Factors Affecting the Specific Productivity of *Pichia pastoris*

Rochelle Aw

Imperial College London
Department of Cell and Molecular Biology

A thesis submitted for the degree of Doctor of Philosophy of Imperial College London and the Diploma of Imperial College London 2012

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.
Declaration

I certify that this thesis represents my own work, unless otherwise stated. All external contributions and any information derived from other sources have been acknowledged accordingly.

Rochelle Aw
Abstract

*Pichia pastoris* has become one of the most popular recombinant protein expression platforms, despite the lack of understanding into the fundamentals of protein expression. Whilst *P. pastoris* exhibits high volumetric productivity, it has a low specific productivity, which can be further reduced by protein-specific problems. This thesis employs several strategies commonly used to increase specific productivity, and assesses their impact on the productivity and cell biology of *P. pastoris*.

Gene dosage has been reported to increase titre; therefore multiple copies of human serum albumin were integrated into *P. pastoris* to assess the correlation between recombinant protein productivity and copy number. Post-transformational vector amplification was used to generate clones containing up to five copies of *HSA*. However it was not possible to correlate copy number and yield as 15 L bioreactor cultures showed significant genetic instability. The mean final copy number was 2.6 ± 1.0. Further work was undertaken to evaluate possible ways to prevent instability, such as different selection methods, mutation of *RAD51* and *RAD52* which are both possible RecA homologs and whether the locus of vector integration plays a part. Integration into the rDNA locus resulted in increased stability with a five copy clone averaging 3.8 ± 1.6. Furthermore, no clones showed complete loss of the integrated vector as observed with integration into the *AOX1* locus.

Additionally, the little understood phenomenon of clonal variation was investigated which has been reported to affect specific productivity. Nine clones, with a range of productivity, were chosen for transcriptomic analysis. Variation between different clones was not uniform, even within the high, mid and low secretor groups. However, the ER associated degradation pathway was consistently upregulated in the high secretors which could be exploited in the future for strain development and selection.
Acknowledgements

My biggest thanks go to Professor David Leak for giving me the opportunity to undertake this PhD and for his continued advice and guidance throughout the project. I would also like to thank Bo Kara and everyone at Fujifilm Diosynth Biotechnologies for their support, intellectual input and for trusting me with their bioreactors. My thanks go to Geraint Barton for his never-ending patience regarding transcriptomic analysis, to Dr. Tom Ellis for his help with FACS and to everyone at the Bacterial Microarray Group at St George’s Hospital, without whom I would not have the transcriptomic data that has featured as such a key component of my thesis.

To the members, past and present, of Lab 326 thank you for the support, encouragement and giggles. I will forever associate you all with “more bass”. My most heartfelt gratitude goes to Dr. Bryn Edwards-Jones and Dr. Karen Polizzi; your guidance, belief and friendship helped carry me through.

To all my friends who have endured four years of science talk, thank you for putting up with me, even when I could no longer differentiate jeans from genes. I would like to thank my parents and my sister for always being there and supporting me unquestionably. Dani, I couldn’t have done this without you, thank you for lighting the way and making sure I followed successfully in your footsteps. I guess you didn’t turn out to be such a bad sister either. To Mum and Dad, no matter how far away you were, I was always aware of your love and I missed you each and every day.

And finally I would like to thank Trev, without whom I would never have had the courage to undertake this. Thank you for believing in me even when I didn’t believe in myself.
## Contents

Declaration ................................................................................................................................. 2
Abstract .................................................................................................................................... 3
Acknowledgements .................................................................................................................. 4
Contents ................................................................................................................................... 5
List of Figures .......................................................................................................................... 11
List of Tables ........................................................................................................................... 15
Abbreviations .......................................................................................................................... 16

1 Introduction ........................................................................................................................ 19
  1.1 Recombinant Proteins for the Biopharmaceutical Industry ............................................... 19
  1.2 Bacteria as Recombinant Expression Platforms ............................................................... 19
      1.2.1 *Escherichia coli* ........................................................................................................ 19
      1.2.2 *Bacillus subtilis* ....................................................................................................... 20
  1.3 Mammalian Cells as Recombinant Expression Platforms .................................................. 20
  1.4 Insect Cells, Plant Cells and Other Organisms .................................................................. 21
      1.4.1 Insect Cells .................................................................................................................. 21
      1.4.2 Green Algae ............................................................................................................... 21
      1.4.3 Transgenic Crops ....................................................................................................... 22
      1.4.4 Filamentous fungi ...................................................................................................... 22
      1.4.5 Other Recombinant Protein Expression Platforms .................................................. 23
  1.5 Yeast as Recombinant Expression Platforms .................................................................... 23
      1.5.1 *Saccharomyces cerevisiae* ....................................................................................... 24
      1.5.2 *Pichia pastoris* .......................................................................................................... 24
  1.6 The History of *P. pastoris* ............................................................................................... 25
  1.7 *P. pastoris* as a Methylotrophic Yeast ............................................................................ 25
      1.7.1 The Alcohol Oxidase (*AOX*) genes .......................................................................... 26
      1.7.2 Methanol Utilisation Phenotypes ............................................................................... 27
  1.8 Strains of *P. pastoris* ...................................................................................................... 27
      1.8.1 Y-11430..................................................................................................................... 28
      1.8.2 X33 ............................................................................................................................ 28
      1.8.3 GS115 ........................................................................................................................ 28
      1.8.4 Protease Deficient Strains .......................................................................................... 29
      1.8.5 The truth about *Komagataella phaffi* and *Komagataella pastoris* ......................... 29
      1.8.6 Humanized *P. Pastoris* strains ............................................................................... 30
  1.9 Expression Vectors for *P. pastoris* .................................................................................. 30
      1.9.1 Promoters ................................................................................................................... 30
      1.9.2 Selectable Markers .................................................................................................... 31
      1.9.3 Secretion Signal ......................................................................................................... 31
  1.10 Homologous recombination ............................................................................................. 32
      1.11 pPICz and pPICzα ...................................................................................................... 34
1.12 Production of Recombinant Proteins in \textit{P. pastoris} .......................................................... 35
1.13 Increasing the yield of \textit{P. pastoris} .................................................................................. 37
  1.13.1 Creating Multiple Copy Clones .................................................................................. 37
1.14 \textit{In-vitro} Multimerisation ............................................................................................... 38
  1.14.1 Post-Transformational Vector Amplification ............................................................... 39
1.15 Unfolded Protein Response (UPR) .................................................................................... 40
1.16 Endoplasmic Reticulum Associated Degradation (ERAD) Pathway ............................ 42
1.17 Human Serum Albumin – The Model Protein ................................................................. 43
1.18 Aims and Objectives .......................................................................................................... 44

2 Materials and Methods ........................................................................................................ 45

  2.1 Strains and Plasmids ........................................................................................................ 45

  2.2 Media ................................................................................................................................ 48
  2.2.1 Miller Lysogeny Broth (LB) Media ............................................................................. 48
  2.2.2 Lennox Lysogeny Broth (LB) Media .......................................................................... 48
  2.2.3 Yeast Extract Peptone Medium (YPD) ....................................................................... 48
  2.2.4 Buffered Glycerol-complex Medium (BMGY) or Buffered-Methanol-complex Medium
      (BMMY) ......................................................................................................................... 48
  2.2.5 Minimal Dextrose Medium (MD) or Minimal Methanol Medium (MM) ..................... 49
  2.2.6 Buffered Minimal Glycerol Medium (BMG) or Buffered Minimal Methanol Medium
      (BMM) ......................................................................................................................... 49
  2.2.7 Tryptone Soy Agar (TSA) .......................................................................................... 49
  2.2.8 \textit{Pichia} Cell Bank Medium (PCMB) ......................................................................... 49
  2.2.9 \textit{Pichia} Freezer Mix (PFM) – Cryopreservation Medium ............................................. 49
  2.2.10 Fermentation Inoculation Media .............................................................................. 49
      2.2.10.1 Yeast Extract/Peptone Media .......................................................................... 49
      2.2.10.2 Glucose Solution .............................................................................................. 49
  2.2.11 Fermentation Media ..................................................................................................... 49

  2.3 Molecular Biology Methods ............................................................................................. 50
  2.3.1 Agarose Gel Electrophoresis ....................................................................................... 50
  2.3.2 Plasmid DNA Isolation ............................................................................................... 51
  2.3.3 Isolation of Yeast Genomic DNA .............................................................................. 51
  2.3.4 DNA quantification ...................................................................................................... 51
  2.3.5 DNA Purification and Concentration ........................................................................... 52
  2.3.6 Restriction Enzymes .................................................................................................... 52
  2.3.7 \textit{HSA} Codon Optimised Gene .................................................................................. 52
  2.3.8 Ligation ....................................................................................................................... 52
  2.3.9 Easy Cloning Vectors .................................................................................................. 52
  2.3.10 Transformation and preparation of competent cells of \textit{E. coli} ..................................... 53
  2.3.11 Collection of cells for RNA Extraction ...................................................................... 53
  2.3.12 Isolation of RNA ........................................................................................................ 54
  2.3.13 Quantification and Quality Checks of RNA ............................................................... 54
  2.3.14 cDNA ......................................................................................................................... 54
  2.3.15 Polymerase Chain Reaction (PCR) ........................................................................... 54
  2.3.16 Gibson Assembly Protocol ....................................................................................... 55
2.3.17 Random-Amplification of Polymorphic DNA (RAPD)-PCR ........................................55
2.3.18 Quantitative Polymerase Chain Reaction (qPCR) ..................................................55
  2.3.18.1 qPCR calculations .................................................................................................56
2.3.19 Quantitative Reverse Transcription (RT)-qPCR..........................................................56
2.4 Transformation of \textit{P. pastoris} by Electroporation.........................................................57
  2.4.1 Preparation of DNA for Transformation ......................................................................57
  2.4.2 Electroporation Protocol ............................................................................................57
2.5 Post-transformational Vector Amplification ......................................................................57
2.6 Preparation of Cell Banks for Fermentation .....................................................................58
  2.6.1 Microbiological Purity Test .........................................................................................58
2.7 Small Scale Heterologous Protein Expression from Recombinant \textit{P. pastoris} Strains ........58
  2.7.1 Expression in 24-well microtitre plates ........................................................................58
  2.7.2 Expression in 50 mL centrifuge tubes ..........................................................................59
  2.7.3 Expression in 250 mL baffled flasks ............................................................................59
2.8 Fermentation in 15 L Bioreactors .....................................................................................59
  2.8.1 Inoculation Preparation 15 L Bioreactors for Fermentation .........................................59
  2.8.2 15 L Fermentation .......................................................................................................59
2.9 SDS-PAGE Protein Gel .....................................................................................................60
  2.9.1 12\% (v/v) Resolving gel ...............................................................................................60
  2.9.2 Stacking gel ..................................................................................................................60
  2.9.3 SDS-PAGE Running Buffer (1x) ..................................................................................60
  2.9.4 1x SDS-PAGE Loading Buffer ...................................................................................60
  2.9.5 Coomassie Blue Stain (1 Litre) ....................................................................................60
  2.9.6 Coomassie Blue Destain (1 Litre) ................................................................................61
  2.9.7 Protein gel electrophoresis ..........................................................................................61
2.10 Albumin Blue Florescence Assay ....................................................................................61
2.11 Southern Blots ..................................................................................................................61
  2.11.1 0.2 M HCl ....................................................................................................................61
  2.11.2 Denaturation Solution ................................................................................................61
  2.11.3 Neutralisation Solution ..............................................................................................62
  2.11.4 1 X Maleic Acid ..........................................................................................................62
  2.11.5 20 X SSC .....................................................................................................................62
  2.11.6 Washing Buffer ...........................................................................................................62
  2.11.7 Detection Buffer .........................................................................................................62
  2.11.8 Preparing the DIG-labelled DNA probe ....................................................................62
  2.11.9 Blotting protocol ........................................................................................................62
2.12 Flow Cytometry Analysis ................................................................................................63
2.13 Statistical Analysis ..........................................................................................................63
2.14 Microarray Analysis .........................................................................................................63
  2.14.1 Gene Expression Analysis ..........................................................................................64
  2.14.2 Pathway Analysis ........................................................................................................64
3 Determining the Effects of using Multi-Copy Clones in \textit{Pichia pastoris} .....................65
  3.1 Introduction ......................................................................................................................65
The Effects of Using a Codon Optimised Gene

Investigation of Genetic Stability was observed with Vectors Integrated into the rDNA Locus

Maintaining Zeocin Selection through Expression Does Not Prevent Genetic Instability

Clones that grow in the presence of 200 μg mL⁻¹ Zeocin but do not show an increase in copy number

Analysis of Secretion Saturation in Multi-Copy Clones

Effects of Gene Copy-Number on induction of UPR

Genetic Instability in Multi-Copy Clones

Discussion

4 Investigation of the Prevalence of Genetic Instability in Multi-Copy Clones of *Pichia pastoris* ................................................................. 92

Introduction................................................................. 92

Current Investigations into Improving Genetic Stability ......................................................... 92

Stability through Integration at the rDNA locus ................................................................. 92

Stability through a KU70 Knockout Strain ................................................................. 93

RecA Homologs in *P. pastoris* ...................................................................................... 93

Aims for Genetic Instability ......................................................................................... 93

Maintaining Zeocin Selection through Expression Does Not Prevent Genetic Instability ...... 94

Using *HIS4* for selection instead of Zeocin does not result in Increased Stability ............ 96

*RADS1* and *RADS2* Knockout Strains are not viable .................................................. 100

Genetic Stability was observed with Vectors Integrated into the rDNA Locus ............... 107

Vector Creation ............................................................... 107

Initial integration into the rDNA locus .............................................................................. 109

Creating Multi-Copy Clones of rDNA Integrated Vectors ............................................... 113

Stability of Multi-Copy Clones ...................................................................................... 118

Discussion ......................................................................................................................... 119

5 The Effects of Using a Codon Optimised Gene for Expression in *Pichia pastoris* .......... 122

Introduction ..................................................................................................................... 122

Codon Optimised Genes in *P. pastoris* ........................................................................... 122

Aims .................................................................................................................................. 122

Single Copy Clones of HSAoptimised gene ........................................................................ 123

Generation of Multi-Copy Clones of the HSAopt Gene in the rDNA locus .................. 125

Vector Creation ............................................................................................................... 125

Initial integration of HSAopt vectors into the rDNA locus ............................................... 126
6 Transcriptomic Analysis of Clonal Variation in *Pichia pastoris* ................................................................. 138

6.1 Introduction .................................................................................................................................................. 138

6.1.1 Evidence of Clonal Variation in Mammalian Cells and Other Organisms .............................................. 138

6.1.2 Clonal Variation in *P. pastoris* .............................................................................................................. 138

6.1.3 Studying Clonal Variation ...................................................................................................................... 139

6.1.4 Aims for the study of Clonal Variation .................................................................................................. 139

6.2 Clonal Variation in HSA expressing strains ............................................................................................... 139

6.2.1 Clonal Variation in Fed-Batch Fermentation ......................................................................................... 139

6.2.2 Range of Variation in a Single Transformation ...................................................................................... 142

6.3 Microarray Analysis .................................................................................................................................... 144

6.3.1 Selecting strains for Microarray Analysis ............................................................................................. 144

6.3.2 Genotyping Microarray Strains ............................................................................................................. 145

6.3.3 Transcript Levels Determined by RT-qPCR ......................................................................................... 147

6.3.4 Microarray Analysis ............................................................................................................................... 148

6.3.4.1 Manual Annotation of the microarray data to reveal differentially expressed genes .................... 149

6.3.4.2 Pathway Analysis ............................................................................................................................ 159

6.3.4.3 KEGG Pathway Mapper for the Visualisation of Significantly Changed Pathways ....................... 163

6.3.5 The anomalous CV7(L) ......................................................................................................................... 173

6.4 Discussion .................................................................................................................................................. 178

7 General Discussion and Future Work ........................................................................................................... 181

7.1 Understanding Genetic Instability ............................................................................................................ 181

7.2 Examining the possibility of a fully stable clone ...................................................................................... 182

7.3 Improving screens for genetic instability ................................................................................................ 183

7.4 Choosing the best protein ........................................................................................................................ 184

7.5 Understanding Clonal Variation ............................................................................................................. 185

7.6 Identifying the underlying cause of clonal variation ................................................................................. 187

7.7 Conclusions ................................................................................................................................................ 187

8 References ..................................................................................................................................................... 188

9 Appendix ....................................................................................................................................................... 201

9.1 *HSA* Codon Optimised DNA Sequence .................................................................................................. 201

9.2 Oligonucleotides Used throughout this Study .......................................................................................... 202

9.3 Zeocin Gene and Promoter Region Sequencing ....................................................................................... 205

9.4 GpaH11_2000B Fermentation Trace ......................................................................................................... 206

9.5 DNA Sequencing of Clonal Variants for Integration Site into the *P. pastoris* genome ......................... 207

9.6 Volcano plots of Microarray Expression Data for Clonal Variant strains compared to X33 ... 209

9.6.1 High Secreting Strains ......................................................................................................................... 209

9.6.2 Mid-Level Secreting Strains ................................................................................................................ 210

9.6.3 Low Secreting Strains ......................................................................................................................... 211

9.7 Manual annotation of Microarray analysis ............................................................................................. 212
List of Figures

Figure 1.1 Methanol Utilisation in *P. pastoris* ................................................................. 26
Figure 1.2 Single Crossover Homologous Recombination ...................................................... 32
Figure 1.3 Double Crossover Homologous Recombination ...................................................... 33
Figure 1.4 Homologous Recombination Pathways in *S. cerevisiae* ........................................ 34
Figure 1.5 pPICz and pPICzα Plasmid Maps ........................................................................ 35
Figure 1.6 Schematic diagram of the method suggested by Invitrogen Corporation for in vitro multimerisation .................................................................................................................. 39
Figure 1.7 A schematic visualisation of the upregulation of the unfolded protein response (UPR) in yeast ........................................................................................................................................ 41
Figure 1.8 Membrane Protein Complexes that Define the ERAD pathways .............................. 42
Figure 2.1 DNA Molecular Marker Size ................................................................................ 50
Figure 3.1 pα-HSA Plasmid Map ....................................................................................... 67
Figure 3.2 Agarose gel confirms the successful construction of the pα-HSA vector .................. 68
Figure 3.3 pPICz-HSA Plasmid Map .................................................................................. 69
Figure 3.4 Agarose gel confirms the successful construction of the pPICz-HSA vector ............. 69
Figure 3.5 Production of HSA after 24 hour production in methanol containing media .......... 71
Figure 3.6 Plates of parental clones isolated on 100 μg mL⁻¹ Zeocin plates and progeny clones post PTVA isolated on 2000 μg mL⁻¹ Zeocin ................................................................. 72
Figure 3.7 Post-PTVA Expression of HSA compared to levels pre-PTVA both Increases and Decreases in Progeny Clones .................................................................................... 73
Figure 3.8 Summary of Copy shows an Increase and Decrease of Copy Number post-PTVA .... 74
Figure 3.9 Southern blot of genomic DNA digested with EcoRV from a selection of parental and progeny clones of PTVA determines a head-to-tail insert orientation ...................................... 76
Figure 3.10 Copy number calculated by qPCR of strains from the parental clone (selected on 100 μg mL⁻¹ Zeocin) and progeny clones which were streaked to single copies indicating a mixed culture. ................................................................................................................................. 78
Figure 3.11 Alignment of DNA sequences within the *TEF1* promoter region from strains selected on 100 μg mL⁻¹ Zeocin or 200 μg mL⁻¹ suggest variation within the promoter regions .......... 80
Figure 3.12 Zeo(R) mRNA Levels from Single Copy Clones Post-PTVA Show an Increase and Decrease in Transcript levels ............................................................................................ 81
Figure 3.13 Titre levels of HSA (mg L⁻¹ DCW⁻¹) during 48 hour fed batch methanol induction .... 82
Figure 3.14 Upregulation of UPR during 48 Hour Fermentation .............................................. 84
Figure 3.15 Colony count (x10⁷) of serial dilutions plated onto either YPD plates (no selective pressure) or 100 μg mL⁻¹ Zeocin YPD plates shows a loss of resistance to Zeocin. ................. 85
Figure 3.16 Genetic instability is prevalent in GpαH11_2000B (5 copy) during a 48 Hour Fermentation ................................................................. 86
Figure 3.17 Genetic Instability is present in the cell banks of GpαH11_2000B (5 copy) determined by qPCR of HSA genes .............................................................................................................. 87
Figure 3.18 Genetic Instability occurs in GpαH1_100 (1 copy), GpαH11_100 (1 copy), GpαH1_2000A (2 copy) and GpαH11_2000A (3 copy) during a 48 hour Fermentation ......................... 89
Figure 4.1 Clones of GpαH11_2000B Grown in the Presence of Constant Zeocin show evidence of Genetic Instability ........................................................................................................... 95
Figure 4.2 pα-Tx3 Plasmid Map .......................................................................................... 96
Figure 4.3 Vector Creation of pIBT3 ................................................................. 98
Figure 4.4 Transformation of pIBT3 into GS115 results in Varying Copy Numbers .......... 99
Figure 4.5 pIBT3 Clones show Genetic Instability post-24 hour Expression ............... 100
Figure 4.6 Construction of the pαTx3R51 and pαTx3R52 Vectors ................................ 101
Figure 4.7 Amplification of the RAD51 gene showed no gene knockout post-integration of the pα-Tx3R51 vector ................................................................. 102
Figure 4.8 pIBT3R51 and pIBT3R52 Plasmid Maps ..................................................... 103
Figure 4.9 Amplification of the RAD51 or RAD52 genes showed no gene knockout post-transformation of the pIBT3R51 and pIBT3R52 vectors ........................... 104
Figure 4.10 Creation RAD51-HIS4 and RAD52-HIS4 Linear DNA fragments ................ 105
Figure 4.11 Amplification of the RAD51 and RAD2 genes showed no gene knockout using a linear DNA fragment ............................................................................. 106
Figure 4.12 pGRzaHSA Plasmid Map ........................................................................ 108
Figure 4.13 pARzaHSA Plasmid Map ................................................................. 108
Figure 4.14 Successful Integration of the pGRzaHSA and pARzaHSA vectors into the rDNA locus ........................................................................................................ 109
Figure 4.15 Integration into the rDNA Locus results in Multi-Copy Clones ................. 110
Figure 4.16 HSA Titre Levels of Clones Integrated into the rDNA Locus containing the HSA gene ................................................................. 111
Figure 4.17 Clones with Integration into the rDNA Locus show Reduced Transcript Levels compared to Clones with Integration in the AOX1 Locus .................................................. 112
Figure 4.18 Comparison of UPR signals from clones Integrated into the rDNA Locus versus Integration into the AOX1 Locus .................................................................................... 113
Figure 4.19 HSA Expression post-PTVA of Vectors Integrated into the rDNA Locus .... 114
Figure 4.20 No correlation between HSA copy number and HSA titre was seen for Clones Integrated into the rDNA Locus ........................................................................ 115
Figure 4.21 HSA Transcript is Increased Post-PTVA for Clones with Vectors Integrated into the rDNA Locus Compared to Clones with Vectors Integrated into the AOX1 Locus ................................................................. 116
Figure 4.22 Gene Expression Levels of UPR Signals Upregulated in Clones Containing Vectors integrated into the rDNA Locus Compared to Integration at the AOX1 Locus ........................................... 117
Figure 4.23 Clones with vectors integrated into the rDNA Locus using both the AOX1 and GAP promoter show evidence of Genetic Instability ........................................... 118
Figure 5.1 pα-HSAopt Plasmid Map ......................................................................... 123
Figure 5.2 HSA Titre of strains containing the HSAopt Gene showed Increased Secretion compared to HSA Gene ........................................................................................................ 124
Figure 5.3 Varying Copy Numbers in Transformed Strains of GpαHopt strains ............. 124
Figure 5.4 pGRzaHSAopt and pARzaHSAopt Plasmid Maps ........................................ 125
Figure 5.5 pARzaHSAopt BglII Plasmid Map ........................................................... 126
Figure 5.6 Colony PCR Indicated Correct Integration of pGRzaHSAopt and pARzaHSAopt vectors into the rDNA locus ......................................................................................... 127
Figure 5.7 Integration into the rDNA Locus results in Multi-Copy Clones .................... 127
Figure 5.8 HSA Titre from Clones Integrated into the rDNA Locus Containing the HSAopt Gene ................................................................. 128
Figure 5.9 HSAopt Transcript Levels of Clones Integrated into rDNA locus were lower than a Single Copy Clone Integrated into the AOX1 Locus ............................................................................ 129
Figure 5.10 Transcript Levels of Genes involved in UPR of Clones Containing the HSAopt Gene Integrated into the rDNA Locus ................................. 129
Figure 5.11 HSA Expression post-PTVA of Multi-Copy Clones containing the HSAopt Gene Integrated into the rDNA Locus .......................................................... 131
Figure 5.12 Multi-copy clones containing the HSAopt Gene produce lower HSA titre that when using the HSA gene........................................................................... 132
Figure 5.13 Quantification of HSA Transcripts of Clones Post-PTVA, Integrated into the rDNA Locus using either the HSAopt gene or HSA gene................................................................. 133
Figure 5.14 Gene Expression levels of UPR signals upregulated in clones containing the HSAopt gene compared to clones containing the HSA gene. .......................................................... 134
Figure 5.15 Instability is Prevalent in Clones with the GAP promoter and HSA optimised Gene.. 135
Figure 5.16 Instability is Prevalent in Clones with the AOX1 promoter and HSA optimised Gene 136
Figure 6.1 HSA transcript levels of single copy clones GpαH1_100 and GpαH11_100. .................. 140
Figure 6.2 HSA transcript levels for single copy clones GpαH1_100 and GpαH11_100. .......... 141
Figure 6.3 UPR upregulation in single copy clones GpαH1_100 and GpαH11_100............... 142
Figure 6.4 Copy number analysis of 23 colonies selected from a single transformation of pα-HSA determine that all are Single Copy Clones .......................................................... 143
Figure 6.5 HSA expression levels of single copy GpαHSA strains.......................................... 143
Figure 6.6 Sequence alignment of the integration site for clones used in microarray analysis... 144
Figure 6.7 HSA titre levels of the nine Strains selected for microarray analysis, grouped into high, medium and low levels. .......................................................................................... 145
Figure 6.8 RAPD-PCR of strains selected for microarray analysis........................................ 146
Figure 6.9 HSA transcript levels after 24 hour expression..................................................... 147
Figure 6.10 Induction of the Unfolded Protein Response (UPR) in strains selected from microarray analysis.................................................................................................................. 148
Figure 6.11 Percentage of Genes Up- or Downregulated in CV2(H) through manual annotation 150
Figure 6.12 Growth Curves of Clonal Variation Strains Grown in Glycerol or Methanol Containing Rich Media........................................................................................................... 151
Figure 6.13 Histogram to Show Forward Scatter of Clonal Variation Strains Grown in Glycerol containing Media to Determine Average Cell Size................................................................. 153
Figure 6.14 Histogram to Show Forward Scatter of Clonal Variation Strains Grown in Methanol containing Media to Determine Average Cell Size................................................................. 154
Figure 6.15 FACS analysis of Clonal Variation Strains and X33 Grown in Glycerol Containing Media .................................................................................................................................. 156
Figure 6.16 FACS analysis of Clonal Variation Strains and X33 Grown in Methanol Containing Media .................................................................................................................................. 157
Figure 6.17 A comparison of Pathways Up- and Downregulated in High Secreting Strains at p≤0.1................................................................................................................................. 160
Figure 6.18 A comparison of Pathways Upregulated and Downregulated in Mid-Level Secreting Strains at p≤0.1.............................................................................................................. 161
Figure 6.19 A comparison of Pathways Upregulated and Downregulated in Low Secreting Strains at p≤0.1 ...................................................................................................................... 162
Figure 6.20 Protein Processing in the Endoplasmic Reticulum Pathway of CV2(H)................... 163
Figure 6.21 Ubiquitin Mediated Proteolysis Pathway in CV7(L) ............................................. 166
Figure 6.22 Ubiquitin Mediated Proteolysis Pathway in CV14(L)............................................ 167
Figure 6.23 Ubiquitin Mediated Proteolysis Pathway in CV5(L) ............................................. 168
Figure 6.24 Ubiquitin Mediated Proteolysis Pathway in CV23(H)............................................ 169
List of Tables

Table 1.1 *P. pastoris* host strains ........................................................................................................... 28
Table 1.2 Overview of Protein Production in *P. pastoris* ..................................................................... 36
Table 2.1 *E. coli* strains used in this study ............................................................................................... 45
Table 2.2 *P. pastoris* strains used in this study ....................................................................................... 46
Table 2.3 Plasmids used in this study ....................................................................................................... 47
Table 2.4 qPCR Primer Efficiencies .......................................................................................................... 56
Table 6.1 Genes Upregulated in High Secretion Strains Only ................................................................. 170
Table 6.2 Genes Downregulated in High Secretion Strains Only ............................................................. 171
Table 6.3 Pathways Significantly (p<0.1) Up- and Downregulated in CV7(L) ................................. 173
Table 9.1 Oligonucleotides ...................................................................................................................... 202
Abbreviations

αMF  alpha-mating factor
μ     Micro
μg    Microgram
μL    Microlitre
μM    Micromolar
°C    Degrees Celsius
ACT1  Actin 1
AFLP  Amplified fragment length polymorphism
AOX1  Alcohol oxidase
ATF6  Activating transcription factor 6
BG    Background Noise
bla   Bleomycin
BMG   Buffered Minimal Glycerol
BMGY  Buffered Glycerol-complex Medium
BMM   Buffered Minimal Methanol
BMMY  Buffered Methanol-complex Medium
BSA   Bovine serum albumin
BSD   Blastidin
CER   Carbon dioxide evolution rate
CHO   Chinese hamster ovary
CSPD  Chemiluminescent Alkaline Phosphate Substrate
Ct    Cycle threshold
CV    Clonal Variation
DCU   Display control unit
DCW   Dry cell weight
H2O   Distilled water
DNA   Deoxyribonucleic acid
DTT   Dithiothreitol
EDTA  Ethylenediaminetetraacetic acid
ELF   Equine lactoferrin
ELISA Enzyme Linked ImmunoSorbent Assay
ER    Endoplasmic reticulum
ERAD  ER associated degradation pathway
FACS  Fluorescence-activated cell sorting
FDA   Food and Drug Administration
fdr   False discovery rate
FLD   Formaldehyde dehydrogenase
g    Gram
GAP   Glyceraldehydes-3-phosphate
GFP   Green fluorescent Protein
GO    Gene ontology
GRAS  Generally Regarded As Safe
h  Hour
HCl  Hydrochloric acid
HIS  Histidine
HSA  Human serum albumin
HSAopt  A codon optimised variation of human serum albumin
HSA-SP  Human serum albumin without the native secretion signal peptide
hSOD  Human Cu, Zn-superoxide dismutase
IRE1  Inositol requiring enzyme 1
IRES  Internal ribosome entry Site
KAR2  A karyogamy gene, equivalent to human BiP
Kb  Kilobase
KDA  Kilodalton
KEGG  Kyoto Encyclopaedia of Genes and Genomes
Kg  Kilogram
KOBAS  KEGG Orthology Based Annotation System
L  Litre
LB  Luria Bertani media
M  Molar
mA  Milliamps
MAPK  Mitogen-activated protein kinase
MD  Minimal Dextrose
mg  Milligram
mL  Millilitre
mM  Millimolar
mRNA  Messenger RNA
Mut  Methanol utilisation
NaCl  Sodium chloride
NaOH  Sodium hydroxide
ng  Nanogram
NHEJ  Non-homologous End Joining
nm  Nanometre
NRRL  Northern Regional Research Laboratories
NoME-Seq  Nucelosome Occupancy and Methylome Sequencing
OD  Optical density
PCBM  Pichia cell bank medium
PCR  Polymerase chain reaction
PDI  Protein disulfide isomerases
PEG  Polyethylene glycerol
PERK  Doubled stranded RNA-activated protein kinase-like ER kinase
PFM  Pichia freezer mix
pH  Measure of acidity/alkalinity in a solution
PI  Propidium iodide
PIMP  Plant-made industrial pharmaceuticals
PIP  Porcine Insulin Precusor
PMP  Plant-made pharmaceuticals
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>qPCR</td>
<td>Quantitative real-time polymerase chain reaction</td>
</tr>
<tr>
<td>RAPD</td>
<td>Random amplification of polymorphic DNA</td>
</tr>
<tr>
<td>RCB</td>
<td>Research cell bank</td>
</tr>
<tr>
<td>RCT</td>
<td>Research Corporation Technology</td>
</tr>
<tr>
<td>rDNA</td>
<td>Ribosomal DNA</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>rpm</td>
<td>Revolutions per minute</td>
</tr>
<tr>
<td>RT-qPCR</td>
<td>Reverse transcription quantitative real-time polymerase chain reaction</td>
</tr>
<tr>
<td>SIBA</td>
<td>Salk Institute Biotechnology/Industrial Associate Inc</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium dodecyl sulphate</td>
</tr>
<tr>
<td>SP</td>
<td>Signal peptide</td>
</tr>
<tr>
<td>SNARE</td>
<td>Soluble NSF adaptor protein receptor</td>
</tr>
<tr>
<td>TAE</td>
<td>Tris-acetate-EDTA buffer</td>
</tr>
<tr>
<td>TBS</td>
<td>Tris buffered saline</td>
</tr>
<tr>
<td>SUMO</td>
<td>Small Ubiquitin-like Modifier</td>
</tr>
<tr>
<td>TE</td>
<td>Tris-EDTA buffer</td>
</tr>
<tr>
<td>TEMED</td>
<td>Tetramethylethylenediamine</td>
</tr>
<tr>
<td>Try1</td>
<td>Trypsinogen 1</td>
</tr>
<tr>
<td>TSA</td>
<td>Tryptone soy agar</td>
</tr>
<tr>
<td>TT</td>
<td>Transcription termination region</td>
</tr>
<tr>
<td>U</td>
<td>Unit</td>
</tr>
<tr>
<td>UPR</td>
<td>Unfolded protein response pathway</td>
</tr>
<tr>
<td>UPRE</td>
<td>Unfolded protein response element</td>
</tr>
<tr>
<td>v/v</td>
<td>volume/volume</td>
</tr>
<tr>
<td>w/v</td>
<td>weight/volume</td>
</tr>
<tr>
<td>YNB</td>
<td>Yeast nitrogen base</td>
</tr>
<tr>
<td>YPD</td>
<td>Yeast peptone dextrose</td>
</tr>
<tr>
<td>Zeo (R)</td>
<td>Zeocin resistance</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Recombinant Proteins for the Biopharmaceutical Industry

It is predicted that by 2013 the protein therapeutic industry will be worth US$160 billion worldwide, with therapeutics making up 70% of the predicted biologics market [1]. Recombinant therapeutic proteins offer an alternative to extracting proteins from the natural hosts and the associated disadvantages of limited sources, possibility of toxins, antigenic rejection and the risk of transmissible spongiform encephalopathies from the host animal [2]. A main drawback to current recombinant therapeutic proteins is cost. A single dose of Herceptin®, a monoclonal antibody used to treat breast cancer on initial diagnosis, costs €39,608 [3-4]. Although monoclonal antibodies are more expensive to manufacture than recombinant proteins, the wholesale cost per cycle of treatment is only US$3672 indicating the huge mark-up in price that pharmaceutical companies rely on [5]. Efforts to develop a cheaper efficient platform process are under way which will deliver results in efficient, humanized and more economical alternatives.

The first recombinant protein manufactured was human insulin made using Escherichia coli in 1978 [6]. By 1982 Genetech Inc. and their licensing partner Eli Lilly were the first to gain approval from the Food and Drug Association (FDA) for a recombinant protein, which they aptly named, Humulin™ [7]. By 2006 over 200 recombinant proteins had been approved by the FDA [8]. In 2009, 151 proteins had been approved for use in the USA and EU, of which 50% were produced using only E. coli and Saccharomyces cerevisiae [9]. The remaining 50% were produced by higher eukaryotic cells such as insect cells (1%), hybridomas (10%), mammalian cells (39%) and transgenic animals (1%) [10]. Evidently E. coli is the most popular choice of expression system accounting for 29.8% of all FDA drugs in 2009 [10]. However, increasingly more complex systems are required to produce larger and more complex proteins.

1.2 Bacteria as Recombinant Expression Platforms

The first recombinant protein was made in a bacterial system mainly due to the ease of genetic manipulation and since then it has become one of the most widely used hosts [11]. There are many advantages to using a bacterial host system such as speed of growth, rapid expression and high product yields [8]. Nonetheless as requirements for protein production become more complex it is apparent that perhaps bacterial systems are not sufficient.

1.2.1 Escherichia coli

The most prominent example of a bacterial expression platform is, as discussed, E. coli. There are many advantages of using E. coli including the ease of genetic manipulation, high growth rates
and the in-depth knowledge that comes from the published genome sequence. In fact, it is in the detailed investigation into the fundamentals of *E. coli* that makes it the most popular microbial host for recombinant proteins. Due to fast growth rates, evaluating the process of producing recombinant protein in *E. coli* can take as little as one week. Development in yeast can take over a month, while in mammalian cells after development of a stable cell line (which in itself can take several years) a minimum of three months is expected [12].

However, *E. coli*, by nature, is a pathogenic strain and is not on the FDA’s list of generally regarded as safe (GRAS) organisms, although lab strains have been genetically modified and deemed non-pathogenic [13]. While most of the strains used for genetic modification have been adapted to reduce pathogenicity there are obvious drawbacks for extensive use. One of the most fundamental disadvantages of using *E. coli* is the lack of post-transformational modification, such as glycosylation [14].

### 1.2.2 Bacillus subtilis

*Bacillus subtilis* is a Gram positive rod-shaped bacteria and a GRAS organism [15]. One of the key advantages to using *B. subtilis* over *E. coli* is that the outer membrane has no lipopolysaccharides (LPS), which are known endotoxins and are pyrogenic in humans and other mammals [11]. *B. subtilis* has naturally high expression capability and yields can reach up to 3 g L$^{-1}$ for the production of α-amylase from *Bacillus amyloliquefaciens* [16]. A further advantage of *B. subtilis* is direct secretion into the extracellular medium; however bottlenecks in the secretory pathway make expressing secretory proteins difficult. Difficulties include degradation of the secretory protein, incorrect folding and poor targeting to the translocase [2, 15]. Furthermore the lack of suitable expression vectors, plasmid instability and the presence of proteases limit the feasibility of *B. subtilis* as a heterologous expression system host.

### 1.3 Mammalian Cells as Recombinant Expression Platforms

The choice of mammalian cells as a recombinant expression platform decreases the risk of immune rejection when producing therapeutic proteins. Without any genetic manipulation mammalian cells are able to efficiently and accurately perform post-translational modifications [8]. The development of Chinese hamster ovary (CHO) cells led to the flourish of proteins produced as biopharmaceuticals; 70% of all therapeutic proteins developed in mammalian systems are produced using CHO cells [17]. Between 2006 and 2010 among the 58 biopharmaceuticals produced were produced from mammalian cells [18]. In 2011 the CHO genomic sequence was published extending the scope of functionality of this recombinant expression platform [19].
Irrespective of all of the advantages that using mammalian cell lines bring, in particular reducing the need for post-translation modifications, the main disadvantage is cost. It has been calculated that in CHO systems for GMP-production media alone costs about €1 per litre [20]. Additionally production time for mammalian cells can last for up to 21 days (compared to bacteria which can express in a single day) [21]. Furthermore, secretion rates are low and cellular stress responses (such as the unfolded protein response) can result in protein degradation leading to reduced yields. Yield from mammalian cells is on average between 1 – 5 g L\(^{-1}\), which is only achieved through extensive optimisation, whereas through the use of E. coli yields can reach 5 g L\(^{-1}\) [21-22].

1.4 Insect Cells, Plant Cells and Other Organisms

1.4.1 Insect Cells

Like mammalian cells insect cells are capable of complex post-translational modifications [8]. The creation of baculovirus vectors made insect cells a good candidate for heterologous protein production. Key advantages lie in the ease of genetic manipulation, stress-resistance, ease of protein folding and, most importantly, higher yields [2, 23]. Transformation of insect cells with baculoviruses have yielded up to 11 g L\(^{-1}\) [24]. Therapeutic vaccines are produced particularly well in insect cells through the use of virus-like particles (VLP), which contain the native virus expressing surface protein of the virus but do not contain the genetic material [2, 25-26]. In 2000 the first therapeutic protein produced from insect cells released onto the market was a veterinary vaccine against swine flu, since then many other veterinary vaccines have been produced using this expression platform [27].

The main disadvantage of insect cells as a recombinant protein expression host is that, despite post-translational modifications, an immune response may occur as the glycosylation may induce antibodies against the recombinant protein.

1.4.2 Green Algae

As a unicellular eukaryotic organism Chlamydomonas reinhardtii has advantages over other expression systems due to its ability to translate and fold proteins effectively and maintain similar growth rates to microbial cells [28-29]. Other advantages include the ease of transformation into the nucleus or chloroplast and the development time between initial transformation to scale up is much shorter than in mammalian cells or in plants [30]. C. reinhardtii like B. subtilis is considered a GRAS organism and as is not susceptible for viral infections or toxic contaminations [31]. There are a number of promoters available for recombinant protein production coupled to the ability to grow phototrophically. The cost of media for this is expression system is similar to that of plants,
roughly US $0.08 L\textsuperscript{-1}, and soluble protein can account for up to 25% of the dry weight [32]. Both recombinant proteins and monoclonal antibodies have been produced in C. reinhardtii [30-31, 33]. However the amount of development into production of recombinant proteins using this system is still relatively limited.

1.4.3 Transgenic Crops

Plants have been used to produce recombinant pharmaceutical products since the late 1980s, with the production of interferon (1986) and serum albumin (1990) [8, 34]. The first therapeutic vaccine was raised against the Hepatitis B surface antigen in 1992 using tobacco plants [35-36]. Tobacco plants are common hosts as expression vectors are easily introduced into the genome using either Agrobacterium tumefaciens or tobacco mosaic virus, [8]. Indeed there are many advantages to using a plant system over other systems such as the reduced risk of contamination with viruses or other pathogens and the low cost of cultivating plants, which only require water, minerals and sunlight. It has even been suggested that the cost of large scale production is cheaper in plants than it is in microbes [37]. Furthermore, plants are able to carry out glycosylation and folding of complex proteins. Titre levels have been reported up to 0.2% of the dry cell weight of the plant or crop. In 2011 human serum albumin (HSA) was produced in Oryza sativa up to 2.75 g Kg\textsuperscript{-1} of rice [38-39]. Disadvantages of transgenic crops include the potential exposure to pesticides, herbicides and toxic plant metabolites [40]. Moreover from initial transformation to the production of reasonable titre (mg to g) development can take up to three years [30].

1.4.4 Filamentous fungi

Filamentous fungi have been a popular choice for producing recombinant proteins due to their ability to secrete high titres of proteins that are bioactive and have the correct post-translational modifications [8]. One of the most favoured of these fungi is Aspergillus niger which has been reported to secrete up to 25 g L\textsuperscript{-1} of glucoamylase [41]. Other proteins produced with Aspergilli include amylases, pectinases, lipases, cellulases, proteases, phytases and xylanases [42]. With the whole genome sequencing being published for both A. niger and A. oryzae the scope for further development has increased [43-44]. A. niger can secrete over ten times as much native protein as yeast [45]. Despite the potential for protein production recombinant protein titres are not always comparable with low yields produced. Typical methods for increasing titre including increasing gene dosage, strong promoters and gene fusions have not consistently improved yield. Transcription has been identified as the main bottleneck in protein production. Nonetheless the
extensive work on vectors that allow for integration of recombinant genes highlights filamentous fungi as a strong tool as a recombinant expression platform [45].

Like all the production systems described, filamentous fungi have disadvantages as a recombinant expression platform as whilst high levels of expression have been obtained with fungal proteins, non-fungal proteins are a much greater challenge [46]. Although the creation of strains containing multiple copies of the gene of interest will result in an increase in gene expression, there is no direct correlation between copy number and titre [45]. Furthermore like mammalian and yeast cells filamentous fungi are known to have an unfolded protein response [47]. The UPR can result in the targeting of misfolded proteins down the degradation pathways thus reducing yield. Additionally production of recombinant protein can be severely inhibited by fungal proteases. Aspergillus nidulans contains 80 protease genes; however investigations to generate new protease deficient strains (but that do not result in lethal mutations) are being undertaken [48]. Perhaps it is fair to say that the strengths of filamentous fungi lie in the production of fungal metabolites and further work is needed to produce non-fungal heterologous proteins with the same efficiency.

1.4.5 Other Recombinant Protein Expression Platforms

A variety of other platforms are currently being developed for production of recombinant proteins. Plant cells, other than green algae and transgenic crops, have been developed for the past 20 years and in the past decade progress has been made using whole plants as well as in vitro plant cells for the development of plant-made pharmaceuticals (PMP) as well as plant-made industrial products (PMIP) [49-51]. Transgenic animals are also an important source for biologically active proteins and development is currently being undertaken to create an effective and cost-efficient method [52-53]

1.5 Yeast as Recombinant Expression Platforms

Similarly to bacterial expression systems one of the key advantages of yeast as a heterologous expression platform is the ability to grow to high cell densities, up to 130 g L⁻¹ (dry cell weight) [54]. Additionally most yeasts are not pathogenic, although Candida albicans is, and they have the capability to secrete efficiently as well as carrying out posttranslational modifications [2]. Several species have been established as industrial expression platform systems including Saccharomyces cerevisiae, Kluyveromyces lactis and Pichia pastoris.
1.5.1 *Saccharomyces cerevisiae*

*Saccharomyces cerevisiae* was originally the preferred choice as a recombinant protein production platform due to the vast wealth of information garnered over decades including a published genome sequence [55-56]. The functionality for *S. cerevisiae* has expanded through the development of the *Saccharomyces* Genome Database, which lists common genes, proteins and is integrated with bioinformatic tools including genome browsers and engine searches [57].

Up until January 2009 18.5% of all approved recombinant pharmaceuticals were generated using *S. cerevisiae* [58]. These proteins include hormones, vaccines and virus-like particles [10]. One of the key advantages of using *S. cerevisiae* is that it has been awarded GRAS status by the FDA [59] and like other eukaryotes, *S. cerevisiae* performs post-translational modifications. Up to 70% of all mammalian proteins that were unable to be expressed in *E. coli* were successfully produced intracellularly using *S. cerevisiae* [60]. Titre has been recorded up to 9 g L\(^{-1}\) for the expression of glucose oxidase from *A. niger* [61].

Unfortunately there are disadvantages with *S. cerevisiae*, predominantly *S. cerevisiae* hyperglycosylates proteins resulting in reduced secretion rates [62]. The N-linked glycans in yeast differ from those of mammalian cells and affects the half life of the proteins once administered [63]. The hyper-mannose N-linked glycans result in binding to mannose receptors in humans and therefore reduces the available protein [64]. Therefore, although *S. cerevisiae* holds potential for the production of intracellularly recombinant proteins or for the production of those that do not require glycosylation, the ability of this yeast to compete with CHO cells is limited.

1.5.2 *Pichia pastoris*

In the last 15 years *P. pastoris* has surpassed *S. cerevisiae* as the preferred yeast recombinant expression system due to the reported high titres produced. *P. pastoris*, unlike other yeast, does not ferment sugar; thus eliminating the toxic fermentative product ethanol. The lack of toxic products results in very high cell densities; dry cell weights (DCW) can reach up to 130 g L\(^{-1}\), similar to *S. cerevisiae* which will reach up to 140 g L\(^{-1}\), though *E. coli* is limited to 50 g L\(^{-1}\) [65-67]. Furthermore *P. pastoris* is a methylotrophic yeast, i.e. capable of using methanol as its sole carbon source. Growth on methanol is inducible by a variety of genes including alcohol oxidase I &II (*AOX1*). *AOX1* is a particularly strong promoter and proteins expressed from this promoter can comprise up to 30% of the biomass [12]. Huge investment has been undertaken to develop this recombinant protein production system. A clear advantage of using *P. pastoris* over *S. cerevisiae* is that it does not glycosylate as extensively; thus reducing the risk of immune activation [68].
\textit{P. pastoris} has flaws that need to be addressed in order to stamp its authority in the pharmaceutical industry. Fundamentally despite high volumetric productivity, which results in the large protein production, cell specific productivity is relatively low. It is hypothesised that this predominantly stems from bottlenecks in the secretory pathway, which, if they can be resolved could lead to \textit{P. pastoris} becoming a key platform for expression of recombinant proteins [69-72].

1.6 The History of \textit{P. pastoris}

In 1969 a Japanese group first described a yeast that was capable of using methanol as its sole carbon source [73]. At the time yeast that could grow on such a low cost carbon source was beneficial and in the 1970s Phillips Petroleum invested millions of dollars into the development of \textit{P. pastoris} as a single cell protein for use in animal food source. However, in the 1970s with the oil crisis the cost of methanol rose extortionately making this an unviable option and other foodstuffs such as soybeans were developed preferentially. In the 1980s the decision was made by Phillips Petroleum and Salk Institute Biotechnology/Industrial Associate Inc (SIBA; California, USA) to invest in using \textit{P. pastoris} for the production of recombinant protein [74-75]. In 1993 the technology was sold to Research Corporation Technology (RCT; Tucson, USA) and simultaneously the license to Invitrogen Corporation (California, USA) to produce an expression kit for easy distribution [75]. It was this decision to create a quick and effective expression kit that guaranteed the success of \textit{P. pastoris} as a heterologous protein expression platform.

1.7 \textit{P. pastoris} as a Methylotrophic Yeast

\textit{P. pastoris}, a methylotrophic yeast, is such a popular choice for heterologous protein expression due to its ability to utilise methanol as the sole carbon source [65, 76]. Using methanol has two main advantages; firstly as a carbon source it is inexpensive and secondly the alcohol oxidase gene (\textit{AOX}) promoter allows for the tightly regulated expression of the recombinant proteins [77] (Fig 1.1).
This complex metabolic pathway contains many genes that have been developed for use of their promoters due to their strength but none more so than the alcohol oxidase genes ($AOX$). The effect of using methanol as the sole carbon source is visible when looking at the cells under an electron microscope as the number of peroxisomes, where the methanol is brought into the cell, accumulates [78]. Peroxsiomes are vacuole-like organelles found in nearly all eukaryotes and are required for methanol assimilation in methylotrophic yeast [79]. Upon the switch from methanol to another carbon source peroxisomes are degraded by an autophagy-like process referred to as pexophagy [80].

### 1.7.1 The Alcohol Oxidase ($AOX$) genes

Growth on methanol is conferred by a specific set of metabolic enzymes, which are regulated according to induction by methanol. The alcohol oxidase genes ($AOX1$ and $AOX2$) are switched off when cells are grown in the presence of glucose or other carbon sources but upon exposure to methanol expression increases to the extent that it contributes up to 30% of the total cellular protein [65, 76, 81]. $AOX$ is the first enzyme in the methanol activation utilisation pathway, but it is encoded by both the $AOX1$ and $AOX2$ [82]. The $AOX1$ and $AOX2$ genes share 92 and 97% nucleotide and amino acid sequence identity, respectively, with the main differences located in
the 5’ end of the protein-coding proportion of the genes and very different promoters [65, 82-83].

The main advantage relates to the strength of the promoter and the fact that it is an inducible and tightly regulated [74]. Cultures can be grown in the presence of a non-methanol carbon source to high cell densities; once this has been achieved induction via methanol activates transcription. In the absence of methanol AOX1 mRNA is not detectable although can reach ~5% of the total mRNA once induced [84]. Furthermore, the strength of the promoter results in high protein titre, even if they are toxic to the cell [65, 74, 76, 84]. The disadvantages to using the AOX1 promoter include that methanol is a toxic and highly flammable substance and as a result new promoters are being investigated that do not rely on this potentially hazardous feed [85].

1.7.2 Methanol Utilisation Phenotypes

Depending on the expression the AOX genes, three methanol utilisation (Mut) phenotypes exist with varying ability to grow in the presence of methanol[86]. Interruption of the AOX1 and/or AOX2 genes will impact the ability to grow on methanol resulting in different Mut phenotypes. The most common phenotype employed by researches is Mut+, in which the AOX1 and AOX2 promoter and gene remain intact [87]. In contrast, MutS (slow) uses the secondary and less-well expressed AOX2 gene and promoter, with the AOX1 gene being either disrupted in the parent strain or during recombinant strain production. While this strain has the ability to grow in the presence of methanol growth is substantially reduced. The final phenotype is Mut− which is unable to grow in the presence of methanol due to a disruption in both the AOX1 and AOX2 genes. The choice of Mut phenotype is often dependent on the protein being produced, for instance hepatitis B surface antigen expresses most efficiently using the MutS strain [74].

1.8 Strains of P. pastoris

As with most industrially relevant species P. pastoris has been genetically modified to create different strains that can be used for recombinant engineering [76] such as knockouts in metabolic pathways for auxotrophic selection, i.e. histidine (HIS4) or arginine (ARG4) [65]. The selection of the expression strain is critical for the successful production of heterologous protein in P. pastoris. All expression strains are derived from the parental NRRL-Y 11430 (Northern Regional Research Laboratories, Peoria, IL; Table 1) [75].
Table 1.1 *P. pastoris* host strains

<table>
<thead>
<tr>
<th>Strain</th>
<th>Genotype</th>
<th>Media Supplementation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y-11430</td>
<td>Wild type</td>
<td>None</td>
<td>NRRL</td>
</tr>
<tr>
<td>X33</td>
<td>Wild type</td>
<td>None</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>GS115</td>
<td>his4</td>
<td>Histidine</td>
<td>[84]</td>
</tr>
<tr>
<td>GS190</td>
<td>arg4</td>
<td>Arginine</td>
<td>[88]</td>
</tr>
<tr>
<td>JC220</td>
<td>ade</td>
<td>Adenosine</td>
<td>[88]</td>
</tr>
<tr>
<td>JC254</td>
<td>ura3</td>
<td>Uridine/Uracil</td>
<td>[88]</td>
</tr>
<tr>
<td>GS200</td>
<td>arg4 his4</td>
<td>Arginine/Histidine</td>
<td>[84]</td>
</tr>
<tr>
<td>KM71</td>
<td>Δaax1::SARG4 his4 arg4</td>
<td>Histidine</td>
<td>[89]</td>
</tr>
<tr>
<td>SMD1168</td>
<td>Δpep4::URA3 his4 ura3</td>
<td>Histidine</td>
<td>[90]</td>
</tr>
<tr>
<td>SMD1163</td>
<td>pep4 prb1 his4</td>
<td>Histidine</td>
<td>[90]</td>
</tr>
</tbody>
</table>

1.8.1 Y-11430

The parental strain Y-11430 has been deposited in the Northern Regional Research Laboratories (NNRL; Peoria, IL). It is from this strain that two of the most popular strains, GS115 and X33, were derived. In 2011 this strain was sequenced in order to determine key aspects of secretion that had yet to be established, such as identifying an alternative to the *S. cerevisiae* alpha pre-pro signal [91].

1.8.2 X33

Similarly to Y-11430, X-33 is also a wild-type strain, although it was created by using GS115 (his4) and re-complementing with the wild-type HIS4 gene [65]. This prototrophic strain is the wild-type strain that is supplied by Invitrogen Corporation. As a wild-type strain there is no media supplementation required for growth.

1.8.3 GS115

GS115 is one of the most common strains used for expression studies in *P. pastoris*; it is a histidinol dehydrogenase (his4) mutant and thus requires histidine supplementation when grown on minimal media. This strain is also supplied by Invitrogen Corporation, when purchasing the *Pichia* Expression Kit. One of the reasons GS115 is now such a popular strain is because in 2009 the genome sequence was released [92]. The presence of a genome sequence has enabled faster development of strategies for heterologous protein expression. The mutation in the HIS4 gene allows for complementation during recombination to allow for selection of vectors containing the HIS4 gene. Complementation selection is minimal media grown in the absence of histidine.
1.8.4 Protease Deficient Strains

An increasingly popular choice for expression strains are the protease deficient strains. These strains reduce the amount of degradation of secreted protein due to knockouts of either the proteinase A (*PEP4*) or proteinase B (*PRB1*) [90, 93]. In the absence of proteinase A, Prb1p is not fully active and as a result SMD1168 is also expected to have a lower proteinase B activity. Like Y-11430 and GS115, SMD1168 was recently sequenced, making it a more popular choice for the protease deficient strains [93]. However, it has been reported that all protease deficient strains are less robust and have low storage viability [94]. SMD1168 can, like the other strains sequenced, be expressed with a Mut⁺ phenotype, whereas KM71 is a Δaox1 knockout (where the AOX1 gene is replaced with ARG4 from *S. cerevisiae*), and thus always results in a MutS or Mut⁻ phenotype.

1.8.5 The truth about *Komagataella phaffi* and *Komagataella pastoris*

In 1995 Yamada *et al.* noted that 18S and 26S sequences from *P. pastoris* strains were significantly different from other methanol assimilating yeast [95]. From this a new genus was proposed, *Komagataella*, to which *P. pastoris* should be added. However, in 1998 Kurtzman rejected a new genus due to the lack of species that would be distinctly varied when looking at phylogenetic trees [96]. Nevertheless, in 2003 Dlauchy *et al.* described *Pichia pseudopastoris*, a species that is closely related to *P. pastoris* and with this it was possible to establish *Komagataella* as a distinct genus [97]. To date gene sequences have confirmed the presence of two species within the *Komagataella* genus; *K. pastoris* and *K. phaffi* [98-99].

Kurtzman undertook the investigation to establish which *P. pastoris* strains belonged to the newly classified species; characterised by phylogenetic trees using sequences from domains 1 and 2 of the nuclear large subunit ribosomal RNA, mitochondrial small unit rRNA, translation elongation factor 1-α (ER-1α) and RNA polymerase I. It was determined that NRRL Y-11431 (a variant wild-type strain) belonged to *K. pastoris*, whereas NRRL Y-11430 (the parental wild type strain of GS115 and X33) belonged to *K. phaffi* [99]. The sequencing data collected by Mattanovich *et al.* (2009) was done so on SMD1168, which is a *K. pastoris* strain [93]. The Invitrogen expression kit uses *K. phaffi* as opposed to *K. pastoris*, purely by coincidence as the development for biotechnological expression was chosen at random.

However, although it has been acknowledged that *P. pastoris* now exists as two separate species under *K. pastoris* and *K. phaffi*, this nomenclature has yet to be adopted into the literature. As a result to keep in line with current publications for the remainder of the thesis, *P. pastoris* will continue to be used.
1.8.6 Humanized *P. Pastoris* strains

In 2006 a major breakthrough was achieved for the use of *P. pastoris* as an expression host for production of glycoproteins for pharmaceutical uses; the advent of a humanized strain, yAS309 [100]. This glycoengineered strain was modified to perform human-like N-glycosylation. The initial step was to remove the α-1,7 mannose extension, which is responsible for the hypermannosylation in yeast. The α-1,2 mannosidase from *Trichoderma reesei* was fused to an HDEL-retention signal [101]. Jacobs *et al.* described a method for creating complex-type N-glycosylation using GlycoSwitch technology, using five GlycoSwitch Vectors [102]. This humanized *P. pastoris* strain was used to successfully produce sialylated human-type recombinant erythropoietin [100].

In 2009 Potgieter *et al.* reported the production of 1 g L\(^{-1}\) of functional monoclonal antibody with uniform N-linked glycans [103]. Furthermore Ye *et al.* (2011) designed an optimal feeding strategy that resulted in up to 1.6 g L\(^{-1}\) of monoclonal antibody in a scaled-up process to a 1200 L scale [104]. The development of a humanized strain allows for the cultivation of commercially available monoclonal antibodies and raises the profile of *P. pastoris* as a true competitor to CHO cells.

1.9 Expression Vectors for *P. pastoris*

When Phillips Petroleum passed over the rights to Invitrogen Corporation in 1993 an aggressive attitude was taken to the creation and distribution of cheap expression kits [105]. One of the key aspects of this was designing a library of expression vectors that could be used under varying conditions. Almost all of the vectors are *E. coli/P. pastoris* shuttle vectors, containing an origin of replication in *E. coli* and a functional selection marker in both organisms.

1.9.1 Promoters

One of the key aspects in vector selection is the choice of promoter. There are several promoters that are frequently used in *P. pastoris*, most commonly AOX1 promoter as previously described [65]. The strength of the promoter is one of its main advantages; however particularly for intracellular proteins the strength of the AOX1 promoter can be detrimental; thus other weaker promoters are sometimes preferred. The AOX2 promoter, which is also inducible by methanol, has been reported to have lower activity than the AOX1 promoter [106]. Strains with a disrupted AOX1 gene contained approximately one third of the AOX activity compared to wild-type strains [82]. Another inducible promoter is the formaldehyde dehydrogenase (FLD1) gene promoter, which is strongly induced using methanol as a carbon source or methylamine as a nitrogen source [107].

While these examples are strong inducible promoters the induction via methanol can be construed as detrimental as some industries (such as the food industry) are unable to use
methanol due to fire and health hazards. Additionally switching from a glycerol feed to a methanol feed during fermentation can be inconvenient; thus other non-utilising methanol promoters have been developed. Of these, the most common one is the glyceraldehydes-3-phosphate dehydrogenase (GAP) gene promoter [85], a constitutive promoter. Unlike the AOX1 promoter driven vectors, which are grown on a glucose carbon source initially to achieve high cell densities before they are induced with methanol, the GAP promoter will start producing proteins at the same time as growing. On methanol the GAP promoter will secrete approximately 50-65% less protein than using the AOX1 promoter [108].

1.9.2 Selectable Markers

In order for the shuttle vectors to work in both E. coli and P. pastoris selection markers must exist that work in both species. One of the most popular antibiotic selection markers is Zeo, using the Sh ble gene from Streptalloteichus hindustanus [109-110]. The reason for its popular nature is the fact that it works in both E. coli and P. pastoris.

Another popular method of selection is to use biosynthetic markers; however such selection will only occur in P. pastoris. As a result the vector will contain an E. coli selection marker, most commonly the Bleomycin (bla) gene, which confers resistance to ampicillin [110-111]. Five biosynthetic markers are currently available commercially these include; HIS4, ADE1 (PR-amidoimidazolesuccinocarboxamide synthase), ARG4 (arginosuccinate lyase), URA3 (orotidine-5’-phosphate decarboxylase) and URA5 (orotate phosphoribosyltransferase) [65].

Other selection methods currently being used include the use of a modified Tn903kan’ gene that allows for direct selection on both P. pastoris using G418 (Geneticin) and kanamycin on E. coli [112-113].

1.9.3 Secretion Signal

P. pastoris secretes very few native proteins therefore purification can be made easier by signalling recombinant proteins to the supernatant. The most commonly used secretion signal is the α-mating factor (α-MF) pre-pro peptide from S. cerevisiae [75]. Following the sequencing of NRRL-Y11430 the P. pastoris α-MF prepro peptide equivalent has been identified but has yet to be evaluated [91, 114]. The secretion signal from the native acid phosphatase (PHO1) isolated from P. pastoris is also frequently used with successful secretion [75]. More recently PHA-E from the plant lectin Phaseolus vulgaris agglutinin has been used for the expression of green fluorescent protein (GFP) as well as two plant lectins [115]. This has been reported to work in
situations where the Kex2 site is not properly cleaved resulting in N-terminus ends of the α-MF signal peptide attached to the recombinant protein [65].

Additionally if the existing secretion signals have not worked it is possible to develop synthetic leader sequences [116-117]. In 2011 Kottmeier et al. undertook an investigation into using three novel secretion signals from hydrophobins of Trichoderma reesei [118]. Due to the length of the existing secretion signals research was undertaken to reduce the amino acid length to less than 25. The group claimed that they were successfully able to include a secretion signal through the use of PCR primers and successfully secrete enhanced green fluorescent protein (EGFP).

1.10 Homologous recombination

One of the advantages of P. pastoris is the highly recombinogenic nature and as such vectors created with regions of homology can often be used to target integration at a specific gene of interest [119]. Depending on the design of a vector, homologous recombination can either result in gene insertion, through a single crossover event, or gene replacement through a double crossover event. A single crossover event can occur in the AOX1 region by linearising the plasmid so that integration only occurs in the 5’ AOX1 region (Fig 1.2).

![Figure 1.2 Single Crossover Homologous Recombination](image)

Integration occurs through linearising the vector in the 5’ AOX1 region to create two copies of the AOX1, each made up partially from the genome AOX1 and partly from the plasmid counterpart.

Single crossovers will create a Mut+ phenotype when used for integration into the AOX1 locus. In order to create gene replacements a double crossover event must be used. In the example of integration into the AOX1, the vector can be cut with BgIII to linearise the plasmid at the very
start of the 5’ AOX1 fragment, the second crossover will occur downstream at the 3’ AOX1 region; thus completely replacing the genome fragment (Fig 1.3).

![Figure 1.3 Double Crossover Homologous Recombination](image)

Integration occurs through linearising the plasmid at the very beginning of the 5’ AOX1 region of homology; thus forcing the second recombination event to occur at the 3’ AOX1 region. This double crossover will create a gene replacement.

For auxotrophic strains regions of homology, such as to the HIS4 gene, this will result in complementation with an additional advantage of being used as a selection method [84, 112]. However, in the example of auxotrophic strains that contain either the GAP promoter or AOX1 promoter integration can occur into these regions. If integration has not occurred at the designed site (based on the location of the linearisation of the vector) then it is important to check other regions where homology exists in the vector [120]. However, non-homologous recombination can occur in *P. pastoris*, and it has been noted that certain loci result in better protein secretion [121].

The homologous recombination pathway is well studied in *S. cerevisiae* and it is this in-depth knowledge that enables parallels to be drawn to *P. pastoris* (Fig 1.4).
Figure 1.4 Homologous Recombination Pathways in *S. cerevisiae*
An induced double stranded break may illicit one of three repair pathways in *S. cerevisiae*; double-strand break repair (DSBR), synthesis-dependent strand annealing (SDSA) and break-induced replication (BIR).

Once the linear piece of DNA containing the gene of interest has integrated successfully into the genome; either through a single or double crossover event then the *P. pastoris* cells are capable of turning into microbial factories in order to produce the recombinant protein.

### 1.11 pPICz and pPICzα

Examples of vectors sold by Invitrogen Corporation are pPICz and pPICzα. These vectors use the *AOX1* promoter and rely on the *Sh ble* gene for resistance to the antibiotic Zeocin. The vectors come in three frames; A, B and C, so the end user can clone the gene of interest into the correct reading frame. The distinguishing difference between the two vectors is the fact that the pPICzα vector contains the *S. cerevisiae* α-MF signal peptide for secretion.
To ensure production of the protein of interest a Kozak consensus sequence is required before the ATG of the protein. However in the case of pPICzα, the Kozak consensus sequence is already integrated into the vector in front of the alpha factor signal peptide [122-124]. Both pPICz and pPICzα are designed for integration into the AOX1 region, although the precise location varies depending on the restriction digest. If the vectors are digested with BglII which is located at the 5’ end of the promoter region, AOX1 gene replacement is possible and a MutS phenotype can be achieved. With all other restriction digest locations a Mut’ phenotype occurs, unless the selecting strain already contains an AOX1 knockout (Fig 1.5).

**Figure 1.5 pPICz and pPICzα Plasmid Maps**

**A)** pPICz. **B)** pPICzα. Both vectors contain the AOX1 promoter, the Sh ble gene for Zeocin resistance (ZeoR), as well as the pUC origin of replication for growth in *E. coli*. pPICzα contains the additional α-MF signal peptide for secretion. Both the BglII and PmeI restriction sites are represented as these are popular sites for linearisation for homologous recombination.

### 1.12 Production of Recombinant Proteins in *P. pastoris*

*LacZ* was the first gene to be expressed in *P. pastoris*, which encodes β-galactosidase [89]. As of 2000 over 200 proteins have been produced in *P. pastoris*, all from a wide range of hosts including bacteria, fungi, protists, plants, animals viruses and humans [125]. In addition to proteins, *P. pastoris* is being used to produce enzymes and antibodies. Proteins may be produced intracellularly, a trait important for the production of membrane proteins, or they can be secreted with the addition of a signal peptide. The flexibility of the *P. pastoris* system has opened the doors for scientists to take advantage of this high secreting strain in order to produce large quantities of protein.

In 2004 human insulin was first manufactured globally by Biocon, who aptly named it INSUGEN® [126]. Since then other proteins produced by *P. pastoris* been added to the global market such as recombinant human serum albumin (HSA), alternatively known as Medway, which is used as a blood expander [127]. However; it was not until 2009 that FDA approved the first protein
produced from *P. pastoris* for therapeutic purposes. Dyax Corporation (Burlington, MA, USA), received approval for their drug KALBITOR® (ecallantide), which is used in the treatment of acute attacks of hereditary angioedema [128].

A brief summary of proteins produced in Table 1.2 highlights the powerful tool that this organism has become.

**Table 1.2 Overview of Protein Production in *P. pastoris***

<table>
<thead>
<tr>
<th>Protein Expressed</th>
<th>Function</th>
<th>Expression Level</th>
<th>Available on the market?</th>
<th>Product Name</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Escherichia coli</em> phytase</td>
<td>Used in animal feed industry to improve phosphorous utilisation</td>
<td>6.4 g L(^{-1})</td>
<td>Phytex LLC (India)</td>
<td>Recombinant Phytase</td>
<td>[129]</td>
</tr>
<tr>
<td><em>Thermus aquaticus</em> YT-I aqualysin I</td>
<td>Heat-stable subtilisin-type serine protease</td>
<td>1 g L(^{-1})</td>
<td>No</td>
<td>N/A</td>
<td>[130]</td>
</tr>
<tr>
<td><em>Trametes versicolor</em> laccase (lcc I)</td>
<td>Phenoloxidase</td>
<td>23.9 U mL(^{-1})</td>
<td>No</td>
<td>N/A</td>
<td>[131]</td>
</tr>
<tr>
<td>Anti-IL6 Receptor</td>
<td>Single domain antibody fragment for rheumatoid arthritis treatment</td>
<td>30 mg L(^{-1})</td>
<td>Albynx (Belgium)</td>
<td>Nanobody® ALX-0061</td>
<td>[132]</td>
</tr>
<tr>
<td>Mouse endostatin</td>
<td>Tumour growth suppressor</td>
<td>133 mg L(^{-1})</td>
<td>No</td>
<td>N/A</td>
<td>[133]</td>
</tr>
<tr>
<td>Anti-HBs Fab fragment</td>
<td>Prevention and treatment of Hepatitis B virus</td>
<td>800 mg L(^{-1})</td>
<td>Shantha/Sanofi (India)</td>
<td>Shanvac™</td>
<td>[134]</td>
</tr>
<tr>
<td>Trypsin</td>
<td>Digestion of proteins</td>
<td>14.4 U mL(^{-1})</td>
<td>Roche Applied Science (Germany)</td>
<td>Recombinant Trypsin</td>
<td>[135]</td>
</tr>
<tr>
<td>Interferon-alpha 2b</td>
<td>Hepatitis C and cancer treatment</td>
<td>200 mg L(^{-1})</td>
<td>Shantha/Sanofi (India)</td>
<td>Shanferon™</td>
<td>[136]</td>
</tr>
<tr>
<td>Interleukin-22</td>
<td>Novel human cytokine important for promoting antimicrobial defence and preventing epithelial damage</td>
<td>100 mg L(^{-1})</td>
<td>No</td>
<td>N/A</td>
<td>[137]</td>
</tr>
<tr>
<td>Insulin</td>
<td>Treatment of diabetes mellitus</td>
<td>250 mg L(^{-1})</td>
<td>Biocon (India)</td>
<td>Insugen®</td>
<td>[138]</td>
</tr>
<tr>
<td>Serum albumin</td>
<td>Binding and transport, colloid osmotic pressure</td>
<td>10 g L(^{-1})</td>
<td>Mitsubishi Tanabe Pharma (Japan)</td>
<td>Medway</td>
<td>[139-140]</td>
</tr>
</tbody>
</table>
1.13 Increasing the yield of *P. pastoris*

A key limiting factor of protein expression in *P. pastoris* is the low cell specific productivity. This aspect is countered via the high volumetric productivity, which is bolstered by the high cell densities that *P. pastoris* is able to grow to. Nevertheless many different techniques have been used to try to increase the cell specific productivity and increase the overall yield of protein. Tactics include genetic manipulation of the secretion pathways in order to reduce stress levels that can result from secreting proteins, expressing the target protein with a folding chaperone such as protein disulfide isomerase, and perhaps most commonly and for certain one of the earliest methods increasing gene dosage [72, 74, 141]. To date there is no systematic procedure that effectively increases yield, yet each of the techniques outlined have been reported to be successful in increasing yield in specific cases. It is becoming more apparent that individual systems require specific optimisation based on the protein being produced.

1.13.1 Creating Multiple Copy Clones

Multi-(gene)copy clones were originally evaluated by Clare *et al.* in 1991 [142], who determined that an increase to 14 copies from a single copy of the tetanus toxin fragment C results in a 10% increase of heterologous protein to total cell protein [76]. Whilst this method was successful for proteins such as tetanus toxin fragment C the production of other proteins via this method were not as successful, such as with the expression of hepatitis B surface antigen [143]. To date there are several methods used to create multi-copy clones. Traditional selection of multi-copy clones was performed by screening large number of colonies for spontaneously occurring colonies with high gene dosage [142, 144-145]. Initially only two resistant markers were capable of generating multi-copy strains by selection on a higher concentration, these were *Sh ble* (Zeocin<sup>®</sup>) and *BSD* (blastidine<sup>®</sup>) ([144]).

Two other selectable markers that can result in the production of multi-copy clones are more laborious and thus are often not chosen as efficient methods. The *FLD1* gene (resulting in resistance to formaldehyde) requires an initial selection on minimal plates containing methylamine chloride as the carbon source [146]. It is then possible to plate on formaldehyde containing plates in order to select multi-copy clones. Of most concern is that this selection marker is only capable of being used with an *fld1* mutant [147].

The second alternative resistance marker is the *Tn903kan<sup>®</sup>* marker which confers resistance to G418; originally like *FLD1* selection, a pre-selection was required, this time on minimal plates lacking in *HIS4*. The vector originally used containing the *Tn930kan<sup>®</sup>* included a *HIS4* gene for complementation on a *HIS4<sup>®</sup>* strain, before selection could occur on G418 plates. However in 2008
Lin-Cereghino established that the bacterial promoter and transcription terminator region was preventing direct selection on G418 and replacing them resulted in a new generation of G418 vectors capable of direct selection in both *E. coli* and *P. pastoris*, removing the need for a preliminary HIS4 complementation selection [112].

Other methods to create multi-copy clones are being expanded at present including an *in-vitro* multimerisation method suggested by Invitrogen Corporation and a method referred to as post-transformational vector amplification (PTVA).

### 1.14 *In-vitro* Multimerisation

The presence of a commercial kit that specifically aims to increase copy number only highlights the importance placed on the production of multi-copy clones in *P. pastoris*. The premise behind this multi-copy kit sold by Invitrogen Corporation is that multi-copy clones can be made with specific vectors either *in-vivo* or *in-vitro* [148]. Briefly, the *in-vivo* method uses one of two vectors; pPIC3.K or pPIC9K. Both vectors contain the HIS4 gene for complementation, a kanamycin gene for *in-vivo* selection and the AOX1 promoter. pPIC3.K produces proteins intracellularly whilst pPIC9K secretes proteins into the supernatant. The *in-vivo* screening method is carried out by selection of His⁺ transformants (generated from complementation with the active HIS4 gene) and then plated onto varying concentration of G418. This method is the same as what was originally described for the *Tn903Kan* gene before Lin-Cereghino *et al.* (2008) made modifications to allow for direct selection [112].

Unlike the *in-vivo* method which can still require the screening of thousands of colonies the *in-vitro* method should create clones with a more directed copy number and a less intensive screening step. Invitrogen Corporation provides the vector pAO815, which contains the HIS4 gene, the AOX1 promoter and the 3’ AOX1 fragment (to allow for double crossovers for gene replacement). The downside to the *in-vitro* method is that the amount of work to create the correct number of copies can be very intense and furthermore the size of the vector may become very large, which can be detrimental to transformation into *P. pastoris*.

The premise behind the *in-vitro* mechanism is that multimers can be created based on creating expression cassettes through digesting a complete vector (with the required gene of interest) with the restriction enzymes BglII and BamHI. The vector is either digested with BamHI, or with both BglII and BamHI. As BglII (AGATCT) and BamHI (GGATCC) restriction sites differ in the first and sixth base only, the restricted DNA has compatible ends. If ligated together in a head-to-tail direction neither restriction site is recreated (GGATCT). However if multimers form in a head-to-
head or tail-to-tail orientation either the BglII or BamHI site is recreated and the fragment will be cut; thus reducing them to the individual vector and expression cassette fragment (Fig 1.6).

![Diagram of in vitro multimerisation method](image)

**Figure 1.6 Schematic diagram of the method suggested by Invitrogen Corporation for in vitro multimerisation**

Initially an expression cassette is isolated by digestion with BglII and BamHI. These expression cassettes are then ligated together with a linearised BamHI vector to create multimers. This can be repeated multiple times, depending on the success rate and the size of the gene of interest.

This *in-vitro* multimerisation method can be repeated several times to obtain a larger number of gene copies. It should be noted however that the pAO815 vector alone is 7.7 Kb so any additional insertions of the expression cassette will mean that this increases significantly. Nonetheless this method offers an invaluable tool in ascertaining the precise number of copies of the gene of interest (considering recombination does not occur elsewhere within the genome either through homologous recombination at the *HIS4* or non-homologous recombination).

### 1.14.1 Post-Transformational Vector Amplification

In 2008 Sunga *et al.* published a new protocol for the production of multi-copy clones entitled post-transformation vector amplification (PTVA) [149]. This method works on the premise
described by Schimke in mammalian systems whereby a stepwise increase in the concentration of antibiotic the methotrexate results in directed selection of cell lines that contain more copies of the antibiotic resistance marker. The theory follows that amplification in the antibiotic resistance gene is mirrored by amplification of the whole vector and produces more copies of the gene of interest [150]. This method has been applied to strains containing vectors expressing both the Sh ble resistance marker (Zeocin) and the Tn309kan’ gene for selection on G418.

For Zeocin the initial transformation was selected on 100 μg mL⁻¹ and the selected colonies streaked to single colonies three times to ensure a pure culture. Clones were then spotted onto YDP plates containing increasing concentrations of antibiotic in a step-wise manner from 100 μg mL⁻¹ to 2000 μg mL⁻¹ and left to recover for 3-5 days at 30°C. The main advantage of the method is the proportion of cells that contain high copy numbers. Traditional methods, such as selecting directly onto high concentrations of antibiotic, produce less than 1% of colonies containing greater than 10 gene copies (known as “jackpot” colonies) through PTVA this percentage was reported to have risen to 5-6% of all of enriched strains tested.

Since the method was first suggested other groups have used this method to obtain multi-copy strains, for instance porcine insulin precursor (PIP) was reported to have been enriched to 52 copies of the gene in a Mut⁺ strain [151]. Furthermore, the method of gene amplification has been used to increase copy number of clones that are directed into the rDNA locus [152].

1.15 Unfolded Protein Response (UPR)

Secretory stress can be monitored by assessing the upregulation of unfolded protein response (UPR); a conserved cellular stress response observed in most eukaryotes [153]. This adaptive cellular response aims to restore homeostasis to the balance between unfolded protein and the capacity of the ER and aims to prevent an accumulation of unfolded protein which can be detrimental to the cell [154].

In higher eukaryotes there are three pathways capable of inducing UPR all activated via transmembrane ER stress sensors; inositol requiring enzyme 1 (IRE1), PKR-like ER kinase (PERK) and activating transcription factor 6 (ATF6) [155-158]. While all three mechanisms show similarities acting through the oligmerisation of the respective transmembrane domains resulting in the recruitment of mRNA to the ER, the precise mechanism of sensing stress is not well understood [154, 159]. While the PERK and ATF6 pathways are found in mammalian systems these systems are not present in yeast; thus indicating the redundancies of the systems [159]. As only the IRE1 pathway is present in yeast it serves that this pathway will be described in detail.
UPR has been studied intensively in *S. cerevisiae* and three gene products have been identified as requirements for activation; Ire1p, Kar2p and Hac1p [71]. Kar2p binds to unfolded protein; therefore upon an increase in unfolded protein, Kar2p will dissociate from Ire1p. As mentioned, Ire1p is a transmembrane domain that will oligomerise upon the dissociation of Kar2p. When Ire1p oligomerises the endoribonuclease activity at the cytoplasmic side splices an intron from *HAC1* mRNA. Once activated Hac1p is produced and acts as a transcription factor to upregulate further UPR genes, such as Kar2p and protein disulfide isomerase (PDI; Fig 1.7) [70].

![Figure 1.7 A schematic visualisation of the upregulation of the unfolded protein response (UPR) in yeast](Modified from Schroder et al. and Guerfal et al. [156, 160])

Recently it has been identified that in addition to activating UPR through accumulation of misfolded protein, unfolded protein can directly induce the UPR by binding to the core ER-luminal domain (cLD) of Ire1p [161].

As UPR has been extensively investigated in *S. cerevisiae* it is of interest to note that subtle differences exist when compared to *P. pastoris*. The most controversial of these is the presence or absence of an intron in *HAC1* and whether this protein is constitutively expressed or induced as is seen in *S. cerevisiae*. In 2010 and 2011 two papers were published on this subject with conflicting views; the first paper by Guerfal et al. suggests that while the splice sites are present in *HAC1* mRNA in *P. pastoris* no evidence exists of an unspliced form in an unstressed condition [160].
Early in 2011 Whyteside et al. published that in addition to the intron site being identified (and was in fact the same site that had been identified by Guerfal et al.) there was clear evidence that an unspliced variant of the HAC1 mRNA was identified [162]. To date it has yet to be clarified as to whether the intron must be spliced for activation or whether the activated Hac1p is always present. Irrespective of these discrepancies in the functionality of HAC1 there is no doubt as to the importance of attempting to curb the effects of the UPR in order to achieve the highest cell specific productivity.

1.16 Endoplasmic Reticulum Associated Degradation (ERAD) Pathway

While initial upregulation of the UPR can be viewed as an important step for the regulation of protein folding, and potentially be advantageous, prolonged induction will result in the upregulation of the endoplasmic reticulum degradation pathway (ERAD) [163-164]. Induction of ERAD results in retrotranslocation of misfolded proteins to the cytosol for degradation by the proteasome [71, 156, 165-166]. ERAD is dependent upon UPR for activation and UPR results in an increase of the ERAD capacity [163].

The fate of misfolded protein assigned to the ERAD is dependent on where the proteins are misfolded. In yeast there are three membrane machineries that define the different types of ERAD pathways; ERAD-L, ERAD-M and ERAD-C [69, 167-169]. Proteins that misfold in the lumen are degraded by ERAD-L, in the membrane ERAD-M and in the cystol ERAD-C [170]. For proteins that misfold in the lumen the E3 ubiquitin ligase Hrd1p initiates the degradation of the protein through association with Hrd3p. The additional requirements of the ERAD-M compared to the ERAD-C are not currently known, though the Der1p and Usa1p are suspected to be excluded (Fig 1.8) [167, 169].

Figure 1.8 Membrane Protein Complexes that Define the ERAD pathways

ERAD pathways involved in the degradation of proteins. A) ERAD-L pathway through the use of the Hrd1p complex. B) ERAD-C pathway utilising the Doa10p complex. This figure is adapted from Carvalho et al. and Denic et al. [167-168]
The induction of ERAD will have detrimental effects on protein production. Therefore increasing gene dosage may be counterproductive if upregulation of UPR and ERAD is evident.

1.17 Human Serum Albumin – The Model Protein

Many different proteins have been expressed extensively in *P. pastoris* including human serum albumin, human trypsinogen, human lysozyme and porcine insulin precursor [74, 87, 151, 162, 171-173]. To investigate secretion saturation an ideal model protein would not put a strain on the secretory pathway and ideally be one that folds well (to prevent the induction of UPR). The more stable a protein the better it folds and the less stress it puts on the secretory pathway; thus reducing induction of UPR [173]. Human serum albumin (HSA) was selected as the model protein for evaluation throughout this thesis. Production up to 10 g L$^{-1}$ has been reported using *P. pastoris* [140]. The high titres achieved have led to many fusion proteins being produced with HSA in order to increase the secretion of complex proteins [174-176].

HSA is the most abundant protein in the body, is soluble, globular and unglycosylated [63, 177]. It is important therapeutically for stabilizing blood volume in burns and shock patients and is a carrier protein for steroids, fatty acids and thyroid hormones [177-179]. Furthermore HSA is used as an excipient for vaccines or therapeutic proteins and as a result demand for this protein is incredibly high [179-181]. It has been reported that due to the expanding use of HSA the worldwide market stands at approximately 500 tonnes a year; in doses of up to 10 g per dose it can be used to treat severe hyperalbuminemia or traumatic shock [39, 172]. The main source of HSA is plasma HSA (pHSA), which is extracted from blood donations; however there are several problems including the shortage of supply and the risk of the spread of pathogenic viruses [140].

HSA has been produced in a range of expression systems including *B. subtilis* but incorrect processing of the protein resulted in immune rejection [182]. More recently successful large scale production of HSA has been produced using *Oryza sativa* yet no testing has been done on its immunological compatibility [39]. Recombinant HSA was first produced in *P. pastoris* by Kobayashi *et al.* in 1998 and since then developed has continued extensively [172, 182-184].

In 2005 Bipha Corporation (Hokkaido, Japan) became the first company worldwide to produce rHSA from *P. pastoris* on an industrial level, calling this Medway [127, 185]. In 2008 the first clinical study was conducted as a comparison between pHSA and rHSA from *P. pastoris* with favourable results towards the use of the recombinant protein [186]. In terms of use as a model protein HSA is an atypical protein, not due to its ease of expression (which is beneficial) but in the fact it is such a highly charged protein, which may have different effects than a less charged and globular protein [187].
1.18 Aims and Objectives

This thesis aims to explore various factors that can affect the specific productivity of recombinant protein expression using the heterologous expression platform \textit{P. pastoris}. Through examining in detail various aspects that have been reported to show increased titre a better system is hoped to be achieved for streamlined production. Aspects to be investigated will include the use of multi-copy clones, an optimised protein and the effects of clonal variation. In order to achieve these aims the following objectives must be met:

- To determine the effects of multi-copy clones on production levels using the model protein HSA.
- To investigate negative effects of using multi-copy clones such as secretion saturation, induction of the unfolded protein response and genetic instability.
- To establish possible solutions to issues of genetic instability.
- To determine whether clonal variation is caused solely by integration of the vector.
- To observe the effects of clonal variation strains on transcript levels through the use of transcriptomic analysis.
## Materials and Methods

### 2.1 Strains and Plasmids

<table>
<thead>
<tr>
<th>Strain</th>
<th>Description</th>
<th>Source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> DH5α</td>
<td>F, φ80lacZΔM15 Δ[lacZYA-argF]U169 deoR recA1 endA1 hsdR17(rK-, mK+) phoA supE44 λ- thi-1 gyrA96 relA1.</td>
<td>Invitrogen</td>
<td>Invitrogen Catalogue</td>
</tr>
<tr>
<td><em>E. coli</em> JM109</td>
<td>endA1, recA1, gyrA96, thi, hsdR17 (rk-, mk+), relA1, supE44, Δ[lac-proAB], [F', traD36, proAB, lacIqZΔM15].</td>
<td>Promega</td>
<td>Promega Catalogue</td>
</tr>
<tr>
<td><em>E. coli</em> BioBlue</td>
<td>recA1, em&gt;endA1 gyrA96 thi-1 hsdR17(rk-, mK+) supE44 relA1 lac [F' proAB lacIqZΔM15 Tn10(Tet')]</td>
<td>Bioline</td>
<td>Bioline Catalogue</td>
</tr>
<tr>
<td><em>E. coli</em> pPICzA</td>
<td>JM109 (pPICzA:: AOX1 Zeo)</td>
<td>Invitrogen</td>
<td>Invitrogen Catalogue</td>
</tr>
<tr>
<td><em>E. coli</em> pPICzαA</td>
<td>JM109 (pPICzA:: AOX α-MF Zeo)</td>
<td>Invitrogen</td>
<td>Invitrogen Catalogue</td>
</tr>
<tr>
<td><em>E. coli</em> pPICz-HSA</td>
<td>JM109 (pPICzA:: HSA)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td><em>E. coli</em> pα-HSA</td>
<td>JM109 (pICzα:: HSA-SP)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td><em>E. coli</em> pαTX3</td>
<td>JM109 (pPICzα:: Try1 x3)</td>
<td>Imperial College</td>
<td>Dr. Bryn Edwards-Jones, unpublished</td>
</tr>
<tr>
<td><em>E. coli</em> pαTX3R51</td>
<td>JM109 (pPICzα:: Try1 x3 RAD51)</td>
<td>This Study</td>
<td>Unpublished.</td>
</tr>
<tr>
<td><em>E. coli</em> pαTX3R52</td>
<td>JM109 (pPICzα:: Try1 x3 RAD52)</td>
<td>This Study</td>
<td>Unpublished.</td>
</tr>
<tr>
<td><em>E. coli</em> pIB2</td>
<td>MC1061 (pIB2:: HIS4 GAP)</td>
<td>Addgene</td>
<td>[188-189]</td>
</tr>
<tr>
<td><em>E. coli</em> pIBTX3</td>
<td>JM109 (pIB2:: AOX1 Try1 x3 )</td>
<td>This Study</td>
<td>Unpublished.</td>
</tr>
</tbody>
</table>
### Table 2.2 *P. pastoris* strains used in this study

<table>
<thead>
<tr>
<th>Strain</th>
<th>Description</th>
<th>Source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-33</td>
<td>Wild type, biotin auxotroph</td>
<td>Invitrogen</td>
<td>Invitrogen Catalogue</td>
</tr>
<tr>
<td>GS115</td>
<td><em>his</em>4</td>
<td>Invitrogen</td>
<td>Invitrogen Catalogue</td>
</tr>
<tr>
<td>SMD1163</td>
<td><em>pep4 prb1 his</em>4</td>
<td>Invitrogen</td>
<td>Invitrogen Catalogue</td>
</tr>
<tr>
<td>GpPICz-HSA</td>
<td>GS115 (<em>pPICz A; HSA</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GpαH</td>
<td>GS115 (<em>pPICzα A; HSA-SP</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GpαHopt</td>
<td>GS115 (<em>pPICzα A; HSAopt</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GpαTx3</td>
<td>GS115 (*pPICz A; <em>TRY1</em> (3 copies))</td>
<td>Imperial College</td>
<td>Dr. Bryn Edwards-Jones, unpublished</td>
</tr>
<tr>
<td>GpαTx3R51</td>
<td>GS115 (*pPICz A; <em>TRY1</em> (3 copies); <em>RAD51</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GpαTx3R52</td>
<td>GS115 (*pPICz A; <em>TRY1</em> (3 copies); <em>RAD52</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GpIBAOXTx3</td>
<td>GS115 (*pIB2; <em>TRY1</em> (3 copies); <em>HIS4</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GpIBTx3R51</td>
<td>GS115 (*pIB2; <em>TRY1</em> (3 copies); <em>RAD51</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GpIBTx3R52</td>
<td>GS115 (*pIB2; <em>TRY1</em> (3 copies); <em>RAD52</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GS115-HIS4-R52</td>
<td>GS115 (<em>HIS4-RAD51</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GS115-HIS4-R52</td>
<td>GS115 (<em>HIS4-RAD52</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GpGRzαHSA</td>
<td>GS115 (<em>pGRzα; HSA-SP; rDNA locus</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GpGRzαHSAopt</td>
<td>GS115 (<em>pGRzα; HSAopt; rDNA locus</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GpARzαHSA</td>
<td>GS115 (<em>pARzα; HSA-SP; rDNA locus</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GpARzαHSAopt</td>
<td>GS115 (<em>pARzα; HSAopt; rDNA locus</em>)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>GpARzαHSAopt BglII</td>
<td>GS115 (*pARzα; HSAopt; rDNA locus, integrated using BglII)</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
</tbody>
</table>

Promega, Southampton, UK
Invitrogen, Paisley, UK
Bioline, London, UK
### Table 2.3 Plasmids used in this study

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Description</th>
<th>Source</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>pPICz A</td>
<td>Expression vector based on AOX1 promoter; MCS; Sh ble; pUC ori; 5’AOX1-MCS-AOX1 TT</td>
<td>Invitrogen</td>
<td>Invitrogen Catalogue</td>
</tr>
<tr>
<td>pPICzα A</td>
<td>Expression vector based on AOX1 promoter; α-mating factor prepro signal sequence; MCS; Sh ble; pUC ori; 5’AOX1-MCS-AOX1 TT</td>
<td>Invitrogen</td>
<td>Invitrogen Catalogue</td>
</tr>
<tr>
<td>pPICz-HSA</td>
<td>Expression vector based on AOX1 promoter; MCS; Sh ble; pUC ori; 5’AOX1-MCS-AOX1 TT HSA containing native secretion signal</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>pa-HSA</td>
<td>Expression vector based on the AOX1 promoter; α-mating factor prepro signal sequence; MCS; Sh ble; pUC ori; 5’AOX1-MCS-AOX1 TT HSA with native secretion signal removed.</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>pa-HSAopt</td>
<td>Expression vector based on the AOX1 promoter; α-mating factor prepro signal sequence; MCS; Sh ble; pUC ori; 5’AOX1-MCS-AOX1 TT Optimised HSA gene with native secretion signal removed.</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>pa-TX3</td>
<td>Expression vector based on the AOX1 promoter; α-mating factor prepro signal sequence; MCS; Sh ble; pUC ori; 5’AOX1-MCS-AOX1 TT TRY1 gene (3 copies)</td>
<td>Imperial College</td>
<td>Dr. Bryn Edwards-Jones, unpublished</td>
</tr>
<tr>
<td>pIB2</td>
<td>Expression vector based on the GAP promoter AOX1 TT; pMB1 ori; AmpR; HIS4</td>
<td>Addgene</td>
<td>[188-189]</td>
</tr>
<tr>
<td>pIBAOX7x3</td>
<td>Expression vector based on the AOX1 promoter; α-mating factor prepro signal sequence; MCS; pUC ori; 5’AOX1-MCS-AOX1 TT; AmpR; HIS4 TRY1 gene (3 copies);</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>pIBTx3R51</td>
<td>Expression vector based on the AOX1 promoter; α-mating factor prepro signal sequence; MCS; HIS4; pUC ori; 5’AOX1-MCS-AOX1 TT TRY1 gene (3 copies); 500 bp sequence identity to RAD51</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>pIBTx3R52</td>
<td>Expression vector based on the AOX1 promoter; α-mating factor prepro signal sequence; MCS; HIS4; pUC ori; 5’AOX1-MCS-AOX1 TT TRY1 gene (3 copies); 500 bp sequence identify to RAD52</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
<tr>
<td>pGRzαHSA</td>
<td>Expression vector on the GAP promoter, α-mating factor prepro signal sequence; MCS;HSA removing minus the native secretion signal; Sh</td>
<td>This Study</td>
<td>Unpublished</td>
</tr>
</tbody>
</table>
ble; pUC ori; rDNA locus for integration

**pGRzaHSAopt**
Expression based on the GAP promoter; α-mating factor prepro signal sequence; MCS; HSA optimized gene; Sh ble; pUC ori; rDNA locus for integration
This Study  Unpublished

**pARzaHSA**
Expression based on the AOX1 promoter; α-mating factor prepro secretion signal; MCS; HSA removing native secretion signal; Sh ble; pUC ori; rDNA locus for integration
This Study  Unpublished

**pARzaHSAopt**
Expression based on the AOX1 promoter; α-mating factor prepro secretion signal; MCS; HSA optimized gene; Sh ble; pUC ori; rDNA locus for integration
This Study  Unpublished

**pARzaHSAopt BglII**
Expression based on the AOX1 promoter; α-mating factor prepro secretion signal; MCS; HSA removing native secretion signal; Sh ble; pUC ori; rDNA locus (integrated at the BglII site) for integration
This Study  Unpublished

Invitrogen, Paisley, UK

### 2.2 Media

#### 2.2.1 Miller Lysogeny Broth (LB) Media
1% (w/v) Peptone au casein (Merck, Nottingham, UK), 0.5% (w/v) Yeast Extract (Merck, Nottingham, UK), 1% (w/v) NaCl. For agar plates 1.5% (w/v) agar included.

#### 2.2.2 Lennox Lysogeny Broth (LB) Media
For pPICz vectors propagated in *E. coli* low salt LB must be used; 1% (w/v) Peptone au casein (Merck, Nottingham, UK), 0.5% (w/v) Yeast Extract (Merck, Nottingham, UK), 0.5% (w/v) NaCl. For agar plates 1.5% (w/v) agar included.

#### 2.2.3 Yeast Extract Peptone Medium (YPD)
1% (w/v) Yeast Extract (Merck, Nottingham, UK), 2% (w/v) Peptone au casein (Merck, Nottingham, UK), 2% (w/v) Dextrose (glucose). For agar plates 2% (w/v) agar included.

#### 2.2.4 Buffered Glycerol-complex Medium (BMGY) or Buffered-Methanol-complex Medium (BMMY)
1% (w/v) Yeast Extract (Merck, Nottingham, UK), 2% (w/v) Peptone au casein (Merck, Nottingham, UK), 100mM potassium phosphate, pH6.0, 1.34% (w/v) Yeast Nitrogen Base (YNB), 4 x 10⁻⁵% (w/v) d-Biotin, 1% (v/v) glycerol or 0.5% (v/v) methanol.
2.2.5 Minimal Dextrose Medium (MD) or Minimal Methanol Medium (MM)
1.34% (w/v) Yeast Nitrogen Base (YNB), $4 \times 10^{-5}$% (w/v) d-Biotin, 2% (w/v) Dextrose (glucose) or 0.5% (v/v) methanol. Plus or minus $4 \times 10^{-3}$% (w/v) Histidine when using a *his4* mutant strain.

2.2.6 Buffered Minimal Glycerol Medium (BMG) or Buffered Minimal Methanol Medium (BMM)
1.34% (w/v) Yeast Nitrogen Base (YNB), $4 \times 10^{-5}$% (w/v) d-Biotin, 1% (v/v) glycerol or 0.5% (v/v) methanol. Plus or minus $4 \times 10^{-3}$% (w/v) Histidine when using a *his4* mutant strain.

2.2.7 Tryptone Soy Agar (TSA)
1.7% (w/v) Peptone au casein (Merck, Nottingham, UK), 0.3% (w/v) Select Soytone (Merck, Nottingham, UK), 0.3% (w/v) Dextrose (glucose), 0.5% (w/v) Sodium Chloride, 0.25% (w/v) Dipotassium Phosphate, 1.5% (w/v) agar.

2.2.8 *Pichia* Cell Bank Medium (PCMB)
3% (w/v) Yeast Extract (Merck, Nottingham, UK), 1.8% Dextrose (glucose), 1.74% (w/v) Dipotassium phosphate, 0.01% (w/v) d-Biotin. Adjusted to pH 6.0 ± 0.2 with 85% (v/v) Phosphoric Acid

2.2.9 *Pichia* Freezer Mix (PFM) – Cryopreservation Medium
8.5% (v/v) Glycerol, 1.5% (w/v) Yeast Extract (Merck, Nottingham, UK), 2.18% (w/v) Dipotassium Phosphate (adjust pH to 6.0 ± 0.2 using 85% (v/v) Phosphoric Acid), 1.8% (w/v) Dextrose (glucose), $2 \times 10^{-4}$% (w/v) d-Biotin.

2.2.10 Fermentation Inoculation Media
2.2.10.1 Yeast Extract/Peptone Media
0.495% (w/v) Yeast Extract (Merck, Nottingham, UK), 0.99% (w/v) Select Soytone (Merck, Nottingham, UK), $4.5 \times 10^{-3}$% (v/v) Polypropylene Glycol 2000. pH adjusted to 6.0 ±0.2 at 20-25°C using 85% (v/v) Phosphoric acid.

2.2.10.2 Glucose Solution
2.2% (w/v) Glucose monohydrate, filter sterilised.

2.2.11 Fermentation Media
A complex fermentation medium was used which is the proprietary intellectual property of Fujifilm Diosynth Biotechnologies.
2.3 Molecular Biology Methods

2.3.1 Agarose Gel Electrophoresis

Agarose gel electrophoresis was employed to separate DNA and RNA fragments based on size. Typically 1% (w/v) agarose gels were used, with the exception of small fragments, such as RAPD-PCR products where a 2.5% (w/v) agarose gel was used instead. Agarose was dissolved in 1x TAE buffer (50X TAE Buffer; 242 g L⁻¹ Tris base, 57.1 mL L⁻¹ glacial acetic acid, 100mL L⁻¹ 0.5 M EDTA, pH 8.0) and heated until all the agarose had melted before being left to cool to approximately 50°C. SYBR® Safe (Invitrogen, Paisley, UK) was diluted 1 in 20,000 in the agarose and poured into a mould and left to solidify.

DNA samples were diluted into 5X Loading Buffer (30% [v/v] glycerol, 0.25% [w/v] bromophenol blue) and loaded onto the gel. 5 μL of molecular weight marker was added to one of the lanes, typically GeneRuler™ 1kb DNA Ladder (Fermentas, York, UK; Fig 2.1) was used.

![Figure 2.1 DNA Molecular Marker Size](image)

5 μL of GeneRuler™ 1 kb DNA ladder run on a 1% (w/v) agarose gel. Size of bands is indicated by annotations (bp) and relative quantity (ng) per each 5 μL.

Agarose gels were typically run in Bio-Rad Wide Mini-Sub Cell GT Tanks at 100V using power supplied from a Bio-Rad Power Pac Basic (Bio-Rad, Hemel Hempstead, UK). DNA was visualised using a short wave-length UV transilluminator and photographed using a Syngene G:Box ChemiHR system (Syngene, Cambridge, UK). The G:Box relies on GeneSnap for the acquisition of pictures and GeneTools for all analytical purposes. Once collected images were printed on a Mitsubishi P93D thermal printer (Syngene, Cambridge, UK).
2.3.2 Plasmid DNA Isolation

Plasmid DNA was isolated from *E. coli* using either a Qiagen Spin Mini-prep Kit or hi-speed Midi-prep kit (Qiagen, Crawley, UK), depending on the amount of DNA required, following the manufacturer’s protocol. Both protocols use alkaline lysis and SDS-precipitation.

2.3.3 Isolation of Yeast Genomic DNA

Genomic DNA was prepared using one of three methods. For qPCR the DNeasy Plant Mini Kit (Qiagen, Crawley, UK) was used according to the manufacturer’s protocol. For RAPD-PCR and other basic PCR, genomic DNA was extracted using MasterPure Yeast DNA purification Kit according to the manufacturer’s protocol (Cambio, Cambridge, UK).

For Southern blots, genomic DNA was isolated based on the method of Philippsen [190]. All centrifugation steps were performed at 4°C at 4000 rpm. A 10 ml culture of *P. pastoris* was grown in YPD overnight at 30°C at 250rpm in 50 mL centrifuge tubes. Cells were centrifuged for 10 minutes, resuspended in 10 ml water and centrifuged for a further 10 minutes. Cells were resuspended in 3 mL of 0.9 M sorbitol, 0.1 M EDTA and 14 mM β-mercaptoethanol. Two hundred units of lyticase (Sigma-Aldrich, Dorset, UK) were added and cells left to form spheroplasts at 37°C for 2 hours. Spheroplasts were centrifuged for 5 minutes and the supernatant discarded. The pellet was resuspended in 3 mL of 50 mM Tris-HCL, 50 mM EDTA, pH 8.0, mixed with 0.3% SDS and incubated at 65°C for 30 minutes. 1 mL 5 M KOAc was added and left on ice for 60 minutes. Samples were centrifuged for 30 minutes, the supernatant transferred to a 15 mL centrifuge tube to which 4 mL ice-cold absolute ethanol was added and the tube centrifuged for 10 minutes. Supernatant was discarded, the pellet washed with 70% (v/v) ethanol and centrifuged for a further 10 minutes. The pellet was resuspended in 300 µL TE, pH 7.5 then 15 µL of 10 mg mL⁻¹ DNase-free RNase was added and the mixture incubated for 30 minutes at 37 °C. 300 µL phenol-chloroform (1:1) was added and mixed by inverting the tube, which was then centrifuged for 10 minutes. Supernatant was removed and transferred to a 1.5 mL eppendorf tube; 15 µL of 3 M NaOH and 700 µL isopropanol was then added. The sample was centrifuged for 10 minutes and the pellet washed in 80% (v/v) ethanol and air dried before resuspending in a final volume of 100 µL TE, pH 7.5.

2.3.4 DNA quantification

DNA and RNA was quantified by absorbance measurements at 260/280 nm. A sample of 5 µL of DNA was added to 45 µL of water and loaded onto a Grenier Bio-One 384 well plate (Scientific Laboratory Supplies, Nottingham, UK). The absorbance was measured at 260 nm and 280 nm using a Synergy HT multi-detection microplate reader (Bio-Tek, Potton, UK). Concentration of
DNA was calculated from the absorbance at 260 nm, while protein contamination was determined by the ratio of absorbance at 260/280 nm.

Quantities were also estimated based on comparison of the fluorescence of bands (of known amount) of the DNA molecular size ladder and the sample of interest, on agarose gels.

2.3.5 DNA Purification and Concentration

DNA was concentrated using the Zymo Clean and Concentrate kit (Cambridge BioScience, Cambridge, UK) following the spin column instructions from the manufacturer’s manual. PCR products were purified using the microcentrifuge protocol in the Zymo Gel Extraction and Purification Kit (Cambridge BioScience, Cambridge, UK).

2.3.6 Restriction Enzymes

Restriction enzymes were purchased from New England Biolabs (Hertfordshire, UK), Promega (Southampton, UK) or Fermentas (York, UK) and digestions were performed according to the manufacturer’s instruction manual, typically at 37°C for 3 hours. Double digests were carried out, where appropriate, using multi-core buffer Tango as provided by Fermentas. Digested products were purified on agarose gels or using spin columns.

2.3.7 HSA Codon Optimised Gene

The HSA gene was codon optimised and synthesised by GenScript (Piscataway, NJ, USA) using the algorithm OptimumGene™ [191].

The complete gene sequence can be seen in Appendix 9.1.

2.3.8 Ligation

Ligation was carried out using T4 DNA Ligase (Fermentas, York, UK) according to the manufacturer’s instructions. The molar ratio of plasmid to insert was typically 3:1, respectively, for cohesive end ligation. The ligation mixture was incubated for 1 hour at room temperature or overnight at 4°C.

2.3.9 Easy Cloning Vectors

pGEM®-T Easy Vector System (Promega, Southampton, UK) was used for T/A cloning according to the manufacturer’s protocols. The reaction was set up as 5 μL 2X rapid ligation buffer, 1 μL pGEM®-T Easy Vector, 3 μL PCR product and 1 μL T4 DNA Ligase, and incubated for 1 hour at room temperature. The ligation mixture was then transformed into JM109 chemically competent cells. The pGEM®-T Easy Vector contains the lacZ gene, which allows for blue white colony screening when spread onto plates containing X-Gal.
For blunt end cloning CloneJET™ PCR Cloning Kit (Fermentas, York, UK) was used. The reaction was set up as per the manufacturer’s protocol; 10 μL 2X Reaction Buffer, 2 μL PCR product, 1μL pJET 1.2/blunt Cloning Vector (50 ng μL⁻¹), 1 μL T4 DNA Ligase and made up to a total of 20 μL with nuclease free water. The mixture was incubated for 5-30 minutes at room temperature. 10 μL of the reaction was used to transform JM109 chemically competent cells. The pJET1.2/blunt vector contains a gene that encodes a lethal product that becomes disrupted upon the successful insertion of blunt ended DNA such as a PCR product; thus only recombinant vectors containing a DNA insert are able to form colonies.

2.3.10 Transformation and preparation of competent cells of *E. coli*

Chemically competent JM109 cells were prepared according to the method described by Chung et al. [192] using TSS (5 g PEG8000, 1.5 mL 1M MgCl₂, 2.5 mL DMSO, LB to 50 mL). An overnight culture in 5 mL of LB was grown at 37°C and diluted the next morning into 50 mL LB in a 250 ml conical flask. OD₆₀₀ was monitored and the cells harvested at an OD₆₀₀ 0.2-0.5. The cells were put into 50 mL centrifuge tubes, incubated for 10 minutes on ice then centrifuged at 4000 rpm at 4°C for 10 minutes before being resuspended in 5 mL chilled TSS. Aliquots of 100 μL were added to ice-cold eppendorf tubes and either used immediately or frozen at -80°C [192-193].

An aliquot of 100 μL cells was thawed on ice for 5 minutes before 1-50 ng (in a volume no greater than 10 μL) of ligated DNA was added and the mixture incubated on ice for 10 minutes. Cells were heat shocked at 42°C for 45 seconds, before returning to ice for a further 2 minutes. 1 mL of LB was added to the microcentrifuge tube and cells were recovered at 37°C with shaking at 250 rpm for 1 hour. Cells were centrifuged and resuspended in 100 μL LB, to be spread on appropriate antibiotic plates and incubated overnight at 37°C.

Chemically competent BioBlue cells (Bioline, London, UK) were transformed according to the manufacturer’s protocol.

2.3.11 Collection of cells for RNA Extraction

In order to store cells for RNA extraction at a later date 3 x 10⁸ cells were aliquoted into 5 mL of RNAlater® (Applied Biosystems, Warrington, UK), this was equivalent to 1 mL of an OD₆₀₀ 10 of *P. pastoris* culture. Cells were then centrifuged for 5 minutes in an Eppendorf 5810-R centrifuge fitted with a A-4-62 rotor (Eppendorf UK Limited, Histone, UK) at 4°C and 9000 x g. The supernatant was removed and cells resuspended in 0.5 – 1 mL RNAlater®.
2.3.12 Isolation of RNA
RNA was isolated using RiboPure – Yeast Kit (Applied Biosystems, Warrington, UK) according to the manufacturers’ instruction manual. RNase-free pipette tips and microcentrifuge tubes were used throughout the protocol along with RNase-free water. Typically $3 \times 10^8$ yeast cells were collected, equivalent to 1 mL culture of $OD_{600} 10$ of *P. pastoris*. DNase I treatment was done according to manufacturer’s protocol except incubation was extended to 1 hour before DNase Inactivation reagent was added to the mixture.

2.3.13 Quantification and Quality Checks of RNA
RNA was quality checked using an Experion Automated Electrophoresis System and the RNA StdSense chip and reagents (Bio-Rad, Hemel Hempstead, UK). RNA was prepared and the chip primed according to the manufacturer’s protocol. Quantification and quality was recorded as generated by Experion software (Bio-Rad, Hemel Hempstead, UK).

2.3.14 cDNA
Cotranscript was prepared using the High-Capacity cDNA Archive Kit (Applied Biosystems, Warrington, UK) according to the manufacturer’s instruction manual. 1 µg RNA was used in a total reaction volume of 20 µL.

2.3.15 Polymerase Chain Reaction (PCR)
Oligonucleotides were purchased from Invitrogen, Paisley, UK, a full list of primers used can be seen in Appendix 9.2. PCR reactions were carried out in sterile 0.2 ml PCR tubes (Thermo Fisher Scientific, Loughborough, UK) in either an Eppendorf Mastercycle Gradient or Bioer Gene Pro thermal cycler (Eppendorf UK, Histon, UK; Alpha Laboratories, Eastleigh, UK). Reactions using Phusion Hot Start II (New England Biolabs, Herefordshire, UK) were set up with 10 µL Buffer, 4 µL 2.5 mM dNTPs, 2.5 µL Primer 1, 2.5 µL Primer 2, 1 µL of template DNA made up to 50 µL with water. An annealing temperature of 60°C was used for all reactions and elongation at 72°C for 15-30 seconds per expected Kb of fragment for 35 cycles. For analytical PCR Biomix Red solution (Bioline, London, UK) or REDTaq Ready Mix (Sigma-Aldrich, Dorset, UK) mix comprised of 10 µL Biomix Red/REDTaq, 1.5 µL Primer 1, 1.5 µL Primer 2, 1 µL template DNA and made up to a total of 20 µL with water. Annealing temperature was adjusted depending on the melting temperature of the primers used and elongation was at 72°C for 1 minute per Kb of expected PCR fragment for 35 cycles.
2.3.16 Gibson Assembly Protocol

Construction of vectors using the Gibson assembly protocol was adapted from Gibson et al. [194-195]. Up to four PCR products were annealed together using the one-step isothermal DNA assembly protocol. Briefly, 5x isothermal reaction buffer was prepared containing 25% PEG-8000, 500 mM Tris-HCL pH7.5, 50 mM MgCl₂, 50 mM DTT, 1mM of each dNTPs and 5 mM NAD. 320 μL of the 5x isothermal reaction buffer was used to create the assembly master mix, with the addition of 0.64 μL 10U μL⁻¹ T5 exonuclease (Epicentre Biotechnologies, Madison, US), 20 μL of 2U μL⁻¹ Phusion DNA polymerase (New England Biolabs, Herefordshire, UK), 160 μL of 40U μL⁻¹ Taq DNA Ligase (New England Biolabs, Herefordshire, UK) with H₂O to a final volume of 1.2 mL. Primers were designed with a 30-50 bp region of sequence identical to that flanking the DNA fragment to which it was to be joined. Once all the relevant fragments had been amplified with the added linker regions, equimolar concentrations of up to 100 ng of DNA from each fragment were added to 15 μL master mix and incubated for one hour at 50°C before being transformed into chemically competent E. coli.

2.3.17 Random-Amplification of Polymorphic DNA (RAPD)-PCR

Random decamer primers were selected and purchased from Invitrogen, Paisley, UK. PCR primers were designed to amplify random segments of genomic DNA using a single primer. Reactions were set-up comprised of 12.5 μL Biomix Red (Bioline, London, UK), 1 μL decamer primer, 1 μL genomic DNA made up to 25 μL with water. Cycling conditions used were an initial denaturation at 95°C for 5 minutes, 94°C for 1 minute, 35°C for 1 minute, 72°C for 2 minutes for 45 cycles followed by a final extension step at 72°C for 5 minutes. PCR reaction products were separated on 2.5% (w/v) agarose gels.

2.3.18 Quantitative Polymerase Chain Reaction (qPCR)

Genomic DNA prepared using Qiagen DNeasy Plant Mini Kit (Qiagen, Crawley, UK), was normalised to 10 ng μL⁻¹. qPCR primers were designed for the amplification of a region of the gene of interest no bigger than 200 bp (Appendix 9.2).

Reactions were set up using 2X SYBR® Green JumpStart Taq Ready Mix (Sigma-Aldrich, Dorset, UK), 1 μL of forward and reverse primers and 3 μL of normalised genomic DNA were made up to 20 μL with water. A Chromo4™ Real-Time Detector using the thermal cycler software Opticon 3 (Bio-Rad, Hemel Hempstead, UK) was used. qPCR settings were: 96°C for 5 minutes and 40 cycles of 96°C for 30 seconds, 55°C for 30 seconds then 72°C for 30 seconds with the plate read at the end of each cycle, and finally 72°C for 10 minutes. A melting curve was recorded from 55°C - 95°C and the plate read at every 0.2°C and held for 2 seconds.
2.3.18.1 qPCR calculations

Data was analysed using the comparative Delta-delta-Ct (ΔΔCt) method and all samples normalised to ACT1 (housekeeping gene). The equations required to calculate fold change are shown below.

ΔΔCt works on the assumption of 100% accuracy for the primer sets. The Pfaffl method, which was implement for this thesis takes into account the specific primer efficiencies [196-197]. To calculate primer efficiencies four 10-fold serial dilutions of genomic DNA were used as a template in qPCR. A standard curve exhibiting Ct values was plotted against log concentration and the slope of the line used to calculate efficiencies. Table 6 outlines primer efficiencies for qPCR primers used.

Table 2.4 qPCR Primer Efficiencies

<table>
<thead>
<tr>
<th>Gene of Interest</th>
<th>Primer 1</th>
<th>Primer 2</th>
<th>Primer Efficiencies (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT1</td>
<td>020-ACT1q1</td>
<td>021-ACT1q2</td>
<td>91.05</td>
</tr>
<tr>
<td>HAC1</td>
<td>067-HAC1q3</td>
<td>066-HAC1-q4</td>
<td>90.24</td>
</tr>
<tr>
<td>KAR2</td>
<td>010-kar2q1</td>
<td>011-kar2q2</td>
<td>87.23</td>
</tr>
<tr>
<td>PDI</td>
<td>012-PDIq1</td>
<td>013-PDIq2</td>
<td>86.96</td>
</tr>
<tr>
<td>HSA</td>
<td>070-HSAq1</td>
<td>071-HSAq2</td>
<td>90.77</td>
</tr>
<tr>
<td>HSAopt</td>
<td>188-HSAoptq3</td>
<td>189-HSAoptq4</td>
<td>97.82</td>
</tr>
<tr>
<td>Zeo (R)</td>
<td>040-Zeoq3</td>
<td>041-Zeoq4</td>
<td>88.23</td>
</tr>
<tr>
<td>TRY1</td>
<td>383-Try1q7</td>
<td>384-TRY1q8</td>
<td>100.67</td>
</tr>
</tbody>
</table>

2.3.19 Quantitative Reverse Transcription (RT)-qPCR

RT-qPCR was carried out using the same protocol as qPCR, using cDNA as opposed to genomic DNA. Of the initial reverse transcription reaction (diluted 1:10), 5 μL of cDNA was used in a 20 μL reaction volume. Prior to conversion to cDNA an RT-qPCR control plate was used to confirm the absence of genomic DNA, using 5 μL of 1 μg RNA diluted in 200 μL water.
2.4 Transformation of *P. pastoris* by Electroporation

2.4.1 Preparation of DNA for Transformation
To achieve site directed homologous recombination in *P. pastoris* the vector was linearised within the region of homology. After confirmation that the plasmid had been digested, by analysis on agarose gel, *P. pastoris* was transformed with 5-10 µg of the linearised DNA.

2.4.2 Electroporation Protocol
The electroporation protocol was modified from the *Pichia* manual (Invitrogen, Paisley, UK). A 5 mL culture of *P. pastoris* in YPD was grown overnight at 250 rpm at 30°C. The overnight culture was used to inoculate 200 mL YPD to an OD$_{600}$ 0.5. The culture was incubated until OD$_{600}$ 1.3-1.5, which took approximately 4 hours. Cells were centrifuged at 4000 rpm for 5 minutes, washed with 200 mL ice-cold autoclaved deionised water, and centrifuged again. Cells were washed with 100 mL water and finally with 10 mL of ice-cold 1 M sorbitol, before being resuspended in 1 mL 1 M sorbitol. Aliquots of 80 µL of cells were incubated on ice for 5 minutes before 5-10 µL of linearised DNA was added and incubated for a further 5 minutes. The cells were pulsed at 2000V, 25 µF, 200 Ω, for approximately 5 milliseconds, using the GenePulser electroporator in 2 mm cuvettes (Bio-Rad, Hemel Hempstead, UK). Immediately 1 mL of ice-cold 1M sorbitol was added to the cuvette and the contents transferred to a sterile eppendorf tube. The transformation was incubated at 30°C without shaking for 1-2 hours. After suitable incubation the transformation was plated onto appropriate selection. The plates were incubated for 3-5 days at 30°C.

2.5 Post-transformational Vector Amplification
Post-transformational vector amplification (PTVA) utilised a gradual step-wise increase in Zeocin concentration to generate multi-copy inserts of heterologous genes and selectable markers in *P. pastoris* [149]. Transformants were initially isolated on 100 µg mL$^{-1}$ Zeocin YPD plates as described. Single colonies were streaked to single colonies on a fresh 100 µg mL$^{-1}$ Zeocin YPD plate and grown for 3-5 days at 30°C. Individual colonies were selected and spotted onto 100 µg mL$^{-1}$ and left to grow under the same conditions. Tolerant strains were then selected by spotting progressively on YPD plates containing 200 µg mL$^{-1}$, 300 µg mL$^{-1}$, 500 µg mL$^{-1}$, 1000 µg mL$^{-1}$ and finally 2000 µg mL$^{-1}$ Zeocin.

The liquid variant of PTVA was adapted to ensure isolation of microbiologically pure samples. Individual colonies isolates from 100 µg mL$^{-1}$ Zeocin plate were inoculated into 5 mL YPD and left to grow for 24 hours. After 24 hours cultures were centrifuged removing the spent media, before resuspending in YPD containing 200 µg mL$^{-1}$ Zeocin. This was continued with every 24 hours.
growth medium being replaced containing increasing concentration of Zeocin as described above. After 24 hours growth on 2000 μg mL⁻¹ Zeocin cultures were plated onto a YPD plate containing 2000 μg mL⁻¹ and left to incubate for 3-5 days at 30°C.

2.6 Preparation of Cell Banks for Fermentation

Two research cell bank (RCB) flasks were prepared containing 250 mL *Pichia* Cell Bank Medium (PCBM) and 100 μg mL⁻¹ Zeocin were inoculated with a single colony from a 100 μg mL⁻¹ Zeocin YPD plate and incubated at 30°C and 250 rpm in an orbital shaker. Cultures were grown for 20 hours until OD₆₀₀ of between 6 and 20 was reached. Samples were transferred to 50 mL sterile centrifuge tubes and spun for 20 minutes at 5000 rpm, at 4°C. The supernatant was discarded and the pellet was resuspended to approximately OD₆₀₀ 10 in Pichia Freezer Mix (PFM) using the following equation:

The resuspended culture was separated into 1 mL aliquots in cyrovials and stored at -80°C.

2.6.1 Microbiological Purity Test

From a single vial of prepared cell bank 10 μL was streaked out onto a TSA and YPD plates in duplicate and incubated at 30°C and 37°C for 72 hours. Fresh samples of culture were analysed by microscopic examination of Gram stained slides. Liquid culture was mixed with ¼ Strength Ringers (Sigma-Aldrich, Dorset, UK) onto a glass side and left to dry at 105°C for 5 minutes. The slide was treated using Gram Stain Reagents according to the manufacturer’s protocol (Sigma-Aldrich, Dorset, UK.). The slide was viewed under oil immersion at 1000x TM. Only vials that indicated pure cultures were used for fermentation.

2.7 Small Scale Heterologous Protein Expression from Recombinant *P. pastoris* Strains

2.7.1 Expression in 24-well microtitre plates

The protocol was adapted from the *Pichia* manual (Invitrogen, Paisley, UK) for growth in 24 deep-well microtitre plates (Whatman, Maidstone, UK) with adhesive gas permeable seals (Thermo Fisher Scientific, Loughborough, UK). Single colonies were used to inoculate 3 mL of BMGY and grown at 30°C, 250 rpm for 24 hours. The OD₆₀₀ was measured and cultures normalised to OD₆₀₀ 10. Cells were centrifuged for 5 minutes at 4000 rpm at room temperature and the supernatant
removed. The cultures were then resuspended with methanol containing BMMY and left to grow for a further 24 hours. Cells were grown for up to 3 days at 250 rpm, 30°C, with 100% methanol added to a final concentration of 0.5% (v/v) every 24 hours.

2.7.2 Expression in 50 mL centrifuge tubes

Expression was carried out as described previously, with the exception that 5 mL of culture was grown in 50 mL centrifuge tubes with the lid loosely attached for gas exchange.

2.7.3 Expression in 250 mL baffled flasks

Expression was carried out as described previously, with the exception that 20 mL of culture was grown in 250 mL baffled flasks (SciLabware Ltd, Staffordshire, UK).

2.8 Fermentation in 15 L Bioreactors

2.8.1 Inoculation Preparation 15 L Bioreactors for Fermentation

Two 1 L flasks per each strain were prepared with a total of 450 mL Yeast Extract/Peptone Media and autoclaved at 121°C for 30 minutes. 50 mL of Glucose Solution was added to the flasks aseptically and placed in a shaking incubator for 1 hour at 30°C and 275 rpm. One RCB vial was removed from the -80°C freezer and left to thaw for 15 minutes. After an hour the flasks were inoculated aseptically with 220 μL of RCB suspension. The flasks were incubated for 20 hours at 30°C and 275 rpm. 10 mL of sample was removed and aseptically aliquoted into two equal volumes. One of the aliquots was used to determine OD_{600} and the other was used to assess microbiological purity by Gram staining. Only cultures that appeared pure were used for fermentations.

2.8.2 15 L Fermentation

Fermentation 15 L vessels (Braun Biostat ED, Sartorius, Epsom, UK) were prepared prior to inoculation. Separate stab ports were prepared for ammonia, phosphoric acid, glycerol/yeast extract feed, antifoam, methanol feed and inoculum and sealed before autoclaving. Vessels were pre-sterilised before addition of 12.5 L fermentation media and re-sterilised at 121°C for 30 minutes. Media volume was drained to 6 L before the addition of 60 mL trace elements solution and 12 mL d-Biotin solution (FDBK-Pichia medium).

Pre-prepared stab ports were steriley inserted into the 15 L vessel via one of the head plate septum for glycerol feed, antifoam, acid, base and inoculum. Additionally a stab port was inserted sub-surface level for the methanol port. Feed reservoirs were prepared for both the glycerol and methanol feeds (100% methanol; Merck, Germany). Fermentation parameters were established.
on the digital control unit (DCU) to 28°C, stirrer at 700 rpm, airflow at 6 slpm and pH 5.0, while the \(pO_2\) controller was set to a minimum point of 30% via cascade 1 to stirrer and cascade 2 to gasmix.

The fermenter was inoculated by transferring 240 mL of shake flask culture. Samples were removed aseptically at intervals to measure wet and dry cell weights, \(OD_{600}\) and to test for microbiological purity. Exhaustion of glycerol in the growth medium was characterised by a sudden fall in carbon dioxide evolution rate (CER) and agitator speed and a rise in \(pO_2\). Glycerol was then fed at 160 g h\(^{-1}\). Once 3900 g glycerol had been fed and \(OD_{600}\) reached 500-600 a 50 mL shot of methanol was added during over 10 minutes, simultaneously the glycerol feed was switched off. Methanol consumption was monitored by the agitator speed, before continuous methanol feed was established at 20 mL h\(^{-1}\). Fermentation was run for 48 hours post-induction. Samples were taken at 0, 2, 4, 6, 24 and 48 hours post-induction for analysis.

### 2.9 SDS-PAGE Protein Gel

#### 2.9.1 12% (v/v) Resolving gel

A 5 mL 12% (v/v) consisted of 2 mL Acrylamide mix (30% v/v acrylamide, 0.8% v/v bis-acrylamide), 1.3 mL Tris-Cl (1.5 M, pH8.8), 50 μL 10% (w/v) SDS, 50 μL 10% (w/v) Ammonium persulphate, 2 μL Tetramethylethylenediamine (TEMED) and 1.5 mL distilled water.

#### 2.9.2 Stacking gel

A 1 mL stacking gel consisted of 170 μL Acrlyamide mix (30% v/v acrylamide, 0.8% v/v bis-acrylamide), 130 μL Tris-Cl (1 M, pH 6.8), 10 μL 10% (w/v) SDS, 10 μL 10% (w/v) Ammonium persulphate, 1 μL TEMED and 680 μL distilled water.

#### 2.9.3 SDS-PAGE Running Buffer (1x)

3 g Tris-base, 14.4 g Glycine, 1g SDS

#### 2.9.4 1x SDS-PAGE Loading Buffer

1x SDS-PAGE loading buffer contained 2.5% (w/v) 2M Tris-HCl pH 6.8, 0.2% (w/v) SDS, 0.6% (v/v) Glycerol, 0.05% (v/v) β-mercaptoethanol and 0.416% (w/v) bromophenol blue in a total volume of 100 ml.

#### 2.9.5 Coomassie Blue Stain (1 Litre)

0.2% (w/v) Coomassie Blue, 7.5% (v/v) acetic acid and 50% (v/v) ethanol made up to 1 L with water.
2.9.6 Coomassie Blue Destain (1 Litre)

10% (v/v) acetic acid, 30% (v/v) ethanol made up to 1 L with water.

2.9.7 Protein gel electrophoresis

Culture supernatant samples were boiled with 5x SDS-PAGE loading buffer for 5 minutes. Cells were centrifuged and supernatant was loaded onto the gel along with 5 μL of unstained protein molecular marker or 10 μL PageRuler™ Prestained Protein ladder (Fermentas, York, UK). Electrophoresis was run in a Mini-PROTEAN® Tetra Cell (Bio-Rad, Hemel Hempstead, UK) at 25 mA for the stacking gel and 35 mA for the resolving gel for approximately one hour in 1X SDS-PAGE Running buffer.

Gels were stained with either Coomassie Blue or SimplyBlue SafeStain (Invitrogen Corporation Paisley, UK). For gels stained with Coomassie Blue, gels were washed briefly in water and then left to stain on a shaker one hour. The gels were then transferred to Coomassie Blue stain for one hour while shaking. For SimplyBlue SafeStain, gels were washed 3 times in deionised water with shaking for 5 minutes. SimplyBlue SafeStain was added and left to stain for one hour. The gel was then washed with deionised water and left to destain for one hour. After an hour the water was changed and left to destain for another hour in deionised water.

Gels with both types of staining were photographed using a G:Box SynGene gel doc (Syngene, Cambridge, UK).

2.10 Albumin Blue Florescence Assay

Albumin concentration was determined using the Albumin Blue Fluorescence assay according to the manufacturer’s instructions (Active Motif, La Hulpe, Belgium). Samples were equilibrated to room temperature prior to use. A HSA standard curve in duplicate was set up using the concentrations 200, 100, 50 25, 12.5, 6.3, 3.2 and 0.0 μg mL⁻¹ in a black microtitre 96-well plate. The fluorescence was measured at excitation 560nm, emission 620nm using Synergy HT multi-detection microplate reader (Bio-Tek, Potton, UK).

2.11 Southern Blots

2.11.1 0.2 M HCl

2.4% (v/v) concentrated HCl made up to 400 mL with water.

2.11.2 Denaturation Solution

2% (w/v) NaOH, 8.7% (w/v) NaCl, made up to 800 mL with water.
2.11.3 Neutralisation Solution
8.76% (w/v) NaCl, 6% (w/v) Tris-Base, 0.2% (v/v) 0.5 M EDTA, pH adjusted to 7.2-7.5 with HCl, made up to 800 mL with water.

2.11.4 1 X Maleic Acid
1.16% (w/v) Maleic Acid, 0.87% (w/v) NaCl, 0.85% (w/v) NaOH pellets, pH adjusted to 7.5 with NaOH, made up with 1 L with water and autoclaved.

2.11.5 20 X SSC
17.5% (w/v) NaCl, 8.8% (w/v) Na-Citrate, pH adjusted to 7.0 with HCl and made up with 1 L with water and autoclaved.

2.11.6 Washing Buffer
0.3% (w/v) Tween-20 in 500 mL maleic acid.

2.11.7 Detection Buffer
1.2% (w/v) Tris base, 0.58% (w/v) NaCl, pH adjusted to 9.5 with HCl made up to 1 L with water.

2.11.8 Preparing the DIG-labelled DNA probe
Template DNA was labelled according to the DIG-High Prime DNA Labelling instruction manual (Roche, West Sussex, UK). Briefly, template DNA (1 µg) was added to 16 µL filtered autoclaved H₂O and denatured by heating in boiling water for 10 minutes and chilled on ice; 4 µL DIG-High prime (Roche, West Sussex, UK) was added and centrifuged briefly followed by incubation for one hour at 37°C. The reaction was stopped by adding 0.2 M EDTA and/or by heating to 65°C for 10 minutes.

2.11.9 Blotting protocol
The DIG-High Prime DNA Labelling and Detection Starter Kit were used for hybridization of DIG-labelled probes for nucleic acid detection according to the manufacturer’s instructions (Roche, West Sussex, UK). All hybridisation steps were carried out in the Techne Hybridisation oven (Bibby Scientific Ltd, Staffordshire, UK).

Briefly, genomic DNA (5-10 mg) was digested overnight with the respective restriction enzymes and visualised on a 1% (w/v) agarose gels. Gels were then washed in 0.25 M HCl for 10 minutes, rinsed in H₂O, soaked in 2-3 gel volumes of denaturing solution for 2 x 15 minutes, rinsed again in H₂O and soaked in 2-3 gel volumes of neutralisation solution for 2 x15 minutes. Gels were blotted onto Amersham Hybond-N’ membrane (GE Healthcare, Little Chalfont, UK) using 20 X
SSC. Post-transfer the membrane was soaked in 0.4 M NaOH for 2-60 minute and washed in 5 X SSC.

Hybridization, stringency washes and detection were carried out according to the manufacturer’s protocol. The membrane was incubated with 1 mL CSPD ready to use (Roche, West Sussex, UK) at room temperature for 5 minutes. Membranes were visualised using the G:Box SynGene gel doc (Syngene, Cambridge, UK).

2.12 Flow Cytometry Analysis

Strains were expressed for 24 hours in BMMY in 250 mL baffled shake flasks as previously described. After 24 hours expression 1 mL of culture was collected through centrifugation and washed in 1x TBS (0.605% Tris, 0.876% NaCl, pH adjusted with 1 M HCl). Samples were washed for a second time before being resuspended in a final volume of 10 mL of buffer to create a 1 in 10 dilution. Samples were stained according to the LIVE/DEAD® FungalLight™ yeast viability kit protocol (Invitrogen Corporation, Paisley, UK). Cells were measured using a FACscan flow cytometer (Becton Dickinson, Oxford, UK). Data was acquired using CellQuest software (Becton Dickinson, Oxford, UK) with samples measured on a high flow rate for 30 seconds. Data was analysed using Cylogic software (CyFlo Ltd, Turku, Finland).

2.13 Statistical Analysis

Statistical analysis was calculated with unpaired T-test or one-way or two-way ANOVA using Prism software (GraphPad, California, USA). Experiments were repeated in biological triplicates unless otherwise stated. All p values of 0.05 or less were considered significant.

2.14 Microarray Analysis

Strains used for microarray analysis were expressed for 24 hours in BMMY in 250 mL baffled shake flasks as previously described. RNA was extracted using the RiboPure Yeast Kit (Applied Biosystems, Warrington, UK) and quantification and quality checks were carried out using the Experion Automated Electrophoresis System and the RNA StdSense chip and reagents (Bio-Rad, Hemel Hempstead, UK). Requirements of samples were 10 μg of RNA, with a RQI rating over 7. Custom probes were designed in accordance with previous experiments (data not shown) on 8x60K slides. Probe coverage was a minimum of nine per gene.

RNA was delivered to the Bacterial Microarray Group at St George’s Hospital for analysis. Cy3-labelled cRNA was prepared from 1 μg total RNA using the Agilent One-Colour Quick Amp Labelling kit according to the manufacturer’s instructions (Agilent Technologies UK Ltd,
Wokingham, UK). One Color Spike-In controls were labelled together with the RNA samples. Purified samples were hybridised to an Agilent 8x60k format Sureprint G3 gene expression custom array and incubated overnight in a rotating oven (Agilent Technologies UK Ltd, Wokingham, UK) at 65°C, 20 rpm. After hybridization, slides were washed for 1 minute at room temperature in GE Wash Buffer 1 and 1 minute at 37°C in GE Wash buffer 2 (Agilent Technologies UK Ltd, Wokingham, UK), placed beneath an Ozone Barrier Slide cover (Agilent Technologies UK Ltd, Wokingham, UK) and scanned immediately, using an Agilent High Resolution Microarray Scanner, at 2 µm resolution. Scanned images were quantified using Feature Extraction software v10.7.3.1.

2.14.1 Gene Expression Analysis

Analysis was carried out by Bioinformatics Support Service, Imperial College London, using the Bioconductor package in the R programming language to identify differentially expressed genes [198]. Each of the clonal variation (CV) strains was contrasted to wild-type X33 in order to determine statistically significant differences. The Empirical Bayes method was applied to identify statistical significance in contrast between gene expression profiles [199]. The false discovery rate (fdr) based on Benjamini and Hochberg’s method, which assumes that all genes are statistically different from one another was set to be less than 5% [200]. Gene functionality was assigned with reference to a created *Pichia pastoris* genome (www.blugen.org/gbrowse-bin/gbrowse/Pichia/).

2.14.2 Pathway Analysis

Pathway analysis was used to identify pathways were significantly up- or downregulated in accordance with the gene expression data. Initially significantly up- or downregulated genes were run through KOBAS (KEGG Orthology Based Annotation System), which assigns genes to pathways based on the KEGG maps specifically for *P. pastoris* [201-203]. Once pathways were identified the KEGG Search & Colour pathway was used to visually map the differentially expressed genes [204-205].
3 Determining the Effects of using Multi-Copy Clones in *Pichia pastoris*

3.1 Introduction

3.1.1 Advantages of Multi-Copy Clones

Low specific productivity is highlighted as a key detriment to using *P. pastoris* as a heterologous expression system. This is somewhat compensated by high cell densities thereby achieving high volumetric productivity [74]. Should specific productivity be increased then the potential of using *P. pastoris* as the preferred heterologous protein expression platform expands enormously. One of the best established methods to achieve this is to increase the number of cognate genes. The approach of using multi-(gene) copy strains to increase the production of heterologous protein has been used since the early 1990s [146, 206], particularly for intracellular expression where expression in high quantities is particularly difficult [87, 171, 207-208].

3.1.2 Traditional Methods for Making Multi-Copy Clones

Traditionally multi-copy strains have been generated by three main methods; sequential transformations, transformation directly on high concentrations of antibiotic and *in vitro* multimerisation [65]. Sequential transformation, which can integrate up to 10 copies of the gene of interest, is achieved through repeatedly transforming a linearised vector into the same locus or into different loci. However the screening process is laborious and this method is only feasible with a number of selection markers being applied [206]. Transformation directly onto high concentrations of antibiotics such as Zeocin and G418 results in low transformation efficiencies with ≤5% resistant clones containing multiple copies of the gene of interest. Furthermore this is not applicable with auxotrophic markers and requires extensive screening with only 1-2% of clones containing “jackpot” colonies of 10 or more gene copies [149]. The final method using *in-vitro* multimerisation creates multi-copies of the gene of interest in an *E. coli* shuttle vector prior to being transformed into *P. pastoris* [148]. Nonetheless due to the size of the pAO815 vector limited number of copies may be integrated and extensive screening is required in *E. coli*. Additionally increased copy number will decrease transformation efficiencies into *P. pastoris* [112].

3.1.3 Post-Transformational Vector Amplification

In 2008 Sunga *et al.* proposed an alternative to the typical methods of creating multimers; post-transformational vector amplification (PTVA) [149]. This method relies on a step-wise increase in Zeocin concentration in order to progressively select colonies that are able to resist higher concentrations. The *Sh ble* gene produces proteins that sequester Zeocin; thus theoretically increasing the number of copies of *Sh ble* gene should result in increased resistance to Zeocin. Sunga
et al. (2008) suggested that PTVA results in the complete amplification in a head-to-tail orientation of the entire vector, which is important in order to obtain a uniform recombinant product.

Advantages of PTVA include an increase in the frequency of “jackpot” colonies to 5-6%, compared to <1% by traditional methods [149]. Most importantly an increase in the amount of recombinant protein was observed. PTVA is becoming increasingly popular for generating multi-copy clones, with up to 15 copies of human Cu/Zn superoxide dismutase (hSOD) and 52 copies of porcine insulin precursor (PIP) reported in P. pastoris [151-152]. Due to the apparent ease and simplicity of this method and the reported successful results, PTVA was adopted in this project in order to create multiple copies of human serum albumin (HSA).

3.1.4 Negative Effects of Multi-Copy Clones
Multi-copy clones when coupled with proteins expressed intracellularly have been reported to increase titres. However, proteins targeted to the secretory pathway and expressed extracellularly show evidence of secretion saturation whereby increased gene copy number does not equate to higher titre [87]. Although secretion saturation does not occur at the same copy number for different proteins the trend of plateauing secretion is the same. Marx et al. (2009) observed that when expressing HSA, integrated into the rDNA locus, expression capped at approximately five copies [152].

In addition to secretion saturation investigations have reported that high copy-number clones induce the unfolded protein response (UPR) [87]. This is of particular interest as induction of UPR can lead to activation of the ER associated degradation pathway (ERAD) [209]. Due to the inclusion of multi-copy clones excess transcript will pass through the secretory pathway; thus increasing the likelihood of ERAD upregulation. Proteins that are poor folders are more likely to induce the UPR and these are the proteins that are liable to activate the ERAD; thus result in lower expression especially in multi-copy strains [173]. HSA is known to express and secrete well from P. pastoris; thus, at low copy number, expression is unlikely to induce UPR. Therefore it is of interest to determine whether UPR upregulation will occur through the inclusion of multiple copies of the HSA gene. The use of multi-copy strains to produce high titre levels using a protein that clearly folds well allows us to investigate secretion saturation as a separate issue from folding stress in order to gain a better understanding of the pressures on the secretory pathway.

3.1.5 Aims for the Investigation of Multi-Copy Clones
- To evaluate the method of Post-Transformational Vector Amplification (PTVA) in making multi-copy strains
To create clones containing multiple copies of the HSA gene and examine the effect of multiple copy clones on:
  o Titre
  o Secretion saturation
  o UPR signals
  o Genetic Instability

3.2 Creating Multi-Copy Clones

3.2.1 Vector Creation

3.2.1.1 pα-HSA

The HSA gene was amplified by PCR using the proof-reading enzyme Phusion Hot Start II and primers 047-HSA2 and 053-HSA4-SP (Appendix 9.2) excluding the region encoding the native signal peptide (SP). The vector pPICzαA (Invitrogen Corporation) contains the sequence encoding the Saccharomyces cerevisiae α-mating factor (MF) signal peptide which, when expressed in frame, can target the HSA protein to the secretion pathway. pPICzαA and the amplified HSA-SP were digested using restriction enzymes EcoRI and NotI. The resulting fragments were ligated together using T4 DNA ligase before being transformed into TSS competent E. coli JM109 cells (Table 2.1) by heat shock and plated onto Lennox LB plates containing 100 μg ml⁻¹ Zeocin. The pPICzαA vector contains a Sh ble gene, expression of which provides resistance to Zeocin in both E. coli and P. pastoris. The pPICzαA vector is 3.6 kb, while the modified HSA gene is 1.8 kb, giving a total size of 5.4 kb (Fig 3.1).

![Figure 3.1 pα-HSA Plasmid Map](image)

**Figure 3.1 pα-HSA Plasmid Map**

pα-HSA is based on the pPICzα vector (Invitrogen, Paisley, UK) containing a truncated HSA gene to remove the SP. pPICzα contains the Sh ble gene, which encodes the Zeocin resistance protein allowing for selection in both E. coli and P. pastoris and the α-MF signal peptide for secretion.
Once the recombinant vector had been successfully transformed into JM109 the plasmid was reisolated using a Qiagen mini-prep kit and a double digest using EcoRI and NotI used to confirm the presence of the HSA-SP gene (Fig 3.2).

![Figure 3.2 Agarose gel confirms the successful construction of the pα-HSA vector](image)

**Figure 3.2 Agarose gel confirms the successful construction of the pα-HSA vector**

1% agarose gel electrophoresis of restriction fragments to confirm the construction of pα-HSA. 1: 1 Kb DNA Ladder (Fermentas); 2: pPICzα A; 3: HSA PCR product; 4: HSA PCR product; 5: pα-HSA 1 digested with EcoRI and NotI; 6: pα-HSA 2 digested with EcoRI and NotI. pPICzα A is 3.6 kb and HSA is 1.8 kb.

### 3.2.1.2 pPICz-HSA

Many heterologous proteins that are secreted using *P. pastoris* are not naturally secreted; thus the importance of using the α-MF secretion signal from *S. cerevisiae*. As HSA is a natively secreted protein, it was of interest to compare the effect of secretion using the native signal peptide and the α-MF signal peptide. Therefore pPICz-HSA was designed to include the intact HSA gene, retaining its native signal sequence, which should still result in the secretion of the HSA protein. The full HSA gene was amplified by PCR using primers 047-HSA2 and 050-HSA3 (Appendix 9.2). The fragment was digested using EcoRI and NotI and ligated to the pPICz vector which had also been digested with EcoRI and NotI (Fig 3.3). The pPICz vector differs from the pPICzα vector in the absence of the α-MF signal peptide, but in all other factors the vectors are the same. The absence of the α-MF signal peptide ensures that any secreted product is a result of the native signal peptide encoded by the HSA gene.
pPICz-HSA contains the full HSA gene ligated together with the pPICz vector, which contains the Sh ble gene encoding a Zeocin resistance protein.

The ligated vector was transformed into E. coli JM109 as described for pα-HSA, reisolated and digested with EcoRI and NotI to confirm the presence of the HSA gene. The pPICz vector is 3.3 kb, while the HSA gene containing the native secretion signal sequence is 1.8 kb (Fig. 3.4).

1% agarose gel electrophoresis of restriction fragments to confirm the construction of pPICz-HSA. 1: 1 Kb DNA Ladder (Fermentas); 2: pPICz A; 3: HSA PCR product; 4: HSA PCR product; 5: pPICz-HSA 1 digested with EcoRI and NotI; 6: pPICz-HSA 2 digested with EcoRI and NotI. pPICz A is 3.3 kb and HSA is 1.8 kb.

### 3.2.2 HSA Production in *P. pastoris*

After successful construction of pPICz-HSA and pα-HSA 5-10 ng was linearised using the restriction enzyme Pmel, which cuts within the 5’ AOX1 promoter in the vector and allows for homologous recombination into the *P. pastoris* genome (Fig 3.1 and 3.3). Homologous recombination occurs through the free DNA termini and results in a single crossover type integration [65]. Linearising the plasmid using Pmel facilitates integration at a pre-determined locus, in this instance AOX1.
Integration of this type results in the Mut⁺ phenotype as two copies of the AOX1 promoter will result from recombination, one driving expression of AOX.

A 200 mL culture of GS115 (his4) was prepared for electroporation as described in the Pichia Expression Kit manual and transformed with linearised pα-HSA or pPICz-HSA [120]. The transformed cells were left to recover at 30°C for 1 hour, without shaking, before being plated on YPD containing 100 μg mL⁻¹ Zeocin. The plates were left to incubate at 30°C for 3-5 days. Individual colonies were selected and streaked onto YPD plates before being re-plated onto YPD 100 μg mL⁻¹ Zeocin plates to ensure that the original colonies were not transiently resistant to Zeocin but were true transformants containing the Sh ble gene. Once colonies had been re-streaked onto YPD plates containing 100 μg mL⁻¹ Zeocin individual colonies were selected for further study.

Individual colonies were used to inoculate a glycerol based medium (BMGY) and left to grow for 24 hours before being induced with methanol (BMMY) in order to evaluate expression of HSA. Due to both constructs containing secreting signal peptides targeting the HSA protein for secretion an evaluation of the supernatant would determine the functionality of the vectors. After 24 hours expression in the methanol containing media cultures, the supernatant was recovered by centrifugation for analysis. A 40 μL sample mixed with 5X Loading buffer was then analysed on a 12% SDS-PAGE gel. One advantage of using P. pastoris as an expression host is that it does not secrete many native proteins; thus the presence of a secretion signal peptide on both constructs makes the target protein easy to identify (Figure 3.5) [65, 74].
Once it had been established that both pPICz-HSA and the pα-HSA had integrated into the genome and the resulting transformants were secreting HSA (a 66.5 KDa protein) it was then possible to proceed with the creation of multi-copy strains.

3.2.3 Making and Selecting Multi-Copy Clones using PTVA

Individual colonies of both GpPICz-HSA and Gpα-HSA were selected from plates containing 100 μg mL⁻¹ Zeocin and streaked onto a fresh plate containing 100 μg mL⁻¹ Zeocin, in order to ensure uniformity and left to incubate for 3-5 days. As described by Sunga et al. each clone was then sequentially spotted onto increasing concentrations of Zeocin, (200, 300, 500, 1000, 2000 μg mL⁻¹) [149]. Any clones that do not have an increase in copy number should be unable to grow in the presence of higher concentrations of Zeocin as they would produce insufficient resistance protein to titrate out the Zeocin (Fig. 3.6).
There was a visible difference in clones spotted onto the YPD 100 μg ml\(^{-1}\) plates and those spotted onto the 2000 μg ml\(^{-1}\) plates in terms of both size and morphology, with growth on the high concentrations considerably reduced (Fig 3.6). Additionally there was a noticeable difference between clones containing either the pa-HSA or pPICz-HSA vector. It would appear that clones containing the pPICz-HSA vector grown in the presence of 2000 μg ml\(^{-1}\) show reduced resistance. Potentially the growth rate may be reduced in clones containing pPICz-HSA as growth on 100 μg ml\(^{-1}\) Zeocin also differs between strains integrated with the different vectors.

As a result of the differences seen between clones containing pa-HSA or pPICz-HSA it was determined that only strains containing the pa-HSA vector were used. As many proteins produced by industry are not natively secreted and the α-MF signal peptide is the preferred choice, the pa-HSA vector may be considered more industrially relevant [75]. For ease of identification clones were named GpaH#_100 (for clones selected on 100 μg ml\(^{-1}\) Zeocin), or GpaH#_2000 for those selected post-PTVA. All twenty-six clones that were selected for enrichment through PTVA were selected for further study.

### 3.3 Testing Multi-Copy Clones

In order to determine the effects of PTVA on amplification from single clones, titre and HSA copy number were determined for both parental and progeny clones. HSA protein concentrations were measured using the Albumin Blue Fluorescence assay (Active Motif). HSA protein titre was recorded for all clones that had been isolated in the presence of 100 μg ml\(^{-1}\) of Zeocin and 2000 μg ml\(^{-1}\) Zeocin.

Figure 3.6 Plates of parental clones isolated on 100 μg ml\(^{-1}\) Zeocin plates and progeny clones post PTVA isolated on 2000 μg ml\(^{-1}\) Zeocin.
The top half of the plates contains single spots of GS115 transformed with pa-HSA, while the bottom half of the plates contain spots of GS115 transformed with pPICz-HSA. A: 100 μg/ml Zeocin plate; B: 2000 μg/ml Zeocin plate.
Zeocin (post PTVA). In order to clearly display whether clones post-PTVA showed an increase or decrease in protein expression, relative expression of HSA was calculated, a percentage above 100% indicates that post-PTVA secretion increased, while a decrease in expression is represented by a percentage of less than 100 (Fig 3.7).

![Graph showing relative HSA expression post-PTVA compared to pre-PTVA levels, with values above 100% indicating increased secretion and below 100% indicating decreased secretion.]

**Figure 3.7 Post-PTVA Expression of HSA compared to levels pre-PTVA both Increases and Decreases in Progeny Clones**

Comparison of the relative expression, measured by Albumin Blue Assay (Active Motif) between original transformants isolated on 100 µg/ml Zeocin plate (set at 100%) and clones isolated after PTVA on 2000 µg/ml Zeocin. Values over 100% indicate a higher level of secretion of progeny clones post-PTVA, while values under 100% indicate a lower level of HSA secretion.

In equal proportions clones post-PTVA showed both increased and decreased HSA titre according to their corresponding parental clones (Fig 3.7). Furthermore there were strains, for instance GpαH15_2000, which showed no change in HSA titre compared to GpαH15_100. In order to determine the relevance of HSA expression levels copy number was calculated for parental and progeny clones. Potentially GpαH15_2000 does not contain increased HSA copy number compared to GpαH15_100; thus explaining the unchanged expression.

Copy number was analysed by qPCR of genomic DNA isolated using the DNeasy Plant mini-prep Kit (Qiagen). DNA was quantified using the Synergy HT multi-detection microplate reader (Bio-Tek) and normalised to 10 ng µL⁻¹ of which 3 µL was added to each reaction. HSA was amplified using primers 070-HSaq1 and 071-HSaq 2 and primers 020-ACT1q1 and 021-ACT1q2 (Appendix 9.2) were used to amplify the housekeeping gene ACT1. Fold change was calculated using the amended delta-delta-CT according to Pfaffl (Fig. 3.8) [196].
Figure 3.8 Summary of Copy shows an Increase and Decrease of Copy Number post-PTVA

Copy number was calculated via ΔΔCt using the qPCR primers 070-HSA-q1 and 071-HSA-q2 to amplifying the HSA gene with comparison to the ACT1 house keeping gene. Copy number was calculated as a comparison between parental clones and progeny clones. Z100 – GpαH#_100 clones; Z2000 – GpαH#_2000 clones.
PTVA successfully generated multi-copy clones; however this was not consistent for all clones analysed (Fig 3.8). In addition to clones showing increased copy number, there is evidence that clones retain a single copy of HSA post-PTVA. qPCR data suggests that some clones post-PTVA show a decrease in copy number in comparison to the parental counterparts. As hypothesised there was no evident increase in copy number within the GpαH15_2000 strain compared to its parental counterpart, which may, in part, explain the equivalent secreted protein titre (Fig 3.7). Clone GpαH16_100 appears to be a multi-copy clone, but there is evidence of a decrease in copy number in the progeny clone post PTVA. Instability observed post-PTVA has been reported for high copy clones (over 10 copies); however stability has not been seen for clones with less than six copies [210]. A decrease in copy number could potentially result in a decrease in protein production, although this does not seem to have been the case with GpαH16_2000 (Fig 3.7).

Sunga et al. highlighted the importance of clones integrating as complete vectors within the AOX1 region [149]. If only the Zeocin resistance marker had inserted into the vector then the clone would not benefit from an increase in copy number of the HSA. Southern blot was used to analyse whether an increase in copy number was as a result of complete amplification of the vector in a head to tail manner [149]. A digest with EcoRV was used to linearise the genomic DNA in the CYC1 transcription termination region. A single copy should produce a band at 8 Kb, while multiple copy clones will show an additional band at approximately 5.4 Kb (Fig. 3.9).
Figure 3.9 Southern blot of genomic DNA digested with EcoRV from a selection of parental and progeny clones of PTVA determines a head-to-tail insert orientation. A) Schematic diagram to depict the sites of digestion using EcoRV in multi-copy colonies in a head-to-tail orientation. B) Samples were digested with EcoRV and probed with DIG-labelled HSA. 1: GpαH1_100; 2: GpαH1_2000; 3: GpαH3_100; 4: GpαH3_2000; 5: GpαH6_100; 6: GpαH16_2000; 7: GpαH21_100; 8: GpαH21_2000; 9: GpαH23_100; 10: GpαH23_2000.

From the five clone pairs selected it is evident that there are multiple copies of the HSA gene in all of the progeny strains, including GpαH16_2000. Interestingly both GpαH16_100 and GpαH21_100 are multi-copy strains, with the progeny strains retaining multiple copies. Should the vectors have integrated in a head-to-head orientation bands of approximately 2.5 Kb and 4.4 Kb would be present. Additionally if bands integrated in a tail-to-tail orientation a 7.8 Kb band and 1.9 Kb band would have been visible. Head-to-tail integration (as suggested by Sunga et al.) would result in an additional 5.4 Kb band. The presence of the band at approximately 6 Kb indicates that multimers have indeed integrated in a head-to-tail manner as suggested by Sunga et al. [149].

The method of spotting cultures raises questions about the purity of each culture. Zeocin resistance occurs through sequestering the antibiotic, but if cultures are growing in spots there is potential that
a form of “protection” may occur. Theoretically, if one colony is producing an excessive amount of Zeocin resistant protein then colonies within the same spot may be able to survive without increased numbers of Zeo(R) genes, and hence lower numbers of HSA genes. Theoretically spots may be mixed cultures; thus colonies existing containing different copy numbers, which may ultimately affect titre. To determine microbiological purity, seven clones were selected from the 2000 μg mL\(^{-1}\) plate and streaked to single colonies on fresh YPD plates containing 2000 μg mL\(^{-1}\) Zeocin. Four to five colonies were then selected from each plate and genomic DNA extracted before copy number was established by qPCR using primers 070-HSAq1 and 071-HSAq2(Fig. 3.10).
Figure 3.10 Copy number calculated by qPCR of strains from the parental clone (selected on 100 μg mL$^{-1}$ Zeocin) and progeny clones which were streaked to single copies indicating a mixed culture.

Each spot grown in the presence of 2000 μg mL$^{-1}$ Zeocin was streaked to single colonies and four to five individual colonies were selected and were named A-E accordingly. Copy number was calculated via the ΔΔCt method using qPCR primers 070HSA-q1 and 071-HSA-q2. A comparison to GpdH1_100 was used to establish copy number using the ACT1 gene as the housekeeping gene.
qPCR results show that spots contain mixed cultures with individual colonies displaying different copy numbers (Fig 3.10). Interestingly, through both qPCR and Southern blot GpαH16_100 appears to be a multi-copy clone, but there is clear evidence through qPCR that post-PTVA not only does copy number decline (Fig 3.8), but some clones revert to containing a single copy of the HSA gene (Fig 3.10). This would imply that multi-copy clones generated through PTVA may not be stable; an issue that requires further investigation.

3.4 Clones that grow in the presence of 2000 μg mL⁻¹ Zeocin but do not show an increase in copy number

Through the stepwise exposure to higher concentrations of Zeocin some clones showed an increase in copy number of the HSA gene, which according to the Southern blot resulted from an increase in copy number of the whole vector. However, some clones exhibited the ability to grow in the presence of 2000 μg mL⁻¹ Zeocin but did not show any increase in HSA copy number. It is possible that mutations have arisen within either the Zeocin resistance gene or the promoter regions that allow for an increase in protein activity or transcription of the Zeocin resistance gene resulting in resistance to higher antibiotic concentrations. In order to evaluate the effect of selecting on high concentrations of Zeocin the TEF1 and EM7 promoter regions, Sh ble gene and the CYC1 transcription terminator were amplified using primers 280-ZeoSeqF and 281-ZeoSeqR (Appendix 9.2) and the PCR product sequenced. Clones GpαH6_100, GpαH6_2000, GpαH14_100, GpαH14_2000, GpαH17_100, GpαH17_2000, GpαH22_100, GpαH22_2000, GpαH25_100 and GpαH25_2000 were selected based on the evidence from qPCR that copy number had not increased post-PTVA. HSA titre levels for the strains varied including some showing no increase in secretion (GpαH17_2000), some showing a decrease in protein secretion (GpαH6_2000, GpαH14_2000) and some that showed an increase in protein secretion (GpαH22_2000, GpαH25_2000) in the progeny clones compared to the parental clones.

pPICzα was included for sequencing, as well as four strains (GpαH1_100, GpαH1_2000A, GpαH11_100 and GpαH11_2000B), which showed an increase in copy number and expression post-PTVA. From the sequencing results it was evident none of the strains selected had a mutation in the Sh ble gene; however there were small variations in the TEF1 promoter region (Fig 3.11; Appendix 9.3).
Figure 3.11 Alignment of DNA sequences within the TEF1 promoter region from strains selected on 100 μg mL^{-1} Zeocin or 2000 μg mL^{-1} suggest variation within the promoter regions

Surprisingly strains Gp\textalpha H1_2000A and Gp\textalpha H11_2000B showed variation in the sequence within the promoter region. It is possible that this variation affected expression levels, adding to the ability to grow at high Zeocin concentrations. To investigate this issue further, mRNA levels of Zeo(R) gene were established from parental and progeny clones using RT-qPCR. The six parental and progeny strains analysed by DNA sequencing were expressed for 24 hours in rich methanol containing media in a 24-well microtitre plate. Parental and progeny strains were grown in the presence of 100 μg mL^{-1} Zeocin and 2000 μg mL^{-1} Zeocin respectively. After 24 hours growth 1 mL of culture was stored in RNA Later® (Applied Biosystems) before the RNA was extracted using the RiboPure Yeast RNA Extraction Kit (Applied Biosystems). Transcript levels of the mRNA from the Zeocin resistant (ZeoR) gene were calculated as relative expression between the parental and progeny clones using primers 022-Zeoq1 and 023-Zeoq2 (Fig. 3.12).
There was a significant difference in Zeo(R) mRNA produced in GpαH11_2000B, a five copy clone post-PTVA, compared with GpαH11_100 (two-tailed t(4) p=0.0062). This was to be expected as Southern blot analysis (Fig 3.9) indicated an amplification of the entire vector, including the Zeocin resistance marker. Predominantly where clones did not show an increase in HSA copy number there was still an increase in Zeo(R) transcript, although with the exception of GPαH22_2000 (two-tailed t(4) p=0.0284) this was not significant (Fig 3.12). GpαH17_100 showed reduced Zeo(R) mRNA compared to the other parental clones; furthermore GpαH17_2000 showed a decrease in Zeo(R) transcript level (two-tailed t(4) p<0.0001). Potentially there is a variation within GpαH17_100 and GpαH17_2000 that affects the ability to grow on high concentration of Zeocin. As there are no apparent differences within the promoter regions or Sh ble gene differences in resistance to Zeocin may be as a result of naturally occurring mutations which affect the permeability of Zeocin.

### 3.5 Analysis of Secretion Saturation in Multi-Copy Clones

Producing multi-copy clones is labour-intensive and potentially expensive (if PTVA is used); therefore if secretion plateaus at increased copy number pharmaceutical companies may decide that it is not worth the investment. To assess evidence for secretion saturation two families were selected for further study; GpαH1 and GpαH11. Clones were selected for further analysis based on copy number;
GpαH1_100 (1 copy), GpαH1_2000A (2 copy), GpαH11_100 (1 copy), GpαH11_2000A (3 copy) and GpαH11_2000B (5 copy). Due to issues of clonal variation [211-212], it was deemed important to analyse clones from family groups to ensure any differences in expression were the consequence of an increase in the HSA gene and not clonal variation.

Each clone was cultured in a 15L bioreactor (Braun Biostat ED) initially with a glycerol fed batch for approximately 18 hours, before induction with methanol once OD₆₀₀ 500-600 had been reached and wet cell weight (WCW) was 400-500 g L⁻¹ (Appendix 9.4). Methanol was fed for 48 hours at 20 mL h⁻¹ with samples taken at 24 hours and 48 hours. HSA titre was analysed from culture supernatants (after removal of cells) using the Albumin Blue fluorescence assay kit (Active Motif). Titre was calculated as mg L⁻¹ g DCW⁻¹ (Fig. 3.13).

![Figure 3.13](image)

**Figure 3.13** Titre levels of HSA (mg L⁻¹ DCW⁻¹) during 48 hour fed batch methanol induction

Samples were taken from 5 L bioreactors pre-induction (0 hours) and 24 hours and 48 hours post-induction with methanol. Error bars indicate represent the Albumin Assay run in triplicate.

All strains showed an increase in the titre of secreted HSA over 48 hours with the exception of GpαH11_2000B (5 copy), which exhibited a lower protein titre at 48 hours compared to that at 24 hours (Fig 3.13). A possible reason for this decrease in expression could be the cells reaching secretion saturation coupled with the induction of the UPR, which then leads to the induction of the ERAD and protease production [213]. Furthermore, GpαH11_2000A (3 copy) appears to have a higher titre of HSA after 48 hours than GpαH11_2000B (5 copy). As these clones were from the same family this suggests that inserting five copies exerts additional stress, thereby inducing the UPR, ultimately leading to protein degradation through the activation of ERAD. Although the initial rate of
expression at 24 hours is higher, expression for 48 hours results in less protein that the equivalent three copy strain. If the ERAD is activated within GpαH11_2000B (5 copy) this can lead to increased cell death, which would result in increased proteases released into the supernatant. This suggests that high copy numbers may actually be detrimental to protein expression after 48 hours expression.

Additionally, when comparing the different families, GpαH1 and GpαH11, there is significant variation in HSA titre secretion between GpαH1_100 and GpαH11_100 after 48 hours (two-tailed t(4) p=0.0007). Furthermore GpαH1_100 has comparable titre to that of GpαH11_2000A (3 copy clone) after 24 hours. However, GpαH1_2000A (2 copy) clone secreted HSA titre equivalent to the parent clone GpαH1_100. We can hypothesise that either an increase in HSA copy to two does not result in increased expression, or the parental strain of GpαH1_100 produces enough protein to reach secretion saturation. This would suggest that secretion saturation in GpαH1 is lower than in GpαH11. This highlights the issue of clonal variation and underlines the importance of screening a variety of clones to select the best secretor.

### 3.6 Effects of Gene Copy-Number on induction of UPR

Increased copy number has been noted to result in upregulation of UPR, which could impact expression titre [87]. In order to investigate whether UPR was upregulated, RT-qPCR was used to analyse the expression of UPR associated genes, such as HAC1, KAR2 and PDI. Samples taken at 0 hours (pre-induction) as well as 24 hours and 48 hours post-induction were analysed by RT-qPCR (Fig. 3.14).
Figure 3.14  Upregulation of UPR during 48 Hour Fermentation

HAC1, KAR2 and PDI genes expression was calculated via ΔΔCt method using qPCR primers 010-KAR2q1, 011-KAR2q2, 012-PDIq1, 013-PDIq2, 067-HAC1q3 and 068-HAC1q4 respectively. ACT1 was the housekeeping gene and results were calculated as a comparison to corresponding 0 hour cultures. Fold change was calculated for samples from GpαH1_100 (1_100), GpαH11_100 (11_100), GpαH1_2000A (1_2000A), GpαH11_2000A (11_2000A) and GpαH11_2000B (11_2000B) taken at 0 hours, 24 hours and 48 hours.
GpαH11_100 shows increased upregulation of UPR compared to GpαH1_100 (Fig 3.14). As GpαH1_100 shows increased titre compared to GpαH11_100, it is possible to hypothesise that there is an additional secretory stress on GpαH11_100 (Fig 3.14), which may result in the induction of ERAD; thus resulting in reduced titre. UPR induction levels in multi-copy clones at 48 hours were higher that the corresponding parental clones (Fig 3.14). The decrease in HAC1 upregulation of GpαH11_2000B (5 copy) after 48 hours suggests that due to decreased protein production (as suggested by titre levels, and potentially due to the induction of ERAD) stress on the secretory pathway will be reduced; thus reducing HAC1 upregulation. The continued upregulation of HAC1 seen after 48 hours for GpαH11_2000A (3 copy) suggests that the high production maintains secretory stress. Potentially if yield was observed until 72 hours GpαH11_2000A (3 copy) may show similar trends as GpαH11_2000B (5 copy) in terms of protein production and UPR stress.

3.7 Genetic Instability in Multi-Copy Clones

To investigate whether a counter-selection against high copy clones existed (which would reduce the stress on the secretory pathway) a population study was carried out on individual colonies selected during the fermentation run to determine copy number post-expression. Samples were taken from 0 hours (pre-induction), 24 and 48 hours post-induction and were serially diluted and plated onto 100 μg mL⁻¹ Zeocin YPD plates as well as YPD plates with no antibiotic. Colony counts were used to compare the ability to grow in the presence of Zeocin (Fig. 3.15).

![Figure 3.15 Colony count (x10^7) of serial dilutions plated onto either YPD plates (no selective pressure) or 100 μg mL⁻¹ Zeocin YPD plates shows a loss of resistance to Zeocin. Samples were taken from the appropriate 15L bioreactor at 0 hrs (pre-induction), 24 hours or 48 hours post-methanol induction. Error bars represent standard error.](image-url)
From the comparison of colonies grown in the presence or absence of Zeocin it appears that, particularly after 48 hours, not all colonies have the ability to grow in the presence of Zeocin, implying they have lost either the Sh ble gene or the entire construct has been lost. This is particularly noticeable in GpαH1_2000A (2 copy) after 48 hours. Interestingly GpαH11_2000B appears to have a reduced ability to grow in the presence of Zeocin even after 0 hours (pre-induction) which is typically 18 hours into the fermentation run. Due to the number of repeats of the AOX1 promoter through integration of five pα-HSA vectors the regions of homology for recombination is significantly increased. This suggests in the absence of secretory stress clones show evidence of instability.

In order to determine the extent of genetic instability individual colonies were analysed for HSA copy number by qPCR. Single colonies were selected from the YPD plates (no selection) from the colony count. It was important to select colonies from the YPD plates and not those containing Zeocin, as this removed any selective pressure against maintaining the integrated vector. Genomic DNA was extracted and qPCR was used to analyse copy number of the HSA gene using ACT1 as a housekeeping gene (Fig. 3.16).

Figure 3.16 Genetic instability is prevalent in GpαH11_2000B (5 copy) during a 48 Hour Fermentation
Copy number was calculated by ΔΔCt using primers 070-HSAq1 and 071-HSAq2 to amplify the HSA gene. ACT1 was used as the housekeeping gene with reference to GpαH1_100, a known one copy clone. Single colonies of GpαH11_2000B (5 copy) were analysed at 0 hours (pre-induction), 24 hours and 48 hours post-induction.

GpαH11_2000B (5 copy) shows varying HSA copy number during the fermentation run (Fig 3.16). Copy number both increased and decreased at both 0 hours and 24 hours. At 0 hours pre-induction HSA copy number was eight in one of the clones analysed. However; after 48 hours post-induction...
copy number was retained at five in only one colony, with all other colonies showing a decrease in copy number. This may be due to secretory stress caused by additional traffic through the secretory pathway, which would result in increased counter-selection against high copy clones. There is evidence that some clones lose the HSA gene completely at both 0 hours and 24 hours. This supports evidence from reduced colonies isolated on YPD plates containing 100 μg mL⁻¹ Zeocin, as the loss of the HSA gene is likely to correspond to a loss of the entire vector therefore showing an inability to grow in the presence of Zeocin (Fig 3.15). Theoretically if selection of GpαH1_2000A (2 copy), GpαH11_2000A (3 copy) and GpαH11_2000B (5 copy) had been grown in the presence of 2000 μg mL⁻¹ Zeocin the number of resistant colonies may have decreased. This raises concerns regarding the stability of the vectors, an issue initially raised with GpαH16_2000 (Fig 3.10).

The presence of variable copy number raises questions on the applicability of multi-copy clones to improve protein production, especially given the variation in copy number prior to induction (Fig 3.16). Pre-induction there should be no protein production stress on the secretory pathway (as the cells are growing in glycerol and the AOX1 promoter is inactive); therefore the loss of copy number may be due to a more fundamental issue. In order to investigate this further, copy numbers from cells banks were examined to determine whether the initial starting culture was a genetically pure culture. To do this a cell bank vial was thawed and spread onto YPD plates and single colonies selected for genomic extraction and copy number was analysed by qPCR (Fig. 3.17).

![Figure 3.17 Genetic Instability is present in the cell banks of GpαH11_2000B (5 copy) determined by qPCR of HSA genes](image_url)

Copy number was analysed via ΔΔCt using qPCR primers 070-HSAq1 and 071-HSAq2 to amplify the HSA gene with comparison to ACT1 as the housekeeping gene. Copy number for individual colonies from GpαH11_2000B (5 copy) were selected from YPD plates and calculated as a comparison to GpαH1_100.
Cell bank cultures showed mixed HSA copy number indicating that the inoculum was not genetically homogeneous (Fig 3.17). Due to the repeat homologous sequences to the AOX1 promoter it is evident that recombination is occurring resulting in both an increase and decrease in HSA copy number. Furthermore, as no significant selective pressure was applied against an increase in copy number (as no expression was occurring) then cells with higher copy numbers were not counter-selected. When preparing the fermentation inoculums selection was maintained at 100 μg mL\(^{-1}\) Zeocin. Potentially if cell banks were prepared maintaining 2000 μg mL\(^{-1}\) Zeocin selective pressure even higher copy number may have been observed.

To confirm that genetic instability was not isolated to GpαH11_2000B (5 copy) copy number was analysed for GpαH1_100, GpαH11_100, GpαH1_2000A and GpαH11_2000A. Single colonies were selected, after serial dilution, from cultures in cell banks, at 0 hours (pre-induction), 24 hours and 48 hours post-induction (Fig. 3.18).
Figure 3.18 Genetic Instability occurs in GpαH1_100 (1 copy), GpαH11_100 (1 copy), GpαH1_2000A (2 copy) and GpαH11_2000A (3 copy) during a 48 hour Fermentation. HSA copy number was analysed via the ΔΔCt method using qPCR primers 070-HSAq1 and 071-HSAq2, with comparison to ACT1 as the house keeping gene. Individual colonies were selected from each fermentation run at the following time points; cell bank, 0 hours (pre-induction), 24 hours and 48 hours post-induction compared to GpαH1_100. A) GpαH1_100, a one copy clone. B) GpαH11_100, a one copy clone. C) GpαH1_2000A, a two copy clone. D) GpαH11_2000A, a three copy clone.

All strains showed mixed HSA copy number from initial cultures in cell banks through to 48 hours post-induction indicating genetic instability (Fig 3.18). This implies that instability is not restricted to GpαH11_2000B (5 copy). Genetic instability will have negative effects on production, as potentially clones with lower copy numbers will produce less protein, which has critical implications to using multi-copy strains as an approach for augmenting protein yields, particularly in industrial settings. Due to instability any conclusions regarding titre levels and copy number may be irrelevant, including those made with reference to secretion saturation.
3.8 Discussion

The use of PTVA to develop multi-copy clones has recently become a preferred method by the scientific community to generate multi-copy clones [149, 152, 210]. However, our investigation into this method has identified issues that have not previously been examined in detail; increase in copy number is not consistent for all clones and genetic instability results in mixed cultures.

Our initial findings showed that clones that exhibited an increased resistance to Zeocin did not consistently result in an amplification of the HSA gene (Fig 3.8). As PTVA works through amplification of the entire vector (Fig 3.9), this implies that no increase of the Sh ble gene occurred. However, it was noted that clones showed an increase in Zeo(R) transcript levels determined by RT-qPCR (Fig 3.12). We can conclude that resistance to high concentrations of Zeocin does not solely occur through amplification of the Sh ble gene. Potentially mutations within the Sh ble gene may increase resistance; however DNA sequencing did not show any differences to the original pPICzα A vector. Conversely, variations were observed within the TEF1 and EM7 promoter regions, similar to differences in GpαH11_2000B (5 copy). Potentially it is these differences that result in increased production of transcript, but further analysis would be required to determine the full effects of these variations. Site-directed mutation within the promoter regions would be able to identify whether specific changes impact Zeo(R) mRNA transcript levels.

GpαH17_2000 was the only strain that showed a decrease in Zeo(R) transcript post-PTVA, yet this did not impact resistance to high concentrations of Zeocin. We hypothesised that a natural variation within the strain may affect the permeability of the cell to Zeocin. This highlights issues of clonal variation, with some strains showing increased natural resistance to the antibiotic. Regardless of the mechanism clones that showed an increased resistance to high concentration of Zeocin despite no increase in Sh ble copy number emphasise the need for a meticulous screening method to determine complete amplification of the entire vector.

Further problems with PTVA arose from the method of spotting colonies, as mixed copy number colonies were observed within a spot. Zeocin resistant proteins bind stoichiometrically to Zeocin antibiotic to inhibit strand cleavage activity. Therefore, if a high copy clone is producing an abundance of resistance protein the antibiotic in the surrounding area may be inhibited; thus allowing for growth of lower copy clones on the high concentrations of Zeocin. For future implementation of PTVA it would be essential to reassess the feasibility of spotting clones onto plates, as mixed copy number spots may be detrimental to yields. Theoretically using a liquid version of PTVA, where cultures are grown in increasing concentrations of Zeocin before being plated onto the highest level of Zeocin YPD plates, would prevent mixed copy number cultures from occurring.
Multi-copy clones were initially created to establish at what copy number secretion saturation occurs using a well-folding protein. Unfortunately it became apparent from analysing 15 L bioreactor samples that genetic instability was prevalent on an individual cell basis. Previous investigations into genetic instability utilize RT-qPCR on a populational level and do not take into consideration the behaviour of individual colonies [214]. The presence of unstable clones means we have been unable to determine the relationship between secretion saturation and UPR through HSA production. A plateau in titre may be due to secretion saturation as suggested by Hohenblum et al. [87], or the counter-selection against high copy clones.

The identification of genetic instability on an individual cell basis raises questions regarding the reliability of PTVA. Zhu et al. showed clones with less than six copies of the PIP gene were stable [210]; however we have observed genetic instability in all of the clones tested including single copy clones. This will have detrimental effects on titre; as if colonies within fermentation cultures contain no copies of the recombinant gene volumetric titre will be reduced. In addition to instability observed through expressing strains evidence of genetic instability was observed under non-expressing conditions. GpαH16_100 was identified as a multi-copy clone but post PTVA GpαH16_2000 progeny clones showed evidence of reversion to single copy clones. Instability under non-inducing conditions was also observed when analysing copy number from cell bank cultures maintaining selection at 100 μg mL⁻¹ Zeocin. This suggests that despite the absence of secretory stress genetic instability is inherent. Due to repeat homologous regions of the AOX1 promoter events of homologous recombination frequency may be increased. If clones are not secreting proteins (and therefore there is no secretory stress) counter-selection may be reduced explaining why clones show both an increase and decrease in copy number. To reduce genetic instability multi-copy clones without multiple homologous repeat regions may be required.

If multi-copy clones are to be used extensively by biopharmaceutical companies it is imperative that our understanding of genetic instability improves, with investigations into the possibility of creating stable clones.
4 Investigation of the Prevalence of Genetic Instability in Multi-Copy Clones of *Pichia pastoris*

4.1 Introduction

Multi-copy clones are prevalent in the production of therapeutic proteins using *Pichia pastoris* and are regarded as highly stable [76, 142, 146, 206]. Determination of copy number is by Southern blot or more recently average copy number by RT-qPCR [152, 210, 215]. However since 1998 evidence of genetic instability has appeared in publications; Ohi et al. reported that a two-copy clone of human serum albumin (HSA) integrated into the HIS4 locus grown for a total of 83 generations (equivalent to approximately 163 hours) lost the foreign gene in 0.01% of cells [216]. However the titre of recombinant protein was not investigated and the effect of genetic instability on protein production remains unknown. In addition in 2002 it was reported that an excess of methanol feed at 4 g L\(^{-1}\) resulted in the loss of a single copy of human chymotrypsinogen-B [217], potentially due to the increased stress caused by the excess induction of the AOX1 promoter.

Generation of multi-copy clones via post-transformational vector amplification (PTVA) has resulted in two papers that analyse the stability of this technique. Zhu et al. integrated up to 52 copies of the porcine insulin precursor (PIP) into GS115 through PTVA, but found only clones with six copies or less were stable [151, 210]. Conversely Marx et al. found clones containing 11 copies of the HSA gene were stable when integrating into the ribosomal DNA (rDNA) locus [152]. In contrast to analysing whole populations by RT-qPCR, our data suggests that individual cell evaluations do not represent stable cultures during expression.

4.1.1 Current Investigations into Improving Genetic Stability

Methods for preventing genetic instability will be investigated in an attempt to produce a more stable multi-copy strain.

4.1.1.1 Stability through Integration at the rDNA locus

In GS115 there are 16 repeats of the rDNA locus [92] and as a result it would be possible that 16 copies of the vector could be inserted into these regions. Furthermore, as rDNA is essential for growth a selective pressure is constantly maintained. This method of using the rDNA locus has also been used in the methyltrophic yeast *H. polymorpha* with similar effects of stable multi-copy clones being achieved [218].
4.1.1.2 Stability through a KU70 Knockout Strain

Naatsaari et al. generated a KU70 knockout strain, preventing the non-homologous end joining (NHEJ) recombination pathway [215]. Non-homologous integration occurs in P. pastoris and is detrimental to protein expression due to different loci having an impact on titre, especially when using auxotrophic strains. Naatsaari et al. showed that a side effect of the KU70 knockout strain, containing seven copies of a GFP based plasmid with an AOX1 promoter, was maintained stability similar to levels discussed by Zhu et al. [210].

4.1.2 RecA Homologs in P. pastoris

RecA strains are used in E. coli to prevent instability of transformed plasmids [219-220]. In eukaryotes the homolog for RecA is RAD51, a protein involved in the recombination pathway in single strand invasion and found in many different eukaryotes [221-223]. Bioinformatic analysis (blastn) revealed that the sequence similarity of RecA and RAD51 was 26.5%. However protein homology (blastp) was 67% [224]. There are two highly homologous regions; the surrounding area of the putative nucleoside triphosphate-binding sequence and the carboxy terminus [222-223]. These regions correspond to domains responsible for recombination, UV resistance and formation of the active oligomers [223]. In mice a mutation in MmRAD51 results in embryonic arrest during early development [225-226]. However in S. cerevisiae a RAD51 knockout is not lethal due to the multiple repair pathways that exist [227-229].

Due to known predicted between S. cerevisiae and P. pastoris it was hypothesised that creating a knockout of a gene involved in the homologous recombination (HR) pathway, post-integration of the vector of interest, would result in a more stable strain. The RAD51 gene is involved in the HR pathway whereas RAD52 affects both the HR and NHEJ pathways (Fig 1.5). We hypothesise that a RAD51 knockout strain may not prevent genetic instability, as it will only affect the HR pathway. Whereas a RAD52 knockout, which affects both recombination pathways, may result in a more stable strain, but could induce lethal mutations. In order for multi-copy clones to reach high titres the issue of stability needs to be addressed. Therefore genetic instability and prevalence of this phenomenon was investigated to generate stable clones.

4.1.3 Aims for Genetic Instability

- To establish whether maintaining constant selection pressure, using Zeocin, affects the prevalence of instability.
- To compare a HIS4 complementation selection method against antibiotic resistance.
- To investigate the genetic stability in P. pastoris of RAD51 or RAD52 knockout strains.
- To evaluate the affects of directed integration into the rDNA locus on genetic instability.

4.2 Maintaining Zeocin Selection through Expression Does Not Prevent Genetic Instability

In the previous chapter genetic instability was shown in all strains grown in 15 L bioreactors, including single copy clones, which questions the stability of clones in general (Fig 3.18). We hypothesised that continued selective pressure would reduce the prevalence of instability of GpαH11_2000B, a five copy clone. Selective pressure was maintained in all media through the presence of 2000 μg mL⁻¹ Zeocin. GpαH11_2000B was inoculated into 3 mL of BMGY in microtitre plates prior to induction of protein expression with methanol for a further 24 hours. Following which serial dilutions were carried out and single colonies isolated on YPD plates (no Zeocin selection). Genomic DNA was extracted from colonies using the Plant DNeasy mini-prep kit (Qiagen) and copy number determined by qPCR using HSA primers. ACT1 was used as the housekeeping gene to determine copy number by ΔΔCt (Fig 4.1).
Figure 4.1 Clones of GpoH11_2000B Grown in the Presence of Constant Zeocin show evidence of Genetic Instability
HSA Copy number was determined by qPCR from individual colonies following 24 hours expression in microtitre plates. 2000 μg mL⁻¹ Zeocin was used to maintain selective pressure. A) 0 hours, pre-induction. B) 24 hours, post induction.

Growth in the presence of 2000 μg mL⁻¹ Zeocin ensured that an overall higher copy number was maintained (Fig 4.1). In the initial glycerol growth phase HSA copy number ranged from two to nine, indicating an increase and decrease relative to the initial five copies present pre-expression.

Following methanol induction copy number ranged between two and six. The decrease in the HSA higher copy number could be due to an increased level of secretory or expression-associated stress. Previously the 15 L bioreactor expression profile in the absence of selective pressure, showed copy number ranging from zero to seven copies (Fig 3.16). While it appears that maintaining selective pressure results in increased genetic stability, as no clones lost the integrated vector completely, this
method is financially unfeasible on an industrial scale due to the high cost of Zeocin (Fig 4.1). Therefore we examined maintaining selection through the use of an auxotrophic selection marker.

### 4.3 Using HIS4 for selection instead of Zeocin does not result in Increased Stability

The auxotrophic selective marker, HIS4, does not require antibiotics and may be a preferred choice for creating stable multi-copy clones. GS115 contains a disrupted HIS4 gene [65, 76], and through complementation the inclusion of an intact HIS4 will result in the strain being able to produce histidine. Therefore any colonies that do not have an intact HIS4 integrated will not be able to grow on minimal media in the absence of histidine. However, the use of minimal media will result in a lower cell density and reduced titre due to an increase in protease activity [230].

Unfortunately, production of multi-copy clones by PTVA requires an increase in Zeocin (or another titratable antibiotic); therefore it is not possible to create multi-copy clones using an auxotrophic marker. For the selective process to work only one copy of the HIS4 gene is required to be integrated. Analysis of multi-copy clones using an auxotrophic marker requires the use of a vector, which contains multiple copies of the gene of interest, to ensure HIS4 integration. The multi-copy vector pα-Tx3 was created using the in-vitro multimerisation method to insert a further two copies of the gene of interest, trypsinogen (Bryn Edwards-Jones, unpublished). Previous experimentations from our lab indicated that pα-Tx3 integration into GS115 resulted in genetic instability under continuous culture (Bryn Edwards-Jones, unpublished). This may be due to the repetition of the three copies of the AOX1 promoter (Fig 4.3) which could lead to recombination in a similar way observed with the PTVA multi-copy strains.

![pα-Tx3 Plasmid Map](image)

**Figure 4.2** pα-Tx3 Plasmid Map

pα-Tx3 contains three copies of the expression cassette; AOX1 promoter, α-MF signal peptide, TRY1 and AOX1 transcription terminator region. These are repeated in tandem, and selection and origin of replication are represented by Zeocin resistance gene Sh ble, and the pUC origin.
pα-Tx3 contains three copies of the TRY1 gene as opposed to the HSA gene as previously used. This was due to the size of the HSA gene (1.8 Kb) which was too big to create multimers in-vitro. For growth on selective minimal media the pα-Tx3 was modified by replacing Zeo(R) resistance gene with HIS4 from pIB2 to create the vector pIBTx3. A restriction digest of pα-Tx3 with BglII and BamHI was used to extract the three expression cassettes which were then ligated together with the pMB1 ori, ampicillin (R) and HIS4 gene of pIB2 (Fig 4.3).
Figure 4.3 Vector Creation of pIBTx3

A) pIB2 Plasmid Map. The pIB2 vector contains the GAP promoter followed by the AOX1 transcription terminator region, and includes both an ampicillin resistance gene for selection in *E. coli* as well as the *P. pastoris HIS4* gene for complementation.

B) 1% Agarose gel of the vector creation of pIBTx3. 1: 1 Kb Ladder (Fermentas); 2: αTx3 digested with BglII and BamHI for extraction of the 7kb Try1 Expression cassette; 3: pIB2 inverse PCR to remove the GAP promoter; 4: pIBTx3 vector; 5: pIBTx3 digested with NheI.

C) pIBTx3 Plasmid Map. The vector pIBTx3 was constructed using a restriction digest of BglII and BamHI on α-Tx3 to release the expression cassette, and pIB2 which had the GAP promoter removed through inverse PCR and the BglII and BamHI restriction sites added to the ends.
pIBTx3 was transformed into *E. coli* JM109 for amplification of the plasmid and reisolated using the Qiagen mini-prep kit. The vector was sequenced to confirm correct construction (data not shown) and 5-10 ng of pIBTx3 was linearised for integration into the *HIS4* locus with NheI before being transformed into GS115. Colonies were selected on minimal dextrose (MD) plates with the absence of histidine for selection. Individual colonies were selected to determine stability by qPCR using the primers 383-Try1q7 and 384-Try1q8 (Fig 4.4; Appendix 9.2).

**Figure 4.4** Transformation of pIBTx3 into GS115 results in Varying Copy Numbers

Copy number was calculated via the ΔΔCt method using qPCR primers (383-Try1q7 and 383-Try1q8) to amplify the *TRY1* gene with comparison to *ACT1* as the housekeeping gene. GpαTx1 strain (darker column) was used as a single copy reference.

qPCR data shows that following transformation the entire construct did not fully integrate into the *HIS4* locus as there is evidence of clones containing a single *TRY1* gene (Fig 4.4). A digest using BglII will cut at the 5’ end of the first *AOX1* promoter region; however due to the size of the vector recombination may occur with the other *AOX1* region of homology. This would result in only one copy of the *AOX1-TRY1* expression cassette being integrated into the genome (Fig 4.4). In addition it would be possible to see two copies integrating of the second and third *AOX1* regions of homology, although this was not the case for the colonies that were tested (Fig 4.4). This highlights the need for an extensive screen if the *in vitro* multimerisation method is to be used.

While instability was evident in vector integration it is essential to know the number of *AOX1-TRY1* expression cassettes when expressed in minimal media. Two clones, GpIBTx3 3-3 and GpIBTx3 3-6, were confirmed to have integrated the entire pIBTx3 vector and therefore selected for further analysis (Fig 4.4). Expression of clones GpIBTx3 3-3 and GpIBTx3 3-6 was examined in 24-well microtitre plates grown for 24 hours in a glycerol based minimal medium (BMG) prior to induction with methanol for a further 24 hours. Post-expression GpIBTx3 3-3 and GpIBTx3 3-6 were serially
diluted and plated to single colonies on YPD plates (no selection) to determine copy number of the TRY1 gene. Individual colonies were selected and genomic DNA extracted using the Plant DNeasy mini-prep kit. TRY1 gene copy number was determined by qPCR using primers 383-Try1q7 and 384-Try1q8 and compared to GpαTx1, a single copy clone (Fig 4.6; Appendix 9.2).

![Figure 4.5 GpIBTx3 Clones show Genetic Instability post-24 hour Expression](image)

Copy number was calculated via the ΔΔCt method using qPCR primers (383-Try1q7 and 383-Try1q8) to amplify the TRY1 gene with comparison to ACT1 as the housekeeping gene. Light grey bars indicate colonies from GpIBTx3-3 dark grey bars represent colonies from GpIBTx3-6.

Post-expression TRY1 copy number ranged from one to three (Fig 4.5). This implies that genetic instability is still prevalent in GpIBTx3 3-3 and GpIBTx3 3-6 despite the maintenance of selection through the absence of histidine. This agrees with the findings by Ohi et al., which showed instability using a HIS4 complemented vector [216]. The in-vitro multimerisation method requires only one copy of HIS4 to maintain selection. Counter-selection against increased TRY1 copy number could result in recombination between the homologous AOX1 promoter regions to reduce the pressure on the secretory pathway. Therefore maintaining auxotrophic strain selection does maintain a level of stability, as no clones lost the plIBTx3 vector completely; however it is not a viable option for the use of creating multi-copy strains unless multiple auxotrophic markers are used.

### 4.4 RAD51- and RAD52- Knockout Strains are not viable

We hypothesised that by creating knockouts of the HR or NHEJ pathways genetic instability could be prevented. We investigated this by knocking out either the RAD51 or RAD52 genes. RAD51 and RAD52 knockout strains were constructed using a modified pα-Tx3 vector (Fig 4.2) that includes homology to 250 bp of 5’ and 3’ ends of the RAD51 or RAD52 genes to respectively create the pαTx3R51 and plIBTx3R52 vectors. The Gibson assembly method was used to anneal fragments together without the use of restriction sites (Fig 4.6).
Figure 4.6 Construction of the pαTx3R51 and pαTx3R52 Vectors

pαTx3R51 and pαTx3R52 were constructed using the Gibson Assembly method in order to add 250 bp of both the 5’ and 3’ end of either the RAD51 or RAD52 gene to the pαTx3 vector. Final linearisation using SpeI allowed for integration into P. pastoris at the RAD51 or RAD52 loci. A) Schematic Representation of the creation of pα-T3R51 using the Gibson Assembly Protocol B) 1% Agarose gel confirming the insertion of genes of the vectors: 1: 1 Kb Ladder (Fermentas); 2: pαTx3 linearised with BglII; 3: 5’ 250 bp RAD51; 4: 3’ 250 bp RAD51; 5: 5’ 250 bp RAD52; 6: 3’ 250 bp RAD52; 7: pαTx3R51; 8: pαTx3R52.
RAD51 and RAD52 homologous regions were ligated into pα-Tx3 and transformed into E. coli for amplification of the plasmid. Reisolated vectors were sequenced to confirm integration of the RAD51 and RAD52 fragments (data not shown). Isolated vectors were linearised with SpeI and transformed into P. pastoris with selection on 100 μg mL⁻¹ Zeocin. In previous transformations it was observed that 100 μL of transformed P. pastoris equated to 100 to 200 transformants. However the maximum number of colonies that resulted from a pαTx3R51 transformation was 10 colonies, while no colonies were isolated from the transformation of the pαTx3R52 vector.

Colony PCR was used to determine successful RAD51 knockout strains in pαTx3R51 transformants. Primers, 264-RAD51 SeqF1 and 265-RAD51 SeqR1 amplified the RAD51 gene to show gene replacement (Fig 4.7; Appendix 9.2).

![Figure 4.7 Amplification of the RAD51 gene showed no gene knockout post-integration of the pα-Tx3R51 vector](image)

Colony PCR on a 1% agarose gel was used to confirm pαTx3R51 integration into RAD51 using primers 264-RAD51 SeqF1 and 265-RAD51 SeqR1. Integration of the vectors will result in a fragment length of approximately 13 Kb, while incomplete gene knockout will result in a 1.1 K b fragment. 1: 1 Kb Ladder (Fermentas); 2-8: Individual colonies from transformation; 9: GS115

Integration of the pαTx3R51 vector into the RAD51 locus would show a 13 Kb band; however a band of this size may not be successfully amplified. However if the pαTx3R51 had not integrated into the RAD51 locus then the primers would amplify the uninterrupted 1.1 Kb RAD51 gene. The presence of the smaller band indicates that the pαTx3R51 vector had not integrated at the RAD51 locus (Fig 4.7). However integration of pαTx3R51 may occur at the AOX1 promoter and transcription terminator homologous regions. During gene replacement 10-20% of all transformants may integrate into the homologous regions as opposed to the RAD51 gene[120].

As no transformants displayed RAD51 or RAD52 gene knockouts this raised questions around the method for selection. Zeocin works through creating double stranded DNA breaks and if the recombination pathway has been compromised by the knockout then it is possible that strains were
unable to recover. While RAD51 mutants are not known to be lethal in *S. cerevisiae*, they are reported to have severe consequences in other eukaryotic systems, such as mouse *mMRAD51* which results in embryonic lethality [221, 223, 225-226].

To investigate whether using Zeocin has detrimental effects on the success of the RAD51 or RAD52 mutants pIBTx3R51 and pIBTx3R52 vectors were created. Both vectors rely on the auxotrophic selection marker *HIS4*. Selection through *HIS4* complementation removes the requirements of the strains to recover from Zeocin induced double stranded breaks. This vector was made using the Gibson assembly protocol, as described for pαTx3R51/52 but using the pIBTx3 vector that had been linearised with BglII for the backbone (Fig 4.8).

Figure 4.8 pIBTx3R51 and pIBTx3R52 Plasmid Maps
pIBTx3R51 and pIBTx3R52 were created using the Gibson Assembly method; whereby 250 bp from the 5' and 3' ends of the RAD51 and RAD52 genes were respectively complemented to the pIBTx3 vector.

The pIBTx3R51 and pIBTx3R52 vectors were transformed into *E. coli* JM109 for amplification and reisolated using the Qiagen mini-prep kit. The isolated vectors were sequenced to ensure integration of the RAD51 and RAD52 homologous fragments (data not shown). Isolated DNA was digested with EcoRV for transformation into *P. pastoris*. Colonies were selected on MD plates with the absence of histidine for selection. Transformation using the vectors pIBTx3R51 and pIBTx3R52 resulted in a maximum of 10 colonies successfully showing complementation of the *HIS4* gene (thus exhibiting growth on minimal media). These figures were similar the pαTx3R51 vector transformation efficiencies. Individual colonies were analysed by colony PCR to determine whether pIBTx3R51 and pIBTx3R52 had integrated into the RAD51 or RAD52 loci to determine if gene knockouts had been successful (Fig 4.9).
Figure 4.9 Amplification of the RAD51 or RAD52 genes showed no gene knockout post-transformation of the pIBTx3R51 and pIBTx3R52 vectors

Colony PCR on a 1% agarose gel was used to confirm integration of pIBTx3R51 and pIBTx3R52 using primers 264-RAD51 SeqF1 and 265-RAD51 SeqR1 or 266-RAD52 SeqF1 and 267-RAD52 SeqR1 respectively. Integration of the vectors would result in a fragment length of approximately 13 Kb, while incomplete gene knockout would result in a 1.1 or 1.2 Kb fragment respectively. **A)** Amplification of the RAD51 gene; **B)** Amplification of the RAD52 gene. 1: 1 Kb Ladder (Fermentas); 2-8: Individual colonies from transformation; 9: GS115

Amplification of the RAD51 and RAD52 genes shows no integration of the pIBTx3R51 or pIBTx3R52 vectors. This implies that the use of Zeocin was not the sole cause of the failure to generate RAD51 or RAD52 gene knockouts. Vector size has an impact in the rate of integration; therefore due to the size at 12.5 Kb it may be impacting successful transformation [112]. Furthermore pIBTx3R51 and pIBTx3R52 integration can occur at several places; AOX1 promoter, AOX1 transcription terminator region, HIS4 and the RAD51 or RAD52 loci. A linearised PCR fragment containing only the RAD51/RAD52 fragments and the HIS4 gene was designed to determine if a smaller fragment results in gene knockouts. By implementing this method, if integration does not occur at the RAD51 or RAD52 locus the only other region of homology for integration would be the HIS4 locus. To create
this linear PCR product the Gibson Assembly method was used to ligate 250 bp from the 5’ and 3’ of RAD51 or RAD52 to flank the HIS4 gene (Fig 4.10).

Figure 4.10 Creation RAD51-HIS4 and RAD52-HIS4 Linear DNA fragments

A) Schematic representation of the creation of RAD51-HIS4 and RAD52-HIS4 fragments. HIS4 was flanked by 250 bp of the 5’ and 3’ regions of homology to the RAD51 or RAD52 genes. B) 1% agarose gel confirming the creation of the RAD51-HIS4 linear DNA fragment 1: 1 Kb Ladder (Fermentas), 2: 250 bp RAD51 gene (5’), 3: HIS4 gene, 4: 250 bp RAD51 gene (3’), 5: RAD51-HIS4 linearised fragment constructed. C) 1% agarose gel confirming the creation of the RAD52-HIS4 linear DNA fragment 1: 1 Kb Ladder (Fermentas), 2: 250 bp RAD52 gene (5’), 3: HIS4 gene, 4: 250 bp RAD52 gene (3’), 5: RAD52-HIS4 linearised fragment constructed.
RAD51-HIS4 and RAD52-HIS4 linear DNA fragments were transformed into GS115 and transformants were selected on MD plates minus histidine for selection. Integration of the RAD51-HIS4 and RAD52-HIS4 linear DNA resulted in a low number of transformants, similar to integration of the pαTx3R51, pIBTx3R51 and pIBTx3R52 vectors. Colonies were analysed by colony PCR using primers 264-RAD51 SeqF1 and 265-RAD51 SeqR1 or 266-RAD52 SeqF1 and 267-RAD52 SeqR1 to determine the presence of RAD51 or RAD52 knockout genes respectively (Fig 4.11).

Figure 4.11 Amplification of the RAD51 and RAD2 genes showed no gene knockout using a linear DNA fragment
Colony PCR on a 1% agarose gel was used to confirm integration of RAD51-HIS4 and RAD52-HIS4 using primers 264-RAD51 SeqF1 and 265-RAD51 SeqR1 or 266-RAD52 SeqF1 and 267-RAD52 SeqR1 respectively. Integration of the linear DNA fragments would result in a fragment length of approximately 4 Kb, while incomplete gene knockout would result in a 1.1 or 1.2 Kb fragment respectively. A) Amplification of the RAD51 gene; B) Amplification of the RAD52 gene. 1: 1 Kb Ladder (Fermentas); 2-8: Individual colonies from transformation; 9: GS115

Amplification of the RAD51 and RAD52 genes showed no integration of the RAD51-HIS4 or RAD52-HIS4 linear fragments (Fig 4.11). Correct integration of the linear DNA fragments would amplify a 4 Kb band. However if the linear fragments had not integrated into the RAD51 or RAD52 loci the primers would amplify the uninterrupted 1.1 Kb RAD51 or 1.2 Kb RAD52 gene respectively. Therefore as integration has not occurred within the RAD51 or RAD52 genes integration may occur within the HIS4 gene due to the region of homology present in the DNA fragment. The colonies were analysed by colony PCR using the primers 392-HIS4-5’ and 337-RAD51-5-HIS or 342-RAD52-5-HIS to determine integration at the 5’ end of the HIS4 gene.
Primers 395-HIS4-3’ and 340-RAD51-3-HIS or 352-RAD52-3-HIS were used to determine integration at the 3’ end of the HIS4 gene. No bands were amplified implying that the RAD51-HIS4 and RAD52-HIS4 linear DNA fragments had not integrated into the HIS4 locus (data not shown).

As the linear fragment only contained regions of homology to two loci it implies that integration is occurring through non-homologous recombination into another locus in the genome. Evidence suggests that non-homologous recombination does occur with regular frequency in P. pastoris [121, 215]. An alternative possibility is that the vector may be integrating in tandem within either the RAD51 or RAD52 gene. If a knockout of RAD51 and RAD52 creates a lethal mutation a gene insertion event may occur as opposed to a gene replacement. Gene replacement occurs with less frequency than a gene insertion event [65, 120]. Unfortunately proving the presence of gene insertion as opposed to gene replacement was not possible in this instance (data not shown).

4.5 Genetic Stability was observed with Vectors Integrated into the rDNA Locus
According to Marx et al. integration of vectors into the rDNA locus results in the production of stable multi-copy clones [152]. Nonetheless stability was determined via analysis on whole populations and does not account for variations on individual cell basis. Therefore genetic instability was investigated on an individual cell basis for vectors integrated into the rDNA locus.

4.5.1 Vector Creation
To emulate Marx et al. the vector pGRzαHSA was constructed using the Gibson assembly method [152]. The α-MF signal peptide, HSA gene, AOX1 transcription terminator and Zeo(R) resistance gene were isolated from the pα-HSA vector. The GAP promoter was extracted from pIB2 and the rDNA locus amplified using the primers 370-rDNA-pUC GibF and 371-rDNA-GAP GibR.(Appendix 9.2; Figure 4.12) [152].
In addition to replicating the vector from Marx et al. the vector, pARzaHSA, replacing the GAP promoter with the AOX1 promoter was created. This allowed comparison to multi-copy clones generated by PTVA using the AOX1 promoter (Fig 3.20). We have hypothesised genetic instability may be due to the volume of protein being passed through the secretory pathway and corresponding secretory stress. The GAP promoter produces less protein than the AOX1 promoter; thus stability observed by Marx et al. may be as a consequence of reduced traffic through the secretory pathway (Fig 4.13).
4.5.2 Initial integration into the rDNA locus

pGRzaHSA and pARzaHSA were transformed into *E. coli* JM109 for amplification of the plasmid and reisolated using the Qiagen mini-prep kit. The reisolated vectors were sequenced (data not shown) before linearisation using SpeI in the rDNA gene. Linearised pGRzaHSA and pARzaHSA were transformed into GS115 and plated on YPD plates containing 100 μg mL\(^{-1}\) Zeocin. Colony PCR was used to confirm that pGRzaHSA and pARzaHSA had integrated into the rDNA locus using primers 352-rDNA IntF and 353-rDNA IntR (Fig 4.14; Appendix 9.2).

Integration of pGRzaHSA and pARzaHSA was confirmed by colony PCR (Fig 4.14). GpGRzaHSA2 shows a brighter band (particularly for amplification of the 5′ integration site) than the other strains analysed. This could suggest the presence of multiple copies of vectors integrated into the rDNA loci. HSA gene copy number was determined by qPCR using primers 070-HSAq1 and 071-HSAq2 and compared to GpαH1_100, a single copy clone (Fig 4.15).
Figure 4.15 Integration into the rDNA Locus results in Multi-Copy Clones

Copy number was calculated via the ΔΔCt method using qPCR primers (70-HSA-q1 and 71-HSA-q2) to amplify the HSA gene with comparison to ACT1 as the housekeeping gene. Light grey bars indicate colonies containing the pARzaHSA vector and dark grey bars represent colonies from containing the GpGRzaHSA vector.

GpGRzaHSA2 was confirmed as a multi-copy clone containing nine copies of the HSA gene, which correlates to the brighter band seen through analysis by colony PCR (Fig 4.14 and 4.15). Previously the highest copy number observed through integration into the AOX1 locus was five copies. Due to the high copy number of GpGRzaHSA2 titre was determined as a comparison to previous multi-copy clones. Clones were expressed for 24 hours in either a glycerol or methanol containing medium depending on the promoter used. Protein levels were determined by the Albumin Blue Fluorescence Assay (Fig 4.16).
GpGRαHSA2 produced 13.9 μg mL\(^{-1}\), which is comparable to titre produced by GpαH1_100, a single copy clone of HSA integrated into the AOX1 locus (Fig 4.16). GpGRαHSA uses the GAP promoter for expression; therefore may not be comparable to GpαH1_100, which uses the AOX1 promoter. Integration into different loci has been known to result in different titre levels, for instance integration into the AOX1 promoter produces more protein than integration into the HIS4 locus [121, 215].

GpGRαHSA and GpARαHSA were analysed for HSA transcript levels to determine whether integration into the rDNA locus varied from clones with vectors integrated into the AOX1 locus. RNA samples were extracted after 24 hours expression in glycerol (GAP promoters) or methanol (AOX1 promoters), using the RiboPure Yeast RNA Extraction kit (Applied Biosystems). HSA gene expression levels were analysed by RT-qPCR using primers 070-HSAq1 and 071-HSAq2 and the ACT1 as the house keeping gene and compared to GpαH1_100, GpαH11_100, GpαH11_2000A (3 copy) and GpαH11_2000B (5 copy) (Fig 4.17; Appendix 9.2).
**Figure 4.17 Clones with Integration into the rDNA Locus show Reduced Transcript Levels compared to Clones with Integration in the AOX1 Locus**

HSA gene expression was calculated via the ΔΔCt method using qPCR primers 70-HSA-q1 and 71-HSA-q2 to amplify the HSA gene with comparison to ACT1 as the housekeeping gene as a comparison to GpαH1_100, a known one copy clone. Statistically significant values of p<0.05 (compared to GpαH1_100) are indicated by a *.

GpARzαHSA1 and GpGRzαHSA2 produced low levels of transcript compared to the strains integrated into the AOX1 locus (Fig 4.17). GpARzαHSA1is a one copy clone therefore HSA transcript should be comparable to GpαH1_100 and GpαH11_100. The low HSA titre suggests a reduced flow of traffic through the secretory pathway, which can be monitored by analysing the HAC1, KAR2 and PDI gene expression levels. Additional stress on the secretory pathway may have a negative effect on protein production. Transcript levels were analysed by RT-qPCR using the ACT1 gene as the housekeeping gene and as a comparison to GpαH1_100, a known one copy clone (Fig 4.18).
GpARzaHSA1 showed reduced upregulation of UPR compared to the single copy clones integrated into the AOX1 (Fig 4.18). This implies that the reduced HSA transcript levels (Fig 4.18) produced by GpARzaHSA resulted in less stress on the secretory pathway (Fig 4.21). GpGRaHSA2 (9 copy) shows a 10 fold upregulation of HAC1 gene transcript levels compared to GpaH11_2000A (3 copy). Therefore despite low HSA transcript levels there is an upregulation of UPR as a result of increased protein production. As GpGraHSA2 uses the GAP promoter and not the AOX1 promoter, 24 hours may not be sufficient for protein production; therefore, for future experiments strains containing the GAP promoter were expressed for 48 hours.

4.5.3 Creating Multi-Copy Clones of rDNA Integrated Vectors

PTVA was used to create multi-copy clones to determine whether integration into the rDNA locus affects titre compared to multi-copy clones integrated into the AOX1 locus. As using PTVA on plates resulted in mixed populations (Fig 3.10) a liquid variant of PTVA was implemented. Colonies were inoculated into 24-well microtitre plates with 3 mL YPD containing 100 μg mL⁻¹ Zeocin. After 24 hours the cultures were centrifuged and resuspended in YPD containing an increased concentration of Zeocin and left to grow for a further 24 hours. This was continued until Zeocin concentration reached 2000 μg mL⁻¹. As with the initial PTVA experiment concentrations of Zeocin used were 100,
200, 300, 500, 1000 and 2000 μg mL\(^{-1}\). After 24 hours growth at the 2000 μg mL\(^{-1}\) Zeocin cultures were serially diluted and plated onto YPD containing 2000 μg mL\(^{-1}\) Zeocin and left to incubate at 30°C for 3-5 days.

Post-PTVA individual colonies were expressed in BMGY and BMMY media for 24 hours for strains containing the AOX1 promoter and for 48 hours in BMGY for strains containing the GAP promoter. Heterologous protein production was visualised on an SDS-PAGE gel using the spent broth from expression to ensure that the α-mating factor (MF) secretion signal was active (Fig 4.19).

![Figure 4.19 HSA Expression post-PTVA of Vectors Integrated into the rDNA Locus](image)

HSA expression was determined by expression for 48 hours in BMGY (GAP promoter) or 24 hours in BMMY (AOX1 promoter) and analysed on a 12% SDS-PAGE gel. 1: GpαH1_2000A; 2: Pre-stained molecular marker (Fermentas); 3: GpGRzαHSA2_2000C; 4: GpARzαHSA1_2000E.

Post-PTVA GpGRzαHSA and GpARzαHSA clones were confirmed to secrete HSA (Fig 4.19). A comparison between copy number and titre was analysed to determine the affects of multi-copy clones integrated into the rDNA locus. Genomic DNA was extracted using the DNeasy Plant mini-prep kit (Qiagen) and qPCR was used to analyse copy number using the primers 070-HSAq1 and 071-HSAq2 compared to GpαH1_100, a known single copy clone (Appendix 9.2). Expression of HSA was quantified using the Albumin Blue Fluorescence Assay (Active Motif) after expression for 24 hours in BMMY (AOX1 promoter) or 48 hours in BMGY (GAP promoter)(Fig 4.20).
Copy number was calculated via the ΔΔCt method using qPCR primers 070-HSAq1 and 071-HSAq2 to amplify the HSA gene with comparison to ACT1 as the housekeeping gene. HSA titre was determined by the Albumin Blue Fluorescence assay (Active Motif) after expression for 24 hours in methanol or 48 hours in glycerol depending on the promoter used. pGRzaHSA (GAP promoter) is represented in blue and pARzaHSA (AOX1 promoter) in purple.

GpGRzaHSA shows 38 copies of the HSA successfully integrated into the rDNA locus. Previously only five copies of the HSA gene had successfully integrated into the AOX1 locus by PTVA. The average copy number of clones integrated with pGRzaHSA was 19 copies, whereas with the pARzaHSA vector this was 11 copies. Variation in average copy number between clones containing the AOX1 promoter and GAP promoter may be due to the limitation in the number of individual clones tested and not due to any differences between the two promoters.

Average expression levels of clones containing the pGRzaHSA and pARzaHSA vectors were 46 and 44 μg mL⁻¹ respectively, two-tailed t(10) p=0.6511 (Fig 4.20). Multi-copy clones resulted in a three-fold increase in titre compared to single copy clones integrated into the AOX1 locus (Fig 4.16 and 4.20). There was no correlation between copy number and titre for either of the vectors analysed with $r^2$ values at 0.0126 and 0.035 (Fig 4.20). As HSA titre did not increase with an increase of HSA genes it implies that secretion saturation has been reached.

HSA transcript levels were analysed by RT-qPCR using primers 070-HSAq1 and 071-HSAq2 to determine whether an increase in copy number increased the amount of transcript that was being
produced, despite no increase in titre. Of the clones that were generated by PTVA only two were selected for further experimentation based on copy number. As with all multi-copy clones selected through PTVA clones were named as follows; GpGRzαHSA#_2000 (GH#_2000) or GpARzαHSA#_2000 (AH#_2000) followed by A-F indicating the relevant colony. The clones containing the highest and lowest number of HSA genes were selected; GpARzαHSA1_2000A (3 copy), GpARzαHSA1_2000C (25 copy), GpGRzαHSA2_2000C (38 copy) and GpαGRzHSA2_2000F (3 copy). GpaH1_100, GpaH11_100, GpaH11_2000A (3 copy) and GpaH11_2000B (3 copy) were included as a comparison to clones integrated into the rDNA locus (Fig 4.22).

![Figure 4.21](image)

**Figure 4.21** HSA Transcript is Increased Post-PTVA for Clones with Vectors Integrated into the rDNA Locus Compared to Clones with Vectors Integrated into the AOX1 Locus

HSA gene expression was calculated via the ΔΔCt method using qPCR primers 070-HSAq1 and 071-HSAq2 to amplify the HSA gene with comparison to ACT1 as the housekeeping gene as a comparison to GpaH1_100, a known one copy clone. PTVA strains are indicated by 2000 and the relevant colony (A-F) chosen. GpARzαHSA1 (AH1), GpGRzαHSA2 (GH2), GpaH1_100 (H1_100), GpaH11_100 (H11_100), GpaH11_2000A (H11_2000A) and GpaH11_2000B (H11_2000B). Statistically significant values of p<0.05 (compared to GpaH1_100) are indicated by a *.

HSA transcript levels of GpARzαHSA strains show a 13 fold increase compared to GpGRzαHSA clones (Fig 4.21). For both GpARzαHSA and GpGRzαHSA clones with the higher copy numbers produced increased HSA transcript compared to the corresponding low copy clones. GpGRzαHSA2_2000F (38 copy) showed a two fold increase in HSA transcript compared to GpGRzαHSA2_2000C (3 copy). However, a 1.8 fold increase of HSA transcript between GpARzαHSA1_2000A (3 copy) and GpARzαHSA1_2000C (25 copy) did not correlate to an increase in titre as respectively each secreted 52 μg mL⁻¹ and 43 μg mL⁻¹ of HSA (Fig 4.24). Furthermore GpGRzαHSA2_2000C (38 copy) secreted 43
μg mL⁻¹ while GpGRzαHSA2_2000F (3 copy) secreted 49 μg mL⁻¹ of protein despite transcript levels being higher in the 38 copy clone. This implies that blockages within the secretory pathway prevent high levels of HSA transcript being converted into protein. Furthermore the decrease in titre of GpGRzαHSA2_2000F (38 copy) and GpARzαHSA1_2000C (25 copy) compared to the low copy clones suggests that degradation may be occurring due to secretory stress in the high copy clones. To determine if secretory stress is upregulated in the high copy clones HAC1, KAR2 and PDI transcript levels were determined by RT-qPCR (Fig 4.22).

Figure 4.22 Gene Expression Levels of UPR Signals Upregulated in Clones Containing Vectors integrated into the rDNA Locus Compared to Integration at the AOX1 Locus HAC1, KAR2 and PDI gene expression were calculated via the ΔΔCt method using qPCR primers 010-KAR2q1, 011-KAR2q2, 012-PDIq1, 013-PDI1q2, 067-HAC1q3 and 068-HAC1qr respectively. ACT1 was the housekeeping gene and fold change was calculated as a comparison to GS115. PTVA strains are indicated by 2000 and the relevant colony (A-F) chosen. GpARzαHSA1 (AH1), GpGRzαHSA2 (GH2). Statistically significant values of p<0.05 (compared to GS115) are indicated by a *. HAC1 was upregulated in all of the multi-copy clones integrated into the rDNA locus compared to vectors integrated into the AOX1 locus (Fig 4.22). PDI and KAR2 transcript levels were similar to clones integrated into the AOX1 locus. Only GpGRzαHSA2_2000F (3 copy) showed increased expression of KAR2 (two tailed t(4) p=0.0016). The significant upregulation of HAC1 in the multi-copy clones integrated into the rDNA locus suggests that UPR is upregulated as a control mechanism to manage the flow of proteins through the secretory pathway [160]. HAC1 transcript levels for GpGRzαHSA2_2000C (38 copy) and GpARzαHSA1_2000C (25 copy) are increased compared to GpGRzαHSA2_2000F (3 copy) and GpARzαHSA1_2000A (3 copy) respectively. Due to the decrease in titre of the high copy clones it can be hypothesised that the high HAC1 induction may be an indication of the activation of the ERAD, which occurs through prolonged UPR induction.
4.5.4 Stability of Multi-Copy Clones

To determine genetic stability of vectors integrated into the rDNA locus cultures were expressed for 24 hours in methanol (for clones containing the AOX1 promoter) or 48 hours in glycerol (for clones containing the GAP promoter). Post-expression cultures were serially diluted and plated onto YPD plates (no selection). Individual colonies were selected post-expression for copy number analysis. Genomic DNA was extracted directly from the colonies and analysed by qPCR using the primers 070-HSAq1 and 071-HSAq2 with comparison to GpαH1_100, a known one copy clone (Fig 4.23).

![Graphs showing copy number analysis](image)

**Figure 4.23 Clones with vectors integrated into the rDNA Locus using both the AOX1 and GAP promoter show evidence of Genetic Instability**

Copy number was calculated via the ΔΔCt method using qPCR primers 070-HSAq1 and 071-HSAq2 to amplify the HSA gene with comparison to ACT1 as the housekeeping gene. In all graphs the first (and darker) column represents the original copy number before expression. A: GpARzαHSA1_2000A (AH1_2000A). B: GpARzαHSA1_2000C (AH1_2000C). C: GpGRzαHSA2_2000C (GH2_2000C). D: GpGRzαHSA2_2000F (GH2_2000F).

Post-expression all four strains analysed showed genetic instability with clones both increasing and decreasing in HSA copy number (Fig 4.23). GpGRzαHSA2_2000C was initially a 38 copy clone, but post expression copy number ranged from two to 28 copies of the HSA gene. Conversely GpGRzαHSA2_2000F was initially as a three copy clone but post-expression copy number ranged from seven to 46 copies. Copy number of GpARzαHSA1_2000A ranged from a single copy clone up to a four copy clone. GpARzαHSA1_2000C showed a decline in copy number to 10 copies of the HSA gene with one colony maintaining the initial copy number of 25.

No clones with vectors integrated into the rDNA locus contained zero copies of the HSA gene as was seen in clones with vectors integrated into the AOX1 locus (Fig 3.16 and 3.18). Multi-copy clones
integrated into the rDNA locus induce stress in the secretory pathway (Fig 4.22), which could act as a counter-selection against stability. However, the stress of a single copy clone on the secretory pathway may be insufficient to act as a counter-selection (Fig 4.18). Whilst investigation on whole populations suggests integration into the rDNA is stable [152] single colony analysis suggests that the repeat sequences of the rDNA locus are recombining post-integration leading to unstable clones.

### 4.6 Discussion

Genetic instability can reduce the efficiency of recombinant expressing strains, which is detrimental to pharmaceutical companies producing biopharmaceuticals as clones may not reach their expressing potential. Moreover development of multi-copy clones is time consuming and potentially costly (especially if PTVA is used with concentrations of Zeocin reaching 2000 μg mL⁻¹). Genetic instability in *P. pastoris* is currently assessed via RT-qPCR or Southern blot on a population level culture to determine copy number. However, by isolating individual colonies and assessing copy number by qPCR we have ascertained that genetic instability is prevalent (Fig 3.16 and 3.18).

To investigate whether maintaining a selective pressure had an effect on genetic instability GpαH11_2000B, a five copy clone, was expressed maintaining a concentration of 2000 μg mL⁻¹ Zeocin. Clones isolated post-expression and analysed by qPCR resulted in colonies with lower HSA copy number (2 copy) and an increased copy number (9 copy) in non-inducing conditions. In inducing conditions copy number ranged from two to six copies (Fig 4.1). This implies that high-copy numbers are counter-selected when grown in inducing conditions, as seen by Zhu et al. [210]. No clones lost the expression vector completely as seen when expressed in the absence of Zeocin (Fig 3.16 and 3.18). Therefore maintaining a selective pressure does increase the stability of multi-copy clones. However, the cost of Zeocin at approximately £210 g⁻¹ is an unrealistic expense for industrial companies.

To determine stability via maintaining selection through an auxotrophic marker the vector pIBTx3 was designed, which relies on the complementation of *HIS4* and contains three copies of the *TRY1* gene. Transformation of the pIBTx3 vector resulted in clones with only one copy of the *AOX1-TRY1* cassette integrated into the genome (Fig 4.4). As larger vectors are more difficult to transform into *P. pastoris* the 12 Kb vector could reduce transformation efficiency [231]. Additionally as three copies of the *AOX1* region of homology are included in the vector the likelihood of integration of only part of the vector is increased. Recombination between the homologous *AOX1* promoter regions would result in a smaller linearised DNA inserting into the genome.
Despite some strains showing partial integration of pIBTx3 clones that contained three TRY1 genes were analysed for stability post-expression. Minimal medium with the absence of histidine was used to maintain the selective pressure. Post-expression clones contained either a single copy, two copies or three copies of the TRY1 gene implying genetic instability (Fig 4.5). However, the maintained stability ensured that no clones lost the vector entirely. This implies that there is counter-selection against clones containing multiple copies of the TRY1. As TRY1 is known the induce UPR [87], multi-copy clones may result in secretory stress leading to the activation of the ERAD. A single copy of TRY1 was maintained through the use of the HIS4 auxotrophic marker. Therefore multiple auxotrophic complementation may be required for the production of stable multi-copy clones, for example the addition a ADE4 knockout, such as strain GS200 [65]. However using strains that require multiple auxotrophic complementation require the use of highly defined minimal media, which could result in lower titre due reduced cell density and increased protease activity [232].

RAD51 and RAD52 mutants were designed to determine whether interrupting the HR or the HR and NHEJ pathways had an effect on genetic instability. It was hypothesised that by simultaneously inserting a vector into the genome and interrupting the recombination pathways regions of homology between the vector and the genome would be prevented from recombining to lose the vector. A variety of attempts to create RAD51 and RAD52 knockout strains were unsuccessful. In 2012 a KU70 knockout strain was successfully developed by Naatsaari et al. designed to interrupt the NHEJ recombination pathway [215]. While RAD51 and RAD52 knockout strains were similar in concept to the KU70 knockout strain both affected the HR pathway. While S. cerevisiae knockout strains of RAD51 are stable there is evidence in other eukaryotic organisms that a RAD51 mutant causes severe deleterious effects [225-226]. We can hypothesise that in P. pastoris interrupting the HR pathway results in a lethal mutation; thus explaining the absence of RAD51 or RAD52 knockout strains.

The last attempt to create stable multi-copy clones was based on an experiment by Marx et al. with integration of multi-copy clones directed into the rDNA locus [152]. Vectors were designed using both the GAP promoter and the AOX1 promoter. Genetic instability was observed for all strains analysed post-expression in non-selective media (Fig 4.23). Stability of the rDNA loci in S. cerevisiae is dependent upon the relocalisation of the DNA double-stranded breaks from the nucleolus to the nucleoplasm, before it associates with the recombination machinery. It is this relocalisation, which relies on the Smc5-Smc6 complex and sumoylation of Rad52p that ensures the stability of the rDNA loci [233]. Potentially integration of the exogenous DNA may affect the relocalisation prior to exposure to the recombination machinery, resulting in increased instability.
Clones containing higher copy numbers of the HSA gene showed increased genetic instability compared to lower copy clones, an effect reported by Zhu et al. [210]. GpGRzählHSA2_2000F (3 copy) and GpARzählHSA1_2000A (3 copy) showed higher titre levels than the high copy clones GpGRzählHSA2_2000C (38 copy) and GpARzählHSA1_2000C (25 copy). Lower copy number clones also showed reduced induction of UPR indicating less stress on the secretory pathway. Additionally saturation of the secretion pathway when expressing HSA has been reported to occur when five copies of the gene are present [152]. Current thoughts are that multi-copy clones are stable when up to six copies of the gene of interest are present [152, 210, 215]. Therefore the creation of high copy clones may be unnecessary for the production of recombinant proteins when expressing HSA. Furthermore as the transformation onto 100 μg mL^{-1} Zeocin resulted in nine copies of the HSA gene high-copy number clones created through the use of PTVA may be unnecessary. While we have been unable to generate clones that show no evidence of genetic instability, the inclusion of low multi-copy clones through integration into the rDNA locus consistently results in the maximum titre of HSA without excess stress on the secretory pathway.
5 The Effects of Using a Codon Optimised Gene for Expression in *Pichia pastoris*

5.1 Introduction

5.1.1 Codon Optimised Genes in *P. pastoris*

Codon optimisation functions on the premise that different organisms translate particular codons more effectively than others [234-235]. Through either point mutations or by buying a genetically synthesised gene it is possible to select codons, which can result in increased titre expressed from the synthesised gene. Development of codon optimisation varies between companies; often it is suggested that replacing rare synonymous codons with favoured codons for the expressing host will result in increased titre [236]. However DNA 2.0 suggest that it is not the selection of more frequently used codons that result in increased titre but tRNAs that are most highly charged during amino acid starvation [237].

Codon optimisation has been implemented in *P. pastoris*; however this method does not consistently result in increased titre [65]. The production of codon optimised human glucocerebrosidase protein, showed a 10.6 fold increase post-optimisation [238]. Additionally the production of equistatin, which was codon optimised to *P. pastoris* resulted in a four to 10-fold increase at protein level [239]. Woo *et al.* suggest that codon optimisation is essential for the production of anti-T cell immunotoxin in *P. pastoris* [240]. Yet whilst there are examples of increases in efficiency through codon optimisation there are counter arguments that suggest that this process is not necessary [59]. Sinclair and Choy suggested that altering the codon usage changes the A/T ratio of the foreign gene, and it was this ratio that resulted in increased expression. This technique appeared to have a similar level of success as that of codon optimisation [238].

To date no codon optimised HSA gene has been evaluated for titre. As HSA secretion reaches 10 g L$^{-1}$ optimisation via media or growth conditions may be sufficient [140]. However, we have observed that production of HSA result in induction of UPR (3.6). Therefore using an optimised gene of interest could potentially reduce secretion stress, as translational processes will be accelerated by readily available tRNAs. A codon optimised HSA gene purchased for GenScript USA Inc. (Piscataway, NJ, USA) was used to investigate both titre of an optimised gene compared to the native HSA for both single and multi-copy clones.

5.1.2 Aims

- To determine the effects of using an optimised HSA gene on titre and UPR
To examine whether the use of an optimised HSA gene affects stability by reducing counter-selection against multi-copy clones integrated into the rDNA locus.

### 5.2 Single Copy Clones of HSAoptimised gene

The HSA optimised (HSAopt) gene purchased from GenScript was synthesized to be codon optimised for *P. pastoris*. GenScript applies the OptimumGene algorithm for developing an optimised gene taking into consideration not only preferred codon usage but translation efficiencies and protein refolding (Appendix 9.1). As with previous HSA vectors pα-HSAopt was formed of a single copy of the HSAopt gene ligated into the pPICzα vector at the EcoRI and NotI sites (Fig 5.1).

![Figure 5.1 pα-HSAopt Plasmid Map](image)

*Figure 5.1 pα-HSAopt Plasmid Map*

Vector design of HSA optimised gene incorporated into the pPICzα vector using EcoRI and NotI. PmeI restriction site linearises the vector for integration into the AOX1 region of the genome.

pα-HSAopt was transformed into JM109 *E. coli* and reisolated using the Qiagen mini-prep kit. The vector was digested with PmeI, which cuts within the AOX1 promoter for linearisation. The linearised vector was transformed into GS115 and plated onto YPD plates containing 100 μg mL⁻¹ Zeocin. HSA titre was determined for single copy clones to determine the effect of using a codon optimised gene. Clones were expressed in 24-well microtitre plates for 24 hours in BMGY before being induced with methanol in BMMY for a further 24 hours. GpαH1_100 and GpαH11_100 strains were expressed simultaneously for reference. The spent media broth was collected through centrifugation and HSA titre measured using the Albumin Blue Fluorescence Assay (Active Motif) (Fig 5.2).
Clones that contained the HSAopt gene showed comparable protein titre compared to GpαH1_100 and GpαH11_100 (Fig 5.2). GpαHopt7 produced over twice as much HSA than either of the GpαH1_100 or GpαH11_100 strains, with past experience suggesting that GpαHopt7 may be a multi-copy clone. Copy number was analysed by qPCR using primers 188-HSAoptq3 and 189-HSAoptq4 compared to the housekeeping gene ACT1 (Appendix 9.2; Fig 5.3).
GpαHopt7 was identified as a multi-copy strain, containing two copies of the HSAopt gene, which may partially account for the higher expression of HSA detected (Fig 5.2). Additionally GpαHopt2 was identified as a multi-copy strain, containing two copies of the HSAopt gene; however titre levels were comparable to that of the HSAopt single copy clones such as GpαHopt4 and the single copy clones GpαH1_100 and GpαH11_100 (Fig 5.2). This variability reinforces the evidence for clonal variation, a problem which should be considered in future experimental designs.

5.3 Generation of Multi-Copy Clones of the HSAopt Gene in the rDNA locus

5.3.1 Vector Creation

Experience of using the HSA native gene to create multi-copy clones suggests that to produce clones with the highest copies (and with increased stability) vectors should be integrated into the rDNA locus (4.5.3). Vectors were designed as described for pGRzaHSA and pARzaHSA (Fig 4.12 and 4.13). The pα-HSAopt vector was modified to incorporate regions encoding the rDNA locus for the pARzaHSAopt vector. For the pGRzaHSAopt vector the GAP promoter replaced the AOX1 promoter (Fig 5.4).

![Figure 5.4 pGRzaHSAopt and pARzaHSAopt Plasmid Maps](image)

The vectors pGRzaHSAopt and pARzaHSAopt were created using the Gibson Assembly Protocol modifying the original pα-HSAopt vector, containing the HSAopt gene. For the pGRzaHSAopt vector the GAP promoter replaced the AOX1 promoter (Fig 5.4).

The pARzaHSAopt vector was constructed with the rDNA sequences integrated between the AOX1 transcription terminator region and the TEF1 promoter region, which differed from the pGRzaHSAopt vector, which integrated the rDNA locus between the AOX1 promoter and the pUC ori. Subsequently, a revised construct was designed (Fig 5.5) to confirm that the design of the vector had no impact on its integration into the genome or on HSA titre. This was possible with vectors containing the HSAopt gene due to the removal of the BglII site from within the sequence (Appendix
9.1), which was an essential restriction site, lending itself to the same arrangement as was seen in the pGRzα vectors (Fig 5.5).

![Diagram of pARzαHSAopt BgIII Plasmid Map](image)

**Figure 5.5 pARzαHSAopt BgIII Plasmid Map**

Vector with the rDNA locus inserted before the AOX1 promoter as compared to the previous pARzαHSAopt vector where it lies between the TEF1 promoter and the AOX1 transcript terminator. This vector was created for comparison to the pGRzαHSAopt vector.

### 5.3.2 Initial integration of HSAopt vectors into the rDNA locus

pGRzaHSAopt, pARzaHSAopt and pARzaHSAopt BgIII were transformed into *E. coli* JM109 for amplification of the plasmid and reisolated using the Qiagen mini-prep kit. The reisolated vectors were sequenced (data not shown) before linearisation using SpeI in the rDNA gene. Linearised pGRzaHSAopt, pARzaHSAopt and pARzaHSAopt BgIII were transformed into GS115 and plated on YPD plates containing 100 μg mL⁻¹ Zeocin. Colony PCR was used to confirm integration into the rDNA locus using primers 352-rDNA IntF and 353-rDNA IntR (Fig 4.14; Appendix 9.2). As seen with the HSA constructs, a band of 3 Kb for both forward and reverse reactions would suggest integration into the rDNA Locus (Fig 5.6).
Figure 5.6 Colony PCR Indicated Correct Integration of pGRzaHSAopt and pARzaHSAopt vectors into the rDNA locus

Colony PCR was carried out to check integration of different transformants into the rDNA locus. **A** Integration at the 5’ end of the vector; **B** Integration at the 3’ end of the vector.

1: 1 Kb DNA Ladder (Fermentas); 2: pARzaHSAopt 3; 3: pARzHSAopt 4; 4: pARzaHSAopt BglII 2; 5: pARzaHSAopt BglII 8; 6: pGRzaHSAopt 3; 7: pGRzaHSAopt 4

Integration of GpGRzaHSAopt, GpARzaHSAopt and GpARαHSAopt BglII into the rDNA locus was confirmed by colony PCR (Fig 5.6). As with the integration of the pARzaHSA and pGRzaHSA vectors (Fig 4.14) there were bands that appeared brighter than others, suggesting the possible presence of multi-copy clones. Copy number was determined by qPCR using primers 188-HSAoptq3 and 189-HSAoptq4 compared to GpaHopt1_100 (Fig 5.7).

Figure 5.7 Integration into the rDNA Locus results in Multi-Copy Clones

Copy number was calculated via the ΔΔCt method using qPCR primers (188-HSAoptq3 and 189-HSAoptq4) to amplifying the HSAopt gene with comparison to ACT1 as the housekeeping gene.
Single and multi-copy clones were isolated from clones transformed onto 100 μg mL⁻¹ Zeocin (Fig 5.7). No clones showed integration of nine copies as seen with GpGRzαHSA2 (Fig 4.15). However, four copies were present in both GpGRzαHSAopt 3 and (Fig 5.7). GpARzαHSAopt BglII 8 was identified as a three copy clone; thus the use of the AOX1 promoter for integration into the rDNA locus was comparable to using the GAP promoter for generating multi-copy clones. Clones were expressed for 24 hours in either a glycerol or methanol containing medium depending on the promoter used. Protein levels were determined by the Albumin Blue Fluorescence assay (Fig 5.8).

![Figure 5.8 HSA Titre from Clones Integrated into the rDNA Locus Containing the HSAopt Gene](image)

HSA titre was determined by the Albumin Blue Fluorescence assay (Active Motif) after 24 hours growth in methanol medium or glycerol medium depending on the promoter used. Error bars represent triplicate biological repeats.

Clones produced titre similar to single copy clones of GpARzαHSA and GpGRzαHSA, which were observed to be lower than clones with vectors integrated into the AOX1 locus (Fig 5.2 and 5.8). Therefore it was of interest to determine HSAopt transcript levels of clones integrated into the rDNA locus. HSAopt transcript levels were analysed by RT-qPCR using primers 188-HSAoptq3 and 189-HSAopt q4 using ACT1 as the housekeeping gene and compared to GpαHopt1_100 (Fig 5.9).
**Figure 5.9 HSAopt Transcript Levels of Clones Integrated into rDNA locus were lower than a Single Copy Clone Integrated into the AOX1 Locus**

HSA gene expression was calculated via the ΔΔCt method using qPCR primers 188-HSAoptq3 and 189-HSAoptq4 as a comparison to ACT1 as the housekeeping gene. Fold change was calculated as a comparison to GpαHopt1_100. Statistically significant values of p<0.05 (compared to GpαHopt1_100) are indicated by a *.

HSAopt gene transcript levels were significantly reduced for clones integrated into the rDNA locus compared to GpαHopt1_100 (Fig 5.9). This correlates to low HSA titre levels observed (Fig 5.8). To observe the impact of low titre on UPR HAC1, KAR2 and PDI gene expression levels were analysed by RT-qPCR as a comparison to GpαHopt1_100, a known one copy clone (Fig 5.10).

**Figure 5.10 Transcript Levels of Genes involved in UPR of Clones Containing the HSAopt Gene Integrated into the rDNA Locus**

HAC1, KAR2 and PDI transcript levels were calculated via the ΔΔCt method using qPCR primers 010-KARq1, 011-KAR2q2, 012-PDIq1, 013-PDI1q3, 067-HAC1q3 and 068-HAC1q4 respectively. ACT1 was the housekeeping gene and resulted were calculated as a comparison to GS115. Statistically significant values of p<0.05 (compared to GS115) are indicated by a *.

129
GpGRzαHSAopt3 and GpGRzαHSAopt4 showed over four-fold induction of KAR2 compared to GpαHopt1_100 (Fig 5.10). PDI also shows significant upregulation in both GpGRzαHSAopt3 (two tailed t(4) p=0.0275) and GpGRzαHSAopt4 (two tailed t(4) p=0.0045). Conversely there is no significant upregulation of HAC1 in any of the strains with vectors integrated into the rDNA locus compared to GpαHopt1_100 (Fig 5.10). Upregulation of UPR signals in GpGRzαHSAopt3 and GpGRzαHSAopt4 suggest that the increase in gene dosage correlates to increased flow through the secretory pathway. However the low levels of HAC1 suggest that a detrimental level of UPR upregulation may not yet be occurring, which is supported by low HSA transcript (Fig 5.9). As determined using GpGRzαHSA expression for 24 hours using the GAP promoter may not be sufficient for protein production; therefore expression was extended to 48 hours for future experiments.

5.3.3 Amplification of rDNA vectors by PTVA to generate Multi-Copy Clones

Multi-copy clones of GpARzαHSA and GpGRzαHSA showed increased titre as well as an indication of secretion saturation (Fig 4.20). The use of a codon-optimised gene might a) affect the level that secretion saturation occurs and b) affect the secretory stress of the cells (through ease of translation due to readily available tRNAs) potentially reducing counter-selection against multi-copy clones.

PTVA was used to create multi-copy clones as described for GpARzαHSA and GpGRzαHSA. Clones were inoculated into YPD containing 100 μg mL⁻¹ Zeocin and after 24 hours growth the medium was changed with an increase in concentration of Zeocin. Stepwise concentrations were 100, 200, 300, 500, 1000 and 2000 μg mL⁻¹ Zeocin. After growth on the maximum concentration of Zeocin cultures were serially diluted and plated onto YPD plates containing 2000 μg mL⁻¹ Zeocin and left to incubate for 5 days at 30°C.

Post-PTVA colonies were expressed in BMGY and BMMY media for 24 hours for the GpARzαHSAopt strains and for 48 hours in BMGY for GpGRzαHSAopt strains. Heterologous protein production was visualised on an SDS-PAGE gel using the spent broth from expression (Fig 5.11).
Figure 5.11 HSA Expression post-PTVA of Multi-Copy Clones containing the HSAopt Gene Integrated into the rDNA Locus

HSA expression was determined by expression for 48 hours in BMGY (GAP promoter) or 24 hours in BMMY (AOX1 promoter) and analysed on a 12% SDS-PAGE gel. 1: GpαH1_100; 2: Pre-stained molecular marker (Fermentas); 3: GpGRzaHSAopt3_2000B; 4: GpGRzaHSAopt4_2000A; 5: GpARzaHSAopt3_2000F; 6: GpARzaHSAopt BglII 2_2000E; 7: GpARzaHSAopt BglII 8_2000D.

Post-PTVA all strains analysed were confirmed to secrete HSA (Fig 5.11). A comparison between copy number and titre was used to determine the advantages of a HSAopt gene versus a native HSA gene. Genomic DNA was extracted using the DNeasy Plant mini-prep kit (Qiagen) and qPCR was used to analyse copy number using the primers 188-HSAoptq3 and 189-HSAoptq4 (Appendix 9.2). Expression of HSA was quantified using the Albumin Blue Fluorescence Assay (Active Motif). For comparison the results from GpGRzaHSA and GpARzaHSA clones were included for reference (Fig 5.12).
Figure 5.12 Multi-copy clones containing the HSAopt Gene produce lower HSA titre that when using the HSA gene

Copy number was calculated via the ΔΔCt method using qPCR primers 188-HSAoptq3 and 189-HSAoptq4 to amplifying the HSAopt gene against ACT1 the housekeeping gene. Copy number was compared to GpoHopt1_100. HSA titre was determined by the Albumin Blue Fluorescence assay (Active motif) after 24 hours expression in methanol (AOX1 promoter) or 48 hours in glycerol (GAP promoter). **A: Vectors integrated containing the GAP promoter.** Turquoise – GpGRzaHSA; Purple – GpGRzaHSAopt. **B: Vectors containing the AOX1 promoter.** Green – GpARzaHSA; Pink – GpARzaHSAopt; Blue – GpARzaHSAopt BglII.

For all clones analysed there was no significant correlation between copy number and titre (Fig 5.12). Furthermore there was no significant difference of expression levels between clones integrated with the AOX1 promoter or the GAP promoter (two tailed t(40) p=0.5481). Nevertheless
there is a consistent decrease in HSA titre when using the HSAopt gene compared to using the native HSA gene (two tailed t(40) p=0.0007). Two clones that were generated by PTVA, containing the highest and the lowest copy number, were selected for further experimentation; GpARzαHSAopt 3_2000E (5 copy), GpARzαHSAopt 3_2000F (21 copy), GpARzαHSAopt BgIII 2_2000D (4 copy), GpARzαHSAopt BgIII 2_2000E (15 copy), GpARzαHSAopt BgIII 8_2000D (7 copy), GpARzαHSAopt BgIII 8_2000F (24 copy), GpGRzαHSAopt 3_2000B (14 copy), GpGRzαHSAopt 3_2000F (22 copy), GpGRzαHSAopt 4_2000D (15 copy) and GpGRzαHSAopt 4_2000E (24 copy).

GpARzαHSAopt and GpGRzαHSAopt clones appear to reach secretion saturation earlier than clones integrated with the HSA gene (Fig 5.12). Therefore it was of interest to compare HSA transcript levels between the two variants. HSAopt transcript levels were analysed by RT-qPCR using primers 188-HSAoptq3 and 189-HSAoptq4 as a comparison to GpαHopt1_100. For comparative purposes HSA transcript levels were included (Fig 5.13).

Figure 5.13 Quantification of HSA Transcripts of Clones Post-PTVA, Integrated into the rDNA Locus using either the HSAopt gene or HSA gene
HSA gene expression was calculated via ΔΔCt method using qPCR primers 188-HSAoptq3 and 189-HSAoptq4 to amplify the HSAopt gene or 070-HSAq1 and 071-HSAq2 to amplify the HSA gene. ACT1 was used as the house keeping gene with comparison to either GpαH1_100 or GpαHopt1_100. Light grey bars indicate clones containing the HSAopt gene and dark grey bars indicate clones containing the HSA gene. Statistically significant values of p<0.05 (compared to GpαH1_100 or GpαHopt1_100) are indicated by a *.

Transcript levels of clones containing the HSAopt gene show increased levels compared to GpαHopt1_100 (Fig 5.13). GpARzαHSAopt BgIII 8_2000D, which shows the lowest fold change indicates a four-fold increase in HSA transcript compared to a known one copy clone integrated into the AOX1 locus. This implies that multi-copy clones successfully result in the increase in HSA transcript levels. GpARzαHSA produces more HSA transcript than the GpARzαHSAopt strains, which correlates to the increased titre observed in the native HSA containing strains (Fig 5.12). However,
GpGRzaHSAopt strains show increased HSA transcript compared to GpGRzaHSA strains, despite average expression being lower. This suggests potential blockages within the secretory pathway of the GpGRzaHSAopt strains that prevent conversion of transcript to folded protein. Therefore secretion stress was examined to evaluate upregulation of UPR. HAC1, KAR2 and PDI transcript levels were determined by RT-qPCR as a comparison to GS115 (Fig 5.14).

Figure 5.14 Gene Expression levels of UPR signals upregulated in clones containing the HSAopt gene compared to clones containing the HSA gene. HAC1, KAR2 and PDI transcript levels were determined by RT-qPCR as a comparison to GS115. PTVA strains are indicated by 2000 and the relevant colony (A-F) chosen. GpARzaHSAopt3 (AO3), GpaARzaHSAopt BglII2 (B2), GpaARzaHSAopt BglII8 (B8), GpGRzaHSAopt3 (GO3), GpGRzaHSAopt4 (GO4), GpARzaHSA1 (AH1), GpGRzaHSA2 (GH2). Statistically significant values of p<0.05 (compared to GS115) are indicated by a *.

HAC1 transcript levels were upregulated in both GpARzaHSA and GpGRzaHSA compared to their counterpart HSAopt clones (Fig 5.14). Conversely the induction of KAR2 and PDI in clones expressing the HSAopt gene using the AOX1 promoter exceeded those with the HSA gene. PDI and KAR2 are both ER chaperones available to assist with folding of the protein [153, 241]. Upregulation of both PDI and KAR2 implies that despite the lack of HAC1 upregulation these folding chaperones are being expressed, potentially to manage excess unfolded protein. Theoretically HAC1 is upregulated in GS115 more substantially than both KAR2 and PDI, which would explain the comparatively low levels of HAC1 induction for clones expressing the HSAopt gene.

5.3.4 Stability of Multi-Copy Clones

Studies using the HSA gene showed that integration into the rDNA locus resulted in reduced levels of instability (with no clones losing the vectors entirely) but did not result in stable strains (Fig 4.23).
Gene transcript levels for UPR signals of GpGRzαHSAopt were lower than GpGRzaHSA (Fig 5.14). It can be hypothesised that reduced stress on the secretory pathway may decrease counter-selection against high copy clones; thus increasing stability. Stability was determined by growth for 48 hours in glycerol before cultures were serially diluted to single colonies on YPD plates (no selection). Genomic DNA was extracted directly from colonies and analysed by qPCR using the primers 188-HSAoptq3 and 189-HSAoptq4 as a comparison to GpαHopt1_100 (Fig 5.15).

Figure 5.15 Instability is Prevalent in Clones with the GAP promoter and HSA optimised Gene
Copy number was calculated via ΔΔCt method using qPCR primers 188-HSAoptq3 and 189-HSAoptq4 to amplify the HSAopt gene. ACT1 was used as the housekeeping gene with a comparison to GpαHopt1_100, a known one copy clone. Dark grey bars represent the original copy number of the clone post-PTVA but before expression. Light grey bars represent individual colonies post expression. A: GpGRzaHSAopt3_2000B (GO3_2000B). B: GpGRzaHSAopt3_2000F (GO3_2000F). C: GpGRzaHSAopt4_2000A (GO4_2000A). D: GpGRzaHSAopt4_2000D (GO4_2000D).

Despite reduced UPR stress (Fig 5.14) GpGRzaHSAopt clones show evidence of genetic instability post-expression (Fig 5.15). This implies that using the HSAopt gene coupled with the GAP promoter does not impact genetic instability. Copy number appears to increase as well as decrease in GpGRzaHSAopt 3_2000B, GpGRzaHSAopt 3_2000F and GpGRzaHSAopt 4_2000A, a trend observed for GpGRzaHSA and GpARzaHSA clones (Fig 4.23). It was also of interest to determine stability of clones containing the AOX1 promoter, as these showed significant upregulation of KAR2 and PDI but not of HAC1. HSAopt copy number was calculated by qPCR using primers 188-HSAoptq3 and 189-HSAoptq4 compared to GpαHopt1_100, a known one copy clone (Fig 5.15).
Instability is Prevalent in Clones with the $AOX1$ promoter and HSA optimised Gene

qPCR was performed on clones isolated from YPD plates post 24 hour expression in methanol containing media. Two clones from each strain isolated on 100 μg mL$^{-1}$ were selected. In all graphs the first (and darker) column represents the original copy number before expression. A: GpARzαHSAopt3_2000E (AO3_2000E). B: GpARzαHSAopt3_2000F (AO3_2000F). C: GpARzαHSAopt(BglII)2_2000D (B2_2000D). D: GpARzαHSAopt(BglII)2_2000E (B2_2000E). E: GpARzαHSAopt(BglII)8_2000D (B8_2000D). F: GpARzαHSAopt(BglII)8_2000F (B8_2000F).

Copy number variation was evident in all clones integrated into the rDNA locus expressed using the $AOX1$ promoter (Fig 5.15). As seen with the GpGRzαHSA, GpARzαHSA and GpGRzαHSAopt no clones investigated lost the HSAopt gene entirely.

5.4 Discussion

The use of an optimised HSA gene for the production of multi-copy clones did not result in increased titre compared to expression using a native HSA gene (Fig 5.12). The GenScript OptimumGene™ algorithm primarily focuses on codon usage bias; however should the DNA 2.0 GeneTPS technology,
which utilises codons that are highly charged during amino acid starvation [237], have been selected instead increased titre may have been observed.

Average secretion levels of GpGRzαHSAopt was 33.0 μg mL⁻¹ compared to GpGRzaHSA which was 46.2 μg ml⁻¹. The same trend was observed for clones expressing via the AOX1 promoter; GpARzaHSAopt average expression was 37.4 μg mL⁻¹ and GpARzaHSA was 44.4 μg mL⁻¹. For GpARzaHSAopt clones reduced titre coincided with reduced HSA transcript when compared to GpARzαHSA clones (Fig 5.13). Furthermore the upregulation of KAR2 and PDI in GpARzαHSAopt strains suggests possible difficulties in protein folding requiring additional chaperones [242]. Codon optimisation potentially alters the bottlenecks within the protein production pathway [236]. Love et al. have indicated that the main bottleneck identified in protein production with P. pastoris is in the secretory pathway [72]. Therefore it is possible to hypothesise that by increasing translational speed one is merely increasing the secretory stress of the cell by adding additional throughput. Furthermore if one adopts the basic view that prolonged induction of UPR results in upregulation of the ER associated degradation pathway (ERAD) then increased secretory stress may explain the reduced titre observed.

Furthermore the use of a codon optimised HSA gene has no impact on the stability of the clones. This may be as a result of maintained secretory stress, which increases the likelihood of counter-selection against high copy clones. It would be of interest to determine the effect of using the HSAopt gene on the ERAD to determine if protein degradation is occurring. Furthermore as prolonged ER stress has been reported to be linked to an increase in cell death in S. cerevisiae, a cell viability assay may determine whether the use of an optimised gene has detrimental effects [243]. In this instance the use of an optimised HSA gene has not resulted in increased specific productivity. However extensive research suggests that codon optimisation can have an impact on protein titre [236, 239, 244-245]. It may be necessary to analyse the advantages of a codon optimised gene on a protein dependent basis.
6 Transcriptomic Analysis of Clonal Variation in Pichia pastoris

6.1 Introduction

6.1.1 Evidence of Clonal Variation in Mammalian Cells and Other Organisms

The phenomenon of clonal variation was first reported for Chinese hamster ovary (CHO) cells in 1977, where it was noted that clones plated onto an agar plate behave differently from one another, with particular reference to colony morphology [246]. This highlighted an underlying heterogeneity of CHO cells that had not previously been identified [247-248]. Clonal variation has also been observed when using CHO cells as a recombinant protein expression platform, evident through differences in titre [249]. In CHO cells much of the variation arises from differences in the integration site or the number of copies of the vector of interest [250]. It has also been discovered that the use of antibiotics, which sometimes have mutagenic properties, increases the level of clonal variation suggesting that differences arise from more than just integration sites and copy number [251]. In addition to clonal variation in mammalian cultures there is evidence that this occurs in other organisms. Since 1981 variation in plant cells generated through transformation procedures to produce transgenic plants has been reported, referred to as somaclonal variation [252]. In plant cells, variation is believed to be predominantly influenced by stress factors, including growth in the presence of antibiotics [253].

6.1.2 Clonal Variation in P. pastoris

The presence of clonal variation in P. pastoris is widely acknowledged in both academia and industry [211, 254-255]. Clonal variation is such an integral part of working with P. pastoris that the Pichia Expression Kit manual provided by Invitrogen Corporation discusses methods to select for the highest secretor, recommending to screen between 6-10 recombinants per phenotype [120]. In industry to select the highest producers typically up to a thousand recombinants are screened [256]. However, this is extremely time-consuming and a major drawback for commercial application; however, at present there are no procedures in place to reduce the effort to find the best secretors. Few investigations have been carried out to examine the origins of clonal variation, with just a single paper describing this variation using amplified-fragment length polymorphism (AFLP) [211]. Clare et al. suggest that the integration event results in the variation in expression levels and in particular the presence of multi-copy clones [254], as seen with mammalian cultures. Yet our experiments have shown when comparing titre levels of single copy clones the presence of clonal variation is still prominent, which suggests that additional factors are at work.
The objective for expression platforms is to produce a strain that expresses the highest titre of recombinant protein possible. High secretors selected by extensive screening reduce the need to subsequently generate multi-copy clones. Furthermore a good secretor can reduce the need for optimisation in regards to media and growth conditions [212]. Hence, due to the highly variable productivity of different recombinant clones of P. pastoris extensive screening of clonal variants is an intrinsic part of any expression programme.

6.1.3 Studying Clonal Variation

For recombinant expression studies clonal variation is primarily reflected in the protein titre obtained with individual clones. This has been examined using various techniques including SDS-PAGE gels, Western blots and ELISAs. However, other underlying factors such as ER stress, indicated by the induction of the unfolded protein response (UPR), which affects the production of the protein of interest, are also useful markers which can be determined by RT-qPCR. The prevalence of clonal variation highlighted in chapters three and four, needs to be investigated in detail and a clearer understanding of the phenomenon could enable the development of methods to simplify the identification of a “best” secretor.

6.1.4 Aims for the study of Clonal Variation

- To establish the extent of clonal variation in P. pastoris
- To investigate the importance of the integration site of the expression vector in relation to clonal variation
- To evaluate the impact of clonal variation on the unfolded protein response and other stress factors
- To identify key indicators of high or low secretors by microarray analysis

6.2 Clonal Variation in HSA expressing strains

6.2.1 Clonal Variation in Fed-Batch Fermentation

From the initial work on multi-copy clones there was evidence for variation in the secretion levels between the two one copy clones that were used for expression in batch-fed fermentation (Fig 6.1).
Figure 6.1 HSA titre levels of single copy clones GpαH1_100 and GpαH11_100.
HSA titre levels were calculated using the Albumin Blue Fluorescence assay (Active Motif). Samples were taken at 0 hours (pre-induction), 24 hours and 48 hours post-induction. Error bars represent technical repeats of the albumin assay. Statistically significant values of p<0.05 are indicated by a *.

GpαH1_100 secreted more HSA than GpαH11_100 when grown under identical conditions, measured at both 24 hours and 48 hours post-induction. After 48 hours GpαH1_100 had secreted 150% more HSA than GpαH11_100. For any industrial process it is clear that GpαH1_100 would have been a better strain for development based on initial secretion level. In order to ascertain whether the differences relate to the level of transcription of the HSA gene, HSA transcript levels were investigated using RT-qPCR (Fig 6.2).
Figure 6.2 HSA transcript levels for single copy clones GpαH1_100 and GpαH11_100. HSA gene expression was calculated via ΔΔCt method using qPCR primers 188-HSAoptq3 and 189-HSAoptq4 to amplify the HSAopt gene or 070-HSAq1 and 071-HSAq2 to amplify the HSA gene. ACT1 was used as the house keeping gene with comparison to GpαH1_100. Transcript levels were determined at 0 hours, 24 hours and 48 hours of fed batch culture. Statistically significant values of p<0.05 (compared to GS115) are indicated by a *.

Surprisingly, although GpαH11_100 produced less than half as much HSA protein than GpαH1_100 (Fig 6.1), HSA transcript levels of GpαH11_100 were actually significantly higher at both 24 hours and 48 hours (two tailed t(4) p<0.0001). This not only suggested that the differences observed were not directly due to transcription rates but that there was considerable variation in the rate of transcription from supposedly identical constructs. The differences in level of secretion were, therefore, post-transcriptional and could be related to secretion stress. The UPR can be taken as an indicator of stress within the secretory pathway and prolonged induction will result in the upregulation of ERAD, which could account for reduced titre through protein degradation [72, 164].

In order to investigate levels of stress within the strains GpαH1_100 and GpαH11_100 were analysed for the induction of the UPR, looking for key indicators such as HAC1, KAR2 and PDI (Fig 6.3).
Overall GpαH1_100 induced the UPR to a lesser extent than GpαH11_100 (Fig 6.3). Specifically the upregulation of PDI in GpαH11_100 after 48 hours showed a two-fold increase compared to GS115. Differences in UPR expression levels once again highlight differences caused by clonal variation.

### 6.2.2 Range of Variation in a Single Transformation

To better understand the range of variation in HSA secretion that occurs after transformation pα-HSA was transformed into GS115 and plated onto 100 μg mL⁻¹ Zeocin plates. Twenty-three colonies were selected and copy number was analysed by qPCR (Fig 6.4).
Copy number analysis of 23 colonies selected from a single transformation of pα-HSA determine that all are Single Copy Clones.

Copy number was calculated via the $\Delta\Delta C_t$ method using qPCR primers (188-HSAoptq3 and 189-HSAoptq4) to amplifying the HSAopt gene with comparison to ACT1 as the housekeeping gene. GpαH1_100 was used as the reference strain.

All of the clones selected were confirmed to contain a single copy of the gene of interest (Fig 6.4).

Expression was carried out in 50 mL centrifuge tubes in BMGY for 24 hours, before being normalised to a standard $OD_{600}$ and induced with methanol containing BMMY to initiate expression of HSA. After 24 hours expression samples were centrifuged and the spent broth analysed using the Albumin Blue Fluorescence Assay (Active Motif; Fig 6.5).
All the clones analysed contained a single copy of the HSA gene (Fig 6.4), yet titre ranged from less than 5 mg L\(^{-1}\), GpaHCV14, to 22.5 mg L\(^{-1}\), clone GpaHCV2 (Fig 6.5); a 4-fold difference in expression level.

This demonstrates that clonal variation is not solely attributable to copy number variation [146]. It has been reported that non-homologous recombination can arise and this could have a negative effect on titre [215]. In order to determine whether the variation in titre observed was a result of differences in plasmid integration into the chromosome, DNA sequencing was used to analyse the region around the sites of integration. Primers 142-AOX Integr F Pmel and 143-AOX Integr R Pmel (Appendix 9.2) were used to sequence upstream and downstream of the integration site (Fig 6.6; Appendix 9.5).

All strains tested showed 100% sequence identity to each other, regardless of the corresponding titre levels, suggesting that in these cases variation was not due to changes at the site of insertion of the vector (Fig 6.6).

6.3 Microarray Analysis

6.3.1 Selecting strains for Microarray Analysis

As copy number or variable integration did not explain the clonal variation in this dataset, further experimentation was done to determine the underlying causes. Microarray analysis was used to highlight transcripts that were either up- or downregulated in correlation with titre levels. Upregulation in association with improved secretion could provide an opportunity to create a
reporter system whereby the relevant promoter is linked to expression of e.g. a fluorescence protein, or alternatively a negative counter-selection might be employed.

For microarray analysis clones were grown in 250 mL baffled flasks and nine strains were selected for further analysis, chosen based on titre levels measured using the Albumin Blue Fluorescence Assay kit (Active Motif). Three high secreting strains were selected; GpαHCV2, GpαHCV18 and GpαHCV23, three mid level secreting strains; GpαHCV8, GpαHCV15 and GpαHCV16 and three low secreting strains GpαHCV5, GpαHCV7 and GpαHCV14. For simplification the strains are referred to as CV# followed by (H) for high, (M) for mid and (L) for low secretors, for example CV2(H) (Fig 6.7).

![Figure 6.7 HSA titre levels of the nine Strains selected for microarray analysis, grouped into high, medium and low levels.](image)

Titre levels of single copy GpαHSA strains expressed for 24 hours in methanol containing media. Samples were grown in 250 mL baffled flasks in triplicate; error bars displayed indicate 95% confidence interval. Green; high secretors, yellow; mid secretors and red; low secretors.

Titre levels ranged from 8.87 mg L⁻¹ with strain CV7(L) to 16.6 mg L⁻¹ CV18(H). Using 95% confidence intervals indicated that grouping the strains based on titre results in three statistically independent sets.

6.3.2 Genotyping Microarray Strains

As a preliminary study, the clonal variation (CV) strains were genotyped in order to investigate whether there were characteristic differences within the genome. Variation is likely to reflect mutations or more major genome rearrangements within the strains. Viader-Salvado and colleagues used a simplified version of amplified-fragment length polymorphism (AFLP) to reveal genetic
differences between transformants [211]. The theory behind the technique is that by using restriction enzymes and adapters it is possible to create and separate out characteristic fragments by gel electrophoresis. Variations in the genome will affect fragment size and hence band pattern, but this depends on where the restriction enzyme cuts. Due to the nature of AFLP there are limitations due to the use of restriction enzymes. Differences may be missed if the restriction enzyme does not cut within that particular fragment. Therefore a similar method was developed; Random-Amplification-Polymorphic-DNA (RAPD) PCR which has been used to study variation within populations [257-259]. Using a random decamer primer (Appendix 9.2) genomic DNA is amplified by PCR and according to where the primers anneal within the genome a fragment pattern is produced, which can be visualised on a high percentage agarose gel.

The nine microarray strains and the wild-type GS115 were analysed using RAPD-PCR, and similar patterns to those described by Viader-Salvado et al. were seen. This method was used to establish whether polymorphisms in the strains selected for microarray analysis were evident (Fig 6.8).

![Figure 6.8 RAPD-PCR of strains selected for microarray analysis.](image)

Nine GpaαHSA strains and the wild-type GS115 were analysed by RAPD-PCR to look for evidence of gross polymorphisms. 2.5% (w/v) agarose gels were used to get complete separation of the small fragments. Ladder; GeneRuler 100 bp Plus DNA Ladder (Fermentas). 1: CV2(H); 2: CV18(H); 3: CV23(H); 4: CV8(M); 5: CV15(M); 6: CV16(M); 7: CV5(L); 8: CV7(L); 9: CV14(L); 10: GS115.

RAPD-PCR revealed differences both between the selected clonal variants and when compared to GS115 (Fig 6.8). The fragment at 750 bp appeared to be of variable abundance and was clearly missing in CV15(M), CV14(L) and GS115. Additionally, the abundance of the fragment at approx 910
bp appeared quite variable. Although RAPD-PCR did not reveal major changes, the differences detected suggest some clonal variation which is neither copy number nor integration site dependent.

### 6.3.3 Transcript Levels Determined by RT-qPCR

As the strains for microarray analysis were selected based solely on secreted titre levels, it was of interest to look for any systematic variation in transcript levels using RT-qPCR of HSA and UPR indicators: HAC1, KAR2 and PDI (Fig 6.9).

![Figure 6.9 HSA transcript levels after 24 hour expression.](image)

Samples were collected after 24 hours induction with methanol in 250 mL baffled flasks. HSA transcript was calculated via the ΔΔCt method using qPCR primers 188-HSAoptq3 and 189-HSAoptq4 as a comparison to ACT1 as the housekeeping gene. Fold change was calculated as a comparison to GpaHopt1_100. Green represents high secretors; yellow mid-level secretors and red low secretors.

Interestingly HSA transcript levels do not indicate a systematic correlation to titre levels across the set of strains analysed. However, CV2(H), CV18(H) and CV23(H) all show a higher transcript level than most of the other clones, consistent with their higher levels of protein production. In contrast, CV7(L) showed comparable transcript levels to the high expressing strains, even though it secreted half the quantity of protein as CV18(H). As the low HSA titre in CV7(L) does not appear to be associated with a reduction in transcription, this suggests that there is a post-transcriptional bottleneck possibly in the secretory pathway, which may result in misfolded protein that is eventually degraded; hence the low level of secreted protein. Induction of the UPR would indicate a stress in the secretory pathway; thus HAC1, KAR2 and PDI were analysed by RT-qPCR to determine whether upregulation was occurring (Fig 6.10).
While HSA transcript levels clearly indicated transcriptional or translational blockages, the results from analysing gene transcript levels of UPR signals do not reveal such a clear story. CV8(M), which showed particularly low levels of HSA transcript, but was a mid-level secretor, also demonstrates higher upregulation of all three UPR genes. An initial induction of UPR can be interpreted as positive for the production of correctly folded proteins. Both KAR2 and PDI act as folding chaperones; therefore an upregulation could increase the flux of correctly folded proteins through the secretory pathway [141, 260-261]. UPR may only be considered a negative trait in protein expression when prolonged upregulation results in the activation of the ERAD [71, 164]. CV7(L), which showed high HSA transcript levels but low levels of protein, did not display upregulated levels of the UPR genes HAC1 and PDI (Fig 6.10). However expression of KAR2 was higher than in other low secreting strains or in the high secreting strains.

6.3.4 Microarray Analysis

Preliminary studies suggest that the groups of high, medium and low secretors do not behave as single cohorts and that, while differences are evident, levels of transcription and UPR do not provide a complete picture. Therefore, to gain a genome-wide view, microarray analysis was carried out on all nine strains, in addition to X33, a wild-type strain. Samples were loaded on five slides of 8x60K, with nine probes per gene. As this was the second generation of P. pastoris microarrays carried out in our lab, the probes were optimised based on previous experiments (data not shown). Each of the
nine strains and X33, were grown in BMGY for 24 hours, before being induced in methanol containing medium BMMY. These were left for a further 24 hours before 1 mL of culture was collected and stored in RNALater (Applied Biosystems). The RiboPure Yeast Kit (Applied Biosystems) was used to extract the RNA and concentration was checked using the Experion RNA StdSens analysis kit (Bio-Rad). In addition to quantifying the amount of RNA a numerical RNA quality indicator (RQI) is derived based on the levels of 18S and 28S RNA. For this experiment a minimum of 330 ng μL⁻¹ was used with a range up to 1955 ng μL⁻¹, while the minimum RQI used was 8.7, with 10 being the best score. The samples were sent to the Bacterial Microarray Group at St. George’s Hospital for analysis. The samples were run on an Agilent High Resolution Microarray scanner having previously been labelled using the One-Colour Quick Amp Labelling Kit (Agilent Technologies UK Ltd).

Data sets were formed as comparisons to the wild-type X33, which with the exception of having an intact histidine pathway made functional through re-complementation, should be identical to GS115 [262]. Three biological replicates (i.e. independent expression) of each of the nine clones were analysed with the mean values being used for comparisons, with a false discovery rate (fdr) threshold set at 0.05 or less. From this data, different behaviours of the clones both individually compared to the wild-type and comparisons between the groups were revealed.

Initially volcano plots were used to portray the levels of significant variation between each of the clonal variation strains in contrast to the wild-type strain (Appendix 9.6). From these plots it was apparent that there were aspects that showed significant variation in all of the different strains in comparison to wild-type with an fdr ≤ 0.05.

6.3.4.1 Manual Annotation of the microarray data to reveal differentially expressed genes

Using the Bioconductor package in R, gene transcripts were quantified for each sample and compared to those for X33. Differences in expression with an fdr ≤0.05 were used to identify differentially expressed genes. Those genes that showed significant variation compared to the wild-type were assigned a category based on their function. The categories used were degradation, stress, secretion, folding, ribosome, mitochondria, DNA replication/repair, peroxisome, biosynthesis, transcription, translation, spliceosome, recycling of material within the cell, cell wall, chromosome/cell cycle, nuclear import/export or nuclear pore, sporulation and cytoskeleton/actin.

For each strain the percentage of genes that were up- or downregulated in each category, compared to the wild-type, were calculated with reference to the total number of genes assigned to each
category. An example using CV2(H) can be seen in figure 6.11, with all other comparisons in Appendix 9.7.

Several trends can be observed through these manual annotations especially when comparing the different groups together (Fig 6.11, Appendix 9.7). All three of the high secreting strains showed a higher percentage of peroxisome-related genes upregulated (approximately 70%), compared to the mid and low level secreting strains, which showed an upregulation of approximately 50%. When *P. pastoris* grows in the presence of methanol, the abundance of peroxisomes within the cell increases in order to process the methanol [77, 263-264]. If the number of genes expressed relating to peroxisome biogenesis is increased then it suggests that these strains may be metabolising more methanol than the wild-type strain, consistent with the metabolic demand of producing a
heterologous protein in high yields. As the mid and low secretors are producing less peroxisomes this could make them energy limited; thus impacting the protein production capabilities. (Appendix 9.7).

Further trends indicate that the high secreting strains showed more genes upregulated in the chromosome or cell cycle pathways compared to the mid and low secretors. This could potentially be related to the growth rate of the strains; therefore growth curves were observed in order to determine whether they were significantly different amongst the clonal variant strains (Fig 6.12).

Figure 6.12 Growth Curves of Clonal Variation Strains Grown in Glycerol or Methanol Containing Rich Media
Growth curves were carried out in baffled flask with samples taken every hour. All nine clonal variation strains were grown, with the addition of X33 as a control. A: Growth curve in glycerol containing media for 24 hours. B: Growth curve from cultures originally grown in glycerol containing media and induced with methanol after 24 hours.
According to the growth curves in both glycerol and methanol containing media there is minimal variation between the strains. In glycerol CV2(H) appears to grow faster than the other strains reaching an OD$_{600}$ of 18.7 at 10 hours, while the other strains showed a maximum of OD$_{600}$ 13.8 after the same time. All CV strains reached a similar maximum growth yield, with a plateauing of growth occurring at approximately 11 hours in glycerol. Growth in methanol will be limited due to competition with protein production for resources. One would expect a strain producing fewer proteins to grow more significantly in methanol. However, variations between the high, mid and low secretors did not appear significant upon induction with methanol.

According to the data from the growth curves there are not enough significant differences that allow differentiation between the three clonal variation groups when grown on methanol. As growth curves do not indicate the size or viability of cultures, fluorescent activated cell sorting (FACS) analysis was used to analyse these aspects. Invitrogen’s FungaLight® Yeast Viability kit compares the number of live cells versus dead cells. Cells were grown for 24 hours in the glycerol containing medium (BMGY), before being induced with methanol containing medium BMMY for 24 hours. Samples of 1 mL were collected from both media types at 24 hours and washed in TBS. Cells were analysed by flow cytometry with 1 μL of SYTO9® and 1 μL of propidium iodide (PI) added for staining according to the manufacturer’s instructions. SYTO9® is a green-fluorescent protein that will bind to nucleic acids, while PI will only penetrate yeast cells with damaged membranes reducing the green fluorescence and indicating damaged or dead cells [265]. Cell size was determined by looking at the forward light scattered reflected in histograms for both cells grown in the presence of glycerol (Fig 6.13) and grown in the presence of methanol (Fig 6.14) [266-268].
Figure 6.13 Histogram to Show Forward Scatter of Clonal Variation Strains Grown in Glycerol containing Media to Determine Average Cell Size

The forward scatter of light determined by flow cytometry on the nine clonal variation strains and X33 (wild-type) when grown in BMGY. The mean of scatter gives an indication of average cell size.
Figure 6.14 Histogram to Show Forward Scatter of Clonal Variation Strains Grown in Methanol containing Media to Determine Average Cell Size

The forward scatter of light determined by flow cytometry on the nine clonal variation strains and X33 (wild-type) when grown in BMMY. The mean of scatter gives an indication of average cell size.
In glycerol cultures the mean forward scatter, with the exception of CV2(H), was similar amongst all of the strains indicating that there is no size variation amongst the strains. As CV2(H) grows faster on glycerol than the other strains it suggests that this strain has a high division rate. Furthermore X33 also indicates a small cell size, which suggests that the remaining CV strains have slower division rates. However post-methanol induction CV5(L), CV7(L) and CV14(L) represented the three highest means in terms of cell size. As the growth rates were comparable to the mid and high secreting strains this implies that an increase in size is accountable for the density of the culture and not an increase in cell numbers (Fig 6.12). This coincides with the high proportion (approximately 40%) of chromosome genes, downregulated in CV7(L) and CV14(L) compared to X33 (Appendix 9.7.7-9.7.8).

Growth curves measured by spectrophotometry may be distorted due to the accumulation of dead cells. Therefore viability staining using FACS would determine the proportion of cells that were alive. Growth in both a glycerol containing medium (BMGY) and a methanol containing medium (BMMY) was carried out for 24 hours to determine viability after each growth phase (Fig 6.15 and 6.16).
Figure 6.15 FACS analysis of Clonal Variation Strains and X33 Grown in Glycerol Containing Media

FACS analysis of clonal variation strains grown for 24 hours in glycerol containing media. FL1_H: SYTO9®, FL2-A: propidium iodide. BG: Background noise. Quadrant displays proportion of live/dead cells and background noise.
FACS analysis of clonal variation strains grown for 24 hours in glycerol containing media.


Figure 6.16 FACS analysis of Clonal Variation Strains and X33 Grown in Methanol Containing Media
All clonal variants of *P. pastoris* exhibit varying degrees of cell death when grown in glycerol as demonstrated by the double staining of SYTO9® and PI, but none exceeded 14% of the population as observed in CV2(H). The least amount of cell death was observed in CV7(L) with 3.4%, while X33 the wild-type strain showed a cell death percentage of 10.7%. However, the proportion of dead cells dramatically increased post-induction with methanol. Although X33 shows a comparable percentage of cell death to that seen when cultured in BMGY, for all of the clonal variation strains the ratio of live to dead cells is considerably reduced with CV23(H) exhibiting an almost 1:1 ratio. A common trend is that the three high secretors indicate a higher level of cell death than the mid or low secretors. CV18(H) shows the least amount of cell death between the three at 31.5%. The only other strain that shows cell death at a comparative level is CV7(L), which has been highlighted as a potentially anomalous strain, with 32.9% exhibiting cell death. Excluding CV7(L) the other strains show a maximum cell death level of 25.5% (CV16(M)), and a low of 18.1% (CV14(L)). This suggests that through the production of recombinant protein increased stress may be detrimental to cell viability. It would have been of interest to observe titre levels in relation to viability over time, as increased cell death should reduce productivity.

Further trends observed through analysing the microarray data suggests that with the exception of CV7(L), which according to the transcript and expression data may be anomalous, the high secretors showed the highest percentage of genes related to stress and secretion upregulated (Fig 6.11, Appendix 9.7). A higher number of upregulated secretion genes might have been expected as better secretion or an amplified secretion apparatus would result in more protein being released into the supernatant. The number of genes involved in secretion which were upregulated in low sectors was half of that seen in the other strains. With regards to transcription and spliceosome related genes, there were twice the number of genes downregulated in these three strains compared to all of the other strains. It would appear for low secretors that through the induction of UPR, or some other feedback control loop, transcription is being reduced in order to preserve resources instead of using these for a function (that of producing proteins) which will only be degraded due to an inability to deal with the strain on the secretory machinery.

Reduced titre can be explained if proteins are being degraded, possibly through ERAD, as there will be fewer proteins to secrete. CV16(M) is a mid-level secretor that shows higher levels of stress on the secretory pathway than that of the other mid-level secreting strains, CV8(M) and CV15(M). This highlights the fact that the differences seen through clonal variation are multi-variate, and that the high, mid and low secretion groupings probably do not correlate with common underlying traits.
Clones CV7(L), CV14(L) and CV16(M) appear to follow similar trends to one another with regards to degradation; almost twice the number of genes were upregulated in degradation compared to the high secreting strains. This suggests that pathways such ERAD or autophagy are upregulated. By looking at examples of degradation genes that are upregulated in these strains it is possible to highlight the pathways that seem to be more active. These include UBX7 which is an ubiquitin-dependent protein which binds to Cdc48p, which is known to be involved in the ERAD associated pathways [269-270]. CKS1 which is involved in recruiting proteins to the proteasome for degradation [271-272] is involved in both the cell cycle and cell death. JDI1 is also a gene upregulated which is involved in ER-associated degradation of misfolded proteins [273-274]. This implies that many different aspects of degradation are upregulated including the ERAD, proteasome and cell death pathways.

The secretion of high levels of a heterologous protein does not necessarily result in an increase in stress on the system; however, an increase in the number of stress genes expressed in high secretors does suggest upregulation of control systems such as UPR [261]. Considering that many of these stress genes are in fact involved in repair or recovery from cellular stress, oxidative stress or osmotic stress, deliberately increased expression of some of these genes could produce a strain that is predisposed to deal with the imposed stress [275-276].

6.3.4.2 Pathway Analysis

Initial analysis has indicated the there is considerable variation in gene expression in the different clonal variants. More specific detail of individual pathways was generated by analysing the microarray data using the program KOBAS, which maps different genes to KEGG pathways [201-203]. KOBAS creates a comprehensive list with significant p-values based on the number of genes which vary within the pathway, as well as a corrected p-value which takes into consideration the total number of genes in the analysed dataset.

Having generated the lists it was possible to determine which pathways were significantly up or downregulated within the three comparison groups; for instance whether there were common pathways that were always upregulated in high secretors. When using a corrected p-value of 0.2 or less no common pathways were observed within the individual high, mid and low secretory groups. Therefore, although a less stringent comparison, it was decided to compare pathways using a p-value of less than or equal to 0.1 to increase the number of pathways analysed. In KOBAS the p-value differs from the corrected p-value as it does not take into consideration the total number of genes that are up- or downregulated within a gene set. When comparing differentially expressed pathways in the high secreting strains it becomes immediately apparent that using this approach increases the
number of pathways significantly up- or downregulated. To visualise significantly varied pathways the online tool Venny was used to create Venn diagrams, and only pathways that had a p-value of 0.1 or less were included (Fig 6.17-6.19) [277].

![Venn diagrams](image)

**Figure 6.17 A comparison of Pathways Up- and Downregulated in High Secreting Strains at p≤0.1**

Venn diagrams to represent the pathways that are significantly upregulated or downregulated for each of the high secreting strains in comparison to X33. A p-value of ≤0.1 was used. A) Pathways that were significantly upregulated. B) Pathways that were significantly downregulated.

Immediately it is apparent that two pathways were upregulated in all of the high secreting strains: oxidative phosphorylation and ribosome biogenesis in eukaryotes pathways. These correspond with the ability to release energy and the production of protein. CV18(H) and CV23(H) have upregulation in three other pathways in common which are peroxisomes, RNA degradation and the spliceosome. There were no pathways that were commonly downregulated in all of the high secreting strains, but there were three pathways that were downregulated in both CV18(H) and CV23(H): steroid biosynthesis, protein processing in the ER and riboflavin metabolism. Additionally there were five pathways downregulated in CV2(H) and CV18(H): lysine biosynthesis, MAPK signalling pathway, alanine, aspartate and glutamate metabolism, phenylalanine, tyrosine and tryptophan biosynthesis and finally starch and sugar metabolism. The downregulation of biosynthesis pathways indicates that the strains have reduced capacity to synthesise these amino acids de novo, which would normally be required for growth and production of protein. Given that the strains are producing high levels of protein this probably reflects a slowdown in growth rate and degradation of endogenous protein and redirection of resource. This could also explain why the RNA degradation and spliceosome pathways are upregulated in order to manage a transient excess of mRNA that is not being used to make cellular protein. This is also consistent with the downregulation of protein
processing pathways, which may be expected after 24 hours growth, and evidence of increased cell death as seen in figure 6.16.

Figure 6.18 A comparison of Pathways Upregulated and Downregulated in Mid-Level Secreting Strains at p≤0.1
Venn diagrams to represent the pathways that are significantly upregulated or downregulated for each of the mid-level secreting strains in comparison to X33. A p-value of ≤0.1 was used. A) Pathways that were significantly upregulated. B) Pathways that were significantly downregulated.

In the mid-level secreting strains two pathways were upregulated in all three strains; ribosome biogenesis (as with high secreting strains) and porphyrin metabolism. While porphyrin metabolism may not be an obvious pathway to be upregulated in P. pastoris, this pathway is linked to steroid biosynthesis and glycine, serine and threonine metabolism. The pathways that were commonly upregulated in CV8(M) and CV15(M) were biosynthesis of secondary metabolites, peroxisome and arginine and proline metabolism. The common pathway between CV8(M) and CV16(M) was purine metabolism, and the common pathways between CV15(M) and CV16(M) were ribosome, oxidative phosphorylation and RNA transport. It appears that the majority of the pathways upregulated relate to metabolism or biosynthesis of metabolites, which would be required for growth or production of proteins.

The single pathway downregulated for all of the mid-level secreting strains was the ubiquitin mediated proteolysis pathway, which encompasses the majority of the ERAD pathway. Only CV8(M) and CV16(M) had other pathways downregulated in common which were MAPK signalling pathway and mismatch repair.
Figure 6.19 A comparison of Pathways Upregulated and Downregulated in Low Secreting Strains at \( p \leq 0.1 \)

Venn diagrams to represent the pathways that are significantly upregulated or downregulated for each of the low secreting strains. A \( p \)-value of \( \leq 0.1 \) was used. A) Pathways that were significantly upregulated. B) Pathways that were significantly downregulated.

Finally for the low secreting strains there was one pathway that was upregulated in all three strains which was the peroxisome pathway, which is to be expected due to the growth on methanol. Three pathways were upregulated in both CV5(L) and CV14(L); purine metabolism, RNA polymerase and ribosome biogenesis. The common pathways upregulated between CV7(L) and CV14(L) were ribosome, biosynthesis of secondary metabolites, cysteine and methionine metabolism and ubiquinone and other terpenoid-quinone biosynthesis pathways. Once again it is apparent that most of the pathways upregulated relate to metabolism or biosynthesis. Nonetheless differing metabolism pathways are upregulated amongst the different groups indicating different requirements of each strain.

For pathways downregulated in all three low secreting strains there was only one common to all which, as with the mid secreting strains, was the ubiquitin mediated proteolysis. Other pathways downregulated in both CV5(L) and CV14(L) were protein processing in the ER as well as glutathione metabolism, the former being particularly important as this implies that protein production is lower than observed in the wild-type strain. This highlights potential blockages within the pathways that could explain the low titre. CV7(L) and CV14(L) had the largest number (8) of common pathways downregulated; cell cycle, MAPK signalling pathway, regulation of autophagy, nucleotide excision repair, mismatch repair, endocytosis, natural killer cell mediated cytotoxicity and SNARE interactions in vesicular transport. *P. pastoris* undergoes a form of autophagy known as pexophagy whereby peroxisomes are degraded in order to release resources for protein production or growth [80, 278]
As autophagy is downregulated in both CV7(L) and CV14(L) compared to X33 it implies that the regulation of resources may not be as efficient in these strains. This may in part be due to a lower growth rate on methanol, as the demand for cellular reorganisation will be reduced if less peroxisomes are generated. Furthermore downregulation of the cell cycle, nucleotide excision repair and mismatch repair suggests a slow in growth rate, potentially due to a lack of resources.

While it would have been interesting to do a comparison of pathways that were up or downregulated in the different groups of high, mid and low secretors it is apparent that within the groups there is no uniformity of pathways that are significantly changed. From these results it is apparent that while different and clear cut titre levels can be determined (Fig 6.7), the causes of clonal variation are variable and as a result it becomes necessary to look at the strain individually instead of grouping them together.

6.3.4.3 KEGG Pathway Mapper for the Visualisation of Significantly Changed Pathways

Once the KOBAS program had determined pathways that were significantly up or downregulated for each clonal variant compared to wild-type it was possible to use the KEGG Mapper Search & Colour pathway tool to create a visual display [202, 204-205]. An fdr threshold of $p \leq 0.05$ was established for gene expression that was significantly different from X33. Genes for which expression was upregulated in a clonal variant were assigned the colour blue and for gene expression that was downregulated, the colour red (Fig 6.20; Appendix 9.8).

![Figure 6.20 Protein Processing in the Endoplasmic Reticulum Pathway of CV2(H)](image)

Pathway of protein processing in the endoplasmic reticulum of strain CV2(H) compared to wild-type. Genes in blue are upregulated, while genes in red are downregulated in the clonal variant strain compared to X33. Green boxes indicate organism-specific pathways.
Once relevant pathways had been mapped for each of the clonal variants it was possible to compare individual genes of pathways which may be significant. PDI, which is an indicator of UPR stress, was downregulated in the three high secretors and the three low secretors compared to the wild-type strain. Conversely it was upregulated in the three mid-level secretors. This could suggest that in the high level strains the level of PDI is lower as the feedback loop that is the UPR has recovered to a homeostatic equilibrium. It is also possible to hypothesise that in the low secreting strains no induction of UPR has occurred; thus explaining the low PDI upregulation. The number of disulfide bonds within a protein may affect the requirements for PDI upregulation. HSA has 17 disulfide bonds and it would be interesting to observe the expression levels of PDI for a protein that either had more or less disulfide bonds [187, 280].

The UPR in mammalian cells functions through three separate pathways; IRE1, PERK and ATF6; however in S. cerevisiae only the IRE1 pathway is present [70-71, 153, 155, 241]. According to the results from the KEGG pathway mapper, all strains with the exception of CV23(H) show downregulation in a PERK – like transcript (Fig 6.23 and Appendix 9.8). The downregulation of this particular gene suggests that there may be a homolog in P. pastoris of the PERK-like transmembrane protein. While S. cerevisiae does not contain a PERK homolog, Caenorhabditis elegans contain both an IRE1 pathway and a PERK pathway [281]. Thus, P. pastoris may resemble C. elegans in this regard, differing from S. cerevisiae in terms of the UPR, a theory which is supported by the difference in the size of the intron of the HAC1 mRNA [162]. Bioinformatic analysis (blastp) revealed that the sequence similarity between P. pastoris and S. cerevisiae PERK-transmembrane protein is 42%. The downregulated PERK-like gene shows 41% sequence identity to the S. cerevisiae gene GCN2, which is also a protein kinase that phosphorylates the alpha-translation initiation factor eIF2 in response to starvation [282]. As none of the strains show significant growth after 24 hours on methanol (growth appears to stop after 8 hours; Fig 6.12), starvation could be occurring due to resources being utilised for protein production. In mammalian cells both GCN2 and PERK contribute to eIF2α phosphorylation which occurs after the activation of the UPR [283].

It is also of interest to note that IRE1 transcript levels were downregulated in the majority of the strains with only CV15(M), CV16(M) and CV23(H) showing no significant variation from wild-type levels (Fig 6.23 and Appendix 9.8). Furthermore none of the strains showed an upregulation of IRE1. Unfortunately HAC1 does not appear on the KEGG pathways, but it was possible to monitor the behaviour when looking at the raw data. CV16(M) strain showed a downregulation of HAC1 compared to wild-type, CV14(L), CV2(H) and CV18(H) showed no difference, while HAC1 in CV5(L), CV7(L), CV8(M), CV15(M) and CV23(H) was upregulated compared to wild-type. A downregulation of
HAC1 nonetheless, does not suggest that there is no UPR being upregulated as X33 will show a baseline level of UPR; furthermore a downregulation could suggest that the strain has returned to homeostasis. PDI expression, which should increase in the presence of misfolded protein to aid folding, shows an upregulation in strains CV5(L), CV7(L), CV8(M), CV15(M) and CV23(H), the same strains that show an increase in HAC1 transcript. All other strains were not significantly different from wild-type. However, no strains showed an increase in KAR2 expression compared to wild-type, with CV7(L), CV8(M) and CV16(M) showing no difference and all the other strains exhibiting a decrease in KAR2 expression. Release of unfolded protein from KAR2 is often the trigger for activation of ERAD [284-285]. Potentially the downregulation of KAR2 is as a result of the length of time that the cultures were expressed for.

Other pathways that were analysed in detail included ERAD, taking into consideration genes such as HRD1, which is the link between ERAD and the unfolded protein response [164]. Two strains showed upregulation of HRD1, CV5(L) and CV23(H), whereas CV7(L) & CV14(L) had reduced expression of HRD1. It is of interest that CV5(L) contrasts with the other low expressing strains, another clear indicator that grouping the strains together based on expression levels probably has no mechanistic correlation. In order to investigate this further the ubiquitin mediated proteolysis pathway was investigated to reveal whether the ERAD was upregulated.

In line with the downregulation of HRD1, a downregulation of approximately half of the genes involved in the proteolysis pathway was observed in CV7(L) (Fig 6.21).
Figure 6.21 Ubiquitin Mediated Proteolysis Pathway in CV7(L)
Pathway of ubiquitin mediated proteolysis in CV7(L) compared to wild-type. Genes in blue are upregulated, while genes in red are downregulated in the clonal variant strain compared to wild-type. Green boxes indicate organism-specific pathways.

CV14(L), which also exhibited downregulation of HRD1, showed downregulation of a large proportion of the pathway, with the exception of 2 genes within the E2 complex (Fig 6.22).
Figure 6.22 Ubiquitin Mediated Proteolysis Pathway in CV14(L)
Pathway of ubiquitin mediated proteolysis in CV14(L) compared to wild-type. Genes in blue are upregulated, while genes in red are downregulated in the clonal variant strain compared to wild-type. Green boxes indicate organism-specific pathways.

While CV5(L) showed upregulation of HRD1 the majority of the ERAD pathway was not upregulated, with a higher proportion of expressed genes than with CV7(L) and CV14(L), showing no variation compared to wild-type (Fig 6.23).
Figure 6.23 Ubiquitin Mediated Proteolysis Pathway in CV5(L)
Pathway of ubiquitin mediated proteolysis in CV5(L) compared to wild-type. Genes in blue are upregulated, while genes in red are downregulated in the clonal variant strain compared to wild-type. Green boxes indicate organism-specific pathways.

CV23(H), which like CV5(L) was upregulated in HRD1 expression, shows upregulation in more steps in the pathway. This was supported by the KOBAS pathway analysis which indicates this pathway was significantly upregulated (Fig 6.24).
When considering the different groups of secretors, the low secretors were all downregulated in a large part of the ERAD pathway, while the high secretors showed upregulation. While CV18(H) was significantly downregulated in the ubiquitin mediated proteolysis pathway (according to the KOBAS analysis) approximately half of the in E1 and E3 subunit genes were upregulated (Appendix 9.9). As ERAD works as a control mechanism to prevent excess stress in the cells, thus reducing the likelihood of cell death, then upregulation of ERAD could potentially be a good indicator of a high expressing strain [261, 286-287]. Skp1, an adaptor protein involved in the multi subunit type E3 complex, is upregulated in all of the high expressing strains, but not in any of the other strains, and as a result may be a useful tool in the identification of a good secretor.
To look for other signals that may be used to identify high secretors the transcript results were analysed to identify genes that were solely up or downregulated in the high secreting strains (Table 6.1 and 6.2).

### Table 6.1 Genes Upregulated in High Secretion Strains Only

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Functionality</th>
<th>Relevant Pathways</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ALG13</strong></td>
<td>Catalytic component of UDP-GlcNAc transferase, required for the second step of dolichyl-linked oligosaccharide synthesis; anchored to the ER membrane via interaction with Alg14p</td>
<td>N-Glycan biosynthesis, Metabolic Pathways</td>
<td>[288]</td>
</tr>
<tr>
<td><strong>VAN1</strong></td>
<td>Component of the mannan polymerase I</td>
<td>N-Glycan biosynthesis</td>
<td>[289]</td>
</tr>
<tr>
<td><strong>CAB4</strong></td>
<td>Pantotheine-phosphate adenyllytransferase (PPAT), which catalyzes the fourth step in the biosynthesis of coenzyme A from pantothenate</td>
<td>Pantothenate and CoA biosynthesis, Metabolic pathways</td>
<td>[290]</td>
</tr>
<tr>
<td><strong>TAF11</strong></td>
<td>Transcription initial factor TFIID subunit 11</td>
<td>Basal transcription factors</td>
<td>[291]</td>
</tr>
<tr>
<td><strong>RAD3</strong></td>
<td>5’ to 3’ DNA helicase involved in nucleotide excision and repair transcription</td>
<td>Basal transcription factors, Nucleotide excision repair</td>
<td>[292]</td>
</tr>
<tr>
<td><strong>SAR1</strong></td>
<td>GTPase; GTP-binding protein of the ARF family, component of COPII coat of vesicles</td>
<td>Protein processing in the endoplasmic reticulum</td>
<td>[293]</td>
</tr>
<tr>
<td><strong>REC8</strong></td>
<td>Meiosis-specific component of sister chromatid cohesion complex</td>
<td>Meiosis</td>
<td>[294]</td>
</tr>
<tr>
<td><strong>NFS1</strong></td>
<td>Cysteine desulfurase involved in iron-sulfur cluster (Fe/S) biogenesis; required for the post-transcriptional thio-modification of mitochondrial and cytoplasmic tRNAs</td>
<td>Thiamine metabolism, Sulphur relay system</td>
<td>[295]</td>
</tr>
<tr>
<td><strong>NSP1</strong></td>
<td>Essential component of the nuclear pore complex, which mediates nuclear import and export RNA transport</td>
<td>RNA transport</td>
<td>[296]</td>
</tr>
<tr>
<td><strong>SKP1</strong></td>
<td>Evolutionarily conserved kinetochore protein that is part of multiple protein complexes, including the SCF ubiquitin ligase complex, the CBF3 complex that binds centromeric DNA, and the RAVE complex that regulates assembly of the V-ATPase</td>
<td>Cell cycle, Ubiquitin mediated proteolysis, Protein processing in the endoplasmic reticulum</td>
<td>[297]</td>
</tr>
<tr>
<td><strong>RPN6</strong></td>
<td>Essential, non-ATPase regulatory subunit of the 26S proteasome lid required for the assembly and activity of the 26S proteasome</td>
<td>Proteasome</td>
<td>[298]</td>
</tr>
<tr>
<td><strong>CMD1</strong></td>
<td>Calmodulin; Ca++ binding protein that regulates Ca++ independent processes and Ca++ dependent processes (stress-activated pathways)</td>
<td>Phosphatidylinositol signalling system</td>
<td>[299]</td>
</tr>
</tbody>
</table>
### Table 6.2 Genes Downregulated in High Secretion Strains Only

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Functionality</th>
<th>Relevant Pathways</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPE1</td>
<td>Ornithine decarboxylase; catalyzes the first step in polyamine biosynthesis;</td>
<td>Arginine and proline metabolism</td>
<td>[300]</td>
</tr>
<tr>
<td></td>
<td>degraded in a proteasome-dependent manner in the presence of excess polyamines</td>
<td>Glutathione metabolism</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Metabolic pathways</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Biosynthesis of secondary metabolites</td>
<td></td>
</tr>
<tr>
<td>TKL1</td>
<td>Transketolase; similar to Tkl2p</td>
<td>Pentose phosphate pathway</td>
<td>[301]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Methane assimilation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Metabolic pathways</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Biosynthesis of secondary metabolites</td>
<td></td>
</tr>
<tr>
<td>ORC6</td>
<td>Subunit of the origin recognition complex, which directs DNA replication by</td>
<td>Cell Cycle</td>
<td>[302]</td>
</tr>
<tr>
<td></td>
<td>binding to replication origins and is also involved in transcriptional</td>
<td>Meiosis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>silencing; phosphorylated by Cdc28p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORC5</td>
<td>Subunit of the origin recognition complex 5</td>
<td>Cell Cycle</td>
<td>[302]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Meiosis</td>
<td></td>
</tr>
<tr>
<td>MTR2</td>
<td>mRNA transport regulator, essential nuclear protein</td>
<td>Ribosome biogenesis</td>
<td>[303]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RNA transport</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>mRNA surveillance pathway</td>
<td></td>
</tr>
<tr>
<td>PFK27</td>
<td>6-phosphofructo-2-kinase, catalyzes synthesis of fructose-2,6-bisphosphate</td>
<td>Fructose and mannose metabolism</td>
<td>[304]</td>
</tr>
<tr>
<td>RPL7A</td>
<td>Protein component of the large (60S) ribosomal subunit</td>
<td>Ribosome</td>
<td>[305]</td>
</tr>
</tbody>
</table>

Considering that over 5000 genes were analysed to determine differences there were few that appear to be solely up- or downregulated in the high copy strains compared to wild-type strain that was not exhibited in the other strains. No clear patterns of pathways upregulated can be discerned, although there is an indication that a number of biosynthetic pathways are upregulated. There are two cell cycle genes that are downregulated in the high copy strains, which correlate with the increase in cell death during growth on methanol observed by FACS analysis (Fig 6.16). Tk1p, like Tkl2p, is required for the biosynthesis of aromatic amino acids [301, 306] as the CV2(H) and CV18(H) show a downregulation of phenylalanine, tyrosine and tryptophan biosynthesis this suggest the downregulation of TKL1 may coincide with the reduced production of aromatic amino acids. As aromatic amino acids in HSAp only constitute 9.2% perhaps this low percentage corresponds to the reduced biosynthesis.
Interestingly there are no genes which are commonly up- or downregulated exclusively in low secreting strains compared to X33 that are not evident in all high or mid-level secretors. This suggests that the causes of low secretors may be more multidimensional and that the evidence we have seen highlighting good secretors may just be as a result of increased titre.

*UBLE1A*, which is involved in the E1 subunit of the ERAD, is only upregulated in high secreting strains with the exception of CV5(L). While CV5(L) is a low secreting strain and shows no obvious differences from the other low expressing genes in terms of titre (Fig 6.9) there are signs that indicate it often behaves as a high secreting strain. This suggests that CV5(L) may be a low secretor due to an overactive ERAD. Contrarily exocytosis of cellular transport vesicles occurs through the SNARE-interaction in vesicular transport pathway. *VAM4*, which is involved in this pathway was upregulated in CV5(L) (Fig 6.25).

Figure 6.25 SNARE-Interactions in Vesicular Transport Pathway in CV5(L)
Pathway of SNARE-Interactions in Vesicular Transport in CV5(L) compared to wild-type. Genes in blue are upregulated, while genes in red are downregulated in the clonal variant strain compared to wild-type. Green boxes indicate organism-specific pathways.
This suggests that CV5(L) has signals similar to the high secretors but that there are blockages elsewhere within the system that result the low titre observed. This highlights the potential problems of creating an indirect screening method by which to select high secretors. Additional screening of titre levels would clearly be required after establishing a reporter system to determine a good secretor.

6.3.5 The anomalous CV7(L)

The titre levels exhibited by CV(L) was 9 mg L⁻¹, which is in the same range as the other low secretors, yet the *HSA* transcript levels was 40-fold that of the other low secretors, comparable to the high secreting strains (Fig 6.7 and 6.9). Clearly, this suggests that this strain has a post-transcriptional blockage in the secretory pathway. What is apparent when looking at CV7(L) is the large number of genes that were significantly up- or downregulated compared to all of the other strains. 3934 genes were differentially regulated compared to wild-type, whereas in CV5(L) only 2124 genes and CV14(L) 3535 genes were regulated significantly differently from X33. The number of pathways up- and downregulated in CV7(L) are higher than in any other strain (Table 6.3).

Table 6.3 Pathways Significantly (p<0.1) Up- and Downregulated in CV7(L)

<table>
<thead>
<tr>
<th>Pathways Upregulated in CV7(L)</th>
<th>Pathways Downregulated in CV7(L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ribosome</td>
<td>Ubiquitin mediated proteolysis</td>
</tr>
<tr>
<td>Biosynthesis of secondary metabolites</td>
<td>Cell cycle - yeast</td>
</tr>
<tr>
<td>Oxidative phosphorylation</td>
<td>MAPK signalling pathway - yeast</td>
</tr>
<tr>
<td>Metabolic pathways</td>
<td>Regulation of autophagy</td>
</tr>
<tr>
<td>Cysteine and methionine metabolism</td>
<td>Nucleotide excision repair</td>
</tr>
<tr>
<td>Lysine biosynthesis</td>
<td>Mismatch repair</td>
</tr>
<tr>
<td>Valine, leucine and isoleucine biosynthesis</td>
<td>Meiosis - yeast</td>
</tr>
<tr>
<td>Phenylalanine, tyrosine and tryptophan biosynthesis</td>
<td>Proteasome</td>
</tr>
<tr>
<td>Alanine, aspartate and glutamate metabolism</td>
<td>DNA replication</td>
</tr>
<tr>
<td>Ubiquinone and other terpenoid-quinone biosynthesis</td>
<td>Endocytosis</td>
</tr>
<tr>
<td>Terpenoid backbone biosynthesis</td>
<td>Natural killer cell mediated cytotoxicity</td>
</tr>
<tr>
<td>Peroxisome</td>
<td>SNARE interactions in vesicular transport</td>
</tr>
<tr>
<td>Sulfur metabolism</td>
<td>Folate biosynthesis</td>
</tr>
<tr>
<td>Aminoacyl-tRNA biosynthesis</td>
<td></td>
</tr>
<tr>
<td>mRNA surveillance pathway</td>
<td></td>
</tr>
<tr>
<td>Tyrosine metabolism</td>
<td></td>
</tr>
<tr>
<td>Phenylalanine metabolism</td>
<td></td>
</tr>
</tbody>
</table>

When looking at the regulation of the autophagy pathway all other strains showed consistent upregulation of *ATG7*, but in CV7(L) this gene was downregulated. The protein ATG7 is involved in cytoplasm to vacuole transport and also plays a role in autophagy. It is predicted to be an E1-like
activating enzyme and essential for the upregulation of autophagy \textsuperscript{[79, 307]}. The majority of genes involved in the CV7(L) autophagy pathway were significantly downregulated. As the ubiquitin mediated pathway was also downregulated (Fig 6.24) this implies that, despite the transcript level, protein production was not excessive, suggesting that either translation was reduced in CV7(L) or perhaps protein was not being correctly targeted to the ER (Fig 6.26).

According to a paper that discusses autophagy post-methanol induction in what is described as the “lag phase autophagy” (LPA), starvation of amino acids could elicit the induction of autophagy \textsuperscript{[308]}. Autophagy is required to recycle amino acids for either the production of proteins or for the cellular reorganisation of the cell in order to deal with a methanol feed (predominantly the increase in peroxisomes) \textsuperscript{[80, 309]}. Perhaps the fact that CV7(L) does not upregulate autophagy (more than wild-type) suggests that this basal ability to recycle amino acids is missing. Thus the cell may be starving, consistent with the FACS data which indicates that after 24 hours on methanol 32.9\% of the cells are non-viable (Fig 6.16).
If the transcript is not being translated efficiently to protein, then perhaps RNA degradation may be occurring. In CV7(L) RNA degradation is upregulated in some aspects, though this appears to be consistent with pathways in all three groups (Fig 6.27).

**Figure 6.27 RNA Degradation Pathway in CV7(L)**
Genes up or down regulated in CV7(L) compared to wild-type X33 in the RNA degradation pathway as mapped by KEGG Mapper. Red genes indicate downregulation compared to wild-type and blue genes indicate upregulation. Green boxes indicate organism-specific pathways.

In support of the suggestion that there is limited translation CV7(L) gene expression in the SNARE interactions in vesicular transport pathway is generally lower than any other strain; a comparison with CV18(H) (vs. X33) highlights the differences that can be seen in the genes that are up or downregulated (Fig 6.28).
Figure 6.28 A Comparison of the SNARE Interactions in Vesicular Transport pathway between CV7 and CV18

Genes up or downregulated in A: CV7(L) to B: CV18(H) compared to wild-type X33 in the SNARE Interactions in Vesicular Transport pathway as mapped by KEGG Mapper. Red indicates genes downregulated, and blue genes upregulated in comparisons to X33. Green boxes indicate organism-specific pathways.

The downregulation of large portions of the SNARE interactions in vesicular transport pathway confirms that there is a reduction in secretion of protein, even compared to the non-recombinant wild-type. This was confirmed by assessing the number of genes up or downregulated for protein export (Fig 6.29).
There were more protein export genes significantly downregulated in CV7(L) than in all of the other strains. This supports the theory that there is a bottleneck between transcription and translation in this particular strain.

Investigation of the pathways upregulated in CV7(L) shows that a large number are related to amino acid biosynthesis. These include the alanine, aspartate and glutamate pathway, the lysine biosynthesis pathway and the cysteine and methionine pathway. This is consistent with this strain being starved for amino acids because of limited autophagy-related turnover which limits translational activity. This hypothesis is supported by manual annotation of the data which indicated that CV7(L) showed 50% of biosynthesis genes upregulated, whereas the other low copy strains showed an average of 22.7% (Appendix 9.7).

UPR and ERAD were not upregulated as protein synthesis was restricted. Thus there was no build up of unfolded protein, which would typically trigger these responses. This also helps explain why SNARE interactions are minimal; there is little recombinant protein around to migrate through the relevant pathways. It would be an interesting experiment to supplement CV7(L) growth medium with excess amino acids, as although these strains were grown in rich media the requirements may not have been met using the standard media. It is also important to note that this strain had been...
growing for 24 hours, a typical time for producing protein and one equivalent to industry standards. A time course would have to be analysed in order to determine whether amino acid starvation occurred pre-induction or early post-induction with methanol.

6.4 Discussion
Clonal variation does not seem to arise from a single attribute such as differences in the secretory pathway. It is suggested that clones vary in protein production for multiple different reasons. From the RT-qPCR results alone the increased transcription of some strains can explain the increase in secretion (Fig 6.7), although this does not explain the reasons for higher transcriptional levels in these strains. However, other strains which show high levels of transcript are not able to successfully translate this into a high level of protein.

The decision to split the clones into three cohorts based on high, medium and low secretion rates proved not to have any basis in terms of common mechanisms and it became apparent that it was more accurate to evaluate each clone on an individual basis. When looking at the specific pathways assigned by the program KOBAS that were up or downregulated there was very little uniformity amongst the groups, even at the lower stringency eventually applied. The corrected p-value takes into consideration the total number of genes that were analysed, which in this case was over 5000. This large number can make it insensitive. For instance, in the proteasome pathway CV2(H) has 17 out of 34 genes that are statistically upregulated in comparison to X33, but using the corrected p-value of ≤0.2 failed to recognise this pathway as upregulated.

Interestingly in both the mid and low secreting strains the ubiquitin mediated proteolysis pathway was downregulated, but this was not the case for the high secreting strains. CV2(H), CV18(H) and CV23(H) all showed upregulation of the gene SKP1 which is involved in the ubiquitin mediated proteolysis pathway, but no other strains show an upregulation of this gene. This immediately highlights a potential difference between the strains that may be of diagnostic use in the future. It, therefore, seems that upregulation of ERAD is a common feature of the high secreting strains, which also show an increased level of cell death after induction with methanol as shown by FACS analysis (Fig 6.16). Induction of the ERAD is clearly a sign of stress which directly or indirectly affects cell viability; an apoptosis like phenomena has been reported in P. pastoris [310-311].

This increased cell death leads to an interesting question regarding the appropriate length of expression. It had been determined that 24 hours resulted in sufficient protein to differentiate between a good secretor and a poor secretor, but perhaps leaving cells for this long explains some of the results that were observed. For instance, UPR was not obviously upregulated for many of the
strains; *HAC1* for instance was only upregulated in CV5(L), CV7(L), CV8(M), CV15(M) and CV23(M), but if a sample had been taken earlier perhaps upregulation of this key UPR signal might have been observed in all of the strains. A paper by Whyteside *et al.* indicates that while upregulation of the UPR can be maintained up to 48 hours post-induction with methanol some signals decrease after 24 hours, as can be seen with the levels of *HAC1* transcript in the wild-type strain [173].

Furthermore as UPR is a feedback control loop it is possible that these strains have effectively controlled the amount of unfolded protein being passed through the system thus returning to a baseline comparable to that of X33 [153]. Continuous upregulation of UPR leads to ERAD activation [164] explaining why ERAD is upregulated in the high secretory strains. It would have been beneficial to analyse these strains over a time course to distinguish whether different signals are prevalent early or late in the expression profile.

The fact that (with the exception of CV7(L)) the high secreting strains exhibited much higher levels of cell death than the low and mid secreting strains when grown in methanol as indicated by the LIVE/DEAD flow cytometry analysis (Fig 6.16) may be a factor that could be developed for creating a more efficient secreting strains. If it was possible to reduce cell death in these strains, productivity should increase, leading to higher titres. This has been investigated by Weis *et al.* who attempted to limit cell death through a strict feeding strategy [310]. It was determined that cell death increased with starvation due to poor mixing of cultures, which may occur in shake flasks as well as micro-titre plates. Therefore a controlled feed would limit starvation in the attempts to limit cell death. Growing cultures of these CVs in a bioreactor may have given more detailed explanations of how the cultures behave on a production scale.

As indicated CV7(L) appeared to be nutritionally starved resulting in a level of cell death comparable to that of the high copy strains (Fig 6.16). If the cells were nutritionally starved it is logical that cell death may occur due to excessive stress on the cells [310]. However, it is not possible to determine precisely what makes it nutritionally starved, except that UPR, ERAD and autophagy are not induced. Upregulation of the biosynthetic pathways should clearly compensate for nutritional starvation.

Clonal variant strains were analysed using microarray analysis in order to comprehensively determine differences related to gene expression profiles. The development of three “*P. pastoris*” sequences was essential for the annotation of the microarray analysis [91-93]. Genome sequencing, whilst extremely comprehensive, results in difficulties when establishing the effects of variation on functionality [312]. Nonetheless, once varying functionalities have been ascertained as key aspects
to be investigated (as carried out through this report) genome sequencing could be used to pinpoint precise mutations within genes to identify areas for strain engineering [312].

All three pathways involved in the UPR appear on the KEGG map irrespective of the fact that only IRE1 is known to exist in P. pastoris [71]. However, a PERK homolog appeared to be consistently downregulated in the CV strains, irrespective that no PERK homology being confirmed in P. pastoris. This raises questions regarding UPR activation in P. pastoris. Details regarding this particular gene indicate that it is a protein kinase involved in the interaction with eIF2, which ultimately works as a translation initiation factor [92]. Bioinformatic analysis reveals that this gene shows 42% sequence identity to PERK and 41% to S. cerevisiae GCN2. A knockout of the PERK gene would determine the loss of function phenotype. Interestingly both PERK and GCN2 are important for the attenuation of the cell cycle progression upon the activation of UPR [281, 283, 313].

In addition to using KOBAS and KEGG pathway mapper it would have been interesting to perform gene ontology (GO) analysis on the data sets in order to provide another dimension. Unfortunately this was not possible as no P. pastoris database for GO analysis is available. A paper from 2011 extensively used GO analysis to investigate methanol utilization, protein chaperone, RNA polymerase and protein secretion or transportation [314]. However, due to the restriction of databases on GO this analysis was done using S. cerevisiae related genes. A similar analysis was not performed in this instance as, while many of the genes between the two species are related, there is an inherent bias towards well studied pathways in S. cerevisiae which may differ in P. pastoris.

Clonal variation in P. pastoris is often attributed to differences in copy number and the integration site of the vector. Studies described in this chapter show that variation occurs due to factors other than vector integration. This raises a question as to where the variation arises from, whether it is from intrinsic differences within the host strain population or whether the transformation or selection method could give rise to the variation seen. Further work will be required to fully understand the extent to which clonal variation occurs; attempts to prevent it occurring would be an interesting area of future research.
7 General Discussion and Future Work

7.1 Understanding Genetic Instability

From initial experiments of expression multi-copy clones showed evidence of genetic instability. Therefore, it was not possible to determine if clones reached secretion saturation or if counter-selection against high copy numbers of the HSA gene was occurring to reduce copy number. The upregulation of UPR indicates that multi-copy clones experience secretory stress. However, evidence of genetic instability in non-inducing conditions (Fig 3.17 and 3.18) indicates that instability is not solely due to secretory stress. Therefore the recombinogenic nature of P. pastoris could result in a baseline level of instability in regions of high homology [82].

Through maintaining selection (either through maintaining Zeocin or through using minimal media in HIS4 complemented strains) a level of stability was achieved. No clones showed evidence of losing the integrated vector entirely with a minimum copy number retained at two, compared to the initial five-copy starting strain whilst maintaining Zeocin (Fig 4.1 and 4.5). Maintaining Zeocin for expression in a 3 mL culture for 2 days at 2000 μg mL⁻¹ equates to approximately £2.8 per sample making this uneconomical. Based on 15L fed-batch fermentations run for 96 hours, the total cost of maintaining 100 μg mL⁻¹ Zeocin a single clone would cost £1050 and for 2000 μg mL⁻¹ Zeocin £21000. Therefore the use of auxotrophic selection markers was examined. Complementation of HIS4 with the integration of pIBTx3 resulted in stability of a single copy of the vector and multi-copy clones were still unstable (Fig 4.5). Generating stable multi-copy clones multiple auxotrophic markers may be more financially viable. Nonetheless the use of auxotrophic markers relies on using minimal media which results in a significant lower titre [232]. Additionally integration into the HIS4 locus results in a greater proportion of false positive colonies; hence an increased screening process in order to ascertain true transformants [215].

RAD51 and RAD52 knockout strains were designed to interrupt the homologous recombination (HR) and the HR and non-homologous end joining (NHEJ) recombination pathways respectively. No viable knockout strains of RAD51 or RAD52 were isolated. Naatsaari et al. created a KU70 knockout strain in P. pastoris to interrupt the NHEJ pathway, producing viable clones that showed evidence of stability with seven copies of the GFP, although based on analysis of the whole culture, [215]. Therefore we can hypothesise that it is the HR pathway in P. pastoris that is essential for viability. This differs from S. cerevisiae, which is able to recover from a RAD51 mutation due to other recombination mechanisms [221, 223]. Bioinformatic analysis (blastp) revealed that protein homology of RAD51 between P. pastoris and S. cerevisiae was 74%. To date no detailed investigation
into recombination pathways in *P. pastoris* have been published; potentially *S. cerevisiae* may be able to recover from *RAD51* due to an additional pathway that is not present in *P. pastoris*.

According to Marx et al. integrating vectors into the rDNA locus results in stable strains as determined on a population level by RT-qPCR [152]. Our analysis of the replicated HSA vector containing the GAP promoter resulted in unstable strains on an individual cell basis. The use of promoter had no implications on the levels of stability when integrating vectors into the rDNA locus. Unlike integration into the *AOX1* locus, no clones in the rDNA locus lost the vector completely (Fig 3.16 and 3.18). Stability under non-inducing conditions was not analysed for clones integrated into the rDNA locus, and it would have been interesting to draw a comparison compared to strains with vectors integrated into the *AOX1* locus. Furthermore a vector which directed cytoplasmic expression of HSA would help identify whether stress on the secretory pathway did have an effect on stability.

### 7.2 Examining the possibility of a fully stable clone

The use of *in-vitro* multimerisation to generate multi-copy clones did not result in full vector stability; however integration of a single copy of the *TRY1* gene remained faithful. Multiple regions of homology will increase the likelihood of trans recombination. Thus it can be hypothesised that relying on only one sequence of homology to the genome and increasing copy number without the addition of further regions of homology may lead to a more stable clone. The use of an internal ribosome entry site (IRES) could be the solution to integrating genes in tandem without repeated regions of homology [315]. However; a recent paper using the IRES system for the production of a light chain and heavy chain antibody reported that production of both proteins was reduced in comparison to using two separate promoters [316].

The generation of a *KU70*’ strain suggests that, through interruption of the NHEJ pathway, multi-copy clones show increased stability [215]. As copy number was analysed by RT-qPCR on a populational level it would be of interest to determine the level of stability of this *KU70*’ knockout strain on an individual cell basis. Furthermore it may be of benefit to combine the *KU70*’ knockout strain with vectors that integrate into the rDNA locus. This could potentially target vectors more specifically to different loci, which may increase stability.

Integration into the rDNA locus may result in more stable clones based on which rDNA locus the vectors are integrated. We have hypothesised that it is the repeat regions of homology that result in the loss of the vector. GS115 contains 16 copies of the rDNA locus in tandem; therefore theoretically up to 16 copies of the vector may integrate into the genome with only two regions of homology within the same locus [92]. No investigation into precise integration site of the rDNA locus has been
undertaken. If multiple vectors integrate into a single rDNA locus instability may be increased compared to single vectors integrating into multiple rDNA loci. In order to determine point of integration the technique of genome walking can be utilised [317]. Additionally it has been reported that rDNA stability in S. cerevisiae is controlled epigenetically and that Sir2p may increase stability by suppressing replication-dependent rDNA recombination [318]. Theoretically overexpressing Sir2p (which naturally occurs in P. pastoris) will increase vector stability.

It is my personal belief that the advent of a fully stable P. pastoris clone will not be possible. As researchers we rely on the recombinogenic nature of P. pastoris to successfully integrate recombinant proteins into the genome. However, we seem unwilling to accept that recombination of vectors out of the genome will occur just as easily. Ultimately if individual cells reach secretion saturation and maximum titre is produced then the development of a stable P. pastoris clone may not be necessary.

### 7.3 Improving screens for genetic instability

All current investigations into genetic instability measure copy number on a population level and not on an individual cell basis [152, 210, 215]. As volumetric productivity when expressing in P. pastoris is high, but specific productivity is low, analysis of genetic instability on an individual cell basis should be examined to ensure the entire culture is secreting protein to the maximum capacity. The method for determining copy number on individual colonies is a time consuming and expensive method. Genomic DNA is extracted from individual colonies before copy number is determined by RT-qPCR. In our investigations it has been impossible to determine the significance of the variation of copy number observed due to limitations in the number of colonies that were analysed due to time constraints. A streamlined process needs to be developed if analysing copy number on individual colonies is to be adopted by the scientific community. Possibilities include using novel equipment such as Bio-Rad’s QX100 droplet system, which calculates on a per molecule scale the concentration of a gene of interest. This method does not rely on quantification using either a housekeeping gene or comparison to a known one copy clone as is required by ΔΔCt [197]. Unfortunately this method still requires the extraction of genomic DNA in order to assess copy number, which is often the most time consuming aspect.

An alternative option would be to create a fluorescence based system for use with FACS, or some other tagged protein which will be visible depending on the number of genes that are present. Mattanovich and Borth reviewed the use of cell sorting to determine concentrations of protein through various organisms, including P. pastoris where Vijayasankaran et al. fused GFP to the promoters of three polyhydroxyalkanoates to determine protein expression levels [319-320]. A
method would need to be determined so that protein production was not the aspect being investigated; potentially adding a fluorescence tag to the antibiotic marker (if multiple copies of the expression cassette are amplified such as in PTVA).

7.4 Choosing the best protein

Many of the aspects that have been investigated in this thesis are dependent on the model protein. Secretion saturation and UPR induction is dependent on folding capabilities of the protein. As the production of HSA results in one of the largest titre recorded for production by *P. pastoris* at 10 g L\(^{-1}\) it is possible to hypothesise that secretion saturation is reached with fewer gene copy numbers due to overloading the secretory pathway [140]. If we consider that stress on the secretory pathway may be linked to genetic stability, different proteins may have an impact on the stability observed. Furthermore it would be of interest to determine whether indicators of high secretors identified through transcriptomic analysis were consistently up- or downregulated through the production of a different protein.

An alternative model protein which could be chosen is glucose oxidase from *A. niger*. The stability of glucose oxidase is high in *S. cerevisiae* and titre has reached 3 g L\(^{-1}\) [321]. We have discussed the advantages of codon bias towards yeast systems and as this protein has yeast origins it may be more suitable for expression. Furthermore it is possible to create a biosensor using glucose oxidase that better monitors secretion levels through an activity assay [322].

Additionally we hypothesised that an optimised HSA gene speeds up the translation efficiency due to ease of translation through readily available tRNAs; thus resulting in increased titre [236]. However, integration of multi-copies of the HSAopt gene into the rDNA locus did not result in increased in titre compared to the use of the native HSA gene, although secretion saturation appears to have been reached. It would appear that using an optimised gene merely shifts the blockage to the secretory pathway; thus increasing stress on the cells through aspects such as UPR. GpARzaHSAopt strains showed on average over a five-fold increase of *KAR2* upregulation and a four-fold increase of *PDI* upregulation compared to GpARzaHSA strains (Fig 4.22 and 5.14). Therefore due to reduced average titre of multi-copy clones containing the HSAopt gene compared to clones containing the HSA gene, codon optimisation in this instance does not increase the specific productivity of *P. pastoris*.

The method of codon optimisation used can have an effect on protein expression. As suggested by Sinclair and Choy when expressing human glucocerebrosidase increasing the A/T ratio is sufficient to increase titre [238]. Optimisation was based on the OptimumGene™ algorithm that takes into consideration not only idea codon usage for *P. pastoris* but also transcriptional and translational
efficiencies [191]. DNA2.0 utilises GeneGPS technology, which is designed on the premise that increased titre is highly affected by the frequency of highly charged tRNAs codons, especially during amino acid starvation [237]. It would be of interest to compare optimised HSA genes synthesised using different algorithms and techniques to ascertain whether the use of a HSAopt gene can have an effect on titre levels. Additionally comparing the success of alternative proteins may help identify under what conditions protein optimisation is beneficial.

7.5 Understanding Clonal Variation

The second aspect of this thesis was to investigate clonal variation through the use of transcriptomic analysis. Selecting high secreting single-copy clones could reduce the requirements for the generation of high copy clones if secretion saturation is reached with a limited number of copies of the gene of interest. Clonal variation was observed with the original GpαH1_100 and GpαH11_100 clones (Fig 6.1), in addition to the variation observed from single colonies isolated from a single transformation (Fig 6.5). To determine the effects of clonal variation DNA microarrays were used to compare gene expression levels of nine single copy HSA strains.

The three published genomes of P. pastoris allow for ease of interpretation of microarray data [91-93]. Gene expression analysis allows for a genome wide view of differences without biasing data towards targeted aspects [323]. Up- or downregulated pathways were established without previously determining functionality related to expression. For instance, all of the mid-level secretion strains showed a significant upregulation of the porphyrin and chlorophyll metabolism pathway when analysed at p ≤ 0.1 despite the apparent insignificance of this pathway for protein production. To supplement microarray data next generation sequencing may determine the cause of variation observed [324]. Potentially strains showing upregulation of particular genes may be due to mutations within the promoter regions. CV7(L) has been identified to be nutritionally starved; however the cause for this is unknown. Genome sequencing would be able to determine whether it is a mutation within a coding region of a gene required for breaking down amino acids for nutrients (such as a ubiquitin gene) or another cause. If the clonal variation strains were sequenced and compared to current genome sequences differences between clones could be identified.

Furthermore genome sequencing removes experimental bias as cross-hybridisation to user-defined sequences is not required as in gene expression experiments [43, 325].

However, genome sequencing has limitations, for instance it does not take into consideration any epigenetic variations between strains. Current techniques to analyse epigenetic variations between strains are being continuously developed. In 2011 Nucelosome Occupancy and Methylome Sequencing (NoME-Seq) was published, which uses high-resolution single molecule analysis to
determine nucleosome positions and DNA methylation [326]. By investigating epigenetic variations between P. pastoris strains it may be possible to achieve a more thorough understanding of the differences between strains.

One of the advantages of microarray analysis compared to whole genome sequencing is that gene expression levels under different conditions can be analysed. Clonal variant strains were analysed for 24 hours in 250 mL flasks; however this does not give an indication of variation over time or with a tightly controlled feeding regime. Theoretically strains may show a different gene expression profile to low secreting strains at an earlier time course; for instance UPR may be significantly upregulated within the first few hours before reaching homeostasis [162]. As the high secreting strains show upregulation of ERAD we can hypothesise that UPR may be more significantly upregulated in these strains, or potentially upregulation may occur earlier. If this is the case identifying strains that show upregulation of UPR quickly after induction may indicate a good secretor.

In P. pastoris clones that secrete well in small scale do not necessarily correlate to clones that secrete well in large scale expression systems [310]. This has been reported to largely be as a result of poor feeding regimes, something that has been tackled by the company Biosilta (Oulu, Finland) who have produced a tablet called Enpresso Y Defined tablets that with the addition of an enzyme, slowly release the carbon source for utilisation by the culture [327-328]. As controlled feeding strategies have been identified as detrimental towards controlling cell death the use of a slow release feeding system may limit the amount of cell death observed (Fig 6.16). Furthermore, as bioreactors use feeding strategies based on the growth profiles of cultures 250 mL flasks with the inclusion of slow releasing media may be more comparable for industrial processes.

Through comparisons between high secreting strains and low secreting strains we have been able to identify signals that could be developed as reporter signals to be used in an easy screening method, such as Skp1 (Table 6.1). If a fluorescent protein was fused to SKP1 then it would be possible to determine when this gene was significantly upregulated via electron microscopy for FACS. Therefore, clones that show an increased fluorescence may be indicators of high secreting strains. Furthermore the microarray analysis has highlighted areas that can be improved to create a more functional strain, such as reducing the cell-death phenomenon in high secreting strains. Upregulation of cell death in the high secreting strains may be as a result of activation of the ERAD. If cell death can be limited (potentially through a more controlled feeding strategy [310]) titre levels in the high secreting strains may be increased. Furthermore it may be possible to encourage strains to upregulate UPR to prevent activation of ERAD, while this may suggest that less protein would be produced, as this system works on a control loop it may prevent cell death from occurring [329].
7.6 Identifying the underlying cause of clonal variation

To understand clonal variation it is essential to determine where differences arise from. There are two possible causes: inherent variability in the host strain or integration of the vector into the genome. By screening the wild-type strain pre-transformation with a heterologous protein it may be possible to determine if variation can be observed. A screen could be developed using a native protein that can be monitored without the insertion of any foreign material. *P. pastoris* only secretes a limited number of native proteins [330]. Therefore it may be possible to monitor membrane proteins, which are advantageous as they are directed through the secretory pathway, to look for differences in the host strain [72]. Custom designed antibodies for immunofluorescence can be raised against membrane specific proteins for *P. pastoris*. It would then be possible to analyse variations by FACS or microscopy to determine whether wild-type cultures are homogeneous [331].

If wild-type cells consistently indicate homogeneous cultures then variation may be as a result of integration of recombinant genes into the genome. In this situation a screen should be applied post-integration of heterologous proteins to determine whether heterogeneity exists within cultures. Screens post-integration can identify indicators of high secretors (such as described for *SKP1*). However if heterogeneous cultures are identified pre-integration of a recombinant gene then wild-type strains need to be monitored more stringently. It may be possible to identify clones with proficiencies for high secretion rates. If these strains are able to be selected then potentially integration of the recombinant gene may lead to higher titres.

7.7 Conclusions

*P. pastoris* has been used for over 20 years as a tool for expressing proteins by the biopharmaceutical industry; however a detailed understanding regarding the molecular biology has often been overlooked. At the 2012 *Pichia* 2012 conference, Alpbach, Austria, it was evident that more groups are delving into the “black box” of *P. pastoris* molecular biology, often funded by pharmaceutical companies. This PhD was a multifaceted approach that started as an investigation into understanding the limitations of *P. pastoris* in terms of secretion saturation and its relationship to the UPR, but became more about understanding the full impact of clonal variation and uncovering the depth that genetic instability occurs. The impact of both clonal variation and genetic instability can have severe detrimental effects on recombinant protein expression. By rigorously screening for clonal variation and ensuring the most stable clones are selected generating clones that produce the highest titre possible may become easier. Furthermore current investigations into expanding the secretory capacity [72] may lead to higher titre making *P. pastoris* the preferential protein expression tool of the future.
8 References

10.1097/COC.0b013e3181931277.


262. Cregg, J. Development of the yeast Pichia pastoris as a recombinant protein production host. in Pichia. 2012. Alpbach, Austria.


9 Appendix

9.1 HSA Codon Optimised DNA Sequence

ATGAAATGGGTACCTTTTATCAGTTTGTTTTGTTTAGTTGCCTCCCTACTCAAGAGGTGTTTTTAGAAGAGATGCACTAAGTCTGAGTGCAGTTGAAGCTTTGAGATGACCTACAGCGTGAGTTGGACATTTCAAGATGCTTGAATGATTGCTGACGCCCGGAGAATTGGACTTGTGCTT

GCACTAAAGTCTGGAAGGGCGGTTGGGCAGAAGTTGCTGGCCGATGACTCTCGACAGATTTGTCTGACGCCTGACGTGGGACATTTCTGACTCCTGAGGCTGAGAAGTTGACGGCTGATGACAGAGCCGACTTGGCAAAGTACATCTGGAAAGTTGAAGGAATGTTGCGAGAAACCTTTGCTTGAAAAGTCTCATTGTATTGCAGAAGTTGAGAACGATGAGATGCCAGCTGACCTTCCTTCATTGGCTGCCGATTTTGTCGAAAGTAAAAGACGTTTGTAAGAATTACGCTGAGGCCAAGGACGTTTTTCAAGGCTGCATCCCAGGCTGCCCTTGGTTTG

ACGCTTTTGCGCAAGATGAGGAGACATCCTGACTTTTTTTTTTATCAGTTTGTTTTGTTTAGTTGCCTCCCTACTCAAGAGGTGTTTTTAGAAGAGATGCACTAAGTCTGAGTGCAGTTGAAGCTTTGAGATGACCTACAGCGTGAGTTGGACATTTCAAGATGCTTGAATGATTGCTGACGCCCGGAGAATTGGACTTGTGCTT
### 9.2 Oligonucleotides Used throughout this Study

Table 9.1 Oligonucleotides

<table>
<thead>
<tr>
<th>Primer Name</th>
<th>5’ – 3’ Sequence</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>001-AOX1F</td>
<td>GACTGGTCCAATTGACAAGC</td>
<td>AOX1 promoter from pPiczA and pPiczαA</td>
</tr>
<tr>
<td>002-AOX1R</td>
<td>GCAATGGCATTTCTGACATCC</td>
<td>AOX1 promoter from pPiczA and pPiczαA</td>
</tr>
<tr>
<td>004-pPIC F</td>
<td>TTT CTA CCG GGT CTG ACG CTC</td>
<td>Amplification of the AOX1 cassette from pPiczA and pPiczαA</td>
</tr>
<tr>
<td>005-pPIC R</td>
<td>TTT GAA GCT ATG GTG TGT GGG G</td>
<td>Amplification of the AOX1 from pPiczA and pPiczαA</td>
</tr>
<tr>
<td>010-KAR2q1</td>
<td>TCA AAG ACG CTG GTG TCA AG</td>
<td>qPCR primer for KAR2 (BiP) 150 bp amplicon</td>
</tr>
<tr>
<td>011-KAR2q2</td>
<td>TAT GCG ACA GCT TCA TCT GG</td>
<td>qPCR primer for KAR2 (BiP) 150 bp amplicon</td>
</tr>
<tr>
<td>012-PDIq1</td>
<td>GCC GTT AAA TTC GGT AAG CA</td>
<td>qPCR primer for PDI 144bp amplicon</td>
</tr>
<tr>
<td>013-PDIq2</td>
<td>TCA GCT CCG TCA CAT CTT TG</td>
<td>qPCR primer for PDI 144bp amplicon</td>
</tr>
<tr>
<td>020-ACT1q1</td>
<td>GCT TGG TTC CAC CCA TCT GT</td>
<td>qPCR primer for ACT1 163 bp amplicon. Housekeeping gene</td>
</tr>
<tr>
<td>021-ACT1q2</td>
<td>TGC ATA CGC TCA GCA ATA CC</td>
<td>qPCR primer for ACT1 163 bp amplicon. Housekeeping gene</td>
</tr>
<tr>
<td>022-Zeoq1</td>
<td>CAA GTT GAC CAG TGC GTG TC</td>
<td>qPCR primer for Sh ble (Zeocin resistance gene) 108 bp amplicon</td>
</tr>
<tr>
<td>023-Zeoq2</td>
<td>GAA GTC GTC CTC CAC GAA GT</td>
<td>qPCR primer for Sh ble (Zeocin resistance gene) 108 bp amplicon</td>
</tr>
<tr>
<td>047-HSA2</td>
<td>AAA GGC GCC GCT TAT AAG CCT AAG GCA GCT TGA C</td>
<td>HSA gene from GS115/HSA into pPiczA/pPiczαA with 3’ NotI restriction site</td>
</tr>
<tr>
<td>050-HSA3</td>
<td>CCC GAA TTC AAAA ATG AAG TGG GTA ACC TTT ATT TCC C</td>
<td>HSA gene from GS115/HSA into pPiczA with 5’ EcoRI restriction site</td>
</tr>
<tr>
<td>053-HSA4-SP</td>
<td>AAA GAA TTC AGG GTG GTT TTT CGT CGA GAT GC</td>
<td>HSA gene from GS115/HSA into pPiczαA removing native signal peptide with 5’ EcoRI restriction site</td>
</tr>
<tr>
<td>067-HAC1q3</td>
<td>CGA CTA CAT TAC TAC AGC TCC ATC A</td>
<td>qPCR primers for HAC1 124 bp amplicon</td>
</tr>
<tr>
<td>068-HAC1q4</td>
<td>TGC TGT AAT GTG TAA AGA TGA ATC C</td>
<td>qPCR primers for HAC1 124 bp amplicon</td>
</tr>
<tr>
<td>070-HSAq1</td>
<td>GTG GTT GAT TGC CTT TGC TCA G</td>
<td>qPCR primers for HSA 138 bp amplicon</td>
</tr>
<tr>
<td>071-HSAq2</td>
<td>GCA TTC ATT TCT CTC AGG TTC TTG</td>
<td>qPCR primers for HSA 138 bp amplicon</td>
</tr>
<tr>
<td>120-optHSA-SP</td>
<td>CCC GAA TTC AGA GGT GTT TTT AGA AGA GAT GC</td>
<td>HSAopt gene without signal peptide with 5’ EcoRI restriction site for cloning into pPiczαA</td>
</tr>
<tr>
<td>121-optHSArev</td>
<td>AAA GCG GCC GCT TAC AAA CCA AGG GCA GCC</td>
<td>HSAopt gene without signal peptide with 5’ EcoRI restriction site for cloning into pPiczαA</td>
</tr>
<tr>
<td>142-AOX integr F Pmel</td>
<td>GAA CAT CAC TCC AGA TGA GGG</td>
<td>Sequencing primers across the AOX1 integration site 50 bp 5’ Pmel</td>
</tr>
<tr>
<td>143-AOX inter R Pmel</td>
<td>CTG GCC GTT AGC ATT TCA ACG</td>
<td>Sequencing primers across the AOX1 integration site 50 bp 3’ Pmel</td>
</tr>
<tr>
<td>146 – RAPD3</td>
<td>AGT CAG CCA C</td>
<td>Random decamer primer for RAPD-PCR</td>
</tr>
<tr>
<td>188-HSAoptq3</td>
<td>TGA CTT TGC AGC TTT CGT TG</td>
<td>qPCR primer for HSAopt 91 bp amplicon</td>
</tr>
<tr>
<td>189-HSAoptq4</td>
<td>GGA TGC AGC AAC CAA TTT CT</td>
<td>qPCR primer for HSAopt 91 bp amplicon</td>
</tr>
<tr>
<td>242-pIB2 inv1</td>
<td>CGG TAC CCG GGT ATC CAC TAG TCT CG</td>
<td>Inverse PCR from pIB2 to remove the GAP promoter</td>
</tr>
<tr>
<td>243- plB2 inv2</td>
<td>AAA AGA TCT GTC AGA GGT TTT CAC CGT CAT CAC C</td>
<td>Inverse PCR from plB2 to remove the GAP promoter</td>
</tr>
<tr>
<td>264- RAD51 SeqF1</td>
<td>CTTCCATGATGGCGTTCTT</td>
<td>Sequencing primer for αTx3R51 250 bp 5' RAD51</td>
</tr>
<tr>
<td>265- RAD51 SeqR1</td>
<td>ATATGGCTCCGGAGTTTTGA</td>
<td>Sequencing primer for αTx3R51 250 bp 3' RAD51</td>
</tr>
<tr>
<td>266- RAD52 SeqF1</td>
<td>AGGCCTAAGGGCAGAGCAAT</td>
<td>Sequencing primer for αTx3R52 250 bp 5' RAD52</td>
</tr>
<tr>
<td>267- RAD52 SeqR1</td>
<td>CAC TGT AGG TTG GGG GCT AA</td>
<td>Sequencing primer for αTx3R52 250 bp 3' RAD52</td>
</tr>
<tr>
<td>280-Zeo SeqF2</td>
<td>GGT AGG GGT TTG GGA AAA TCA TTC G</td>
<td>Sequencing primer for TEF1, EM7, Sh ble, CYC1 TT</td>
</tr>
<tr>
<td>281-Zeo SeqR2</td>
<td>GCG GCC TTT TTA CGG TTC CTG G</td>
<td>Sequencing primer for TEF1, EM7, Sh ble, CYC1 TT</td>
</tr>
<tr>
<td>293 - GibR51-plB2</td>
<td>TAT CAC GAG GCC CTT TCG TCA ATT GCA GTG CTG ATC ACT AAT CAA GTT</td>
<td>Gibson Primer pIBTx3R51: Forward attaching RAD51 3' to pIB2-HIS4</td>
</tr>
<tr>
<td>294-GibR51-3-5-EcoRV</td>
<td>ATG ACT TCA TGG TTC GAC ATG ATA TCT TAG TCA TCA TCC TCT GAG GGG TCC</td>
<td>Gibson Primer pIBTx3R51: Reverse attaching RAD51 3' to RAD51 5' with an EcoRV site</td>
</tr>
<tr>
<td>295-GibR51-5-EcoRV</td>
<td>AAA GAT ATC ATG TCG AAC CAT GAA GTC ATA GAG G</td>
<td>Gibson Primer pIBTx3R51: Forward RAD51 5' plus EcoRV site (5')</td>
</tr>
<tr>
<td>296-GibR52-plB2</td>
<td>TAT CAC GAG GCC CTT TCG TCA GGT GGA AGA AAA GGT CTA AAG CCC</td>
<td>Gibson Primer pIBTx3R52: Forward attaching RAD52 3' to RAD52 5' with an EcoRV site</td>
</tr>
<tr>
<td>297-GibR52-3-5-EcoRV</td>
<td>AAA GAT ATC GAC GCT GAG C</td>
<td>Gibson Primer pIBTx3R52: Forward RAD52 5' plus EcoRV site (5')</td>
</tr>
<tr>
<td>331-RAD51-5</td>
<td>ATG TCG AAC CAT GAA GTC ATA GAG G</td>
<td>Gibson Primer 51-HIS: Forward RAD51 5'</td>
</tr>
<tr>
<td>337-RAD51-5-HIS</td>
<td>TGC GTA TGC AGG TAG CAA GGG AAA TGT CAT TAA CAG TAA GCA GAG CTC TTT TAG GAG T</td>
<td>Gibson Primer 51-HIS: Reverse RAD51 5' (250 bp) with HIS4 homology</td>
</tr>
<tr>
<td>338-HIS4-RAD51-5</td>
<td>ATA CTC CTA AAA GAG CTC TGC TTA CTG TTA ATG ACA TTT CCC TTG CTA CCT GC</td>
<td>Gibson Primer 51-HIS: Forward HIS4 with homology to RAD51 5' (250 bp)</td>
</tr>
<tr>
<td>339-HIS4-RAD51-3</td>
<td>ACA ACT TGA TTA GTG ATC AGC ACT GCA ATT TTA AAT AAG TCC CAG TTT CTC CAT ACG AAC C</td>
<td>Gibson Primer 51-HIS: Reverse HIS4 with homology to RAD51 3' (250 bp)</td>
</tr>
<tr>
<td>340-RAD51-3-HIS</td>
<td>GTT CGT ATG GAA AAA CTG GGA CTT ATT TAA AAT TGC AGT GCT GAT CAC TAA TCA AG</td>
<td>Gibson Primer 51-HIS: Forward RAD51 3' (250bp) with HIS4 Homology</td>
</tr>
<tr>
<td>341-RAD52-5</td>
<td>ATG TCT TTT GAT GAC GGT GAG C</td>
<td>Gibson Primer 52-HIS: Forward RAD52 5'</td>
</tr>
<tr>
<td>342-RAD52-5-HIS</td>
<td>TGC GTA TGC AGG TAG CAA GGG AAA TGT CAT GAA TAA CAG ACG AAA GTC CCA TTG A</td>
<td>Gibson Primer 52-HIS: Reverse RAD52 5' with HIS4 homology</td>
</tr>
<tr>
<td>343-HIS4-RAD52-5</td>
<td>AAA TTT CAA TGG GAC TTT CGT CTC TTA TTC ATG ACA TTT CCC TTG CTA CCT GC</td>
<td>Gibson Primer 52-HIS: Forward HIS4 with homology to RAD52 5'</td>
</tr>
<tr>
<td>344-HIS4-RAD52-3</td>
<td>TTC GTA GGA TCT ACA GTT CTT CTT AAA GAA TAA AAT AAG TCC CAG TTT CTC CAT ACG AAC C</td>
<td>Gibson Primer 52-HIS: Reverse HIS4 with homology to RAD52 3'</td>
</tr>
<tr>
<td>345-RAD52-3-HIS</td>
<td>GTT CGT ATG GAA AAA CTG GGA CTT ATT TAA TCT TTT AAG AAG AAC TGT AGA TCC TAC GAA</td>
<td>Gibson Primer 52-HIS: Forward RAD52 3' with homology to HIS4</td>
</tr>
<tr>
<td>346-RAD52-3</td>
<td>TTA ATT CGA AGC TGG AGA GTT</td>
<td>Gibson Primer 52-HIS: Reverse RAD52 3'</td>
</tr>
<tr>
<td>Oligonucleotides                                                                haustion</td>
<td>Application</td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------------------------------------------</td>
<td>---------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>353-rDNA IntF</td>
<td>TTC CCC GTT ACC CGT AGA AAC C</td>
<td>Forward primer to check integration into rDNA locus</td>
</tr>
<tr>
<td>354-rDNA IntR</td>
<td>TGT TCA CCC ACT AAG GAA CG</td>
<td>Reverse Primer to check integration into rDNA locus</td>
</tr>
<tr>
<td>366-GAP-rDNA GibF</td>
<td>TAA CCT GTC TCA CGA CGG TCT</td>
<td>Gibson Primer pGRz: Forward GAP + rDNA homology</td>
</tr>
<tr>
<td></td>
<td>AAA CCC AGC AGA TCC TTT TTT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GTA GAA ATG TCT TGG</td>
<td></td>
</tr>
<tr>
<td>367-GAP-aMF GibR</td>
<td>AGT AAA AAT TGA AGG AAA TCT</td>
<td>Gibson Primer pGRz: Reverse GAP + alpha MF homology</td>
</tr>
<tr>
<td></td>
<td>CAT CGT TTC ATA GTT CAA TTG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATT GAA ATA GGG</td>
<td></td>
</tr>
<tr>
<td>368-amF-GAP GibF</td>
<td>TGT CCC TAT TTC AAT CAA TTG AAC</td>
<td>Gibson Primer pGRz: Forward alpha MF + GAP homology</td>
</tr>
<tr>
<td></td>
<td>AAC TAT GAA ACG ATG AGA TTT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCT TCA AT</td>
<td></td>
</tr>
<tr>
<td>369-pUC-rDNA GibR</td>
<td>TTT CTG CCC TAT CAA CTT TCG ATG</td>
<td>Gibson Primer pGRz: Reverse pUC + rDNA homology</td>
</tr>
<tr>
<td></td>
<td>GTA GGA GAT CTC ATG ACC AAA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ATC CCT TAA C</td>
<td></td>
</tr>
<tr>
<td>370-rDNA-pUC GibF</td>
<td>CTC ACG TTA AGG GAT TTT GGT</td>
<td>Gibson Primer pGRzα: Forward rDNA + pUC homology</td>
</tr>
<tr>
<td></td>
<td>CAT GAG ATC TCC TAC CAT CGA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAG TTG ATA GGG C</td>
<td></td>
</tr>
<tr>
<td>371-rDNA-GAP</td>
<td>ACA CCA AGA CAT TTC TAC AAA</td>
<td>Gibson Primer pGRzα: Reverse rDNA + GAP homology</td>
</tr>
<tr>
<td></td>
<td>AAA GGA TCT GCT GGG TTT AGA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCG TCG TGA</td>
<td></td>
</tr>
<tr>
<td>372-rDNA-AOX1TT GibF</td>
<td>AG AAG ATT AAG TGA GAC CTT</td>
<td>Gibson Primer pARzα: Forward rDNA + homology to AOX1 TT</td>
</tr>
<tr>
<td></td>
<td>CGT TTG TGG TCG TCC TAG CAT CGA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAG TTG ATA GGG C</td>
<td></td>
</tr>
<tr>
<td>373-rDNA-TEF1 GibR</td>
<td>ACA TTT TAG AGC TAT GTT GGG</td>
<td>Gibson Primer pARzα: Reverse rDNA + homology to TEF1 promoter</td>
</tr>
<tr>
<td></td>
<td>GTG ATG CAC GCT GGG TTT AGA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCG TCG TGA G</td>
<td></td>
</tr>
<tr>
<td>374-rDNA-pUC GibF</td>
<td>TCA CGT TAA GGG ATT GTG GTC</td>
<td>Gibson Primer pARzαHAopt BglII: Forward rDNA + Homology to pUC</td>
</tr>
<tr>
<td></td>
<td>ATG AGA TCA TCC TAC CAT CGA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAG TTG ATA GGG C</td>
<td></td>
</tr>
<tr>
<td>375-rDNA-AOX1P GibR</td>
<td>ATT CAA CCT TTC GTC TTT GGA</td>
<td>Gibson Primer pARzαHAopt BglII: Reverse rDNA + Homology to AOX1 promoter</td>
</tr>
<tr>
<td></td>
<td>TAG ATC GCT GGG TTT AGA CCG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TCG TGA G</td>
<td></td>
</tr>
<tr>
<td>383-TRY1q7</td>
<td>TCC CTC ATC AAC GAA CAG TG</td>
<td>qPCR primers for TRY1 133 bp amplicon</td>
</tr>
<tr>
<td>384-TRY1q8</td>
<td>GGA TGA TCT TGG CTG CAT TG</td>
<td>qPCR primers for TRY1 133 bp amplicon</td>
</tr>
</tbody>
</table>

Oligonucleotides were purchased from Invitrogen, Paisley, UK.
9.3 Zeocin Gene and Promoter Region Sequencing

9.1 Alignment of DNA sequences within the TEF1 promoter region from strains selected on 100 µg mL\(^{-1}\) Zeocin or 2000 µg mL\(^{-1}\) shows variation within the promoter regions

DNA sequencing of TEF1 and EM7 promoter, Sh ble gene and CYC1 transcription termination region of clones that showed increased resistance to Zeocin but did not show an increase in copy number. Both parental and progeny (post-PTVA) clones are present to show where the variation occurs. 1_100, 11_100 and their progeny are included as positive controls as the progeny clones show an increase in copy number.
9.4 GpαH11_2000B Fermentation Trace

Example fermentation trace from 15 L fermentation. The example show is from pαH11_2000B, the five copy strain. A) Fermentation trace of temperature, stirrer speed, PO₂ percentage, pH and airflow. B) Carbon dioxide evolution rate (CER) and Oxygen uptake rate (OUR). 1: Initiation of fed-batch. 2: Glycerol-feed turned off and methanol shot given. 3: Methanol feed turned on.

9.2 Fermentation Traces GpαH11_2000B
9.5 DNA Sequencing of Clonal Variants for Integration Site into the P. pastoris genome
9.9.3 Sequence alignment of the integration site for clones used in microarray analysis.
Sequence alignment of the region around the Pmel directed integration site of clones GpaHCV 2, 5, 7, 8, 14, 15, 16, 18 and 23. Pmel recognises and cuts at the sequence GTTAAAC, located at 467 and highlighted in yellow. Sequences both upstream and downstream of the integration site were identical.
9.6 Volcano plots of Microarray Expression Data for Clonal Variant strains compared to X33

9.6.1 High Secreting Strains

9.4 Volcano plots for High Secreting Strains CV2(H), CV18(H) and CV23(H) at fdr ≤ 0.05 compared to X33
9.6.2 Mid-Level Secreting Strains

9.5 Volcano plots for Mid-Level Secreting Strains CV8(H), CV15(H) and CV16(H) at fdr ≤ 0.05 compared to X33
9.6.3 Low Secreting Strains

Volcano plots for Low Secreting Strains CV5(H), CV7(H) and CV14(H) at fdr ≤ 0.05 compared to X33
9.7 Manual annotation of Microarray analysis

9.7.1 CV18(H)

Figure 9.7 Percentage of Genes Up- or Downregulated in CV18(H) through manual annotation compared to wild-type X33.
9.7.2 CV23(H)

9.8 Percentage of Genes Up- or Downregulated in CV23(H) through manual annotation compared to wild-type X33.
9.7.3 CV8(M)

9.9 Percentage of Genes Up- or Downregulated in CV8(M) through manual annotation compared to wild-type X33.
9.7.4 CV15(M)

9.10 Percentage of Genes Up- or Downregulated in CV15(M) through manual annotation compared to wild-type X33.
9.7.5 CV16(M)

9.11 Percentage of Genes Up- or Downregulated in CV16(M) through manual annotation compared to wild-type X33.
9.7.6 CV5(L)

9.12 Percentage of Genes Up- or Downregulated in CV5(L) through manual annotation compared to wild-type X33.
9.7.7 CV7(L)

9.13 Percentage of Genes Up- or Downregulated in CV7(L) through manual annotation compared to wild-type X33.
9.7.8 CV14(L)

9.14 Percentage of Genes Up- or Downregulated in CV14(L) through manual annotation compared to wild-type X33.
9.8 Protein Processing in the Endoplasmic Reticulum Pathways

9.8.1 High Copy Strains

Genes in blue are upregulated, while genes in red are downregulated in the clonal variant strain compared to X33. A) CV18(H); B) CV23(H). Green boxes indicate organism-specific pathways.
9.8.2 Mid-Level Secreting Strains

Genes in blue are upregulated, while genes in red are downregulated in the clonal variant strain compared to X33. A) CV8(M); B) CV15(M); C) CV16(M). Green boxes indicate organism-specific pathways.

9.16 Protein Processing in the Endoplasmic Reticulum Pathways in Mid-Level Secreting Strains
9.8.3 Low-Level Secreting Strains

Genes in blue are upregulated, while genes in red are downregulated in the clonal variant strain compared to X33. A) CV5(L); B) CV7(L); C) CV14(L). Green boxes indicate organism-specific pathways.

9.17 Protein Processing in the Endoplasmic Reticulum Pathways in Low Secreting Strain
9.9 Ubiquitin Mediated Proteolysis Pathway in High Secreting Strains

Genes in blue are upregulated, while genes in red are downregulated in the clonal variant strain compared to wild-type.  

A) CV2(H);  
B) CV18(H). Green boxes indicate organism-specific pathways.