-scale production means for the manufacturing of lentiviral vectors.** *Current Gene Therapy*, *10*(6):474–86.

Scott, L. J. (2015). **Alipogene Tiparvovec: A Review of Its Use in Adults with Familial Lipoprotein Lipase Deficiency**. *Drugs*, *75*(2):175–182.

Segura, M. de las M., Garnier, A., Di Falco, M.R., Whissell, G., Meneses-Acosta, A., Arcand, N. & Kamen, A. (2008). **Identification of Host Proteins Associated With Retroviral Vector Particles by Proteomic Analysis of Highly Purified Vector Preparations.** *Journal of Virology* 82(3):1107-17.

Segura, M. de las M., Kamen, A., Lavoie, M.-C., & Garnier, A. (2007). **Exploiting heparin-binding properties of MoMLV-based retroviral vectors for affinity chromatography**. *Journal of Chromatography B, 846*(1):124–131.

Segura, M. de las M., Kamen, A., Trudel, P., & Garnier, A. (2005). **A novel purification strategy for retrovirus gene therapy vectors using heparin affinity chromatography**. *Biotechnology and Bioengineering*, *90*(4):391–404.

Segura, M. de las M., Mangion, M., Gaillet, B., & Garnier, A. (2013). **New developments in lentiviral vector design, production and purification**. *Expert Opinion on Biological Therapy*, *13*(7):987–1011.

Selkirk, S. (2004). **Gene therapy in clinical medicine**. *Postgraduate Medical Journal*, 80(948):560-570.

Sena-Esteves, M., Tebbets, J. C., Steffens, S., Crombleholme, T., & Flake, A. W. (2004). **Optimized large-scale production of high titer lentivirus vector pseudotypes**. *Journal of Virological Methods*, *122*(2):131–139.

Seng, P., Drancourt, M., Gouriet, F., La Scola, B., Fournier, P., Rolain, J. M., & Raoult, D. (2009). **Ongoing Revolution in Bacteriology: Routine Identification of Bacteria by Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry**. *Clinical Infectious Diseases*, *49*(4):543–551.

Slepushkin, V., Chang, N., & Cohen, R. (2003). **Large-scale purification of a lentiviral vector by size exclusion chromatography or Mustang Q ion exchange capsule**.

Smith, S. L., & Shioda, T. (2009). **Advantages of COS-1 monkey kidney epithelial cells as packaging host for small-volume production of high-quality recombinant lentiviruses**. *Journal of Virological Methods*, *157*(1):47–54.

Sommer, C. A., Stadtfeld, M., Murphy, G. J., Hochedlinger, K., Kotton, D. N., & Mostoslavsky, G. (2009). **Induced Pluripotent Stem Cell Generation Using a Single Lentiviral Stem Cell Cassette**. *Stem Cells*, *27*(3):543–549.

Stewart, H. J., Fong-Wong, L., Strickland, I., Chipchase, D., Kelleher, M., Stevenson, L., Thoree, V., McCarthy, J., Ralph, G. S., Mitrophanous, K. A., & Radcliffe, P. A. (2011). **A Stable Producer Cell Line for the Manufacture of a Lentiviral Vector for Gene**

Therapy of Parkinson's Disease. *Human Gene Therapy*, *22*(3):357–369.

Stewart, H. J., Leroux-Carlucci, M. A., Sion, C. J. M., Mitrophanous, K. A., & Radcliffe, P. A. (2009). **Development of inducible EIAV-based lentiviral vector packaging and producer cell lines**. *Gene Therapy*, *16*(6):805–814.

Stolberg, S. G. (1999). **The biotech death of Jesse Gelsinger.** *The New York Times Magazine*, 136–140, 149–150.

Stone, K. L., Deangelis, R., LoPresti, M., Jones, J., Papov, V. V., & Williams, K. R. (1998). **Use of liquid chromatography-electrospray ionization-tandem mass spectrometry (LC-ESI-MS/MS) for routine identification of enzymatically digested proteins separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis**. *Electrophoresis*, *19*(6):1046–1052.

Stornaiuolo, A., Piovani, B. M., Bossi, S., Zucchelli, E., Corna, S., Salvatori, F., Mavilio, F., Bordignon, C., Rizzardi, G. P., & Bovolenta, C. (2013). **RD2-MolPack- *Chim3,* a Packaging Cell Line for Stable Production of Lentiviral Vectors for Anti-HIV Gene Therapy**. *Human Gene Therapy Methods*, *24*(4):228–240.

Strayer, D. S. (2000). **SV40-based gene therapy vectors: turning an adversary into a friend.** *Current Opinion in Molecular Therapeutics*, *2*(5):570–8.

Takahashi, K., & Yamanaka, S. (2006). **Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors**. *Cell*, *126*(4):663–676.

Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., & Matsuo, T. (1988). **Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry**. *Rapid Communications in Mass Spectrometry*, *2*(8):151–153.

Tom, R., Bisson, L., & Durocher, Y. (2008). **Culture of HEK293-EBNA1 Cells for Production of Recombinant Proteins.** *CSH Protocols*, *2008*:pdb.prot4976.

Tsao, Y.-S., Condon, R., Schaefer, E., Lio, P., & Liu, Z. (2001). **Development and improvement of a serum-free suspension process for the production of recombinant adenoviral vectors using HEK293 cells.** *Cytotechnology*, *37*(3):189–198.

Tudó, G., Monté, M. R., Vergara, A., López, A., Hurtado, J. C., Ferrer-Navarro, M., Vila, J.,

& Gonzalez-Martin, J. (2015). **Implementation of MALDI-TOF MS technology for the identification of clinical isolates of Mycobacterium spp. in mycobacterial diagnosis**. *European Journal of Clinical Microbiology & Infectious Diseases*, *34*(8):1527–1532.

van der Loo, J. C. M., & Wright, J. F. (2016). **Progress and challenges in viral vector manufacturing.** *Human Molecular Genetics*, *25*(R1):R42-52.

Veloo, A. C. M., Elgersma, P. E., Friedrich, A. W., Nagy, E., & van Winkelhoff, A. J. (2014). **The influence of incubation time, sample preparation and exposure to oxygen on the quality of the MALDI-TOF MS spectrum of anaerobic bacteria**. *Clinical Microbiology and Infection*, *20*(12):O1091–O1097.

Vestal, M. L. (2009). **Modern MALDI time-of-flight mass spectrometry**. *Journal of Mass Spectrometry*, *44*(3):303–317.

Volpers, C., & Kochanek, S. (2004). **Adenoviral vectors for gene transfer and therapy**. *The Journal of Gene Medicine*, *6*(S1):S164–S171.

Walker, J. E., Chen, R. X., McGee, J., Nacey, C., Pollard, R. B., Abedi, M., Bauer, G., Nolta, J. A., & Anderson, J. S. (2012). **Generation of an HIV-1-Resistant Immune System with CD34+ Hematopoietic Stem Cells Transduced with a Triple-Combination Anti-HIV Lentiviral Vector**. *Journal of Virology*, *86*(10):5719–5729.

Wheeler, J.X., Jones, C., Thorpe, R. & Zhao, Y. (2007). **Proteomics Analysis of Cellular Components in Lentiviral Vector Production Using Gel-LC-MS/MS**. *Proteomics Clinical Application* 1(2):224-30.

Williams, T. L., Andrzejewski, D., Lay, J. O., & Musser, S. M. (2003). **Experimental factors affecting the quality and reproducibility of MALDI TOF mass spectra obtained from whole bacteria cells**. *Journal of the American Society for Mass Spectrometry*, *14*(4):342–351.

Wilson, J. M. (2005). **Gendicine: The First Commercial Gene Therapy Product; Chinese Translation of Editorial**. *Human Gene Therapy*, *16*(9):1014–1015.

Wirth, T., Samaranayake, H., Pikkarainen, J., Määttä, A.-M., & Ylä-Herttuala, S. (2009). **Clinical trials for glioblastoma multiforme using adenoviral vectors.** *Current*

Opinion in Molecular Therapeutics, *11*(5):485–92.

Witting, S. R., Li, L.-H., Jasti, A., Allen, C., Cornetta, K., Brady, J., Shivakumar, R., & Peshwa, M. V. (2012). **Efficient large volume lentiviral vector production using flow electroporation.** *Human Gene Therapy*, *23*(2):243–9.

Wold, H. (1966, January 1). **Estimation of principal components and related models by iterative least squares**.

Wold, S., Esbensen, K., & Geladi, P. (1987). **Principal component analysis**. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3):37–52.

Wold, S., Kettaneh-Wold, N., MacGregor, J. F., & Dunn, K. G. (2009). **Batch Process Modeling and MSPC**. *Comprehensive Chemometrics*, 163–197.

Wold, S., Sjöström, M., & Eriksson, L. (2001). **PLS-regression: a basic tool of chemometrics**. *Chemometrics and Intelligent Laboratory Systems*, *58*(2):109–130.

Wu, X., Li, Y., Crise, B., & Burgess, S. M. (2003). **Transcription Start Regions in the Human Genome Are Favored Targets for MLV Integration**. *Science*, *300*(5626):1749–1751.

Zhang, X., Scalf, M., Berggren, T. W., Westphall, M. S., & Smith, L. M. (2006). **Identification of mammalian cell lines using MALDI-TOF and LC-ESI-MS/MS mass spectrometry**. *Journal of the American Society for Mass Spectrometry*, *17*(4):490–499.

Zielske, S. P., Reese, J. S., Lingas, K. T., Donze, J. R., & Gerson, S. L. (2003). **In vivo selection of MGMT(P140K) lentivirus–transduced human NOD/SCID repopulating cells without pretransplant irradiation conditioning**. *Journal of Clinical Investigation*, *112*(10):1561–1570.

Zufferey, R. (2002). **Production of lentiviral vectors.** *Current Topics in Microbiology and Immunology*, *261*:107–21.

Zufferey, R., Dull, T., Mandel, R. J., Bukovsky, A., Quiroz, D., Naldini, L., & Trono, D. (1998). **Self-inactivating lentivirus vector for safe and efficient in vivo gene delivery.** *Journal of Virology*, *72*(12):9873–80.

Appendix 1

Variable	Type
Working volume	setpoint
cell line	setpoint
DO2 setpoint	setpoint
Temp setpoint	setpoint
pH setpoint	setpoint
agitation [rpm]	setpoint
Tip speed	derived
P/V	derived
pH post INOC	online
pH post TFX	online
pH post IND	online
pH at HRV	online
pH at HRV48	online
pCO2 post INOC	offline
pCO2 post TFX	offline
pCO2 post IND	offline
pCO2 at HRV	offline
pCO2 at HRV48	offline
pO2 post INOC	offline
pO2 post TFX	offline
pO2 post IND	offline
pO2 at HRV	offline
pO2 at HRV48	offline
lactate post INOC	offline
lactate post TFX	offline
lactate post IND	offline
lactate post HRV	offline
lactate post HRV48	offline
glucose post INOC	offline

Variable	Type
glucose post TFX	offline
glucose post IND	offline
glucose post HRV	offline
glucose post HRV48	offline
glutamine post INOC	offline
glutamine post TFX	offline
glutamine post IND	offline
glutamine post HRV	offline
glutamine post HRV48	offline
glutamate post INOC	offline
glutamate post TFX	offline
glutamate post IND	offline
glutamate post HRV	offline
glutamate post HRV48	offline
VCN post INOC	offline
VCN post TFX	offline
VCN post IND	offline
VCN post HRV	offline
VCN post HRV48	offline
Viability post INOC	offline
Viability post TFX	offline
Viability post IND	offline
Viability post HRV	offline
Viability post HRV48	offline
FACS at HRV	analysis
FACS at HRV48	analysis
RNA at HRV	analysis
RNA at HRV48	analysis

Table 11: Full list of variables initially considered for statistical analysis of process development for suspension-based LVV production in Chapter 2.2.2.
Time points refer to: INOC – Final inoculation; TFX – Transfection; IND – Sodium Butyrate addition; HRV – first harvest; HRV48 – second harvest as in the process flow diagram (Figure 7);
Types refer to: setpoint – process setpoint, determined by operator; derived – value calculated from other parameters; online – parameter measured as part of process monitoring, using bioreactor sensors; offline - parameters measured after sampling, using standalone equipment; analysis – variable obtained from an analytical assay after process is finished