The
University
Of
Sheffield.

Department of Chemical and Biological Engineering

University of Sheffield

Thesis Submitted for the Degree of Doctor of Philosophy (PhD)

CHO Cell Genetic Instability: From Transfection to Stable Cell Line

By:

Joseph Cartwright

February 2016

Declaration

I, Joseph Cartwright, declare that I am the sole author of this thesis and that the results presented within are a product of my own efforts and achievements. Where this is not the case, it has been clearly stated. The work within this thesis has not been previously submitted for any other degrees.

Table of Contents

Acknowledgements

I would firstly like to thank my supervisor, Professor David James, for his guidance and support throughout the project and for giving me the opportunity to study for my PhD. I would also like to thank Pfizer for funding the project and for providing support throughout. In particular, I would like to thank Kurt Droms and Karin Anderson for their advice and guidance. I am also grateful to the EPSRC for funding the project.

A special mention has to go to all past and present PhD students, post-docs and technicians of the James lab who were always a great support, both professionally and personally. These thanks are extended to all the staff and PhD students in the department. In particular I would like to thank Dr. Joseph Longworth, Dr. Ben Thompson, and Philip Lobb for project help and guidance. Thank you to Dave Wengraf, the unsung hero of the James lab. To Darren Geoghegan and Katie Syddall who were in it with me from the beginning, the experience would not have been the same without you.

A big thank you to all my friends and family, who have been supportive, encouraging and patient during my PhD. I was never short of people to go for a beer with when I needed it, or someone to give me a bit of perspective when I was agonising over a bad day in the lab.

Finally, thank you to Sam, for everything.

List of Figures

List of Tables

List of Abbreviations

>1 filter	Mutations filtered occurring more than once
2FI	2-factor interaction
320-26	Electroporation using 320 V, 26 ms, exponential decay waveform
A	Adenine
ACD	Average Cell Diameter
Add	Additional material of unknown origin
Amp	Ampicillin
ANOVA	Analysis of Variance
ASCII	American Standard Code for Information Interchange
BHK	Baby Hamster Kidney
BLASR	Basic Local Alignment with Successive Refinement
C	Cytosine
C-GFP	C-terminal GFP fusion
CCD	Central Composite Design
CCS	Ciruclar Consensus Sequencing
CD-CHO	Chemically defined CHO cell Media
CHO	Chinese Hamster Ovary
CHO-S	Commercially available suspension-adapted cell line
CHO269M	Pfizer CHOK1SV Cell line
CHOK1SV	Suspension Variant from CHOK1 Parental Cell Line
CMV	Cytomegalovirus
CpG	C – phosphate – G
CsCl	Caesium Chloride
Der	Derived Chromosome
DHFR	Dihydrofolate Reductase
diH$_2$O	Deionised water
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic acid
DoE	Design of Experiments
E. coli	Escherichia coli
EDTA	Ethylenediaminetetraacetic acid

FACS	Fluorescence-activated cell sorting
Fc	Fragment Crystalisable
FDA	Food and Drug Administration
FLP-FRT	Flippase/FLP recombination target
FSC	Forward Scatter
FSR	Fusion stable reporter
G	Guanine
GCN	Gene Copy Number
GFP	Green fluorescent protein
GS	Glutamine Synthetase
HC	Heavy Chain
HCl	Hydrochloric Acid
HEK	Human Embryonic Kidney
HSV	Herpes Simplex Virus
HT-supplement	Sodium Hypoxanthine and Thymidine supplement
IgG_2	Immunoglobulin G dimer
Indel	Insertion / Deletion
Iso	Iso Chromosome
Kan	Kanamycin
LB	Lysogeny Broth
LC	Light Chain
mAb	Monoclonal Antibody
Mar	Marker Chromosome
MCS	Multiple Cloning Site
MFU	Median Fluorescence Unit
MMR	Mismatch repair
mRNA	Messenger RNA
MSS	Model Summary Statistics
MSX	Methionine Sulfoximine
MTX	Methotrexate
Neo	Neomycin
NGS	Next Generation Sequencing

NS	Nearly-Stable
NS0	Non-secreting myeloma cell line
NTP	Nucleoside Triphosphate
OFAT	One factor at a time
ORF	Open reading frame
OriP	Origin of DNA replication (mammalian)
P. pastoris	*Pichia pastoris*
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PEI	Polyethylenimine
pH	Potential Hydrogen
PMT	Photomultiplying Tube
Poly(A)	Polyadenylation
PRESS	Predicted Residual Sum of Squares
PTM	Post-translational Modification
pUC Ori	Origin of DNA replication (bacterial)
Q filter	Quality score filter
qP	Cell specific productivity
QV / Q score	Quality Value / Quality Score
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
ROI	Read of Insert
RSM	Response Surface Model
S	Stable
S. cerevisiae	*Saccharomyces cerevisiae*
SAM	Sequence Alignment/Map
SAM	Sequence Alignment/Map
SDS	Sodium Dodecyl Sulfate
SGE	Stable Gene Expression
SMRT	Single Molecule Real-time
SMSS	Sequential Model Sum of Squares
SNP	Single Nucleotide Polymorphism
SS	Semi-Stable

SSC	Side Scatter
SV40	Simian Vacuolating virus 40
T	Thymine
TAE	Tris-acetate-EDTA
TE	Tris-EDTA
TGE	Transient Gene Expression
tPA	Tissue plasminogen activator
Tris	Trisaminomethane
tRNA	Transfer RNA
UPR	Unfolded Protein Response
UV	Ultraviolet
VCD	Viable cell density
z	Z group chromosome
ZMW	Zero-mode Waveguide
τ	Time Constant

Abstract

Chinese hamster ovary (CHO) cells are the predominant host cell type used in the production of recombinant therapeutic proteins. They are chosen as hosts, because of their ability to create, fold and modify proteins in a manner that makes them compatible with the human immune system. Moreover, CHO cells are tried and tested model organisms for bioprocess platforms, meaning regulatory body approval for new therapeutics is relatively easy to achieve. CHO cells are inherently genetically unstable, which can lead to a decline in productivity and poses a threat to product quality heterogeneity of stable cell lines. The primary aim of this thesis was to characterise genomic instability of a CHOK1SV cell line and measure directly the impact this genetic instability has on the fidelity of recombinant plasmid copies. The impact of this would be two-fold: Firstly, an accurate quantification of genetic instability type and frequency would be established. Secondly, the techniques used to characterise genetic instability would be evaluated as tools for the detection of instability in cell line development processes.

Microsatellite analysis and karyotype analysis were used to assess CHO cell genomic instability at the base pair / gene copy number (GCN) level and the chromosome level respectively. Microsatellites were found to be effective markers for genetic drift and cell line relatedness. However, there was no substantial evidence of microsatellite mutational change, and so it could not be concluded that microsatellites are an effective marker for deficient DNA replication / DNA damage or mismatch repair. Microsatellite change did not correlate with changes in GCN or cell specific productivity (qP). There was substantial evidence of chromosomal aberration from Karyotype analysis, which showed considerable levels of aneuploidy and chromosome breakage/fusion events. It was concluded that CHO cells have an inherent chromosomal instability and that karyotyping is a promising tool for genetic instability cell line development assessments. However, there was no substantial association found between changes in CHO karyotype and changes in qP or GCN.

In order to generate a stable GFP cell line for the investigation of recombinant plasmid genetic instability it was necessary to optimise an electroporation protocol. Preliminary experiments indicated that standard industry conditions were suboptimal and so a Design of Experiments (DoE) – based strategy was used to optimise electroporation. Final optimal conditions (termed 320-26) improved transfection efficiency by 17%.

The final results chapter outlines a novel single-molecule real time (SMRT) sequencing analysis platform, which maximises the sensitivity of the technology, enabling mutation calling from individual molecules at a 0.01% frequency. One mutation was present at high levels throughout the study, a C → T transition in the bacterial origin of replication, which is assumed to have originated from the original plasmid stock. There was no evidence of mutations arising in plasmid cloning or as a result of the pre-integration CHO cell environment. Substantial levels of point mutation were found in recombinant plasmid copies. Mutations were randomly distributed along the length of the plasmid and were apparently not influenced by natural selection. G and C residues were mutated to a greater extent than A and T residues, with G.C → A.T transitions predominating. This final assessment of CHO cell genetic instability shows the requirement for product quality checks during cell line development.

This page is intentionally left blank

Chapter 1

Introduction

This chapter will present the wider subject knowledge surrounding the work presented in this thesis in order to provide context and reason for it. A summary of biopharmaceutical industry development, production processes and example achievements are given to highlight how advances have been made, processes have been optimised and some of the areas in which processes can still be improved upon. The chapter is written to broadly introduce the biopharmaceutical industry with a skewed focus towards the concepts investigated and discussed in this thesis and outlines how advancement of these areas could lead to the production of better drugs, more quickly and cheaply. A brief review of the more specific material surrounding each chapter will be presented in more detail at the start of each chapter.

1.1. The Biopharmaceutical Industry

Biological sources have long been exploited for therapeutic use, such as the use of the smallpox virus by Edward Jenner in 1796 to combat cowpox, which established vaccination therapy as a medical treatment (Baxby, 1999); the serendipitous discovery of penicillin in Staphylococcus by Alexander Fleming in 1928 marking the advent of antibiotic medicine (Ligon, 2004); the therapeutic potential of naturally occurring proteins such as insulin and antibodies (Walsh, 2000). These biological sources have been shaped by millions of years of evolution, and harnessing them can offer a novelty and high degree of specificity to medical treatment.

Cell cultivation methods have been developed over the last century to such an extent that they can be used as production factories for these biologics. The creation of permanent and immortal cell lines, which are able to be grown and phenotypically manipulated in sub-culture has enabled the progression of large-scale industrial bioprocesses (Kretzmer, 2002). The development of mammalian cell culture on this scale was largely driven by the need for human viral vaccines in the 1950s and has continued to be the primary cell type used for the production of biological therapeutics. This is because the specific protein folding and modification systems they employ are compatible with human cellular components and immune system (Butler, 2005, Dinnis and James, 2005).

Initially only products native to cell type could be produced, so just a small range of usable molecules were obtainable and only at the low concentrations yielded naturally. Therefore, only a limited number of therapies could be established (Kretzmer, 2002; Walsh, 2000). However, during the early 1970’s techniques were developed to covalently link DNA molecules regardless of their base-pair sequence, giving rise to recombinant DNA technology. Insertion of target DNA into mammalian cell hosts became possible, facilitating the linkage of exogenous and endogenous DNA within the cell (Lobban and Kaiser, 1973, Kretzmer, 2002). Moreover, the fusion of continuously proliferating myeloma cells with antibody producing lymphocytes gave rise to hybridoma cells capable of both continuous proliferation and antibody production (Kretzmer, 2002, Kohler and Milstein, 1975). Through genetic engineering or fusion, using specific antibody-producing lymphocytes, many more proteins could be produced and on a larger scale, which meant that recombinant therapeutic proteins found greater medical application. The first recombinant therapeutic protein to be made available from recombinant DNA technology was human insulin (Humulin, Genentech) for diabetes treatment in 1982, produced in *Escherichia coli* (*E. coli*). However, many therapeutic proteins have a higher, cell type-specific, structural and molecular complexity than insulin, and so need to be cultivated within a mammalian host; the first of these products was tissue plasminogen activator (tPA) in 1987, which is an anticoagulant primarily used in the treatment of heart attack and stroke (Butler, 2005, Kretzmer, 2002, Pineda et al., 2012).

Furthermore, engineering strategies have enabled proteins to be refined by modification, which led to the production of more therapeutically efficient products. For example, changes to the sequence of insulin stopped the interaction of insulin molecules with each other, thus creating a faster acting and more efficacious product (Kretzmer, 2002, Walsh, 2000, Olsen et al., 1996). Since the development of these technologies and the identification of more biomolecules with potential therapeutic applications a wider range of biologics have been produced in sufficient quantities to allow their medical application (Walsh, 2000).

The modern definition of a biopharmaceutical is an engineered protein or nucleic acid which can be used for in vivo diagnostic or therapeutic purposes (Walsh, 2002). The biopharmaceutical industry is currently thriving with 212 products on the market (Walsh, 2014). The top ten products in the USA are presented in Table 1.1a. In the USA alone sales in 2012 reached $63.6 billion, which was an 18.2% increase from 2011 (Aggarwal, 2014). This illustrates the scale of growth in this industry. The major targets of these therapeutic products are cancer, infectious diseases, autoimmune disorders and cardiovascular disease (Walsh, 2005). A wide range of therapeutic molecules (Table 1.1b) are used, the five most common being monoclonal antibodies (mAbs), hormones, growth factors and fusion proteins and cytokines. In particular, monoclonal antibodies, which generate $24.6 billion in US sales (approximately 39% of total biopharmaceutical sales), dominate the biopharmaceutical market (Aggarwal, 2014, Dinnis and James, 2005).

A		B		C	
Product	Sales ($ Billions)	Product Type	Sales ($ Billions)	Company	Sales ($ Billions)
Humira	4.6	mAb	24.6	Roche	13.2
Lantus	4.51	Hormones	16.1	Amgen	12.9
Enbrel	3.9	Growth Factors	8.1	Sanofi	5.1
Remicade	3.6	Fusion Proteins	5.8	Novo Nordisk	4.9
Rituxan	3.5	Cytokines	4.9	J&J	4.7
Neulastsa	3.5	Therapeutic Enzymes	1.4	Abbott	4.6
Novolog	2.97	Blood Factors	1.2	Biogen Idec	3.9
Avastin	2.8	Recombinant Vaccines	1.1	Lilly	3.6
Humalog	2.08	Anti-coagulants	0.4	BMS	1.7
Herceptin	1.9			Merck	0.9

Table 1.1. Biopharmaceutical Sales of Top Selling Products
Therapeutics are given in terms of therapeutic names (a), product types (b) and biotech companies (c). (Adapted from Aggarwal, 2014)

Nearly 50% of new biopharmaceutical products being approved are biosimilars (Walsh, 2010), which are alternative versions of already existing products. When patents on biopharmaceutical products expire, competing biotech companies (top ten – Table 1.1c) are permitted to create their own version of a product. In some cases drugs are engineered to be more efficient than the original and can often be produced more cheaply. These drugs are called biobetters (Barbosa, 2011). Furthermore, the release of this information can advance general understanding and lead to the discovery of novel products (Covic and Kuhlmann, 2007, Mellstedt et al., 2008). The first of these products was Omnitrope (Sandoz), a biosimilar of the human growth hormone somatroptin (Moran, 2008).

Biological therapeutics, such as mAbs, have aided the treatment of a large number of conditions and had a positive impact on the quality of life of many patients. Clearly, there is a high demand to make therapeutic proteins cheaper, more efficient and of high quality to ensure that success in treatment can continue to be improved upon and become as widespread as possible (Dinnis and James, 2005, Shukla and Thommes, 2010).

1.2. Recombinant Protein Expression: Expression Systems

The production of biologics by biopharmaceutical companies is governed by certain aspects of the production process, such as cost-effectiveness, efficacy, effectiveness, time to market and safety, amongst others. Therefore it is important to use expression systems flexible enough to provide a manufacturing platform capable of fulfilling all of these criteria for multiple biologics at an individual level (Ferrer-Miralles et al., 2009, Li et al., 2010). Due to the large variation in recombinant proteins with potential therapeutic functions and the additional complexity of protein folding and post-translational modifications (PTMs), it is unlikely that there will be a naturally occurring expression system capable of making all biologics. Different expression systems are metabolically diverse from one another. Therefore particular expression systems are better adapted for particular applications (Andersen and Krummen, 2002, Ferrer-Miralles et al., 2009). The cell types harnessed for biopharmaceutical production show great amenability to a range of culture conditions and desirable phenotypes, through both adaptive evolution and engineering techniques. This enables the production of a vast amount of biopharmaceuticals from a single organism (Mohan et al., 2008, Davies et al., 2013).

1.2.1. Non-mammalian Systems and Important Characteristics.

Prokaryotes have been utilised as biologic expression platforms for many applications, such as the production of Humulin by *E. coli*. Much of our initial understanding of molecular biology was centered around *E. coli,* so it is extremely well characterised. Therefore, our understanding of molecular genetics and the development of genetic tools for engineering were established in a prokaryotic background and so generating an engineered production organism is relatively straightforward. Moreover, it is easy to rapidly culture bacteria and produce large yields of recombinant product. Simple molecules such as hormones, interferons and interleukins are amongst the approved therapeutic products synthesised by *E. coli* (Ferrer-Miralles et al., 2009). However, generally, their ability to produce complex humanised proteins is limited, because they naturally process proteins differently to a eukaryotic cell and so lack the ability to carry out complex eukaryotic processes. A humanised protein must be folded in the correct conformation and attain the correct PTMs, such as acetylation, carboxylation, amidation, glycosylation and phosphorylation. Such modifications affect the efficacy of a protein through properties such as specificity, stability and activity (Walsh and Jefferis, 2006). The differences between proteins produced by prokaryotes and eukaryotes is enough to cause an immunogenic reaction when a potential therapy is administered, because the immune system would likely recognise these differences and elicit an immune response (Ferrer-Miralles et al., 2009).

Glycosylation is the most influential PTM in terms of therapeutic specificity, because it is the most commonly found PTM in eukaryotic organisms, with over 50% of all human proteins being glycosylated (Walsh and Jefferis, 2006). Protein glycosylation affects protein folding, secretion, degradation, cell signaling, immune function and transcription, so is likely to have a significant impact on a proteins therapeutic function. The potential variation in glycosylation profiles makes it a more varied and consequently more complicated attribute than the proteome itself, which means that each organism's glycosylation profile can be extremely specific (Lauc et al., 2010). Therefore, it is essential to make sure a host expression system is capable of producing a recombinant protein with a glycosylation profile compatible with humans so it does not provoke an immune response (Ferrer-Miralles et al., 2009). Protein glycosylation pathways do exist in prokaryotes, and these can be engineered into, and implemented, in an *E. coli* system. However, there are distinct differences between this form of glycosylation and that which occurs in a mammalian system. If prokaryotes could be engineered to produce humanized glycosylation forms then they would likely come to the fore in biopharmaceutical production (Abu-Qarn et al., 2008, Valderrama-Rincon et al., 2012).

Therefore, for the time being, eukaryotes are better candidates for the production of therapeutic proteins, especially complex ones, because their metabolism allows them to produce these proteins with the correct specificity in structure and PTMs so not to elicit an immune response (Walsh and Jefferis, 2006, Ferrer-Miralles et al., 2009, Andersen and Krummen, 2002). The eukaryotic production systems able to carry out the protein folding and PTMs needed to produce humanised proteins are yeast, insect, plants and mammalian cells (Walsh, 2006). Plants can be utilised as production vehicles for recombinant proteins both in the form of transgenic plants and plant cell culture. Commercially, plants have been able to successfully produce animal proteins. Recombinant plant technology offers high yields, low cost, low chance of pathogen contamination and the protein can be produced in storage organs such as seeds to ease purification (Sharp and Doran, 2001, Giddings et al., 2000). However plant-based recombinant technology is less developed than other expression systems and attaining regulatory approval for engineered plants is a challenge. Until a robust, tested and trusted infrastructure is in place it is unlikely that plants will challenge mammalian cells as a production platform (Hellwig et al., 2004, Fischer et al., 2012).

Yeasts, like plants, are able to produce high yields of recombinant protein at a low cost. Furthermore, like *E. coli* they exhibit quick growth and are extremely well characterised and understood, because they formed the basis of our understanding of the eukaryotic cell cycle, amongst other processes. The two most utilised strains for recombinant protein production are *Saccharomyces cerevisiae* (*S. cerevisiae*) and *Pichia pastoris* (*P. pastoris*) (Demain and Vaishnav, 2009). The glycan structure in mammalian and yeast cells is the same as it arrives at the Golgi. However, the mammalian Golgi elicits various trimming and extension reactions, resulting in a sialylated glycan structure. On the other hand, rather than trim, yeast adds further mannose groups thus resulting in recombinant protein unsuitable for therapeutic use. Despite this, *P. pastoris* is a promising expression system. Through a series of engineering strategies it is capable of producing proteins with humanized glycosylation profiles. This along with its good growth characteristics and protein secretory mechanisms makes *P. pastoris* a capable production system. It has already been successfully engineered to produce proteins such as insulin precursor, interleukin 2 and tumour necrosis factor amongst others (Macauley-Patrick et al., 2005, Demain and Vaishnav, 2009, Berlec and Strukelj, 2013, Hamilton and Gerngross, 2007).

Whilst these expression systems have all shown promise they are not yet producing to the same quality or quantity as the industry standard of mammalian cells (Dinnis and James, 2005). Non-mammalian cells are more likely to stimulate an immune response, because of their lack of specificity in PTMs (Raju, 2003). For example, plants consistently add α1,3-fucose and β1,3-xylose sugars, which elicit immunogenic responses in humans (Walsh and Jefferis, 2006). Furthermore, there needs to be further development and understanding before these alternative expression systems could offer a potential replacement to mammalian cells. For example, *P. pastoris,* which is arguably the best non-mammalian production system, still needs a great deal of process optimisation. Yields produced are still three to five-fold less than the gold standard of mammalian cell systems and the heterogeneity and stability of glycosylation is still something that needs to be proven in its consistency. However, it is believed that yeast systems will reach these standards, the confidence of which is reinforced by the endorsement of the technology by Merck & Co by taking over Glycofi technology in 2006 (Beck et al., 2010). Although these technologies show promise, it is mammalian systems that predominate the production of humanised therapeutic proteins, despite being expensive and slow in comparison to alternative systems (Demain and Vaishnav, 2009). Moreover, it is likely that process outputs would need to show considerable improvements for companies to consider the replacement of the mammalian systems for which the industry has been moulded upon.

1.2.2. Mammalian Expression systems

Mammalian cells currently dominate the biopharmaceutical market with 60-70% of recombinant therapeutic proteins being produced by mammalian cell culture. To put this into context, biopharmaceutical sales currently constitute 27% of total drug sales and are growing at a rate 7-fold higher than the pharmaceutical sales overall (Wurm, 2004, O'Callaghan and James, 2008, Walsh, 2014, Aggarwal, 2014). Therefore mammalian cell culture is a hugely important platform in the drug market. As described previously this is largely due to their ability to correctly fold and assemble large, complex molecules and carry out the appropriate PTMs to make a protein suitable for therapeutic application in humans both in terms of their therapeutic activity and safety. Also, as higher eukaryotes, mammalian cells are able to recognise secretion signal sequences in the recombinant gene and the mammalian cell machinery is able to mediate the successful secretion of the recombinant gene product (Barnes et al., 2000, Page, 1988). A great amount of research and development has, and continues, to be carried out on mammalian cell culture, cell biology and cell engineering. There are a variety of mammalian cell types currently being used and developed for recombinant protein production, including Chinese Hamster Ovary (CHO), Mouse Myeloma (NS0), Baby Hamster Kidney (BHK) and Human Embryonic Kidney (HEK-293) (Wurm, 2004). The choice of cell line is largely down to its ease of large-scale culture, high growth rates, cell specific productivity (qP), titers and their ability to produce a efficacious and safe product (O'Callaghan and James, 2008). CHO expression systems are the most widely used, which is due not only to their protein folding and PTMs, but also to their ability to be cultured quickly and robustly on a large scale, their simplicity in transfection and recombinant gene integration, and their ease of product approval by the FDA (Jayapal et al., 2007, Wurm and Hacker, 2011, Wurm, 2004).

Cell line engineering is an area of research and development that has resulted in process improvements in the manufacturing of biological therapeutics in mammalian cells in terms of increased recombinant gene expression, product quality and cell attribute improvement. For example increased sialylation was achieved by increased expression of sialyl transferase and the production of non-fucosylated products by creating FUT8 knockout cell lines. These changes have led to the production of specific and more efficacious proteins, increasing their therapeutic potential (Zhu, 2012, Bork et al., 2009, Shields et al., 2002, Iida et al., 2006, Wong et al., 2010). In another example, engineering against late cell culture conditions that can induce apoptosis, such as nutrient and oxygen depletion and the accumulation of harmful bi-products, was achieved by overexpression of Bcl family members and E1B-19K. This significantly increased mAb productivity and created cells that are more robust to these conditions (Dorai et al., 2010, Dinnis and James, 2005, Zhu, 2012). In a further example, engineering strategies have been used to improve production of difficult to express proteins. This could serve to increase the number of proteins with therapeutic potential finding commercial application. Pybus et al. (2014) varied the mAb LC: HC ratio and the expression of foldases, chaperones and unfolded protein response (UPR) transactivators to subvert UPR induction, thus increasing mAb productivity in a product specific manner. In general, as the understanding of the mechanistic processes of the cell increases more engineering targets can be identified and researched (Dinnis and James, 2005).

1.3. Recombinant Protein Expression: The Process in Mammalian Cells

1.3.1. Stable Gene Expression

Stable gene expression (SGE) is the term used to describe permanent expression of a recombinant protein by a cell host. This is achieved by introducing a plasmid DNA vector containing the recombinant gene of interest, which facilitates its integration into the host organism's genome. Subsequently, highly expressing clonal populations are generated, expanded and used for screening and further analysis (Makrides, 1999).

There is a very well established platform (Figure 1.1) for the stable expression and subsequent production of recombinant proteins in mammalian cells (Jayapal et al., 2007, Wurm, 2004). After a potential therapeutic product has been identified its DNA sequence is determined and the gene of interest is inserted into a plasmid along with a selection gene, which will give recombinant cells a survival advantage to ensure their propagation. Carefully optimised genetic regulatory elements are included to govern gene expression. Mammalian cells are transfected with multiple copies of this plasmid and a small number of these will integrate into the mammalian host genome (Wurm, 2004, Li et al., 2010). After a brief recovery period a selection agent is administered to the cells so that only cells with the selection gene, and subsequently the recombinant gene, survive. Therefore non-recombinants are gradually removed from the population. After selection the cell population will consist of a heterogeneous stable pool of expressing cells. Clonal populations are made by isolating single cell survivors, which are cultivated and expanded into, theoretically, homogenous cell lines. The clonal cell lines are tested for attributes desirable in a recombinant protein-expressing cell line. These attributes include high qP, growth characteristics in shaking flask and bioreactor conditions, and product quality. Eventually one cell line is taken forward for long term large-scale production and a cell bank is generated and frozen for future use. This process can take more than 6 months. Despite the success of this platform research, development and optimisation continue to improve this process (Wurm, 2004, Jayapal et al., 2007, Li et al., 2010, Kim et al., 2012, Birch and Racher, 2006). During this process clonal cell lines are expanded over multiple passages and go on to be cultured in laboratory-scale bioreactors to assess and optimise growth in these conditions.

Eventually, the cultures are scaled up to industrial bioreactor size for commercial production (Jayapal et al., 2007).

This thesis focuses on upstream processes. However, for completeness, a brief summary of downstream processes is given here. It is important that optimised upstream methods are followed by efficient extraction and purification of recombinant proteins from cell culture. A large proportion of yield can be gained or lost by effective downstream methods and as a result have a large impact on manufacturing costs (Shukla et al., 2007). The final product must be free from any impurities of the cell, bioreactor or the purification procedure itself. These impurities include protein A, media components, DNA, host cell protein, viruses and endotoxins (Shukla et al., 2007, Kelley, 2007). The downstream process will differ with each product, but there is a common industrial approach used. Briefly, the standard platform for mAb purification is as follows: Cells and cell debris are removed through centrifugation and depth filtration, which is referred to as cell culture harvesting. After this the mAb is captured directly by protein A affinity chromatography, binding specifically to the Fc region of the antibody and removes cell impurities such as DNA and host cell protein. This provides more than 98% purity in a single step and is responsible for a large reduction in volume. Elution is carried out using low pH, serving as a viral inactivation step. The solution is neutralised before the polishing steps. Polishing typically consists of ion-exchange chromatographic techniques that help remove leftover impurities. After a viral filtration step an ultrafiltration/diafiltration process mediates the transfer of the product into its formulation buffer (Shukla et al., 2007, Shukla and Thommes, 2010).

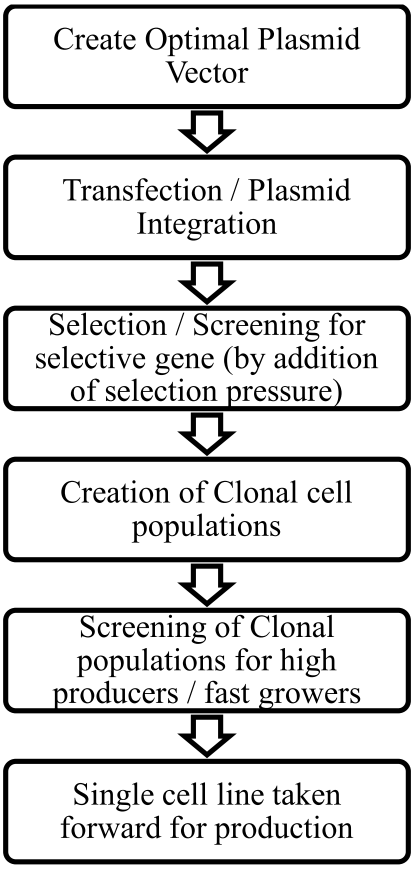

Create Optimal Plasmid Vector
Transfection / Plasmid Integration
Selection / Screening for selective gene (by addition of selection pressure)
Creation of Clonal cell populations
Screening of Clonal populations for high producers / fast growers
Single cell line taken forward for production

Figure 1.1: Stable Cell Line Generation
The figure briefly summarises the stable cell line generation process, as described in the text.

1.3.2. Transient Gene Expression

Transient gene expression (TGE) is an alternative method to generate producing cells, which offers very quick production of small amounts of protein rather than slow production of large amounts of protein. It is the quickest and least expensive way of producing recombinant protein (Wurm, 2004, Makrides, 1999). In TGE the ability of the multiple plasmid copies to produce recombinant protein extrachromosomally is utilised in order to rapidly assess aspects of the production process such as vector design and product efficacy. Therefore, this process is quick because there is no need to screen for successful genome integrations. TGE lasts around 10 days, because expression is rapidly lost when the plasmid copy number becomes diluted due to cell division and lost in line with plasmid half-life (Rita Costa et al., 2010, Barnes et al., 2003, Baldi et al., 2007). Generally, TGE is used for initial analysis and characterisation of the cell line, recombinant protein and the plasmid vector used to express it, so that the process can be reviewed before taking a system and product into long-term cell culture. For example, different combinations of vector elements such as promoters and enhancers can be tested and optimised (Makrides, 1999). Process evaluation using the TGE process can take as little as three days and the parameters, which are evaluated as having been most successful, are taken on to produce long term stable cell lines (Wurm, 2004). TGE is also being developed as a recombinant protein production method in its own right, being able to produce milligram to gram quantities in just a few days via large-scale transfection processes (Derouazi et al., 2004, Wurm, 2004, Zhu, 2012). If it can be done on a larger scale TGE can be used more in product process development meaning that SGE is not needed until the later stages of the bioprocess. This means that development can be done more quickly and is less resource intensive (Steger et al., 2015).

HEK293 and CHO cells are the most commonly used cell lines for TGE, with HEK293 cells having the ability to produce the highest titers of recombinant protein. However even between two mammalian cell types, such as HEK293 and CHO, growth characteristics can differ and varied products can be manufactured from the same construct, because of the specific processing that occurs in an organism (e.g. PTMs). Most manufacturing is done via SGE and ideally the host system should be consistent throughout the production process. Due to the fact that CHO produce the majority of recombinant products via SGE, a great amount of research and development is trying to improve yields from TGE in CHO so that consistency can be maintained for the best producing mammalian cell type. For example, engineered CHO cell lines expressing T antigen and the presence of genetic elements such as OriP or SV40 Ori allows the prolonged episomal presence of the plasmid. Also, culturing cells with DMSO and in hypothermic conditions has helped raise the yields achieved through TGE in CHO cells (Agrawal et al., 2013, Makrides, 1999, Wurm, 2004).

1.3.3. Expression Vector and Selection System

The expression vector primarily used for gene expression in mammalian systems is a DNA plasmid, which exists extrachromosomally and is designed to contain various elements that enhance transcription and translation (Wurm, 2004). A plasmid is often linearised before transfection to enhance integration efficiency, but this is not essential, as it will be linearised in the nucleus before integration (Rita Costa et al., 2010, Wurm, 2004). Typically, the plasmid will contain strong promoter and enhancer elements upstream of the gene of interest to drive its high expression. The function of a promoter is to be bound by transcription factors to initiate transcription. One such promoter is the cytomegalovirus (CMV) promoter and is the most widely used to drive strong expression in industrial platform processes. Promoters can be constitutively active, induced or repressed if a finer level of control is required over gene expression. Enhancer sequences are positioned further upstream and influence the activity of the promoter (Rita Costa et al., 2010, Makrides, 1999). Genetic elements are often included to elicit desirable RNA processing and stability. For example the SV40 Poly (A) tail is included to increase RNA stability as well as its role in transcription termination (Rita Costa et al., 2010). Moreover, genes in plasmid vectors do not contain introns like a regular gene would, but one is usually inserted to ensure transport of the mRNA from the nucleus to the cytoplasm, increasing the rate of translation. This is because during pre-mRNA splicing into mRNA the exon junction complex is added, which is thought to enhance the mRNA's transport from the nucleus into the cytoplasm (Tange et al., 2004, Wurm, 2004). Different organisms have a specific tRNA pool in their cells, which means that some anticodons are more common than others. Therefore to optimise translation gene sequences are usually codon-optimised to enhance protein production (Wurm, 2004). Sequences will also be altered so they do not contain cryptic poly(A) tails or splice sites and will be carefully assembled in such a way to avoid unwanted RNA folding (Birch and Racher, 2006). The expression of the selection gene will be driven by a weak promoter to make the selection process more stringent. Thus any given cell will need to contain recombinant plasmid copies capable of high expression. One example of this is the SV40 promoter (Kim et al., 2011, Rita Costa et al., 2010). The plasmid also contains bacterial elements such as antibiotic resistance gene and origin of replication for plasmid replication in a bacterial host prior to mammalian cell introduction (Birch and Racher, 2006, Rita Costa et al., 2010).

In monoclonal antibody assembly the reaction kinetics are greatly influenced by the stoichiometric ratio of the heavy and light chains. To ensure efficient ratios the delivery of each gene can be optimised. Typically, two methods have been used to attempt this: In the first, two plasmids can be used each containing one of the light or heavy chains and the ratio is maintained by proportion of each plasmid going into the cell via co-transfection. The problem with this method is that the plasmids will integrate into different genomic regions that are capable of different levels of gene expression. This is known as the position effect and will be discussed later in this section. Therefore, gene delivery of optimal gene ratios does not necessarily lead to optimal ratios of gene expression (Wurm, 2004, Rita Costa et al., 2010). The second method uses a single vector containing both genes, which can be under the same promoter or promoters with slightly different expression capabilities to try and encourage an optimal ratio for any given antibody. The problem with this method is that there is not yet a diverse enough range of readily available promoters for use in these vectors. However, recent studies have identified a range of synthetic promoters that could offer bespoke stoichiometric gene expression for any given protein (Brown et al., 2014). Moreover, the plasmid sequence often undergoes recombination resulting in gene loss. The choice of method differs in different systems and with different products (Rita Costa et al., 2010, Kim et al., 2011).

A common system used industrially utilises the glutamine synthetase (GS)- vector (Figure 1.4), in which the gene for the GS enzyme is used as the selection gene (Barnes et al., 2000). For monoclonal antibody production this plasmid contains codon-optimised genes for the antibody light and heavy chains and for GS. It also contains the strong viral CMV promoter upstream of the light and heavy chains and the weaker SV40 promoter upstream of the GS gene. Poly(A) tails are positioned downstream of each gene and introns are included to facilitate mRNA processing as mentioned previously. The β-lactamase ampicillin resistance gene (Amp) and the bacterial origin of replication are included for selection and replication in bacteria prior to transfection (Kim et al., 2011, Brown et al., 1992). GS is an enzyme which catalyses the production of glutamine from glutamate and ammonia and this is the only enzyme capable of glutamine synthesis in the cell. Therefore, cells cultured in media lacking glutamine will have more efficient growth when they have obtained the plasmid vector sequence (Jun et al., 2006). NS0 cells are often used with this system because they do not produce glutamine. CHO cells, on the other hand, can produce endogenous glutamine. However, the application of methionine sulphoximine (MSX), an inhibitor of GS, means this system is still applicable to CHO cells. The sequential addition of MSX to the cultured cells steadily increases the cell's need for more GS to overcome MSX's inhibitive effect, which causes the GS gene to be amplified within surviving cell population. Therefore the sequential addition of MSX indirectly amplifies the gene copy number of the mAb genes, which will result in the generation of more cells capable of producing higher amounts of recombinant protein. This is because cells, which do not contain amplified copies of the recombinant construct will not survive. (Jun et al., 2006, Barnes et al., 2001, Brown et al., 1992). More recently, CHO GS-knockout cell lines have been generated in order to prevent endogenous GS production. This removes the reliance upon the selection pressure to generate productive cell lines and has led to shorter process development times and higher levels of production (Fan et al., 2012).

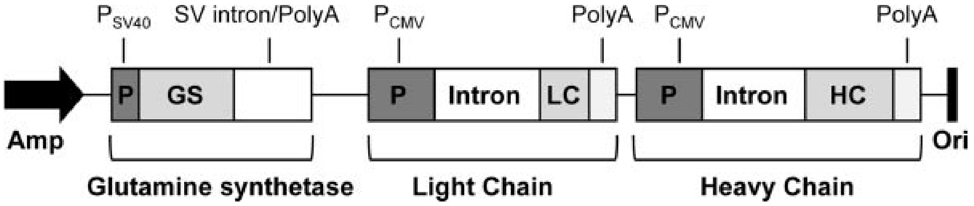


Figure 1.2: GS vector
Each gene contains the coding region, intron and poly(A) tail. The SV40 viral promoter is used for the GS gene, whereas the CMV promoter is used for the light chain (LC) and heavy chain (HC) genes. The ampicillin resistance gene and bacterial origin of replication elements are contained for bacterial selection and replication. (Taken, with permission, from Kim et al., 2011).

Another system utilizes the dihydrofolate reductase (DHFR) gene, which codes for an enzyme involved in nucleotide metabolism. In this case, specific CHO cell lines have been engineered to be DHFR-deficient. In the same way as MSX in the GS system, methotrexate (MTX) concentration in cell culture is sequentially increased in the DHFR system. This inhibits the production of hypoxanthine and thymidine, which are essential to the cell. Therefore, the DHFR and recombinant gene are amplified within the surviving cell population due to this treatment. The GS system is favoured because only one round of amplification is needed, so the process only takes 3 months, whereas the DHFR system, needing multiple rounds of amplification to achieve the required productivity, is a 6 month process (Barnes et al., 2000, Barnes et al., 2001, Jun et al., 2006, Wurm, 2004).

1.3.4. The Position Effect

The site at which plasmid DNA integrates into the mammalian cell genome has a large impact on the expression of the recombinant gene, which subsequently has a large impact on recombinant protein production. This is known as the position effect, which is largely due to epigenetic effects (Wurm, 2004). Epigenetic characteristics are heritable components of an organism's genome, which can change expression, but are not coded by the DNA sequence itself. Gene expression depends on the DNA sequence of a coding regions regulatory elements, such as the promoter and enhancers, their activation and the structure of the chromosomal location at which those DNA sequences are located (Wolffe and Matzke, 1999). Indeed, the structure of chromatin can be open and easily accessible to transcription factors (euchromatin) or condensed by various modifications and bound proteins, leaving it far less accessible to transcription factors (Richards and Elgin, 2002, Mutskov and Felsenfeld, 2004).

Therefore, the location at which a recombinant gene integrates with the host genome can greatly influence its level of expression (Wurm, 2004). There has been successful development of approaches to combat negative position effects (Wurm, 2004). For example boundary elements, such as insulators, are DNA sequences that surround the coding region and prevent interaction with outside effectors of expression, such as enhancers and heterochromatin. Therefore, these coding regions can function as an independent genetic unit within a chromosome (Geyer, 1997). Anti-repressor elements can be used to flank coding regions and stop the spread of heterochromatic features like methylation and hypoacetylation in order to preserve gene expression. Boundary elements such as these have been shown to enable the stable and long term expression of recombinant genes (Kwaks et al., 2003, Wurm, 2004). That being said, integration within heterochromatic regions and certain regions of euchromatin will still cause dampened or no expression (Lattenmayer et al., 2006). Therefore it is widely accepted that the successful development of gene integration targeting methods, whereby the plasmid DNA is targeted to a specific transcriptionally active genomic location (approximately 0.1% of the CHO genome), could make high gene expression more consistent. This could greatly reduce the need for time consuming and expensive selection and screening steps (Lattenmayer et al., 2006, Zhou et al., 2010). Recombinases can be used to recombine sequences inserted into the plasmid with sequences in the genome known to be located in highly expressed areas (Wurm, 2004). Research in this field has shown promise, as shown by Zhou et al. (2010). In this work a reporter plasmid was transfected and single-copy gene expression was selected for in order to generate a cell population with transcriptionally active insertion sites. Next, a second plasmid, containing the gene of interest, was targeted to the initial insertion site using the FLP-FRT system. Successfully targeted integrants contained a second selection gene, reconstituted from sequences from both plasmids. Therefore, selection of desirable integrants could propagate the gene of interest within a transcriptionally active site. Finally, the DHFR system was used for gene amplification to produce high producing clones after very few rounds of amplification (Zhou et al., 2010).

1.3.5. Transfection

Transfection is the general term given to the introduction of nucleic acids into host cells and is used to promote expression of an exogenous product. Transfection can be termed stable or transient depending on whether the non-self DNA is expressed permanently (as described previously) or for a short period of time, respectively. Broadly, transfection methods are categorized into three types: biological, chemical and physical. None of these methods are considered the best for all systems, as each have their advantages in different situations (Wurm, 2004, Kim and Eberwine, 2010).

Biological methods of transfection are carried out through viral delivery. Clearly viruses, by nature, have an evolved inherent ability to introduce foreign DNA to a host and typically this is done with a high transfection efficiency (Kim and Eberwine, 2010, Douglas, 2008). Despite this efficiency, gene delivery has moved away from viral-mediated transfection methods due to safety concerns with viral toxicity, manufacturing limitations and plasmid size constraints (Douglas, 2008, Mehier-Humbert and Guy, 2005).

Chemical methods of transfection rely on the interaction of positively charged chemicals and negatively charged DNA, leading to the formation of DNA-chemical complexes. Examples of this include DNA-Calcium phosphate co-precipitation, cationic lipid complexes, such as with lipofectamine, and cationic polymer based transfection such as with polyethylenimine (PEI). In each case the ratio of DNA to the chemical of choice needs to be optimised for any given system (Douglas, 2008, Rita Costa et al., 2010). These complexes are able to form electrostatic interactions with the cell membrane, possibly with the help of cell surface proteins and other moieties, to enter the cell via endocytosis. The exact mechanism for this is yet to be elucidated. Indeed, the mechanism by which the chemical-DNA complexes leave the endosomes is also yet to be discovered. In the case of lipofection it is thought the complexes bind or destabilize the membrane in order for translocation to take place (Rita Costa et al., 2010, Rehman et al., 2013). Whereas, it is thought that PEI soaks up protons within the endosome due to its large buffering capacity, leading to an increase in endosomal pH. This causes an osmotic swelling of the endosome due to the rapid influx of protons and chloride ions, which subsequently causes it to burst and release the PEI-DNA complexes into the cytosol. Moreover, it is postulated that the buffering capacity of PEI protects PEI-DNA complexes once they reach the liposomes by neutralizing the lysosomal compartment. Therefore the nucleases, which are active at a low pH, do not degrade the complexed DNA. It is also believed that PEI may facilitate the entry of DNA vector into the nucleus (Rita Costa et al., 2010, Tait et al., 2004, Akinc and Langer, 2002). The calcium phosphate method is relatively cheap, can be applicable to many cell types and generates cells with high productivity (Rita Costa et al., 2010). However, despite being able to transfect a high plasmid copy number, the efficiency with which this method can create recombinant cell lines is low (0.05-0.1%), even with attempts at increasing efficiency with DMSO. Moreover, it cannot be used in serum-free processes, such as with CHO cells, which is the largest biologic producer. Therefore, the calcium phosphate method it is not as widely used in the biopharmaceutical processes described in this chapter (Rita Costa et al., 2010, Chenuet et al., 2008). Lipofection and PEI mediated transfection are simple techniques, which can be carried out in serum or serum-free conditions with high transfection efficiencies. PEI is the preferred choice for large-scale bioprocesses due to its comparatively low cost (Rita Costa et al., 2010, Rehman et al., 2013, Baldi et al., 2007, Reed et al., 2006).

There are a variety of physical transfection methods used for gene delivery. For example, mechanical methods such as microinjection and particle bombardment have proven useful in single cell and tissue work respectively. However the laborious, costly and low throughput nature of these techniques are amongst the reasons they have not been popularized in the bioprocesses described in this review (Mehier-Humbert and Guy, 2005). Electroporation is a simple and very quick method for gene delivery into the host cell. It involves subjecting the cells to a pulsed electric field in order to disrupt the membrane potential (voltage gradient) across the plasma membrane. As a result, aqueous pores are created and exist transiently in the lipid bilayer, through which plasmid DNA can enter the cell (Canatella et al., 2001, Rita Costa et al., 2010). Electroporation is the transfection methodology used in this thesis and will be discussed in detail in the next section.

Different transfection procedures are typically used for transient gene expression and stable gene expression. Electroporation is a commonly used transfection methodology for stable gene expression due to its ease, cost and potential for high-throughput. DNA-PEI polymers are more commonly used for transient transfection. The main reason for this difference is that electroporation typically transfects DNA into milliliter quantities of cell culture, which can be cloned and scaled up. On the other hand transient gene expression is short-lived, because extrachromosomal plasmid DNA becomes diluted and eventually lost. Therefore, to fulfill high yield needs production must be instantaneously large-scale. PEI mediated transfection can be carried out on a large scale and immediately yield large volumes of transiently producing cells and as a result is predominantly used to fill this niche (Zhu, 2012). However, recent advances in flow electroporation technology, as with the MaxCyte® transfection system, allow for scalable electroporation to take place that offers a closed, sterile and disposable system (Fratantoni et al., 2003). Cells are suspended in a buffer and electroporation can be optimised on a relatively large scale at high efficiencies resulting in gram quantities of antibody (Fratantoni et al., 2004, Steger et al., 2015), which is in line with other leading transient systems (Bandaranayake and Almo, 2014).

1.3.6. Electroporation

Electroporation is a transfection methodology, which uses an electric field pulse(s) to transiently permeabilise the plasma membrane of a cell. This process is utilised in order to introduce molecules such as DNA into the cell (Gehl, 2003). For this to happen the transmembrane potential needs to reach a threshold level, in which it is estimated that the electrical field across the membrane is approximately 10^8 V/m for a standard membrane width of 5 nm. To achieve this the minimum electrical field that needs to be applied is reported to be around 0.2-1V (Chen et al., 2006). When the threshold is reached the structure of the membrane is reconfigured and pores form, through which molecules can travel into the cell. Confirmation of these pores has been achieved by electron microscopy (Chen et al., 2006, Bio-Rad, n.d.). In the case of DNA, loading of the cell occurs through electrophoretic movement rather than by osmosis, because DNA is negatively charged. Therefore, there is a more direct relationship between the intensity of the electric field and the efficiency of DNA transfection than with uncharged molecules (Gehl, 2003, Sukharev et al., 1992). Furthermore, DNA interacts with the plasma membrane and helps facilitate pore formation during electroporation (Spassova et al., 1994, Gehl, 2003, Escoffre et al., 2009). In this study DNA is linearised to promote higher levels of genome integration. Linear DNA has lower transfection efficiencies than circular and supercoiled DNA and so may need stronger optimal electroporation conditions (Schmidt et al., 2004). As stated, the transmembrane potential must be increased for the destabilisation of the membrane to take place. A variety of factors have an impact on this (Gehl, 2003). One of these factors is electrical field strength. This is the measurement of electrical intensity within the electroporation chamber and is affected by the voltage applied to the chamber and the distance between the two electrodes. This is summarised by equation 1.1, where E is electric field strength (V/cm), V is Voltage and d is the distance between electrodes (cm) (Gehl, 2003, Bio-Rad, n.d.).

$$E = V/d \qquad \text{Equation 1.1. Electric Field Strength}$$

Also, cells with different radii have different transmembrane potential thresholds. Larger field strengths are needed to permeabilise smaller cells (Escoffre et al., 2009, Gehl, 2003). The angle between the membrane and the electrode (i.e. the electric field) also affects the dynamics of electroporation. The inside of the cell is negatively charged. Therefore the pole of the cell facing the anode will be permeablised first and to a greater extent, because this is where the transmembrane potential will be exceeded earliest. The pole facing the cathode will be permeabilised second and to a lesser extent. Even though overall permeabilisation is greater at the cell pole facing the anode, DNA enters the cell to a greater extent at the pole facing the cathode due to the direction of electrophoretic forces. The permeabilised area increases in size with higher field strengths and the extent of permeabilisation within this area is determined by the duration, and number of pulses (Gehl, 2003, Escoffre et al., 2009). The temperature at which electroporation is carried out affects the dynamics of the transfection process. A lower temperature may help increase cell viability due to the heating effect caused during electroporation. Moreover, the process by which pores are resealed would be slowed and so DNA, potentially, has longer to enter the cell. However, a higher temperature would facilitate pores to reseal more quickly, which might in turn increase overall cell viability. Moreover, differences in temperature result in differences in conductivity and subsequently sample resistance. Therefore, it is important to use an optimal temperature which strikes a balance between these characteristics (Bio-Rad, n.d.). Although the extent to which these factors impact on electroporation are reasonably well defined, the exact mechanisms by which the membrane is destabilised and DNA traverses the membrane are yet to be fully elucidated (Bio-Rad, n.d., Escoffre et al., 2009).

Typically, there are two waveform types that are used for DNA electroporation: exponential decay and square wave (Jordan et al., 2008, Jordan et al., 2013) (Figure 1.3). In exponential decay electroporation the voltage rapidly increases to a peak and decreases exponentially over time (Equation 1.2.):

$$V_t = V_0 \left[e^{-\left(\frac{t}{RC}\right)}\right]$$

Equation 1.2. Exponential Decay Waveform

Where Vt is voltage at time = t (msec), V_0 is the voltage upon discharge, R is circuit resistance (ohms) and C is circuit capacitance (μF). The time voltage takes to decrease is dependent upon the capacitance and resistance of the circuit (Jordan et al., 2013, Bio-Rad, n.d., Jordan et al., 2007). The total resistance is a product of the resistance of the electroporation system being used and the resistance of the sample. The sample resistance is affected by a number of factors. Essentially, these factors impact on the overall consistency of the sample being electroporated. They include sample volume, temperature, inter-electrode gap, ionic-strength of the extracellular medium, conductivity of the cell membrane and cytoplasm, cell density and the purity, concentration and size of nucleic acid being transfected. The resistance will impact on the transmembrane potential and as a result the voltage delivery parameters required to destabilise the membrane (Escoffre et al., 2009, Jordan et al., 2007). The resistance of the electroporation system being used can be set manually, and in the case of this work, is in line with manufacturer instructions. The capacitance of the circuit describes the ability of the circuit to store electric charge and is used as the changeable variable when experimentally altering the length of an exponential pulse. This is also manually set (Bio-Rad, n.d.). The time constant (τ) (Equation 1.3.), given in milliseconds, is the term used to describe the rate of voltage decay and is given as the time taken for the pulse to reach approximately 37% (1/e) of its initial intensity, which is derived from equation 1.2. This is the standard measure of pulse length for exponential decay electroporation (Bio-Rad, n.d., Jordan et al., 2007).

$\tau = \mathrm{R} \times \mathrm{C}$ Equation 1.3. Time constant

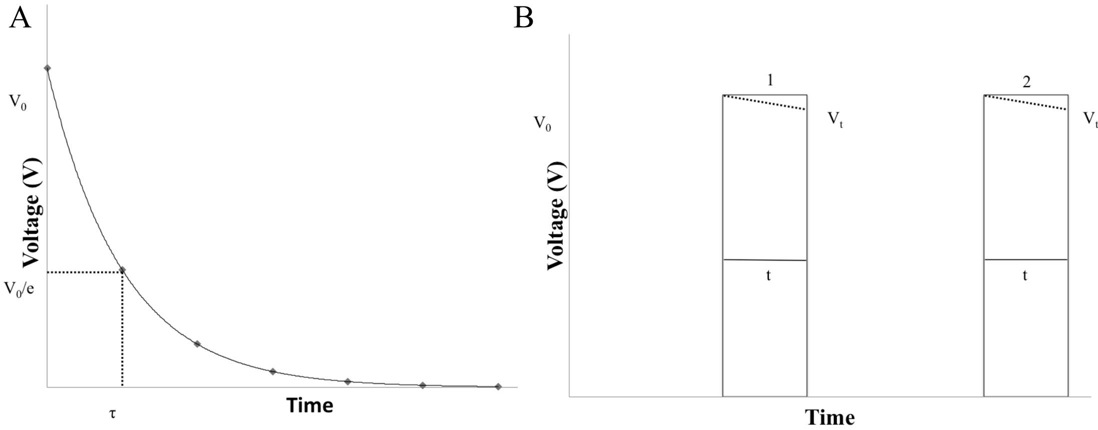


Figure 1.3. Electroporation Waveforms

A) This plot depicts the decay of an exponential pulse derived from Equation 1.2, whereby the voltage is decreasing at an exponential rate, influenced by the capacitance and resistance of the circuit. The time constant (τ) is given as the numerical measurement of pulse length (Equation 1.3.).
B) This plot depicts the square wave waveform, derived from Equation 1.4., with two pulses. A voltage is discharged for a determined amount of time (t). The pulse droop is represented by the dotted line and is derived from Equation 1.5.

(This figure is adapted from Bio-Rad (n.d.), page 47, Figure 4.1.)

Square wave electroporation involves the active truncation of a pulse, which is maintained at the same voltage for a set amount of time and provides the option of delivering multiple pulses (Equation 1.4.) (Bio-Rad, n.d., Jordan et al., 2008).

$$\ln(V_0 - V_t) = \frac{t}{R\ x\ C}$$ Equation 1.4. Square Wave Waveform

For square wave electroporation the pulse length is not given as the time constant, but is instead given as an actual pulse length in milliseconds that has been set manually with the electroporation device. The pulse truncation gives a squared waveform rather than the curved waveform of an exponential decay pulse. In reality the voltage at the end of a square wave pulse is always slightly less than the initial voltage. This slight voltage decay is referred to as the droop (%) and is largely influenced by the resistance and capacitance of the circuit, as well as the time set for the pulse length (Equation 1.5.) (Jordan et al., 2007, Bio-Rad, n.d.).

$$Droop = \frac{(V_0 - V_t)}{V_0}$$ Equation 1.5. Square Wave Droop

1.4. CHO Cell Genetic Instability

The Chinese Hamster (*Cricetulus griseus*) has long been used as a laboratory example specimen. In 1957 Theodore Puck isolated and cultured cells from the ovary of a Chinese Hamster. They were found to be robust, quick and easy to culture and so CHO cells became an established immortal cell line. Genetic instability has always been an inherent feature of CHO cells and they were often used as a model system in studying karyotype heterogeneity and chromosomal aberrations (Jayapal et al., 2007).

Immortal mammalian cell lines are typically genetically heterogeneous (Wurm, 2004). In the cell culture environment, as opposed to a mammalian cell's natural environment in the organ of the organism itself, the selection pressures are different. Initially, the only genes under evolutionary constraint are those that influence cell growth and viability. Therefore, many genes, which do not have a great influence on these growth characteristics, become neutral in the context of evolution. Subsequently, these genes are no longer fixed by natural selection, meaning that when mutations occur they may be more likely to remain in the subsequent generations. These genes will become polyallelic and survival of alleles will be random. This is known as genetic drift (Kimura, 1955, Kimura, 1979). This inherent ability to develop genetic heterogeneity allows for the straightforward and quick evolution of cells towards particular phenotypes, which are desirable for the process of producing recombinant proteins, by imposing particular constraints. Indeed, these genomes are relatively malleable and so can be moulded to fit many purposes. For example, cells have been evolved to be cultured without serum with high cell densities and viabilities, which is desirable because of the potential immunogenic contaminants found in serum (Sinacore et al., 2000). Cells have also been adapted to be able to grow in the presence of compounds such as lactate and ammonia, so that when they are produced as bi-products of the production process cells are not affected by their toxicity (Prentice et al., 2007). The use of a selection gene and an inhibitor (discussed in the section 1.3.3) is another example of exploiting this rapid evolution to produce cells which have more expressive or greater number of copies of the recombinant gene to achieve higher yields of recombinant protein production (Wurm, 2004). Through many generations of cell culture and adaptive evolutionary engineering strategies, a number of phenotypically and genetically distinct CHO cell lines have been created, which exhibit drastic genetic differences to the original Chinese hamster genome (Derouazi et al., 2006, Wurm, 2013).

In the process of stable cell line generation a cloning step is carried out to create homogenous populations of cells, through the generation of new populations from a single cell. Despite this process there is a great deal of phenotypic variability observed between cells in these apparently clonal cell populations, because rapid phenotypic drift generates a mixed population (Barnes et al., 2006). Genetic heterogeneity is a relatively uncontrollable and unpredictable phenomenon, which can greatly affect host cell performance in the production process (Kim et al., 2011). When a population of cells is evolved towards a particular phenotype, such as protein production or to optimise growth characteristics, it stands to reason that the cells selected for use in the production process are those cells that achieve the desirable phenotype first and can do it most efficiently (Figure 1.4). To achieve this change in phenotype there has to be a change in the genetic elements of the cell capable of causing differential expression. The selected cells have achieved this change in genetic elements first and so are likely to be the most genetically unstable. Therefore, potentially, instability itself is selected for and so perhaps it is no surprise that during long-term culture cells tend to deviate from what is desirable, because they are inherently unstable (Heller-Harrison et al., 2009). Alternatively, this could be due to a particular cell acquiring "high-producing" mutations where other cells have acquired less high-producing mutations, because mutation is random. However, if cells are heterogeneous for the many attributes tested, then it is likely that cells are also heterogeneous in terms of genetic stability. It is likely that the properties of a high-producer are attributable to both of these factors. This theory is supported by findings from Liu et al. (2010), in which a dysfunctional state of DNA mismatch repair was induced for the purpose of creating a pool of genetically diverse cells for subsequent phenotypic selection. Inherent instability is a desirable characteristic in the generation of a cell line, but becomes undesirable in the latter stages where desirable phenotypes can be lost. A better understanding of instability is required before this problem can be screened for or solved.

Indeed, CHO cells are believed to have a so-called "mutator" phenotype (Kim et al., 2011). In particular, CHO cells are very karyotypically unstable in the form of homologous recombination-based rearrangements, especially in response to gene amplification steps (Yoshikawa et al., 2000, Derouazi et al., 2006). Instability has also been seen through the loss of recombinant gene copies (Kim et al., 2011), and at the base pair level (Zhang et al., 2015), which has been shown to contain a plethora of single nucleotide polymorphisms (SNPs) (Lewis et al., 2013). A large number of cell doublings are required to create a working cell bank of a recombinant protein-producing cell line that is suitable for the start of long-term cell culture, and then subsequently to scale up cell numbers for production processes. The inherent instability of these cell lines often causes productivity to be greatly decreased or even lost during this period, which can subsequently lead to rejection of cell lines for production purposes (Heller-Harrison et al., 2009, Barnes et al., 2003). Clearly this is unwanted, because a lot of resources have gone into a cell line's development (Barnes et al., 2003). Changes in productivity have been firmly correlated with changes in the transcript level of the recombinant gene, which can be a result of changes in gene expression or changes in recombinant gene copy number (Yang et al., 2010, Kim et al., 2011).

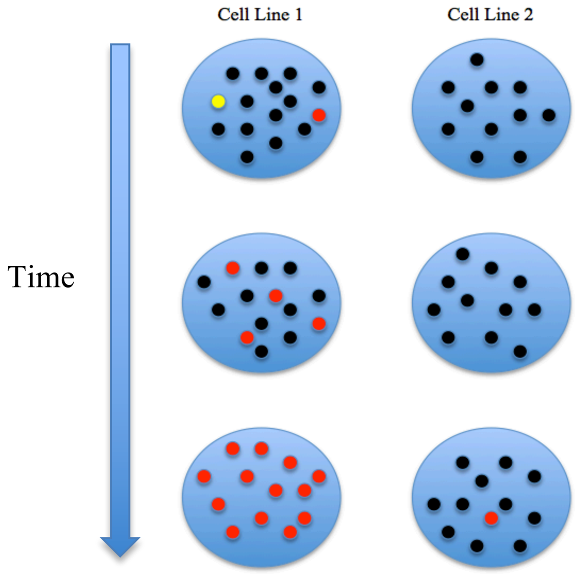


Figure 1.4. Selection of Genetic Instability Phenotype
The schematic illustrates three populations of two different cell lines at different times over the course of a screening process for a desirable phenotype (red). Cell line 1 is more genetically unstable, so acquires mutations more quickly, which generates genetic heterogeneity. Some of these mutations are lost (yellow - neutral) or retained (red – desirable) through random sampling or selection. Cell Line 2 is more genetically stable so acquires mutations at a slower rate and as a result will take longer to achieve the desirable phenotype. Cell line 1 is more likely to be chosen for production processes, but may be more likely to lose productivity further down the line, because of its inherent genetic instability. Perhaps Cell line 2 would be more likely to retain a desirable phenotype once it has been achieved.

Epigenetic regulation is responsible for some of this change in gene expression. For example, it has been shown that DNA methylation correlates with loss in protein productivity. Specific CpG islands within the CMV recombinant gene promoter can become methylated in regions used as transcription binding sites, which has the result of diminishing gene expression (Yang et al., 2010, Kim et al., 2011). Some studies show that loss in productivity is almost solely down to loss in gene expression through methylation (Yang et al., 2010), whereas others show that the predominant cause is recombinant gene loss (Barnes et al., 2007). Kim et al. (2011) showed that a reduction in recombinant gene copy number has been correlated to loss in productivity. In this study instability was present in high and low producing cell lines, such at gene copies of the heavy chain, light chain and GS gene were uneven, despite the initial 1:1:1 ratio in the plasmid vector. In the productively unstable cell lines light chain genes were lost to a greater proportion than heavy chain and GS genes. Potentially, this is because the light chain gene is surrounded by more repetitive sequences, so a homologous recombination event is more likely to happen around this gene than the others (Kim et al., 2011).

The position effect could influence both of these factors that cause changes in gene expression. Plasmid integration near inactive regions can make the transgenic region itself become inactive through silencing (Wurm, 2004). The position effect could also impact on gene copy number; For a plasmid to become integrated there needs to be a gap created by genomic breakage for the plasmid sequence to integrate. There are certain hot spots for DNA damage and subsequently for areas creating these genomic gaps. Therefore, plasmids could be more likely to insert into a region prone to genomic breakage and thus be more at risk of rearrangements. Insertion sites are likely to have different levels of inherent stability and capacity for gene expression (Denissenko et al., 1997, Barnes et al., 2007, Kim et al., 2011).

Once some cells within the population have acquired lower productivity attributes it is thought they have a growth advantage over high producing cells because of the lessened metabolic burden of not producing recombinant protein. Therefore, these low producing cells can take over the population because of their growth advantage, causing the cell line's overall production to decline. If genetic instability can be understood and controlled then this phenomenon can be prevented (Barnes et al., 2007).

This instability does not only impact upon cell productivity, but can also have adverse effects on product quality. During the cell line development process cell lines are assessed for product quality attributes, such as protein aggregation, charge variants, glycosylation variants, and sequence variants, in line with regulatory body requirements (Ren et al., 2011, Zhu, 2012). It is crucial that these attributes remain consistent to ensure the safety and efficacy of a recombinant therapeutic product (Zhang et al., 2015). An underlying instability can lead to phenotypic heterogeneity in all if these quality attributes (Ren et al., 2011, Davies et al., 2013). Other than sequence variants, which will be analysed in this thesis (chapter 5), an example of one of these influential phenotypes is glycosylation. It has been shown that cell lines become heterogeneous in N-glycan processing of recombinant products, which can have an impact on the pharmacokinetics and biological function of a recombinant protein (van Berkel et al., 2009, Zhu, 2012, Davies et al., 2013). Sequence variants have been discovered in a large proportion of clonal cell lines, and have been shown to directly cause changes to the amino acid sequence of a recombinant protein (Zhang et al., 2015). Evaluation of these product quality attributes on product efficacy and immunogenicity presents a technical challenge and so if cell lines carrying these undesirable attributes can be identified early in cell line development, they can be eliminated as a candidate cell lines for production processes (Zhang et al., 2015, Davies et al., 2013).

1.5. Advancements and Future Directions

Advancements in the production of biological therapeutics can be measured in different ways, such as increased product titers, increased cell qP, product quality consistency, time to market and the variety of products able to be produced by a given system, amongst others. This chapter has already summarised some of the key areas in which changes have been, and continue to be, made. For example; vector design for an increased and tailored production, cell engineering strategies to boost productivity and growth, transfection method variety and optimization, advancements in TGE for more insightful and faster screening processes, research into targeted integration for more consistent and predictable levels of gene expression, improvement of downstream methodologies for higher titers and product purity, and improvements in gene selection systems such as the GS knockout cell line. These improvements have already led to volumetric productivity being increased from 0.5 to 2-10 g/L in large-scale bioprocesses (Datta et al., 2013).

1.5.1. Systems Biology and Omics technology

The overall concept of systems biology is the shift from looking at biological organisms purely at the molecular level to investigating whole-organism biology. Clearly, molecular techniques are needed to study processes mechanistically and in detail, but the idea is to integrate all of these defined isolated reactions and processes into a working model of a dynamic cellular network (Westerhoff and Palsson, 2004). As well as integrated analysis, development of high-throughput technologies has allowed the study of cellular functions on a global scale in which large datasets can be analysed together. The term 'omics' is used to describe this (Westerhoff and Palsson, 2004, Kildegaard et al., 2013). Omics includes the study of the entirety of a cell's genes (genomics), mRNA (transcriptomics), proteins (proteomics), metabolism (metabolomics), metabolic flux (fluxomics) and glycosylation profiles (glycomics) (Datta et al., 2013, Kildegaard et al., 2013). All of this information together allows for a better understanding of cellular complexity in a way that is more than just a sum of its parts, but as the interacting and ever-changing environment that it is. Discoveries here can lead to a wide range of useful engineering targets to facilitate cell line improvements (Kildegaard et al., 2013; Westerhoff and Palsson, 2004).

1.5.2. Synthetic Biology

Synthetic biology aims to apply understanding of genetic elements and their interactions to the engineering of novel genetic constructs that offer novel or improved functionality to a host (Lienert et al., 2014). Logical parallels were derived from electrical circuit design such that genetic circuits could be built in a similar modular fashion with functional components such as switches, oscillators and feedback loops (Khalil and Collins, 2010, Lienert et al., 2014). These so-called building blocks can be taken from different organisms and combined in a way that would not occur naturally to create truly novel functions. This can be achieved through the interaction of different genes and recombinant proteins, and through the creation of advanced proteins that contain sequence components from different origins (Purnick and Weiss, 2009, Lienert et al., 2014). In one study a library of synthetic promoters was created based upon a bioinformatics sequence analysis of promoter sequence abundance. It was determined, via the use of synthetic reporters, that promoters designed in this fashion could reach expression at twice the level of the CMV promoter and could consistently and precisely control gene expression in CHO cells over two orders of magnitude. Through doing this the importance of different promoter sequence components were accurately defined (Brown et al., 2014).

1.5.3. Screening Tool

There is a rising demand for a wider variety of therapeutics that can be produced in abundance. Therefore, as our understanding and capacity for process optimisation increases it is important that there is the capability of assessing production platform attributes in a high throughput manner quickly and cheaply (Browne and Al-Rubeai, 2009). For example, an essential step in the production process is the transition from a heterogeneous pool of producing cells to the generation of clonal cell lines, which can be assessed for desirable attributes. Initially, this was achieved by limited dilution cloning methods, which are slow and laborious (Browne and Al-Rubeai, 2007). Fluorescence-activated cell sorting (FACS) methods allowed for a more high-throughput process and enabled the selection of cells by their productivity through the assessment of cell surface protein expression, saving time in the clonal screening process (Browne and Al-Rubeai, 2009). More recently clone picking has been automated through the use of mechanical systems such as the ClonePix from Genetix. The ClonePix quantifies secreted protein immoblised in semi-solid medium on a single cell level, providing a better indication of total cellular protein than protein expressed on the cell surface as with FACS (Nakamura and Omasa, 2015, Browne and Al-Rubeai, 2009). After clonality has been established multiple clones are grown and assessed for their growth and productivity characteristics in a high-throughput plate format. The best of these clones are taken forward for expansion and further testing (Le et al., 2015, Noh et al., 2013). Cell line stability and heterogeneity as well as product quality are also key attributes of concern at this stage and will be discussed separately in chapters 4 and 5.

TGE, as discussed previously, is an extremely useful process in which process parameters can be optimised quickly and cheaply, because it is not as laborious or expensive as SGE. This means that new candidates and their variants can be tested, different cell lines compared, vectors can be varied and optimised and different media formulations can be analysed in a high-throughput manner. This is an extremely useful platform in predicting how processes will function in stable production (Pham et al., 2006, Baldi et al., 2007, Andersen and Krummen, 2002). Clearly, the development of screening tools in TGE processes can help streamline the production process.

1.6. Project Aims

This chapter has summarized the platforms and bioprocesses utilised for the production and recovery of recombinant therapeutic proteins with a focus on genetic instability. As described, there are still many gaps in our biological and process knowledge, the understanding of which can facilitate the improvement and optimisation of these processes. As our knowledge base widens the number of options for bioprocesses increases. For example, a wider variety of proteins can be produced, through a number of engineering strategies, different vector designs, via different transfection technologies and using different selection methods. Options are also increased in that bioprocess characteristics can be more acutely tested and analysed at each stage of the process. Clearly it is important that we have the ability to test these attributes efficiently.

This thesis focuses on the characterisation and understanding of three aspects of the bioprocesses described above and the potential application of the findings through more optimised methodologies or potential bioprocess screening tools.

Chapter 3 discusses the effect of CHO cell genetic instability and heterogeneity on therapeutic protein production bioprocesses. The chapter aims to characterise the extent of this stability at the base pair and chromosomal level and demonstrate the potential need for a screening tool for genetic stability of clonal protein-producing cell lines.

Chapter 4 shows the optimisation of electroporation, which was needed for the generation of stable GFP cells in chapter 5. In doing this it was discovered that standard industry conditions could be vastly improved in a product and platform specific manner using design of experiments (DoE) methodology.

Chapter 5 discusses the difficulties in maintaining product quality throughout the production bioprocess, specifically in the form of sequence point mutation. Firstly, the chapter aims to assess recombinant DNA sequence integrity at different stages in generating a GFP-producing cell population. The second aim is to validate an alternative analysis of the Pacific Biosciences PacBio RSII single-molecule sequencing platform to facilitate a higher resolution of mutation detection.

This page is intentionally left blank

Chapter 2

Materials and Methods

This chapter provides a detailed description of the materials and methods used to complete the experiments described in results chapters three, four and five.

Microbial work and molecular biology techniques were carried out in a separate lab to mammalian cell culture to ensure cell culture sterility. Any materials or vessels to be used in culturing of mammalian cells were sterilized with 70% ethanol and work was conducted within a laminar flow hood. Materials used were of high purity and, where necessary, underwent appropriate filtering and autoclaving procedures.

2.1. CHO Cell Culture

2.1.1. Cell Culture Maintenance

CHOK1SV derived suspension cells (cell line CHO269M, Pfizer, NY, USA) were cultured in CD-CHO medium (Thermo Fisher Scientific, MA, USA) supplemented with 6mM L-glutamine (Thermo Fisher Scientific, MA, USA) in vented Erlenmeyer flasks (Corning, Surrey, UK). Cell culture volumes used were 20-25% of Erlenmeyer flask total volume. Flasks were incubated at a temperature of 37 °C, in 5% (v/v) CO_2 and shaking at 140 rpm. Cells were routinely subcultured at a seeding density of 0.2 x 10^6 cells/mL on a 3-4 day schedule. A Vi-Cell cell viability analyser (Beckman-Coulter, High Wycombe, UK) was used to determine the average cell viability, concentration and diameter via an automated Trypan Blue exclusion assay in which non-viable cells are permanently stained. Cells were subcultured up to a maximum of 25 times in order to minimise genetic diversity, apart from for the generation of stable cell lines (detailed in section 2.6)

The cell culture growth characteristics; cell doubling time (equation 2.1.), generation number (equation 2.2.) and cell specific growth rate (μ) (equation 2.3.) were calculated using the equations below:

$$\text{Cell Doubling Time} = \frac{(t_f - t_0)\ ln2}{\ln(VCD)t_f - \ln(VCD)t_0} \qquad \text{Equation 2.1.}$$

$$\text{Generation Number} = \frac{t_f \cdot \ln(VCD)t_f - \ln(VCD)t_0}{(t_f - t_0)ln2} \qquad \text{Equation 2.2.}$$

$$\mu = \frac{\ln(VCD)t_f - \ln(VCD)t_0}{(t_f - t_0)} \qquad \text{Equation 2.3.}$$

Where *t* is time, *f* is final and VCD is viable cell density.

2.1.2. Cryopreservation and Cell Bank Generation

Master and working cell banks were created for the CHO269M cell line received from Pfizer (NY, USA); Two days after subculture (mid-exponential phase) cells were pelleted by centrifugation at 130 x g for 8 minutes and resuspended at a concentration of 1 x 10^7 cells/mL in CD-CHO media containing 7.5% DMSO (Sigma Aldrich, Dorset, UK). 1 mL aliquots were assorted into NUNC cryovials (ThermoFisher Scientific, MA, USA) and stored in a “Mr. Frosty” container (Nalgene, Roskilde, Denmark), filled with 100% isopropanol, at -80 °C overnight to allow slow freezing of cell solutions. Cryovials were then transferred to a liquid nitrogen freezer (-196 °C) for long-term storage. To revive cells from liquid nitrogen storage, cells were rapidly thawed at 37 °C. Subsequently, the cell solution was added to 30 mL of pre-warmed media and a sample taken for determination of viability and VCD. Cells were then incubated at standard culture conditions. These cells are labeled “Day 0”. Cells are subcultured after two days of subculture and subsequently follow the standard subculture regime. Cells are acclimatised to these conditions for three subcultures before being used for any experimental work.

2.2. Plasmid DNA Amplification and Preparation

2.2.1. Transformation and Plasmid Amplification

A phCMV C-GFP FSR Vector (Genlantis, CA, USA) plasmid was transformed into Library Efficiency® DH5α™ *Escherichia coli* (*E. coli*) competent cells (Thermo Fisher Scientific, MA, USA); DH5α™ cells were thawed on ice and mixed with 25 ng of plasmid DNA, incubated for 30 minutes on ice, heat shocked at 42 °C for 45 seconds and then returned to ice incubation for a further 2 minutes. Cells were then diluted 1:10 in LB-Broth (Thermo Fisher Scientific, MA, USA) and incubated for 1 hour at 37 °C. The cells were then spread on to LB-Agar (Thermo Fisher Scientific, MA, USA) plates containing Kanamycin (Sigma Aldrich, Dorset, UK) at a concentration of 50 ug/mL. Plates were incubated at 37 °C overnight. A colony was picked and used to inoculate 5 mL LB-Broth containing 50 ug/mL Kanamycin to generate a starter culture, which was incubated at 37 °C, 200 rpm for 8 hours. Starter cultures were then used to inoculate larger volumes of Kanamycin containing LB-Broth for bulk amplification, which were incubated at 37 °C, shaken at 200 rpm for 12-16 hours.

2.2.2. Plasmid Extraction and Purification from *E. coli*

A Gigaprep kit (Qiagen, Manchester, UK) was used to lyse *E. coli* cells and purify amplified plasmid DNA, following the manufacturers protocol. Briefly; kit buffers are used between centrifugation steps to lyse cells via alkaline lysis and precipitate a large proportion of cellular components. The remaining supernatant is applied to an anion exchange resin column, which binds plasmid DNA and the remaining impurities are removed through wash steps. The plasmid DNA is then eluted using nuclease free water (Thermo Fisher Scientific, MA, USA) for short-term storage or Tris-EDTA buffer (Thermo Fisher Scientific, MA, USA) for long-term storage, both at -20 °C.

2.2.3. Caesium Chloride Extraction from Transfected Mammalian Cells

Cells were pelleted by centrifugation at 2500 x g for 5 minutes, washed in PBS (Sigma Aldrich, Dorset, UK), resuspended in 250 uL of a resuspension solution (50 mM Tris-HCl - Thermo Fisher Scientific, MA, USA; 10 mM EDTA - Thermo Fisher Scientific, MA, USA; 100 ug/mL RNase – QIAGEN, Manchester, UK) and lysed with 250 uL 1.2% SDS (Sigma Aldrich, Dorset, UK) supplemented with 20 uL Proteinase K (Thermo Fisher Scientific, MA, USA). The solution was mixed by inversion and incubated at room temperature for 5 minutes before adding 350 uL precipitation solution (3M CsCl - Sigma Aldrich, Dorset, UK; 1M potassium acetate - Sigma Aldrich, Dorset, UK; 0.67M acetic acid - Thermo Fisher Scientific, MA, USA). The precipitation solution was mixed by inversion, incubated for 15 minutes on ice and centrifuged at 15,000 x g for 15 minutes. Supernatant was applied to a Miniprep column (Qiagen, Manchester, UK) and centrifuged at maximum speed for 1 minute. 750 uL wash buffer (80 mM potassium acetate - Sigma Aldrich, Dorset, UK; 10 mM Tris-HCl ph7.5 - Thermo Fisher Scientific, MA, USA; 40 uM EDTA - Thermo Fisher Scientific, MA, USA; 60% ethanol - Thermo Fisher Scientific, MA, USA; diH$_2$O) was added to the column and centrifuged at maximum speed for 1 minute. DNA was eluted from the column using 50 uL nuclease-free water (Thermo Fisher Scientific, MA, USA) via a final centrifugation at maximum speed for one minute. Samples were pooled using the miVac DNA concentrator (Genevac Ltd, Ipswich, UK).

2.2.4. BluePippin Purification

Validation of the Blue Pippin system (instrument and reagents - Sage Science, MA, USA) was carried out with the assistance of demonstration from a Sage Science representative, Will Deacon. Purification was carried out using pulsed field electrophoresis cassette BLF7150, which uses a 0.75% agarose gel and an external S1 marker. The instrument was set to purify DNA of a target length of 5.3 kb, with a maximum range of purification between 4.25 kb and 6.35 kb. Target DNA was purified after 145 minutes of running the gel. For purification of DNA samples to undergo sequencing, Blue Pippin purification was carried out by GATC Biotech (Konstanz, Germany), which targeted 5 kb DNA fragments, with a maximum range of purification of 3 kb.

2.2.5. Restriction Digestion of Plasmid DNA

500 ug plasmid DNA, 1x CutSmart™ Buffer (New England Biolabs, UK) diH_2O and 2000U AflII restriction endonuclease (New England Biolabs, UK) were mixed and incubated for 2 hours at 37 °C. The endonuclease was denatured by incubating the restriction solution at 65 °C for 25 minutes. An ethanol precipitation was carried out to purify the linearised plasmid DNA, which was resuspended in Tris-EDTA buffer (Thermo Fisher Scientific, MA, USA) at a concentration of 1.3 mg mL^{-1} for storage at -20 °C.

2.3. Post-preparation Assessments of Plasmid DNA

2.3.1. Agarose Gel Electrophoresis

Plasmid DNA was run on 0.8% agarose Tris-acetate-EDTA (TAE) (Sigma Aldrich, Dorset, UK) gels mixed with ethidium bromide (Sigma Aldrich, Dorset, UK) for ~90 minutes alongside a Hyperladder I (Bioline, UK) molecular weight ladder. Images were taken under UV light using a Biospectrum Imaging System (UVP, CA, USA).

2.3.2. Nanodrop Quantification of DNA

A Nanodrop 2000 (Thermo Scientific, MA, USA) was used to determine DNA concentration and purity. The Beer Lambert Law (Equation 2.4) calculates the absorbance of a DNA sample, which in turn can be used to calculate the concentration of DNA samples (Equation 2.5), using light path length and extinction coefficient of DNA (0.02 ug ml^{-1} cm^{-1}) at a wavelength of 260 nm.

$$A = \in bc \qquad \text{Equation 2.4}$$

$$c = \frac{A_{260}}{0.02} \qquad \text{Equation 2.5}$$

Where A is absorbance, ε is molar extinction coefficient (L mol^{-1} cm^{-1}), b is path length (cm) and c is concentration (mol L^{-1}). The purity of samples was determined by the 260/280 ratio, where a ratio between 1.8 and 1.9 was indicates purity.

2.4. Electroporation

CHO cells were centrifuged at 130 x g for 8 minutes two days after subculture, at which point they had reached a VCD between 0.8 x 10^6 and 1.2 x 10^6. Cell pellets were then washed with 20 mL CD-CHO (Thermo Fisher Scientific, MA, USA) and centrifuged again at 130 x g for 8 minutes. Cells were resuspended in pre-warmed media (CD-CHO, L-Glutamine - Thermo Fisher Scientific, MA, USA) at a concentration of 1.68 x 10^6 cells mL^{-1}. 40 uL (50 ug) linearised phCMV C-GFP plasmid DNA (Genlantis, CA, USA) in Tris-EDTA (Thermo Fisher Scientific, MA, USA) buffer was added to a 4 mm electroporation cuvette (Bio-Rad Laboratories, CA, USA) followed by 595 uL cell solution (1 x 10^6 cells). During preliminary parameter optimisations to determine the constant conditions of the optimisation process, cells were electroporated using standard Pfizer Conditions: 300 V, 900 uF, exponential decay pulse. Post-optimisation, electroporation was conducted using 320-26 conditions: 320 V, 26 ms, exponential decay (time constant protocol). All electroporations were carried out on the Gene Pulser Xcel electroporation system (Bio-Rad Laboratories, CA, USA). Electroporated cells are immediately diluted with 500 uL media and transfered to a 6-well plate (ThermoFisher Scientific, MA, USA) containing 2 mL pre-warmed media for static, humidified incubation at 37 °C and 5% (v/v) CO_2 for 24 hours.

2.5. Generation of Stable GFP Cells

The phCMV C-GFP plasmid contains a neomycin resistance gene. Therefore G418, a neomycin analogue, can be used to select for cells with genome-integrated plasmid, so that over time the cell culture will be populated only by stably producing GFP cells. Protocols were derived from a combination of the electroporation protocol optimised above, Lonza reference guides and an in-house GFP stable cell line generation protocol (Lonza, 2012).

G418 is known to have batch to batch and cell line to cell line inconsistencies, so a dose response study was carried out to ascertain the minimum concentration needed to prevent cell growth for each batch used. A dose response study was set up in which CHO cells were subcultured at 0.2 x 10^6 cells/mL into 50 mL Cultiflasks (Sigma Aldrich, Dorset, UK) containing CD-CHO (ThermoFisher Scientific, MA, USA), 6mM L-glutamine (ThermoFisher Scientific, MA, USA) and G418 disulphate salt (Sigma Aldrich, Dorset, UK) at concentrations spanning 0-1.5 mg/mL. Cultiflasks were incubated at 37 °C, 5% (v/v) CO_2, shaking at 170 rpm. Cell viability and concentration were determined daily for 5 days.

Electroporation was carried out using 320-26 conditions, using 1 x 10^7 cells and transferred into T-75 flasks (ThermoFisher Scientific, MA, USA) for static, humidified incubation. After 24 hours, 0.8 mg/mL and 0.9 mg/mL G418 disulphate salt (Sigma Aldrich, Dorset, UK) was added to cultures for G418 batches 1 and 2 respectively. G418 is known to be relatively unstable at 37 °C, so during this static incubation phase, media was replaced every 3-4 days. When cultures reached a viable cell density of > 0.5 x 10^6 cells/mL, cells were transferred to 30 ml Erlenmeyer flasks with 0.2 x 10^6 cells/mL for standard shaking conditions, supplemented with G418 and 1x HT supplement (ThermoFisher Scientific, MA, USA). Cell viability and VCD was measured using the Vicell and GFP fluorescence was recorded by flow cytometry after each subculture. Fluorescence-activated cell sorting (FACS), carried out at the University of Sheffield Flow cytometry core facility, was used twice for the top 90% and 20% of GFP positive cells respectively in order to generate a high-producing cell population.

2.6. Flow Cytometry

Cells were centrifuged at 130 x g for 8 minutes and resuspended in PBS (Sigma Aldrich, Dorset, UK) for flow cytometric analysis. An Attune® Autosampler (ThermoFisher Scientific, MA, USA) flow cytometer and Attune® Autosampler software (ThermoFisher Scientific, MA, USA) was used to analyse cell samples for GFP fluorescence via excitiation with a 488 nm laser, and detection with a 530/30 band pass filter. Photomultiplying tube (PMT) sensors were optimized at 900 mV for GFP detection and 1200 and 2400 for forward scatter (FSC) and side scatter (SSC) respectively. A viable cell population was gated in accordance with FSC and SSC. Non-transfected cells were used to measure cell auto-fluorescence and used to set a bi-marker gate to distinguish between GFP and non-GFP producing cells, so that 99% of cells were in the negative gate in a non-transfected cell sample (Figure 2.1). For each sample 10,000 cells were measured.

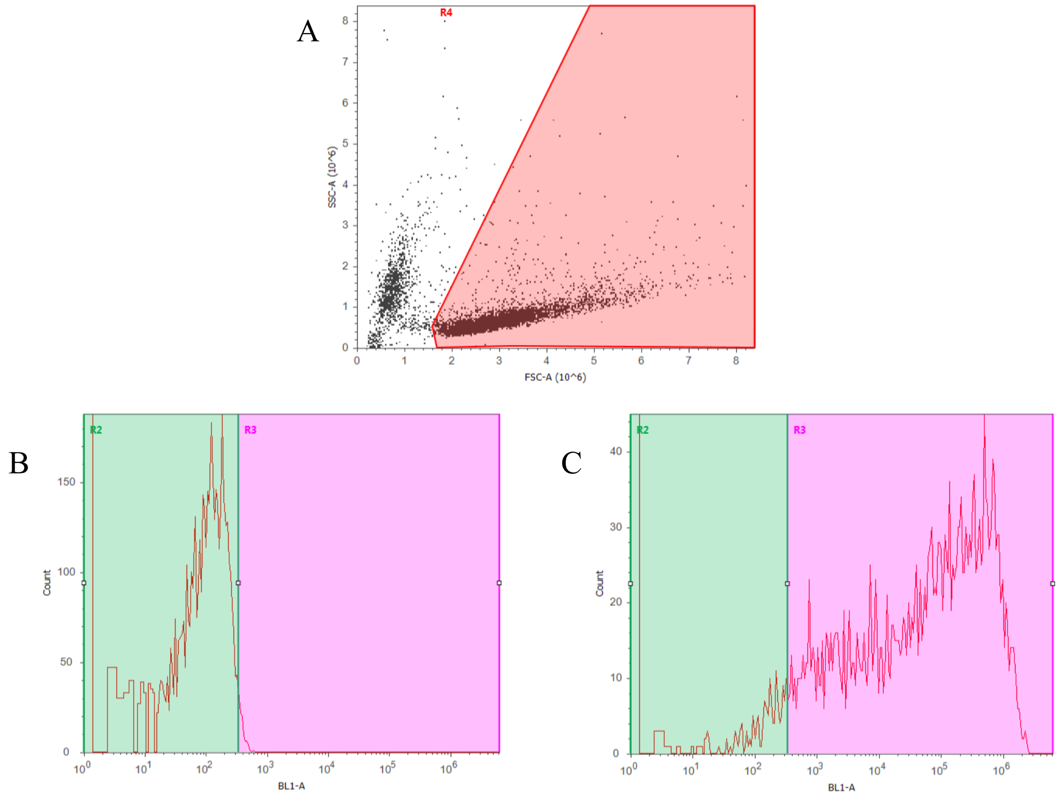


Figure 2.1. Flow Cytometry Gating Example

Viable cells were gated using FSC and SSC (A). Bi-marker gates were set according to cell auto-fluorescence of negative controls (B) to ascertain the percentage of cells that are fluorescing in positive samples (C).

2.7. Response Surface Methods

Design expert 9.0.4 was used to design and analyse experiments for the optimisation of electroporation parameters. For one preliminary experiment for the optimisation of sample volume was carried out using a one factor RSM model. The remaining RSM optimisations were carried out using rotatable central composite designs. For each model the data is analysed in this order: 1. In terms of the response range ratio, which reveals if any data transformations would make data easier to interpret. In this case a

Box-Cox plot for power transforms would instruct on which type of transformation to carry out. 2. A fit summary is presented in terms of a sequential model sum of squares (SMSS), lack of fit tests and model summary statistcs (MSS): standard deviation, R squared, adjusted R squared, predicted R squared and predicted residual sum of squares (PRESS). This fit summary suggests which type of model best fits the data in terms of polynomials. 3. Based on this information the appropriate model is set to fit the data. 4. An ANOVA describes the significance of the model and the significance of each factor within this model. Moreover, the lack of fit is again presented. 5. Diagnostic tests are viewed and analysed to check for residual abnormalities. 6. Graphical representations of the models are plotted for visualising the response surface. 7. The optimisation function then utilises the predictive capacity of the model to generate an optimal set of parameters for a given set of requirements.

2.8. Microsatellite Analysis

2.8.1. Stable Cell Line Generation – 2

Ten GS-CHOK1SV cell lines (B1-B10), producing recombinant mAb were generated by Peter M. O'Callaghan and Minsoo Kim as desbribed in Kim et al. (2011), according to standard methodology (Porter et al., 2010). CHOK1SV (Lonza Biologics) cells were electroporation with a linearised GS vector containing a mAb light and heavy chain (LC and HC). Cells containing genome-integrated plasmid were selected for by 50 uM methionine sulphoximine (MSX; Sigma-Aldrich, Dorset, UK). Clones were made by capillary cloning (B3-B10) or by FACS-facilitated single cell sorting (B1 and B2). Cell lines B1 and B2 expressed IgG_1 mAb, whereas cell lines B3-B10 expressed a range of different IgG_2 mAbs. Cell lines B4, B5, B6, B8, B9 and B10 were transfected with a codon-optimised HC sequence along its entire length, whereas cell lines B1, B2, B3 and B7 were transfected with a non-codon optimised HC sequence. LC and GS genes were identical throughout.

2.8.2. Cell Culture

Cells were cultured by Peter M. O’Callaghan and Minsoo Kim in the conditions described in (Kim et al., 2011). Briefly, cell lines B1-B10 were subcultured using a 3-4 day regime, in which they were seeded at 0.3 x 10^6 viable cells/mL. CD-CHO medium (ThermoFisher Scientific, MA, USA) with a supplement of 25 uM MSX (Sigma-Aldrich, Dorset, UK) was used. Cells were incubated at 36.5 °C. All other cell culture conditions are in line with those described in section 2.1.1.

2.8.3. Microsatellites and Primers

Peter O’callaghan and Claire Bennett identified and designed primers for six microsatellites in the CHO genome (Table 2.1).

Microsatellite	Sequence	Forward Primer	Reverse Primer	Source
10.1	$(CA)_n$	GCCTAGGCTCAAAC AAGCAC (20)	TATAAGACACAAG TAGTGAGTG (22)	(Aquilina et al., 1994)
11.1	$(CA)_n$	TTTTCCAAGTATGTG CTTCCCTG (20)	AAACAAGGTTCAG TGGGATAGC (22)	(Aquilina et al., 1994)
21.1	$(CA)_n$	TTTCCCAAAGAAGTC ATATGCC (22)	CCTTCCTGCAATCT CAAGATG (21)	(Aquilina et al., 1994)
GNAT2	$(TTC)_n$	CAATGTTACTCTATC CCATCCTGG (24)	GTAAGGCTCCTGTC TGTGAGACAG (24)	(Baron et al., 1996)
GT-23	$(CA)_n$	ATCTGAAGTTAAAAT GAAGTTG (22)	CTCTGTGGGTATGC ACATAG (20)	(Hinz and Meuth, 1999)
BAT25	(T)n	GAGGAGTGCCACAA ATCAAAGCTAG (25)	CCCAGATTTTCAGA TTTTAACCATG (25)	(Liu et al., 2010)

Table 2.1. Microsatellites and Primers
The table contains a list of the microsatellites used in this study, their base composition, the forward and reverse primer sequences used for each microsatellite and the literary source, which provided a previous example of microsatellite use.

2.8.4. Sample Preparation

Genomic DNA samples from 1 x 10^8 cells were prepared using the Agilent DNA extraction kit (Agilent Technologies, CA, USA) according to manufacturer instructions. PCR was carried out using the Hot Start Taq plus kit (QIAGEN, Manchester, UK) according to manufacturer instructions, using the primers shown in table 2.1 to amplify microsatellite DNA.

2.8.5. Capillary Gel Electrophoresis

This was outsourced to Steven Haynes of the Core Genomics Facility at the University of Sheffield Medical School. The following was a protocol provided: 1 ul of PCR amplified sample was mixed with 8.7 ul of formamide and 0.3 ul of LIZ600 size standard per sample, which was then transferred to plates and centrifuged to the bottom of each well. The plate was then transferred on to the heating block, heated to 95°C, for 3 minutes. The plate was then incubated on ice for 5 minutes. The 3730 genetic analyser (ThermoFisher Scientific, MA, USA) was used to separate fragments by size using automoated capillary gel electrophoresis.

2.8.6. Statistical Analysis in R

Statistical analysis using ANOVAs, Tukey's multicomparisons tests, F tests, power transformations (Box-Cox plots), T-tests and graphical representation was carried out using R software.

2.9. Karyotype Analysis

Genomic samples were prepared as described in section 2.9.4. Karyotype analysis was outsourced to Duncan Baker of the Sheffield's Children's Hospital genetic diagnostics service in which for each cell line 30 cell squashes were viewed by giemsa staining, and karyotypes were noted when they existed in 3 or more of the cells within this squash, because this is the number thought to be enough to represent a new clone of cells.

2.10. Single Molecule DNA Sequencing

2.10.1. Sample Preparation

The linearised stock and transfected / non-integrated samples discussed in chapter 5 were prepared using protocols described in sections 2.2.2 and 2.2.4 respectively. The transfected / non-integrated sample underwent an additional purification step using BluePippin (Sage Science) technology as described in section 2.2.5. For integrated genomic recombinant plasmid DNA samples genomic DNA was prepared using a Blood and Cell Culture DNA kit (QIAGEN, Manchester, UK) according to manufacturer protocols. Briefly, cells were centrifuged, washed and resuspended in PBS (Sigma Aldrich, Dorset, UK). Cells were then lysed and DNA purified using the buffers and Genomic-tip 20/G column provided. Recombinant plasmid DNA was then amplified via PCR. Primers (Table 2.2) were designed using SnapGene software (GSL Biotech LLC, Chicago, USA) to amplify the plasmid sequence in quarters to generate ~1.25 kb fragments.

Fragment	Forward Primer	Reverse Primer
1	TTAAGGCGTAAATTGTAAGCGTTAATATTTG	CGCTTCAGTGACAACGTCGAG
	CAATAGGCCGAAATCGGCAAAATCC	CAATAGCAGCCAGTCCCTTCC
2	GCTCGACGTTGTCACTGAAGC	GGAAGGGACTGGCTGCTATTG
	CACTAGAAGGACAGTATTTGGTATCTGC	GTGGCCTAACTACGGCTACAC
3	GAGCTACCAACTCTTTTTCCGAAGG	GAATCCGCGTTCCAATGCAC
	GGTTTGTTTGCCGGATCAAGAG	CGTTCCAATGCACCGTTCC
4	GTGCATTGGAACGCGGATTC	GATACATTGATGAGTTTGGACAAACCAC
	GGAACGGTGCATTGGAACG	GATACATTGATGAGTTTGGACAAACCACAAC

Table 2.2. phCMV C-GFP Plasmid Primers
(Sigma Aldrich, Dorset, UK)

The PCR mix contained: 250 ng plasmid DNA, 0.5 ul Phusion high fidelity DNA polymerase (New England Biolabs, UK), 1 ul NTP mix (New England Biolabs, UK), 10 ul Polymerase Buffer (New England Biolabs, UK), 2.5 ul of forward and reverse primers, diH_2O to make final volume 50 ul. The Veriti 96-well thermal cycler (ThermoFisher Scientific, MA, USA) was used for PCR. Samples were heated to 98 °C for 30 seconds, then cycled 40 times through 98 °C for 10 seconds, 65.4 °C for 30 seconds and 72 °C for 38 seconds, followed by a final heating of 72 °C for 10 minutes before being held at 4 °C. Amplified DNA fragments were then purified using a QIAquick PCR purification kit (QIAGEN, Manchester, UK) according to manufacturer instructions. Briefly, DNA is purified using a series of centrifugation steps facilitated by the use of the buffers and spin column provided. The success of the PCR and purification was checked by agarose gel electrophoresis. Resulting samples were resuspended in Tris-Hcl (pH 8.0) buffer (Sigma Aldrich, Dorset, UK) and pooled together for DNA sequencing.

2.10.2 PacBio RSII SMRT Sequencing

Single molecule real time (SMRT) sequencing was outsourced to GATC Biotech (Konstanz, Germany). Briefly, samples are ligated to hairpin adapter sequences to create a SMRTbell template. Individual SMRTbell templates are sequenced by a single polymerase to generate sequence reads containing multiple versions of the template (see explanation in chapter 5). Sequencing is conducted using a PacBio RSII instrument (Pacific Biosciences, CA, USA).

2.10.3 SMRT Sequencing Analysis

Primary analysis was outsourced to Phillip Lobb of Pacific Biosciences (CA, USA) who generated consensus sequences from individual molecules. Secondary analysis. BLASR software was used to align these consensus sequences to the reference sequence. R was used to call mutations and comment on coverage. The script can be found in Figure A26. Details of this analysis can be found in chapter 5.

This page is intentionally left blank

Chapter 3

CHO Cell Genomic Instability and Heterogeneity

3.1. Introduction

3.1.1. Chapter Summary

This chapter provides further introduction to the subject of genetic instability, which was discussed in section 1.5. The chapter focuses on the inherent genomic instability and heterogeneity of CHO cells, and so looks into genetic changes on a global scale. The development of methodologies that are capable of characterising and quantifying this genomic instability would be extremely useful for cell line development platforms, because it would enable the detection and elimination of cell lines with a predisposition to genetic instability and so reduce the chance that production cell lines suffer declines in productivity over long-term cell culture. This study aimed to characterise genomic instability at the base pair and gene copy number level through microsatellite analysis, and at the chromosomal level using karyotype analysis. Ten monoclonal antibody-producing cell lines, which had previously been shown to suffer changes in cell productivity as a result of changes in recombinant gene copy number, were used so that the genetic changes discovered in this study could be directly compared with changes in productivity and gene copy number found in a previous study (Kim et al., 2011).

There was significant microsatellite allelic variation between the ten cell lines, and there were marginal changes in microsatellite allele frequencies across different generations of individual cell lines. However, this variation could only be attributed to genetic drift, rather than mutational change, and so the study did not provide sufficient evidence to suggest that microsatellites could be used as markers for mutational change. There was substantial karyotypic change found in this study, both in the form of changes in chromosome number and breakage / fusion events. This genetic instability was not shown to directly correlate with changes in productivity or gene copy number, but it was concluded that karyotyping could be a useful tool to eliminate genetically unstable cell lines during cell line development.

3.1.2. Forms of Genetic Instability

If a cell is genetically unstable it undergoes genomic changes at a higher rate than a normal cell would, which can come in a variety of forms. There can be: sequence changes involving base substitution, insertion or deletion of one or a few nucleotides, gene copy loss, chromosome number changes from the loss or gain of a chromosomes resulting in aneuploidy, chromosome breakage resulting in loss of chromosome parts, chromosome translocations where two chromosomes fuse, and gene amplification (Lengauer et al., 1998). Cell proliferation is a tightly regulated process with many processes to coordinate. One of these aspects is DNA replication and segregation. DNA needs to be replicated accurately and efficiently segregated in order to maintain genomic integrity throughout many generations (Aguilera and Gomez-Gonzalez, 2008). There are many DNA damage sense and repair pathways and mechanisms to ensure this is the case and if it is not done efficiently mutations and aberrations occur (Jackson, 2002). Clearly this can have its disadvantages, but on the other hand for selection or genetic drift to drive evolution there has to be genetic variation. Therefore mutation is needed for the evolution of cell lines towards desired phenotypes (Hastings et al., 2009, Aguilera and Gomez-Gonzalez, 2008, Sinacore et al., 2000). Unfortunately, information on the specific causes of genetic instability in CHO cells is lacking.

In the case of CHO cells and developing a producing cell line, genetic instability is an attribute that should be closely monitored. Protein folding, PTMs, protein expression and amino acid sequence are some of the key attributes which could be affected by genetic instability, which could have implications for product quality as well as gene expression (O'Callaghan and James, 2008). It has been shown that loss in recombinant gene copy number correlates with a decline in cell specific productivity. Perhaps it is this underlying genetic instability of CHO cells that causes recombinant gene loss and causes observed losses in productivity (Kim et al., 2011). Markers of genetic instability can be used to characterise the extent and type of genetic instability of a given cell line, which include the measure of chromosomal instability, point mutations and the cells response to DNA damage (Lengauer et al., 1998, Jackson, 2002). This study involves the investigation into chromosomal instability and microsatellite instability, which can be used to estimate changes at the base pair level, changes in gene copy number and be used for cell line identification.

Chromosomal instability is a hallmark of the cancer phenotype, a marker of an unstable cell and has been shown to propagate further genetic instability (Mitelman et al., 2007). Chromosomal instability has been shown to cause defects in a wide range of cellular functions such as protein synthesis, protein folding, changes in cellular metabolism, gene expression, cell proliferation and increases in point mutations (Gordon et al., 2012). One form of chromosome instability is aneuploidy, which is the alteration in chromosome number and involves the loss or gain of chromosomes in a daughter cell compared to its mother cell. This is predominantly due to a decline in mitotic fidelity, meaning that the cell is less able to carry out equal chromosome segregation (Thompson and Compton, 2011, Lengauer et al., 1998). Another form of chromosomal instability results in the rearrangement of chromosomes, which can come in the form of deletions, insertions, translocations, duplications, inversions, the formation of isochromosomes and the formation of marker chromosomes. These types of changes result from breakage and fusion of chromosomes (Thompson and Compton, 2011). Chromosome aberrations can cause changes in gene copy number and gene expression, which will inevitably influence cell homeostasis (Thompson and Compton, 2011, Gordon et al., 2012). CHO cells are known for their chromosomal instability, so it is a logical marker to use when measuring the genetic instability of a potential producing cell line (Derouazi et al., 2006).

Microsatellites are short (1-6 nucleotide) DNA motifs repeated in tandem and are interspersed throughout the genome. They are very common, highly variable sequences with many length-based polymorphisms, and are a popular genetic marker (Ellegren, 2004). Mutations causing changes in the number of repeats, and thus causing polymorphic lengths of microsatellite, are relatively frequent. They occur through a mechanism called slippage (Figure 3.1.). Due to the repetitive and homologous nature of microsatellites, complementary strands can misalign after denaturation during DNA replication. This can cause expansion or contraction of repeats depending on the orientation of the misalignment relative to the template strand, because the DNA polymerase does not synthesise the microsatellite to a length consistent with the template (Lai and Sun, 2003). If this mistake is not recognised by the DNA mismatch repair (MMR) systems then the new allele is carried through to subsequent generations. This causes a large amount of variation within populations. Microsatellite polymorphism is commonly used as measure of relatedness between subjects and can be used as a method of cell line identification. Moreover, microsatellite slippage is more common than other base pair-level mutation. Therefore, it is a sensitive marker of MMR fidelity and so can be used as a proxy for all genetic instability at this level i.e. base substitution and insertion / deletion mismatches, as well as gene copy number changes. Base-pair level mismatches can have a wide range of deleterious effects on cellular metabolism and so studying their frequency can give useful information on the genetic stability of a given cell line and its ability to sense and repair that damage (Lengauer et al., 1998, Kurzawski et al., 2004, Lai and Sun, 2003, Kunkel and Erie, 2005, Aquilina et al., 1994, Yu et al., 2015).

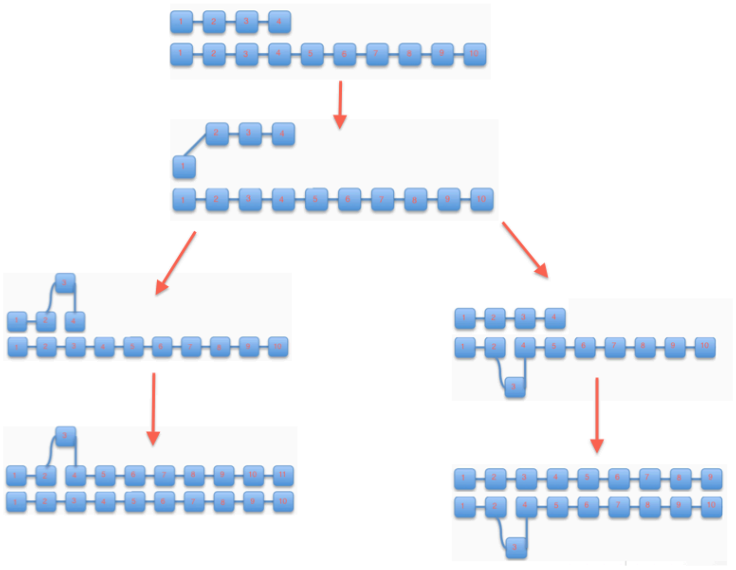


Figure 3.1. Replication slippage.
Each numbered block represents one repeat of a microsatellite. The figure illustrates how DNA strands can become misaligned and as a result the microsatellite can undergo expansion (left) or contraction (right).

3.1.3. Chapter Aims and Hypotheses

This investigation aims to characterise the extent of genetic change at the base pair, copy number and chromosome level through microsatellite and karyotype analysis. Hypotheses:

- There would be significant nucleotide-level change over long-term cell culture.
- There would be significant karyotype-level change over long-term cell culture.
- This genetic change would correlate with observed changes in cell specific productivity and gene copy number
- Heterogeneity would be present between cell lines and would be seen to develop over time.

3.2 Results

The hypothesis generated by Kim et al (2011) was that repetitive sequences within the GS vector are subject to homologous recombination-based gene loss. This was supported by the fact that light chain genes, which are surrounded by more repetitive elements, were lost to a greater extent than heavy chain and GS genes. Vector design, genomic location of recombinant gene integration (position effect) and underlying cell line genomic stability are all postulated to influence this phenomenon. The work presented here aims to build on the work of Kim et al (2011), with an investigation into the hypothesis that the underlying background genetic instability is significant and could strongly influence recombinant gene loss and, subsequently, a decline in qP. The same ten (B1-B10) GS-CHOK1SV mAb-producing cell lines used in the Kim et al. (2011) study, sampled at the same low and high generation numbers, were used to analyse cell line genetic instability at the base pair and chromosome level. A brief summary of the workflow of the experiments and analysis carried out in this chapter is provided in figure 3.2.

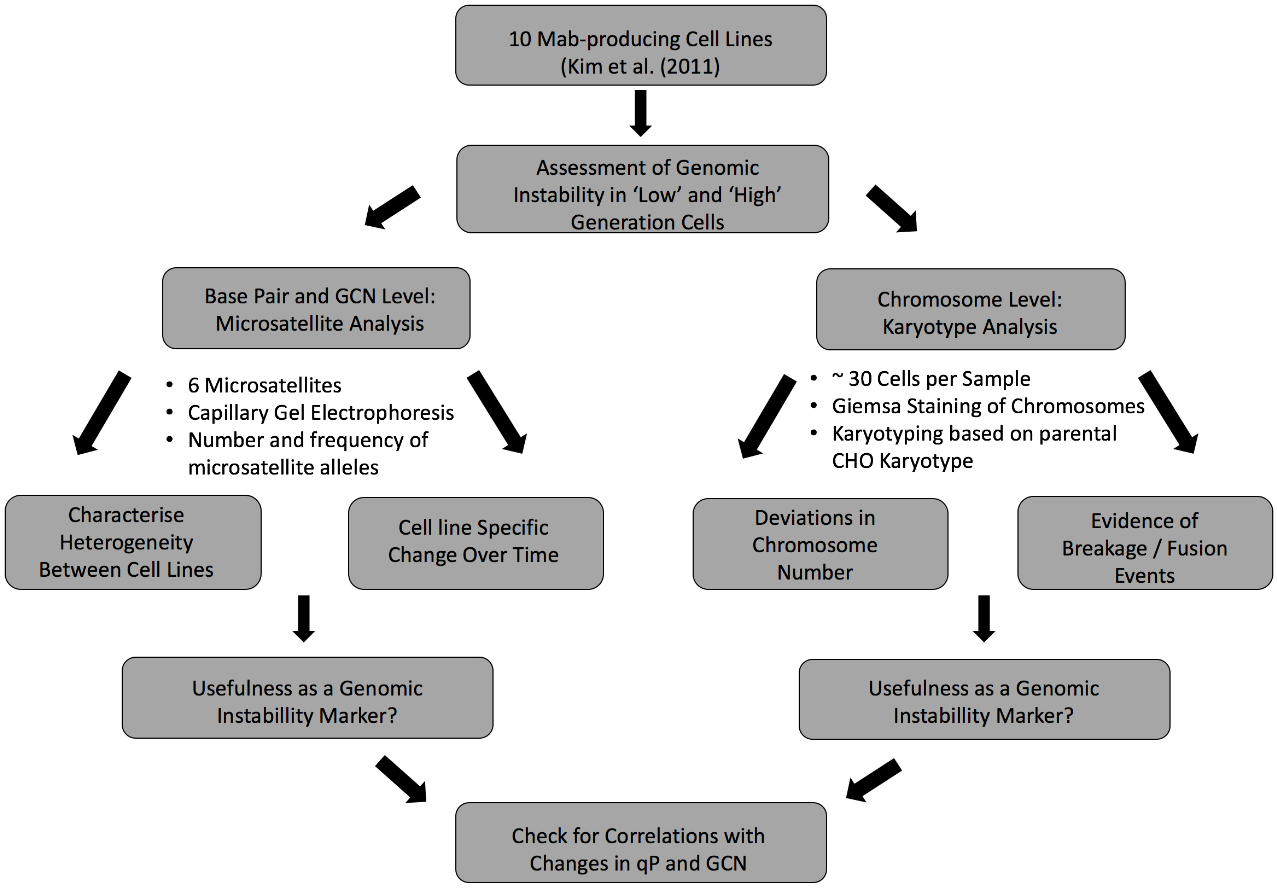

10 Mab-producing Cell Lines
(Kim et al. (2011)
Assessment of Genomic Instability in 'Low' and 'High' Generation Cells
Base Pair and GCN Level: Microsatellite Analysis
Chromosome Level: Karyotype Analysis
• 6 Microsatellites
• Capillary Gel Electrophoresis
• Number and frequency of microsatellite alleles
• ~ 30 Cells per Sample
• Giemsa Staining of Chromosomes
• Karyotyping based on parental CHO Karyotype
Characterise Heterogeneity Between Cell Lines
Cell line Specific Change Over Time
Deviations in Chromosome Number
Evidence of Breakage / Fusion Events
Usefulness as a Genomic Instabillity Marker?
Usefulness as a Genomic Instabillity Marker?
Check for Correlations with Changes in qP and GCN

Figure 3.2. Chapter 3 Workflow
The flow chart begins with the chapter premise, which originates from the 10 cell lines studied by Kim et al. (2011) and the need to assess their genomic instability. This is done at the base pair / GCN level by microsatellite analysis and at the chromosome level by karyotype analysis. Microsatellite analysis involved the study of allelic heterogeneity amongst the 10 cell lines and how allele frequency changes over time. Karyotyping considers changes in chromosome number and form. Both of these tools are then assessed in their ability to report on genomic instability as well as their correlation with changes to qP and GCN.

3.2.1. Microsatellite Analysis

Microsatellite instability was used as a marker of overall genetic instability at the base pair level to assess the heterogeneity of these cell lines and their genetic stability over time. Six microsatellites (GNAT2, 10.1, 21.1, 11.1, GT-23 and BAT25) were used as markers to genetically characterise each of the ten cell lines (B1-B10). Samples from each cell line were taken at both a low and high number of generations after cloning. A summary of the exact sampling generations, production stability and gene copy number (GCN) is shown in Table 3.1. Microsatellites were amplified by PCR and analysed by capillary gel electrophoresis in order to determine the extent of microsatellite polymorphism in each cell line. Genemapper® (Applied Biosystems) and Peak Scanner® (Applied Biosystems) software were used to determine the number (number of peaks) and frequency (peak height) of alleles for each microsatellite (Figure 3.3). Peak height is sample specific and so cannot be compared between different samples. Therefore, peak heights were converted to percentages to normalise the data. This profile of different allele frequencies in a given cell line will hereafter be referred to as the allele frequency distribution. Table 3.2 shows the number of alleles for each microsatellite. Each microsatellite was sampled in duplicate, except for GT-23, which was sampled in triplicate. Due to sampling errors there is no data available for microsatellite 11.1 in cell line B6.

	Generation Number		Change in qP		Recombinant GCN Change (%)			Recombinant GCN Change (rate) (generation^{-1} x 10^3)		
Cell Line	Low	High	Percentage	Rate (generation^{-1} x 10^2)	HC	LC	GS	HC	LC	GS
B1	20	72	- 32.6	- 0.80	- 17.9	- 48.1	- 13.6	- 3.8	- 12.6	- 2.8
B2	20	84	- 24.7	- 0.46	- 11.5	- 38.9	- 1.8	- 1.9	- 7.7	- 0.3
B3	16	57	- 3.7	- 0.09	0	8.7	- 5.6	0.0	2.0	- 1.4
B4	23	103	- 23.8	- 0.31	- 15.4	- 34.9	- 12.9	- 2.1	- 5.3	- 1.7
B5	19	95	0.0	0.00	- 20.0	- 31.3	- 22.8	- 2.9	- 4.9	- 3.4
B6	14	82	- 1.8	- 0.03	- 11.1	- 9.6	- 15.9	1.6	- 1.5	2.2
B7	2	77	- 13.9	- 0.22	- 13.8	- 7.9	- 19.2	- 2.0	- 1.1	- 2.8
B8	16	76	- 44.4	- 0.93	- 43.2	- 57.5	- 45.9	- 9.4	- 14.2	- 10.2
B9	12	93	- 70.7	- 1.47	- 72.8	- 83.2	- 81.0	- 15.9	- 21.8	- 20.3
B10	11	92	6.3	0.07	24.1	3.0	19.7	2.7	0.4	2.2

Table 3.1. Gene Copy Number and qP Changes in Cell Lines B1-B10
The data in this table is adapted from data generated in (Kim et al., 2011). The table shows the generation number for each cell line, referred to as "low" or "high", changes in qP both as a percentage and rate, and changes in gene copy number for the heavy chain, light chain and GS gene in terms of percentage and rate.

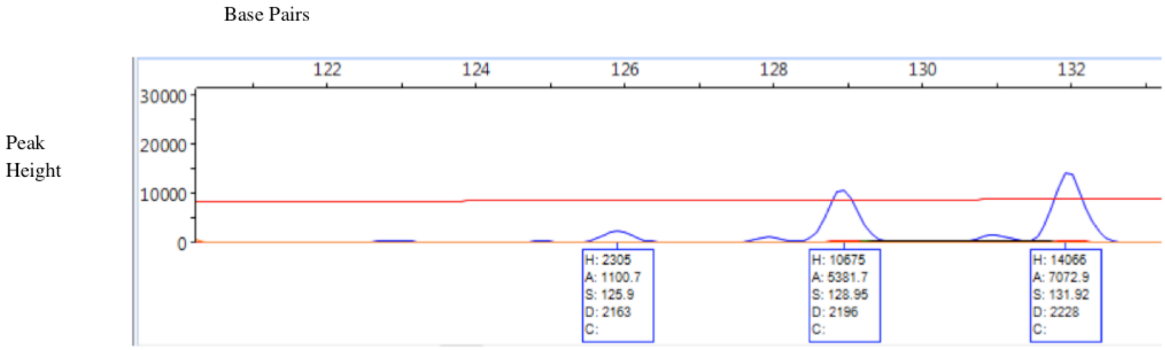


Figure 3.3. Peak Scanner Software Allele Frequency Determination
The figure illustrates how the allele frequencies for each microsatellite were determined using Peak Scanner® software. The three main peaks show that there are three alleles of the microsatellite GNAT2 (126bp, 129bp, 132bp). The peak height (H) determines the frequency of each allele (2305, 100675, 14066) in this cell line (B3 High). These were converted to percentages to enable comparisons between samples (8.5%, 39.5%, 52%).

Microsatellite	Number of Alleles
GNAT2	3
10.1	4
21.1	3
11.1	6
GT-23	6
BAT25	4

Table 3.2. Number of Alleles per Microsatellite
The table shows how many alleles were detected for each microsatellite as determined by Genemapper® (Applied Biosystems) and Peak Scanner® (Applied Biosystems) software.

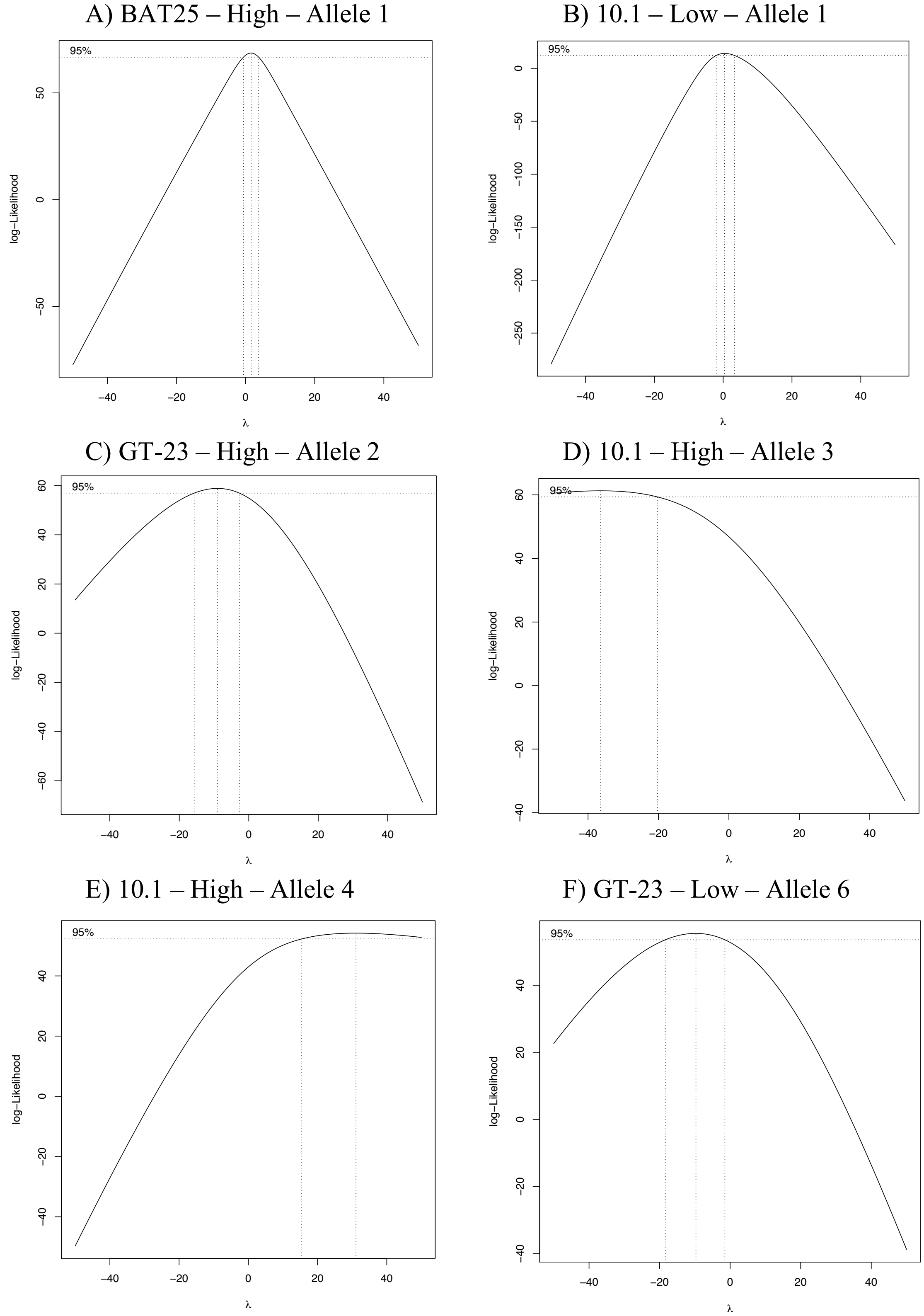
3.2.1.1. Microsatellite Heterogeneity Between Cell lines

This first set of analyses gives an insight into the variation between cell lines within a given generation, which provides information on genetic drift and cell line heterogeneity. Subsequently, it was investigated whether any observed heterogeneity is seen to increase over long-term cell culture by comparing the variance at low and high generation numbers.

It was decided that this analysis would be carried out on an allele-by-allele basis and so the dataset was split into 52 data subsets by categories of microsatellite, generation and allele. So, for example, one subset contained all the percentage values of total GNAT2 microsatellite copies for allele 1 in each cell line at low generation number (GNAT2 – Low – Allele 1: 20 data points, two for each cell line). Before analysing the variance within these data subsets, it was necessary to establish whether their residuals were normally distributed in order to determine whether to carry out a parametric or non-parametric variance test. A Shapiro-Wilk test was carried out to test for normality (Table A41), which shows data is normally distributed when $p > 0.05$. 6 of the 52 data subset residuals were not normally distributed and so Box-Cox plots for data transforms (Figure 3.4) were generated to ascertain the power transformation most likely to yield normal residuals in each case. The following power transformations were carried out:

- BAT25 – High – Allele $1^{1.57}$
- 10.1 – Low – Allele $1^{0.44}$
- GT-23 – High – Allele 2^{-9}
- 10.1 – High – Allele $3^{-36.42}$
- 10.1 – High – Allele $4^{31.1}$
- GT-23 – Low – Allele $6^{-9.73}$

Using these transformed percentage values the Shapiro-Wilk test was used again to check for data subset normality (Table A42). The data transformations resulted in five out of the six data subset residuals being normally distributed. However, one data subset (10.1 – High – Allele 4) still had non-normally distributed residuals. To assess microsatellite heterogeneity between cell lines, one-way ANOVAs were conducted for all data subsets to assess differences in allele frequency distribution, except for data subset 10.1 – High – Allele 4, which was assessed using a non-parametric equivalent rank test, the Kruskal-Wallis one-way analysis of variance. The p-values were then adjusted using a Benjamini Hochberg adjustment, to nullify the type I error risk from using multiple ANOVAs.



A) BAT25 – High – Allele 1

B) 10.1 – Low – Allele 1

C) GT-23 – High – Allele 2

D) 10.1 – High – Allele 3

E) 10.1 – High – Allele 4

F) GT-23 – Low – Allele 6

Figure 3.4: Box-Cox Plots for Power Transforms: Non-Normal Microsatellite Data
The plots show the value of lambda (directly under peak) for a power transformation most likely to yield a normally distributed dataset.

Table 3.3 shows the p-values from all 52 variances tests, highlighting those that were significant ($p < 0.05$). 4 out of the 6 microsatellites (GNAT2, 21.1, 11.1, BAT25) show significant variance in allele frequency distributions between the low generation cell lines and between high generation cell lines. Figure 3.5 contains plots illustrating the allele frequency distributions, in which cell lines are represented by differentially coloured plot lines in individual plots for all microsatellite-generation number combinations. The plots show that the significant microsatellite variation revealed by the ANOVAs (GNAT2, 21.1, 11.1 and BAT25) is not randomly distributed.

A) Low Generation						
Microsatellite	Allele 1	Allele 2	Allele 3	Allele 4	Allele 5	Allele 6
GNAT2 *	1.07E-07*	1.09E-10*	2.45E-11*			
10.1	0.738	0.738	0.738	0.738		
21.1 *	1.45E-04*	1.52E-08*	5.18E-12*			
11.1 *	5.46E-05*	2.78E-07*	1.59E-08*	1.05E-06*	5.48E-07*	2.47E-07*
GT-23	0.854	0.854	0.854	0.854	0.854	0.854
BAT25 *	6.80E-09*	7.03E-12*	7.31E-12*	2.32E-09*		
B) High Generation						
Microsatellite	Allele 1	Allele 2	Allele 3	Allele 4	Allele 5	Allele 6
GNAT2 *	8.87E-11*	6.08E-11*	8.12E-12*			
10.1	0.603	0.603	0.499	0.603		
21.1 *	9.09E-09*	1.23E-12*	6.59E-15*			
11.1 *	4.51E-06*	4.77E-06*	3.01E-06*	0.002*	1.61E-05*	6.01E-07*
GT-23	0.060	0.094	0.094	0.094	0.094	0.094
BAT25 *	1.24E-11*	5.49E-11*	1.12E-11*	2.69E-12*		

Table 3.3: Microsatellite Polymorphism: Variance Between Cell Lines
The table contains the p-values from ANOVA tests describing the microsatellite allelic variance between B1-B10 cell lines at a low (A) and high (B) generation number. * represents significant variance.

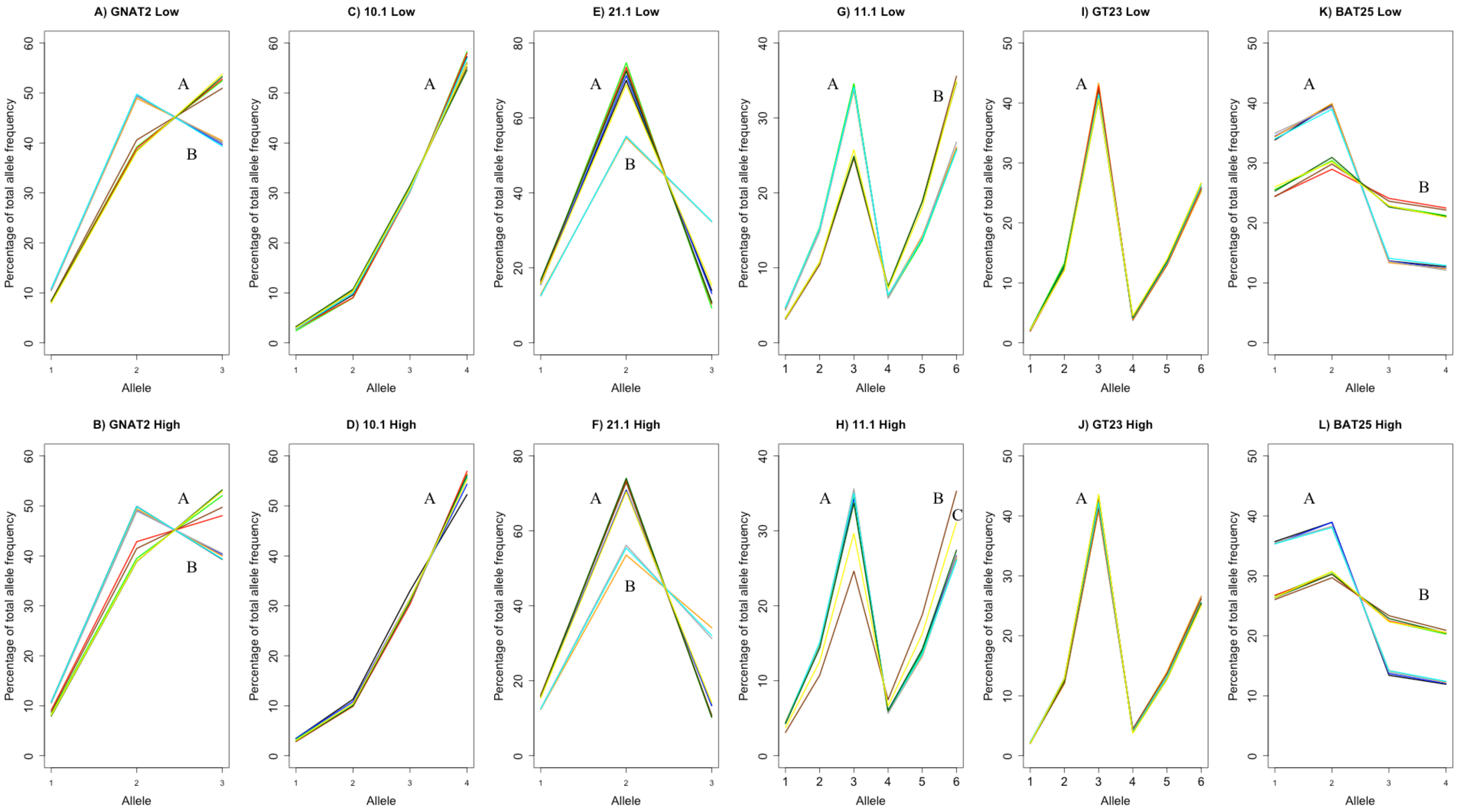


Figure 3.5: Allele Frequency Distribution

The plots show allele frequency for each cell line: Colours: B1 (black), B2 (blue), B3 (green), B4 (orange), B5 (dark gray), B6 (red), B7 (brown), B8 (cyan), B9 (dark green), B10 (yellow). The letters represent different clusters of cell lines that are similar in allele frequency distribution.

Instead, it is due to an apparent clustering phenomenon, whereby subgroups of cell lines have similar allele frequency distributions, which differ significantly to the allele frequency distributions of the other subgroup(s). Cell lines that belong to the same cluster with one microsatellite do not necessarily belong to the same cluster for other microsatellites. However, it is noteworthy that cell lines B1 and B2 as well lines B4, B5, and B8 are always the same cluster. Significantly variable microsatellites present the following clusters:

- GNAT2: Low and High
  - A) B3, B6, B7, B9, B10.
  - B) B1, B2, B4, B5, B8.
- 21.1: Low and High
  - A) B1, B2, B3, B6, B7, B9, B10.
  - B) B4, B5, B8.
- 11.1: Low
  - A) B1, B2, B3, B4, B5, B8.
  - B) B7, B9, B10.
- 11.1: High
  - A) B1, B2, B3, B4, B5, B8, B9.
  - B) B7
  - C) B10
- BAT25: High and Low
  - A) B1, B2, B4, B5, B8.
  - B) B3, B6, B7, B9, B10.

The clusters within these variable microsatellites remain the same from low to high generation, except for microsatellite 11.1, in which cell line B9 appears to change from cluster B to cluster A and cell line B10 forms its own cluster (C) in high generation cells. The plots illustrating those microsatellites that exhibited no significant microsatellite variation in the ANOVAs (10.1 and GT-23) show all 10 cell lines in the same cluster. As well as clustering, the shapes of clusters shift between low and high generations, which is indicative of change over long-term cell culture. For example, GNAT2 cluster A cell lines are more widely spread in the high generation plot in comparison with the low generation plot.

To further analyse the ANOVA results a Tukey's multiple comparisons test was carried out to give a cell line by cell line breakdown of comparisons for each microsatellite. In the case of data subset 10.1 – High – Allele 4 a Kruskal Nemenyi test was carried out, which can be used in the same manner as a Tukey's test for non-parametric data. These results are presented in Tables 3.4-3.9, whereby each comparison is represented by how many of the total alleles for each microsatellite were significantly different ($p < 0.05$) between cell lines. The tables support the conclusions drawn from Figure 3.6, whereby the significant variation shown by the ANOVAs is not randomly distributed, but is mostly down to an apparent clustering phenomenon, in which subgroups of cell lines have similar allele frequency distributions, which differ significantly to the allele frequency distributions of other subgroups. However, these tables also show that there is significant variation within these clusters as represented by the black and red dashed borders for clusters A and B respectively. Moreover, these tables show that, on a cell line-by-cell line basis, there is no significant variation between low and high generation cell lines for microsatellites 10.1 and GT-23. Microsatellites GNAT2, 21.1, 11.1 and BAT25 had significant generational allelic changes over long-term cell culture (orange shaded boxes), having 6, 11, 15 and 3 changes in the number of cell line-by-cell line allelic differences respectively. These allelic changes appear to be randomly distributed in microsatellites GNAT2, 21.1 and BAT25, whereas the changes appear exclusively in cell lines 9 and 10 in microsatellite 11.1. This supports the conclusions drawn from figure 3.6 and indicates that microsatellite change may have led to a breakaway of these cell lines from the initial clustering identified in low generation cell lines. Box plots illustrating these cell line-by-cell line differences are included in the appendix (Figures A27-32). Interestingly, these box plots reveal that the size of residuals for microsatellites 10.1 and GT-23 could be the reason for finding a lack of significant variance in the ANOVAs.

3.4 **GNAT2** (3 Alleles)

Cell Lines		Low									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
High	**1**	■	0	3	0	0	3	3	0	3	3
	2	0	■	3	0	0	3	3	0	3	3
	3	3	3	■	3	3	3	0	3	0	0
	4	0	0	3	■	0	3	3	0	3	3
	5	0	0	3	0	■	3	3	0	3	3
	6	3	3	3	3	3	■	1	3	0	0
	7	3	3	2	3	3	1	■	3	2	2
	8	0	0	3	0	0	3	3	■	3	3
	9	3	3	1	3	3	3	3	3	■	0
	10	3	3	0	3	3	3	3	3	0	■

3.5 **10.1** (4 Alleles)

Cell Lines		Low									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
High	**1**	■	0	0	0	0	0	0	0	0	0
	2	0	■	0	0	0	0	0	0	0	0
	3	0	0	■	0	0	0	0	0	0	0
	4	0	0	0	■	0	0	0	0	0	0
	5	0	0	0	0	■	0	0	0	0	0
	6	0	0	0	0	0	■	0	0	0	0
	7	0	0	0	0	0	0	■	0	0	0
	8	0	0	0	0	0	0	0	■	0	0
	9	0	0	0	0	0	0	0	0	■	0
	10	0	0	0	0	0	0	0	0	0	■

3.6 **21.1** (3 Alleles)

Cell Lines		Low									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
High	**1**	■	0	2	3	3	1	1	3	1	0
	2	0	■	1	3	3	1	1	3	1	0
	3	2	2	■	3	3	0	0	3	0	2
	4	3	3	3	■	0	3	3	0	3	3
	5	3	3	3	2	■	3	3	0	3	3
	6	2	2	0	3	3	■	0	3	0	2
	7	2	2	0	3	3	0	■	3	0	1
	8	3	3	3	2	0	3	3	■	3	3
	9	2	2	0	3	3	0	0	3	■	1
	10	0	0	2	3	3	2	2	3	2	■

3.7 **11.1** (6 Alleles)

Cell Lines		Low									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
High	**1**	■	0	0	0	0	╲	6	0	6	6
	2	0	■	0	0	0	╲	6	0	6	6
	3	0	0	■	0	0	╲	6	0	6	6
	4	0	0	0	■	0	╲	6	0	6	6
	5	0	0	0	0	■	╲	6	0	6	6
	6	╲	╲	╲	╲	╲	■	╲	╲	╲	╲
	7	6	6	6	6	6	╲	■	6	0	0
	8	0	0	0	0	0	╲	6	■	6	6
	9	0	0	0	0	0	╲	6	0	■	0
	10	5	5	5	5	5	╲	5	5	5	■

3.8 **GT-23** (6 Alleles)

Cell Lines		Low									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	1		0	0	0	0	0	0	0	0	0
	2	0		0	0	0	0	0	0	0	0
	3	0	0		0	0	0	0	0	0	0
	4	0	0	0		0	0	0	0	0	0
High	**5**	0	0	0	0		0	0	0	0	0
	6	0	0	0	0	0		0	0	0	0
	7	0	0	0	0	0	0		0	0	0
	8	0	0	0	0	0	0	0		0	0
	9	0	0	0	0	0	0	0	0		0
	10	0	0	0	0	0	0	0	0	0	

3.9 **BAT25** (4 Alleles)

Cell Lines		Low									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	1		0	4	0	0	4	4	0	4	4
	2	0		4	0	0	4	4	0	4	4
	3	4	4		4	4	2	0	4	0	0
	4	0	0	4		0	4	4	0	4	4
High	**5**	0	0	4	0		4	4	0	4	4
	6	4	4	0	4	4		0	4	2	2
	7	4	4	0	4	4	0		4	0	0
	8	0	0	4	0	0	4	4		4	4
	9	4	4	0	4	4	0	0	4		0
	10	4	4	0	4	4	0	0	4	0	

Tables 3.4-3.9: Tukey's Multiple Comparisons Tests
The tables provide the number of significantly different alleles for each cell line-cell line comparison at low and high generation numbers for each microsatellite. Low and high cell line comparisons are shown above and below the blacked-out cells respectively. The orange highlighted values represent the allele number values that have changed over long-term cell culture. Black and red dashed borders represent variations within A and B clusters respectively.

So far the analysis has indicated that cell line microsatellite heterogeneity may have increased from low to high generation. For example, the ANOVA p-values generally decreased from low to high generations, figure 3.5 showed greater dispersion of cell lines for higher generations and generally Tukey's tests revealed that higher generation cell lines were more significantly different than lower generation cell lines. To confirm the validity of these inferences, an F Test was carried out to compare variances between low and high generation cell lines. The F test results are summarised in table 3.10. Generally, there was no significant difference in variances between low and high generations, except on two occasions with 10.1-Allele 3 and GT-23-Allele 1. Variance in microsatellite 10.1 and GT-23 was deemed not to be significant, so significant changes in variance here were not counted. This indicates that variance has not changed over long-term cell culture.

Microsatellite	Allele 1	Allele 2	Allele 3	Allele 4	Allele 5	Allele 6
GNAT2	0.738	0.673	0.690			
10.1	0.211	0.117	0.031*	0.987		
21.1	0.787	0.971	0.974			
11.1	0.109	0.069	0.397	0.860	0.348	0.146
GT-23	0.0398*	0.318	0.056	0.206	0.304	0.162
BAT25	0.911	0.524	0.724	0.705		

Table 3.10: F Test for Variance Comparison Between Generations
The table contains p-values generated from comparing variances between allele frequencies across low and high generations by F Tests (Table A43 contains the full set of p-values with variance ratios). * represents a significant change in variance.

However, it has already been established that the main cause of significant variation between cell lines is due to the clustering phenomenon described previously and it could be the case that significant changes in variance between low and high generation were being masked by this large source of variation. Therefore, F Tests were carried out to assess the significant differences in variances between the clusters of cell lines identified in the low generation. The results of these F Tests are summarised in table 3.11. This method was more able to identify significant variance differences between generations and the results show that heterogeneity appears to increase over long term cell culture. Differences in variance appears to be partially present for all microsatellites, apart from microsatellite 11.1 for which cluster B has significant differences in variance for all alleles. This supports the greater amount of change determined previously in microsatellite 11.1 in comparison to other microsatellites. Again, for those variances deemed not significant by ANOVAs (10.1, GT-23, significant changes in variance should not be counted.

Microsatellite	Cluster	Allele 1	Allele 2	Allele 3	Allele 4	Allele 5	Allele 6
GNAT2	A	0.088	0.086	0.049*			
	B	0.060	0.768	0.952			
10.1	A	0.211	0.117	0.031*	0.987		
21.1	A	0.007*	0.267	0.340			
	B	0.193	0.385	0.103			
11.1	A	0.006*	0.785	0.115	0.819	0.275	0.058
	B	0.007*	0.002*	0.001*	0.003*	0.008*	0.001*
GT-23	A	0.040*	0.317	0.056	0.206	0.304	0.162
BAT25	A	0.115	0.939	0.874	0.165		
	B	0.011*	0.330	0.478	0.007*		

Table 3.11: F Test for Variance Comparison Between Generations by Cluster
The table contains p-values generated from comparing variances between allele frequencies across low and high generations by cluster, using F Tests (Table A44 contains the full set of p-values with variance ratios). * represents a significant change in variance.

3.2.1.2. Cell Line-specific Microsatellite Changes Over Time

A more direct analysis, using T-TESTs, was carried out to assess the differences in allelic frequency distributions between low and high generations of individual cell lines. A Benjamini Hochberg p value adjustment was carried out to minimise type I error. A p-value less than 0.05 indicated a significant allele frequency distribution difference between early and late generations of a cell line. Table 3.12 shows the results of the T-TEST in terms of how many alleles per cell line changed significantly for each microsatellite.

		Cell Line										Changed Cell Lines (%)				
												By Number		Weighted		Stability
Microsatellite	Alleles	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	R	N	R	N	
GNAT2	3	0	0	0	0	0	3	0	0	0	0	10	11.9	10	11.9	SS
10.1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	S
21.1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	SS
11.1	6	0	0	0	0	0	0	0	0	2	0	10	10	3.33	3.33	NS
GT-23	6	0	1	0	0	0	0	0	0	0	0	10	12.7	2.8	3.5	NS
BAT25	4	0	1	0	0	0	0	0	0	0	0	10	12.7	10	12.7	SS
Cell Line	R	0	33.3	0	0	0	16.7	0	0	16.7	0					
Change (%)	N	0	42.2	0	0	0	19.9	0	0	16.7	0					
Weighted	R	0	6.9	0	0	0	16.7	0	0	5.6	0					
Change (%)	N	0	8.7	0	0	0	19.9	0	0	5.6	0					
Stability		S	NS	S	S	S	SS	S	S	NS	S					

Table 3.12: T tests for Cell line-specific Microsatellite Changes over Time

The table summarises cell line specific microsatellite allele frequency percentage comparisons and microsatellite-specific change between low and high generations. Cell line microsatellite stability is represented by percentage of microsatellite change and by weighted percentage change, which takes into account allele number. Percentages are given in raw (R) or normalised (N) (by generation number) form. Microsatellite stability is represented in the same manner. Stability categories based on normalised weighted percentages: (S – 0%), nearly stable (NS – 0-10%) and semi-stable (SS – 10-20%).

The percentage of cell line change in microsatellite was calculated using the number of microsatellites showing any instability (Cell Line Change (%)) and the weighted percentage of cell line microsatellite change was calculated by averaging the percentage of significantly unstable alleles for each microsatellite (Weighted Change (%)). These percentage values (R) were then normalised (N) for generation by using the highest generation number as a reference point (B9 and B10 – 81 generations) Cell lines were categorized into stable (S – 0% change: B1, B3, B4, B5, B7, B8, B10), nearly stable (NS – 0-10%: B2, B9) and semi-stable (SS – 10-20%: B6) groups based on their normalised weighted percentage changes. As can be seen in Table 3.12 there was only a small amount of significant cell line-specific microsatellite change between low and high generation, which ranging between 0 – 19.9% weighted normalised percentage. Changes exhibited some cell line specificity, indicating that some cell lines were more genetically stable than others. It should be noted that figure 3.5 shows little change between early and late generations in allele frequency distribution, which is in line with the level of change shown in the T Tests. Individual microsatellite instability was calculated in the same manner, using percentage of cell lines (By Number) exhibiting change and a weighted (Weighted) percentage using an average of significantly unstable alleles. Microsatellites differed in their stability, ranging between 0-17% weighted normalised percentage, which would indicate that genetic instability is microsatellite (locus) specific. It should be noted that there were many T Test p – values that fell within the 0.05-0.1 range, meaning that they were nearly deemed to show significance (Table A45). More repeats may have revealed a higher level of microsatellite change.

A correlation analysis was carried out using Pearson’s product moment correlation coefficient to establish whether the microsatellite instability observed in this study correlates with the qP and GCN changes observed by Kim et al. (2011) (Table 3.1). Both weighted and non-weighted percent changes in microsatellite were used for this. Table 3.13 contains the p-values from these correlation analyses, which show that there is no significant correlation between the observed microsatellite instability and changes in qP or GCN.

	Normalised Microsatellite Change	Weighted Normalised Microsatellite Change
Rate of HC GCN Change	0.688	0.950
Rate of LC GCN Change	0.525	0.946
Rate of GS GCN Change	0.763	0.895
Rate of average GCN Change	0.635	0.971
Rate of qP Change	0.543	0.919

Table 3.13. Microsatellite Stability Correlation Analysis
The table contains p-values from Pearson's product moment correlation coefficients when comparing microsatellite changes with changes in GCN and qP.

3.2.2. Karyotype Analysis

CHO cells are known for having an unstable karyotype, which perhaps is not surprising considering it was originally isolated and cultured to study different forms of chromosomal aberration, amongst other things. Chromosomes can change in number to form aneuploid cells, or can change in form via breakage and fusion events with different chromosomes. The karyotypes of cell lines B1-B10 both at low and high generation were attained by viewing giemsa-stained cell squashes of cells from these populations. Approximately 30 cell squashes were analysed per sample. This was carried out by the Sheffield's Children's Hospital. Each chromosome was characterised and annotated in line with the methodology used by (Derouazi et al., 2006), which follows criteria set out in the established system for the karyotyping of CHO cells (Ray and Mohandas, 1976) and the International Standard Committee on Human Cytogenetic Nomenclature (Mitelman, 1995). Karyotypes of the ten cell lines were compared to the karyotype of parental CHO cell lines (Figure A33) as a standard. Chromosomes were identified and are referred to using the following nomenclature:

- Numeric: According to wild type hamster chromosomes.
- Derived (der(y)): Structurally rearranged chromosome derived from a known chromosome, where y = the name of the known chromosome type. In the case of chromosomal fusion, resulting in a chromosome made from two known

chromosomes, y is given as the known chromosome that is the largest constituent of the derived chromosome.

- Isochromosome (Iso): Chromosome made from two identical arms of known origin.
- Additional material of unknown origin (Add(y)): A chromosome in which chromosome y has fused with unidentifiable chromosomal fragment(s).
- Z: Specific groups of morphologically altered chromosomes that have been previously identified in CHO cells.
- Marker (Mar): Unclassifiable chromosome.

The karyotype of the parental cell line contains 19 chromosomes. Even in the relatively small sample size of ~30 cells per culture sample, cell lines B1-B10 clearly deviate from the standard 19 chromosomes per cell (Table 3.14). Indeed, only three culture samples (B7 Low, B9 High and B10 Low) show complete homogeneity in chromosome number. This aneuploidy indicates that DNA replication is error prone, either in the form chromosome number moderation during mitosis or in the form of chromosome breakage or fusion events that generate modified chromosomes. Figure 3.6 shows the composite karyotype containing every chromosome type seen in cell lines B1-B10 within this investigation, differentially colour-labeled according to whether they exist in wild type hamster cells (blue), are considered as common CHO chromosomes (i.e. parental - red), or are chromosomes novel to this investigation (black) (NB. “novel” chromosomes were counted as chromosomes not seen in the parental CHO karyotype and chromosome duplication events that appear to have occurred during stable cell line generation within this investigation). Figure 3.6 shows that these cell lines have undergone vast chromosomal change and show that this collection of cell lines have diverged from the parental cell line karyotype with 18 novel chromosomes being presented here, both in the form of duplications of common CHO chromosomes and modified chromosomes. There are 4 cases of novel (black) chromosome duplication, as indicated by the ‘+’ symbol. Novel chromosomes labeled ‘add’, ‘iso’, ‘der’ or ‘Mar’ represent those chromosomes that are generated as a result of breakage and fusion events. Marker (Mar) chromosomes in particular are postulated to derive from multiple breakage or fusion events, because their constituents cannot be recognised as a previously seen chromosome. This shows that improper segregation moderation and breakage / fusion detection and repair are a consistent phenotype of these cell lines. A full list of cell line karyotypes can be found in table A46.

	Number of Metaphase Cells with Chromosome Number n								
Cell Line	17	18	19	20	21	22	23	24	25
B1 Low		1	28	1					
B1 High	1	3	20	5	1				
B2 Low		1	25	4					
B2 High		3	27						
B3 Low		1	27						1
B3 High		2	28						
B4 Low			28	2					
B4 High			28	2					
B5 Low		2	27	1					
B5 High			29	1					
B6 Low	1	1	26	1	1				
B6 High		2	26	4	1				
B7 Low			30						
B7 High			15	14	1				
B8 Low			17	2	1				
B8 High		8	21	2	1				
B9 Low	1	3	21	2					
B9 High			30						
B10 Low			26						
B10 High		1	28	1					

Table 3.14: Chromosome Number in Cell Lines B1-B10.
The table contains the chromosome numbers from the ~30 cell squashes obtained from cell culture samples of cell lines B1-B10 at both low and high generation numbers.

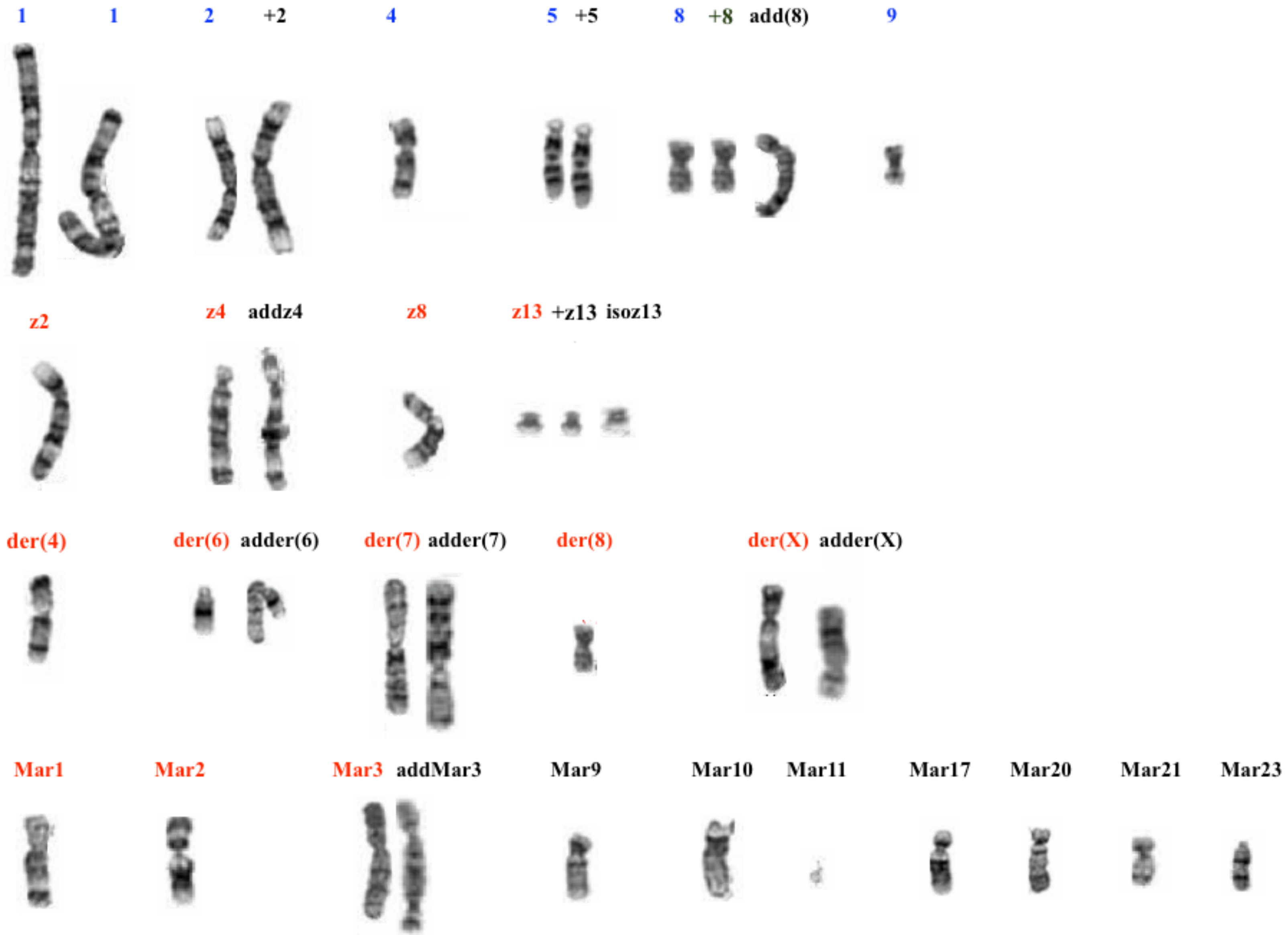


Figure 3.6: Composite CHO Karyotype from Cell Lines B1-B10

The figure contains all chromosomes identified in the investigation. (Colours and terminology described in text)

Table 3.15 provides a summary of the chromosomal differences between the parental cell line and cell lines B1-B10. It contains all the novel chromosomes that were generated during the course of this investigation and also chromosomes that were present in the parental karyotype, but absent in some of cell lines B1-B10 (i.e. they have been lost - distinguished by *). Therefore, the table gives an impression of how unstable each cell line is in terms of how many abnormal chromosomal observations it contained (crosses) and how many changes in karyotype were observed between low and high generation numbers (orange highlight). In the instances where there was more than one karyotype observed in a given population of cells, the karyotype subpopulations are distinguished numerically aside the cell line-generation label. Data from Kim et al. (2011) regarding changes in qP are included in the right-hand columns. This table demonstrates that there were a large amount of abnormal chromosomes generated and that there have been many changes from the parental karyotype. There were no cell lines that maintained the parental karyotype in either generation. Moreover, 70% of cell lines B1-B10 showed changes in karyotype from low to high generation number. This indicates that CHO cells are largely unstable at the chromosome level, which has caused genetic heterogeneity between and within clonal cell lines. Interestingly, the cell lines that did not show any karyotypic changes between low and high generations showed some of the lower changes in qP and the two cell lines that demonstrated the largest changes in karyotype exhibited the largest changes in qP. However, this data is qualitative, and so firm conclusions regarding correlations cannot be made for potential cause and effect relationships. It is perhaps noteworthy that all cell lines lacking the marker10 chromosome showed no karyotypic changes in long-term culture, whereas add8 chromosome was found to be present in all of these non-changing cell lines. Moreover, chromosomes 1(x2), 2, der(4), 5, 8 / add8, 9, Z13 / isoZ13, Z8, Z4 /addZ4, Z2, marker1 and marker3 were found in all cell lines, so could perhaps be essential or contain essential elements. Also, no structural changes to original CHO chromosomes (blue) were observed in chromosomes 1, 2, 5 and 9, which may indicate that they contain essential genes and so if changed could have lethal results. Indeed, no duplication events are observed for chromosomes 1 and 9, which may indicate that changes in gene balance of these chromosomes cannot be tolerated.

Cell Line	Marker 2*	Marker 9	Marker 10	Marker 11	Add 8	Marker 17	AddMar 3	Add z4	Marker 21	(+) 2	Marker 20	Iso z13	Marker 23	(+) 5	Add er (6)	Adder (7)	Adder (X)	(+) z13	Der 6*	(+) 8	Der 7*	qP Change (%)	qP Change (rate)
B1 Low		☒	☒																☒		☒	-32.6	-0.8
B1 High 1		☒	☒	🟧☒															☒		☒		
B1 High 2		☒	☒																☒		☒		
B2 Low		☒	☒																☒		☒	-24.7	-0.46
B2 High 1	🟧☒	🟧	🟧															🟧☒	☒		☒		
B2 High 2		☒	☒																☒		☒		
B3 Low		☒			☒	☒													☒		☒	-3.7	-0.09
B3 High		☒			☒	☒													☒		☒		
B4 Low		☒			☒	☒	☒												☒		☒	-23.8	-0.31
B4 High		☒			☒	☒	☒												☒		☒		
B5 Low		☒			☒	☒	☒												☒		☒	0	0
B5 High		☒			☒	☒	☒												☒		☒		
B6 Low	☒							☒	☒										☒		☒	-1.8	-0.03
B6 High 1	☒		🟧☒					☒	☒										☒		☒		
B6 High 2	🟧		🟧☒					☒	☒										☒		☒		
B7 Low			☒							☒	☒								☒		☒	-13.9	-0.22
B7 High 1			☒							☒	☒								☒		☒		
B7 High 2			☒							🟧	☒								☒		☒		
B8 Low 1			☒			☒	☒	☒				☒							☒		☒	-44.4	-0.93
B8 Low 2			☒			☒	☒	☒				☒	☒						☒		☒		
B8 High 1			☒			☒	☒	☒				☒	🟧						🟧		☒		
B8 High 2			☒			☒	☒	☒				☒							☒		☒		
B9 Low	☒		☒					☒				☒		☒	☒							-70.7	-1.47
B9 High 1	☒		🟧					☒				☒		☒	☒					🟧☒			
B9 High 2	☒		🟧					☒				☒		🟧	☒	🟧☒				🟧☒			
B10 Low	☒		☒											☒	☒							6.3	0.07
B10 High	☒		☒											☒	☒		🟧☒						

Table 3.15: Cell Lines B1-B10 – Differences to Parental Karyotype

The table contains all the chromosomal differences between cell lines B1-B10 and the parental karyotype (Explained in text)

Table 3.16 presents the karyotype data slightly differently, whereby all the cell line populations / subpopulations with the same karyotype are grouped into the same 'cluster'. "L" and "H" refer to low and high generation cell lines respectively. Cell line sub populations with different karyotypes are distinguished in the same numerical format as in table 3.15.

Cluster	Cell Line Subpopulations
1	B1-L, B2-L, B1-H-1, B2-H-1
2	B3-L, B3-H
3	B4-L, B5-L, B4-H, B5-H
4	B6-L
5	B7-L, B7-H-1
6	B8-L-1, B8-H-1
7	B8-L-2
8	B9-L
9	B10-L
10	B1-H-2
11	B2-H-2
12	B6-H-1
13	B6-H-2
14	B7-H-2
15	B8-H-2
16	B9-H-1
17	B9-H-2
18	B10-H

Table 3.16. Unique Karyotype Clusters
The table combines together all cell line subpopulations with the same karyotype. Clusters that appear in both low and high generation cell lines, only low generation cell lines or only high generations are distinguished by orange, white and blue shading respectively.

The table shows that there were 18 distinct karyotypes present within cell lines B1-B10 over the course of this investigation. Five of these karyotypes were only present in low generation cell lines (white), five karyotypes were present in both low and high generation cell lines (orange), and nine karyotypes are only present in high generation cell lines (blue). Therefore, in total there were fourteen distinct changes of karyotype between low and high generations (5 lost, 9 gained). Again, this supports the conclusion of gross genetic change at the chromosome level and demonstrates how this has led to an increased heterogeneity between cell lines B1-B10.

The observed changes in karyotype and microsatellite frequency distribution were compared and there appears to be no apparent commonalities in terms of instability. For example, cell line B8 had several abnormal chromosomes in low generation cells and displayed karyotypic changes from low to high generation, but it was completely stable in terms of microsatellites. On the other hand, cell line B3 did not change in karyotype from low to high generation, but showed significant microsatellite instability (8.7% normalized weighted change).

3.3. Discussion

Phenotypic instability has been observed in the form of a decline in recombinant protein productivity and generation of product variants over long-term cell culture, which is a common and costly trait of current bioprocess platforms. Unfortunately, at present there are no predictive tools capable of indicating whether a given cell line may go on to show these undesirable traits (Kim et al., 2011, Derouazi et al., 2006; Zhang et al., 2015). Clearly, it would be a great benefit if a cell line could be marked as stable or unstable in the developmental stages of testing new recombinant therapeutic candidates and their production, rather than investing time and resources into a cell line before discovering that it is productively unstable. This would save time, money and increase the overall efficiency of bioprocesses in terms of time to market and gaining consistent production titers. For such a tool to be put into place, there needs to be a firm understanding of the traits that can cause a cell to decrease in its productive capacity. Detectable markers of these instability-related traits, that are consistent and can be called with confidence, need to be established to make it possible to efficiently evaluate and predict the relative stability of candidate cell lines.

Broadly speaking, the molecular basis of production instability has most commonly been attributed to recombinant gene loss and a decline in the transcription of the recombinant gene (Kim et al., 2011). The relationship of gene copy number with productivity is relatively straightforward, whereby loss in gene copies correlates with a loss in productivity. This relationship may not be strictly linear, because the location of recombinant plasmid insertion dictates the expression capabilities of a given construct, but the correlation has been established. A decline in transcription of the recombinant gene is more complex. This can be due to methylation-based transcriptional silencing, changes to expression-determining sequences (promoter, open-reading frame and enhancer elements), translocation events to less active chromosomal regions, changes to other elements that impact upon transcription (transcription factors) or global transcriptional regulation changes.

3.3.1. Microsatellite Analysis

This study investigated microsatellite instability as a genetic marker for all mismatch repair-related changes, such as point mutations, insertions and deletions (Kunkel and Erie, 2005). Moreover, the slippage mechanism by which microsatellites are altered is similar to the proposed mechanism of recombinant gene loss suggested by Kim et al. (2011), whereby repetitive elements of sequence within the plasmid vector are subject to homologous recombination-based events, causing gene loss. Here, microsatellite instability was used to assess cell line-specific changes over time (two generational time points) and the relatedness of cell lines B1-B10, which were derived from the same parental cell line to measure developed heterogeneity.

Microsatellite changes were analysed through the measurement of allele frequency distributions on an allele-by-allele basis. ANOVAs showed that amongst the low generation cell lines microsatellites GNAT2, 21.1, 11.1 and BAT25 showed significant variation in allele frequency distributions. It was shown that separate clusters of microsatellite-based relatedness were predominantly responsible for this observed variation, which indicates that most of this heterogeneity may have been derived from the cell population of the parental cell line. These clusters were not the same in their cell line content for all microsatellites. However, cell lines B1 and B2 as well as cell lines B4, B5, and B8 were always in the same cluster, which is indicative of a closer level of relatedness between these cell lines compared to others. All cell lines remained within the same cluster from low to high generation number, except for microsatellite 11.1. In microsatellite 11.1 cell line B9 changed from cluster B to cluster A and cell line B10 formed its own cluster, C, over long-term cell culture.

A cell line-by-cell line analysis (Tukey's multicomparisons test) confirmed the presence of these clusters, but also revealed that were was significant variance between some cell lines within certain clusters, which indicates development of heterogeneity that cannot exclusively be attributed to parental cell line derivation. Moreover, in cases where cell line-to-cell line comparisons changed in the number of significantly different alleles from low to high generation number, these changes were predominantly to an increased number of differences, which again would indicate an increasing heterogeneity over long-term cell culture.

F tests were carried out between clusters at different generational time points to statistically measure for a change in heterogeneity over time. There was a significant generational difference between cluster variances, especially for microsatellite 11.1, which supports the conclusion that heterogeneity had developed over long-term cell culture, with microsatellite 11.1 showing the most dramatic change.

T tests were used to determine cell line-specific changes in allele frequency distribution over long-term cell culture. These tests showed that there were minimal significant cell line-specific and microsatellite-specific changes in allele frequency distributions over long-term cell culture. Whilst this shows that cell lines differ in their level of stability, it is difficult to draw any firm conclusions from a dataset reporting so little change. Further study into these microsatellites with many more repeats may generate results that show more significant change. This is supported by the number of p – values generated from T Tests, which could be seen as 'nearly significant' (p = 0.05-0.1). Moreover, the fact that the different microsatellites showed different levels of instability supports the idea that genomic location impacts upon stability. Therefore, if microsatellites as stability markers were validated then they could elucidate 'stable' targets for targeted recombinant DNA integration. The changes observed in microsatellite allele frequency did not correlate with changes in recombinant gene copy number or changes in cell specific productivity, which indicates that these microsatellites could not be used as a predictor of gene copy and production instability in these cell lines. However, this is not surprising given the small amount of change that was observed.

This study has shown that microsatellite allele frequencies vary marginally, but significantly, over time and so, with further validation, could be used as a general marker of genetic instability. However, the variation was not an effective tool for predicting recombinant product stability. There were significant cell-line specific differences in microsatellite changes, which would indicate that microsatellites can distinguish stability between cell lines. Most of the change reported here is likely to be as a result of the slow but progressive nature of genetic drift, which gradually causes cell lines to differ in their allelic frequency distributions. Essentially, this is just an effect of random sampling over the generations (Kimura, 1955, Kimura, 1979). Therefore, microsatellites are a useful marker of allele frequency changes over time. Only microsatellite 11.1 showed signs of replication slippage occurrence, because of the more dramatic cluster changes observed. However, this cannot be concluded definitively, because no novel alleles of a different microsatellite length were detected, but rather a putative slippage event occurred causing a microsatellite change to a length that had already been seen. Therefore, no conclusive evidence was given that microsatellites could be used as a marker for base pair substitution.

The fact that there was no correlation between microsatellite changes and changes in GCN or qP would indicate that these microsatellites are not a reliable marker of genetic instability at the gene copy number level and that base pair level changes did not significantly impact gene expression in these cell lines. However, the genomic location of a microsatellite has an impact on its stability just as the integration site of a recombinant plasmid has an effect on its production stability (Barnes et al., 2007). Therefore, given the fact that the genomic location of these microsatellites is not known and that their genomic context is likely to be different to that of integrated plasmid DNA, perhaps it is the case that microsatellites can be markers for overall genomic instability (Kurzawski et al., 2004), but cannot predict stability of integration sites specifically. Furthermore, six microsatellites may not be enough to confidently assess instability at the base-pair level for a whole genome. This study showed that microsatellites adeptly showed the relatedness between different cell lines, as demonstrated by Yu et al., (2015). The allele frequency distribution plot (Figure 3.5) is the best illustration of this. These six microsatellites were able to characterise the ten cell lines through the adherence to frequency distribution clusters. Perhaps the use of more microsatellites would enable an exclusive identification pattern for each cell line. Overall, this study has highlighted the ways in which microsatellites can be analysed for markers of genetic instability in commercial cell lines and provides a useful platform for processing of future datasets, which might be more elucidating.

3.3.2. Karyotype Analysis

Cell lines B1-B10 were also assessed for their generational differences and cell line heterogeneity at the chromosome level through karyotype analysis. Both low and high generation cell lines were shown to be heterogeneous in terms of chromosome number, exhibiting a range of 17-25 chromosomes per cell. In all cases the modal chromosome number was 19, which was the parental cell line chromosome number. Cell lines B1-B10 at low and high generations all contained chromosomes that were not present in the parental cell line. There were a total of 18 of these chromosomes generated within the cell culture period of this study. 70% of cell lines showed karyotype changes over long-term cell culture, with a total of 14 distinct karyotype changes over long-term cell culture. It is difficult to compare GCN and qP instability with this genetic instability, because this data is qualitative. The number of chromosomal changes seen here does not necessarily correlate with changes in productivity, because it is difficult to quantify the impact of any single chromosomal aberration. Therefore, it is not feasible to directly project the chromosomal changes seen here on to the phenotypic changes observed these cell lines. Indeed, the genes that are affected by these aberrations and the subsequent downstream affects are not easy to interpret. Genetic and epigenetic effects can impact gene expression when genes are moved into different genomic locations (Gordon et al., 2012).

Clearly this study has shown that these CHO cell lines are extremely unstable at the chromosome level, which is a hallmark of immortal and cancer cell lines. It has been shown that chromosomal instability begets further chromosomal instability (Duesberg et al., 1998), so it is perhaps no surprise that changes were seen over long-term cell culture in cell lines that had already undergone karyotype change. In some cases chromosomal instability has been shown to be a predecessor for gene mutation and enzyme imbalance (Duesberg et al., 1998), which could also lead to production instability over long-term cell culture. As previously stated, CHO cells were initially used for the investigation of chromosomal aberrations (Jayapal et al., 2007) and since the start of their use in industrial bioprocesses have been manipulated, engineered and evolved towards desirable phenotypes, potentially at the cost of genetic fidelity (Sinacore et al., 2000, Heller-Harrison et al., 2009). Therefore, this instability is likely to be an inherent feature of all CHO cell lines. This may contribute to production instability, because hotspots of the CHO genome for DNA double-strand breaks are more likely to be integration targets for plasmid DNA. This could be a source of instability further down the line.

It may be possible to engineer or evolve cell lines towards phenotypes that exhibit less genetic instability, but this is challenging, because the underlying mechanisms behind it are not fully understood. Practically speaking, assessments that enable the early detection of genetic instability may allow for the selection of cell lines less likely to undergo drastic genetic changes throughout the production process. Also, it has been shown here that some genomic regions including whole chromosomes, such as chromosome 1, are somewhat immune to the chromosomal instability presented here and microsatellite analysis has shown that some loci may be less prone to base pair change than others. Therefore, targeting plasmid insertion to these relatively stable regions may lead to a cell line more able to keeping its productive capacity. However, this may be a simplistic view, because recombinant protein production relies upon many genes, in terms of the level and the fidelity of their products, which are likely to be situated in different genomic loci.

3.3.3. Conclusion

This study aimed to characterise and quantify CHO cell genomic instability at the base pair and gene copy level through microsatellite analysis and at the chromosomal level using karyotype analysis, for the assessment of their validity as tools in the cell line development process to minimise phenotypic instability. Overall, significant allelic variation in microsatellites could only be attributed to genetic drift, rather than mutational change, and so in this format is not suitable for assessing global instability. Potential studies, outlined in section 3.3.4, may provide more insight into the usefulness of microsatellites for this purpose. On the other hand, karyotype analysis showed that there is substantial change at the chromosome level, both in terms of chromosome number and breakage / fusion events. This high level of chromosome instability did not directly correlate with changes in qp or GCN, but it was concluded that karyotype analysis could be used to eliminate unstable cell lines during the cell line development process.

3.3.4. Future Work

This study has shown that microsatellites may be able to be utilised as a marker for genetic instability for mismatch repair related instability, such as point mutations, insertions and deletions. However, as stated in section 3.3.1, six microsatellites cannot fully diagnose instability at this level. A more comprehensive microsatellite instability analysis of the genome may allow for an increased resolution in investigating genomic instability at this level, in which a higher number of microsatellites would be used. A large number of microsatellites would need to be identified and characterized in terms of genomic location to ensure that genome coverage is as comprehensive as is possible.

Section 3.3.1 also highlighted the difficulty in correlating a genome-wide state of instability with a locus-specific instability such as plasmid-related gene expression. Therefore, even if a large number of microsatellites were identified that covered the whole genome at an informative resolution, it would be difficult to validate their use as a genetic instability marker by investigating the stability of a single locus (i.e. an insertion site). Perhaps instead of using recombinant DNA expression to validate microsatellite instability a more global analysis using transcriptomics could be used, because logically a marker for global stability can be better verified by a global output. If transcriptomic analysis was carried out on cell lines B1-B10 at low and high generation then a quantification of overall gene expression change could be determined. If this change was to correlate with microsatellite instability using a comprehensive genome-wide array of microsatellites then this would validate microsatellite instability as a marker for gene expression instability and it would heavily indicate that mismatch repair, or a lack thereof, impacts significantly upon gene expression. From this, an array of microsatellites could be used to assess cell line instability in the developmental stages of the production process with an aim to weed out unstable candidates. Furthermore, the transcriptomic data could be used to analyse genes known to be involved in the regulation of DNA replication and its fidelity to see whether expression rates differ from what might be expected. This could lead to engineering or evolution-based strategies to generate more stable cell lines.

A stated above, the experimental format used in this study is perhaps not the best indicator of production stability, because this phenomenon is likely to be locus specific. One use of microsatellite instability analysis for a locus-specific purpose could be to design a plasmid vector carrying a recombinant protein that was also carrying microsatellites. If it is indeed a reliable marker for genetic instability, microsatellite change could be used as a tool, in this setting, to more accurately assess whether observed decline in recombinant protein production is due in any part to mismatch repair fidelity and gene loss through repeat induced recombination events. Moreover, this system could be used to assess the stability of a given integration site with the aim of identifying sites for targeted integration efforts. As well as microsatellites, this probing plasmid could be littered with other types of repetitive sequences that might better imitate the repetitive nature of a plasmid that is used commercially, such as with the GS vector system, to assess whether repetitive sequences are responsible for recombinant gene loss.

Another avenue of research could be to sequence cell lines, such as B1-B10, which show production instability to ascertain whether there is evidence of recombination-based gene loss around repetitive elements in recombinant plasmid DNA. Also, sequencing may identify point mutations in the recombinant plasmid sequence in elements that could affect gene expression (promoters and enhancers) or elements that may affect processing downstream, such as translation or protein folding. Chapter 5 provides an in depth analysis of mutation in recombinant plasmid DNA.

This study also showed that the CHO cell karyotype is extremely unstable and changeable. As stated in section 3.3.2 it is difficult to draw correlations between a changeable karyotype phenotype and changes in productivity. Each change in karyotype may have a unique impact on the productivity phenotype and so to gain a true understanding of how a given chromosomal change impacts upon recombinant protein expression then a single cell analysis of protein production is required, which could be done through techniques such as FACS single cell sorting. Perhaps if the location of genomic insertion was ascertained through methods such as fluorescent in-situ hybridization (FISH) and compared with the observed chromosomal changes then it could be established whether changes in productivity could be attributed to changes in genomic location. However, this may be a reductive theory, because changes in productivity could be a result of the changes in gene expression of other influential gene products, which are spread throughout the genome. Again, a transcriptomic analysis could assess globally for changes in gene expression for a correlation analysis with changes in productivity and it could be determined whether genes responsible for the regulation of chromosomal stability have changed in their gene expression. Indeed, sequencing may even uncover mutations in these genes.

If further analysis led to the confirmation of the conclusions in this study, that the CHO cell karyotype is unstable and could be responsible for global genetic instability, then the implementation of a high-throughput karyotyping system into the bioprocesses involved in the production of recombinant proteins may be able to be used as a predictive tool for genetic instability of cell lines with the aim of eliminating unstable candidate cell lines from the developmental process. Moreover, it may be possible to evolve or engineer cell lines towards more karyotypically stable phenotypes.

As mentioned above, chapter 5 provides an in depth analysis of point mutations in recombinant plasmid DNA. This study required the generation of stable CHO cells to acquire recombinant plasmid DNA. Therefore, it was necessary to optimise an electroporation protocol to facilitate the transfection of plasmid DNA. Preliminary analysis revealed that industry electroporation conditions could be vastly improved upon and so chapter 4 shows the DoE-based electroporation optimization carried out.

Chapter 4

Electroporation Optimisation Using DoE Methodology

4.1. Introduction

4.1.1. Chapter Summary

As mentioned in section 3.3.4, it was necessary to generate a stable CHO cell line in order to investigate the fidelity of recombinant plasmid DNA and to assess whether there is a substantial level of DNA mutation that could impact upon product quality. Preliminary experiments revealed that the standard electroporation conditions used in industry were suboptimal. Therefore, it was decided that a comprehensive optimisation process would help provide more effective electroporation conditions for the generation of stable CHO cells and could also provide a framework for future, product-specific, transfection optimisation for CHO bioprocesses.

This chapter demonstrates the effectiveness of DoE methodology for the optimisation of bioprocess-related protocols and how it offers a higher level of precision and insight as to how different parameters contribute towards the experimental output. The results showed that an increase in the level of electroporation parameters (voltage, pulse length, DNA load) increased transfection efficiency and decreased cell viability. This inverse relationship of transfection efficiency and cell viability was found to be somewhat predictive and was utilized in the optimisation process. The DoE strategy was to start with a wide range in electroporation parameters and to gradually narrow towards an optimal region of the design space. This narrow region was then experimentally tested to yield the final, optimal set of electroporation conditions. These conditions (320-26) increased transfection efficiency by ~17% compared to standard industrial conditions, without a substantial detriment to cell health. The optimal conditions could then be taken forward to generate a stable CHO cell pool. It was concluded that DoE, or other modelling methodologies, could be used in the same manner demonstrated here to quickly optimise electroporation for the generation of producing stable cell lines in a product-specific manner.

4.1.2. DoE for Electroporation Optimisation

All the variables discussed in section 1.3.6 should be considered when optimising an electroporation protocol. Different cell types and applications will have different optimal conditions for electroporation (Jordan et al., 2008). Typically, two output factors need to be maximised when optimising electroporation: Transfection efficiency, a marker of protein expression, and cell viability (Pucihar et al., 2011), which is decreased by DNA electroporation-mediated apoptosis (Shimokawa et al., 2000). There is an inverse correlation between the two, because stronger conditions will facilitate greater membrane permeabilisation (i.e. DNA entering the cell), but at a greater cost to the health and recovery of a population of cells. Therefore an optimal trade-off needs to be made to ensure the maximum transfection efficiency without compromising cell viability (Andreason and Evans, 1989). For each new biopharmaceutical product being developed, the cellular reaction to electroporation may change in terms of transfection efficiency or cell viability. For example cell types will differ in their tolerance to electroporation parameters (Jordan et al., 2008), the metabolic burden on the cell may vary from product to product (Kim et al., 2011), or vector types and sizes can be interchangeable each having a different effect on an electroporation process and gene expression (Jordan et al., 2007, Wurm, 2004). Each of these factors will impact on cell viability and transfection efficiency and so electroporation parameters could be adapted to cater for the different features of each new product – cell – vector combination. Median fluorescence will be used as a secondary measurement of gene expression in this study, which is a measure of expression intensity rather than expression by cell number. It has a more variable output than transfection efficiency and so is less reliable for comparing parameter settings. Median fluorescence is more valuable when considering transient expression systems, in which immediate high levels of expression are required. Average cell diameter (ACD) will be used as a secondary assessment of cell health, because electroporation causes cells to shrink through loss of cellular content, which is likely to be a stressor (Chang and Reese, 1990).

Typically, an optimisation procedure like this would be carried out using a one factor at a time (OFAT) approach in which one factor is varied while the others are kept constant to measure its effect on the system. All factors are independently measured in this manner. Alternatively, DoE methodology mathematically models the response in a multifactorial manner and statistically analyses the model for significance. It offers a better estimate at optimal conditions with fewer experimental runs and all factors can be tested simultaneously. Furthermore, DoE offers insight into how different factors interact within a system, which OFAT fails to do. DoE methodology is a proven tool for the optimisation of transfection methods. Two examples of which are the optimisation of PEI-mediated transfection for transient production by (Thompson et al., 2012) and the optimisation of microporation by (Madeira et al., 2010). Design Expert 9.0.4 software was used to facilitate DoE experimental design and analysis.

As described above there are many variables that contribute to the efficiency of an electroporation protocol, both within the sample itself and by the electroporation device parameters that are set. A complete DoE analysis would first assess all of these variables in a factorial design, whereby all factors would be varied simultaneously at high and low levels to determine whether they have a significant impact on the response. Subsequently, these high-impact variables would be taken forward using response surface methods (RSM) to give a three-dimensional map response in which the output can be visualised in detail. However, the number of variables that would need to be analysed by an initial factorial design is extensive. The literature (section 1.3.6) has already defined how these factors interact in a typical electroporation system and can indicate which factors have the greatest effect on transfection efficiency and cell viability (equations above). Moreover, the interactive nature of these factors means that a balance of factor levels is needed, which could be in the form of a number of optimal sets of parameters. For example, a sample with a low resistance would require a different voltage and pulse length to a sample with a high resistance. Varying sample resistance or parameter settings to calculated extents could achieve the same balance and subsequently the same transfection efficiency and cell viability. Therefore, it is unnecessary to vary all influential factors to discover an optimal output. Furthermore, in reality, the factors affecting the samples response to electroporation, including DNA vector, cell type, media and recombinant protein, will have already been carefully designed for each product. So, to apply DoE optimisation to electroporation universally, it would be more practical to optimise with electroporator parameter settings (voltage, pulse length, waveform) to achieve this balance, rather than to factor electroporation into design of sample components. Therefore, because of this logistical practicality and the level of definition electroporation already has, it was decided to proceed directly to RSM based methods of analysis with only a subset of factors.

The electroporation factors investigated in this work were voltage, pulse length, waveform and to a lesser extent, DNA load. Other factors were kept constant throughout the study. The work demonstrated in this chapter uses Central Composite Designs (CCDs) (Figure 4.1.) to model electroporation responses. In a CCD each factor is measured at two initial levels, the low factorial and high factorial, which determine the boundaries of the design space being investigated. Center points are measured repeatedly to estimate the pure error of the model, and to estimate the curvature of the responses. Two levels outside of the design space are measured for each factor to enable the model to fully estimate the quadratic nature of the system in terms of each factor individually. These are called the low and high axial factors. CCDs can only adequately model up to and including quadratic terms, because the number of experimental runs is not enough for anything higher and so leaves cubic and quartic terms aliased. The procedure for analyzing these statistical models is clearly outlined by the design expert software. Firstly, diagnostics are carried out regarding normality and suggestions are made for data transformation and data point elimination, which might lead to a more accurate interpretation of the data. A fit summary is then provided, using sequential model sum of squares (SMSS) and model summary statistics (MSS), which suggests the order of model to be used. A model is then fit and is subsequently analysed using an

ANOVA to identify the experimental factors which have a significant impact on the response variable. It also provides statistics such as: Lack of fit, which informs the user if the model fits the data to an acceptable level of statistical significance; R-squared, which informs the user of the proportion of variance in the response that can be explained by the model; The predicted R-squared, which informs the user on the accuracy of the model in terms of its predictive capacity; 'Adeq Precision', which informs the user as to whether the signal to noise ratio of the response is strong enough for the model to adequately model the design space (>4). The model response to the independent variables can then be visualized using a response surface plot. The model terms and the response plot give the user a clear idea of the type and intensity of the relationship of the independent variables and the response, and indeed whether any of the independent variables interact in terms of their relationship with the response. Lastly, the optimisation function, which uses an inbuilt desirability function within the software, can then be used to combine response models to provide the user with a final set of optimal independent variable levels to use for future use, according to the priorities and thresholds set regarding the importance of each of the independent variables. So, for example, transfection efficiency might be given a higher priority than median fluorescence and cell viability can be set at a minimum value of 65% when determining optimal conditions.

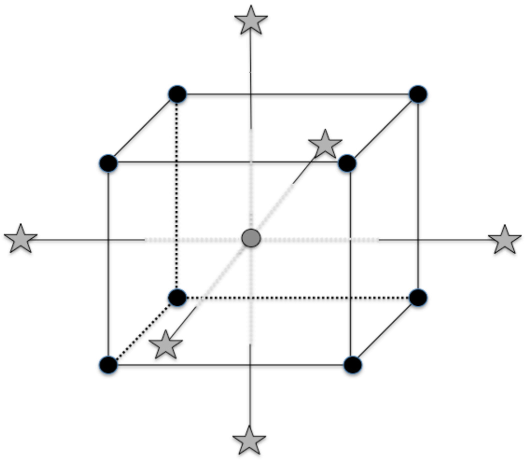


Figure 4.1. Central Composite Design
This figure illustrates a 3-factor CCD. Each dimension of the cube represents a different factor in terms of factorials (black dots), center points (grey dot) and axial points (stars). Figure adapted from Anderson and Whitcomb (2005).

4.1.3 Chapter Aims and Hypothesis

The hypotheses of the investigation were that:

- A balance would need to be met between applied voltage and pulse length to give maximal transfection efficiency whilst maintaining high cell viabilities and that these optimal parameters would vary with waveform.
- DoE methods would be able to identify a number of parameters that met this balance and, ideally, would identify those that were more optimal than others.
- DNA load would also need to be balanced in the same manner, with increased loads enabling higher transfection efficiencies with a cost to cell viability.
- The optimal parameters determined by DoE methodology would achieve higher transfection efficiencies than industrial parameter settings (Pfizer conditions).
- The optimal parameters generate would be used to generate stable CHO cell pools in future work.

4.2. Results

It was decided that the investigation would involve a succession of RSM-based experiments in which the factor level ranges would progressively narrow towards narrow optimal range. This optimal range would then be tested to ascertain the most optimal parameter settings. The phCMV-CGFP plasmid (Figure 4.2.) was linearised using restriction enzyme AflII. Gene expression responses were analysed via GFP fluorescence detection by flow cytometry in terms of transfection efficiency (percentage cells expressing GFP) and median fluorescence (level of GFP expression). Cell health measurements were taken in the form of cell viability (%) and average cell diameter (ACD) (um), which were assessed using a ViCell. All assessments were carried out 24 hours after electroporation.

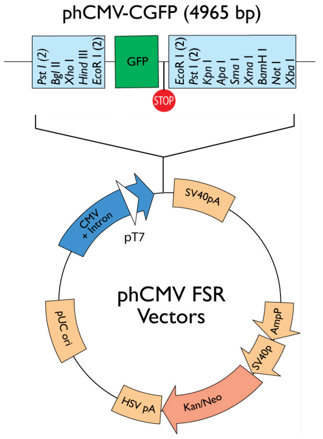


Figure 4.2. phCMV C-GFP Vector
The vector contains a GPF ORF surrounded by the plasmid multiple cloning site (MCS). The GFP ORF is flanked by a CMV promoter and an SV40 polyA tail. Genes coding for Kanamycin and Neomycin (Kan/Neo) are included for bacterial and mammalian selection respectively. The Kan/Neo open reading frame is under the AmpP and SV40p promoters and followed by the HSV PolyA tail. pUC ori is included for bacterial replication.

The factors tested were field strength, pulse length / time constant, DNA load (initial RSM only), waveform and pulse number (square wave only). Field strength is typically measured in V/cm, but will hereafter be referred to in terms of its voltage unless specifically stated (Equation 1.1. can be used to calculate actual field strength), because this is the measurement set on the electroporation device. Other electroporation variables described in section 1.3.6 were kept constant: The distance between electrodes was kept at 0.4 cm; The media and cell type were used in line with Pfizer standard protocols as described in chapter 2; DNA was suspended in TE buffer and consistently administered in 40 ul; All experiments were carried out at room temperature.

This optimisation is directed towards application in stable cell generation processes and so the responses are analysed differently to how they would be for transient expression optimisation. Clearly, for both TGE and SGE a high transfection efficiency is desirable, but it is more crucial in a TGE setting that cell recovery is fast, because of the short window for production. Whereas, with SGE a fast recovery is less crucial, because inevitably only one cell is used as a source to generate a new cell line. Therefore a greater compromise on post-electroporation cell viability was accepted here, because it would mean more vector copies have the chance to integrate with the CHO genome. In this study a cell viability lower than 50% was used as a cut off for conditions that were deemed to be too harsh (Canatella and Prausnitz, 2001).

4.2.1. Cell Number Optimisation

Cell number is another factor that affects sample resistance. A standard Pfizer electroporation protocol for generating stable cell lines involves the electroporation of 1 x 10^7 cells, whereas other protocols and instruction manuals (Terefe et al., 2008, Lonza, 2009, Bio-Rad, n.d.) describe processes using 1-2 x 10^6 cells. Clearly, this optimisation procedure needs to be catered towards bettering existing protocols for stable cell line generation in an industrial setting, but lower cell densities are more practical for enabling a high-throughput optimisation process. Therefore a preliminary experiment was carried out to test the effect of cell number, using Pfizer standard pulse settings, on transfection efficiency, median fluorescence and cell viability (Figure 4.3A, 4.3B and 4.3C respectively). One-way ANOVAs followed by Tukey's multicomparisons tests were used to call significant variation between means and pairwise variation between conditions respectively. ANOVAs showed significant variation between means for all three responses ($p < 0.0001$). There was no significant difference in transfection efficiency or median fluorescence when using 1 x 10^6 cells or 1 x 10^7 cells for electroporation. However, there was a small, but significant increase in viability when using 1 x 10^7 cells (74%) compared to 1 x 10^6 cells (69.5%) ($p < 0.05$). The cell number taken forward for subsequent experiments was 1 x 10^6 cells to enable a more high-throughput approach with the caveat that cell viability would be a slight underestimate of standard conditions. When cell number was changed to 1 x 10^6 cells, but the cell-to-DNA ratio was kept the same as Pfizer conditions by changing DNA load to 5 ug, transfection efficiency and median fluorescence were significantly ($p < 0.05$) lower (by 46.7% and ~12.8-fold respectively) and cell viability was significantly ($p < 0.05$) higher (by 10.4%) than Pfizer standard conditions. This indicates that a consistent cell to DNA ratio is not necessarily an important factor to balance, but rather that DNA concentration in a given volume is more influential. Therefore, despite the 10-fold change to cell number compared to Pfizer standard conditions, the DNA load in optimisation would be held at standard levels (50 ug).

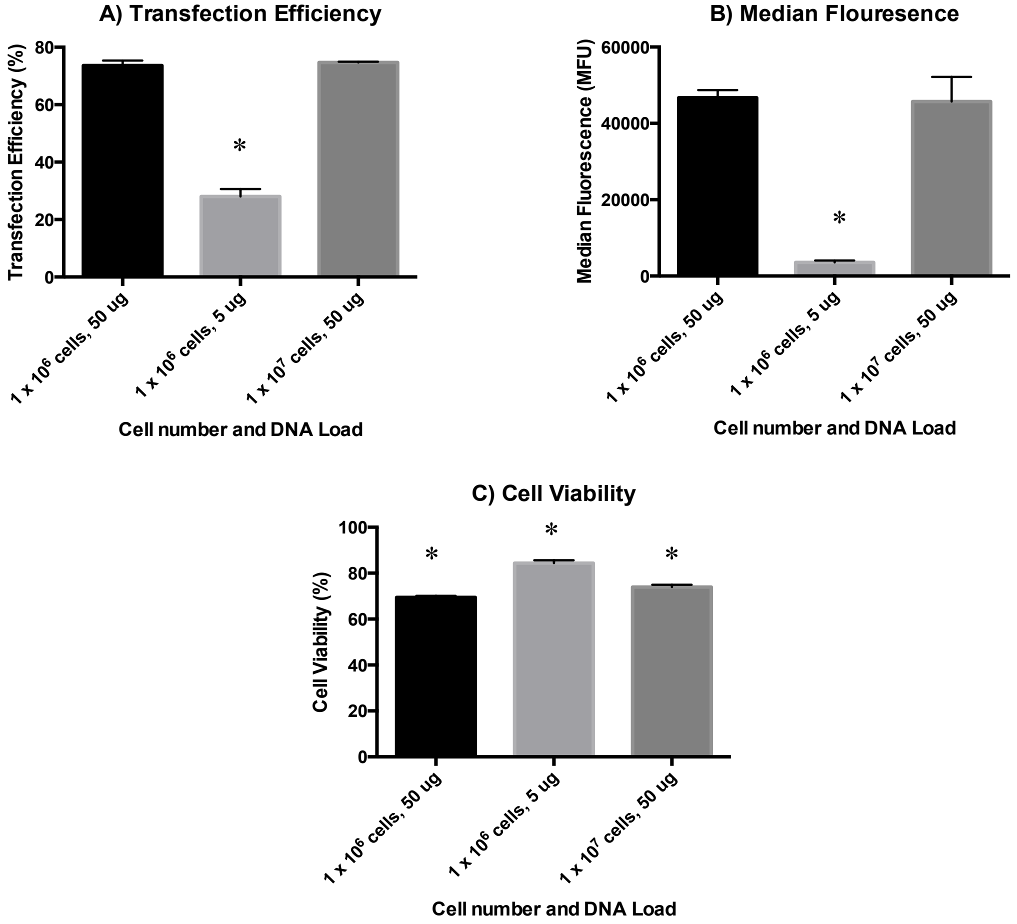


Figure 4.3. Cell Number Optimisation
Cell number (1 x 10^6 or 1 x 10^7) and DNA load (5 ug or 50 ug) were varied and responses were measured for A) Transfection efficiency, B) Median Fluorescence and C) Cell Viability. * relates to the significant differences referred to the in the text.

4.2.2. Sample volume Optimisation

Sample volume is another factor effecting sample resistance. In standard Pfizer conditions 700 ul of sample is used for electroporation, whereas the Bio-Rad gene pulser Xcell standard conditions for mammalian cell electroporation uses 400 ul (Bio-Rad, n.d.). A one-factor RSM experiment was carried out to determine the sample volume to be used in this study, in which the design space to be tested spanned between 400-800 ul (factor A). The electroporation parameters set were in line with Pfizer standard conditions. The Design Expert software analysis interface guides the user through analysis.

	Response Model Summary						
Response	Transform (λ)	Model Order	Model Terms	Lack of Fit (p)	Adjusted R^2	Predicted R^2	Adeq Precision
Transfection Efficiency	--	Cubic	A^2 A^3	0.7171	0.9458	0.8786	12.686
Cell Viability	--	Quadratic	A A^2	0.5227	0.7749	0.5513	7.101

Table 4.1: Sample Volume – Response Model Outputs
The table contains the lambda value for data transformation, the order of the model, the significant terms of the model, the lack of fit p-value of the model, the adjusted R-Squared, the predicted R-squared and the "Adec Precision".

Models were generated for the transfection efficiency and cell viability response to changes in sample volume:

Transfection Efficiency = + 65.81 + 4.23*A – 5.91*A^2 – 12.24A^3 (Coded Factors)

Cell Viability = + 82.33 + 2.51*A – 4.14*A^2 (Coded Factors)

The models were deemed to fit the data and describe an acceptable proportion of the response variance. Data residuals were normally distributed. The models and their response surfaces show (Figure 4.4 and 4.5) that an increase in sample volume results in very little change in transfection efficiency (approximately constant at 65%) until volumes exceed 650 ul, at which point transfection efficiency starts to decline. Cell viability increases with sample volume from 400 ul to 650 ul, at which point cell viability starts to decrease. Cubic and quadratic terms were the most influential for transfection efficiency and cell viability respectively. Transfection efficiency and cell viability were maximized with equal priority using the optimisation function, which recommended an optimal sample volume of 649.97 ul. Therefore, it was decided to proceed using 650 ul sample volume in future experiments. A summary of important model statistical outputs are shown in table 4.1 and further information is provided in tables A1-A4 and figures A1 and A2.

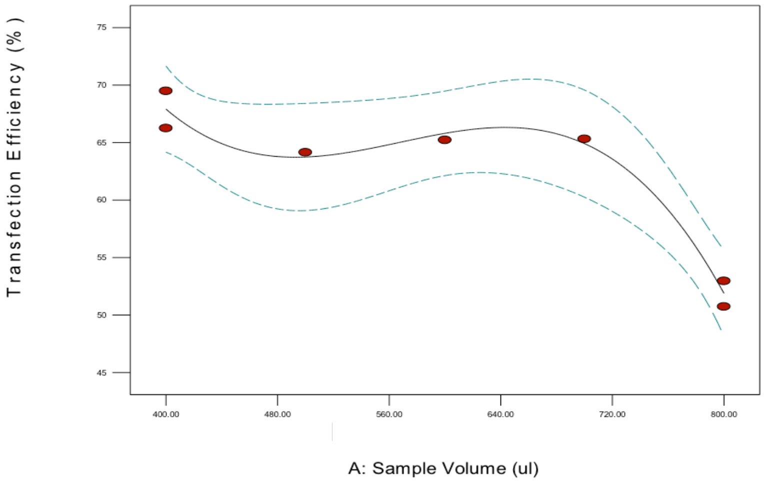


Figure 4.4. Sample Volume: Transfection Efficiency:
The graph shows transfection efficiency change with sample volume. The black line represents the data trend, red dots represent the actual data points and the blue dotted lines represent the 95% confidence range for each point.

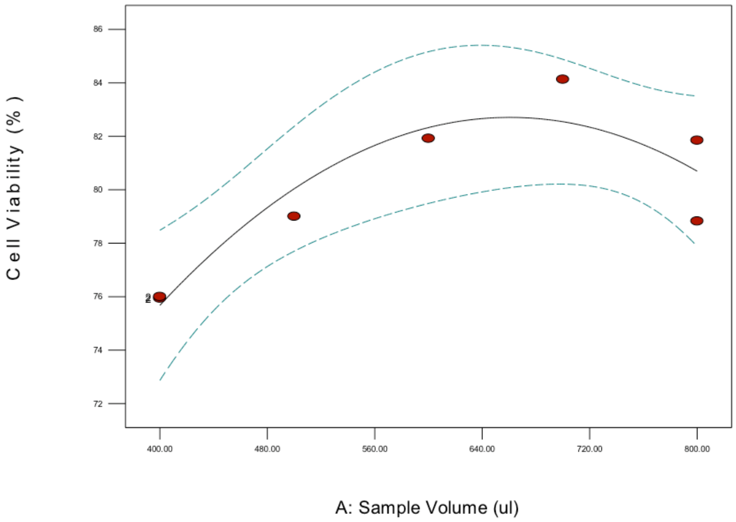


Figure 4.5. Sample volume: Cell Viability
The graph shows the cell viability response to changing sample volume. The black line represents the data trend, red dots represent the actual data points and the blue dotted lines represent the 95% confidence range for each point.

Due to the optimisation of cell number and sample volume and other factors influencing sample resistance being kept constant, we could assume that sample resistance was relatively consistent throughout the investigation. The approximate resistance for each sample was 30 ohms.

4.2.3. Electroporation Optimisation: Wide Parameters

A review of Bio-Rad protocol optimisations (Terefe et al., 2008), other literature and Pfizer standard conditions (see section 2.5.) was carried out to determine the initial electroporation parameter ranges to be investigated with the idea of starting with wide ranges in order to completely characterise how these parameters effect the CHO cell response at their extreme levels. The aim was to discover areas of the design space that yield high transfection efficiencies and gene expression, whilst maintaining a high cell viability. Voltage, pulse length and DNA load are numerical factors, whereas waveform is a categorical factor. It was decided that experiments for exponential decay and square wave electroporation would be conducted side-by-side rather than integrated into the same CCD.

4.2.3.1. Exponential Decay Wide

A three-factor (Voltage, pulse length, DNA concentration), two level, rotatable CCD was set up with the Design Expert software, inputting the axial values instead of factorial values for practicality (keeping factors above 0). This generated a 20-run experiment including 8 factorial points, 6 center points and 6 axial points. The levels for each factor are laid out in Table 4.2. The four responses modeled were transfection efficiency, median fluorescence, cell viability and ACD.

Factor	Name	Units	-1 Factorial	+1 Factorial	Center	- a	+ a
A	Field Strength	V	89.05	320.95	205	10	400
B	Pulse Length	ms	8.91	32.09	20.5	1	40
C	DNA Load	ug/mL	41.34	159	100.5	1	200

Table 4.2. Initial Exponential Decay Parameter Ranges
The table shows the parameters and their unit ranges used in the experiment, including the factorial, center and axial (a) points. '+' and '-' refer to upper and lower respectively.

The following models for the responses analysed had a significant lack of fit:

- Transfection Efficiency – F value = 28.31, p = 0.0009.
- Average Cell Diameter – F value = 19.79, p = 0.0026

This means that the model could not sufficiently describe the relationship between the experimental factors and the responses with any statistical significance. This is likely to be a result of the large range in experimental parameters investigated. In large design spaces such as this, different areas of the design space could have vastly different responses, causing the response variation to be large. With so few values within a large design space being experimentally tested this is problematic, because the experimental design does not have the resolution to model the response adequately. Moreover, only a small area of this large design space will be 'useful' for transfection, which means that responses will drastically change within this area. This means the variance is different for different areas of the design space. The center points of a CCD are used to infer the pure error of a model, but cannot do so adequately here because pure error will not be consistent (Figure 4.6.). However, these models report detection signals above a threshold that would be expected from noise alone ('Adeq Precision' statistic > 4), and still seem to explain some of the design space variance. Therefore, the models can still be used, with caution, to spot associations and indications as to how these factors impact responses. This was done through ANOVA "significance" values that were instead called as indicative or associative. Furthermore, the models were still used to derive a set of narrower parameter ranges for future experiments. A narrower parameter range, closer to the optimal range for transfection, is much more likely to be modeled effectively. This is because the center point-based estimation of pure error is more likely to be reflective of design space variance as a whole. Even though significantly fitted models were generated for median fluorescence and cell viability, it is still true that a large design space provides a low resolution analysis. It might be that these two responses are more straightforward in their relationship with the experimental factors, or that the experimentally tested values may have been positioned more optimally for these responses by chance. Due to the significant lack of fit and low resolution of these wide range CCDs, only general trends will be commented upon and model details, such as adjusted R-squared and predicted R-squared, will not be used in these analyses.

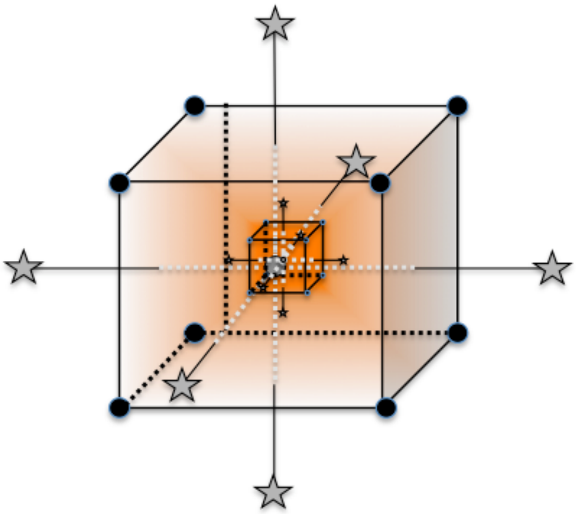


Figure 4.6. Variance Inconsistency in a Large Design Space
The schematic illustrates the inconsistency in variance over the large design space, in which intense colour represents greater variance. The smaller cube illustrates the narrower design space designed using conclusions based upon the larger design space output. This smaller design space is more likely to be consistent in response variance.

Although the accuracy is compromised by the large design space, as reflected by the lack of fit in two of the responses, there are still clear trends to be seen in the data from this CCD. Table 4.3 contains the transformation and model information for the four responses in this experiment. Transfection efficiency, median fluorescence and cell viability data were transformed according to the lambda values in table 4.3 and outliers were removed from the median fluorescence response according to advisory software diagnostics (Figure A5). The following models were generated:

$$(\text{Transfection Efficiency})^{0.69} = + 8.09 + 5.86*A + 2.48*B + 1.75*C + 2.01*AB + 1.08*AC - 0.77*BC$$

(Coded Factors)

$$(\text{Median Fluorescence})^{-0.02} = + 1.08 + 0.019*A + 7.690E\text{-}003*B + 5.688E\text{-}003*C + 7.655E\text{-}003*AB + 4.596E\text{-}003*AC + 4.308E\text{-}004*BC + 3.904E\text{-}003*A^2 - 2.695E\text{-}003*B^2 - 3.225E\text{-}003*C^2$$

(Coded factors)

$(\text{Cell Viability})^{2.86} = +4.379\text{E}+005 - 1.494\text{E}+005*A - 68511.94*B - 34325.26*C - 76478.43*AB - 35674.56*AC + 23039.14*BC - 65982.96*A^2 - 8770.35*B^2 - 5173.46*C^2$

(Coded Factors)

$\text{Average Cell Diameter} = +15.14 - 1.16*A - 0.55*B - 0.18*C - 0.86*AB - 0.22*AC + 0.11*BC - 0.73*A^2 - 0.065*B^2 - 0.026*C^2$

(Coded factors)

Response	Response Model Summary						
	Transform (λ)	Model Order	Model Terms	Lack of Fit (p)	Adjusted R^2	Predicted R^2	Adeq Precision
Transfection Efficiency	0.69	2FI	A B C AB	0.0009	0.8605	0.6307	16.498
Median Fluorescence	-0.02	Quadratic	A B C AB AC A^2 B^2 C^2	0.0732	0.9885	0.9524	43.225
Cell Viability	2.86	Quadratic	A B C AB AC A^2	0.1008	0.9685	0.8843	26.035
ACD	--	Quadratic	A B AB A^2	0.0026	0.8505	0.4142	12.103

Table 4.3 Exponential Decay: Wide – Response Model Outputs
The table contains the lambda value for data transformation, the order of the model, the significant terms of the model, the lack of fit p-value of the model, the adjusted R-Squared, the predicted R-squared and the “Adec Precision”.

Despite model inaccuracies, there are clear trends within the dataset. An increase in all independent variables leads to an increase in the gene expression responses and a decrease in cell health responses. Field strength appears to have the largest impact on the response, followed by pulse length and then DNA load. There appears to be interaction between field strength and pulse length and between field strength and DNA load. This design space yields transfection efficiencies between 1% and 80% and cell viabilities ranging from 1% to 100%. Gene expression appears to peak sharply at a specific point within the design space, with very little activity being detected below ~260 V for a duration of ~17 ms. ACD ranges from 10 um to 16 um. Because these models are not completely informative response surfaces are only included in the appendix, along with fit summary statistics, the ANOVA statistics and normal distribution plots (Tables A5-A12 and Figures A3-A10).

The numerical optimisation function of the design expert software was used as a guide to generate a narrower set of electroporation parameters to be tested. In doing this criteria can be set for each factor and response and the model will provide predictions based upon these desired criteria inputs. DNA load was kept constant at 76.92 ug/ml (50 ug – Pfizer conditions) and transfection efficiency and cell viability were both maximised with a minimum cut off of ~60%. The suggestion given was to use electroporation parameters of 309.09 V and 32.09 ms. To further analyse the cell response to a design space centering around these suggested conditions, cell viability was investigated in more detail. Voltages in a range of 260 – 400 V and pulse lengths of 27 ms, 32 ms and 37 ms were tested and viability was measured 24 hours post-electroporation (Figure 4.7).

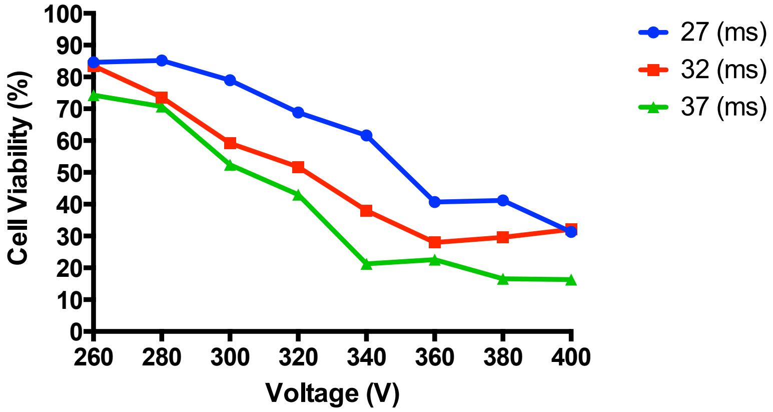


Figure 4.7. Exponential Decay: Cell Viability Optimisation
Cell viability was assessed in response to incremental changes in voltage (260-400 V) at three different pulse lengths (27, 32 and 37 ms).

So far this study has agreed with others of its kind in that transfection efficiency and cell viability have inverse responses electroporation. Therefore, it is logical to postulate that transfection efficiency and cell viability could have a direct inverse correlation, such that changes in either could predict the electroporation response of the other. For this reason it was decided that the parameter settings resulting in cell viability changes here could be used to guide new experimental parameter ranges for subsequent CCD designs. Taking the software optimisation and viability study into account, centering the next response surface model parameter around 310 V was deemed appropriate. The upper limit was set to 360 V, because this is where viabilities were below 50% at each pulse length in the viability study. Therefore the next CCD design ranged between 260-360 V in its axial points, centering around 310 V. The initial indication from the software optimisation function was to use a pulse length of ~32 ms. However, the viability plot shows that the viability around the 310 V center is lower (~55%) than model prediction at this pulse length. Therefore, the lower pulse length of 27 ms was used as the center point, spanning a range of 24-30 ms (axial points).

4.2.3.2. Square Wave Wide

Due to the issues encountered using a large design space, the wide square wave parameters were only analysed for transfection efficiency and cell viability to help derive a narrow parameter range for further analysis. Square wave electroporation offers the option to use more than one pulse and so this analysis will compare electroporation with one or two square wave pulses. All other factors had the same ranges as with the exponential decay analysis and were repeated for one and two pulses. Table 4.4 shows the levels used for each factor.

Factor	Name	Units	-1 Factorial	+1 Factorial	Center	- a	+ a
A	Field Strength	V	89.05	320.95	205	10	400
B	Pulse Length	ms	8.91	32.09	20.5	1	40
C	DNA Load	ug / mL	41.34	159	100.5	1	200
D	Pulses	Numerical	One or two pulses used				

Table 4.4. Initial Square Wave Parameter Ranges
The table states the parameters and their unit ranges used in the experiment, including the factorial, center and axial (a) points. '+' and '-' refer to upper and lower respectively.

Both the models for the transfection efficiency and cell viability responses had significant lack of fit:

- Transfection Efficiency – F value = 36.5, $p < 0.0001$
- Cell Viability – F value = 3.01, $p = 0.0418$

Therefore, as explained in the previous section, statistical significance cannot be derived from these models, but instead just associative inference.

Response	Response Model Summary Transform (λ)	Model Order	Model Terms	Lack of Fit (p)	Adjusted R^2	Predicted R^2	Adeq Precision
Transfection Efficiency	0.19	Quadratic	A C A^2 B^2 C^2	<0.0001	0.7350	0.4291	10.242
Cell Viability	2.49	Quadratic	A B C D, AB BC A^2	0.0418	0.9425	0.8795	25.974

Table 4.5 Square Wave: Wide – Response Model Outputs
The table contains the lambda value for data transformation, the order of the model, the significant terms of the model, the lack of fit p-value of the model, the adjusted R-Squared, the predicted R-squared and the "Adec Precision".

Although the accuracy is compromised by the large design space, as reflected by the lack of fit in the two responses, there are still clear trends to be seen in the data from this CCD. Table 4.5 contains the transformation and model information for the two responses in this experiment. Transfection efficiency and cell viability data were transformed according to the lambda values in table 4.5 and outliers were removed from the cell viability response according to advisory software diagnostics (Figure A13). The following models were generated:

$$(\text{Transfection Efficiency})^{0.19} = +\ 2.1 + 0.43{*}A + 0.086{*}B + 0.15{*}C + 0.036{*}D - 0.075{*}AB + 0.039{*}AC + 2.128E\text{-}003{*}AD + 0.019{*}BC - 0.011{*}BD + 5.846E\text{-}003{*}CD - 0.15{*}A^2 - 0.16{*}B^2 - 0.14{*}C^2$$

(Coded Factors)

$$(\text{Cell Viability})^{2.49} = +\ 67716.96 - 28638.97{*}A - 12326.14{*}B - 4854.42{*}C - 3331.83{*}D - 13377.86{*}AB - 2575.12{*}AC - 641.53{*}AD + 3920.6{*}BC + 1199.41{*}BD - 895.35{*}CD - 8409.94{*}A^2 + 1347.79{*}B^2 - 691.88{*}C^2$$

(Coded Factors)

As with the exponential decay CCD, transfection efficiency increases with an increase inthe electroporation parameters field strength, DNA load, and additionally, pulse number. However, transfection efficiency peaks around the midrange of pulse length delivery, which would indicate that the optimum pulse length was in the region of 20 ms. Cell viability, again, decreases with an increase in all independent variables, which is the inverse response to transfection efficiency. Within this design space, modeled transfection efficiency ranged from ~10% to 120%, which illustrates the model lack of fit, because transfection efficiency cannot exceed 100%. However, one conclusion that could be made was that 80% of the tested points in the design space that yielded high transfection efficiencies were those in which 2 pulses were administered. Therefore, square wave protocols from this point onwards would be carried out with 2 pulses only. Cell viability ranged from ~10% to 100%. The independent variables appeared to have the following order of influence: Field strength > pulse length > DNA load > pulse number. Again, because of the lack of fit if these models, the response surfaces are only included in the appendix, along with the fit summary statistics, the ANOVA statistics and normal distribution plots (Tables A13-A16 and Figures A11-A14).

The software optimisation function was used to guide the next set of experimental parameters. DNA was kept constant at 75.92 ug/mL (50ug). Transfection efficiency and cell viability were maximised with minimum threshold values of 60%. The software predicted that 302.98 V, 15.44 ms with two pulses were optimal conditions for the criteria given. As with exponential decay, a viability response study (Figure 4.8) was undertaken to probe further into these conditions. Two square wave pulses were used with field strength and pulse length ranging 260-400 V and 10-20 ms respectively. The response study was in agreement with the software in terms of voltage (~300 V for the center), but for pulse length 15 ms caused too much cell death, so an additional experimental run using 12.5 ms was used to determine a more optimal center. It was decided that a pulse length center point of 11.5 ms would be used with the prediction that its response would fall between the 10 ms and 12.5 ms responses. The axial ranges of field strength and pulse length were to be set at 271.7-328.3 V and 8-15 ms respectively.

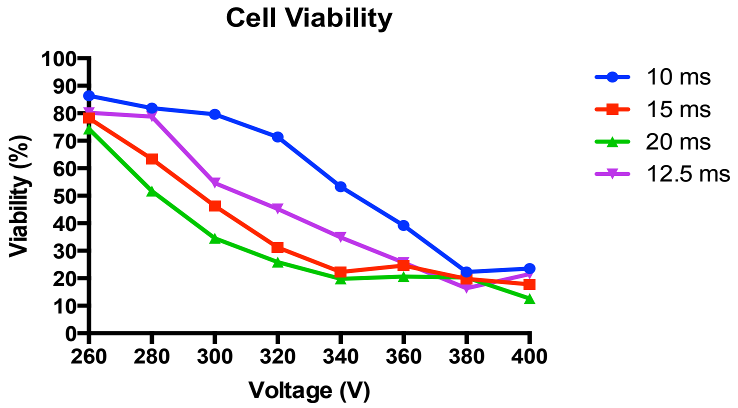


Figure 4.8. Square Wave: Cell Viability Optimisation
Cell viability was assessed in response to incremental changes in voltage (260-400 V) at four different pulse lengths (10, 12.5, 15 and 20 ms).

4.2.4. Electroporation Optimisation: Narrow Parameters

4.2.4.1. Exponential Decay Narrow - 1:

A two-factor (field strength and pulse length), two-level, rotatable CCD was devised using the parameter ranges determined in the section 4.2.3.1. DNA load was kept constant at 50 ug. The factors and their ranges are laid out in Table 4.6 This generated a 13-run experiment, measuring the four responses: transfection efficiency, median fluorescence, cell viability and ACD. The aim of this CCD was to provide insight into the electroporation parameters that yield optimal transfection responses. The hypothesis was that this parameter range would provide a higher resolution analysis around the dynamic range of optimal responses and that this set of models would better fit the data than with the previous wide range analysis.

Factor	Name	Units	-1 Factorial	+1 Factorial	Center	- a	+ a
A	Field Strength	V	274.6	345.4	310	260	360
B	Pulse Length	ms	24.88	29.12	27	24	30

Table 4.6. Exponential Decay: Narrow Parameter Ranges - 1
The table states the parameters and their unit ranges used in the experiment, including the factorial, center and axial (a) points. ‘+’ and ‘-‘ refer to upper and lower respectively.

Response	Response Model Summary						
	Transform (λ)	Model Order	Model Terms	Lack of Fit (p)	Adjusted R^2	Predicted R^2	Adeq Precision
Transfection Efficiency	3	Quadratic	A AB A^2 B^2	0.4359	0.9536	0.8887	19.046
Median Fluorescence	0.84	Quadratic	AB A^2 B^2	0.5121	0.9465	0.8717	15.38
Cell Viability	--	Quadratic	A B A^2	0.2363	0.9681	0.9071	26.43
ACD	--	Linear	A	0.8799	0.8724	0.8375	19.066

Table 4.7 Exponential Decay Narrow – 1: Response Model Outputs
The table contains the lambda value for data transformation, the order of the model, the significant terms of the model, the lack of fit p-value of the model, the adjusted R-Squared, the predicted R-squared and the "Adec Precision".

Table 4.7 contains the transformation and model information for the two responses in this experiment. Transfection efficiency and median fluorescence data were transformed according to the lambda values in table 4.7 and outliers were removed from the median fluorescence response according to advisory software diagnostics (Figure A16). The following models were generated:

$$(\text{Transfection Efficiency})^{3} = + 6.636E{+}005 - 35018.98*A + 17119.18*B - 53342.77*AB - 2.168E{+}005*A^2 - 35059.3*B^2$$

(Coded factors)

$$(\text{Median Fluorescence})^{0.84} = +24951.96 - 1628.93*A + 66.79*B - 6115.63*AB - 10220.63*A^2 - 4262.72*B^2$$

(Coded Factors)

$$\text{Cell Viability} = + 63.29 - 24.19*A - 4.73*B - 1.72*AB - 8.23*A^2 - 2.47*B^2$$

(Coded Factors)

$$\text{Average Cell Diameter} = +12.83 - 1.32*A + 0.055*B$$

The statistics displayed in table 4.7, along with supplementary information in Tables A17-A24 and Figures A15-A18 , were used to assess the models. The models predicting the four responses are all deemed to significantly fit the data (lack of fit) and explain a large proportion of variance (R^2) in the response. Statistically, the predictive capacity of these models is deemed to be high (Predicted-R^2), which validates the accuracy of the model and its usefulness in describing the response in the given design space. Transfection efficiency (Figure 4.9a) appears to increase with field strength up until ~ 305 V, at which point it starts to decrease. Field strength and pulse length interact in their effect on transfection efficiency, which is can be seen by a positive correlation of pulse length with transfection efficiency at low voltages, but a negative correlation at high voltages. Transfection efficiency ranges from ~70% to ~85% in this CCD. The same result is seen in the median fluorescence response (Figure 4.9b), which shows a peak in expression around the middle of the design space (305 V, 27 ms). The cell health responses, cell viability and ACD are both negatively correlated with field strength and cell viability is also negatively correlated with pulse length (Figures 4.9c and 4.9d) and their predicted range is from ~20% to ~80% and ~11 um to -14 um respectively according to the response surface plots. This could perhaps be the reason for the change in response-factor associations in transfection efficiency and median fluorescence responses, such that with harsher electroporation conditions the health of the cell is diminished to the point that its capacity for protein production is lessened. Again, these models indicate that field strength has a larger impact than pulse length on the electroporation response.

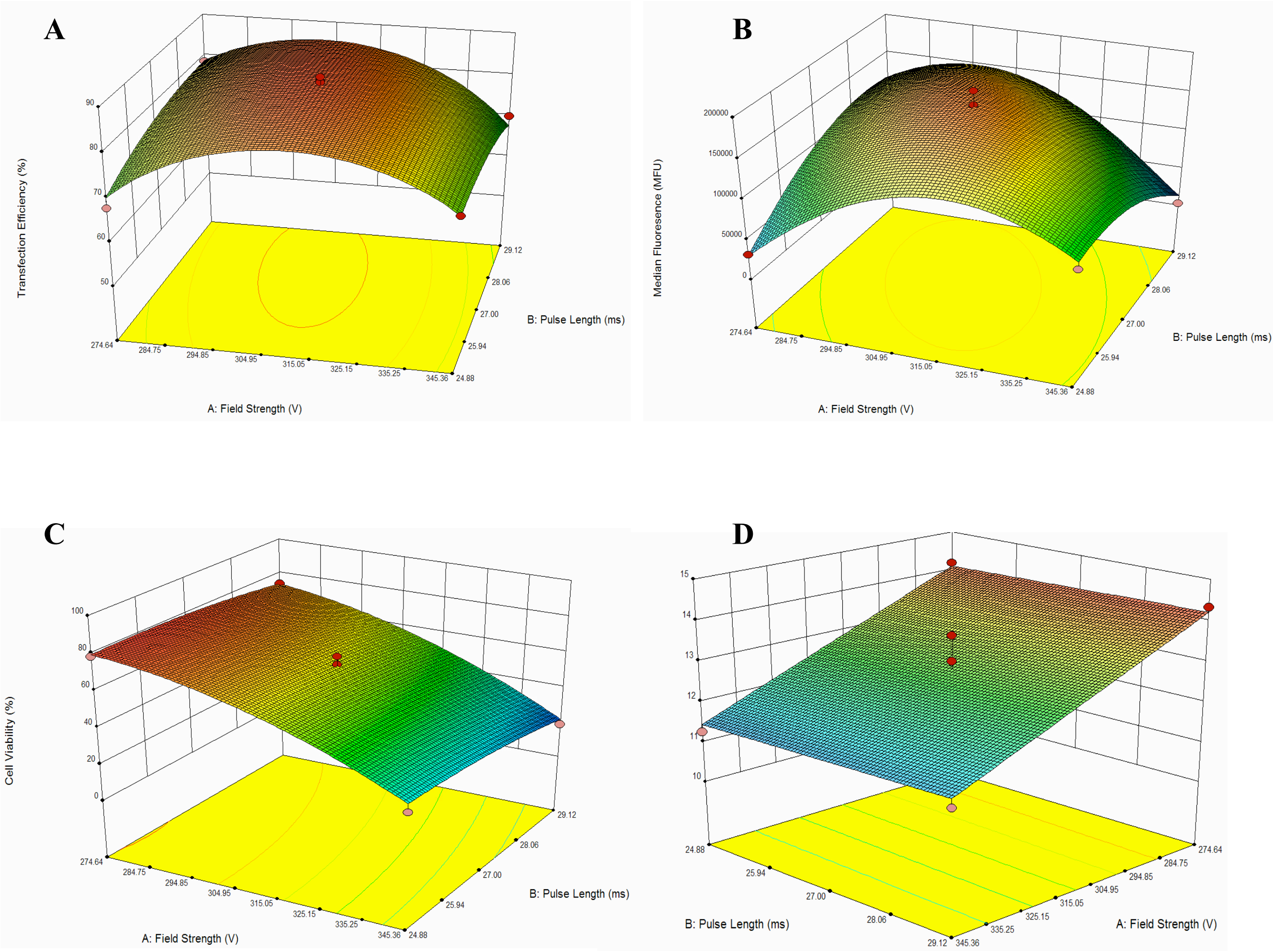


Figure 4.9. Exponential Decay: Narrow 1 –Response Surfaces
The response surface depicts the relationship between the transfection efficiency (A), median fluorescence (B), cell viability (C) and ACD (D) with both experimental factors; field strength and pulse length.

4.2.4.2. Square wave Narrow

A two-factor (field strength and pulse length), two-level, rotatable CCD was devised using the parameter ranges determined in the previous section. DNA load was kept constant at 50 ug and pulse number was kept constant at 2. The factors and their ranges are laid out in Table 4.8 This generated a 13-run experiment, measuring the four responses: transfection efficiency, median fluorescence, cell viability and ACD. The aim of this CCD was to provide insight into the electroporation parameters that yield optimal transfection. The hypothesis was that this parameter range would provide a higher resolution analysis around the dynamic range of optimal responses and that this set of models would better fit the data than with the previous wide range analysis.

Factor	Name	Units	-1 Factorial	+1 Factorial	Center	- a	+ a
A	Field Strength	V	280	320	300	271.7	328.3
B	Pulse Length	ms	11.5	9.03	11.5	8	15

Table 4.8. Square Wave: Narrow Parameter Ranges
The table states the parameters and their unit ranges used in the experiment, including the factorial, center and axial (a) points. '+' and '-' refer to upper and lower respectively.

Table 4.9 contains the transformation and model information for the two responses in this experiment. Transfection efficiency and median fluorescence data were transformed according to the lambda values in table 4.9. The following models were generated:

$$(\text{Transfection Efficiency})^{-2.55} = + 64747.49 + 15441.15*A + 11525.43*B$$

(Coded Factors)

$$(\text{Median Fluorescence})^{0.17} = + 6.78 + 0.78*A + 0.46*B - 0.051*AB - 0.16*A^2 - 0.15*B^2$$

(Coded Factors)

$$\text{Cell Viability} = + 77.93 - 12.03*A - 12.36*B - 7.36*AB - 4.5*A^2 - 6.08*B^2$$

(Coded Factors)

$$(\mathrm{ACD})^{-3} = +\ 3.943\mathrm{E}\text{-}004 + 6.706\mathrm{E}\text{-}005{*}A + 1.061\mathrm{E}\text{-}004{*}B + 8.037\mathrm{E}\text{-}005{*}AB + 2.971\mathrm{E}\text{-}005{*}A^2 + 1.424\mathrm{E}\text{-}005{*}B^2$$

(Coded Factors)

Response	Response Model Summary						
	Transform (λ)	Model Order	Model Terms	Lack of Fit (p)	Adjusted R^2	Predicted R^2	Adeq Precision
Transfection Efficiency	2.55	Linear	A B	0.2934	0.8484	0.7765	17.132
Median Fluorescence	0.17	Quadratic	A B A^2 B^2	0.4934	0.9586	0.9062	23.426
Cell Viability	--	Quadratic	A B AB A^2 B^2	0.1259	0.9565	0.8578	22.359
ACD	-3	Quadratic	A B AB A^2	0.3659	0.9678	0.9173	29.605

Table 4.9. Square Wave Narrow: Response Model Outputs
The table contains the lambda value for data transformation, the order of the model, the significant terms of the model, the lack of fit p-value of the model, the adjusted R-Squared, the predicted R-squared and the "Adec Precision".

The statistics displayed in table 4.9, along with supplementary information in tables A25-A32 and figures A19-A22, were used to assess the models. The models predicting the four responses are all deemed to significantly fit the data (lack of fit) and explain a large proportion of variance (R^2) in the response. Statistically, the predictive capacity of these models is deemed to be high (Predicted-R^2), which validates the accuracy of the model and its usefulness in describing the response in the given design space. Increases in field strength and pulse length result in higher transfection efficiency and median fluorescence, but lower cell viabilities and ACD (Figure 4.10a-d). Predicted transfection efficiency, ranged from ~65% to ~90%, cell viability ranged from ~30% to ~85%, and ACD ranged from ~11 um to ~15 um according to the response surface plots. Again field strength had larger impact upon the response than did pulse length and there was significant interaction between the two independent variables.

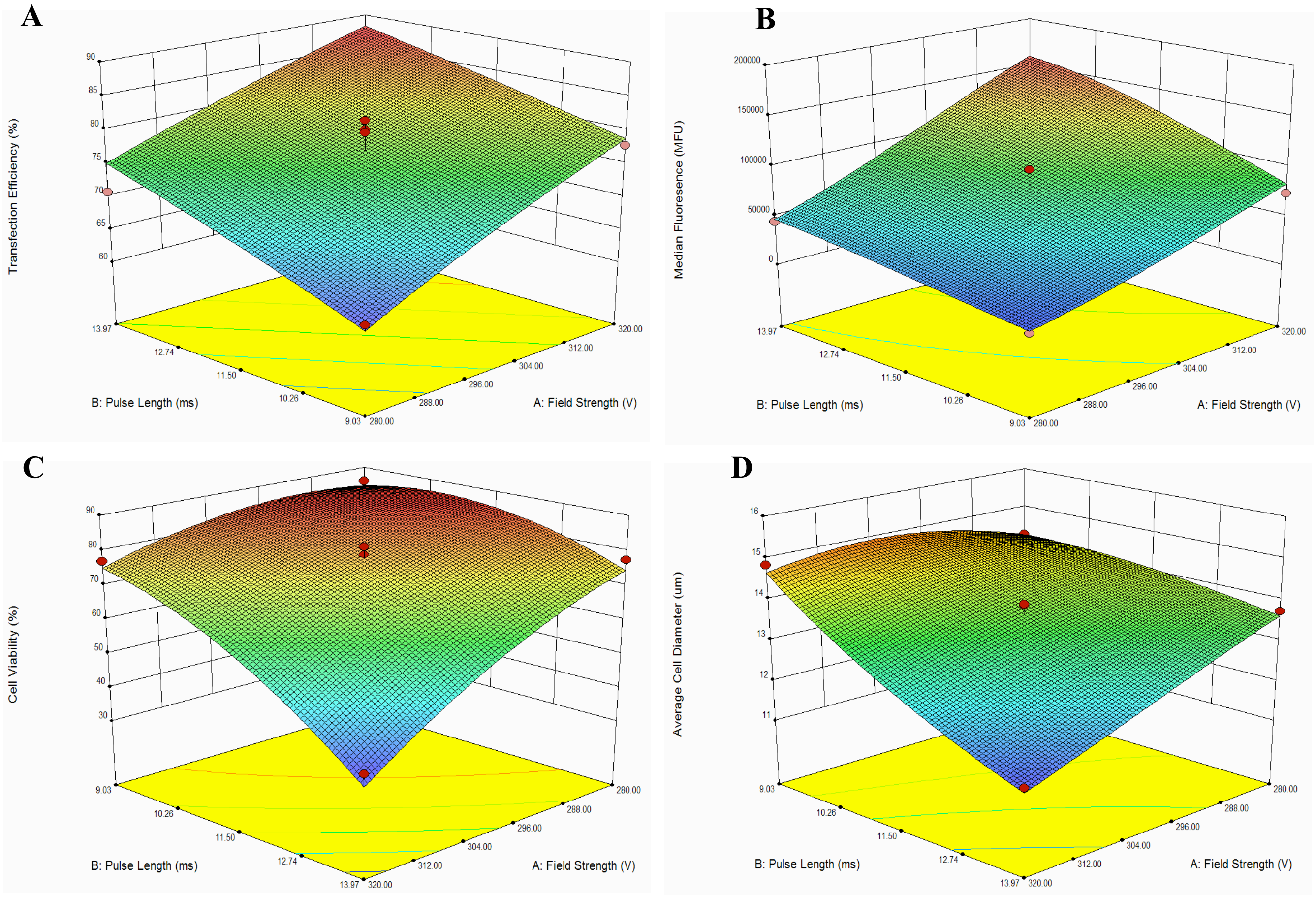


Figure 4.10. Square Wave: Narrow –Response Surfaces
The response surface depicts the relationship between the transfection efficiency (A), median fluorescence (B), cell viability (C) and ACD (D) with both experimental factors; field strength and pulse length.

4.2.4.3. Optimisation - 1

After using the optimisation function for the exponential decay narrow dataset, in which transfection efficiency and cell viability were maximised with minimum threshold values of 80% and 60% respectively, the suggested voltage for optimal responses was ~ 300 V. Given this output it was decided to run another RSM model-based experiment for exponential decay electroporation using the same voltage range as the narrow square wave experiment, because it would allow for a better comparison between the two waveforms in terms of actual voltages and data range resolution. Therefore, as well as the above criteria for transfection efficiency and cell viability, using the optimisation function, voltage was set at 300 V for optimisation to generate a new center point for pulse length. The suggested pulse length was 26-27 ms, so it was decided to center the new experiment around 300 V and 26 ms. The factors and their ranges are laid out in table 4.10.

4.2.4.4. Exponential Decay Narrow – 2

Factor	Name	Units	-1 Factorial	+1 Factorial	Center	- a	+ a
A	Field Strength	V	280	320	300	271.7	328.3
B	Pulse Length	ms	24	28	26	23.17	28.83

Table 4.10. Exponential Decay: Narrow Parameter Ranges - 2
The table states the parameters and their unit ranges used in the experiment, including the factorial, center and axial (a) points. '+' and '-' refer to upper and lower respectively.

Table 4.11 contains the transformation and model information for the two responses in this experiment. Transfection efficiency and median fluorescence data were transformed according to the lambda values in table 4.9. The following models were generated:

$$(\text{Transfection Efficiency})^{3} = + 6.016E{+}005 + 1.590E{+}005{*}A + 63563.37{*}B + 27385.12{*}AB - 30514.23{*}A^{2} - 19458.31{*}B^{2}$$

(Coded Factors)

$$(\text{Median Fluorescence})^{2.53} = + 2.637E{+}012 + 8.610E{+}011{*}A + 4.782E{+}011{*}B - 4.477E{+}010{*}AB - 4.688E{+}011{*}A^2 - 6.313E{+}011{*}B^2$$

(Coded Factors)

$$(\text{Cell Viability})^{1.5} = + 583.94 - 95.73{*}A - 81.07{*}B - 11.92{*}AB - 39.00{*}A^2 + 3.29{*}B^2$$

(Coded Factors)

$$(\text{ACD})^{1.75} = + 102.08 - 8.15{*}A - 8.64{*}B - 2.89{*}AB - 5.29{*}A^2 +0.13{*}B^2$$

(Coded Factors)

	Response Model Summary						
Response	Transform (λ)	Model Order	Model Terms	Lack of Fit (p)	Adjusted R^2	Predicted R^2	Adeq Precision
Transfection Efficiency	3	Quadratic	A B A^2	0.1143	0.9665	0.8891	27.035
Median Fluorescence	2.53	Quadratic	A B A^2 B^2	0.1931	0.8323	0.4902	9.598
Cell Viability	1.5	Quadratic	A B A^2	0.4198	0.9281	0.8248	17.748
ACD	1.75	Quadratic	A B A^2	0.4027	0.8838	0.7120	14.608

Table 4.11. Exponential Decay Narrow 2: Response Model Outputs
The table contains the lambda value for data transformation, the order of the model, the significant terms of the model, the lack of fit p-value of the model, the adjusted R-Squared, the predicted R-squared and the "Adec Precision".

The statistics displayed in table 4.11, along with supplementary information in tables A33-A40 and figures A23-A26 , were used to assess the models. The models predicting the four responses are all deemed to significantly fit the data (lack of fit) and explain a large proportion of variance (R^2) in the response. Statistically, the predictive capacity of these models is deemed to be high (Predicted-R^2), which validates the accuracy of the model and its usefulness in describing the response in the given design space. However, this was not true for median fluorescence, whereby the model had a diminished predicted R-squared (0.4902) compared to the other responses. Increases in field strength and pulse length result in higher transfection efficiency and median fluorescence, but lower cell viabilities and ACD (Figure 4.11a-d). Predicted transfection efficiency, ranged from ~70% to ~92%, cell viability ranged from ~55% to ~80%, and

ACD ranged from ~12 um to ~15 um according to the response surface plots. Again field strength had larger impact upon the response than did pulse length and there was significant interaction between the two independent variables. This lack of interaction could potentially be explained by the increase in the lower ends of the cell viability response. In previous CCDs, in which harsher conditions led to extreme lows in cell viability, the interaction between field strength and pulse length was more substantial. One explanation for this could be that low cell viabilities prohibit protein production and that the combination of high voltages and longer pulse lengths diminish cell viability to such an extent that the correlations between higher levels of electroporation and gene expression is reversed.

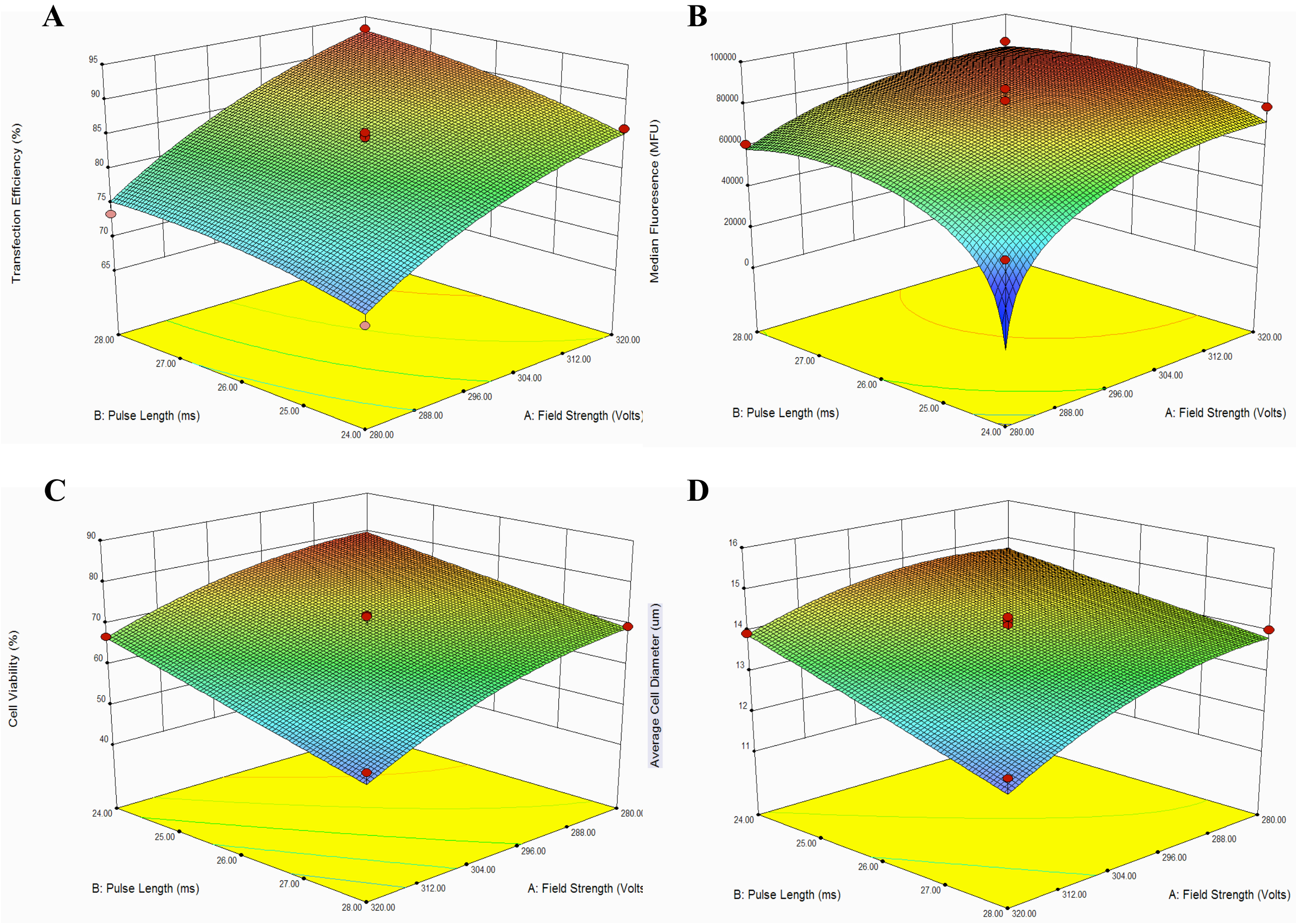


Figure 4.11. Exponential Decay: Narrow 2 –Response Surfaces
The response surface depicts the relationship between the transfection efficiency (A), median fluorescence (B), cell viability (C) and ACD (D) with both experimental factors; field strength and pulse length.

4.2.4.5. Optimisation – 2

The design expert software optimisation function was used to determine the optimal electroporation conditions for exponential decay and square wave waveforms. Transfection efficiency was maximised, with a minimum threshold value of 80% and viability was targeted towards 65% with a minimum threshold value of 60%. For the exponential decay waveform the optimisation function suggested using 310.8 V and 25.9 ms for field strength and pulse length respectively. The software predicted a transfection efficiency of 87.7% and cell viability of 65% using these conditions. For the square wave waveform the optimisation function suggested using 320 V and 11 ms for field strength and pulse length respectively. The software predicted a transfection efficiency of 82.8% and a cell viability of 65% using these conditions. A second set of criteria, in which cell viability was sacrificed for higher transfection efficiency, was then tested. Transfection efficiency was maximised with a minimum threshold value of 90% and cell viability targeted towards 55% with a minimum threshold value of 50%. For the exponential decay waveform the software suggested using 317.8 V and 27.3 ms for field strength and pulse length respectively. A prediction of 91.6% transfection efficiency and 55% cell viability was given for these criteria. For the square wave waveform no solutions were offered when using these criteria. The highest achievable predicted transfection efficiency with this viability setting was 86%, when using 320 V and 12.73 ms for field strength and pulse length respectively. Therefore, it was concluded that the exponential decay waveform was better suited for this platform and would be taken forward for use in future experiments.

4.2.5. Optimal Electroporation Conditions Testing

Design of Experiments software allowed for a complete dissection of the electroporation response across a wide range of parameter settings and led to the elucidation of parameter settings, which were predicted to result in highly efficient transfection. However, model predictions are not sufficient and outputs need to be tested. The software optimisation function was used to predict transfection responses using two sets of criteria for exponential decay electroporation:

1. > 80% transfection efficiency and 65% cell viability = ~310 V and ~26 ms.
2. > 90% transfection efficiency and 55% cell viability = ~318 V and ~27.5 ms.

In order to test the model and decide upon optimal conditions to take forward a more traditional OFAT approach was used to investigate this small range of parameter settings. Field strengths of 310 V, 315 V and 320 V with pulse lengths of 25 ms, 26 ms, 27 ms and 28 ms were tested. It was also important to experimentally test this range of settings, because the resolution of the electroporation device is such that it cannot precisely achieve input settings and exact conditions can vary from sample to sample. Therefore the predicted optimum settings may differ from the actual optimum settings. In terms of the fluorescence characteristics, transfection efficiency (Figure 4.12a) and median fluorescence (Figure 4.12b), there appears to be a general upward trend with increased electroporation strength, with the 320 V – 26 ms setting (hereafter referred to as 320-26) transfecting the most cells and having the highest gene expression. Fluorescence characteristics decrease with harsher settings than 320-26. A one-way ANOVA showed the differences among the means were statistically significant for both transfection efficiency and median fluorescence (both < 0.0001) and a Tukey's multiple comparisons test showed that 320-26 was significantly superior to all other conditions (all $p < 0.05$). In terms of cell health measurements, cell viability and average cell diameter, there does not appear to be a significant trend in the results. A one-way ANOVA showed the differences among the means were statistically significant for both cell viability and ACD (both $p < 0.0001$), but a Tukey's multiple comparisons test showed that there is no significant difference between the 320-26 and any other condition (all $p > 0.05$). Therefore, there is no cell health disadvantage to using these superior transfection conditions, which had an average transfection efficiency of 93.7% and average cell viability of 56.3%. Interestingly, cell viability only decreases by ~10% for mock transfection compared to the negative control, which means that DNA is the main contributor to low cell viability rather than electroporation intensity itself.

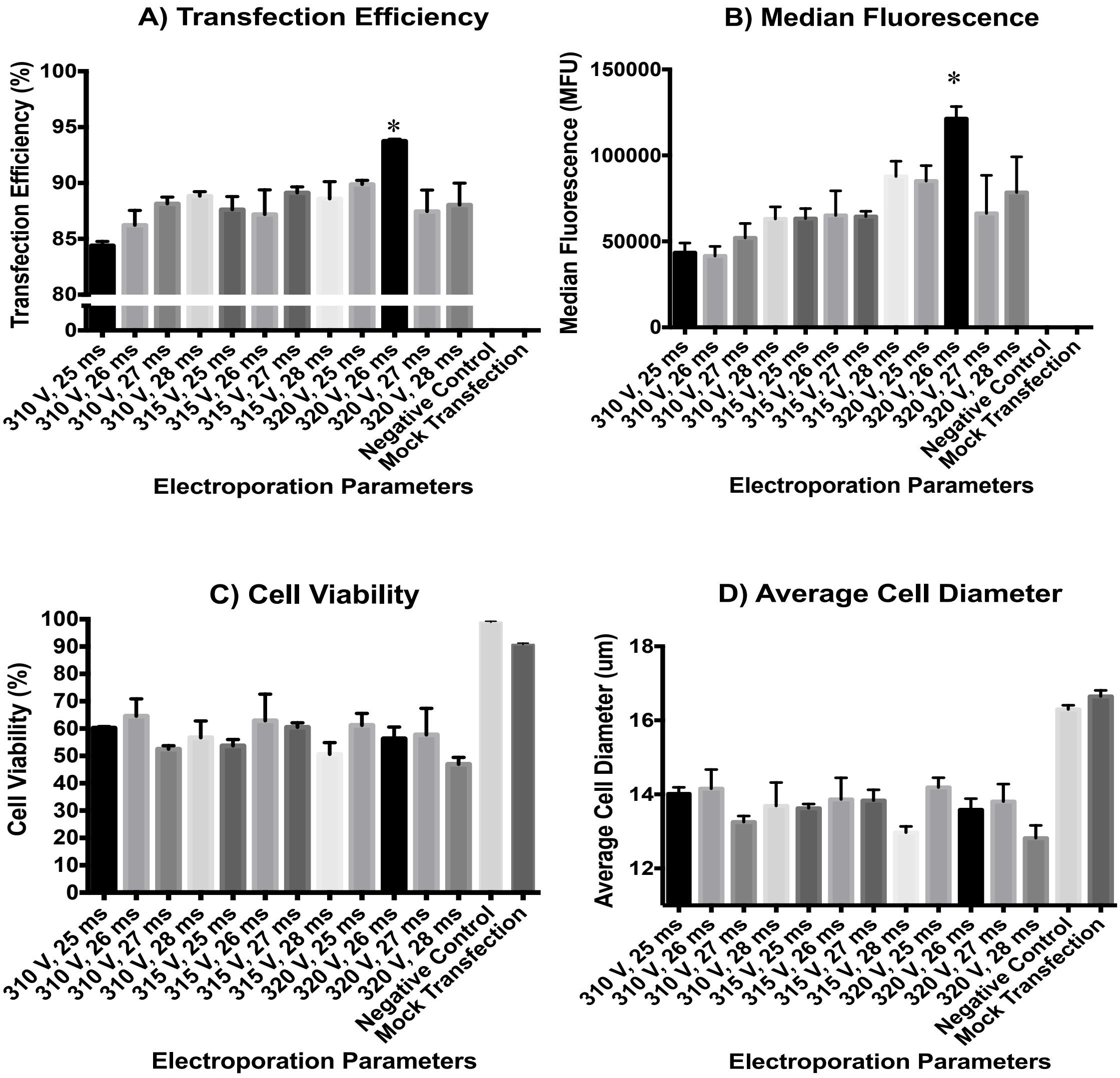


Figure 4.12. Electroporation Optimal Range OFAT

The figure shows the electroporation responses in terms of A) Transfection Efficiency (Y-axis altered for clarity between bars), B) Median Fluorescence, C) Cell Viability, and D) Average Cell Diameter. The field strengths tested were 310 V, 315 V and 320 V. The pulse lengths tested were 25 ms, 26 ms, 27 ms and 28 ms. Mock transfections were run at the harshest condition (320 V, 28 ms). * relates to a significant difference of the 320-26 condition.

320-26 was then compared to Pfizer conditions and to conditions used for transfection in the generation of stable cell lines, using 1 x 10^7 cells instead of 1 x 10^6 cells (referred to as '320-26 scaled-up' conditions). Each response was analysed by an ANOVA followed by a Tukey's multiple comparisons test (significant when $p < 0.05$). All responses had a significant difference between sample means ($p < 0.0001$ for transfection efficiency, cell viability and ACD, $p = 0.0001$ for median fluorescence). There was no significant difference in transfection efficiency (Figure 4.13a) between 320-26 and 320-26 scaled-up conditions and both 320-26 conditions were significantly higher than Pfizer conditions (~75%) by ~17%. There was a significant decrease in median fluorescence (Figure 4.13b) between 320-26 and 320-26 scaled-up conditions (0.7-fold), but both were significantly higher than with Pfizer conditions (3.6-fold and 2.5-fold respectively). Cell viability (Figure 4.13c) for 320-26 scaled-up conditions (75.8%) and Pfizer conditions (82.3) are both significantly higher than 320-26 (64%), but are not significantly different from one another, meaning the increase in transfection efficiency does not come at the cost of decreased cell viability compared to Pfizer conditions. ACD (Figure 4.13d) is significantly increased in 320-26 scaled-up conditions (14.4 um) compared to 320-26 (13.3 um) and significantly lower than with Pfizer conditions (15.1 um). Therefore, despite not having a significant difference in cell viability from Pfizer conditions, using scaled-up 320-26 does appear to have a significant physiological impact on the cell, which may indicate a slight decrease in cell health compared to Pfizer conditions.

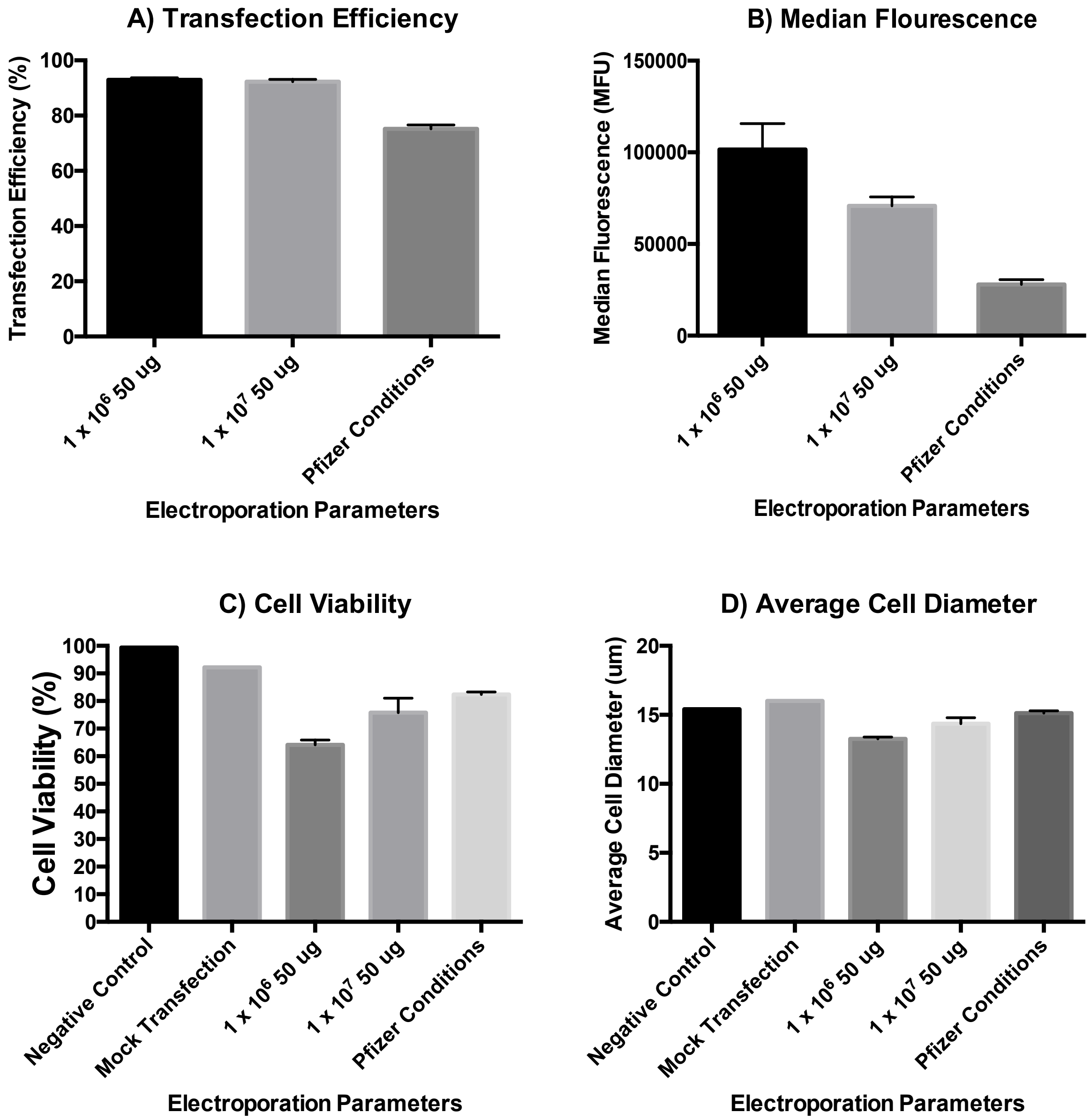


Figure 4.13. 320-26 Scale-up and Pfizer Conditions Comparison
This figure illustrates the differences in electroporation responses for 320-26 scale up to the stable cell line-generating cell number (1 x 10^7) and compares these optimised conditions to Pfizer conditions. The responses analysed are A) transfection efficiency, B) median fluorescence, C) cell viability, and D) ACD. Mock transfections were electroporated using 320-26.

4.3. Discussion

The aim of this chapter was to investigate an industrial cell line response to varying electroporation conditions in order to improve industrial standard electroporation conditions. Field strength, pulse length, waveform and initially DNA load and pulse number (square wave only) were the variable factors tested. The response was analysed in the form of transfection efficiency (percentage of cells producing GFP), median fluorescence (intensity of GFP expression), cell viability (percentage of live cells), and ACD (a marker of physiological stress). The hypothesis was that DoE methodologies would provide a more complete analysis of electroporation and as a result would be more able to identify the optimal dynamic range of activity for the generation of optimal electroporation protocols. Firstly, the results will be discussed in terms of the success of using DoE methodologies for optimisation purposes and then, in terms of the effects of the experimental factors on the cell response.

4.3.1. DoE in Process Optimisation

The Design Expert 9.0.4 software package offers an easy to use interface for mathematical modeling of a predefined design space, in which a response is measured against all experimental factors simultaneously (Anderson and Whitcomb, 2005, Anderson, 2007). This study utilised CCDs whereby two levels for each factor were tested, which defines the limits of the design space. A central point is repeated multiple times in order to estimate the pure error of the output and provides information on response curvature. Axial points (values outside of the design space) are also tested to provide more information on the response within the design space, which provides factor-specific information on response curvature. The output provides information on factor interaction as well as individual factor effects in the form of a response surface that can be visualised in 3D and statistically validated. This utilises the information provided by experimental runs and the subsequent model to provide a prediction for individual responses throughout the whole design space. The software optimisation function can then be used to integrate the response models and suggest optimal parameter settings based on criteria of the users choosing (Anderson and Whitcomb, 2005, Anderson, 2007).

As was seen in this work with the initial wide parameter setting experiments, DoE methodologies are not completely accurate when faced with a large design space. This is likely to be due to two reasons. Firstly, a relatively small proportion of the design space showed high levels of activity in terms of transfection responses. A large design space means it is likely that this activity will not be described accurately, because experimentally tested values are too far apart to provide a high-resolution analysis.

Secondly, the center points are used to estimate the pure error of the whole design space. If there are pockets of the design space that show more activity than other areas, then pure error will not be consistent throughout and thus cannot be estimated accurately from testing only one point (Box and Draper, 1959). These traits of a large design space clearly had a large impact on the model lack of fit in this study. Despite the fact that such models were defined statistically as being ineffective predictive tools, the outputs were still able to sufficiently guide the next set of experiments in the form of narrow range CCDs. This guidance was tweaked via assistance from viability response experiments. It was found that the optimisation function suggested settings that were too strong, causing more cell death than the models had predicted, and so suggested pulse lengths were altered accordingly for subsequent narrow CCD experiments. The error estimation and resolution problems faced in the initial experiments were apparently minimised in these narrow range design spaces, such that all subsequent models were deemed to fit the data well, statistically. For square wave electroporation the wide CCD output and cell viability responses study generated narrow parameter ranges that seemed to describe the dynamic range of activity for electroporation with reasonable resolution. For exponential decay electroporation, two narrow range CCDs were needed. The CCD output from wide parameter settings and the cell viability response study provided slightly wider parameter settings than with square wave electroporation. Therefore, the initial attempt at a narrow parameter range was used as a guide to generate a second narrow range CCD, which had a similar resolution to the square wave experiment. Investigating each waveform using similar parameter range settings enabled better comparison between the two. Two sets of criteria, in which cell viability was sacrificed to varying degrees for increased transfection efficiency, were then used to generate a final set of parameter settings to be tested using a OFAT approach. These final settings were all capable of higher transfection efficiencies than with Pfizer standard conditions and one setting was identified as being significantly better than the others (320-26).

Clearly, this study shows the benefits of using DoE-based modeling for process optimisation. It provides information on factor-response relationships in the form of relationship order and factor interaction and it does this with fewer experimental runs than would be needed with a OFAT approach. Moreover, the study delivered a new parameter setting, which was a significant (~17%) improvement on Pfizer industrial standard settings, which arguably could not have been done using OFAT methodology, especially in this timeframe. However, as stated previously, this approach is not without its limitations. Pure error must be consistent throughout the design space and the sparse distribution of experimentally tested points in large design spaces could mean useful information is missed. Moreover, the strategy used in this study involved the generation of CCDs, in which ranges became progressively narrower. Whilst, this resulted in successful optimisation, a strategy in which all of this data could be included in a single model might be more informative. Furthermore, using an iterative model that could be built upon and fine tuned with further data might increase its predictive capacity and would mean the model could be more applicable to optimisation processes involving different components. This will be discussed further in section 4.3.3.

4.3.2. The Electroporation Response

Before starting the optimisation process it was important to ensure that all factors that could impact on electroporation responses were kept constant at appropriate levels. These were factors that could impact on sample resistance. For the most part conditions for these variables were ascertained from Bio-rad protocols (Bio-Rad, n.d.) and Pfizer standard conditions. However, these sources gave contradictory information on cell number and sample volume and so it was necessary for them to be optimised before proceeding. It was found that a ten-fold decrease cell density did not have a significant impact on electroporation responses besides causing a slight decrease in cell viability and so it was decided to proceed using 1 x 10^6 cells to enable more high-throughput experiments to be a carried out. This cell number was then tested with a ten-fold decrease in DNA load, causing transfection efficiency and cell viability to have a significant decrease and increase respectively. This shows that the cell-to-DNA ratio is not an important factor to keep constant for electroporation, but rather the concentration of DNA in a given sample volume. However, as stated in section 1.3.6, the cell membrane interacts with DNA and DNA then actively enters the cell by electrophoresis. Therefore, fewer cells provide less membrane surface for DNA to interact with and may result in a greater number of DNA molecules interacting with each cell. This is supported by figure 4.13B, in which median fluorescence is higher at a lower cell density. A greater number of DNA molecules per cell could lead to a greater number of integration events per cell, which would be advantageous in electroporation procedures for the generation of high-producing stable cell lines. So perhaps the use of lower cell numbers would better suit the needs of industrial bioprocesses. It was also found that a lower sample volume decreased cell viability and did not affect transfection efficiency until sample volume exceeded ~ 700 ul, with slight peaks at ~ 400 ul and ~ 550 ul. So it was decided to proceed with a sample volume of 650 ul.

The DoE results throughout generally agreed with the hypothesis that stronger electroporation conditions and higher DNA loads are positively correlated with transfection efficiency and high gene expression, but negatively correlated with cell viability and average cell diameter. As expected, for optimisation of electroporation, a tradeoff was needed between DNA entry to the cell and cell health (Andreason and Evans, 1989). Field strength and pulse length are the experimental factors that control the intensity of electric charge delivered to the sample and they do this in different ways. Field strength controls the membrane surface area that becomes permeabilised and pulse length and pulse number control the extent of permeabilisation within this area (Gehl, 2003, Escoffre et al., 2009). Clearly, both an increased permeabilised area and extent of permeabilisation will be influential in transfection. A combination of these two factors is needed to facilitate successful transfection of DNA. Both of these factors impact on the transfection response by altering the plasma membrane, which means they are linked. Therefore a balance needs to be found to ensure their additive effect is not too harsh. Indeed, this study confirmed their interaction through modeling. Generally, field strength had a larger effect on the responses than pulse length, meaning that the surface area of permeabilisation is more influential in transfection than permeabilisation intensity. However, permeabilisation intensity, or more specifically pore diameter, must reach a certain level to facilitate the transfection of DNA molecules of a particular size, so pulse length must remain high enough for transfection to occur. For these reasons field strength was considered a higher priority and pulse length was optimised around it. Transfection for the purposes of stable gene expression is less concerned with immediate cell viability, because the cell population is given time to recover. Subsequently, desirable cells are selected for clonal cell line generation. Perhaps with transient gene expression, in which cells are needed to be actively producing recombinant protein sooner, a higher priority would be given to pulse length and pulse number to ensure higher immediate cell viabilities. However, transient electroporation would be with circular DNA, which is not as difficult to transfect

(Schmidt et al., 2004), and so conditions are not required to be as strong to reach high transfection efficiencies.

As expected, DNA load was positively correlated with gene expression and showed toxicity to CHO cells (Winterbourne et al., 1988). Moreover, it was shown to interact with field strength (area of cell permeabilisation), which supports the idea that reducing cell number may increase the number of DNA molecules entering the cell and, subsequently, integration events. However, it is difficult to draw firm conclusions for the effect of DNA load, because it was only investigated in a wide design space, in which models were not fit with significance. Analysis of the effect of pulse number on square wave electroporation showed no significant relationship with transfection efficiency. However, the response surface and individual data points indicated that two pulses led to cells having a higher transfection efficiency than with one pulse. Pulse number was negatively correlated with cell viability. Again, it is difficult to draw firm conclusions on pulse number from this study, because it was only investigated in the large design space.

The final CCDs for exponential decay and square wave electroporation allowed for a direct comparison of the two waveforms through the use of models that significantly fit the data. Two criteria were used for optimisation: maximising transfection efficiency with a minimum threshold value of 80%, whilst targeting cell viability to 65% with a minimum threshold value of 60%, and; maximising transfection efficiency with a minimum threshold value of 90%, whilst targeting cell viability to 55% with a minimum threshold value of 50%. The results showed that exponential decay was the superior waveform in this study, predicting a peak transfection efficiency of 91.6% using 317.8 V and 27.3 ms. A narrow range of values derived from this final exponential decay CCD were experimentally tested to ascertain which electroporation parameter setting was optimal. All of these parameter settings achieved transfection efficiencies > 84%. The 320-26 condition achieved 93.7% transfection efficiency, only 2.1% higher than model prediction, which indicates that model prediction was accurate. Median fluorescence was also significantly higher than other parameter settings when using this condition. There appeared to be no significant differences in cell health responses when testing these settings, which indicates that there is no health disadvantage in using this condition. When the 320-26 parameter settings were applied to 1 x 10^7 cells (as used in stable transfection), there was no significant change in transfection efficiency and cell health characteristics were marginally improved. The optimised conditions were superior to Pfizer conditions.

This study agrees with the literature in terms of the inverse correlation between transfection efficiency and cell viability (Andreason and Evans, 1989). It was shown that the transfection response shows a dramatic increase in activity when experimental factors reached a certain threshold level. These thresholds were approximately the same for both transfection efficiency and cell viability (260 V, 17 ms for exponential decay electroporation). Therefore, it could be likely that transfection efficiency and cell viability are extremely linked and that their inverse relationship could be used as a predictive tool. This was the case in the first round of optimisations, whereby a cell viability response study helped provide a set of electroporation parameters for the next set of experiments. The resulting design spaces covered the dynamic range of optimal transfection efficiency well. Therefore, in this case, cell viability was an accurate predictor of optimal transfection efficiency. If the relationship between transfection efficiency and cell viability were to be more thoroughly defined then cell viability may be able to be used as a predictor of transfection efficiency in the optimisation of electroporation for new expression systems. This would be advantageous, because it would greatly reduce the workload in an electroporation optimisation procedure by minimising the need for protein expression assays. If electroporation optimisation procedures were to be implemented into the development of new biopharmaceuticals it would increase the number of plasmid copies entering the host cell. In the case of stable cell line generation this is advantageous, because it could increase the number of integration events and subsequently the number of high producing clones detected in screening. Therefore, optimised electroporation might lead to integration of more plasmid copies in to desirable genomic locations. In the case of transient gene expression, conditions could be discovered that may increase gene expression without having a diminishing effect on cell viability.

4.3.3. Future Work

This study provides evidence that industrial standard conditions for electroporation can be vastly improved by using modeling approaches. Moreover, these approaches allow for a more global explanation as to how electroporation factors interact in their impact on the cell response. However, the optimised conditions derived here are likely to be unique to this system. So for these optimisation strategies to find commercial application, the conclusions found here need to be consistent across all potential permutations of the bioprocesses. For example, a change in vector size would impact on plasmid DNA entry into the cell via electrophoresis. Larger vectors may need stronger electroporation conditions to enter the cell, which may come at the cost of decreased cell viability. Also, different vector designs will be capable of variable levels of gene expression, which would impact upon transfection efficiency and the level of gene expression per cell. The recombinant product will also impact on the optimisation process. Some products will be more of a metabolic burden than others, which will impact on cell viability and growth. Whereas, other proteins are more difficult to express, which will impact on gene expression capabilities of a given system. In addition, different cell lines are likely to have different reactions to electroporation, in terms of gene expression and health characteristics and so may need to be uniquely optimised.

For further investigation of electroporation parameter settings, analyses similar to those carried out in this study should be conducted, but with more experimentally tested points to provide a higher resolution analysis. Indeed, higher levels of experimental repetition and an increase in the number of experimentally tested points would increase experimental accuracy and the estimation of systematic error. In particular, square wave electroporation and DNA load in this study were not tested fully and a more detailed study may reveal that altering their input values would have a positive impact on transfection. Indeed, it could be the case that different bioprocess conditions (as mentioned above) might be more suited to electroporation settings that are different to what would be predicted by this study.

As mentioned in section 4.2.1, the DoE methodology used with the design expert software may not be the most efficient and informative modeling method to do this, because a model which cannot be integrated or built upon is limited. Moreover, as described in section 4.1.2 and shown in these results, the factors influencing transfection are interactive and so a single model that fully integrates all variable aspects of electroporation and bioprocess variations would be more informative and have a higher accuracy in predicting optimal parameters across the complete range of bioprocess needs. By utilising a modeling strategy that is open ended, the predicted response across the entire design space could be experimentally tested and the results fed back into the model to improve it. This data-rich and iterative process would lead to a more accurate, experimentally tested and predictive model. This is something the design expert software is unable to do.

All of these variations would paint a detailed picture as to how each factor in a new biopharmaceutical system might affect the electroporation response. When a new product is being developed a new combination of cell line, vector and product type will need to be tested. The model would use this information to provide predicted optimal electroporation conditions and responses to them. Then an experimental test, centered around this prediction, would be carried out to ascertain the actual optimal set of electroporation parameters. Each new product that was tested would provide more data for the model to improve is accuracy. Eventually, a database of information could be generated, containing electroporation responses to all previous permutations of the bioprocess, which could serve as a useful repository for future optimisations.

A fully integrative model such as this would provide a complete analysis into how bioprocess factors and electroporation parameters interact, providing useful insight into their relationships. Arguably, the most useful definition generated by the model could be the relationship between transfection efficiency and cell viability. If this relationship were to be accurately defined then only cell viability may need to be measured in an optimisation process. An end product for this modeling system could be in the form of an electroporation 96-well plate, in which cells are tested with new vectors and products against many electroporation conditions for a high resolution assessment. The viability response to these conditions could then be used to predict the relative gene expression response and provide optimal parameter settings.

As mentioned at the start of the chapter, the primary purpose of this chapter was to optimise an in-house electroporation protocol in order to generate a stable GFP CHO cell pool and so the immediate future work to be carried out is to generate these stable pools and carry out mutational analysis on recombinant plasmid DNA.

Chapter 5

Plasmid DNA Mutation Analysis

5.1. Introduction

5.1.1 Chapter Summary

This chapter takes a different approach to investigating the genetic instability phenomenon described in CHO cells. In chapter 3, two whole-genome methods were used to analyse CHO cell genomic instability at the base pair, gene copy and chromosomal level. The work carried out was not able to validate microsatellite analysis as a potential marker for instability detection at the base pair level. Even though further work with microsatellites might have yielded more informative results it was decided to use DNA sequencing as a tool to measure base pair change directly. Despite the importance of CHO cell whole-genome stability this chapter focuses on the fidelity of recombinant DNA specifically and the potential threat of sequence variants to product quality, as well as providing a commentary on base pair change as a whole.

As will be described in detail in the next section sequence variants, resulting from non-synonymous DNA mutations are a threat to product quality and, in some cases, are estimated to be present in approximately a quarter of protein-producing clones. The aim of this chapter was to develop a secondary analysis tool for PacBio SMRT sequencing, which would enable a higher sensitivity in mutation calling compared to the sensitivity being reported in the literature. This DNA sequencing platform would then be used to sequence plasmid DNA at various points in the process for generating stable GFP pools, which previously has only been carried out on clonal or nearly clonal cell populations. This would enable a more comprehensive characterization and of the frequency, type and biases of the mutation that is seen in recombinant DNA.

Development of the secondary analysis platform for SMRT sequencing allowed mutations to be called from single DNA molecules at coverages reaching 10,000X, meaning that mutation detection was carried out to a 0.01% level. Apart from one mutation originating from the manufacturer, no or very little mutation was detected in plasmid stocks or DNA that had been transfected into CHO cells, but not integrated into the genome. A high level of low frequency mutation was detected in recombinant DNA, such that approximately a quarter of all plasmid copies contained at least one mutation. The mutations detected were predominantly in C and G base pairs (85%), but there were no positional biases, with an even distribution of mutation being detected across the length of the plasmid. Mutation was deemed to be unaffected by natural selecetion.

5.1.2. Sequence Variants

This study is focused on sequence variants as a product quality attribute and their identification in heterogeneous cell lines, in which sequence variants are likely to be present at low frequencies. Many studies have identified sequence variants in recombinant products through peptide mapping, mass spectrometry, capillary isoelectric focusing and other protein analytical techniques. These variants have been shown to derive from DNA level mutations (Harris et al., 1993, Ren et al., 2011, Zhang et al., 2015) and amino acid misincorporation during protein synthesis (Wen et al., 2009, Yu et al., 2009). Mostly, these sequence variants have been identified through first establishing the mutation at the protein level, which can then be used to target the culpable DNA mutation at the corresponding locus. Cell line Transcripts are routinely reverse-transcribed to cDNAs for sanger sequencing analysis, but this is a relatively low resolution sequencing technology and is not likely to detect low level sequence variants (Zhang et al., 2015). Next-generation sequencing (NGS) has been used to identify sequence variants at the DNA level, but again these studies were targeted towards regions corresponding protein sequences that are known to be polymorphic (Zeck et al., 2012, Victoria et al., 2010). To our knowledge, Zhang et al. (2015) carried out the first NGS-based analysis for novel sequence variant identification in recombinant protein-producing CHO cells. Using RNA-seq this group were able to successfully identify low level sequence variants, some of which were confirmed as being generated during long term cell culture. More than 25% of cell lines were shown to carry sequence variants. Vector stock sequencing, also using NGS (usually carried out by low resolution sanger sequencing), confirmed that these mutations did not originate from plasmid stocks.

Zhang et al. (2015) were able to establish that at least one of the detected mutations was derived from a replication error during long-term cell culture. This means that the mutation event occurred after plasmid integration into the CHO genome. This supports the ideas discussed in chapter 3 regarding CHO cells having a mutator phenotype, which was shown here to extend to changes at the base pair level. Indeed, it has been shown previously that CHO cells are extremely prone to mutation at the base pair level. In one study it was shown that over 300,000 new SNPs were detected in the generation of the C0101 mAb-producing cell line from its CHO-S parent (Lewis et al., 2013). The Zhang et al. (2015) study was unable to determine whether some of the observed sequence variants derived from changes before genome integration. Various studies have shown that plasmid DNA sequences being transfected into mammalian cells undergo variety of changes such as deletions, insertions and point mutations prior to genome integration. This has been observed in monkey, mouse and human cells (Hauser et al., 1987, Lebkowski et al., 1984). Studies have indicated that the cause of this plasmid DNA instability results from damaging agents both in the cytosol (Lechardeur et al., 1999) and in the nucleus (Lebkowski et al., 1984). It is noteworthy that point mutations predominantly occur at G:C base pairs, which could indicate towards their source of origin (Miller et al., 1984, Hauser et al., 1987). To our knowledge there have been no studies to investigate the potential mutation of transfected DNA before genome integration.

5.1.3. Single Molecule Sequencing

This study uses PacBio RS II Single-Molecule Real-Time (SMRT) sequencing to further study sequence variants in recombinant CHO cells at the DNA level. This technology utilises zero-mode waveguides (ZMWs), which are nanoholes 70 nm in diameter (McCarthy, 2010, Levene et al., 2003). The small size prohibits light waves from traversing the ZMW, which leads to only the bottom of the ZMW (20-30 nm) being illuminated. A DNA polymerase molecule is fixed to the bottom of the ZMW and a single DNA molecule is used as a sequencing template (McCarthy, 2010, Gupta, 2008, Levene et al., 2003). Nucleotides labeled with different fluorophores are incorporated into the synthesised DNA strand which, because incorporation occurs at the bottom of the ZMW, is detected by laser illumination. The incorporated nucleotide is bound for the time (milliseconds) it takes to create a phosphodiester bond, which is a greater amount of time than other, non-bound, nucleotides might diffuse in and out of the detection volume (microseconds). This enables the distinct detection of the incorporated nucleotide (Gupta, 2008, McCarthy, 2010). The fluorophores are attached to the DNA phosphate group as opposed to the base, which is the point of attachment for most sequencing technologies. This means that before the next base can be incorporated, the fluorophore must be cleaved. Therefore, an efficient system is achieved, whereby bases are detected quickly one at a time, allowing for a more definitive distinction between bases.

There are tens of thousands of ZMWs per sequencing reaction, allowing for a high coverage and single molecule analysis (Gupta, 2008). The PacBio SMRT technology is such that a consensus sequence can be called from a single ZMW, which means that a consensus sequence is generated from a single DNA molecule. This is made possible by circular consensus sequencing (CCS) of a SMRTbell template (Figure 5.1a) (Roberts et al., 2013, Travers et al., 2010). The SMRTbell template consists of the linear, double stranded target sequence (insert template), which is ligated to looped, single stranded hairpin adapters at both ends. Sequencing primers hybridise with the adapter sequence and a strand-displacing polymerase facilitates the sequencing of the SMRTbell template, whereby the template is sequenced as a single-stranded circle until the polymerase detaches naturally (Figure 5.1b) (Travers et al., 2010). The current P6-C4 chemistry allows a polymerase to sequence for an average of 10-15 kb (so-called read length) before strand displacement with some reactions reaching ~ 60 kb, which means multiple rounds of this circular sequence can be completed (Rhoads and Au, 2015). The resulting sequencing read (Figure 5.1c) is comprised of sense and antisense strand sequences, interspersed with adapter sequences. Both sense and antisense sequences are then used as individual sequence subreads to generate a consensus sequence (Figure 5.1d). The utilisation of both sense and antisense information helps eliminate sequence context-based sequencing errors. The number of passes of a given template molecule is defined as the number of subreads used to generate the consensus sequence, which is determined by template length and read length (Travers et al., 2010). A single pass of the SMRT template has a high median error rate of ~11%, but the level of error is significantly lowered with each pass (Korlach, 2013, Travers et al., 2010).

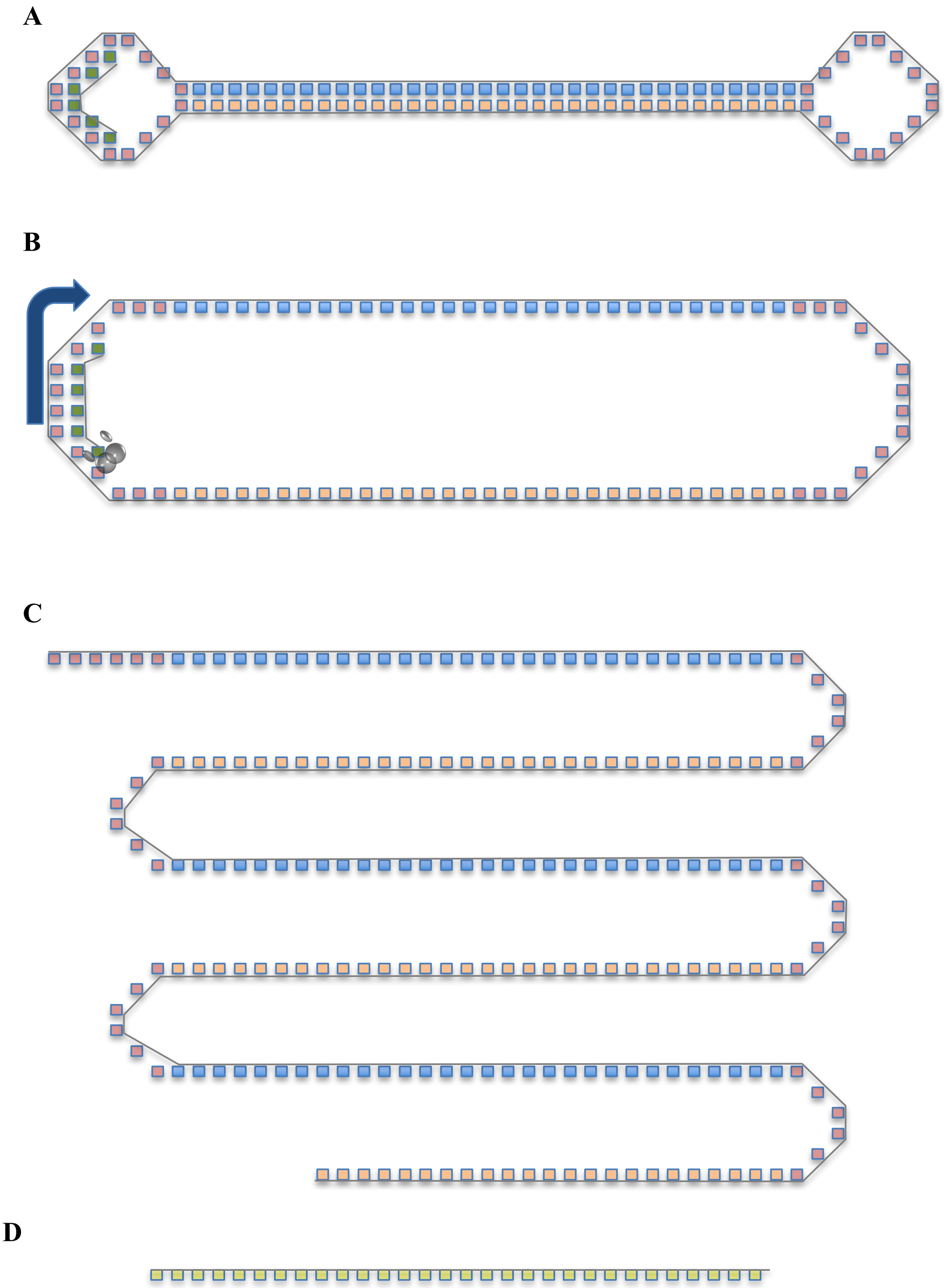


Figure 5.1. Circular Consensus Sequencing
The figure illustrates the SMRTbell template (a) and how, by the use of a strand-displacing DNA polymerase (grey) and a primer (green) complementary to the hairpin adapter (red), it is sequenced in a circular fashion as a single-stranded molecule. The resulting sequence is comprised of alternating sense (blue) and antisense (orange) sequences, interspersed with hairpin adapter sequences (c). A consensus sequence (d) (yellow) is generated from these subreads.

SMRT sequencing does not require PCR, the technology has been shown not to have sequence bias, and there is no signal degradation over time, which means lower error rates are achieved and that any errors are randomly distributed along the template sequence. This means that CCS can successfully overcome the single pass error rate of ~11%. Circular consensus accuracy increases with pass number, but this relationship starts to reach a plateau around 5 or 6 passes where accuracy starts to level off towards QV40 (Phred-type quality value) (99.99%) (Travers et al., 2010). > 99.999% (> Quality Value 60 – QV60) accuracy can be achieved with this technology (Travers et al., 2010, Korlach, 2013, Roberts et al., 2013). The top level of accuracy is achieved by forming a final consensus sequence from a combination of the CCS consensus outputs. However, in order to analyse the sequence output from individual molecules, this study did not combine ROIs, so that low-level variants could be detected beyond the 1% frequency detection limit reported by Pacific Bioscience (CA, USA) (Dilernia et al., 2015).

5.1.4. Chapter Aims

- Develop a high resolution SMRT sequencing analysis platform for point mutations.
    - Build consensus sequences from individual DNA molecules by using only high template pass numbers.
    - Eliminate sequencing and other error to ensure maximum accuracy
- Investigate the assumption that plasmid stock DNA does not contain sequence variants.
- Determine the extent of mutation in transfected / non-integrated plasmid DNA, to establish whether the CHO cell cytoplasmic or nuclear environment is mutagenic to plasmid DNA.
- Assess and characterise the extent and type of mutations that occur in recombinant plasmid DNA during the generation and long term cell culture of a GFP stable CHO cell line, including:
    - Mutation frequency.
    - Mutation Distribution across the plasmid in terms of nucleotide position and potential biases towards coding and non-coding sequences.
    - The type of nucleotide changes.
    - Assessing the level of synonymous and non-synonymous mutations.

5.2 Results

This study investigated three potential sources of point mutation: Plasmid DNA stocks, the pre-integration cellular environment and the genomic environment (Samples: Low and High generation). The phCMV C-GFP plasmid (Genlantis) was used again here to assess this genetic instability. Figure 5.2. depicts the process by which stably producing CHO cells are generated and highlights (red arrows) the time points within this process that DNA samples were taken for SMRT sequencing analysis. The plasmid stock analysis aimed to reveal any errors that were present from initial synthesis of the plasmid and errors that may have been introduced during cloning in *E. coli* DH5α cells. The general assumption (Zhang et al., 2015) is that plasmid stocks do not carry point mutations, so as well as verifying this assumption, this sample will likely serve as a negative control for mutation to give an estimate of error levels in this novel analysis platform. The investigation into point mutations in DNA prior to integration will reveal whether the cytosolic, nuclear or electroporation environment is mutagenic. Any mutations present here are likely to be extremely rare, because plasmid DNA is not replicated in this environment, as opposed to the other samples, in which DNA had been replicated by *E. coli* or mammalian cell genomic replication. Therefore it was necessary to use a method of DNA extraction without the use of PCR, because PCR-based errors could present as a false positive for point mutation. Finally, the two genomic samples taken at two time points over long-term cell culture aimed to reveal whether the fidelity and in vivo error rates of the CHO polymerase and mismatch repair system are responsible for introducing point mutations over long-term cell culture. Samples were sequenced by GATC biotech (Konstanz, Germany).

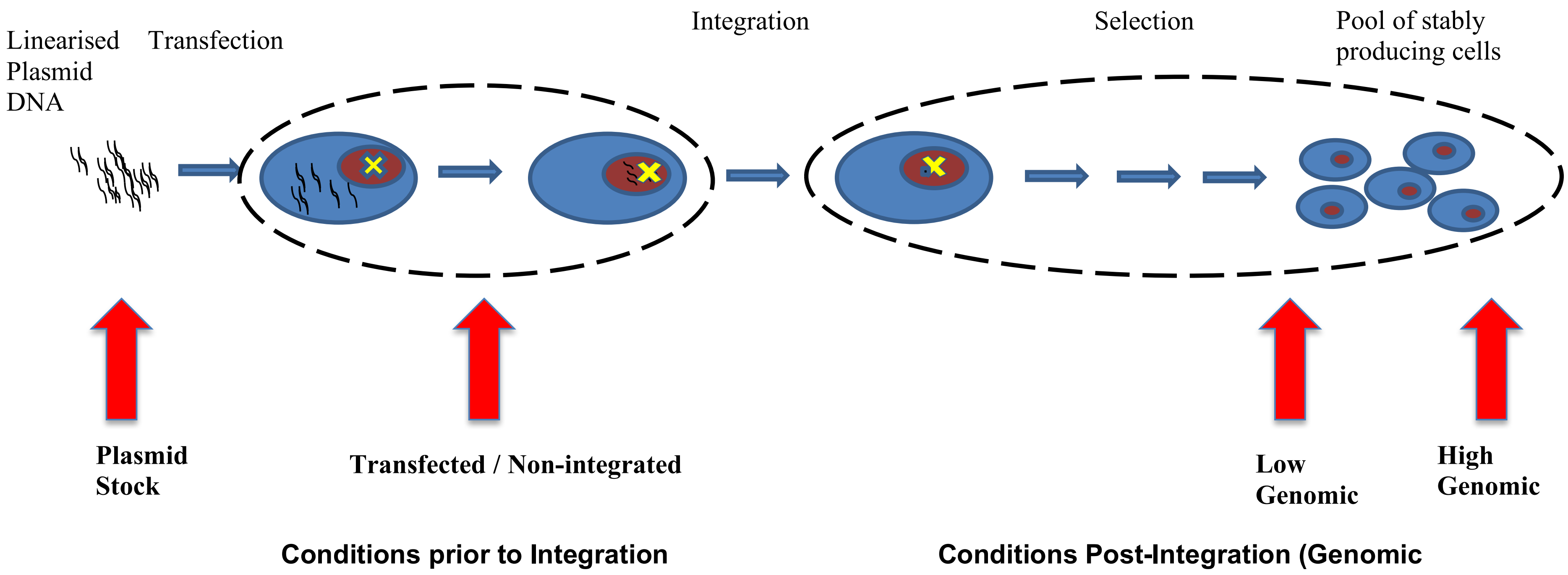

Linearised Plasmid DNA
Transfection
Integration
Selection
Pool of stably producing cells
Plasmid Stock
Transfected / Non-integrated
Low Genomic
High Genomic
Conditions prior to Integration
Conditions Post-Integration (Genomic

Figure 5.2 Stable Pool Generation
The illustration shows the process by which stable CHO cells are generated. Firstly, linearised plasmid DNA is transfected into cells. Some of this plasmid DNA will be present in the nucleus and an extremely small proportion of plasmid molecules will integrate into the host genome. Cells will then be treated with a selection agent to enrich the cell population for cells containing integrated plasmid DNA. This results in the generation of a pool of stably producing cells. The red arrows highlight the time points at which DNA samples were taken for SMRT sequencing analysis. The two arrows pointing towards the stable pool of cells represent the two cell culture time points that samples were taken. Each arrow is labeled with the sample name.

5.2.1. Sequencing Analysis Platform Workflow

Figure 5.3 shows the workflow and decision making process for this analysis platform and is described in detail in the text below.

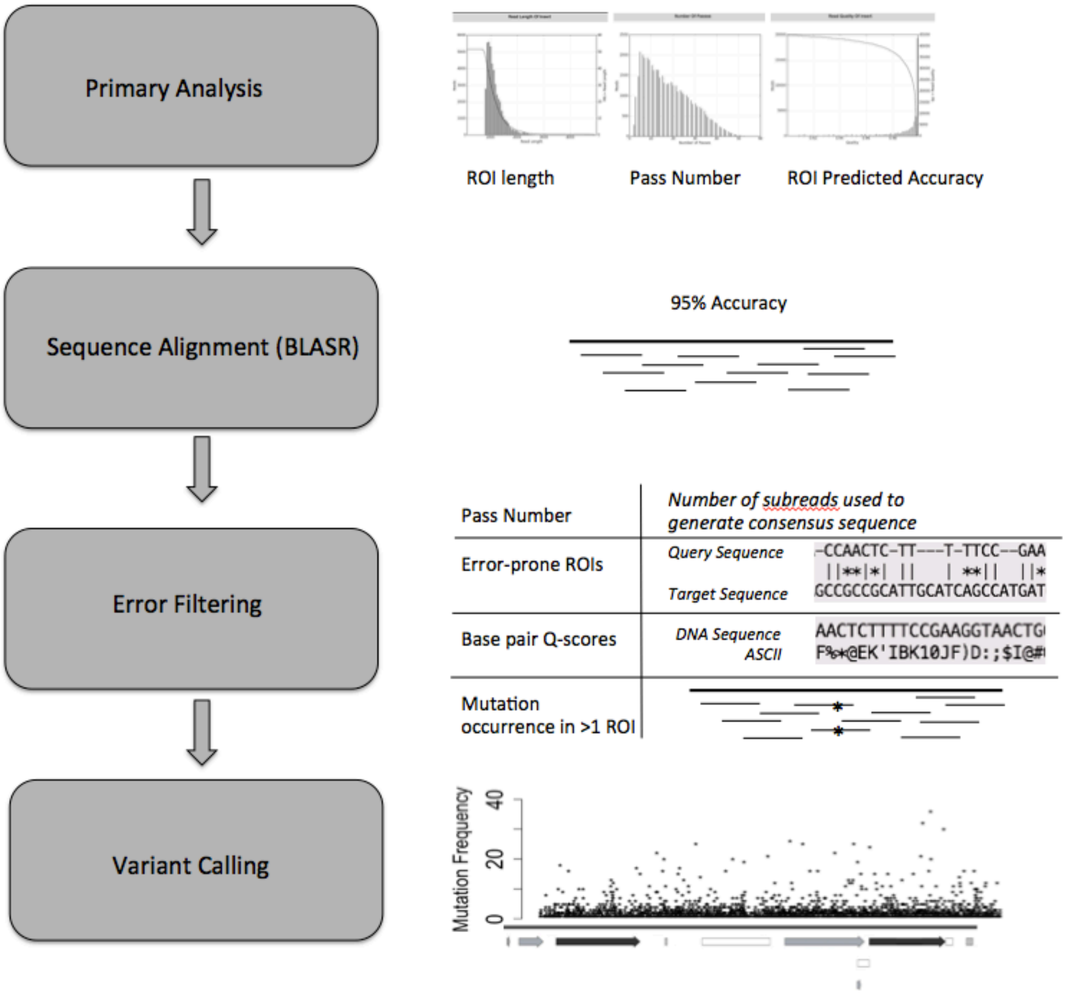


Figure 5.3: Sequencing Analysis Platform Workflow
In primary analysis, subreads are used to generate ROIs from single DNA molecules and filtered by length, at multiple pass numbers and predicted accuracy. ROIs are then aligned to the reference sequence with 95% minimum identity. Error removal is carried out assessing the effect of increased pass number, error-prone ROI removal, Phred (Q) score, and then positional and base pair biases are removed by only counting mutations occurring in more than one ROI. Nucleotide differences compared to the reference sequence, which pass these filtering criteria are then called as variants.

Primary analysis was carried out by Philip Lobb (Pacific Biosciences, CA, USA). This involves the generation of consensus sequences from each ZMW, whereby a so-called read of insert (ROI) is generated from the total number of subreads from each well.

ROIs that were < 800 bp in length or had a predicted accuracy < 90% were eliminated from analysis in order to reduce the abundance of error-prone ROIs. This dataset was then provided to us after 0, 5, 10, 15 and 20 – pass filter permutations in FASTA and FASTQ formats, so that an in-house assessment of pass number error reduction could be conducted. Other important statistics generated at this stage include average ROI length, read qualities and average number of passes.

Each dataset was aligned to the reference plasmid sequence using the BLASR sequence alignment tool. A ROI was only aligned when it showed a minimum of 95% identity to the reference sequence to allow for further error-prone ROI elimination. The BLASR output for subsequent sequence processing was in the human readable format, whereas the output for the processing of ASCII (American Standard Code for Information Interchange) characters relating to a quality score for each given base was in the SAM (Sequence Alignment/Map) format. Processing of the aligned sequences and subsequent analysis was carried out in R.

SMRT sequencing errors are predominantly indel miscalls (Carneiro et al., 2012), so this platform would be likely to show inaccuracies when calling insertions or deletions. Therefore, only point mutations were to be assessed. Each ROI was then aligned against the reference to enable a total coverage count, mutation count and mutation type to be scored at each plasmid position. Upon visual inspection of ROIs containing multiple mutations it was found that these mutations were located in small regions of these ROIs, which also contained multiple insertions and deletions (Figure A25). These error-prone regions were deemed to more likely be a result of individual ZMW error, rather than genuine mutation. Therefore, ROIs containing >3 mismatches were removed from subsequent analyses. A pass number error filtering step was imposed at this point and is explained in section 5.2.2. The ASCII characters were converted to Phred quality (Q) scores (Equation 5.1).

$$Phred\ Quality\ Score = ASCII\ DEC\ operating\ system\ number - 33 \quad \text{Equation 5.1}$$

Nucleotides with Phred score of < Q25 (99.7% accuracy) (Fichot and Norman, 2013) were eliminated from further analysis to ensure high accuracy in base calling (Q score filter). The mutations called here were used to comment upon mutation frequency.

However, a further filter that eliminated mutations only present in one ROI (“>1” filter) was imposed to comment upon base pair and positional baises of the mutations detected, to ensure that errors unque to one ZMW were eliminated.

The data will be presented here in terms of estimations of error, mutation frequency, nucleotide change type, plasmid position, sequence bias, and mutational impact on protein sequences. The coding for this platform can be found in figure A26.

5.2.2. Estimation of Removed Error

As described previously, SMRT sequencing allows for a consensus sequence to be derived from multiple passes of a single DNA molecule. The accuracy of this consensus sequence increases with the number of passes used to derive it. However, this effect of increased accuracy reaches a plateau after a certain number of passes and so there is a tradeoff between increasing accuracy by using a high pass number, and the loss of useful data caused by the pass number filter being too strict. This plateau threshold has previously been reported at 5 or 6 passes (Travers et al., 2010). However, because this study is not building a consensus between different molecules, a greater importance was imposed upon single molecule accuracy. Therefore an analysis of error elimination through pass number filtering was carried out in order to determine which pass number dataset to use for this analysis (Figure 5.4). Datasets for all 5-pass filters were analysed using the secondary analysis platform outlined in the previous section. As stated previously, it was assumed that the plasmid stock sample would not contain large amounts of mutation, which was confirmed by this analysis. Therefore, it was used as a mutation negative control / representation of error to determine the number of passes to proceed with in this data analysis. There is a clear decrease in the number of observed mutations with increasing pass number. This decline is steep until 10 passes, at which point the trend starts to plateau. The sharp change in the gradient of decline indicates that this is the point at which the phenomenon of sequencing error elimination by increased pass number stops. The more gradual decline seen between 10 and 20 passes

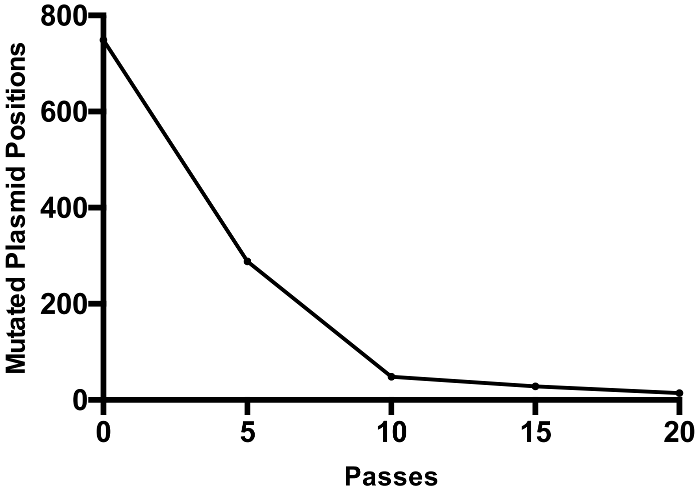


Figure 5.4. Pass Number Effect
The figure shows the number of mutated plasmid positions (out of 4966 bp of the total plasmid) that were shown to have at least one mutation across all ROIs for 0, 5, 10, 15 and 20 passes.

was assumed to be due to the loss of mutation calls from a decreasing number of ROIs that meet the filter criteria (i.e. decline in coverage). The average number of passes for this sample was 17.8. Therefore, it is extremely unlikely that the sharp change in gradient was due to an abrupt change in sequence coverage. It was decided to proceed with the 10 – pass filter for subsequent analysis, because this is the pass number that seemed to meet the accuracy – coverage loss tradeoff described above. This pass number analysis clearly shows the large amount of error from SMRT sequencing that needs to be filtered out by imposing a pass filter.

Removal of further error was facilitated by the Q score and > 1 filters. Figure 5.5 shows the number of mutated plasmid positions (normalised by average sample coverage) that were called as mutated for the three differently filtered datasets, for all four samples. Both filters greatly reduce the number of mutations being called for each sample. These filters, especially the >1 filter, are strict and it is likely that genuine mutations will not be called as a result. However, this is a necessary precaution to ensure that any trends that are found in these data are as genuine and error-free as possible. There is clearly more mutation in plasmid DNA that has been integrated into the CHO genome when compared to the plasmid stock and transfected / non-integrated samples. This will be discussed later in the chapter.

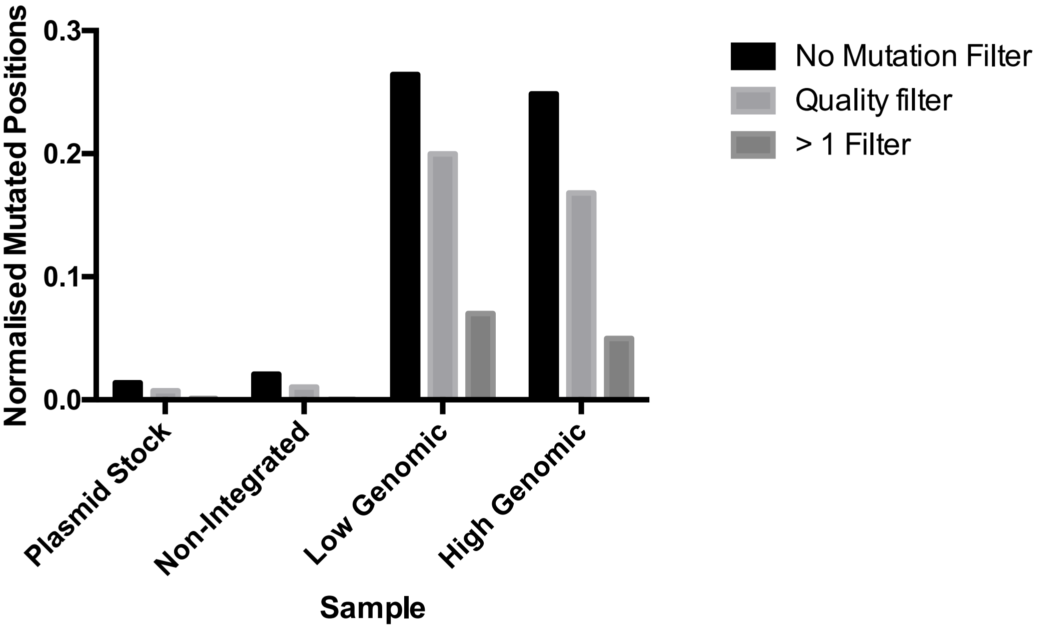


Figure 5.5. Error Filters

The figure shows the reduction in normalised mutated plasmid positions for the plasmid stock, transfected / non-integrated, Low genomic and High genomic samples using a 10 – pass filter after the imposition of the Q score and >1 filters.

Two other filters used to reduce potential error were the 95% minimum percentage identity in the BLASR alignment and the elimination of ROIs that contained more than 3 mismatches. These did not have a large impact on the results. The percentage identity filter reduced ROIs from 29775 to 29540, from 15196 to 15063, from 43191 to 41965, and from 41902 to 40968 in the plasmid stock, transfected / non-integrated, Low genomic and High genomic samples respectively. The >3 mismatch filter reduced ROIs from 29540 to 29533, from 15063 to 15060, from 41965 to 41910, and from 40968 to 40924 in the plasmid stock, transfected / non-integrated, Low genomic and High genomic samples respectively.

5.2.3. Mutation Analysis of Linearised Plasmid DNA Stocks

The phCMV C-GFP plasmid vector (Genlantis) was amplified using Library Efficiency DH5α E. *coli* cells, purified using a GigaPrep kit (QIAGEN) and linearised using restriction enzyme AflII, as described in chapter 2. Plasmid fragmentation and SMRT sequencing of linearised plasmid DNA was carried out by GATC biotech. The fragmentation step prior to sequencing selected 1 kb fragments for sequencing in order to increase the number of sequencing passes per molecule. Primary sequencing analysis by Philip Lobb (Pacific Biosciences) generated a 10 – pass – filtered dataset containing 29705 ROIs, an average ROI length of 1119, a mean ROI quality of 0.9941 and a mean of 23.485 passes. BLASR alignment software aligned 29540 ROIs to the reference sequence with a minimum percentage identity of 95%. The number of ROIs was decreased to 29533 after fragments containing more than 3 mutations were excluded. These ROIs were taken forward to secondary sequencing analysis. Figure 5.6 shows the sequencing coverage of the plasmid in the linearised plasmid stock sample. The mean coverage of this sample was 6600, ranging from 0 to 9041. Aside from two outlier bases, covered 1 and 0 times respectively at positions 3434 and 4653, the minimum coverage was 3426. Coverage decreases from the start to the end of the plasmid sequence and spikes around base pair ~800 and ~4000.

Figure 5.7a shows the complete collection of point mutations detected by the secondary sequencing analysis platform in terms of plasmid location and frequency. Overall there were 92 mutated plasmid positions detected. One of these point mutations, a C → T transition in the bacterial origin of replication (position 2539), is present in 6783 of 6788 fragments (6754 out of 6758 after filtering). We assume here that a mutation called at this frequency is genuine. As can be seen, the other detectable mutated plasmid positions in this sample have a much lower mutation frequency. Figure 5.7b shows the same dataset, but scaled in for examination of the low frequency mutations. After Q score filtering (Figure 5.7c) only 48 mutated plasmid positions were detected. With the exclusion of the mutation detected at position 2539, there were 47 mutated plasmid positions, which had an accumulation of 58 mutation events.

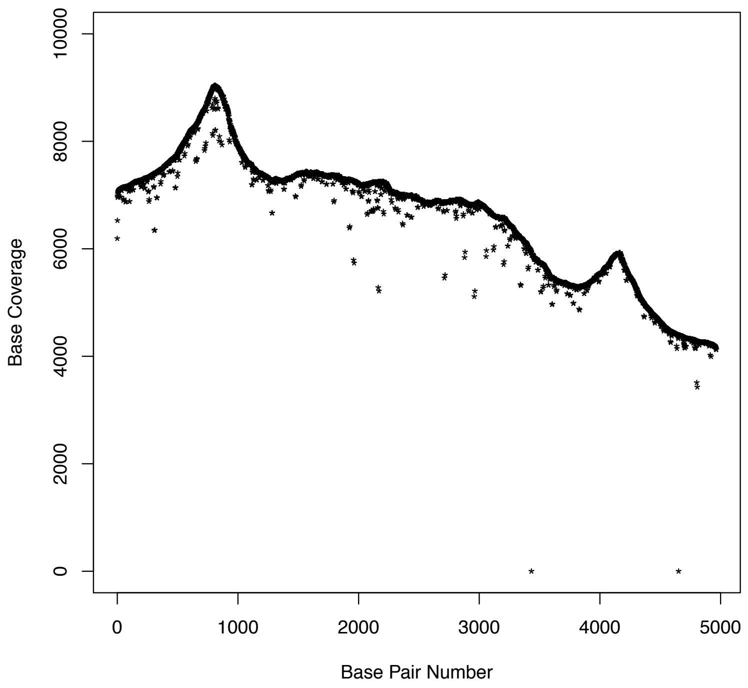


Figure 5.6. Plasmid Stock Sample Coverage
The figure illustrates the coverage of each base pair across the 4966 bp – long GFP plasmid in the linearised plasmid stock sample.

After the data was >1 filtering (Figure 5.7d) only 8 mutated bases were detected. Excluding mutation 2539, 7 mutated plasmid positions were detected, which had an accumulation of 16 mutation events. The total number of called bases that passed the quality score filter was 32,416,625. Therefore, depending on filtering stringency, the mutation rates within the low frequency mutation dataset were 1 in 5.6 x 10^5 and 1 in 2.0 x 10^6 for the Q score and >1 filters respectively. Whilst it is possible that some of these point mutations could be genuine, this mutation frequency will be used as an estimate of error for this sequencing analysis platform. The overall conclusion was that there was genuine mutation found in the plasmid stock sample (position 2539), which was present in nearly all ROIs covering this base. Within the low frequency mutations, although some of these mutations could be genuine, they are more likely to be representative of systematic error in this sequencing platform.

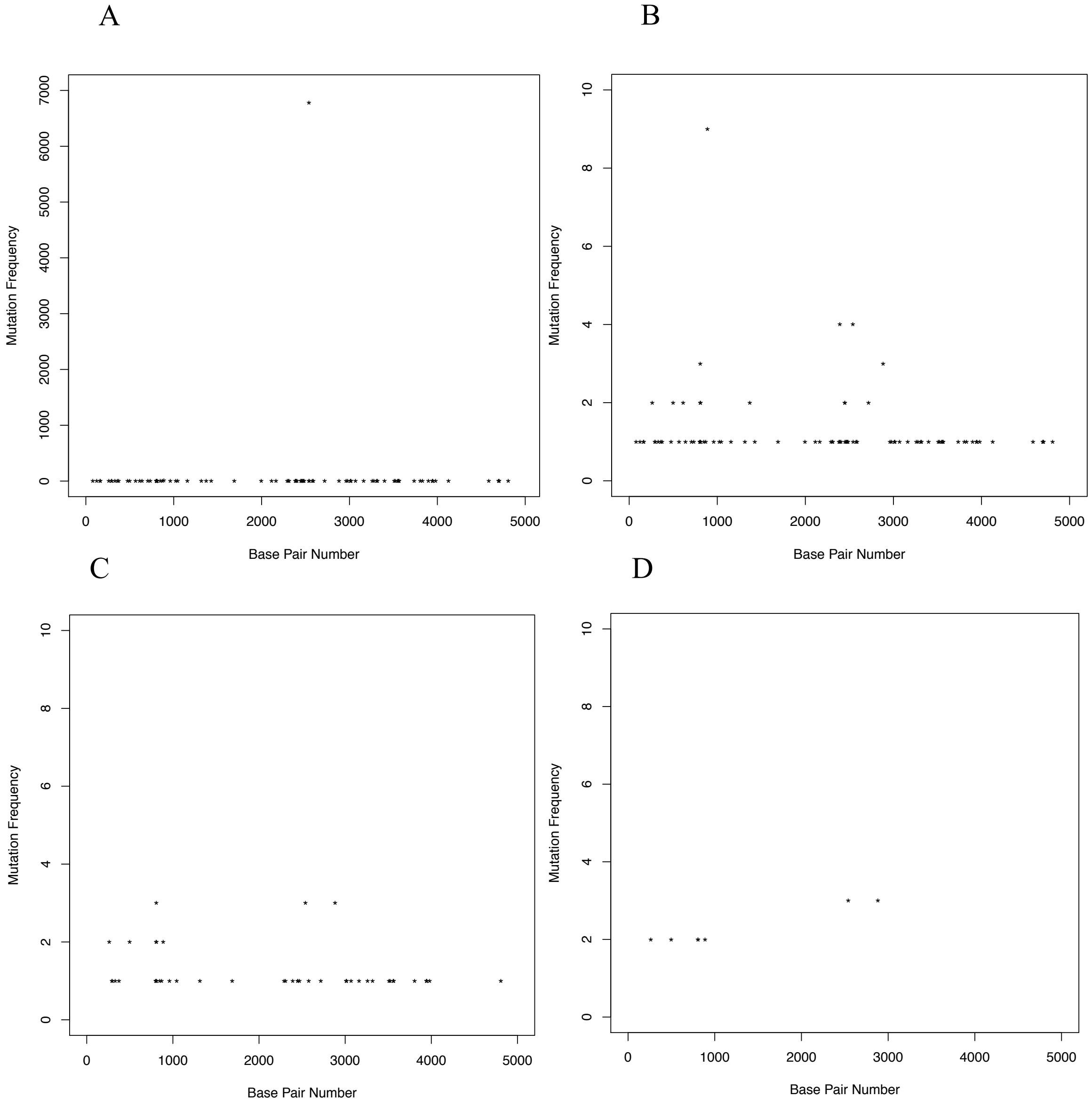


Figure 5.7. Plasmid Stock Mutation Frequency

This figure shows the frequency and locations of detected point mutations in the plasmid stock sample. All observed (A), low frequency (B), low frequency quality filtered (C) and low frequency quality filtered and >1 filtered (D) point mutations are shown.

5.2.4. Mutation Analysis of Transfected Non-Integrated Plasmid DNA

Plasmid DNA was used from the same plasmid stock as was used in section 5.2.3 and transfected into CHO269M cells using 320-26 electroporation conditions. A modified Hirt method protocol (Section 2.2.4) was used to extract linearised plasmid DNA from CHO cells 24 hours after transfection as devised by (Arad, 1998). Agarose gel electrophoresis was used to confirm the successful extraction of the 5 kb plasmid DNA molecules and to assess whether plasmid DNA remains intact in the mammalian cell environment (Figure 5.8). The modified Hirt method successfully extracted plasmid DNA from CHO cells. However, samples also contained, what appear to be, large fragments of genomic contaminant DNA and unidentified smaller DNA fragments. Controls demonstrate that these smaller fragments are only present when DNA (linear or circular) is transfected into CHO cells and that the electric current alone does not cause this phenomenon.

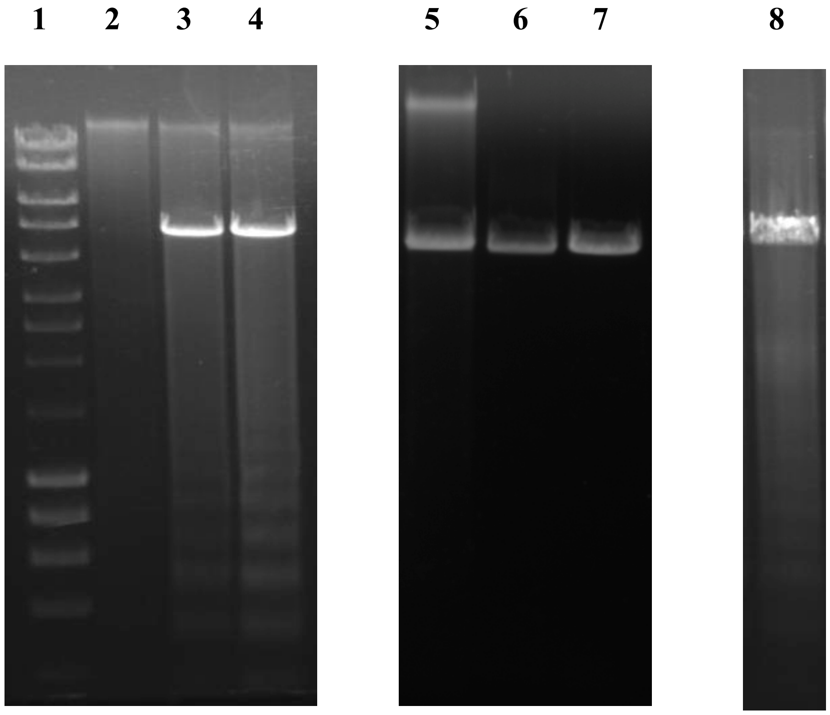


Figure 5.8. Transfected DNA Purification
The figure shows gel images of purified plasmid DNA from CHO cells, containing hyperladder I (Lane 1), a mock transfection contol (Lane 2), two purified plasmid samples (Lanes 3 and 4), a mock transfection + spiked plasmid DNA control (Lane 5), DNA in 320-26 conditions (Lane 6), plasmid DNA (Lane 7), and transfected non-linearised plasmid DNA (Lane 8).

Therefore, the electric current alone is not responsible for plasmid or genomic fragmentation. These smaller fragments could be the result of plasmid digestion or fragmentation in the CHO cell environment or could be a fragmented genomic DNA occurs after DNA transfection. This will be discussed further in section 5.3. Clearly it is undesirable to sequence DNA samples containing DNA that may not be plasmid DNA, because it will lead to reduced sequence coverage. Therefore it was necessary to purify the plasmid from this unidentified DNA. SMRT sequencing requires that samples have not been in contact with DNA intercalating agents in their preparation, meaning common gel extraction techniques cannot be used for purification. Blue Pippin technology offers automated gel purification without the need for intercalating agents (Sage Science, MA, USA). To validate the use of BluePippin technology for this purpose, it was tested in-house. A target purification size of 5.3 kb was used and DNA within the maximum range of 4.25 kb and 6.35 kb was collected. This approach was successful in removing any visually identifiable (by agarose gel electrophoresis) contaminants from the plasmid sample (Figure 5.9).

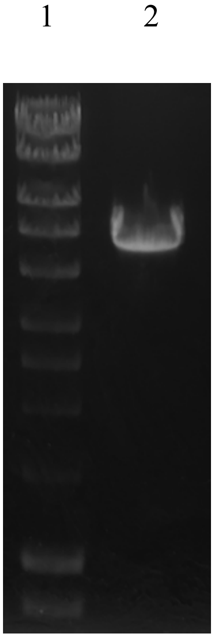


Figure 5.9. BluePippin Purification
The figure shows the agarose gel image of a Blue Pippin purified CHO plasmid extract (Lane 2) with Hyperladder I (Lane 1).

Therefore, Blue Pippin purification was carried out by GATC Biotech (Konstanz, Germany) before sample fragmentation. Unfortunately, GATC Biotech (Konstanz, Germany) were unable to use the Blue Pippin instrument at the same resolution as was carried out in the validation of the technology. A 5 kb target purification was carried out, but with a wider range of 3 kb around this target. SMRT sequencing was carried out under the same conditions as with the previous sample.

Primary analysis filtering for ROIs with a minimum of 10 passes, 99% predicted accuracy and a minimum length of 800 bp generated 30,824 ROIs, with a mean length of 1473 bp, a mean quality of 0.9958 and a mean pass number of 20.012. BLASR alignment software aligned 15,063 ROIs to the reference sequence with a minimum percentage identity of 95%. This is approximately half of the total number of ROIs generated from the primary sequencing analysis. Therefore, it is likely that the Blue Pippin purification step was not efficient in removing genomic contaminant DNA. Blue Pippin purification with the size range used in the validation study (Figure 5.9) may have reduced the amount of non-plasmid DNA in the sample. However, it might be the case that there are genomic fragments that are too close in size to plasmid DNA to allow for complete purification using this method. The number of ROIs was decreased to 15,060 after fragments containing more than 3 mutations were excluded. The remaining ROIs were taken forward to secondary sequencing analysis.

Figure 5.10 shows the sequencing coverage of the plasmid in the non-integrated / transfected plasmid DNA sample. The mean coverage of this sample was 4319, ranging from 0 to 4919. Two bases with low coverage in the plasmid stock sample, at positions 3434 and 4653, were covered 0 times in this sample. This could be a result of the polymerase having difficulty reading these particular nucleotides within this specific sequence. Aside from these outlier bases the minimum coverage was 2975. As opposed to the plasmid stock sample, which showed a gradual decrease in coverage across the plasmid length, there was no detectable increase or decline in coverage in this sample. There are two clear spikes in coverage in line with the coverage spikes seen in the linearised plasmid stock sample at ~ 800 bp and 4000 bp respectively. The coverage in this sample was less than in the plasmid stock sample, which is presumably due to the apparent presence of contaminating DNA that the Blue Pippin instrument failed to remove. This would indicate that the unidentified contaminant DNA was not fragmented plasmid DNA.

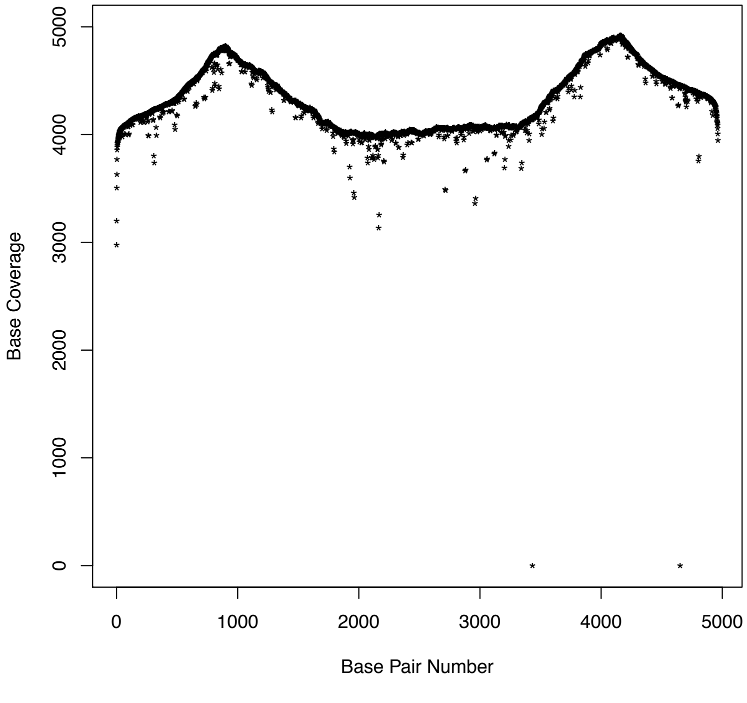


Figure 5.10. Transfected / Non-Integrated DNA Sample Coverage
The figure illustrates the coverage of each base pair across the 4966 bp – long GFP plasmid in the non-integrated / transfected plasmid DNA sample.

Figure 5.11a shows the complete collection of mutated plasmid positions detected by the secondary sequencing analysis platform in terms of plasmid location and frequency. Overall there were 90 mutated plasmid positions detected. As was seen in the plasmid stock sample, a C → T transition in the bacterial origin of replication (position 2539), was present in 3986 of 3988 fragments (3974 out of 3974 after filtering). Again, we assume here that a mutation called at this frequency is genuine. As can be seen, the other detectable mutated plasmid positions in this sample have a much lower frequency. Figure 5.11b shows the same dataset, but scaled in for examination of the low frequency mutations. After Q score filtering (Figure 5.7c) only 45 mutated plasmid positions were detected. With the exclusion of the mutation detected at position 2539, there were 44 mutated plasmid positions, which had an accumulation of 45 mutation events. After >1 filtering (Figure 5.7d) only 2 mutated bases were detected. Excluding mutation 2539, 1 mutated plasmid positions were detected, which was observed twice in total. The total number of called bases that passed the Q score filter was 21,246,529. Therefore, depending on filtering stringency, the mutation rates within the low frequency mutation dataset were 1 in 4.7 x 10^5 and 1 in 1.1 x 10^7 for the quality score and >1 filters respectively. As with the plasmid stock sample, it is possible that some of these point mutations could be genuine, but it is more likely that this mutation frequency represents an estimate of error for this sequencing and analysis platform. It should be noted that the coverage for the non-integrated transfected sample was considerably less (65.4%) than the coverage for the plasmid stock sample and so low frequency mutations are less likely to be detected. Figure 5.5 shows mutation levels after being normalised for coverage, in which the mutation frequency in the non-integrated transfected sample was marginally higher than for the plasmid stock sample, but not to an extent that indicates this is due to anything other than random sampling. Both of these samples are relatively low in comparison with the genome-integrated samples. Therefore, the conclusion was that there was not convincing evidence that non-genomic cellular environment caused mutation of the plasmid DNA.

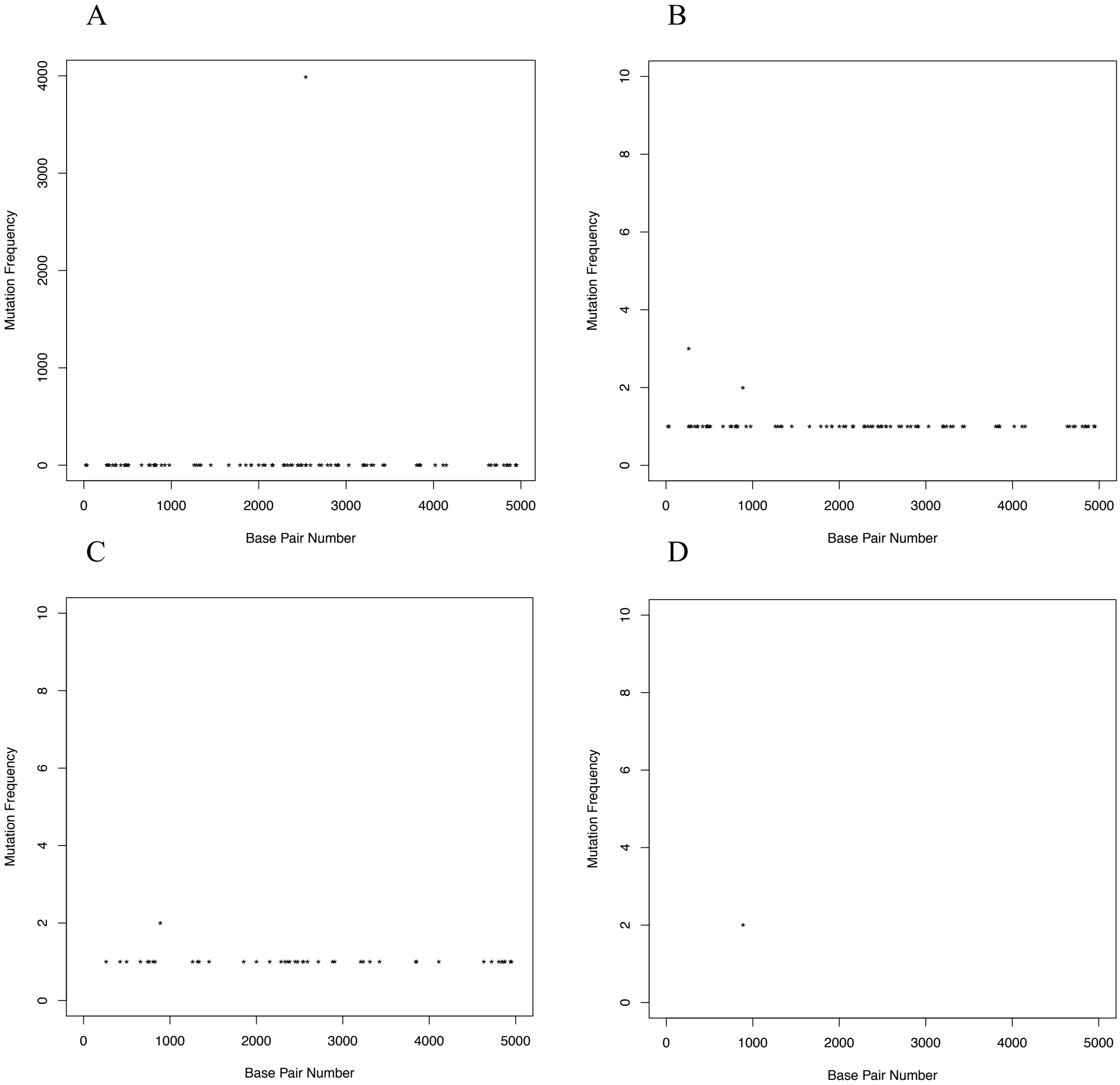


Figure 5.11. Transfected / Non-Integrated Plasmid Mutation Frequency
This figure shows the frequency and locations of detected point mutations in the transfected / non-integrated plasmid sample. All observed (A), low frequency (B), low frequency quality filtered (C) and low frequency quality filtered and >1 filtered (D) point mutations are shown.

5.2.5. Stable GFP Cell Line Generation

In order to investigate the occurrence of point mutation of integrated plasmid DNA, GFP stable cell lines were generated. CHO269M cells were transfected using 320-26 electroporation conditions and cells containing integrated plasmid DNA were selected using the neomycin analogue G418 in order to generate a population of plasmid-containing cells. The Kanamycin / Neomycin resistance gene on the phCMV C-GFP vector provides resistance against G418 and thus selects for cells containing integrated plasmid DNA, which is detectable through GFP measurements by flow cytometry. Some studies have shown that G418 alone is not sufficient to facilitate selection and so FACS was used as a supplementary technique to increase the number of recombinants in the population (Zhang et al., 2006). FACS was carried out by Kay Hopkinson at the University of Sheffield flow cytometry core facility.

Due to the batch – to – batch variation in G418 disulphate stocks, it was necessary to carry out a dose response experiment for each batch. Two batches were used during the selection process. G418 concentrations 0.1-0.2 mg/mL above the concentration which led to complete cell death after 8 days of batch culture were selected for cell line selection (Lonza, 2012). Cell viability and VCD were used in making this decision. For batch 1 it was decided to proceed using 0.8 mg/mL G418, (Figure 5.12) and for batch 2 it was decided to proceed using 0.9 mg/mL (Figure 5.13).

A brief summary of transfection, the selection process and cell culture of the stable cell line is as follows. 1 x 10^7 CHO269M cells were transfected with 50 ug of phCMV C-GFP plasmid using 320-26 electroporation conditions and then immediately transferred into T75 flasks containing 40 ml media (detailed in section 2.6). T75 flasks were incubated in a humidified static incubator. After 24 hours recovery (Day1) transfection efficiency (94%), cell viability (84%) and VCD (0.16 x 10^6 cells.ml) were in line with optimised values presented in chapter 4, and so G418 was added to begin recombinant cell selection. Cells were transferred into E125 flasks for shaking incubation on day 7, from which time they were passaged on a standard 3-4 day regime.

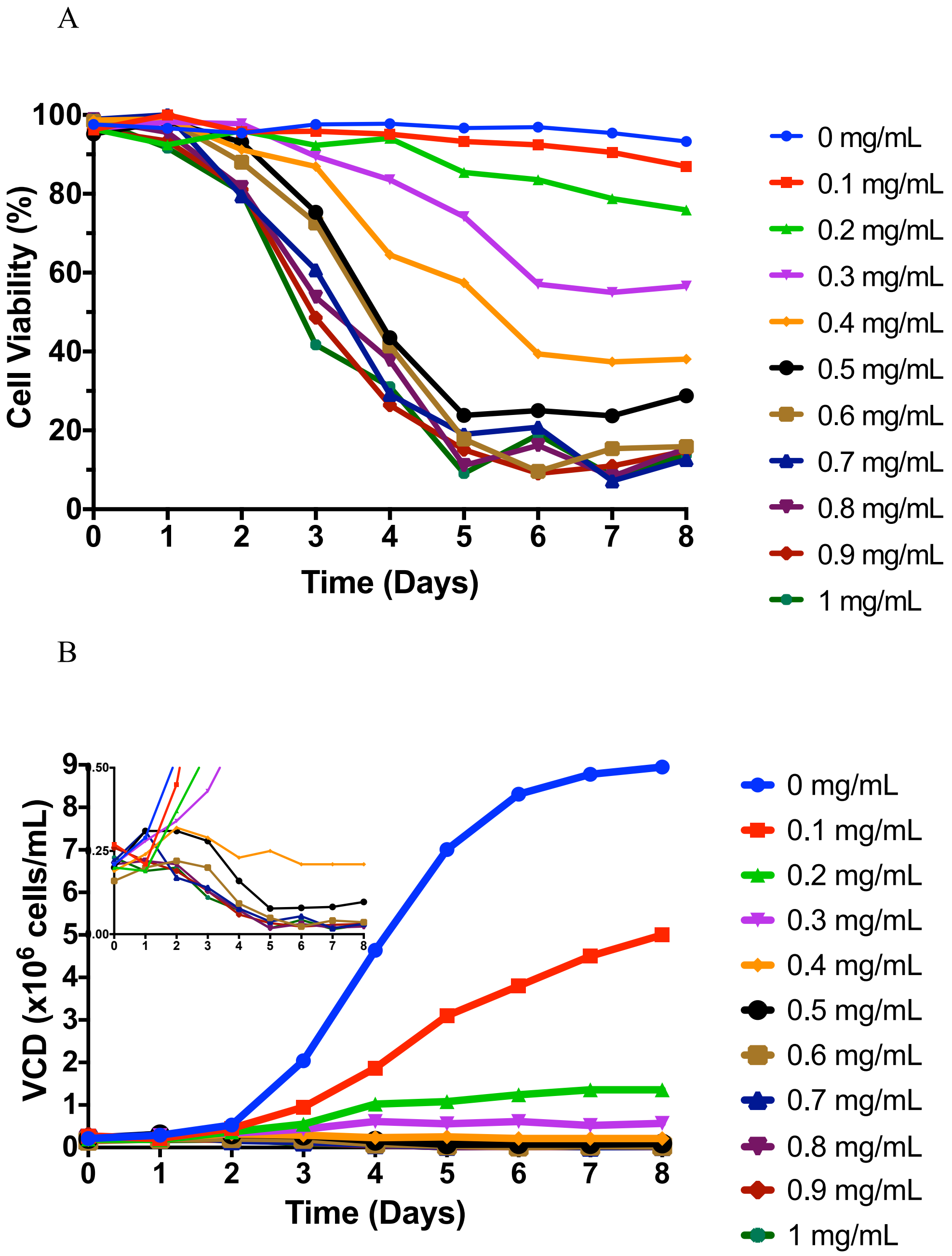


Figure 5.12. G418 Dose Response: Batch 1
The figure shows the cell viability (A) and VCD (B) response to a range of G418 concentrations ranging from 0 – 1 mg/mL over 8 days of batch culture.

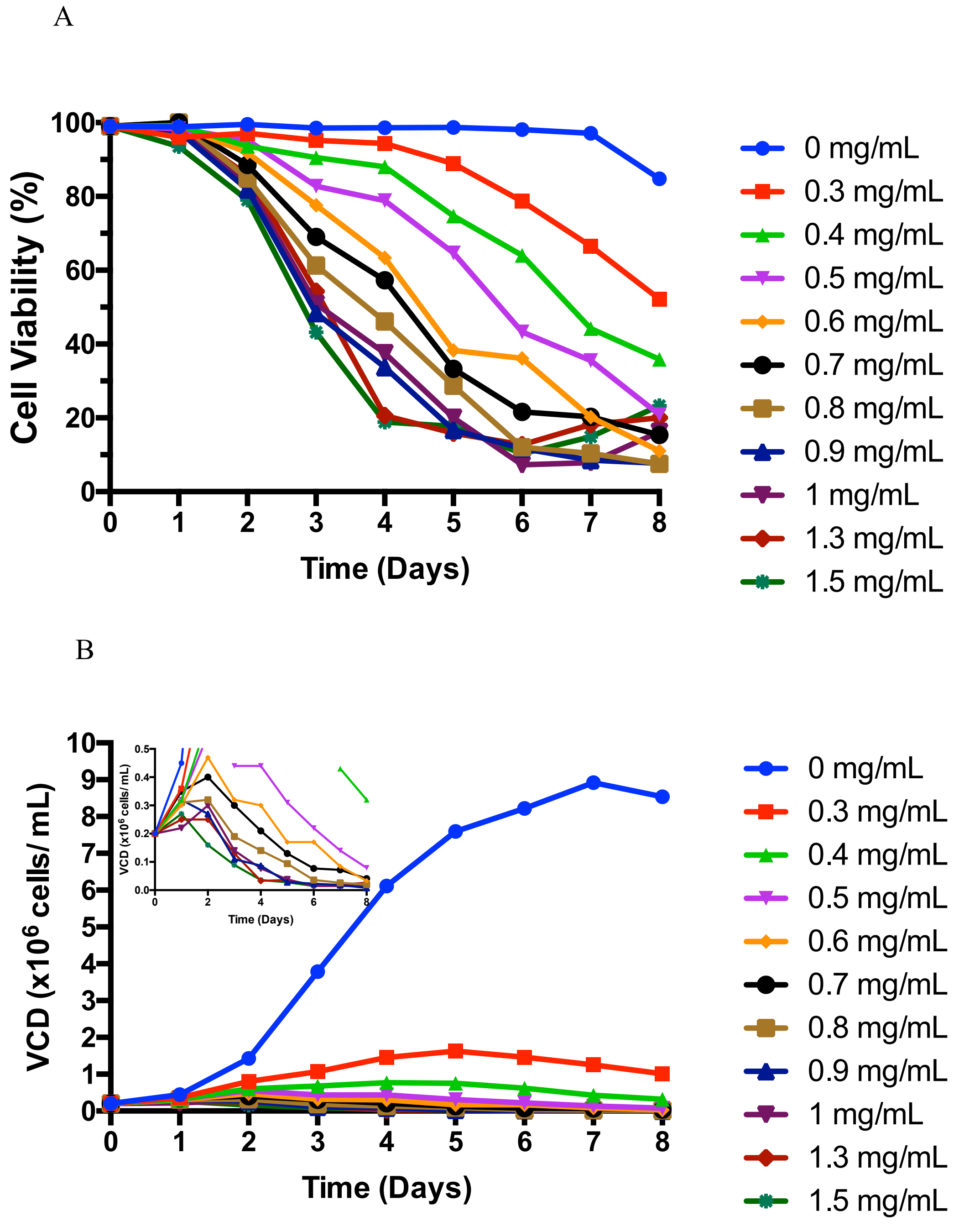


Figure 5.13. G418 Dose Response: Batch 2
The figure shows the cell viability (A) and VCD (B) response to a range of G418 concentrations ranging from 0 – 1.5 mg/mL over 8 days of batch culture.

Cells were sorted for GFP production using a low threshold (top ~90% of GFP positive cells) on day 39 and then at a higher threshold (top ~20% of GFP positive cells) on day 46. A cell bank was made using cryopreservation protocols (section 2.1.2) from stable cells on day 59, which was used to generate the "Low" generation sample for DNA sequencing. Cells were cultured until day 126, at which point cell banks were made and samples taken for the "High" generation sample for DNA sequencing. Figure 5.14 shows VCD, cell viability and GFP positive cell measurements over the course of stable cell line generation and cell culture. As can be seen VCD is slow to increase at the start of the selection process, because of the growth inhibition of non-recombinants. Another dip in VCD can also be seen around the two FACS events. This is because the FACS imposes a strict population bottleneck, which reduces the population of cells and so time is needed for VCD to return to normal levels. Cell viability is initially seen to be lower, because of electroporation recovery and cell selection. Viability then returns to higher levels, but appears to oscillate during each cell subculture. This is due to an apparent culture artifact of G418 – containing media, such that viabilities are counted as lower towards days 3 and 4 and when cells are replenished with fresh media viabilities return to normal levels (~98%) and so were assumed to be healthy. Inspection of Vi-Cell images reveals artifacts within the culture that are called as dead cells. Initial GFP positive cell measurements were high due to transient gene expression, which subsides in line with plasmid degradation and dilution. GFP positive measurements then remained consistent at ~7% in line with the expectation that G418 may not be a sufficient selector (Zhang et al., 2006). The first round of FACS led the GFP positive cell measurements ~60% and the second round of FACS led to GFP positive cell measurements ~93%. Once this was determined to be stable, the Low generation cell bank was generated. GFP positive cell measurements remained fairly consistent, apart from a slight decrease around day 85. When high generation samples were taken GFP positive cell measurement was ~90%. Generation number for low and high samples was ~57 and ~133 respectively.

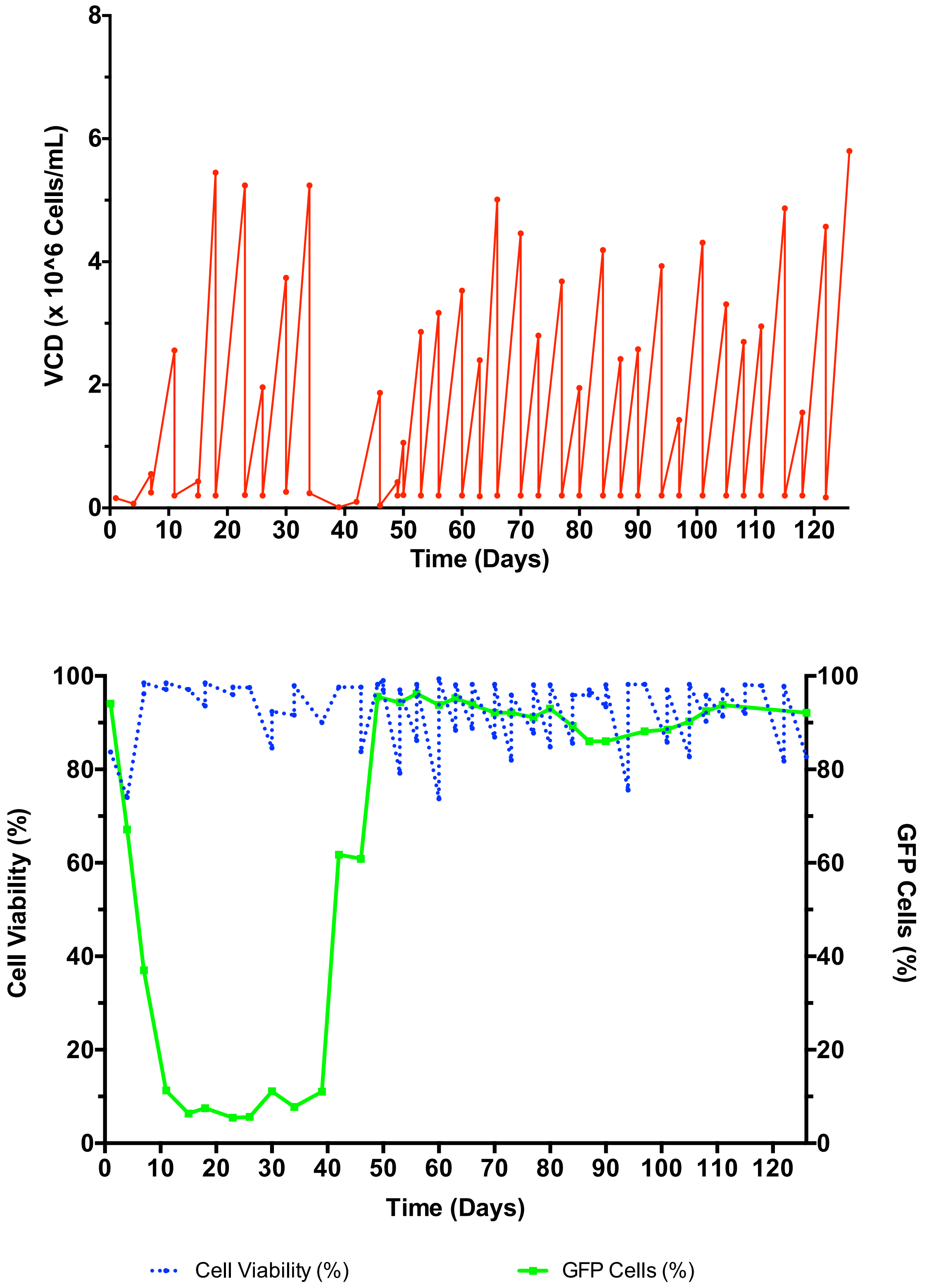


Figure 5.14. GFP Stable Cell Line Generation
The figure shows cell growth (VCD - A) along with cell viability (%) and GFP positive cells (%) (B) over the 126 day selection and culture period of the stable GFP cell line.

5.2.6. Genome – Integrated Plasmid: Low Generation

Genomic DNA was extracted from low generation stable GFP cells using a Blood and Cell Culture DNA kit (QIAGEN, Manchester, UK). In order to provide a sufficient quantity of recombinant plasmid DNA that was free from other CHO genomic DNA, it was necessary to carry out PCR. The fragmentation process carried out by GATC biotech prior to DNA sequencing could not be carried out on PCR products. Therefore, to ensure the sequencing of plasmid templates with sizes allowing for multiple passes it was decided to carry out four separate PCR's, which were designed to amplify four overlapping plasmid regions (~1.3 kb) covering the entire plasmid length. PCR was carried out using the Phusion High Fidelity DNA polymerase (New England Biolabs, UK) and the subsequent samples were quantified using a Nanodrop, so that the four PCR products could be pooled together in equal quantities into one sample. SMRT sequencing of this sample was carried out by GATC Biotech. Primary analysis filtering for ROIs with a minimum of 10 passes, 99% predicted accuracy and a minimum length of 800 bp generated 41,500 ROIs, with a mean length of 1338 bp, a mean quality of 0.9935 and a mean pass number of 22.306. BLASR alignment software aligned 41,965 ROIs to the reference sequence with a minimum percentage identity of 95%. The number of ROIs was decreased to 41,910 after fragments containing more than 3 mutations were excluded. These ROIs were taken forward to secondary sequencing analysis.

Figure 5.15 shows the sequencing coverage of the plasmid in the Low generation DNA sample. The mean coverage of this sample was 10,525, ranging from 0 to 23,957. The coverage here is clearly different to the coverage in the plasmid stock and non-integrated transfected samples. The pooling together of four separate PCR reactions resulted in four predominant plasmid sequence coverage frequencies. The coverage at the start of the sequence (positions 1-77) is approximately a 3-fold lower than the rest of the sequence from the same PCR reaction. The overlapping regions between the separate PCR-based sequences result in spikes of coverage, where plasmid regions are being covered by two PCR templates. Again, the coverage of plasmid positions 3,434 and 4,653 are extremely low, being covered 0 and 4 times respectively. The vast majority of the plasmid positions within this data reside within the four predominant PCR-based frequency populations, which range in averages from 6,370 to 13,768.

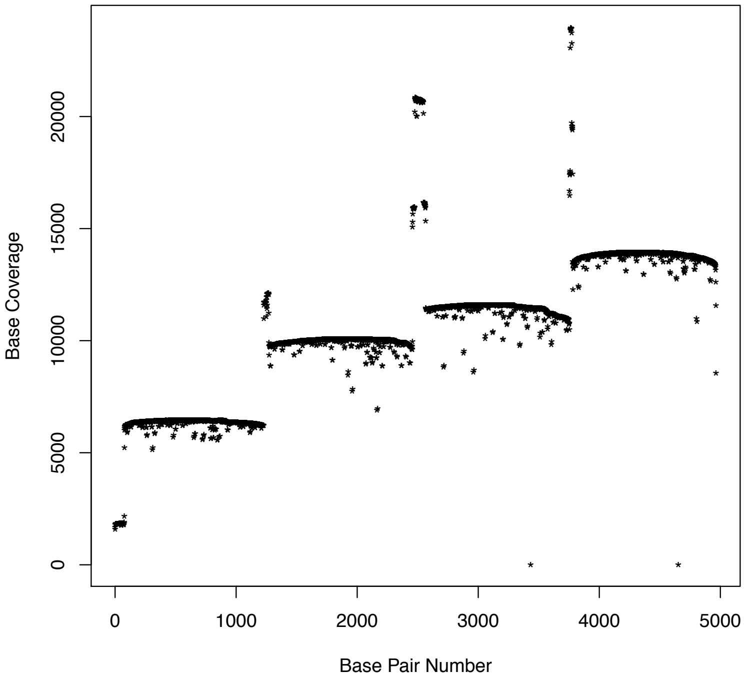


Figure 5.15. Low Generation Sample Coverage
The figure illustrates the coverage of each base pair across the 4966 bp – long GFP plasmid in the low generation recombinant plasmid DNA sample.

Figure 5.16a shows the complete collection of point mutations detected by the secondary sequencing analysis platform in terms of plasmid location and frequency in the low generation genomic sample. Overall there were 2783 mutated plasmid positions detected. As was seen in previous samples, a C → T transition in the bacterial origin of replication (position 2539), was present in 16098 of 20676 fragments (16013 out of 20016 after filtering). Again, we assume here that a mutation called at this frequency is genuine. As can be seen, the other detectable mutated plasmid positions in this sample have a much lower frequency. Figure 5.16b shows the same dataset, but scaled in for examination of the low frequency mutations. After quality score filtering (Figure 5.7c) 2104 mutated plasmid positions were detected. With the exclusion of the mutation detected at position 2539, there were 2103 mutated plasmid positions, which had an accumulation of 4214 mutation events. After the data was filtered for mutations occurring more than once (Figure 5.7d) only 739 mutated bases were detected. Excluding mutation 2539, 738 mutated plasmid positions were detected, which had an accumulation of 2456 mutation events. Mutation seems to be randomly

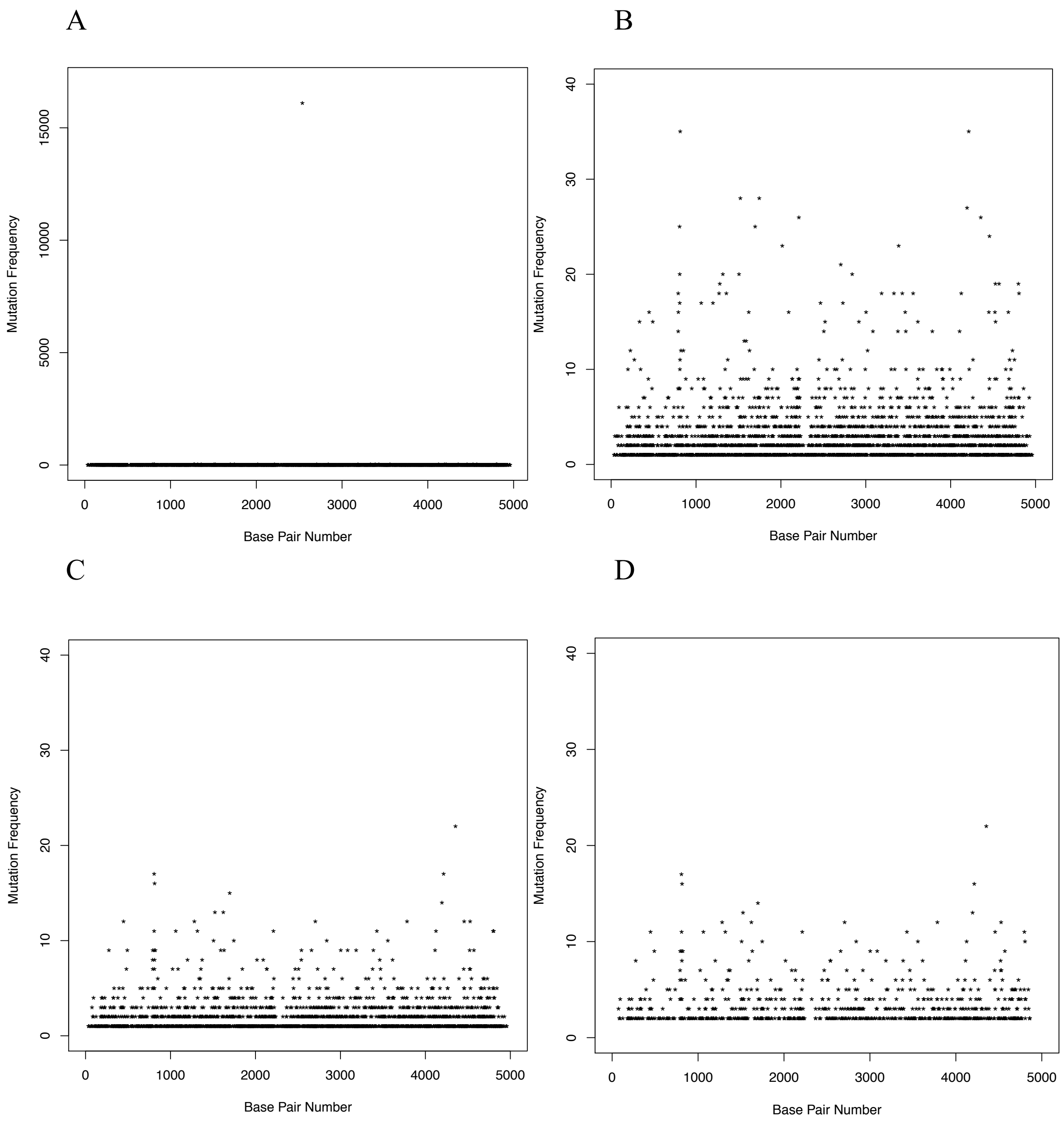


Figure 5.16 Low Generation Recombinant Plasmid Mutation Frequency
This figure shows the frequency and locations of detected point mutations in the low generation recombinant plasmid sample. All observed (A), low frequency (B), low frequency quality filtered (C) and low frequency quality filtered and >1 filtered (D) point mutations are shown.

distributed along the full range of the plasmid. The total number of called bases that passed the quality score filter was 51,635,389. Therefore, depending on filtering stringency, the mutation rates within the low frequency mutation dataset were 1 in 1.2 x 10^4 and 1 in 2.1 x 10^4 for the quality score and >1 filters respectively. These mutation rates are approximately 47-fold and 95-fold higher than those seen in the plasmid stock negative control for the Q score filter and >1 filter mutation rates respectively. There are mutations in more plasmid positions and with higher frequencies than seen in the plasmid stock control. This is strong evidence of mutations occurring during the generation of the stable GFP cell line and subsequent cell culture for ~57 generations. The coverage of plasmid bases in this sample was considerably higher than the plasmid stock negative control, but this does not impact on the differences found between the two datasets (Figure 5.5). Depending on filter stringency, the mutation rates given here suggest that 1 in every 2.4 or 4.2 of the 5 kb plasmids used here contain a point mutation.

As stated previously, the >1 – filtered dataset is more likely to result in the detection of genuine mutations, because it overcomes the sources of error that are unique to individual ZMWs as well as being filtered for quality. The Q score – filtered dataset was considered sufficient to comment on mutation frequencies, but not for drawing conclusions regarding the type of the mutations detected, because inaccuracies here may skew the results. Therefore, only the >1 – filtered dataset will be used for this purpose.

Table 5.1 contains all the genetic elements of the phCMV C-GFP plasmid and the percentage of mutations that fall within each element from the >1 – filtered dataset. Where two or more elements overlap, a separate element is designated so not to count mutations more than once. If mutation is assumed to be random, then a base within one genetic element is equally likely to be mutated as a base within another genetic element. Therefore, the longer a genetic element, the more likely it is that it will have been mutated at some point along its length. In order to determine whether mutation is targeted towards particular genetic elements, mutation percentage was normalised to element length to correct for this potential bias. All but two of the sequence types noted here were mutated. The two mutation-free sequences were the polyadenylation signal sequences for the Kan / Neo GFP genes, which perhaps are conserved through natural selection. However, this may also be due to polyadenylation signals being short and are less likely to be hit by random mutation. Of the mutated elements, there appears to be no substantial difference between coding and non-coding DNA, which may be an indication of random mutation, not greatly affected by natural selection. MCS's in particular appear to be more heavily mutated than other sequences.

Sequence Element	Mutation Frequency (%)	Mutation Frequency (Normalised by element length)
pAmp	0.4	13.79
pSV40	4.7	20.43
Kan / Neo	16.9	21.26
HSV_TK_PolyA	0	0
Puc_Ori	13.7	21.27
phCMV + Intron	14.9	21.72
phCMV + Intron + MCS1	0.9	18.75
pT7	0.3	18.75
MCS1	1.5	30
GFP ORF	18.4	25.56
MCS2	3.1	44.29
SV40 PolyA	0	0
Non-Coding	25.2	15.68

Table 5.1. Low Generation Sample: Mutated Genetic Elements
The table contains the percentage of mutations that fall within each genetic element of the phCMV C-GFP plasmid and the normalized mutation value relative to the length of each genetic element. The sequence elements are as follows: Ampicillin resistance gene promoter, SV40 promoter, Kanamycin / Neomycin resistance gene, HSV Thymidine Kinase polyadenylation signal, pUC origin of replication, Human CMV promoter enhancer and intron, overlapping region of Human CMV promoter enhancer and intron plus the multiple cloning site upstream of the GFP open reading frame, T7 promoter priming for sequencing, multiple cloning site upstream of GFP open reading frame, GFP open reading frame, multiple cloning site downstream of GFP open reading frame, SV40 polyadenylation signal sequence and non-coding DNA.

Table 5.2 shows the frequency of each type of point mutation, along with a total frequency of changed nucleotides, from the >1 – filtered dataset. There is a clear predominance in point mutations of G and C nucleotides, showing 41.96% and 43.98% of changes respectively. More specifically, by far the most frequent types of changes are G.C → A.T transitions (C → T (24.9%), G → A (19.22%)) and C.G → A.T transversions (C → A (18.54%), G → T (22.6%)). The GC content of the plasmid reference sequence is 50.7%, so will not have influenced these results.

		To				
		A	T	C	G	Total
From	A	--	0.41	0.41	7.44	8.26
	T	0.81	--	4.74	0.27	5.82
	C	18.54	24.9	--	0.54	43.98
	G	19.22	22.6	0.14	--	41.96

Table 5.2. Low Generation Sample: Nucleotide Changes
The table shows the percentages of each type of nucleotide change seen within this dataset, the sum of which are used to give the total percentage change for each nucleotide.

The phCMV C-GFP plasmid contains two open reading frames (Kan / Neo and GFP). The >1 – filtered dataset was used to determine whether the observed DNA point mutations were synonymous or non-synonymous in terms of the resulting amino acids coded for. For the Kan / Neo open reading frame there were 105 (76%) non-synonymous changes and 33 (24%) synonymous changes and for the GFP open reading frame there were 115 (75%) non-synonymous changes and 38 (25%) synonymous changes. Given that the probability of a random mutation causing a synonymous or non-synonymous change is 24% and 76% respectively (generated by mathematical simulation), the data in this sample show that the observed amino acid changes are random. Despite these probabilities mutation studies, generally, do not usually uncover point mutations in line with the ratio of synonymous to non-synonymous mutations observed here. This is because non-synonymous mutations are more likely to be deleterious and result in changes that prohibit the natural selection of these mutation-containing genes, and so it is more common to find synonymous mutations. Therefore, this would indicate that the mutations found in this study are not under the influence of natural selection. This is likely to be due to their extremely low frequency. It is highly likely that a given cell will contain more than one copy of recombinant plasmid DNA, because this is a trait that will be selected for through G418 resistance and FACS events and so if one of these plasmid copies contains a mutation that effects phenotype then it can be compensated for by other, unchanged, plasmid copies.

5.2.7. Genome-Integrated Plasmid: High Generation Number

High generation DNA samples were prepared using the same protocols as with the low generation sample, in which genomic DNA was purified, four recombinant plasmid DNA regions were amplified through PCR and pooled together into one sample. SMRT sequencing of this sample was carried out by GATC Biotech. Primary analysis filtering for ROIs with a minimum of 10 passes, 99% predicted accuracy and a minimum length of 800 bp generated 40,315 ROIs, with a mean length of 1336 bp, a mean quality of 0.9935 and a mean pass number of 21.936. BLASR alignment software aligned 40,968 ROIs to the reference sequence with a minimum percentage identity of 95%. The number of ROIs was decreased to 40,924 after fragments containing more than 3 mutations were excluded. These ROIs were taken forward to secondary sequencing analysis.

Figure 5.17 shows the sequencing coverage of plasmid DNA in the high generation DNA sample. The mean coverage of this sample was 10,253, ranging from 0 to 22,570. The coverage seen here is clearly different to the coverage seen in the plasmid stock and non-integrated transfected samples. Again, the pooling together of four separate PCR reactions resulted in four predominant plasmid sequence coverage frequencies. The coverage at the very start of the sequence (positions 1-77) is approximately a 2-fold lower than the rest of the sequence from the same PCR reaction. The overlapping regions between the separate PCR-based sequences result in spikes of coverage, because these plasmid regions are being covered by two PCR templates. Again, the coverage of plasmid positions 3,434 and 4,653 are extremely low, being covered 0 and 2 times respectively. The vast majority of the plasmid positions within this data reside within the four main PCR-based frequency populations seen in figure 5.17, which range in averages from 7,521 to 13,290.

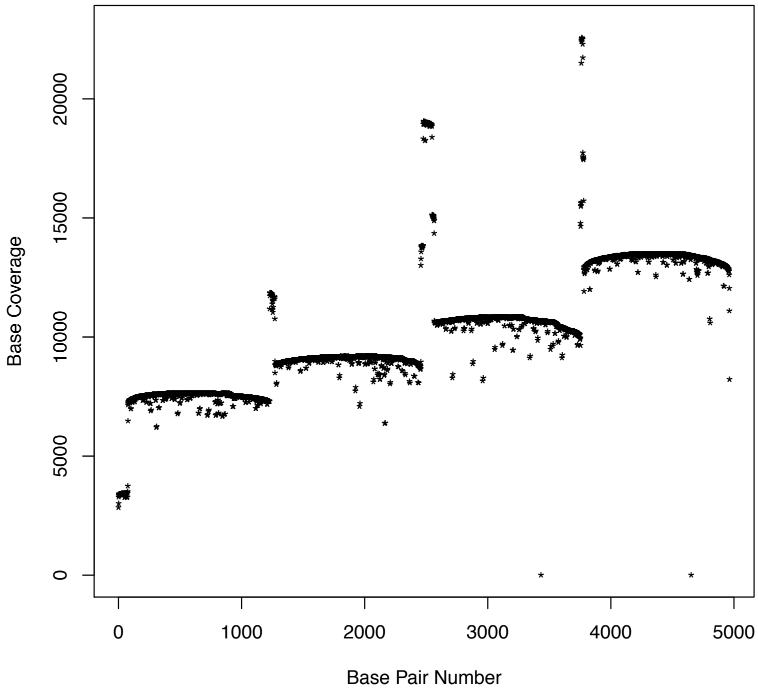


Figure 5.17. High Generation Sample Coverage
The figure illustrates the coverage of each base pair across the 4966 bp – long GFP plasmid in the high generation recombinant plasmid DNA sample.

Figure 5.18a shows the complete collection of point mutations detected by the secondary sequencing analysis platform in terms of plasmid location and frequency in the high generation genomic sample. Overall there were 2550 mutated plasmid positions detected. As was seen in previous samples, a C → T transition in the bacterial origin of replication (position 2539), was present in 15,010 of 18,097 fragments (14,922 out of 18,746 after filtering). Again, we assume here that a mutation called at this frequency is genuine. As can be seen, the other detectable mutated plasmid positions in this sample have a much lower frequency. Figure 5.18b shows the same dataset, but scaled in for examination of the low frequency mutations. After quality score filtering (Figure 5.18c) 1724 mutated plasmid positions were detected. With the exclusion of the mutation detected at position 2539, there were 1723 mutated plasmid positions, which had an accumulation of 3095 mutation events. After the data was filtered for mutations occurring more than once (Figure 5.7d) only 512

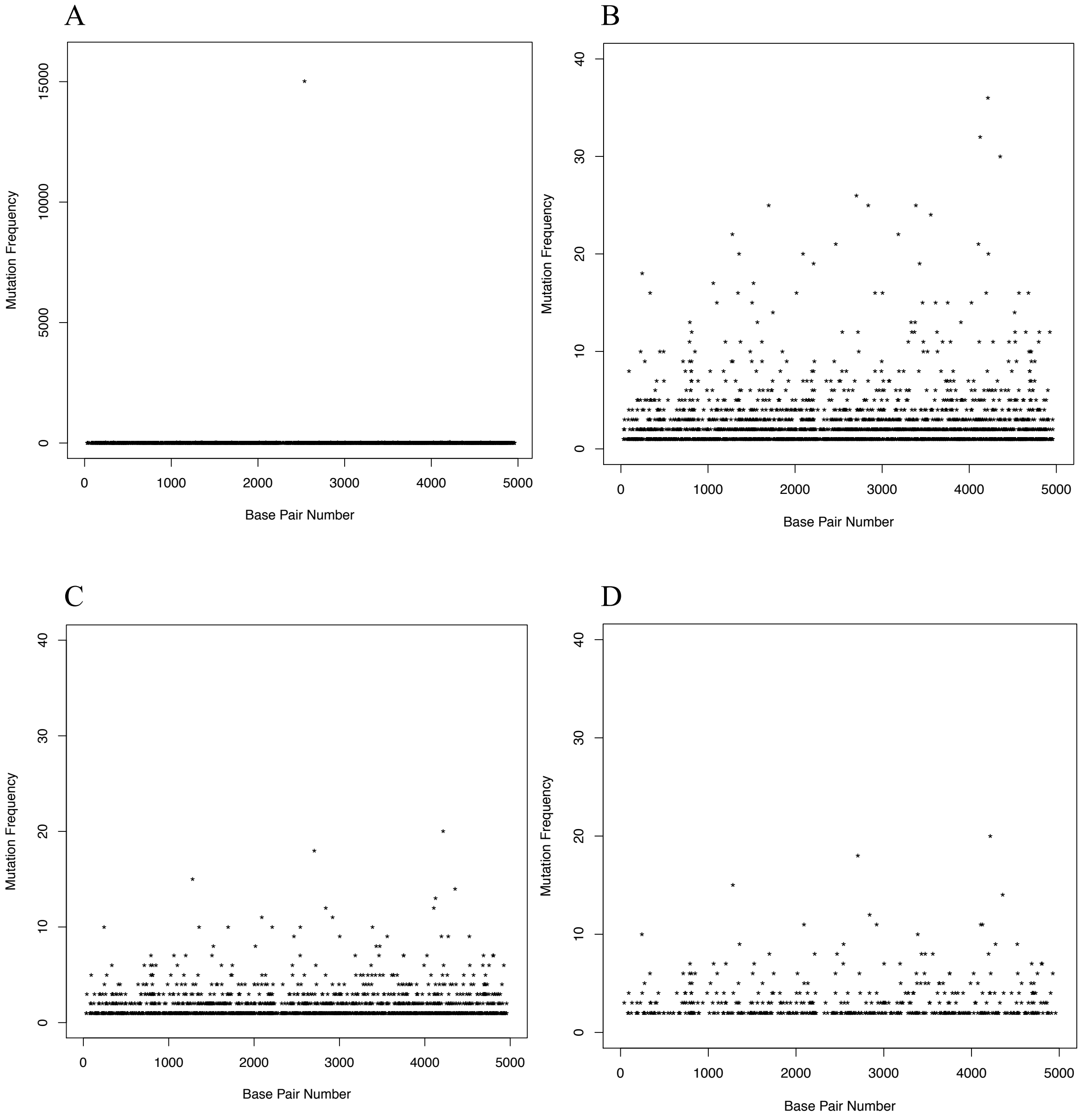


Figure 5.18 High Generation Recombinant Plasmid Mutation Frequency
This figure shows the frequency and locations of detected point mutations in the high generation recombinant plasmid sample. All observed (A), low frequency (B), low frequency quality filtered (C) and low frequency quality filtered and >1 filtered (D) point mutations are shown.

mutated bases were detected. Excluding mutation 2539, 511 mutated plasmid positions were detected, which had an accumulation of 1590 mutation events. Mutation seems to be randomly distributed along the full range of the plasmid. The total number of called bases that passed the quality score filter was 50,279,121. Therefore, depending on filtering stringency, the mutation rates within the low frequency mutation dataset were 1 in 1.6 x 10^4 and 1 in 3.2 x 10^4 for the Q score and >1 filters respectively. These mutation rates are approximately 35-fold and 63-fold higher than those seen in the plasmid stock negative control for the Q score filter and >1 filter mutation rates respectively. Again, there are mutations in more plasmid positions and with higher frequencies than seen in the plasmid stock control. Mutation frequencies for both filters are approximately 1.3-fold lower in the high generation same when compared to the low generation sample. Figure 5.5 confirms that these trends are still apparent when mutation frequencies are normalised by sequence coverage. By number, using the >1 filter, there are 227 more mutated plasmid positions in the low generation sample when compared to the high generation sample. This difference can be broken down into 251 maintained mutated positions, 528 lost mutation positions and 278 gained mutation positions. This is further evidence of mutations occurring during ~76 generations between sampling over long-term cell culture. Depending on filter stringency, the mutation rates given here suggest that 1 in every 3.2 or 6.4 of these 5 kb plasmids contain a point mutation. The average rate of mutation found across the two genomic samples is 1 in 4 plasmids (5 kb).

Table 5.3 contains all the genetic elements of the phCMV C-GFP plasmid and the percentage of mutations that fall within each element from the >1 – filtered high generation dataset. As with the low generation dataset, where two or more elements overlap, a separate element is designated so not to count mutations more than once. Mutation percentage was normalised to element length to correct for the potential bias of element sequence length. In this sample mutations were detected in all element types, apart from the Kan / Neo polyadenylation sequence. Again, there was not a substantial difference in mutation frequencies between coding and non-coding DNA. MCS 2 had a substantially higher mutation frequency than other plasmid sequence elements.

Sequence Element	Mutation Frequency (%)	Mutation Frequency (Normalised by element length)
pAmp	0.2	6.9
pSV40	6.8	29.57
Kan / Neo	15.6	19.62
HSV_TK_PolyA	0	0
Puc_Ori	14.8	22.98
phCMV + Intron	17	24.78
phCMV + Intron + MCS1	0.6	12.5
pT7	0.6	37.5
MCS1	1	20
GFP ORF	14.3	19.86
MCS2	3.7	52.86
SV40 PolyA	0.8	15.69
Non-Coding	24.6	15.31

Table 5.3. High Generation Sample: Mutated Genetic Elements
The table contains the percentage of mutations that fall within each genetic element of the phCMV C-GFP plasmid and the normalised mutation value relative to the length of each genetic element. The sequence elements are as follows: Ampicillin resistance gene promoter, SV40 promoter, Kanamycin / Neomycin resistance gene, HSV Thymidine Kinase polyadenylation signal, pUC origin of replication, Human CMV promoter enhancer and intron, overlapping region of Human CMV promoter enhancer and intron plus the multiple cloning site upstream of the GFP open reading frame, T7 promoter priming for sequencing, multiple cloning site upstream of GFP open reading frame, GFP open reading frame, multiple cloning site downstream of GFP open reading frame, SV40 polyadenylation signal sequence and non-coding DNA.

Table 5.4 shows the frequency of each type of point mutation, along with a total frequency of changed nucleotides, from the >1 – filtered high generation dataset. Again there is a clear predominance in point mutations of G and C nucleotides, showing 42.19% and 41.21% of changes respectively. Upon closer inspection, the frequency of mutation types seen here differ from those in the low generation sample. G.C → A.T transitions were predominant, but G.C → A.T transition (C → T (29.3%) and G → A (29.69%)) mutations were more common than C.G → A.T transversion (C → A (11.13%) and G → T (11.91%)) mutations. Again, it should be noted that the GC content of the plasmid reference sequence is 50.7%, so will not have influenced these results.

		To				
		A	T	C	G	Total
From	A	--	0.2	1.37	7.62	9.19
	T	0.59	--	6.84	0	7.43
	C	11.13	29.3	--	0.78	41.21
	G	29.69	11.91	0.59	--	42.19

Table 5.4. High Generation Sample: Nucleotide Changes
The table shows the percentages of each type of nucleotide change seen within this dataset, the sum of which are used to give the total percentage change for each nucleotide.

The >1 – filtered high generation dataset was used to determine whether the observed DNA point mutations were synonymous or non-synonymous in terms of the resulting amino acid sequences coded for by the two open reading frames, Kan / Neo and GFP. For the Kan / Neo open reading frame there were 54 (67%) non-synonymous changes and 27 (33%) synonymous changes and for the GFP open reading frame there were 55 (71%) non-synonymous changes and 22 (29%) synonymous changes. The ratio of synonymous to non-synonymous mutations deviates slightly more from the ratio expected from completely random mutation (24%:76%) than the low generation sample, but still vastly deviates from ratios commonly found in mutational studies, which indicates that natural selection has not solely impacted upon the mutation frequencies observed here. However, the increase in the percentage of synonymous mutations may indicate that natural selection is slowly acting upon this population, but this could be a result of random fluctuations between samples.

5.2.7. PCR-based error

Overall, the results here have shown that point mutations predominantly occur after plasmid integration. However, sample preparation of genome-integrated samples involved PCR, whereas the pre-integration samples did not. Therefore, PCR error needed to be eliminated as a potential source of these observed mutations. The reported error rate of the Phusion polymerase when using the High Fidelity buffer is 1 in 4.4 x 10^7 (Ingman and Gyllensten, 2009). Using the ThermoFisher Scientific online PCR Fidelity Calculator (Thermo Fisher Scientific, n.d.), with inputs of the length of the PCR product (an average of 1338 and 1336 bp for the low and high generation samples respectively) and the number of PCR cycles (40) used, the approximate PCR error rate was calculated for the two genomic samples. It was calculated that approximately 2.35% of ROIs in the low and high generation samples would contain 1 error. This percentage was used in comparison with the Q-score-filtered dataset. For the low generation recombinant sample, an error rate of 2.35% in 51,635,389 bases from ROIs with an average length of 1338 would yield 907 PCR-originating point mutations. For the high generation recombinant sample, an error rate of 2.35% in 50,279,121 bases from ROIs with an average length of 1336 would yield 884 PCR-originating point mutations. The mutation frequencies observed in these samples are 4.7-fold and 3.5-fold higher than the estimated level of PCR-based errors, for low and high generations respectively, and so are likely to be a result of genuine occurrences of point mutation. Even though the majority of mutations uncovered in this study are likely not to be a result of PCR error, PCR-based errors may still be frequent enough to skew the dataset. A previous study regarding the fidelity of the Phusion polymerase, which confirmed the reported manufacturer error rate, revealed that errors were predominantly transitions (~60%) rather than transversions (Kinde et al., 2011) and further manufacturer in-house data has revealed a predominance of C → T and G → A transitions (personal correspondence with New England Bioscience technical support), which is the same predominance shown in this study. However, various studies have shown that these types of mutations are also predominant in CHO and other mammalian cell DNA replication (Dejong et al., 1988, Gojobori et al., 1982, Hauser et al., 1987). Therefore, although it may be exacerbated by PCR, the trends found in base-pair bias are likely to be genuine trends of point mutation occurrence in CHO cells over long-term cell culture.

5.3. Discussion

5.3.1 Summary and Conclusions

As stated in the introduction to this chapter, recombinant protein-producing CHO cell lines have been shown to produce product variants in the form of amino acid sequence changes. Many of these changes have been attributed to non-synonymous point mutations in the recombinant DNA sequence (Harris et al., 1993, Ren et al., 2011). A number of these point mutations have been shown to originate during long-term cell culture of stable cell lines after the plasmid, coding for the protein of interest, has integrated into the host genome (Zhang et al., 2015). Other studies have shown that point mutations were found to occur in plasmid DNA immediately after transfection into mammalian cells, before genome integration (Hauser et al., 1987, Lebkowski et al., 1984, Lechardeur et al., 1999). Therefore, it is possible that point mutation-derived product variants in recombinant CHO cell lines could result from DNA polymerase replication error of genomic DNA or the potential mutative environment of the cell cytosol or nucleus.

To our knowledge only one study (Zhang et al., 2015) has investigated point mutations in recombinant CHO cell lines using NGS without prior knowledge of sequence variants. The Zhang et al. (2015) study was carried out on 11 CHO cell populations, derived from limited dilution transfectant, clones, and subclones, in which 3 mutations were identified. So, although the Zhang et al (2015) study provides insight into recombinant DNA point mutations in CHO cell populations and a novel use for RNA-seq in mutation identification, the restrictions in cell heterogeneity (dilution, cloning and subcloning) limit the number of observable mutations. The use of clonal, or nearly clonal, cell populations in these types of studies ensure the frequency of a unique mutation is high enough to be detected by NGS technologies, because it ensures that DNA samples contain many copies of the same ‘version’ of a plasmid. There have been reports of Illumina-based sequencing detecting mutations at < 5% frequency (Spencer et al., 2014) and the Pacific Biosciences lower limit to PacBio standard variant calling is reportedly 1% (Dilernia et al., 2015). Previous studies, presumably, have been devised around these reported detection sensitivities. Without the imposition of cell heterogeneity restrictions (e.g. in a non-diluted transfectant pool), many more recombinant plasmid ‘versions’ would be sequenced. Indeed, the frequency of any given mutation would be lower, but there would be a higher number of unique mutations present. This was the premise behind the analysis platform devised in this study, which can detect these low frequency mutations and provide a more in-depth characterisation of them. This study aimed to push the limits of SMRT sequencing by maximising the accuracy in the sequencing of individual molecules. Consensus sequence generation between molecules was avoided, so that rare mutations were not diluted to the extent that they could not be detected.

Sequencing was carried out on DNA from linearised plasmid stocks, transfected but not integrated linearised plasmid, and genome integrated plasmid from two time points in long-term cell culture (low and high generation). SMRT sequencing was carried out on fragments (through fragmentation or PCR) of the ph-CMV C-GFP vector. Primary analysis generated ROIs with a minimum predicted accuracy of 99% and a minimum length of 800 bp. This was carried out for a range of minimum pass numbers (0, 5, 10, 15, 20) required to generate a consensus sequence. Using BLASR, sequences were aligned to a plasmid reference sequence with 95% matched identity, generating data on sequences, sequence coverage, and sequence quality. A novel secondary analysis platform was then used to report all called nucleotides at each plasmid position, using various stringencies of error-eliminating filters (Removal or error-prone ROIs, Q score filtering and > 1 filtering). Plasmid mutation was then assessed for frequency, position, type and impact on amino acid sequences. Final analysis and conclusions were drawn from the 10 – pass datasets, because they delivered the highest coverage from the datasets deemed to have low error frequencies.

The average coverage of samples varied: linearised plasmid stock – 6,600; transfected / non-integrated plasmid – 4,319; Low genomic sample – 10,525; High genomic sample – 10,253. These values of coverage are derived from 10 – pass ROIs and so arguably are more accurate than 1x coverage in other sequencing methods. The discrepancy between total ROIs (30,824) and aligned ROIs (15,063) in the transfected / non-integrated sample is the reason for the lower coverage seen in this sample. This was due to the inability of the Blue Pippin instrument (Sage Science, MA, USA) to remove non-plasmid DNA from the sample. Carrying out Blue Pippin purification using the same conditions as the validation study would be more likely to remove a greater proportion of non-plasmid DNA and result in increased sequence coverage. Coverage from PCR-derived samples was noticeably different from non-PCR-derived samples, in that the four PCR-fragments had distinct coverages, presumably due to the their relative concentrations within the pooled samples. Interestingly, two plasmid positions (3434 and 4653) were consistently covered at low frequencies (ranging 0 to 4), which could be due to an inherent issue with sequencing at this position.

All samples were found to have a high frequency C → T transition in the bacterial origin of replication (plasmid position 2539) in > 99.9% of fragments covering this position. The same plasmid stock was used throughout this study. We assume here that this mutation was present in the initial plasmid stock from the manufacturer. However it is possible that the mutation originated from a DNA replication error during E. *coli* cell divisions during plasmid cloning, but the error would have had to of occurred during an extremely early plasmid replication.

Other than the 2539 mutation, the observed mutation frequency in the linearised plasmid stock sample was extremely low (Q score filter: 1 in 5.6 x10^5, >1 filter: 1 in 2 x 10^6). Although it is possible that here we are observing genuine low frequency mutation as a result of rare E. *coli* DNA replication errors, it was deemed more likely that these were representative of false positive call rates within this sequencing analysis platform. Therefore, this sample was used as a negative control sample for point mutations in this study.

The level of mutation observed in the transfected / non-integrated plasmid sample (Q score filter: 1 in 4.7 x10^5, >1 filter: 1 in 1.1 x 10^7) did not substantially surpass the level seen in the negative control and so the conclusion drawn here is that the pre-integration cellular environment did not cause point mutations in plasmid DNA. However, previous studies investigating the putative mutagenic environment of a mammalian cell utilised protocols, in which transfected plasmid extracts were then transformed into a bacterial host to identify mutations. It was hypothesised that mutations are a result of DNA damage in the mammalian cell environment, such as Cytosine damination, depurination of Guanine residues or through nuclease attack (Hauser et al., 1987, Lebkowski et al., 1984). These transformed DNA molecules are presumably replicated or transcribed by a DNA polymerase before assessing the DNA for mutation. Theoretically, a mutation will only be present once this DNA damage is misread by a replicase or polymerase. In this study the DNA in the transfected / non-integrated sample was deliberately left unamplified due to concerns that PCR-based errors may be at a greater frequency than mutation itself, which may have only been present as a single copy. However, perhaps there were DNA damage events, which had, in essence, marked a given nucleotide for point mutation, but there was a lack of replication to consolidate this change before sequencing and so they were left undetected. The PacBio sequencing polymerase will not have served this purpose, because a DNA damage repair step in sample preparation removes DNA damages such as cytosine deamination and oxidative damages, so that the polymerase does not stall during sequencing (Pacific-Bioscience, 2010). Therefore, these mutations were unlikely to have been detected in the transfected / non-integrated sample in this experimental design. It might be the case that some mutations detected in the genome-integrated samples (Low and High) were caused by pre-integration damage. So, although the sequencing of this sample determined that there is no observable point mutation occurring before genome integration, it was unable to address the hypothesis that DNA is somehow marked for mutation upon replication.

As mentioned in chapter 4 in regards to cell viability and average cell diameter responses to transfection, electroporation of plasmid DNA has a substantial impact on cell health in that it is known to cause apoptosis, which is presumed to be due to a cellular response in line with the response to a viral attack (Shimokawa et al., 2000). One observation of this apopotic response is genomic DNA fragmentation, which gives rise to gel banding patterns not dissimilar to the unidentified contaminant DNA in the transfected non-integrated sample (Nagata, 2000, Ioannou and Chen, 1996). This is a heavy indication that a proportion of cells in this study were undergoing apoptosis. Not all cells undergo apoptosis-mediated cell death as a result of DNA electroporation, but it might be the case that cells elicit a response as a result of electroporation stress. Indeed, it could be worthwhile to investigate the global cellular response to electroporation. Mammalian cells are known to detect the presence of foreign DNA and have been shown to silence transfected plasmid DNA (Orzalli and Knipe, 2014). Furthermore, the redox state of the cell is known to change as a result of apoptosis (Slater et al., 1996, Bustamante et al., 1997). Changes such as this to the cellular environment could play a role in the putative mutations that occur as a result of pre-integration damage, meaning point mutation is an indirect cellular response to electroporation.

The level of mutation observed in the genome-integrated plasmid copies was considerably greater than in the linearised plasmid stock negative control. Mutation frequency was higher in the low generation sample (Q score filter: 1 in 1.2 x 10^4, >1 filter: 1 in 2.1 x 10^4) than in the high generation sample (Q score filter: 1 in 1.6 x 10^4, >1 filter: 1 in 3.2 x 10^4). The mutations observed here were predominantly observed between 1 and 20 times and were shown to be well above the level of mutation expected from PCR-based errors alone. Indeed, the assumption that these mutations are genuine is made more likely by the fact that the 11% error rate of a single pass in SMRT sequencing is predominantly due to indel errors (Carneiro et al., 2012). Upon closer inspection, this difference was a result of hundreds of mutation gains and losses, and so it is difficult to establish whether the difference in mutation frequency between these two samples is due anything other than random fluctuations of observed mutations in a given sample. The data here clearly show strong evidence of mutation in recombinant plasmid DNA, which are most likely a result of DNA replication errors. Generally, it would appear that there is no evidence to strongly suggest that mutation is anything other than randomly distributed across the plasmid, with genetic elements and non-coding regions showing no observable difference in mutation frequency. Mutations were observed in all genetic elements other than the polydenylation signal sequence (HSV_TK_PolyA) for the Kanamycin / Neomycin resistance gene, which could be a result of sequence conservation through natural selection. On the other hand, this sequence is only 19 bp long and may not have been mutated due to the random distribution of mutations across the length of the plasmid. MCS sequences appeared to be mutated to a greater extent than other sequences, but again, this could be down to chance. There was a clear bias in the type of mutation seen in these samples. G and C residues (~85%) were mutated to a far greater extent than A and T residues (~15%). In the low generation sample G.C → A.T (19.22%, 24.9%) transitions and G.C → T.A (22.6%, 18.54%) transversions were the predominant mutations observed, whereas in the high generation sample the G.C → A.T (29.69%, 29.3%) transitions became more predominant than G.C → T.A (11.91%, 11.13%) transversions. A and T residue changes also showed a higher level of transition mutation than transversion mutation. The rates of mutation type seen here are in line with mutation occurrences reported in other mammalian cells, both as a result of genome replication and pre-integration mutation (Dejong et al., 1988, Gojobori et al., 1982, Hauser et al., 1987).

The observed point mutations were then used to determine the subsequent amino acid sequences of the Kan / Neo and GFP ORFs. The Kan / Neo ORF was subject to 138 and 81 mutations, of which 76% and 67% were non-synonymous changes, for low and high generation numbers respectively. The GFP ORF was subject to 153 and 77 mutations, of which 75% and 71% were non-synonymous changes, for low and high generation numbers respectively. Generally speaking, in most mutation studies the rate of synonymous mutation is far higher than the rate of non-synonymous mutation, because non-synonymous mutations are likely to be deleterious and as such are selected against evolutionarily. On the other hand synonymous mutations are neutral, or at least nearly neutral, and so their rate of prevalence and fixation is subject only to random genetic drift (Nei and Gojobori, 1986, Kimura, 1979). Indeed, a recent study into CHO cell SNPs revealed that only 0.15% of discovered SNPs were non-synonymous (Lewis et al., 2013). The raw probabilities of the occurrence of non-synonymous and synonymous mutations are 76% and 24% respectively. The mutations identified in this study seem to adhere closely to the raw probabilities of non-synonymous and synonymous mutation occurrence and are apparently not being affected by natural selection. This is most likely explained by the extremely low frequency that these mutations reside within the total population. It is likely that many of the cells harvested for recombinant plasmid contain more than one copy of plasmid DNA, because these cells are more likely to have been included in the high producers that were selected during FACS. Therefore, if one of these copies contained a non-synonymous point mutation, any deleterious affects could be compensated by other gene copies. Moreover, after these sorting events the only genes on which a selection pressure is imposed code for elements influencing cell growth. Therefore, after FACS, changes to the GFP ORF are not influenced by natural selection. In theory, the Kan / Neo ORF sequence should be constantly fixed by natural selection, because it is essential to the growth and survival of the cell in G418 media. However, as was shown during stable cell line generation, G418 selection was not sufficient for cell line selection. Either, cells had become resistant to G418 irrespective of plasmid copies or the resistance achieved by a proportion of cells could provide resistance to many of the remaining cells of the population. This could due to resistance protein secretion. Therefore, as long as there is plentiful supply of resistance protein within the population, cells can tolerate deleterious mutations.

In summary, this study has shown that ~25% of the plasmid copies used in this study were mutated over long-term cell culture and that there was no evidence of mutation occurring before integration. Due to their low frequency, natural selection does not impact strongly on the prevalence or fixation of these mutations, which means they can reside anywhere along the length of the plasmid and result in non-synonymous changes more often than would be expected (~72% of the time). G and C residues were found to be mutated more frequently than A and T residues, with G.C → A.T transitions being predominant. This appears to be in line with mutation patterns that have been found to occur in other studies into mammalian cell mutation. The novel analysis platform used in this study adeptly identified mutations at a resolution beyond what is generally reported in NGS studies, using careful and logical elimination wherever possible. Due to the necessity for high resolution accuracy is sacrificed, despite this error elimination. However, the conclusions here were made using trends on the dataset as a whole, which adds a certain level of confidence to the findings. This study has confirmed the need for sequence variant screens in cell line development. Despite the success of this high-resolution platform, it is far more practical in terms of cost and time to screen clonal cell line candidates, which need lower resolution sequencing technologies. However, this platform could find other avenues for application, such as checking the homogeneity of gene therapy DNA stocks or for a higher resolution analysis of cancer genetic heterogeneity.

5.3.2. Future Work

The DNA sequencing secondary analysis platform outlined in this chapter has been shown to effectively detect extremely rare mutations. However, there are experiments that could be carried out to further validate its efficacy. The calculations to rule out PCR-based error in this study showed that the mutation detected in the low and high genomic samples was genuine. However, a plasmid stock negative control, which has undergone PCR would more effectively quantify the exact level of PCR-based error that made it through the error filters put in place. Changes to the PCR process, such as the use of less PCR cycles or the use of a more high fidelity DNA polymerase, such as Q5 polymerase (New England Biolabs, UK), would also help quantify this source of error more accurately.

Further validation of this platform could be carried out through mutagenesis studies, whereby DNA mutations are deliberately induced to different extents, using techniques such as UV radiation or error-prone DNA polymerases. Different samples would have different levels of random mutation, which, in theory, should be quantified using this analysis platform. Moreover, a study could be conducted using a similar format to (Spencer et al., 2014), in which a DNA template is synthesised with a range of known mutations along its length in comparison to the non-mutated reference. Through dilution, samples are then made from these sequences, with varying proportions of the mutated version. This would offer a precise evaluation of platform accuracy. Although this would not involve the discovery of unknown mutations, it would offer a more accurate insight into the top end of resolution that can be achieved using this platform.

A mutation detection analysis of the dataset used in this study with the Pacific Bioscience variant caller would provide an accurate evaluation of the difference in resolution between the platform devised in this study and the standard platform used for SMRT sequencing.

As discussed in this chapter, the experimental setup in this study may not have been sufficient to identify mutations that were caused as a result of DNA damage before plasmid integration into the host genome, because the DNA used was unreplicated and was subjected to DNA repair before sequencing. A future study could consist of purifying plasmid DNA from CHO cells as it was done in this chapter, but then transforming the DNA into E. *coli* DH5α cells for replication, which was shown to be relatively error-free in the sequencing of the plasmid stock sample. If multiple clones from this transformation were pooled together to prepare DNA for sequence then a large collection of these putative mutations could be detected.

Finally, as was discussed in this chapter, it is difficult to discern whether an individual mutation discovered in this study is genuine or a result of sequencing or PCR-based error. To characterise genuine mutations, a number of clones or extremely diluted cultures could be generated from the working cell banks of the low and high generation stable GFP cell line samples. These clones / cultures would contain a much smaller number of plasmid versions compared to the whole cell population. Sequencing of the plasmid DNA derived from these cultures would lead to the identification of genuine mutations, because they are present at a much higher frequency.

This page is intentionally left blank.

Chapter 6

Concluding Remarks

This chapter will give a brief summary of the findings, conclusions and the impact of the work presented in this thesis.

6.1. Chapter 3 – Genomic Instability

Genetic instability is an inherent feature of CHO cells lines. The lack of evolutionary constraint within the cell culture environment leads to genetic drift within the CHO genome, whereby genomic sequences that do not directly influence growth characteristics are not heavily influenced by natural selection in terms of their consistency through generations of cell culture (Kim et al., 2011, Kimura, 1955, Kimura, 1979). Therefore allele frequencies will gradually change and the propagation of, potentially detelerious, genetic changes is more likely. This instability means that CHO cells can be moulded into cell factories with a range of desirable phenotypes, which is put to good use through evolution and engineering strategies in the generation of commercial cell lines (Sinacore et al., 2000, Prentice et al., 2007). However, this phenotype, whilst desirable for these evolutionary strategies, becomes problematic in the long-term cell culture of productive cell lines. Phenotypic drift causes these desirable cell lines to deviate from the phenotypes by which they were once selected. Indeed, this instability means that it is difficult to maintain consistent phenotypes for the duration of the production process. Despite undergoing cloning procedures, cell heterogeneity an inherent feature of CHO cell lines, which often leads to a decline in productivity and concerns over product quality (Barnes et al., 2006, Kim et al., 2011, Ren et al., 2011, Davies et al., 2013). CHO cells have been said to have a mutator phenotype (Kim et al., 2011), which has been shown to be the case at the chromosome level (Yoshikawa et al., 2000, Derouazi et al., 2006), through recombinant gene copy loss (Kim et al., 2011), and at the base pair level through the appearance of sequence variants and a plethora of SNPs (Zhang et al., 2015, Lewis et al., 2013). If understanding of these genetic changes was further elucidated, then there could be potential for engineering strategies to generate more stable cell lines, such as to bolster proof reading capabilities or to select slower adapting cell lines in an attempt to select for genetic fidelity. On the other hand, instability may not be trait of cells in culture that is easy, or even possible, to eliminate. In this case, efficient screening tools to quickly identify unstable or error-containing cells lines may be able to eliminate candidate cell lines from production pipelines. In this chapter genetic instability was measured at the base pair level, via microsatellite analysis and at the chromosomal level via karyotype analysis.

The microsatellite analysis showed the slow, progressive change in allele frequencies by genetic drift and allowed for the relatedness of cell lines to be established through microsatellite allele similarities and differences. There was an indication, but no conclusive evidence, of a physical change to microsatellite length through replication slippage. Therefore, it could not be concluded that this selection of microsatellites were able to be used as a successful marker for changes at the base pair level. There was no correlation between cell line genetic drift and changes in cell specific productivity. Microsatellites differed in their level of change, which shows that different genomic loci are more changeable than others. For microsatellite analysis to be validated as a useful marker and screening tool for base pair level genetic instability and drift, a greater number of microsatellites, spanning the whole genome, at a high resolution would need to be used.

Karyotype analysis revealed that chromosomal instability is substantial, with changes in chromosome number and chromosome breakage / fusion events both contributing to this instability. Over long-term cell culture 70% of cell lines were shown to change in karyotype, which included the generation of 18 chromosome types that were not seen in parental cell lines. Karyotype analysis is not quantitative, so we were unable to establish whether chromosomal instability correlated with observed changes to cell specific productivity. Some chromosomes or chromosomal regions, such as chromosome 1, remained unaltered for the duration of the study, which could be due to an evolutionary conservation effect. Perhaps targeted integration to these, stable, regions might lead to greater phenotypic stability in important production process attributes. Again, further study here may lead to evolution, engineering or screening / selection strategies to facilitate the use of more stable cell lines for production pipelines.

6.2. Chapter 4 – Electroporation Optimisation

Chapter 4 presented a complete optimisation of plasmid DNA delivery into CHOK1SV cells by electroporation. Electroporation is a key part of the bioprocess, because it marks the start point of the generation of a stably producing cell line. It is also used in bioprocess development, whereby new therapeutics are tested for performance attributes in transient production platforms (Jayapal et al., 2007, Wurm, 2004, Makrides, 1999). Therefore, techniques that deliver the ability to fine tune this process for the bespoke requirements of any given therapeutic production platform could be put to good use in an academic or industrial setting. The need for bespoke parameters become apparent when comparing the requirements of different stages of bioprocesses. For example, in the generation of a stable cell line an increase in plasmid copy numbers entering the cell could lead to an increased number of integration events, which in turn could lead to a greater probability of generating high producing cell lines from a cloning procedure. Moreover, with TGE, increasing the number of plasmid copies entering the cell will increase the level of plasmid copies capable of gene expression. However, during the SGE process, cells are allowed the time to recover from electroporation during recombinant cell selection and enrichment, whereas in transient platforms cells are required to achieve high culture densities and productivities immediately (Wurm, 2004, Rita Costa et al., 2010). Therefore, optimum TGE platforms will require higher levels of cell viability and growth post-electroporation, whereas a lag time in electroporation recovery might be a worthwhile sacrifice in SGE platforms. Moreover, cells are typically transfected with linear DNA for the generation of stable cell lines, whereas TGE is carried out with circular plasmid. Linear DNA is more difficult to transfect, and so electroporation parameters will likely differ between the two platforms (Schmidt et al., 2004).

When optimising a bioprocess for a new therapeutic candidate, in many cases a number of variables will differ compared to other therapeutics, such as vector elements and size, cell type, product expression and product impact on growth and viability (Wurm, 2004, Jordan et al., 2007, Jordan et al., 2008, Kim et al., 2011). Therefore, an electroporation optimisation platform that enables the quick and easy assessment of protocol permutations will allow for the easy implementation of bespoke conditions for each new candidate. This study clearly provides such a platform. Using a simple DoE strategy, a range of parameters (310 – 320 V, 25-28 ms, exponential day time constant protocol) resulting in positive range of transfection response activity, was discovered for the phCMV C-GFP plasmid being used. After this range was tested experimentally, one parameter setting was clearly seen to offer the best response (320-26). These conditions resulted in an improvement of 17% transfection efficiency, which was achieved without greatly sacrificing on the health of the transfected cells. These optimised conditions were shown to be successful in chapter 5 when generating a stable cell line for DNA sequencing analysis. Not only were conditions improved, but a DoE analysis allowed for the interactive nature of the different electroporation parameters to be identified. Indeed, field strength, pulse length and DNA load were all found to interact in their effect on the transfection response. Moreover, the relationship between transfection efficiency and cell viability was reasonably well defined, to the extent that cell viability alone was able to successfully predict a design space that would yield a high level of gene expression. If this work was to be taken further, whereby a number of different protein products, cell types and DNA vectors were used then, not only would the relationships discussed in this study be more acutely understood, but a certain level of predictability may be possible for the optimisation of future platforms. For example, the optimisation process for a new therapeutic gene, contained within a well defined vector of a particular size, being transected into a well characterised cell type could be started within a much narrower range of electroporation parameters, because a model-based information repository could accurately provide the predicted parameter range that would yield positive results. Indeed, this narrow range of parameters may only need to be assessed using a cell viability output, because the relationship between cell viability and gene expression could be characterised to the extent that it is completely predictive.

A scenario such as this would lead to a high-throughput and cost-effective platform for electroporation optimisation.

6.3 Chapter 5 – Recombinant DNA Sequence Analysis

Regulatory bodies require that the therapeutics produced by bioprocess platforms are of a certain quality. Therefore, product variants, such as aggregates, charge variants, glycosylation variants and sequence variants must be reduced to minimal levels, because of concerns over product safety and efficacy (Zhang et al., 2015, Ren et al., 2011, Zhu, 2012). As discussed in chapter 4, genetic instability is a regularly observed phenomenon in CHO cells, and this is seen to manifest in point mutations. These point mutations have been shown to occur in recombinant DNA (Zhang et al., 2015) and in CHO genomic DNA, through the appearance of SNPs (Lewis et al., 2013). Non-synonymous point mutations in recombinant DNA cause sequence variants, which result in unwanted heterogeneous protein products. Mostly, these sequence variants have been identified at the protein level, and traced back to DNA sequence changes (Zeck et al., 2012, Victoria et al., 2010). Zhang et al. (2015) used NGS to identify DNA point mutations without prior knowledge of protein sequence changes, but this was only carried out in clonal or diluted cell populations. Therefore, only a small range of mutations were identified, and so detailed information on mutation position, type and raw frequency is lacking. The reported resolution of NGS does not allow for analysis on non-diluted or non-clonal cell populations, because mutations need to be at a certain frequency within a DNA sample to be detected (Spencer et al., 2014, Dilernia et al., 2015).

In this study SMRT sequencing was used with an altered analysis platform, in which high-coverage CCS reads were used in order to generate information on point mutations from individual molecules. Various filtering strategies were employed to eliminate error-prone ROIs and individual nucleotide reads. One point mutation, a C → T transition in the bacterial origin of replication, was found to be present at high levels in all samples, which was presumed to have been present in the initial plasmid stock received from the manufacturer, or was a result of a point mutation occuring in an early generation of bacterial cloning. Other than this mutation, it was concluded that plasmid stocks showed no substantial evidence of mutation. The low frequency changes observed in the plasmid stock sample were used as a base level of error for this sequencing analysis platform. There was no evidence of mutation in samples derived from transfected, non-integrated, plasmid DNA. However, further investigation might reveal that DNA is damaged within this pre-integration period, but converted into a mutation upon DNA replication, and so would not be called as a mutation in the experimental platform used here. Other studies have shown that point mutation of plasmid DNA within this period does occur in mammalian cells (Hauser et al., 1987, Lebkowski et al., 1984), so this might be a worthwhile avenue for research. A substantial level of low-frequency point mutation was covered after sequencing recombinant DNA, sampled from two time points in long-term cell culture. On average, 25% of 5 kb plasmid molecules were found to contain at least one point mutation. Mutations were found to be randomly distributed along the length of the plasmid sequence, showing no bias towards coding or non-coding localisation. 85% of point mutations occurred with G and C nucleotides, with G.C → A.T transitions being the predominant type of change observed. This bias is in line with mutation frequency observations of mammalian cell DNA replication (Dejong et al., 1988, Gojobori et al., 1982). On average, within the two plasmid open reading frames, Kan / Neo and GFP, 72.25% of mutations were non-synonymous. This proportion of non-synonymous mutations is in line with the raw probability of a non-synonymous mutation occurring, rather than with the proportion of non-synonymous mutations found in nature (Lewis et al., 2013). The results presented here indicate that natural selection does not greatly impact upon these low-frequency point mutations, but rather that their existence and prevalence is random.

Overall, this chapter showed the preliminary validation of a novel SMRT sequencing secondary analysis platform in the identification of low-frequency mutations from individual DNA molecules. This validation could be built upon with a small set of quantitative controls. Moreover, protein sequence variant-causing DNA point mutations were characterised at a frequency and resolution that, to our knowledge, has not been seen previously.

6.4 Future Directions for Genetic Instability

As has been discussed throughout this thesis, genetic instability of CHO cells poses a threat to cell line development processes and biopharmaceutical production. This instability causes phenotypic drift in cell lines that have been carefully selected for attributes suitable for bioprocesses, such as fast growth rates and high productivity. Instability, gives rise to heterogeneous cell populations, which is clearly an undesirable trait for a 'clonal' cell line. One form of phenotypic drift commonly encountered is a decline in cell productivity over long-term cell culture (Wurm, 2004). This has been shown to be due to epigenetic changes as well as genetic changes, such as changes in recombinant gene copy number (Kim et al., 2011). The seemingly random nature of a cell line's disposition to decline in productivity makes it extremely difficult, if not impossible, to predict. Moreover, it is not the case that genetic instability can be traced back to a specific point mutation in DNA replication / repair machinery or a common chromosome breakage, but rather genetic instability seems to be an almost inevitable global attribute of an immortalized cell line and so prediction or elimination of genetic instability is not straightforward. This is because an immortalized cell line growing in culture is almost in a state of evolutionary freefall, whereby the only genes to be monitored by natural selection are those which contribute to growth and cell division. Other genes, which do not directly impact upon growth and division, are neutral, or at least nearly neutral, in the context of evolution and so are relatively free to change. Therefore, continuous cell culture facilitates an environment in which DNA replication becomes a process that is not constrained to high standards of fidelity and so over time DNA replication becomes an error-prone process and genetic change is commonplace (Kimura, 1955, Kimura, 1979). Conceivably, this process is quickened by cell line evolution and engineering strategies guide cells towards desirable attributes, such as growth, productivity, growth in serum-free media, and adaptation to growth in a late-stage culture environment (Sinacore et al., 2000; Prentice et al., 2007). This is because genetic instability is likely to also be a heterogeneous phenotype and so when a particular cell is selected for a desirable trait it is because that cell has changed genetically to present this phenotype. Therefore, genetic change is being selected for and so the process of selection is likely to increase the likelihood of a genetic instability phenotype. It is perhaps unsurprising that these types of cells would lose the ability to produce recombinant protein, because these cells are simply adapting towards a more desirable phenotype, such that they are able to thrive and grow in a given environment without the metabolic burden of producing a complex protein, such as a Mab (Kim et al., 2011).

The karyotype analysis in chapter 3 and sequencing analysis in chapter 5 illustrate the high frequency and randomness of this genetic instability phenotype at the chromosomal and sequence level respectively. It seems unfeasible that such a global and consistent phenomenon could be targeted by any direct genetic engineering strategy that might attempt to reconstitute a cells ability to accurately segregate chromosomes upon cell division, limit chromosome form changes or increase the fidelity of DNA replication, because it is likely to be a phenotype that has a different origin in any given case and is likely to persist regardless of any tinkering to gene content. It seems far more pertinent to try and development genetically stable cell populations through selection strategies, because selection, as opposed to engineering, is likely to draw upon a whole-cell-based solution. Of course, the ability to select for a pool of genetically stable cells would depend upon a set of robust selection markers for genetic stability. One such marker, as proven in chapter 3, is cell karyotyping (Derouazi et al., 2006). Selection for cells that are less changeable in their karyotype could be a method for generating cell populations better able to maintain a homogenous cell number and that are less subject to changes in chromosome form. As well as a method for generating novel, genetically stable cell lines, periodic karyotype screens during cell line development could serve as a quality control step to prevent or detect the onset of chromosomal instability. Chapter 5 showed that NGS can serve as a selection marker for the fidelity of DNA replication and DNA damage repair. Perhaps a strategy involving the selection of cell populations containing fewer of the low frequency mutations detected in this study would serve to generate cell populations with an improved accuracy in DNA replication. Moreover, sequencing throughout cell line development could serve as a useful supplementary tool for protein sequencing methods to ensure that product quality is maintained. Despite progress being made in enabling high-throughput sequencing of recombinant DNA at a cheaper price (Zhang et al., 2015), NGS is an expensive and relatively time consuming process and so development of cheaper tools, using markers that could stand as proxy for point mutation would make this a much more feasible ambition. Chapter 3 attempted to do this using microsatellite analysis, but was unable to prove its worth as a marker for genomic instability. However, as mentioned in section 3.3.4 further investigation with a larger number of microsatellites, or microsatellites within a recombinant plasmid may be more informative.

As mentioned above, it is could be the case that genetic instability within CHO cell lines is an inevitable bi-product of continuous cell culture and so perhaps attempts to generate cell lines that have a higher level of inherent stability is a futile exercise. Therefore, perhaps a more promising direction would be to accept the unstable landscape of the CHO genome and try to work around it. For example, the karyotype analysis in chapter 3 found that chromosome 1 was unchanged throughout the study and it was postulated that this likely to be because it contains essential genes. There is progress being made into targeted integration of plasmid DNA into genomic sites that are more likely to facilitate high gene expression (Wurm, 2004). Perhaps attempts to target genetically stable sites would lead to the development of cell lines more likely to maintain consistent productivity over long-term cell culture. Strategies like this, in combination with regular quality control measures, such as the karyotype and sequence screens mentioned above, would help to decrease genomic instability manifesting in changes to product yields or quality.

This page is intentionally left blank.

Reference List

ABU-QARN, M., EICHLER, J. & SHARON, N. 2008. Not just for Eukarya anymore: protein glycosylation in Bacteria and Archaea. *Current opinion in structural biology,* 18, 544-50.

AGGARWAL, R. S. 2014. What's fueling the biotech engine-2012 to 2013. *Nature biotechnology,* 32, 32-9.

AGRAWAL, V., YU, B., PAGILA, R., YANG, B., SIMONSEN, C. & BESKE, O. 2013. A High-Yielding, CHO-K1–Based Transient Transfection System.

Rapid Production for Therapeutic Protein Development. *Bioprocess International,* 11, 28-35.

AGUILERA, A. & GOMEZ-GONZALEZ, B. 2008. Genome instability: a mechanistic view of its causes and consequences. *Nature reviews. Genetics,* 9, 204-17.

AKINC, A. & LANGER, R. 2002. Measuring the pH environment of DNA delivered using nonviral vectors: implications for lysosomal trafficking. *Biotechnology and Bioengineering,* 78, 503-8.

ANDERSEN, D. C. & KRUMMEN, L. 2002. Recombinant protein expression for therapeutic applications. *Current opinion in biotechnology,* 13, 117-123.

ANDERSON, M. J. & WHITCOMB, P. J. 2005. *RSM simplified: optimizing processes using response surface methods for design of experiments*, Productivity Press.

ANDERSON, M. J. A. W., P.J. 2007. *DOE simplified: practical tools for effective experimentation*, CRC Press.

ANDREASON, G. L. & EVANS, G. A. 1989. Optimization of electroporation for transfection of mammalian cell lines. *Analytical biochemistry,* 180, 269-75.

AQUILINA, G., HESS, P., BRANCH, P., MACGEOCH, C., CASCIANO, I., KARRAN, P. & BIGNAMI, M. 1994. A mismatch recognition defect in colon carcinoma confers DNA microsatellite instability and a mutator phenotype. *Proceedings of the National Academy of Sciences of the United States of America,* 91, 8905-9.

ARAD, U. 1998. Modified Hirt procedure for rapid purification of extrachromosomal DNA from mammalian cells. *Biotechniques,* 24, 760-+.

BALDI, L., HACKER, D. L., ADAM, M. & WURM, F. M. 2007. Recombinant protein production by large-scale transient gene expression in mammalian cells: state of the art and future perspectives. *Biotechnology Letters,* 29**,** 677-84.

BANDARANAYAKE, A. D. & ALMO, S. C. 2014. Recent advances in mammalian protein production. *FEBS letters,* 588**,** 253-60.

BARBOSA, M. D. 2011. Immunogenicity of biotherapeutics in the context of developing biosimilars and biobetters. *Drug discovery today,* 16**,** 345-53.

BARNES, L. M., BENTLEY, C. M. & DICKSON, A. J. 2000. Advances in animal cell recombinant protein production: GS-NS0 expression system. *Cytotechnology,* 32**,** 109-23.

BARNES, L. M., BENTLEY, C. M. & DICKSON, A. J. 2001. Characterization of the stability of recombinant protein production in the GS-NS0 expression system. *Biotechnology and Bioengineering,* 73**,** 261-70.

BARNES, L. M., BENTLEY, C. M. & DICKSON, A. J. 2003. Stability of protein production from recombinant mammalian cells. *Biotechnology and Bioengineering,* 81**,** 631-9.

BARNES, L. M., BENTLEY, C. M., MOY, N. & DICKSON, A. J. 2007. Molecular analysis of successful cell line selection in transfected GS-NS0 myeloma cells. *Biotechnology and Bioengineering,* 96**,** 337-48.

BARNES, L. M., MOY, N. & DICKSON, A. J. 2006. Phenotypic variation during cloning procedures: analysis of the growth behavior of clonal cell lines. *Biotechnology and Bioengineering,* 94**,** 530-7.

BARON, B., FERNANDEZ, M. A., CARIGNON, S., TOLEDO, F., BUTTIN, G. & DEBATISSE, M. 1996. GNAI3, GNAT2, AMPD2, GSTM are clustered in 120 kb of Chinese hamster chromosome 1q. *Mammalian genome : official journal of the International Mammalian Genome Society,* 7**,** 429-32.

BAXBY, D. 1999. Edward Jenner's inquiry; a bicentenary analysis. *Vaccine,* 17**,** 301-307.

BECK, A., COCHET, O. & WURCH, T. 2010. GlycoFi's technology to control the glycosylation of recombinant therapeutic proteins. *Expert opinion on drug discovery,* 5**,** 95-111.

BERLEC, A. & STRUKELJ, B. 2013. Current state and recent advances in biopharmaceutical production in Escherichia coli, yeasts and mammalian cells. *Journal of industrial microbiology & biotechnology,* 40**,** 257-74.

BIO-RAD n.d. Gene Pulser XcellTM Electroporation System: Instruction Manual.

BIRCH, J. R. & RACHER, A. J. 2006. Antibody production. *Advanced Drug Delivery Reviews,* 58**,** 671-685.

BORK, K., HORSTKORTE, R. & WEIDEMANN, W. 2009. Increasing the sialylation of therapeutic glycoproteins: the potential of the sialic acid biosynthetic pathway. *Journal of Pharmaceutical Sciences,* 98**,** 3499-508.

BOX, G. E. P. & DRAPER, N. R. 1959. A BASIS FOR THE SELECTION OF A RESPONSE-SURFACE DESIGN. *Journal of the American Statistical Association,* 54**,** 622-654.

BROWN, A. J., SWEENEY, B., MAINWARING, D. O. & JAMES, D. C. 2014. Synthetic promoters for CHO cell engineering. *Biotechnology and Bioengineering,* 111**,** 1638-47.

BROWN, M. E., RENNER, G., FIELD, R. P. & HASSELL, T. 1992. Process development for the production of recombinant antibodies using the glutamine synthetase (GS) system. *Cytotechnology,* 9**,** 231-6.

BROWNE, S. M. & AL-RUBEAI, M. 2007. Selection methods for high-producing mammalian cell lines. *Trends in Biotechnology,* 25**,** 425-432.

BROWNE, S. M. & AL-RUBEAI, M. 2009. Selection Methods for High-Producing Mammalian Cell Lines. *Cell Engineering, Vol 6: Cell Line Development,* 6**,** 127-151.

BUSTAMANTE, J., TOVAR, A., MONTERO, G. & BOVERIS, A. 1997. Early redox changes during rat thymocyte apoptosis. *Archives of Biochemistry and Biophysics,* 337**,** 121-128.

BUTLER, M. 2005. Animal cell cultures: recent achievements and perspectives in the production of biopharmaceuticals. *Applied microbiology and biotechnology,* 68**,** 283-291.

CANATELLA, P. J., KARR, J. F., PETROS, J. A. & PRAUSNITZ, M. R. 2001. Quantitative study of electroporation-mediated molecular uptake and cell viability. *Biophysical journal,* 80**,** 755-64.

CARNEIRO, M. O., RUSS, C., ROSS, M. G., GABRIEL, S. B., NUSBAUM, C. & DEPRISTO, M. A. 2012. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *Bmc Genomics,* 13.

CHANG, D. C. & REESE, T. S. 1990. Changes in Membrane-Structure Induced by Electroporation as Revealed by Rapid-Freezing Electron-Microscopy. *Biophysical journal,* 58**,** 1-12.

CHEN, C., SMYE, S. W., ROBINSON, M. P. & EVANS, J. A. 2006. Membrane electroporation theories: a review. *Medical & Biological Engineering & Computing,* 44**,** 5-14.

CHENUET, S., MARTINET, D., BESUCHET-SCHMUTZ, N., WICHT, M., JACCARD, N., BON, A. C., DEROUAZI, M., HACKER, D. L., BECKMANN, J. S. & WURM, F. M. 2008. Calcium phosphate transfection generates mammalian recombinant cell lines with higher specific productivity than polyfection. *Biotechnology and Bioengineering,* 101**,** 937-45.

COVIC, A. & KUHLMANN, M. K. 2007. Biosimilars: recent developments. *International Urology and Nephrology,* 39**,** 261-266.

DATTA, P., LINHARDT, R. J. & SHARFSTEIN, S. T. 2013. An 'omics approach towards CHO cell engineering. *Biotechnology and Bioengineering,* 110**,** 1255-71.

DAVIES, S. L., LOVELADY, C. S., GRAINGER, R. K., RACHER, A. J., YOUNG, R. J. & JAMES, D. C. 2013. Functional heterogeneity and heritability in CHO cell populations. *Biotechnology and Bioengineering,* 110**,** 260-274.

DEJONG, P. J., GROSOVSKY, A. J. & GLICKMAN, B. W. 1988. SPECTRUM OF SPONTANEOUS MUTATION AT THE APRT LOCUS OF CHINESE-HAMSTER OVARY CELLS - AN ANALYSIS AT THE DNA-SEQUENCE LEVEL. *Proceedings of the National Academy of Sciences of the United States of America,* 85**,** 3499-3503.

DEMAIN, A. L. & VAISHNAV, P. 2009. Production of recombinant proteins by microbes and higher organisms. *Biotechnology Advances,* 27**,** 297-306.

DENISSENKO, M. F., CHEN, J. X., TANG, M. S. & PFEIFER, G. P. 1997. Cytosine methylation determines hot spots of DNA damage in the human P53 gene. *Proceedings of the National Academy of Sciences of the United States of America,* 94**,** 3893-8.

DEROUAZI, M., GIRARD, P., VAN TILBORGH, F., IGLESIAS, K., MULLER, N., BERTSCHINGER, M. & WURM, F. M. 2004. Serum-free large-scale transient transfection of CHO cells. *Biotechnology and Bioengineering,* 87**,** 537-45.

DEROUAZI, M., MARTINET, D., BESUCHET SCHMUTZ, N., FLACTION, R., WICHT, M., BERTSCHINGER, M., HACKER, D. L., BECKMANN, J. S. & WURM, F. M. 2006. Genetic characterization of CHO production host DG44 and derivative recombinant cell lines. *Biochemical and Biophysical Research Communications,* 340**,** 1069-77.

DILERNIA, D. A., CHIEN, J.-T., MONACO, D. C., BROWN, M. P. S., ENDE, Z., DEYMIER, M. J., YUE, L., PAXINOS, E. E., ALLEN, S., TIRADO-RAMOS, A. & HUNTER, E. 2015. Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing. *Nucleic Acids Research,* 43.

DINNIS, D. M. & JAMES, D. C. 2005. Engineering mammalian cell factories for improved recombinant monoclonal antibody production: Lessons from nature? *Biotechnology and Bioengineering,* 91**,** 180-189.

DORAI, H., ELLIS, D., KEUNG, Y. S., CAMPBELL, M., ZHUANG, M., LIN, C. & BETENBAUGH, M. J. 2010. Combining high-throughput screening of caspase activity with anti-apoptosis genes for development of robust CHO production cell lines. *Biotechnology progress,* 26**,** 1367-81.

DOUGLAS, K. L. 2008. Toward development of artificial viruses for gene therapy: a comparative evaluation of viral and non-viral transfection. *Biotechnology progress,* 24**,** 871-83.

DUESBERG, P., RAUSCH, C., RASNICK, D. & HEHLMANN, R. 1998. Genetic instability of cancer cells is proportional to their degree of aneuploidy. *Proceedings of the National Academy of Sciences of the United States of America,* 95**,** 13692-13697.

ELLEGREN, H. 2004. Microsatellites: simple sequences with complex evolution. *Nature reviews. Genetics,* 5**,** 435-45.

ESCOFFRE, J. M., PORTET, T., WASUNGU, L., TEISSIE, J., DEAN, D. & ROLS, M. P. 2009. What is (Still not) Known of the Mechanism by Which Electroporation Mediates Gene Transfer and Expression in Cells and Tissues. *Molecular biotechnology,* 41**,** 286-295.

FAN, L., KADURA, I., KREBS, L. E., HATFIELD, C. C., SHAW, M. M. & FRYE, C. C. 2012. Improving the efficiency of CHO cell line generation using glutamine synthetase gene knockout cells. *Biotechnology and Bioengineering,* 109**,** 1007-15.

FERRER-MIRALLES, N., DOMINGO-ESPIN, J., CORCHERO, J. L., VAZQUEZ, E. & VILLAVERDE, A. 2009. Microbial factories for recombinant pharmaceuticals. *Microbial cell factories,* 8**,** 17.

FICHOT, E. B. & NORMAN, R. S. 2013. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome,* 1.

FISCHER, R., SCHILLBERG, S., HELLWIG, S., TWYMAN, R. M. & DROSSARD, J. 2012. GMP issues for recombinant plant-derived pharmaceutical proteins. *Biotechnology Advances,* 30**,** 434-9.

FRATANTONI, J. C., DZEKUNOV, S., SINGH, V. & LIU, L. N. 2003. A non-viral gene delivery system designed for clinical use. *Cytotherapy,* 5**,** 208-10.

FRATANTONI, J. C., DZEKUNOV, S., WANG, S. & LIU, L. N. 2004. A Scalable Cell-Loading System for Non-Viral Gene Delivery and other Applications. *Bioprocess. J.,* 3**,** 49-54.

GEHL, J. 2003. Electroporation: theory and methods, perspectives for drug delivery, gene therapy and research. *Acta physiologica Scandinavica,* 177**,** 437-47.

GEYER, P. K. 1997. The role of insulator elements in defining domains of gene expression. *Current opinion in genetics & development,* 7**,** 242-8.

GIDDINGS, G., ALLISON, G., BROOKS, D. & CARTER, A. 2000. Transgenic plants as factories for biopharmaceuticals. *Nature biotechnology,* 18**,** 1151-1155.

GOJOBORI, T., LI, W. H. & GRAUR, D. 1982. PATTERNS OF NUCLEOTIDE SUBSTITUTION IN PSEUDOGENES AND FUNCTIONAL GENES. *Journal of Molecular Evolution,* 18**,** 360-369.

GORDON, D. J., RESIO, B. & PELLMAN, D. 2012. Causes and consequences of aneuploidy in cancer. *Nature reviews. Genetics,* 13**,** 189-203.

GUPTA, P. K. 2008. Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology,* 26**,** 602-611.

HAMILTON, S. R. & GERNGROSS, T. U. 2007. Glycosylation engineering in yeast: the advent of fully humanized yeast. *Current opinion in biotechnology,* 18**,** 387-92.

HARRIS, R. J., MURNANE, A. A., UTTER, S. L., WAGNER, K. L., COX, E. T., POLASTRI, G. D., HELDER, J. C. & SLIWKOWSKI, M. B. 1993. ASSESSING GENETIC-HETEROGENEITY IN PRODUCTION CELL-LINES - DETECTION BY PEPTIDE-MAPPING OF A LOW-LEVEL TYR TO GLN SEQUENCE VARIANT IN A RECOMBINANT ANTIBODY. *Bio-Technology,* 11**,** 1293-1297.

HASTINGS, P. J., LUPSKI, J. R., ROSENBERG, S. M. & IRA, G. 2009. Mechanisms of change in gene copy number. *Nature reviews. Genetics,* 10**,** 551-64.

HAUSER, J., LEVINE, A. S. & DIXON, K. 1987. Unique pattern of point mutations arising after gene transfer into mammalian cells. *The EMBO journal,* 6**,** 63-7.

HELLER-HARRISON, R., CROWE, K., COOLEY, C., HONE, M., MCCARTHY, K. & LEONARD, M. 2009. Managing Cell Line Instability and Its Impact During Cell Line Development. *Biopharm International***,** 16-+.

HELLWIG, S., DROSSARD, J., TWYMAN, R. M. & FISCHER, R. 2004. Plant cell cultures for the production of recombinant proteins. *Nature biotechnology,* 22**,** 1415-22.

HINZ, J. M. & MEUTH, M. 1999. MSH3 deficiency is not sufficient for a mutator phenotype in Chinese hamster ovary cells. *Carcinogenesis,* 20**,** 215-20.

IIDA, S., MISAKA, H., INOUE, M., SHIBATA, M., NAKANO, R., YAMANE-OHNUKI, N., WAKITANI, M., YANO, K., SHITARA, K. & SATOH, M. 2006. Nonfucosylated therapeutic IgG1 antibody can evade the inhibitory effect of serum immunoglobulin G on antibody-dependent cellular cytotoxicity through its high binding to FcgammaRIIIa. *Clinical cancer research : an official journal of the American Association for Cancer Research,* 12**,** 2879-87.

INGMAN, M. & GYLLENSTEN, U. 2009. SNP frequency estimation using massively parallel sequencing of pooled DNA. *European Journal of Human Genetics,* 17**,** 383-386.

IOANNOU, Y. A. & CHEN, F. W. 1996. Quantitation of DNA fragmentation in apoptosis. *Nucleic Acids Research,* 24**,** 992-993.

JACKSON, S. P. 2002. Sensing and repairing DNA double-strand breaks. *Carcinogenesis,* 23**,** 687-96.

JAYAPAL, K. R., WLASCHIN, K. F., HU. W-S. & YAP, M. G. S. 2007. Recombinant protein therapeutics from CHO cells - 20 years and counting. *Cell Engineering Progress,* 103**,** 40-47.

JORDAN, C. A., NEUMANN, E. & SOWERS, A. E. 2013. *Electroporation and electrofusion in cell biology*, Springer Science & Business Media.

JORDAN, E., TEREFE, J. & UGOZZOLI, L. 2007. Optimization of electroporation conditions with the Gene Pulser MXcell™ electroporation system. *Bio-Rad Bulletin,* 5622.

JORDAN, E. T., COLLINS, M., TEREFE, J., UGOZZOLI, L. & RUBIO, T. 2008. Optimizing electroporation conditions in primary and other difficult-to-transfect cells. *Journal of biomolecular techniques : JBT,* 19**,** 328-34.

JUN, S. C., KIM, M. S., HONG, H. J. & LEE, G. M. 2006. Limitations to the development of humanized antibody producing Chinese hamster ovary cells using glutamine synthetase-mediated gene amplification. *Biotechnology progress,* 22**,** 770-80.

KELLEY, B. 2007. Very large scale monoclonal antibody purification: the case for conventional unit operations. *Biotechnology progress,* 23**,** 995-1008.

KHALIL, A. S. & COLLINS, J. J. 2010. Synthetic biology: applications come of age. *Nature reviews. Genetics,* 11**,** 367-79.

KILDEGAARD, H. F., BAYCIN-HIZAL, D., LEWIS, N. E. & BETENBAUGH, M. J. 2013. The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology. *Current opinion in biotechnology,* 24**,** 1102-7.

KIM, J. Y., KIM, Y. G. & LEE, G. M. 2012. CHO cells in biotechnology for production of recombinant proteins: current state and further potential. *Applied microbiology and biotechnology,* 93**,** 917-930.

KIM, M., O'CALLAGHAN, P. M., DROMS, K. A. & JAMES, D. C. 2011. A Mechanistic Understanding of Production Instability in CHO Cell Lines Expressing Recombinant Monoclonal Antibodies. *Biotechnology and Bioengineering,* 108**,** 2434-2446.

KIM, T. K. & EBERWINE, J. H. 2010. Mammalian cell transfection: the present and the future. *Analytical and bioanalytical chemistry,* 397**,** 3173-8.

KIMURA, M. 1955. Solution of a Process of Random Genetic Drift with a Continuous Model. *Proceedings of the National Academy of Sciences of the United States of America,* 41**,** 144-50.

KIMURA, M. 1979. Model of effectively neutral mutations in which selective constraint is incorporated. *Proceedings of the National Academy of Sciences of the United States of America,* 76**,** 3440-4.

KINDE, I., WU, J., PAPADOPOULOS, N., KINZLER, K. W. & VOGELSTEIN, B. 2011. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America,* 108**,** 9530-9535.

KOHLER, G. & MILSTEIN, C. 1975. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature,* 256**,** 495-497.

KORLACH, J. 2013. Understanding Accuracy in SMRT® Sequencing. http://www.pacb.com/wp-content/uploads/2015/09/Perspective_UnderstandingAccuracySMRTSequencing1.pdf.

KRETZMER, G. 2002. Industrial processes with animal cells. *Applied microbiology and biotechnology,* 59**,** 135-142.

KUNKEL, T. & ERIE, D. 2005. DNA mismatch repair. *Annual Review of Biochemistry,* 74**,** 681-710.

KURZAWSKI, G., SUCHY, J., DEBNIAK, T., KLADNY, J. & LUBINSKI, J. 2004. Importance of microsatellite instability (MSI) in colorectal cancer: MSI as a diagnostic tool. *Annals of Oncology,* 15**,** 283-284.

KWAKS, T. H., BARNETT, P., HEMRIKA, W., SIERSMA, T., SEWALT, R. G., SATIJN, D. P., BRONS, J. F., VAN BLOKLAND, R., KWAKMAN, P., KRUCKEBERG, A. L., KELDER, A. & OTTE, A. P. 2003. Identification of anti-repressor elements that confer high and stable protein production in mammalian cells. *Nature biotechnology,* 21**,** 553-8.

LAI, Y. & SUN, F. 2003. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular biology and evolution,* 20**,** 2123-31.

LATTENMAYER, C., LOESCHEL, M., STEINFELLNER, W., TRUMMER, E., MUELLER, D., SCHRIEBL, K., VORAUER-UHL, K., KATINGER, H. & KUNERT, R. 2006.

Identification of transgene integration loci of different highly expressing recombinant CHO cell lines by FISH. *Cytotechnology,* 51, 171-82.

LAUC, G., ESSAFI, A., HUFFMAN, J. E., HAYWARD, C., KNEZEVIC, A., KATTLA, J. J., POLASEK, O., GORNIK, O., VITART, V., ABRAHAMS, J. L., PUCIC, M., NOVOKMET, M., REDZIC, I., CAMPBELL, S., WILD, S. H., BOROVECKI, F., WANG, W., KOLCIC, I., ZGAGA, L., GYLLENSTEN, U., WILSON, J. F., WRIGHT, A. F., HASTIE, N. D., CAMPBELL, H., RUDD, P. M. & RUDAN, I. 2010. Genomics meets glycomics-the first GWAS study of human N-Glycome identifies HNF1alpha as a master regulator of plasma protein fucosylation. *PLoS genetics,* 6, e1001256.

LE, H., VISHWANATHAN, N., JACOB, N. M., GADGIL, M. & HU, W. S. 2015. Cell line development for biomanufacturing processes: recent advances and an outlook. *Biotechnology Letters,* 37, 1553-1564.

LEBKOWSKI, J. S., DUBRIDGE, R. B., ANTELL, E. A., GREISEN, K. S. & CALOS, M. P. 1984. Transfected DNA Is Mutated in Monkey, Mouse, and Human-Cells. *Molecular and cellular biology,* 4, 1951-1960.

LECHARDEUR, D., SOHN, K. J., HAARDT, M., JOSHI, P. B., MONCK, M., GRAHAM, R. W., BEATTY, B., SQUIRE, J., O'BRODOVICH, H. & LUKACS, G. L. 1999. Metabolic instability of plasmid DNA in the cytosol: a potential barrier to gene transfer. *Gene Therapy,* 6, 482-497.

LENGAUER, C., KINZLER, K. W. & VOGELSTEIN, B. 1998. Genetic instabilities in human cancers. *Nature,* 396, 643-9.

LEVENE, M. J., KORLACH, J., TURNER, S. W., FOQUET, M., CRAIGHEAD, H. G. & WEBB, W. W. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science,* 299, 682-686.

LEWIS, N. E., LIU, X., LI, Y., NAGARAJAN, H., YERGANIAN, G., O'BRIEN, E., BORDBAR, A., ROTH, A. M., ROSENBLOOM, J., BIAN, C., XIE, M., CHEN, W., LI, N., BAYCIN-HIZAL, D., LATIF, H., FORSTER, J., BETENBAUGH, M. J., FAMILI, I., XU, X., WANG, J. & PALSSON, B. O. 2013. Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome. *Nature Biotechnology,* 31, 759-+.

LI, F., VIJAYASANKARAN, N., SHEN, A. Y., KISS, R. & AMANULLAH, A. 2010. Cell culture processes for monoclonal antibody production. *mAbs,* 2, 466-79.

LIENERT, F., LOHMUELLER, J. J., GARG, A. & SILVER, P. A. 2014. Synthetic biology in mammalian cells: next generation research tools and therapeutics. *Nature reviews. Molecular cell biology,* 15**,** 95-107.

LIGON, B. L. 2004. Penicillin: its discovery and early development. *Seminars in pediatric infectious diseases,* 15**,** 52-7.

LIU, X., LIU, J., WILLIAMS WRIGHT, T., LEE, J., LIO, P., DONAHUE-HJELLE, L., RAVNIKAR, P. & FLORENCE WU, F. 2010. Isolation of Novel High-Osmolarity Resistant CHO DG44 Cells After Suspension of DNA Mismatch Repair. *Bioprocess International,* 8**,** 68-76.

LOBBAN, P. E. & KAISER, A. D. 1973. Enzymatic End-to-End Joining of DNA Molecules. *Journal of Molecular Biology,* 78**,** 453-&.

LONZA 2009. Optimized Protocol for Suspension CHO Clones - Lonza.

LONZA 2012. Guideline for Generation of Stable Cell Lines: Technical Reference Guide.

MACAULEY-PATRICK, S., FAZENDA, M. L., MCNEIL, B. & HARVEY, L. M. 2005. Heterologous protein production using the Pichia pastoris expression system. *Yeast,* 22**,** 249-70.

MADEIRA, C., RIBEIRO, S. C., TURK, M. Z. & CABRAL, J. M. S. 2010. Optimization of gene delivery to HEK293T cells by microporation using a central composite design methodology. *Biotechnology Letters,* 32**,** 1393-1399.

MAKRIDES, S. C. 1999. Components of vectors for gene transfer and expression in mammalian cells. *Protein expression and purification,* 17**,** 183-202.

MCCARTHY, A. 2010. Third Generation DNA Sequencing: Pacific Biosciences' Single Molecule Real Time Technology. *Chemistry & Biology,* 17**,** 675-676.

MEHIER-HUMBERT, S. & GUY, R. H. 2005. Physical methods for gene transfer: improving the kinetics of gene delivery into cells. *Advanced Drug Delivery Reviews,* 57**,** 733-53.

MELLSTEDT, H., NIEDERWIESER, D. & LUDWIG, H. 2008. The challenge of biosimilars. *Annals of Oncology,* 19**,** 411-419.

MILLER, J. H., LEBKOWSKI, J. S., GREISEN, K. S. & CALOS, M. P. 1984. Specificity of mutations induced in transfected DNA by mammalian cells. *The EMBO journal,* 3**,** 3117-21.

MITELMAN, F. 1995. *ISCN 1995: an international system for human cytogenetic nomenclature (1995): recommendations of the International Standing Committee on Human Cytogenetic Nomenclature, Memphis, Tennessee, USA, October 9-13, 1994*, Karger Medical and Scientific Publishers.

MITELMAN, F., JOHANSSON, B. & MERTENS, F. 2007. The impact of translocations and gene fusions on cancer causation. *Nature reviews. Cancer,* 7**,** 233-45.

MOHAN, C., KIM, Y. G., KOO, J. & LEE, G. M. 2008. Assessment of cell engineering strategies for improved therapeutic protein production in CHO cells. *Biotechnology Journal,* 3**,** 624-30.

MORAN, N. 2008. Fractured European market undermines biosimilar launches. *Nature biotechnology,* 26**,** 5-6.

MUTSKOV, V. & FELSENFELD, G. 2004. Silencing of transgene transcription precedes methylation of promoter DNA and histone H3 lysine 9. *The EMBO journal,* 23**,** 138-49.

NAGATA, S. 2000. Apoptotic DNA fragmentation. *Experimental Cell Research,* 256**,** 12-18.

NAKAMURA, T. & OMASA, T. 2015. Optimization of cell line development in the GS-CHO expression system using a high-throughput, single cell-based clone selection system. *Journal of bioscience and bioengineering,* 120**,** 323-9.

NEI, M. & GOJOBORI, T. 1986. SIMPLE METHODS FOR ESTIMATING THE NUMBERS OF SYNONYMOUS AND NONSYNONYMOUS NUCLEOTIDE SUBSTITUTIONS. *Molecular Biology and Evolution,* 3**,** 418-426.

NOH, S. M., SATHYAMURTHY, M. & LEE, G. M. 2013. Development of recombinant Chinese hamster ovary cell lines for therapeutic protein production. *Current Opinion in Chemical Engineering,* 2**,** 391-397.

O'CALLAGHAN, P. M. & JAMES, D. C. 2008. Systems biotechnology of mammalian cell factories. *Briefings in functional genomics & proteomics,* 7**,** 95-110.

OLSEN, H. B., LUDVIGSEN, S. & KAARSHOLM, N. C. 1996. Solution structure of an engineered insulin monomer at neutral pH. *Biochemistry,* 35**,** 8836-8845.

ORZALLI, M. H. & KNIPE, D. M. 2014. Cellular Sensing of Viral DNA and Viral Evasion Mechanisms. *Annual Review of Microbiology, Vol 68,* 68**,** 477-492.

PACIFIC-BIOSCIENCE 2010. Template Preparation and Sequencing Guide. *In:* BIOSCIENCE, P. (ed.).

PAGE, M. J. 1988. Expression of foreign genes in Mammalian cells. *Methods in molecular biology,* 4**,** 371-84.

PHAM, P. L., KAMEN, A. & DUROCHER, Y. 2006. Large-scale transfection of mammalian cells for the fast production of recombinant protein. *Molecular biotechnology,* 34**,** 225-37.

PINEDA, D., AMPURDANES, C., MEDINA, M. G., SERRATOSA, J., TUSELL, J. M., SAURA, J., PLANAS, A. M. & NAVARRO, P. 2012. Tissue plasminogen activator induces microglial inflammation via a noncatalytic molecular mechanism involving activation of mitogen-activated protein kinases and Akt signaling pathways and AnnexinA2 and Galectin-1 receptors. *Glia,* 60**,** 526-40.

PORTER, A. J., RACHER, A. J., PREZIOSI, R. & DICKSON, A. J. 2010. Strategies for selecting recombinant CHO cell lines for cGMP manufacturing: improving the efficiency of cell line generation. *Biotechnology progress,* 26**,** 1455-64.

PRENTICE, H. L., EHRENFELS, B. N. & SISK, W. P. 2007. Improving performance of mammalian cells in fed-batch processes through "bioreactor evolution". *Biotechnology progress,* 23**,** 458-64.

PUCIHAR, G., KRMELJ, J., REBERSEK, M., NAPOTNIK, T. B. & MIKLAVCIC, D. 2011. Equivalent Pulse Parameters for Electroporation. *Ieee Transactions on Biomedical Engineering,* 58**,** 3279-3288.

PURNICK, P. E. & WEISS, R. 2009. The second wave of synthetic biology: from modules to systems. *Nature reviews. Molecular cell biology,* 10**,** 410-22.

RAJU, T. S. 2003. Glycosylation variations with expression systems and their impact on biological activity of therapeutic immunoglobulins. *Bioprocess International***,** 44-53.

RAY, M. & MOHANDAS, T. 1976. PROPOSED BANDING NOMENCLATURE FOR CHINESE-HAMSTER CHROMOSOMES (CRICETULUS-GRISEUS). *Cytogenetics and Cell Genetics,* 16**,** 83-91.

REED, S. E., STALEY, E. M., MAYGINNES, J. P., PINTEL, D. J. & TULLIS, G. E. 2006. Transfection of mammalian cells using linear polyethylenimine is a simple and effective means of producing recombinant adeno-associated virus vectors. *Journal of virological methods,* 138**,** 85-98.

REHMAN, Z. U., ZUHORN, I. S. & HOEKSTRA, D. 2013. How cationic lipids transfer nucleic acids into cells and across cellular membranes: Recent advances. *Journal of Controlled Release,* 166**,** 46-56.

REN, D., ZHANG, J., PRITCHETT, R., LIU, H., KYAUK, J., LUO, J. & AMANULLAH, A. 2011. Detection and identification of a serine to arginine sequence variant in a therapeutic monoclonal antibody. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences,* 879**,** 2877-2884.

RHOADS, A. & AU, K. F. 2015. PacBio Sequencing and Its Applications. *Genomics, proteomics & bioinformatics*.

RICHARDS, E. J. & ELGIN, S. C. 2002. Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. *Cell,* 108**,** 489-500.

RITA COSTA, A., ELISA RODRIGUES, M., HENRIQUES, M., AZEREDO, J. & OLIVEIRA, R. 2010. Guidelines to cell engineering for monoclonal antibody production. *European journal of pharmaceutics and biopharmaceutics : official journal of Arbeitsgemeinschaft fur Pharmazeutische Verfahrenstechnik e.V,* 74**,** 127-38.

ROBERTS, R. J., CARNEIRO, M. O. & SCHATZ, M. C. 2013. The advantages of SMRT sequencing. *Genome Biology,* 14.

SCHMIDT, H. M., ZUMBANSEN, M., WITTIG, R., BLAICH, S., BROWN, L., LYER, S., POUSTKA, A., MOLLENHAUER, J. & NIX, M. 2004. Use of Nucleofector® Technology to Establish Stably Expressing Cell Lines. Koln: amaxa biosystems.

SCIENTIFIC, Thermo Fisher *PCR Fidelity Calculator* [Online]. Available: https://www.thermofisher.com/uk/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/pcr-fidelity-calculator.html [Accessed].

SHACKLETON, M., QUINTANA, E., FEARON, E. R. & MORRISON, S. J. 2009. Heterogeneity in cancer: cancer stem cells versus clonal evolution. *Cell,* 138**,** 822-9.

SHARP, J. M. & DORAN, P. M. 2001. Characterization of monoclonal antibody fragments produced by plant cells. *Biotechnology and Bioengineering,* 73**,** 338-46.

SHIELDS, R. L., LAI, J., KECK, R., O'CONNELL, L. Y., HONG, K., MENG, Y. G., WEIKERT, S. H. & PRESTA, L. G. 2002. Lack of fucose on human IgG1 N-linked oligosaccharide improves binding to human Fcgamma RIII and antibody-dependent cellular toxicity. *The Journal of biological chemistry,* 277**,** 26733-40.

SHIMOKAWA, T., OKUMURA, K. & RA, C. 2000. DNA induces apoptosis in electroporated human promonocytic cell line U937. *Biochemical and Biophysical Research Communications,* 270**,** 94-99.

SHUKLA, A. A., HUBBARD, B., TRESSEL, T., GUHAN, S. & LOW, D. 2007. Downstream processing of monoclonal antibodies--application of platform approaches. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences,* 848**,** 28-39.

SHUKLA, A. A. & THOMMES, J. 2010. Recent advances in large-scale production of monoclonal antibodies and related proteins. *Trends in Biotechnology,* 28**,** 253-61.

SINACORE, M. S., DRAPEAU, D. & ADAMSON, S. R. 2000. Adaptation of mammalian cells to growth in serum-free media. *Molecular biotechnology,* 15**,** 249-57.

SLATER, A. F. G., STEFAN, C., NOBEL, I., VANDENDOBBELSTEEN, D. J. & ORRENIUS, S. 1996. Intracellular redox changes during apoptosis (vol 3, pg 57, 1996). *Cell Death and Differentiation,* 3**,** 446-446.

SPASSOVA, M., TSONEVA, I., PETROV, A. G., PETKOVA, J. I. & NEUMANN, E. 1994. Dip Patch-Clamp Currents Suggest Electrodiffusive Transport of the Polyelectrolyte DNA through Lipid Bilayers. *Biophysical Chemistry,* 52**,** 267-274.

SPENCER, D. H., TYAGI, M., VALLANIA, F., BREDEMEYER, A. J., PFEIFER, J. D., MITRA, R. D. & DUNCAVAGE, E. J. 2014. Performance of Common Analysis Methods for Detecting Low-Frequency Single Nucleotide Variants in Targeted Next-Generation Sequence Data. *Journal of Molecular Diagnostics,* 16**,** 75-88.

STEGER, K., BRADY, J., WANG, W., DUSKIN, M., DONATO, K. & PESHWA, M. 2015. CHO-S antibody titers >1 gram/liter using flow electroporation-mediated transient gene expression followed by rapid migration to high-yield stable cell lines. *Journal of biomolecular screening,* 20**,** 545-51.

SUKHAREV, S. I., KLENCHIN, V. A., SEROV, S. M., CHERNOMORDIK, L. V. & CHIZMADZHEV, Y. A. 1992. Electroporation and Electrophoretic DNA Transfer into Cells - the Effect of DNA Interaction with Electropores. *Biophysical journal,* 63**,** 1320-1327.

TAIT, A. S., BROWN, C. J., GALBRAITH, D. J., HINES, M. J., HOARE, M., BIRCH, J. R. & JAMES, D. C. 2004. Transient production of recombinant proteins by Chinese hamster ovary cells using polyethyleneimine/DNA complexes in combination with microtubule disrupting anti-mitotic agents. *Biotechnology and Bioengineering,* 88**,** 707-21.

TANGE, T. O., NOTT, A. & MOORE, M. J. 2004. The ever-increasing complexities of the exon junction complex. *Current opinion in cell biology,* 16**,** 279-84.

TEREFE, J., PINEDA, M., JORDAN, E., COLLINS, M., UGOZZOLI, L. & RUBIO, T. 2008. Transfection of Mammalian Cells Using Preset Protocols on the Gene Pulser MXcellTM Electroporation System. Bio-Rad Laboratories, Inc.

THOMPSON, B. C., SEGARRA, C. R. J., MOZLEY, O. L., DARAMOLA, O., FIELD, R., LEVISON, P. R. & JAMES, D. C. 2012. Cell line specific control of polyethylenimine-mediated transient transfection optimized with "Design of experiments" methodology. *Biotechnology progress,* 28**,** 179-187.

THOMPSON, S. L. & COMPTON, D. A. 2011. Chromosomes and cancer cells. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology,* 19**,** 433-44.

TRAVERS, K. J., CHIN, C. S., RANK, D. R., EID, J. S. & TURNER, S. W. 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic acids research,* 38.

VALDERRAMA-RINCON, J. D., FISHER, A. C., MERRITT, J. H., FAN, Y. Y., READING, C. A., CHHIBA, K., HEISS, C., AZADI, P., AEBI, M. & DELISA, M. P. 2012. An engineered eukaryotic protein glycosylation pathway in Escherichia coli. *Nature chemical biology,* 8**,** 434-6.

VAN BERKEL, P. H. C., GERRITSEN, J., PERDOK, G., VALBJORN, J., VINK, T., VAN DE WINKEL, J. G. J. & PARREN, P. W. H. I. 2009. N-Linked Glycosylation is an Important Parameter for Optimal Selection of Cell Lines Producing Biopharmaceutical Human IgG. *Biotechnology Progress,* 25**,** 244-251.

VAN STEENSEL, B. 2011. Chromatin: constructing the big picture. *The EMBO journal,* 30**,** 1885-95.

VICTORIA, J. G., WANG, C., JONES, M. S., JAING, C., MCLOUGHLIN, K., GARDNER, S. & DELWART, E. L. 2010. Viral Nucleic Acids in Live-Attenuated Vaccines: Detection of Minority Variants and an Adventitious Virus. *Journal of Virology,* 84**,** 6033-6040.

WALSH, G. 2000. Biopharmaceutical benchmarks. *Nature biotechnology,* 18**,** 831-833.

WALSH, G. 2002. Biopharmaceuticals and biotechnology medicines: an issue of nomenclature. *European Journal of Pharmaceutical Sciences,* 15**,** 135-138.

WALSH, G. 2005. Biopharmaceuticals: recent approvals and likely directions. *Trends in Biotechnology,* 23**,** 553-558.

WALSH, G. 2006. Biopharmaceutical benchmarks 2006. *Nature biotechnology,* 24**,** 769-U5.

WALSH, G. 2010. Biopharmaceutical benchmarks 2010. *Nature biotechnology,* 28**,** 917-924.

WALSH, G. 2014. Biopharmaceutical benchmarks 2014. *Nature biotechnology,* 32**,** 992-1000.

WEN, D., VECCHI, M. M., GU, S., SU, L., DOLNIKOVA, J., HUANG, Y.-M., FOLEY, S. F., GARBER, E., PEDERSON, N. & MEIER, W. 2009. Discovery and Investigation of Misincorporation of Serine at Asparagine Positions in Recombinant Proteins Expressed in Chinese Hamster Ovary Cells. *Journal of Biological Chemistry,* 284**,** 32686-32694.

WESTERHOFF, H. V. & PALSSON, B. O. 2004. The evolution of molecular biology into systems biology. *Nature biotechnology,* 22**,** 1249-52.

WINTERBOURNE, D. J., THOMAS, S., HERMONTAYLOR, J., HUSSAIN, I. & JOHNSTONE, A. P. 1988. Electric Shock-Mediated Transfection of Cells - Characterization and Optimization of Electrical Parameters. *Biochemical Journal,* 251**,** 427-434.

WOLFFE, A. P. & MATZKE, M. A. 1999. Epigenetics: regulation through repression. *Science,* 286**,** 481-6.

WONG, A. W., BAGINSKI, T. K. & REILLY, D. E. 2010. Enhancement of DNA uptake in FUT8-deleted CHO cells for transient production of afucosylated antibodies. *Biotechnology and Bioengineering,* 106**,** 751-63.

WURM, F. 2013. CHO Quasispecies - Implications for Manufacturing Processes. *Processes,* 1**,** 296-311.

WURM, F. M. 2004. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nature biotechnology,* 22**,** 1393-8.

WURM, F. M. & HACKER, D. 2011. First CHO genome. *Nature biotechnology,* 29**,** 718-20.

YANG, Y., MARIATI, CHUSAINOW, J. & YAP, M. G. 2010. DNA methylation contributes to loss in productivity of monoclonal antibody-producing CHO cell lines. *Journal of biotechnology,* 147**,** 180-5.

YOSHIKAWA, T., NAKANISHI, F., OGURA, Y., OI, D., OMASA, T., KATAKURA, Y., KISHIMOTO, M. & SUGA, K. 2000. Amplified gene location in chromosomal DNA affected recombinant protein production and stability of amplified genes. *Biotechnology progress,* 16**,** 710-5.

YU, M., SELVARAJ, S. K., LIANG-CHU, M. M. Y., AGHAJANI, S., BUSSE, M., YUAN, J., LEE, G., PEALE, F., KLIJN, C., BOURGON, R., KAMINKER, J. S. & NEVE, R. M. 2015. A resource for cell line authentication, annotation and quality control. *Nature,* 520**,** 307-+.

YU, X. C., BORISOV, O. V., ALVAREZ, M., MICHELS, D. A., WANG, Y. J. & LING, V. 2009. Identification of Codon-Specific Serine to Asparagine Mistranslation in Recombinant Monoclonal Antibodies by High-Resolution Mass Spectrometry. *Analytical Chemistry,* 81**,** 9282-9290.

ZECK, A., REGULA, J. T., LARRAILLET, V., MAUTZ, B., POPP, O., GOEPFERT, U., WIEGESHOFF, F., VOLLERTSEN, U. E. E., GORR, I. H., KOLL, H. & PAPADIMITRIOU, A. 2012. Low Level Sequence Variant Analysis of Recombinant Proteins: An Optimized Approach. *Plos One,* 7.

ZHANG, S., BARTKOWIAK, L., NABISWA, B., MISHRA, P., FANN, J., OUELLETTE, D., CORREIA, I., REGIER, D. & LIU, J. 2015. Identifying low-level sequence variants via next generation sequencing to aid stable CHO cell line screening. *Biotechnology Progress,* 31**,** 1077-1085.

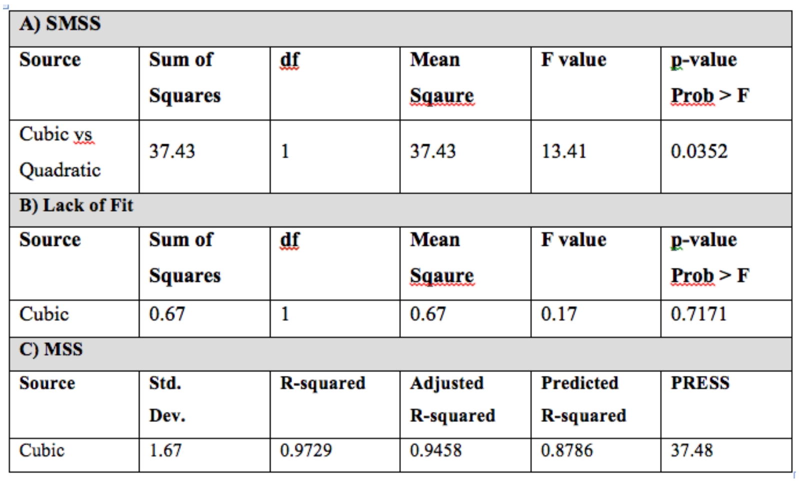
ZHANG, S., LIU, W., HE, P., GONG, F. & YANG, D. 2006. Establishment of stable high expression cell line with green fluorescent protein and resistance genes. *Journal of Huazhong University of Science and Technology. Medical sciences = Hua zhong ke ji da xue xue bao. Yi xue Ying De wen ban = Huazhong keji daxue xuebao. Yixue Yingdewen ban,* 26**,** 298-300.

ZHOU, H., LIU, Z. G., SUN, Z. W., HUANG, Y. & YU, W. Y. 2010. Generation of stable cell lines by site-specific integration of transgenes into engineered Chinese hamster ovary strains using an FLP-FRT system. *Journal of biotechnology,* 147**,** 122-9.

ZHU, J. 2012. Mammalian cell protein expression for biopharmaceutical production. *Biotechnology Advances,* 30**,** 1158-70.

This page is intentionally left blank.

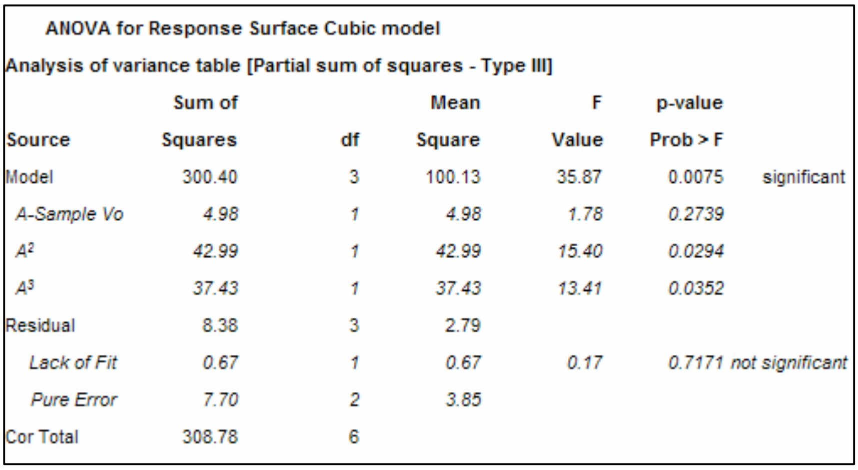
Appendix



A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Cubic vs Quadratic	37.43	1	37.43	13.41	0.0352
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Cubic	0.67	1	0.67	0.17	0.7171
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Cubic	1.67	0.9729	0.9458	0.8786	37.48

Table A1: Sample Volume Transfection Efficiency Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS.



ANOVA for Response Surface Cubic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	300.40	3	100.13	35.87	0.0075	significant
A-Sample Vo	*4.98*	*1*	*4.98*	*1.78*	*0.2739*	
A^2	*42.99*	*1*	*42.99*	*15.40*	*0.0294*	
A^3	*37.43*	*1*	*37.43*	*13.41*	*0.0352*	
Residual	8.38	3	2.79			
Lack of Fit	*0.67*	*1*	*0.67*	*0.17*	*0.7171*	*not significant*
Pure Error	*7.70*	*2*	*3.85*			
Cor Total	308.78	6				

Table A2: Sample volume Transfection Efficiency ANOVA table

Table of ANOVA output statistical terms and values.

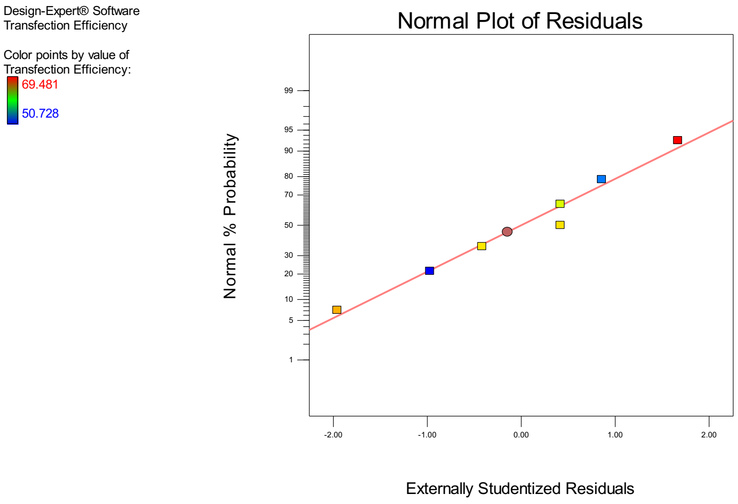
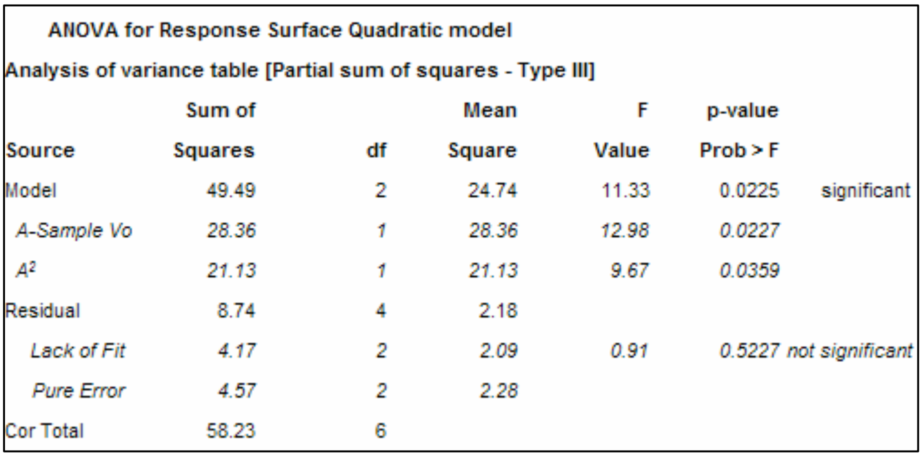


Figure A1: Sample Volume Transfection Efficiency Normal Plot of Residuals

A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Quadratic vs Linear	21.13	1	21.13	9.67	0.0359
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Quadratic	4.17	2	2.09	0.91	0.5227
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Quadratic	1.48	0.8499	0.7749	0.5513	26.13

Table A3: Sample Volume Cell Viability Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

	Sum of		Mean	F	p-value	
Source	Squares	df	Square	Value	Prob > F	
Model	49.49	2	24.74	11.33	0.0225	significant
A-Sample Vo	*28.36*	*1*	*28.36*	*12.98*	*0.0227*	
A^2	*21.13*	*1*	*21.13*	*9.67*	*0.0359*	
Residual	8.74	4	2.18			
Lack of Fit	*4.17*	*2*	*2.09*	*0.91*	*0.5227*	*not significant*
Pure Error	*4.57*	*2*	*2.28*			
Cor Total	58.23	6				

Table A4: Sample volume Cell Viability ANOVA table

Table of ANOVA output statistical terms and values.

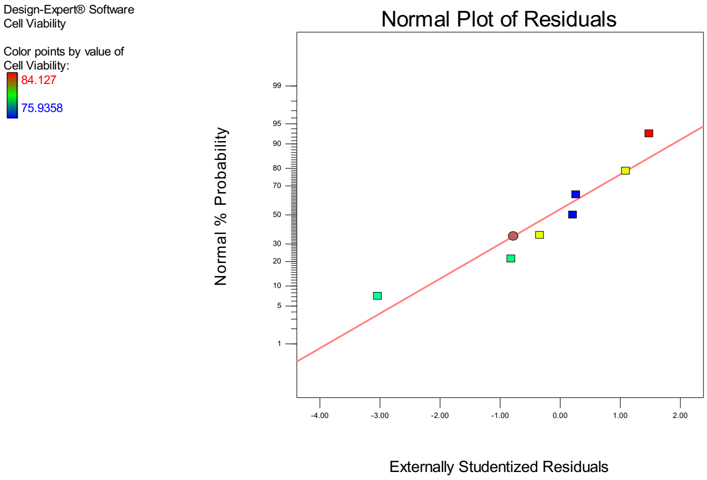
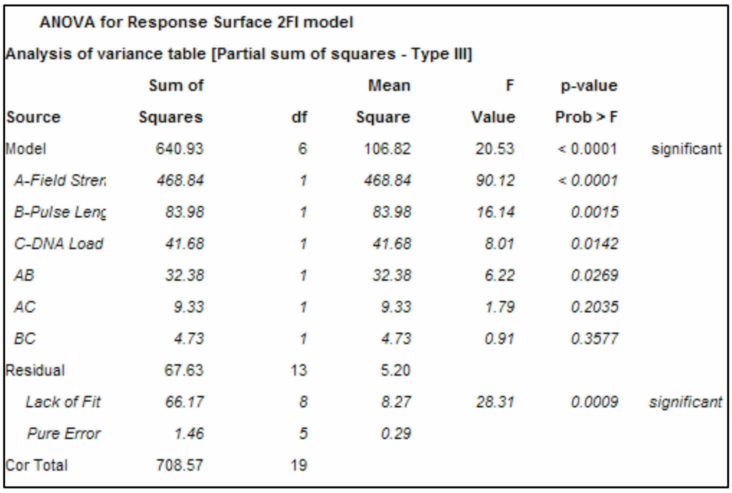


Figure A2: Sample Volume Cell Viability Normal Plot of Residuals

A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
2FI vs Linear	46.44	3	15.48	2.98	0.0707
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
2FI	66.17	8	8.27	28.31	0.0009
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
2FI	2.28	0.9045	0.8605	0.6307	261.70

Table A5: Exponential Decay: Wide – Transfection Efficiency Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface 2FI model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	640.93	6	106.82	20.53	< 0.0001	significant
A-Field Stren	*468.84*	*1*	*468.84*	*90.12*	*< 0.0001*	
B-Pulse Leng	*83.98*	*1*	*83.98*	*16.14*	*0.0015*	
C-DNA Load	*41.68*	*1*	*41.68*	*8.01*	*0.0142*	
AB	*32.38*	*1*	*32.38*	*6.22*	*0.0269*	
AC	*9.33*	*1*	*9.33*	*1.79*	*0.2035*	
BC	*4.73*	*1*	*4.73*	*0.91*	*0.3577*	
Residual	67.63	13	5.20			
Lack of Fit	*66.17*	*8*	*8.27*	*28.31*	*0.0009*	*significant*
Pure Error	*1.46*	*5*	*0.29*			
Cor Total	708.57	19				

Table A6: Exponential Decay: Wide – Transfection Efficiency ANOVA table

Table of ANOVA output statistical terms and values.

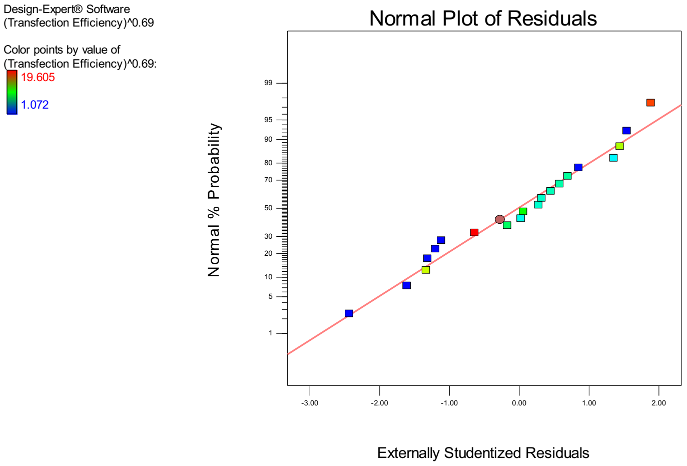
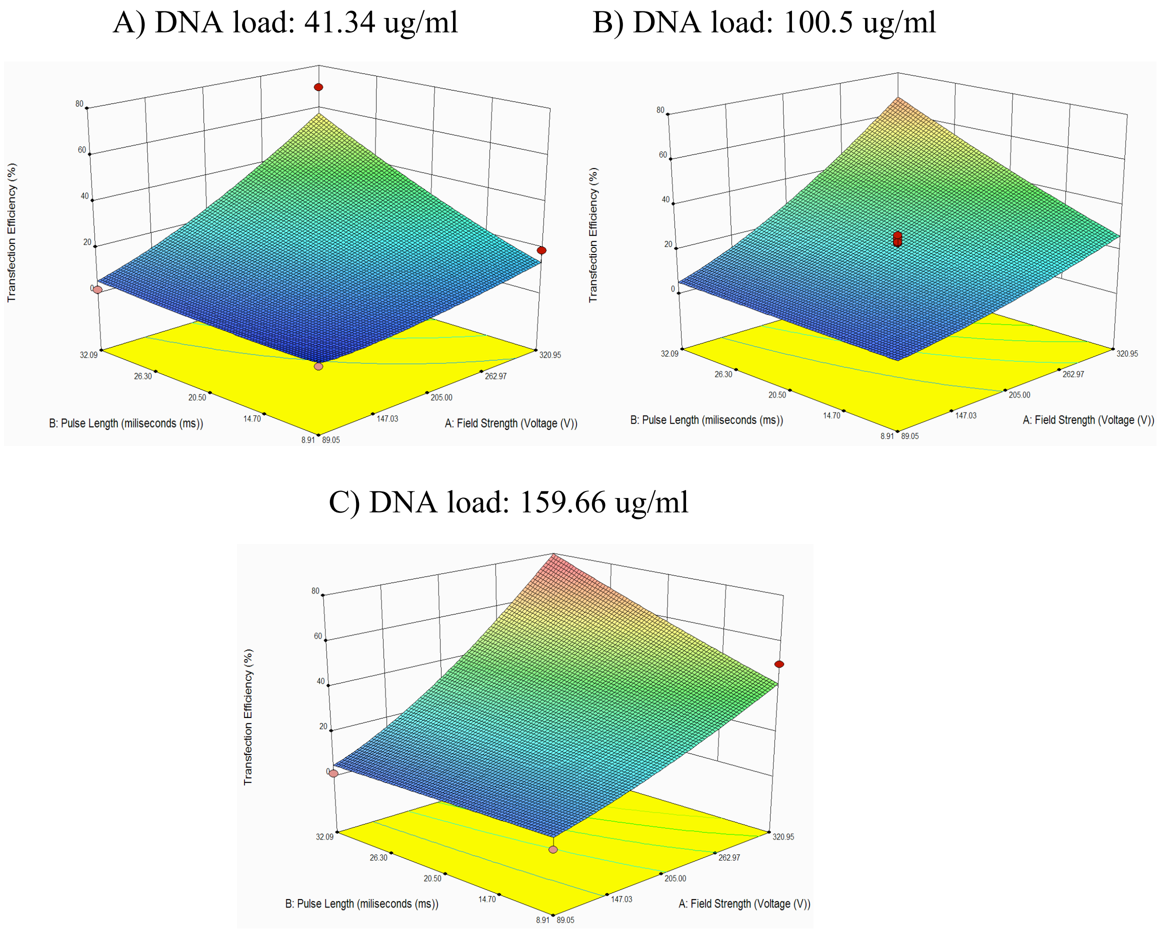


Figure A3: Exponential decay: Wide – Transfection Efficiency Normal Plot of Residuals

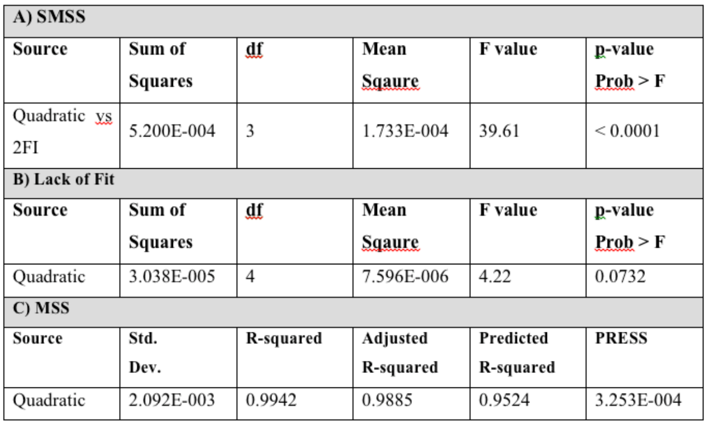


A) DNA load: 41.34 ug/ml

B) DNA load: 100.5 ug/ml

C) DNA load: 159.66 ug/ml

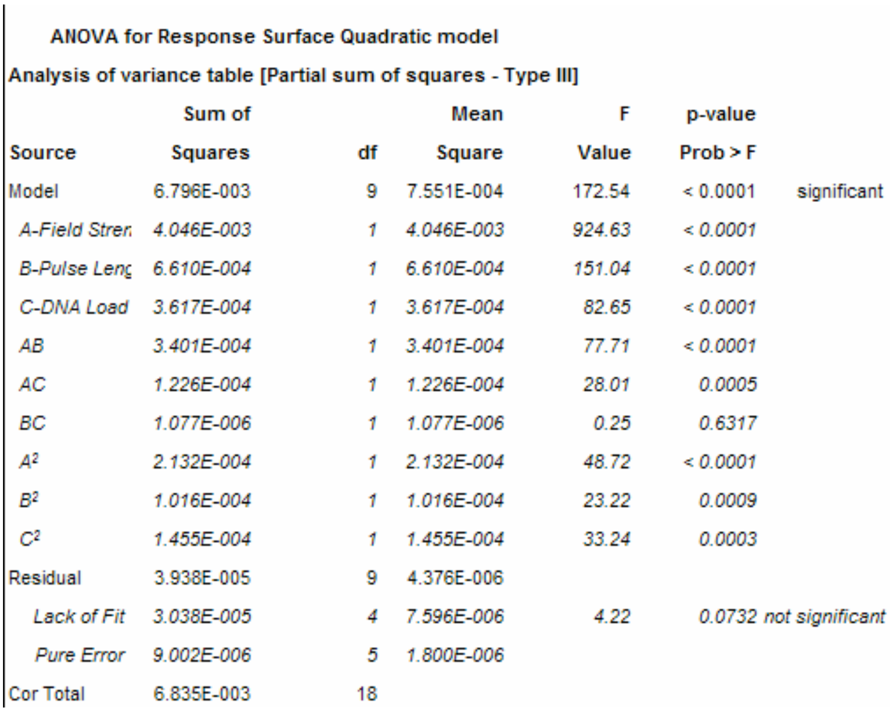
Figure A4. Exponential Decay: Wide – Transfection Efficiency Response Surface
Response surfaces of the transfection efficiency response to changes in field strength and pulse length at different levels of DNA load: Low Factorial (A), Center point (B) and Upper Factorial (C).



A) SMSS					
Source	Sum of Squares	df	Mean Sqaure	F value	p-value Prob > F
Quadratic vs 2FI	5.200E-004	3	1.733E-004	39.61	< 0.0001
B) Lack of Fit					
Source	Sum of Squares	df	Mean Sqaure	F value	p-value Prob > F
Quadratic	3.038E-005	4	7.596E-006	4.22	0.0732
C) MSS					
Source	Std. Dev.	R-squared	Adjusted R-squared	Predicted R-squared	PRESS
Quadratic	2.092E-003	0.9942	0.9885	0.9524	3.253E-004

Table A7: Exponential Decay: Wide – Median Fluorescence Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	6.796E-003	9	7.551E-004	172.54	< 0.0001	significant
A-Field Stren	*4.046E-003*	*1*	*4.046E-003*	*924.63*	*< 0.0001*	
B-Pulse Leng	*6.610E-004*	*1*	*6.610E-004*	*151.04*	*< 0.0001*	
C-DNA Load	*3.617E-004*	*1*	*3.617E-004*	*82.65*	*< 0.0001*	
AB	*3.401E-004*	*1*	*3.401E-004*	*77.71*	*< 0.0001*	
AC	*1.226E-004*	*1*	*1.226E-004*	*28.01*	*0.0005*	
BC	*1.077E-006*	*1*	*1.077E-006*	*0.25*	*0.6317*	
A^2	*2.132E-004*	*1*	*2.132E-004*	*48.72*	*< 0.0001*	
B^2	*1.016E-004*	*1*	*1.016E-004*	*23.22*	*0.0009*	
C^2	*1.455E-004*	*1*	*1.455E-004*	*33.24*	*0.0003*	
Residual	3.938E-005	9	4.376E-006			
Lack of Fit	*3.038E-005*	*4*	*7.596E-006*	*4.22*	*0.0732*	*not significant*
Pure Error	*9.002E-006*	*5*	*1.800E-006*			
Cor Total	6.835E-003	18				

Table A8: Exponential Decay: Wide – Median Fluorescence ANOVA table

Table of ANOVA output statistical terms and values.

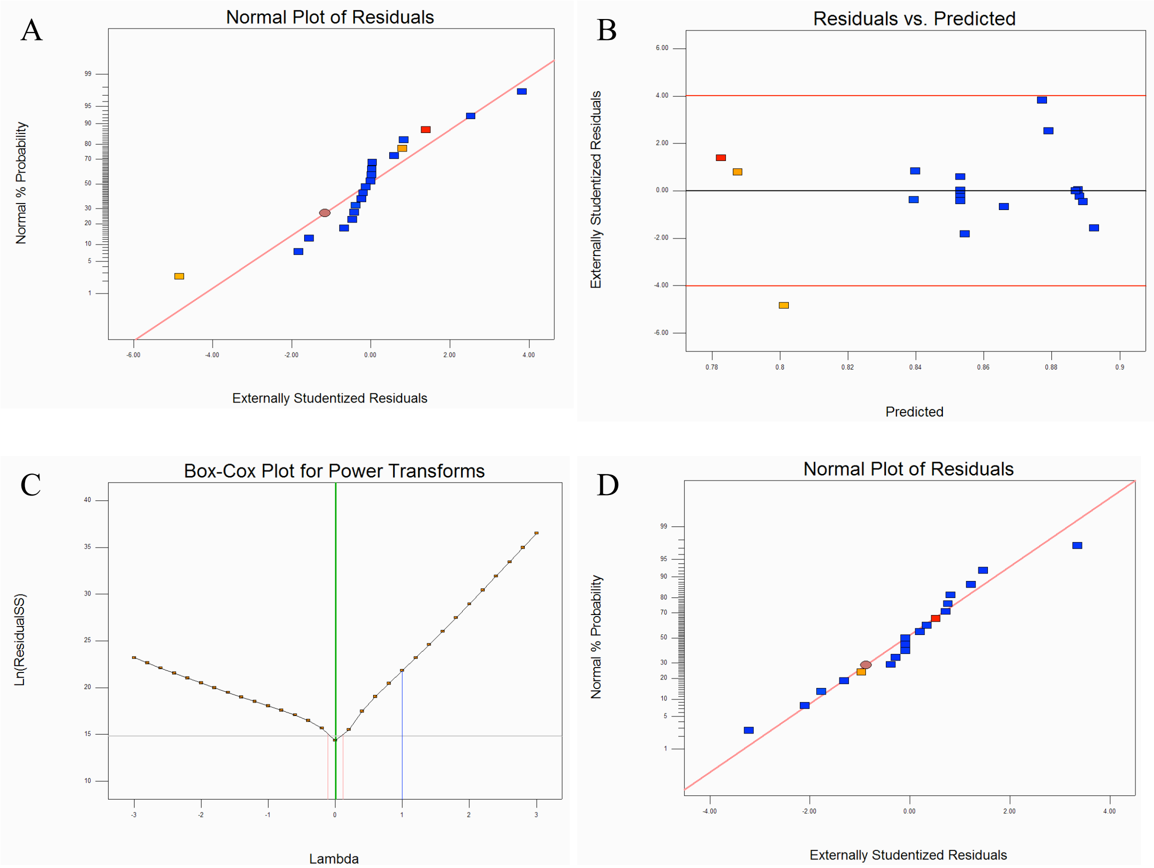
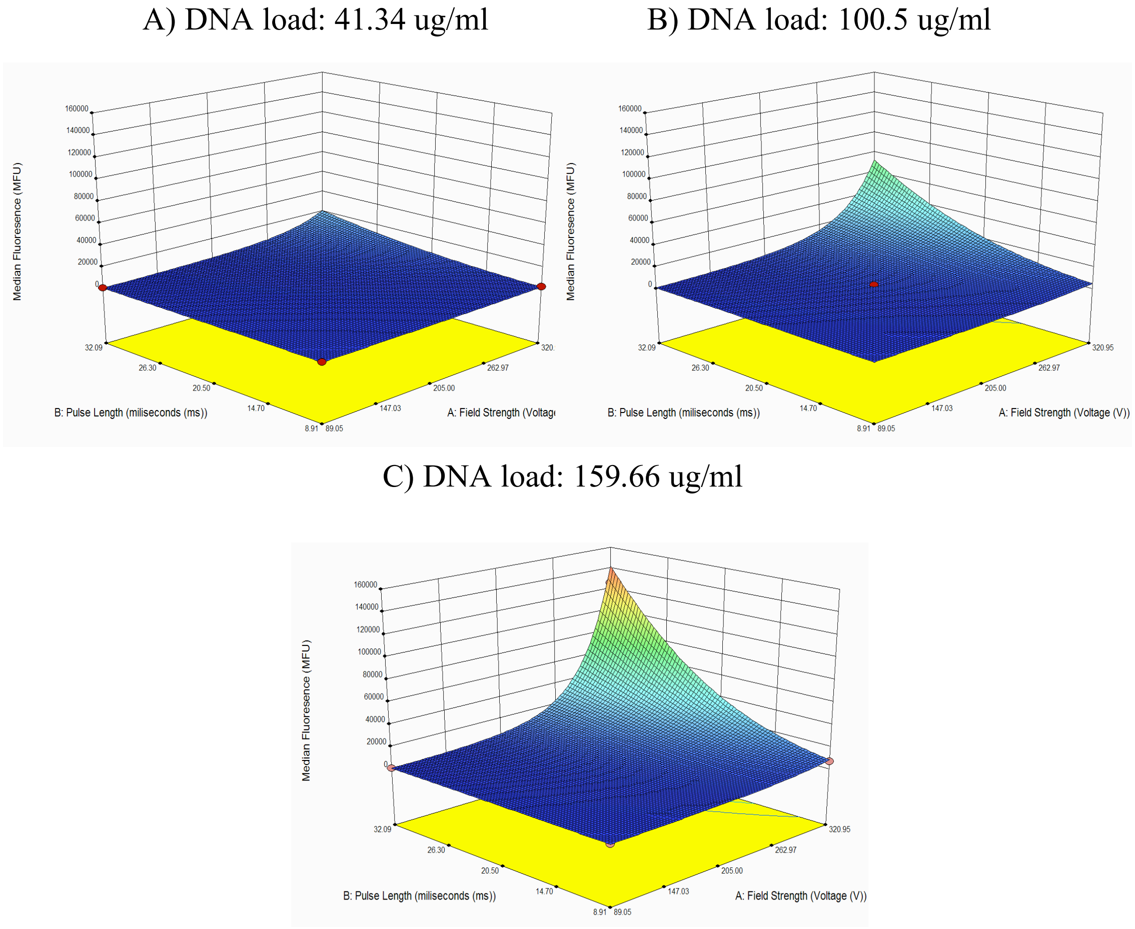


Figure A5. Exponential Decay: Wide – Median Fluorescence Data Manipulation
The figure shows the identification of non-normality (A) and the outlier responsible (B) as highlighted by a data point falling outside of a threshold level difference (red line). The Box-Cox plot (C) highlights a recommended transformation (green line). After ignoring the outlier and transformation the data residuals are normally distributed (D).



A) DNA load: 41.34 ug/ml

B) DNA load: 100.5 ug/ml

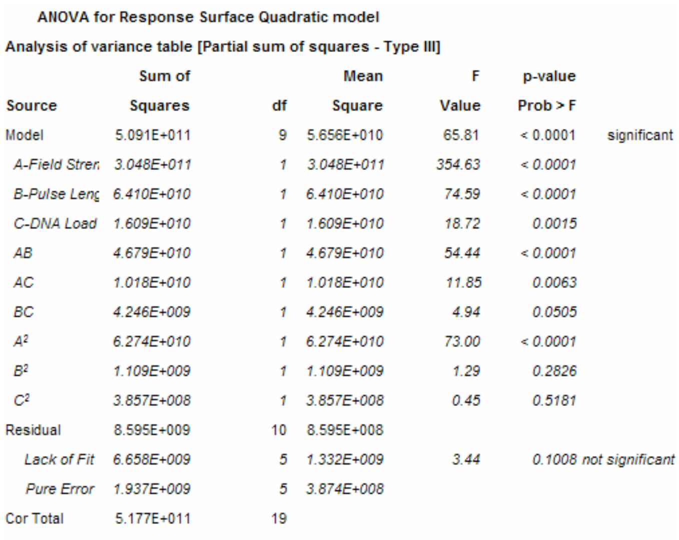
C) DNA load: 159.66 ug/ml

Figure A6. Exponential Decay: Wide – Median Fluorescence Response Surface Response surfaces of the median fluorescence response to changes in field strength and pulse length at different levels of DNA load: Low Factorial (A), Center point (B) and Upper Factorial (C).

A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Quadratic vs 2FI	6.285E+010	3	2.095E+010	24.38	< 0.0001
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Quadratic	6.658E+009	5	1.332E+009	3.44	0.1008
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Quadratic	29316.74	0.9834	0.9685	0.8843	5.989E+010

Table A9: Exponential Decay: Wide – Cell Viability Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	5.091E+011	9	5.656E+010	65.81	< 0.0001	significant
A-Field Stren	*3.048E+011*	*1*	*3.048E+011*	*354.63*	*< 0.0001*	
B-Pulse Leng	*6.410E+010*	*1*	*6.410E+010*	*74.59*	*< 0.0001*	
C-DNA Load	*1.609E+010*	*1*	*1.609E+010*	*18.72*	*0.0015*	
AB	*4.679E+010*	*1*	*4.679E+010*	*54.44*	*< 0.0001*	
AC	*1.018E+010*	*1*	*1.018E+010*	*11.85*	*0.0063*	
BC	*4.246E+009*	*1*	*4.246E+009*	*4.94*	*0.0505*	
A^2	*6.274E+010*	*1*	*6.274E+010*	*73.00*	*< 0.0001*	
B^2	*1.109E+009*	*1*	*1.109E+009*	*1.29*	*0.2826*	
C^2	*3.857E+008*	*1*	*3.857E+008*	*0.45*	*0.5181*	
Residual	8.595E+009	10	8.595E+008			
Lack of Fit	*6.658E+009*	*5*	*1.332E+009*	*3.44*	*0.1008*	*not significant*
Pure Error	*1.937E+009*	*5*	*3.874E+008*			
Cor Total	5.177E+011	19				

Table A10: Exponential Decay: Wide – Cell Viability ANOVA table

Table of ANOVA output statistical terms and values.

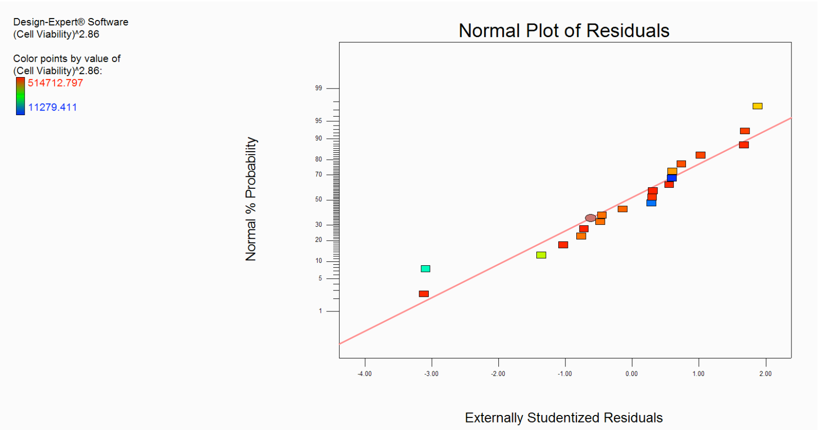
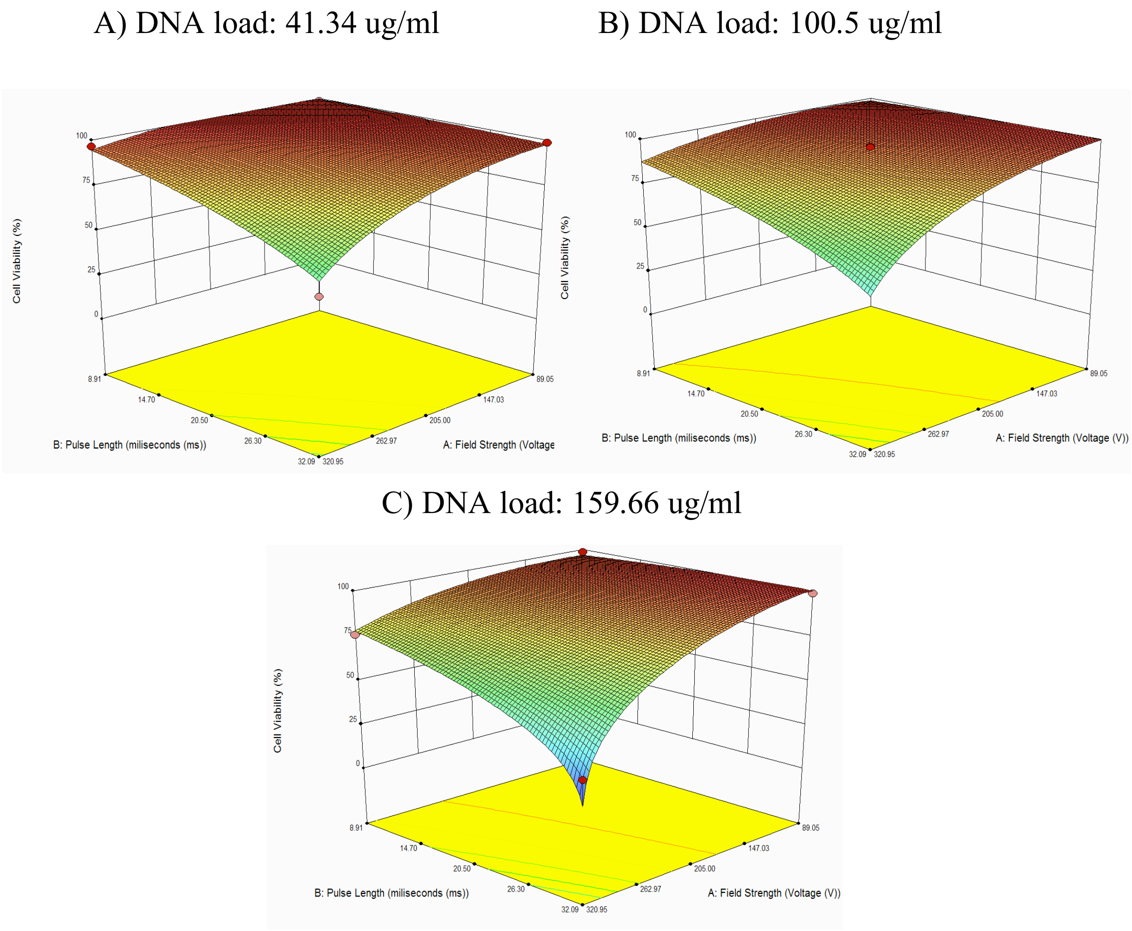


Figure A7: Exponential decay: Wide – Cell Viability Normal Plot of Residuals



A) DNA load: 41.34 ug/ml

B) DNA load: 100.5 ug/ml

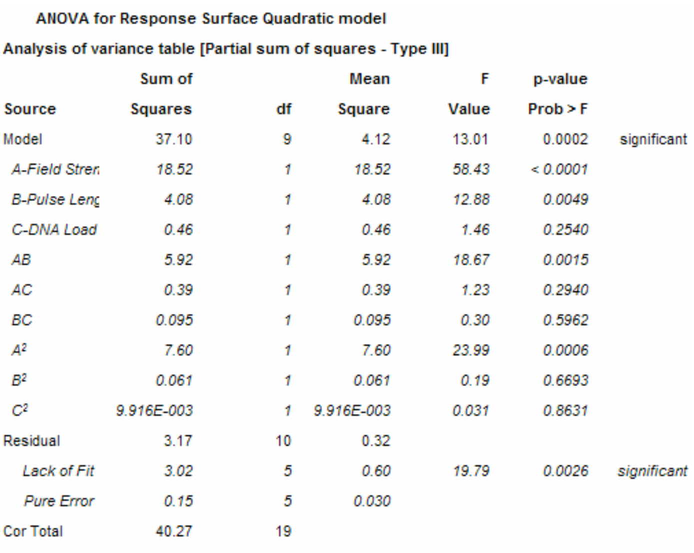
C) DNA load: 159.66 ug/ml

Figure A8. Exponential Decay: Wide – Cell Viability Response Surface
Response surfaces of the cell viability response to changes in field strength and pulse length at different levels of DNA load: Low Factorial (A), Center point (B) and Upper Factorial (C).

A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Quadratic vs 2FI	7.63	3	2.54	8.03	0.0051
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Quadratic	3.02	5	0.6	19.79	0.0026
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Quadratic	0.56	0.9213	0.8505	0.4142	23.59

Table A11: Exponential Decay: Wide – ACD Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

	Sum of		Mean	F	p-value	
Source	Squares	df	Square	Value	Prob > F	
Model	37.10	9	4.12	13.01	0.0002	significant
A-Field Stren	*18.52*	*1*	*18.52*	*58.43*	*< 0.0001*	
B-Pulse Leng	*4.08*	*1*	*4.08*	*12.88*	*0.0049*	
C-DNA Load	*0.46*	*1*	*0.46*	*1.46*	*0.2540*	
AB	*5.92*	*1*	*5.92*	*18.67*	*0.0015*	
AC	*0.39*	*1*	*0.39*	*1.23*	*0.2940*	
BC	*0.095*	*1*	*0.095*	*0.30*	*0.5962*	
A^2	*7.60*	*1*	*7.60*	*23.99*	*0.0006*	
B^2	*0.061*	*1*	*0.061*	*0.19*	*0.6693*	
C^2	*9.916E-003*	*1*	*9.916E-003*	*0.031*	*0.8631*	
Residual	3.17	10	0.32			
Lack of Fit	*3.02*	*5*	*0.60*	*19.79*	*0.0026*	*significant*
Pure Error	*0.15*	*5*	*0.030*			
Cor Total	40.27	19				

Table A12: Exponential Decay: Wide – ACD ANOVA table

Table of ANOVA output statistical terms and values.

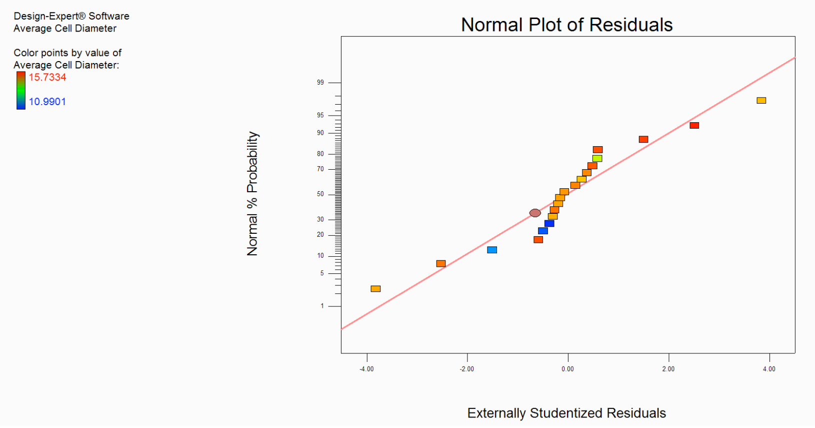
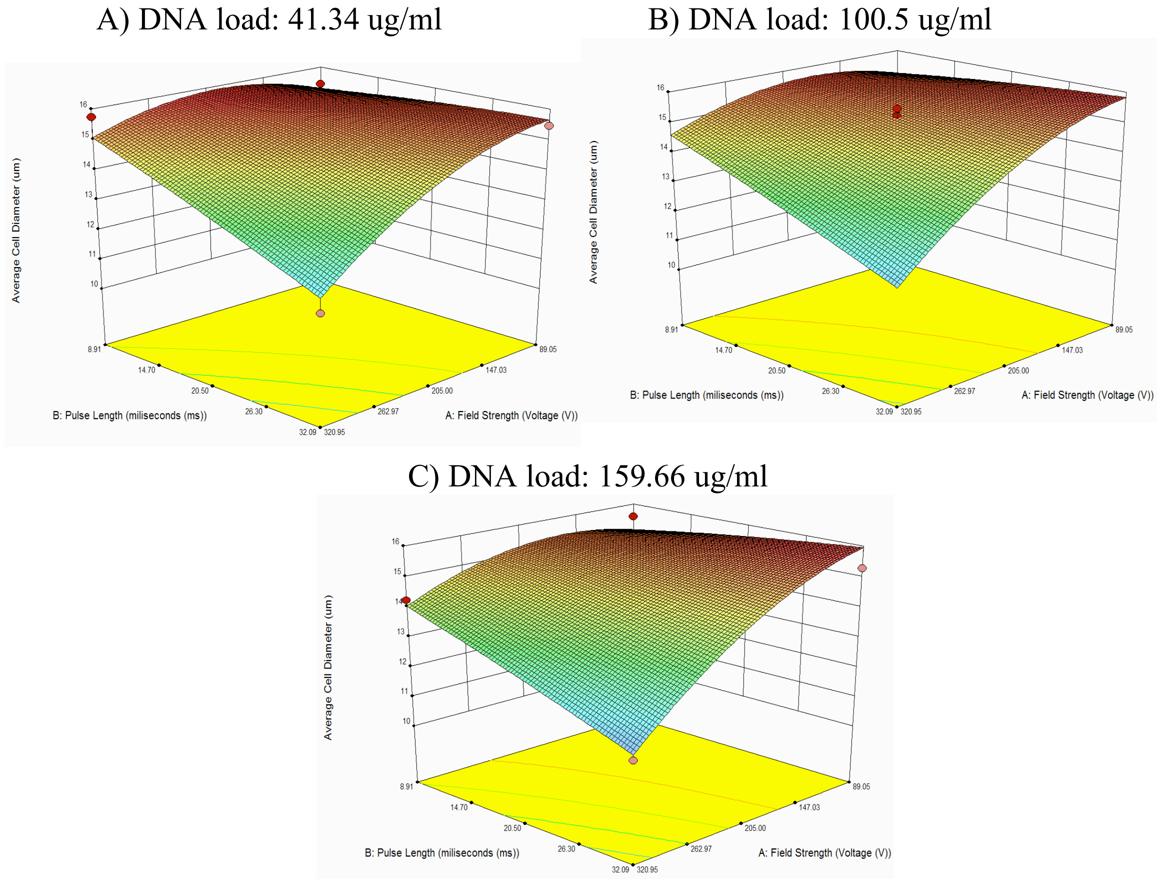


Figure A9: Exponential decay: Wide – ACD Normal Plot of Residuals



A) DNA load: 41.34 ug/ml

B) DNA load: 100.5 ug/ml

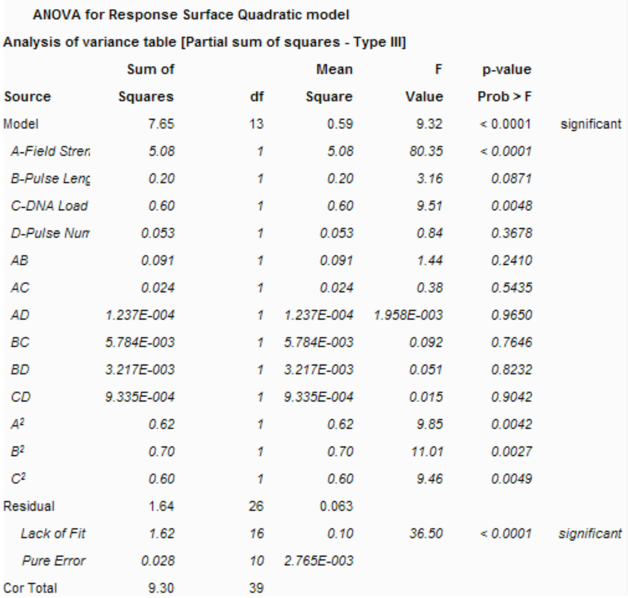
C) DNA load: 159.66 ug/ml

Figure A10. Exponential Decay: Wide – ACD Response Surface
Response surfaces of the ACD response to changes in field strength and pulse length at different levels of DNA load: Low Factorial (A), Center point (B) and Upper Factorial (C).

A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Quadratic vs 2FI	1.6	3	0.53	8.44	0.0004
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Quadratic	1.62	16	0.1	36.5	< 0.0001
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Quadratic	0.25	0.9923	0.7350	0.4291	5.31

Table A13: Square Wave: Wide – Transfection Efficiency Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	7.65	13	0.59	9.32	< 0.0001	significant
A-Field Stren	*5.08*	*1*	*5.08*	*80.35*	*< 0.0001*	
B-Pulse Leng	*0.20*	*1*	*0.20*	*3.16*	*0.0871*	
C-DNA Load	*0.60*	*1*	*0.60*	*9.51*	*0.0048*	
D-Pulse Num	*0.053*	*1*	*0.053*	*0.84*	*0.3678*	
AB	*0.091*	*1*	*0.091*	*1.44*	*0.2410*	
AC	*0.024*	*1*	*0.024*	*0.38*	*0.5435*	
AD	*1.237E-004*	*1*	*1.237E-004*	*1.958E-003*	*0.9650*	
BC	*5.784E-003*	*1*	*5.784E-003*	*0.092*	*0.7646*	
BD	*3.217E-003*	*1*	*3.217E-003*	*0.051*	*0.8232*	
CD	*9.335E-004*	*1*	*9.335E-004*	*0.015*	*0.9042*	
A²	*0.62*	*1*	*0.62*	*9.85*	*0.0042*	
B²	*0.70*	*1*	*0.70*	*11.01*	*0.0027*	
C²	*0.60*	*1*	*0.60*	*9.46*	*0.0049*	
Residual	1.64	26	0.063			
Lack of Fit	*1.62*	*16*	*0.10*	*36.50*	*< 0.0001*	*significant*
Pure Error	*0.028*	*10*	*2.765E-003*			
Cor Total	9.30	39				

Table A14: Square Wave: Wide – Transfection Efficiency ANOVA table

Table of ANOVA output statistical terms and values.

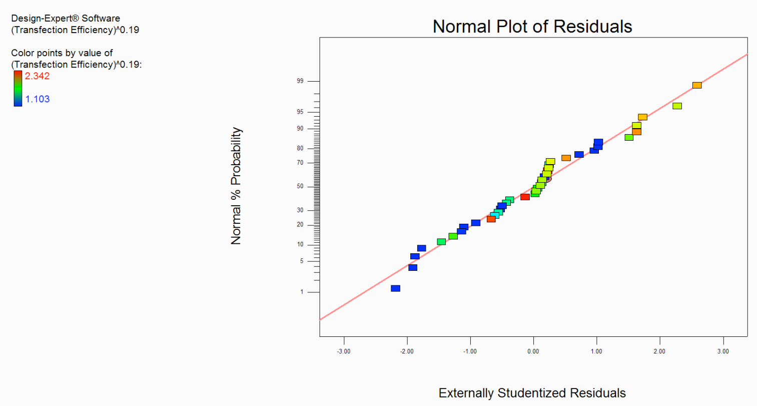
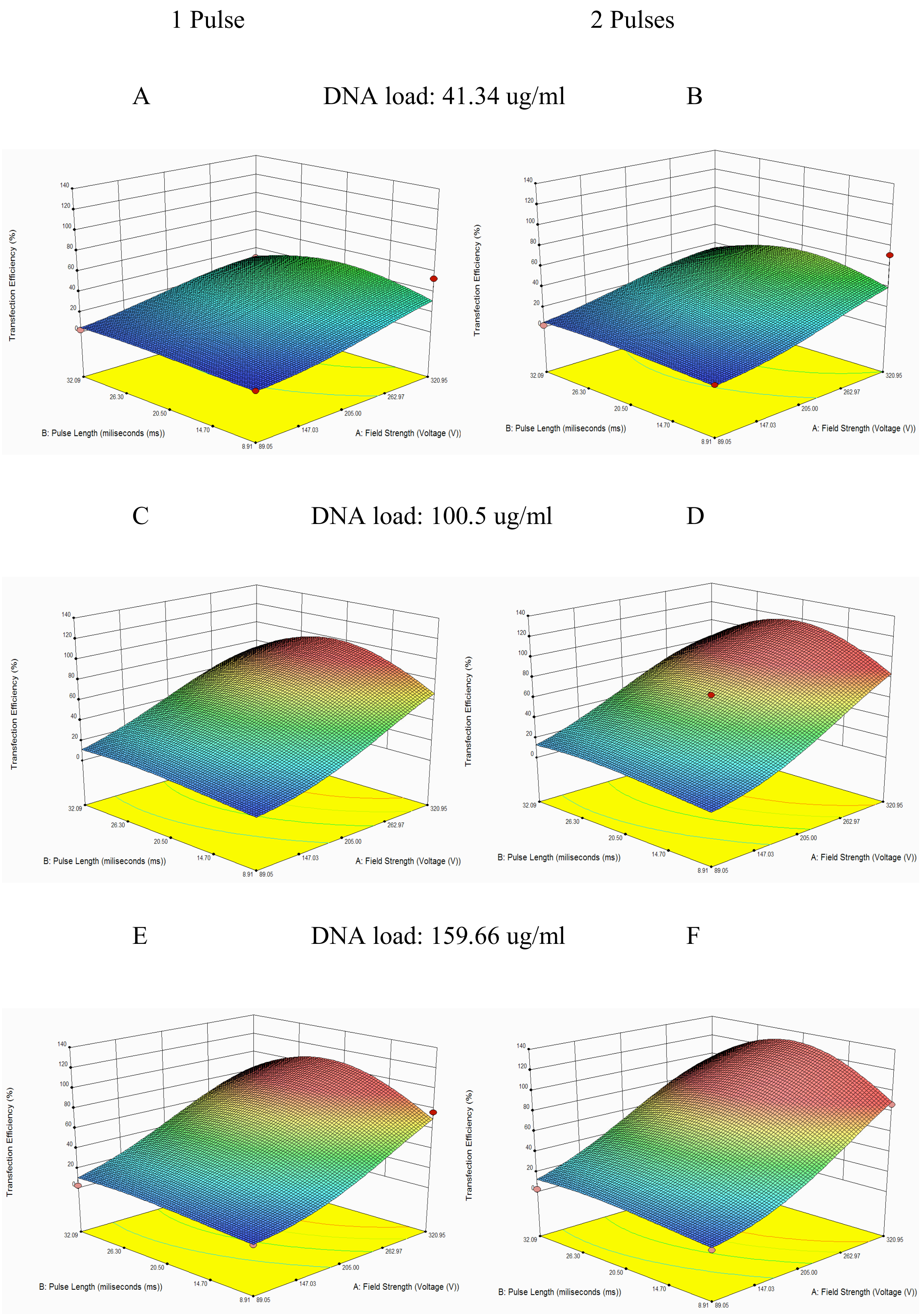


Figure A11: Square Wave: Wide - Transfection Efficiency Normal Plot of Residuals



1 Pulse

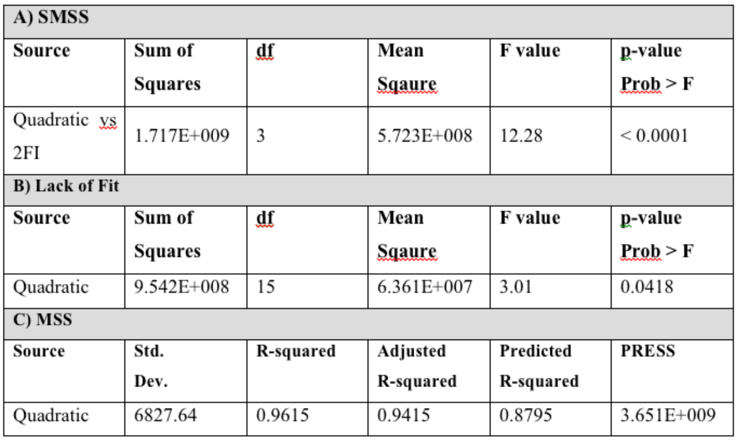
2 Pulses

A DNA load: 41.34 ug/ml B

C DNA load: 100.5 ug/ml D

E DNA load: 159.66 ug/ml F

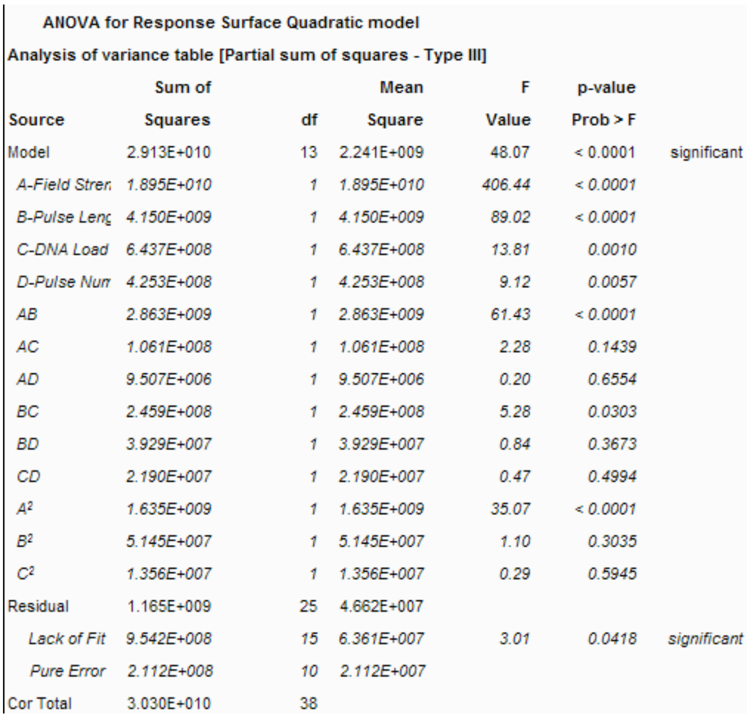
Figure A12. Square Wave: Wide – Transfection Efficiency Response Surface
Response surfaces of the transfection efficiency response to changes in field strength, pulse length, at different levels of DNA load (A&B, C&D, E&F) and with one (A,C,E) or two (B,D,F) pulses.



A) SMSS					
Source	Sum of Squares	df	Mean Sqaure	F value	p-value Prob > F
Quadratic vs 2FI	1.717E+009	3	5.723E+008	12.28	< 0.0001
B) Lack of Fit					
Source	Sum of Squares	df	Mean Sqaure	F value	p-value Prob > F
Quadratic	9.542E+008	15	6.361E+007	3.01	0.0418
C) MSS					
Source	Std. Dev.	R-squared	Adjusted R-squared	Predicted R-squared	PRESS
Quadratic	6827.64	0.9615	0.9415	0.8795	3.651E+009

Table A15: Square Wave: Wide – Cell Viability Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	2.913E+010	13	2.241E+009	48.07	< 0.0001	significant
A-Field Stren	1.895E+010	1	1.895E+010	406.44	< 0.0001	
B-Pulse Leng	4.150E+009	1	4.150E+009	89.02	< 0.0001	
C-DNA Load	6.437E+008	1	6.437E+008	13.81	0.0010	
D-Pulse Nun	4.253E+008	1	4.253E+008	9.12	0.0057	
AB	2.863E+009	1	2.863E+009	61.43	< 0.0001	
AC	1.061E+008	1	1.061E+008	2.28	0.1439	
AD	9.507E+006	1	9.507E+006	0.20	0.6554	
BC	2.459E+008	1	2.459E+008	5.28	0.0303	
BD	3.929E+007	1	3.929E+007	0.84	0.3673	
CD	2.190E+007	1	2.190E+007	0.47	0.4994	
A^2	1.635E+009	1	1.635E+009	35.07	< 0.0001	
B^2	5.145E+007	1	5.145E+007	1.10	0.3035	
C^2	1.356E+007	1	1.356E+007	0.29	0.5945	
Residual	1.165E+009	25	4.662E+007			
Lack of Fit	9.542E+008	15	6.361E+007	3.01	0.0418	significant
Pure Error	2.112E+008	10	2.112E+007			
Cor Total	3.030E+010	38				

Table A16: Square Wave: Wide – Cell Viability ANOVA table

Table of ANOVA output statistical terms and values.

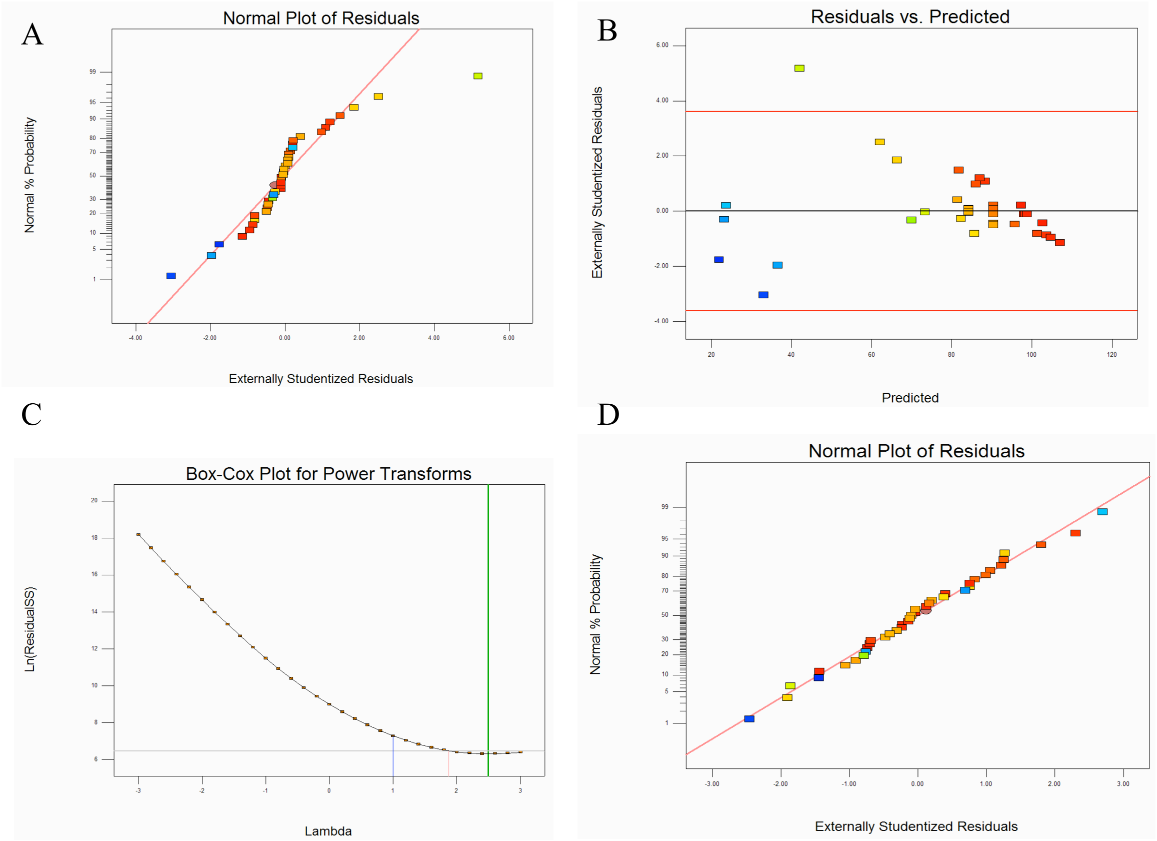


Figure A13. Square Wave: Wide – Cell Viability Data Manipulation
The figure shows the identification of non-normality (A) and the outlier responsible (B) as highlighted by a data point falling outside of a threshold level of residual (red line). The Box-Cox plot (C) highlights a recommended transformation (green line). After ignoring the outlier and the power transformation the data is normal (D).

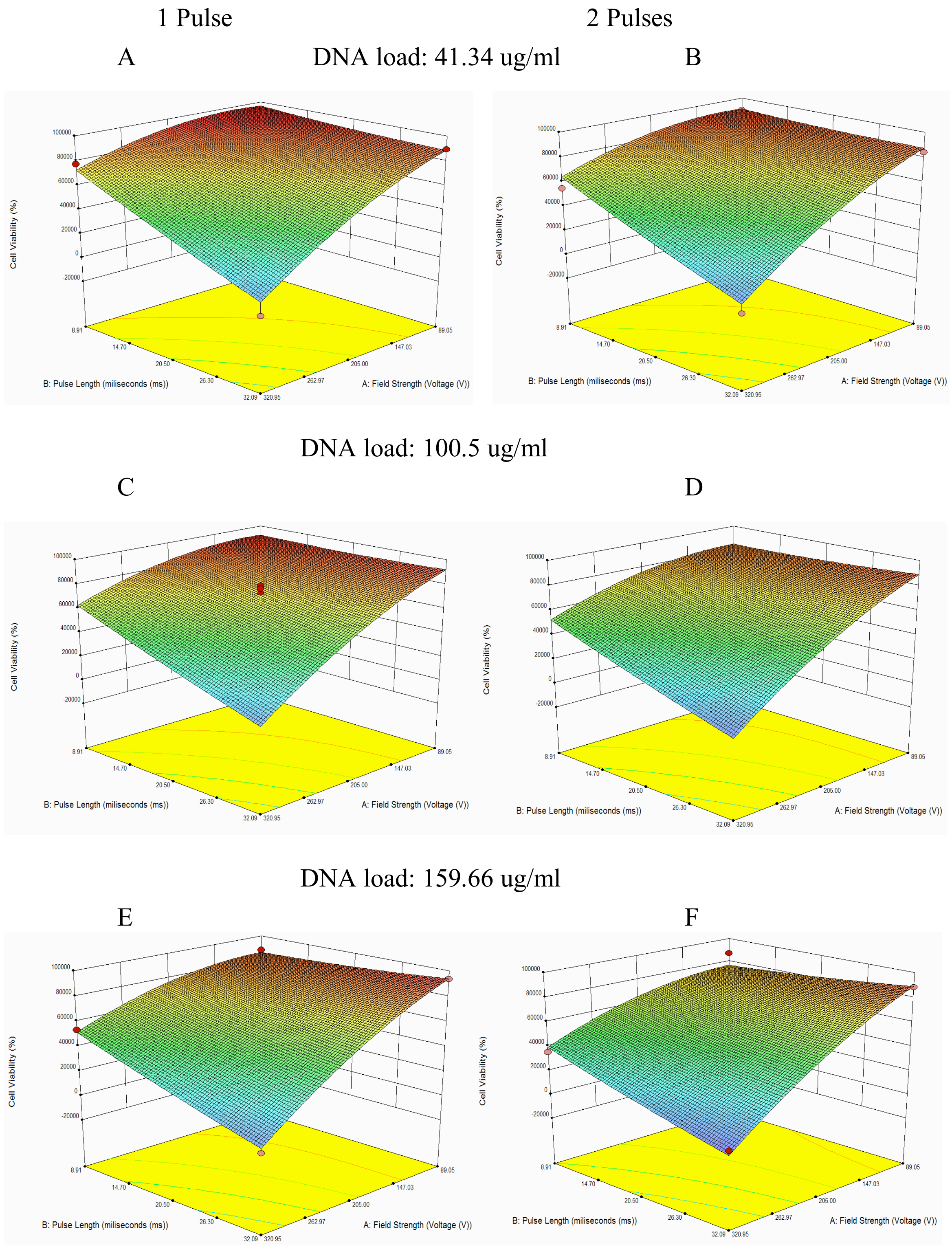
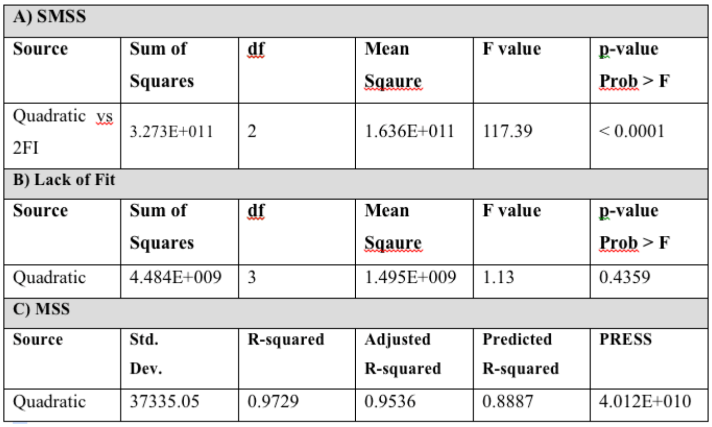


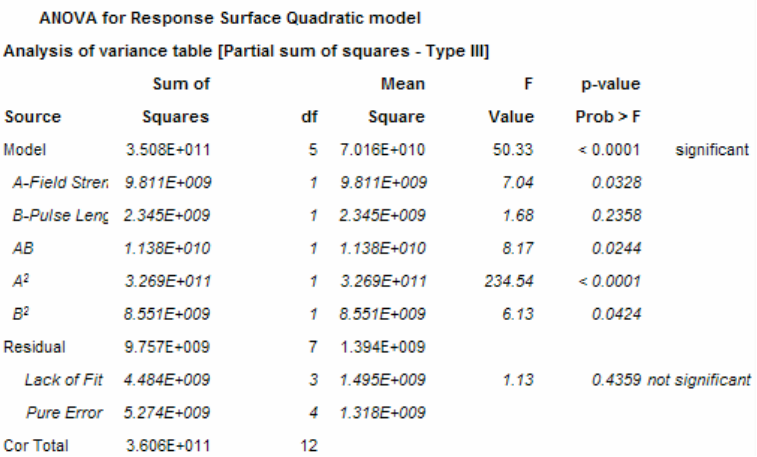
Figure A14. Square Wave: Wide – Cell Viability Response Surface
Response surfaces of the cell viability response to changes in field strength, pulse length, at different levels of DNA load (A&B, C&D, E&F) and with one (A,C,E) or two (B,D,F) pulses. Transformation had to be carried out manually, so the Y-axis is the transformed data scale.



A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Quadratic vs 2FI	3.273E+011	2	1.636E+011	117.39	< 0.0001
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Quadratic	4.484E+009	3	1.495E+009	1.13	0.4359
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Quadratic	37335.05	0.9729	0.9536	0.8887	4.012E+010

Table A17: Exponential Decay: Narrow – Transfection Efficiency Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	3.508E+011	5	7.016E+010	50.33	< 0.0001	significant
A-Field Stren	*9.811E+009*	*1*	*9.811E+009*	*7.04*	*0.0328*	
B-Pulse Leng	*2.345E+009*	*1*	*2.345E+009*	*1.68*	*0.2358*	
AB	*1.138E+010*	*1*	*1.138E+010*	*8.17*	*0.0244*	
A²	*3.269E+011*	*1*	*3.269E+011*	*234.54*	*< 0.0001*	
B²	*8.551E+009*	*1*	*8.551E+009*	*6.13*	*0.0424*	
Residual	9.757E+009	7	1.394E+009			
Lack of Fit	*4.484E+009*	*3*	*1.495E+009*	*1.13*	*0.4359*	*not significant*
Pure Error	*5.274E+009*	*4*	*1.318E+009*			
Cor Total	3.606E+011	12				

Table A18: Exponential Decay: Narrow – Transfection Efficiency ANOVA table

Table of ANOVA output statistical terms and values.

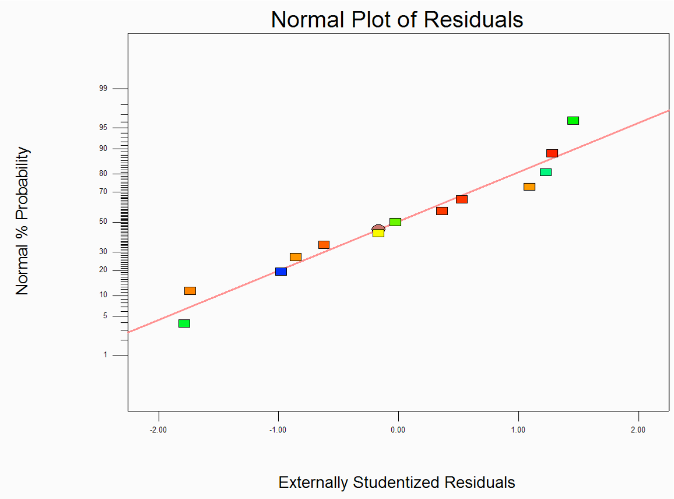
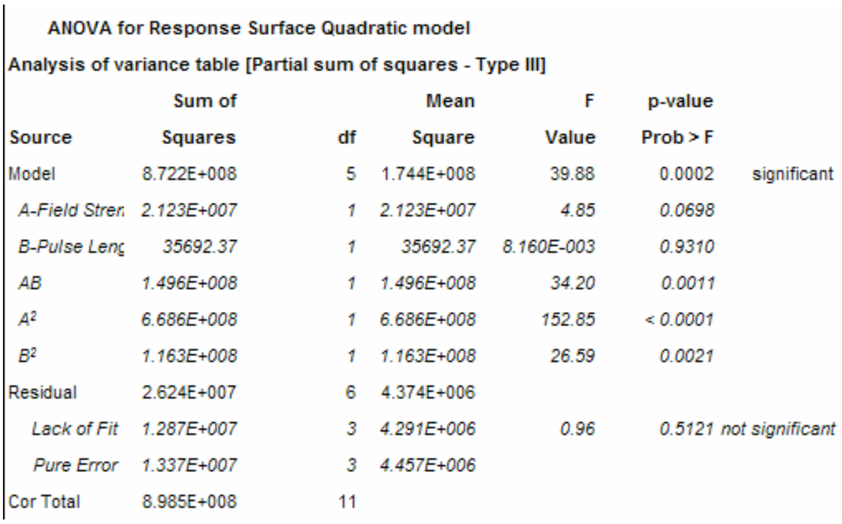


Figure A15: Exponential Decay: Narrow – Transfection Efficiency Normal Plot of Residuals

A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Quadratic vs 2FI	7.014E+008	2	3.507E+008	80.18	< 0.0001
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Sqaure**	**F value**	**p-value Prob > F**
Quadratic	1.287E+007	3	4.291E+006	0.96	0.5121
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Quadratic	2091.36	0.9708	0.9465	0.8717	1.53E+008

Table A19: Exponential Decay: Narrow – Median Fluorescence Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	8.722E+008	5	1.744E+008	39.88	0.0002	significant
A-Field Strer	*2.123E+007*	*1*	*2.123E+007*	*4.85*	*0.0698*	
B-Pulse Leng	*35692.37*	*1*	*35692.37*	*8.160E-003*	*0.9310*	
AB	*1.496E+008*	*1*	*1.496E+008*	*34.20*	*0.0011*	
A²	*6.686E+008*	*1*	*6.686E+008*	*152.85*	*< 0.0001*	
B²	*1.163E+008*	*1*	*1.163E+008*	*26.59*	*0.0021*	
Residual	2.624E+007	6	4.374E+006			
Lack of Fit	*1.287E+007*	*3*	*4.291E+006*	*0.96*	*0.5121*	*not significant*
Pure Error	*1.337E+007*	*3*	*4.457E+006*			
Cor Total	8.985E+008	11				

Table A20: Exponential Decay: Narrow – Median Fluorescence ANOVA table

Table of ANOVA output statistical terms and values.

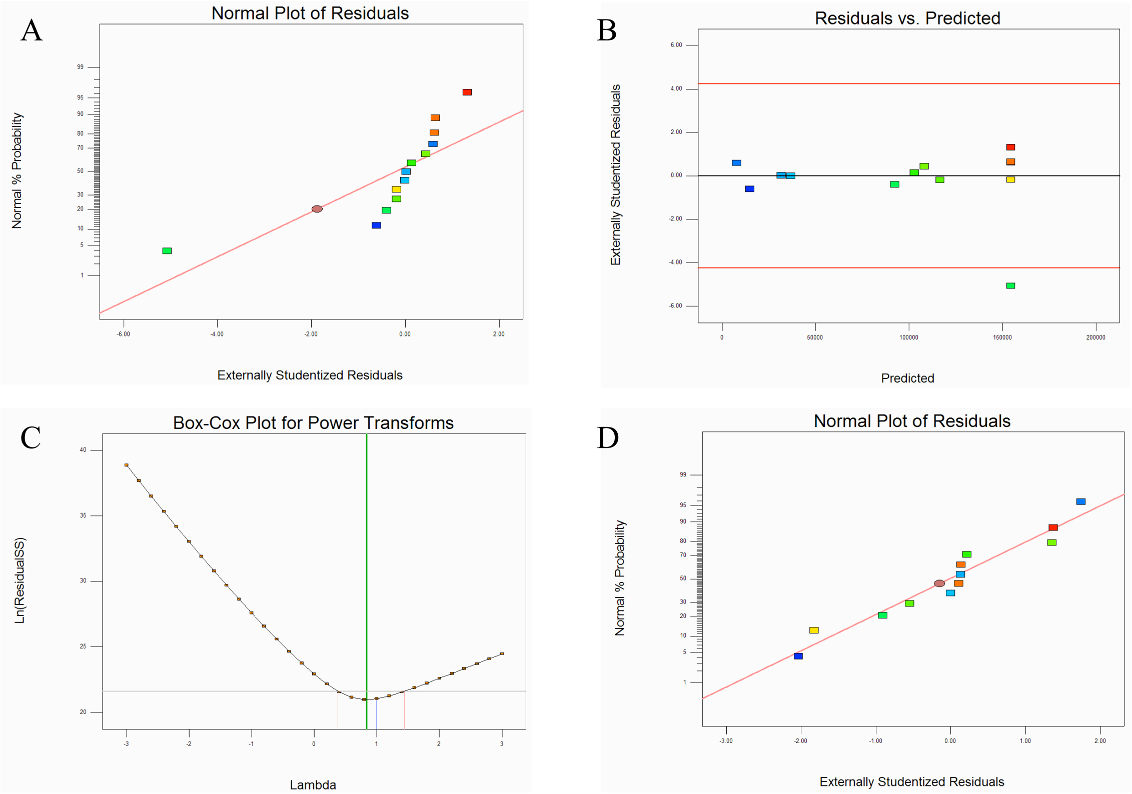


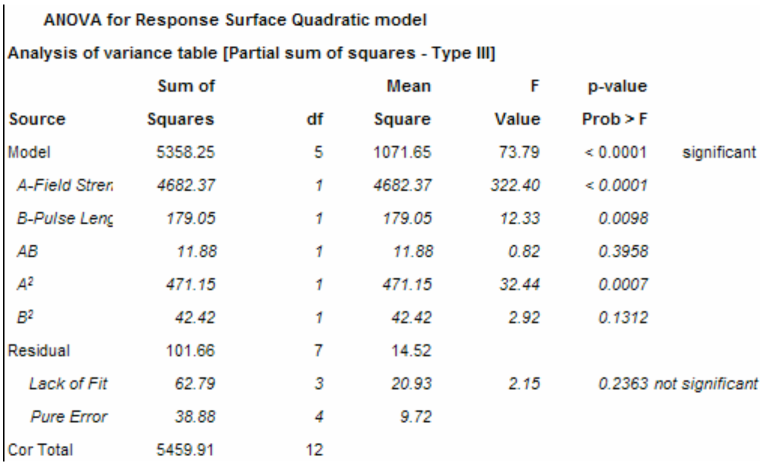
Figure A16. Exponential Decay: Narrow – Median Fluorescence Data Manipulation

The figure shows the identification of non-normality (A) and the outlier responsible (B) as highlighted by a data point falling outside of a threshold level of residual (red line). The Box-Cox plot (C) highlights a recommended transformation (green line). After ignoring the outlier and transformation the data is normal (D).

A) SMSS					
Source	Sum of Squares	df	Mean Sqaure	F value	p-value Prob > F
Quadratic vs 2FI	484.94	2	242.47	16.70	0.0022
B) Lack of Fit					
Source	Sum of Squares	df	Mean Sqaure	F value	p-value Prob > F
Quadratic	62.79	3	20.93	2.15	0.2363
C) MSS					
Source	Std. Dev.	R-squared	Adjusted R-squared	Predicted R-squared	PRESS
Quadratic	3.81	0.9814	0.9681	0.8071	507.22

Table A21: Exponential Decay: Narrow – Cell Viability Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	5358.25	5	1071.65	73.79	< 0.0001	significant
A-Field Stren	*4682.37*	*1*	*4682.37*	*322.40*	*< 0.0001*	
B-Pulse Leng	*179.05*	*1*	*179.05*	*12.33*	*0.0098*	
AB	*11.88*	*1*	*11.88*	*0.82*	*0.3958*	
A²	*471.15*	*1*	*471.15*	*32.44*	*0.0007*	
B²	*42.42*	*1*	*42.42*	*2.92*	*0.1312*	
Residual	101.66	7	14.52			
Lack of Fit	*62.79*	*3*	*20.93*	*2.15*	*0.2363*	*not significant*
Pure Error	*38.88*	*4*	*9.72*			
Cor Total	5459.91	12				

Table A22: Exponential Decay: Narrow – Cell Viability ANOVA table

Table of ANOVA output statistical terms and values.

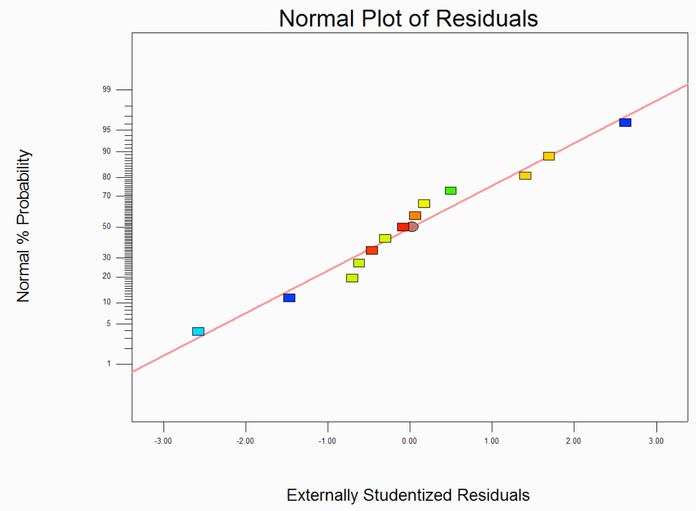
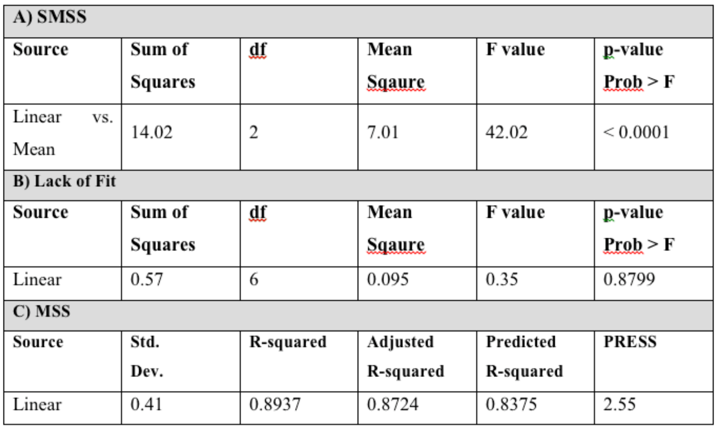


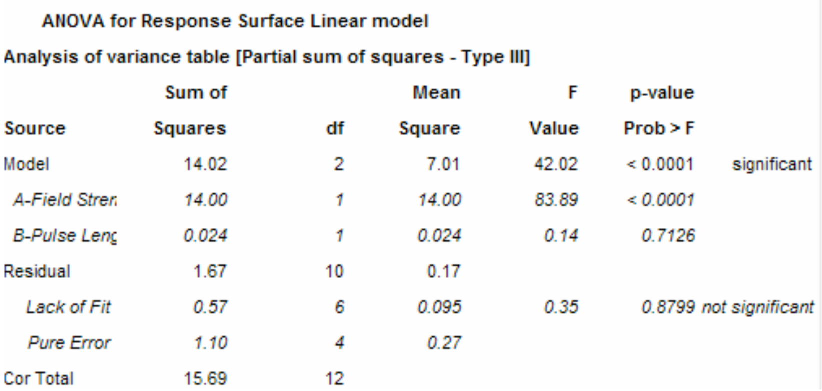
Figure A17: Exponential Decay: Narrow – Cell Viability Normal Plot of Residuals



A) SMSS					
Source	Sum of Squares	df	Mean Sqaure	F value	p-value Prob > F
Linear vs. Mean	14.02	2	7.01	42.02	< 0.0001
B) Lack of Fit					
Source	Sum of Squares	df	Mean Sqaure	F value	p-value Prob > F
Linear	0.57	6	0.095	0.35	0.8799
C) MSS					
Source	Std. Dev.	R-squared	Adjusted R-squared	Predicted R-squared	PRESS
Linear	0.41	0.8937	0.8724	0.8375	2.55

Table A23: Exponential Decay: Narrow – ACD Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface Linear model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	14.02	2	7.01	42.02	< 0.0001	significant
A-Field Stren	*14.00*	*1*	*14.00*	*83.89*	*< 0.0001*	
B-Pulse Leng	*0.024*	*1*	*0.024*	*0.14*	*0.7126*	
Residual	1.67	10	0.17			
Lack of Fit	*0.57*	*6*	*0.095*	*0.35*	*0.8799*	*not significant*
Pure Error	*1.10*	*4*	*0.27*			
Cor Total	15.69	12				

Table A24: Exponential Decay: Narrow – ACD ANOVA table

Table of ANOVA output statistical terms and values.

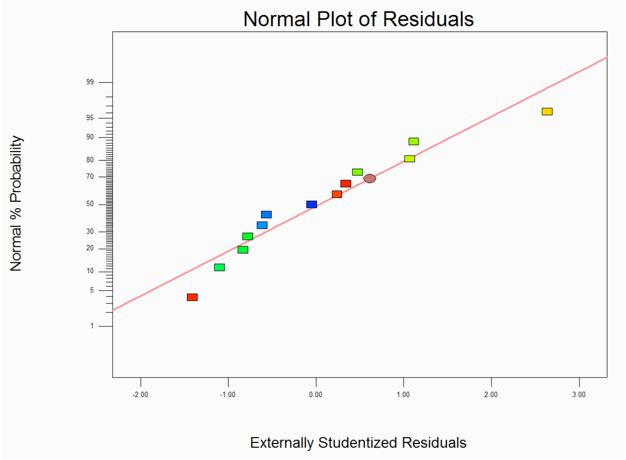
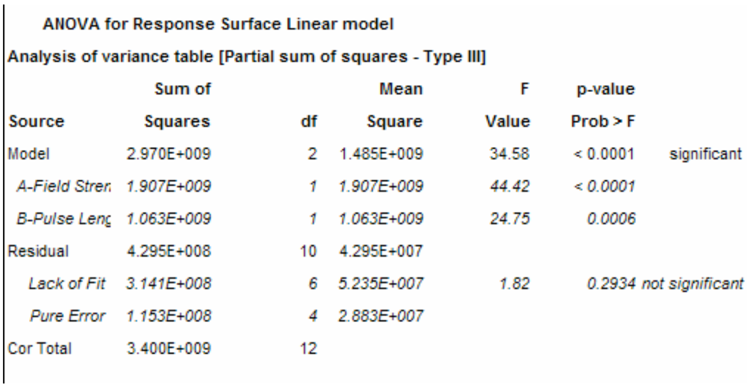


Figure A18: Exponential Decay: Narrow – ACD Normal Plot of Residuals

A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Linear vs. Mean	2.970E+009	2	1.485E+009	34.58	< 0.0001
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Linear	3.141E+008	6	5.235E+007	1.82	0.2934
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Linear	6553.26	0.8737	0.8484	0.7765	7.597E+008

Table A25: Square Wave: Narrow – Transfection Efficiency Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface Linear model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	2.970E+009	2	1.485E+009	34.58	< 0.0001	significant
A-Field Stren	*1.907E+009*	*1*	*1.907E+009*	*44.42*	*< 0.0001*	
B-Pulse Leng	*1.063E+009*	*1*	*1.063E+009*	*24.75*	*0.0006*	
Residual	4.295E+008	10	4.295E+007			
Lack of Fit	*3.141E+008*	*6*	*5.235E+007*	*1.82*	*0.2934*	*not significant*
Pure Error	*1.153E+008*	*4*	*2.883E+007*			
Cor Total	3.400E+009	12				

Table A26: Square Wave: Narrow – Transfection Efficiency ANOVA table

Table of ANOVA output statistical terms and values.

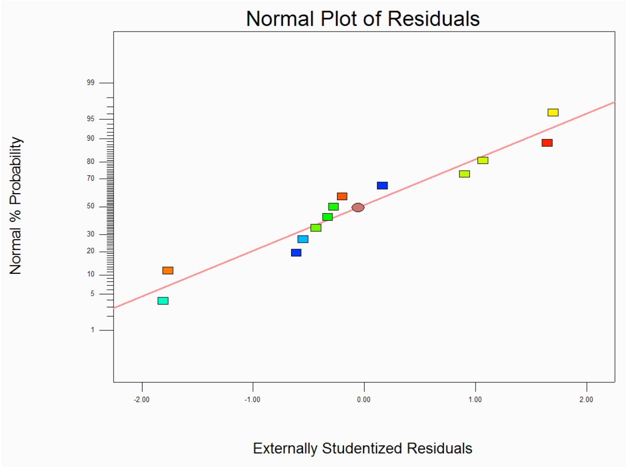
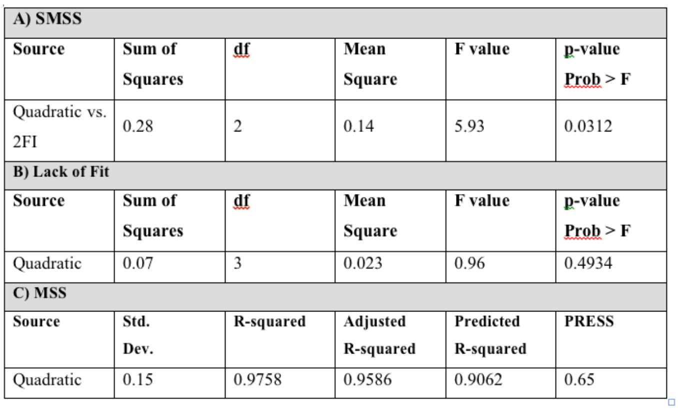


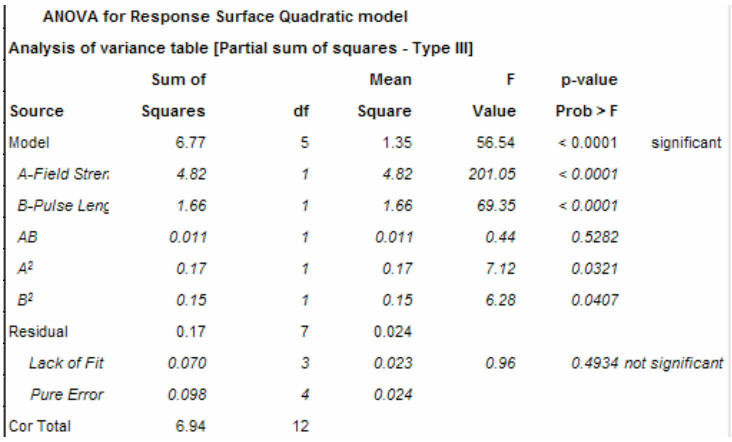
Figure A19: Square Wave: Narrow – Transfection Efficiency Normal Plot of Residuals



A) SMSS					
Source	Sum of Squares	df	Mean Square	F value	p-value Prob > F
Quadratic vs. 2FI	0.28	2	0.14	5.93	0.0312
B) Lack of Fit					
Source	Sum of Squares	df	Mean Square	F value	p-value Prob > F
Quadratic	0.07	3	0.023	0.96	0.4934
C) MSS					
Source	Std. Dev.	R-squared	Adjusted R-squared	Predicted R-squared	PRESS
Quadratic	0.15	0.9758	0.9586	0.9062	0.65

Table A27: Square Wave: Narrow – Median Fluorescence Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	6.77	5	1.35	56.54	< 0.0001	significant
A-Field Strer	*4.82*	*1*	*4.82*	*201.05*	*< 0.0001*	
B-Pulse Leng	*1.66*	*1*	*1.66*	*69.35*	*< 0.0001*	
AB	*0.011*	*1*	*0.011*	*0.44*	*0.5282*	
A^2	*0.17*	*1*	*0.17*	*7.12*	*0.0321*	
B^2	*0.15*	*1*	*0.15*	*6.28*	*0.0407*	
Residual	0.17	7	0.024			
Lack of Fit	*0.070*	*3*	*0.023*	*0.96*	*0.4934*	*not significant*
Pure Error	*0.098*	*4*	*0.024*			
Cor Total	6.94	12				

Table A28: Square Wave: Narrow – Median Fluorescence ANOVA table

Table of ANOVA output statistical terms and values.

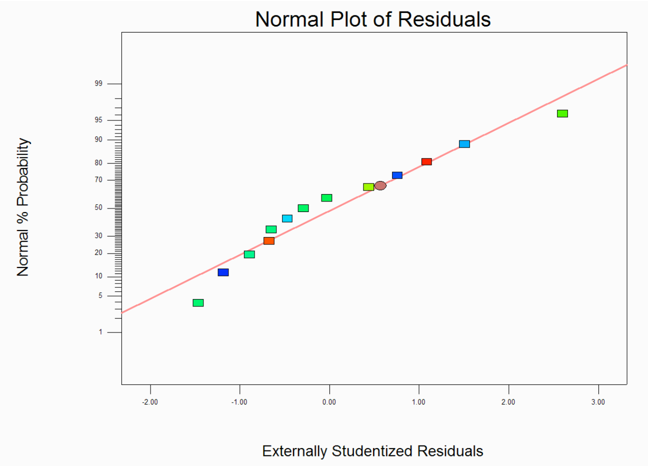
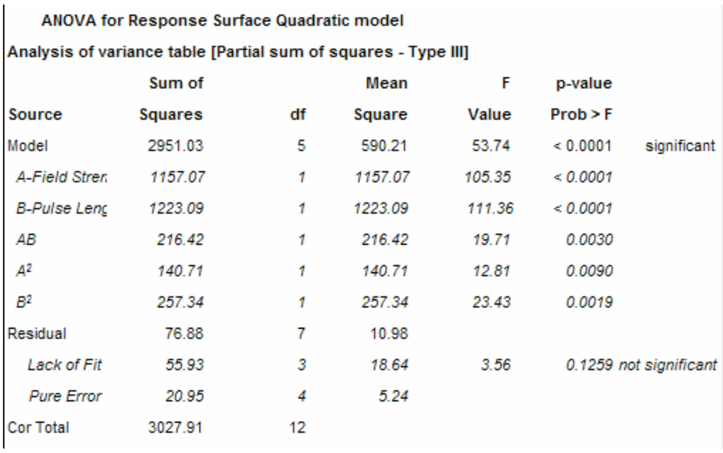


Figure A20: Square Wave: Narrow – Median Fluorescence Normal Plot of Residuals

A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Quadratic vs. 2FI	354.44	2	177.22	16.14	0.0024
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Quadratic	55.93	3	18.64	3.56	0.1259
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Quadratic	3.31	0.9765	0.9565	0.8578	430.45

Table A29: Square Wave: Narrow – Cell Viability Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS.



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	2951.03	5	590.21	53.74	< 0.0001	significant
A-Field Stren	*1157.07*	*1*	*1157.07*	*105.35*	*< 0.0001*	
B-Pulse Leng	*1223.09*	*1*	*1223.09*	*111.36*	*< 0.0001*	
AB	*216.42*	*1*	*216.42*	*19.71*	*0.0030*	
A^2	*140.71*	*1*	*140.71*	*12.81*	*0.0090*	
B^2	*257.34*	*1*	*257.34*	*23.43*	*0.0019*	
Residual	76.88	7	10.98			
Lack of Fit	*55.93*	*3*	*18.64*	*3.56*	*0.1259*	*not significant*
Pure Error	*20.95*	*4*	*5.24*			
Cor Total	3027.91	12				

Table A30: Square Wave: Narrow – Cell Viability ANOVA table

Table of ANOVA output statistical terms and values.

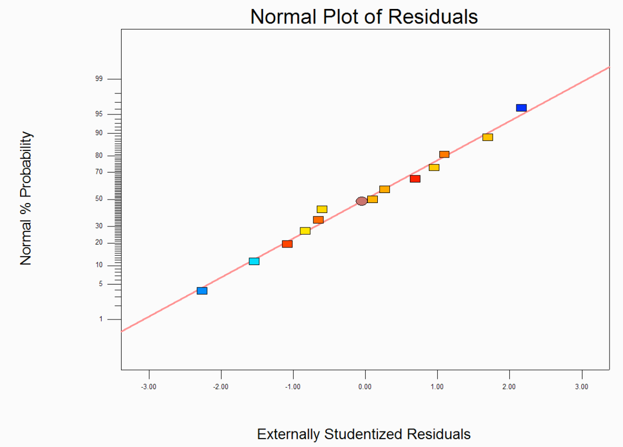
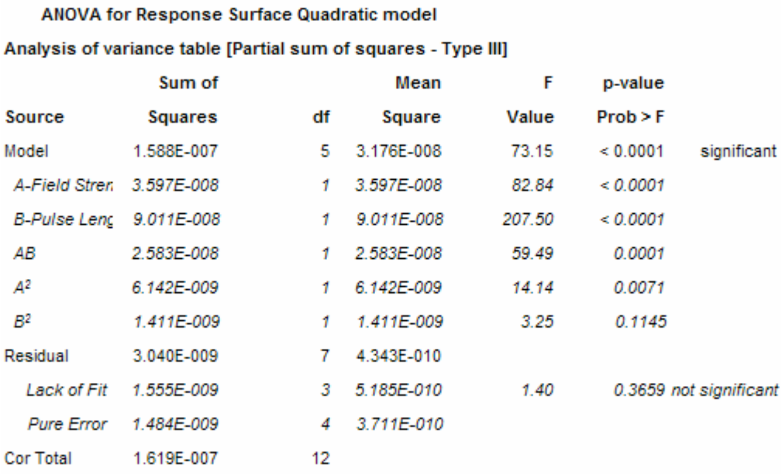


Figure A21: Square Wave: Narrow – Cell Viability Normal Plot of Residuals

A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Quadratic vs. 2FI	6.903E-009	2	3.451E-009	7.95	0.0158
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Quadratic	1.555E-009	3	5.185E-010	1.40	0.3659
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Quadratic	2.084E-005	0.9812	0.9678	0.9173	1.338E-008

Table A31: Square Wave: Narrow – ACD Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS.



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	**Sum of Squares**	**df**	**Mean Square**	**F Value**	**p-value Prob > F**	
Model	1.588E-007	5	3.176E-008	73.15	< 0.0001	significant
A-Field Stren	*3.597E-008*	*1*	*3.597E-008*	*82.84*	*< 0.0001*	
B-Pulse Leng	*9.011E-008*	*1*	*9.011E-008*	*207.50*	*< 0.0001*	
AB	*2.583E-008*	*1*	*2.583E-008*	*59.49*	*0.0001*	
A²	*6.142E-009*	*1*	*6.142E-009*	*14.14*	*0.0071*	
B²	*1.411E-009*	*1*	*1.411E-009*	*3.25*	*0.1145*	
Residual	3.040E-009	7	4.343E-010			
Lack of Fit	*1.555E-009*	*3*	*5.185E-010*	*1.40*	*0.3659*	*not significant*
Pure Error	*1.484E-009*	*4*	*3.711E-010*			
Cor Total	1.619E-007	12				

Table A32: Square Wave: Narrow – ACD ANOVA table

Table of ANOVA output statistical terms and values.

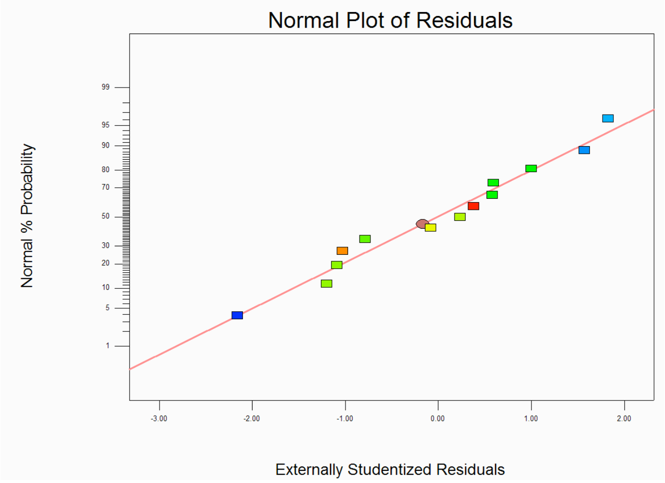
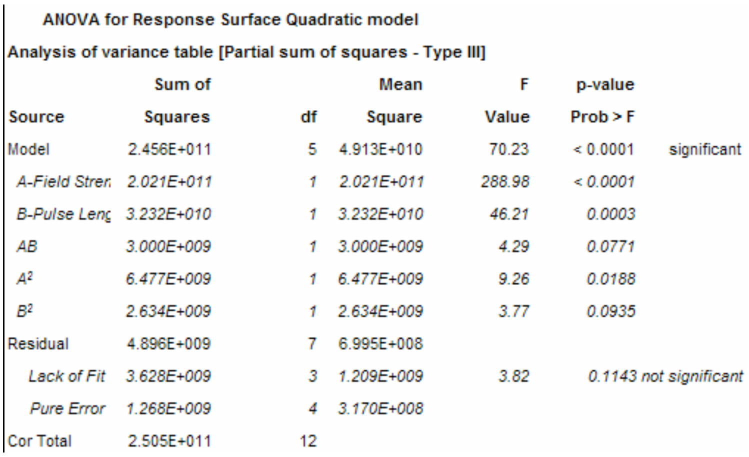


Figure A22: Square Wave: Narrow – ACD Normal Plot of Residuals

A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Quadratic vs. 2FI	8.173E+009	2	4.086E+009	5.84	0.0322
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Quadratic	3.628E+009	3	1.209E+009	3.82	0.1143
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Quadratic	26447.76	0.9805	0.9665	0.8891	2.778E+010

Table A33: Exponential Decay: Narrow 2 – Transfection Efficiency Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS.



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	2.456E+011	5	4.913E+010	70.23	< 0.0001	significant
A-Field Stren	*2.021E+011*	*1*	*2.021E+011*	*288.98*	*< 0.0001*	
B-Pulse Leng	*3.232E+010*	*1*	*3.232E+010*	*46.21*	*0.0003*	
AB	*3.000E+009*	*1*	*3.000E+009*	*4.29*	*0.0771*	
A^2	*6.477E+009*	*1*	*6.477E+009*	*9.26*	*0.0188*	
B^2	*2.634E+009*	*1*	*2.634E+009*	*3.77*	*0.0935*	
Residual	4.896E+009	7	6.995E+008			
Lack of Fit	*3.628E+009*	*3*	*1.209E+009*	*3.82*	*0.1143*	*not significant*
Pure Error	*1.268E+009*	*4*	*3.170E+008*			
Cor Total	2.505E+011	12				

Table A34: Exponential Decay: Narrow 2 – Transfection Efficiency ANOVA table

Table of ANOVA output statistical terms and values.

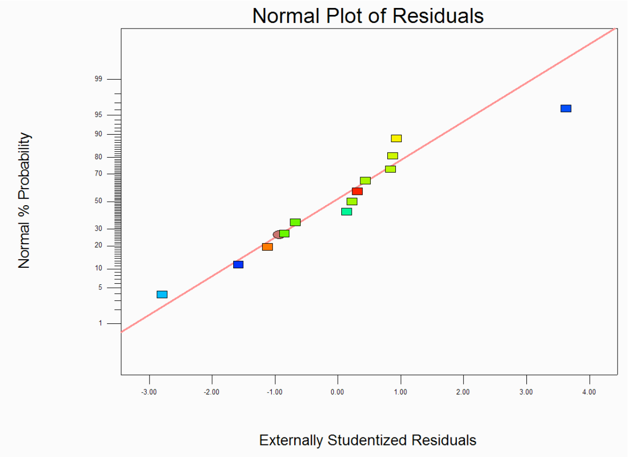
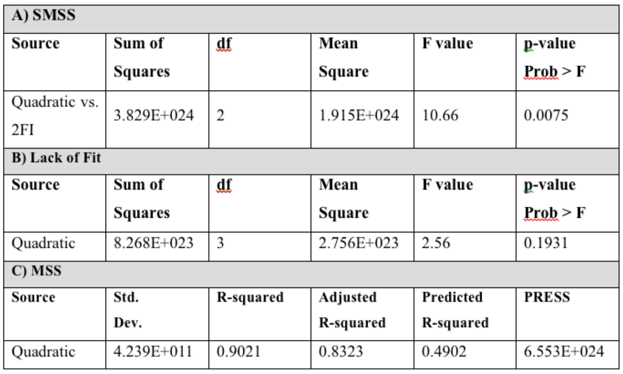


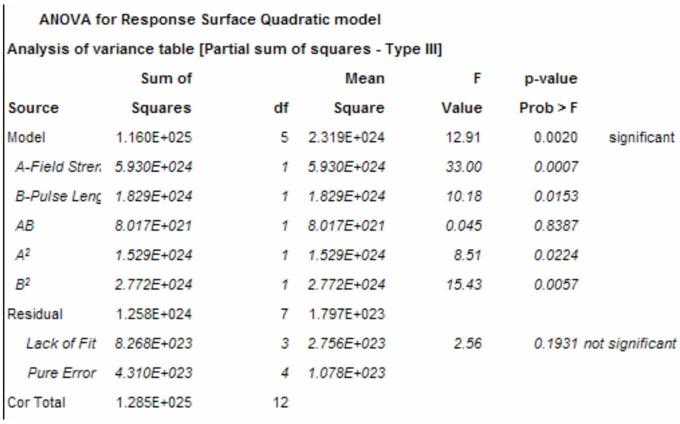
Figure A23: Exponential Decay: Narrow 2 – Transfection Efficiency Normal Plot of Residuals



A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Quadratic vs. 2FI	3.829E+024	2	1.915E+024	10.66	0.0075
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Quadratic	8.268E+023	3	2.756E+023	2.56	0.1931
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Quadratic	4.239E+011	0.9021	0.8323	0.4902	6.553E+024

Table A35: Exponential Decay: Narrow 2 – Median Fluorescence Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS.



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	1.160E+025	5	2.319E+024	12.91	0.0020	significant
A-Field Strer	*5.930E+024*	*1*	*5.930E+024*	*33.00*	*0.0007*	
B-Pulse Leng	*1.829E+024*	*1*	*1.829E+024*	*10.18*	*0.0153*	
AB	*8.017E+021*	*1*	*8.017E+021*	*0.045*	*0.8387*	
A^2	*1.529E+024*	*1*	*1.529E+024*	*8.51*	*0.0224*	
B^2	*2.772E+024*	*1*	*2.772E+024*	*15.43*	*0.0057*	
Residual	1.258E+024	7	1.797E+023			
Lack of Fit	*8.268E+023*	*3*	*2.756E+023*	*2.56*	*0.1931*	*not significant*
Pure Error	*4.310E+023*	*4*	*1.078E+023*			
Cor Total	1.285E+025	12				

Table A36: Exponential Decay: Narrow 2 – Median Fluorescence ANOVA table

Table of ANOVA output statistical terms and values.

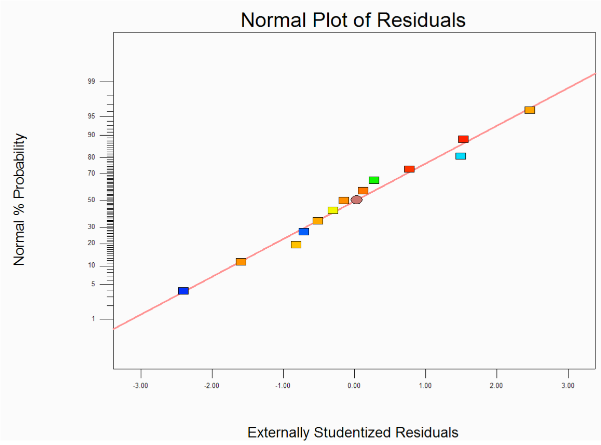
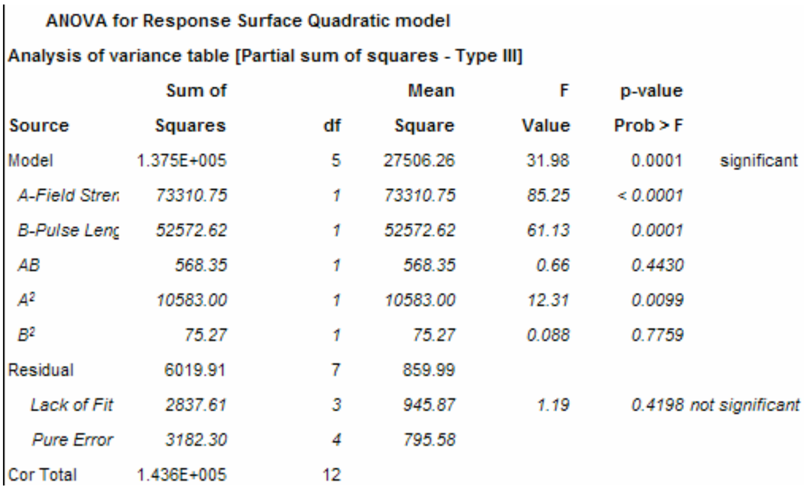


Figure A24: Exponential Decay: Narrow 2 – Median Fluorescence Normal Plot of Residuals

A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Quadratic vs. 2FI	11079.61	2	5539.8	6.44	0.0259
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Quadratic	2837.61	3	945.87	1.19	0.4198
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Quadratic	29.33	0.9581	0.9281	0.8248	25150.92

Table A37: Exponential Decay: Narrow 2 – Cell Viability Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS.



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	1.375E+005	5	27506.26	31.98	0.0001	significant
A-Field Strer	*73310.75*	*1*	*73310.75*	*85.25*	*< 0.0001*	
B-Pulse Leng	*52572.62*	*1*	*52572.62*	*61.13*	*0.0001*	
AB	*568.35*	*1*	*568.35*	*0.66*	*0.4430*	
A²	*10583.00*	*1*	*10583.00*	*12.31*	*0.0099*	
B²	*75.27*	*1*	*75.27*	*0.088*	*0.7759*	
Residual	6019.91	7	859.99			
Lack of Fit	*2837.61*	*3*	*945.87*	*1.19*	*0.4198*	*not significant*
Pure Error	*3182.30*	*4*	*795.58*			
Cor Total	1.436E+005	12				

Table A38: Exponential Decay: Narrow 2 – Cell Viability ANOVA table

Table of ANOVA output statistical terms and values.

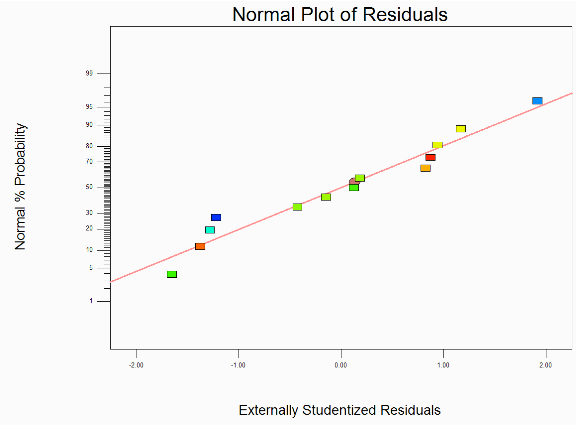
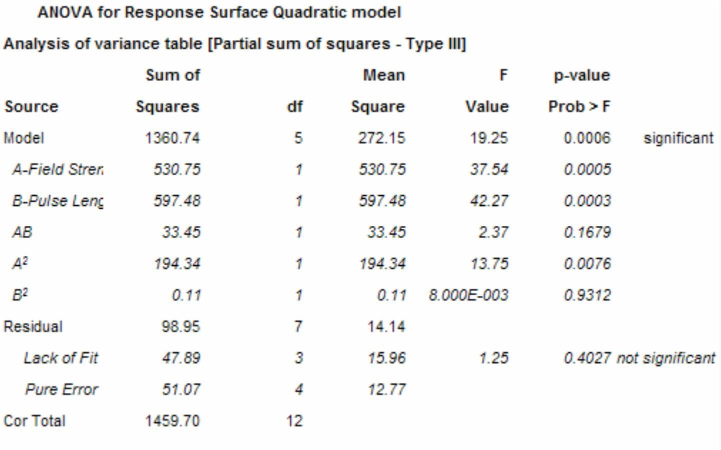


Figure A25: Exponential Decay: Narrow 2 – Cell Viability Normal Plot of Residuals

A) SMSS					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Quadratic vs. 2FI	199.07	2	99.53	7.04	0.0211
B) Lack of Fit					
Source	**Sum of Squares**	**df**	**Mean Square**	**F value**	**p-value Prob > F**
Quadratic	47.89	3	15.96	1.25	0.4027
C) MSS					
Source	**Std. Dev.**	**R-squared**	**Adjusted R-squared**	**Predicted R-squared**	**PRESS**
Quadratic	3.76	0.9322	0.8838	0.7120	420.33

Table A39: Exponential Decay: Narrow 2 – ACD Fit Summary

Showing suggested model fit summary data for A) SMSS, B) Lack of fit and C) MSS.



ANOVA for Response Surface Quadratic model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	1360.74	5	272.15	19.25	0.0006	significant
A-Field Stren	*530.75*	*1*	*530.75*	*37.54*	*0.0005*	
B-Pulse Leng	*597.48*	*1*	*597.48*	*42.27*	*0.0003*	
AB	*33.45*	*1*	*33.45*	*2.37*	*0.1679*	
A^2	*194.34*	*1*	*194.34*	*13.75*	*0.0076*	
B^2	*0.11*	*1*	*0.11*	*8.000E-003*	*0.9312*	
Residual	98.95	7	14.14			
Lack of Fit	*47.89*	*3*	*15.96*	*1.25*	*0.4027*	*not significant*
Pure Error	*51.07*	*4*	*12.77*			
Cor Total	1459.70	12				

Table A40: Exponential Decay: Narrow 2 – ACD ANOVA table

Table of ANOVA output statistical terms and values.

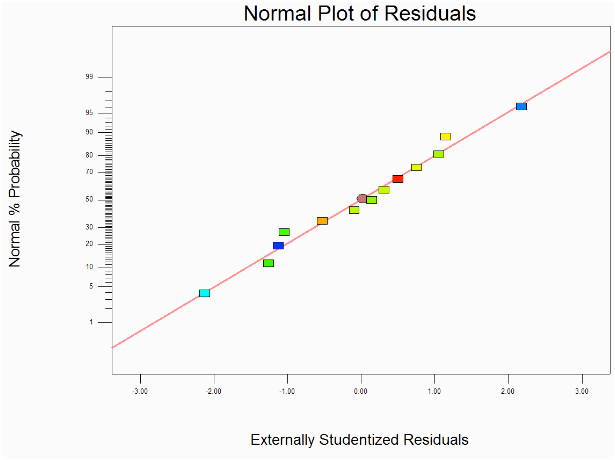


Figure A26: Exponential Decay: Narrow 2 – ACD Normal Plot of Residuals

MS	Passage	Allele	W-statistic	P-value	Normally_Distributed
BAT25	High	1	0.895646285	0.034199133	FALSE
GNAT	High	1	0.955992723	0.467212562	TRUE
GT23	High	1	0.971771309	0.588696645	TRUE
MS.10.1	High	1	0.936382757	0.204689835	TRUE
MS.11.1	High	1	0.899093069	0.055405989	TRUE
MS.21.1	High	1	0.994787502	0.999987622	TRUE
BAT25	Low	1	0.96528812	0.653911998	TRUE
GNAT	Low	1	0.980223251	0.936951501	TRUE
GT23	Low	1	0.95515801	0.2318533	TRUE
MS.10.1	Low	1	0.897088704	0.036380816	FALSE
MS.11.1	Low	1	0.984519137	0.984522183	TRUE
MS.21.1	Low	1	0.978412725	0.91199442	TRUE
BAT25	High	2	0.997320982	0.999999986	TRUE
GNAT	High	2	0.953212688	0.418500308	TRUE
GT23	High	2	0.927206732	0.041412929	FALSE
MS.10.1	High	2	0.94470569	0.293768813	TRUE
MS.11.1	High	2	0.985465071	0.98880964	TRUE
MS.21.1	High	2	0.968981519	0.733234678	TRUE
BAT25	Low	2	0.944752948	0.294362489	TRUE
GNAT	Low	2	0.948800272	0.349245936	TRUE
GT23	Low	2	0.976101988	0.715171958	TRUE
MS.10.1	Low	2	0.952588214	0.408093338	TRUE
MS.11.1	Low	2	0.901202793	0.060299815	TRUE
MS.21.1	Low	2	0.949498502	0.359541652	TRUE
BAT25	High	3	0.973529861	0.82695169	TRUE
GNAT	High	3	0.961268927	0.569501156	TRUE
GT23	High	3	0.986332045	0.957634808	TRUE
MS.10.1	High	3	0.761632736	0.000244157	FALSE
MS.11.1	High	3	0.942050863	0.314051686	TRUE
MS.21.1	High	3	0.97229543	0.802428827	TRUE
BAT25	Low	3	0.949802048	0.364094987	TRUE
GNAT	Low	3	0.971701901	0.790342261	TRUE
GT23	Low	3	0.947660783	0.1463017	TRUE
MS.10.1	Low	3	0.990947843	0.999047513	TRUE
MS.11.1	Low	3	0.989889003	0.998695	TRUE
MS.21.1	Low	3	0.97305754	0.817676017	TRUE
BAT25	High	4	0.924725022	0.122193565	TRUE
GT23	High	4	0.980311649	0.833781078	TRUE
MS.10.1	High	4	0.883302838	0.020295718	FALSE
MS.11.1	High	4	0.991767305	0.999688775	TRUE
BAT25	Low	4	0.940775101	0.247997664	TRUE
GT23	Low	4	0.940063733	0.091318917	TRUE
MS.10.1	Low	4	0.933929475	0.183729819	TRUE
MS.11.1	Low	4	0.981347061	0.962832437	TRUE
GT23	High	5	0.969105742	0.515004932	TRUE
MS.11.1	High	5	0.980370064	0.953750386	TRUE
GT23	Low	5	0.949635713	0.165302526	TRUE
MS.11.1	Low	5	0.971079941	0.817737857	TRUE
GT23	High	6	0.978254395	0.77748186	TRUE
MS.11.1	High	6	0.988504169	0.997050654	TRUE
GT23	Low	6	0.92262145	0.031384188	FALSE
MS.11.1	Low	6	0.956221066	0.530639633	TRUE

Table A41: Shapiro Wilk Test for Normality 1

MS	Passage	Allele	W-statistic	P-value	Normally_Distributed
BAT25	High	1	0.914534023	0.077807028	TRUE
GNAT	High	1	0.955992723	0.467212562	TRUE
GT23	High	1	0.971771309	0.588696645	TRUE
MS.10.1	High	1	0.940145989	0.241308698	TRUE
MS.11.1	High	1	0.899093069	0.055405989	TRUE
MS.21.1	High	1	0.994787502	0.999987622	TRUE
BAT25	Low	1	0.96528812	0.653911998	TRUE
GNAT	Low	1	0.980223251	0.936951501	TRUE
GT23	Low	1	0.95515801	0.2318533	TRUE
MS.10.1	Low	1	0.990640822	0.993910437	TRUE
MS.11.1	Low	1	0.984519137	0.984522183	TRUE
MS.21.1	Low	1	0.978412725	0.91199442	TRUE
BAT25	High	2	0.997320982	0.999999986	TRUE
GNAT	High	2	0.953212688	0.418500308	TRUE
GT23	High	2	0.96471322	0.641622583	TRUE
MS.10.1	High	2	0.937627722	0.216181175	TRUE
MS.11.1	High	2	0.985465071	0.98880964	TRUE
MS.21.1	High	2	0.968981519	0.733234678	TRUE
BAT25	Low	2	0.944752948	0.294362489	TRUE
GNAT	Low	2	0.948800272	0.349245936	TRUE
GT23	Low	2	0.976101988	0.715171958	TRUE
MS.10.1	Low	2	0.952588214	0.408093338	TRUE
MS.11.1	Low	2	0.901202793	0.060299815	TRUE
MS.21.1	Low	2	0.949498502	0.359541652	TRUE
BAT25	High	3	0.973529861	0.82695169	TRUE
GNAT	High	3	0.961268927	0.569501156	TRUE
GT23	High	3	0.986332045	0.957634808	TRUE
MS.10.1	High	3	0.944002359	0.285058017	TRUE
MS.11.1	High	3	0.942050863	0.314051686	TRUE
MS.21.1	High	3	0.97229543	0.802428827	TRUE
BAT25	Low	3	0.949802048	0.364094987	TRUE
GNAT	Low	3	0.971701901	0.790342261	TRUE
GT23	Low	3	0.947660783	0.1463017	TRUE
MS.10.1	Low	3	0.990947843	0.999047513	TRUE
MS.11.1	Low	3	0.989889003	0.998695	TRUE
MS.21.1	Low	3	0.97305754	0.817676017	TRUE
BAT25	High	4	0.924725022	0.122193565	TRUE
GT23	High	4	0.980311649	0.833781078	TRUE
MS.10.1	High	4	0.892273963	0.029615721	FALSE
MS.11.1	High	4	0.991767305	0.999688775	TRUE
BAT25	Low	4	0.940775101	0.247997664	TRUE
GT23	Low	4	0.940063733	0.091318917	TRUE
MS.10.1	Low	4	0.933929475	0.183729819	TRUE
MS.11.1	Low	4	0.981347061	0.962832437	TRUE
GT23	High	5	0.969105742	0.515004932	TRUE
MS.11.1	High	5	0.980370064	0.953750386	TRUE
GT23	Low	5	0.949635713	0.165302526	TRUE
MS.11.1	Low	5	0.971079941	0.817737857	TRUE
GT23	High	6	0.978254395	0.77748186	TRUE
MS.11.1	High	6	0.988504169	0.997050654	TRUE
GT23	Low	6	0.95041019	0.173390193	TRUE
MS.11.1	Low	6	0.956221066	0.530639633	TRUE

Table A42: Shapiro Wilk Test for Normality 2

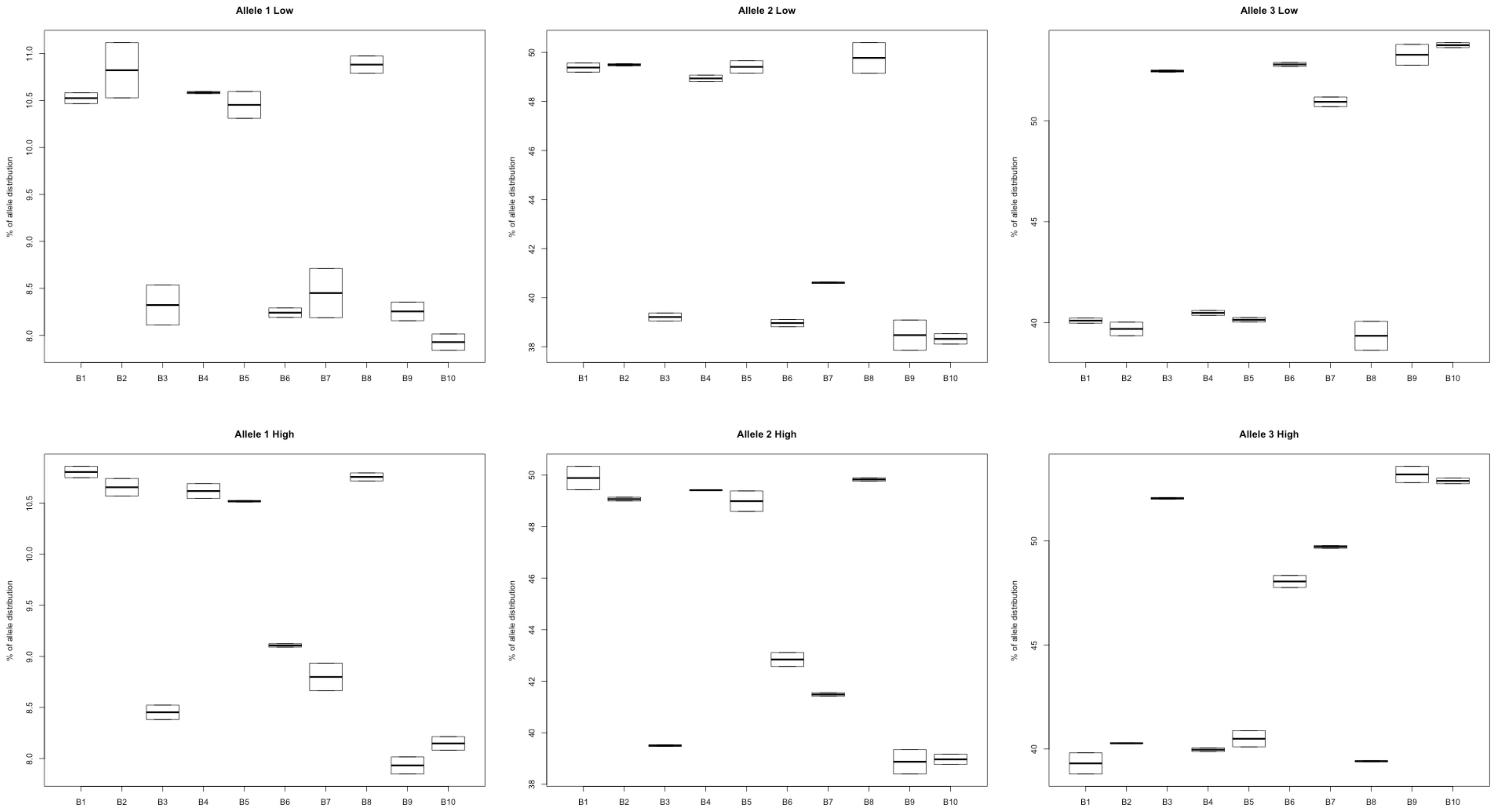


Figure A27: GNAT2 Box Plots for Allele Percentage

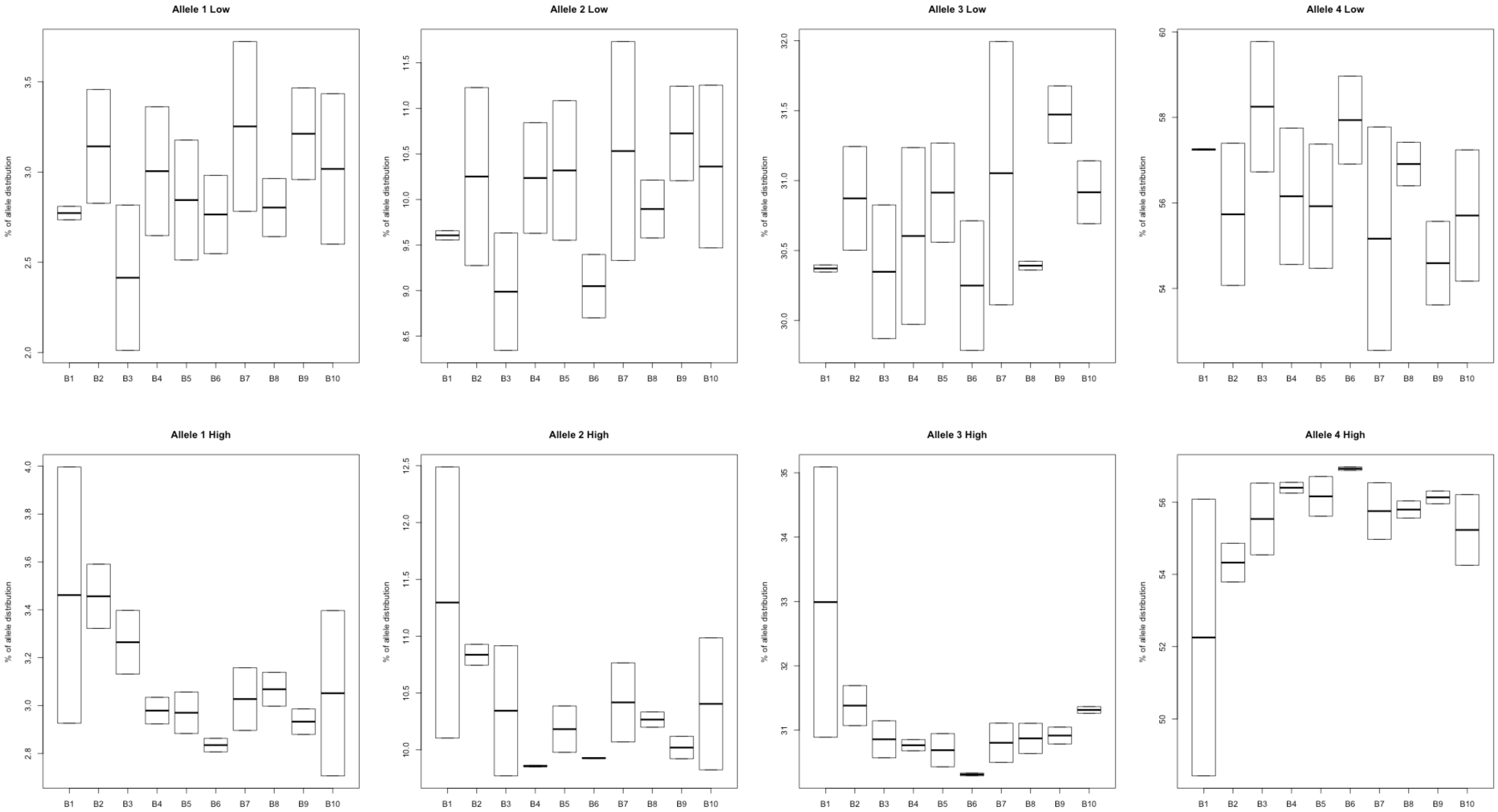


Figure A28: 10.1 Box Plots for Allele Percentage

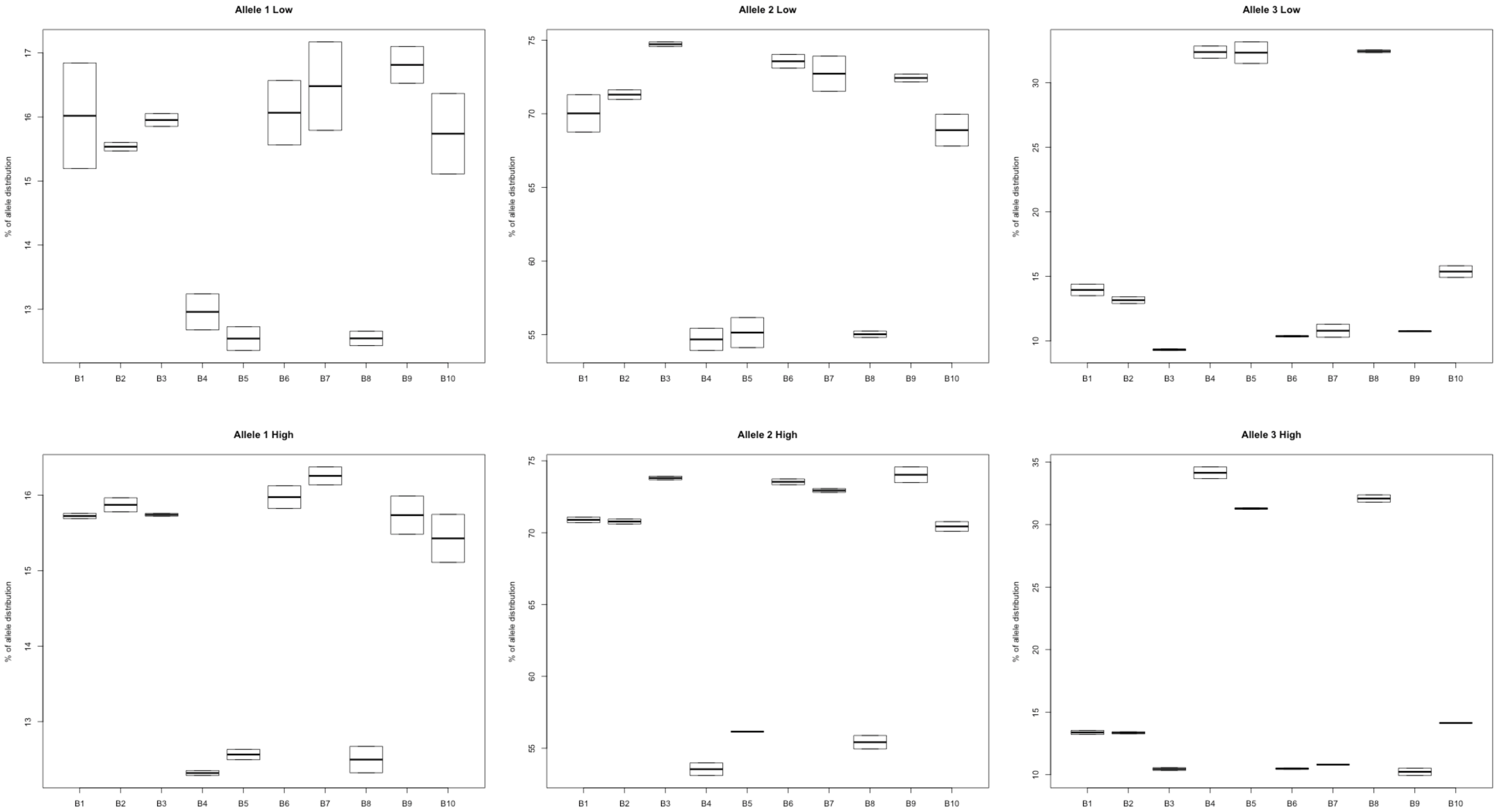


Figure A29: 21.1 Box Plots for Allele Percentage

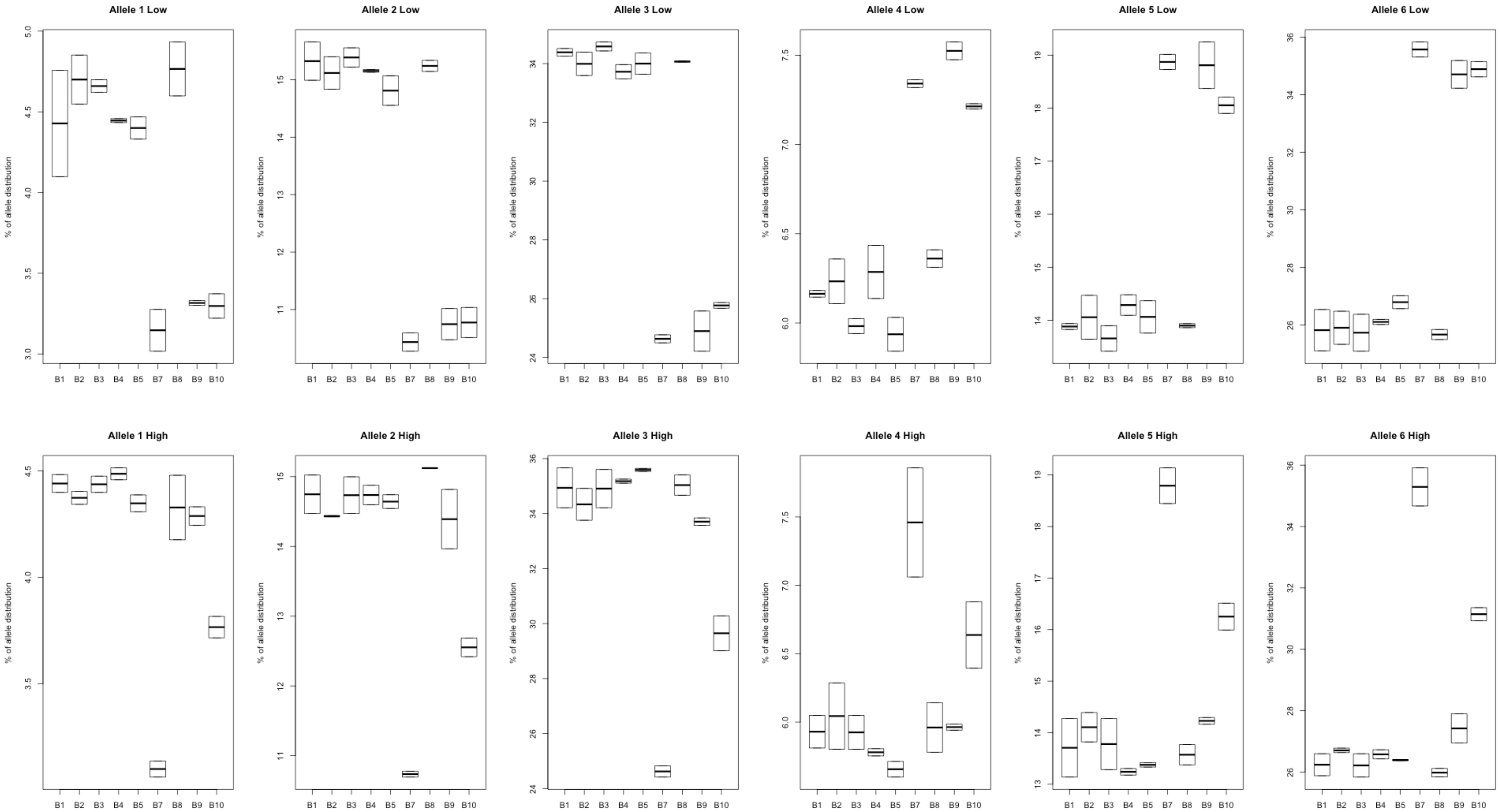


Figure A30: 11.1 Box Plots for Allele Percentage

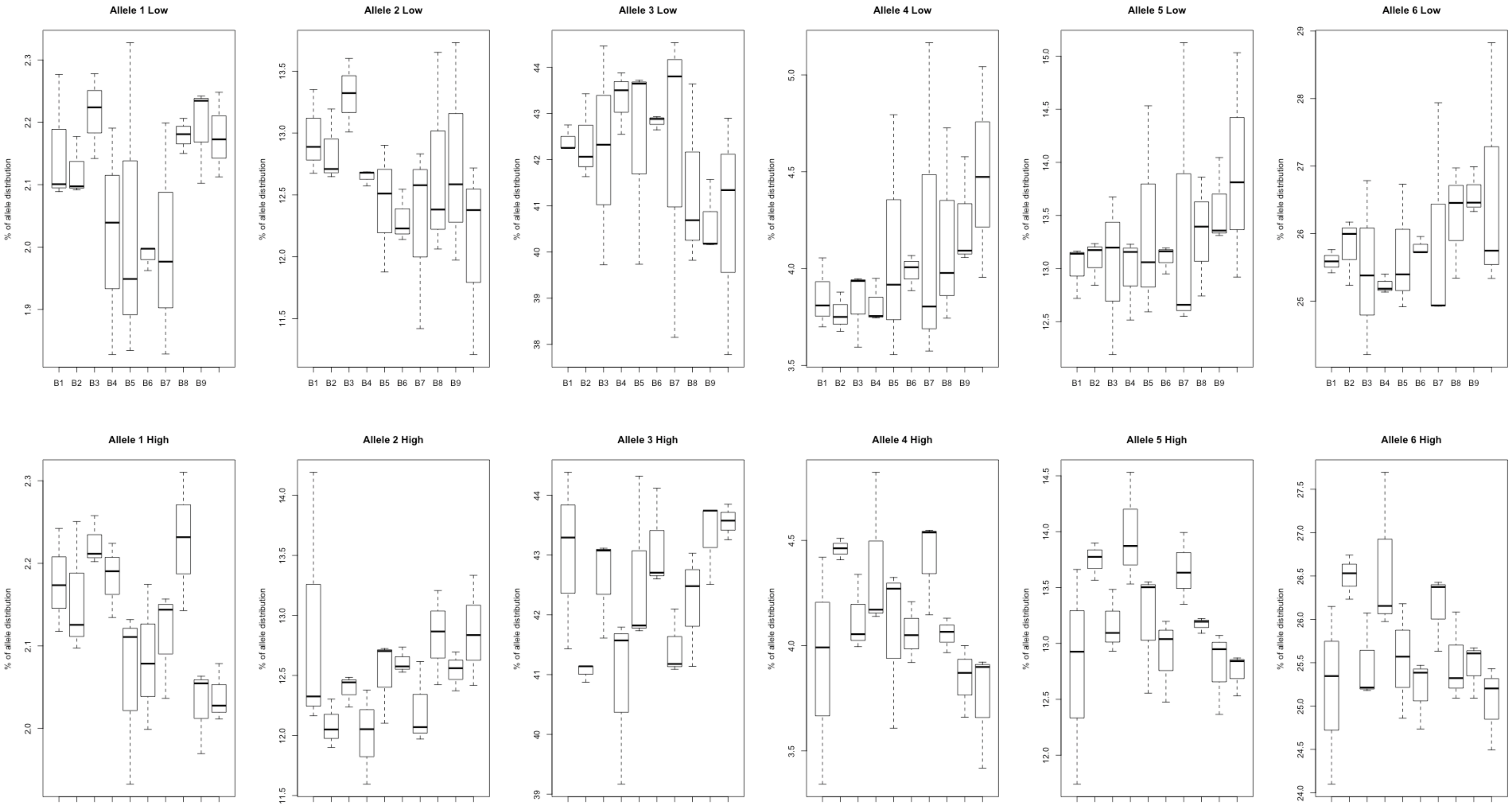


Figure A31: GT-23 Box Plots for Allele Percentage

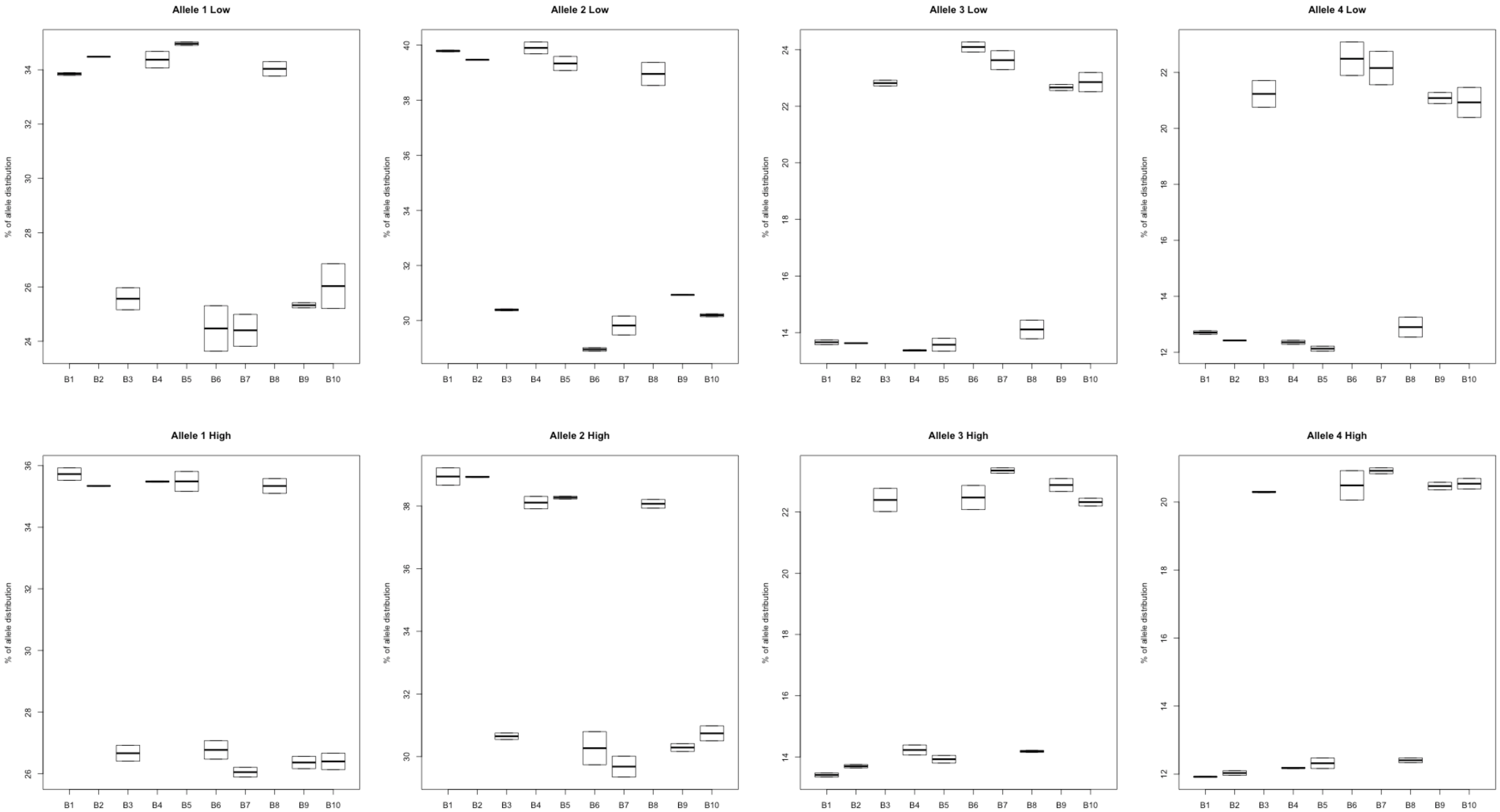


Figure A32: BAT25 Box Plots for Allele Percentage

		Variance	
Microsatellite	Allele	Ratio	P-value
BAT25	1	1.053519542	0.910714013
GNAT	1	1.168340239	0.738009019
GT23	1	2.180148051	0.039847612
MS.10.1	1	1.794863861	0.211487385
MS.11.1	1	2.221055888	0.109459607
MS.21.1	1	1.13393719	0.78694436
BAT25	2	1.345976545	0.523510817
GNAT	2	1.216823155	0.673184772
GT23	2	1.456597188	0.316784346
MS.10.1	2	2.089612921	0.116870231
MS.11.1	2	2.481139237	0.069350219
MS.21.1	2	0.983106051	0.970765138
BAT25	3	1.178428529	0.724121937
GNAT	3	1.203642153	0.690330612
GT23	3	2.060195742	0.056227015
MS.10.1	3	0.358176031	0.030520912
MS.11.1	3	1.519576052	0.396990224
MS.21.1	3	0.984898839	0.973890062
BAT25	4	1.192579588	0.704995678
GT23	4	1.610278191	0.205559438
MS.10.1	4	1.007585909	0.987030567
MS.11.1	4	1.091005427	0.859576507
GT23	5	1.47114654	0.304239528
MS.11.1	5	1.589758002	0.348423574
GT23	6	1.692691268	0.162388698
MS.11.1	6	2.062425191	0.145617444

Table A43: F Test for Variance Comparison

Microsatellite	Allele	Cluster	Variance Ratio	P-value
BAT25	1	B	6.383817457	0.010896608
GNAT	1	B	3.792108002	0.059966717
MS.11.1	1	B	0.055724458	0.006565282
MS.21.1	1	B	3.526929437	0.192822575
BAT25	2	B	1.960644011	0.330298003
GNAT	2	B	0.816529564	0.767626887
MS.11.1	2	B	0.034675437	0.00215426
MS.21.1	2	B	0.437317369	0.385135243
BAT25	3	B	1.629828318	0.478147605
GNAT	3	B	0.959709717	0.952157213
MS.11.1	3	B	0.028750666	0.001376396
MS.21.1	3	B	0.200640004	0.102565016
BAT25	4	B	7.278209634	0.006822434
MS.11.1	4	B	0.03878704	0.002810799
MS.11.1	5	B	0.061016215	0.008091691
MS.11.1	6	B	0.02484157	0.000968192
BAT25	1	A	3.02136691	0.115060971
GNAT	1	A	0.300187716	0.087627839
GT23	1	A	2.180148051	0.039847612
MS.10.1	1	A	1.794863861	0.211487385
MS.11.1	1	A	6.04734176	0.005890824
MS.21.1	1	A	5.009211792	0.006584751
BAT25	2	A	0.949020826	0.939151236
GNAT	2	A	0.298619757	0.086328269
GT23	2	A	1.456597188	0.316784346
MS.10.1	2	A	2.089612921	0.116870231
MS.11.1	2	A	1.183059249	0.785348098
MS.21.1	2	A	1.881562855	0.267458264
BAT25	3	A	0.896757427	0.873710087
GNAT	3	A	0.247367721	0.049372885
GT23	3	A	2.060195742	0.056227015
MS.10.1	3	A	0.358176031	0.030520912
MS.11.1	3	A	0.371301268	0.115100147
MS.21.1	3	A	1.7198411	0.340459352
BAT25	4	A	2.63592155	0.164966524
GT23	4	A	1.610278191	0.205559438
MS.10.1	4	A	1.007585909	0.987030567
MS.11.1	4	A	0.868286166	0.818976664
GT23	5	A	1.47114654	0.304239528
MS.11.1	5	A	0.506560007	0.2746497
GT23	6	A	1.692691268	0.162388698
MS.11.1	6	A	3.32393671	0.058192121

Table A44: F Test for Variance Comparison by Cluster

Microsatellite	Cell.line	Allele	P Value
BAT25	2	1	0.012674149
BAT25	2	2	0.058691898
GT23	2	4	0.012387279
GT23	2	5	0.072504019
GNAT	3	3	0.074524779
MS.21.1	3	2	0.099746573
MS.21.1	3	3	0.07776738
GNAT	6	1	0.023002864
GNAT	6	2	0.023002864
GNAT	6	3	0.023002864
GNAT	7	2	0.083405626
GT23	9	3	0.072477286
GT23	9	6	0.072477286
MS.11.1	9	1	0.056022967
MS.11.1	9	2	0.056022967
MS.11.1	9	3	0.056022967
MS.11.1	9	4	0.040308138
MS.11.1	9	5	0.056022967
MS.11.1	9	6	0.04256647
MS.11.1	10	6	0.057723343

Table A45: TTEST Results ($p < 0.1$)

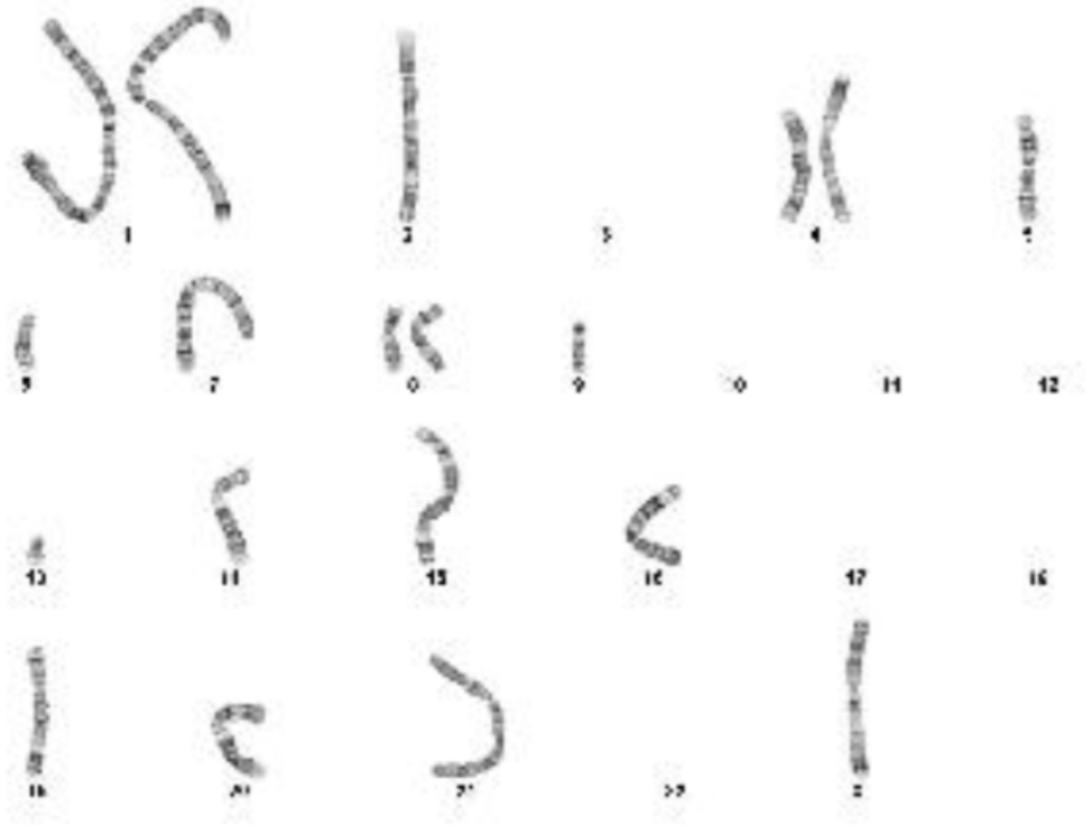
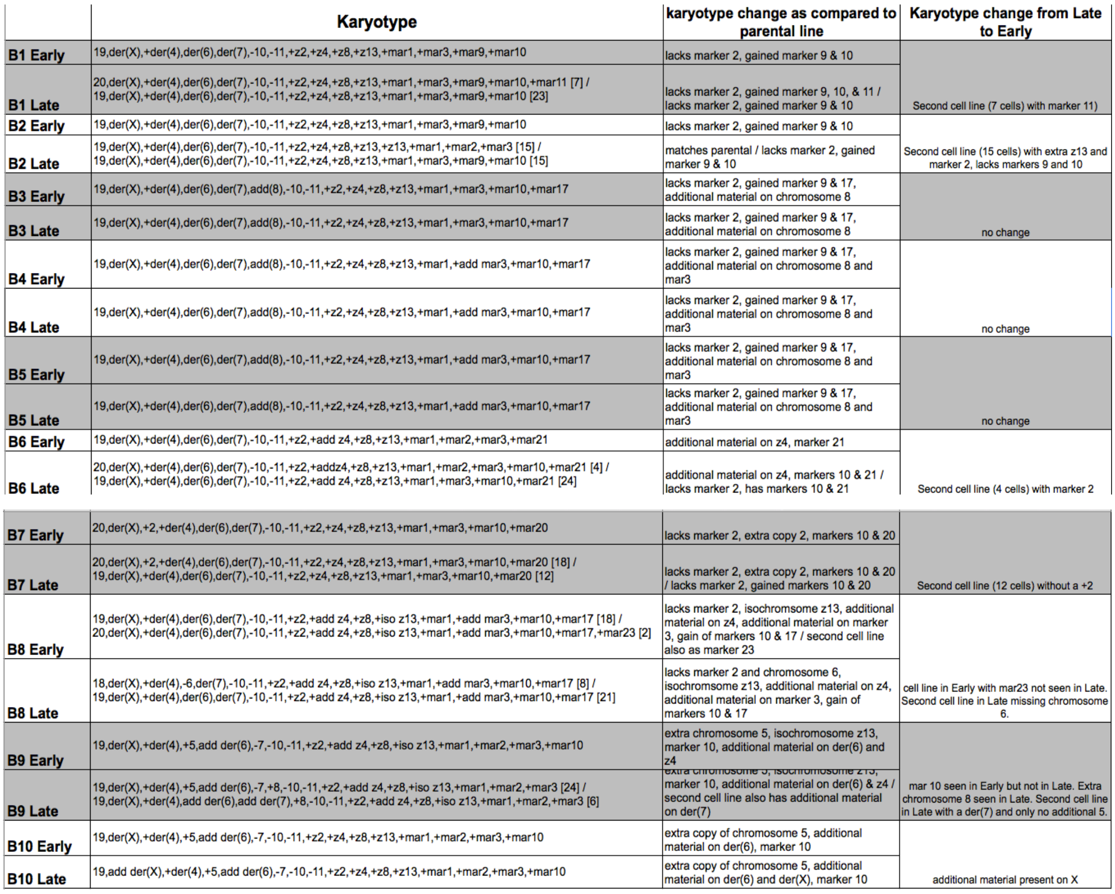


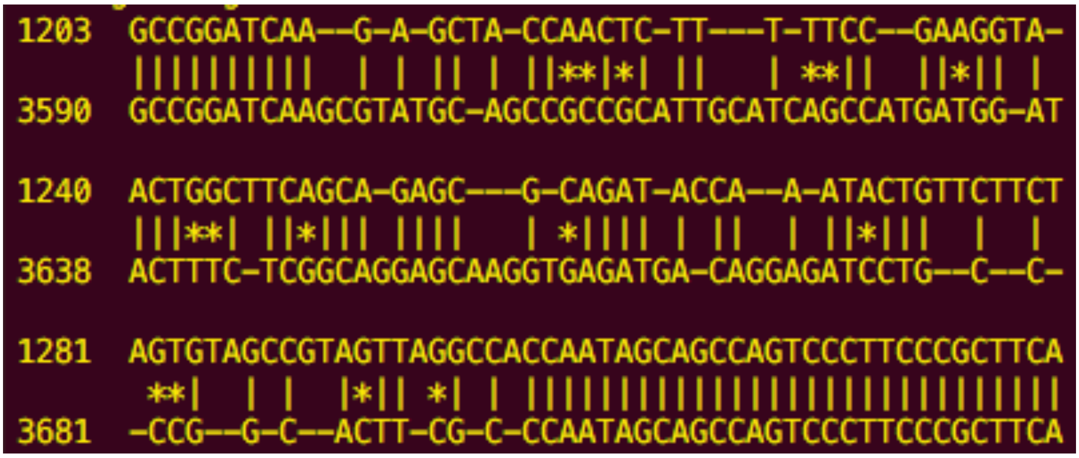
Figure A33: Parental Cell Line Karyotype
Karyotype:
19: 1,+1,2,4,5,8,9, der(X), +der(4), der(6), der(7), +der(8),+z13,+z4, +z8, +z2, +Mar1, +Mar2, +Mar3.



	Karyotype	karyotype change as compared to parental line	Karyotype change from Late to Early
B1 Early	19,der(X),+der(4),der(6),der(7),-10,-11,+z2,+z4,+z8,+z13,+mar1,+mar3,+mar9,+mar10	lacks marker 2, gained marker 9 & 10	Second cell line (7 cells) with marker 11)
B1 Late	20,der(X),+der(4),der(6),der(7),-10,-11,+z2,+z4,+z8,+z13,+mar1,+mar3,+mar9,+mar10,+mar11 [7] / 19,der(X),+der(4),der(6),der(7),-10,-11,+z2,+z4,+z8,+z13,+mar1,+mar3,+mar9,+mar10 [23]	lacks marker 2, gained marker 9, 10, & 11 / lacks marker 2, gained marker 9 & 10	
B2 Early	19,der(X),+der(4),der(6),der(7),-10,-11,+z2,+z4,+z8,+z13,+mar1,+mar3,+mar9,+mar10	lacks marker 2, gained marker 9 & 10	Second cell line (15 cells) with extra z13 and marker 2, lacks markers 9 and 10
B2 Late	19,der(X),+der(4),der(6),der(7),-10,-11,+z2,+z4,+z8,+z13,+z13,+mar1,+mar2,+mar3 [15] / 19,der(X),+der(4),der(6),der(7),-10,-11,+z2,+z4,+z8,+z13,+mar1,+mar3,+mar9,+mar10 [15]	matches parental / lacks marker 2, gained marker 9 & 10	
B3 Early	19,der(X),+der(4),der(6),der(7),add(8),-10,-11,+z2,+z4,+z8,+z13,+mar1,+mar3,+mar10,+mar17	lacks marker 2, gained marker 9 & 17, additional material on chromosome 8	no change
B3 Late	19,der(X),+der(4),der(6),der(7),add(8),-10,-11,+z2,+z4,+z8,+z13,+mar1,+mar3,+mar10,+mar17	lacks marker 2, gained marker 9 & 17, additional material on chromosome 8	
B4 Early	19,der(X),+der(4),der(6),der(7),add(8),-10,-11,+z2,+z4,+z8,+z13,+mar1,+add mar3,+mar10,+mar17	lacks marker 2, gained marker 9 & 17, additional material on chromosome 8 and mar3	no change
B4 Late	19,der(X),+der(4),der(6),der(7),add(8),-10,-11,+z2,+z4,+z8,+z13,+mar1,+add mar3,+mar10,+mar17	lacks marker 2, gained marker 9 & 17, additional material on chromosome 8 and mar3	
B5 Early	19,der(X),+der(4),der(6),der(7),add(8),-10,-11,+z2,+z4,+z8,+z13,+mar1,+add mar3,+mar10,+mar17	lacks marker 2, gained marker 9 & 17, additional material on chromosome 8 and mar3	no change
B5 Late	19,der(X),+der(4),der(6),der(7),add(8),-10,-11,+z2,+z4,+z8,+z13,+mar1,+add mar3,+mar10,+mar17	lacks marker 2, gained marker 9 & 17, additional material on chromosome 8 and mar3	
B6 Early	19,der(X),+der(4),der(6),der(7),-10,-11,+z2,+add z4,+z8,+z13,+mar1,+mar2,+mar3,+mar21	additional material on z4, marker 21	Second cell line (4 cells) with marker 2
B6 Late	20,der(X),+der(4),der(6),der(7),-10,-11,+z2,+addz4,+z8,+z13,+mar1,+mar2,+mar3,+mar10,+mar21 [4] / 19,der(X),+der(4),der(6),der(7),-10,-11,+z2,+add z4,+z8,+z13,+mar1,+mar3,+mar10,+mar21 [24]	additional material on z4, markers 10 & 21 / lacks marker 2, has markers 10 & 21	
B7 Early	20,der(X),+2,+der(4),der(6),der(7),-10,-11,+z2,+z4,+z8,+z13,+mar1,+mar3,+mar10,+mar20	lacks marker 2, extra copy 2, markers 10 & 20	Second cell line (12 cells) without a +2
B7 Late	20,der(X),+2,+der(4),der(6),der(7),-10,-11,+z2,+z4,+z8,+z13,+mar1,+mar3,+mar10,+mar20 [18] / 19,der(X),+der(4),der(6),der(7),-10,-11,+z2,+z4,+z8,+z13,+mar1,+mar3,+mar10,+mar20 [12]	lacks marker 2, extra copy 2, markers 10 & 20 / lacks marker 2, gained markers 10 & 20	
B8 Early	19,der(X),+der(4),der(6),der(7),-10,-11,+z2,+add z4,+z8,+iso z13,+mar1,+add mar3,+mar10,+mar17 [18] / 20,der(X),+der(4),der(6),der(7),-10,-11,+z2,+add z4,+z8,+iso z13,+mar1,+add mar3,+mar10,+mar17,+mar23 [2]	lacks marker 2, isochromsome z13, additional material on z4, additional material on marker 3, gain of markers 10 & 17 / second cell line also as marker 23	cell line in Early with mar23 not seen in Late. Second cell line in Late missing chromosome 6.
B8 Late	18,der(X),+der(4),-6,der(7),-10,-11,+z2,+add z4,+z8,+iso z13,+mar1,+add mar3,+mar10,+mar17 [8] / 19,der(X),+der(4),der(6),der(7),-10,-11,+z2,+add z4,+z8,+iso z13,+mar1,+add mar3,+mar10,+mar17 [21]	lacks marker 2 and chromosome 6, isochromsome z13, additional material on z4, additional material on marker 3, gain of markers 10 & 17	
B9 Early	19,der(X),+der(4),+5,add der(6),-7,-10,-11,+z2,+add z4,+z8,+iso z13,+mar1,+mar2,+mar3,+mar10	extra chromosome 5, isochromosome z13, marker 10, additional material on der(6) and z4	mar 10 seen in Early but not in Late. Extra chromosome 8 seen in Late. Second cell line in Late with a der(7) and only no additional 5.
B9 Late	19,der(X),+der(4),+5,add der(6),-7,+8,-10,-11,+z2,+add z4,+z8,+iso z13,+mar1,+mar2,+mar3 [24] / 19,der(X),+der(4),add der(6),add der(7),+8,-10,-11,+z2,+add z4,+z8,+iso z13,+mar1,+mar2,+mar3 [6]	extra chromosome 5, isochromosome z13, marker 10, additional material on der(6) & z4 / second cell line also has additional material on der(7)	
B10 Early	19,der(X),+der(4),+5,add der(6),-7,-10,-11,+z2,+z4,+z8,+z13,+mar1,+mar2,+mar3,+mar10	extra copy of chromosome 5, additional material on der(6), marker 10	additional material present on X
B10 Late	19,add der(X),+der(4),+5,add der(6),-7,-10,-11,+z2,+z4,+z8,+z13,+mar1,+mar2,+mar3,+mar10	extra copy of chromosome 5, additional material on der(6) and der(X), marker 10	

Table A46: List of Karyotypes.

Table containing all the karyotypes seen in the investigation and which cell line / generation they were in. The 'karyotype' column presents chromosomes that differ from wild type hamster. In cases where there are subpopulations of a cell line with differing karyotypes, both are listed with a '/' to separate them. Numbers at the start of each karyotype refer to the number of chromosomes there are. Numbers in square brackets show how many cells had a particular karyotype. The 'karyotype change as compared to parental line' column shows how cell lines differed from the standard CHO karyotype. The 'karyotype change from late to early' column shows differences between early and late generation cell lines.



```
1203  GCCGGATCAA--G-A-GCTA-CCAACTC-TT---T-TTCC--GAAGGTA-
      ||||||||||  | | || | ||**|*| ||   | **||  ||*|| |
3590  GCCGGATCAAGCGTATGC-AGCCGCCGCATTGCATCAGCCATGATGG-AT

1240  ACTGGCTTCAGCA-GAGC---G-CAGAT-ACCA--A-ATACTGTTCTTCT
      |||**| ||*||| ||||   | *|||| | ||  | ||*|||  |  |
3638  ACTTTC-TCGGCAGGAGCAAGGTGAGATGA-CAGGAGATCCTG--C--C-

1281  AGTGTAGCCGTAGTTAGGCCACCAATAGCAGCCAGTCCCTTCCCGCTTCA
       **|  | |  |*|| *| | |||||||||||||||||||||||||||||
3681  -CCG--G-C--ACTT-CG-C-CCAATAGCAGCCAGTCCCTTCCCGCTTCA
```

Figure A34: Example Heavily Mutated ROI Region
When ROIs were found to contain many mismatches, generally these matches were found to be within error-prone regions, containing both mismatches and indels. It was assumed that this phenomenon was likely to be due to sequencing error in a given ZMW and so a threshold filter was imposed, whereby ROIs with >3 mismatches were eliminated from analysis.

Figure A35: R scripts for the SMRT secondary analysis platform. # precedes explanatory information or acts to split up separate script phases. Red script refers to file names or directory names unique to this study and should be changed when using a different computer or file name.

Figure A35a: The following R script is used to convert BLASR Human readable format output into a useable CSV file containing information regarding query sequence, matches, target sequence, ROI name and positional information.

```
##########################################################################
##########################################################################
# This Script converts BLASR output into readable csv file
# Inputs: 1) CSV file the raw output from BLASR with header limited to Query,
QueryRange, TargetRange
# Output: 1) CSV named in input 2 that contains the query header information and
concatonated
# sequences for target query and match.
##########################################################################
##########################################################################

##########################################################################
# Set working directory and clear the working directory and load required packages.
setwd("/Users/josephcartwright/Google Drive/shared folder-JLongworth &
JCartwright/R scripts")
rm (list=ls())
##########################################################################

##########################################################################
# Load in inputs and set input variables
data=read.csv("BLASR CSVs/High_20pass.csv",head=F)
output_name="processed_data/New CSVs/new_High_20_800_2.csv"
##########################################################################

##########################################################################
#Set some empty variables and counting variables used in out for loop
Query_list=c()
QueryRange_list=c()
TargetRange_list=c()
Query_seq_list=c()
Match_seq_list=c()
Target_seq_list=c()
Full_Query_seq_list=c()
Full_Match_seq_list=c()
Full_Target_seq_list=c()
l=1
Query_count=0
Current_Query_count=0
##########################################################################
```

```
###########################################################################
# Initiate for loop taking data line by line. Assuming the start of a sequence entry extract
the header information into variable lists. Skipping though the next two header lines
with an empty else if recognise the start of the sequence section by setting query count
being out of sync and set l to 1 to identify the start of the three lines of sequence match
information.
for(i in 1:nrow(data)){
   if (grepl("Query:",data[i,])){
   Query_list=c(Query_list,as.character(data[i,]))
   QueryRange_list=c(QueryRange_list,as.character(data[(i+1),]))
   TargetRange_list=c(TargetRange_list,as.character(data[(i+2),]))
   Query_count=Query_count+1
  }else if(grepl("QueryRange:",data[i,])){
  }else if(grepl("TargetRange:",data[i,])){
  }else if(l==1 & Query_count!=Current_Query_count){
###########################################################################
```

```
###########################################################################
# Add the concatenated sequence derived from the previous sequence entry to the list of
all sequence entries and empty these variables.
   Full_Query_seq_list=c(Full_Query_seq_list,Query_seq_list)
   Full_Match_seq_list=c(Full_Match_seq_list,Match_seq_list)
   Full_Target_seq_list=c(Full_Target_seq_list,Target_seq_list)
   Query_seq_list=c()
   Match_seq_list=c()
   Target_seq_list=c()
###########################################################################
```

```
###########################################################################
# Increase Current Query count ready to identify next sequence entry. Using Reg
expression to identify the start point of the sequence information rather than the numeric
point as this varies with the number whether the total sequence length is greater than
1000 or less. Last two lines of this section fix an error occuring if there are none of any
of the possible start point characters.
   Current_Query_count=Current_Query_count+1
   first_character=c(regexpr("A",as.character(data[i,])),
                regexpr("C",as.character(data[i,])),
                regexpr("T",as.character(data[i,])),
                regexpr("G",as.character(data[i,])),
                regexpr("-",as.character(data[i,])))
   first_character[first_character==-1]=NA
   first_character=min(first_character,na.rm = T)
###########################################################################
```

```
###########################################################################
# Having prepped for concatonation of the three sequences the first sequence is added to
the appropriate variable then progresses on through the else if for the rest of the three
sequence as Query_count was set to equal current query count. This continues until a
new sequence entry is recognised by the header information.
```

```
Query_seq_list=paste(Query_seq_list,substring(as.character(data[i,]),first_character),se
p="")
            l=2
  }else if(l==1 &
Query_count==Current_Query_count){Query_seq_list=paste(Query_seq_list,
                              substring(as.character(data[i,]),first_character),sep="")
             l=2
  }else if(l==2){Match_seq_list=paste(Match_seq_list,
                              substring(as.character(data[i,]),first_character),sep="")
            l=3
  }else if(l==3){Target_seq_list=paste(Target_seq_list,
                               substring(as.character(data[i,]),first_character),sep="")
            l=1
  }}
############################################################################

############################################################################
#Add the final concatenated sequence to the list of sequneces.
Full_Query_seq_list=c(Full_Query_seq_list,Query_seq_list)
Full_Match_seq_list=c(Full_Match_seq_list,Match_seq_list)
Full_Target_seq_list=c(Full_Target_seq_list,Target_seq_list)
############################################################################

############################################################################
# Bind together all the extracted data into a dataframe and store as named input 2
new_data=cbind(Query_list,QueryRange_list,TargetRange_list,Full_Query_seq_list,Ful
l_Match_seq_list,
            Full_Target_seq_list)
write.csv(new_data,output_name)
############################################################################
```

Figure A35b: The following R script converts CSV information generated by the script in figure A27b into three matrices for sequence, match and quality information, respectively. The matrices contain information for individual nucleotides in individual matrix cells.

```
########################################################################
########################################################################
# This Script creates a binary system for mismatches and aligns to whole plasmid
sequence so that mutations can be counted at each position and creates a query sequence
matrix for base calling
# Inputs: 1) Formatted output from 'alignment conversion to fasta' R script (Figure
A27a)
#           2) List of Target strand orientations for each query sequence
#           3) SAM output from BLASR, containing Quality score information
# Output: 1) CSV named in input 3 that contains the match/mismatch matrix
#            2) CSV named in input 3 that contains the sequence matrix
#            3) CSV named in input 3 that contains the Quality matrix
########################################################################
########################################################################

########################################################################
# Set working directory and clear the working directory and load required packages.
setwd("/Users/josephcartwright/Google Drive/shared folder-JLongworth &
JCartwright/R scripts")

rm (list=ls())
########################################################################

########################################################################
# Load in inputs and set input variables & bind target strand data to main dataframe
data.1=read.csv("processed_data/New CSVs/new_Ecoli_15_800_2.csv")
data2=read.csv("Target CSVs/Ecoli_15passTARGET.csv", header=F)
data3=cbind(data.1,data2[,3])
names(data3)=c(names(data.1),"Target_Strand")
FASTQ=read.delim("FASTQ BLASR/Ecoli_15_Q",header = F,quote="")
FASTQ=FASTQ[5:nrow(FASTQ),]
data=cbind(data3,FASTQ[,11])
names(data)=c(names(data3),"Q-scores")
output_name="processed_data/Matrix CSVs/Ecoli_15/matrix_Ecoli_15_800_2.csv"
output_name2="processed_data/Matrix
CSVs/Ecoli_15/QUERY_matrix_Ecoli_15_800_2.csv"
output_name3="processed_data/Matrix
CSVs/Ecoli_15/Q_matrix_Ecoli_15_800_2.csv"
########################################################################

########################################################################
# Taking the data line by line create variables a and b containing the sequence for the
'Query' and 'Match' respectively. These are then separated into strings with one base
reported per element
# If query sequence is in opposite orientation it is converted into the proper orientation
and
```

```
# complement and match data is reversed
Perc=seq(from=1, to=nrow(data),by=100)
Perc2=c(Perc,nrow(data))
i=1
j=1
full_matrix=matrix(,nrow=nrow(data),ncol=4965)
full_matrix2=matrix(,nrow=nrow(data),ncol=4965)
full_matrix3=matrix(,nrow=nrow(data),ncol=4965)
matrix=matrix(,nrow=1,ncol=4965)
matrix2=matrix(,nrow=1,ncol=4965)
matrix3=matrix(,nrow=1,ncol=4965)
for(i in 1:nrow(data)){
  a=as.character(data[i,5])
  b=as.character(data[i,6])
  c=as.character(data[i,9])
  if (data[i,8]==0){
    a=as.character(substring(a,c(1:nchar(a)),c(1:nchar(a))))
    b=as.character(substring(b,c(1:nchar(b)),c(1:nchar(b))))
    c=as.character(substring(c,c(1:nchar(c)),c(1:nchar(c))))}
  if (data[i,8]==1){
    a=as.character(rev(substring(a,c(1:nchar(a)),c(1:nchar(a)))))
    b=as.character(rev(substring(b,c(1:nchar(b)),c(1:nchar(b)))))
    c=as.character(substring(c,c(1:nchar(c)),c(1:nchar(c))))}
  if (data[i,8]==1){
    a=(unname(sapply(a, switch,  "A"="T.", "T"="A.","G"="C.","C"="G.","-"="-")))}
  if (data[i,8]==0){
    a=unname(sapply(a, switch,  "A"="A.", "T"="T.","G"="G.","C"="C.","-"="-"))}

###########################################################################

###########################################################################
# having created stings for each sequence this loop looks at each element in the
sequence and asks the question is it a match thus scoring 0, a mismatch scoring 1, a
deletion scoring NA, or a insertion scoring nothing (to avoid matrix misalignment) this
score is built up into x_list.
# The same information is used to create y_list and z_list, containing the query
sequence without insertions.
x_list=c()
for(n in 1:  length(b)){
  #i=1}
  if (b[n]=="|"){x=0
  }  else if (b[n]=="*"){x=1
  }  else if (b[n]==" "&a[n]=="-"){x=NA
  }  else if (b[n]==" "&a[n]!="-"){x="X"}

  if (is.na(x)){x_list=c(x_list,x)
  }  else if (x==0|x==1){
  x_list=c(x_list,x)}
}
```

```
 y_list=c()
 for(n in 1:  length(b)){
   #i=1}
   if (b[n]=="|"){y=(a[n])
   }  else if (b[n]=="*"){y=(a[n])
   }  else if (b[n]==" "& a[n]=="-"){y=NA
   }  else if (b[n]==" "& a[n]!="-"){y="Y"}

   if (is.na(y)){y_list=c(y_list,y)
   }  else if (y=="A."|y=="T."|y=="C."|y=="G."){
     y_list=c(y_list,y)}
 }

 z_list=c()
 k=1
 for(n in 1:  length(b)){
   if (b[n]=="|"){z=(c[k]);k=k+1
   }  else if (b[n]=="*"){z=(c[k]);k=k+1
   }  else if (b[n]==" "& a[n]=="-"){z=NA
   }  else if (b[n]==" "& a[n]!="-"){z="NULL"}

   if (is.na(z)){z_list=c(z_list,z)
   }  else if (z!="NULL"){
     z_list=c(z_list,z)}
 }
###############################################################################

###############################################################################
# Start and end positions relative to the target are calculated for the query sequence and
matches.  This determines the number of NA's to be added to the front and the end of
the sequence to give full length comparable to the full target sequence. The full
sequence length is made of 'head' NA's at the start of the sequence, the sequence itself,
and then 'tail' NA's until the end of the sequence. This is carried out for matches and
query sequences to generate full matrices for all match and sequencing data

start=c()
end=c()
 if (data[i,8]==0){
  start=as.numeric(strsplit(as.character(data[i,4]),"  ")[[1]][4])
  end=as.numeric(strsplit(as.character(data[i,4]),"  ")[[1]][6])}
 if (data[i,8]==1){
   start=4965-as.numeric(strsplit(as.character(data[i,4]),"  ")[[1]][6])
   end=4965-as.numeric(strsplit(as.character(data[i,4]),"  ")[[1]][4])}


head_NA=start
tail_NA=4965-end
matrix=c(rep(NA,head_NA),x_list,rep(NA,tail_NA))
matrix2=c(rep(NA,head_NA),as.character(y_list),rep(NA,tail_NA))
matrix3=c(rep(NA,head_NA),as.character(z_list),rep(NA,tail_NA))
```

```
full_matrix[i,]=matrix
full_matrix2[i,]=matrix2
full_matrix3[i,]=matrix3

if (i==Perc2[j]){
  print(i/nrow(data)*100,digits = 3);j=j+1}
}

full_data_frame=as.data.frame(cbind(data[,2],full_matrix))
full_data_frame2=as.data.frame(cbind(data[,2],full_matrix2))
full_data_frame3=as.data.frame(cbind(data[,2],full_matrix3))
###############################################################################

###############################################################################
# Save the compiled matrices to csv files
write.csv(full_data_frame,output_name)
write.csv(full_data_frame2,output_name2)
write.csv(full_data_frame3,output_name3)
###############################################################################
```

Figure A35c: The following R script removes error-prone ROIs from the analysis pipeline and provides a preliminary analysis on the extent of fragment mutation

```
############################################################################
############################################################################
# This Script creates a modified matrices with removed error prone fragments and
analyses the number of fragments that are mutated
# Inputs: 1) Sequence matrix
#             2) Match/mismatch matrix
#             3) Quality matrix
# Output: 1) Sequence matrix with error-prone fragments removed
#              2) Match/mismatch matrix with error-prone fragments removed
#              3) Quality matrix with error-prone fragments removed
#              4) Statistics regarding the frequency of mutations in the fragments
############################################################################
############################################################################

############################################################################
# Set working directory and clear the working directory and load required packages.
setwd("/Users/josephcartwright/Google Drive/shared folder-JLongworth &
JCartwright/R scripts")
rm (list=ls())
############################################################################

############################################################################
# Load in inputs and set input variables
data=read.csv("processed_data/Matrix CSVs/High_10/matrix_High_10_800_2.csv")
Query.data=read.csv("processed_data/Matrix
CSVs/CHO_10/QUERY_matrix_CHO_10_800_2.csv")
Q.data=read.csv("processed_data/Matrix
CSVs/CHO_10/Q_matrix_CHO_10_800_2.csv")
output_name="processed_data/Matrix CSVs/CHO_10/matrix_CHO_10_800_3.csv"
output_name2="processed_data/Matrix
CSVs/CHO_10/Query_matrix_CHO_10_800_3.csv"
output_name3="processed_data/Matrix
CSVs/CHO_10/Q_matrix_CHO_10_800_3.csv"
output_name4="processed_data/Mutated Fragment
Data/MUTFRAG_CHO_10_800_2.csv"
############################################################################

############################################################################
#Remove unwanted columns
# Remove all rows whose sum is greater than 3 - fragments seem to show sequencing
error
data2=data[,-c(1,2)]
rSUMS=rowSums(data2,na.rm=T)
data3=cbind(data,rSUMS)
data4=data3[data3[,4968]<=3,]
data5=data4[,-4968]

Query.data2=Query.data[,-c(1,2)]
Query.data3=Query.data2[data3[,4968]<=3,]

Q.data2=Q.data[,-c(1,2)]
Q.data3=Q.data2[data3[,4968]<=3,]
############################################################################

############################################################################
# Write altered dataset to csv
write.csv(data5,output_name)
write.csv(Query.data3,output_name2)
write.csv(Q.data3,output_name3)
############################################################################

############################################################################
# View deleted fragments in terms of number of detected mutations
delsum=sort(data3[,4968], decreasing = T)
head(delsum, n = 10)
############################################################################

############################################################################
# calculate number mutated fragments, their % of total and the number of fragments with 1,2 or 3 mutations. Bind together.
Mutated_Fragments=nrow(data4[data4[,4968]>0,])
Percent_Mutated_Fragments=Mutated_Fragments/nrow(data4)*100
Mutated_Fragments1=nrow(data4[data4[,4968]==1,])
Mutated_Fragments2=nrow(data4[data4[,4968]==2,])
Mutated_Fragments3=nrow(data4[data4[,4968]==3,])
MUTFRAG=cbind(Mutated_Fragments,Percent_Mutated_Fragments,Mutated_Fragments1,Mutated_Fragments2,Mutated_Fragments3)
############################################################################

############################################################################
# write csv for mutated fragment data
write.csv(MUTFRAG,output_name4)
```

Figure A35d: The following R script generates statistics and plots for mutation frequency along the length of the plasmid sequence. It also calculates nucleotide coverage

```
############################################################################
############################################################################
# This Script generates statistics and plots for plasmid mutation frequency and plasmid
position
# Inputs: 1) Match/mismatch matrix
# Output: 1) Table containing number of mutations and coverage at each target position.
#            2) Statistics regarding plasmid mutation
#            3) Plots for plasmid mutation and coverage
############################################################################
###########################################################################

############################################################################
# Set working directory and clear the working directory and load required packages.
setwd("/Users/josephcartwright/Google Drive/shared folder-JLongworth &
JCartwright/R scripts")
rm (list=ls())
############################################################################

############################################################################
# Load in inputs and set input variables
data=read.csv("processed_data/Matrix CSVs/High_10/matrix_High_10_800_3.csv")
output_name="processed_data/Mutation Tables/mutations_MOD_High_10_800_2.csv"
output_name2="processed_data/Mutated Plasmid
Data/MUTPLAS_MOD_High_10_800_2.csv"
output_name3="processed_data/Mutated Plasmid Data/Av_COV_High_10_800_2.csv"
PDFPath = "/Users/josephcartwright/Google Drive/shared folder-JLongworth &
JCartwright/R scripts/processed_data/Plots/plots_MOD_High_10_800_2.pdf"
############################################################################

############################################################################
#Remove unwanted columns
data2=data[,-c(1:3)]
############################################################################

############################################################################
# Calculate the sum for all base pairs across the plasmid length.
# Create data frame of sums where 0 = NA for clarity in plots
# Create subset dataframe of this, only including obserbed mutations
# Calculate the Coverage at each base pair position
SUMS=colSums(data2,na.rm=T)

COVER=c()
for (i in 1:ncol(data2)){
  x=data2[,i]
  cove=length(na.omit(x))
  COVER=c(COVER,cove)}
Av.cov=mean(COVER)
```

```
write.csv(Av.cov,output_name3)
plasmid=c(1:4965)
MFREQ=as.data.frame(cbind(SUMS,plasmid,COVER))
MFREQ[MFREQ==0]=NA

Mutations=as.data.frame(MFREQ[complete.cases(MFREQ),])
names(Mutations)=c("Mutations","Base number","Coverage")

Coverage_T_Test=t.test(MFREQ[,3],Mutations[,3])
Coverage_Mean_Difference=mean(MFREQ[,3],na.rm = T)-mean(Mutations[,3])
#############################################################

#############################################################
#calculate minimum and maximums of SUMS and COVER datasets to establish y axis
limits for plots
max(SUMS)
min(SUMS)
max(COVER)
min(COVER)
head(sort(MFREQ[,1],decreasing=T))
#############################################################

#############################################################
# Plot coverage and sums and write out to pdf
# For sum plots create one for overall and another for lower frequencies (higher
resolution plot)
# Write mutations to csv
pdf(file=PDFPath)
par(mfrow=c(1,1))
plot(COVER,pch="*",ylim=c(0,11000),ylab="Base Coverage",xlab="Base Pair
Number")
plot(MFREQ[,1]~MFREQ[,2],pch="*",ylim=c(1,7000),ylab="Mutation
Frequency",xlab="Base Pair Number")
plot(MFREQ[,1]~MFREQ[,2],pch="*",ylim=c(1,30),ylab="Mutation
Frequency",xlab="Base Pair Number")
dev.off()
write.csv(Mutations,output_name)
#############################################################

#############################################################
# Calculate the number of mutated positions of plasmid
# Does coverage differ between mutated positions compared to all positions
# Normalise number of mutated positions by sequence coverage
# Write this data to csv
Mutated_Positions=nrow(Mutations)
Coverage_Mutation_Significance=Coverage_T_Test$p.value
Normalised_Mutated_Positions=Mutated_Positions/mean(na.omit(MFREQ$COVER))
MUTPLAS=cbind(Mutated_Positions,Coverage_Mean_Difference,Coverage_Mutation
_Significance,Normalised_Mutated_Positions)
write.csv(MUTPLAS,output_name2)
```

Figure A35e: The following R script generates an annotated list of mutations with mutation frequencies for all, Q score filtered and >1 filtered mutation sets.

```
##########################################################################
##########################################################################
# This Script provides an annotated list of observed mutations for all, Q score filtered
and >1 filtered data
# Inputs: 1) sequence matrix
#          2) Quality matrix
#          3) GFP target sequence
#          4) Quality character --> Phred score key
#          5) List of mutations from mutation frequency table
#          6) Mutation table generated in Analysis 2
# Output: 1) Base type count at each plasmid position and corresponding target
sequence
#           2) Updated Mutation table containing the base changes.
#           3) Updated Mutation table containing the base changes (Q score filtered).
#           4) Updated Mutation table containing the base changes (>1 filtered).
##########################################################################
##########################################################################

##########################################################################
# Set working directory and clear the working directory and load required packages.
setwd("/Users/josephcartwright/Google Drive/shared folder-JLongworth &
JCartwright/R scripts")
rm (list=ls())
##########################################################################

##########################################################################
# Load in inputs and set input variables
data=read.csv("processed_data/Matrix
CSVs/High_10/QUERY_matrix_High_10_800_3.csv")
data3=read.csv("processed_data/Matrix
CSVs/High_10/Q_matrix_High_10_800_3.csv")
GFP=read.csv("processed_data/GFP.csv",header = F, stringsAsFactors = F, colClasses
= c("character"))
Mutations=read.csv("processed_data/Mutation
Tables/mutations_MOD_High_10_800_2.csv")
FASTQ.CHAR=read.csv("FASTQ.VALUES.csv")
output_name="processed_data/Base counts/BASE_High_10_800.csv"
output_name2="processed_data/Mutation
Tables/FINALmutations_MOD_High_10_800_2.csv"
output_name3="processed_data/Mutation
Tables/FINALmutations_MOD.Q_High_10_800_2.csv"
output_name4="processed_data/Mutation
Tables/FINALmutations_MOD.Q_>1_High_10_800_2.csv"
PDFPath = "/Users/josephcartwright/Google Drive/shared folder-JLongworth &
JCartwright/R scripts/processed_data/Plots/plots_Q_High_10_800.pdf"
##########################################################################

##########################################################################
```

```
# remove information that is not sequence
data2=(data[,-1])
qmat=(data3[,-1])
rm(data)
rm(data3)
##########################################################################

##########################################################################
# Alter the quality score matrix, so that each quality character is replaced by
corresponding Phred score value

FASTQ.CHAR=as.matrix(FASTQ.CHAR)

QMAT2=as.matrix(qmat)
QMAT3=matrix(,nrow=nrow(qmat),ncol=ncol(qmat))
for (j in 1:nrow(QMAT2)){
  QMATM=QMAT2[j,]
  for (i in 1:nrow(FASTQ.CHAR)){
    x=FASTQ.CHAR[i,1]
    y=FASTQ.CHAR[i,2]
    QMATM=replace(QMATM,QMATM==x,y)
  }
  QMAT3[j,]=QMATM
  print(j)
}

### Anything with Phred score lower than 25 (99.5% accuracy) changed to NA
### Corresponding base in nucleotide matrix replaced with NA so it is not counted.

data3=as.matrix(data2)
QMAT4=QMAT3
QMAT4[QMAT4<25]=NA

for (i in 1:nrow(data3)){
  x=data3[i,]
  x[is.na(QMAT4[i,])]=NA
  data3[i,]=x
  print(i)
}
##########################################################################

##########################################################################
# Base_count table created from number of A's, T's, C's or G's at each postion, along
with the target sequence.
# Save this as CSV file
Adenosine=c()
Thymine=c()
Cytosine=c()
Guanine=c()
ad=c()
```

```
th=c()
gu=c()
cy=c()
for (i in 1:ncol(data2)){
    ad=as.data.frame(summary(as.factor(data3[,i])))["A.",1]
    th=as.data.frame(summary(as.factor(data3[,i])))["T.",1]
    cy=as.data.frame(summary(as.factor(data3[,i])))["C.",1]
    gu=as.data.frame(summary(as.factor(data3[,i])))["G.",1]
  Adenosine=cbind(Adenosine,ad)
  Thymine=cbind(Thymine,th)
  Cytosine=cbind(Cytosine,cy)
  Guanine=cbind(Guanine,gu)}

GFP=as.matrix(GFP[1,])

Base_counts=rbind(Adenosine,Thymine,Cytosine,Guanine,GFP)
rownames(Base_counts)=c("A","T","C","G","Seq")
colnames(Base_counts)=c(1:4965)

write.csv(Base_counts,output_name)
###########################################################################

###########################################################################
# Create an updated mutation table containing the target base changed and the base it
has changed to create 3 of these tables:
#         1) Containing all mutations observed
#         2) Mutations removed by quality filtering
#         3) Mutations occuring only once removed
# plots for these
# save to CSV

Basenames=as.numeric(colnames(Base_counts))

Targ=c()
Target=c()
for (i in 1:nrow(Mutations)){
  Targ=Base_counts[5,Basenames[Mutations[i,3]]]
  Target=rbind(Target,Targ)}
Mutations2=cbind(Mutations,Target[,1])

Empty.changes=matrix(0,nrow=nrow(Mutations2),ncol=4)
colnames(Empty.changes)=c("A","T","C","G")
for (i in 1:nrow(Mutations)){
  if(is.na(Base_counts[1,Basenames[Mutations[i,3]]])==F & Mutations2[i,5]!="A"){
Empty.changes[i,1]=as.numeric(Base_counts[1,Mutations[i,3]])}
  if(is.na(Base_counts[2,Basenames[Mutations[i,3]]])==F & Mutations2[i,5]!="T"){
Empty.changes[i,2]=as.numeric(Base_counts[2,Mutations[i,3]])}
  if(is.na(Base_counts[3,Basenames[Mutations[i,3]]])==F & Mutations2[i,5]!="C"){
Empty.changes[i,3]=as.numeric(Base_counts[3,Mutations[i,3]])}
```

```
  if(is.na(Base_counts[4,Basenames[Mutations[i,3]]])==F & Mutations2[i,5]!="G"){
Empty.changes[i,4]=as.numeric(Base_counts[4,Mutations[i,3]])}}

Mutations2=cbind(Mutations2,Empty.changes)
Mutations.2=rowSums(Mutations2[,c(6:9)])
Mutations2=cbind(Mutations2[,c(1:2)],Mutations.2,Mutations2[,c(3:9)])
Mutations3=Mutations2[,c(7:10)]
Mutations3[Mutations3==1]=0
Mutations.3=rowSums(Mutations3)
Mutations3=cbind(Mutations2[,c(1:3)],Mutations.3,Mutations2[,c(4:6)],Mutations3)
Mutations4=Mutations3[Mutations3[,3]>0,]
Mutations5=Mutations4[Mutations4[,4]>0,]

for (i in 1:nrow(Mutations3)){
  if (Mutations3[i,5]==2539)
    RM=i}
for (i in 1:nrow(Mutations4)){
  if (Mutations4[i,5]==2539)
    RM2=i}
for (i in 1:nrow(Mutations5)){
  if (Mutations5[i,5]==2539)
    RM3=i}

YLIM1=round(sort(Mutations3[,2],decreasing=T)[2]+5,-1)

pdf(file=PDFPath)
par(mfrow=c(1,1))
plot(Mutations3[-RM,2]~Mutations3[-RM,5],xlim=c(1,5000),ylim=c(0,YLIM1),xlab =
"Base Pair Number", ylab = "Mutation Frequency",pch="*")
plot(Mutations4[-RM2,3]~Mutations4[-RM2,5],xlim=c(1,5000),ylim=c(0,YLIM1),xlab
= "Base Pair Number", ylab = "Mutation Frequency",pch="*")
plot(Mutations5[-RM3,4]~Mutations5[-RM3,5],xlim=c(1,5000),ylim=c(0,YLIM1),xlab
= "Base Pair Number", ylab = "Mutation Frequency",pch="*")
dev.off()

write.csv(Mutations3,output_name2)
write.csv(Mutations4,output_name3)
write.csv(Mutations5,output_name4)
########################################################################
```

Figure A35f: The following R script calculates the amount of mutation in each element of the plasmid. It also calculates the proportion of each type of mutation that was observed. It then generates various statistics regarding mutation frequency and an overall mutation rate.

```
##########################################################################
##########################################################################
# This Script calculates the percentage of mutation that falls within each genetic
element of the plasmid sequence and calculates the number of each mutation type. Then
mutation information is normalised by the average coverage of the sample. Overall
mutation rates are then calculated.
# Inputs: 1) Mutation table for all mutations
#           2) Mutation table for Q score filtered mutations
#           3) Mutation table for >1 filtered mutations
#           4) Base counts table
#           5) Average coverage for the sample
# Output: 1) A table containing the percentage mutation of each plasmid genetic
element
#            2) A table containing the percentage of each mutation type
#            3) A table containing mutation frequency information
##########################################################################
##########################################################################

##########################################################################
# Set working directory and clear the working directory and load required packages.
setwd("/Users/josephcartwright/Google Drive/shared folder-JLongworth &
JCartwright/R scripts")
rm (list=ls())
##########################################################################

##########################################################################
# Load in inputs and set input variables
Mutation1=read.csv("processed_data/Mutation
Tables/FINALmutations_MOD_High_10_800_2.csv")
Mutation2=read.csv("processed_data/Mutation
Tables/FINALmutations_MOD.Q_High_10_800_2.csv")
Mutation3=read.csv("processed_data/Mutation
Tables/FINALmutations_MOD.Q_>1_High_10_800_2.csv")
Base_counts=read.csv("processed_data/Base counts/BASE_High_10_800.csv")
Av.cov=read.csv("processed_data/Mutated Plasmid
Data/Av_COV_High_10_800_2.csv")
output_name1="processed_data/Mutation Annotation/High_10_position.csv"
output_name2="processed_data/Mutation Annotation/High_10_Bases.csv"
output_name3="processed_data/Mutated Plasmid
Data/High_10_MUTPLAS_NEW.csv"
##########################################################################

##########################################################################
# Create dataframe with plasmid annotations
# Use dataframe to create a table regarding the plasmid positions percentages of
mutations
```

```
# Standardise these numbers by dividing by the number of bases in a given element
Emat=matrix(,nrow=4965,ncol=12)
colnames(Emat)=c("pAmp","pSV40","Kan/Neo","HSV_TK_PolyA","Puc_Ori","phCM
V_and_Intron","phCMV_and_Intron_MCS.1","pT7","MCS.1","GFP_ORF","MCS.2","
SV40_PolyA")
Emat[c(527:555),1]="pAmp";Emat[c(639:868),2]="pSV40";Emat[c(990:1784),3]="Ka
n/Neo";Emat[c(2020:2038),4]="HSV_TK_PolyA";Emat[c(2369:3012),5]="Puc_Ori";
Emat[c(3153:3838),6]="phCMV_and_Intron";Emat[c(3839:3852),7]="phCMV_and_In
tron_MCS.1";Emat[c(3853:3868),8]="pT7";Emat[c(3869:3902),7]="phCMV_and_Intro
n_MCS.1";
Emat[c(3903:3952),9]="MCS.1";Emat[c(3953:4672),10]="GFP_ORF";Emat[c(4673:47
42),11]="MCS.2";
Emat[c(4878:4928),12]="SV40_PolyA"

Element=matrix(,nrow=nrow(Mutation3),ncol=1)
for (i in 1:nrow(Mutation3)){
  x=Emat[Mutation3[i,6],]
  Y=c()
  for (j in 1:ncol(Emat)){
    y=c()
    if (is.na(x[j])==F){
      y=x[j]}
      Y=c(Y,y)}
  if (length(Y)==0){
    Y="Non-coding"}
  Element[i,]=Y}
Mutation.A=cbind(Mutation3,Element)
names(Mutation.A)=c(names(Mutation3),"Element")

Element_percentages=matrix(,nrow=3,ncol=13)
colnames(Element_percentages)=c("pAmp","pSV40","Kan/Neo","HSV_TK_PolyA","P
uc_Ori","phCMV_and_Intron","phCMV_and_Intron_MCS.1","pT7","MCS.1","GFP_O
RF","MCS.2","SV40_PolyA","Non-coding")

for (i in 1:ncol(Element_percentages)){
Element_percentages[1,i]=round(summary(Mutation.A$Element)[colnames(Element_p
ercentages)[i]]/nrow(Mutation.A)*100,digits=1)}

Element_percentages[1,][is.na(Element_percentages[1,])]=0
Elengths=matrix(,nrow=1,ncol=12)
for (i in 1:ncol(Elengths)){
  Elengths[i]=length(na.omit(Emat[,i]))}
NC=4965-sum(Elengths)
Elengths2=cbind(Elengths,NC)
Element_percentages[2,]=Elengths2

Enorm=Element_percentages[1,]/Element_percentages[2,]*1000
Element_percentages[3,]=Enorm
Element_percentages=round(Element_percentages,digits=2)
rownames(Element_percentages)=c("Percentage","Element_Bases","Normalised")
```

```
write.csv(Element_percentages,output_name1)
############################################################################

############################################################################
# Create a table illustrating the percentage mutation frequency of each base change type
Base_percentages=matrix(,nrow=4,ncol=5)
rownames(Base_percentages)=c("A","T","C","G")
colnames(Base_percentages)=c("A","T","C","G","Total")

Mutation_base=Mutation3[9:12]
Mutation_base[Mutation_base==0]=NA

Y=c()
for (i in 1:nrow(Mutation_base)){
  x=Mutation_base[i,]
  y=c()
  for (j in 1:length(x)){
  if (is.na(x[j])==F){
    y=colnames(Mutation_base)[j]}}
    Y=rbind(Y,y)}
colnames(Y)=c("Change")
Mutation3=cbind(Mutation3,Y[,1])

A_to_T=Mutation3[Mutation3[,8]=="A" &
Mutation3[,13]=="T",];Base_percentages[1,2]=round(nrow(A_to_T)/nrow(Mutation3)*
100,digits=2)
A_to_C=Mutation3[Mutation3[,8]=="A" &
Mutation3[,13]=="C",];Base_percentages[1,3]=round(nrow(A_to_C)/nrow(Mutation3)
*100,digits=2)
A_to_G=Mutation3[Mutation3[,8]=="A" &
Mutation3[,13]=="G",];Base_percentages[1,4]=round(nrow(A_to_G)/nrow(Mutation3)
*100,digits=2)
T_to_A=Mutation3[Mutation3[,8]=="T" &
Mutation3[,13]=="A",];Base_percentages[2,1]=round(nrow(T_to_A)/nrow(Mutation3)
*100,digits=2)
T_to_C=Mutation3[Mutation3[,8]=="T" &
Mutation3[,13]=="C",];Base_percentages[2,3]=round(nrow(T_to_C)/nrow(Mutation3)*
100,digits=2)
T_to_G=Mutation3[Mutation3[,8]=="T" &
Mutation3[,13]=="G",];Base_percentages[2,4]=round(nrow(T_to_G)/nrow(Mutation3)
*100,digits=2)
C_to_A=Mutation3[Mutation3[,8]=="C" &
Mutation3[,13]=="A",];Base_percentages[3,1]=round(nrow(C_to_A)/nrow(Mutation3)
*100,digits=2)
C_to_T=Mutation3[Mutation3[,8]=="C" &
Mutation3[,13]=="T",];Base_percentages[3,2]=round(nrow(C_to_T)/nrow(Mutation3)*
100,digits=2)
C_to_G=Mutation3[Mutation3[,8]=="C" &
Mutation3[,13]=="G",];Base_percentages[3,4]=round(nrow(C_to_G)/nrow(Mutation3)
*100,digits=2)
```

```
G_to_A=Mutation3[Mutation3[,8]=="G" &
Mutation3[,13]=="A",];Base_percentages[4,1]=round(nrow(G_to_A)/nrow(Mutation3)
*100,digits=2)
G_to_T=Mutation3[Mutation3[,8]=="G" &
Mutation3[,13]=="T",];Base_percentages[4,2]=round(nrow(G_to_T)/nrow(Mutation3)*
100,digits=2)
G_to_C=Mutation3[Mutation3[,8]=="G" &
Mutation3[,13]=="C",];Base_percentages[4,3]=round(nrow(G_to_C)/nrow(Mutation3)
*100,digits=2)

Base_percentages[1,5]=sum(na.omit(Base_percentages[1,c(1:4)]))
Base_percentages[2,5]=sum(na.omit(Base_percentages[2,c(1:4)]))
Base_percentages[3,5]=sum(na.omit(Base_percentages[3,c(1:4)]))
Base_percentages[4,5]=sum(na.omit(Base_percentages[4,c(1:4)]))
write.csv(Base_percentages,output_name2)
#########################################################################

#########################################################################
#Create a table summarising the mutation frequencies observed and normalise using
coverage.
Mutated_positions=nrow(Mutation1)
Mutated_positions_Q=nrow(Mutation2)
Mutated_positions_1=nrow(Mutation3)
Mutation_number_Q=sum(Mutation2$Mutations.2)
Mutation_number_1=sum(Mutation3$Mutations.3)
Mutated_positions_norm=Mutated_positions/Av.cov[1,2]
Mutated_positions_Q_norm=Mutated_positions_Q/Av.cov[1,2]
Mutated_positions_1_norm=Mutated_positions_1/Av.cov[1,2]
Mutation_number_Q_norm=Mutation_number_Q/Av.cov[1,2]
Mutation_number_1_norm=Mutation_number_1/Av.cov[1,2]
Plasmid_mutations=cbind(Mutated_positions,Mutated_positions_Q,Mutated_positions_
1,Mutated_positions_norm,Mutated_positions_Q_norm,Mutated_positions_1_norm)
write.csv(Plasmid_mutations,output_name3)
#########################################################################

#########################################################################
# Overall mutation rates
zxc=Base_counts[c(1:4),c(2:4966)]
z=0
for (i in 1:nrow(zxc)){
  x=zxc[i,]
  y=0
  for (j in 1:ncol(x)){
    if (is.na(as.numeric(as.character(x[1,j])))==F){
      y=y+as.numeric(as.character(x[1,j]))}}
  z=z+y}

Q_mut_rate=z/(sum(Mutation2$Mutations.2)-Mutation2[Mutation2[,6]==2539,4])
once_mut_rate=z/(sum(Mutation3$Mutations.3)-Mutation3[Mutation3[,6]==2539,4])
```

Figure A35f: The following R script calculates the number of synonymous and non-synonymous mutations in the GFP and Kan / Neo ORFs. It then calculates the general probability of a non-synonymous or synonymous mutation occurring.

```
#########################################################################
#########################################################################
# This Script calculates the percentage of synonymous and non-synonymous mutations.
It then simulates the raw probability of these occuring for comparison
# Inputs: 1) Base_counts table
#          2) Mutation table (>1 filtered)
#          3) Codon sequence key
#########################################################################
#########################################################################

#########################################################################
# Set working directory and clear the working directory and load required packages.
setwd("/Users/josephcartwright/Google Drive/shared folder-JLongworth &
JCartwright/R scripts")
rm (list=ls())
#########################################################################

#########################################################################
# Load in inputs and set input variables
Mutation1=read.csv("processed_data/Mutation
Tables/FINALmutations_MOD.Q_>1_High_10_800_2.csv")
Base_counts=read.csv("processed_data/Base counts/BASE_High_10_800.csv")
Codons=read.csv("Amino acid codons.csv",header = F)
names(Codons)=c("Amino_acid","1","2","3")
#########################################################################

#########################################################################
# Create dataframes for open reading frame positions
Kan_Neo=seq(990:1784)+989
GFP=seq(3953:4672)+3952

# Create mutation dataframe for Kan/Neo gene only
Kan_mut=c()
for (i in 1:length(Kan_Neo)){
  x=Mutation1[Mutation1[,6]==Kan_Neo[i],]
  Kan_mut=rbind(Kan_mut,x)}

# Create mutation dataframe for GFP gene only
GFP_mut=c()
for (i in 1:length(GFP)){
  x=Mutation1[Mutation1[,6]==GFP[i],]
  GFP_mut=rbind(GFP_mut,x)}

# Isolate ORF sequences
Plasmid=Base_counts[5,c(2:4966)]
Kan_seq=Plasmid[c(990:1784)]
GFP_seq=Plasmid[c(3953:4672)]
```

```
names(Kan_seq)=c(Kan_Neo[1:length(Kan_Neo)])
names(GFP_seq)=c(GFP[1:length(GFP)])
############################################################################

############################################################################
#KAN/NEO GENE
# Split Kan/Neo ORF into codons by row
Kan_pos=seq(0,length(Kan_seq)-1, by=3)
Kan_cod=matrix(,nrow=length(Kan_pos),ncol=3)
for (i in 1:length(Kan_pos)){
  x=as.matrix(Kan_seq[c((1+Kan_pos[i]):(3+Kan_pos[i]))])
  Kan_cod[i,]=x}
Kan_cod=as.data.frame(Kan_cod)

# Annotate each codon with amino acid it codes
b=c()
for (i in 1:nrow(Kan_cod)){
  x=Kan_cod[i,1]
  y=Kan_cod[i,2]
  z=Kan_cod[i,3]
  a=as.character(Codons[Codons[,2]==x & Codons[,3]==y & Codons[,4]==z,1])
  b=rbind(b,a)
  }
  Kan_cod=cbind(Kan_cod,b[,1])

# Change all base annotations that are 0 to NA in Kan_mut dataframe
Kan_mut.x=Kan_mut[,c(9:12)]
Kan_mut.x[Kan_mut.x==0]=NA
Kan_mut[,c(9:12)]=Kan_mut.x

# add in extra rows to Kan_mut where two mutation types are seen and label each
change - named Kan_mut2
Kan_mut2=c()
for (j in 1:nrow(Kan_mut)){
z=Kan_mut[j,c(9:12)]
y=c()
for (i in 1:length(z)){
  if (is.na(z[i])==F){
    x=z[i]
    y=c(y,x)}}

if (length(y)==1){w=Kan_mut[j,]}
if (length(y)==2){w=rbind(Kan_mut[j,],Kan_mut[j,])}
if (length(y)==3){w=rbind(Kan_mut[j,],Kan_mut[j,],Kan_mut[j,])}
if (length(y)==4){w=rbind(Kan_mut[j,],Kan_mut[j,],Kan_mut[j,],Kan_mut[j,])}

N=names(y)
u=cbind(w,N)
Kan_mut2=rbind(Kan_mut2,u)
}
```

```
#create a matrix containing all mutated versions of the Kan/Neo seqeunce
Kan_changes=matrix(,nrow=nrow(Kan_mut2),ncol=ncol(Kan_seq))
for (i in 1:nrow(Kan_mut2)){
  Kan_changes[i,]=as.matrix(Kan_seq)}
colnames(Kan_changes)=names(Kan_seq)
for (i in 1:nrow(Kan_mut2)){

Kan_changes[i,colnames(Kan_changes)==Kan_mut2[i,6]]=as.character(Kan_mut2[i,13
])}

# Create a numerical position dataframe to mirror codons
Kan_Neo2=as.data.frame(Kan_Neo)
Kan_Neo3=t(Kan_Neo2)
Kan_pos=seq(0,length(Kan_seq)-1, by=3)
Kan_Neo4=matrix(,nrow=length(Kan_pos),ncol=3)
for (i in 1:length(Kan_pos)){
  x=as.matrix(Kan_Neo3[c((1+Kan_pos[i]):(3+Kan_pos[i]))])
  Kan_Neo4[i,]=x}
Kan_Neo4=as.data.frame(Kan_Neo4)

# Create a dataframe containing reference amino acids and the amino acid seen as a
result of mutation
# Append this to Kan_mut dataframe and create an extra Synonymous vs Non-
synonymous column
Kan_amino_changes=c()
colnames(Kan_amino_changes)=c(colnames(w))
for (i in 1:nrow(Kan_changes)){
  a=matrix(,nrow=length(Kan_pos),ncol=3)
  b=Kan_changes[i,]
  for (j in 1:length(Kan_pos)){
    c=as.matrix(b[c((1+Kan_pos[j]):(3+Kan_pos[j]))])
    a[j,]=c}
  a=as.data.frame(a)

  d=c()
  z=c()
  for (k in 1:nrow(a)){
    e=a[k,1]
    f=a[k,2]
    g=a[k,3]
    h=as.character(Codons[Codons[,2]==e & Codons[,3]==f & Codons[,4]==g,1])
    z=rbind(z,h)
  }
  a=cbind(a,z[,1])

  y=as.character(a[Kan_Neo4[,1]==Kan_mut2[i,6] | Kan_Neo4[,2]==Kan_mut2[i,6] |
Kan_Neo4[,3]==Kan_mut2[i,6],4])
```

```
  x=as.character(Kan_cod[Kan_Neo4[,1]==Kan_mut2[i,6] |
Kan_Neo4[,2]==Kan_mut2[i,6] | Kan_Neo4[,3]==Kan_mut2[i,6],4])
  w=cbind(x,y)
  colnames(w)=c("Reference","Sample")

  Kan_amino_changes=rbind(Kan_amino_changes,w)}

Kan_mut3=cbind(Kan_mut2,Kan_amino_changes)


Kan_change_type=c()
for (i in 1:nrow(Kan_mut3)){
  if (as.character(Kan_mut3[i,14])==as.character(Kan_mut3[i,15])){
    x="Synonymous"}
  else{x="Non-Synonymous"}
  Kan_change_type=rbind(Kan_change_type,x)}

Kan_mut3=cbind(Kan_mut3,Kan_change_type)

summary(Kan_mut3$Kan_change_type)
#####################################################################

#####################################################################
#GFP GENE
# Split GFP ORF into codons by row
GFP_pos=seq(0,length(GFP_seq)-1, by=3)
GFP_cod=matrix(,nrow=length(GFP_pos),ncol=3)
for (i in 1:length(GFP_pos)){
  x=as.matrix(GFP_seq[c((1+GFP_pos[i]):(3+GFP_pos[i]))])
  GFP_cod[i,]=x}
GFP_cod=as.data.frame(GFP_cod)

# Annotate each codon with amino acid it codes
b=c()
for (i in 1:nrow(GFP_cod)){
  x=GFP_cod[i,1]
  y=GFP_cod[i,2]
  z=GFP_cod[i,3]
  a=as.character(Codons[Codons[,2]==x & Codons[,3]==y & Codons[,4]==z,1])
  b=rbind(b,a)
}
GFP_cod=cbind(GFP_cod,b[,1])

# Change all base annotations that are 0 to NA in GFP_mut dataframe
GFP_mut.x=GFP_mut[,c(9:12)]
GFP_mut.x[GFP_mut.x==0]=NA
GFP_mut[,c(9:12)]=GFP_mut.x

# add in extra rows to GFP_mut where two mutation types are seen and label each
change - named GFP_mut2
```

```
GFP_mut2=c()
for (j in 1:nrow(GFP_mut)){
  z=GFP_mut[j,c(9:12)]
  y=c()
  for (i in 1:length(z)){
    if (is.na(z[i])==F){
      x=z[i]
      y=c(y,x)}}

  if (length(y)==1){w=GFP_mut[j,]}
  if (length(y)==2){w=rbind(GFP_mut[j,],GFP_mut[j,])}
  if (length(y)==3){w=rbind(GFP_mut[j,],GFP_mut[j,],GFP_mut[j,])}
  if (length(y)==4){w=rbind(GFP_mut[j,],GFP_mut[j,],GFP_mut[j,],GFP_mut[j,])}

  N=names(y)
  u=cbind(w,N)
  GFP_mut2=rbind(GFP_mut2,u)
}


#create a matrix containing all mutated versions of the GFP seqeunce
GFP_changes=matrix(,nrow=nrow(GFP_mut2),ncol=ncol(GFP_seq))
for (i in 1:nrow(GFP_mut2)){
  GFP_changes[i,]=as.matrix(GFP_seq)}
colnames(GFP_changes)=names(GFP_seq)
for (i in 1:nrow(GFP_mut2)){

GFP_changes[i,colnames(GFP_changes)==GFP_mut2[i,6]]=as.character(GFP_mut2[i,1
3])}

# Create a numerical position dataframe to mirror codons
GFP2=as.data.frame(GFP)
GFP3=t(GFP2)
GFP_pos=seq(0,length(GFP_seq)-1, by=3)
GFP4=matrix(,nrow=length(GFP_pos),ncol=3)
for (i in 1:length(GFP_pos)){
  x=as.matrix(GFP3[c((1+GFP_pos[i]):(3+GFP_pos[i]))])
  GFP4[i,]=x}
GFP4=as.data.frame(GFP4)

# Create a dataframe containing reference amino acids and the amino acid seen as a
result of mutation
# Append this to GFP_mut dataframe and create an extra Synonymous vs Non-
synonymous column
GFP_amino_changes=c()
colnames(GFP_amino_changes)=c("Reference","Sample")
for (i in 1:nrow(GFP_changes)){
  a=matrix(,nrow=length(GFP_pos),ncol=3)
  b=GFP_changes[i,]
  for (j in 1:length(GFP_pos)){
    c=as.matrix(b[c((1+GFP_pos[j]):(3+GFP_pos[j]))])
    a[j,]=c}
  a=as.data.frame(a)

  d=c()
  z=c()
  for (k in 1:nrow(a)){
    e=a[k,1]
    f=a[k,2]
    g=a[k,3]
    h=as.character(Codons[Codons[,2]==e & Codons[,3]==f & Codons[,4]==g,1])
    z=rbind(z,h)
  }
  a=cbind(a,z[,1])

  y=as.character(a[GFP4[,1]==GFP_mut2[i,6] | GFP4[,2]==GFP_mut2[i,6] |
GFP4[,3]==GFP_mut2[i,6],4])
  x=as.character(GFP_cod[GFP4[,1]==GFP_mut2[i,6] | GFP4[,2]==GFP_mut2[i,6] |
GFP4[,3]==GFP_mut2[i,6],4])
  w=cbind(x,y)
  colnames(w)=c("Reference","Sample")

  GFP_amino_changes=rbind(GFP_amino_changes,w)}

GFP_mut3=cbind(GFP_mut2,GFP_amino_changes)


GFP_change_type=c()
for (i in 1:nrow(GFP_mut3)){
  if (as.character(GFP_mut3[i,14])==as.character(GFP_mut3[i,15])){
    x="Synonymous"}
  else{x="Non-Synonymous"}
  GFP_change_type=rbind(GFP_change_type,x)}

GFP_mut3=cbind(GFP_mut3,GFP_change_type)

summary(GFP_mut3$GFP_change_type)

############################################################################

############################################################################
# Calculate the probability of synonymous vs non-synonymous mutations

Codons2=Codons
Codons3=Codons
Codons4=Codons
Codons5=Codons

Total=c()
for (i in 1:nrow(Codons2)){
  x=Codons2[i,]
  for (j in 2:4){
    r=c();Codons3=Codons;Codons4=Codons;Codons5=Codons
    if
(as.character(x[1,j])=="A"){Codons3[i,j]="T";Codons4[i,j]="C";Codons5[i,j]="G"}
    if
(as.character(x[1,j])=="T"){Codons3[i,j]="A";Codons4[i,j]="C";Codons5[i,j]="G"}
    if
(as.character(x[1,j])=="C"){Codons3[i,j]="T";Codons4[i,j]="A";Codons5[i,j]="G"}
    if
(as.character(x[1,j])=="G"){Codons3[i,j]="T";Codons4[i,j]="C";Codons5[i,j]="A"}

a=as.character(Codons3[i,2]);b=as.character(Codons3[i,3]);c=as.character(Codons3[i,4]
)

d=as.character(Codons4[i,2]);e=as.character(Codons4[i,3]);f=as.character(Codons4[i,4]
)

g=as.character(Codons5[i,2]);h=as.character(Codons5[i,3]);k=as.character(Codons5[i,4]
)
    l=as.character(Codons[Codons[,2]==a & Codons[,3]==b & Codons[,4]==c,1])
    m=as.character(Codons[Codons[,2]==d & Codons[,3]==e & Codons[,4]==f,1])
    n=as.character(Codons[Codons[,2]==g & Codons[,3]==h & Codons[,4]==k,1])
    if (as.character(Codons3[i,1])==l){o="Synonymous"}else{o="Non-Synonymous"}
    if (as.character(Codons4[i,1])==m){p="Synonymous"}else{p="Non-Synonymous"}
    if (as.character(Codons5[i,1])==n){q="Synonymous"}else{q="Non-Synonymous"}
    r=rbind(o,p,q)
    Total=rbind(Total,r)}}

Syn=0
Non=0
for (i in 1:nrow(Total)){
  if (Total[i,1]=="Non-Synonymous"){Non=Non+1}else{Syn=Syn+1}}
```