



Provided by the author(s) and University of Galway in accordance with publisher policies. Please cite the published version when available.

Title	Mapping protein architecture in centromeres of <i>Equus asinus</i>
Author(s)	Masterson, Teri Anne
Publication Date	2016-11-28
Item record	<a href="http://hdl.handle.net/10379/6344">http://hdl.handle.net/10379/6344</a>

Downloaded 2024-05-14T13:57:32Z

Some rights reserved. For more information, please see the item record link above.





**Mapping protein architecture in centromeres of *Equus asinus***

A thesis presented to the National University of Ireland, Galway for the degree of Doctor of Philosophy

by

Teri Anne Masterson B.Sc  
Center for Chromosome Biology  
Department of Biochemistry  
National University of Ireland, Galway

November 2016

**Chair of Biochemistry:** Prof. Noel F. Lowndes  
**Head of Department:** Dr. Michael P. Carty  
**PhD Academic Supervisor:** Prof. Kevin F. Sullivan

## Table of Contents

List of Figures .....	iii
List of tables .....	vi
List of commands .....	xi
Acknowledgements .....	xii
Abbreviations.....	xiii
Abstract .....	xv
Chapter1- Introduction.....	1
1.1 Centromere identity .....	1
1.2 CENP-A .....	1
1.3 CCAN.....	2
1.4 Chromosomal Passenger Complex (CPC).....	4
1.4.1 INCENP .....	5
1.4.2 Aurora B.....	5
1.4.3 Survivin.....	6
1.4.4 Borealin/Dasra B.....	7
1.4.5 Chromosomal Passenger complex localization.....	8
1.5 Cohesin .....	11
1.5.1 Cohesin loading.....	11
1.5.2 Cohesin removal .....	13
1.6 Neocentromeres and satellite free centromeres.....	15
1.7 The equid model system.....	16
1.8 Transposable elements .....	20
1.8.1 Transposable elements at the centromere .....	22
Research objectives.....	23
Chapter 2 Materials and Methods.....	24
2.1 Materials – Wet lab .....	24
2.1.1 Chemical reagents and consumables.....	24
2.1.2 Molecular biology reagents, strains and equipment.....	24
2.1.3 Protein methods .....	27
2.1.4 Tissue culture reagents .....	28
2.2 Methods – Wet Lab.....	28
2.2.1 Nucleic Acid techniques .....	28
2.2.2 Protein techniques.....	31
2.2.3 Cell culture .....	36
2.2.4 Immunofluorescence Microscopy.....	39
2.3 Materials – Dry lab .....	40
2.3.1 Hardware.....	40
2.3.2 Software .....	40
2.4 Methods – Dry Lab .....	40
2.4.1 Quality control of sequenced reads.....	40
2.4.2 Generation of the EquDonk2.0 hybrid genome.....	41
2.4.3 Alignment of reads to the genome using Bowtie2.....	41
2.4.4 File conversion and indexing .....	42
2.4.5 Normalising data .....	42
2.4.5 Visualising data .....	43
2.4.6 MACS peakcalling.....	43
2.4.7 Read count extraction .....	43

2.4.8 Relative abundance of CENP-A at centromere domains .....	44
2.4.9 Identification of centromeres in the Guanzhong donkey .....	44
2.4.10 Analysis of repetitive sequences.....	44
2.4.11 Schematic representation of centromere comparisons.....	45
Chapter 3 Preparation of CENP-A antibody .....	47
3.1 Introduction .....	47
3.2 Preparation of CENP-A gene .....	47
3.3 Protein expression and purification.....	47
3.3.1 Solubility assay.....	48
3.3.2 Inclusion bodies.....	49
3.6 CENP-A ChIPSeq.....	58
3.6.1 FRiP (fraction of reads in peaks) analysis.....	64
3.7 CENP-A Correlation .....	65
3.8 Relative abundance of CENP-A at satellite-free centromeres .....	67
3.9 Discussions .....	69
Chapter 4 Interindividual and interspecies centromere comparison.....	71
4.1 Introduction .....	71
4.2 Domain comparison .....	72
4.3 Domain analysis .....	159
4.4 Discussion .....	167
Chapter 5 - Toward the identification of the Inner Centromere .....	169
5.1 Introduction .....	169
5.2 Antibody identification .....	171
5.3 Antibody characterization .....	172
5.3.1 Cohesin-Smc1 .....	173
5.3.2 CPC-Aurora B .....	174
5.3.3 CPC- Survivin.....	176
5.3.4 CPC- Borealin .....	178
5.4 Preparation of mitotically enriched cell populations .....	179
5.4.1 Mitotic arrest.....	180
5.4.2 Synchrony and release.....	186
5.5 ChIPSeq using mitotically enriched populations .....	193
5.5.1 CENP-A ChIPSeq .....	193
5.5.2 Cohesin-Smc1 ChIPSeq.....	199
5.5.3 CPC-Aurora B, Survivin and Borealin .....	205
5.6 Discussion .....	208
Chapter 6 - Conclusion.....	211
Chapter 7 References .....	215
Chapter 8 Appendices .....	230
<i>Appendix I</i> .....	230
<i>Appendix II</i> .....	241

## List of Figures

Figure 1.1 Protein architecture of a kinetochore-microtubule attachment site.....	4
Figure 1.2 The Chromosomal Passenger complex (CPC).....	8
Figure 1.3 Recruitment of the CPC to the inner centromere.....	11
Figure 1.4 Cohesin throughout the cell cycle.....	15
Figure 1.5 Evolutionary mechanism for neocentromere formation.....	18
Figure 1.6 Variable centromere positions between different horse individuals.....	19
Figure 1.7 Transposable elements.....	21
Figure 3.1 Expression and solubility of recombinant horse CENP-A.....	49
Figure 3.2 Nickel affinity purified CENP-A fractions.....	50
Figure 3.3 CENP-A sera characterization.....	52
Figure 3.4 Examination of the efficiency of Sera purification using CNBr activated Sephacrose.....	54
Figure 3.5 Purified Sera characterization.....	56
Figure 3.6 ChIP qPCR analysis of CENP-A immunoprecipitation.....	57
Figure 3.7 Horse CENP-A ChIPSeq.....	59
Figure 3.8 Centromere profile across the 16 donkey unique sequence centromeres.....	61
Figure 3.9 CENP-A ChIPSeq Profile comparison in immortalized and primary donkey fibroblasts.....	63
Figure 3.10 CENP-A ChIPSeq comparison Eca11/Eas13 and Eca13/Eas14.....	64
Figure 3.11 FRIP scores for CENP-A ChIPSeq in primary and immortalized cell lines .....	65
Figure 3.12 Correlation of CENP-A ChIPSeq from primary and immortalized donkey fibroblasts.....	66
Figure 3.13 Relative abundance of CENP-A in the immortalised cell line.....	68
Figure 3.14 Relative abundance of CENP-A in primary donkey fibroblasts.....	68
Figure 4.1 EAS 4 centromere donkey comparison.....	74
Figure 4.2 EAS 4 centromere donkey versus horse comparison.....	78
Figure 4.3 EAS 5 centromere donkey comparison.....	80
Figure 4.4 EAS 5 centromere donkey versus horse comparison.....	83
Figure 4.5 EAS7 centromere donkey comparison.....	85
Figure 4.6 EAS 7 centromere donkey versus horse comparison.....	88
Figure 4.7 EAS 8 centromere donkey comparison.....	90
Figure 4.8 EAS 8 centromere donkey versus horse comparison.....	93
Figure 4.9 EAS 9 centromere donkey comparison.....	95
Figure 4.10 EAS9 centromere donkey versus horse comparison.....	98
Figure 4.11 EAS10 centromere donkey comparison.....	100
Figure 4.12 EAS10 centromere donkey versus horse comparison.....	103
Figure 4.13 EAS11 centromere donkey comparison.....	105
Figure 4.14 EAS11 centromere donkey versus horse comparison.....	108
Figure 4.15 EAS12 centromere donkey comparison.....	110
Figure 4.16 EAS12 centromere donkey versus horse comparison.....	113
Figure 4.17 EAS13 centromere donkey comparison.....	115
Figure 4.18 EAS13 centromere donkey versus horse comparison.....	118
Figure 4.19 EAS14 centromere donkey comparison.....	120
Figure 4.20 EAS14 centromere donkey versus horse comparison.....	123
Figure 4.21 EAS16 centromere donkey comparison.....	125
Figure 4.22 EAS16 centromere donkey versus horse comparison.....	128
Figure 4.23 EAS18 centromere donkey comparison.....	130
Figure 4.24 EAS18 centromere donkey versus horse comparison.....	133

Figure 4.25 EAS19 centromere donkey comparison.....	135
Figure 4.26 EAS19 centromere donkey versus horse comparison.....	138
Figure 4.27 EAS27 centromere donkey comparison.....	140
Figure 4.28 EAS27 centromere donkey versus horse comparison.....	143
Figure 4.29 EAS30 centromere donkey comparison.....	145
Figure 4.30 EAS30 centromere donkey versus horse comparison.....	148
Figure 4.31 EASX centromere donkey comparison.....	150
Figure 4.32 EAS X centromere donkey versus horse comparison.....	153
Figure 4.33 Sequence analysis of repetitive elements across domains that donkey CENP-A reads mapped to.....	159
Figure 4.34 Sequence analysis of CENP-A mapped domains compared to the whole genome.....	160
Figure 5.1 (A) <i>Centromere Organisation</i> .....	170
(B). <i>Centromeric higher order chromatin structure</i> .....	170
Figure 5.2 Alignment of protein sequences antibody's were raised against to the horse homolog.....	172
Figure 5.3 Smc1 characterization in donkey fibroblasts.....	174
Figure 5.4 Aurora B characterization in donkey fibroblasts.....	175
Figure 5.5 Survivin characterization in donkey fibroblasts.....	177
Figure 5.6 Borealin characterization in donkey fibroblasts.....	178
Figure 5.7 Population distribution of asynchronous versus Nocodazole arrested donkey fibroblasts.....	181
Figure 5.8 Flow cytometric analysis of population distributions in drug treated donkey fibroblasts.....	183
Figure 5.9 Survivin staining on metaphase chromosomes following treatment with BI3536 and Taxol.....	185
Figure 5.10 Flow cytometric analysis of serum starved (1% serum) and released donkey fibroblasts with propidium iodide DNA staining.....	188
Figure 5.11 Graphical representation of population distribution after serum starvation and release.....	189
Figure 5.12 Double Thymidine block analysis.....	191
Figure 5.13 Graphical representation of population distribution after thymidine block, release and second thymidine block.....	192
Figure 5.14 Proportion of cells distributed throughout the cell cycle in an asynchronous population and following release from a single thymidine block (CENP-A ChIPSeq).....	194
Figure 5.15 Mitotically enriched CENP-A and asynchronous CENP-A profile comparison.....	196
Figure 5.16 Correlation analysis of mitotically enriched CENPA ChIPSeq and an asynchronous ChIPSeq.....	198
Figure 5.17 Proportion of cells distributed throughout the cell cycle in an asynchronous population and following release from a single thymidine block (Smc1 ChIPSeq).....	199
Figure 5.18 5 Mb window view of Smc1 ChIPSeq compared with mitotic CENP-A ChIPSeq.....	201
Figure 5.19 Motif logos associated with regions of Smc1 binding.....	202
Figure 5.20 Distribution of cells in an asynchronous population and the mitotically enriched population used in Aurora B and Survivin ChIPSeq.....	205

Figure 5.21 5 Mb window view of CPC ChIPSeq compared with mitotic CENP-A ChIPSeq.....	207
Figure 6.1 Neocentromere formation in the equids.....	214

## List of tables

Table 2.1 Gateway Plasmids.....	25
Table 2.2 qPCR primers.....	25
Table 2.3 Primary Antibodies.....	26
Table 2.4 Secondary antibodies.....	27
Table 2.5 Gateway cloning LR reaction.....	30
Table 2.6 qPCR conditions.....	31
Table 2.7 Software.....	40
Table 3.1 A280 values and concentration of eluted purified CENP-A antibody.....	54
Table 3.2 Sequence details for Horse CENP-A ChIPSeq.....	58
Table 3.3 Sequence details for donkey CENP-A ChIPSeq.....	60
Table 3.4 Centromere domain size in the primary and immortalized fibroblasts.....	62
Table 3.5 Spearman correlative values for the CENPA binding domain in primary and immortalized donkey fibroblasts.....	66
Table 4.1 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS4.....	75
Table 4.2 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi 933836246 gb JREZ01000325.1  (EAS4) compared with whole genome levels.....	76
Table 4.3 Summary of repetitive elements that span the centromere of EquDonk (EAS4) .....	76
Table 4.4 Summary of sequence variation between EquDonk and EquCab on Eas4.....	79
Table 4.5 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS 5.....	81
Table 4.6 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi 933833362 gb JREZ01000925.1  (EAS5) compared with whole genome levels.....	81
Table 4.7 Summary of repetitive elements that span the centromere of EquDonk (EAS5) .....	82
Table 4.8 Summary of sequence variation between EquDonk and EquCab at the EAS5 centromere.....	84
Table 4.9 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS7.....	86
Table 4.10 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi 933835286 gb JREZ01000511.1 (EAS7) compared with whole genome levels.....	86
Table 4.11 Repetitive elements across the EAS7 centromere EquDonk compared with whole genome levels.....	87
Table 4.12 Summary of sequence variation between EquDonk and EquCab on EAS 7.....	89
Table 4.13 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS8.....	91
Table 4.14 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi 933832210 gb JREZ01001199.1  (EAS8) compared with whole genome levels.....	91
Table 4.15 Summary of repetitive elements across that span the centromere of EquDonk (EAS8) .....	92



Table 4.16 Summary of sequence variation between EquDonk and EquCab at the EAS8 centromere.....	94
Table 4.17 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS9.....	96
Table 4.18 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi 933836078 gb JREZ01000366.1  (EAS9) compared with whole genome levels.....	96
Table 4.19 Repetitive elements across the EAS9 centromere EquDonk compared with whole genome levels.....	97
Table 4.20 Summary of sequence variation between EquDonk and EquCab at the EAS9 centromere.....	99
Table 4.21 Summary of sequence variation between EquDonk and the Guanzhong on EAS10.....	101
Table 4.22 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi 933836905 gb JREZ01000195.1  (EAS10) compared with whole genome levels.....	101
Table 4.23 Repetitive elements across the EAS10 centromere EquDonk compared with whole genome levels.....	102
Table 4.24 Summary of sequence variation between EquDonk and EquCab on EAS10.....	104
Table 4.25 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS11.....	106
Table 4.26 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi 933835325 gb JREZ01000504.1  & gi 933831881 gb JREZ01001282.1  combined (EAS11) compared with whole genome levels. ....	106
Table 4.27 Repetitive elements across the EAS11 centromere EquDonk compared with whole genome levels.....	107
Table 4.28 Summary of sequence variation between EquDonk and EquCab on Eas11.....	109
Table 4.29 Summary of sequence variation between EquDonk and the Guanzhong donkey on Eas12.....	111
Table 4.30 Repetitive elements across the EAS12 centromere EquDonk compared with whole genome levels.....	111
Table 4.31 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi 933837599 gb JREZ01000107.1  & gi 933833617 gb JREZ01000871.1  (EAS12) compared with whole genome levels.....	112
Table 4.32 Summary of sequence variation between EquDonk and EquCab on EAS12.....	114
Table 4.33 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS13.....	116
Table 4.34 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi 933838084 gb JREZ01000066.1  (EAS13) compared with whole genome levels.....	116
Table 4.35 Repetitive elements across the EAS13 centromere EquDonk compared with whole genome levels.....	117

Table 4.36 Summary of sequence variation between EquDonk and EquCab on Eas13.....	119
Table 4.37 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS14.....	121
Table 4.38 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi 933833808 gb JREZ01000828.1  (EAS14) compared with whole genome levels.....	121
Table 4.39 Repetitive elements across the EAS14 centromere EquDonk compared with whole genome levels.....	122
Table 4.40 Summary of sequence variation between EquDonk and EquCab on EAS14.....	124
Table 4.41 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS16.....	126
Table 4.42 Summary of repetitive elements across the EAS16 centromere EquDonk compared with whole genome levels.....	126
Table 4.43 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi 933836929 gb JREZ01000191.1  (EAS16) compared with whole genome levels.....	127
Table 4.44 Summary of sequence variation between EquDonk and EquCab at EAS 16.....	129
Table 4.45 Summary of sequence variation between EquDonk and the Guanzhong donkey at EAS18.....	131
Table 4.46 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi 933835538 gb JREZ01000464.1  (EAS18) compared with whole genome levels.....	131
Table 4.47 Repetitive elements across the EAS18 centromere EquDonk compared with whole genome levels.....	132
Table 4.48 Summary of sequence variation between EquDonk and EquCab on EAS18.....	134
Table 4.49 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS19.....	136
Table 4.50 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi 933831780 gb JREZ01001308.1  (EAS19) compared with whole genome levels.....	136
Table 4.51 Repetitive elements across the EAS19 centromere EquDonk compared with whole genome levels.....	137
Table 4.52 Summary of sequence variation between EquDonk and EquCab on EAS19.....	139
Table 4.53 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS27.....	141
Table 4.54 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi 933836537 gb JREZ01000266.1  (EAS27) compared with whole genome levels.....	141
Table 4.55 Repetitive elements across the EAS27 centromere EquDonk compared with whole genome levels.....	142
Table 4.56 Summary of sequence variation between EquDonk and EquCab on EAS27.....	144

Table 4.57 Summary of sequence variation between EquDonk and the Guanzhong donkey at EAS30.....	146
Table 4.58 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi 933838627 gb JREZ01000029.1  (EAS30) compared with whole genome levels.....	146
Table 4.59 Repetitive elements across the EAS30 centromere EquDonk compared with whole genome levels.....	147
Table 4.60 Summary of sequence variation between EquDonk and EquCab at EAS30.....	149
Table 4.61 Summary of sequence variation between EquDonk and the Guanzhong donkey at EASX.....	151
Table 4.62 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi 933835280 gb JREZ01000512.1  (EASX) compared with whole genome levels.....	151
Table 4.63 Repetitive elements across the EASX centromere EquDonk compared with whole genome levels.....	152
Table 4.64 Summary of sequence variation between EquDonk and EquCab on EASX.....	154
Table 4.65 Sequence analysis of EquDonk centromeres compared to the Guanzhong donkey and horse orthologous regions.....	155
Table 4.66 Sequence analysis of the horse and Guanzhong donkey.....	158
Table 4.67 Summary of differences in the CENP-A mapped domains in EquDonk, the Guanzhong donkey and EquCab.....	165
Table 5.1 Summary details of antibodies used in ChIPSeq experiments.....	172
Table 5.2 qPCR primers flanking the CENP-A binding domain of chromosome 30.....	179
Table 5.3 Distribution of cells throughout the cell cycle.....	181
Table 5.4 Pharmacological agents employed to achieve arrest in the donkey fibroblasts.....	182
Table 5.5 Percentage distribution of cells throughout the cycle after treatment with pharmacological agents.....	184
Table 5.6 Distribution of cells in an asynchronous and in serum starved and release populations.....	189
Table 5.7 Distribution of cells throughout the cell cycle in asynchronous and thymidine treated cells.....	192
Table 5.8 Distribution of cells in an asynchronous population and the mitotically enriched population used in CENP-A ChIPSeq.....	194
Table 5.9 Mitotically enriched CENP-A Sequence details.....	194
Table 5.10 Mitotically enriched CENP-A reads dropped after trimming.....	194
Table 5.11 Spearman correlative values for the CENPA binding domain of the mitotically enriched CENPA ChIP and the asynchronous CENPA ChIP.....	197
Table 5.12 Distribution of cells in an asynchronous population and the mitotically enriched population used in Smc1 ChIPSeq.....	199
Table 5.13 Mitotically enriched Smc1 Sequence details.....	200
Table 5.14 Mitotically enriched Smc1 reads dropped after trimming.....	200
Table 5.15 Sequences identified from the meme output (Figure 5.19 A) present at 5MB windows containing the unique sequence donkey	

centromeres.....	204
Table 5.16 Sequences identified from the meme output present at 5Mb windows containing the unique sequence donkey centromeres.....	204
Table 5.17 Distribution of cells in asynchronous and mitotically enriched populations.....	205
Table 5.18 Details of reads obtained for ChIPSeq of the CPC subunits.....	206
Table 5.19 Mitotically enriched CPC reads dropped after trimming.....	206

## List of commands

Command 2.1 Build reference genome.....	41
Command 2.2 Alignment of paired end reads.....	41
Command 2.3 Convert SAM to BAM.....	42
Command 2.4 Sorting of BAM file.....	42
Command 2.5 Indexing of BAM file.....	42
Command 2.6 Normalization using Deeptools. ....	42
Command 2.7 Plotting bedgraphs in R.....	43
Command 2.8 MACS peak calling.....	43
Command 2.9 SAMtools mpileup read extraction.....	43
Command 2.10 Extracting columns from a file.....	44
Command 2.11 Converting a SAM file to a fastq file.....	44
Command 2.12 Sequence extraction.....	44
Command 2.13 Repeatmasker command.....	45
Command 2.14 R command example for generating schematic representation of centromere domains.....	46

## **Acknowledgements**

Firstly I would like to thank my supervisor, Prof. Kevin Sullivan, for giving me the opportunity to undertake my PhD in his research group. His support, guidance and encouragement over the last number of years have been invaluable. I would also like to thank our collaborator Prof. Elena Giulotto and members of her lab in Pavia, who welcomed me into their group and made me feel very at home. I would also like to thank everyone in the Center for Chromosome Biology for providing such a great environment to work in. Its all hands on deck and be it PhD student or PI, everyone tries to help in anyway they can. I'm going to miss working with ye!

I would like to thank my two favourite Donegalers, Joe and Micheal for all the craic and nights out. Ye were really the light at the end of the tunnel when the lab work and bioinformatics wasn't going so great. I'd like to thank the girls Caroline, Rebecca, Roisin and Devon for their support, encouragement and wine nights.

Last but not least id like to thank my family, my parents Margaret and John and my sisters Karen, Katie and Meabh. I have no doubt that without you're love and encouragement I would not be where I am today. Thank you for listening to me lament about the PhD and for reassuring me that one-day I'd get there.

## Abbreviations

APC	Anaphase promoting complex
BAM	Binary sequence alignment
BIR	Baculoviral IAP repeat
Bp	Base pair
BSA	Bovine serum albumin
C-terminus	Carboxy terminus
CAD	CENP-A distal
CATD	CENP-A targeting domain
CCAN	Constitutive Centromere Associated Network
Cdk	Cyclin-dependent kinase
CENP	Centromere protein
ChIP	Chromatin Immunoprecipitation
ChIPSeq	Chromatin Immunoprecipitation sequencing
CNBr	Cyanogen bromide
CPC	Chromosomal Passenger Complex
CR	Centromeric retrotransposon
CREST	Calcinosis, Raynaud's phenomenon, Esophageal dysmotility, Sclerodactyly and Telangiectasia
CTCF	CCCTC-binding factor
DAPI	4' , 6-diamidino-2-phenylindole
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
EAS	Equus asinus
ECA	Equus caballus
ECI	Enhanced Chemiluminescence
EDTA	Ethylenediaminetetraacetic acid
EGS	Ethylene glycol bis(succinimidyl succinate)
FBS	Fetal bovine serum
FRIP	Fraction of reads in peaks
HJURP	Holliday junction recognition protein
HP1	Heterochromatin Protein 1
HRP	Horseshoe peroxidase
hTERT	human telomerase reverse transcriptase
IAP	Inhibitor of Apoptosis
ICEN	Interphase centromere complex
IgG	Immunoglobulin G
IGV	Integrative genomics viewer
INCENP	Inner centromere protein
ITR	Inverted terminal repeat
Kb	Kilobase
kDa	Kilodaltons
LINE	Long Interspersed Nuclear Element
LTR	Long terminal repeat
MACs	Model-based Analysis of ChIP-Seq
Mb	Megabase
mRNA	Messenger ribonucleic acid
N-terminus	Amino terminus
NAC	Nucleosome Associated Complex

Ni-NTA	Nickle-Nitrilotriacetic acid
NP-40	Nonidet-P40
ORF	Open reading frame
PBS	Phosphate-buffered saline
PBS-TX	Phosphate-buffered saline-Triton X-100
Pds5	Precocious dissociation of sister's protein 5
PP2A	Serine/Threonine phosphatase 2A
PVDF	Polyvinylidene Fluoride membrane
qPCR	Real-time polymerase chain reaction
RBS	Ribosomal binding site
SAC	Spindle assembly checkpoint
SAM	Sequence alignment/Map format
Scc	Sister chromatid cohesion protein
SD	Standard deviation
SDS	Sodium dodecyl sulphate
Sgo	Shugoshin
SINE	Short Interspersed Nuclear Element
Smc	Structural maintenance of chromosome
SNP	Single nucleotide polymorphism
TAE	Tris acetate EDTA
TAP	Tandem affinity purification
TBS	Tris-buffered saline
TPRT	Target primed reverse transcription
Wap1	Wings apart like protein 1



## **Abstract**

The centromere is a genetic locus present once per chromosome that specifies the site of kinetochore formation and is vital for chromosomal segregation. With the exception of *Saccharomyces cerevisiae*, whose ‘point’ centromeres are a defined 125bp sequence and Trypanosomes, most other eukaryotic centromeres are determined epigenetically. Eukaryotic centromeres are typically associated with highly repetitive, tandem repeats of alpha satellite DNA, varying greatly in span. Reports of instances of neocentromere formation, whereby the centromere has moved to a new non repetitive region of the chromosome has been reported in humans, equids, primates, birds and rice supporting the epigenetic status of centromere identity, independent of DNA sequence. The common feature shared by almost all centromeres is the presence of the “epigenetic placeholder” histone H3 variant CENP-A. Centromeres and pericentric heterochromatin have been shown to contain an abundance of transposable elements in a number of phylogenetic species. Transposable elements, such as Long Interspersed Nuclear Elements (LINEs) have been implicated in the recruitment of CENP-A to new genomic loci forming neocentromeres. Transposons play a large role in shaping the genome, from maize to humans, these ‘jumping genes’ have been shown to play a role in gene regulation and genomic evolution. In this thesis we utilize the *Equus asinus*, which contains 16 naturally occurring unique sequence centromeres to gain insight into centromere dynamics. We identify inter-individual and interspecies sequence anomalies associated with these unique sequence centromeres as well as start the process of identifying the location of the inner centromere at the linear one-dimensional primary sequence level.

## **Chapter1- Introduction**

### **1.1 Centromere identity**

The centromere was first described by Walter Flemming in the 1800s, as the primary constriction of the chromosome (Flemming, 1882). The centromere is a genetic locus present once per chromosome that specifies the site of kinetochore formation and is vital for chromosomal segregation. Centromeres are defined by a 125bp sequence in the case of *Saccharomyces cerevisiae* (Fitzgerald-Hayes, Clarke, & Carbon, 1982; Steven Henikoff & Henikoff, 2012) but in the vast majority of other eukaryotes the centromere is determined epigenetically. Trypanosomes are an exception, these unicellular protozoa lack CENP-A as well as the majority of kinetochore forming proteins (Echeverry, Bot, Obado, Taylor, & Kelly, 2012). Generally, centromeres are associated with highly repetitive tandemly repeated sequences, that make dissecting the centromeric sequences very cumbersome, even with the most modern techniques (Steven Henikoff, 2001). Centromere associated sequences are highly divergent even within closely related species (Lamb & Birchler, 2003). In conjunction with this, the movement of a centromere temporarily or stably to non repetitive regions (Craig, Wong, Lo, Earle, & Choo, 2003; Wade et al., 2009) has solidified the argument that centromere location is determined independent of DNA sequence. Centromere location is ultimately defined by the incorporation of the histone H3 variant CENP-A (De Rop, Padeganeh, & Maddox, 2012; Hooser et al., 2001; Palmer, O'Day, Wener, Andrews, & Margolis, 1987).

### **1.2 CENP-A**

CENP-A was first identified in 1985 using autoimmune sera from patients suffering from scleroderma CREST syndrome (Moroi, Peebles, Fritzler, Steigerwald, & Tan, 1980). Immunoblotting and immunostaining techniques established that three centromeric proteins: CENP-A, CENP-B and CENP-C were recognized by sera from these patients (Earnshaw & Rothfield, 1985). Subsequent nuclei extractions and micrococcal nuclease digestions determined CENP-A to be a histone associated with nucleosome particles (Palmer et al., 1987). CENP-A is a histone H3 variant with a divergent amino terminus and a relatively conserved C-term histone fold domain (HFD) sharing 60% homology to histone H3 which is responsible for CENP-A

targeting to the centromere (Kevin F. Sullivan, Mirko Hechenberger, 1994). Within the HFD, the CENP-A targeting domain (CATD) which is comprised of the loop 1 and  $\alpha 2$  helix was found to be sufficient for centromere targeting (Black et al., 2007; Shelby, Vafa, & Sullivan, 1997). A chimeric H3<sup>CATD</sup> protein, whereby the CATD was substituted into H3 was capable of assembly on centromeric chromatin as well as adopting a more compact conformation with H4 compared to CENP-A: H4 (Black et al., 2007). While the CENP-A CATD is sufficient for centromere targeting, centromere maintenance and CCAN (Constitutive Centromere Associated Network) recruitment requires both CENP-A N- and C- term (Folco et al., 2015; Logsdon et al., 2015).

Centromere inheritance is critical for transmission of the genome and requires CENP-A nucleosomes to maintain the epigenetic mark. CENP-A synthesis is not coupled with DNA replication and takes place in G1 (Jansen, Black, Foltz, & Cleveland, 2007; Shelby, Monier, & Sullivan, 2000). At mitosis the CENP-A occupancy at the centromere is half the full complement, following dilution in S phase and the incorporation of H3.3 'placeholders' (Dunleavy, Almouzni, & Karpen, 2011). These 'placeholders' are replaced with CENP-A in the subsequent G1 cycle. CENP-A assembly requires Holliday junction recognition protein (HJURP), a dedicated histone chaperone that contains an N terminal CENP-A binding domain (Shuaib, Ouararhni, Dimitrov, & Hamiche, 2010) and the Mis18 complex.

### **1.3 CCAN**

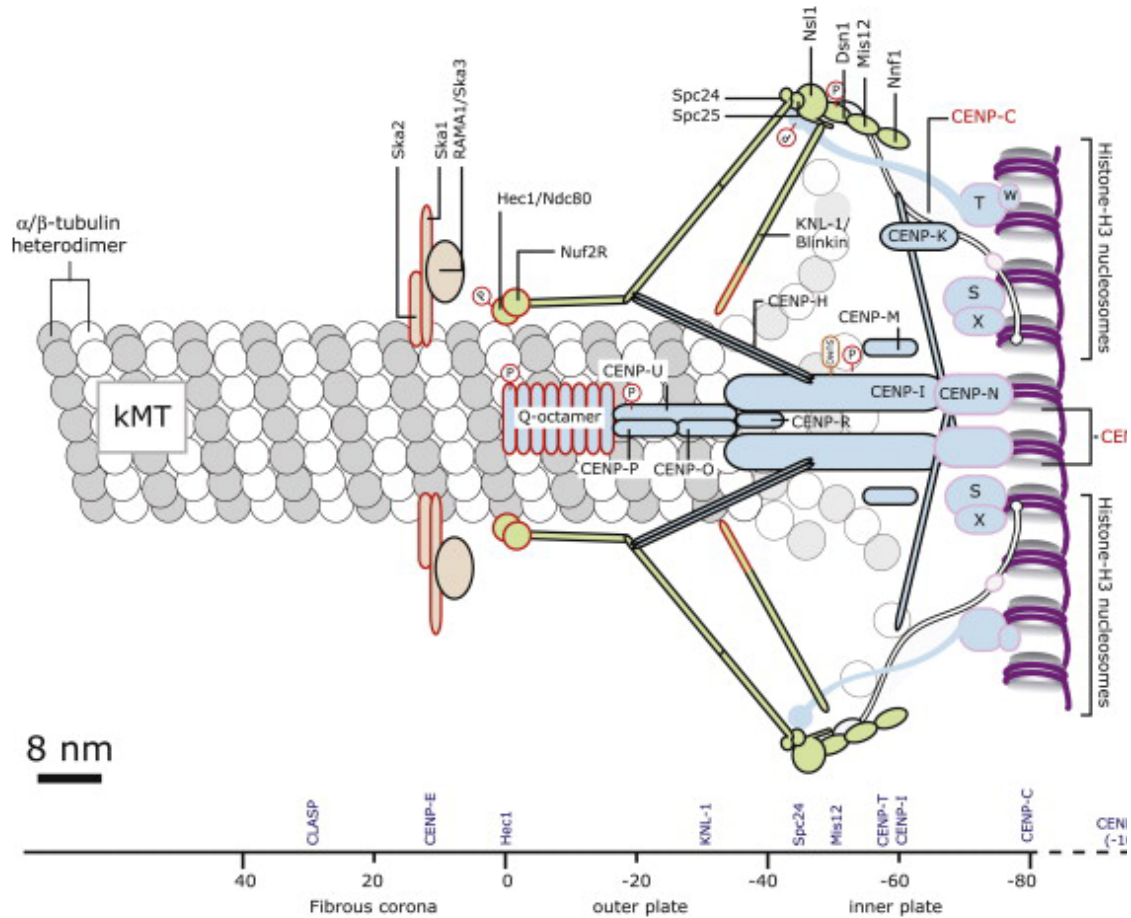
The CCAN was originally identified as proteins that affinity purified with CENP-A containing centromeric nucleosome. Initially, the CCAN was named CENP-A Nucleosome Associated Complex (CENP-A NAC) comprised of CENPs -U, -C, -H, -M, -N, -T and CENP-A Distal (CAD) comprised of CENP's -K, -L, -O, -P, -Q, -R and -S which assembled on the CENP-A NAC, following identification by tandem affinity purification (TAP) in HeLa cells (Foltz et al., 2006). Many of the CCAN subunits were identified by isolation of the interphase centromere complex (ICEN) from HeLa cells (Izuta et al., 2006). Proteomic analysis of proteins pulled down in a CENP-A native CHIP, identified 40 ICEN associated proteins, seven with previously unknown function. siRNA knockdowns showed their importance in chromosome segregation (Izuta et al., 2006). The inner kinetochore components CENPH/I, were

identified as a multisubunit complex. FLAG/GFP tagged CENP-H and CENP-I were generated in DT40 cell lines, completely replacing the endogenous protein (Okada et al., 2006). Affinity purification of these proteins identified 5 previously unknown centromere proteins CENP-K, -L, -M, -O and -P. Parallel experiments in HeLa cells with LAP tagged CENP-H, -O and -P, identified four more centromere proteins, CENP-N, -Q, -R and -U. Combined, purifications from human and chicken lines identified a complex consisting of 11 interacting proteins, taken together this complex was divided into 3 functional groups based on localization: CENP-O, CENP-M and CENP-N class proteins (Okada et al., 2006).

Following on from proteomic analysis, the relationship of centromere-associated proteins was investigated and it was demonstrated that some kinetochore subunits bound centromeric nucleosomes. CENP-N was shown to bind exclusively to the histone H3 variant CENP-A. Mutations of CENP-N impaired its recruitment and subsequent recruitment of CENP-H, -I and -K. CENP-T and CENP-W recruitment were shown to be interdependent as knockdown of either protein abolishes recruitment of the other (Hori et al., 2008). CENP-T/W were shown to bind H3 nucleosomes in the vicinity of CENP-A, but not CENP-A. CENP-C is also a putative DNA binding protein that associates with histone H3 (Hori et al., 2008). CENP-C interaction with H3 and CENP-A provides a platform for connecting the centromere to outer kinetochore. CENP-N, -H, -I have also been shown to play a role in the incorporation of CENP-A in centromeric chromatin (Carroll, Silva, Godek, Jansen, & Straight, 2009; Okada et al., 2006).

Once assembled on the centromere, the CCAN acts as the platform for assembly of the outer kinetochore Figure 1.1. The KMN network, comprised of KNL1, the Mis12 complex and the Ndc80 complex, is essential for kinetochore microtubule interactions. The N terminus of CENP-C is associated with Nnf1, a subunit of the Mis12 complex, while the C terminus is associated with CENP-A. The N terminus of CENP-T is associated with the Ndc80 complex, while the C terminus, in a complex with CENP-W, is associated with H3 (Tanaka, 2013). Both CENP-T and CENP-W C-termini contain a histone fold domain and form a heterotetramer with CENP-S and -X, that binds centromeric DNA (Takeuchi et al., 2014). Both CENP-C and CENP-T serve as linkers physically connecting the centromere to the outer kinetochore. Over ~100 proteins that associate with the centromere and kinetochore have been identified

(Ohta et al., 2010) and proteomic analysis is still being carried in order to establish the relationship of these proteins (Samejima et al., 2015).



**Figure 1.1 Protein architecture of a kinetochore-microtubule attachment site.** KMN network is shown in green and the Ska complex in orange, forming the microtubule binding interface. The CCAN is shown in blue, with those outlined in pink binding DNA or histones. CENP-C in white, links CENP-A to the outer kinetochore. Proteins are drawn to scale based on their molecular weights (McAinsh & Meraldi, 2011).

#### 1.4 Chromosomal Passenger Complex (CPC)

Successful cell division depends on the accurate segregation of sister chromatids to daughter cells. These events are regulated by the competing actions of phosphatases and protein kinases. The CPC is comprised of INCENP, Aurora B, Survivin and Borealin (Figure 1.2) and together these proteins shift to different locations during the cell cycle where they mediate critical mitotic events such as activation of the spindle assembly checkpoint, correction of chromosome-microtubule attachment errors and construction and regulation of the apparatus that drives cytokinesis.

#### **1.4.1 INCENP**

INCENP (INner CENtromere Protein) was first discovered in a screen of monoclonal antibodies that were raised against mitotic chromosome scaffold proteins fractionated from chicken cell extracts (Cooke, Heck, & Earnshaw, 1987). Since then INCENP homologs have been found in many organisms from yeast to human. INCENP is comprised of two functional domains, the conserved IN box C-terminal domain that binds and activates Aurora B (Bishop & Schuniacher, 2002) and the N-terminal which forms a three-helix bundle with Borealin N-terminal and Survivin C-terminal as well as playing a critical role in centromere targeting (Xu et al., 2009). CPC localization in interphase is also mediated by INCENP binding HP1 (Ainsztein, Kandels-Lewis, Mackay, & Earnshaw, 1998; Kang et al., 2011).

Aurora B and Cdk1, the cyclin which controls mitotic entry and exit, regulates INCENP. In budding yeast, the INCENP homolog Sli15 is phosphorylated by the Aurora B homolog IpI-1 and Cdk, halting spindle midzone association before the onset of anaphase (Nakajima et al., 2011). In yeast Sli15 plays a critical role in the Spindle Assembly Checkpoint (SAC), a mechanisms that delays the onset of anaphase until all chromosomes are aligned and attached to the mitotic spindle, Sli15 is phosphorylated by Cdk at multiple sites triggering SAC activation (Mirchenko & Uhlmann, 2010). Anaphase onset triggers the dephosphorylation of these sites and prevents reactivation of the SAC following sister chromatid separation (Mirchenko & Uhlmann, 2010).

#### **1.4.2 Aurora B**

Aurora B is a member of a highly conserved family of Serine-Threonine kinases, first discovered in *Drosophila* while screening for mutants with defective spindle poles (Glover, Leibowitz, Mclean, & Parry, 2014). This family has three members; Aurora A, associated with spindle poles and plays a role in mitotic entry, spindle assembly and centrosome function, Aurora B, located in the inner centromere at the start of mitosis and then traverses to the spindle midzone, equatorial cortex and midbody. Aurora B functions in kinetochore-microtubule attachment, cohesion, spindle checkpoint and cytokinesis (Carmena, Ruchaud, Earnshaw, Building, & Road, 2009), in contrast little is known about Aurora C, however it has been shown to interact with INCENP and share a similar localization pattern to Aurora B as well as complementing Aurora B function (Sasai et al., 2004).

Aurora kinases along with cyclin dependent kinases and polo like kinases orchestrate the overall cell cycle progression (Carmena et al., 2009). Aurora B activation is a two-step process involving INCENP binding. Initially Aurora B binds the IN box of INCENP, triggering low kinase activity, subsequent Aurora B phosphorylation of the INCENP C-terminal TSS motif (threonine-serine-serine) as well as autophosphorylation at threonine 232 results in the fully activated kinase (Sessa et al., 2005). Aurora B activity is also directly regulated by two other kinases, Chk1 and tousel-like kinase (TLK1) (Carmena et al., 2009). During mitosis Chk1 phosphorylates Aurora B at serine 331, this phosphorylation is essential for optimal INCENP TSS phosphorylation and Survivin association, following phosphorylation, Aurora B then subsequently translocates to kinetochores in prometaphase cells (Petsalaki, Akoumianaki, Black, Gillespie, & Zachos, 2011). Aurora B activation also requires PLK1 phosphorylation of survivin (Chu et al., 2011). In the case of *C. elegans* TLK1 phosphorylates Aurora B in prophase/prometaphase, increasing kinase activity in an INCENP dependent manner (Han, Riefler, Saam, Mango, & Schumacher, 2005).

In prophase, Aurora A and Aurora B are shown to cooperate leading to inner centromeric accumulation of Aurora B by Aurora A phosphorylation of CENP-A Ser 7. During late prophase, this Aurora B accumulation plays an important role in maintaining CENP-A Ser7 phosphorylation, ensuring timely execution of cytokinesis (Kunitoku et al., 2003; Zeitlin, Shelby, & Sullivan, 2001).

### ***1.4.3 Survivin***

Survivin was initially characterized as member of the inhibitor of apoptosis family (IAP), that accumulated in G2 and negatively influenced apoptosis in mitosis (Chandele, Prasad, Jagtap, Shukla, & Shastry, 2004). The IAP family traditionally are known to contain from one to three baculoviral IAP repeat (BIR) domains, a RING domain and a caspase recruitment domain (CARD) (Deveraux & Reed, 1999). Survivin is unique since it only contains a single BIR domain repeat and lacks a ring finger and CARD domain (Yue et al., 2008). Survivin overexpression is a recurrent hallmark of many invasive cancers. Survivin makes aberrant cells less susceptible to chemotherapeutic agents and increases the chances of tumor recurrence (Fukuda & Pelus, 2006; Kapellos et al., 2013). In instances of cancer, survivin is not expressed in

a cell cycle dependent manner with survivin detectable in interphase. Recent evidence shows that targeting cytoplasmic survivin to the nucleus leads to its degradation in a *cdh1* dependent manner without abergating mitotic localization as well as increasing cell susceptibility to chemotherapeutics (Connell, Colnaghi, & Wheatley, 2008).

The precise mechanism of how survivin inhibits apoptosis remains to be understood but reports show survivin directly inhibits caspases 3, 7 and 9 (Dohi, Beltrami, Wall, Plescia, & Altieri, 2004; Tamm et al., 1998). Survivin association with Smac and Diablo in the cytosol in interphase has been shown to indirectly inhibit the caspase cascade. Mitochondrial proteins, Smac and Diablo, are IAP antagonists and are released in the presence of apoptotic stimuli, promoting caspase activity (Du, Fang, Li, Li, & Wang, 2000; Verhagen et al., 2000). Survivin seems to serve as a sponge binding Smac and Diablo, preventing their inhibition and allowing cascade interaction and subsequent cell survival.

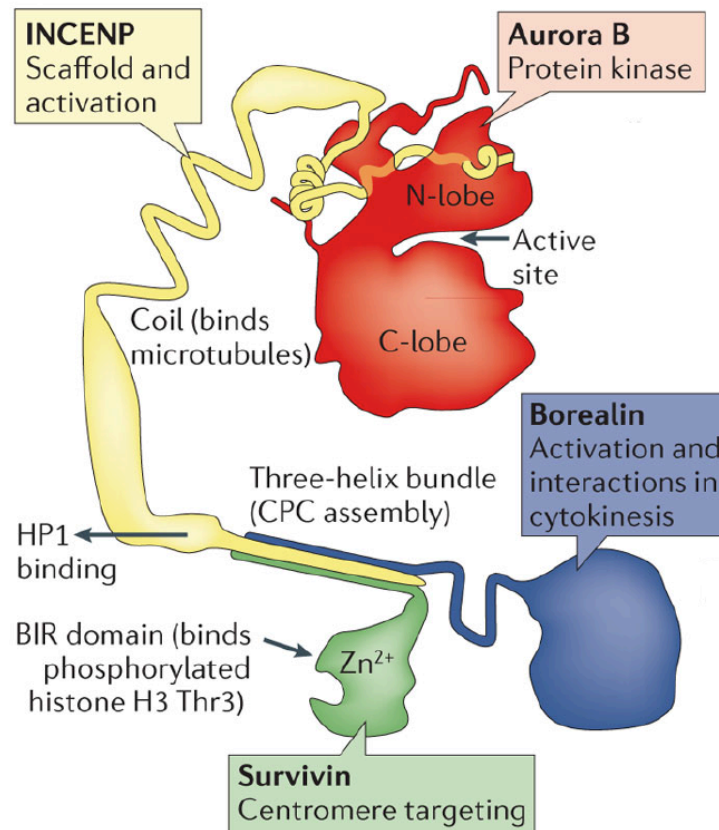
Survivin ubiquitination plays a critical role in CPC centromeric binding and chromosome segregation. The de-ubiquitinating enzyme, hFAM modulates survivin and Aurora B localization to centromeres. Survivin is ubiquitinated through two lysine residues Lys48 and Lys63, hFAM deubiquitination of Lys63 is required for centromeric dissociation of survivin, the ubiquitin binding protein Ufd1 ubiquitinates lys63 and is required for centromeric survivin association (Vong, Cao, Li, Iglesias, & Zheng, 2005).

#### ***1.4.4 Borealin/Dasra B***

Borealin or Dasra B were both discovered in 2004 in two separate studies (Gassmann et al., 2004; Sampath et al., 2004). A second protein Dasra A was also identified in the same study. Similar to other CPC components, Borealin is phosphorylated by different kinases. Borealin targeting to centromeres is mediated by Cdk1 phosphorylation which allows interaction with Shugoshin 1 and 2 (Tsukahara, Tanno, & Watanabe, 2010). Borealin is also phosphorylated by monopolar spindle (Mps1) at Thr230, a modification that modulates dimerization as well influencing Aurora B activity (Jelluma, Dansen, Sliedrecht, Kwiatkowski, & Kops, 2010). The N terminal of Borealin interacts with INCENP and Survivin to form a three helical bundle that serves as the localization platform of the CPC (Jeyaprkash et al., 2007). There is no Borealin ortholog present in yeast, but a Borealin like subunit has been identified. Nbl1p shares the same localization pattern as the CPC and is essential for CPC



loading and accurate chromosome segregation. There is homology between Nblp1 and the N terminal of Borealin (Nakajima et al., 2009). In the case of *C. elegans* there is a distantly related Borealin like subunit CSC-1. Borealin is conserved among vertebrates with a paralogue observed in chicken, *X. laevis*, and zebrafish. Knockdown of borealin by RNAi lead to defects in CPC localization, spindle attachment and cytokinesis (Gassmann et al., 2004).



**Figure 1.2 The Chromosomal Passenger complex (CPC)** The CPC is comprised of Aurora B, INCENP, survivin and Borealin. The functional domains for each subunit are shown (Carmena, Wheelock, Funabiki, & Earnshaw, 2012).

#### 1.4.5 Chromosomal Passenger complex localization

INCENP serves as the main CPC loading platform and together with Borealin and Survivin, modulates the localization and activity of the kinase component Aurora B (Jeyaprakash et al., 2007). The chromosomal passenger complex traverses around the cell in a cell cycle dependent manner, here I will discuss the discrete localization pattern of the complex as shown in Figure 1.3.

##### 1.4.5.1 Interphase

Histone H3 Ser10 is a substrate of Aurora B kinase, this modification is required for accurate chromosome segregation. In interphase cells, this modification is visible by

immunofluorescence, illustrating Aurora B activity (Hayashi-Takanaka, Yamagata, Nozaki, & Kimura, 2009). Heterochromatin protein 1 (HP1) plays a key role in recruitment of CPC to heterochromatin. INCENP associates with HP1 through its PXXVL motif. This motif is not required for centromeric INCENP localisation. Centromeric localisation of the CPC is dependent on Borealin's interaction with HP1 (X. Liu et al., 2014). Reports show that human Mis14, a member of the Mis12 complex, is responsible for the assembly of the KNL network and Bub1 and BubR1 and directly interacts with HP1, ensuring the association of HP1 in the vicinity of the centromere (Kiyomitsu, Iwasaki, Obuse, & Yanagida, 2010).

HP1 localisation is mediated by Aurora B. HP1 is initially recruited to H3K9me3 sites where it regulates functions such as chromatin packaging and gene expression. Following the onset of mitosis HP1 dissociates from this methylation site due to the phosphorylation of H3S10 by Aurora B (Fischle et al., 2005).

#### ***1.4.5.2 Mitosis***

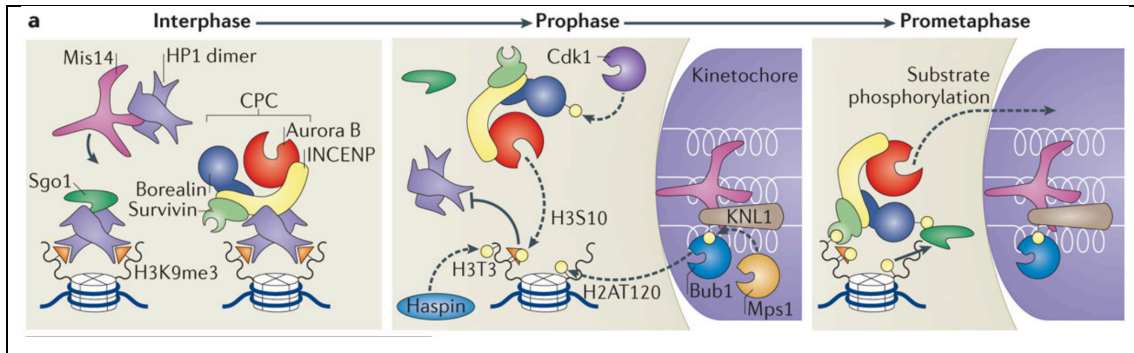
Recruitment of the CPC to the inner centromere by the onset of mitosis is governed by two kinases: Haspin and Bub1 which phosphorylate H3Thr3 and H2AThr120 respectively. Haspin activity is dependent on phosphorylation by Aurora B. Haspin activity is also mediated by the cohesin regulator Pds5. Shugoshin, which protects centromeric cohesin, association with the inner centromere is dependent on the phosphorylation of H2AT120. This Bub1-shugoshin interdependence is also important for concentrating H3T3 at the inner centromere, thereby concentrating the CPC there also (Tanno et al., 2015; F. Wang et al., 2011). Haspin antagonists PP1 $\gamma$ /Repo-Man phosphatase dephosphorylates H3Thr3 at the chromosome arms further concentrating the CPC at the centromere. Centromeric H3Thr3 remains unaffected because of Aurora B phosphorylation of Repo-Man on Ser893, preventing recruitment (Qian, Beullens, Lesage, & Bollen, 2013). Aurora B mediated removal of cohesin from the chromosome arms also serves to concentrate the CPC at the centromere (Dai, Sullivan, & Higgins, 2006).

Centromeric targeting of the CPC is dependent on the amino terminal of INCENP (Ainsztein, Kandels-Lewis, Mackay, & Earnshaw, 1998b). INCENP/Survivin interaction is also through this domain and fusion experiments in the absence of NH2 INCENP domain, show that Survivin is sufficient to target the complex to the

centromere even in the absence of Borealin (Vader, Kauw, Medema, & Lens, 2006). However Borealin has a critical role in the CPC localization in the naturally occurring mechanism, it is only in INCENP-Survivin fusion experiments that its function is dispensable, suggesting Borealin has a role in INCENP-Survivin interaction. Borealin has also been shown to directly bind chromatin making it the likely anchor for CPC centromeric localization (Klein, Nigg, & Gruneberg, 2006). A study carried out in *Xenopus* egg extracts, show that long non coding RNA, transcribed from the centromere, binds the CPC, is important for CPC maintenance at the inner centromere as well playing a role in CPC regulation. Upon transcription inhibition, Aurora B localisation to the inner centromere was decreased by 50%, while centromere intensity of H2AT120 and H3T3 remained unaffected. In transcription inhibited cells, ~50% failed to form bipolar attachments. Transcription, at least in the case of *Xenopus*, plays a critical role in CPC localisation and function (Blower, 2016).

#### ***1.4.5.3 Anaphase onset***

At the onset of anaphase, Aurora B is ubiquitinated, contributing to the active removal of the CPC from the anaphase chromosomes. Inhibition of Repo-Man by Aurora B phosphorylation is reversed by PP2A (Qian et al., 2013). PP1 $\gamma$ /Repo-Man dependent dephosphorylation of H3Thr3 and H2AT120 halts CPC recruitment to the centromere (Qian, Lesage, Beullens, Van Eynde, & Bollen, 2011). As highlighted previously Borealin is phosphorylated by Cdk1, targeting it to H2AT120. At the onset of anaphase Cdk1 levels decrease, this is also likely to play a role in CPC dissociation. Targeting of the CPC to the spindle midzone and subsequently the equatorial cortex requires the interaction of INCENP and Aurora B with MKLP2 (Gruneberg, Neef, Honda, Nigg, & Barr, 2004). Targeting of the CPC to the central spindle requires Cdk1 mediated phosphorylation of INCENP on Thr59 and at multiple residues of MKLP2 (Hümmer & Mayer, 2009). MKLP2 has a critical role in CPC relocalisation since RNAi knockdown prevents CPC accumulation and results in failed cytokinesis (Gruneberg et al., 2004). Full Aurora B kinase activity is also required for translocation to the spindle midzone (Xu et al., 2009).



**Figure 1.3 Recruitment of the CPC to the inner centromere:** The CPC is targeted to heterochromatin in interphase by INCENP interaction with HP1, Mis12 is required for proper HP1 localisation. In prophase, Aurora B (red) phosphorylates H3S10, displacing HP1 from adjacent H3K9me3. Histone tails are phosphorylated by Haspin and Bub1, providing a platform for CPC recruitment to the inner centromere. The BIR domain of survivin binds H3T3 while borealin phosphorylated by Cdk1 binds Sgo1, which then interacts with H2AT120 (Carmena et al., 2012).

## 1.5 Cohesin

Cohesin is a highly conserved multisubunit protein complex which mediates cohesion, keeping sister chromatids together from S phase until the onset of anaphase. Cohesin has also been implicated in gene regulation, DNA repair, chromosome condensation and homolog pairing (Peters, Tedeschi, & Schmitz, 2008). Cohesin consists of six subunits; the structural maintenance of chromosome (Smc) proteins Smc1 and Smc3 (also called the stromal antigen (SA) in animal cells), sister chromatid cohesion protein 1 (Scc1), Scc3, precocious dissociation of sister's protein 5 (Pds5) and wings apart like protein 1 (Wapl). Another protein Sororin, plays a role in stabilization of DNA associated cohesin (Nishiyama, Sykora, Huis in 't Veld, Mechtler, & Peters, 2013). The main body of the complex consists of Smc1, Smc3 and Scc1, which form a tripartite ring that entraps the DNA. The Smc subunits are rod shaped proteins with a hinge on one end and an ATPase 'head' domain on the other, which is bound by Scc1 (Haarhuis, Elbatsh, & Rowland, 2014). The function of Scc3 and Pds5 in cohesin maintenance is poorly understood.

### 1.5.1 Cohesin loading

Cohesin assembly takes place prior to its recruitment (Losada, A. et al 1998), it then opens up allowing the entrapment of DNA. Potentially the DNA could enter the cohesin ring in three ways, through the 'hinge' domain between Smc1 and Smc3 or the 'head' domain between either Smc1 and Scc1 or Smc3 and Scc1. *S. cerevisiae* fusion experiments locking the interfaces between these subunits show that DNA enters the ring complex at the Smc1 and Smc3 interface (Gruber et al., 2006) (Figure 1 B). This

has also shown to be the case in humans (Buheitel & Stemmann, 2013). Cohesin loading is dependent on the heterodimeric Scc2/Scc4 loader complex but how loading is facilitated is unknown. It has been proposed that Smc mediated ATPase activity is facilitated by Scc2/Scc4, causing the hinge domain to open (Arumugam et al., 2003). Conversely, the loader may promote DNA entrapment through chromatin remodelling creating an accessible template for cohesin loading (Kogut, Wang, Guacci, Mistry, & Megee, 2009). Conserved Scc2 orthologues have been identified in multiple organisms, fission yeast (Mis14), *Drosophila* (Nipped-B), frogs (XSc2) and human (Nipped-B like, NIPBL) (Kogut et al., 2009). Scc4 orthologues have also been identified, fission yeast (Ssl3) (Bernard et al., 2006) and metazoans (Seitan et al., 2006).

The timing of cohesin loading differs between animal cells and yeast. In animal cells, cohesin loading occurs in telophase whereas in yeast, due to extended separate activity, recruitment occurs in G1 (Haarhuis et al., 2014). ChIP microarray analysis of cohesin distribution in budding yeast show predominant enrichment between convergent RNA polymerase II transcribed genes (Glynn et al., 2004). Further ChIP analysis revealed a preference for AT rich binding sites, every 10-15kb along the chromosome arms which correlates with intergenic regions (Blat & Kleckner, 1999). ChIP analysis of the cohesin loader Scc2-Scc4, show no particular overlap with cohesin binding sites, rather an overlap with the promotor regions of strongly expressed genes. Cohesin is loaded and appears to slide along the chromatin in an RNA pol II transcriptionally dependent manner (Lengronne et al., 2004). In contrast, the cohesin loader subunit Nipped B, in *Drosophila*, has an almost identical distribution to that of cohesin. Here, Nipped B is associated with actively transcribed RNA polymerase II regions (Misulovin et al., 2008). In mammalian cells, the cohesin loader NIPBL is found at the promotor regions of expressed genes, with cohesin also detectable at these sites. In the case of humans, the bulk of cohesin is colocalises with the CCCTC-binding factor (CTCF). The CCCTC-binding factor (CTCF) is a highly conserved DNA binding protein, involved in numerous diverse cellular functions: gene activation, gene repression, the maintenance of genomic imprinting, chromatin insulator function and X chromosome inactivation (Filippova, 2007; Ohlsson, Renkawitz, & Lobanenkov, 2001). CTCF acts as an insulator protein that blocks enhancer-promoter interactions. Cohesin is required for CTCF's insulator activity,

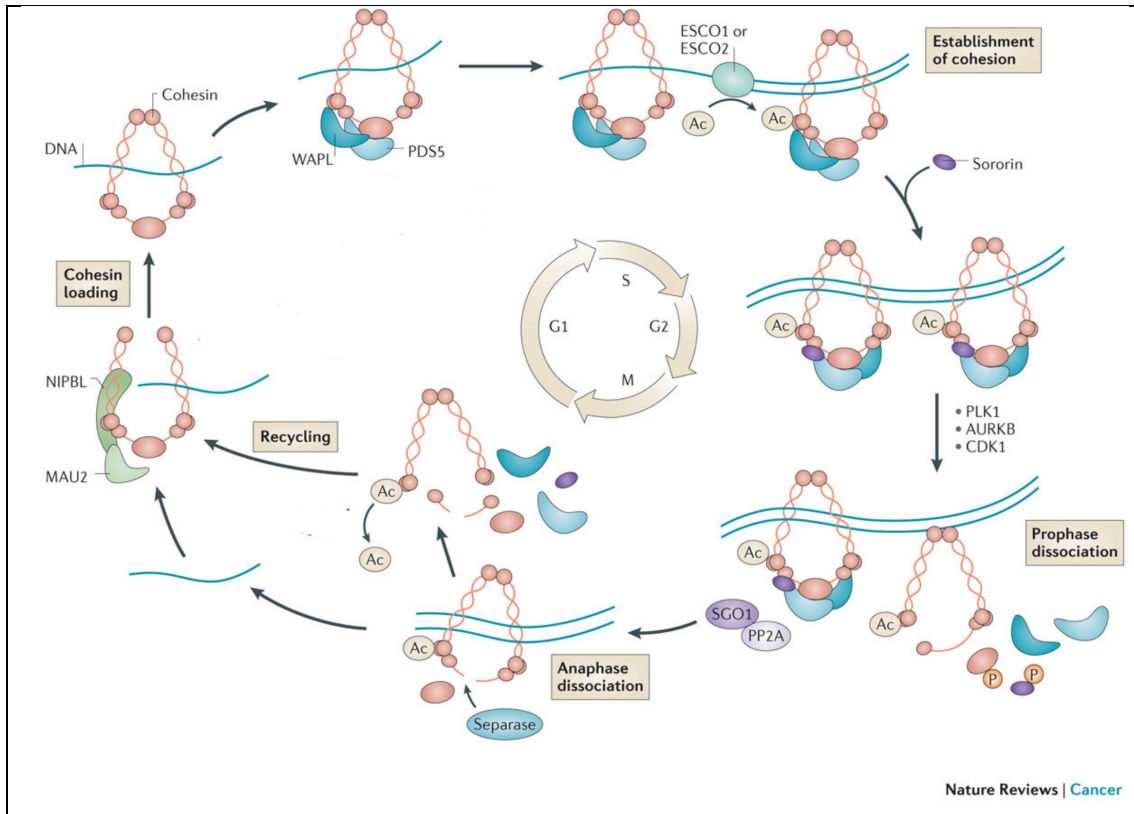
since cohesin or CTCF knockdowns interfere with the enhancer blocking activity of CTCF resulting in both up and downregulation of genes as reported by luciferase reporter assays (Wendt et al., 2008).

### *1.5.2 Cohesin removal*

In vertebrates, cohesin is removed from sister chromatids by two distinct mechanisms, the prophase pathway and separase cleavage (Waizenegger, Hauf, Meinke, & Peters, 2000). This two step process is unique to metazoans since yeast utilize only separase dependent cleavage of cohesin. Following the onset of mitosis, removal of cohesin from chromosome arms is mediated by the mitotic kinases PLK1 (Silke Hauf et al., 2005), Aurora B and Cdk1 (Nishiyama et al., 2013). PLK1 phosphorylates Scc1 and Scc3 (Silke Hauf et al., 2005), while Aurora B and Cdk1 phosphorylates Sororin destabilising the interaction with Pds 5 (Nishiyama et al., 2013), leading to the release of cohesin from chromosome arms. The prophase pathway, is mediated by the cohesin antagonist Wapl, which binds Pds5 (Kueng et al., 2006) causing the dissociation of cohesin from chromosome arms through the Smc3-Scc1 exit gate (Buheitel & Stemmann, 2013) most likely by regulating ATPase activity. Centromeric cohesin remains intact and is protected by shugoshin. Shugoshin, meaning guardian spirit, protects centromeric cohesin in two ways, the first involves Sgo1 binding the serine/threonine phosphatase 2A (PP2A), which counteracts Sororin and Scc3 phosphorylation (Kitajima et al., 2006), secondly shugoshin competes with Wapl by binding directly to cohesin (Hara et al., 2014). Centromeric cohesin is crucial for the cohesion that holds sister chromatids together until the satisfaction of the spindle assembly checkpoint (SAC). Sgo1 recruitment is dependent on two distinct phosphorylations, H2A by Bub1 which recruits Sgo1 to the centromere (Kawashima, Yamagishi, Honda, Ishiguro, & Watanabe, 2010) and the Cdk1 dependent phosphorylation of Sgo1 which allows it to bind the chromosome arms (H. Liu, Rankin, & Yu, 2013). SAC is composed of two groups of proteins mitotic arrest deficient proteins (Mad) and budding uninhibited by benzimidazole (Bub) (Zhou, Yao, & Joshi, 2002). The SAC senses microtubule binding errors and inhibits the anaphase promoting complex or cyclosome (APC/C), delaying anaphase. APC activation requires the binding of cdc20 (Fang, Yu, & Kirschner, 1998), which MAD 2, 3 and BubR1 subunits bind to and block APC association. Bub1 phosphorylates cdc20 and catalytically inhibits APC, thus by inhibiting APC/C<sup>cdc20</sup> association centromeric

cohesin remain intact until amphitelic attachment (Tang, Sun, Harley, Zou, & Yu, 2004). Once the spindle assembly checkpoint is satisfied, the APC ubiquitinates securin, an inhibitor of separase causing the cleavage of Scc1 and sister chromatid separation (S Hauf, Waizenegger, & Peters, 2001).

This two step cohesin removal pathway is important for three key events: Sister DNA decatenation, correction of erroneous microtubule-kinetochore attachment and preservation of cohesin rings for reloading. Catenations are a normal consequence of DNA replication and can function as a kind of cohesion. DNA topoisomerase allevates these catenations (Nitiss, 2009), however cohesin has been shown to play a role in maintenance and removal of catenations (L. H.-C. Wang, Mayer, Stemmann, & Nigg, 2010). Cohesin's role in the correction of erroneous microtubule-kinetochore attachment lies in the recruitment of the CPC by two mechanisms. The first is dependent on bub1 phosphorylation of H2AT120 which recruits Sgo1 to the centromere, cdk1 then phosphorylates Borealin which binds the coiled coil region of Sgo1 (Tsukahara et al., 2010). The second involves phosphorylation of H3T3 by Haspin which is mediated by the Pds5 subunit, this creates a platform for Survivin binding (Yamagishi, Honda, Tanno, & Watanabe, 2010). The two step cohesin removal mechanism allows for the preservation of cohesin for reloading in subsequent cycles. After the prophase pathway the Scc1 subunit is still intact and can be reloaded onto DNA, separase cleaved cohesin however must bind to an uncleaved Scc1 subunit before DNA loading can occur. The recycling of cohesin is important for timely reloading in the subsequent cell cycle (Haarhuis et al., 2014).



**Figure 1.4 Cohesin throughout the cell cycle.** Cohesin loading takes place in G1 and is aided by the NIPBL/MAU2 heterodimer. The Smc1 and Smc3 hinge domains dissociate allowing the entry of DNA. During DNA replication, SMC3 is acetylated by ESCO1 and ESCO2 and sororin is recruited displacing WAPL. In prophase, Plk1 phosphorylates the SA subunit while aurora B and Cdk1 phosphorylates sororin, causing cohesin dissociation from chromosome arms. Sgo1 and protein phosphatase PP2A accumulate at the centromere, inhibiting this phosphorylation and thereby protecting centromeric cohesin until anaphase. Rad21 is cleaved by separase in anaphase, causing dissociation of centromeric cohesin, which is then recycled and used in the subsequent G1 phase, following the removal of acetyl groups from SMC3 (Losada, 2014).

### 1.6 Neocentromeres and satellite free centromeres

The movement of centromeres from satellite containing regions to satellite free regions both endogenously and artificially sealed the epigenetic status of the centromere (Craig et al., 2003; Wade et al., 2009). In 1993 the first ‘new’ centromere or neocentromere was discovered (Voullaire, Slater, Petrovic, & Choo, 1993) and since then over 90 instances have been described in humans (Kalitsis & Choo, 2012). Cytogenetic screens in humans have shown that neocentromere formation rescues acentric chromosomes formed as a result of two types of chromosomal rearrangement: inverted duplication of the chromosome arm, leading to an unbalanced karyotype or interstitial deletion resulting in a balanced karyotype with linear or circular chromosomes (Marshall, Chueh, Wong, & Choo, 2008). Sites of neocentromere formation share some common traits, typically neocentromeres form in euchromatic regions but surprisingly some of these have been found to be associated with HP1, suggesting the neocentromere still carries its ‘heterochromatic imprint’ (Saffery et al., 2000). There are two ways to interpret neocentromere



formation: a) it's a random event and that a selection process determines which neocentromeres will be stably transmitted or b) the presence of neocentromeric "hotspot" whereby specific regions of the genome is more favorable for neocentromere formation. This latter hypothesis is particularly supported due to presence of recurrent neocentromere observation on a number of chromosomal loci including 3q, 15q and in particular 13q which has 16 instances described (Alonso, Hasson, Cheung, & Warburton, 2010). A study involving human neocentromeres show the existence of latent centromeres common with primate ancestors that persistently bear the potential to seed neocentromere formation (Ventura et al., 2004). More instances of neocentromere formation have been reported in other mammals, with 9 examples in the macaques (Ventura et al., 2007), 16 instances in the donkey (Piras et al., 2010) and one example of chromosome 9 in the orangutan (human chromosome 12) (Locke et al., 2011) as well as in birds (Zlotina et al., 2012) and plants.

Chromosome engineering techniques have been employed in a bid to discern if neocentromere formation is biased toward particular types of sequences. In DT40 cells, the centromere of chromosome Z was conditionally removed and in surviving clones, neocentromeres were found to form across all regions of the chromosome (p-telomere, p-arm, metacentric, q -arm and q-telomere). However, 76% were found to be metacentric, suggesting a preference for neocentromere formation at this domain. ChIPSeq was carried out and it was found that newly formed neocentromeres, irrespective of their position were remarkably similar in size to the endogenous centromere. Sequence analysis of neocentromere domains showed a higher than average %AT content and no enrichment of DNA transposons (Shang et al., 2013).

### **1.7 The equid model system**

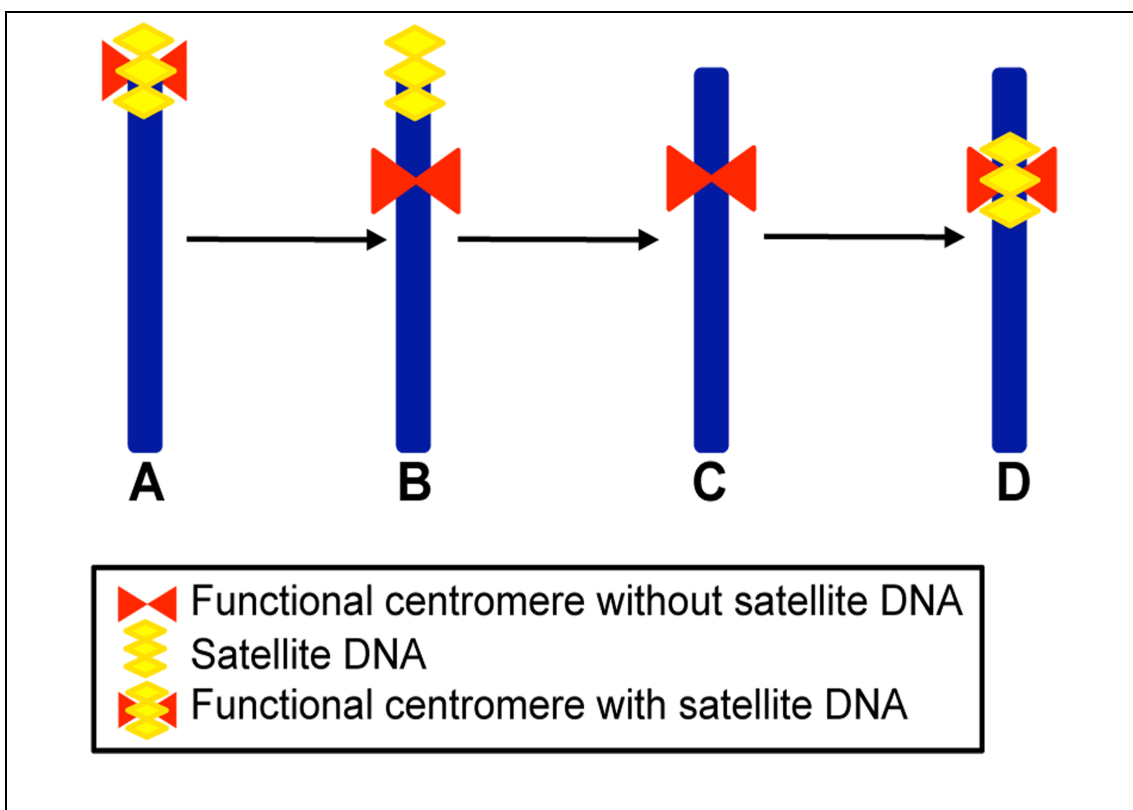
The *Perissodactyla*, also known as the odd-toed ungulates, are comprised of the Equidae family (horses, asses and zebras), the Tapiridae (tapirs) family and the Rhinocerotidae (rhinos) family (Steiner & Ryder, 2011). In the equidae family, eight species are still in existence today, two horses (*E. caballus* and *E. przewalskii*), two Asiatic asses (*E. kiang* and *E. hemionus*), one African ass (*E. asinus*) and three zebras (*E. grevyi*, *E. burchelli* and *E. zebra*). Approximately 4 - 4.5 million years ago, the Equus species diverged from a common ancestor which is remarkable fast in evolutionary terms (Orlando et al., 2013). The equids have adapted to extremely

diverse environments from the arid African savannahs to the Sakha republic, known for its extremely severe and cold environment. Genetic adaptations have been identified that reprogram transcription of genes involved in carbohydrate and lipid metabolism, hair development and limb morphogenesis. Perhaps the most interesting and fastest example of evolutionary adaptation is that of the subarctic Yakutian horse. Introduced to the region between the 13<sup>th</sup>-15<sup>th</sup> century this hairy, short-limbed breed adapted to the year round deep snow and permafrost soil. The earliest archeological evidence for domestication of the horse is 5,500 years ago, where pottery with equine milk traces and fossil remains of bit worn teeth indicative of harnessing was found in Kazakhstan (Orlando, 2015).

Present day horses have extremely diverse mitochondrial DNA, suggesting that during domestication mares were constantly restocked from the wild, while conversely almost complete homogeneity is observed in the paternal Y chromosome suggesting that during domestication a limited number of stallions were used. In the case of the donkey, there are two main mitochondrial groups indicative of two independent domestication sources (Orlando, 2015). Karyotypic diversification and variable diploid numbers between species;  $2n=32$  zebra and  $2n=66$  Przewalski's horse (Yang et al., 2003) have confounded attempts to identify modes of chromosome change across the species, using traditional chromosome banding analysis. Chromosome painting techniques have shown numerous centric fusions, centric fissions, tandem fusions and a small number of inversions are responsible for the karyotypic difference between the three species (Yang et al., 2003).

Centromeric repositioning within the genus *Equus* has also occurred, whereby the centromere shifts to a new region on the chromosome without displacement of gene order. Systematic marker order comparison across the equids demonstrate at least nine centromeric repositioning events. One of these took place in the horse (chromosome 11) and eight in the donkey (chromosome 8, 9, 11, 13, 15, 16, 18, 19) (Carbone et al., 2006; Piras et al., 2010). Using a cytogenetic approach Piras et al. 2010 investigated the distribution of tandemly repeated satellite arrays on metaphase chromosomes in the horse, donkey and two zebra individuals (*E. grevyi* and *E. burchelli*). The results showed that many functional centromeres were devoid of detectable satellite sequences by FISH with one instance in the horse on chromosome 11, eighteen instances in the donkey on chromosomes 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 22, 26, 27 and several examples in the zebra. In some instances satellite repeats

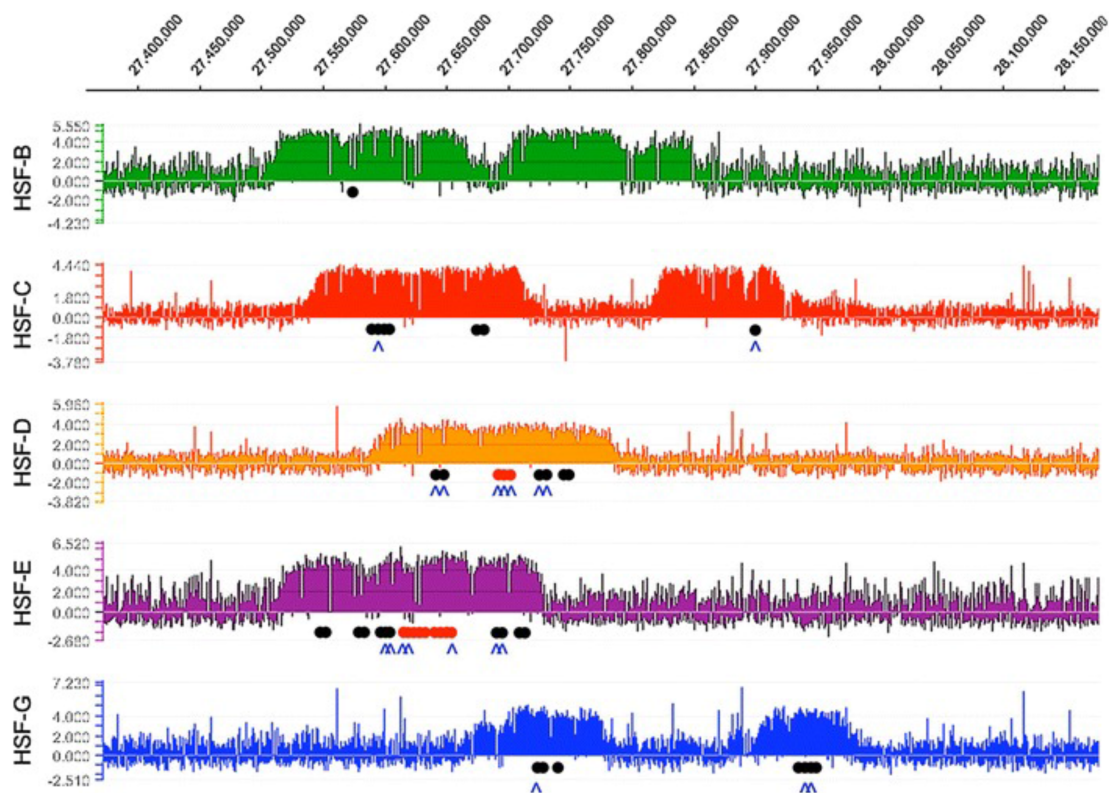
were observed at non centromeric termini (telocentric), supporting the concept, that the new “immature” centromeres have yet to mature into satellite containing loci and that these satellite domains are the ancient remnants of now inactive terminal centromeres. Piras et al. show a possible mechanism for the formation of an evolutionary new centromere from an acrocentric ancestral chromosome as shown in Figure 1.5. They suggest that the centromere shifts to a new position lacking satellite DNA, leaving the satellite DNA at the old centromere location Figure 1.5 B. The satellite DNA of the old centromere is lost, Figure 1.5 C and finally the neocentromere acquires satellite DNA, becoming a ‘mature’ centromere, Figure 1.5 D.



**Figure 1.5 Evolutionary mechanism for neocentromere formation.** (A) shows the ancestral acrocentric chromosome with the satellite containing centromere. (B) The functional centromere moves to a new region devoid of satellite sequence, leaving the satellite DNA sequences at the terminal position corresponding to the old centromere site. (C) Over time the satellite sequence of the old centromere is lost. (D) The neocentromere fully “matures” gaining satellite sequence (Piras et al., 2010).

It appears that in the equids, given the presence of so many functional satellite free centromeres, there is no requirement for centromeric satellite accumulation once the neocentromere has formed. ChIP-on-ChIP analysis of the centromeric domain on horse chromosome 11 using antibodies against CENP-A and CENP-C showed two distinct peaks of hybridization spanning 136kb and 99kb respectively (Wade et al. 2009). The detection of two domains of binding indicates two possibilities; this

organization is shared by both chromosome homologs or each peak signifies the centromeric domain on the two different homologs of chromosome 11. Driven by this observation, Purgato et al., 2015, extended this analysis to five individual horses. CENP-A was immunoprecipitated and the DNA was hybridized to a tiling array that spanned the centromere of horse chromosome 11. Each horse individual had a distinct CENP-A binding domain, as shown in Figure 1.6, located across a 500kb region and exhibited either two defined peaks (HSF-B, HSF-C, HSF-G) or a single peak (HSF-D, HSF-E). Single nucleotide polymorphism analysis to identify heterozygous nucleotide position was employed. In the case of the two CENP-A domains being present on both homologs, the ChIP would contain a similar amount of the two SNPs, in contrast, should the two homologues contain different CENP-A domains, only one of the SNP sequences would be enriched in the ChIP fraction. SNP analysis showed that HSF-D and HSF-E peak profile was a result of the partial overlap of two distinct peaks. In the instance of HSF-B, HSF-C and HSF-G, SNP analysis indicated that each homolog contained a CENP-A binding domain (Purgato et al., 2015).



**Figure 1.6 Variable centromere positions between different horse individuals.** Informative SNPs are shown as black dots, a single nucleotide is enriched in the immunoprecipitated DNA. Red dots, both SNPs are present in the immunoprecipitated DNA (Purgato et al., 2015).

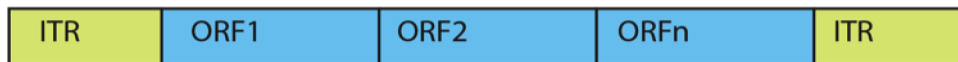
The variation in centromere profile and position between individuals illustrates the plasticity of the chromosome 11 centromere and furthers the notion that CENP-A binding is independent of DNA sequence. The “centromere sliding” shown by Purgato et al. provides a ‘snapshot’ of the ongoing evolution of the horse lineage. The satellite free equid centromeres offer a powerful model system, with naturally occurring, stably present loci that can be used to study the maturation of neocentromeres as well as examination of the architecture of the centromere at the molecular level.

### **1.8 Transposable elements**

Initially thought of as “junk DNA”, transposable elements are the most abundant class of genetic material in higher eukaryotes, accounting for ~40% of the human genome. Discovered in the 1940s, Barbara McClintock proposed their regulatory role in gene expression which was largely dismissed until relatively recently (Pray, 2008). Transposable elements include both transposons and retrotransposons, the latter include long interspersed nuclear repeats (LINEs) and short interspersed nuclear repeats (SINEs) as shown in Figure 1.7. Transposons move by a cut and paste mechanism and contain an inverted terminal repeat (ITR), acting as a cis element during integration and at least two open reading frames, which encode the transposase activity. Retrotransposons operate by a copy and paste mechanism whereby they reverse transcribe an RNA intermediate using their reverse transcriptase activity, before integrating the copy into the genome. Retrotransposons are divided into two classes referring to their self propagated mobility, autonomous and nonautonomous (Carnell & Goodman, 2003). Autonomous transposons are defined as Long Terminal Repeat (LTR) and non-LTR retrotransposons, which apart from lacking an env gene are structurally similar to retroviruses. LTR retrotransposons contain gag and pol genes, with gag encoding capsid-like protein and pol which encodes protease, reverse transcriptase, RNASE H and integrase activities (Wong & Choo, 2004). Non-LTR elements, such as LINEs contain an internal RNA pol II promoter, two ORFs, one with RNA binding activity the other encoding endonuclease and reverse transcriptase activity. Both autonomous types are mobile, yet their method of recombination differs. LTR retrotransposons are transcribed into RNA, reverse transcribed into DNA and recombined into genomic DNA. Non-LTR retrotransposons such as LINEs,

utilize a target primed reverse transcription mechanism (TPRT) (Cost, Feng, Jacquier, & Boeke, 2002). The full-length protein is transcribed into mRNA and is then reverse transcribed via its own reverse transcriptase and integrated into the genome using TPRT. The majority of L1 retrotransposons are inactive due to truncations, inversions or point mutations (Ostertag & Kazazian, 2001), resulting in the genome being littered with inactive LINES. SINEs are short ~100-400bp nonautonomous elements, that contain an internal RNA pol III promoter and do not encode protein. The movement of SINEs are dependent on LINES, however LINES are more typically associated with AT rich, gene poor genomic regions, while SINEs are associated with high GC, gene rich loci. LINES and SINEs both have a 3' poly A tail (Wong & Choo, 2004).

### Transposons



### Autonomous

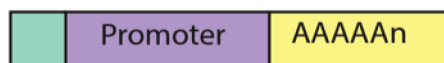


LTR retrotransposon



Non LTR retrotransposon (LINE)

### Nonautonomous



SINE

Figure 1.7 Transposable elements adapted from Carnell & Goodman 2003

### ***1.8.1 Transposable elements at the centromere***

Centromeres are normally associated with repetitive DNA, with repeat unit lengths surprisingly similar between organisms: primates 171bp, the fish *Spaus aurata* 186bp, the insect *Chironomus pallidicittatus* 155bp, rice 168bp and in both *Arabidopsis* and maize its 180bp. What is notable about these similar lengths is that it is corresponding closely to nucleosomal unit lengths (S Henikoff, Ahmad, & Malik, 2001). Phylogenetic analysis of centromere associated DNAs, carried out by Melters et al. across 282 species, 78 of which were plants and 204 were animals from 16 different phyla showed the sequences did not share common characteristics, except in the case of very closely related species eg. Primates, grasses etc. Centromeric sequences are rapidly evolving as a result of a number of mutational processes, including unequal exchange, transposition excision, as well as centromere inactivation and relocation to a genomic region of entirely unique sequence. This presents the centromere paradox, why not a single conserved centromeric sequence? (Steven Henikoff, 2001) More puzzling yet, how do similar sequences become associated with each centromere in a given species? One possible mechanism is gene conversion, a phenomenon widespread at maize centromeres, which facilitates the exchange of sequences among chromosomes (Shi et al., 2010). Another possibility is the incorporation of transposable elements, most strikingly, centromere specific transposable elements.

A Ty3/Gypsy class of centromere specific retrotransposons, a subset of the CR (centromeric retrotransposon) family, was identified in grasses. This family of retrotransposon is highly conserved and has been identified at the centromeres of rice, maize and barley (Jiang, Birchler, Parrott, & Dawe, 2003). In maize, a centromeric specific transposable element (CRM), is found interspersed with a 156bp centromeric satellite repeat (CentC), with both sequences capable of binding CENH3 (Jin et al., 2004). Similarly a centromeric specific rice retrotransposon (CRR) is found to be interspersed with satellite CentO repeats, that also associates with CENH3 (H. Yan & Jiang, 2007). A phenomenon recently observed in maize, shows the movement of centromere, as a direct result of domestication. A partial or complete loss of centromeric CentC repeats is observed causing cenH3 to expand into or jump to the closest domain with adequate stability. This domain is then subsequently invaded with centromeric retrotransposon 2 at a rate, which would occupy a centromere-sized domain (1.8Mb) in 20,000-95,000 yrs. The domestication of maize has driven selection of centromere-linked genes, which were formed by the relocation of active

centromeres by hemicentric inversions to formerly euchromatic regions. A theory put forward by Schneider et al, suggests CentC is a relic of previous neocentromere formation and retrotransposon invasion in the grass species. They postulate since the vast majority of eukaryotes contain centromeric tandem repeats rather than transposable elements, that retrotransposons are a transient stage of centromere evolution (Schneider, Xie, Wolfgruber, & Presting, 2016). A similar mechanism of evolutionary new centromere formation could be the case in the Equids. In the instances of primate centromere repositioning, which have been implicated in being a major mechanism of speciation and karyotype divergence, there is an increase LINE L1 retrotransposons at the new CENP-A binding domain. In the case of the mardel 10 chromosome, there was a 2.5 fold increase in L1 retrotransposons, with 4 instances of full-length L1 elements present. siRNA of the L1 transcripts led to reduced CENP-A incorporation and impaired mitotic function of the centromere (Chueh, Northrop, Brettingham-Moore, Choo, & Wong, 2009). Retrotransposons, particularly LINE L1 elements appear to play a role in neocentromere formation, at least in the instance of the mardel 10 chromosome.

### **Research objectives**

Given the abundance of satellite free centromeres, the equid model organism provides a unique tool for the study of centromeres. Centromere domains can be mapped with great accuracy providing an opportunity to identify particular DNA sequences associated with evolutionary new centromeres. The presence of unique sequence centromeres also provides a platform for determining protein architecture with respect to associated DNA.

The specific objectives of this thesis were:

- To generate an equid specific CENP-A antibody for use in ChIPSeq
- To determine centromeric domains in a donkey individual and compare these with another donkey and a horse to identify sequence variation and transposable elements present
- Identify the DNA associated with inner centromere proteins and map their linear one-dimensional primary structure with respect to the CENP-A binding domain



## **Chapter 2 Materials and Methods**

### **2.1 Materials – Wet lab**

#### ***2.1.1 Chemical reagents and consumables***

All reagents and chemicals used were purchased from Sigma Aldrich (Arklow, Co. Wicklow, Ireland) or Fisher Scientific (Ballycoolin, Dublin, Ireland), unless otherwise stated. All solutions were prepared with Milli-Q purified water and autoclaved at 121°C for 15 min where applicable. General and sterile laboratory plasticware was purchased from Sarstedt Ltd (Sinnottstown Lane, Drinnagh, Wexford, Ireland) and glassware from Fisher Scientific (Ballycoolin, Dublin, Ireland).

#### ***2.1.2 Molecular biology reagents, strains and equipment***

Reagents used in gateway cloning were purchased from Gibco-Invitrogen Life Technologies (Paisley, UK). Restriction enzymes, DNA molecular weight markers and protein molecular weight markers were obtained from NEB (New England Biolabs, ISIS Ltd, Unit 1&2, Ballywaltrim Business Centre, Boghall Road, Bray, Co. Wicklow, Ireland). DNA gels were made with ultrapure agarose from Gibco-Invitrogen Life Technologies (Paisley, UK). Agarose gels were run in Owl Separation System tanks using a Fischer Scientific POWER 608 powerpack and analysed with a Multi Image Light Cabinet (ChemiImager 5500, Alpha Innotech). DNA purification from gels was carried out using Qiaquick Gel Extraction kit from Qiagen (Crawley, UK). DNA plasmid purification was carried out with either Qiagen Mini Prep from Qiagen (Crawley, UK) or NucleoBond® Xtra Midi Prep from Macherey-Nagel (GmbH & Co. KG Neumann-Neander-Straße, 6-852355, Düren). Purification of ChIP DNA was performed using the PCR clean up kit from Qiagen (Crawley, UK). For crosslinking cells, EGS (ethylene glycol bis-succinimidyl succinate) was purchased from Fisher Scientific Ireland Ltd (Ballycoolin, Dublin 15) and 16% Paraformaldehyde was purchased from Electron Microscopy Sciences (1560 Industry Rd, Hatfield, PA 19440, United States). Disruption of bacterial cells and probe sonication of crosslinked horse and donkey chromatin was carried out using a Branson Digital Sonifier® Cell disrupter 250. For water bath chromatin shearing, 15ml Biorupter® plus TPX tubes from Diagenode (Liege science park, Rue Bois Saint-Jean, 34102 Seraing (Ougrée), Belgium) were used in a diagenode Bioruptor® UCD-200.

DNA transformations and plasmid preparations were performed using Escherichia coli Top 10 cells and protein expression was performed in BL21AI cells, Gibco-Invitrogen Life Technologies (Paisley, UK). BL21AI contains a T7 RNA polymerase under the control of an arabinose inducible promoter, ensuring no leaky basal expression of the recombinant protein. Deficient in the outer membrane aspartyl protease OmpT, BL21AI cells have reduced degradation of heterologously expressed proteins.

Plasmid	Source	Use
pENTR4	Invitrogen	Bacterial gateway entry vector, gene flanked L arms for recombination into pDEST17
pDest17	Invitrogen	Bacterial gateway expression vector with R arms and Histag.

**Table 2.1 Gateway Plasmids**

### Real time PCR

Real time FAST SYBR 2x mastermix was purchased from Biosciences (3 Charlemont Terrace, Crofton Road, Dun Laoghaire, Co Dublin, Ireland). qPCR was carried out using a DNA engine OPTICON 2 instrument. Primers were designed in the centromeric region of EquDonk2.0 chromosome 30, the centromeric region of horse chromosome 11 and in the equid single copy gene PRKC. Primers were ordered from Eurofins Genomics (Anzinger Str. 7a 85560 Ebersberg, Germany) and are listed in the table below.

Region	Primer Forward (5'→3')	Primer Reverse (3'→5')	PCR fragment length (bp)
Eca Cen11	CAGCAAGGCATTTCCAGTGA	CATGCAAGACAAGGAGGAACG	130 bp
PRKC	TGGAGCAAAAGCAGGTGGTA	ATCGTCATCTGGAGTGAGCTG	116 bp
Eas Cen 30	CACTACCCTGGCACTGCGA	TGGATGTCACGGTAGGCAATG	103 bp

**Table 2.2 qPCR primers**

### Antibodies

Antibodies used in this study are listed in Table 2.3 (primary antibodies) and Table 2.4 (secondary antibodies) below. Secondary antibodies for immunofluorescence, preadsorbed to remove cross-reacting anti-IgG antibodies, were obtained from Jackson ImmunoResearch Europe Ltd with the indicated fluorochrome (table 2.4).

HRP-coupled antibodies and protein A were obtained from Jackson ImmunoResearch Europe Ltd and Merck Millipore respectively (table 2.4).

Reactivity	Host	Dilution WB	Dilution IF	ChIP	Source
<b>CENP-A</b>	Sheep IgG	1:5000 5% milk TBST	1:100	1ul/1x10 <sup>6</sup> cells	This work
<b>CTCF</b>	Rabbit IgG	1:1000 5% milk TBST	1:100	1ul/1x10 <sup>6</sup> cells	Merck Millipore 07-727
<b>Smc1</b>	Rabbit IgG	1:1000 5% milk TBST	-	1ul/1x10 <sup>6</sup> cells	Bethyl Laboratories A300-055A
<b>Aurora B</b>	Mouse IgG	1:1000 1% BSA PBST	1:100	1ul/1x10 <sup>6</sup> cells	BD biosciences 611082
<b>Survivin</b>	Rabbit IgG	1:100 5% milk TBST	1:1000	1ul/1x10 <sup>6</sup> cells	Novus Biologicals NB500-201
<b>Borealin</b>	Mouse IgG	1:1000 1% BSA PBST	1:100	1ul/1x10 <sup>6</sup> cells	MBL M147-3
<b>CREST</b>	Human IgG	-	1:100	-	Kevin F Sullivan

**Table 2.3 Primary Antibodies**

Reactivity	Conjugation	Host	Dilution IF	Dilution WB	Source
Anti Rabbit IgG	Cy5	Goat	1:100	-	Jackson ImmunoResearch Europe Ltd.
Anti Rabbit IgG	Tritc	Donkey	1:100	-	Jackson ImmunoResearch Europe Ltd.
Anti Mouse IgG	Tritc	Goat	1:100	-	Jackson ImmunoResearch Europe Ltd.
Anti Mouse IgG	Fitc	Goat	1:100	-	Jackson ImmunoResearch Europe Ltd.
Anti Human IgG	Fitc	Goat	1:100	-	Jackson ImmunoResearch Europe Ltd.
Anti Sheep IgG	Tritc	Rabbit	1:100	-	Jackson ImmunoResearch Europe Ltd.
Protein A	Horseraddish peroxidase (HRP)	-	-	1:20,000	Merck Millipore
Anti Mouse IgG	Horseraddish peroxidase (HRP)	Goat	-	1:20,000	Jackson ImmunoResearch Europe Ltd.
Anti Rabbit IgG	Horseraddish peroxidase (HRP)	Goat	-	1:20,000	Jackson ImmunoResearch Europe Ltd.
Anti Sheep IgG	Horseraddish peroxidase (HRP)	Goat	-	1:20,000	Jackson ImmunoResearch Europe Ltd.

Table 2.4 Secondary antibodies

### 2.1.3 Protein methods

SDS PAGE was carried out in the Invitrogen X-Cell Sure Lock Gel electrophoresis System using Precast Novex NuPAGE Gels. Wet transfer was performed in the Invitrogen X-Cell II blot module (supplied by Biosciences, 3 Charlemont Terrace, Crofton Road, Dun Laoghaire, Dublin). Hydrophobic Polyvinylidene Fluoride membrane (PVDF) was purchased from Merck Millipore (Tullagreen Carrigtwohill, County Cork, Ireland). Enhanced Chemiluminescence (ECL) reagent was purchased from Perkin Elmer (Unit G13 Calmount Park Ballymount, Dublin 12, Ireland) and X-ray film was obtained from Agfa (Vantage West, Great West Road, Brentford,

Middlesex TW8 9AX, United Kingdom). Ni-NTA agarose beads for affinity purification of His-tagged protein were purchased from Qiagen (Crawley, UK) and Bio-Rad Poly-Prep® Chromatography Columns were obtained from Fannin (South County Business Park, Leopardstown, Dublin 18). For cell lysis, complete mini EDTA protease inhibitor tablets were purchased from Roche (Clarecastle, Co. Clare).

#### ***2.1.4 Tissue culture reagents***

All reagents and chemicals for cell culture were purchased from Sigma-Aldrich Ireland Ltd. (Arklow, Co. Wicklow, Ireland), unless otherwise stated. Fetal bovine serum (FBS) and phosphate-buffered saline (PBS) were obtained from Lonza (Slough UK). 10X Trypsin EDTA was purchased from Gibco-Invitrogen Life Technologies (Paisley, UK). For hTERT immortalized cell lines, media was supplemented with the G418 to a final concentration of 400µM/ml purchased from Sigma-Aldrich. For populations enriched in mitosis, cells were treated with a single thymidine block and released into media supplemented with 2'-deoxycytidine both cell culture grade reagents purchased from Sigma Aldrich (Arklow, Co. Wicklow). Culturing of cells was carried out in a Class II Bio-safety cabinet.

## **2.2 Methods – Wet Lab**

### ***2.2.1 Nucleic Acid techniques***

#### **2.2.1.1 Preparation of competent Top10 cells**

Competent Top10 cells were plated on non selective agar and incubated overnight at 37°C. For the starter culture, a colony was added to 20mls LB and incubated overnight. The starter culture was then diluted 1:50 in LB and grown at 37 °C for 2 hours. When the culture reached an OD<sub>600</sub> of 0.5, the culture was pelleted at 6000xg for 10 minutes at 4 °C. The pellet was resuspended in 50ml ice cold 0.1M CaCl<sub>2</sub> per 100ml starter culture and incubated on ice for 30 minutes. Cells were pelleted again and resuspend in 10mls ice cold 0.1M CaCl<sub>2</sub> with 15% glycerol per 100ml starter culture. Finally cells were aliquoted and stored at -80 °C.

#### **2.2.1.2 E.coli transformations**

E.coli cells prepared above were transformed with ligated DNA by heat shock at 42°C for 40 seconds and allowed recover for 1hour with agitation at 37°C. Cells were plated on LB agar with the appropriate antibiotic and grown overnight at 37°C.

#### **2.2.1.3 E.coli DNA extraction**

A single colony of transformed E.coli described above was picked and inoculated in 2mls LB (Mini-prep) or 200mls LB (Midi-prep) with the appropriate antibiotic. This culture was grown overnight and purified as per the manufactures instructions.

#### **2.2.1.4 Gene synthesis**

Horse CENP-A sequence was obtained from ENSEMBL, ENSECAP00000013849 and codon optimised for expression in E.coli using an algorithm written by Dr Andrew Flaus. The sequence was flanked by a 5' BamHI site and a 3' NotI site for ease of cloning and orientation. The gene was synthesized by Eurofins Genomics (Ebersberg, Germany). The CENP-A gene was supplied in a pEx-A ampicillin resistant plasmid.

#### **2.2.1.5 Restriction digestion**

1µg of pEX-A CENP-A was digested with BamHI HF and NotI HF in buffer 4 supplemented with BSA. The digestion was incubated at 37°C for 3 hours. The digest was then loaded onto an agarose gel as described below and the 436bp band corresponding to CENP-A was gel extracted and purified.

#### **2.2.1.6 Agarose gel analysis**

1% agarose gels were prepared with 1x TAE (40mM Tris, 20mM Acetate and 1mM EDTA pH 8.6) and 0.1mg/ml ethidium bromide unless otherwise stated. 2log DNA ladder was used as DNA marker. Gels were run at 100Volts until the desired resolution was achieved.

#### **2.2.1.7 Gel extraction**

DNA was resolved on an agarose gel as described above. The band of the desired size was visualised by transillumination and excised using a clean scalpel and placed in a 2ml eppendorf tube. The DNA was purified from the agarose using the gel extraction kit as per the manufacturers protocol.

#### **2.2.1.8 Gateway cloning**

Gateway cloning was carried out as per manufacturers instructions. pENTR4 was digested as outlined in Section 2.2.1.5. CENP-A was then ligated into pENTR4 using T4 DNA ligase. The ligation was then transformed into Top10 E.coli (Section 2.2.1.2). Following transformation, colonies were picked and grew in 2mls LB with

50µg/ml Kanamycin. Vectors were purified by mini-prep, screened by restriction digest and sent for sequencing to GATC (Gottfried-Hagen-Straße 20, 51105 Cologne, Germany). Once the insert was verified, CENP-A was recombined into pDEST17 by LR recombination, with the appropriate controls, as outlined in the table 2.5 and incubated at 25 °C for 16 hours. Post recombination, 1 µl of proteinase K was added to each reaction and incubated for 10 minutes at 37°C, to stop the reaction. The recombined pDest17 CENP-A was transformed into Top 10 E.coli and plated on LB Ampicillin 100µg/ml, colonies were picked and screened by restriction digest.

	Sample	Negative Control	Positive Control
<b>pENTR4 CENP-A (100ng/µl)</b>	3µl	3µl	-
<b>pDEST17 (150ng/µl)</b>	1µl	1µl	1µl
<b>pENTR –gus (50ng/µl)</b>	-		2µl
<b>TE buffer, pH8</b>	4µl	6µl	5µl
<b>LR Clonase</b>	2µl	-	2µl

**Table 2.5 Gateway cloning LR reaction**

### **2.2.1.9 ChIP qPCR**

ChIP enrichment was measured using real-time PCR (qPCR) under the conditions outlined in Table 2.6. qPCR was carried out using SYBR green dye to measure the amount of amplified product at the end of each cycle. SYBR green is an intercalating agent, which binds DNA base pairs, making it specific for double-stranded DNA thereby SYBR green intensity is directly proportional to the amount of product formed.

To calculate the relative enrichment in the ChIP DNA, 1% of the input was taken from the total ChIP and adjusted to 100% by subtracting 6.644, which is the amount of cycles taken to reach 100%. Amplification was carried out using centromere-associated primers for horse and donkey as well as the negative PRKC single copy gene control. The enrichment of the ChIP relative to the input was calculated using the following equation:  $100 * 2^{(Adjusted\ input - C_t\ (IP))}$ .

Step	Temperature (°C)	Duration (secs)	Cycles
<b>Polymerase activation</b>	95	20	1
<b>Denature</b>	95	3	40
<b>Anneal/Extend</b>	60	30	

Table 2.6 qPCR conditions

## 2.2.2 Protein techniques

### 2.2.2.1 Donkey cell extract preparation

Cells were harvested, counted and split into  $5 \times 10^6$  aliquots. The cells were washed twice with 500 $\mu$ l of buffer A (10mM Hepes, 10mM KCl, 0.5mM DTT). Cells were resuspended in 40 $\mu$ l/ $5 \times 10^6$  cells buffer B (10mM Hepes, 10mM KCl, 0.5mM DTT, 1% NP40) and incubated on ice on the rocker for 10 minutes. A fraction of cells were taken at this point for whole cell extract. For the cytoplasmic fraction, lysates were spun at maximum speed for two minutes and the supernatant aliquoted. The remaining nuclear pellet was resuspended with 110 $\mu$ l buffer C/ $5 \times 10^6$  cell equivalents (20mM Hepes, 20% Glycerol, 500mM KCl, 0.2mM EDTA, 0.5mM PMSF, 0.5mM DTT and 1.5mM MgCl<sub>2</sub>) and incubated on ice on the rocker for 15 minutes. Nuclei were spun at maximum speed for 10 minutes at 4°C, the supernatant was removed and stored as the nuclear extract. The pellet was then resuspended with Micrococcal digestion buffer (15mM Tris pH8, 15mM NaCl, 60mM KCl, 1mM CaCl<sub>2</sub>, 1mM DTT, 0.2mM PMSF, 0.15mM Spermine, 0.5mM Spermidine, protease inhibitor cocktail) at a concentration of 100,000 cells/ $\mu$ l. Micrococcal nuclease was added at a concentration of 0.005u/ $\mu$ l of cell suspension, the digest was incubated at 37°C for 20 minutes. To stop the reaction, EGTA was added to a final concentration of 10mM and vigorously pipetted to ensure complete chelation of the CaCl<sub>2</sub> ions. NaCl<sub>2</sub> was added to a final concentration of 300mM. The digest is then spun down at maximum speed for 10 minutes and the supernatant removed (chromatin fraction). All fractions were supplemented with 4x LSB to 1x final concentration and 100mM DTT. Samples were boiled for 10 minutes and stored at -20°C until subsequent SDS-PAGE/Western blot analysis.

### 2.2.2.2 SDS-polyacrylamide Gel electrophoresis

SDS-PAGE was carried using two different buffer systems depending on protein size and desired resolution. The gel was run using either NuPAGE® MES buffer (50 mM MES, 50 mM Tris base, 1 mM EDTA, 0.1% (w/v) SDS, pH 7.3) which gives a maximum resolution of 188kDa and a minimum of 3kDa or MOPS buffer (50 mM



MOPS, 50 mM Tris base, 0.1% SDS, 1 mM EDTA pH 7.7) giving a maximum resolution of 191kDa and a minimum of 14kDa. 500µl of NuPAGE antioxidant was added to the inner chamber immediately prior to loading samples, to maintain proteins in a reduced state. SeeBlue® Plus2 prestained protein standard was used in all SDS-PAGE experiments. Samples were run at 200Volts until the dye front reached the bottom of the gel.

#### **2.2.2.3 Wet transfer**

Sponges and whatman paper were soaked in transfer buffer (10% MeOH, 25mM Bicine, 25mM Bis-tris, 1mM EDTA, 0.05mM Chlorobutanol). PVDF membrane was activated by soaking in 100% methanol for 5 minutes. The membrane was immersed in deionised water for 1 minute and stored in transfer buffer until transfer. Proteins were transferred from the resolved gel onto the PVDF membrane at 30volts for 1 hour. Following transfer, efficiency was tested by staining the membrane with Ponceau S (1% Ponceau S, 5% acetic acid) for 15 minutes at room temperature with gentle agitation.

#### **2.2.2.4 Western blot**

Conditions were optimised for each antibody in table 2.3. The membrane was blocked in either TBST (Tris-buffered saline, 0.1% Tween 20) or PBST (phosphate-buffered saline, 0.1% Tween 20) with 5% Milk at room temperature with gentle agitation for 1 hour. The blocking solution was replaced with either 1% BSA in PBS or 5% Milk in TBST solution along with the antibody probe. The membrane was incubated on the roller at 4°C overnight. After incubation, the membrane was washed three times in TBST or PBST and the appropriate secondary antibody was applied in either 1% BSA in PBST or 3% milk in TBST solution. The membrane was incubated for one hour at room temperature with gentle agitation. Excess antibody was removed from the membrane by washing and Enhanced Chemiluminescent (ECL) reagent was washed over the membrane. Excess ECL was removed from the membrane before exposure to photographic film for between 0.1-1hour. The film was then developed.

#### **2.2.2.5 Coomassie staining**

Proteins were resolved by electrophoresis as previously outlined. The gel was then immersed in Coomassie stain (0.1% Coomassie Brilliant blue (Bio-Rad Laboratories), 50% Methanol and 10% Glacial acetic acid) for 3-4 hours at room temperature with gentle agitation. When a desired level of staining was achieved the excess coomassie

was removed from the gel with destaining solution (50% Methanol,40% Acetic Acid) and kimwipes were also added to the solution. The gel was incubated at room temperature with gentle agitation for one hour. The gel was then scanned and a digital image saved.

#### **2.2.2.6 Protein expression in E.coli**

The pDest17 CENP-A construct was transformed into BL21AI competent cells. A single colony was picked and inoculated into 5mls LB with ampicillin. The culture was grown overnight at 37°C. This 'starter culture' was used to inoculate 1 liter of LB amp which was grown until the  $A_{600}$  reached 0.5. A 1ml sample was taken for protein analysis of the uninduced culture. The culture was then induced with L-arabinose to final concentration of 0.2% for 4 hours. Another 1ml sample was taken for protein analysis and the rest of the culture was pelleted and stored at -20°C.

#### **2.2.2.7 Solubility Assay**

Protein expression was induced in BL21AI as described in Section 2.2.2.6. The pellet was resuspended in native extraction buffer (100mM Tris pH8, 10% Glycerol, 500mM NaCl, 1% NP40, 100mM DTT, protease inhibitor cocktail). Cells were sonicated at maximum amplitude for 8 minutes, 30seconds on/off in probe sonicator. Cells were spun down at 500xg and a sample of the supernatant was taken for protein analysis.

#### **2.2.2.8 Inclusion body prep**

Cell cultures were grown up and induced as outlined in Section 2.2.2.6 and stored at -20°C. The pellet was thawed for 30 minutes at 37°C. The pellet was resuspended in ice cold wash buffer (50mM Tris pH 7.5, 100mM NaCl, 1mM EDTA pH8, 1mM Benzamidine, 5mM Beta-mercaptoetanol) to a volume of 30mls/1g pellet. Cells were aliquoted into 50ml falcon tubes and kept on ice. Cells were lysed with a probe sonicator at an amplitude of 40% for two minutes, 5 seconds on/10 seconds off. The sonicated suspension was then transferred to nalgene tubes and centrifuged for 15 minutes at 4°C and 23000xg (JA17 rotor). The supernatant was discarded and the pellet was resuspended in wash buffer and spun down again. In the same manner the pellet was resuspended twice in Triton wash buffer (wash buffer with 1% Triton). The pellet was then washed a further two times with wash buffer. In order to solubilize the inclusion bodies, 0.5mls DMSO was added to the pellet and incubated on a roller at room temperature for 30 minutes. Unfolding buffer (7M Guanidium-HCl, 20mM Tris

pH7.5, 10mM DTT) was then added to the solubilized inclusion bodies and incubated for a further hour on the roller at room temperature. The inclusion bodies were then spun at 35000xg for 20 minutes. The supernatant was stored at -20°C until purification.

#### **2.2.2.9 Nickel affinity purification**

CENP-A inclusion bodies prepared above were diluted 1:20 with binding buffer (7M Urea, 50 mM Tris pH 7.5, 50 mM NaCl). The diluted protein was incubated with 500µl of Ni-NTA Agarose beads on a rotator for 1hour at room temperature. The resin mixture was loaded onto a polyprep column and the flow through collected. A sample was taken for SDS-PAGE analysis of protein binding. The resin was washed with 20mls of wash buffer (7M Urea, 50mM Tris pH7.5, 50mM NaCl, 20mM Imidazole) and the fractions were collected for protein analysis. Finally the protein was eluted from the column with 10mls elution buffer (7M Urea, 50mM Tris pH7.5, 50mM NaCl, 500mM Imidazole). The eluted protein was collected in 1ml fractions, 10ul of which was taken for protein analysis.

#### **2.2.2.10 Antibody production in Sheep**

The cleanest CENP-A fractions were pooled and dialyzed into PBS overnight. CENP-A formed a fine precipitate in the absence of urea. The CENP-A precipitate was sent to the Scottish National Blood Transfusion Service (Castlelaw Building, Pentlands Science Park, Penicuik, Midlothian, EH26 0PZ). A maximum of 1mg of protein was administered to the animal over three consecutive immunizations. Four bleeds were obtained, pre-immunization and post immunization bleed 1, 2 and 3. Should the animal show immunogenic response specific to the antigen further bleeds can be requested. The sera were aliquoted and characterized by western blot, immunofluorescence, ChIP-qPCR and ChIPSeq.

#### **2.2.2.11 Antibody affinity purification**

Nickle affinity purified CENP-A protein was dialyzed overnight into coupling buffer (8M Urea, 0.1M carbonate buffer, 0.5M NaCl pH 8.3). 0.5g of CNBr-activated Sepharose 4B beads were weighed out and activated in 100mls ice cold 1mM HCl for 15 minutes. 1g of freeze dried beads swell to 3.5mls (5-10mg protein/ml beads). After sedimentation, the supernatant was removed and the beads were washed three times in coupling buffer, 3mins 1000rpm. Dialyzed CENP-A was diluted 1:100 with coupling buffer, added to the activated sepharose beads and incubated on a rotator for 2hours at

room temperature. Beads were spun down, washed twice in coupling buffer and incubate in 8M urea, 0.1M Tris pH8 for 2 hours to preserve the activity of remaining active groups. Beads were then washed three times in 8M urea, 0.1M Tris pH 8, three times in acetate buffer (8M Urea, 0.1M Na Acetate, 0.5M NaCl pH 4) and three times in PBS. 1ml of beads were resuspended in 10mls of PBS and 10mls of antiserum was added, the suspension was incubated overnight on a rotator. The beads were washed three times in PBS and loaded into a polyprep column and washed again with PBS until  $A_{280}$  was 0. Bound antibody was eluted with 200mM glycine pH2.8 and collected in 1ml fractions in eppendorf tubes containing 27 $\mu$ l 3M Tris-HCl pH 8.8 and 100 $\mu$ l 3M KCl.  $A_{280}$  of the elutions were measured and concentrated fractions were pooled and dialyzed overnight into PBS. Purified sera was then aliquoted and snap frozen.

#### **2.2.2.12 Chromatin Immunoprecipitation**

Formaldehyde or EGS and Formaldehyde crosslinked cells (See sections 2.2.3.4 and 2.2.3.5) were thawed on ice for 20 minutes and resuspended by gentle flicking of the tube. The pellet was resuspended in ChIP lysis buffer (0.25% SDS, 50mM Tris-HCl pH8, 10mM EDTA, protease inhibitor cocktail (Complete Ultra Mini Tablets, Roche)) to a volume of 1ml/ $20 \times 10^6$  cells for the waterbath sonicator and a volume of 650 $\mu$ l/ $20 \times 10^6$  for the probe sonicator. In the case of the waterbath sonicator chromatin was sheared at maximum intensity for 60 cycles 30secs on/off. Water was changed every 5 cycles to ensure the bath stayed at 4°C. For the probe sonicator, chromatin was pulsed for 10seconds at output 3 for 24 cycles. 500,000 cell equivalents were taken for DNA analysis. Two sonication methods were employed since the horse CENP-A ChIPSeq was performed in the University of Pavia, where there was no access to a waterbath sonicator. Attempts to recapitulate the probe sonication in the Center for Chromosome Biology proved unsuccessful and waterbath sonication was then employed.

Sheared chromatin was spundown at maximum speed in a microcentrifuge for 10 minutes and the supernatent carefully removed without disturbing the pellet. The chromatin was diluted 1:3 with alternative ChIP dilution buffer (0.5% Nonidet P-40, 10mM Tris-HCl pH7.5, 2.5mM MgCl<sub>2</sub>, 150mM NaCl, protease inhibitor cocktail). Diluted chromatin was precleared with 100 $\mu$ l Protein G beads per  $50 \times 10^6$  cells (GE Healthcare Life Sciences, Amersham Place, Little Chalfont, Buckinghamshire, HP7 9NA UK), which were blocked with IgG free BSA and E.coli genomic DNA at 4°C

for 2 hours in an orbital shaker. After preclearing an input sample was taken and stored at 4°C for subsequent purification and analysis. The chromatin was then spun down at 500xg and aliquoted into 1.5ml eppendorf tubes, the appropriate amount of antibody was added (table 2.3) and incubated overnight on the orbital shaker at 4°C. Appropriate amount of protein G sepharose beads was added to each tube to bind the antibody and samples were incubated for a further 4 hours on the orbital shaker at 4°C.

ChIP samples were spun down at 1000xg and the supernatant removed. Beads were washed 5 times with 1ml of ice cold ChIP wash buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA pH 8, 150mM NaCl, 20mM Tris-HCl pH8) followed by 1ml of ChIP final wash buffer (0.1% SDS, 1% Triton X-100, 2mM EDTA pH8, 500mM NaCl, 20mM Tris-HCl pH8).

The immunocomplexes were eluted from the beads with 240ul ChIP elution buffer (1% SDS, 100mM NaHCO<sub>3</sub>, 40ug/ml RNase A). The precleared input sample was processed in the same manner making up to the same final concentration of the ChIP elution buffer. Both the beads and the input were incubated at room temperature for 15 minutes. To allow optimal RNaseA activity, immunocomplexes were incubated at 37°C for 1 hour, followed by proteinase K digestion at 55°C for 2 hours. Decrosslinking was carried out in a 65°C waterbath overnight.

DNA was purified using the PCR clean up kit as per manufacturers instructions. DNA concentrations were measured using the Qubit dsDNA High Sensitivity kit (ThermoFischer Scientific) and relative enrichment was examined by qPCR. DNAs were shipped to IGA technology services (Via Jacopo Linussio, 51, 33100 Udine UD, Italy) for library preparation and sequencing.

### ***2.2.3 Cell culture***

#### **2.2.3.1 Cell lines**

Horse and donkey fibroblast cell lines were obtained from Prof. Elena Giulotto (*Università di Pavia*). Horse cell lines were primary. Donkey fibroblasts were immortalized by transfection with ATCC human telomerase reverse transcriptase (hTERT) (Vidale et al., 2012).

### **2.2.3.2 Culture conditions**

Fibroblasts were cultured in Dulbecco's Modified Eagle Medium (DMEM) F12 Ham supplemented with 1x Non essential amino acids, 2mM L-glutamine, 1% Penicillin-streptomycin, 1mM Sodium pyruvate, 0.348% Sodium bicarbonate, 10% Foetal bovine Serum and 10% Horse Sera. Immortalised Donkey fibroblast were grown in identical media supplemented with G418. All cells grew as adherant cultures.

Cells were typically grown in 100mm x 20mm or 150mm x 25mm culture dishes and incubated in 5% CO<sub>2</sub> at 37°C. Cells were subcultured approximately every two days, when they reached approximately 80% confluence: cells were first washed in phosphate-buffered saline (PBS) and then treated with 1X Trypsin EDTA for 3 min to detach the cells. The cells were then resuspended in fresh medium and centrifuged at 1,200 r.p.m (300g) for 5 minutes. Cells were then gently resuspended in 5ml fresh medium. Total cell number was determined using a hemocytometer and cells were plated  $1.2 \times 10^4$  per cm<sup>2</sup>.

### **2.2.3.3 Mitotic enrichment**

Donkey fibroblasts were plated in 15 cm dishes, grown to 50% confluence and treated with 2mM thymidine (200mM stock in serum free media, Sigma) for 16 hours. Addition of thymidine causes some cells to accumulate G1/S transition while others will be cycling in S phase. Cells were washed twice in serum free media and released into complete media supplemented with 24µM Deoxycytidine. After 6 hours cells were harvested, EGS and formaldehyde crosslinked (Section 2.2.3.5) and analysed by flow cytometry.

### **2.2.3.4 Formaldehyde crosslinking**

To crosslink the proteins to DNA, cells were harvested as outlined previously, counted and crosslinked in PBS containing 1% paraformaldehyde on a shaker at 500rpm and 25°C for 9 minutes. Unreacted formaldehyde was then quenched with 125mM glycine, and a further incubation of 10 minutes at 25°C was performed. Cells were pelleted at 500xg and washed three times in ice cold PBS. Cells were then frozen at -80°C until downstream ChIP processing.

### **2.2.3.5 EGS crosslinking**

EGS (ethylene glycol bis-succinimidyl succinate) crosslinking occurs through the amine reactive NHS-ester ends of a 12-atom spacer arm. Use of a longer spacer arm is critical for mapping proteins that are not directly associated with chromatin. EGS

powder was equilibrated to room temperature and a 25mM stock in DMSO was made immediately prior to crosslinking. Cells were resuspended in PBS and EGS was added to a final concentration of 1mM. Cells were incubated on a roller at room temperature for 25 minutes before formaldehyde crosslinking as above.

#### **2.2.3.6 Cryopreservation**

Cells were stored in liquid nitrogen at a concentration of  $1.5 \times 10^6$ /ml in 45% horse sera, 45% fetal bovine serum and 10% DMSO. Cells were harvested at 70% confluence by trypsination and counted. Cells were then resuspended to a final concentration of  $1.5 \times 10^6$  cells/ml and aliquoted into labelled cryovials. Cells were frozen at a rate of  $1^\circ\text{C}$  per minute, to ensure membrane integrity, in a Mr Frosty® freezing container (Nalgene, Rochester, NY, USA) containing 250mls of Isopropanol overnight at  $-80^\circ\text{C}$ . Cells were then transferred to liquid nitrogen for long term storage.

#### **2.2.3.7 Resuscitation**

Cells were resuscitated by rapid thawing at  $37^\circ\text{C}$  for one minute. Cells were then suspended in 20mls of prewarmed culture media and plated in a 10cm dish. The cells were incubated at  $37^\circ\text{C}$  5%  $\text{CO}_2$  overnight and the media was changed the following day.

#### **2.2.3.8 Flow cytometry**

Cell cycle distribution was analysed by flow cytometry. Cells were counted and filtered using CellTrics® (Sysmex, Bornbarch 1, 22848 Norderstedt), cells were then spun down at 1200 rpm for 5 minutes. Cells were fixed in 70% ethanol and stored at  $-20^\circ\text{C}$  until analysis. Fixed cells were thawed on ice and dispersed by vortexing while adding PBS. Cells were spun down as above and 200 $\mu\text{l}$  propidium iodide (PI/RNase Staining Buffer, BD Pharmingen, 550825) was added to the pellet. Samples were incubated in the dark at  $4^\circ\text{C}$  for 30minutes before analysis on the BD Accuri C6 flow cytometer. Data generated was analysed using Modfit by Verity. Doublets were excluded by gating. Cell cycle distribution was visualized by histogram, plotting cell count versus PI intensity.

## **2.2.4 Immunofluorescence Microscopy**

### **2.2.4.1 Metaphase spreads**

Mitotic cells were harvested by elutriation. Cells were swollen in 75mM KCl for 30 minutes at 37°C. Sucrose was added to a final concentration of 25mM and cells were incubated at room temperature for 15 minutes. Swollen cells were spun down at 1250 rpm for 10 minutes using the cytopsin. The cells were then fixed to the slides with either 100% ice cold methanol or 4% Paraformaldehyde in 1x PBS.

### **2.2.4.2 Fixation**

Donkey cells were grown on glass coverslips in four well dishes. When the cells were 70% confluent, the media was removed and cells were gently washed in PBS. Depending on antibodies used, cells were fixed with either 100% ice cold methanol or 4% Paraformaldehyde in 1x PBS.

### **2.2.4.3 Immunofluorescence protocol - Metaphase spreads**

Post fixation, spread quality was examined by DAPI staining. After spread preparation described above, the coverslips were removed from slides by incubation in 2xSSC (from a 20xSSC stock-3M NaCl, 0.3M sodium citrate pH 7) at 37°C for 5 minutes. The spreads were permeabilized in PBS-Tween20 (0.05%) at room temperature for 10 minutes. Primary antibody was diluted in PBS-Tween20 and incubated for 1 hour at 37°C. Slides were washed twice with PBS-Tween20 at 37°C. Secondary antibody was diluted as the primary and slides were incubated in the dark for 1 hour at 37°C. Slides were washed twice in PBS-Tween20 at room temperature. Slowfade with DAPI was dropped onto the slides and coverslip was mounted. Coverslips were sealed with nail varnish and stored at 4 °C in the dark.

### **2.2.4.4 Immunofluorescence protocol - Coverslips**

Donkey cells were grown on glass coverslips in four well dishes. When the cells were 70% confluent, the media was removed and cells were gently washed in PBS. Cells were fixed as outlined above. Fixed cells were washed twice for 2 minutes in PBS followed by two 3 minute washes in PBS-TX (PBS, 0.1% Triton X-100). Cells were then blocked in 1% BSA PBS-TX for 15 minutes. Primary antibody was diluted in 1% BSA-PBS-TX and incubated on the coverslips for 1 hour at 37 °C. Cells were washed twice for 10 minutes in PBS-TX and incubated with a fluorescently conjugated secondary antibody in BSA-PBS-TX at 37 °C for 1 hour. Cells were then



washed once in PBS-TX, once in PBS and once in water before being air dried and mounted on slides with Slowfade and DAPI.

#### **2.2.4.5 Immunofluorescence imaging**

All microscopy was carried out using a Deltavision Core system (Applied Precision) controlling an interline charge-coupled device camera mounted on an inverted microscope (Olympus). For each sample images were collected at either 1x1 or 2x2 binning using a 60x oil objective at 0.2µm z sections. Immunofluorescence images were subject to iterative constrained deconvolution and maximum intensity projection using the SoftWoRx software (Applied Precision). ImageJ software was used to analyze images.

### **2.3 Materials – Dry lab**

#### **2.3.1 Hardware**

The computer used for Bioinformatic analyses was an iMac running iOS X 10.8.5 with a 3.4GHz Intel Core i7 processor and 32GB RAM.

#### **2.3.2 Software**

Software packages used throughout this thesis include:

Software	Version	Utility
<b>Trimmomatic</b>	0.33	Quality Control
<b>FastQC</b>	0.10.1	Quality Control
<b>Bowtie2</b>	2.1.0	Short read alignment
<b>SAMtools</b>	0.1.19	Alignment file manipulation
<b>BEDtools</b>	2.22.1	Alignment file manipulation
<b>Deeptools</b>	2.0.1	Normalisation
<b>MACS</b>	2.0	Peak caller
<b>R</b>	3.1.3	Data plotting
<b>IGV</b>	2.3.36	Genome browser

Table 2.7 Software

### **2.4 Methods – Dry Lab**

#### **2.4.1 Quality control of sequenced reads**

The quality of sequences generated from the illumina platform were examined using FastQC. FastQC shows the number of reads generated as well as per base sequence quality statistics. Trimmomatic is used through the command line and was used to

trim poor quality reads and to filter reads using sequencing adapters.

#### **2.4.2 Generation of the EquDonk2.0 hybrid genome**

(Joseph G.W. McCarter<sup>1</sup>, Federico Cerutti<sup>2</sup>, Riccardo Gamba<sup>2</sup>, Solomon Nergadze<sup>2</sup>, Francesca Piras<sup>2</sup>, Elena Giulotto<sup>2</sup> and Kevin F. Sullivan<sup>1</sup>)

<sup>1</sup>*Centre for Chromosome Biology, National University of Ireland, Galway*

<sup>2</sup>*Dipartimento di Genetica e Microbiologia, Università di Pavia, Pavia, Italy)*

The horse and donkey genome share 99% sequence identity. Due to the absence of a donkey genome the “EquDonk” genome was created, allowing donkey ChIPSeq data of centromere associated proteins to be mapped with greater accuracy. The EquDonk hybrid genome was created by *de novo* assembly of ChIP and Input datasets from a donkey (Asino Nuovo, name of the donkey individual whose fibroblasts were used in this experiment) ChIPSeq. The assembled centromere sequences were then spliced into the corresponding region of the horse genome.

#### **2.4.3 Alignment of reads to the genome using Bowtie2**

Bowtie2 was used to build an indexed reference genome and align the paired end reads to the genome. The indexed reference genome is in a binary form resulting in a smaller memory footprint. Genomes used in this study include EquDonk2.0, horse (EquCab2) and the Guanzhong donkey. In order to build the indexed reference genome the following command was used:

```
Bowtie2-build -f chr1.fa,chr2.fa,chr3.fa...chrN.fa Genome_bowtie2_index
```

**Command 2.1 Build reference genome**

In this case, -f denotes the input fasta files while Genome\_bowtie2\_index is the name given to the indexed reference genome. Once the indexed reference genome has been built, the trimmed ChIPSeq reads can then be aligned to it using the following command:

```
Bowtie2 -x Genome_bowtie2_index -p8 -1 reads_R1.fastq -2 reads_R2.fastq -S aligned_reads_Genome.sam
```

**Command 2.2 Alignment of paired end reads**

The above command is the bowtie2 default alignment setting. -x indicated the genome to which the reads are aligned, -p8 is a multithreading option and ensures that all possible computer processors are used for a faster alignment. -1 and -2 represent the first and second paired end reads and -S outputs the file in SAM format, followed by the output file name in this case aligned\_reads\_Genome.sam.

#### 2.4.4 File conversion and indexing

Once reads have been aligned to the genome the outputted format is Sequence Alignment format (SAM). This is a tab delimited text format, with both a header and alignment section, showing alignment information such as mapping position. Using SAMtools, a utility that readily allows manipulation of high throughput sequencing data, files in SAM format are converted to their binary equivalent ie. BAM (Binary Alignment/Map) see command 2.3. The compressed BAM files are then sorted by chromosome number and indexed to allow access to specific intervals or locations within the aligned sequence.

```
Samtools view -bSo aligned_reads_Genome.sam aligned_reads_Genome.bam
```

**Command 2.3 Convert SAM to BAM**

In this command, the ‘view’ function converts SAM to BAM, while `-bSo` says binary SAM output. Once in BAM format the reads are sorted by chromosome number, instead of genomic locations using the following command:

```
Samtools sort aligned_reads_Genome.bam aligned_reads_Genome.sorted
```

**Command 2.4 Sorting of BAM file**

Similar to the indexing of the genome carried out in Section 2.4.3, the BAM file is indexed (command 2.5) so that specific locations on the alignment can be searched for more readily.

```
Samtools index aligned_reads_Genome.sorted.bam
```

**Command 2.5 Indexing of BAM file**

#### 2.4.5 Normalising data

Deeptools is a suite of programs for analysis and normalization of next generation sequences. In these studies the ‘bamCompare’ function was used to normalize ChIP reads to input reads. As the name suggests, this function compares two BAM files based on the number of mapped reads as default. Throughout this thesis, the method of normalized employed was subtractive, based on the RPKM (reads per kilobase million) See command 2.6. Files were outputted in a .bedgraph or .bigwig file.

```
bamCompare -b1 ChIP.sorted.bam -b2 Input.sorted.bam --outFileName normalized.bigwig  
-outFileFormat bedgraph --scaleFactorsMethod readCount --ratio subtract --  
normalizeUsingRPKM --binSize 10 --numberOfProcessors max
```

**Command 2.6 Normalization using Deeptools.**

In the above command `-b1` and `-b2` indicate the ChIP and Input in BAM format respectively. The `scaleFactorsMethod` is the method used to scale the samples, in this

case readcount was employed. Generally the genome is partitioned into bins, in this case 10bp, and the read count per bin is counted and a summary is outputted.

#### 2.4.5 Visualising data

Data was visualized using IGV (Integrative Genomics Viewer- Broad Institute) (Thorvaldsdóttir, Robinson, & Mesirov, 2013) and was plotted in R. Before visualization on IGV, normalized data in bedgraph format were converted to a tiled data file (.tdf) using IGV tools, which is essentially a binary bedgraph allowing for faster display in IGV. For plotting bedgraphs in R, the *Sushi* package was used with the following command:

```
pdf ("File_name.pdf")
par (mar=c (0.1, 5, 0.1, 5), oma=c (4,0,4,0))
plotBedgraph (bedgraph, chrom, chromstart, chromend, color= "gray", transparency=1,
lwd = 0.01, linecolor = "black", range = c (0,1000))
mtext ("Heading",side=3,col="black",line=0.5, cex=0.2)
mtext ("Y-axis label",side=2,col="gray40",line=0.5, cex=0.2)
mtext ("X-axis label", side=1, col="gray40", line=0.5, cex=0.2)
axis (1, col="gray40", col.axis="gray40",col.ticks="gray40", cex.axis=0.4, lwd=0.5)
axis (2, col = "gray40", col.axis = "gray40", col.ticks = "gray40", cex.axis=0.4,
lwd=0.5)
dev.off ()
```

**Command 2.7 Plotting bedgraphs in R**

#### 2.4.6 MACS peakcalling

MACS (Model based analysis for ChIPSeq) (Zhang et al., 2008) is used for identifying regions of enrichment in ChIPSeq data. Peaks are called using the command:

```
Macs2 callpeak -t chip.sorted.bam -c input.sorted.bam --call-summits -n subpeaks
```

**Command 2.8 MACS peak calling**

#### 2.4.7 Read count extraction

The read counts from the CENP-A binding domains were extracted using the SAMtools mpileup command. The genomic loci is specified, here in chrN from 5,000-200,000nt, and the read counts are extracted from the sorted BAM file using the following command:

```
Samtools mpileup -r chrN: 5,000-200,000 alignment.sorted.bam >
chrN_mpileup_5000to200000.txt
```

**Command 2.9 SAMtools mpileup read extraction**

The SAMtools mpileup command output contains information such as base qualities, read bases and alignment mapping qualities, which are unnecessary for the purpose of viewing the file. In order to extract just the necessary information, in this case column

2 and column 4, which contain the genomic position and the read count value, the following command was used:

```
Awk '{print$2, $4}' chrN_mpileup_5000to200000.txt > chrN_mpileup_5000to200000.prn
```

**Command 2.10** Extracting columns from a file

#### ***2.4.8 Relative abundance of CENP-A at centromere domains***

The relative abundance of CENP-A at centromeres was calculated by extracting the read counts at centromere domains as outlined in Section 2.4.7. The average CENP-A abundance was calculated by adding the read counts of each of the centromeres on the autosomes and dividing by the number of domains. The read count of each individual centromere was then divided by the average and the relative abundance of CENP-A across each of the satellite free centromeres was plotted.

#### ***2.4.9 Identification of centromeres in the Guanzhong donkey***

The CENP-A ChIPSeq reads from Asino Nuovo as mentioned in Section 2.4.2 were aligned to the Guanzhong donkey genome using bowtie2 in the same manner as outlined in Section 2.4.3. The domains which the CENP-A reads mapped to were identified by direct inspection using IGV (Thorvaldsdóttir et al., 2013), Section 2.4.5. The regions which the CENP-A ChIPSeq mapped were extracted. The BAM file was then converted back to a SAM file using the SAMtools function view, described in Section 2.4.4. The SAM file was then converted into a fastq file using command 2.11. The fastq sequences were then aligned to the EquDonk genome and the corresponding centromere was identified.

```
cat file_name.sam | grep -v ^@ | awk 'NR%2==1 {print "@"$1"\n"$10"\n+\n"$11}' > file_name.fastq
```

**Command 2.11** Converting a SAM file to a fastq file

#### ***2.4.10 Analysis of repetitive sequences***

The sequences associated with the donkey CENP-A ChIPSeq reads in the Guanzhong donkey, EquDonk and EquCab were extracted using the following command:

```
samtools faidx Genome.fa "CHR":X-Y > Genome_Chrr
```

**Command 2.12** Sequence extraction

Analysis of repetitive sequences across these domains was performed with Repeatmasker (Smit, Hubley, & Green, 2013) using the following command:

```
Repeatmasker Genome_Chr > Genome_Chr_Repeatmasked
```

#### Command 2.13 Repeatmasker command

This command outputted a summary table, which showed the abundance of repetitive elements within the domains specified.

#### ***2.4.11 Schematic representation of centromere comparisons***

Sequences extracted using command 2.12 were then blasted (Altschul, Gish, Miller, Myers, & Lipman, 1990) against each other and an excel file of the alignments were generated. The genomic coordinates were identified by inputting 30bp of the aligned sequence into the *find motif* function of IGV. Regions of insertion were identified by gaps in the alignment, duplications were identified when genomic coordinates of one genomic assembly mapped to the same coordinates on the other genomic assembly more than once. Once regions of homology, sequence insertion and duplication were established, their coordinates were saved in separate text files, which were then used in R to generate the schematic using the command below:

```

EquDonk_CenX <-read.table ("/file_path/Equdonk_CenX_homologous_regions.txt")
EquDonk_CenX_Insertions <-read.table ("/file_path/Equdonk_CenX_insertions.txt")
EquDonk_CenX_Duplications <-read.table ("/file_path/Equdonk_CenX_insertions.txt")
**Import horse and Guanzhong donkey sequence in the same manner

pdf ("File_name.pdf")
start<- EquDonk_CenX [1:9,1]
end<- EquDonk_CenX[1:9,2]
plot (x= start, y= end, type= 'n', bty= 'n', yaxt= 'n',ylab= '', xlab= 'Position',
xlim= range (c (26286390, 26535316)), ylim= c (0, 1), bty="o")
rect (xleft= start, xright= end, ybottom= 0.1, ytop= 0.2, border= NA, col="blue")
par (new=T)
start<- Guanzhong_CenX[1:9,1]
end<- Guanzhong_CenX[1:9,2]
plot (x= start, y= end, type= 'n', bty= 'n', yaxt= 'n',ylab= '', xlab= '', xlim=
range (c (0, 248926)), ylim= c (0, 1), yaxt="n")
rect (xleft= start, xright= end, ybottom= 0.6, ytop= 0.7, border= NA, col="blue")
axis (3)
par (new=T)
start<- EquDonk_CenX_Insertions[1:2,1]
end<- EquDonk_CenX_Insertions [1:2,2]
plot (x= start, y= end, type= 'n', bty= 'n', yaxt= 'n',ylab= '', xlab= '', xlim=
range (c (26286390, 26535316)), ylim= c (0, 1), bty="o")
rect (xleft= start, xright= end, ybottom= 0.1, ytop= 0.2, border= NA, col="red")
par (new=T)
start<- Guanzhong_CenX_Insertions[1:5,1]
end<- Guanzhong_CenX_Insertions [1:5,2]
plot (x= start, y= end, type= 'n', bty= 'n', yaxt= 'n',ylab= '', xlab= '', xlim=
range (c (0, 248926)), ylim= c (0, 1), yaxt="n")
rect (xleft= start, xright= end, ybottom= 0.6, ytop= 0.7, border= NA, col="red")
axis (3)
par (new=T)
start<- Guanzhong_CenX_Duplications[1:4,1]
end<- Guanzhong_CenX_Duplications [1:4,2]
plot (x= start, y= end, type= 'n', bty= 'n', yaxt= 'n',ylab= '', xlab= '', xlim=
range (c (0, 248926)), ylim= c (0, 1), yaxt="n")
rect (xleft= start, xright= end, ybottom= 0.6, ytop= 0.7, border= NA,
col="aliceblue")
axis (3)
par (new=T)
start<- EquDonk_CenX_Duplications[1:5,1]
end<- EquDonk_CenX_Duplications[1:5,2]
plot (x= start, y= end, type= 'n', bty= 'n', yaxt= 'n',ylab= '', xlab= '', xlim=
range (c (26286390, 26535316)), ylim= c (0, 1), bty="o")
rect (xleft= start, xright= end, ybottom= 0.1, ytop= 0.2, border= NA,
col="aliceblue")
dev.off ()

```

**Command 2.14 R command example for generating schematic representation of centromere domains**

## **Chapter 3 Preparation of CENP-A antibody**

### **3.1 Introduction**

ChIPSeq is a technique that is employed throughout this thesis and the utility of this technique depends largely on the antibodies used to pull down the target protein and associated DNA. Without highly specific affinity reagents, the ChIPSeq will be noisy or perhaps not work at all. In this chapter, I will discuss the generation of an equine optimised sheep CENP-A sera, its characterisation, purification and application in ChIPSeq. Using this purified polyclonal antibody, I will quantify the abundance of CENP-A at the donkey satellite free centromeres.

### **3.2 Preparation of CENP-A gene**

In order to prepare an antibody reactive to equine CENP-A, horse CENP-A sequence was first identified using the Ensembl genome browser, ENSECAP00000013849. A codon optimization procedure was carried out on the coding sequence to ensure optimal expression in *E.coli* strains. Restriction sites were incorporated for ease of cloning and a synthetic CENP-A gene was ordered (Eurofins Genomics).

The synthetic horse CENP-A gene was liberated from its commercial backbone by restriction digest, gel purified and ligated into the Gateway entry vector pENTR4. pENTR4-CENP-A served as the entry clone for recombining CENP-A into any Gateway expression vector. N terminally Histidine-tagged pDEST17 was chosen as the expression vector. This vector already contained an ATG start codon upstream of the 6Xhis tag as well as a Shine-Dalgarno RBS (ribosomal binding site) upstream of the ATG ensuring optimal translation initiation in *E.coli*. BL21-AI were chosen as the expression strain due to their tightly controlled T7 RNA polymerase expression and their deficiency in Ion and OmpT proteases, reducing degradation of heterologous expressed proteins (Bilgimol et al., 2015; Studier, 2005).

### **3.3 Protein expression and purification**

CENP-A expression was regulated by L-arabinose induction. Expression of T7 RNA polymerase in the host strain BL21-AI is regulated by the araBAD promoter ( $P_{\text{BAD}}$ ), which in turn is regulated by the product of the AraC gene (Ogden, Haggerty, Stoner, Kolodrubetz, & Schleif, 1980; R Schleif, 1992). L-arabinose forms a complex with the transcriptional regulator AraC, prior to L-arabinose induction, the AraC dimer forms a 210bp loop by linking the  $O_2$  and  $I_1$  half sites of the araBAD operon. L-

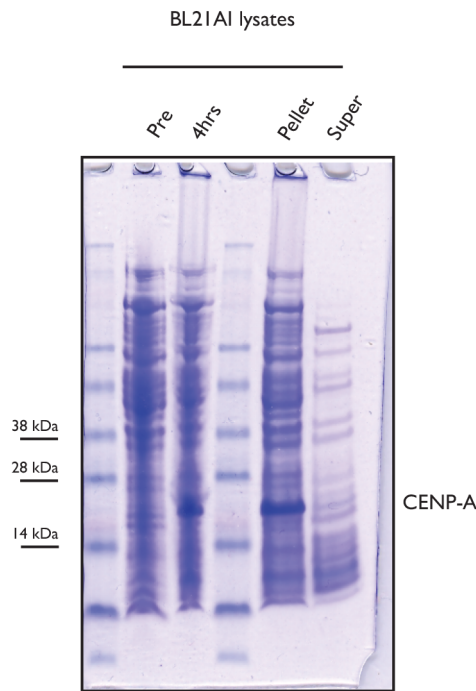


arabinose binds AraC resulting in the release of the DNA loop from the O<sub>2</sub> and I<sub>1</sub> sites and its association with the I<sub>2</sub> site, triggering transcription activation. This process is mediated by the cAMP activator protein (CAP)-cAMP which stimulates the binding of AraC to I<sub>1</sub> and I<sub>2</sub> (Robert Schleif, 2010).

Cell lysates, both before and after L-arabinose induction, were prepared and examined by SDS PAGE and Coomassie staining, Figure 3.1. A band corresponding in size to that of CENP-A was observed in the commasie gel 4 hours post induction, indicating no leaky expression before induction and robust CENP-A expression after L-arabinose addition.

### ***3.3.1 Solubility assay***

CENP-A solubility was examined post induction by SDS PAGE and Coomassie staining. Cell lysates were resuspended in native extraction buffer and sonicated, the lysates were spun down, with soluble protein found in the supernatant (super) and insoluble proteins remaining in the pellet, Figure 3.1. Recombinant CENP-A was found to be insoluble remaining in the pellet. Efforts to yield soluble CENP-A, inducing cells for a shorter time, inducing cultures at lower temperatures and expressing CENP-A as a pDEST15 GST fusion protein, failed to result in a soluble CENP-A fraction.



**Figure 3.1 Expression and solubility of recombinant horse CENP-A.** To test expression, transformed cells were grown to mid log phase and induced with L-arabinose. Samples were taken before induction and four hours post, spun down, resuspended in Lamelli sample buffer and boiled for 10 minutes (lysates). SDS PAGE analysis of lysates before induction (pre) shows no obvious leaky CENP-A expression. After induction (4hrs), a band of approximately 16kDa is observed, corresponding in size to CENP-A. Examination of CENP-A solubility shows the protein remaining in the pellet, with little or no soluble fraction in the supernatant (super).

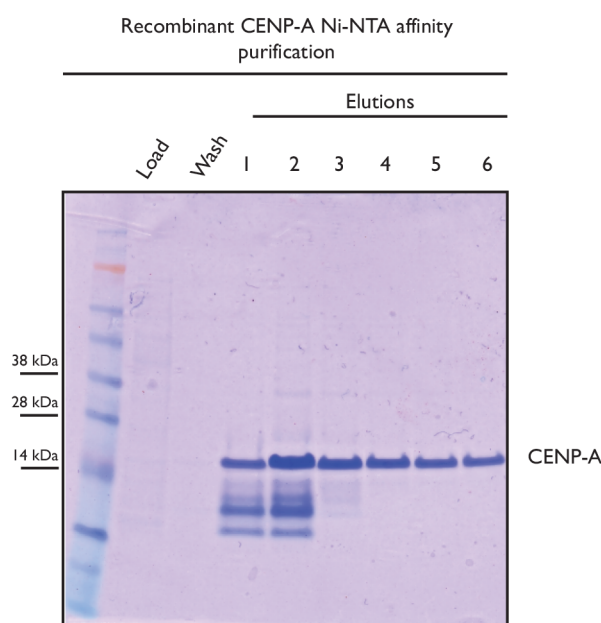
### 3.3.2 Inclusion bodies

Recombinant CENP-A was accumulating as insoluble aggregates or inclusion bodies within the bacterial cells. Formation of inclusion bodies in *E.coli* often occurs when recombinant protein is expressed at high levels (Kane & Hartley, 1988) which is why a shorter induction time was implemented in a bid to yield soluble protein. Inclusion bodies are densely packed, denatured proteins that have no biological activity. A denatured antigen as a target for immunogenic response raises questions about the seras ability to recognize the antigen in native biochemical applications, which is why the techniques described in Section 3.3.1 were employed in a bid to yield soluble protein. Given, the inability to express soluble recombinant CENP-A protein, the inclusion bodies were isolated and purified for immunization. In terms of isolation and obtaining a pure protein population, inclusion bodies have advantages when compared with soluble proteins: inclusion bodies are less readily degraded and resistant to cellular proteases, they are bigger and more dense than cellular contaminants allowing isolation by differential centrifugation and the homogeneity of the protein means fewer purification steps to obtain pure protein (Singh & Panda, 2005).

CENP-A inclusion bodies were isolated by sonication of cell lysates and multiple washes in buffer (50mM Tris pH 7.5, 100mM NaCl, 1mM EDTA pH8, 1mM Benzamidine, 5mM Beta-mercaptoetanol) supplemented with 1% Triton-X, which dissolved cell membranes and solubilizes membrane proteins, eliminating any soluble material in the fraction. The inclusion bodies were solubilized in buffer containing 7M Guandium-HCl, 20mM Tris pH7.5, 10mM DTT and DMSO.

### 3.3.3 CENP-A purification

Solubilised CENP-A inclusion bodies were diluted 1:20 in a buffer containing 7M Urea, 50mM Tris pH7.5 and 50mM NaCl and incubated with Ni-NTA agarose beads on a roller at room temperature, to allow binding of the 6-His CENP-A to the Ni-NTA beads. The suspension was then loaded onto a column and all the fractions were collected (load, wash and elutions). Fractions 4, 5 and 6 were pooled, Figure 3.2, and dialysed into PBS resulting in a fine precipitate. The precipitate was sent for sheep immunization.



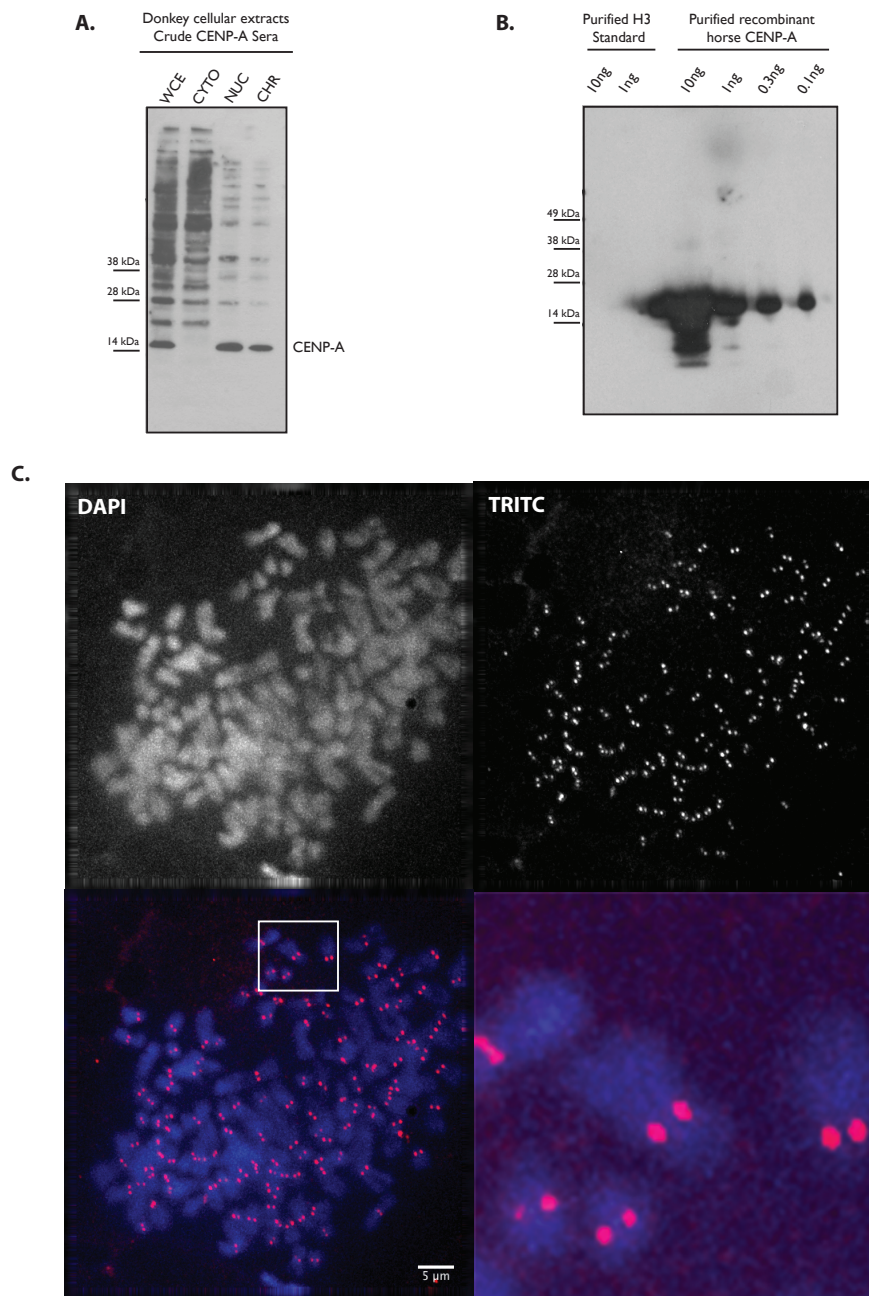
**Figure 3.2 Nickel affinity purified CENP-A fractions.** SDS PAGE analysis of fractions collected from the affinity purification of recombinant CENP-A shows no protein loss in the load or wash fractions. The first 3 elution fractions, appear to show degradation of the protein, for this reason, the cleanest and most intact protein fractions 4, 5 and 6 were pooled and sent for sheep immunization.

### 3.4 Antibody characterization

Five batches of sheep sera were received: 1x pre immunization, 4x post immunization (bleed 1, 2, 3 and 4). Donkey cell extracts were prepared by lysing fibroblasts (whole cell extract/WCE) and spinning the lysate down, the soluble fraction contained the

cytoplasm (CYTO) and the pellet contained the insoluble fraction and nuclei (NUC). A fraction of the nuclei were digested with micrococcal nuclease yielding a chromatin (CHR) fraction. Antibody response was only observed with the third and fourth bleed, with data from bleed three only shown in this thesis. A band corresponding in size to CENP-A was observed in the whole cell, nuclear and chromatin extracts, with no detectable signal in the cytoplasmic fraction as expected for CENP-A, Figure 3.3, A. Western blot analysis of the cell lysates shows high levels of background. To examine the specificity of the serum, affinity for histone H3 was examined to determine if there was cross reactivity, Figure 3.3 B. Western blot analysis shows the serum is specific for CENP-A with no detectable affinity for histone H3.

The serum's application in immunofluorescence was also characterized, Figure 3.3 C. Immunofluorescence was carried on metaphase spreads of donkey fibroblasts; mitotic cells were gathered from an asynchronous population by elutriation, swollen in KCl and spun onto slides using a cytospin. The chromosomes were fixed using 100% MeOH. Both bleed 3 and 4 showed comparable results with distinct double foci visible at the primary constriction of the chromosomes and no significant background.

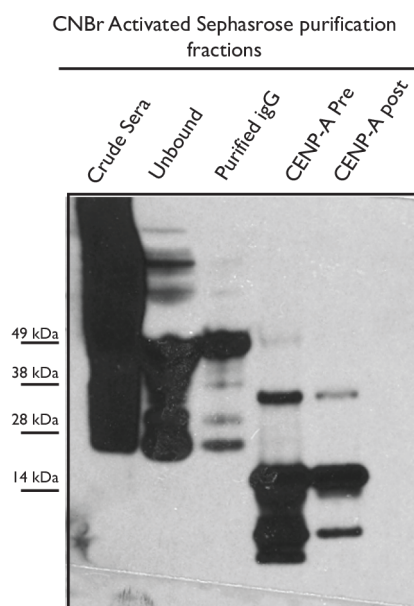


**Figure 3.3 CENP-A sera characterization.** A) Western blot analysis was carried out using 1:5000 dilution of serum on donkey cell extracts showing a band corresponding in size to CENP-A in the whole cell (WCE), nuclear (nuc) and chromatin (chr) fractions, with no signal observed in the cytoplasm (cyto) as expected for CENP-A. B) To examine sera affinity for recombinant CENP-A and Histone H3, western blot was carried out against purified protein at different concentrations. 0.1ng of CENPA produced a readily detectable band, no cross-reactivity was seen with histone H3 at this serum dilution. C) Immunofluorescence using a 1:250 dilution of the serum on metaphase spreads show punctate foci at the primary constriction of the chromosome, indicating the sera recognizes CENP-A.

### **3.5 Antibody affinity purification**

The serum shows reactivity with CENP-A but western blot analysis shows non-specific binding to other proteins, particularly in the whole cell and cytoplasmic extracts. In order to obtain a reagent suitable for ChIP-qPCR and ChIPSeq the antibody was affinity purified. Antigen affinity purification results in pure antibodies with the least amount of cross reactivity depending on the method used. One can expect that only ~ 1-10% of the antibodies in polyclonal antisera are specific for the immunized antigen, the other >90% are irrelevant host derived antibodies.

The purified CENP-A antigen was coupled to cyanogen bromide activated sepharose beads in a carbonate buffer with 0.5M salt, as described in Section 2.2.2.11. The coupling reaction occurs through primary amines. Buffers such as Tris and other buffers containing amino groups are avoided, as these couple to the sepharose. High salt concentration minimizes the formation of protein aggregates and stops protein-protein adsorption. The efficiency of CENP-A coupling to the beads was examined in two ways. Direct examination of the protein solution in coupling buffer before (CENP-A pre) and after coupling (CENP-A post) showed that CENP-A was somewhat depleted by coupling, indicating that coupling was successful, Figure 3.4. Secondly, antibody was affinity purified using the CENP-A-sepharose matrix. The serum before affinity purification (Crude Sera), the unbound fraction (Unbound) and after purification (purified IgG) were all examined by western blot, Figure 3.4. While only a fraction of IgG applied to the column was bound (Purified IgG), a bound antibody was eluted from the column with 200mM glycine pH2.8 after extensive washing in PBS.



**Figure 3.4 Examination of the efficiency of Sera purification using CNBr activated Sepharose.** Fractions were taken before and after coupling of recombinant CENP-A to the CNBr beads (CENP-A pre/CENP-A post) and shows that the beads bound the CENP-A protein, with significantly less protein present in the CENP-A post (adsorption fraction). Analysis of the serum purification (Crude Sera, Unbound, Purified Sera) shows that a large proportion of IgG was removed in the wash steps (unbound) when compared with the IgG in the purified fraction. The purified IgG appears relatively clean, with the bigger bands seen in the unbound fraction not present.

Elution	A280	Mg/ml
1	0.946	0.72
2	2.611	1.98
3	2.223	1.69
4	0.912	0.69
5	0.391	0.29
6	0.177	0.134
7	0.134	0.102
8	0.165	0.125
9	0.099	0.075

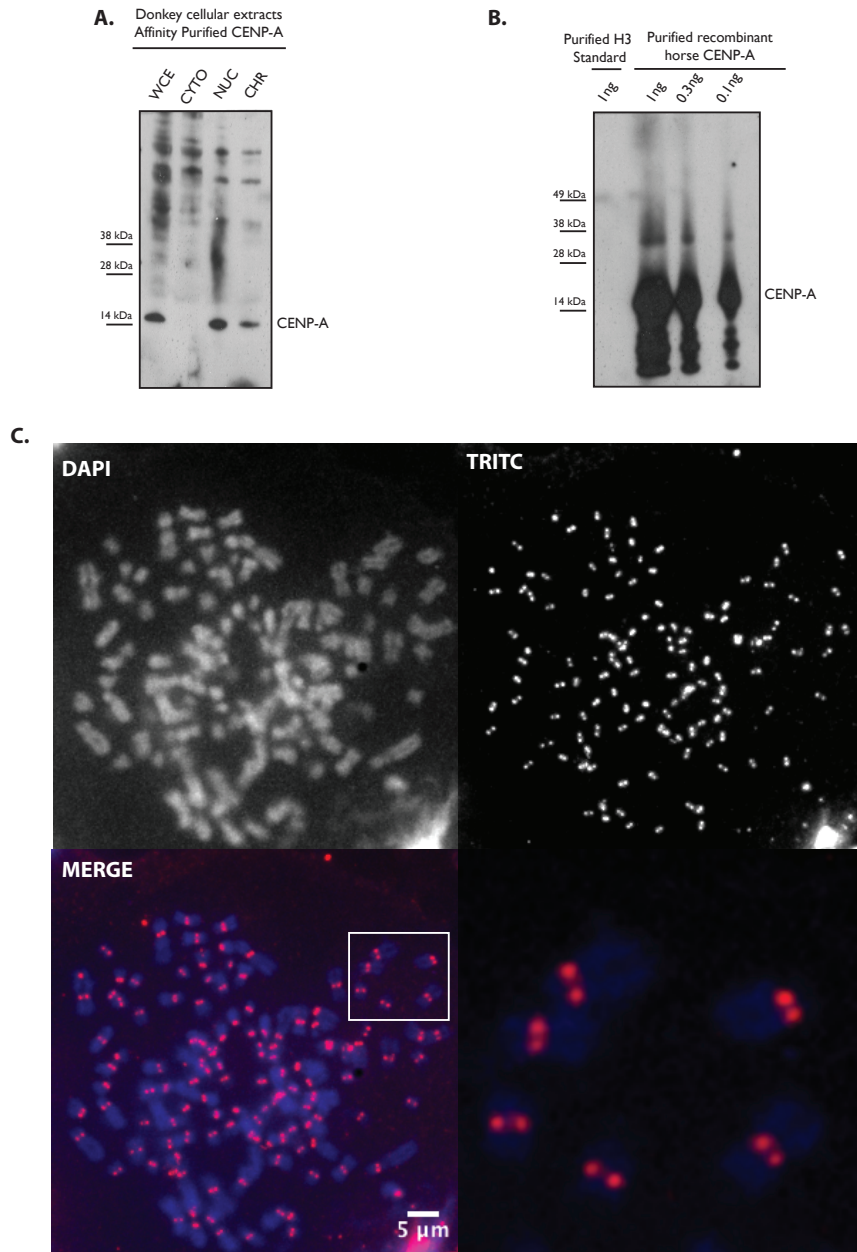
**Table 3.1 A280 values and concentration of eluted purified CENP-A antibody**

The OD<sub>280</sub> of the eluted fractions was measured and the mg/ml concentration was estimated using the formula  $OD_{280}/\epsilon \times \text{molecular weight}$  ( $\epsilon = 1.36$  for IgG). The most concentrated elutions 1, 2, 3 and 4 were pooled and dialyzed into PBS overnight at 4°C overnight. The antibody was then aliquoted, snap frozen and stored at -80°C.

### ***3.5.1 Validation of the affinity purified Sera***

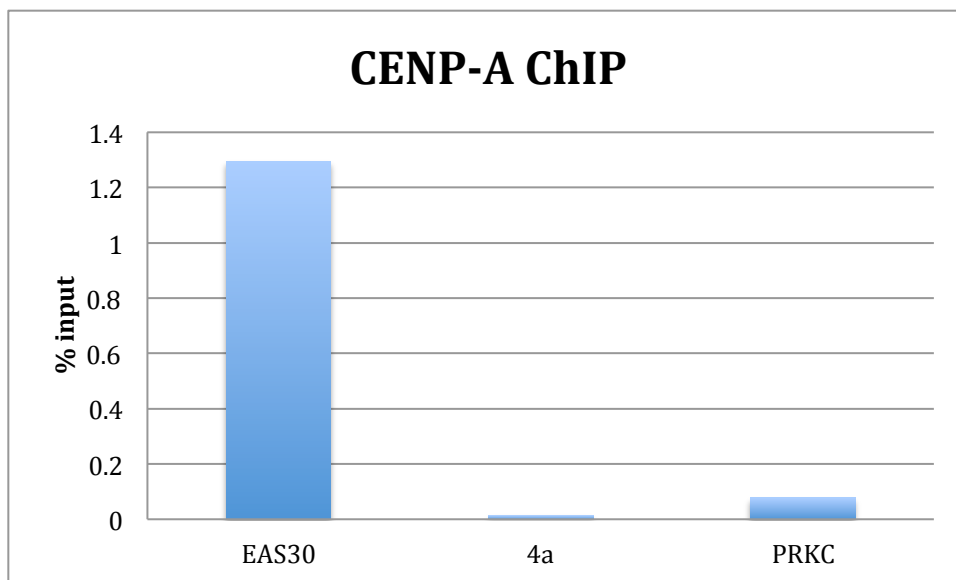
The affinity purified CENP-A was characterized in the same manner as the crude serum. Western blot analysis shows that the antibody remains reactive against protein corresponding in size to CENP-A, as well as producing a significantly cleaner blot with decreased levels of background, Figure 3.5 A. To examine antibody cross reactivity with histone H3 western blot analysis was performed using a H3 standard of 1ng and dilutions of purified recombinant horse CENP-A, 1ng, 0.3ng and 0.1ng, Figure 3.5 B. No signal was detected in the histone H3 fraction and the antibody readily detects the recombinant CENP-A dilutions, validating the antibody affinity for CENP-A and application for use in dissecting centromere organisation the equids. Immunofluorescence analysis on donkey metaphase spreads shows the antibody localization at the primary constriction of the chromosome, with two punctate foci clearly visible on paired sister chromatids, Figure 3.5 C. No other chromosomal staining is observed.





**Figure 3.5 Purified Sera characterization.** A) Western blot analysis was performed, using 1:2000 dilution of the 0.68mg/ml purified antibody, on donkey cell extracts and showed decreased levels of background when compared to the crude sera, with a band corresponding in size to CENP-A in the whole cell (WCE), nuclear (nuc) and chromatin (chr) fractions. B) The antibody affinity for recombinant histone H3 and CENP-A was examined by western blot, with no affinity for histone H3 observed. C) Immunofluorescence using a 1:100 dilution of the affinity purified antibody on metaphase spreads, showed decreased levels of background when compared with the crude sera, with punctate foci clearly visible at the primary constriction of the chromosome, indicating the sera recognizes CENP-A.

The antibody was validated for use in ChIP using a qPCR assay. 1ul of 0.68mg/ml affinity purified antibody was used per  $1 \times 10^6$  chromatin cell equivalents. Donkey fibroblasts were formaldehyde crosslinked and sheared by sonication. The chromatin was isolated by centrifugation and the antibody was incubated with the chromatin overnight at 4 °C to allow CENP-A binding. Protein G sepharose beads were then added to recover the antibody and antibody associated complexes. The DNA was decrosslinked at 65°C and subsequently proteinase K and RNaseA treated. The DNA was then cartridge purified using the *Qiagen PCR purification kit* and analyzed for centromeric enrichment by qPCR. qPCR was carried out using primers within the unique sequence centromere of donkey chromosome 30 (Eas30), a single copy gene (PRKC) and at the horse neocentromere associated sequence on chromosome 11 (4a). Enrichment can be calculated as the ratio of centromere recovery to other single copy sequences in the ChIP experiment. Averaging the 4a and PRKC values, enrichment of centromere sequence is 28.7 fold in this experiment. As shown in Figure 3.6, clear centromeric enrichment can be seen at chromosome 30 with very little signal present in the negative control region. This suggests that the antibody is specific to the centromere and is suitable for use in ChIPSeq.



**Figure 3.6 ChIP qPCR analysis of CENP-A immunoprecipitation.** There is centromeric enrichment at EAS30, with a recovery of 1.29%. The levels of background are low at the horse Chr11 centromere, 86 times less than the EAS30 centromere and the single copy gene, PRKC ~17 times less.

### 3.6 CENP-A ChIPSeq

To examine antibody application in ChIPSeq, ChIP was initially carried on primary horse fibroblasts (HSFG-Horse Skin Fibroblasts G). This was carried out in the same manner highlighted above for the ChIP qPCR. The DNA was sent to *IGA Technology Services* for sequencing. 80ul of crude sera was used for  $100 \times 10^6$  cells.

File name	Input/ChIP	Sequence length (bp)	Reads
<b>1_ACACGA_L004_R1_001.fastq</b>	Input	150	23660527
<b>1_ACACGA_L004_R2_001.fastq</b>	Input	150	23660527
<b>2_GTGGCC_L004_R1_001.fastq</b>	ChIP	150	17872205
<b>2_GTGGCC_L004_R2_001.fastq</b>	ChIP	150	17872205

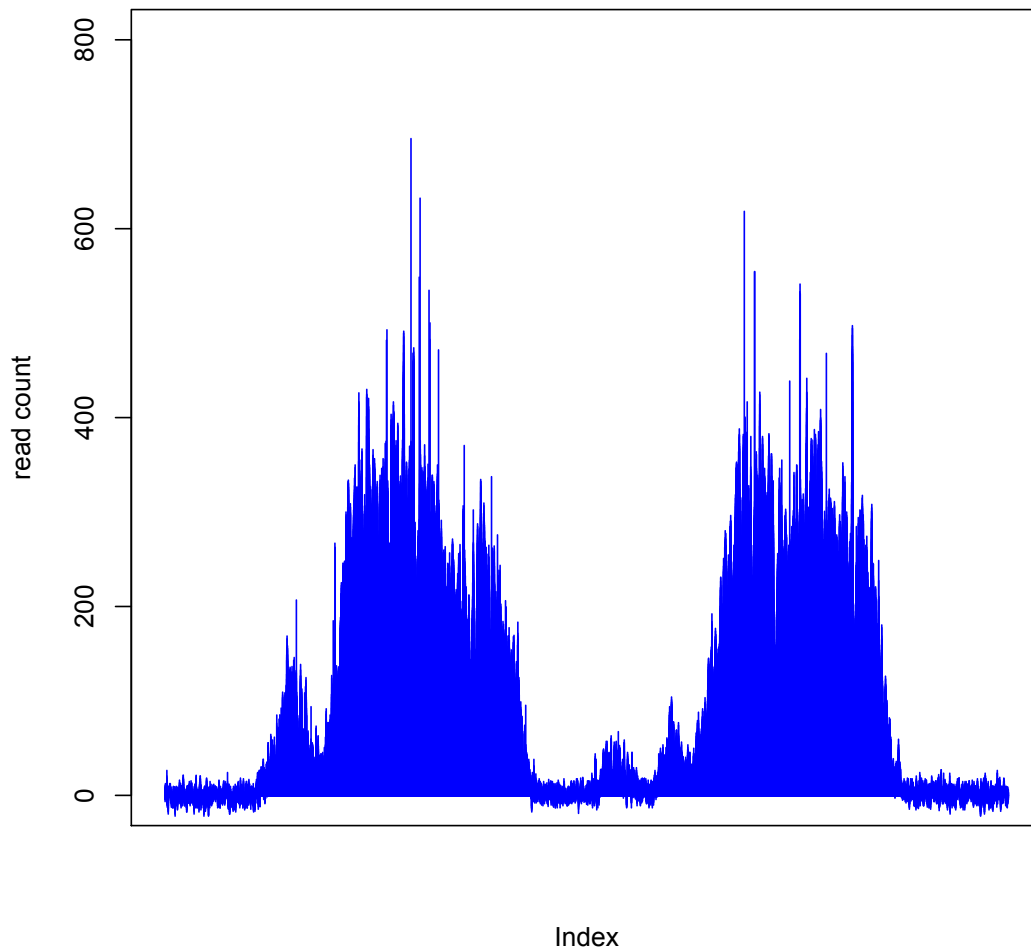
Table 3.2 Sequence details for Horse CENP-A ChIPSeq.

Read quality was examined using *FASTQC*. Poor quality reads were trimmed before alignment using *Trimmomatic*. Since the reads generated were paired end, both libraries R1 and R2 were trimmed together, so that in case of one end being bad quality, the other end was also trimmed. The reads were scanned in four base sliding window and any window containing a Phred score of less than 15 was excluded. Any trimmed reads less than 75bp were also excluded. The trimmed reads were aligned to horse genome *EquCab2.0*. The quality of the immunoprecipitation was examined by measuring signal enrichment at centromeres using FRIP (fraction of reads in peaks). FRIP calculates the percentage of reads that are associated with significantly enriched domains ie. the unique sequence centromere, compared to rest of the genome. The FRIP score for the HSFG CENP-A ChIPSeq was 2.07% well above the 1% threshold that defines a valid ChIPSeq experiment.

For visualization of the aligned data, normalization was performed against the input DNA using the *bamcompare* parameter in the *Deeptools* suite as described in Section 2.4.5 and the output was visualized using the R package *Sushi*.

Centromere enrichment was observed at horse chromosome 11, with two Gaussian like peaks observed, Figure 3.7. Based on Purgato et al., 2015, the two peaks correspond to the centromere on each homolog. Taken together these experiments validate the utility of this antibody in ChIPSeq.

### HSFG CENP-A ChIP Chr11



**Figure 3.7 Horse CENP-A ChIPSeq.** Centromeric enrichment is observed on the horse neocentromere at chromosome 11. Two peaks are observed corresponding to the centromere domain on each of the homologs of chr11, thereby validating the utility of this antibody in ChIPSeq.

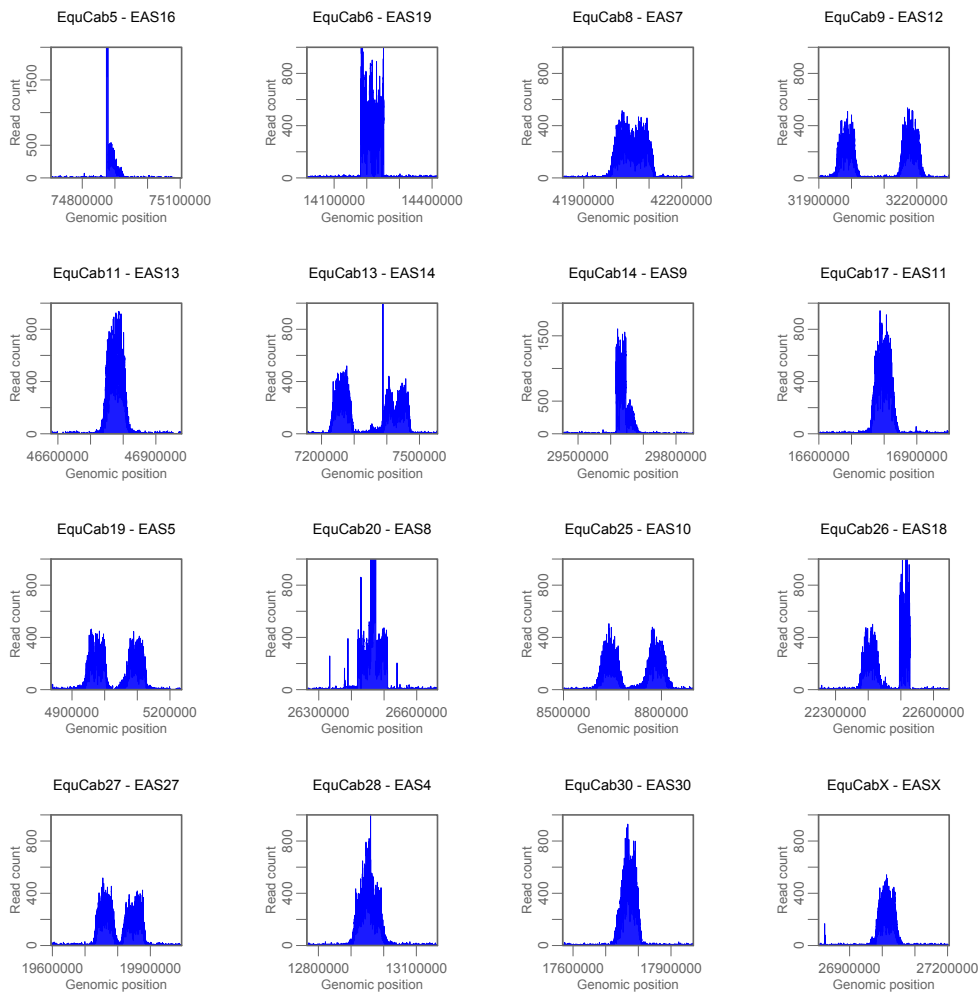
To further examine the immunoprecipitation utility of this antibody, ChIPSeq was carried out in the same manner as described previously using immortalized donkey fibroblasts. In this experiment, 100ul of affinity-purified antibody (described in Section 2.2.2.11) was used per  $100 \times 10^6$  cells. The sequence details are shown in table 3.3, the sequencing reaction was carried out across two lanes for both the ChIP (L001, L005) and Input (L001, L003).

The same bioinformatics pipeline as described above was employed for the immortalized donkey CENP-A ChIPSeq analysis. The reads generated in this experiment were mapped to the “EquDonk” genome. EquDonk is a hybrid genome described in section 2.4.2, whereby de novo assembly of centromere domains was carried out using donkey CENP-A ChIPSeq reads followed by insertion into the

corresponding region in the horse genome (work of Joseph GW McCarter and collaborators in Pavia). This allows proper alignment to bonafied donkey centromere sequences in a genomic context. Centromere enrichment was observed across all 16 unique sequence centromeres, with a number of different distributions observed, Figure 3.8. Each centromere is labeled according to its chromosome of origin on the horse scaffold (Eca) and the corresponding donkey chromosome (Eas). A single gaussian-like profile was observed in the case of Eca11/Eas13, Eca17/Eas11, Eca30/Eas30 and EcaX/EasX, with profiles showing the highest signal intensity in the center of the domain with gradual dissipation approaching the domain boundaries. There were also examples of multi domain CENP-A binding in the case of Eca9/Eas12, Eca13/Eas14, Eca19/Eas5, Eca25/Eas10, Eca26/Eas18 and Eca27/Eas27. The presence of two domains of CENP-A enrichment corresponds to the CENPA binding domain on each homolog. Some of the domains also contained a spike like profile, as well as a Gaussian-like distribution in the case of Eca5/Eas16, Eca14/Eas9, Eca20/Eas8 and Eca26/Eas18. Eca6/Eas19 contained a spike like distribution, with a strong CENP-A signal that does not dissipate approaching the centromere boundary. These centromeres show evidence of sequence amplification as seen by corresponding spikes in the input reads and direct analysis (E. Giulotto, unpublished).

File name	Input/ChIP	Sequence length (bp)	Reads
<b>1 TCGGATTC L001 R1 001.fastq</b>	Input	150	5232871
<b>1 TCGGATTC L001 R2 001.fastq</b>	Input	150	5232871
<b>1 TCGGATTC L003 R1 001.fastq</b>	Input	125	5731115
<b>1 TCGGATTC L003 R2 001.fastq</b>	Input	125	5731115
<b>3 GAACCTTC L001 R1 001.fastq</b>	ChIP	150	8591664
<b>3 GAACCTTC L001 R2 001.fastq</b>	ChIP	150	8591664
<b>3 GAACCTTC L005 R1 001.fastq</b>	ChIP	125	4943930
<b>3 GAACCTTC L005 R2 001.fastq</b>	ChIP	125	4943930

**Table 3.3** Sequence details for donkey CENP-A ChIPSeq.



**Figure 3.8 Centromere profile across the 16 donkey unique sequence centromeres.** Note the different read count scales on the y-axis.

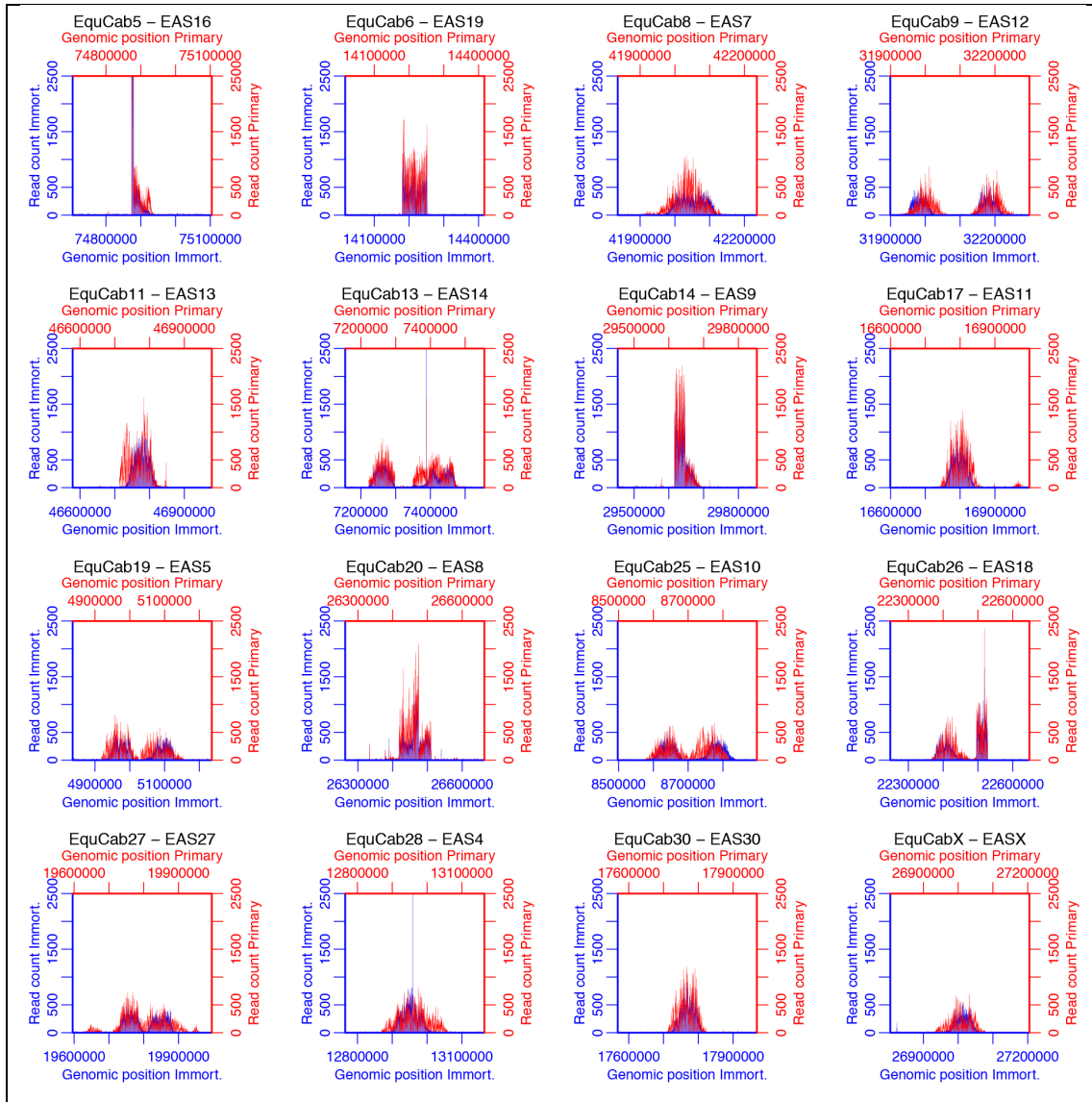
The unique sequence centromeres in the donkey were originally characterized by ChIPSeq, using a peptide CENP-A antibody in primary donkey fibroblasts (Nergadze et al, in preparation). This CENP-A ChIPSeq dataset was compared against the ChIPSeq data generated using the sheep CENP-A antibody in the immortalized donkey fibroblasts derived from the primary cell line used in that study. In order to examine the co-localization of the two CENP-A datasets, the ChIPSeq reads were superimposed, Figure 3.9. The immortalized read counts are shown in blue, while the primary are shown in red.

Superimposition of the CENP-A ChIPSeq data from the immortalized and the primary fibroblasts shows that CENP-A tends to occupy a larger footprint in primary fibroblasts (Table 3.4). The centromere boundaries were defined by direct inspection of the centromere domains using IGV (Thorvaldsdóttir et al., 2013) and are shown in

appendice II. The average size of the CENP-A binding domain in the primary cell line is 120kb or 117kb for individual alleles. In the immortalized cell line, the average centromere domain is 91kb or 93kb for individual alleles. This observation supports the “founder effect” hypothesis. In the primary cell line, the ChIPSeq shows the CENP-A binding domain of a collection of heterogenous fibroblasts. The immortalized cell line was derived from a single cell clone from this population and shows the propagation of one diploid pair of centromere molecules from the original mother cell. These data also show that centromere position is tightly regulated since the centromere domain in the immortalized cell line does not recapitulate the broader distribution observed in the primary cells after ~35 population doublings. Figure 3.10 details Eca11/Eas13 and Eca13/Eas14 centromere superimpositions showing the different centromere domain size in the two cell lines.

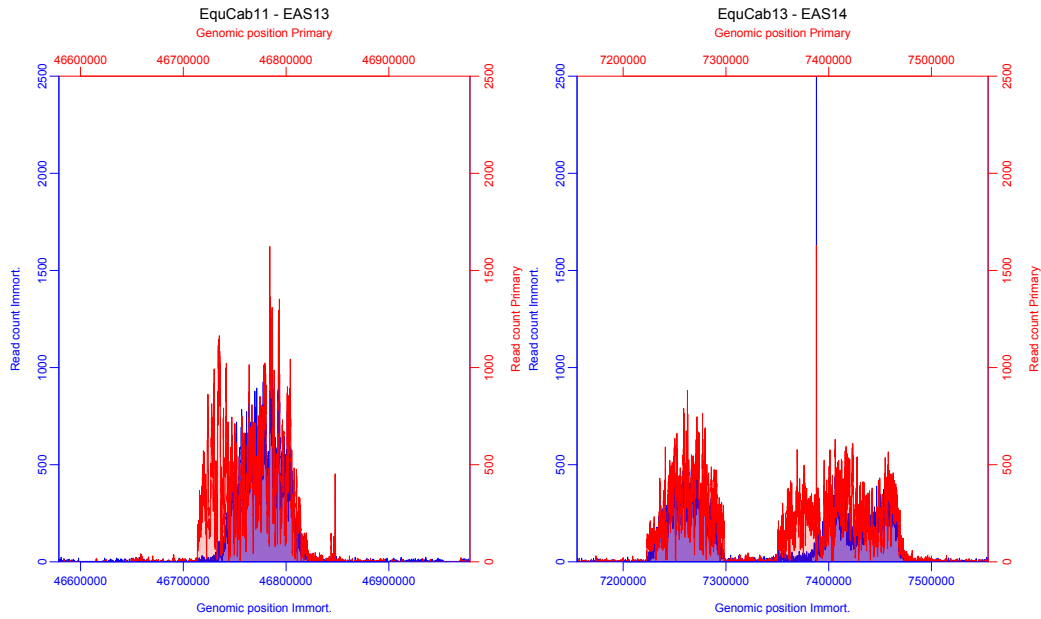
Cen	Peak	Primary domain span (kb)	Immortalised domain span (kb)	Difference (kb)
Eca5/Eas16	1	66.1	56.2	-9.9
Eca6/Eas19	1	71.4	71.4	0.0
Eca8/Eas7	1	220.1	148.8	-71.4
Eca9/Eas12	1	122.4	78.0	-44.4
Eca9/Eas12	2	111.7	84.4	-27.3
Eca11/Eas13	1	137.1	98.5	-38.6
Eca13/Eas14	1	75.7	74.9	-0.7
Eca13/Eas14	2	126.3	125.1	-1.1
Eca14/Eas9	1	75.5	75.5	0.0
Eca17/Eas11	1	117.4	88.9	-28.6
Eca19/Eas5	1	107.2	96.1	-11.1
Eca19/Eas5	2	129.0	99.8	-29.2
Eca20/Eas8	1	92.2	92.2	0.0
Eca25/Eas10	1	122.2	89.1	-33.0
Eca25/Eas10	2	114.2	114.2	0.0
Eca26/Eas18	1	110.3	91.6	-18.6
Eca26/Eas18	2	32.6	32.6	0.0
Eca27/Eas27	1	220.2	84.5	-
Eca27/Eas27	2	-	78.5	-57.2
Eca28/Eas4	1	194.1	134.8	-59.2
Eca30/Eas30	1	114.3	91.0	-23.2
EcaX/EasX	1	151.5	99.3	-52.2
AVERAGE		120	91	-24
Individual alleles		117	93	-25

**Table 3.4 Centromere domain size in the primary and immortalized fibroblasts.** This table shows the centromere span in both cell lines with the average span and the span of the individual alleles shown in pink.



**Figure 3.9 CENP-A ChIPSeq profile comparison in immortalized and primary donkey fibroblasts.** Immortalised (blue) and primary (red) ChIPSeq superimposition shows the distribution of CENP-A signal across the centromere domains

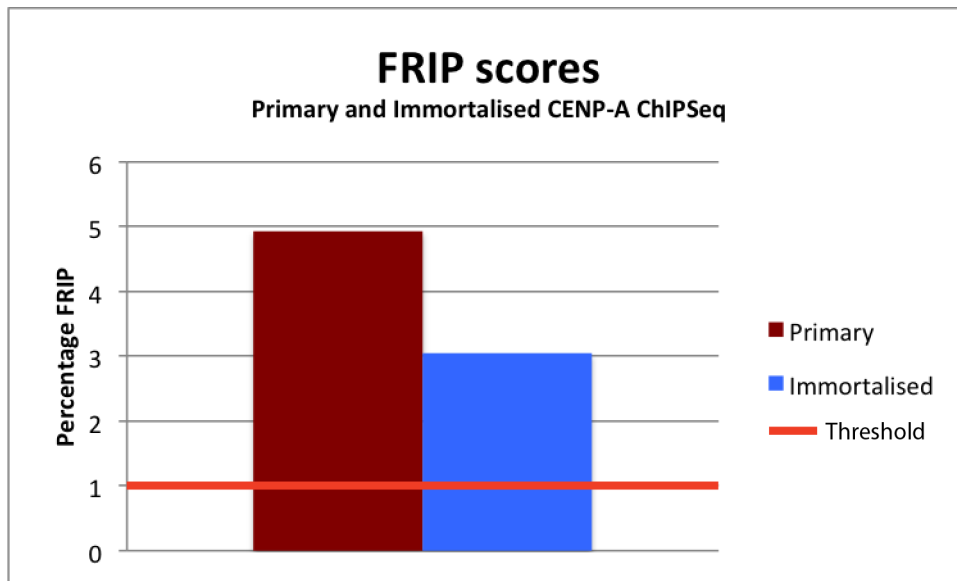




**Figure 3.10** CENP-A ChIPSeq comparison Eca11/Eas13 and Eca13/Eas14 Immortalised (blue) and primary (red) ChIPSeq superimposition shows the spread of the CENP-A signal in the primary cell line.

### 3.6.1 FRiP (fraction of reads in peaks) analysis

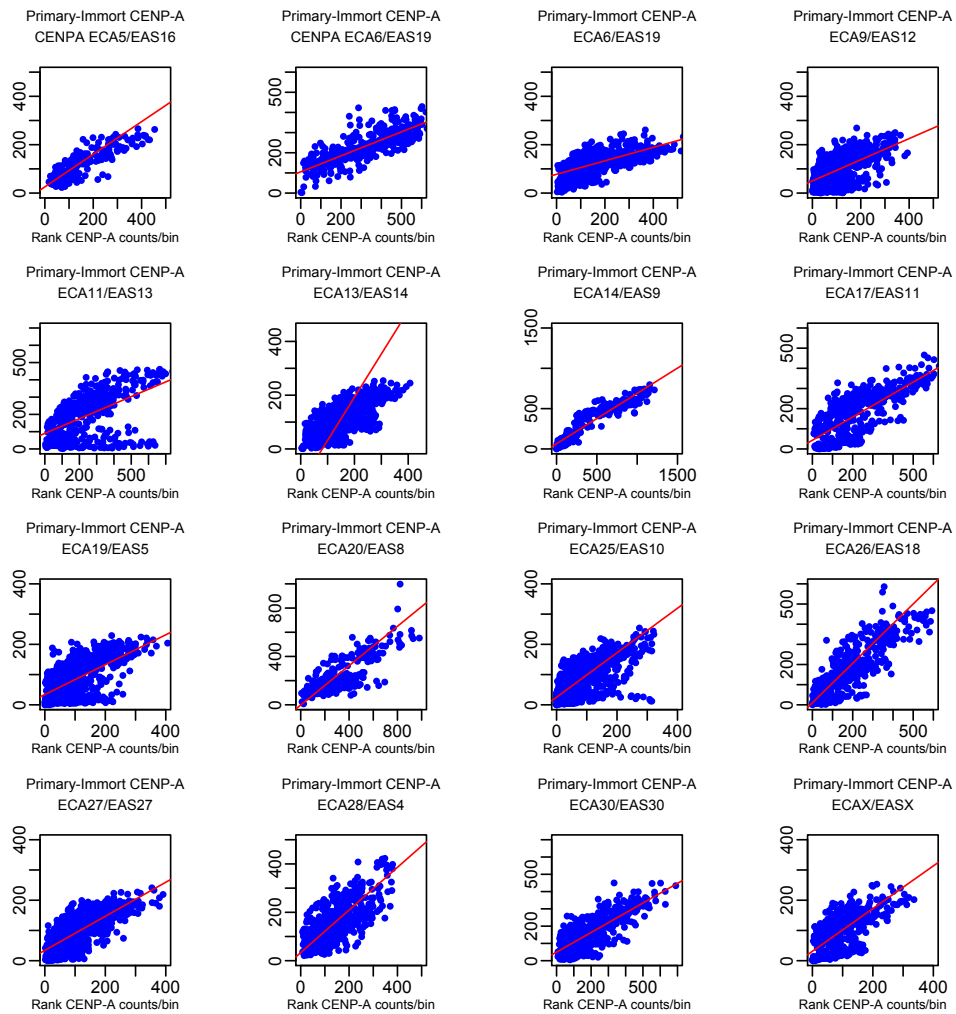
FRiP is used to examine the quality of the immunoprecipitation and is performed by measuring the signal enrichment at centromere domains. The percentage of reads in both the immortalized and primary CENP-A ChIPSeq datasets that fall within the centromere domains, shown in Figure 3.9, were calculated. The reads at the centromere were then divided by the reads across the genome and multiplied by 100 to express them as a percentage. FRiP scores of 4.939% and 3.05% were obtained for the primary and immortalised datasets (Figure 3.11), which was well above the 1% threshold for a successful ChIP. These data show that the antibody developed here performs quite well in ChIPSeq applications.



**Figure 3.11** FRIP scores for CENP-A ChIPSeq in primary and immortalized cell lines. FRIP scores for both the primary and immortalized data sets are well above the 1% threshold indicating a high signal to noise ratio and a successful immunoprecipitation.

### 3.7 CENP-A Correlation

To further characterize the centromere positioning between the two cell lines, a correlative approach was taken. ChIPSeq data from the two ChIP experiments were binned into 200bp windows and using the *deeptools* function *multiBamSummary*, the read coverage for the centromeres was calculated. The correlations were carried out in R using the Spearman algorithm to generate a correlogram scatter plot for each centromere Figure 3.12 and output the rho values for each centromere as shown in Table 3.5. To do this, the data in the immortalized cell line was sorted by rank and the corresponding rank for that bin in the primary cell line was plotted as the x-axis. In profiles that are quantitatively very similar the rank of a particular bin will be very similar in both datasets. In cases where the distribution of CENP-A has changed, correlation in rank will be degraded. The correlogram scatterplots for each centromere show that the CENP-A signal in both the immortalized and primary cell line is correlated. The Spearman values in Table 3.5 confirms this with ECA14/EAS9 showing a very high positive correlation ( $>0.9$ ), Eca5/Eas16, Eca6/Eas19, Eca8/Eas7, Eca17/Eas11, Eca20/Eas8, Eca26/Eas18, Eca27/Eas27 and Eca30/Eas30 (0.7-0.9) showing a high positive correlation and Eca9/Eas12, Eca11/Eas13, Eca13/Eas14, Eca19/Eas5, Eca25/Eas10, Eca28/Eas4 and EcaX/EasX showing a moderate correlation (Mukaka, 2012).



**Figure 3.12 Correlation of CENP-A ChIPSeq from primary and immortalized donkey fibroblasts.**

<b>Centromere</b>	<b>Rho Value</b>
ECA5/EAS16	0.8971069
ECA6/EAS19	0.8326019
ECA8/EAS7	0.7054842
ECA9/EAS12	0.5711706
ECA11/EAS13	0.5408591
ECA13/EAS14	0.6872
ECA14/EAS9	0.9514065
ECA17/EAS11	0.7713945
ECA19/EAS5	0.6069282
ECA20/EAS8	0.7304303
ECA25/EAS10	0.5823275
ECA26/EAS18	0.8491163
ECA27/EAS27	0.7669011
ECA28/EAS4	0.6393635
ECA30/EAS30	0.7556224
ECAX/EASX	0.6120112

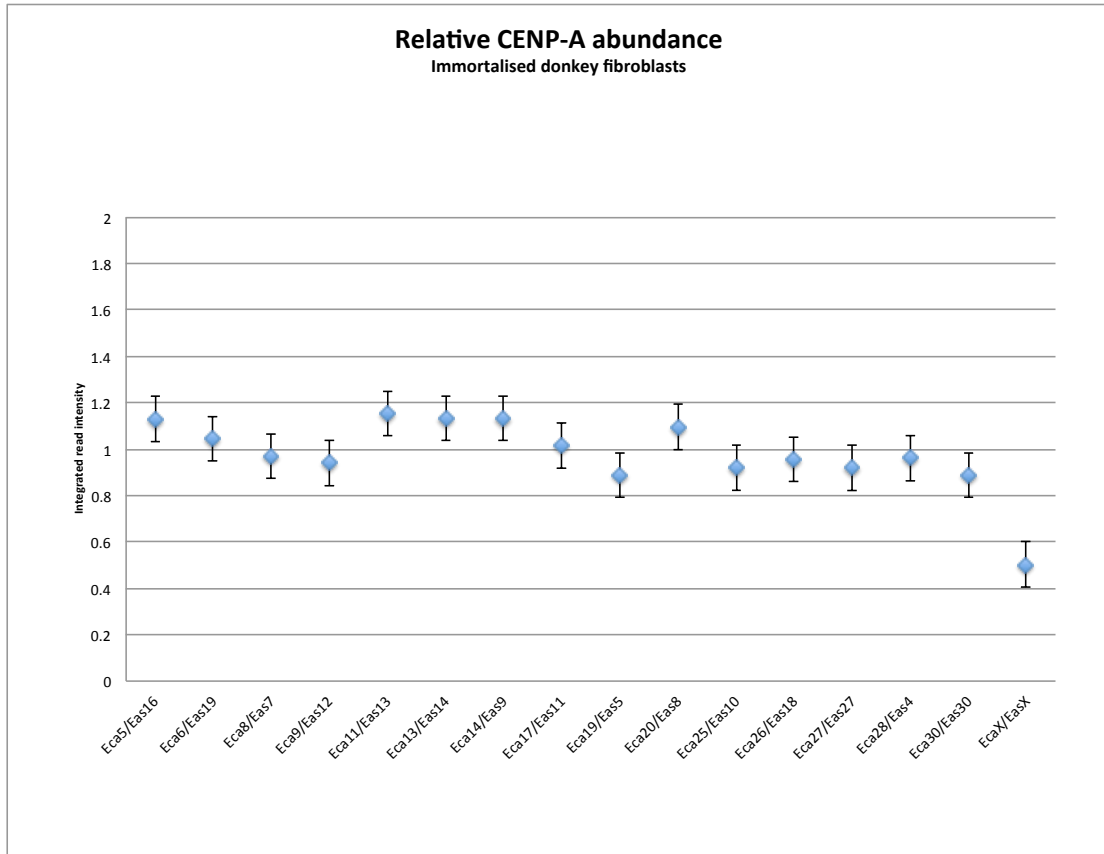
**Table 3.5 Spearman correlative values for the CENPA binding domain in primary and immortalized donkey fibroblasts**

Taken together, the correlation results and the superimposition of the CENP-A reads show that both datasets primarily co-occupy the same domains with a smaller centromeric footprint observed in the immortalised cell line. This observation is congruent with the founder effect from a heterogeneous initial population and shows that centromere position is tightly conserved during mitotic propagation within the immortalized cell line.

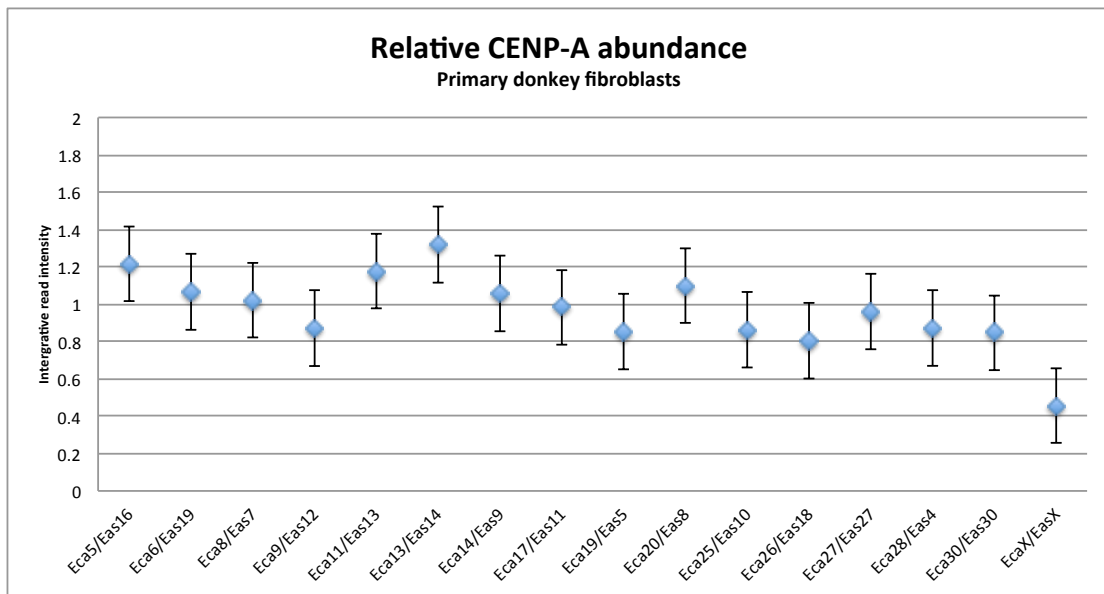
### **3.8 Relative abundance of CENP-A at satellite-free centromeres**

The abundance of CENP-A associated DNA at unique sequence centromeres was examined in both the primary and immortalized donkey fibroblasts and showed a remarkable uniformity across both cell lines. By comparing the relative abundance of CENP-A at satellite-free centromeres in the donkey we can ask about the fidelity of CENP-A maintenance on unrelated DNA sequences.

To further characterize the antibody, the relative abundance of CENP-A across the autosomes was measured. To do this, the integrated read counts within each of the individual donkey CENP-A binding domains on autosomes was summed and divided by the number of autosomes to determine the average CENP-A associated DNA signal per centromere. The data were then normalized by dividing the integrated counts at each individual centromere by the average. The relative abundance of CENP-A at the centromeres of the immortalized and primary cell lines is shown in Figures 3.13 and 3.14 respectively. The error bars represent the standard deviation in CENP-A signal. A uniform distribution can be seen in both datasets, with the ChIPSeq from the immortalized cell line (std dev 0.09) showing a tighter CENP-A distribution than the primary cell line (std dev 0.2). The haploid X chromosome was excluded from these analyses, since the fibroblasts used in this study are from a male donkey. The abundance of CENP-A at the X chromosome in the primary and immortalized cell lines are 45.62% and 50.27%, ~50% the signal found for autosomes and consistent with the haploid representation of the X chromosome in a male.



**Figure 3.13 Relative abundance of CENP-A in the immortalised cell line** The relative abundance of CENP-A at each centromere domain shows a uniform distribution with error bars depicting the standard deviation (0.09). The haploid X chromosome is 50% that of the mean abundance across the other diploid centromeres



**Figure 3.14 Relative abundance of CENP-A in primary donkey fibroblasts** The relative abundance of CENP-A at each centromere domain shows a uniform distribution with error bars depicting the standard deviation (0.2). The haploid X chromosome is 50% that of the mean abundance across the other diploid centromeres

### **3.9 Discussions**

I have successfully generated an equid optimised affinity purified CENP-A antibody that has application in western blot, immunofluorescence, immunoprecipitation and ChIPSeq. Despite the non-specific bands present in the western blot, the immunofluorescence and ChIPSeq are remarkably clean with little background and notably high FRIP scores of 2.07% and 3.05%.

The antibody was used to validate the immortalized fibroblasts for ChIPSeq experiments and to ask whether centromere location stable as a result of prolonged culturing. The CENP-A associated domains in both the primary and immortalized fibroblasts occupied the same overall binding domains, with the immortalized cell line containing on average a ~20% smaller CENP-A footprint than the corresponding centromere in the primary cell line. This observation is in line with the founder effect hypothesis, whereby the ChIPSeq on the primary cell line was carried out on a heterogenous population of fibroblasts while the immortalized cell line was generated from a single clone from this population and the CENP-A ChIPSeq shows the tight conservation of centromere position during mitotic propagation.

The relative abundance of CENP-A at the centromeres in both cell lines were notably uniform. This indicates that there is an optimal level of CENP-A that is maintained for centromere function. Studies (Bodor et al., 2014) demonstrated through three independent methods that the abundance of CENP-A at centromeres is relatively uniform and in two fold excess of that required to recruit kinetochore complexes. CENP-A abundance at individual centromeres however ranged over a 2-3 fold span. Analysis of CENP-A abundance at defined centromeres showed that alpha satellite array size is positively correlated with CENP-A occupancy (Sullivan, Boivin, Mravinac, Song, & Sullivan, 2011). Reports have shown that the abundance of CENP-A at unique sequence centromeres is lower than that observed at satellite containing domains (Amor et al., 2004; Irvine et al., 2005). The human unique sequence centromere on chromosome 4 (PD-NC4) contains 16% less CENP-A than satellite containing centromeres, while in the case of the human Y chromosome, which contains alphoid DNA but lacks CENP-B there is 18% less CENP-A than the satellite containing average (Fachinetti et al., 2015). These DNA sequence dependent differences in CENP-A binding underscore the remarkable uniformity of CENP-A abundance at satellite free centromeres.

This data, Figure 3.13 and 3.14, shows that CENP-A nucleosomes are stabilized independent of CENP-B binding maintaining a uniform abundance across all satellite free centromeres. The CENP-A domains in the immortalized cell line occupy a smaller centromere footprint and the levels of CENP-A in these cells are more tightly regulated than the primary cell line. This observation is independent of the signal to noise ratio, since the CENP-A ChIPSeq in the primary cell line has a higher FRIP score than the immortalized dataset. This again is in agreement of the founder effect whereby the immortalized cell line indicates that centromere position and CENP-A abundance is tightly regulated and maintained during mitotic propagation. Thus, the immortalized fibroblasts were used in further experiments throughout this thesis.

## Chapter 4 Interindividual and interspecies centromere comparison

### 4.1 Introduction

Centromere formation can occur on any type of DNA sequence, however evolutionary preferences across long established centromeres in many organisms, containing arrays of satellite repeats as well as transposable elements, are apparent (Cheng et al., 2002; Rudd & Willard, 2004; Sun, Le, Wahlstrom, & Karpen, 2003; Wolfgruber et al., 2009). In Metazoans and plants, transposable elements are the most common class of genetic material, accounting for 44 % of the human genome (Mills, Bennett, Iskow, & Devine, 2007). In both *Drosophila* and *Arabidopsis*, pericentric and centric heterochromatin shows an enrichment for transposable elements (Kaminker et al., 2002; Kapitonov & Jurka, 1999) but unlike humans do not contain dominant repeat classes such as SINEs and LINEs. In the case of Poaceae, a large family of grasses, a Ty3-gypsy derived family of retrotransposons are present exclusively at centromeres. The conservation of centromeric sequence between plant species that diverged tens of millions of years ago suggest a role in evolutionary adaptation (Langdon et al., 2000). Similar highly conserved centromeric retrotransposons have also been identified in many cereals including maize, barley and rye (Jiang et al., 2003). The maize centromeric retrotransposon (CRM) is interspersed with a 156bp satellite repeat, known as CentC, which can span up to 2Mb in maize centromeres. ChIPSeq of CENH3 pulls down CRM elements as efficiently as CentC, indicating that it has a functional role in the centromere (Jiang et al., 2003).

The centromere associated protein, CENP-B is recruited to centromeric chromatin through its binding motif, the ‘CENP-B box’ which is conserved across humans, mouse, ferret, giant panda, tree shrews and gerbils despite having otherwise unrelated satellite sequences (Kipling & Warburton, 1997). There are some incidences of centromeres void of CENP-B and its binding motif (Ohzeki, Nakano, Okada, & Masumoto, 2002) that are functional but have a higher frequency of mis-segregation (Fachinetti et al., 2015), illustrating its importance in chromosome fidelity. CENP-B shares a striking homology to the pogo family of retrotransposons, in particular Tigger elements. Given the degree of similarity of the CENP-B box to Tigger2, this raises the possibility that CENP-B has the ability to induce nicks and promote centromere recombination (Kipling & Warburton, 1997). CENP-B also shares similarity to the pogo related Tc1/mariner elements (Casola, Hucks, & Feschotte,



2008). This suggests that CENPB has evolved from the transposon family, indicating that transposons may be related to centromere formation.

Given the unquestionable diversity and abundance of transposable elements, as well as their mobility and versatility there is no doubt that they serve an important role in the evolutionary process. Mounting evidence suggest that transposons play a critical role in the centromeric function and architecture observed in many organisms today.

#### **4.2 Domain comparison**

In order to examine features at the sequence level of chromatin capable of “centromerization”, CENP-A ChIPSeq reads from immortalized donkey fibroblasts (Asino Nuovo) generated in Chapter 3 were mapped to the EquCab genome and recently published Guanzhong donkey (Huang et al., 2015) genomes and compared with the EquDonk assembly. This allows the study of interspecies and inter-individual sequences at a domain, that according to the Immortalised donkey cell line has the propensity to facilitate CENP-A binding and centromere formation. The horse, EquCab genome, contains a single unique sequence centromere on chromosome 11. The domains, which the EquDonk CENP-A ChIPSeq reads map to, are not bona fide centromere domains in the horse. The immortalized donkey fibroblasts used in the CENP-A ChIPSeq experiments were from a European donkey, while the Guanzhong donkey is native to China. Given the tendency of centromeres to “slide” in this genus (Purgato et al., 2015) as well as the continental diversification of these two donkey individuals, it cannot be confirmed that the loci which the EquDonk CENP-A ChIPSeq reads map to is centromeric in the Guanzhong donkey. Nevertheless, the comparison of the underlying sequences present at the loci which the donkey centromeres map to provides a useful comparator for identification of sequence anomalies and repetitive elements that may be related to neocentromere formation. Centromere associated DNA sequences of satellite free centromeres identified by ChIPSeq will be compared between two donkey individuals as well as the horse.

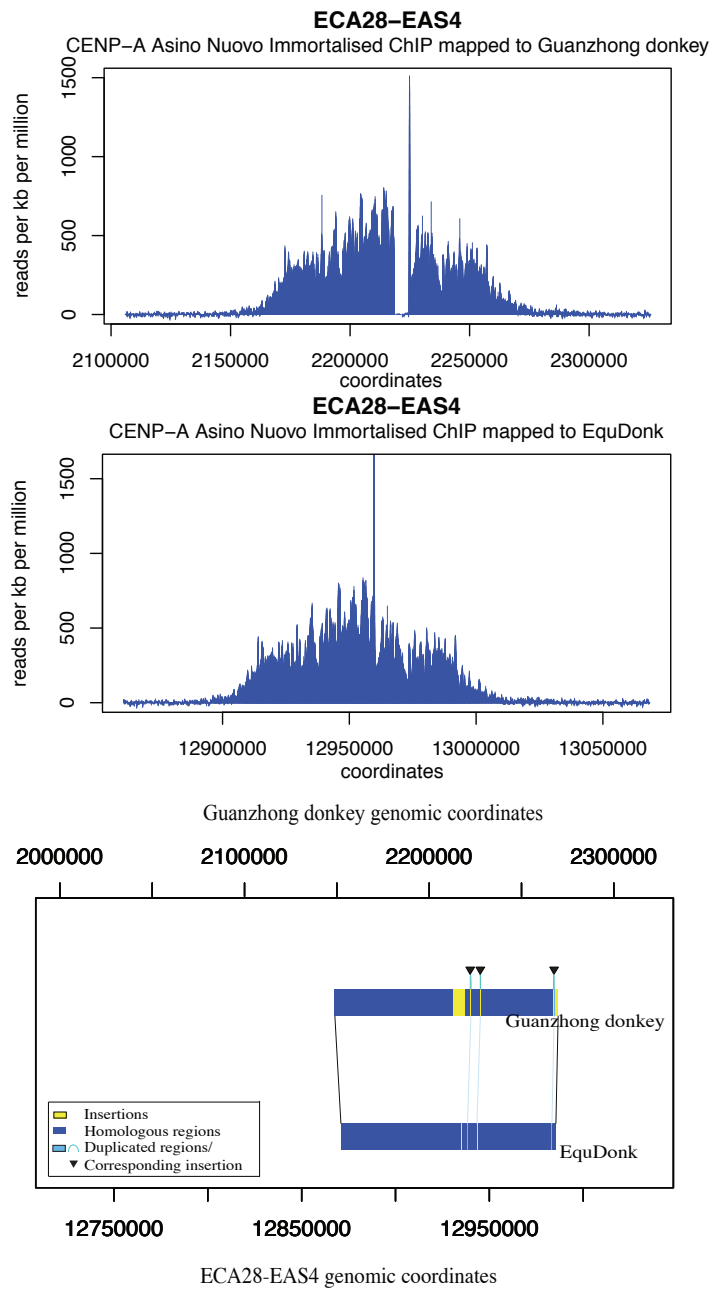
The Guanzhong donkey was sequenced using a whole genome shotgun strategy generating a 2,391,051,217bp genome consisting of 2,167 scaffolds and total sequence coverage of 42.4 fold. From the scaffold sequences, an indexed genome was generated using the *bowtie2 build* function to which the EquDonk CENP-A ChIPSeq reads were aligned. The alignment was visualized using the *Intergrative Genomics Viewer* from the BROAD institute (Robinson et al., 2011; Thorvaldsdóttir et al., 2013)

and the contigs were scanned for regions of CENP-A enrichment. The reads were extracted from domains, which CENP-A mapped to and were subsequently mapped back to the EquDonk assembly to establish which centromere the domains corresponded to. The alignments were then normalized using *Deeptools* and subsequently visualized using the R package *Sushi*. Sequence differences between the EquDonk centromere and the corresponding loci in the Guanzhong donkey were examined. *Repeatmasker* (Smit et al., 2013) was used to identify repetitive elements across the entire CENP-A mapped domain in both donkey individuals as well as identification of repetitive element in regions of sequence insertion and duplication.

A similar approach was adopted for analysis of sequences in the EquCab domains, following on from work by Dr. Federico Cerutti and collaborators in the Giulotto lab at the University of Pavia, Lombardy, Italy. Their work involved CENP-A ChIPSeq using a peptide antibody in a primary donkey cell line, mapping the reads to the EquCab genome and searching for sequence “peculiarities” between EquDonk and the corresponding EquCab loci. The study carried out in this chapter examines CENP-A ChIPSeq using the antibody generated in Chapter 3 for comparison of two donkey individuals. For a direct relatable comparison to the horse, I recapitulated our collaborators work, using CENP-A ChIPSeq reads from immortalized donkey fibroblasts for the EquDonk, Guanzhong donkey and horse (EquCab) mapping.

In these mapping experiments, there is an abundance of LINE elements associated with regions which the CENP-A ChIPSeq reads map to in both the horse and the two donkey individuals. Analysis of repetitive elements present in regions of sequence insertion and duplication is also carried out in this chapter. Details of repetitive elements associated with the horse domains are shown in appendix 1. In the following section each centromere is considered in turn. I first examine the donkey sequences and compare the two donkey individuals followed by a comparison of the corresponding domain in the horse.

## The EAS4 Centromere



**Figure 4.1 EAS 4 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (top) and EquDonk (middle). Schematic representation of sequence features of EAS4 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey (bottom).

## Sequence features of EAS4 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	1	3	-	-
Guanzhong donkey	4	3	-	-

**Table 4.1 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS4.**

The centromere function of donkey chromosome 4 was mapped to Guanzhong donkey contig gi|933836246|gb|JREZ01000325.1|: 2,153,677-2,280,289, its peak profile showed a Gaussian-like distribution with a gap signifying an insertion in the Guanzhong genome, absent from the EquDonk centromere. There were a number of sequence variations between the two donkey centromere domains, as illustrated in the schematic. EquDonk contained one insertion spanning a mere 7bp, while the Guanzhong donkey contained 4 insertions spanning 6399bp, 172bp, 221bp and 40bps. There were three instances of single copy sequence in EquDonk that were duplicated in the Guanzhong donkey, duplicated sequences were shown in light blue and the corresponding duplication was signified as an insertion (yellow).

### *Repetitive elements across the EAS4 centromere domain in the donkey*

The distribution of SINEs at this domain in the Guanzhong donkey was 2.1%, 1.57% less than that observed across the whole genome (3.67%). There was a notable increase in LINEs, with 29.88% of CENP-A mapped domain containing these elements when compared to 21.96% across the genome, in particular, LINE1 elements which increase from whole genome level of 16.09% to 25.59%. LTR element abundance dropped by 1.02%, when compared to the whole genome. There was also a 0.51% increase in DNA elements in this domain with TcMar-Tigger levels increasing by 2.01%, while hAT-Charlie levels dropped by 1.41%.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	19	3198	2.1	3.67
ALUs	0	0	0	0
MIRs	19	3198	2.1	3.63
<b>LINEs:</b>	59	45604	29.88	21.69
LINE1	29	39054	25.59	16.09
LINE2	27	6261	4.1	4.9
L3/CR1	3	289	0.19	0.5
<b>LTR elements:</b>	23	8339	5.46	6.48
ERV_L	4	1144	0.75	2.19
ERV_L-MaLRs	10	3794	2.49	2.72
ERV_classI	8	2667	1.75	1.17
ERV_classII	0	0	0.00	0
<b>DNA elements:</b>	19	6607	4.33	3.82
hAT-Charlie	6	829	0.54	1.95
TcMar-Tigger	6	4494	2.94	0.93

**Table 4.2 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi|933836246|gb|JREZ01000325.1| (EAS4) compared with whole genome levels**

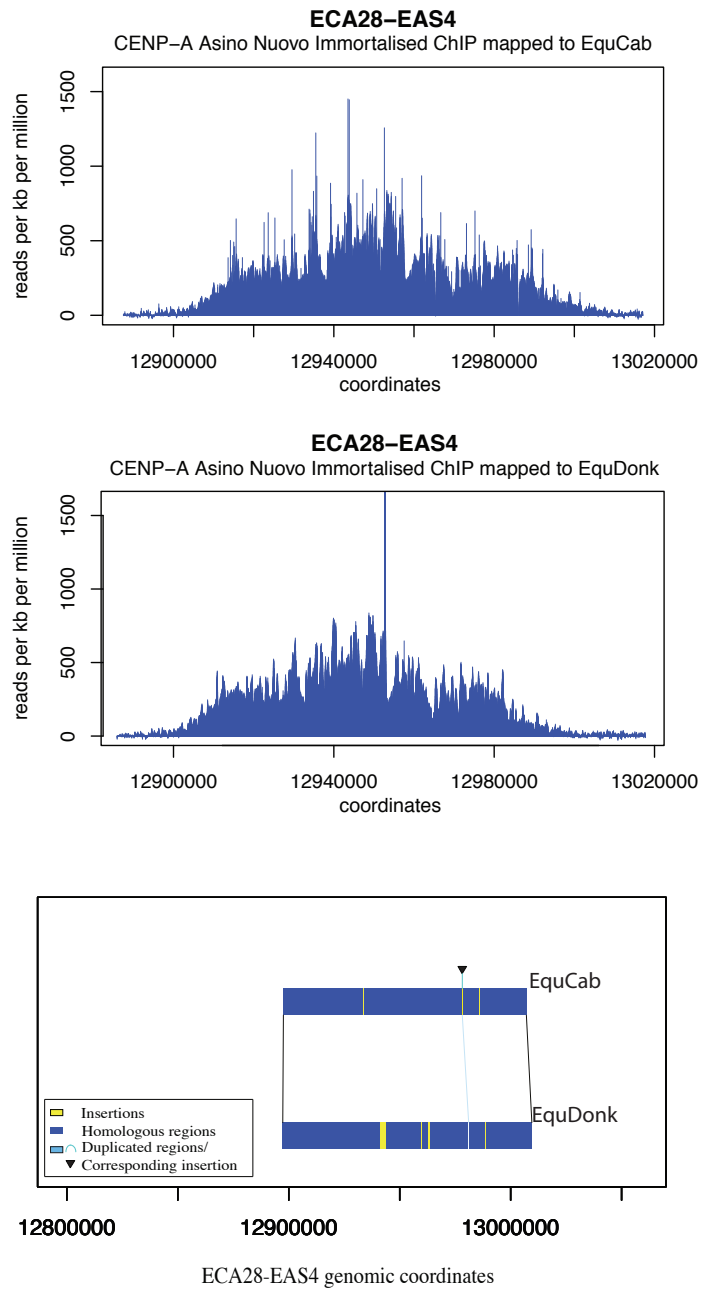
Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	12	2132	1.86	3.67
ALUs	0	0	0	0
MIRs	12	2132	1.86	3.63
<b>LINEs:</b>	40	34306	29.95	21.69
LINE1	25	30543	26.67	16.09
LINE2	14	3695	3.23	4.9
L3/CR1	1	68	0.06	0.5
<b>LTR elements:</b>	18	6522	5.69	6.48
ERV_L	4	1144	1	2.19
ERV_L-MaLRs	6	2711	2.37	2.72
ERV_classI	8	2667	2.33	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	15	6083	5.31	3.82
hAT-Charlie	2	305	0.27	1.95
TcMar-Tigger	6	4494	3.92	0.93

**Table 4.3 Summary of repetitive elements that span the centromere of EquDonk (EAS4)**

Comparison of two donkey individuals revealed overall conservation in repetitive sequence properties. Analysis of repetitive elements in regions of sequence variation between the two donkeys showed that the Guanzhong donkey insertion spanning 6399bp from 2,218,482-2,224,881nt contained four instances of LINE/L2s, from the L2a and L2c subfamilies and one instance of LINE/CR1 from the L3 subfamily. This region also contained two examples of SINE/MIR, LTR/ERV\_L-MaLR and hAT-Charlie. When these regions of insertion are taken together, the overall abundance of LINEs (13.23%) and LTR elements (4.60%) are below the whole genome average while SINEs (3.88%) and DNA elements (4.34%) are above the genomic average. No repetitive elements were detected in regions of duplication in the Guanzhong donkey. There were also no instances of repetitive elements in EquDonk inserted or duplicated

regions. The Guanzhong donkey centromere spanned 12,076bp longer than the EquDonk centromere (114536bp) and shared 99% sequence identity across the CENP-A binding domains. Levels of repetitive elements across the entire EquDonk EAS4 centromere and the Guanzhong orthologous domain remained comparable, with SINE abundance slightly less in EquDonk (0.24%), while levels of DNA elements in the Guanzhong donkey were less than that seen in EquDonk (0.98%).

## The EAS4 Centromere



**Figure 4.2 EAS 4 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the horse orthologous domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS4 centromere region in EquDonk compared to the orthologous region in EquCab (bottom).

## Sequence features of EAS4 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	4	1	-	-
EquCab	4	1	-	-

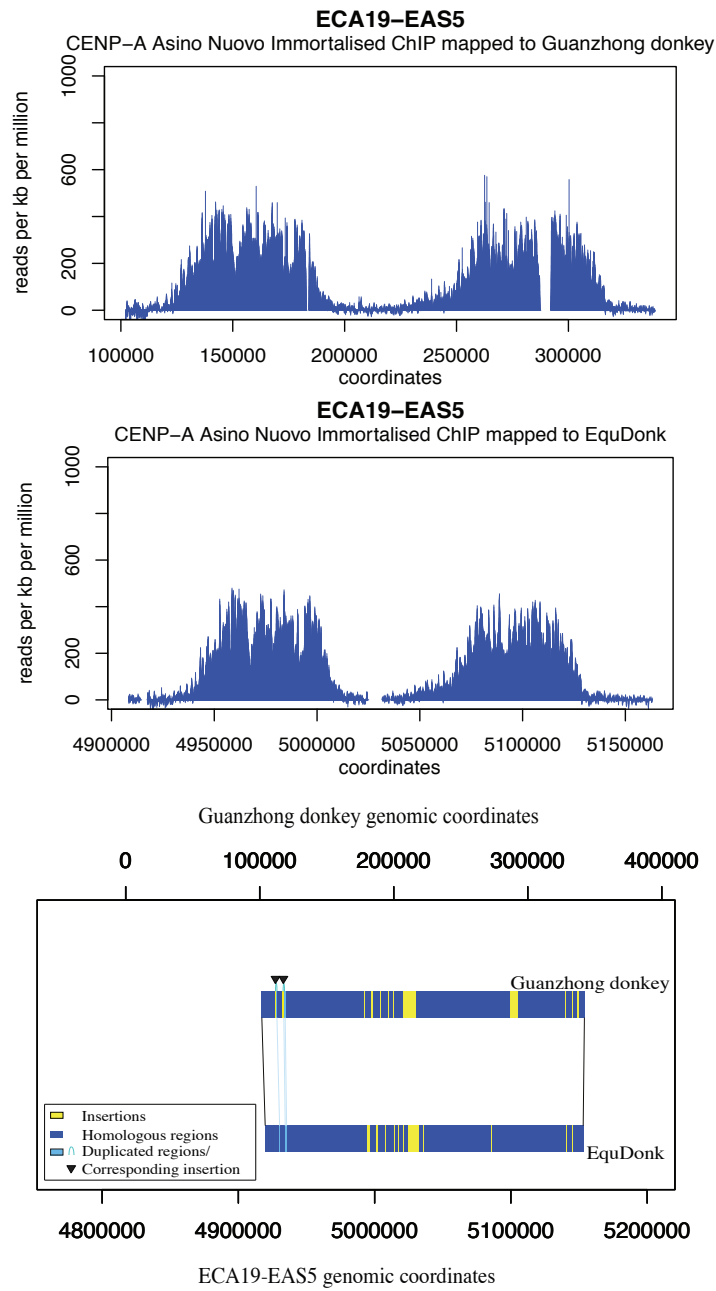
**Table 4.4 Summary of sequence variation between EquDonk and EquCab on Eas4**

CENP-A ChIPSeq reads from donkey chromosome 4 were mapped to the horse genomic coordinates, 12,897,478-13,007,113nt on chromosome 28. A gaussian-like distribution was observed in both horse and donkey profiles. The schematic shows regions of sequence divergence between both species with insertions showed in yellow and duplications shown in light blue. There are four instances of sequences present in the EquDonk CENP-A binding domain but absent from corresponding loci in EquCab, spanning 2344bp, 265bp, 490bp and 88bp. Similarly there are four regions of sequence insertion in the EquCab domain spanning 432bp, 65bp, 28bp and 37bp. The duplicated regions spanned 37bp in both species (light blue).

Analysis of repetitive elements present in regions of insertions showed that in EquDonk there were two novel instances of LINE/L1 from subfamilies L1M1 and L1MEf as well as two instances of LTR/ERV (LTR16A, ERV3-16A3\_I-int) repetitive elements. The overall abundance of repetitive elements in regions of insertion for LINES is 15.30%, 6.39% lower than the genomic average while LTR element abundance is 20.04%, 13.56% higher than the genomic average. In the EquCab insertion domains, there was one instance of LINE/L1 from L1M3 subfamily as well as short simple repeats. EquCab inserted regions were combined and the overall abundance of LINES was 48.01%, 24.42% higher than the genomic average. The EquDonk centromere spanned 114,536bp slightly larger than the horse orthologous region, which spanned 109,636bp and shared 98% sequence identity.



## The EAS5 Centromere



**Figure 4.3 EAS 5 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS5 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey (bottom).

### Sequence features of EAS5 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	11	2	-	-
Guanzhong donkey	14	2	-	-

**Table 4.5 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS 5**

The CENP-A ChIP reads from EAS5 were mapped to the Guanzhong donkey genome contig gi|933833362|gb|JREZ01000925.1|: 111,522-329,322nt. Two peak profiles were observed with a number of discontinuities, the largest spanning 1399bp from 182,750-184,149 and 4972bp from 287,257-292,234. The Schematic shows the inserted sequences (yellow) present in both species, with eleven examples present at the EquDonk centromere and fourteen instances at the corresponding Guanzhong donkey domain. Sequence spanning 498bp and 929bp present in single copy in EquDonk (shown in light blue) were found to be duplicated in the Guanzhong donkey. For each one copy is shown in light blue while the second copy is shown in yellow with a black inverted triangle.

### *Repetitive elements across the EAS5 centromere domain in the donkey*

The abundances of SINEs at this Guanzhong donkey domain were 2.21% less than that found across the whole genome. There was a 4.04% increase in overall LINES present in this domain with a 6.12% increase in L1 elements but a 1.77% decrease in L2 elements. There was little difference in the overall abundance of LTR elements but again ERV class I showed a slight increase of 1.38%. There was a reduction in hAT-Charlie and TcMar-Tigger dropping from 1.95% and 0.93% across the whole genome to 0.55% and 0.65% respectively at this domain.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	16	2306	1.06	3.67
ALUs	0	0	0	0
MIRs	15	2248	1.03	3.63
<b>LINES:</b>	69	56051	25.73	21.69
LINE1	48	48373	22.21	16.09
LINE2	19	6811	3.13	4.9
L3/CR1	2	867	0.4	0.5
<b>LTR elements:</b>	36	15223	6.99	6.48
ERV L	13	4979	2.29	2.19
ERV L-MaLRs	14	4598	2.11	2.72
ERV_classI	8	5550	2.55	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	23	4340	1.99	3.82
hAT-Charlie	9	1203	0.55	1.95
TcMar-Tigger	7	1419	0.65	0.93

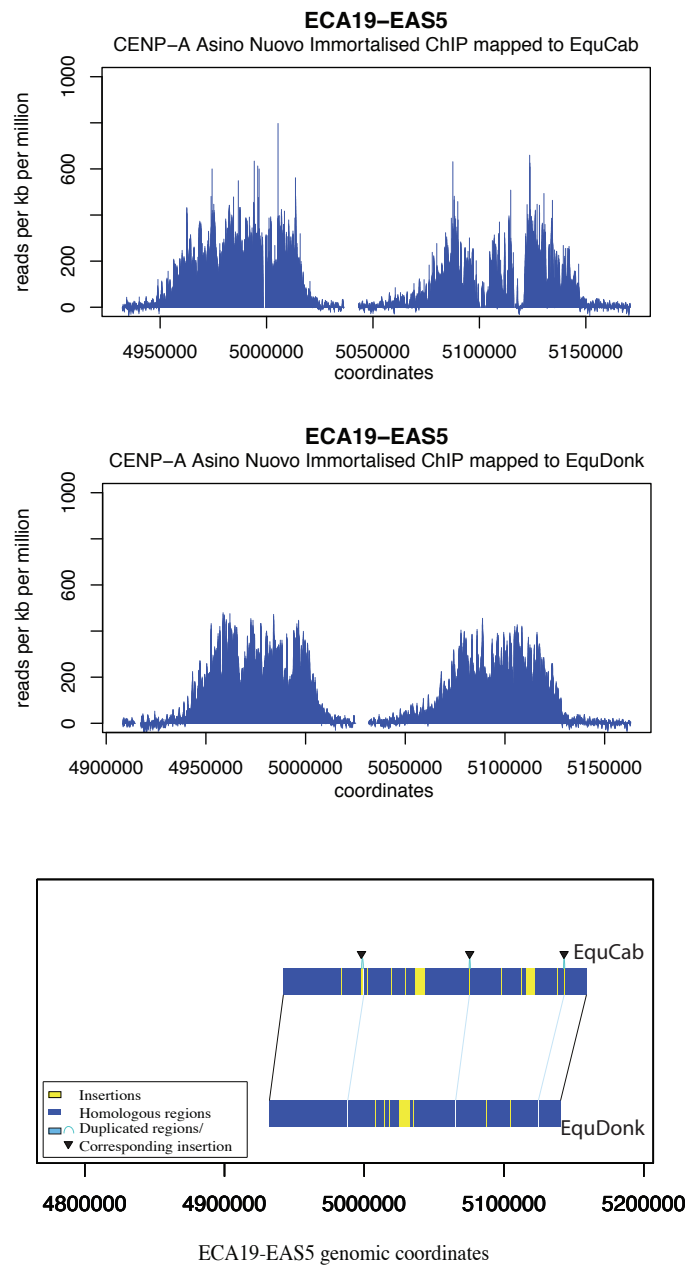
**Table 4.6 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi|933833362|gb|JREZ01000925.1| (EAS5) compared with whole genome levels**

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	19	2654	1.17	3.67
ALUs	0	0	0	0
MIRs	18	2596	1.14	3.63
<b>LINEs:</b>	69	56411	24.81	21.69
LINE1	50	49370	21.72	16.09
LINE2	17	6174	2.72	4.9
L3/CR1	2	867	0.38	0.5
<b>LTR elements:</b>	41	17695	7.78	6.48
ERV1	14	5394	2.37	2.19
ERV1-MaLRs	14	4596	2.02	2.72
ERV_classI	11	7534	3.31	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	21	3823	1.68	3.82
hAT-Charlie	7	980	0.43	1.95
TcMar-Tigger	6	956	0.42	0.93

**Table 4.7 Summary of repetitive elements that span the centromere of EquDonk (EAS5)**

Analysis of repetitive elements present in the inserted sequences in EquDonk showed three instances of LTR/ERV1 and two instances of LINE/L1 from subfamilies L1MC and L1Meh as well as a single SINE. The overall abundance of repetitive elements in these regions for SINEs (0.64%), LINEs (3.92%) and LTR elements (8.80%) are lower than the genomic average. In inserted regions of the Guanzhong donkey, there were five instances of LINE/L2 elements from L2, L2a and L2b subfamilies two instances of LINE/L1 from L1MA7 and L1Meh subfamilies as well as TcMar-Tigger, SINEs and hAT-Charlie elements present. Abundance of repetitive elements at these domains are also lower than the whole genomic average for SINEs (0.45%), LINEs (6.55%) and DNA elements (1.24%). A simple 28bp TGAA repeat was present in the duplicated regions. The Guanzhong donkey orthologous region was 9533bp smaller than the EAS5 CENP-A binding domain sharing 99% sequence identity.

## The EAS5 Centromere



**Figure 4.4 EAS 5 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the horse orthologous domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS5 centromere region in EquDonk compared to the orthologous region in EquCab (bottom).

## Sequence features of EAS5 centromere

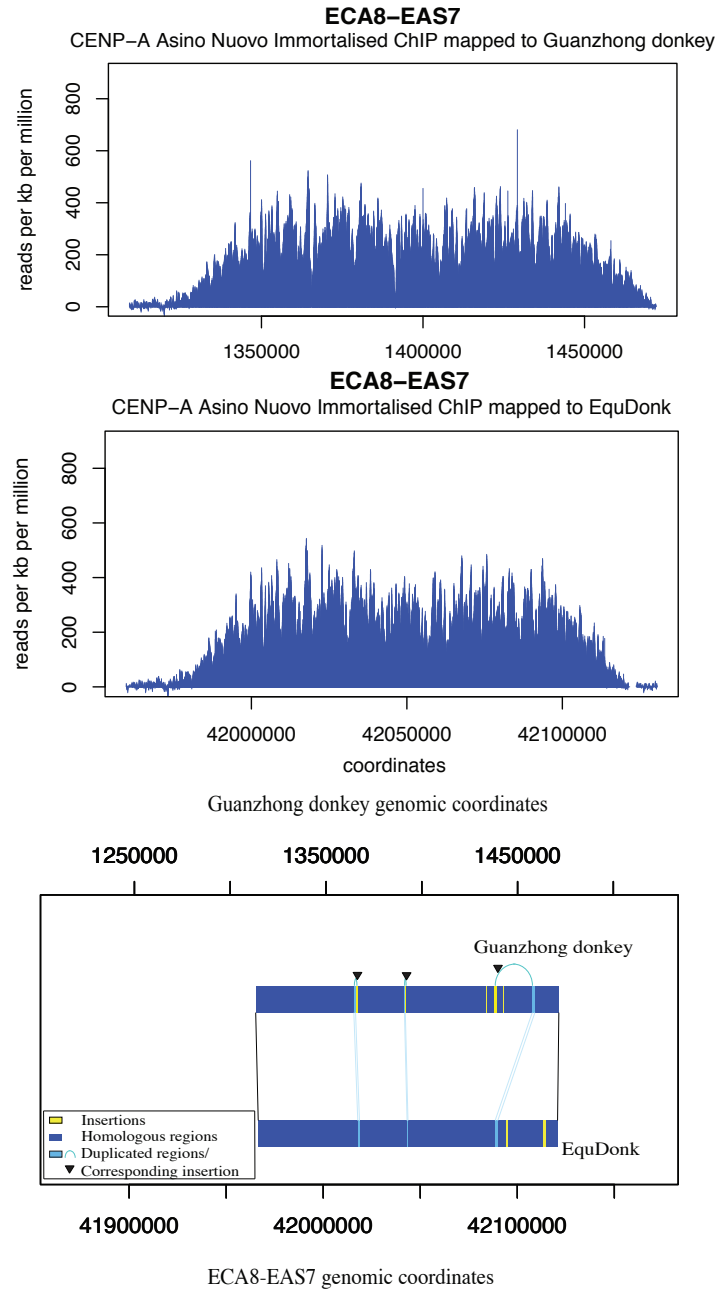
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	7	3	-	-
EquCab	15	3	-	-

**Table 4.8 Summary of sequence variation between EquDonk and EquCab at the EAS5 centromere**

The centromere function of donkey chromosome 5 was mapped to the horse reference region chr19: 4,950,000-5,150,000nt. The schematics show variation between the two species at the sequence level. EquDonk shows seven instances of instances of sequence insertion, the largest of which 7810bp is located between the CENP-A binding domains. Given the manner in which the EquDonk genome was generated using reads obtained from donkey CENP-A ChIP, the sequences between the domains of CENP-A association are less well defined. There are twelve instances of inserted sequences in the EquCab domain, ranging in size from 15bp to 7048bp. There are three cases of single copy sequences spanning 15bp, 18bp and 66bp in EquDonk that are duplicated in the horse domain, illustrated by an inverted triangle.

Analysis of the repetitive elements at the EquCab 5170bp insertion spanning from 5,116,071-5,121,246 shows three instances of LINE/L1 elements from the L1M subfamily, the largest spanning 3212bp, else where in the other insertion regions there were further instances of LINE/L1 from L1MC subfamily as well as LTR/ERV elements. Taken together the overall abundance of LINEs (28.83%) is higher than the genomic average while the abundance of LTR elements (0.57%) is lower. The inserted sequences in EquDonk contained a single SINE (1.02%), simple repeat and tRNA elements. No repetitive elements were detected in the duplicated sequences. Both domains were a similar size with the EAS5 domain spanning 227,334bp and the horse orthologous region spanning 218,656bp. The regions shared 97% sequence identity.

## The EAS7 Centromere



**Figure 4.5 EAS7 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS7 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey (bottom).

## Sequence features of EAS7 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	4	2	-	-
Guanzhong donkey	7	2	-	-

**Table 4.9 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS7**

The Eas7 centromere was mapped to the Guanzhong donkey contig gi|933835286|gb|JREZ01000511.1|: 1,190,428-1,471,883nt. EquDonk contains four instances of insertion while the Guanzhong donkey contains 7 inserted sequences. Sequences spanning 1993bp, 877bp and 462bp present in a single copy in EquDonk (light blue) are duplicated in the Guanzhong donkey with one copy show in light blue and the duplication shown in yellow with a black inverted triangle.

### *Repetitive elements across the EAS7 centromere domain in the donkey*

SINE abundance at the EAS7 centromere orthologous region in the Guanzhong donkey was decreased by 1.57% compared to the genome. There was an increase of 14.93% in LINE/L1s almost doubling whole genome levels (16.09%). LTR elements remained virtually unchanged with an increase of 1.59% in ERV class I. There was a 45% drop in hAT-Charlie elements and a 0.6% decrease in TcMar-Tigger elements.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	23	3200	2.1	3.67
ALUs	0	0	0	0
MIRs	22	3154	2.07	3.63
<b>LINEs:</b>	61	56003	36.67	21.69
LINE1	39	47368	31.02	16.09
LINE2	19	7603	4.98	4.9
L3/CR1	3	1032	0.68	0.5
<b>LTR elements:</b>	21	10859	7.11	6.48
ERV_L	8	3074	2.01	2.19
ERV_L-MaLRs	10	3338	2.19	2.72
ERV_classI	2	4221	2.76	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	14	2854	1.87	3.82
hAT-Charlie	11	2296	1.5	1.95
TcMar-Tigger	2	501	0.33	0.93

**Table 4.10 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi|933835286|gb|JREZ01000511.1 (EAS7) compared with whole genome levels**

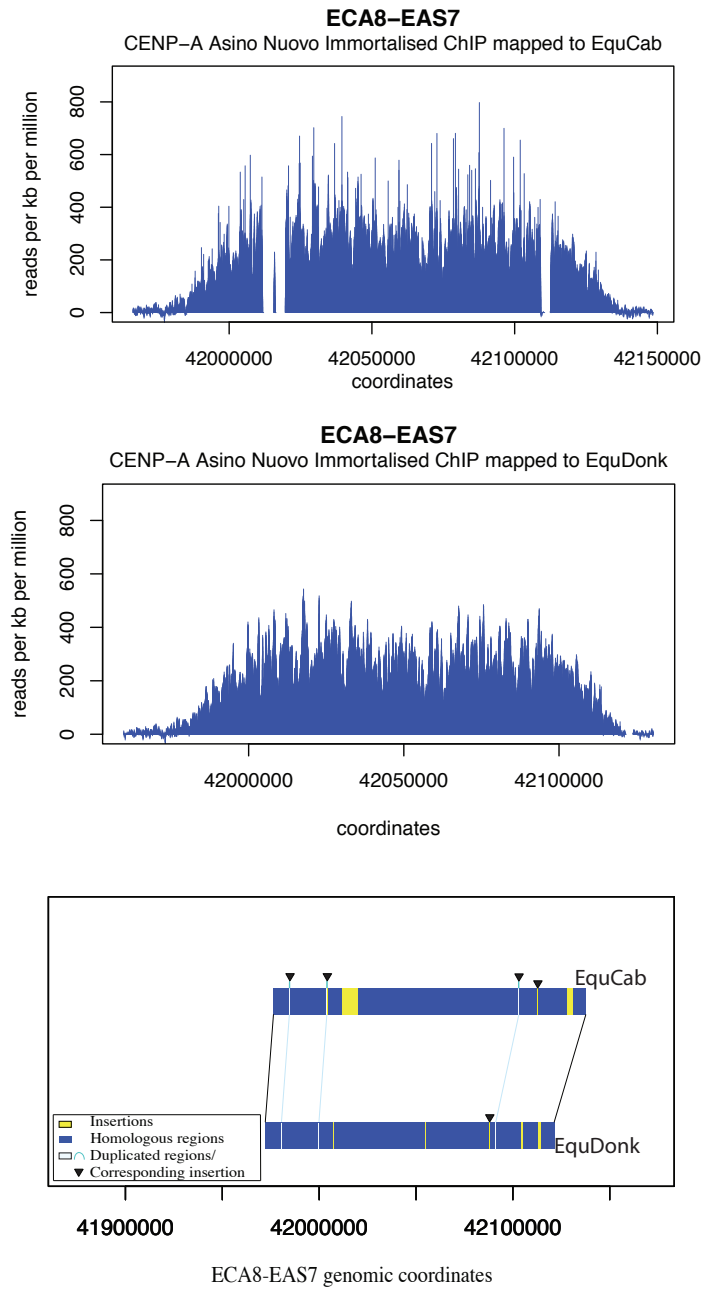
Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	22	3064	1.94	3.67
ALUs	0	0	0	0
MIRs	21	3018	1.91	3.63
<b>LINES:</b>	65	56230	35.64	21.69
LINE1	41	47549	30.14	16.09
LINE2	22	7863	4.98	4.9
L3/CR1	2	818	0.52	0.5
<b>LTR elements:</b>	23	12151	7.7	6.48
ERV1	10	3935	2.49	2.19
ERV1-MaLRs	10	3338	2.12	2.72
ERV_classI	3	4878	3.09	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	14	2854	1.81	3.82
hAT-Charlie	11	2296	1.46	1.95
TcMar-Tigger	2	501	0.32	0.93

**Table 4.11** Repetitive elements across the EAS7 centromere EquDonk compared with whole genome levels

Analysis of repetitive elements present in domains of insertion in the Guanzhong donkey showed two instances of LINES/L1 from subfamily L1M2, in the same proximity. Also present was one example of a SINE, LTR and tRNA. In the Guanzhong donkey duplicated regions, the only repetitive element present was a LINE/L1. The abundance of LINES in duplicated and inserted sequences was 42.86% well above the genomic average, while levels of SINEs (1.92%) and LTR (4.32%) elements were below. In the EquDonk insertion regions, the only repeat class present were three LINE/L1 elements from subfamilies L1M2, L1MA3 and L1M2, spanning 536bp, 179bp and 349bp respectively. Similarly in the EquDonk duplicated region, there was a single LINE/L1 from sub family L1M4. The abundance of repetitive elements across these domains in Equdonk also showed an abundance of LINES (39.83%). The EquDonk centromere domain spanned 157,777bp while in the Guanzhong donkey it spanned 152,713bp.



## The EAS7 Centromere



**Figure 4.6 EAS 7 centromere donkey versus horse comparison.** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the horse orthologous domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS7 centromere region in EquDonk compared to the orthologous region in EquCab (bottom).

## Sequence features of EAS7 centromere

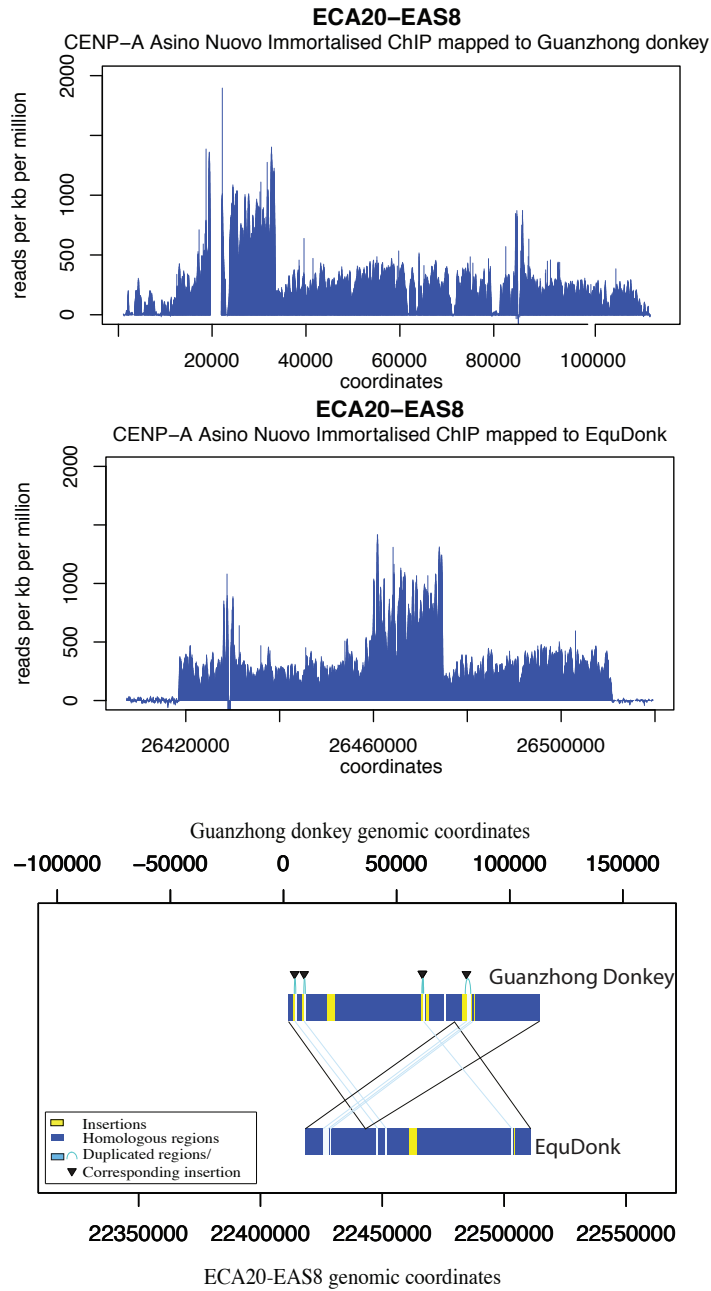
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	6	4	-	-
EquCab	8	4	-	-

**Table 4.12** Summary of sequence variation between EquDonk and EquCab on EAS 7

Donkey CENP-A ChIP reads from EAS7 were mapped to the horse orthologous region chr8: 41,976,329-42,138,526. Inserted sequences in EquDonk range in size from 46bp to 1216bp while in EquCab insertions range from 37bp to 7966bp. Duplicated sequences are small, 44bp, 105bp, 106bp and 174bp.

The centromere domain in EquDonk spanned 157,777bp while in the corresponding horse loci, including insertions it spanned 162,198bp. Analysis of repetitive elements in inserted sequences and regions of duplication in EquCab showed an abundance of LINES both L1 and L2 elements from subfamilies L2c, L2b, L1ME1, L1MD, L1M1, L1M4c, L1M2 and L1MC ranging in size from 58bp to 1979bp as well as single copies of SINE/MIR, hAT-Charlie and TcMar-Tigger. Taken together the abundance of LINES at these domains is 46.51%, well above the genomic average, as are DNA elements (4.54%) while SINE abundance is decreased (1.22%). Repetitive elements in EquDonk inserted/duplicated regions identified a single LINE/L1 (L1M2) spanning 350bp comprising 12.20% of the sequence and simple DNA repeats.

## The EAS8 Centromere



**Figure 4.7 EAS 8 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS8 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey (bottom)

### Sequence features of EAS8 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	2	6	-	-
Gunazhong Donkey	14	6	-	-

**Table 4.13** Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS8

The CENP-A reads from EAS8 map to the Guanzhong donkey contig gi|933832210|gb|JREZ01001199.1|: 1,100-113,305 giving a broad gapped irregular shaped peak distribution containing a spiked domain. The peak profile in EquDonk gives a similar profile but the spike domain is centered. The shift in spike domain in EquDonk is explained in the schematic where rearrangement between both individuals is observed. Regions of sequence divergence as well as sequence duplication can also be seen. There are two instances of sequence insertion in EquDonk of 197bp and 3244bp in size. In the Guanzhong donkey there are 14 instances of insertions ranging in size from 21bp to 3272bp. Duplicated sequences vary in size from 70bp to 2034bp.

#### *Repetitive elements across the EAS8 centromere domain in the donkey*

The abundance of SINEs at this Guanzhong donkey domain was 2.37% less than whole genome levels. There was an increase of 10.13% in overall LINE abundance, with L1 elements rising from 16.09% to 27.8% while L2 levels decreased by 0.88%. Overall LTR element abundance was down by 2.29%, with a decrease in ERVL (0.24%), ERVL-MaLRs (0.63%) and conversely a 1.07% increase in ERV class I. There was also a decrease in hAT-Charlie (1.5%) and TcMar-Tigger elements (0.59%) at this domain.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	11	1795	1.6	3.67
ALUs	0	0	0	0
MIRs	11	1795	1.6	3.63
<b>LINEs:</b>	46	35706	31.82	21.69
LINE1	28	31196	27.8	16.09
LINE2	18	4510	4.02	4.9
L3/CR1	0	0	0	0.5
<b>LTR elements:</b>	12	4703	4.19	6.48
ERVL	5	2189	1.95	2.19
ERVL-MaLRs	5	2343	2.09	2.72
ERV_classI	1	116	0.1	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	7	967	0.86	3.82
hAT-Charlie	3	508	0.45	1.95
TcMar-Tigger	3	381	0.34	0.93

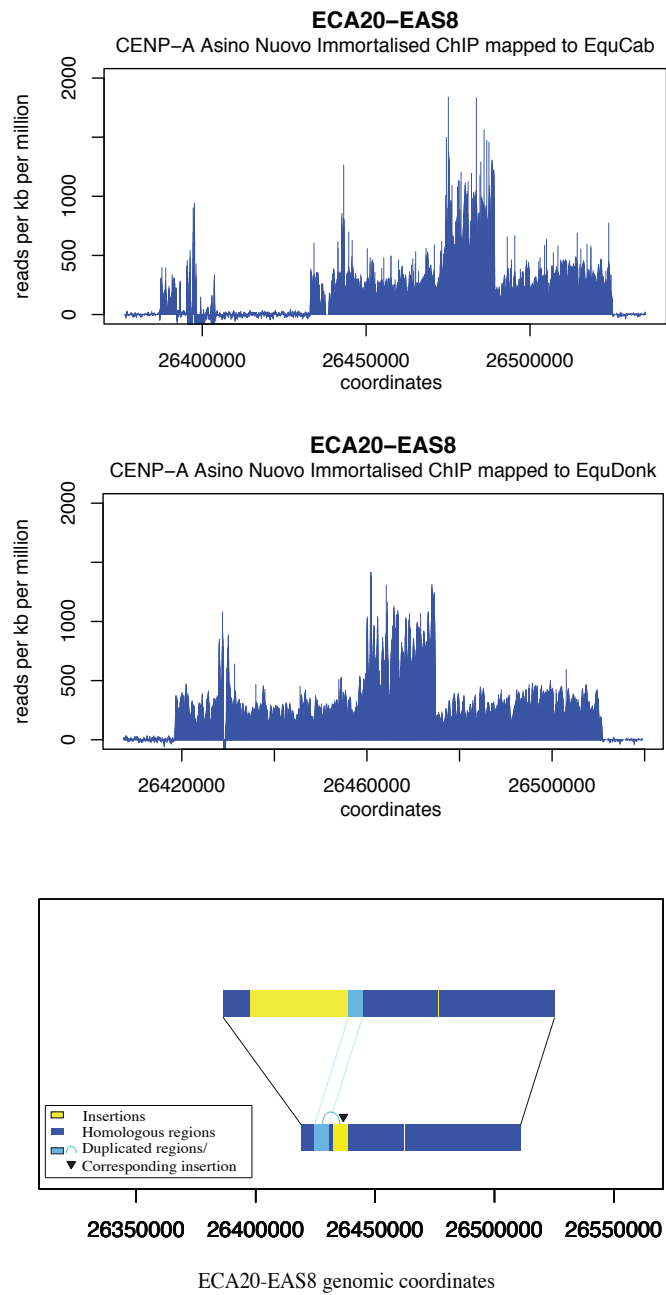
**Table 4.14** Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi|933832210|gb|JREZ01001199.1| (EAS8) compared with whole genome levels

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	10	1727	1.85	3.67
ALUs	0	0	0	0
MIRs	10	1727	1.85	3.63
<b>LINEs:</b>	32	26182	28.03	21.69
LINE1	19	23040	24.66	16.09
LINE2	13	3142	3.36	4.9
L3/CR1	0	0	0	0.5
<b>LTR elements:</b>	12	5155	5.52	6.48
ERV1	5	2321	2.48	2.19
ERV1-MaLRs	5	2343	2.51	2.72
ERV_classI	1	436	0.47	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	6	905	0.97	3.82
hAT-Charlie	3	508	0.54	1.95
TcMar-Tigger	2	319	0.34	0.93

**Table 4.15 Summary of repetitive elements across that span the centromere of EquDonk (EAS8)**

Examination of repetitive elements in regions of insertion in EquDonk showed two cases of LINE/L2 from L2 and L2a subfamilies, SINE/MIR from the MIRb subfamily and LTR/ERV1-MaLR from the MLT1B and MLT1D subfamilies. In EquDonk duplicated regions there were two cases of LINE/L1 from the L1M5 and L1ME4a subfamilies others repeats present included low complexity and simple repeats. The repetitive elements across these domains show the abundance of SINEs (2.68%) and LINEs (13.38%) is below the genomic average, while abundance of LTR elements (10.63%) has increased. In the Guanzhong donkey, the insertion and duplicated regions contained six instances of LINE/L2 from the L2 and L2a subfamily, six instances of LINE/L1 from L1ME3B, 3A, L1M3, L1ME3A subfamilies, while the only other elements present were simple repeats. The overall abundance of LINEs in these domains were 32.15%, 10.43% higher than the genomic average. This domain was 18791bp larger in the Guanzhong donkey spanning 112206bp while the EAS8 centromere in EquDonk spanned 93415bp, these domains shared 99% sequence identity.

## The EAS8 Centromere



**Figure 4.8 EAS 8 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the horse orthologous domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS8 centromere region in EquDonk compared to the orthologous region in EquCab (bottom)

## Sequence features of EAS8 centromere

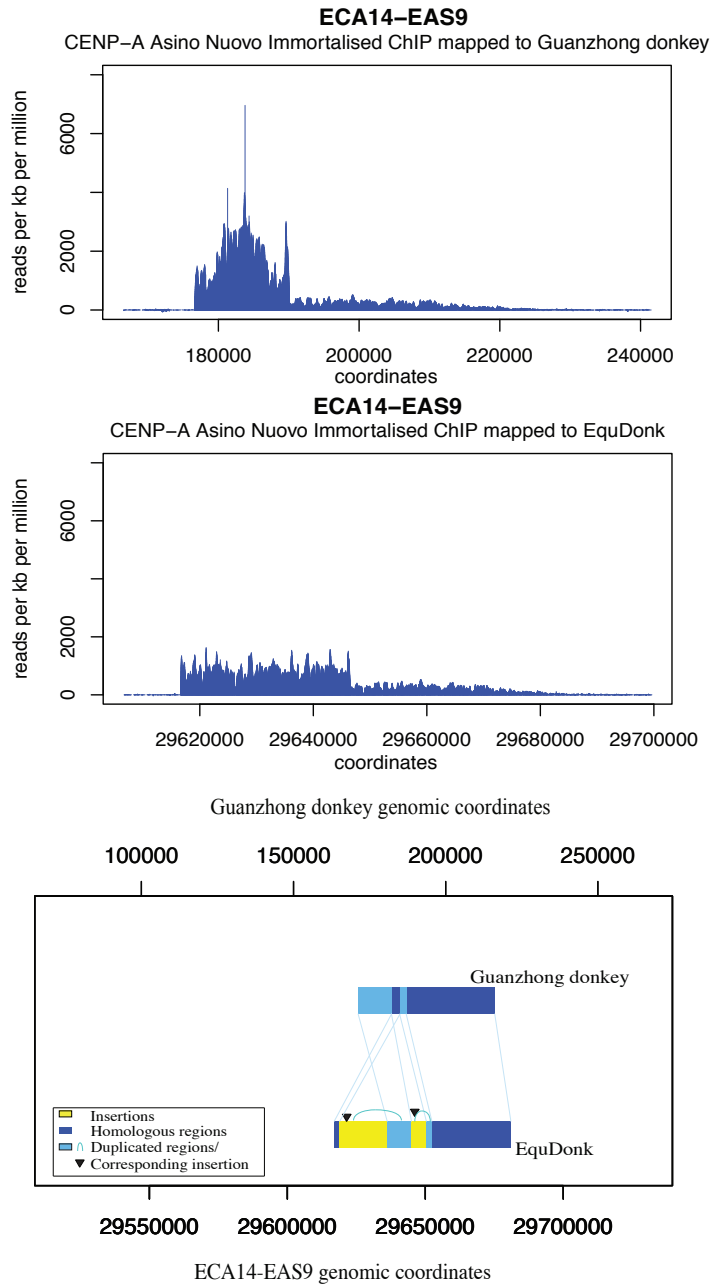
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	1	1	-	-
EquCab	3	1	-	-

**Table 4.16** Summary of sequence variation between EquDonk and EquCab at the EAS8 centromere

The EAS8 centromere was mapped to horse coordinates 26,386,390-26,525,316nt on chromosome 20, a broad irregular shaped peak profile with a spiked domain in the center of the enrichment profile was observed. A large non-homologous domain between 26438697-26476627nt spanning 40,755bp was present in the horse and missing from the EquDonk assembly. Upstream from the large insertion domain in EquCab there was a 6262bp single copy sequence that was duplicated in EquDonk.

Repetitive sequence analysis of the EquCab inserted sequences showed an abundance of LINES, particularly L1 elements with 23 instances from subfamilies L1MA, L1ME, L1M4 and L1M5. There were four instances of L2 elements from subfamilies L2a and L2c. Also present in fewer instances were SINEs/MIR, LTR/ERV1-MaLR (MLT1C2) and hAT-Charlie elements. The horse duplicate sequences contained, one instance of LINE/L1 from L1ME subfamily, a SINE/MIR and an LTR/ERV1-MaLR (MLT1C2) element. The overall abundance of LINES in these domains was higher than the genomic average at 34.36%, while the levels of SINEs (0.81%), LTR (1.68%) and DNA elements (0.18%) were reduced. Analysis of repetitive elements in the EquDonk duplicated and inserted regions showed the same elements as in the EquCab duplication, a strong enrichment in LINES, twenty three copies of L1 (L1MA9, L1Med, L1M4c, L1M3de, L1Meh, L1ME3A), four instances of L2 (L2a, L2c), along with two SINE/MIR copies, two hAT-Charlie copies and an LTR/ERV1-MaLR (MLT1C2). The overall abundance of LINES (13.38%) and SINEs (2.68%) in this domain was below while LTR element (10.63%) levels were above the genomic average. The centromere domain in EquDonk spanned 93415bp while in EquCab, including the non-homologous region, the domain spanned 138927bp and shared 98% sequence identity.

## The Eas9 Centromere



**Figure 4.9 EAS 9 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS9 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey (bottom)



## Sequence features of EAS9 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	3	2	-	-
Gunazhong Donkey	-	3	-	-

**Table 4.17** Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS9

CENP-A ChIPSeq reads from EAS9 were mapped to the Guanzhong donkey contig gi|933836078|gb|JREZ01000366.1|: 176,538-221,420nt, an irregular shaped peak profile was observed with a large spike domain spanning 13kb in the guanzhong donkey. As seen in the schematic the sequence under the spike enrichment is duplicated in EquDonk, giving a more dispersed 30kb spike peak.

### *Repetitive elements across the EAS9 centromere domain in the donkey*

There were five times less SINEs at this Guanzhong donkey domain when compared to levels observed across the whole genome. LINE levels were almost doubled, with no L2 elements present and LINE/L1 increasing from 16.09% to 42.82%. A 4.16% decrease is observed in LTR elements. In contrast there was a 1.9% increase in DNA elements due to a 4.43% increase in TcMar-Tigger, while hAT-Charlie levels dropped by 1.59%.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	3	321	0.72	3.67
ALUs	0	0	0	0
MIRs	3	321	0.72	3.63
<b>LINEs:</b>	24	19269	42.93	21.69
LINE1	23	19221	42.82	16.09
LINE2	0	0	0	4.9
L3/CR1	1	48	0.11	0.5
<b>LTR elements:</b>	4	1042	2.32	6.48
ERV1	4	1042	2.32	2.19
ERV1-MaLRs	0	0	0	2.72
ERV_classI	0	0	0	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	4	2567	5.72	3.82
hAT-Charlie	1	162	0.36	1.95
TcMar-Tigger	3	2405	5.36	0.93

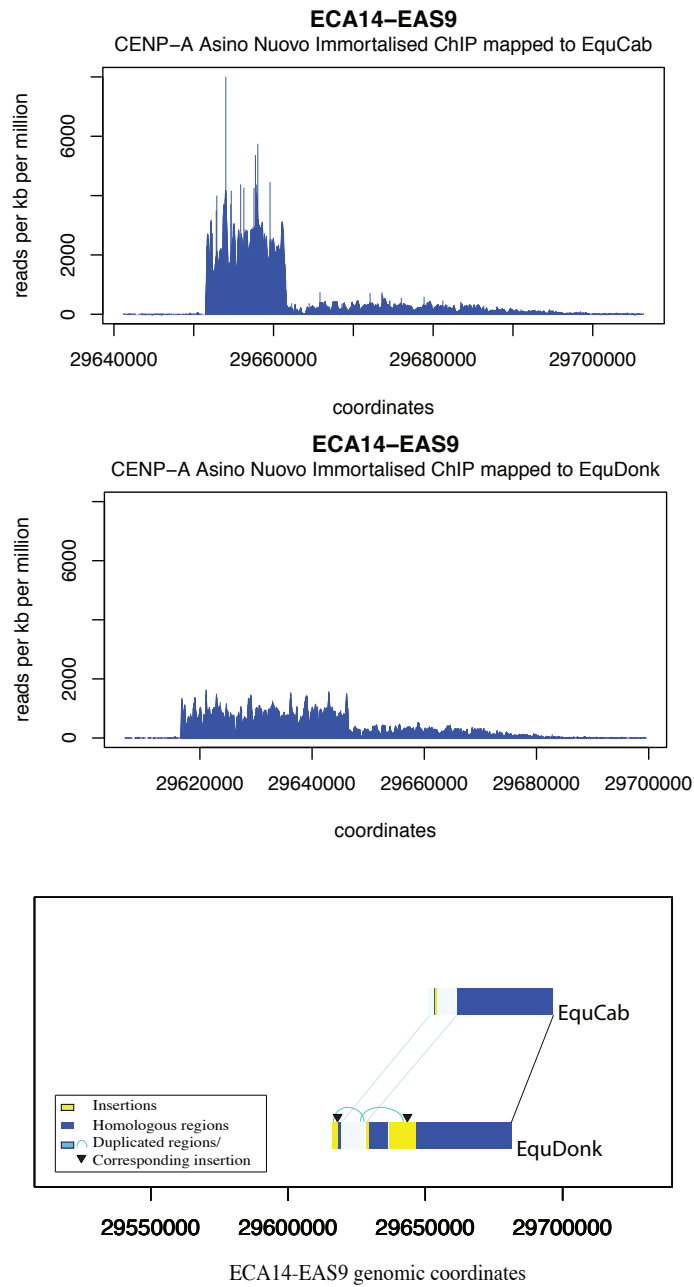
**Table 4.18** Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi|933836078|gb|JREZ01000366.1| (EAS9) compared with whole genome levels

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	2	208	0.28	3.67
ALUs	0	0	0	0
MIRs	2	208	0.28	3.63
<b>LINEs:</b>	34	36636	49.7	21.69
LINE1	33	36588	49.63	16.09
LINE2	0	0	0	4.9
L3/CR1	1	48	0.07	0.5
<b>LTR elements:</b>	5	1289	1.75	6.48
ERV1	5	1289	1.75	2.19
ERV1-MaLRs	0	0	0	2.72
ERV_classI	0	0	0	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	5	5107	6.93	3.82
hAT-Charlie	1	162	0.22	1.95
TcMar-Tigger	4	4945	6.71	0.93

**Table 4.19** Repetitive elements across the EAS9 centromere EquDonk compared with whole genome levels

Repetitive elements in the EquDonk duplicated and inserted regions included eight examples of LINE/L1 from subfamilies L1MC, L1ME3D, L1MB8 and CL1MC1, two SINE/MIR copies from MIRc and MIRb. Taken together the levels of SINEs (8.06%) in these domains are 4.39% higher than the genomic average, LINE abundance (23.87%) is also above the genomic average. In the Guanzhong donkey duplicated regions there was an abundance of LINE/L1 elements with ten examples from L1ME3D, L1MCa, C L1MC and L1MB8 subfamilies, there was also two instances of SINE/MIR. Taken together, the LINE abundance at these domains was 45.25%, 23.59% higher than the genomic average, while SINE levels (2.32%) were reduced. Given the copy number variation in EquDonk spike domain, the entire centromere spans 73715bp, 28832bp more than the Guanzhong donkey orthologous domain, the sequences share 99% identity.

## The EAS9 Centromere



**Figure 4.10 EAS9 centromere donkey versus horse comparison.** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the horse orthologous region (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS9 centromere region in EquDonk compared to the orthologous region in EquCab (bottom)

## Sequence assembly and features of EAS9 centromere

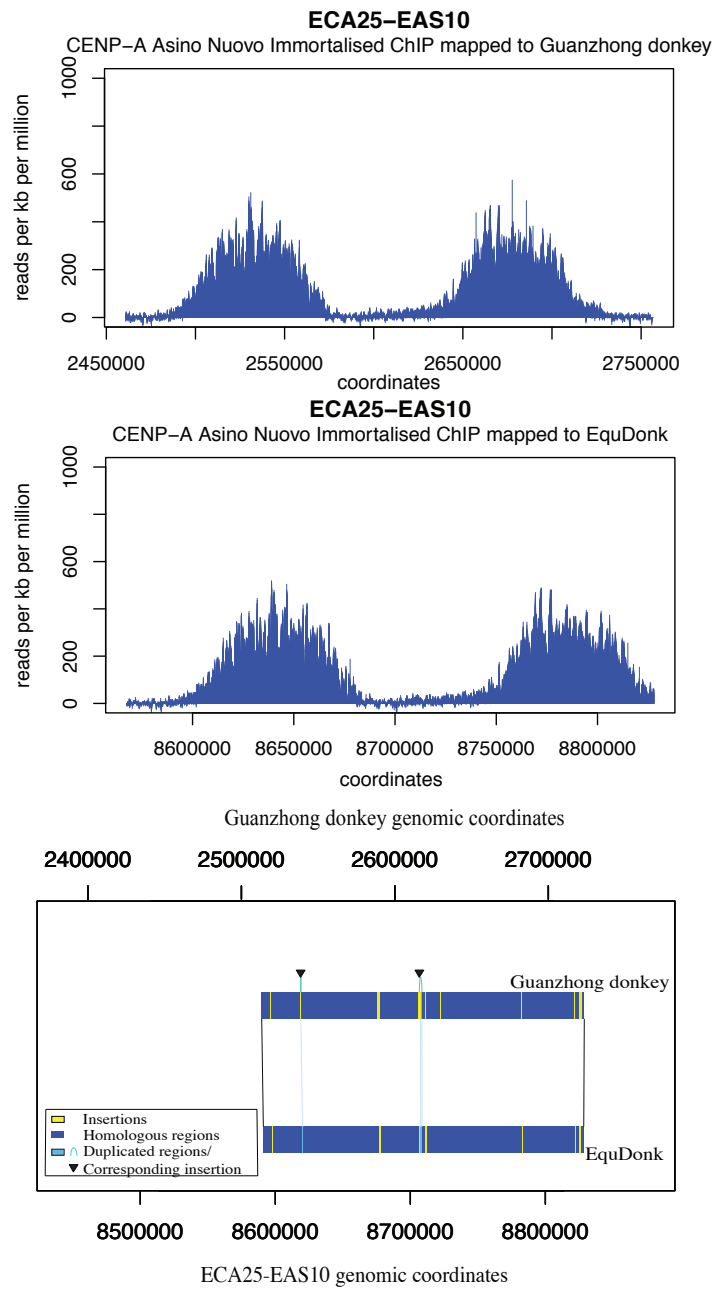
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	5	3	-	-
EquCab	1	3	-	-

**Table 4.20** Summary of sequence variation between EquDonk and EquCab at the EAS9 centromere

The centromere function of EAS9 were mapped to the horse genomic coordinates chr14: 29,651,149-29,696,370 nt giving an irregular shaped peak profile that spanned 45222bp with a spike domain spanning 10kb at the start of the centromere. In Equdonk there is duplicated sequence present in the spike that is only present once in EquCab, giving EquDonk a broader 30kb spike peak illustrated in the schematic. The duplicated sequences spanned 1686bp, 2025bp and 7224bp.

Analysis of repetitive elements in the inserted and duplicated domains in EquDonk showed an abundance of LINE/L1s from L1ME3D, L1MB8, L1MC1 and L1MC subfamilies with 14 instances. Also present in these domain LTR elements as well as simple repeats. The levels of LINEs across these domains are 51.49%, 29.8% higher than the genomic average, the abundance of LTR elements is lower than the genomic average (1.09%). In duplicated and inserted sequences of the horse there were 7 instances of LINE/L1s from L1ME3D, L1MB8, L1MC1 and L1MC subfamilies and one instance of LTR/ERV1 (LTR16B2) and a larger overall CENPA associated domain of 73715bp. Combined these sequences contained 49.92% LINEs compared to the genomic average of 21.69%, while 2.10% of the sequences contained LTR elements. Taken into the account the duplicated sequences present in the EquDonk spike peak, the domains spans 73715bp while the corresponding horse loci spans 45222bp.

## The EAS10 Centromere



**Figure 4.11 EAS10 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS10 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey

## Sequence features of EAS10 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	6	1	-	-
Guanzhong donkey	8	1	-	-

**Table 4.21** Summary of sequence variation between EquDonk and the Guanzhong on EAS10

The centromere domain of EAS10 was mapped to the Guanzhong donkey contig gi|933836905|gb|JREZ01000195.1|: 2,482,794-2,734,376nt. There were six instances of sequences variation in the EquDonk domain spanning between 12bp and 1353bp, a single copy 43bp sequence in EquDonk was duplicated in the Guanzhong donkey. There were eight instances of sequence insertion in the guanzhong domain spanning between 10bp and 2200bp.

### *Repetitive elements across the EAS10 centromere domain in the donkey*

The abundance of SINEs was reduced by 1.13% at this Guanzhong donkey domain when compared to levels observed across the whole donkey genome. Overall LINE levels were increased by 7.61%, with L1 elements rising by 10.71% while L2 elements decreased by 2.77%. A decrease in was also observed LTR (0.51%), with all class levels decreasing and DNA elements (2.51%).

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	40	6394	2.54	3.67
ALUs	0	0	0	0
MIRs	40	6394	2.54	3.63
<b>LINEs:</b>	116	73711	29.3	21.69
LINE1	87	67433	26.8	16.09
LINE2	24	5370	2.13	4.9
L3/CR1	4	688	0.27	0.5
<b>LTR elements:</b>	41	15029	5.97	6.48
ERV	8	3940	1.57	2.19
ERV-L-MaLRs	19	5561	2.21	2.72
ERV_classI	13	5387	2.14	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	18	3289	1.31	3.82
hAT-Charlie	9	1445	0.57	1.95
TcMar-Tigger	1	346	0.14	0.93

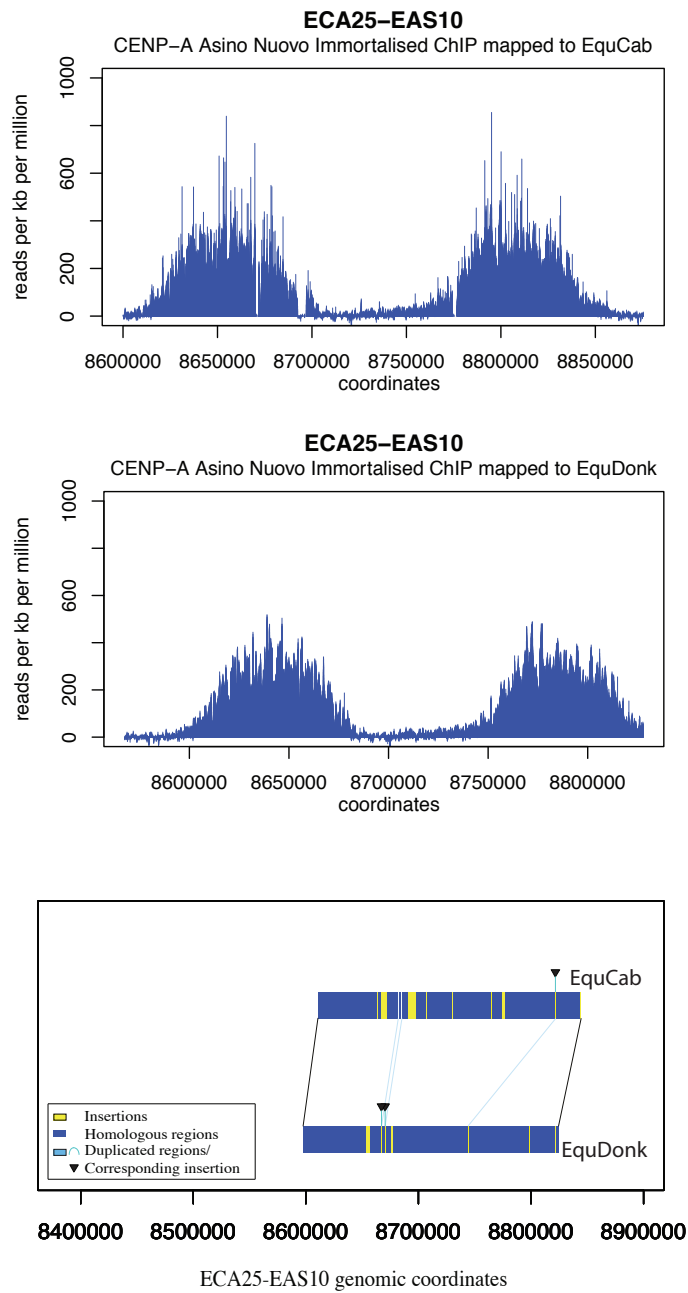
**Table 4.22** Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi|933836905|gb|JREZ01000195.1| (EAS10) compared with whole genome levels

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	39	6301	2.55	3.67
ALUs	0	0	0	0
MIRs	39	6301	2.55	3.63
<b>LINEs:</b>	108	70851	28.63	21.69
LINE1	81	65023	26.28	16.09
LINE2	22	4920	1.99	4.9
L3/CR1	4	688	0.28	0.5
<b>LTR elements:</b>	41	14967	6.05	6.48
ERV1	9	4028	1.63	2.19
ERV1-MaLRs	19	5552	2.24	2.72
ERV_classI	13	5387	2.18	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	18	3278	1.32	3.82
hAT-Charlie	9	1445	0.58	1.95
TcMar-Tigger	1	346	0.14	0.93

**Table 4.23** Repetitive elements across the EAS10 centromere EquDonk compared with whole genome levels

There were no repetitive elements present in regions of duplication in either donkey individual. Analysis of inserted regions in EquDonk showed the sole repetitive element as LINE/L1, with five copies from both the L1M2 and L13 subfamilies. The abundance of LINEs across all the inserted sequences was 71.42%, 3.29 times than the whole genomic average. Similarly in the Guanzhong donkey regions of insertion all repetitive elements observed were LINEs, four cases of L1 from subfamilies L1M3 and HAL1 and one case of L2 from subfamily L2a. The overall abundance of LINEs at these domains was 70.76%, 3.26 times higher than the genomic average. The EAS10 centromere domain in EquDonk was similar in size (247430bp) to the Guanzhong donkey orthologous domain (251582bp) with which it also shared high sequence identity (99%).

## The EAS10 Centromere



**Figure 4.12 EAS10 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the horse orthologous domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS10 centromere region in EquDonk compared to the orthologous region in EquCab



## Sequence features of EAS10 centromere

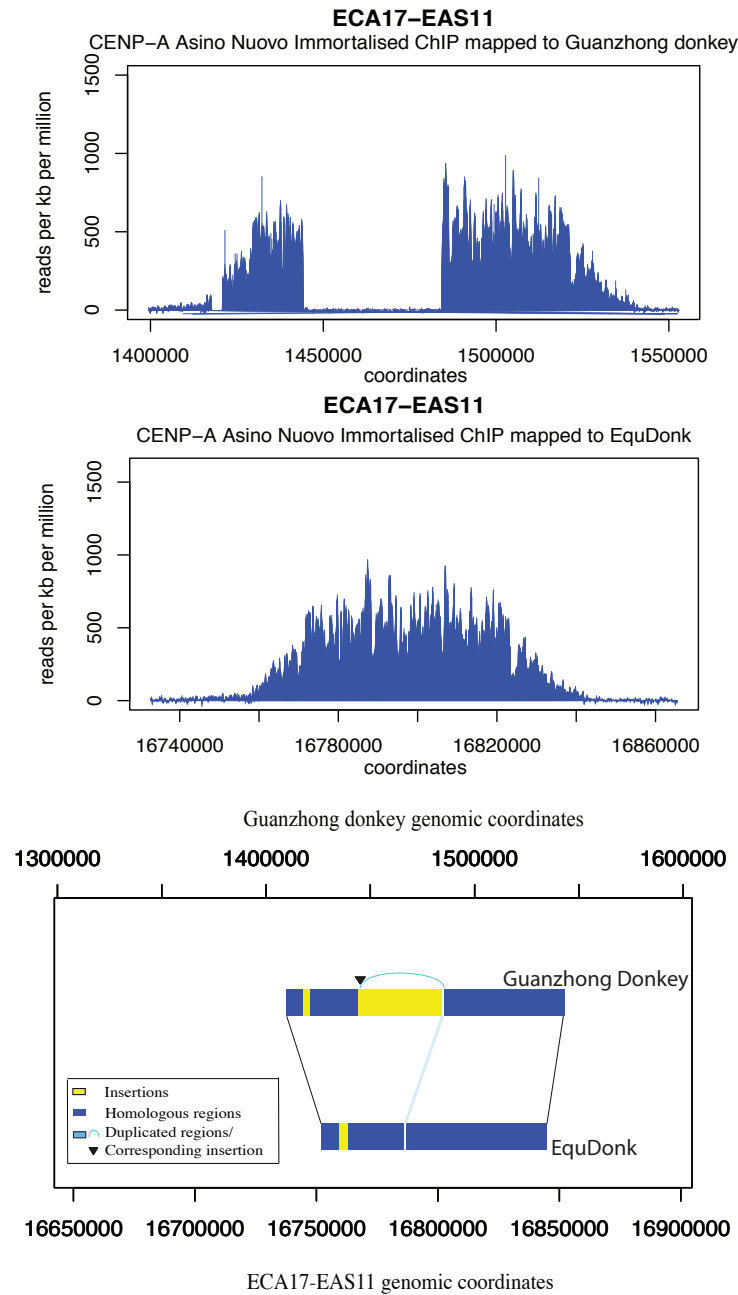
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	8	2	-	-
EquCab	10	4	-	-

**Table 4.24 Summary of sequence variation between EquDonk and EquCab on EAS10**

The EAS10 centromere domain was mapped to the horse orthologous region chr25: 8,609,949-8,865,465nt. There were 8 instances of sequences present in EquDonk but absent from EquCab spanning between 15bp and 3165bp. Regions of duplicated spanned between 15bp and 38bp. In EquCab there were ten examples of sequence insertion spanning between 15bp and 6920bp.

Analysis of repetitive sequences present in regions of insertion in both EquDonk and EquCab showed enrichment in LINE/L1, with three instances in EquDonk from subfamilies L1M5 and L1M2 and fifteen instances in EquCab from subfamilies L1MEc, L1M2, L1P4, L1MC4 and L1MA9. In regions of duplication, both genomes contained only simple repeats present. Taken together these domains in EquDonk contained sequences that were 22.17% LINEs slightly above the genomic average while in the horse the sequences 49.92% were LINEs, 2.3 times higher than the genomic average. The EquCab orthologous domain was 8087bp larger than the EAS10 centromere domain, which spanned 247430bp. The domains shared 98% sequence similarity.

## The Eas11 Centromere



**Figure 4.13 EAS11 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS11 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey

## Sequence features of EAS11 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	1	1	-	-
Guanzhong Donkey	3	1	-	-

**Table 4.25 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS11**

The EAS11 centromere domain was mapped to the orthologous region of the Guanzhong donkey which was originally on two different contigs that were combined gi|933835325|gb|JREZ01000504.1|: 39,591 - 92,854nt and gi|933831881|gb|JREZ01001282.1|: 18,249 - 75,463nt. A single gaussian peak was observed in both donkey individuals with a large insertion spanning 39532bp which is either a bonafied insertion or may be a result of “stitching” the two contigs together. The Guanzhong donkey contained two other sequence insertions spanning 633bp and 3244bp and a 633bp duplication present in one copy at the EquDonk centromere. EquDonk contained one sequence insertion spanning 3313bp.

### *Repetitive elements across the EAS11 centromere domain in the donkey*

The abundance of SINEs was reduced by 1.37% at this Guanzhong donkey domain when compared to levels observed across the whole donkey genome. Overall LINE levels were increased by 13.88%, with L1 elements rising by 14.92% while L2 elements decreased by 0.97%. An increase was observed LTR (5.71%) and DNA elements (0.44%).

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	21	3647	2.30	3.67
ALUs	0	0	0	0
MIRs	21	3647	2.30	3.63
<b>LINEs:</b>	61	56458	35.57	21.69
LINE1	41	49219	31.01	16.09
LINE2	15	6239	3.93	4.9
L3/CR1	2	276	0.17	0.5
<b>LTR elements:</b>	25	19347	12.19	6.48
ERV1	7	5837	3.68	2.19
ERV1-MaLRs	9	3779	2.38	2.72
ERV_classI	8	9591	6.04	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	17	6763	4.26	3.82
hAT-Charlie	10	3074	1.94	1.95
TcMar-Tigger	2	2597	1.64	0.93

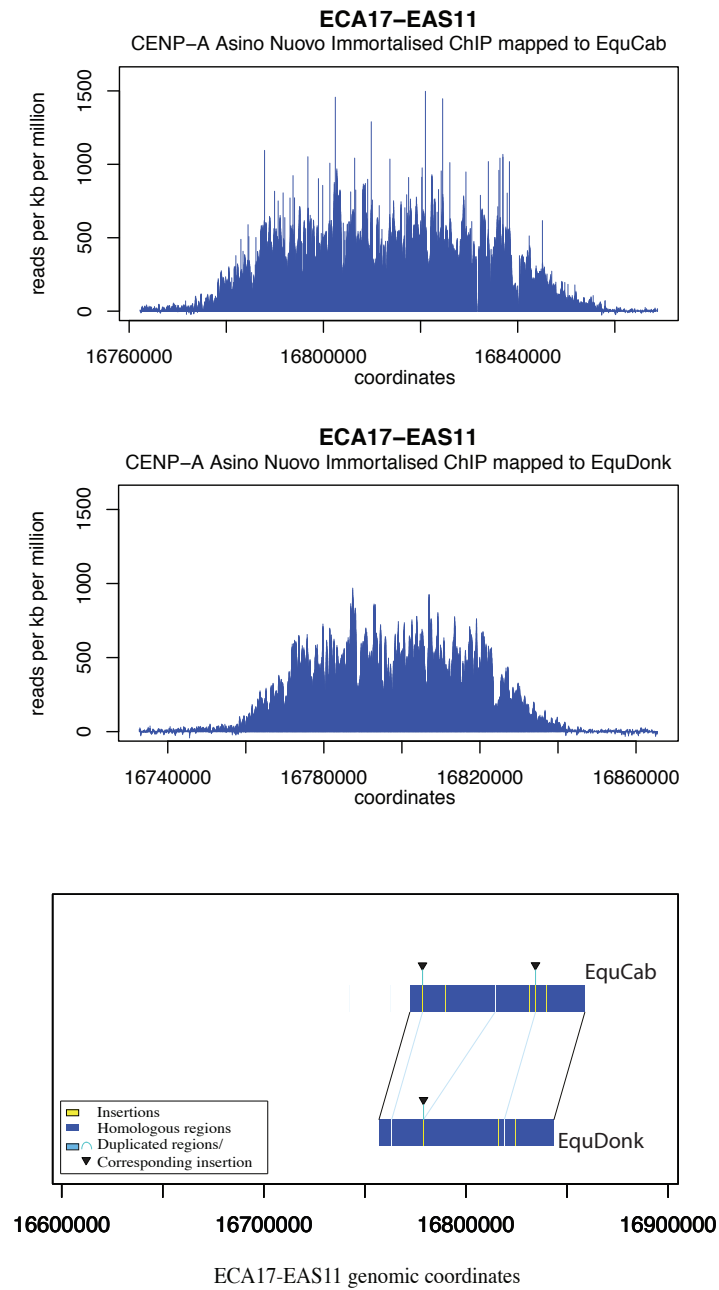
**Table 4.26 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contig gi|933835325|gb|JREZ01000504.1| & gi|933831881|gb|JREZ01001282.1| combined (EAS11) compared with whole genome levels.**

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	11	1837	1.97	3.67
ALUs	0	0	0	0
MIRs	11	1837	1.97	3.63
<b>LINEs:</b>	39	36453	39.18	21.69
LINE1	23	30177	32.44	16.09
LINE2	14	6151	6.61	4.9
L3/CR1	1	52	0.06	0.5
<b>LTR elements:</b>	14	9086	9.77	6.48
ERV1	5	5459	5.87	2.19
ERV1-MaLRs	6	2130	2.29	2.72
ERV_classI	2	1357	1.46	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	13	3974	4.27	3.82
hAT-Charlie	6	2537	2.73	1.95
TcMar-Tigger	2	327	0.35	0.93

**Table 4.27** Repetitive elements across the EAS11 centromere EquDonk compared with whole genome levels

Analysis of repetitive sequences present at regions of insertion at the EquDonk centromere showed one copy of a LINE/L1 element from the L1MA9 family spanning 3307bp. A L1 element was also present in the duplicated sequence spanning 478bp. The abundance of LINEs across these sequences was 96.22%. Repetitive elements in the Guanzhong donkey insertion domains contained 25 examples of LINE/L1 from families L1MB5, L1MC3, HAL1b, L1M3c, L1m4, L1M2, L1MA6 L1M3, L1MA9 and L1MD. Also present were four copies of LTR/ERV1, six SINEs, LINE/L3s and DNA elements. The overall abundance of LINEs (43.19%), LTR elements (8.33%) and DNA elements were higher than the genomic average while SINE levels (2.61%) were lower. Analysis of duplicated domains show two copies of LINE/L1s from the L1M3de family.

## The EAS11 Centromere



**Figure 4.14 EAS11 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the horse orthologous domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS11 centromere region in EquDonk compared to the orthologous region in EquCab (bottom)

## Sequence features of EAS11 centromere

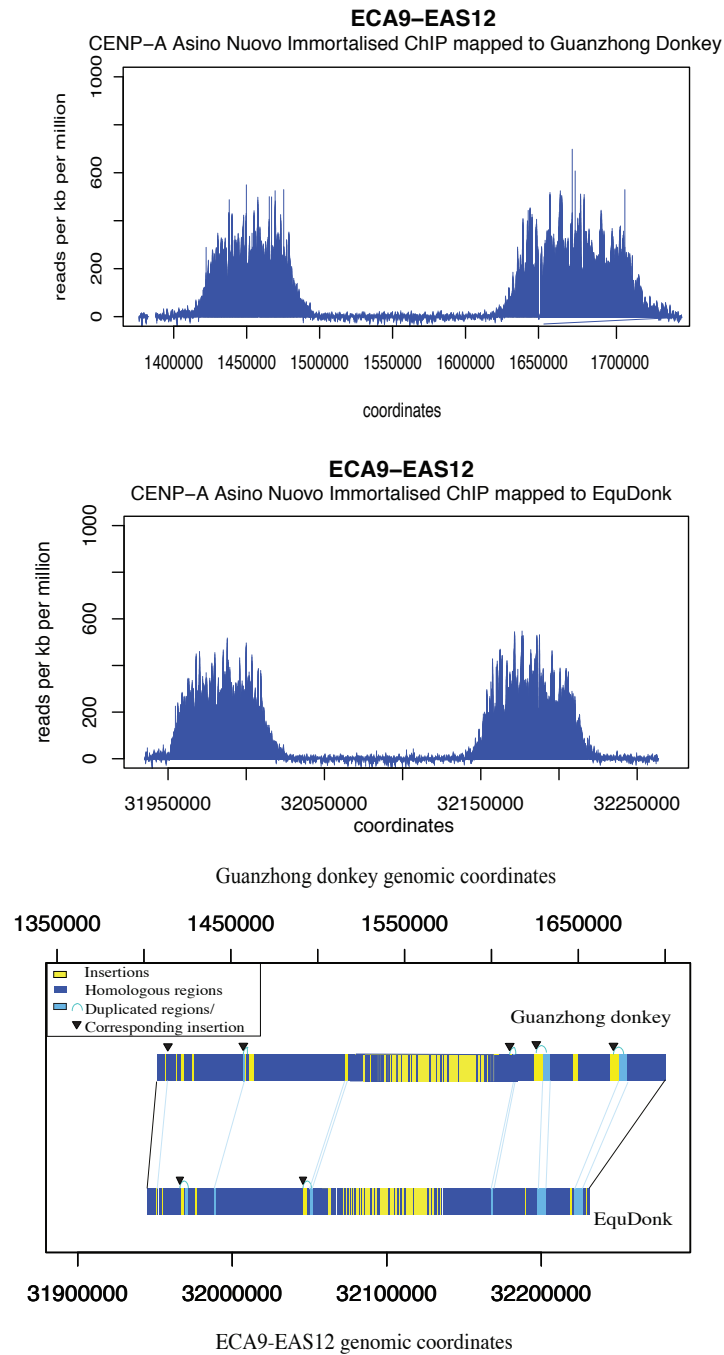
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	3	3	-	-
EquCab	6	3	-	-

**Table 4.28** Summary of sequence variation between EquDonk and EquCab on Eas11

The CENP-A ChIPSeq reads from EAS11 were mapped to the horse genomic coordinates 16,772,244-16,858,830nt on chromosome 17. EquDonk contains three instances of sequence insertion spanning 40bp, 129bp and 193bp. There are sequences present once in EquDonk that are duplicated in the horse genome, spanning 30bp and 73bp. A 40bp sequence present once in EquCab is duplicated in EquDonk. There are six examples of sequence insertion in EquDonk ranging from 19bp to 150bp.

No repetitive elements were present in the EquCab inserted or duplicated domains. In EquDonk, two instances of repetitive elements were present in the inserted sequences, a LINE/L1 from the L1ME1 subfamily and a single hAT-Charlie, no repetitive elements were present in duplicated domains. These domains taken together contain 24.27% LINE elements and 37.57% DNA elements an increase of 2.68% and 33.9% respectively compared to whole genome abundance. This domain spanned 86587bp in the horse and 93033bp in EquDonk with 92% sequence identity.

## The Eas12 Centromere



**Figure 4.15 EAS12 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS12 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey (bottom)

## Sequence features of EAS12 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	33	4	-	-
Guanzhong Donkey	37	5	-	-

**Table 4.29** Summary of sequence variation between EquDonk and the Guanzhong donkey on Eas12

The functional domain of the EAS12 centromere was mapped to two guanzhong donkey contigs gi|933837599|gb|JREZ01000107.1| & gi|933833617|gb|JREZ01000871.1| the contigs were combined. There a number of sequence insertions in both assemblies as well as four cases of single copy sequence in EquDonk that was duplicated in the Guanzhong donkey and five cases of single copy sequence in the Guanzhong donkey that was duplicated in EquDonk.

### *Repetitive elements across the EAS12 centromere domain in the donkey*

Analysis of repetitive elements across the Eas12 centromere orthologous domain in Guanzhong donkey showed a decrease in SINE abundance (1.14%). Overall LINE levels increased by 2.49% when compared to whole genome levels with L1 increasing by 2.86% while L2 levels decreased by 0.42%. LTR element abundance remained similar to whole genome levels (rising by 0.48%), as do DNA elements (decreasing by 0.01).

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	28	3972	1.29	3.67
ALUs	0	0	0	0
MIRs	28	3972	1.29	3.63
<b>LINEs:</b>	131	123747	40.15	21.69
LINE1	106	119148	38.65	16.09
LINE2	23	4145	1.34	4.9
L3/CR1	1	243	0.08	0.5
<b>LTR elements:</b>	50	17291	5.61	6.48
ERV1	19	7364	2.39	2.19
ERV1-MaLRs	15	5165	1.68	2.72
ERV_classI	12	3565	1.16	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	27	9486	3.08	3.82
hAT-Charlie	13	3099	1.01	1.95
TcMar-Tigger	5	2476	0.8	0.93

**Table 4.30** Repetitive elements across the EAS12 centromere EquDonk compared with whole genome levels

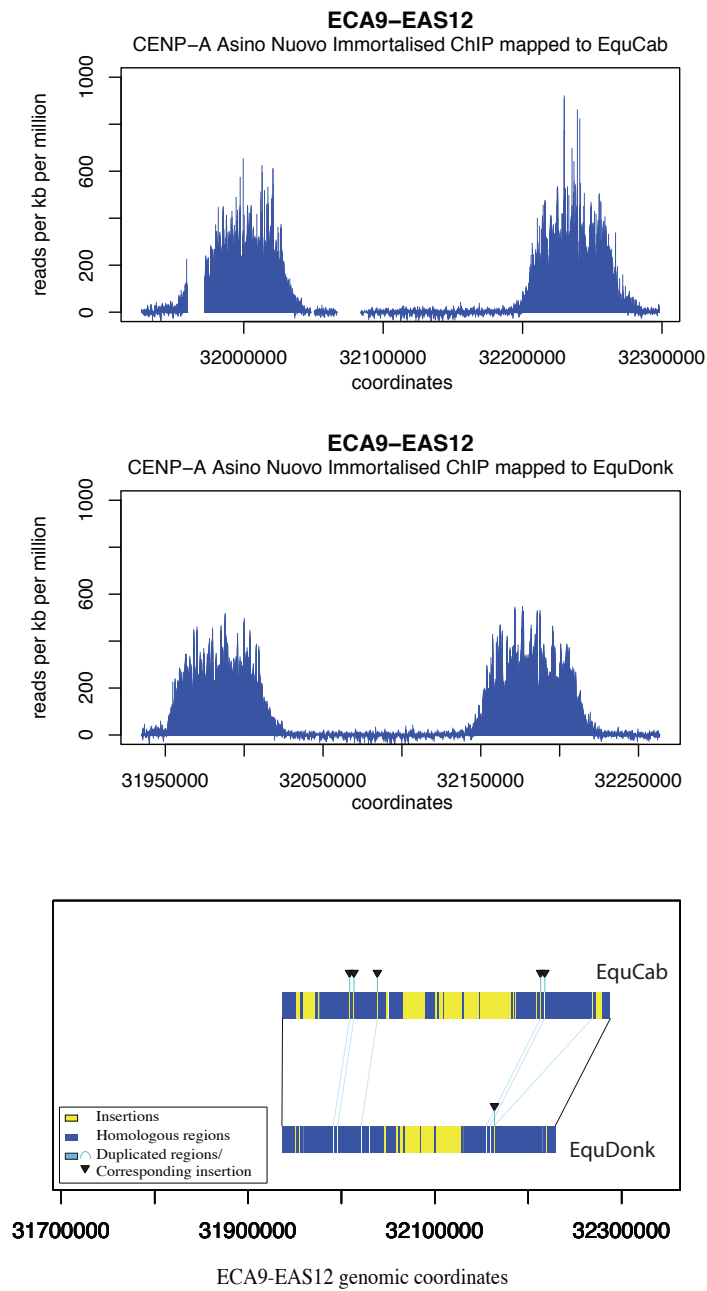


Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	935	145891	2.53	3.67
ALUs	0	0	0	0
MIRs	914	142948	2.48	3.63
<b>LINEs:</b>	2124	1391957	24.18	21.69
LINE1	1169	1090776	18.95	16.09
LINE2	809	257633	4.48	4.9
L3/CR1	98	28741	0.5	0.5
<b>LTR elements:</b>	898	400452	6.96	6.48
ERV1	303	164778	2.86	2.19
ERV1-MaLRs	343	134421	2.34	2.72
ERV_classI	175	77625	1.35	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	862	219443	3.81	3.82
hAT-Charlie	424	94654	1.64	1.95
TcMar-Tigger	203	64377	1.12	0.93

**Table 4.31 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi|933837599|gb|JREZ01000107.1| & gi|933833617|gb|JREZ01000871.1| (EAS12) compared with whole genome levels**

There was an abundance of LINE/L1s in the inserted regions of the Guanzhong donkey from subfamilies L1MD, L1M4, L1MC L1MB, L1ME, L1MA, L1M and L1M3. There was two instances of L2 elements, both from subfamily L2c as well as SINE/MIR, TcMar-Tc2, LTR/ERV1 and hAT-Tip100 present. In the duplicated regions there were 28 instances of LINE/L1, four examples of L2 and one example of LINE/RTE-BovB. Other repetitive elements include LTR/ERV1, SINEs and hAT-Charlie. Analysis of the EquDonk insertion domains also showed an abundance of LINE/L1 elements with 47 instances as well as examples L2 elements, SINEs/MIR, TcMar-Tc2, LTR/ERV1, hAT-Charlie and Hat-Tip100. In the duplicated domains, there was also an enrichment of LINEs along with examples of LTR/ERV1, SINEs, hAT-Charlie and hAT-Tip100.

## The Eas12 Centromere



**Figure 4.16 EAS12 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the horse orthologous region (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS12 centromere region in EquDonk compared to the orthologous region in EquCab

## Sequence assembly of EAS12 centromere

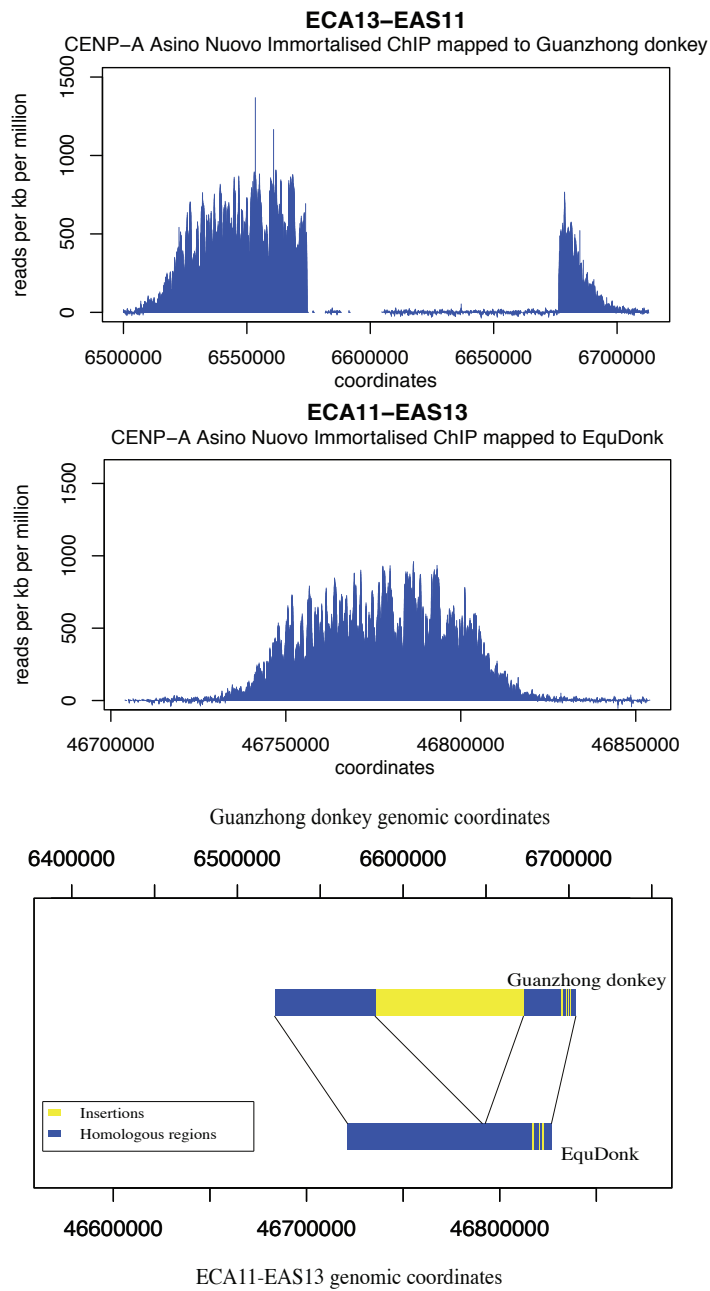
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	12	9	-	-
EquCab	26	9	-	-

**Table 4.32** Summary of sequence variation between EquDonk and EquCab on EAS12

The EAS12 centromere was mapped to horse chromosome 9: 31,936,755-32,287,869nt. Two distinct Gaussian peak profiles were observed in both species with a 12425bp insertion between coordinates 31971838-31975348nt in EquCab. The schematic illustrates that the majority of inserted sequences between both individuals are outside the CENP-A binding domain. Since EquDonk was assembled using CENP-A ChIPSeq reads the fidelity of this sequence is not expected to be high. Nonetheless given the centromeres ability to ‘slide’ in the Equids, this domain was taken into consideration for inspection of repetitive element. Twelve examples of sequence insertion was observed at the EquDonk centromere domain spanning between 5bp and 26575bp, there were eight single copy sequences present in EquDonk that were duplicated in EquCab and one single copy EquCab sequence that was duplicated in EquDonk. The twenty-six cases of sequence insertion in the EquCab domain ranged in size from 6bp to 32117bp.

Analysis of repetitive elements in the EquCab inserted sequences showed 57 examples of LINE/L1 elements, given the abundance virtually all the subfamilies. There were also instances, albeit substantially less than LINE/L1, of L2, SINE/MIR, LTR/ERV1, hAT-Tip100, hAT-Charlie and TcMar. In regions of duplication there were two instances of LINE/L1, both from L1MB subfamilies and two instances of LTR/ERV1. These regions taken together contained 0.75% SINEs, 40.76% LINES, 8.56% LTR elements and 2.21% DNA elements. In EquDonk insertion regions there was also an abundance of LINE/L1s, with 64 instances, from L1M, L1MB, L1mC, L1mE, L1MD and L1M4 subfamilies. Also present in these regions were LTR/ERV1, hAT-Tip100, TcMar and SINES/MIR. In the inserted and duplicated domains there was one example of LINE/L1 from subfamily L1MB and one example of LTR/ERV1. Combined these regions contain 0.41% SINEs, 41.68% LINES, 5.23% LTR elements and 2.97% DNA elements.

## The EAS13 Centromere



**Figure 4.17 EAS13 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS8 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey

### Sequence features of EAS13 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	2	-	-	-
Guanzhong donkey	3	-	-	-

**Table 4.33 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS13**

The functional domain of the EAS13 centromere was mapped to the guanzhong donkey orthologous domain gi|933838084|gb|JREZ01000066.1|: 6,503,425-6,710,816nt, giving a single gaussian shaped peak. A large insertion spanning 102256bp was present in the guanzhong donkey but absent from the EquDonk assembly, two other sequence insertions spanning 1038bp and 2505bp were also observed. The large insertion shares homology with the horse sequence, and the deletion appears to be unique to EquDonk, this will be further discussed in SECTION. There was no sequence duplication between the two individuals. EquDonk contained two sequence insertions spanning 820bp and 2381bp.

#### *Repetitive elements across the EAS13 centromere domain in the donkey*

There was a 1.16% decrease in SINEs at the Guanzhong donkey EAS13 centromere orthologous region when compared to levels observed across the entire donkey genome. Overall LINE presence had increased by 18.03%, with L1 elements (20.54%) accounting for the increase while L2 level had decreased by 2.31%. LTR elements in this domain had increase by 1.95%, with ERVL levels doubling by 2.23%. DNA element levels were down by 2.78%, with hAT-Charlie (1.21%) and TcMar-Tigger (0.7%) respectively.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	31	5199	2.51	3.67
ALUs	0	0	0	0
MIRs	31	5199	2.51	3.63
<b>LINEs:</b>	114	82368	39.72	21.69
LINE1	82	75971	36.63	16.09
LINE2	24	5372	2.59	4.9
L3/CR1	8	1025	0.49	0.5
<b>LTR elements:</b>	42	17485	8.43	6.48
ERVL	19	9171	4.42	2.19
ERVL-MaLRs	17	5484	2.64	2.72
ERV_classI	5	2185	1.05	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	12	2151	1.04	3.82
hAT-Charlie	9	1531	0.74	1.95
TcMar-Tigger	2	472	0.23	0.93

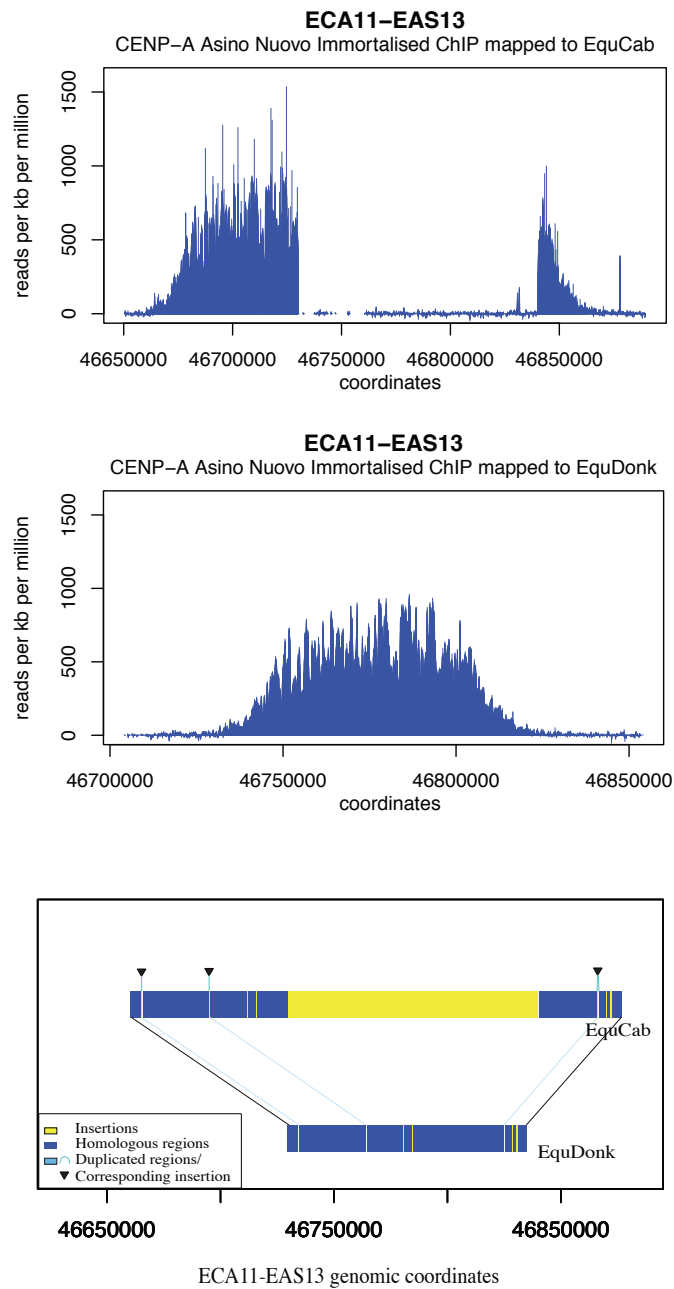
**Table 4.34 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi|933838084|gb|JREZ01000066.1| (EAS13) compared with whole genome levels**

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	14	2411	2.1	3.67
ALUs	0	0	0	0
MIRs	14	2411	2.1	3.63
<b>LINEs:</b>	56	45346	39.52	21.69
LINE1	40	42575	37.1	16.09
LINE2	11	2153	1.88	4.9
L3/CR1	5	618	0.54	0.5
<b>LTR elements:</b>	18	7876	6.86	6.48
ERV_L	8	4379	3.82	2.19
ERV_L-MaLRs	9	3324	2.9	2.72
ERV_classI	1	173	0.15	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	5	753	0.66	3.82
hAT-Charlie	3	475	0.41	1.95
TcMar-Tigger	2	278	0.24	0.93

**Table 4.35** Repetitive elements across the EAS13 centromere EquDonk compared with whole genome levels

Analysis of repetitive elements in the Guanzhong donkey inserted sequences showed a large array of repetitive elements including LINEs (L1, L2, CR1), SINEs (MIR), LTRs and DNA elements. The overall of abundance of SINEs is 2.60%, 1.07% less than whole genomic levels, LINE abundance is 37.96%, 16.27% higher than whole genomic levels, LTR element (11.82%) abundance is also higher while levels of DNA elements (1.58%) is reduced. The EquDonk Eas13 centromere spanned 114,752bp while the orthologous domain in the Guanzhong donkey excluding the large insertion spanned 105,135bp and shared 99% identity with EquDonk. Repetitive elements in the smaller insertions of both individuals showed a single LINE/L1 from subfamily L1ME3Cz and an LTR/ERV\_L-MaLR element. The abundance of repetitive elements in domains of insertion and duplication in EquDonk showed a decrease in LINEs (19.61%) and an increase in LTR elements (11.02%) compared to whole genome levels.

## The EAS13 Centromere



**Figure 4.18 EAS13 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the orthologous horse domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS8 centromere region in EquDonk compared to the orthologous region in EquCab

## Sequence features of EAS13 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	4	4	-	-
EquCab	12	4	-	-

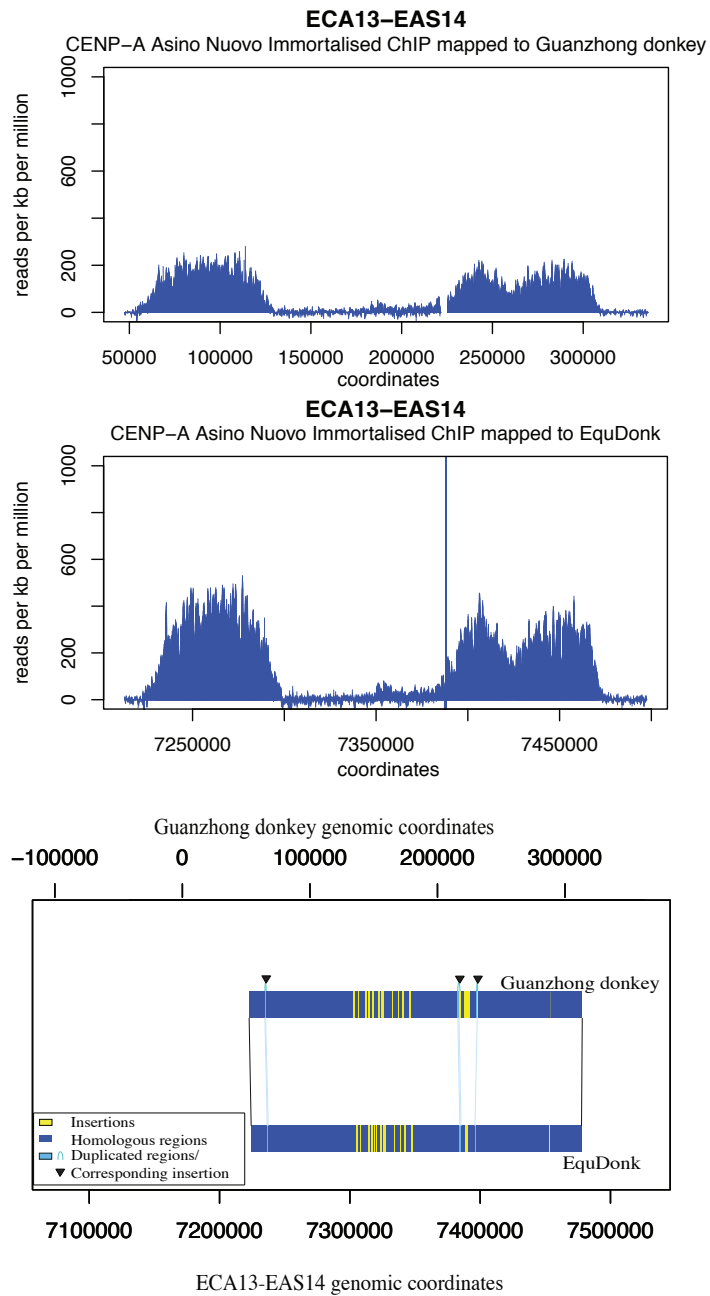
**Table 4.36** Summary of sequence variation between EquDonk and EquCab on Eas13

CENP-A ChIPSeq reads from the EAS13 centromere was mapped to the horse orthologous region chr11:46,660,406-46,879,576nt giving a gaussian like peak with a large sequence spanning 110229 bp present in the horse but not in the Equdonk assembly. There were 11 other instances of insertion in the EquCab domain ranging in size from 30bp to 852bp. In EquDonk there were four cases of sequence insertion ranging in size from 10bp to 757bp. There were four single copy sequences spanning 38bp, 42bp, 74bp and 81bp that were duplicated in EquCab.

Examination of repetitive elements present in the 110229 bp horse insertion showed the presence of LINES (L1, L2, CR1), SINEs (MIR), LTR elements (ERVL, ERV1, MaLR), hAT-Tip100, hAT-Charlie and TcMar-Tigger. Analysis of the smaller insertions showed three repetitive elements were present, a LINE/L1 (L1ME3Cz), LTR/ERVL (MER21) and a SINE (MIR). The overall abundance of elements in these domains showed an increase in SINEs (10.73%), LINES (36.62%) and LTR elements (20.23%) when compared to whole genome abundance. The inserted sequences in EquDonk contained one example of a 627bp LINE/L1 (L1ME3C) and one example of an LTR/ERVL-MaLR. Combined LINES occupied 36.62% of the inserted sequences while LTR elements occupied 20.23%. The Equdonk centromere spanned 114752bp while the EquCab orthologous domain without the insertion spans 108,942bp and shared 99% sequence identity.



## The EAS14 Centromere



**Figure 4.19 EAS14 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS14 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey

## Sequence features of EAS14 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	14	3	-	-
Guanzhong Donkey	20	3	-	-

**Table 4.37** Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS14

The centromere of EAS14 was mapped to the Guanzhong donkey orthologous domain gi|933833808|gb|JREZ01000828.1|: 49,410-315,452nt. There were fourteen examples of sequence insertion in EquDonk ranging in size from 8bp to 2094bp. There were three single copy sequences present once in EquDonk spanning 125bp, 429bp and 998bp that were duplicated in the Guanzhong donkey. In the Guanzhong donkey there were twenty examples of insertions ranging in size from 6bp to 1945bp.

### *Repetitive elements across the EAS14 centromere domain in the donkey*

Overall SINE abundance at the Guanzhong donkey orthologous region has decreased (1.46%) when compared to levels observed across the whole genome. Conversely the abundance of LINEs has increased by 7.91%, with an increase in L1 (11.1%) and a decrease in L2 (2.61%). LTR element levels have also increased (+2.05%), with a decrease in ERVL (-0.75%), ERVL-MaLR (-1.15%), ERV classI (-1.87%). DNA element abundance had also decreased (0.95%) compared to levels observed across the genome.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	41	5887	2.21	3.67
ALUs	0	0	0	0
MIRs	39	5770	2.17	3.63
<b>LINEs:</b>	113	78738	29.6	21.69
LINE1	83	72331	27.19	16.09
LINE2	29	6087	2.29	4.9
L3/CR1	0	0	0	0.5
<b>LTR elements:</b>	59	22685	8.53	6.48
ERVL	14	3819	1.44	2.19
ERVL-MaLRs	31	10285	3.87	2.72
ERV_classI	11	8082	3.04	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	47	7623	2.87	3.82
hAT-Charlie	32	5154	1.94	1.95
TcMar-Tigger	3	745	0.28	0.93

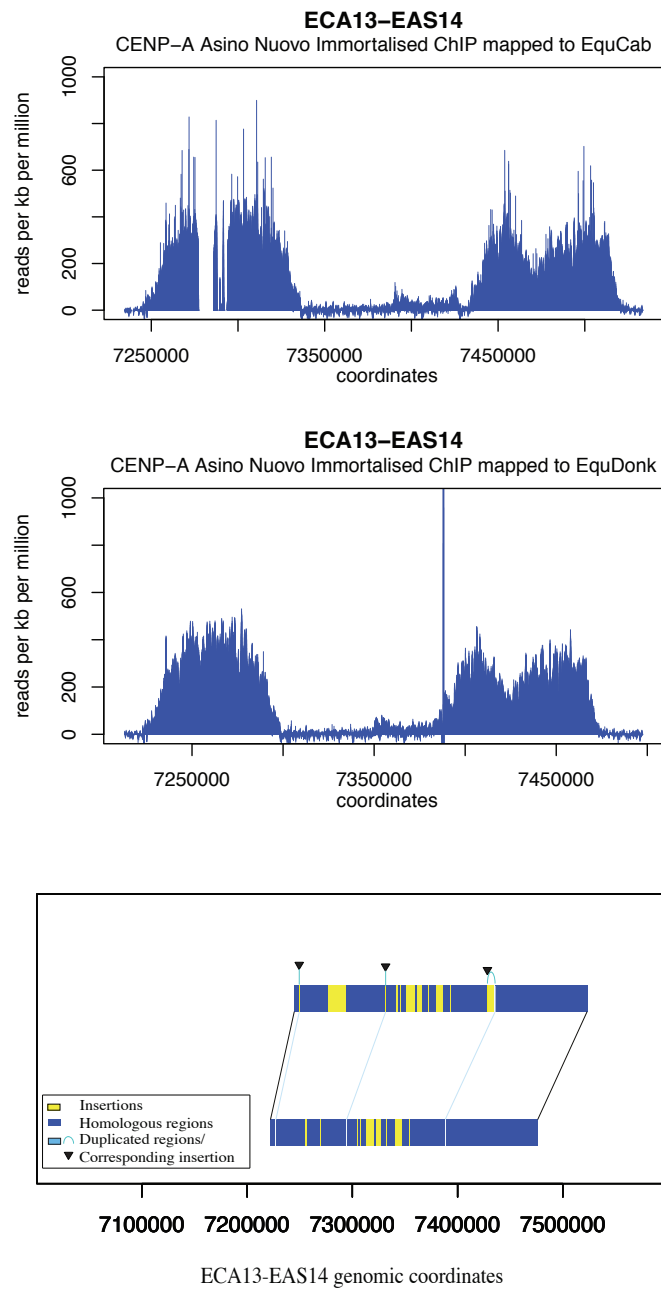
**Table 4.38** Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi|933833808|gb|JREZ01000828.1| (EAS14) compared with whole genome levels

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	41	5984	2.26	3.67
ALUs	0	0	0	0
MIRs	39	5867	2.22	3.63
<b>LINEs:</b>	110	78395	29.6	21.69
LINE1	79	72133	27.24	16.09
LINE2	30	5942	2.24	4.9
L3/CR1	0	0	0	0.5
<b>LTR elements:</b>	58	21488	8.11	6.48
ERVL	14	3446	1.3	2.19
ERVL-MaLRs	30	9597	3.62	2.72
ERV_classI	11	7946	3	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	47	7466	2.82	3.82
hAT-Charlie	32	5153	1.95	1.95
TcMar-Tigger	3	589	0.22	0.93

**Table 4.39** Repetitive elements across the EAS14 centromere EquDonk compared with whole genome levels

Analysis of repetitive element in the EquDonk inserted sequences showed nine instances of LINEs, eight L1s (HAL1, L1M3c, L1M3, L1M2) and one L2 (L2d) along with five examples of LTRs, three ERVL-MaLR and two ERV1, as well as a single SINE (MIR) and hAT-Tip100. Combined these domains contained less SINEs (3.67%), LINEs (20.99%) and DNA elements (1.34%) than observed across the whole genome while LTR element (11.72%) levels had increased. In the Guanzhong donkey insertion domain, there were seven instances of LINEs, 6 L1s (HAL1, L1ME2z and L1M3c) and one L2 (L2d), 6 examples of LTRs (ERV1, ERVL-MaLR, Gypsy) and an example of SINE (MIR) and hAT-Tip100 along with an A-rich low complexity repeat. Combined regions of insertion and duplication showed a decrease in SINEs (1.19%), LINEs (13.52%) and DNA elements (0.99%) while LTR elements (13.67%) showed an increase compared to whole genome levels. In the duplicated regions in EquDonk there was a LINE/L1 (L1MA6) and LTR/ERVL-MaLR similarly in the Guanzhong donkey the same elements are present in duplication. The EAS14 centromere domain spans 264813bp in EquDonk and 1230bp less in the Guanzhong donkey (266043bp) as well as sharing 99% sequence identity.

## The Eas14 Centromere



**Figure 4.20 EAS14 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the orthologous horse domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS14 centromere region in EquDonk compared to the orthologous region in EquCab

## Sequence features of EAS14 centromere

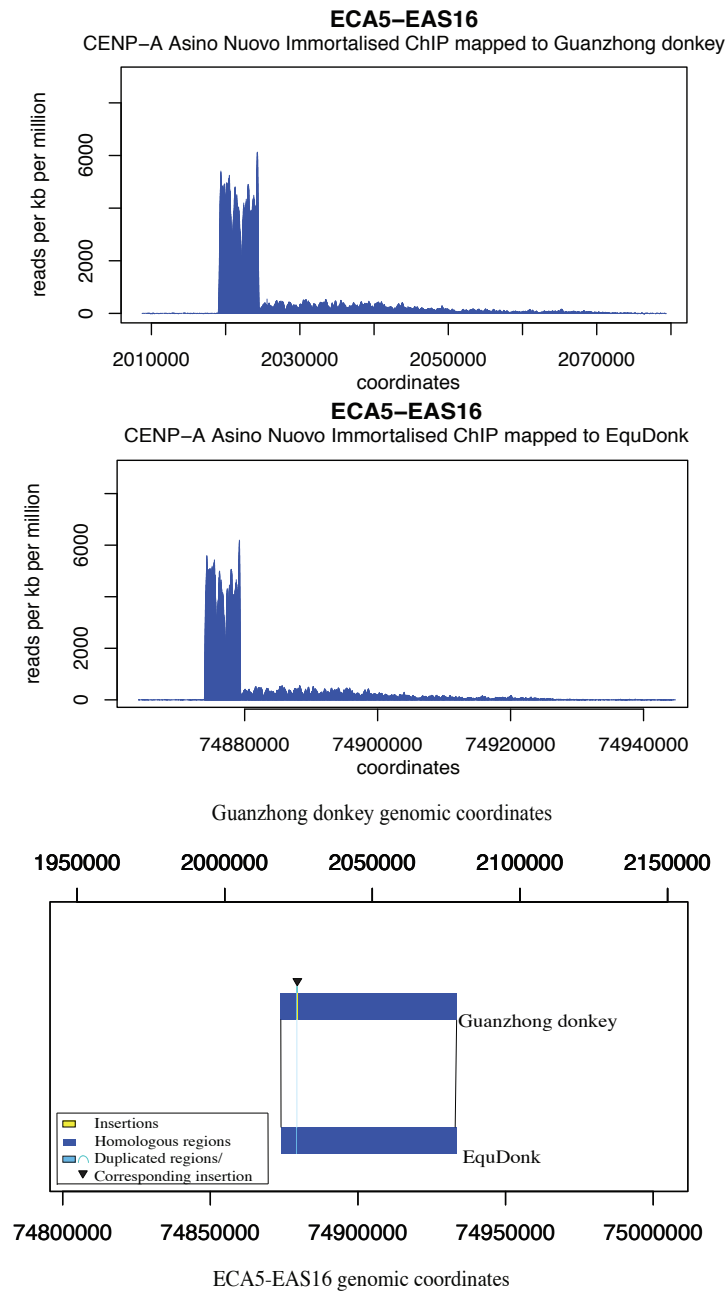
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	9	3	-	-
EquCab	12	3	-	-

**Table 4.40 Summary of sequence variation between EquDonk and EquCab on EAS14**

Reads from the EAS14 centromere domain were mapped to the horse orthologous domain chr13: 7,242,115-7,538,167nt. EquCab contained 12 sequence insertions relative the corresponding EAS14 centromere, spanning between 71bp and 16863bp. There were 3 sequences present once in EquDonk that were duplicated in EquCab. There were also sequence insertions in the Equdonk centromere spanning from 20bp to 7377bp.

Further analysis of the EquCab insertion sequences showed an abundance of LINE/L1s with 36 instances present from various subfamilies (L1M2, L1M1, L1M3c, L1MEd, L1MD2, L1ME3Cz, HAL1, L1Mc, L1ME2z and L1MA9), four examples of L2 (L2a, L2c and L2d), ten instances of SINEs (MIR), seventeen examples of LTRs (ERVL-MaLR, ERV1 and Gypsy) and two examples of DNA elements (hAT-Tip100, hAT-Ac) as well as simple repetitive elements. When combined SINEs occupied 2.44% of these domains, while LINEs occupied 38.76%, LTR elements and DNA elements occupied 9.89% and 1.64% respectively. In the EquDonk insertion domains, there was also a notable increase in LINEs with 16 instances, eleven L1 (HAL1, L1M4c, L1ME2z, L1MD2, L1M3c and L1M4) and five L2 (L2a, L2d). LTR elements were also abundant at this domain, with twelve examples of ERV1, ERVL-MaLR, ERVL. There were nine instances of SINEs (eight MIR, one tRNA) and three instances of DNA elements (hAT-Charlie, hAT-Tip100). These domains taken together show a decrease in SINEs (3.06%), LINEs (18.56%) and DNA elements (2.49%) while there was an increase in LTR elements (12.43%) compared to abundance observed across the whole genome. Repetitive elements in duplicated regions in both genomes showed the presence of LINE/L1s (L1M3c, L1M2) as well as simple repeats. The EquDonk centromere spanned 264813bp while the EquCab orthologous domain including non-homologous sequences spanned 296053bp and shared 98% sequence identity with the EquDonk EAS14 centromere.

## The EAS16 Centromere



**Figure 4.21 EAS16 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS16 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey

## Sequence features of EAS16 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	-	1	-	-
Guanzhong donkey	1	1	-	-

**Table 4.41 Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS16**

The EAS16 centromere function was mapped to the Guanzhong donkey domain gi|933836929|gb|JREZ01000191.1|: 2,018,727-2,069,363nt. There was a 59bp sequence present in a single copy in EquDonk that was duplicated in EquCab.

### *Repetitive elements across the EAS16 centromere domain in the donkey*

SINE abundance at this domain had decreased by 2.24% when compared to levels observed across the donkey genome. LINE levels were almost doubled at this domain, with L1 elements increasing by 19.88% and L2 increasing by 1.14%. LTR element abundance decreased by 1.78% while hAT-Charlie and TcMar-Tigger levels dropped by 3.35%.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	4	725	1.39	3.67
ALUs	0	0	0	0
MIRs	4	725	1.39	3.63
<b>LINEs:</b>	26	22239	42.7	21.69
LINE1	18	19177	36.82	16.09
LINE2	8	3062	5.88	4.9
L3/CR1	0	0	0	0.5
<b>LTR elements:</b>	7	2379	4.57	6.48
ERV1	4	1364	2.62	2.19
ERV1-MaLRs	3	1015	1.95	2.72
ERV_classI	0	0	0	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	2	237	0.46	3.82
hAT-Charlie	1	194	0.37	1.95
TcMar-Tigger	1	43	0.08	0.93

**Table 4.42 Summary of repetitive elements across the EAS16 centromere EquDonk compared with whole genome levels**

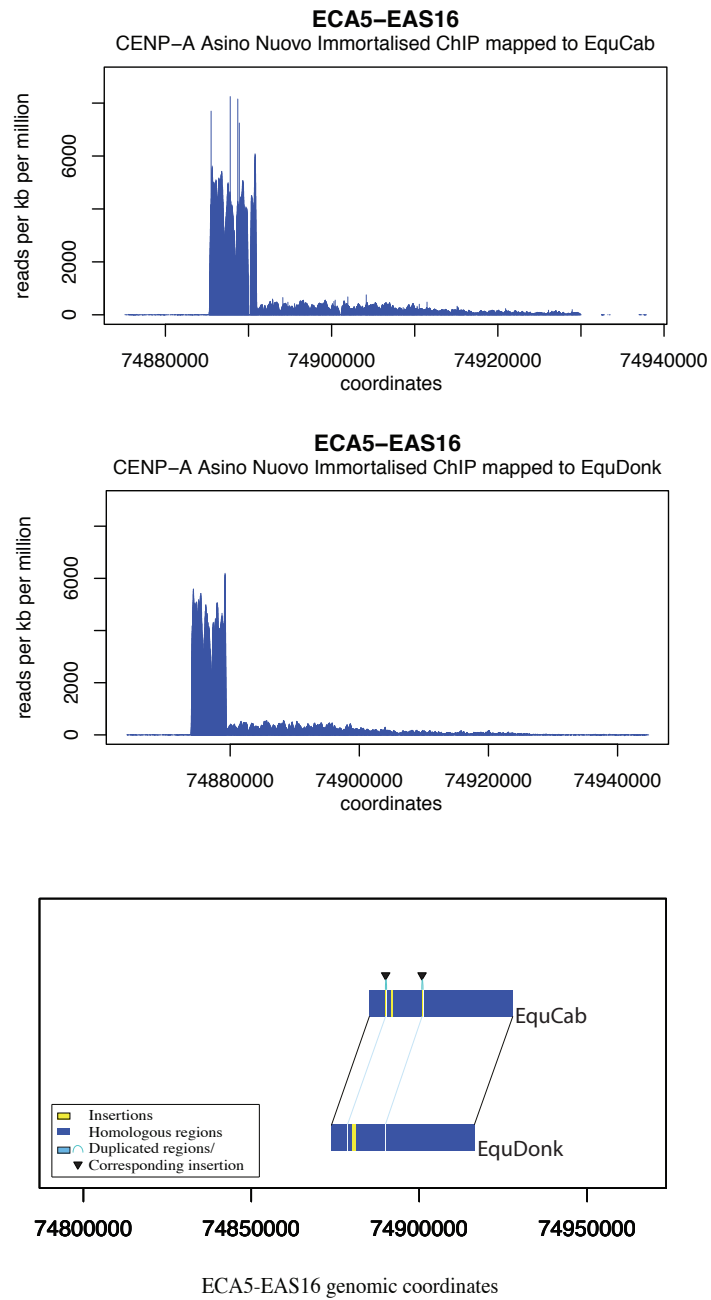
Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	4	725	1.43	3.67
ALUs	0	0	0	0
MIRs	4	725	1.43	3.63
<b>LINEs:</b>	2	21274	42.01	21.69
LINE1	4	18216	35.97	16.09
LINE2	8	3058	6.04	4.9
L3/CR1	0	0	0	0.5
<b>LTR elements:</b>	7	2379	4.7	6.48
ERV1	4	1364	2.69	2.19
ERV1-MaLRs	3	1015	2	2.72
ERV_classI	0	0	0	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	2	237	0.47	3.82
hAT-Charlie	1	194	0.38	1.95
TcMar-Tigger	1	43	0.08	0.93

**Table 4.43 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi|933836929|gb|JREZ01000191.1| (EAS16) compared with whole genome levels**

No repetitive elements were found in the inserted or duplicated sequences in either genome. The domain span of the two individuals was comparable with EquDonk spanning 50637bp and Guanzhong donkey spanning 52082bp.



## The EAS16 Centromere



**Figure 4.22 EAS16 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the orthologous horse domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS16 centromere region in EquDonk compared to the orthologous region in EquCab

### Sequence features of EAS16 centromere

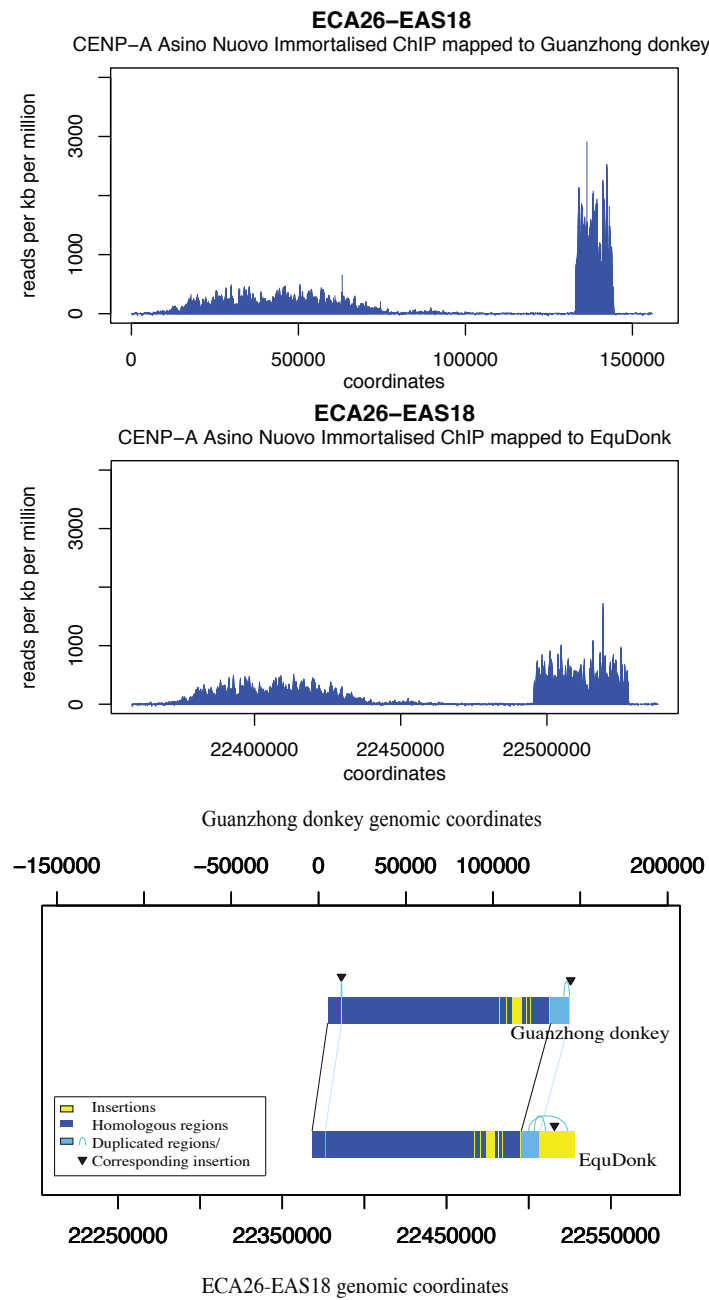
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	1	2	-	-
EquCab	5	2	-	-

**Table 4.44 Summary of sequence variation between EquDonk and EquCab at EAS 16**

The centromeric function of donkey chromosome 16 was mapped to the horse orthologous domain on chromosome 5:74,885,145-74,927,853. EquCab contained five cases of insertion spanning between 11bp and 324bp. The EquDonk EAS16 centromere contained one insertion spanning 970bp. There were two single copy sequences in EquDonk spanning 11bp and 17bp that were duplicated in EquCab.

There were small inserted sequences in both individuals and duplicated domains in EquCab. Analysis of these domains showed relatively few repetitive elements with a 367bp LTR/ERV1 (LTR16E1) in EquDonk inserted sequence, occupying 36.76% of the combined inserted and duplicated sequences and a simple (ATTT)<sub>n</sub> repeat in the EquCab inserted domain. The EAS16 horse orthologous domain (42709bp) was 9,373bp shorter than the CENPA binding domain in EquDonk (52082bp) and shared 99% sequence identity.

## The EAS18 Centromere



**Figure 4.23 EAS18 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS18 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey

## Sequence features of EAS18 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	8	5	-	-
Guanzhong Donkey	8	5	-	-

**Table 4.45 Summary of sequence variation between EquDonk and the Guanzhong donkey at EAS18**

The CENP-A binding domain of EAS18 was mapped to the Guanzhong donkey contig gi|933835538|gb|JREZ01000464.1|: 7,977-145,853nt giving a two peak profile, a gaussian peak (98,345bp) and a spike peak (11,566bp), while in EquDonk there was a broader spike peak spanning 32,876bp as a result of sequence duplication. There were eight instances of sequence insertion in both EquDonk and the Guanzhong donkey spanning between 15bp - 11,078bp and 7bp - 5378bp respectively. There were 4 single copy sequences present in the EquDonk centromere that were duplicated in the Guanzhong donkey, while the Guanzhong donkey contained a single copy sequence spanning 11078bp that was duplicated in EquDonk.

### *Repetitive elements across the EAS18 centromere domain in the donkey*

There was 2.75 times less SINEs at the Guanzhong donkey EAS18 centromere orthologous domain when compared to whole genomic levels. Overall LINE abundance increased (0.8%), with a rise in L1 (4.16%) and a decrease in L2 (3.04%). LTR element abundance also showed an increase of 3.14%, with ERVL (1.56%), ERVL-MaLRs (0.92%) and ERV classI (0.96%) levels all rising. 4.96 times less DNA elements were observed at this Guanzhong donkey domain compared to whole genome levels.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	12	1827	1.33	3.67
ALUs	0	0	0	0
MIRs	12	1827	1.33	3.63
<b>LINEs:</b>	40	31002	22.49	21.69
LINE1	28	27967	20.28	16.09
LINE2	11	2566	1.86	4.9
L3/CR1	0	0	0	0.5
<b>LTR elements:</b>	29	13261	9.62	6.48
ERVL	11	5165	3.75	2.19
ERVL-MaLRs	13	5020	3.64	2.72
ERV_classI	4	2938	2.13	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	8	1057	0.77	3.82
hAT-Charlie	5	580	0.42	1.95
TcMar-Tigger	1	134	0.1	0.93

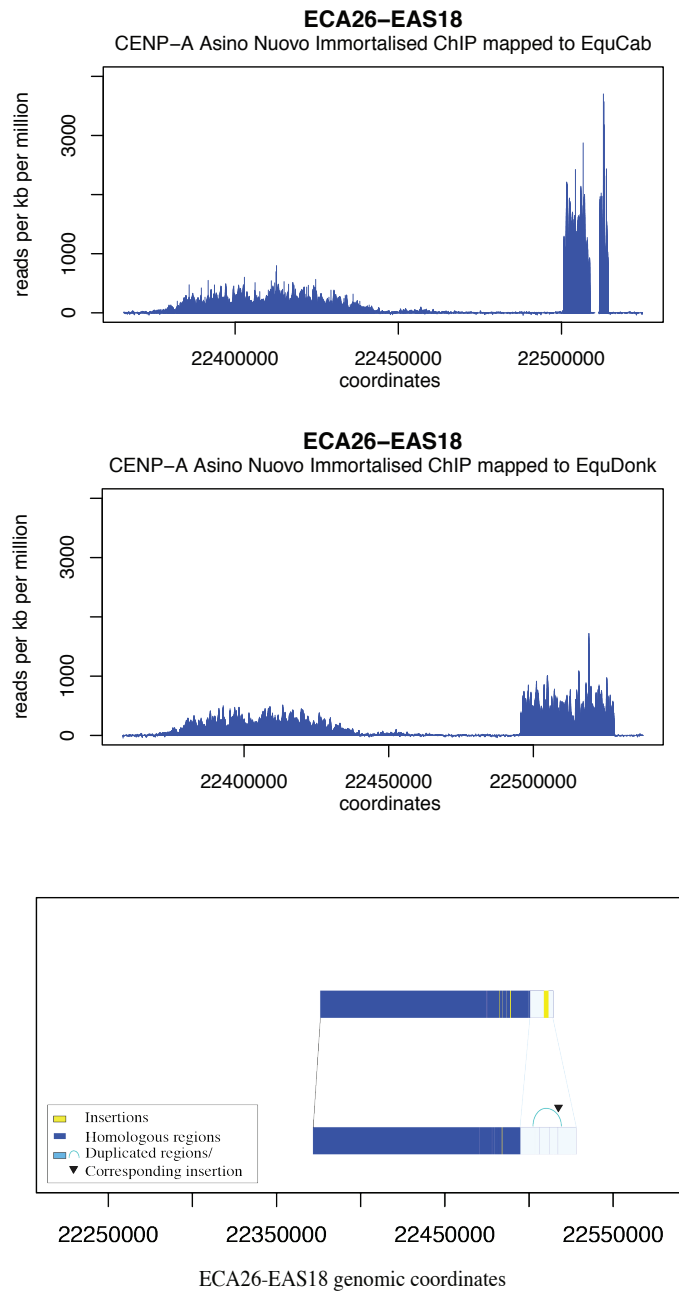
**Table 4.46 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi|933835538|gb|JREZ01000464.1| (EAS18) compared with whole genome levels**

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	12	1827	1.13	3.67
ALUs	0	0	0	0
MIRs	12	1827	1.13	3.63
<b>LINEs:</b>	58	45993	28.35	21.69
LINE1	45	42610	26.26	16.09
LINE2	12	2914	1.8	4.9
L3/CR1	0	0	0	0.5
<b>LTR elements:</b>	32	14181	8.74	6.48
ERV1	11	5125	3.16	2.19
ERV1-MaLRs	14	5076	3.13	2.72
ERV_classI	6	3842	2.37	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	11	1394	0.86	3.82
hAT-Charlie	8	917	0.57	1.95
TcMar-Tigger	1	134	0.08	0.93

**Table 4.47** Repetitive elements across the EAS18 centromere EquDonk compared with whole genome levels

Analysis of repetitive elements across regions of insertion in EquDonk show four instances of LINE/L1s (L1M3, L1M1) and one example of LTR/ERV1-MaLR (MLT1J) while in the Guanzhong donkey, there were also four instance of LINE/L1s (L1M1) and two examples of LTR/ERV1-MaLR (MLT1J, MLT1D). In the duplicated domains of the Guanzhong donkey, there were eight cases of LINE/L1 (L1M1, L1ME3A, L1MA7, L1Meg, L1MB2 subfamilies) one case of LINE/L2 (L2a) and two cases of LTR, ERV1, ERV1 (LTR40A1, LTR31). In the EquDonk, there were twenty-seven examples of LINE/L1s (L1M1, L1ME3A, L1MA7, L1Meg, L1MB2, L1M3, L1MA7), three cases of L2 (L2a), four instances of LTR elements, ERV1, ERV1 (LTR31, LTR40A1) as well as three examples of the DNA element hAT Charlie (MER33). The domains combined in the Guanzhong donkey show an abundance of LINEs (58.89%), increasing by 37.2% when compared to whole genome levels while abundance of LTR (3.88%) and DNA (0.83%) elements showed a decrease. In EquDonk LINEs occupied 59.45% of inserted and duplicated sequences, while LTR and DNA elements occupied 3.97% and 1.22% respectively. The EAS18 centromere in EquDonk spans 162256bp while considering the lack of the large sequence duplication in the spike peak the Guanzhong donkey orthologous domain spans 137877bp and shares 99% sequence identity.

## The EAS18 Centromere



**Figure 4.24 EAS18 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the orthologous horse domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS18 centromere region in EquDonk compared to the orthologous region in EquCab

## Sequence features of EAS18 centromere

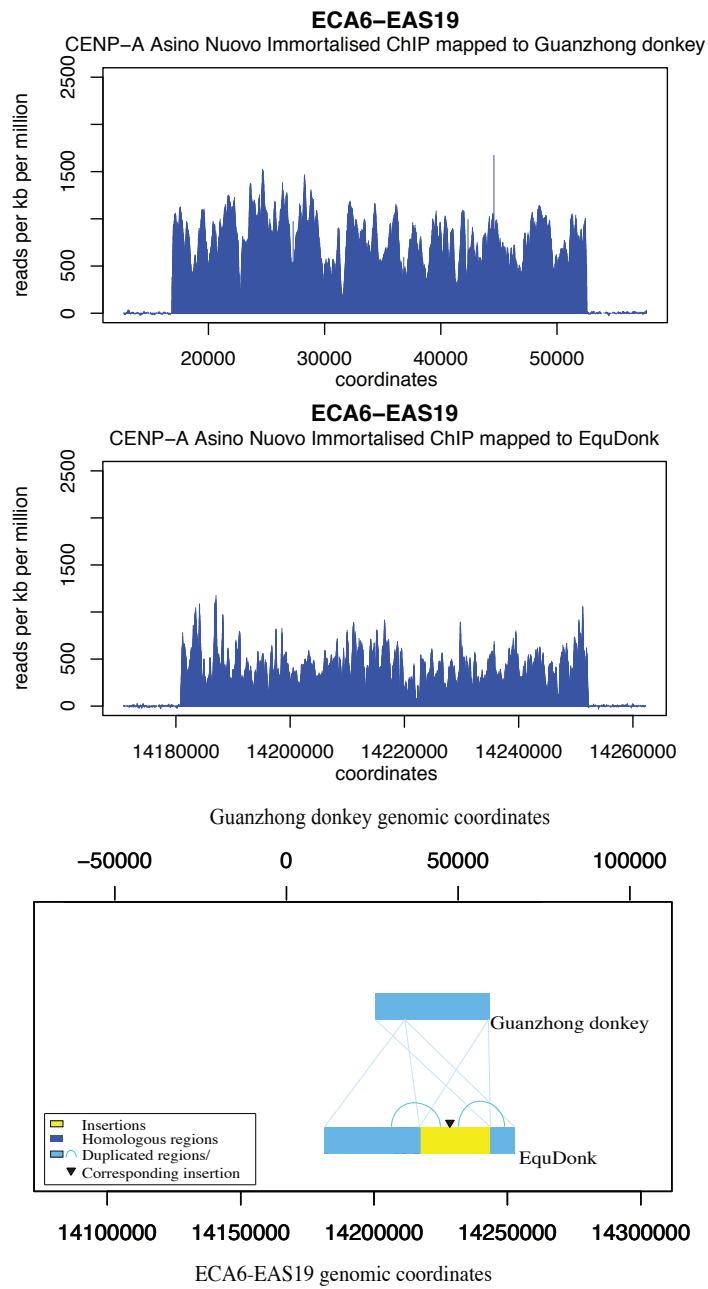
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	5	1	-	-
EquCab	6	1	-	-

**Table 4.48** Summary of sequence variation between EquDonk and EquCab on EAS18

The functional domains of the EAS18 centromere were mapped to the horse genomic region chr26: 22,375,934-22,514,838nt. EquDonk contained five insertions ranging in size from 13bp to 368bp. The spike domain in the Equdonk genome was larger than the horse domain due to sequence duplication as depicted in the illustration. EquCab also contained five other smaller insertions, ranging in size from 82bp to 466bp.

Analysis of repetitive elements within duplicated sequences in the horse domain showed enrichment in LINE/L1 with 14 cases (L1M1, L1ME3A, L1MA9, L1MEg, L1M4, L1ME2z, L1MB2 and L1M3), one instance of LINE/L2 (L2a), a LTR/ERV1 element (LTR31) and three DNA elements; two TcMar-Tigger and one hAT-Charlie. In the 2726bp EquCab inserted sequence there were five examples of LINE/L1 (L1MCc, L1M4, L1ME2z) and two cases of DNA elements/TcMar (Tigger1). Taken together these domains contain an increase in LINEs (50.05%) and a decrease in LTR (3.54%) and DNA (2.89%) elements compared to whole genome levels. In the EquDonk duplicated domain, there were twenty six cases of LINE/L1 (L1MB2, L1M3, L1M1, L1ME3A, L1MA7, L1MEg, L1MDa, L1MA6, L1M5), three examples of LINE/L2 (L2a), three of LTR/ERVL (MER33) and three of hAT Charlie. Combined there was an increase in LINEs (56.56%) at these domains and a decrease in LTR (4.11%) and DNA elements (1.31%) compared to whole genome levels. The centromere domain in Equdonk spanned 162256bp while in EquCab it spanned 138905bp and shared 99% sequence identity.

## The EAS19 Centromere



**Figure 4.25 EAS19 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS19 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey



## Sequence features of EAS19 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	1	1	-	-
Guangzhong Donkey	-	1	-	-

**Table 4.49 Summary of sequence variation between EquDonk and the Guangzhong donkey on EAS19**

The centromere function of EAS19 was mapped to the corresponding loci in the Guangzhong donkey gi|933831780|gb|JREZ01001308.1|: 16,703-52,720nt. The domain spanned 73043bp in EquDonk and 36,017bp in the Guangzhong donkey, the schematic showed duplication of sequence in EquDonk accounting for the larger domain size.

### *Repetitive elements across the EAS19 centromere domain in the donkey*

Analysis of repetitive elements across the whole EAS19 centromere orthologous Guangzhong donkey domain showed an increase in SINEs (0.28%) while LINE (13.99%) and LTR (5.26%) abundance was decreased when compared to whole donkey genomic levels. A small increase was observed in DNA elements (0.11%) with TcMar-Tigger elements completely absent at this domain.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	8	1423	3.95	3.67
ALUs	0	0	0	0
MIRs	8	1423	3.95	3.63
<b>LINEs:</b>	16	2775	7.7	21.69
LINE1	3	878	2.44	16.09
LINE2	11	1688	4.69	4.9
L3/CR1	1	129	0.36	0.5
<b>LTR elements:</b>	1	439	1.22	6.48
ERV1	1	439	1.22	2.19
ERV1-MaLRs	0	0	0	2.72
ERV_classI	0	0	0	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	1416	1416	3.93	3.82
hAT-Charlie	1416	1416	3.93	1.95
TcMar-Tigger	0	0	0	0.93

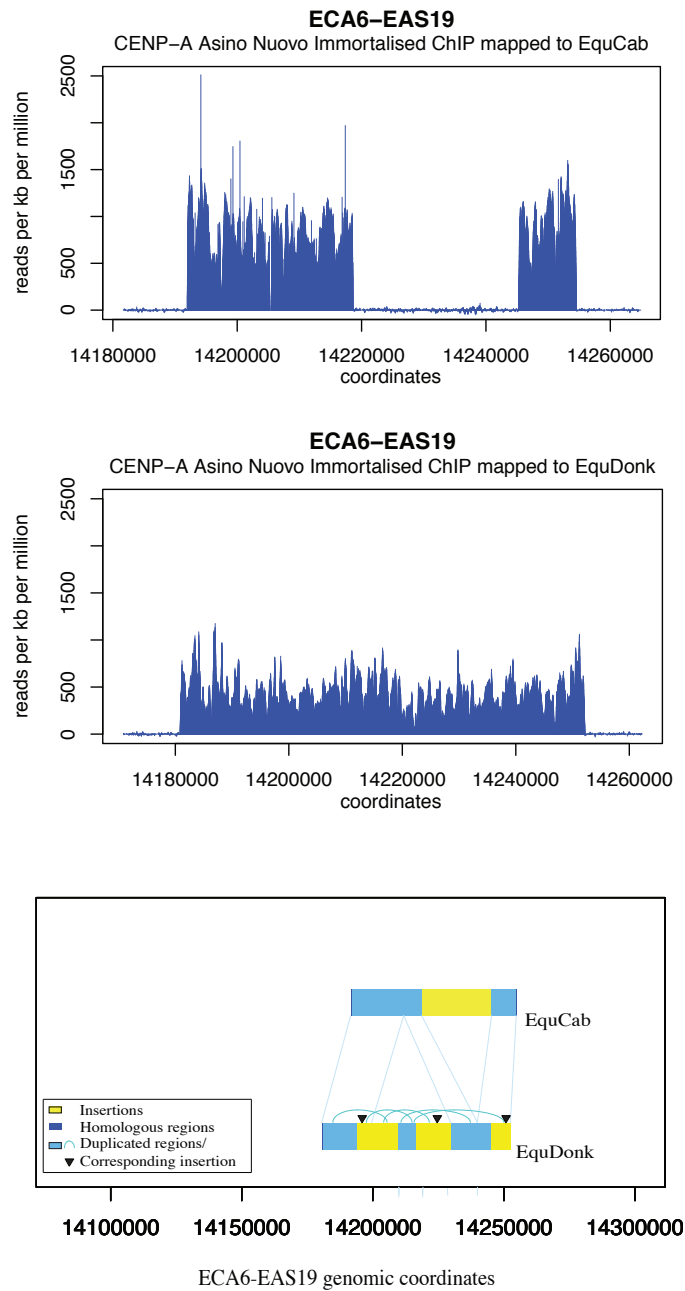
**Table 4.50 Summary of repetitive elements that span the CENP-A binding domain of Guangzhong donkey contigs gi|933831780|gb|JREZ01001308.1| (EAS19) compared with whole genome levels**

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	19	2654	3.82	3.67
ALUs	0	0	0	0
MIRs	18	2596	3.82	3.63
<b>LINEs:</b>	69	56411	8.84	21.69
LINE1	50	49370	3.63	16.09
LINE2	17	6174	4.69	4.9
L3/CR1	2	867	0.35	0.5
<b>LTR elements:</b>	41	17695	1.20	6.48
ERV_L	14	5394	1.20	2.19
ERV_L-MaLRs	14	4596	0	2.72
ERV_classI	11	7534	0	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	21	3823	3.57	3.82
hAT-Charlie	7	980	3.57	1.95
TcMar-Tigger	6	956	0	0.93

**Table 4.51** Repetitive elements across the EAS19 centromere EquDonk compared with whole genome levels

In domains of duplication in the Guanzhong donkey, ten instances of LINEs were present two L1s (L1M5, L1MC3), seven L2 (L2c, L2d2, L2a, L2) and one LINE/RTE-BovB (MamRTE1). Also present were eight copies of hAT-Charlie and eight cases of SINE/MIR (MIRb, MIR3, MIR). In the EquDonk inserted and duplicated domain there were twenty three cases of LINEs, four L1s (L1M5, L1MC3), seventeen cases of L2 (L2c, L2d2, L2a, L2), one LINE/RTE-BovB (MamRTE1) and one LINE/CR1 (L3), fourteen instances of SINE/MIR (MIR3, MIRb, MIR), eight examples of hAT-Charlie and two instances of LTR/ERV\_L (MLT2B4).

## The Eas19 Centromere



**Figure 4.26 EAS19 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the orthologous horse domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS19 centromere region in EquDonk compared to the orthologous region in EquCab

## Sequence features of EAS19 centromere

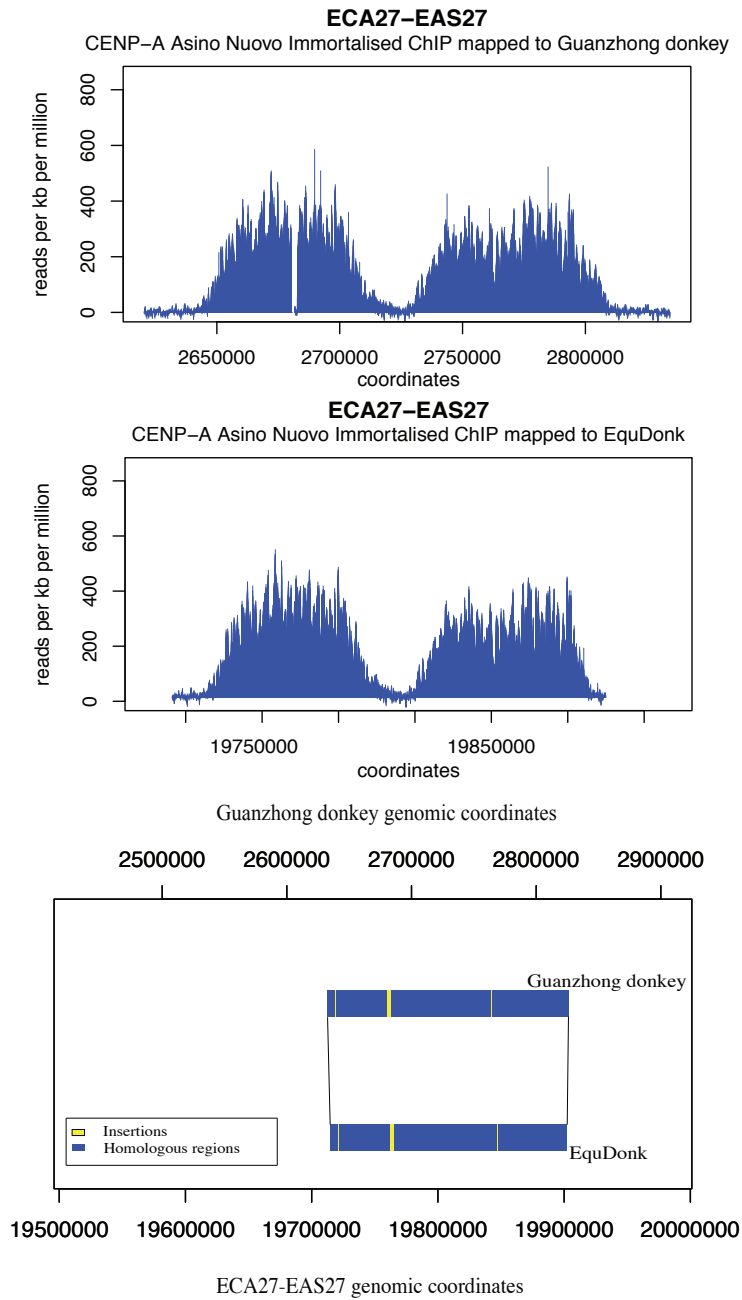
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	3	3	-	-
EquCab	1	2	-	-

**Table 4.52 Summary of sequence variation between EquDonk and EquCab on EAS19**

The EAS19 centromere was mapped to the corresponding domain in the horse genome chr6: 14,191,649-14,254,876nt, a peak profile of two narrow peaks was observed with a 26697bp insertion between 14218611- 14245308nt not present in EquDonk. The EquDonk Eas19 centromere domain is comprised of repeated segments of sequence, present in a single copy in EquCab. The CENP-A binding domain in EquDonk spans 73043bp while in EquCab including the large non homologous sequence the domain spans 63228bp and share 99% sequence identity.

Analysis of repetitive elements in the 26697bp EquCab inserted sequence showed ten instances of LINE/L1 (L1M3, L1M2, L1MA6, L1MA8, L1ME3G, HAL1ME), five copies of hAT-Tip100, seven cases of LTRs: ERVL (MLT2B4, LTR41, LTR79, LTR16A), ERVL-MaLR (MLT1I) and Gypsy (MamGypLTR2b) and two copies of SINEs (MIR, MIRb). Analysis of regions of duplication in the horse shows nine copies of SINEs/MIR (MIR3, MIRb, MIR), sixteen cases of LINEs, two L1 (L1M5, L1MC3), twelve L2 (L2d, L2d2, L2a, L2), one of LINE/RTE-BovB (MamRTE1) and one Cr1 (L3). There are also eight instances of hAT-Charlie. While the Equdonk duplicated and inserted sequences contains all the same repetitive elements displayed in EquCab but in greater abundance given the sequence over representation.

## The EAS27 Centromere



**Figure 4.27 EAS27 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS27 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey

## Sequence features of EAS27 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	3	-	-	-
Guanzhong donkey	3	-	-	-

**Table 4.53** Summary of sequence variation between EquDonk and the Guanzhong donkey on EAS27

The CENP-A binding domain of EAS27 was mapped to the Guanzhong donkey orthologous region gi|933836537|gb|JREZ01000266.1|: 2,630,403-2,824,355nt. EquDonk contained three instances of insertion spanning 9bp, 26bp and 2657bp. The Guanzhong donkey contained three insertions spanning 10bp, 94bp and 2503bp.

### *Repetitive elements across the EAS27 centromere domain in the donkey*

The abundance of SINEs had decreased by 1.6% at this loci in the Guanzhong donkey while LINE levels had increased by 10.31% with both L1 (10.05%) and L2 (0.84%) increasing. LTR levels had also increased by 0.6%, with ERVL-MaLRs (0.06%) and ERV class I (1.43%) increasing while ERVL levels were down (1.32%). DNA element abundance was reduced by 0.79% with a drop in hAT-Charlie.95%) and a rise in TcMar-Tigger (0.25%) when compared with whole genome levels.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	25	4007	2.07	3.67
ALUs	0	0	0	0
MIRs	25	4007	2.07	3.63
<b>LINEs:</b>	86	62057	32	21.69
LINE1	59	50707	26.14	16.09
LINE2	25	11129	5.74	4.9
L3/CR1	2	221	0.11	0.5
<b>LTR elements:</b>	35	13737	7.08	6.48
ERVL	7	1684	0.87	2.19
ERVL-MaLRs	13	5400	2.78	2.72
ERV_classI	10	5037	2.6	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	22	5885	3.03	3.82
hAT-Charlie	7	1944	1	1.95
TcMar-Tigger	8	2280	1.18	0.93

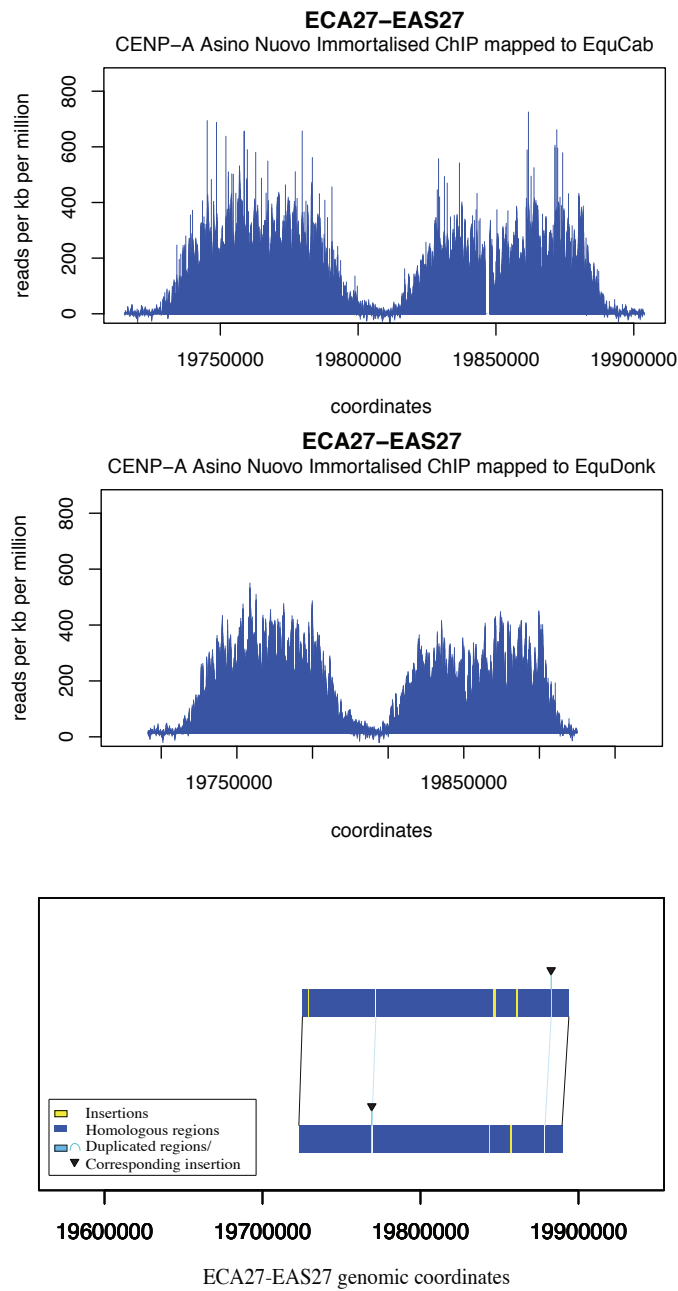
**Table 4.54** Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi|933836537|gb|JREZ01000266.1| (EAS27) compared with whole genome levels

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	24	3717	2.15	3.67
ALUs	0	0	0	0
MIRs	24	3717	2.15	3.63
<b>LINEs:</b>	74	59012	34.11	21.69
LINE1	49	48899	28.26	16.09
LINE2	23	9892	5.72	4.9
L3/CR1	2	221	0.13	0.5
<b>LTR elements:</b>	23	8000	4.62	6.48
ERV1	6	1667	0.96	2.19
ERV1-MaLRs	10	3752	2.17	2.72
ERV_classI	2	965	0.56	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	22	5346	3.09	3.82
hAT-Charlie	8	2153	1.24	1.95
TcMar-Tigger	8	1614	0.93	0.93

**Table 4.55** Repetitive elements across the EAS27 centromere EquDonk compared with whole genome levels

Repetitive elements in the 2503bp insertion in the Guanzhong donkey was examined, a 1085bp LINE/L1 (L1M3) was identified. In EquDonk, there was a single LINE/L1 (L1MB1) present in an insertion. This line occupied 98.63% of the inserted sequence. The CENP-A binding domain of EAS27 in EquDonk spanned 173002bp while in the Guanzhong donkey it spanned 193953bp and shared 99% homology.

## The EAS27 Centromere



**Figure 4.28 EAS27 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the orthologous horse domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS27 centromere region in EquDonk compared to the orthologous region in EquCab



### Sequence features of EAS27 centromere

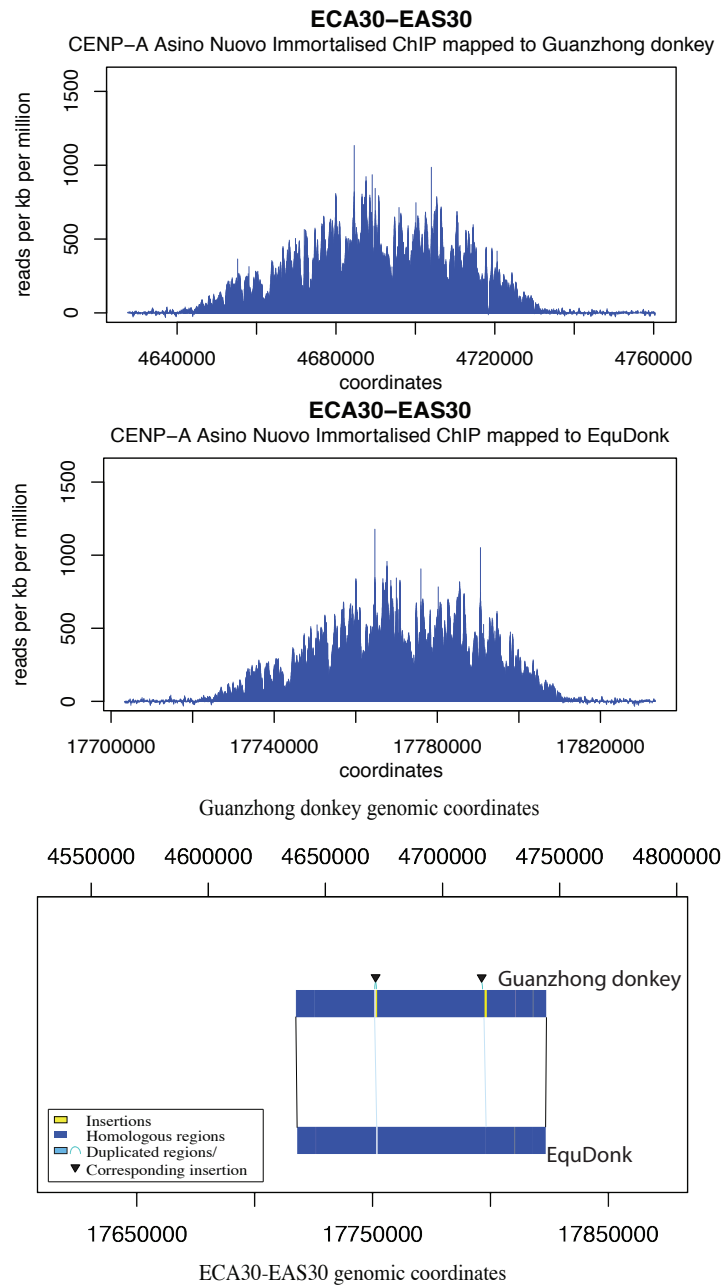
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	3	2	-	-
EquCab	5	2	-	-

**Table 4.56 Summary of sequence variation between EquDonk and EquCab on EAS27**

The centromeric function of EAS27 was mapped to the horse chr27: 19,725,276-19,893,921nt. There were instances of EquDonk sequence absent for the EquCab domain spanning 74bp, 107bp and 1000bp. There was a 14bp single copy sequence that was duplicated in the EquCab domain. EquCab contained five inserted sequences spanning between 14bp and 1576bp. There was 74bp sequence was duplicated in EquCab.

Sequence divergence was observed between both individuals with two instances of LINE/L1 (L1M3, L1MCa) in the EquDonk inserted sequences and two instances of LINE/L1 (L1MCc, L1MCa) as well as a SINE/MIR (MIRc) present in the EquCab inserted sequences. Taken together LINES occupied 71.27% of divergent sequence in EquDonk while in EquCab the LINES occupied 33.08% of sequence while SINES occupied 3.12%. The EAS27 centromere in EquDonk spanned 173002bp while in EquCab it spanned 168646bp and shared 98% sequence identity.

## The EAS30 Centromere



**Figure 4.29 EAS30 centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS30 centromere region in EquDonk compared to the orthologous region in Guanzhong donkey

## Sequence features of EAS30 centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	2	2	-	-
Guanzhong donkey	5	2	-	-

**Table 4.57 Summary of sequence variation between EquDonk and the Guanzhong donkey at EAS30**

CENP-A ChIPSeq reads from EAS30 were mapped to the Guanzhong donkey contig gi|933838627|gb|JREZ01000029.1|: 4637367-4750415nt. EquDonk contained two sequence insertions spanning 37bp and 76bp and two single copy sequences spanning 606bp and 14bp that were duplicated in the Guanzhong donkey. The Guanzhong donkey contained 5 insertions spanning between 14bp and 896bp.

### *Repetitive elements across the EAS30 centromere domain in the donkey*

Analysis of repetitive elements at the Guanzhong donkey domain corresponding to the EAS30 centromere showed a decrease in SINEs (2.79%). LINE levels had increased (2.03%), with L1 levels increased (0.6%) and L2 (0.87%) levels decreased. LTR abundance was also up rising by 2.62% with ERVL-MaLRs (2%), ERV classI (1.9%) increased while ERVL levels were down (1.03%). The abundance of DNA elements was also reduced by 0.42% compared to that observed across the genome.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	7	986	0.88	3.67
ALUs	0	0	0	0
MIRs	7	986	0.88	3.63
<b>LINEs:</b>	39	26612	23.72	21.69
LINE1	26	22090	19.69	16.09
LINE2	13	4522	4.03	4.9
L3/CR1	0	0	0	0.5
<b>LTR elements:</b>	21	10216	9.1	6.48
ERVL	3	1306	1.16	2.19
ERVL-MaLRs	10	5299	4.72	2.72
ERV_classI	7	3444	3.07	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	9	3813	3.4	3.82
hAT-Charlie	5	1796	1.6	1.95
TcMar-Tigger	2	1430	1.27	0.93

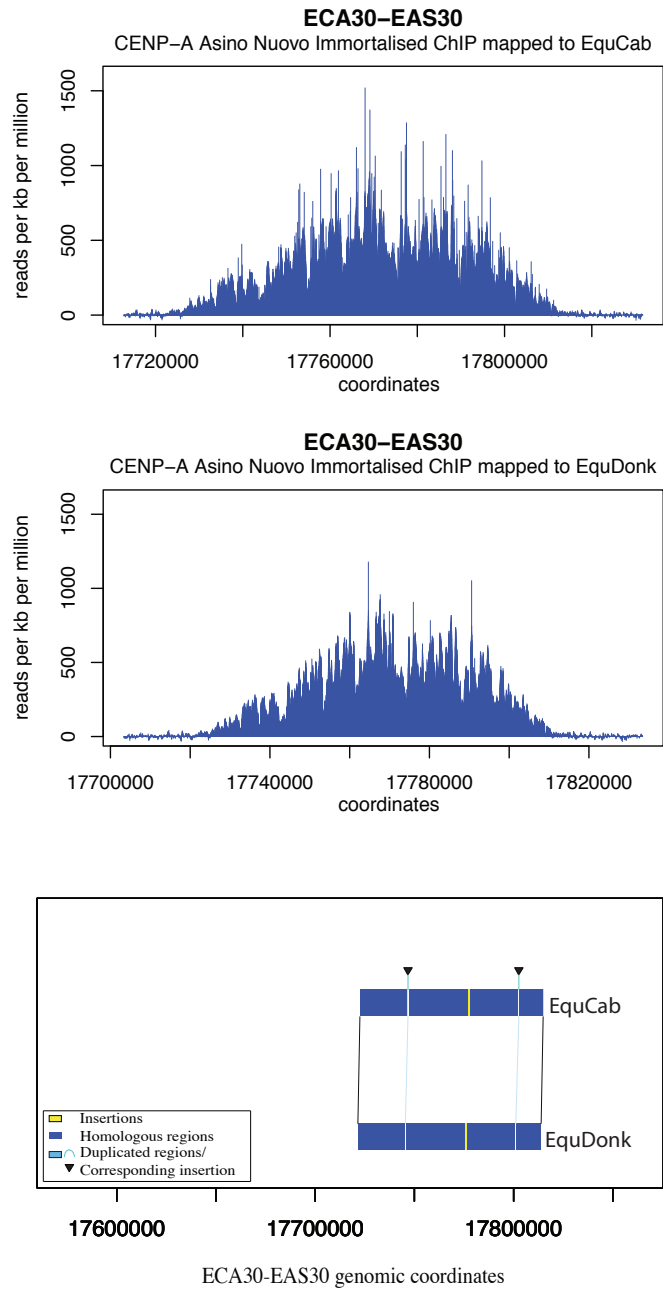
**Table 4.58 Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi|933838627|gb|JREZ01000029.1| (EAS30) compared with whole genome levels**

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	7	986	1.06	3.67
ALUs	0	0	0	0
MIRs	7	986	1.06	3.63
<b>LINEs:</b>	31	19841	21.37	21.69
LINE1	20	16029	17.26	16.09
LINE2	11	3812	4.11	4.9
L3/CR1	0	0	0	0.5
<b>LTR elements:</b>	18	9048	9.74	6.48
ERV1	3	1341	1.44	2.19
ERV1-MaLRs	10	4784	5.15	2.72
ERV_classI	4	2584	2.78	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	10	3881	4.18	3.82
hAT-Charlie	6	1662	1.79	1.95
TcMar-Tigger	2	1504	1.62	0.93

**Table 4.59** Repetitive elements across the EAS30 centromere EquDonk compared with whole genome levels

Analysis of repetitive elements within domains of insertion in Equdonk showed two copies of LINE/L1 (L1MA9) while in the Guanzhong donkey one LINE/L1 (L1M2) was present along with two LTR/ERV1 elements (MER34C). Combined, 54.71% of these sequences contained LINEs in EquDonk while in the Guanzhong donkey LINEs and LTR elements occupied 37.53% and 13.76% respectively of these sequences. The centromere domain in EAS30 Equdonk spanned 92848bp while in the Guanzhong donkey it spanned 112210bp and shared 99% sequence identity.

## The EAS30 Centromere



**Figure 4.30 EAS30 centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the orthologous horse domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EAS30 centromere region in EquDonk compared to the orthologous region in EquCab

### Sequence assembly and features of EAS30 centromere

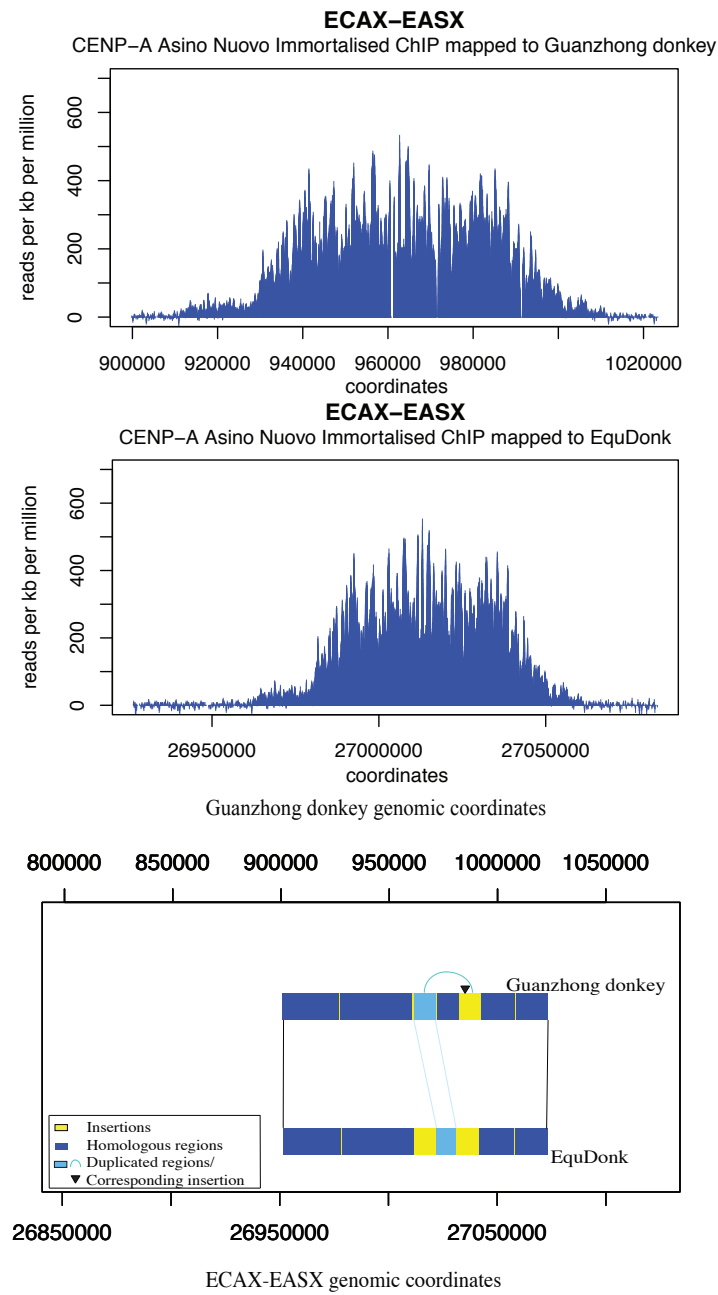
Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	1	2	-	-
EquCab	5	2	-	-

**Table 4.60 Summary of sequence variation between EquDonk and EquCab at EAS30**

The functional domain of the EAS30 centromere was mapped to the horse genome chr30:17,722,654-17,821,655nt. There were two single copy sequences both spanning 17bp in Equdonk that were duplicated in EquCab. EquDonk contained one insertion spanning 794bp. EquCab contained 5 insertions spanning between 17bp and 697bp.

Analysis of repetitive elements at Equdonk regions of insertion showed the presence of LINE/L1 (L1MCb) likewise in the Guanzhong donkey a LINE/L1 (L1MCb) was also present. Combined the abundance of LINEs at these domains in EquDonk was 61.61% while in EquCab it was 57.18%. The EquDonk Eas30 centromere spanned 92848bp while the corresponding EquCab domain spanned 99002bp and shared 98% sequence identity.

## The EASX Centromere



**Figure 4.31 EASX centromere donkey comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to Guanzhong Donkey (TOP) and EquDonk (middle). Schematic representation of sequence features of EASX centromere region in EquDonk compared to the orthologous region in Guanzhong donkey

## Sequence features of EASX centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	4	1	-	-
Guanzhong donkey	5	1	-	-

**Table 4.61** Summary of sequence variation between EquDonk and the Guanzhong donkey at EASX

The CENP-A binding domain of EAS30 was mapped to the Guanzhong donkey contig gi|933835280|gb|JREZ01000512.1|: 909,690-1,013,340nt. The schematic shows duplication of a 9916bp sequence in the Guanzhong donkey and two block of non-homologous inserted sequence in EquDonk spanning 10444bp and 10467bp.

### *Repetitive elements across the EASX centromere domain in the donkey*

Analysis of repetitive elements across the EASX centromeric orthologous region in the Guanzhong donkey showed a drop of 3.07% in overall SINE abundance when compared to whole genome levels. LINE abundance was increased by 16.05%, with L1 and L3/CR1 rising by 19.75% and 0.25% respectively, while L2 levels dropped by 3.75%. LTR elements were also increased, with overall abundance rising by 1.89%, ERVL levels more than doubled increasing by 2.59% while ERVL-MaLRs (0.04%) and ERV class I (0.26%) levels were down. DNA element abundance was decreased slightly at this domain, dropping by 0.51%.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	6	627	0.6	3.67
ALUs	0	0	0	0
MIRs	5	514	0.5	3.63
<b>LINEs:</b>	54	39115	37.74	21.69
LINE1	44	37145	35.84	16.09
LINE2	6	1193	1.15	4.9
L3/CR1	4	777	0.75	0.5
<b>LTR elements:</b>	21	8679	8.37	6.48
ERVL	8	4957	4.78	2.19
ERVL-MaLRs	9	2779	2.68	2.72
ERV_classI	4	943	0.91	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	13	3428	3.31	3.82
hAT-Charlie	8	1961	1.89	1.95
TcMar-Tigger	2	841	0.81	0.93

**Table 4.62** Summary of repetitive elements that span the CENP-A binding domain of Guanzhong donkey contigs gi|933835280|gb|JREZ01000512.1| (EASX) compared with whole genome levels

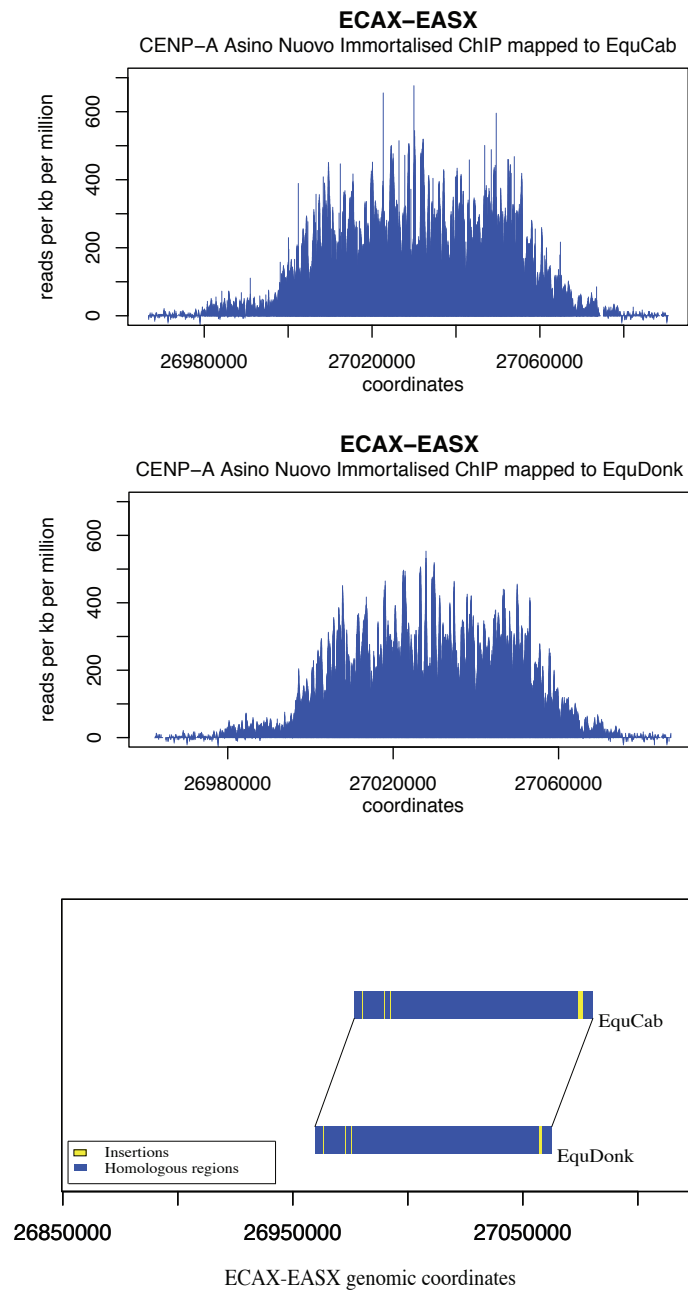


Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	7	734	0.7	3.67
ALUs	0	0	0	0
MIRs	6	621	0.6	3.63
<b>LINEs:</b>	54	39871	38.29	21.69
LINE1	44	37900	36.4	16.09
LINE2	6	1194	1.15	4.9
L3/CR1	4	777	0.75	0.5
<b>LTR elements:</b>	21	9284	8.92	6.48
ERV1	9	5893	5.66	2.19
ERV1-MaLRs	8	2448	2.35	2.72
ERV_classI	4	943	0.91	1.17
ERV_classII	0	0	0	0
<b>DNA elements:</b>	14	3507	3.37	3.82
hAT-Charlie	9	2040	1.96	1.95
TcMar-Tigger	2	841	0.81	0.93

**Table 4.63** Repetitive elements across the EASX centromere EquDonk compared with whole genome levels

Analysis of repetitive elements in the Equdonk inserted and duplicated domains shows 25 cases of LINEs, 19 of L1 (L1ME4a, L1Meg, L1MDa, L1MD, L1Mec, L1MA9) and three of L2 (L2d2), three of LINE/CR1 (L3), four copies of LTR/ERV1-MaLR (MLT1A0, MLT1K, MLT1B), five instances of DNA elements, four hAT-Charlie and one hAT-Tip100. Combined the abundance of LINEs in these regions was 34.27% well above the genomic average, LTR element abundance was 3.81%, below the genomic average while DNA elements levels were 3.58% comparable with whole genome abundance. In the Guanzhong donkey, inserted and duplicated regions, there was an abundance of LINEs, twelve L1 (L1Mec, L1MA9, L1M4c, L1Meg, L1Mda), one L2 (L2d2), four LTR/ERV1-MaLR (MLT1b, MLT1A0, MLT1K), DNA elements; four instances of hAT-Charlie and one of hAT-Tip100. Taken together the overall abundance of LINEs in these domains is 32.98%, 11.29% higher than whole genome levels, the relative abundance of LTR and DNA elements are 4.34% and 4.95% respectively In EquDonk the EASX centromere spanned 104120bp while in the Guanzhong donkey the corresponding loci spanned 103651bp and shares 99% sequence identity.

## The EASX Centromere



**Figure 4.32 EAS X centromere donkey versus horse comparison** Peak profiles of donkey CENP-A ChIPSeq reads mapped to the orthologous horse domain (TOP) and EquDonk (middle). Schematic representation of sequence features of EASX centromere region in EquDonk compared to the orthologous region in EquCab

### Sequence features of EASX centromere

Genome	Insertions	Duplications	Deletions	Inversions
EquDonk	4	-	-	-
EquCab	4	-	-	-

**Table 4.64 Summary of sequence variation between EquDonk and EquCab on EASX**

The centromere function of EASX was mapped to the horse chrX: 26,976,728-27,080,475nt.

Analysis of repetitive elements in inserted sequences in EquDonk show an enrichment in LINEs, with three copies of L1 (L1MB8, L1MC2, L1MD) and one copy of L2 (L2a), while in EquCab two instances of L2 were observed (L2a). The abundance of LINEs at inserted regions was 48.97%, 27.28% higher than whole genome levels. The overall abundance of LINEs at these domain in the horse were 25.15%, 3.56% higher than the whole genome average. Both these domains span similar sizes in EquDonk and EquCab measuring 104120bp and 103748bp respectively, as well as sharing 98% sequence identity.

Region	Size (bp)	GC%	SINEs %	LINES %	LTRs %	DNA elements%	Low complexity %	Identity with EquDonk %
Eas4 CEN	114536	35.16	1.86	29.95	5.69	5.31	0.19	
chr28: 12,897,478-13,007,113nt	109636	35.03	2.01	30.63	5.34	5.38	0.27	98
gi 933836246 gb JREZ01000325.1 :2,144,613-2,297,220 (Rv Complement)	126612	35.16	2.10	29.88	5.46	4.33	0.21	99
Eas5 CEN	227334	34.57	1.17	24.81	7.78	1.68	0.23	
chr19:4,942,106-5,160,761	218656	34.88	1.03	27.16	7.25	1.47	0.17	97
gi 933833362 gb JREZ01000925.1 :111,522-329,322	217801	34.48	1.06	25.73	6.99	1.99	0.24	99
Eas7 CEN	157777	35.38	1.94	35.64	7.7	1.81	0.08	
chr8:41,976,329-42,138,526	162198	35.36	1.97	38.92	6.57	2.32	0.18	98
gi 933835286 gb JREZ01000511.1 :1,190,428-1,471,883	152713	35.26	2.1	36.67	7.11	1.87	0.09	99
Eas8 Cen	93415	37.25	1.85	28.03	5.52	0.97	0.37	
chr20:26,386,390-26,525,316	138927	36.76	1.53	32.26	4.15	0.58	0.24	98
gi 933832210 gb JREZ01001199.1 :1,100-113,305 (Rv Complement)	112206	37.24	1.6	31.82	4.19	0.86	0	99
Eas9 Cen	73715	36.49	0.28	49.7	1.75	6.93	0	
chr14:29,651,149-29,696,370	45222	36.83	0.21	47.93	1.38	5.83	0	98
gi 933836078 gb JREZ01000366.1 :176,538-221,420	44883	36.89	0.72	42.93	2.33	5.72	0	99

Eas10 Cen	247430	34.31	2.55	28.63	6.05	1.32	0.16	
chr25:8,609,949-8,865,465	255517	34.47	2.45	28.96	5.86	1.34	0.18	98
gi 933836905 gb JREZ01000195.1 : 2,482,794-2,734,376 (Rv Complement)	251582	34.36	2.54	29.30	5.97	1.31	0.13	99
Eas11 Cen	93033	35.39	1.97	39.18	9.77	4.27	0.13	
chr17:16,772,244-16,858,830	86587	35.33	2.14	42.13	9.36	3.34	0.12	92
gi 933835325 gb JREZ01000504.1 :39 ,591-92,854 & gi 933831881 gb JREZ01001282.1 :18 ,249-75,463 (Rv Complement & Combined)	1560035	37.02	2.42	27.30	6.85	3.99	0.15	99
Eas12 Cen	308246	35.39	1.29	40.15	5.61	3.08	0.08	
chr9:31,936,755-32,287,869	351115	35.47	1.28	41.1	6.53	2.74	0.16	98
gi 933837599 gb JREZ01000107.1 :1, 395,612-1,658,211 & gi 933833617 gb JREZ01000871.1 :40 6,270-470,252 combined	5755608	36.97	2.53	24.18	6.96	3.81	0.19	99
Eas13	114752	38.12	2.1	39.52	6.86	0.66	0.14	
chr11:46,660,406-46,879,576	219171	38.99	2.28	36.73	9.99	1.08	0.05	99
gi 933838084 gb JREZ01000066.1 :6, 503,425-6,710,816	207391	38.70	2.51	39.72	8.43	1.04	0.08	99
Eas14 Cen	264813	39.27	2.26	29.60	8.11	2.86	0.15	
chr13: 7,242,115-7,538,167	296053	39.38	2.10	33.46	8.53	2.56	0.17	98
gi 933833808 gb JREZ01000828.1 : 49,410-315,452	266043	39.24	2.21	29.60	8.53	2.87	0.15	99

Eas16 Cen	52082	37.88	1.39	42.7	4.57	0.46	0.2	
chr5:74,885,145-74,927,853	42709	37.93	1.72	45.01	3.81	0.55	0	99
gi 933836929 gb JREZ01000191.1 :2,018,727-2,069,363	50637	37.87	1.43	42.01	4.7	0.47	0.2	99
Eas18	162256	34.54	1.13	28.35	8.74	0.86	0.3	
chr26:22,375,934-22,514,838	138905	34.72	1.43	23.64	9.51	0.99	0.23	98
gi 933835538 gb JREZ01000464.1 :7,977-145,853	137877	34.62	1.33	22.49	9.62	0.77	0.3	99
Eas19	73043	35.81	3.82	8.84	1.2	3.57	0	
chr6:14,191,649-14,254,876	63228	37.01	3.16	22.17	3.6	3.86	0.06	99
gi 933831780 gb JREZ01001308.1 :16,703-52,720 (rv Complement)	36,017	35.78	3.95	7.7	1.22	3.93	0	99
Eas27 Cen	173002	35.4	2.15	34.11	4.62	3.09	0.21	
chr27:19,725,276-19,893,921	168646	35.5	2.17	33.57	4.7	2.96	0.15	98
gi 933836537 gb JREZ01000266.1 :2,630,403-2,824,355	193953	35.53	2.07	32	7.08	3.03	0.2	99
Eas30 Cen	92848	35.11	1.06	21.37	9.74	4.18	0.32	
chr30:17,722,654-17,821,655	99002	35.11	1	24.1	8.91	4.16	0.39	98
gi 933838627 gb JREZ01000029.1 :4637367-4750415	112210	35.2	0.88	23.72	9.1	3.4	0.35	99
EasX Cen	104120	34.2	0.7	38.29	8.92	3.37	0.11	
chrX:26,976,728-27,080,475	103748	34.33	0.56	37.92	8.23	3.27	0.05	98
gi 933835280 gb JREZ01000512.1 :909,690-1,013,340	103651	34.12	0.6	37.74	8.37	3.31	0.11	99

**Table 4.65 Sequence analysis of EquDonk centromeres compared to the Guanzhong donkey and horse orthologous regions**

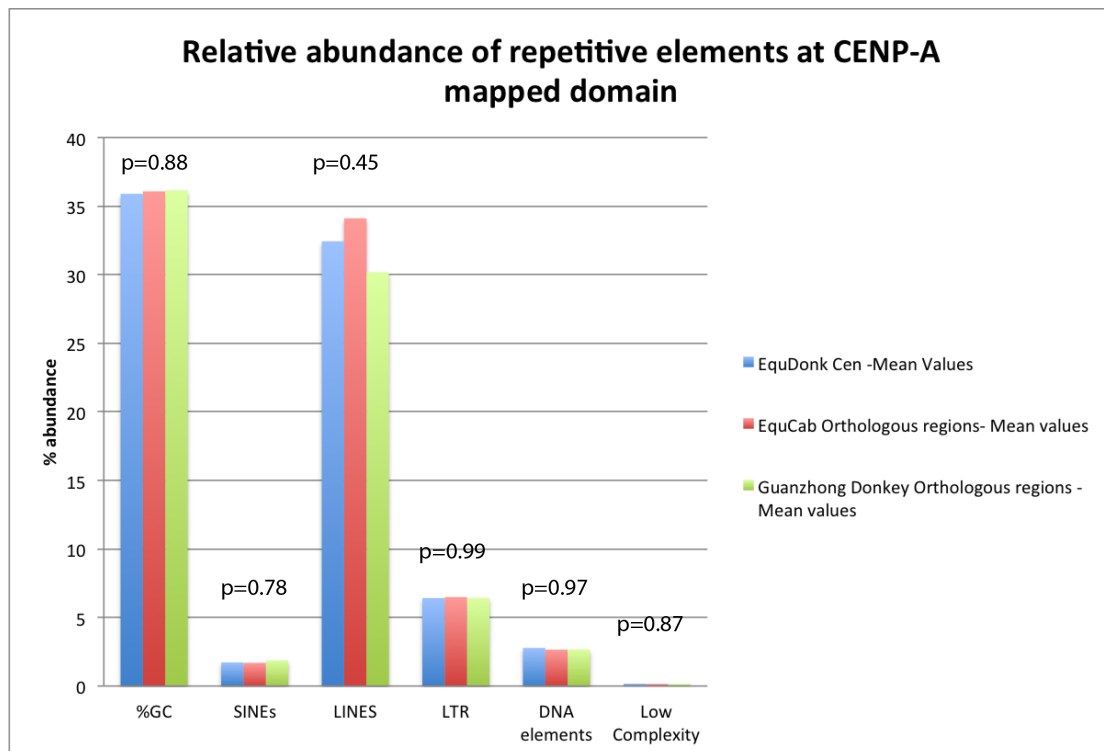
Sequence analysis was performed using the RepeatMasker software ( on each EquDonk, Guanzhong donkey and horse (EquCab2.0) centromeric region. These regions were identified by mapping donkey CENP-A ChIPSeq to the respective genomes. The enriched centromeric sequences were extracted. The above table shows the GC content and the abundance of the repetitive elements: SINEs, LINEs, LTRs, DNA transposable elements and low complexity regions as well as the respective identities to the EquDonk genome.

<b>Region</b>	<b>Size (bp)</b>	<b>GC%</b>	<b>SINEs%</b>	<b>LINES%</b>	<b>LTRs%</b>	<b>DNA elements%</b>	<b>Low complexity %</b>
EquCab	2484532062	41.50	3.53	21.59	6.29	3.67	0.19
Guanzhong donkey	2391034547	41.28	3.67	21.69	6.48	3.82	0.19

**Table 4.66 Sequence analysis of the horse and Guanzhong donkey.** Sequence analysis was performed using repeatmasker software the entire EquCab and Guanzhong donkey genomes. The above table shows the GC content and the abundance of the repetitive elements: SINEs, LINEs, LTRs, DNA transposable elements and low complexity regions

### 4.3 Domain analysis

The abundance of repetitive elements across the domains that the CENP-A ChIPSeq reads map to in EquDonk, the Guanzhong donkey and EquCab are highly similar. The abundance of LINES present at the majority of these domains is above the genome wide average, while the levels of SINEs are depleted in the majority of these domains. Analysis of the mean percentage of repetitive elements across the CENP-A mapped domains in the three individuals showed no significant difference in abundance, with p values of 0.88, 0.78, 0.45, 0.99, 0.97 and 0.87 reported for GC, SINEs, LINES, LTR elements, DNA elements and low complexity repeats respectively. This indicated that centromeres arise in these domains rather than driving significant genome rearrangement after they form.

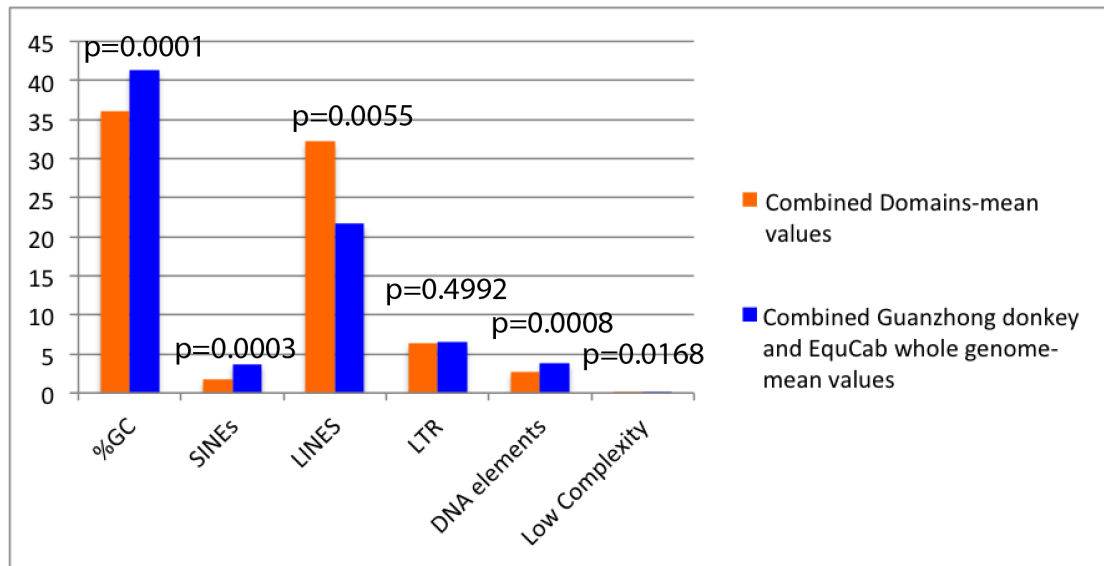


**Figure 4.33 Sequence analysis of repetitive elements across domains that donkey CENP-A reads mapped to.** The average percentage of each class of repetitive sequence at the CENP-A mapped domains in Equdonk (blue), EquCab (red) and the Guanzhong donkey (green). The bars represent the mean percentage of each repeat class. One-way anova was performed and the p values for each repeat class are shown indicating that the difference in CENP-A mapped domains are not statistically significant. ( $p > 0.05$ )

Since there was no statistical difference in the mean levels of repetitive element classes from each domain, the mean values were averaged and compared against the whole genome mean values of the Guanzhong donkey and EquCab. Student t-tests were carried out to determine the statistical differences in repeat class abundance. Figure 4.34 shows that there is significant differences in abundance of all repeat



classes with the exception of LTR elements ( $p=0.4992$ ), at the CENP-A mapped domains compared to the rest of the genome.



**Figure 4.34 Sequence analysis of CENP-A mapped domains compared to the whole genome.** The mean percentage of each class of repetitive sequence at CENP-A mapped domains in the three Equid individuals is shown in orange, while the mean percentage across the whole genome in the Guanzhong donkey and EquCab are shown in blue. Statistical analysis was carried out using the Students t test and the p values for each comparison are shown. The difference between repeat class abundance at the CENP-A mapped domains compared with the whole genomes are statistically significant for all repeat classes with the exception of the LTR elements ( $p>0.4992$ ), indicating that there is a statistically significant difference between the CENP-A mapped domains and the rest of the genome ( $p>0.05$ ).

#### *Eas4 centromere domain analysis*

There is a 6kb sequence insertion in the Guanzhong donkey absent from the EquDonk and EquCab assemblies. This sequence contained 4.8% SINES, 16.10% LINES, 5.10% LTR and 4.81% DNA elements. This sequence was unique to the Guanzhong donkey, suggesting divergence between the two donkey individuals.

#### *Eas5 centromere domain analysis*

There are two gaps in the Guanzhong donkey spanning 1396bp and 4971bp in similar positions and of similar size to gaps observed in the horse. Further analysis of sequences within these gaps reveals undetermined sequences (Ns). Analysis of these gaps in EquCab reveal no repetitive elements in the 1326bp gap while the larger 4845bp gap contains 88.17% LINES.

#### *Eas7 centromere domain analysis*

There are 3 instances of insertion unique to the horse genome that is absent from the two donkey individuals spanning between 3.5-4kb. These sequences are enriched in LINE elements, which occupy 46.08% of the non-homologous horse sequence.

#### *Eas8 centromere domain analysis*

There was sequence rearrangement in the Guanzhong donkey that was absent from the horse and EquDonk assemblies. This rearrangement is unique to the Guanzhong donkey as horse shares the same profile as EquDonk with the exception of a 34kb insertion unique to the horse. There are also a number of small sequence insertions and duplication unique to the Guanzhong donkey combined with the sequence rearrangement show divergence within this donkey individual. There is a single copy sequence present once in the horse that is duplicated in EquDonk. Combined with the large sequence insertion, there is an enrichment of LINEs at these domains in the horse (34.36%).

#### *Eas9 centromere domain analysis*

There is evidence of genomic amplification in the Equdonk EAS9 centromere domain when compared to the corresponding regions in the Guanzhong donkey and the horse assemblies. A single copy sequence, in the Guanzhong donkey is present in three copies in EquDonk. This is also the case when comparing EquDonk to the corresponding horse domain. The abundance of LINEs at this single copy domain is 45.25% significantly higher than the genomic average and the combined domains of sequence insertion and duplication in the EquDonk assembly (23.87%). This genomic amplification suggests that this centromere domain maybe accumulating “repetitive sequences” in the Asino Nuovo individual and is perhaps the beginning of centromere “maturation” whereby the unique sequence centromere gains repetitive ‘satellite’ sequences (Piras et al., 2010).

#### *Eas10 centromere domain analysis*

There are instances of small insertions unique to the horse that is absent from the two donkey individuals. These insertions show enrichment in LINEs (49.92%). With the exception of small regions of non-homologous sequences that are enriched in LINEs the donkey domains are similar to each other.

#### *Eas11 centromere domain analysis*

There is a ~40kb insertion at this domain unique to the Guanzhong donkey as well as a smaller 633bp insertion, that was absent from both the EquDonk and EquCab assemblies. The CENP-A reads from this domain mapped to two different contigs in this assembly. The contigs were merged and the large sequence insertion is either an assembly error or a bona fide divergence of the Guanzhong donkey. The abundance of LINEs, DNA elements and LTR elements were increased in these inserted sequences when compared to whole genome averages. There were small sequence insertions and duplications when comparing the EquDonk and EquCab domains but overall these regions were highly similar.

#### *Eas12 centromere domain analysis*

The peak profiles across all three individuals are similar. There were a number of instances of sequence insertion and duplication in the Guanzhong donkey relative to the EquDonk assembly that showed an abundance of LINEs. EquCab also contained a number of insertions and duplications relative to the EquDonk assembly that also showed enrichment in LINEs.

#### *Eas13 centromere domain analysis*

There is a large insertion present in both the Guanzhong donkey and the horse relative to the EquDonk EAS13 centromere. This domain spans 101kb in the Guanzhong donkey and spans a comparable 109kb in the horse. Blast alignment of the Guanzhong donkey and EquCab inserted sequences show that these domains share 99% sequence identity. This insertion sequence show enrichment in LINEs and LTR elements. The absence of this sequence from the EquDonk assembly shows divergence unique to this donkey individual while the horse and the Guanzhong donkey domains are congruous. Thus, in this case, a large deletion is associated with the Asino Nuovo individual.

#### *Eas14 centromere domain analysis*

Both donkey individuals share a similar profile with some small regions of sequence insertions and duplication with decreased LINE abundance but increased LTR elements in both donkey individuals. There were instances of sequence insertion and duplication unique to the horse in this domain, the largest of which spanned 16863bp.

Analysis of repetitive elements across these regions showed an increase in LINE and LTR elements.

#### *Eas16 centromere domain analysis*

This domain in both donkey individuals and the horse are highly similar with small instances of sequence insertion and duplication.

#### *Eas18 centromere domain analysis*

There is evidence of sequence amplification at this domain in EquDonk, absent from the Guanzhong donkey and horse assemblies. Single copy sequence enriched in LINES in the Guanzhong donkey is duplicated in Equdonk, like in the case of Eas9. This sequence amplification could potentially be the beginning of ‘satellite’ accumulation. In the horse there is a unique sequence spanning 466bp within this duplicated domain that is absent from both donkey individuals.

#### *Eas19 centromere domain analysis*

The Eas19 centromere domain has an unusual profile with sharp boundaries. In Equdonk there is evidence for genomic amplification when compared to the corresponding Guanzhong donkey and horse domains. The abundance of LINES across this domain in both donkey individuals is notably low (7.63%-8.84%) when compared to whole genome abundance. LINE abundance across the entire horse domain is comparable with whole genome levels due to the large sequence insertion. Single copy sequences in the Guanzhong donkey are present in triplicate in the corresponding EquDonk domain. While there is an insertion in the horse sequence spanning 25kb that is unique to the horse, the sequence that the donkey CENP-A reads map to in the horse are duplicated in the Equdonk assembly.

#### *Eas27 centromere domain analysis*

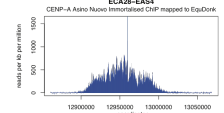
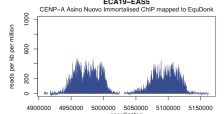
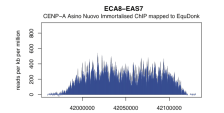
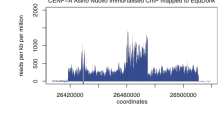
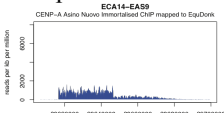
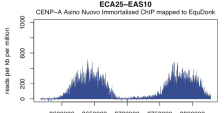
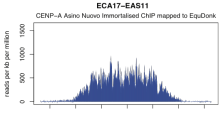
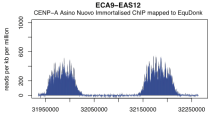
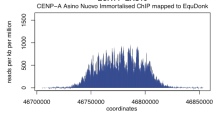
The Eas27 ChIPSeq read show a similar distribution across the two donkey individuals and the horse domain with regions of small sequence insertion, which in the all three individuals showed LINE enrichment.

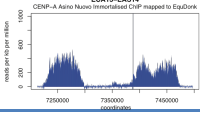
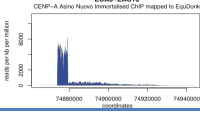
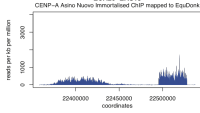




#### *Eas30 centromere domain analysis*

This domain is similar across the Equdonk, Guanzhong donkey and the horse with small instances of sequence insertion and duplication. In both donkey individuals and horse there is enrichment in LINES in regions of insertion and in the case of the Guanzhong donkey enrichment in LTR elements also.

#### *EasX centromere domain analysis*

Both these domain in the horse and in Equdonk occupy a similar profile, with instances of small sequence insertion. There is sequence absence and duplication unique to the Guanzhong donkey. There are two sequences spanning 10444bp and 10467bp present in Equdonk absent from the Guanzhong assembly while there is a 9916bp sequence present in single copy in EquDonk that is duplicated in the Guanzhong donkey. Given the absence of these sequence changes in both the horse and Equdonk individual, this sequence change is unique to the Guanzhong donkey.

Centromere	Profile (EquDonk)	Guanzhong insertions	EquDonk insertions relative to the Guanzhong donkey	EquCab insertions	EquDonk insertions relative to EquCab
<b>Eas4</b>	<b>Gaussian</b> ECA29-EAS4 CENP-A Asino Nuovo Immunolabelled ChIP mapped to EquDonk 	~6kb (16.1% LINEs)	-	-	2.3kb (15.30% LINEs)
<b>Eas5</b>	<b>Multi-domain</b> ECA19-EAS5 CENP-A Asino Nuovo Immunolabelled ChIP mapped to EquDonk 	~1.4kb (Ns) ~5kb (Ns)	-	~1.4kb ~5kb ~7kb (88.17% LINEs)	-
<b>Eas7</b>	<b>Gaussian</b> ECA8-EAS7 CENP-A Asino Nuovo Immunolabelled ChIP mapped to EquDonk 	1.9kb (42.86% LINEs)	~2kb (39.83% LINEs)	~3.5kb ~3.7kb ~4kb ~8kb (46.51% LINEs)	1.2kb (12.20% LINEs)
<b>Eas8</b>	<b>Complex</b> ECA20-EAS8 CENP-A Asino Nuovo Immunolabelled ChIP mapped to EquDonk 	2kb 3.2kb (32.15% LINEs)	3.2kb (13.38% LINEs)	34kb (34.36% LINEs)	6.2kb (13.38% LINEs)
<b>Eas9</b>	<b>Complex</b> ECA14-EAS9 CENP-A Asino Nuovo Immunolabelled ChIP mapped to EquDonk 	-	Sequence duplication	-	Sequence duplication
<b>Eas10</b>	<b>Multi-domain</b> ECA25-EAS10 CENP-A Asino Nuovo Immunolabelled ChIP mapped to EquDonk 	2.2kb (70.76% LINEs)	1.3kb (71.42% LINEs)	6.9kb 4kb 2kb (49.92% LINEs)	3.1kb (22.17% LINEs)
<b>Eas11</b>	<b>Gaussian</b> ECA11-EAS11 CENP-A Asino Nuovo Immunolabelled ChIP mapped to EquDonk 	~40kb ~3.2kb (43.19% LINEs)	3.3kb (96.22% LINEs)	-	-
<b>Eas12</b>	<b>Multi-domain</b> ECA8-EAS12 CENP-A Asino Nuovo Immunolabelled ChIP mapped to EquDonk 	~3kb (32.8% LINEs)	-	~2.5kb ~5kb (40.76% LINEs)	-
<b>Eas13</b>	<b>Gaussian</b> ECA11-EAS13 CENP-A Asino Nuovo Immunolabelled ChIP mapped to EquDonk 	~101kb (99% sequence identity to horse insertion) ~2.5kb	~2.3kb (19.61% LINEs)	~109kb (99% sequence identity to Guanzhong donkey	-

		~1kb (37.96% LINES)		insertion (36.62% LINES)	
<b>Eas14</b>	<b>Multi-domain</b> 	~2kb (13.52% LINES)	~2kb (20.99% LINES)	17kb (38.76% LINES)	-
<b>Eas16</b>	<b>Complex</b> 	-	-	-	-
<b>Eas18</b>	<b>Complex</b> 	~5.4kb (58.89% LINES)	~11kb Sequence Duplication ~5kb (59.45% LINES)	-	Sequence duplication
<b>Eas19</b>	<b>Complex</b> 	-	Sequence duplication	~27kb	Sequence duplication
<b>Eas27</b>	<b>Multi-domain</b> 	~2.5kb (43.34% LINES)	~2.6kb (98.63% LINES)	~1.6kb (33.08% LINES)	1kb (71.27% LINES)
<b>Eas30</b>	<b>Gaussian</b> 	-	-	-	-
<b>EasX</b>	<b>Gaussian</b> 	~9.9kb	~10kb ~10kb	-	-

**Table 4.67 Summary of differences in the CENP-A mapped domains in EquDonk, the Guanzhong donkey and EquCab.** Only insertions greater than 1kb and that fall within the CENP-A mapped domains (peaks) are shown, while the abundance of LINES documented is including all insertions in the domain (ie. >1kb, <1kb).

#### 4.4 Discussion

Centromeres are generally associated with AT-rich sequences (Eichler, 1999) and this is the case in donkey individuals where the mean GC content across the centromere domains is ~5.4% less than that observed across the entire donkey genome. SINEs are generally associated with GC rich euchromatic genomic region and LINEs with AT-rich heterochromatic genomic regions (Schmitz, 2012). Given the similar abundance and stability of these transposable elements in the EquDonk, the predicted Guanzhong donkey centromere domains as well as the EquCab domains, where it is known there is no centromere function, this suggests it is the underlying AT rich DNA sequence that is facilitating the accommodation of transposable elements and their presence has nothing to do with the centromere per se. Given that there is no full length LINEs in either species' centromere associated sequences it is unlikely that active transposition has driven these sequence changes. An interesting approach to examine the centromere seeding capabilities of these domains in EquCab would involve CRISPR excision of the satellite containing centromere and subsequent studies of new centromere seeding.

There are instances of sequence divergence between the two donkey individuals. In the Guanzhong donkey, there are notable sequence insertion and duplication unique to this donkey individual at the Eas4, Eas11, Eas12 and EasX, suggesting divergence from the European donkey *Asino Nuovo*. There is also an enrichment of LINEs at these regions of insertion with the exception of Eas4. There is an instance of domain rearrangement unique to the Guanzhong donkey at Eas8 not observed in the horse or Equdonk assemblies. There appears to be an absence of genomic amplification in this donkey individual when compared to the corresponding domain in Equdonk. Significant sequence duplication is observed in the Equdonk centromere domains on Eas9 and Eas19 absent from the Guanzhong donkey and horse assemblies. Equdonk shows evidence of divergence at the Eas13 centromere domain. There is a large sequence deletion relative to the horse and Guanzhong donkey assemblies spanning 109kb and 101kb respectively with increased LINE abundance when compared to whole genome levels. Comparison of these domains from the horse and Guanzhong donkey shows high homology (~99%) and the fact that these domains are congruent indicates that this is a bona fide deletion in the EquDonk individual. The presence of sequence amplification in the Equdonk centromere domains on Eas9 and Eas19 relative to the horse and Guanzhong donkey assemblies is suggestive of centromere



maturation as proposed by Piras et al., 2010. In this model, the centromere moves from a satellite-containing region to a satellite free domain where overtime it becomes 'mature' upon accumulation of satellite DNA. Comparison of the horse to both donkey individuals shows instances of substantial sequence insertion (>1kb) unique to the horse at the centromere domain in Eas5, Eas7, Eas14 and Eas19 indicating divergence between the two species. What is not clear is whether such sequences are insertions in the horse genome or deletions in the donkey genome. Comparison to other equid species would be required to resolve this.

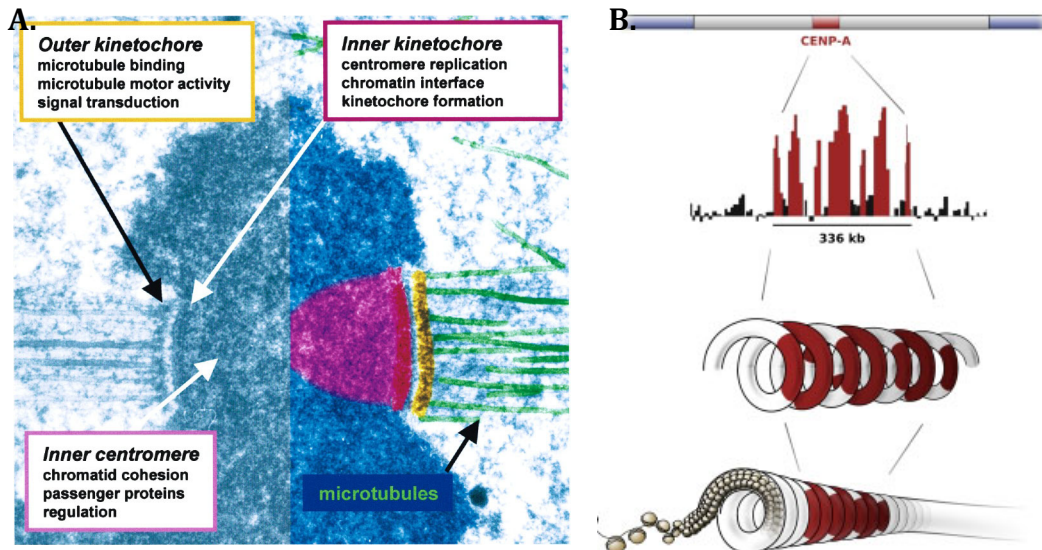
It is not clear whether the regions in which the CENP-A ChIPSeq reads map to in the Guanzhong donkey are bona fide centromere domains in this individual due to centromere sliding. To further validate this, it would be necessary to carry out ChIPSeq using fibroblasts from this donkey individual and compare the proposed CENP-A binding domains to the actual centromeres. The sequence divergence between the two species starts to address evolutionary questions associated with the equids. The remarkably fast divergence of this genus from a common ancestor (Orlando et al., 2013) and the presence of a suite of naturally occurring unique sequence centromeres make this a very versatile model organism for addressing questions about the fundamentals of centromere biology and evolution.

## Chapter 5 - Toward the identification of the Inner Centromere

### 5.1 Introduction

The location of proteins associated with the inner centromere at the linear one-dimensional primary structure DNA level has remained a puzzle mainly due to the repetitive nature of these domains, rendering ChIPSeq methods of mapping impossible. The localization of inner centromere proteins, like in the case of the centromere, is epigenetically defined, independent of DNA sequence (Bassett et al., 2010). Recruitment of the Chromosomal Passenger complex (CPC) to inner centromere is dependent on the phosphorylation of two histone tails, H3Thr3 by Haspin which binds the BIR domain of Survivin (Kelly et al., 2010) and H2AThr120 by Bub1 kinase (Kitagawa & Lee, 2015; Yamagishi et al., 2010). H2AThr120 phosphorylation is also required for shugoshin, Sgo1 and Sgo2, recruitment to the inner centromere which interacts with the CPC subunit Borealin (Carmena et al., 2012) as well as protecting centromeric cohesin from degradation (Watanabe & Kitajima, 2005). The chromosomal passenger complex traverses throughout the cell cycle in a phosphorylation dependent manner, until recruitment to the inner centromere at prometaphase where it regulates mitotic events including correction of chromosome-microtubule attachment errors (Carmena et al., 2012). Sister chromatids are held together by cohesin, which upon phosphorylation dissociates from chromosome arms and becomes concentrated at the inner centromere during mitosis (Dai et al., 2006).

Figure 5.1 shows the volume of DNA within the inner centromere (A) and the possible chromatin-folding configuration at the centromere (B). Given the large volume of DNA not associated with CENPA containing chromatin shown in Figure 5.1, the question of where the inner centromere is with respect to the kinetochore at a one-dimensional resolution remains elusive.



**Figure 5.1** A) *Centromere Organisation*: Electron microscopy shows the symmetric bipolar organization of a mitotic chromosome with the spindle attached. The condensed chromosome has been sectioned along the spindle axis plane. The compartments are shown in colour on the right chromatid: the inner centromere (violet), the inner kinetochore (red) and the outer kinetochore, site of microtubule attachment (yellow) (Cleveland, Mao, & Sullivan, 2003). B) *Centromeric higher order chromatin structure*: The centromere of the mardel (10) human neocentromere is depicted in red, ChIPSeq analysis showed that the CENP-A binding domain had a periodic distribution and occupied less than one tenth of the constricted DNA. Also shown is the possible CENP-A distribution within the supercoiled DNA (Marshall, Marshall, & Choo, 2008).

The availability of 16 unique sequence centromeres in the donkey model system provides a unique platform for gaining insight into the primary configuration of the inner centromere. Given the possible chromatin folding configuration illustrated in Figure 5.1 B, we theorize that the inner centromere will flank the CENP-A binding domain, showing domains of enrichment, either up or downstream of the core centromere. In order to dissect and map the inner centromere a suite of antibodies against subunits of the chromosomal passenger complex and cohesin must be identified, their utility within the equid system examined and use in ChIPSeq established. As well as the identification of useable antibodies, a further crux in mapping the inner centromere is the need for a mitotic population, where the inner centromere associated proteins, the chromosomal passenger complex and cohesin, are associated exclusively at this domain. A mitotic population of cells ensures minimization of background and clear inner centromere resolution. A third and final vital consideration is crosslinking methods. While cohesin is directly associated with chromatin, the chromosomal passenger complex subunits are associated with the inner centromere through binding of modified histones by survivin (Kitagawa & Lee, 2015). Survivin is tightly associated with INCENP and Borealin in a three-helix bundle as discussed in Section 1.4.1. For this reason, the length of the crosslinker arm

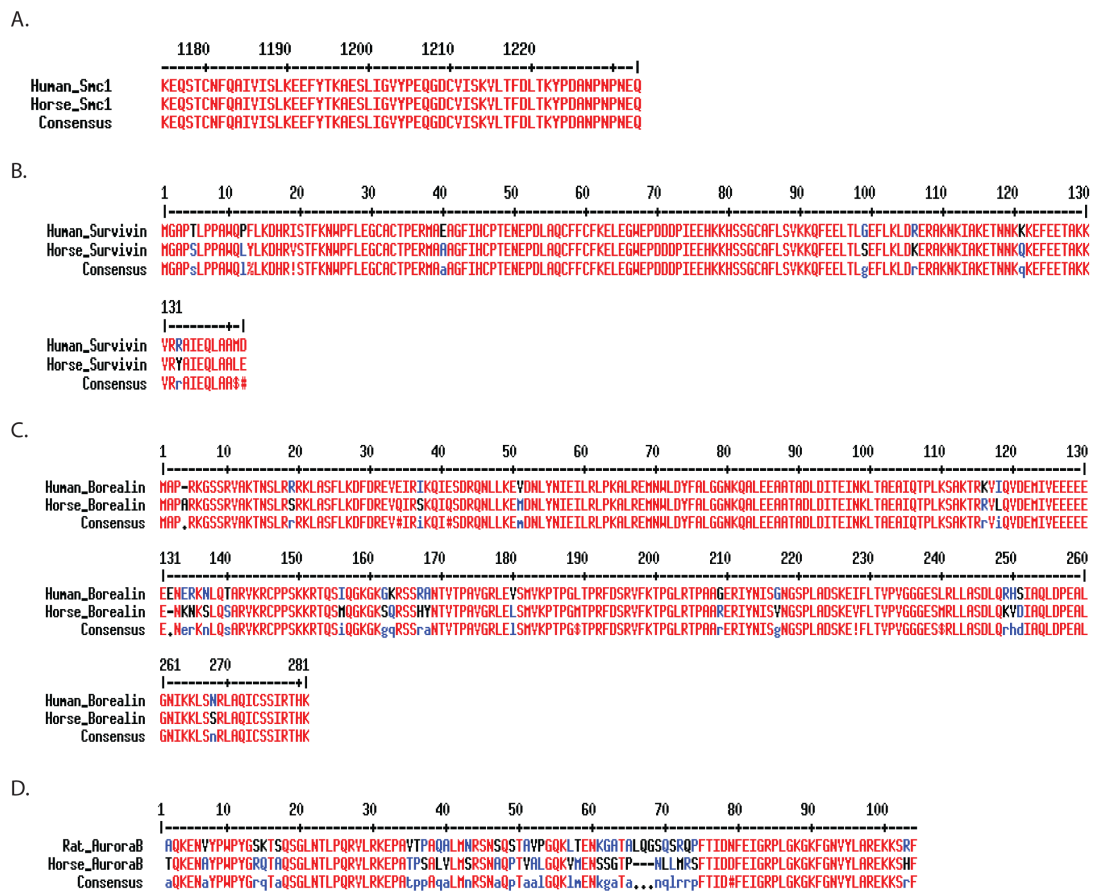
must be taken into consideration to ensure adequate fixing of the complexes to nearby DNA (Zeng, Vakoc, Chen, Blobel, & Berger, 2006).

## **5.2 Antibody identification**

In order to obtain the most suitable antibodies for use in identification of the inner centromere, a literature search was carried out and antibodies against cohesin and CPC subunits were identified. A number of suitable candidates were identified, described in this thesis are the antibodies that recognized the donkey target proteins and immunoprecipitated the target protein as verified by western blot.

For cohesin, a rabbit polyclonal antibody against Smc1 was selected (Bethyl Laboratories Inc A300-055A). The antibody was raised against human Smc1A Isoform 1 (NCBI reference sequence NP\_006297.2), recognizing the region between residue 1175 and the C-terminus. This sequence shared 100% identity with horse Smc1 as illustrated in Figure 5.2 A. This Smc1 antibodies utility in ChIP (Banerjee, Kim, & Kim, 2014) and Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) (Downen et al., 2014) has also been characterized making it a promising candidate for mapping the cohesin complex to the inner centromere. For the chromosomal passenger complex, antibodies against Survivin, Borealin and Aurora B were selected. A rabbit polyclonal antibody raised against recombinant full length human survivin [UniProt# O15392] was selected (Novus Biologicals NB500-201). Human survivin shares 92% (131/142) sequence identity to its horse homolog, Figure 5.2 B. This antibody was also selected based on its utility in immunoprecipitation (Fortugno et al., 2002). Survivin is a strong candidate for ChIPSeq given its direct association with Histone H3 through a histone phosphorylation at threonine 3. For Borealin, a mouse monoclonal antibody raised against recombinant full length human protein was selected (MBL International Incorporation # M147-3). Horse Borealin shares 90% sequence identity (253/281) to its human homolog, shown in Figure 5.2 C. This antibody's utility in immunoprecipitation has also been characterized as detailed by the manufacturer. A mouse monoclonal Aurora B antibody, raised against rat AIM1 from residue 2-124, was selected (BD Transduction Laboratories™, cat no. 611082). Rat aurora B shares 74% (90/122) sequence identity with its horse homolog (Figure 5.2 D). This antibody was also selected on the basis of its utility in immunoprecipitation (Chen et al., 2003).

A summary of the antibodies selected for use in mapping the inner centromere compartment is shown in Table 5.1.



**Figure 5.2 Alignment of protein sequences antibodies were raised against to the horse homolog.** A) Smc1 alignment of horse to human Smc1 between residue 1175 and the C-terminus shows 100% homology. B) Protein alignment of full length human to horse survivin, the sequences share 92% identity with 131 of 142 amino acids matching. C) Alignment of Human Borealin protein sequence to Horse Borealin shows that 253 out of 281 amino acids are identical. D) Alignment of Rat aurora B protein sequence from residues 2-124aa to horse aurora B. Sequences share 74% homology with 90/122 residues matching.

Antibody	Host	Clonality	Sequence Identity (%)
Aurora B	Mouse	Monoclonal	74
Borealin	Mouse	Monoclonal	90
Survivin	Rabbit	Polyclonal	92
Smc1	Rabbit	Polyclonal	100

**Table 5.1 Summary details of antibodies used in ChIPSeq experiments**

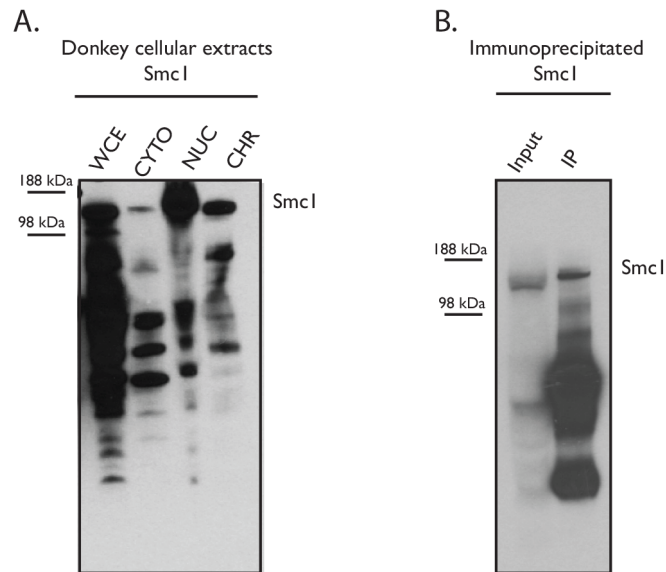
### 5.3 Antibody characterization

In order to characterize the application of these antibodies in the donkey both western blot and immunofluorescence analyses were carried out. Donkey cell extracts were prepared as described in Section 3.4 and were used in western blot analysis to both examine antibody cross reactivity with donkey and to evaluate their specificity. The antibodies were further characterized by immunofluorescence analysis on metaphase

spreads. Spreads were prepared from an asynchronous population as described in Section 3.4. Spreads were fixed using either 100% methanol or 4% formaldehyde. To determine the inner centromere location, spreads were costained with the CENP-A antibody generated in Chapter 3 in the case of methanol fixation and with CREST sera when fixed with formaldehyde. Once cross reactivity was verified, the antibodies utility in immunoprecipitation was examined. Donkey fibroblasts were crosslinked using both EGS (ethylene glycol bis (succinimidyl succinate)) and formaldehyde. EGS has a 16.1 Å spacer arm while formaldehyde has a spacer arm of 2 Å. Studies have suggested that combining the two agents is useful for crosslinking proteins that indirectly bind DNA (Zeng et al., 2006). Crosslinked cells were sheared by sonication and the chromatin isolated by centrifugation. Chromatin was precleared using protein G beads, the antibody was added and incubated with the chromatin overnight to ensure adequate binding. The antibody was recovered by addition of protein G beads. The immunoprecipitation was examined by western blot, which showed the antibody light (23kDa), heavy chain (50kDa) and if the immunoprecipitation was successful a band corresponding in size to the protein the antibody was raised against.

### ***5.3.1 Cohesin-Smc1***

To characterize the Smc1 antibody, western blot analysis was carried out at a 1:1000 dilution of the 1mg/ml antibody. The antibody recognizes a donkey protein of the expected molecular weight by western blot analysis, Figure 5.3 A, with a 170kDa band present in all four cellular extracts. No immunofluorescence signal could be detected in metaphase spreads, using either methanol or formaldehyde fixation. This suggests the accessibility of the protein in these preparations was compromised or that in a metaphase spread, where all the protein has been cleaved with the exception of at the centromere, there is not enough protein for the antibody to bind and yield a detectable signal. Given the clear cross reactivity observed by western blot analysis, antibody application in immunoprecipitation (IP) was tested using 1ug of antibody per  $1 \times 10^6$  chromatin cell equivalents. Figure 5.3 B shows a band corresponding in size to Smc1 recovered by IP, indicating that the antibody is suitable for use in ChIPSeq.

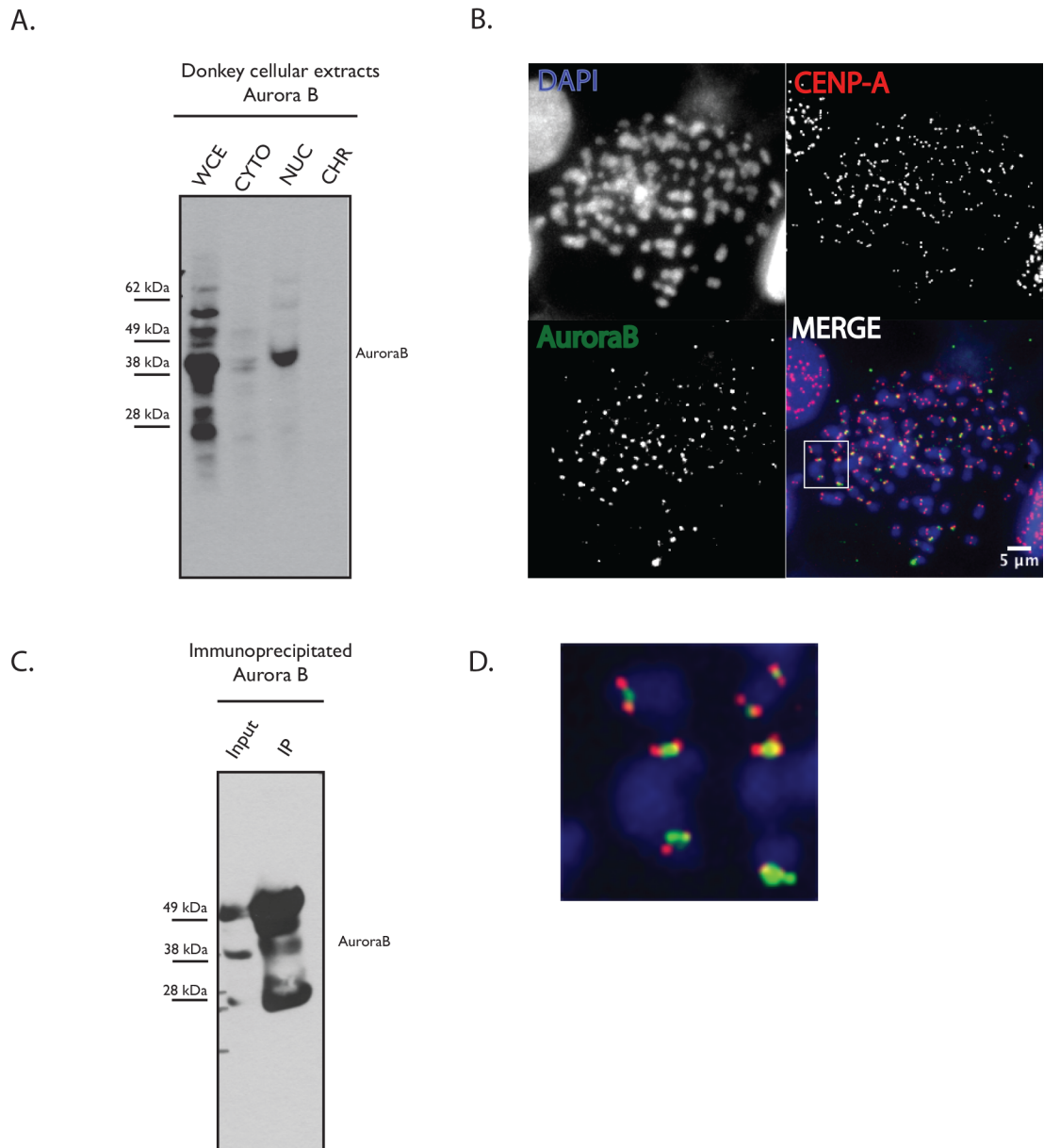


**Figure 5.3 Smc1 characterization in donkey fibroblasts.** Western blot analysis of cellular extracts (A): a band corresponding in size to that of Smc1 (170kDa) was observed in the whole cell extract (WCE), cytoplasm (CYTO), nuclear (NUC) and the chromatin fraction (CHR). Western blot analysis of Smc1 immunoprecipitation shows that Smc1 was solubilized and pulled down.

### 5.3.2 CPC-Aurora B

Aurora B was validated for use in the donkey system by western blot and immunofluorescence (Figure 5.4). Western blot analysis was carried out using a 1:1000 dilution of the 250ug/ml antibody. The antibody shows cross reactivity, with Aurora B signal present in both the whole cell extract and the nuclear fractions, a small fraction was also present in the cytoplasm. Aurora B was not detectable in the chromatin fraction, suggesting that it was inefficiently solubilized using this method (Figure 5.4 A). Aurora B distribution in donkey metaphase chromosomes was also examined by immunofluorescence as shown in Figure 5.4 B, D. By costaining with CENP-A (TRITC) and Aurora B (FITC), the inner centromere localization of Aurora B can be detected at some metaphase chromosomes, between paired CENP-A foci. In other instances, Aurora B appears to spread into the CENP-A binding domain. Studies of Aurora B distribution at the human neocentromere PD-NC4 shows altered spatial distribution (Bassett et al., 2010), perhaps the spreading of aurora B into the CENP-A binding domain is a result of the absence of satellite sequences. In a bid to address this question, immuno-FISH was employed using whole genomic DNA as a probe to show the satellite containing centromeres. However, using this method, CENP-A or Aurora B signal was undetectable at the primary constriction. ChIPSeq can be used to address the question of Aurora B spreading as protein association with the CENP-A binding domain can readily be examined at the 16 donkey unique sequence

centromeres. The application of this antibody in immunoprecipitation was examined, using 500ng of antibody per  $1 \times 10^6$  chromatin cell equivalents. A 41kDa band, corresponding in size to Aurora B was detectable in the IP fraction, Figure 5.4 C. Given the cross-reactivity of this antibody and its application in immunoprecipitation as determined by western blot, this antibody was selected for use in ChIPSeq.

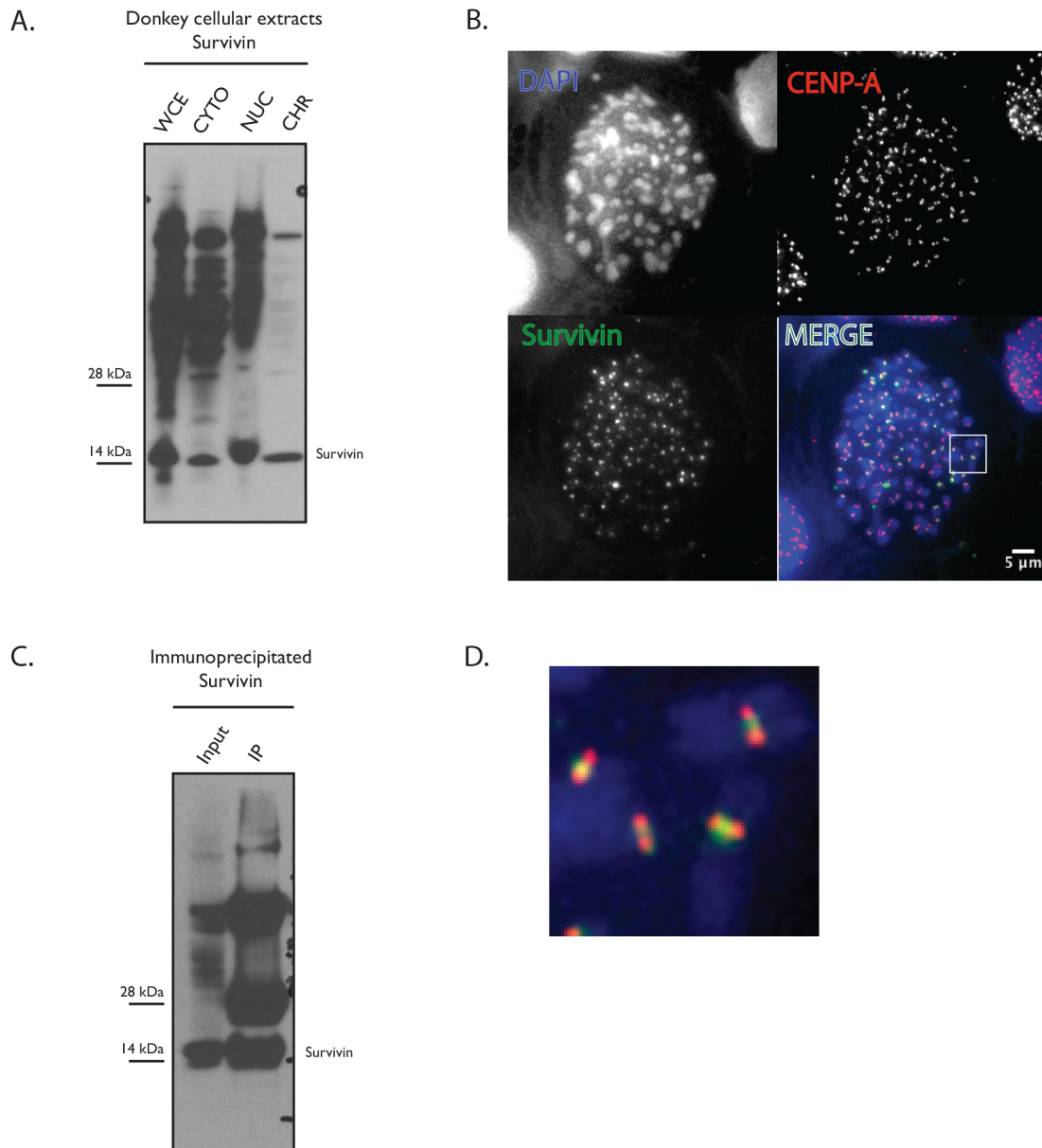


**Figure 5.4 Aurora B characterization in donkey fibroblasts.** **A)** Western blot analysis of cell extracts shows a band corresponding in size to that of Aurora B (41kDa) in the whole cell extract (WCE), a small fraction was also present in the cytoplasm (CYTO), Aurora B was abundant in the nuclear (NUC) fraction but was not detectable in the chromatin fraction (CHR). **B)** Immunofluorescence analysis shows Aurora B (FITC) present at the inner centromere of some chromosomes. **C)** Western blot analysis of Aurora B immunoprecipitation shows that the antibody pulls down a protein corresponding in size to Aurora B suggesting its suitability in ChIPSeq. **D)** Further inspection of the distribution of Aurora B at the inner centromere shows spreading of Aurora B into the CENP-A binding domain, some instances of exclusively inner centromere association.



### **5.3.3 CPC- Survivin**

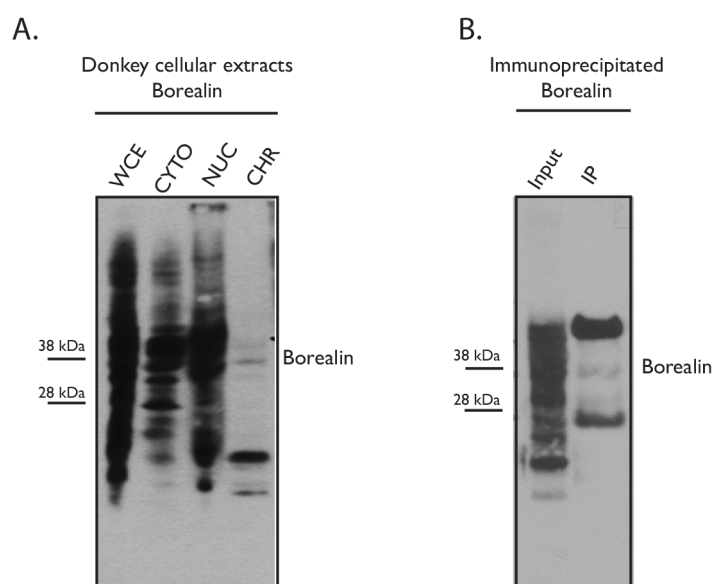
Characterisation of the survivin antibody was performed by western blot, immunofluorescence and immunoprecipitation (Figure 5.5). Western blot was performed using a 1:1000 dilution of the 1mg/ml survivin antibody. Figure 5.5 A, shows antibody reactivity against the donkey protein, with a 16.5kDa band present in the whole cell, cytoplasmic, nuclear and chromatin fractions. Immunofluorescence shows a similar distribution to that of Aurora B, with instances of Survivin spreading into the CENP-A binding domain, discrete inner centromere localization as well as centromeres with no detectable Survivin signal (Figure 5.5 B, D). When considering both Aurora B and Survivin localisation, immunofluorescence suggests that the localization of CPC subunits, may be perturbed in the absence of satellite sequences. Immunoprecipitation was carried out using 1ug of antibody per  $1 \times 10^6$  chromatin cell equivalents and a 16.5kDa protein was observed by western blot (Figure 5.5 C), suggesting utility in ChIPSeq.



**Figure 5.5 Survivin characterization in donkey fibroblasts.** **A)** Western blot analysis of cell extracts shows a band corresponding in size to that of Survivin (16.5kDa) in the whole cell extract (WCE), cytoplasm (CYTO), nuclear (NUC) and in the micrococcal nuclease digested chromatin (CHR) fractions. **B) D)** Immunofluorescence shows the different distributions of Survivin at the inner centromere. In some instances Survivin spreads into and colocalises with the CENP-A binding domain, at other chromosomes, Survivin is localised exclusively between paired CENP-A foci while at other inner centromeres, Survivin signal is undetectable. **C)** Western blot analysis of Survivin immunoprecipitation shows that the protein is pulled down, with a 16.5kDa band present in the IP fraction.

### 5.3.4 CPC- Borealin

The suitability of the Borealin antibody for use in the donkey was characterized by western blot using a 1:1000 dilution of the 1mg/ml Borealin antibody. The antibody was shown to cross-react with the donkey protein, a 35kDa band detectable in all four cellular fractions (Figure 5.6 A). The preparation of the chromatin fraction appears to inefficiently solubilize Borealin, with a reduced level of protein detectable by western blot. Western blot also shows high background signal. Immunoprecipitation was performed using 1ug of antibody per  $1 \times 10^6$  chromatin cell equivalents, and band corresponding in size to borealin was observed by western blot (Figure 5.6 B), suggesting the antibody is useful for ChIPSeq experiments. No Borealin signal was detectable on metaphase spreads fixed with methanol or formaldehyde. Borealin is a monoclonal antibody perhaps the epitope is lost in the spread preparation, rendering the antibody unable to bind to the protein.



**Figure 5.6 Borealin characterization in donkey fibroblasts.** **A)** Western blot analysis of cell extracts show a band corresponding in size to that of Borealin (35kDa) in the whole cell extract (WCE), cytoplasm (CYTO), nuclear (NUC) and in the micrococcal nuclease digested chromatin (CHR) fractions. **B)** Western blot analysis of Borealin immunoprecipitation shows that the protein is pulled down suggesting that the antibody is useful for ChIPSeq experiments

The three antibodies against the CPC subunits Aurora B, Survivin and Borealin as well as the cohesin antibody, Smc1, shows cross reactivity with the donkey by western blot analysis of cellular extracts and in the case of Aurora B and survivin, immunofluorescence. These antibodies have been characterized by ChIP western blot and deemed suitable candidates for use in ChIPSeq.

An antibody's application in ChIP is usually validated by qPCR. The sheep CENP-A antibody described in Chapter 3 was validated for use in ChIP by a set of three primers: EAS30, a primer pair that amplifies a 103bp region in the unique sequence centromere of chromosome 30, EcaCen11, a primer pair amplifying a region in the horse unique sequence centromere and PRKC, a primer pair within the single copy PRKC gene. Use of these primers did not detect any enrichment for cohesin or the CPC subunits. We hypothesize that the inner centromere associated DNAs flank the CENP-A binding domain, for this reason primers were designed against regions flanking the CENP-A binding domain. Immunofluorescence also shows some overlap with the CENP-A binding, therefore primers were also designed within the centromere region. Table 5.2 shows the primers that were characterized by initial PCR and yielded a single band. These primers are designed to amplify short sequences of ~100bp, however due to the lack of specific information about the location of the CPC and cohesin associated sequences it is not straightforward to design an adequate qPCR screening strategy. For this reason, western blot analysis was used to characterize antibody application in ChIP.

Primer	Sequence	Distance from centromere	Location	Amplicon size (bp)	Side
Eas30_F1 Fw EAS30_F1 Rv	ACTGGCTTTGGGTCTAATGG CTCCCTACTTTGACCCTTGC	~7kb	17,716,878- 17,716,992	114	5'
Eas30_F2 fw Eas30_F2 rv	GCATTATGAGTGCCCAGAGG TGCTACCATTCTCCATTGC	within	17,809,699- 17,809,800	101	3'
Eas30_F3 Fw Eas30_F3 rv	AACAAGACCCACCAACATGC TGGTTTGCCGTTATCTTGG	~3.8kb	17,814,572- 17,814,695	123	5'

**Table 5.2 qPCR primers for inner centromere identification on chromosome 30.**

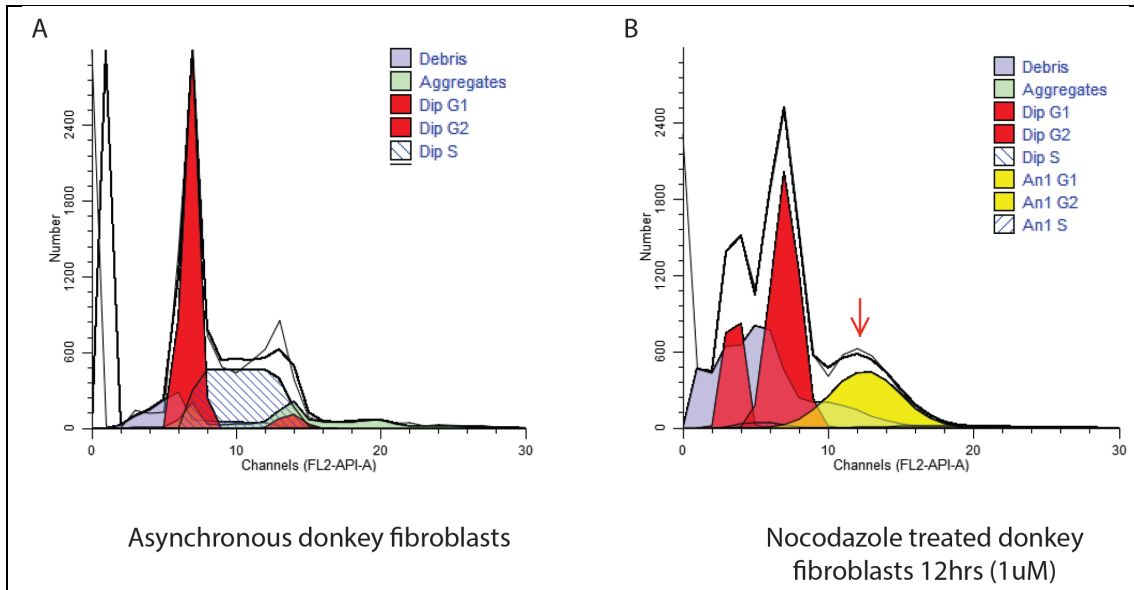
#### **5.4 Preparation of mitotically enriched cell populations**

The inner centromere is a compartment formed specifically on mitotic chromosomes, in order to be able to map its location a mitotic population of cells must be generated. To do this, a number of different approaches based on cell cycle inhibition strategies, including mitotic arrest and synchrony/release were carried out. For mitotic arrest, the utility of a number of different pharmacological agents was examined. Synchrony and release methods were employed using serum starvation and thymidine arrest, in order

to establish an optimum method for generation of a mitotic population. Cell cycle analysis was performed by flow cytometry and population analysis was carried out using *ModFit LT* from *Verity*. Cells were prepared for flow analysis by fixation in 70% ethanol. Flow analysis was carried out on propidium iodide stained cells using the BD Accuri.

#### **5.4.1 Mitotic arrest**

For initial characterization of a mitotically arrested population of donkey fibroblasts, cells were arrested with 1 $\mu$ M of Nocodazole for 12 hours. 1 $\mu$ M of Nocodazole has been shown to induce mitotic arrest in many cell lines (Blajeski, Phan, Kottke, & Kaufmann, 2002) and given the ~24hour doubling time of the donkey fibroblasts, it was anticipated that approximately half the population would be arrested in mitosis after 12 hours. Figure 5.7 and table 5.3 shows the distribution of cells throughout the cell cycle in an asynchronous population and distribution of cells after a 12 hour Nocodazole arrest. In the asynchronous population 100% of the cells are diploid, with 59.14%, 6.99% and 33.88% in G1, G2/M and S respectively. In the Nocodazole treated cells 69.60% of the population are diploid while the remaining 30.40% have executed mitosis despite the absence of polymerized microtubules and become tetraploid. This suggests the donkey fibroblasts have a weak mitotic checkpoint. The accumulation of diploid cells with a G2/M DNA content (72.94%) suggests that the checkpoint is initially activated and weakens during prolonged activation. “Mitotic Slippage” is the term given to the progression of cells to interphase without chromosome segregation, through a mechanism of Cyclin B1 degradation in the absence of checkpoint inactivation (Brito, Yang, & Rieder, 2008). This could be the mechanism for bypassing the mitotic checkpoint in the donkey cells.



**Figure 5.7 Population distribution of asynchronous versus Nocodazole arrested donkey fibroblasts.** The diploid populations are shown in red, while tetraploid populations are shown yellow. S phase is shown in white with blue lines. **A)** In the asynchronous populations the majority of the cells have a G1 DNA content and no evidence of tetraploidy. **B)** In the nocodazole arrested cells the majority of the population are in diploid G2/M but display tetraploidy.

Treatment	Diploid%	G1	G2/M	S	Tetraploid %	G1	G2/M
Asynchronous	100	59.14	6.99	33.88	-	-	-
Nocodazole	69.60	24.62	72.94	2.44	30.40	93.31	6.69

**Table 5.3 Distribution of cells throughout the cell cycle.** Cells cycling normally in an asynchronous population, show no evidence of tetraploidy and the majority of the population is found at G1. In nocodazole treated cells, there are two cell populations, diploid (69.60%) and tetraploid (30.40%).

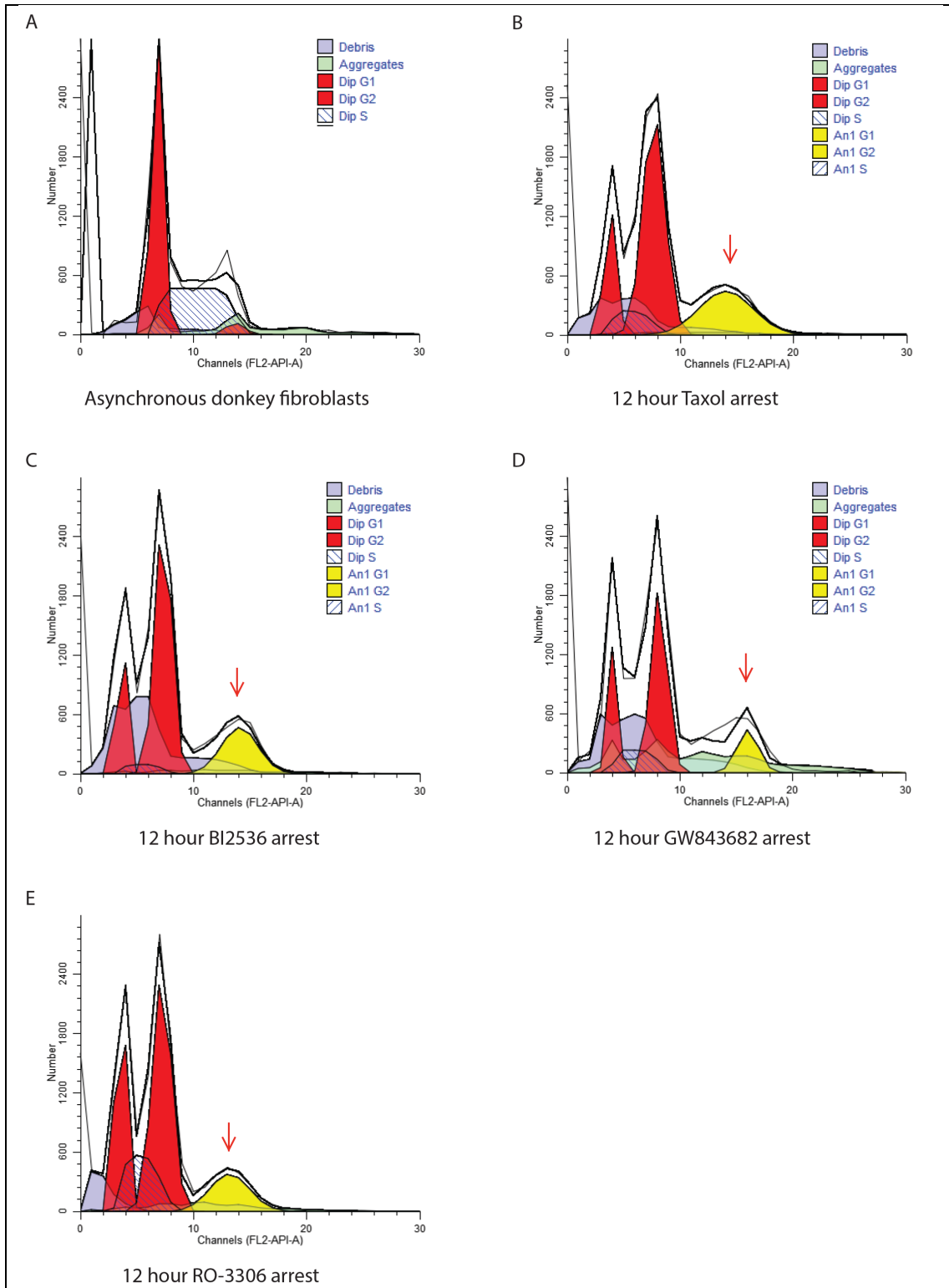
While use of nocodazole arrest suggests a weak mitotic checkpoint, a number of inhibitors cause mitotic arrest through different pharmacological mechanisms. The great utility of pure mitotic populations for analysis of the inner centromere led to several inhibitors being examined (table 5.4) to empirically determine if any could be utilized for these experiments. Paclitaxel (taxol) results in a mitotic arrest by stabilization of microtubules and prevention of microtubule depolymerization. Cells have been shown to contain “near” normal bipolar spindles and chromosomes align at the metaphase plate normally (Weaver, 2014). BI2536 is a PLK1 inhibitor that binds PLK1, inhibiting its activity, subsequently resulting in loss of  $\gamma$ -tubulin recruitment to centrosomes, a key complex required in microtubule nucleation (Haren, Stearns, & Lüders, 2009). Plk1 has a role in mitosis regulation by cross talk with cell cycle mediators, where its involved in centrosome maturation, spindle formation, chromosome alignment and cytokinesis (Hartsink-Segers et al., 2013). Plk1 is also required for cohesin removal from chromosome arms (Giménez-Abián et al., 2004)

rendering it impractical for use in mapping cohesin to the inner centromere. GW843682X is an ATP-competitive inhibitor of both PLK1 and PLK3 and yields the same downstream effects as BI2536. RO-3306 is an ATP competitive selective inhibitor of Cdk1, arresting cells in late G2 and can subsequently be washed out allowing cells to cycle synchronously into mitosis (Vassilev, 2006). Cells were treated for 12 hours with the drugs and concentrations outlined in Table 5.4, it was anticipated that approximately half the population should be arrested at the given checkpoint. Drug concentrations employed in these experiments were taken from publications (Ikui, Chia-Ping, Matsumoto, & Horwitz, 2005; Lu et al., 2010; Vassilev, 2006).

<b>Drug</b>	<b>Stage</b>	<b>Mechanism</b>	<b>Concentration</b>
Taxol	M	Microtubule stabilization	50nM
Nocodazole	M	Inhibition of microtubule polymerization	1uM
BI2536	M	PLK1 inhibitor	9nM
GW843682X	M	PLK1 and PLK3 inhibitor	1uM
RO-3306	Late G2	ATP-competitive inhibitor of CDK1	9uM

**Table 5.4 Pharmacological agents employed to achieve arrest in the donkey fibroblasts**

Cells were treated with the drugs and concentrations shown in table 5.4 for 12 hours in an effort to accumulate approximately half the population in mitosis. Figure 5.8 and Table 5.5 show the distribution of cells throughout the cell cycle. Tetraploid populations were observed after all four drug treatments. This renders mitotic arrest using pharmacological agents impractical for these experiments. Further, analysis of the effect of pharmacological agents on Survivin localization shows a more diffuse localization pattern with the protein spreading along chromosome arms in many cases (Figure 5.9). Taking into account the perturbed localization of Survivin as well as the presence of tetraploid populations, the use of pharmacological agents was not further employed for generating mitotic populations. In a bid to achieve a more physiologically normal enriched mitotic population, a synchrony and release approach was examined.

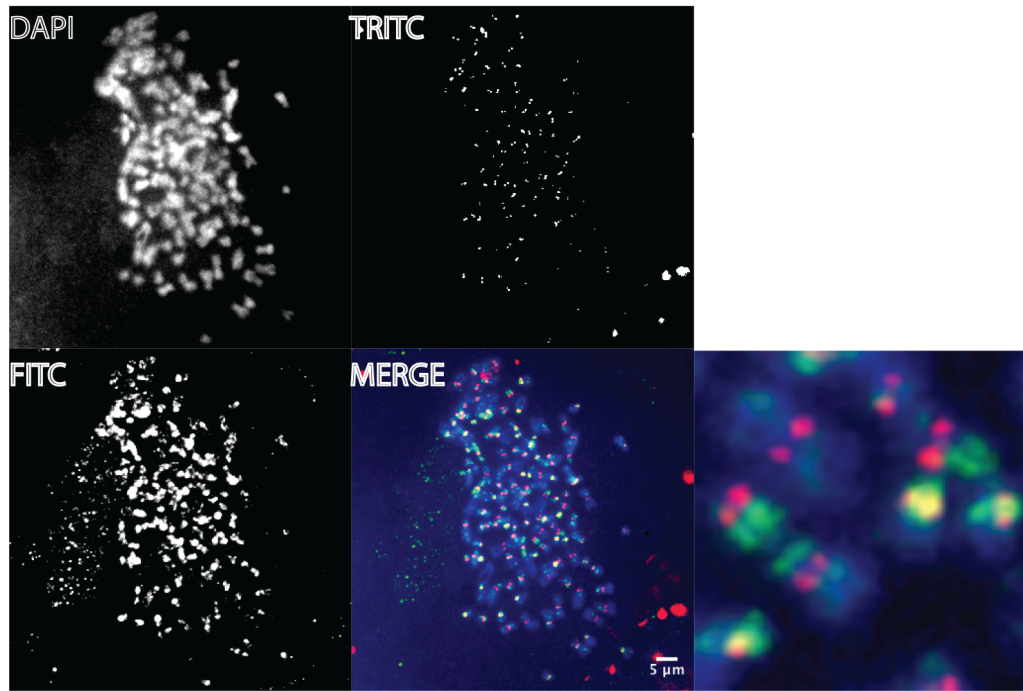


**Figure 5.8** Flow cytometric analysis of population distributions in drug treated donkey fibroblasts. Diploid populations are shown in red, while tetraploid populations are shown in yellow. **A)** The distribution of cells in an asynchronous population. In the four drug treated samples **B)** taxol, **C)** BI2536, **D)** GW843682 and **E)** RO-3306, cells appear to accumulate at the G/M boundary before by passing the checkpoint and becoming tetraploid.

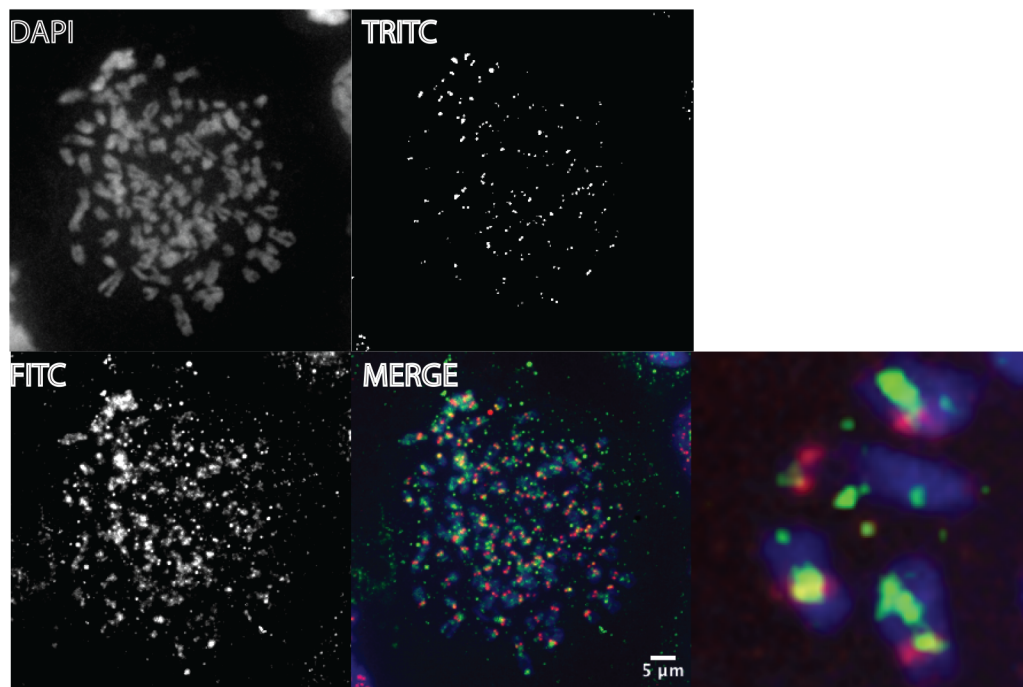


<b>Treatment</b>	<b>Diploid%</b>	<b>G1</b>	<b>G2/M</b>	<b>S</b>	<b>Tetraploid %</b>	<b>G1</b>	<b>G2/M</b>
Asynchronous	100	59.14	6.99	33.88	-	-	-
Taxol	75.77	21.41	66.95	11.64	24.23	98.18	1.81
BI2536	78.45	23.94	70.85	5.20	21.55	99.53	0.47
GW843682	80.89	25.58	58.67	15.75	19.11	74.5	25.50
RO-3306	86.00	28.21	51.04	20.75	14.00	100	-

**Table 5.5 Percentage distribution of cells throughout the cycle after treatment with pharmacological agents**



**BI3536 treatment**



**Taxol treatment**

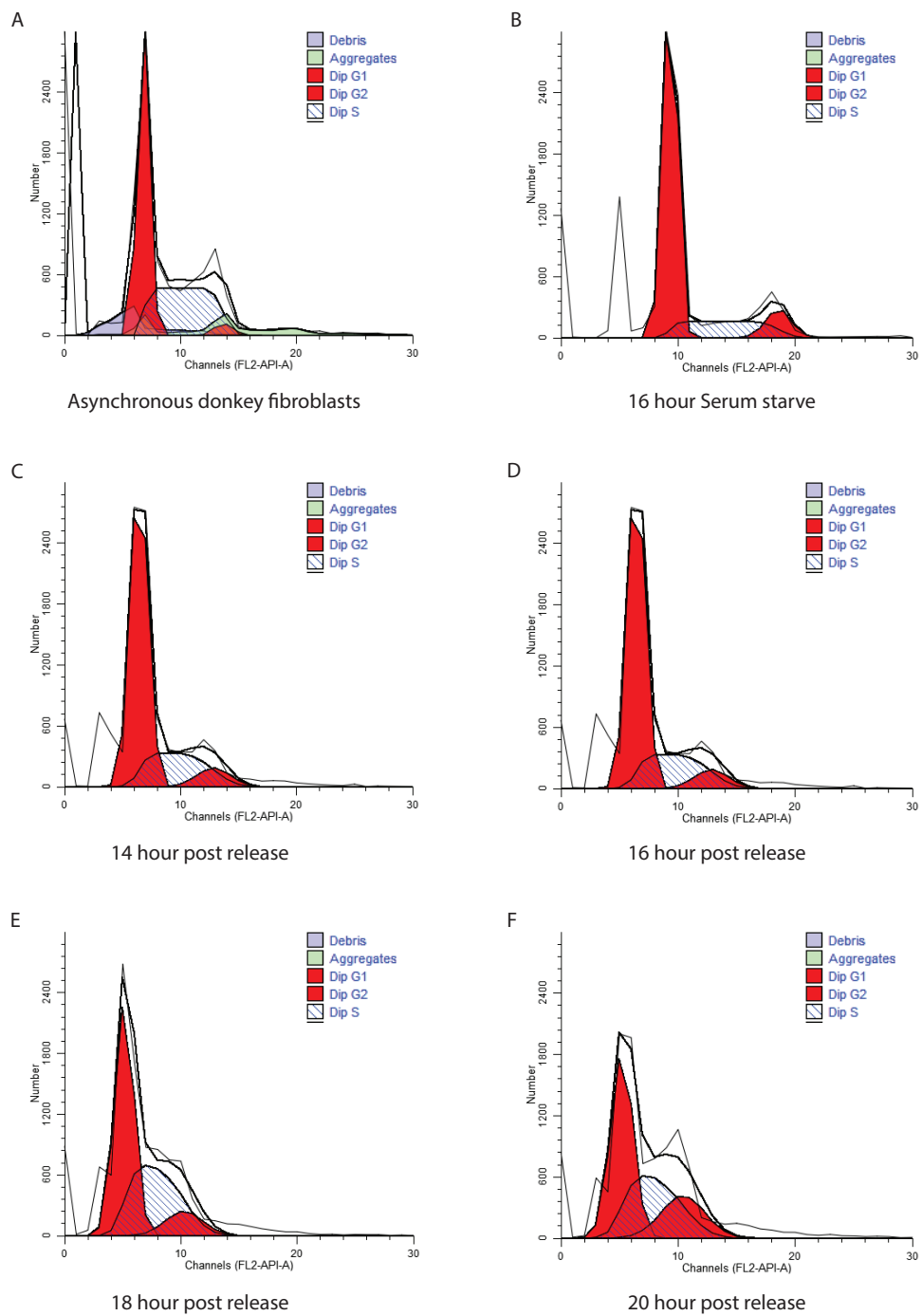
**Figure 5.9 Survivin staining on metaphase chromosomes following treatment with BI3536 and Taxol.** The localization of Survivin appears to be perturbed on chromosomes treated with BI3536 (top) and Taxol (bottom) with non-specific inner centromere localization and spreading along chromosome arms

#### ***5.4.2 Synchrony and release***

Given the weak mitotic checkpoint point in the donkey fibroblasts and the perturbed localization of Survivin following the use of mitotic inhibitors, the practicality of G1 arrest and release was examined. The utility of serum starvation and thymidine block were tested to generate a G1 or G1/S phase arrest that could subsequently be synchronously released allowing for the accumulation of mitotic populations. Serum starvation, is essentially withdrawal of mitogenic factors which results in induced quiescence; cells are arrested with a DNA content equivalent to G1. Upon the addition of serum, cells should exit their quiescent state and continue cycling. This method could be useful depending on the ability of the donkey cells to successfully exit the quiescent state. The rationale of the thymidine block is that high concentrations of thymidine arrest cells in S phase through inhibition of ribonucleotide reductase (Bootsma, Budke, & Vos, 1964). Ribonucleotide reductase is involved in the biosynthesis of pyrimidines and inhibition halts DNA synthesis in S phase by depleting pools of deoxycytidine-5-triphosphate (dCTP) halting cells at the G1/S transition. A double thymidine block is often utilized to generate a synchronous population. The first thymidine block imposed should result in approximately half the population distributed throughout S phase while the other half is arrested at the beginning of S phase. The cells are then released allowing population accumulation in early G2 and G2/M. A second thymidine block is then imposed and the entire population of cells should be arrested at the beginning of S phase. Release of this population results in synchronous entry into S phase.

In the serum starvation experiments, cells were cultured in complete media containing 1% serum for 16 hours. Extended periods of starvation were avoided, to minimize permanent quiescence and apoptosis. Cells were then released into complete media containing 20% serum and cells were harvested every two hours, 14 hours after release to monitor mitotic progression. Figure 5.10 and Table 5.6 shows the distribution of cells throughout the cycle following addition of complete media. The progression of cells through the cell cycle after serum starvation is delayed, 14 hours after release, 68.22% of the population have a G1 DNA content with the cells entering S phase (24.33%) in a non uniform manner. 16 hours after release 65.17% of the population are in G1 with 27.52% in S phase. 18 and 20 hours post complete media addition shows a similar trend with cells leaving G1, entering S phase and G2/M in a non-synchronous manner. The constant presence of a strong of the G1 peak suggests

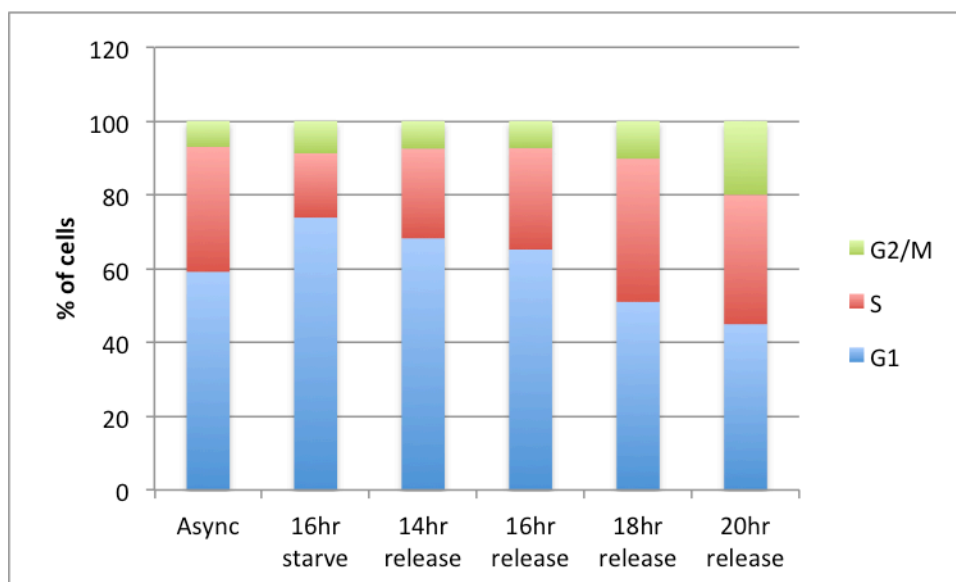
that many of the cells remain in a quiescent state following the addition of whole media. The distribution of cells throughout the serum starve and release are also shown in a stacked column chart, with G1 shown in blue, S phase in red and G2/M in green, Figure 5.11. Since the addition of complete media results in populations exiting the quiescent state in a non-synchronous manner with no significant accumulation of G2/M, this method was not employed further.



**Figure 5.10** Flow cytometric analysis of serum starved (1% serum) and released donkey fibroblasts with propidium iodide DNA staining. **A**) The distribution of cells throughout an asynchronous population. After a 16 hour starvation (**B**), cells were released into full media containing 20% FBS. Cells were analyzed 14 hours (**C**), 16 hours (**D**), 18 hours (**E**) and 20 hours (**F**) post release.

	Diploid%	G1	G2/M	S
Asynchronous	100	59.14	6.99	33.88
16 hour Serum starve	100	73.84	8.73	17.43
14 hour post release	100	68.22	7.45	24.33
16 hour post release	100	65.17	7.32	27.52
18 hour post release	100	51	10.18	38.82
20 hour post release	100	44.94	19.94	35.11

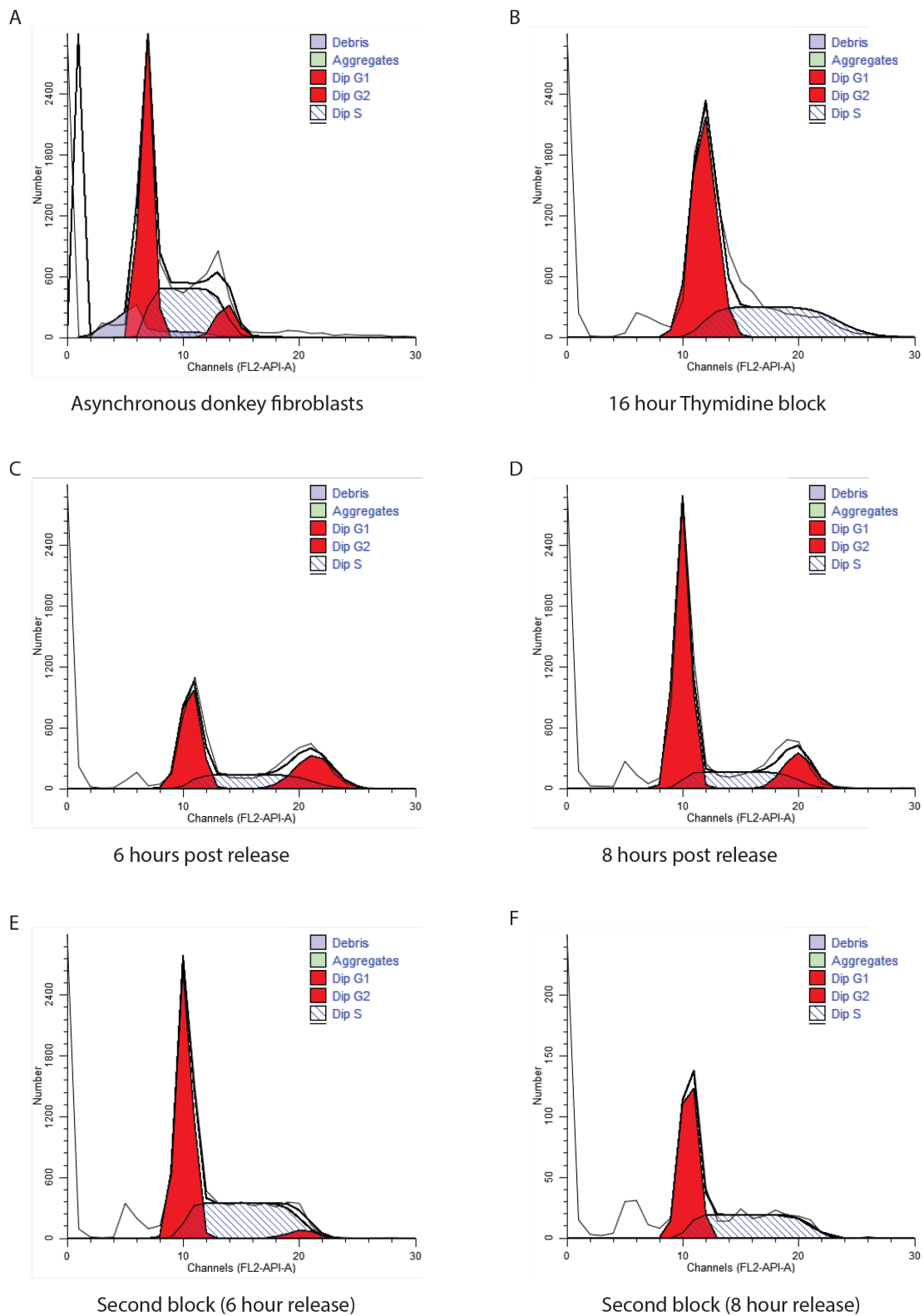
**Table 5.6 Distribution of cells in an asynchronous and in serum starved and release populations**



**Figure 5.11 Graphical representation of population distribution after serum starvation and release**

Given the poor synchrony achieved using serum starvation the utility of a thymidine arrest strategy in generating an enriched mitotic population was explored. Cells were prepared for flow analysis as described in Section 5.4. The first thymidine block imposed should last the equivalent of  $G_2 + M + G_1$ . The distribution of cells in the asynchronous population in Table 5.6 shows that 66.13% are in G1, G2/M. Given the 24hour doubling time, the length of G1, G2/M is approximately 16hours. Cells were treated with 2mM thymidine for 16 hours and released into complete media containing 24uM deoxycytidine, restoring dCTP levels. Following the initial thymidine block, approximately half the cell population should be distributed throughout S phase. In order to calculate the optimum release time which should be the equivalent of S-phase, cells were allowed cycle for 6 and 8 hours respectively post initial thymidine block. A second 16 hour thymidine block was then imposed which should result in cells being synchronously blocked at the beginning of S phase. Figure 5.12 and Table 5.7 show the distributions of cells throughout the cell cycle in an asynchronous population and in the thymidine treated populations. After the initial thymidine block, Figure 5.12 B, 62.73% of the cells were in G1 and 37.27% in S

phase. Cells were then released to allow populations arrested at G1/S to reenter the cell cycle. Figure 5.12 C & Table 5.7 shows the distribution of cells 6 hours post release from the thymidine block with 44.66% in G1, 26.40% in G2/M and 28.93% in S phase. 8 hours post release 63.81% of the population are in G1, 14.29% in G2/M and 21.70% in S (Figure 5.12 D, Table 5.7). A second 2mM thymidine block was imposed on the cells after the 6 and 8 hour release and cells were analyzed at the end of the second 16 hour thymidine block. Surprisingly, Figure 5.12 E & F and Table 5.7, show that the second thymidine block was ineffective, the cells were not synchronous arrested at G1/S with almost half the cells still in S phase. Figure 5.13 shows a stacked column chart showing the population distributions after thymidine treatment with G1 (blue), S phase (red) and G2/M (green).

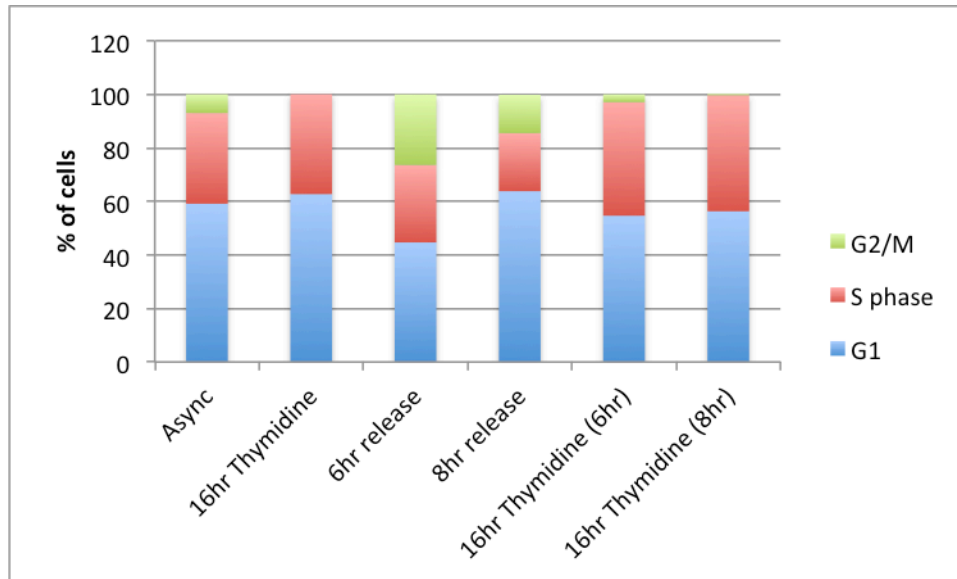


**Figure 5.12 Double Thymidine block analysis.** The distribution of cells in an asynchronous population is shown in A. The distribution of cells after a 16hour 2mM Thymidine block (B), shows the majority of the cells are in G1 with approximately one third in S phase. Following release into complete media containing 24uM deoxycytidine for six (C) and eight hours (D) respectively, cells exit G1 and traverse through the cycle. A second thymidine block was imposed after the six (E) and eight hour (F) release and population distributions were examined.



	Diploid%	G1	G2/M	S
Asynchronous	100	59.14	6.99	33.88
16 hour Thymidine block	100	62.73	-	37.27
6 hour post release	100	44.66	26.40	28.93
8 hour post release	100	63.81	14.29	21.70
Second block (6 hour release)	100	54.66	2.94	42.40
Second block (8 hour release)	100	56.31	0.33	43.36

**Table 5.7 Distribution of cells throughout the cell cycle in asynchronous and thymidine treated cells**



**Figure 5.13 Graphical representation of population distribution after thymidine block, release and second thymidine block**

Implementing a double thymidine block over this time frame does not generate a synchronous population. Further efforts to optimize the thymidine synchrony of the cells with a shorter thymidine block, also failed to yield a synchronous population. A proportion of the population appears to become permanently quiescent with a G1 DNA content following release from the thymidine block, regardless of the length of the block imposed. As well as this, a substantial S-phase population broke through the second arrest. The accumulation of cells in G2/M six hours post release from the single thymidine block is 3.77 times greater (26.40%) than that observed in the asynchronous population (6.99%). Given the accumulation of cells in G2/M 6 hours post thymidine release, it was decided that this may be enough to see inner centromere enrichment by ChIPSeq.

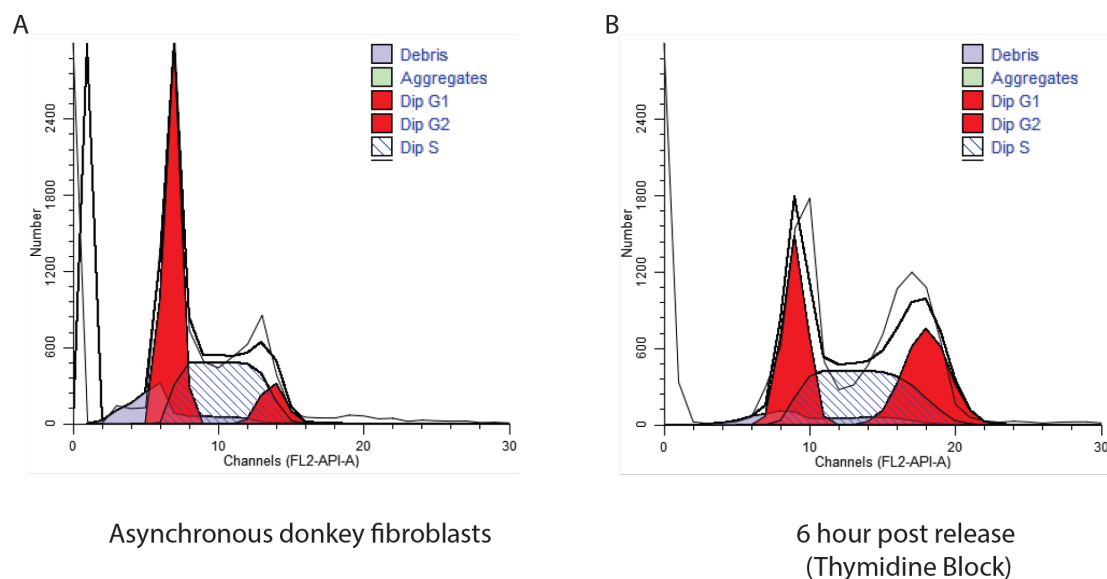
## **5.5 ChIPSeq using mitotically enriched populations**

Mitotically enriched ChIPSeq was carried out by imposing a single 16 hour thymidine block and releasing cells for 6 hours allowing G2/M accumulation. Cells were crosslinked with EGS and formaldehyde, a combination shown to be useful for crosslinking proteins that indirectly bind DNA (Zeng et al., 2006). Chromatin was prepared by sonication, precleared with protein G beads antibody was then added and incubated overnight. The antibody and protein associated DNA was recovered by the addition of protein G beads and the DNA was purified and sent for sequencing. The ChIPSeq reads were processed, normalized and visualised as described in Section 3.6.

### **5.5.1 CENP-A ChIPSeq**

In mitosis, there is half the full complement of CENP-A at the centromere. The CENP-A nucleosome has been shown to be rigid (Black et al., 2004), but in both satellite containing and satellite free centromeres wraps DNA less tightly than canonical histones (Hasson et al., 2013). CENP-A has been shown to be exclusively localized to the chromosome surface in the inner kinetochore plate (Marshall, Marshall, et al., 2008; Warburton et al., 1997). An alternative hypothesis is that a chromatin remodeling process enriches the kinetochore surface in CENP-A nucleosomes. Several models for the folding of CENP-A associated chromatin fibers have been proposed including amphipathic helices/loops and the boustrophedon model whereby the CENP-A nucleosomes are clustered on the outer surface of the chromatin (Fukagawa & Earnshaw, 2014). In order to establish if a change in the kinetochore competent CENP-A binding domain could be detected, ChIPSeq was carried out on a mitotically enriched population of cells using the sheep CENP-A antibody generated in Chapter 3.

Mitotically enriched populations were generated by a single thymidine block and release method described in section 5.4.2 and ChIP was performed on EGS and formaldehyde crosslinked cells as described previously (section 2.2.2.12). Figure 5.14 and table 5.8 show the distributions of cells in an asynchronous population and in a mitotically enriched population, with cells with a G2/M DNA content increasing from 6.99% to 30.07%.



**Figure 5.14 Proportion of cells distributed throughout the cell cycle in an asynchronous population and following release from a single thymidine block (CENP-A ChIPSeq).** A) shows the distribution of cells throughout an asynchronous population and B) shows the distribution of cells used in the mitotically enriched CENP-A ChIPSeq, 6 hours post thymidine block release

	Diploid%	G1	G2/M	S
Asynchronous	100	59.14	6.99	33.88
6 hour post release	100	30.93	30.07	39.00

**Table 5.8 Distribution of cells in an asynchronous population and the mitotically enriched population used in CENP-A ChIPSeq**

The CENP-A ChIPSeq read details are shown in table 5.9 with read lengths of between 125-150bp. The sequencing reactions were run in two lanes, L001 and L005. Paired end sequencing was employed and R1 and R2 refer to the either end. Read quality was examined and sequences were processed as described in Sections 2.4 and 3.6. Table 5.10 shows the percentage of reads dropped after trimming. Overall this experiment resulted in high quality sequence data.

File name	Input/ChIP	Sequence length (bp)	Reads
4_AGAGGATG_L001_R1_001.fastq	Input	150	7649200
4_AGAGGATG_L001_R2_001.fastq	Input	150	7649200
4_AGAGGATG_L005_R1_001.fastq	Input	125	6443430
4_AGAGGATG_L005_R2_001.fastq	Input	125	6443430
5_ACGTTCT_L001_R1_001.fastq	ChIP	150	5988558
5_ACGTTCT_L001_R2_001.fastq	ChIP	150	5988558
5_ACGTTCT_L003_R1_001.fastq	ChIP	125	6484692
5_ACGTTCT_L003_R2_001.fastq	ChIP	125	6484692

**Table 5.9 Mitotically enriched CENP-A Sequence details**

Library	Reads dropped (%)
Mitotic CENPA/Borealin Input L001	0.61
Mitotic CENPA/Borealin Input L005	1.03
Mitotic CENPA ChIP L001	0.68
Mitotic CENPA ChIP L003	0.95

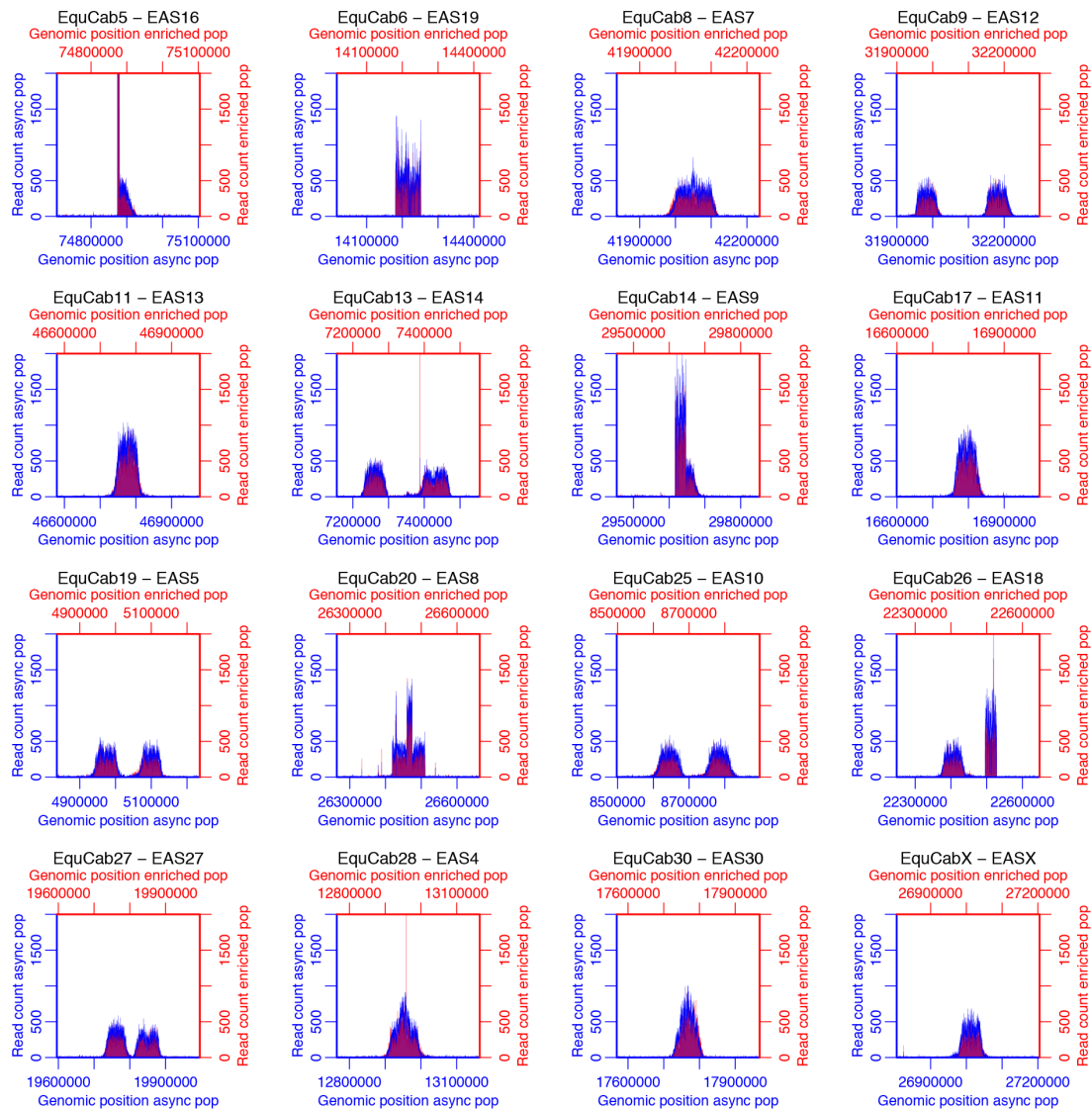
**Table 5.10 Mitotically enriched CENP-A reads dropped after trimming**

#### **5.5.1.2 FRiP (fraction of reads in peaks) analysis**

FRiP was performed as described in section 3.6.1 and the proportion of reads at the centromere domains were measured. The FRiP score for mitotically enriched CENP-A was 4.44% indicating that the ChIPSeq was successful. The ENCODE consortium scrutinizes experiments with FRiP values less than 1%.

#### **5.5.1.3 Comparison of ChIPSeq in a mitotically enriched population versus an asynchronous population**

To get a comparable picture of the CENP-A distribution between a mitotically enriched population and an asynchronous population, the two CENP-A datasets were superimposed as shown in Figure 5.15. The asynchronous population is shown in blue and the enriched population is shown in red. The enrichment profiles and domain boundaries across the two data sets appear highly similar.



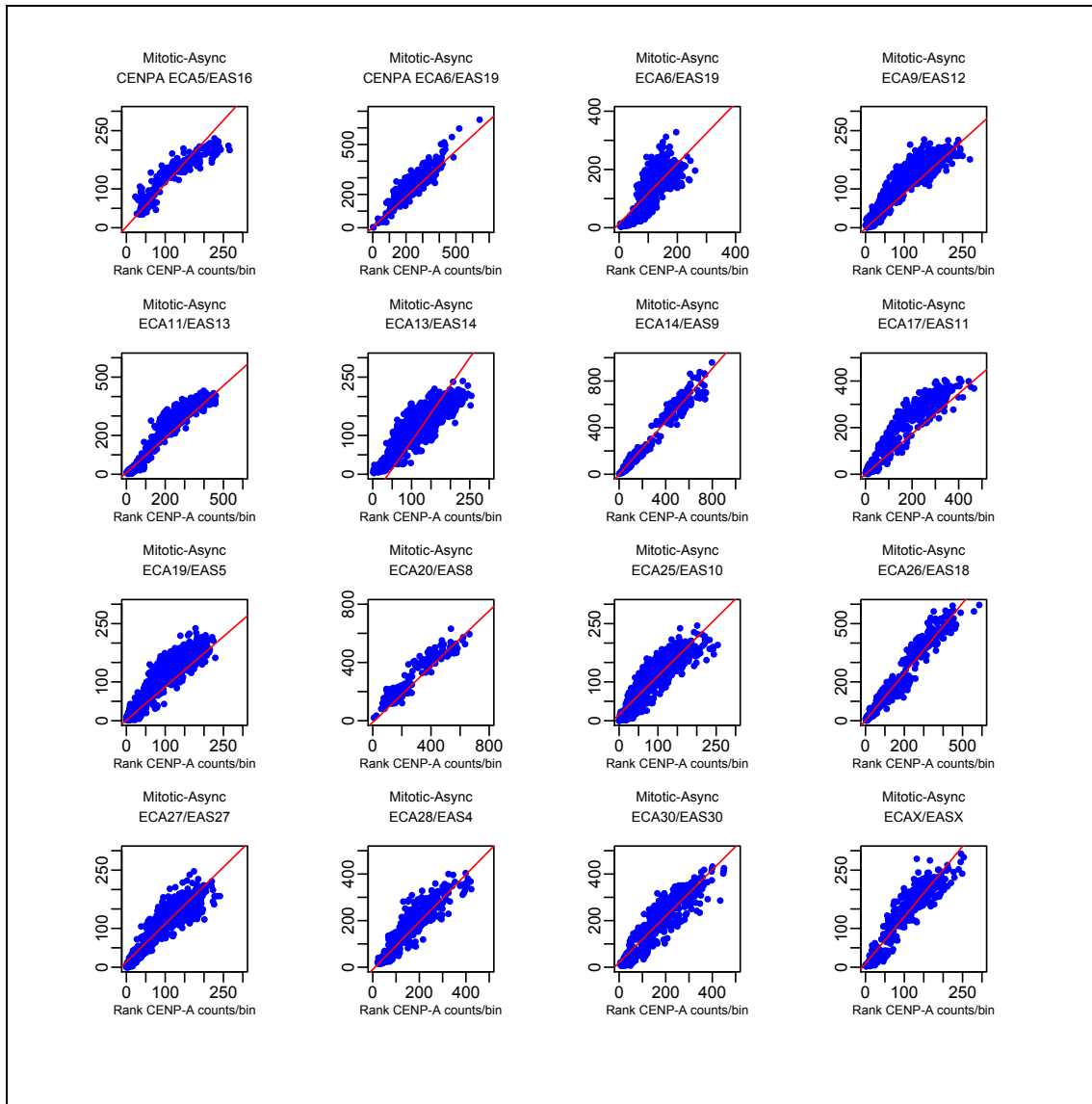
**Figure 5.15 Mitotically enriched CENP-A and asynchronous CENP-A profile comparison.** The centromere profiles for mitotically enriched CENP-A and asynchronous CENP-A were superimposed for all 16 unique sequence centromeres. The distribution of signal from both data sets are alike, showing highly similar boundaries.

A correlative approach was adopted to investigate the relationship between the mitotically enriched CENP-A ChIP and the asynchronous ChIP. To do this, each centromere domain was isolated and the read coverage for the centromeric regions in the CENPA asynchronous ChIP and mitotic ChIP, were calculated using the deeptools function “multiBamSummary”. This function computes the read coverage across multiple sorted bam files at given genomic regions, in this case in 200bp bins. The coverage per bin was correlated using the Spearman algorithm in R as described in Section 3.7. The rho values, correlation coefficients, for each centromere are shown in Table 5.11 while Figure 5.16 shows the correlogram scatter plots. This analysis shows high correlation across all centromere domains.

The Spearman values for all the centromeres in table 5.11 indicate a high (0.7-0.89) or very high (>0.9) positive correlation between the two CENP-A ChIP datasets (Mukaka, 2012). This indicates that across the centromere there is little difference in CENP-A distribution at the level of resolution in this experiment. In particular, the data are more highly correlated, showing much less evidence for CENP-A redistribution within the centromere than observed in comparison of data from two cell sources (Figure 3.13, Table 3.5)

<b>Centromere</b>	<b>Rho Value</b>
ECA5/EAS16	0.9554974
ECA6/EAS19	0.8603125
ECA8/EAS7	0.7591478
ECA9/EAS12	0.9084629
ECA11/EAS13	0.9560609
ECA13/EAS14	0.8686827
ECA14/EAS9	0.9747969
ECA17/EAS11	0.9399136
ECA19/EAS5	0.9252572
ECA20/EAS8	0.914095
ECA25/EAS10	0.8008305
ECA26/EAS18	0.9664674
ECA27/EAS27	0.9043324
ECA28/EAS4	0.9065447
ECA30/EAS30	0.9156731
ECAX/EASX	0.9334477

**Table 5.11 Spearman correlative values for the CENPA binding domain of the mitotically enriched CENPA ChIP and the asynchronous CENPA ChIP.**

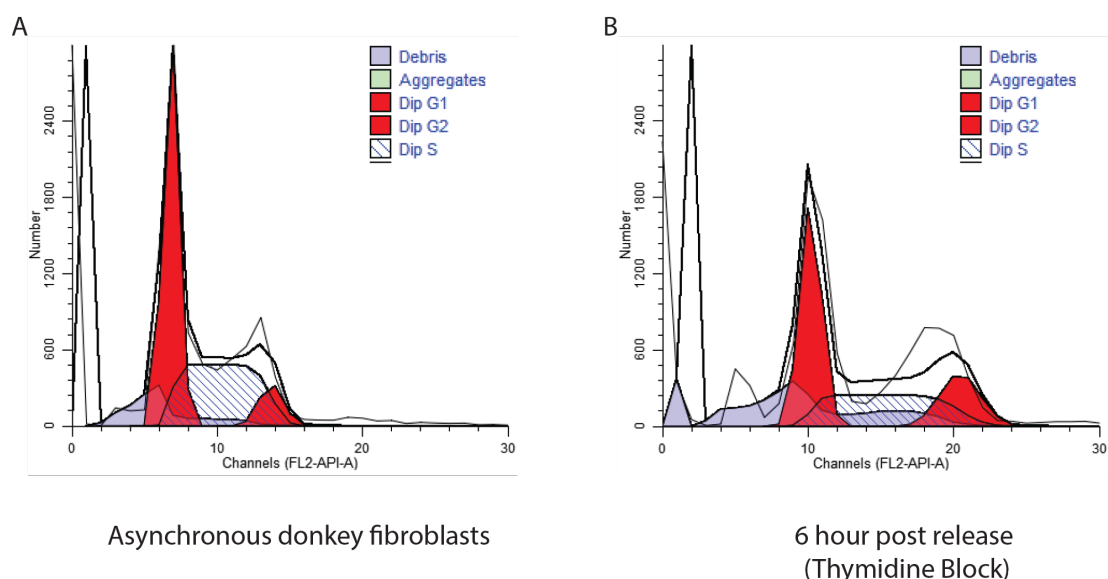


**Figure 5.16** Correllogram analysis of mitotically enriched CENPA ChIPSeq and an asynchronous ChIPSeq. Spearman correlation shows the CENP-A mitotically enriched and CENP-A asynchronous signal intensities.

Taking into account the superimposed CENP-A ChIP and the correlative analysis, the CENP-A distribution in both experiments co-occupy the same centromeric domains. This data shows that CENP-A distribution remains the same in a mitotically enriched population indicating that there is no redistribution or remodelling of CENP-A chromatin associated with a kinetochore competent centromere detectable at the level of resolution of this experiment. Indeed, given high proportion of cells in G1 (30.93%) and S phase (39.00%) perhaps this is providing too much noise in the data to discern changes in the chromatin structure. However given the fact 30% of the cells have a G2/M content we assume that if there was a significant redistribution of CENP-A the centromeres would not be so highly correlated to the asynchronous centromere domains.

### 5.5.2 Cohesin-Smc1 ChIPSeq

Mitotically enriched populations of cells were generated for Smc1 ChIPSeq. In this experiment  $50 \times 10^6$  cells were used. Cells were crosslinked and ChIPed in the same manner as described in section 5.5. Figure 5.17 and Table 5.12 show the distribution of cells in an asynchronous population and in the mitotically enriched population. The portion of cells with a G2/M DNA content in the enriched population is 2.78 times greater than that observed in the asynchronous population. The sequence details for the Smc1 reads are shown in Table 5.13. ChIPSeq reads were processed as described in Sections 2.4 and 3.6. The sequencing reactions for both the ChIP and the input were run in two lanes L006 and L007. Read quality was examined and the proportion of reads dropped are shown in Table 5.14.



**Figure 5.17** Proportion of cells distributed throughout the cell cycle in an asynchronous population and following release from a single thymidine block (Smc1 ChIPSeq).

	Diploid%	G1	G2/M	S
Asynchronous	100	59.14	6.99	33.88
6 hour release (post Thymidine block)	100	45.49	19.50	35.01

**Table 5.12** Distribution of cells in an asynchronous population and the mitotically enriched population used in Smc1 ChIPSeq



File name	Input/ChIP	Sequence length (bp)	Reads
12 AGCTAGTG L006 R1 001.fastq	Input	125	20781017
12 AGCTAGTG L006 R2 001.fastq	Input	125	20781017
12 AGCTAGTG L007 R1 001.fastq	Input	125	17921908
12 AGCTAGTG L007 R2 001.fastq	Input	125	17921908
13 AGGTCTGT L006 R1 001.fastq	ChIP	125	31363550
13 AGGTCTGT L006 R2 001.fastq	ChIP	125	31363550
13 AGGTCTGT L007 R1 001.fastq	ChIP	125	13281578
13 AGGTCTGT L007 R2 001.fastq	ChIP	125	13281578

**Table 5.13 Mitotically enriched Smc1 Sequence details**

Library	Reads dropped (%)
Mitotic Smc1 Input L006	1.01
Mitotic Smc1 Input L007	1.51
Mitotic Smc1 ChIP L006	1.83
Mitotic Smc1 ChIP L007	2.11

**Table 5.14 Mitotically enriched Smc1 reads dropped after trimming**

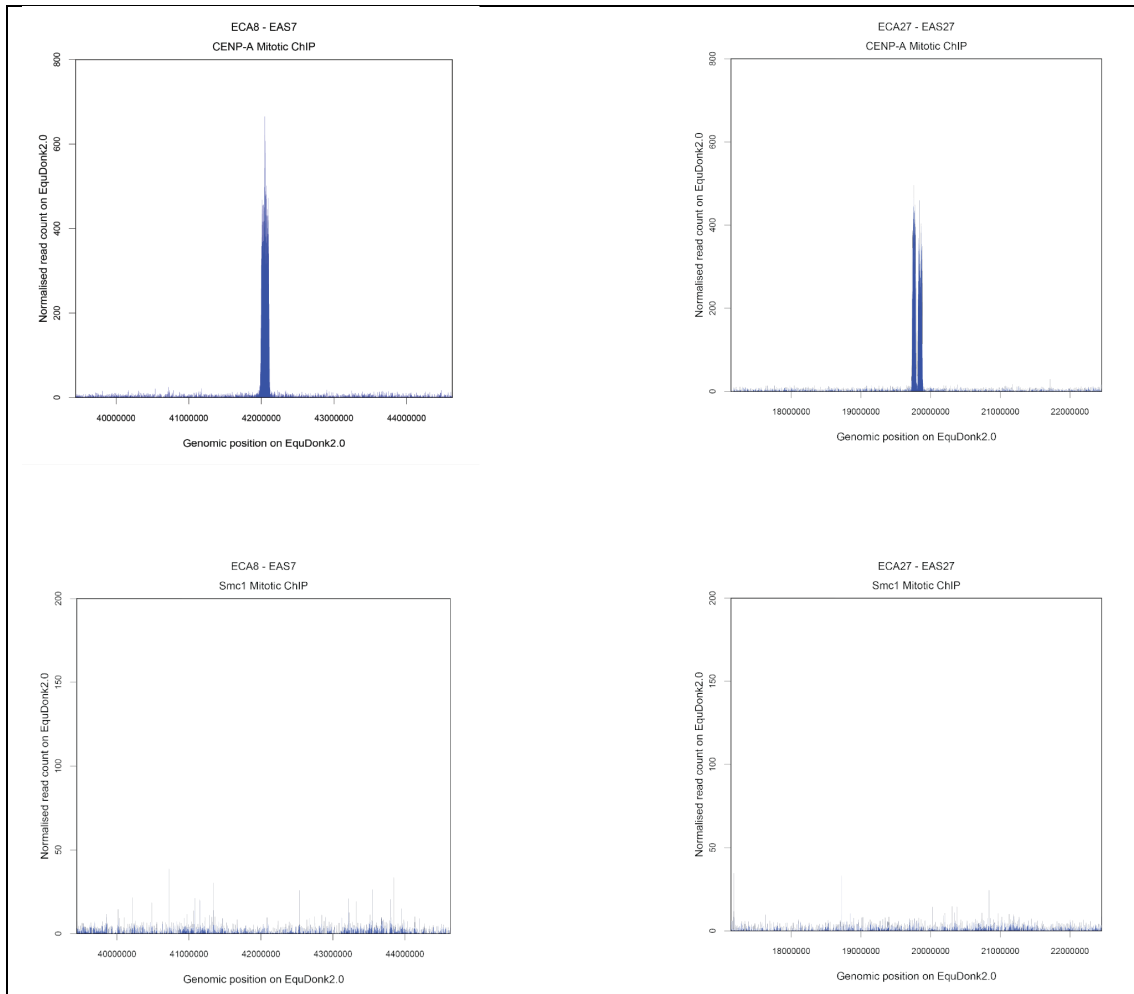
### 5.5.2.1 FRiP (fraction of reads in peaks) analysis

FRiP analysis was carried out using MACS2 (<https://github.com/taoliu/MACS/>) with the *callpeak* function (Zhang et al., 2008) to identify regions of enrichment or peaks across the whole genome. Reads within these peaks were counted and divided by the reads across the entire sorted bam file. A FRiP score of 1.22% was observed, above the 1% threshold for a viable ChIP.

### 5.5.2.2 Visualization

Visualisation of cohesin distribution was performed in R as described in Section 3.6. For comparison purposes the CENP-A ChIPSeq data from Section 5.5 was also shown in Figure 5.18. Since the location of the inner centromere in a 1 dimensional conformation is unknown the alignments are shown in 5Mb windows to discern if any enrichment is observed in regions flanking the CENP-A binding domain.

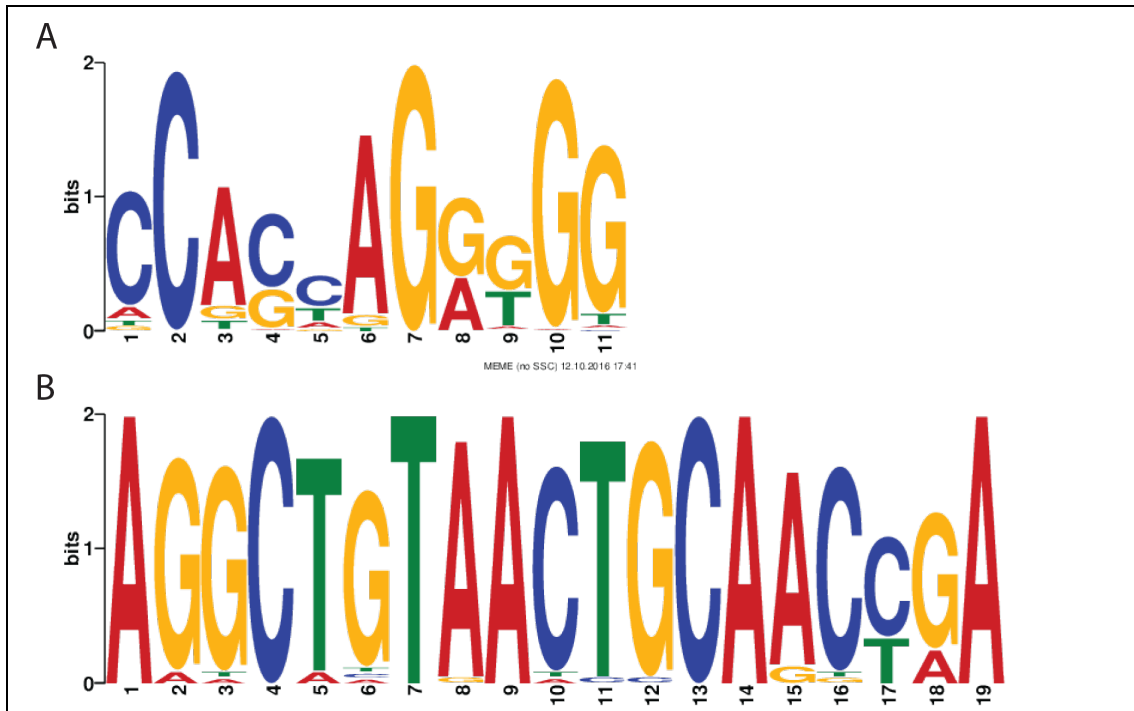
FRiP analysis shows the ChIP was successful, yet no enrichment was seen at the centromere or centromere periphery Figure 5.18, suggesting that the enrichment of cells with a G2/M DNA content (19.50%) in this experiment is insufficient to see distinctive signal at the inner centromere.



**Figure 5.18** 5 MB window view of Smc1 ChIPSeq compared with mitotic CENP-A ChIPSeq (note different scales).

### 5.5.2.3 Motif analysis

To further characterize the Smc1 ChIPSeq motif analysis was carried out. The sorted Smc1 ChIP and Input Bam file were converted to bed format using the *bamtobed* function in *bedtools*. Smc1 peaks were then called using the SISR peak calling software (Site Identification from Short Sequence Reads) (Jothi, Cuddapah, Barski, Cui, & Zhao, 2008; Narlikar & Jothi, 2012). SISR outputted a file containing the Chromosome name, start and end position, NumTags (number of reads supporting the identified binding site), fold enrichment and p value. The DNA sequence between the genomic coordinates identified were extracted and *MEME* (motif based sequence analysis tool) (Bailey et al., 2009) was used to identify motifs within the peaks identified. Using SiSSR 5826 peaks were identified, of these 110 were associated within a 5MB window containing the CENP-A binding domain across 14 of the 16 satellite free domains (no peaks detected at Eca30/Eas30 or EcaX/EasX centromere domain or flanks). Peaks averaged at 38 bp in span.



**Figure 5.19 Motif logos associated with regions of Smc1 binding:** There are 2941 Motif A sites present in the Smc1 binding domains (E-value 8.3e-3768). Motif A is sequence commonly associated with the CCCTC-binding factor (CTCF) and a transcription repressor known to colocalise with cohesin. There are 35 instances of Motif B within the Smc1 associated domains (E-value 7.5e-125) the sequences shares homology with donkey ectodysplasin A2 receptor (EDA2R) mRNA and hydroxysteroid 17 beta dehydrogenase 7 mRNA.

Motif analysis (Figure 5.19) showed 2941 instances (E-value 8.3e-3768) of a motif commonly associated with CTCF (Essien et al., 2009) within the Smc1 binding domains. The CCCTC-binding factor (CTCF) is a highly conserved DNA binding protein, involved in numerous diverse cellular functions: gene activation, gene repression, the maintenance of genomic imprinting, chromatin insulator function and X chromosome inactivation (Filippova, 2007; Ohlsson et al., 2001). In mammals, cohesin is found to accumulate at regions also associated with CTCF binding (Parelho et al., 2008). Cohesin is required for CTCF's insulator activity (Wendt et al., 2008).

The meme output file was examined and 45 motifs corresponding to Figure 5.19 A, were identified in the 5MB window containing the CENP-A binding domain, as shown in Table 5.15. Work from Giulotto et al. (unpublished) show that donkey satellite free centromeres occur in gene deserts, so it is not surprising that none of the motifs identified in Table 5.15 occur in centromere domains. The association of cohesin with this particular motif plays a role in transcription.

Sequence	Chromosome	Coordinates
CCAGAAGAGGG	Eca5	77044751-77044791
ACACTAGAGGG	Eca6	12362871-12363031
CCCCTAGAGGG	Eca6	13313711-13313891
CCAGTAGATGT	Eca6	15805231-15805331
CCACGAGGGGT	Eca8	39854611-39854671
CCACTAGATGG	Eca8	43546771-43546891
TCAGTAGGTGG	Eca8	43955211-43955251
GCAGTAGGTGG	Eca8	43215811-43215851
CCACCAGGGGG	Eca11	44890371-44890411
CCAGCAGGGGG	Eca11	44513531-44513591
CCACGAGGTGG	Eca11	44576271-44576311
CCACAAGATGG	Eca11	45021531-45021631
CCAGCAGGGGG	Eca11	45263971-45264011
CCAGCAGAGGG	Eca11	45927611-45927671
ACAGCAGGGGG	Eca11	45215171-45215231
CCAGCAGGGGG	Eca11	46319191-46319251
CCAGCAGAGGG	Eca11	47328311-47328391
CCAGCAGAGGG	Eca13	6923691-6923731
CCACGAGGGGG	Eca13	7815511-7815571
CCACCAGGCGG	Eca13	7909471-7909511
CCAGCAGGTGG	Eca13	8346011-8346051
ACACTAGAGGG	Eca13	8535631-8535671
CCAGCAGGAGG	Eca13	8594111-8594171
CCGCCAGGGGG	Eca13	8647251-8647291
CCGCCAGGTGG	Eca13	8768331-8768371
CCACCAGGGGG	Eca17	14478851-14478911
CCACCAGGGGG	Eca17	15409051-15409091
CCACTAGATGG	Eca17	19255171-19255211
CCACTAGAGGG	Eca19	2969131-2969171
CCAGCAGGGGG	Eca20	24066731-24066851
CCAGCAGAGGG	Eca20	26197571-26197611
CCAGCAGAGGG	Eca20	26197571-26197611
CCAGAAGAGGG	Eca20	28805911-28805951
GCACTGGAGGG	Eca25	6627391-6627451
CCACTAGATGT	Eca25	7070811-7070851
CCAGAAGGTGG	Eca25	7867171-7867231
CCGCCAGGGGG	Eca26	23290251-23290291
CCAGCAGGGGG	Eca26	23374451-23374491
CCACTAGGAGG	Eca28	10605231-10605271
CCACCAGAGGG	Eca28	11923891-11924011
GCACAAGAAGG	Eca28	15122351-15122431
CAGCCAGGGGG	Eca28	15163811-15163851

CCACAAGGGGG	Eca28	15535551-15535591
ACGCTAGGTGG	Eca28	15538511-15538551

**Table 5.15 Sequences identified from the meme output (Figure 5.19 A) present at 5MB windows containing the unique sequence donkey centromeres**

The second motif identified has homology to equus asinus ectodysplasin A2 receptor mRNA and equus asinus hydroxysteroid (17-beta) dehydrogenase 7 (HSD17B7) mRNA. Ectodysplasin A2 receptor is a member of the tumor necrosis factor family, that activate a range of intracellular signaling pathways (M. Yan et al., 2000) while HSD17B7 is involved in the biosynthesis of sex steroids (Vihko, Isomaa, & Ghosh, 2001) and cholesterol (Marijanovic et al., 2003). Two instances of this motif were identified in the 5MB centromere window of Eca8 and Eca17 as shown in Table 5.16.

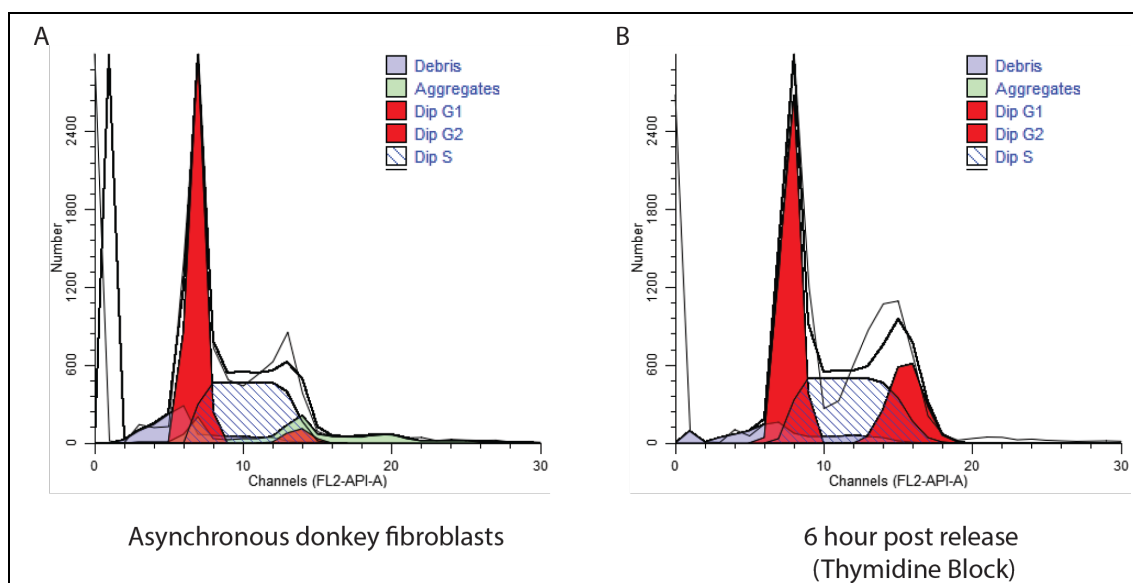
Sequence	Chromosome	Coordinates
AGGCTGTAAGTCAACCAA	Eca8	43546771-43546891
AGACTGTAAGTCAACCGA	Eca17	19255171-19255211

**Table 5.16 Sequences identified from the meme output (Figure 5.19 B) present at 5MB windows containing the unique sequence donkey centromeres**

The identification of these motifs within the mitotically enriched Smc1 datasets, illustrates the levels of whole genomic cohesin binding present in the data. Flow analysis of the population of cells used in this ChIPSeq experiment shows a high proportion of cells with a G1 and S DNA content, therefore it is not surprising that the top motif hit is a CTCF associated motif. Of the 5826 peaks identified none were colocalised with the CENP-A binding domain. A total of 45 motif instances were found in the 5Mb regions surrounding centromeres (Table 5.15). The lack of enhanced cohesin binding at the centromere and centromere periphery shows clear evidence that a pure population of mitotic cells is required to identify and map cohesin associated inner centromere.

### 5.5.3 CPC-Aurora B, Survivin and Borealin

The distribution of cells used for Aurora B and Survivin ChIPSeq experiments are shown in Figure 5.20. The proportion of cells with a G2/M DNA content are 2.16% times that of an asynchronous population. Borealin ChIPSeq was performed on the same population of cells harvested for the mitotically enriched CENP-A ChIPSeq Section 5.5, with a G2/M population 2.59 times greater than an asynchronous population



**Figure 5.20** Distribution of cells in an asynchronous population and the mitotically enriched population used in Aurora B and Survivin ChIPSeq.

	Diploid%	G1	G2/M	S
Asynchronous	100	59.14	6.99	33.88
6 hour release (post Thymidine block)	100	43.51	18.15	38.34

**Table 5.17** Distribution of cells in asynchronous and mitotically enriched populations

The sequence details of the CPC reads are shown below in table 5.18. The sequencing reactions for both the ChIPs and the inputs were run in two lanes L001 and L003. Read quality was examined and reads were aligned and normalized as described previously. The percentage of reads dropped after trimming are shown in Table 5.19.

File name	Input/ChIP	Sequence length (bp)	Reads
14_GATGGAGT_L001_R1_001.fastq	AuroraB/Survivin Input	150	7943195
14_GATGGAGT_L001_R2_001.fastq	AuroraB/Survivin Input	150	7943195
14_GATGGAGT_L003_R1_001.fastq	AuroraB/Survivin Input	125	8273014
14_GATGGAGT_L003_R2_001.fastq	AuroraB/Survivin Input	125	8273014
15_CTAGCTCA_L001_R1_001.fastq	Aurora B ChIP	150	9931412
15_CTAGCTCA_L001_R2_001.fastq	Aurora B ChIP	150	9931412
15_CTAGCTCA_L003_R1_001.fastq	Aurora B ChIP	125	12863435
15_CTAGCTCA_L003_R2_001.fastq	Aurora B ChIP	125	12863435
7_ACGAATCC_L001_R1_001.fastq	Survivin ChIP	150	8305798
7_ACGAATCC_L001_R2_001.fastq	Survivin ChIP	150	8305798
7_ACGAATCC_L003_R1_001.fastq	Survivin ChIP	125	14865905
7_ACGAATCC_L003_R2_001.fastq	Survivin ChIP	125	14865905
6_CACAGGAA_L001_R1_001.fastq	Borealin ChIP	150	9656908
6_CACAGGAA_L001_R2_001.fastq	Borealin ChIP	150	9656908
6_CACAGGAA_L003_R1_001.fastq	Borealin ChIP	125	13782384
6_CACAGGAA_L003_R2_001.fastq	Borealin ChIP	125	13782384

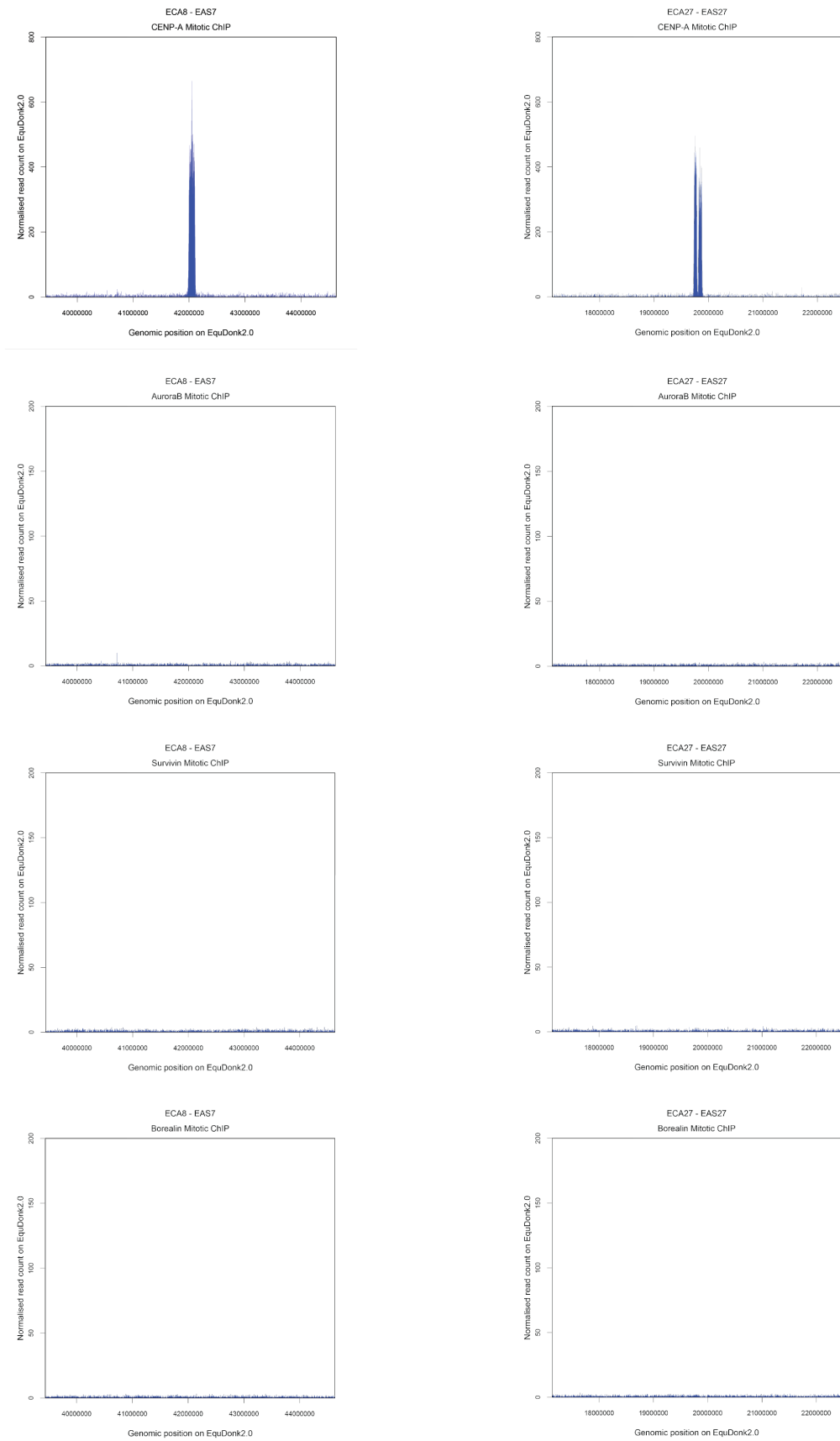
**Table 5.18** Details of reads obtained for ChIPSeq of the CPC subunits

Library	Reads dropped (%)
Mitotic AuroraB/Survivin Input L001	0.63
Mitotic AuroraB/Survivin Input L003	0.91
Mitotic AuroraB ChIP L001	0.74
Mitotic AuroraB ChIP L003	1.02
Mitotic Survivin ChIP L001	0.69
Mitotic Survivin ChIP L003	0.96
Mitotic Borealin ChIP L001	0.64
Mitotic Borealin ChIP L003	0.95

**Table 5.19** Mitotically enriched CPC reads dropped after trimming

### 5.5.3.1 Visualisation of data

To view the distribution of the Aurora B, Borealin and Survivin in the donkey, datasets aligned to EquDonk were normalized using *deeptools* and visualized using the R package *Sushi* as described in Sections 2.4 and 3.6. Figure 5.21 shows a 5Mb window containing the centromere domain. No enrichment can be seen at the centromere or centromere periphery, suggesting similarity to the case for cohesin, a pure mitotic population is required to visualize the inner centromere. To further evaluate the ChIP, FRiP analysis was performed.



**Figure 5.21** 5 Mb window view of CPC ChIPSeq compared with mitotic CENP-A ChIPSeq (note different scale).



### 5.5.3.2 FRiP (fraction of reads in peaks) analysis

As described previously in Section 5.6.1, FRiP was used to measure the enrichment of signal across the genome. FRiP analysis carried out on Aurora B, Borealin and Survivin datasets show values of 0.0173, 0.06158 and 0.0116 respectively, well below the 1% threshold, suggesting that the immunoprecipitations failed. Since the CPC traverses throughout the cell cycle, initial observations from visualization of the alignments suggested that given the proportion of cells with G1 and S DNA content, that the levels of background signal were too high to identify signal associated with the inner centromere.

## 5.6 Discussion

Given the inability to map the inner centromere associated DNA in a one dimensional DNA configuration, it is clear the steps that need to be taken to address pitfalls in the experiments: 1. A pure mitotic population is absolutely necessary to map the inner centromere compartment. 2. ChIP needs to become more efficient requiring fewer cells, making work with mitotic populations feasible. 3. Crosslinking methods need to be optimized, as it is apparent that characterizing immunoprecipitation by western blot alone is insufficient for determining suitability for ChIPSeq.

The thymidine arrest and release protocol needs to be refined, since there is variability in the proportions of cells with G2/M DNA content from harvest to harvest. This is more than likely due to the large-scale drug treatment and subsequent harvesting of many dishes. Given the nature and scale of the procedure timing of harvest can vary slightly from batch to batch depending on how quickly the cells are gathered. To minimize this variability, the ChIP needs to become more efficient requiring fewer cells.

How many cells are required for a ChIPSeq experiment to be successful? Illumina recommends that a minimum of 10ng of DNA be used in sequencing library construction. At approximately 0.1% of the genome, approximately  $2 \times 10^7$  cells are required to have 10ng of centromeric DNA in the experiment. However, good coverage of centromeres can be obtained at 1-4% FRIP as shown in the preceding experiments. At 2% FRIP there would be 0.2ng of centromeric DNA in a 10ng library, corresponding to  $4 \times 10^5$  cells. Since immunoprecipitation is not 100% efficient, this estimate can be revised to  $2 \times 10^6$  cells for a 20% efficient immunoprecipitation. In principle, an immunoprecipitation of this scale should

provide adequate coverage of enriched sequences in the centromere domains under consideration. Obtaining pure mitotic cells at this scale is feasible with fluorescence activated cell sorting.

FRiP analysis showed that the CPC immunoprecipitations failed to yield DNA. Western blot analysis of the immunoprecipitation showed that for all three subunits Figure 5.4 C, 5.5 C and 5.6 C, the protein was immunoprecipitated. Given the antigens were recovered the likely reason for failure to recover DNA is inadequate crosslinking. It is apparent that crosslinking methods need to be validated more rigorously, since the protein is not pulling down DNA. In this chapter, two crosslinking steps were utilized employing amine reactive crosslinkers. Formaldehyde is a commonly used crosslinker that acts through primary amines, crosslinking proteins to DNA and other macromolecules with its 2Å spacer arm (Zeng et al., 2006). In the case of the CPC, where Survivin is associated with the chromatin through a histone modification, a longer spacer arm is required to crosslink the proteins to adjacent DNA. The use of the bifunctional NHS-ester crosslinker, EGS was examined given its spacer arm length of 16.1Å and its published utility in immunoprecipitation of GATA-1 cofactors FOG-1 and MTA-1 where crosslinking with formaldehyde alone failed (Zeng et al., 2006). It is clear from the FRiP scores obtained with the CPC ChIPSeq that this crosslinking method needs to be further examined and optimised. In conjunction with this, a gentler approach when shearing the crosslinked chromatin could be employed, such as micrococcal nuclease digestion. However, in our hands this method of shearing crosslinked cells was extremely inefficient at solubilizing chromatin. As the situation stands, perhaps optimizing conditions and carrying out the cohesin ChIPSeq, since cohesin is directly associated with chromatin, is the most logical method of mapping the inner centromere. In this way the inner centromere domain can be identified and qPCR primers can be identified specific for the region, assuming cohesin and the CPC co-occupy the same domain.

For mapping the CENP-A binding domain at mitosis a pure mitotic population, would provide greater resolution to establish if there was any redistribution or remodeling of chromatin associated with a kinetochore competent centromere. The superimposition and correlative analysis show that there is no significant difference in the centromeres of an asynchronous population and the mitotically enriched population. We postulate despite the high proportion of cells in G1 (30.93%) and S phase (39.00%) that a

significant remodeling would still be discernible. Considering that the “background” in this experiment (G1 + S) is two times the potential signal, a change of at least 2-3 fold in distribution would be required to observe a difference. Clearly acquisition of data from a pure mitotic population would allow direct comparison of mitotic and interphase CENP-A distribution.

## **Chapter 6 - Conclusion**

Centromeres are essential chromosomal loci that direct chromosome segregation during cell division. Despite their highly conserved function, the DNA associated with centromeres is highly variable in evolution. In metazoans and vertebrates in particular centromeres are established on highly repetitive satellite DNA arrays, which hinder detailed molecular analysis of their chromatin organization. As centromere identity in most eukaryotes is determined through the epigenetically controlled assembly of CENP-A, knowledge of the molecular organization of centromeric chromatin is essential for understanding centromere identity and function. In this respect, the distinctive equid system introduced in this thesis provides a novel model system for dissecting the architecture of centromeres in a mammalian organism. The novel contributions of this body of work are 1. Examination of centromere stability during mitotic propagation. Centromeres in the donkey fibroblasts are inheritably stable during prolonged periods of culturing and CENP-A abundance is tightly regulated. 2. Comparison of CENP-A associated domains in two donkey individuals. Identified enrichment of LINEs and AT rich sequences at these domains and showed instances of divergence at these loci between the two individuals.

An equid optimized CENP-A antibody was generated over the course of this work that has application in immunofluorescence, western blot, immunoprecipitation and ChIPSeq. This antibody was used to examine centromere distribution in an immortalized donkey cell line and quantify CENP-A abundance at the unique sequence centromeres in the donkey. Centromeres in the immortalized cell line were found to have a ~20% smaller CENP-A footprint when compared to the primary cell line. Given that the immortalized cell line was derived as a single cell clone from the heterogeneous primary fibroblasts, this shows the tight conservation of centromere position during prolonged culturing. The overall abundance of CENP-A at the unique sequence centromeres in the immortalized cell line showed, a tighter uniformity (std. dev 0.09) when compared to the primary cells (std. dev 0.2). Taken together, these observations are indicative of the founder effect and shows that the centromere position and CENP-A abundance in this immortalized cell line is tightly maintained and regulated. This also indicates a maintenance mechanism independent of DNA sequence (Sullivan et al. 2011) and independent of CENP-B association (Fachinetti et al. 2015).

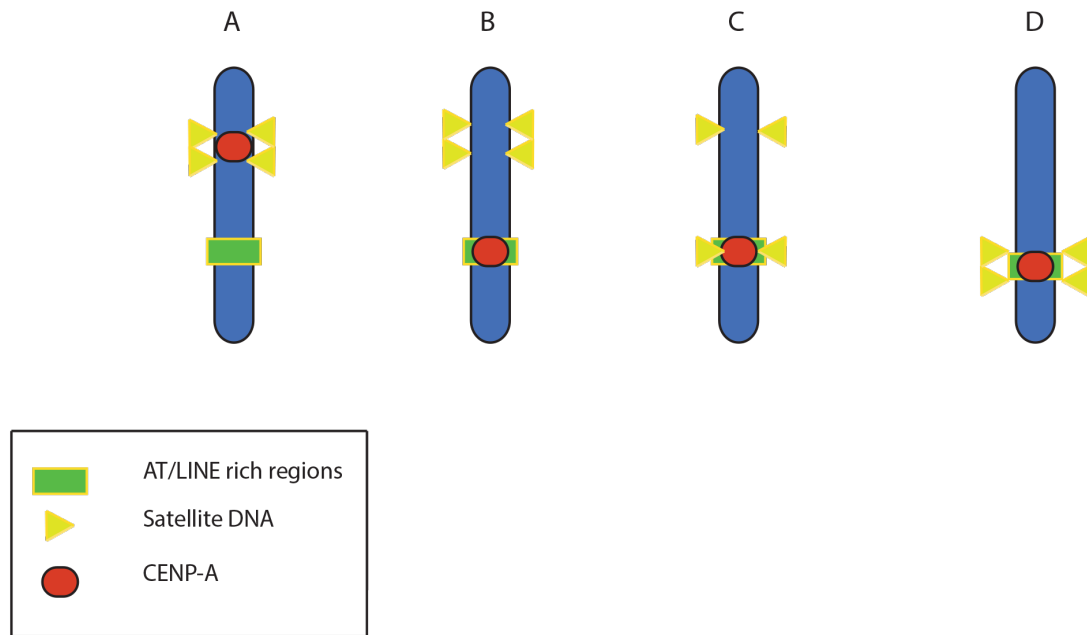
The underlying DNA sequence which CENP-A ChIPSeq reads map to in two donkey individuals, Asino Nuovo (EquDonk) and the Guanzhong donkey as well as the horse was examined. There was an abundance of LINEs at these domains in both the donkeys and the horse. Since the corresponding loci in the horse lack centromere function, we postulate that centromeres arise in these domains instead of driving genomic rearrangement after formation. There was also an absence of full length LINEs in the horse and donkey individuals at these loci, suggesting that active transposition has not driven these sequence changes. While it can only be assumed, due to centromere sliding (Purgato et al. 2015), that the loci in which the CENP-A ChIPSeq reads map to in the Guanzhong donkey are functional centromeres, there are number of notable difference between the individuals, particularly at the Eas8, Eas9, Eas13 and Eas19 centromere domains. There is evidence for genomic amplification in the EquDonk assembly as single copy sequences are duplicated at the Eas9 and Eas19 centromere domains when compared to the corresponding loci in the Guanzhong donkey and the horse. This could potentially be the early steps in “maturation” of the Asino Nuovo centromere whereby it accumulates “repetitive sequences” (Piras et al. 2010). In the case of the Eas8 centromere domain, there is large genomic rearrangement in the Gunazhong donkey when compared to the corresponding loci in Asino Nuovo and the horse. At the Eas13 centromere domain, there is a large deletion in Asino Nuovo (> 100kb) when compared to the Guanzhong donkey and the horse. There are also a number of sequences unique to the horse when compared to two donkey individuals. While clear divergence can be seen across the three individuals in this study, in order to discern whether sequences are deletions in the horse genome or insertions in the donkey a comprehensive comparison with other equid species is required.

The first steps in identifying the linear one-dimensional primary DNA structure of the inner centromere have been addressed in this body of work. Although the experiments presented here did not allow identification of inner centromere-associated DNA sequences, key requirements for successfully completing this approach has been identified: the need for a pure mitotic population, a more efficient ChIP method and efficient crosslinking of proteins indirectly associated with DNA. To fully discern if there is any redistribution or remodeling of CENP-A associated with the kinetochore competent centromere, as with the inner centromere mapping, a pure mitotic population would be required. While the correlative analysis of the mitotically

enriched CENP-A ChIP and the asynchronous ChIP showed no significant difference between either dataset, we estimate that given the levels of background in the experiment (G1 + S) a minimum of a 2-3 fold change in apparent abundance would be required to overcome the ‘background’.

Taken together, the results presented in this thesis provide a basis for further detailed molecular analysis of mammalian centromeres to provide both a physical map of centromere protein distribution as well as a quantitative framework for analysis of CENP-A regulation.

A critical understanding of what drives centromere formation and relocation is vital for understanding the remarkably fast speciation of the equids as well as the development of new treatments of diseases such as cancer. Chromosome instability, as a result of aberrant centromere and kinetochore function, can result in chromosome missegregation, which in turn leads to aneuploidy and chromosomal rearrangements. Neocentromere formation has been observed in both lipomatous tumours and acute myeloid leukemia and is likely to be found in other tumors but this has not been published due to the infrequent karyotyping of solid tumors (Amor & Choo 2002). In many cancers, HJURP and CENP-A are overexpressed and are beginning to be used as prognostic markers (Tomonaga et al. 2003; Montes de Oca et al. 2015). Centromere function is conserved across eukaryotes yet is surprisingly fluid on an evolutionary timescale, occupying different DNA sequences and chromosomal loci between closely related species. A possible mechanism for neocentromere formation as deduced from analysis of the CENP-A binding domains in the EquDonk assembly is shown in Figure 6.1. Our findings fit the hypothesis that the original centromere function is altered or compromised and that this drives neocentromere formation at a new chromosomal loci rich in AT and LINE sequences. Satellite sequences gradually accumulate at the new centromere through duplication of existing sequences. Satellite sequences at the old centromere are gradually lost due to the absence of selection pressure.



**Figure 6.1 Neocentromere formation in the equids.** The centromere is associated with satellite sequences (A). Centromere function is altered resulting in movement to a more favourable chromosomal loci rich in AT and LINE sequences (B). The neocentromere gradually accumulates satellite sequences while satellite sequences at the old centromere are lost (C)(D). Adapted from Amor & Choo 2002.

The presence of naturally occurring unique sequence centromeres in the equids allows for experimental manipulation to provide a clearer understanding of what defines a centromere and drives CENP-A deposition. The chromosomal architecture of the mammalian centromere can be investigated using chromatin conformation capture methods to identify how the centromere chromatin fiber is organized and gain a clearer insight into the functional architecture of the centromere. Genome editing in conjunction with CENP-A ChIPSeq will allow for examination of the centromere in response to genomic DNA alterations. The introduction or deletion of sequences at the CENP-A binding domain will allow for investigation of CENP-A redistribution. Unique sequence centromeres could also be targeted with full length LINEs or satellite sequences to observe the molecular response to various DNA substrates of the centromeres. These approaches will allow for a deeper understanding of centromere function and identity.

## Chapter 7 References

- Ainsztein, A. M., Kandels-Lewis, S. E., Mackay, A. M., & Earnshaw, W. C. (1998a). INCENP centromere and spindle targeting: Identification of essential conserved motifs and involvement of heterochromatin protein HP1. *Journal of Cell Biology*. <http://doi.org/10.1083/jcb.143.7.1763>
- Ainsztein, A. M., Kandels-Lewis, S. E., Mackay, A. M., & Earnshaw, W. C. (1998b). INCENP centromere and spindle targeting: Identification of essential conserved motifs and involvement of heterochromatin protein HP1. *Journal of Cell Biology*, 143(7), 1763–1774. <http://doi.org/10.1083/jcb.143.7.1763>
- Alonso, A., Hasson, D., Cheung, F., & Warburton, P. E. (2010). A paucity of heterochromatin at functional human neocentromeres. *Epigenetics & Chromatin*, 3, 6. <http://doi.org/10.1186/1756-8935-3-6>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J. Mol. Biol.*, (215), 403–410. [http://doi.org/10.1016/S0022-2836\(05\)80360-2](http://doi.org/10.1016/S0022-2836(05)80360-2)
- Amor, D. J., Bentley, K., Ryan, J., Perry, J., Wong, L., Slater, H., & Choo, K. H. A. (2004). Human centromere repositioning “in progress”. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 6542–6547. <http://doi.org/10.1073/pnas.0308637101>
- Arumugam, P., Gruber, S., Tanaka, K., Haering, C. H., Mechtler, K., & Nasmyth, K. (2003). ATP Hydrolysis Is Required for Cohesin’s Association with Chromosomes. *Current Biology*, 13(22), 1941–1953. <http://doi.org/10.1016/j.cub.2003.10.036>
- Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., ... Noble, W. S. (2009). MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research*, 37(SUPPL. 2). <http://doi.org/10.1093/nar/gkp335>
- Banerjee, A. R., Kim, Y. J. ung, & Kim, T. H. oon. (2014). A novel virus-inducible enhancer of the interferon-?? gene with tightly linked promoter and enhancer activities. *Nucleic Acids Research*, 42(20), 12537–12554. <http://doi.org/10.1093/nar/gku1018>
- Bassett, E. A., Wood, S., Salimian, K. J., Ajith, S., Foltz, D. R., & Black, B. E. (2010). Epigenetic centromere specification directs aurora B accumulation but is insufficient to efficiently correct mitotic errors. *Journal of Cell Biology*, 190, 177–185. <http://doi.org/10.1083/jcb.201001035>
- Bernard, P., Drogat, J., Maure, J. F., Dheur, S., Vaur, S., Genier, S., & Javerzat, J. P. (2006). A Screen for Cohesion Mutants Uncovers Ssl3, the Fission Yeast Counterpart of the Cohesin Loading Factor Scc4. *Current Biology*, 16(9), 875–881. <http://doi.org/10.1016/j.cub.2006.03.037>
- Bilgimol, J. C., Suthakaran, P., Sankaranarayanan, S., Musti, M., Kalimuthu, S., Ganesan, M., & Sadananda, R. M. (2015). An overview of the parameters for recombinant protein expression in Escherichia coli. *Cell Science & Therapy*, 6(5), 1–7. <http://doi.org/10.4172/2217-7013.1000221>
- Bishop, J. D., & Schuniacher, J. M. (2002). Phosphorylation of the carboxyl terminus of inner centromere protein (INCENP) by the Aurora B kinase stimulates Aurora B kinase activity. *Journal of Biological Chemistry*. <http://doi.org/10.1074/jbc.C200307200>
- Black, B. E., Foltz, D. R., Chakravarthy, S., Luger, K., Woods, V. L., & Cleveland, D. W. (2004). Structural determinants for generating centromeric chromatin. *Nature*, 430(6999), 578–582. <http://doi.org/10.1038/nature02766>
- Black, B. E., Jansen, L. E. T., Maddox, P. S., Foltz, D. R., Desai, A. B., Shah, J. V., &



- Cleveland, D. W. (2007). Centromere Identity Maintained by Nucleosomes Assembled with Histone H3 Containing the CENP-A Targeting Domain. *Molecular Cell*, 25(2), 309–322. <http://doi.org/10.1016/j.molcel.2006.12.018>
- Blajeski, A. L., Phan, V. A., Kottke, T. J., & Kaufmann, S. H. (2002). G1 and G2 cell-cycle arrest following microtubule depolymerization in human breast cancer cells. *Journal of Clinical Investigation*, 110(1), 91–99. <http://doi.org/10.1172/JCI200213275>
- Blat, Y., & Kleckner, N. (1999). Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell*, 98(2), 249–259. [http://doi.org/10.1016/S0092-8674\(00\)81019-3](http://doi.org/10.1016/S0092-8674(00)81019-3)
- Blower, M. D. (2016). Centromeric Transcription Regulates Aurora-B Localization and Activation. *Cell Reports*, 15(8), 1624–1633. <http://doi.org/10.1016/j.celrep.2016.04.054>
- Bodor, D. L., Mata, J. F., Sergeev, M., David, A. F., Salimian, K. J., Panchenko, T., ... Jansen, L. E. T. (2014). The quantitative architecture of centromeric chromatin. *eLife*, 2014(3). <http://doi.org/10.7554/eLife.02137>
- Bootsma, D., Budke, L., & Vos, O. (1964). Studies on Synchronous Division of Tissue Culture Cells Initiated By Excess Thymidine. *Experimental Cell Research*, 33(1–2), 301–309. [http://doi.org/10.1016/S0014-4827\(64\)81035-1](http://doi.org/10.1016/S0014-4827(64)81035-1)
- Brito, D. A., Yang, Z., & Rieder, C. L. (2008). Microtubules do not promote mitotic slippage when the spindle assembly checkpoint cannot be satisfied. *Journal of Cell Biology*, 182(4), 623–629. <http://doi.org/10.1083/jcb.200805072>
- Buheitel, J., & Stemann, O. (2013). Prophase pathway-dependent removal of cohesin from human chromosomes requires opening of the Smc3-Scc1 gate. *The EMBO Journal*, 32(5), 666–76. <http://doi.org/10.1038/emboj.2013.7>
- Carbone, L., Nergadze, S. G., Magnani, E., Misceo, D., Francesca Cardone, M., Roberto, R., ... Giulotto, E. (2006). Evolutionary movement of centromeres in horse, donkey, and zebra. *Genomics*, 87, 777–782. <http://doi.org/10.1016/j.ygeno.2005.11.012>
- Carmena, M., Ruchaud, S., Earnshaw, W. C., Building, M. S., & Road, M. (2009). Making the Aurora glow: regulation of Aurora A and B kinase function by interacting proteins. *Current Opinion in Cell Biology*, 21(6), 796–805.
- Carmena, M., Wheelock, M., Funabiki, H., & Earnshaw, W. C. (2012). The chromosomal passenger complex (CPC): from easy rider to the godfather of mitosis. *Nature Reviews. Molecular Cell Biology*, 13, 789–803. <http://doi.org/10.1038/nrm3474>
- Carnell, A. N., & Goodman, J. I. (2003). The long (LINEs) and the short (SINEs) of it: Altered methylation as a precursor to toxicity. *Toxicological Sciences*. <http://doi.org/10.1093/toxsci/kfg138>
- Carroll, C. W., Silva, M. C. C., Godek, K. M., Jansen, L. E. T., & Straight, A. F. (2009). Centromere assembly requires the direct recognition of CENP-A nucleosomes by CENP-N. *Nature Cell Biology*, 11(7), 896–902. <http://doi.org/10.1038/ncb1899>
- Casola, C., Hucks, D., & Feschotte, C. (2008). Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. *Molecular Biology and Evolution*, 25(1), 29–41. <http://doi.org/10.1093/molbev/msm221>
- Chandele, A., Prasad, V., Jagtap, J. C., Shukla, R., & Shastry, P. R. (2004). Upregulation of Survivin in G2/M Cells and Inhibition of Caspase 9 Activity Enhances Resistance in Staurosporine-Induced Apoptosis. *Neoplasia (New York,*

- N.Y.*), 6(1), 29–40.
- Chen, J., Jin, S., Tahir, S. K., Zhang, H., Liu, X., Sarthy, A. V., ... Ng, S. C. (2003). Survivin enhances aurora-B kinase activity and localizes aurora-B in human cells. *Journal of Biological Chemistry*, 278(1), 486–490. <http://doi.org/10.1074/jbc.M211119200>
- Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C. R., Gu, M., ... Jiang, J. (2002). Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *The Plant Cell*, 14(8), 1691–1704. <http://doi.org/10.1105/tpc.003079>
- Chu, Y., Yao, P. Y., Wang, W., Wang, D., Wang, Z., Zhang, L., ... Yao, X. (2011). Aurora B kinase activation requires survivin priming phosphorylation by PLK1. *Journal of Molecular Cell Biology*. <http://doi.org/10.1093/jmcb/mjq037>
- Chueh, A. C., Northrop, E. L., Brettingham-Moore, K. H., Choo, K. H. A., & Wong, L. H. (2009). LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. *PLoS Genetics*, 5. <http://doi.org/10.1371/journal.pgen.1000354>
- Cleveland, D. W., Mao, Y., & Sullivan, K. F. (2003). Centromeres and kinetochores: From epigenetics to mitotic checkpoint signaling. *Cell*. [http://doi.org/10.1016/S0092-8674\(03\)00115-6](http://doi.org/10.1016/S0092-8674(03)00115-6)
- Connell, C. M., Colnaghi, R., & Wheatley, S. P. (2008). Nuclear survivin has reduced stability and is not cytoprotective. *Journal of Biological Chemistry*, 283(6), 3289–3296. <http://doi.org/10.1074/jbc.M704461200>
- Cooke, C. a., Heck, M. M. S., & Earnshaw, W. C. (1987). The inner centromere protein (INCENP) antigens: Movement from inner centromere to midbody during mitosis. *Journal of Cell Biology*. <http://doi.org/10.1083/jcb.105.5.2053>
- Cost, G. J., Feng, Q., Jacquier, A., & Boeke, J. D. (2002). Human L1 element target-primed reverse transcription in vitro. *EMBO Journal*, 21(21), 5899–5910. <http://doi.org/10.1093/emboj/cdf592>
- Craig, J. M., Wong, L. H., Lo, A. W. I., Earle, E., & Choo, K. H. A. (2003). Centromeric chromatin pliability and memory at a human neocentromere. *EMBO Journal*, 22(10), 2495–2504. <http://doi.org/10.1093/emboj/cdg232>
- Dai, J., Sullivan, B. A., & Higgins, J. M. G. (2006). Regulation of Mitotic Chromosome Cohesion by Haspin and Aurora B. *Developmental Cell*, 11(5), 741–750. <http://doi.org/10.1016/j.devcel.2006.09.018>
- De Rop, V., Padeganeh, A., & Maddox, P. S. (2012). CENP-A: The key player behind centromere identity, propagation, and kinetochore assembly. *Chromosoma*, 121(6), 527–538. <http://doi.org/10.1007/s00412-012-0386-5>
- Deveraux, Q. L., & Reed, J. C. (1999). IAP family proteins - Suppressors of apoptosis. *Genes and Development*, 13(3), 239–252. <http://doi.org/10.1101/gad.13.3.239>
- Dohi, T., Beltrami, E., Wall, N. R., Plescia, J., & Altieri, D. C. (2004). Mitochondrial survivin inhibits apoptosis and promotes tumorigenesis. *Journal of Clinical Investigation*, 114(8), 1117–1127. <http://doi.org/10.1172/JCI200422222>
- Downen, J. M., Fan, Z. P., Hnisz, D., Ren, G., Abraham, B. J., Zhang, L. N., ... Young, R. A. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, 159(2), 374–387. <http://doi.org/10.1016/j.cell.2014.09.030>
- Du, C., Fang, M., Li, Y., Li, L., & Wang, X. (2000). Smac, a mitochondrial protein that promotes cytochrome c-dependent caspase activation by eliminating IAP inhibition. *Cell*, 102(1), 33–42. [http://doi.org/10.1016/S0092-8674\(00\)00008-8](http://doi.org/10.1016/S0092-8674(00)00008-8)

- Dunleavy, E. M., Almouzni, G., & Karpen, G. H. (2011). H3.3 is deposited at centromeres in S phase as a placeholder for newly assembled CENP-A in G<sub>1</sub> phase. *Nucleus (Austin, Tex.)*, 2(2), 146–157. <http://doi.org/10.4161/nucl.2.2.15211>
- Earnshaw, W. C., & Rothfield, N. (1985). Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. *Chromosoma*, 91, 313–321. <http://doi.org/10.1007/BF00328227>
- Echeverry, M. C., Bot, C., Obado, S. O., Taylor, M. C., & Kelly, J. M. (2012). Centromere-associated repeat arrays on *Trypanosoma brucei* chromosomes are much more extensive than predicted. *BMC Genomics*, 13(1), 29. <http://doi.org/10.1186/1471-2164-13-29>
- Eichler, E. E. (1999). Repetitive conundrums of centromere structure and function. *Human Molecular Genetics*. <http://doi.org/10.1093/hmg/8.2.151>
- Essien, K., Vigneau, S., Apreleva, S., Singh, L. N., Bartolomei, M. S., & Hannenhalli, S. (2009). CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome Biology*, 10(11), R131. <http://doi.org/10.1186/gb-2009-10-11-r131>
- Fachinetti, D., Han, J. S., McMahan, M. A., Ly, P., Abdullah, A., Wong, A. J., & Cleveland, D. W. (2015). DNA Sequence-Specific Binding of CENP-B Enhances the Fidelity of Human Centromere Function. *Developmental Cell*, 33(3), 314–327. <http://doi.org/10.1016/j.devcel.2015.03.020>
- Fang, G., Yu, H., & Kirschner, M. W. (1998). Direct binding of CDC20 protein family members activates the anaphase-promoting complex in mitosis and G<sub>1</sub>. *Molecular Cell*, 2(2), 163–171. [http://doi.org/10.1016/S1097-2765\(00\)80126-4](http://doi.org/10.1016/S1097-2765(00)80126-4)
- Filippova, G. N. (2007). Genetics and Epigenetics of the Multifunctional Protein CTCF. *Current Topics in Developmental Biology*, 80(7), 337–360. [http://doi.org/10.1016/S0070-2153\(07\)80009-3](http://doi.org/10.1016/S0070-2153(07)80009-3)
- Fischle, W., Tseng, B. S., Dormann, H. L., Ueberheide, B. M., Garcia, B. a, Shabanowitz, J., ... Allis, C. D. (2005). Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature*, 438(7071), 1116–1122. <http://doi.org/10.1038/nature04219>
- Fitzgerald-Hayes, M., Clarke, L., & Carbon, J. (1982). Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs. *Cell*, 29(1), 235–244. [http://doi.org/10.1016/0092-8674\(82\)90108-8](http://doi.org/10.1016/0092-8674(82)90108-8)
- Flemming, W. (1882). Zellsubstanz, kern und zelltheilung. *F.C.W. Vogel, Leipzig*, 419. Retrieved from <http://books.google.com/books?id=ndYcngEACAAJ&pgis=1\nhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Zellsubstanz,+kern+und+zelltheilung#0>
- Folco, H. D., Campbell, C. S., May, K. M., Espinoza, C. A., Oegema, K., Hardwick, K. G., ... Desai, A. (2015). The CENP-A N-tail confers epigenetic stability to centromeres via the CENP-T branch of the CCAN in fission yeast. *Current Biology*, 25(3), 348–356. <http://doi.org/10.1016/j.cub.2014.11.060>
- Foltz, D. R., Jansen, L. E. T., Black, B. E., Bailey, A. O., Yates, J. R., & Cleveland, D. W. (2006). The human CENP-A centromeric nucleosome-associated complex. *Nature Cell Biology*, 8(5), 458–69. <http://doi.org/10.1038/ncb1397>
- Fortugno, P., Wall, N. R., Giodini, A., O'Connor, D. S., Plescia, J., Padgett, K. M., ... Altieri, D. C. (2002). Survivin exists in immunochemically distinct subcellular pools and is involved in spindle microtubule function. *Journal of Cell Science*, 115(Pt 3), 575–585.

- Fukagawa, T., & Earnshaw, W. C. (2014). The centromere: Chromatin foundation for the kinetochore machinery. *Developmental Cell*.  
<http://doi.org/10.1016/j.devcel.2014.08.016>
- Fukuda, S., & Pelus, L. M. (2006). Survivin, a cancer target with an emerging role in normal adult tissues. *Molecular Cancer Therapeutics*, 5(5), 1087–1098.  
<http://doi.org/10.1158/1535-7163.MCT-05-0375>
- Gassmann, R., Carvalho, A., Henzing, A. J., Ruchaud, S., Hudson, D. F., Honda, R., ... Earnshaw, W. C. (2004). Borealin: A novel chromosomal passenger required for stability of the bipolar mitotic spindle. *Journal of Cell Biology*, 166(2), 179–191. <http://doi.org/10.1083/jcb.200404001>
- Giménez-Abián, J. F., Sumara, I., Hirota, T., Hauf, S., Gerlich, D., De La Torre, C., ... Peters, J. M. (2004). Regulation of sister chromatid cohesion between chromosome arms. *Current Biology*, 14(13), 1187–1193.  
<http://doi.org/10.1016/j.cub.2004.06.052>
- Glover, D. M., Leibowitz, M. H., Mclean, D. A., & Parry, H. (2014). 1-S2.0-0092867495903747-Main, 81, 1–11. Retrieved from papers://4c48fe13-7c69-4b65-a0fa-467dcdedc3f/Paper/p16767
- Glynn, E. F., Megee, P. C., Yu, H. G., Mistrot, C., Unal, E., Koshland, D. E., ... Gerton, J. L. (2004). Genome-wide mapping of the cohesin complex in the yeast *Saccharomyces cerevisiae*. *PLoS Biology*, 2(9).  
<http://doi.org/10.1371/journal.pbio.0020259>
- Gruber, S., Arumugam, P., Katou, Y., Kuglitsch, D., Helmhart, W., Shirahige, K., & Nasmyth, K. (2006). Evidence that Loading of Cohesin Onto Chromosomes Involves Opening of Its SMC Hinge. *Cell*, 127(3), 523–537.  
<http://doi.org/10.1016/j.cell.2006.08.048>
- Gruneberg, U., Neef, R., Honda, R., Nigg, E. A., & Barr, F. A. (2004). Relocation of Aurora B from centromeres to the central spindle at the metaphase to anaphase transition requires MKlp2. *Journal of Cell Biology*, 166(2), 167–172.  
<http://doi.org/10.1083/jcb.200403084>
- Haarhuis, J. H. I., Elbatsh, A. M. O., & Rowland, B. D. (2014). Cohesin and Its Regulation: On the Logic of X-Shaped Chromosomes. *Developmental Cell*, 31(1), 7–18. <http://doi.org/10.1016/j.devcel.2014.09.010>
- Han, Z., Riefler, G. M., Saam, J. R., Mango, S. E., & Schumacher, J. M. (2005). The *C. elegans* Tousled-like kinase contributes to chromosome segregation as a substrate and regulator of the Aurora B kinase. *Current Biology*, 15(10), 894–904. <http://doi.org/10.1016/j.cub.2005.04.019>
- Hara, K., Zheng, G., Qu, Q., Liu, H., Ouyang, Z., Chen, Z., ... Yu, H. (2014). Structure of cohesin subcomplex pinpoints direct shugoshin-Wapl antagonism in centromeric cohesion. *Nature Structural & Molecular Biology*, 21(10), 864–870.  
<http://doi.org/10.1038/nsmb.2880>
- Haren, L., Stearns, T., & Lüders, J. (2009). Plk1-dependent recruitment of  $\gamma$ -tubulin complexes to mitotic centrosomes involves multiple PCM components. *PLoS ONE*, 4(6). <http://doi.org/10.1371/journal.pone.0005976>
- Hartsink-Segers, S. A., Exalto, C., Allen, M., Williamson, D., Clifford, S. C., Horstmann, M., ... Den Boer, M. L. (2013). Inhibiting Polo-like kinase 1 causes growth reduction and apoptosis in pediatric acute lymphoblastic leukemia cells. *Haematologica*, 98(10), 1539–1546.  
<http://doi.org/10.3324/haematol.2013.084434>
- Hasson, D., Panchenko, T., Salimian, K. J., Salman, M. U., Sekulic, N., Alonso, A., ... Black, B. E. (2013). The octamer is the major form of CENP-A nucleosomes

- at human centromeres. *Nature Structural & Molecular Biology*, 20(6), 687–95. <http://doi.org/10.1038/nsmb.2562>
- Hauf, S., Roitinger, E., Koch, B., Dittrich, C. M., Mechtler, K., & Peters, J. M. (2005). Dissociation of cohesin from chromosome arms and loss of arm cohesion during early mitosis depends on phosphorylation of SA2. *PLoS Biology*, 3(3), 0419–0432. <http://doi.org/10.1371/journal.pbio.0030069>
- Hauf, S., Waizenegger, I. C., & Peters, J. M. (2001). Cohesin cleavage by separase required for anaphase and cytokinesis in human cells. *Science (New York, N.Y.)*, 293(5533), 1320–1323. <http://doi.org/10.1126/science.1061376>
- Hayashi-Takanaka, Y., Yamagata, K., Nozaki, N., & Kimura, H. (2009). Visualizing histone modifications in living cells: Spatiotemporal dynamics of H3 phosphorylation during interphase. *Journal of Cell Biology*, 187(6), 781–790. <http://doi.org/10.1083/jcb.200904137>
- Henikoff, S. (2001). The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA. *Science*, 293(5532), 1098–1102. <http://doi.org/10.1126/science.1062939>
- Henikoff, S., Ahmad, K., & Malik, H. S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. *Science (New York, N.Y.)*, 293(5532), 1098–1102. <http://doi.org/10.1126/science.1062939>
- Henikoff, S., & Henikoff, J. G. (2012). “Point” Centromeres of *Saccharomyces* Harbor Single Centromere-Specific Nucleosomes. *Genetics*, 190(4), 1575–1577. <http://doi.org/10.1534/genetics.111.137711>
- Hooser, A. A. Van, Ouspenski, I. I., Gregson, H. C., Starr, D. A., Yen, T. J., Goldberg, M. L., ... Brinkley, B. R. (2001). Specification of kinetochore-forming chromatin by the histone H3 variant CENP-A. *Journal of Cell Science*, 114(Pt 19), 3529–3542. Retrieved from <http://jcs.biologists.org/content/114/19/3529.long>
- Hori, T., Amano, M., Suzuki, A., Backer, C. B., Welburn, J. P., Dong, Y., ... Fukagawa, T. (2008). CCAN Makes Multiple Contacts with Centromeric DNA to Provide Distinct Pathways to the Outer Kinetochore. *Cell*, 135(6), 1039–1052. <http://doi.org/10.1016/j.cell.2008.10.019>
- Huang, J., Zhao, Y., Bai, D., Shiraigol, W., Li, B., Yang, L., ... Dugarjaviin, M. (2015). Donkey genome and insight into the imprinting of fast karyotype evolution. *Scientific Reports*, 5, 14106. <http://doi.org/10.1038/srep14106>
- Hümmer, S., & Mayer, T. U. (2009). Cdk1 Negatively Regulates Midzone Localization of the Mitotic Kinesin Mklp2 and the Chromosomal Passenger Complex. *Current Biology*, 19(7), 607–612. <http://doi.org/10.1016/j.cub.2009.02.046>
- Ikui, A. E., Chia-Ping, H. Y., Matsumoto, T., & Horwitz, S. B. (2005). Low concentrations of taxol cause mitotic delay followed by premature dissociation of p55CDC from Mad2 and BubR1 and abrogation of the spindle checkpoint, leading to aneuploidy. *Cell Cycle*, 4(10), 1385–1388. <http://doi.org/2061> [pii]
- Irvine, D. V., Amor, D. J., Perry, J., Sirvent, N., Pedoutour, F., Choo, K. H. A., & Saffery, R. (2005). Chromosome size and origin as determinants of the level of CENP-A incorporation into human centromeres. *Chromosome Research*, 12(8), 805–815. <http://doi.org/10.1007/s10577-005-5377-4>
- Izuta, H., Ikeno, M., Suzuki, N., Tomonaga, T., Nozaki, N., Obuse, C., ... Yoda, K. (2006). Comprehensive analysis of the ICEN (Interphase Centromere Complex) components enriched in the CENP-A chromatin of human cells. *Genes to Cells*, 11(6), 673–684. <http://doi.org/10.1111/j.1365-2443.2006.00969.x>

- Jansen, L. E. T., Black, B. E., Foltz, D. R., & Cleveland, D. W. (2007). Propagation of centromeric chromatin requires exit from mitosis. *Journal of Cell Biology*, *176*(6), 795–805. <http://doi.org/10.1083/jcb.200701066>
- Jelluma, N., Dansen, T. B., Sliedrecht, T., Kwiatkowski, N. P., & Kops, G. J. P. L. (2010). Release of Mps1 from kinetochores is crucial for timely anaphase onset. *Journal of Cell Biology*, *191*(2), 281–290. <http://doi.org/10.1083/jcb.201003038>
- Jeyaprasath, A. A., Klein, U. R., Lindner, D., Ebert, J., Nigg, E. A., & Conti, E. (2007). Structure of a Survivin-Borealin-INCENP Core Complex Reveals How Chromosomal Passengers Travel Together. *Cell*, *131*(2), 271–285. <http://doi.org/10.1016/j.cell.2007.07.045>
- Jiang, J., Birchler, J. A., Parrott, W. A., & Dawe, R. K. (2003). A molecular view of plant centromeres. *Trends in Plant Science*. <http://doi.org/10.1016/j.tplants.2003.10.011>
- Jin, W., Melo, J. R., Nagaki, K., Talbert, P. B., Henikoff, S., Dawe, R. K., & Jiang, J. (2004). Maize centromeres: organization and functional adaptation in the genetic background of oat. *The Plant Cell*, *16*(3), 571–581. <http://doi.org/10.1105/tpc.018937>
- Jothi, R., Cuddapah, S., Barski, A., Cui, K., & Zhao, K. (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Research*, *36*(16), 5221–5231. <http://doi.org/10.1093/nar/gkn488>
- Kalitsis, P., & Choo, K. H. A. (2012). The evolutionary life cycle of the resilient centromere. *Chromosoma*. <http://doi.org/10.1007/s00412-012-0369-6>
- Kaminker, J., Bergman, C., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., ... Celniker, S. (2002). The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biology*, *3*(12), 1–20. <http://doi.org/10.1186/gb-2002-3-12-research0084>
- Kane, J. F., & Hartley, D. L. (1988). Formation of recombinant protein inclusion bodies in *Escherichia coli*. *Trends in Biotechnology*, *6*(5), 95–101. [http://doi.org/10.1016/0167-7799\(88\)90065-0](http://doi.org/10.1016/0167-7799(88)90065-0)
- Kang, J., Chaudhary, J., Dong, H., Kim, S., Brautigam, C. a, & Yu, H. (2011). Mitotic centromeric targeting of HP1 and its binding to Sgo1 are dispensable for sister-chromatid cohesion in human cells. *Molecular Biology of the Cell*. <http://doi.org/10.1091/mbc.E11-01-0009>
- Kapellos, G., Polonifi, K., Farmakis, D., Spartalis, E., Tomos, P., Aessopos, A., ... Mantzourani, M. (2013). Overexpression of Survivin Levels in Circulation and Tissue Samples of Lung Cancer Patients, *3480*, 3475–3480.
- Kapitonov, V. V., & Jurka, J. (1999). Molecular paleontology of transposable elements from *Arabidopsis thaliana*. *Genetica*, *107*(1–3), 27–37. <http://doi.org/10.1023/A:1004030922447>
- Kawashima, S. a, Yamagishi, Y., Honda, T., Ishiguro, K., & Watanabe, Y. (2010). Phosphorylation of H2A by Bub1 prevents chromosomal instability through localizing shugoshin. *Science (New York, N.Y.)*, *327*(5962), 172–177. <http://doi.org/10.1126/science.1180189>
- Kelly, A. E., Ghenoiu, C., Xue, J. Z., Zierhut, C., Kimura, H., & Funabiki, H. (2010). Survivin Reads Phosphorylated Histone H3 Threonine 3 to Activate the Mitotic Kinase Aurora B. *Science*, *330*(6001), 235–239. <http://doi.org/10.1126/science.1189505>
- Kevin F. Sullivan, Mirko Hechenberger, K. M. (1994). Human CENP-A Contains a Histone H3 Related Histone Fold Domain That Is Required for Targeting to the Centromere. *The Journal of Cell Biology*, *127*(3), 581–592.

- <http://doi.org/10.1083/jcb.127.3.581>
- Kipling, D., & Warburton, P. E. (1997). Centromeres, CENP-B and Tigger too. *Trends in Genetics*. [http://doi.org/10.1016/S0168-9525\(97\)01098-6](http://doi.org/10.1016/S0168-9525(97)01098-6)
- Kitagawa, M., & Lee, S. H. (2015). The chromosomal passenger complex (CPC) as a key orchestrator of orderly mitotic exit and cytokinesis. *Frontiers in Cell and Developmental Biology*, 3(March), 14. <http://doi.org/10.3389/fcell.2015.00014>
- Kitajima, T. S., Sakuno, T., Ishiguro, K., Iemura, S., Natsume, T., Kawashima, S. a., & Watanabe, Y. (2006). Shugoshin collaborates with protein phosphatase 2A to protect cohesin. *Nature*, 441(7089), 46–52. <http://doi.org/10.1038/nature04663>
- Kiyomitsu, T., Iwasaki, O., Obuse, C., & Yanagida, M. (2010). Inner centromere formation requires hMis14, a trident kinetochore protein that specifically recruits HP1 to human chromosomes. *Journal of Cell Biology*, 188(6), 791–807. <http://doi.org/10.1083/jcb.200908096>
- Klein, U. R., Nigg, E. A., & Gruneberg, U. (2006). Centromere targeting of the chromosomal passenger complex requires a ternary subcomplex of Borealin, Survivin, and the N-terminal domain of INCENP. *Molecular Biology of the Cell*, 17, 2547–2558. <http://doi.org/10.1091/mbc.E05-12-1133>
- Kogut, I., Wang, J., Guacci, V., Mistry, R. K., & Megee, P. C. (2009). The Scc2/Scc4 cohesin loader determines the distribution of cohesin on budding yeast chromosomes. *Genes and Development*. <http://doi.org/10.1101/gad.1819409>
- Kueng, S., Hegemann, B., Peters, B. H., Lipp, J. J., Schleiffer, A., Mechtler, K., & Peters, J. M. (2006). Wapl Controls the Dynamic Association of Cohesin with Chromatin. *Cell*, 127(5), 955–967. <http://doi.org/10.1016/j.cell.2006.09.040>
- Kunitoku, N., Sasayama, T., Marumoto, T., Zhang, D., Honda, S., Kobayashi, O., ... Hirota, T. (2003). CENP-A phosphorylation by Aurora-A in prophase is required for enrichment of Aurora-B at inner centromeres and for kinetochore function. *Developmental Cell*, 5(6), 853–864. [http://doi.org/10.1016/S1534-5807\(03\)00364-2](http://doi.org/10.1016/S1534-5807(03)00364-2)
- Lamb, J. C., & Birchler, J. a. (2003). The role of DNA sequence in centromere formation. *Genome Biology*, 4, 214. <http://doi.org/10.1186/gb-2003-4-5-214>
- Langdon, T., Seago, C., Mende, M., Leggett, M., Thomas, H., Forster, J. W., ... Jenkins, G. (2000). Retrotransposon evolution in diverse plant genomes. *Genetics*, 156(1), 313–325.
- Lengronne, A., Katou, Y., Mori, S., Yokobayashi, S., Kelly, G. P., Itoh, T., ... Uhlmann, F. (2004). Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature*, 430(6999), 573–578. <http://doi.org/10.1038/nature02742>
- Liu, H., Rankin, S., & Yu, H. (2013). Phosphorylation-enabled binding of SGO1-PP2A to cohesin protects sororin and centromeric cohesion during mitosis. *Nature Cell Biology*, 15(1), 40–9. <http://doi.org/10.1038/ncb2637>
- Liu, X., Song, Z., Huo, Y., Zhang, J., Zhu, T., Wang, J., ... Yao, X. (2014). Chromatin protein HP1?? interacts with the mitotic regulator borealin protein and specifies the centromere localization of the chromosomal passenger complex. *Journal of Biological Chemistry*, 289(30), 20638–20649. <http://doi.org/10.1074/jbc.M114.572842>
- Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V, Muzny, D. M., ... Wilson, R. K. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, 469(7331), 529–33. <http://doi.org/10.1038/nature09687>
- Logsdon, G. A., Barrey, E. J., Bassett, E. A., DeNizio, J. E., Guo, L. Y., Panchenko,

- T., ... Black, B. E. (2015). Both tails and the centromere targeting domain of CENP-A are required for centromere establishment. *Journal of Cell Biology*, 208(5), 521–531. <http://doi.org/10.1083/jcb.201412011>
- Losada, A., Hirano, M., and Hirano, T. (1998). Identification of Protein Complexes Required for Efficient Sister Chromatid Cohesion. *Genes Dev*, 12. <http://doi.org/10.1091/mbc.E03-08-0619>
- Losada, A. (2014). Cohesin in cancer: chromosome segregation and beyond. *Nature Reviews. Cancer*, 14(6), 389–93. <http://doi.org/10.1038/nrc3743>
- Lu, B., Mahmud, H., Maass, A. H., Yu, B., van Gilst, W. H., de Boer, R. A., & Silljé, H. H. W. (2010). The Plk1 Inhibitor BI 2536 temporarily arrests primary cardiac fibroblasts in mitosis and generates aneuploidy In Vitro. *PLoS ONE*, 5(9). <http://doi.org/10.1371/journal.pone.0012963>
- Marijanovic, Z., Laubner, D., Moller, G., Gege, C., Husen, B., Adamski, J., & Breitling, R. (2003). Closing the gap: identification of human 3-ketosteroid reductase, the last unknown enzyme of mammalian cholesterol biosynthesis. *Molecular Endocrinology (Baltimore, Md.)*, 17(9), 1715–1725. <http://doi.org/10.1210/me.2002-0436>
- Marshall, O. J., Chueh, A. C., Wong, L. H., & Choo, K. H. A. (2008). Neocentromeres: New Insights into Centromere Structure, Disease Development, and Karyotype Evolution. *American Journal of Human Genetics*. <http://doi.org/10.1016/j.ajhg.2007.11.009>
- Marshall, O. J., Marshall, A. T., & Choo, K. H. A. (2008). Three-dimensional localization of CENP-A suggests a complex higher order structure of centromeric chromatin. *Journal of Cell Biology*, 183, 1193–1202. <http://doi.org/10.1083/jcb.200804078>
- McAinsh, A. D., & Meraldi, P. (2011). The CCAN complex: Linking centromere specification to control of kinetochore-microtubule dynamics. *Seminars in Cell and Developmental Biology*. <http://doi.org/10.1016/j.semcdb.2011.09.016>
- Melters, D. P., Bradnam, K. R., Young, H. a, Telis, N., May, M. R., Ruby, J. G., ... Chan, S. W. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, 14(1), R10. <http://doi.org/10.1186/gb-2013-14-1-r10>
- Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics*. <http://doi.org/10.1016/j.tig.2007.02.006>
- Mirchenko, L., & Uhlmann, F. (2010). Sli15INCENP dephosphorylation prevents mitotic checkpoint reengagement due to loss of tension at anaphase onset. *Current Biology*, 20(15), 1396–1401. <http://doi.org/10.1016/j.cub.2010.06.023>
- Misulovin, Z., Schwartz, Y. B., Li, X. Y., Kahn, T. G., Gause, M., MacArthur, S., ... Dorsett, D. (2008). Association of cohesin and Nipped-B with transcriptionally active regions of the *Drosophila melanogaster* genome. *Chromosoma*, 117(1), 89–102. <http://doi.org/10.1007/s00412-007-0129-1>
- Moroi, Y., Peebles, C., Fritzler, M. J., Steigerwald, J., & Tan, E. M. (1980). Autoantibody to centromere (kinetochore) in scleroderma sera. *Proceedings of the National Academy of Sciences of the United States of America*, 77(3), 1627–31. <http://doi.org/10.1073/pnas.77.3.1627>
- Mukaka, M. (2012). Statistic Corner A guide to appropriate use of Correlation coefficient in medical research.pdf. *Malawi Medical Journal*, 24(3)(September), 69–71. <http://doi.org/10.1016/j.cmpb.2016.01.020>
- Nakajima, Y., Cormier, A., Tyers, R. G., Pigula, A., Peng, Y., Drubin, D. G., &



- Barnes, G. (2011). Ipl1/Aurora-dependent phosphorylation of Sli15/INCENP regulates CPC-spindle interaction to ensure proper microtubule dynamics. *Journal of Cell Biology*. <http://doi.org/10.1083/jcb.201009137>
- Nakajima, Y., Tyers, R. G., Wong, C. C. L., Yates, J. R., Drubin, D. G., & Barnes, G. (2009). Nbl1p: a Borealin/Dasra/CSC-1-like protein essential for Aurora/Ipl1 complex function and integrity in *Saccharomyces cerevisiae*. *Molecular Biology of the Cell*. <http://doi.org/10.1091/mbc.E08-10-1011>
- Narlikar, L., & Jothi, R. (2012). ChIP-Seq data analysis: Identification of Protein-DNA Binding Sites with SISSRs Peak-Finder. *Methods in Molecular Biology*, 802, 305–322. [http://doi.org/10.1007/978-1-61779-400-1\\_20](http://doi.org/10.1007/978-1-61779-400-1_20)
- Nishiyama, T., Sykora, M. M., Huis in 't Veld, P. J., Mechtler, K., & Peters, J.-M. (2013). Aurora B and Cdk1 mediate Wapl activation and release of acetylated cohesin from chromosomes by phosphorylating Sororin. *Proceedings of the National Academy of Sciences of the United States of America*, 110(33), 13404–9. <http://doi.org/10.1073/pnas.1305020110>
- Nitiss, J. L. (2009). DNA topoisomerase II and its growing repertoire of biological functions. *Nature Reviews. Cancer*, 9(5), 327–337. <http://doi.org/10.1038/nrc2608>
- Ogden, S., Haggerty, D., Stoner, C. M., Kolodrubetz, D., & Schleif, R. (1980). The *Escherichia coli* L-arabinose operon: binding sites of the regulatory proteins and a mechanism of positive and negative regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 77(6), 3346–50. <http://doi.org/10.1073/pnas.77.6.3346>
- Ohlsson, R., Renkawitz, R., & Lobanenko, V. (2001). CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends in Genetics*, 17(9), 520–527. [http://doi.org/10.1016/S0168-9525\(01\)02366-6](http://doi.org/10.1016/S0168-9525(01)02366-6)
- Ohta, S., Bukowski-Wills, J. C., Sanchez-Pulido, L., Alves, F. de L., Wood, L., Chen, Z. A., ... Rappsilber, J. (2010). The Protein Composition of Mitotic Chromosomes Determined Using Multiclassifier Combinatorial Proteomics. *Cell*, 142(5), 810–821. <http://doi.org/10.1016/j.cell.2010.07.047>
- Ohzeki, J., Ichirou, Nakano, M., Okada, T., & Masumoto, H. (2002). CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. *Journal of Cell Biology*, 159(5), 765–775. <http://doi.org/10.1083/jcb.200207112>
- Okada, M., Cheeseman, I. M., Hori, T., Okawa, K., McLeod, I. X., Yates, J. R., ... Fukagawa, T. (2006). The CENP-H-I complex is required for the efficient incorporation of newly synthesized CENP-A into centromeres. *Nature Cell Biology*, 8(5), 446–57. <http://doi.org/10.1038/ncb1396>
- Orlando, L. (2015). Equids. *Current Biology*, 25(20), R973–R978. <http://doi.org/10.1016/j.cub.2015.09.005>
- Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., ... Willerslev, E. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, 499(7456), 74–8. <http://doi.org/10.1038/nature12323>
- Ostertag, E. M., & Kazazian, H. H. (2001). Biology of Mammalian L1 Retrotransposons. *Annual Review of Genetics*, 35, 501–538. <http://doi.org/10.1146/annurev.genet.35.102401.091032>
- Palmer, D. K., O'Day, K., Wener, M. H., Andrews, B. S., & Margolis, R. L. (1987). A 17-kD centromere protein (CENP-A) copurifies with nucleosome core particles and with histones. *Journal of Cell Biology*, 104, 805–815. <http://doi.org/10.1083/jcb.104.4.805>

- Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., ... Merckenschlager, M. (2008). Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. *Cell*, *132*(3), 422–433. <http://doi.org/10.1016/j.cell.2008.01.011>
- Peters, J. M., Tedeschi, A., & Schmitz, J. (2008). The cohesin complex and its roles in chromosome biology. *Genes and Development*. <http://doi.org/10.1101/gad.1724308>
- Petsalaki, E., Akoumianaki, T., Black, E. J., Gillespie, D. A. F., & Zachos, G. (2011). Phosphorylation at serine 331 is required for Aurora B activation. *Journal of Cell Biology*, *195*(3), 449–466. <http://doi.org/10.1083/jcb.201104023>
- Piras, F. M., Nergadze, S. G., Magnani, E., Bertoni, L., Attolini, C., Khoriauli, L., ... Giulotto, E. (2010). Uncoupling of satellite DNA and centromeric function in the genus *Equus*. *PLoS Genetics*, *6*. <http://doi.org/10.1371/journal.pgen.1000845>
- Pray, L. (2008). Transposons: The jumping genes. *Nature Education*, *1*(1), 204. Retrieved from <http://www.nature.com/scitable/topicpage/transposons-the-jumping-genes-518>
- Purgato, S., Belloni, E., Piras, F. M., Zoli, M., Badiale, C., Cerutti, F., ... Giulotto, E. (2015). Centromere sliding on a mammalian chromosome. *Chromosoma*, *124*(2), 277–287. <http://doi.org/10.1007/s00412-014-0493-6>
- Qian, J., Beullens, M., Lesage, B., & Bollen, M. (2013). Aurora B defines its own chromosomal targeting by opposing the recruitment of the phosphatase scaffold Repo-Man. *Current Biology*, *23*(12), 1136–1143. <http://doi.org/10.1016/j.cub.2013.05.017>
- Qian, J., Lesage, B., Beullens, M., Van Eynde, A., & Bollen, M. (2011). PP1/repo-man dephosphorylates mitotic histone H3 at T3 and regulates chromosomal aurora B targeting. *Current Biology*, *21*(9), 766–773. <http://doi.org/10.1016/j.cub.2011.03.047>
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. <http://doi.org/10.1038/nbt.1754>
- Rudd, M. K., & Willard, H. F. (2004). Analysis of the centromeric regions of the human genome assembly. *Trends Genet*, *20*, 529–533. Retrieved from [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list\\_uids=15475110](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15475110)
- Saffery, R., Irvine, D. V., Griffiths, B., Kalitsis, P., Wordeman, L., & Choo, K. H. (2000). Human centromeres and neocentromeres show identical distribution patterns of >20 functionally important kinetochore-associated proteins. *Human Molecular Genetics*, *9*(2), 175–185. <http://doi.org/10.1093/hmg/9.2.175>
- Samejima, I., Spanos, C., De Lima Alves, F., Hori, T., Perpelescu, M., Zou, J., ... Earnshaw, W. C. (2015). Whole-proteome genetic analysis of dependencies in assembly of a vertebrate kinetochore. *Journal of Cell Biology*, *211*(6), 1141–1156. <http://doi.org/10.1083/jcb.201508072>
- Sampath, S. C., Ohi, R., Leismann, O., Salic, A., Pozniakovski, A., & Funabiki, H. (2004). The chromosomal passenger complex is required for chromatin-induced microtubule stabilization and spindle assembly. *Cell*, *118*, 187–202. <http://doi.org/10.1016/j.cell.2004.06.026>
- Sasai, K., Katayama, H., Stenoién, D. L., Fujii, S., Honda, R., Kimura, M., ... Sen, S. (2004). Aurora-C kinase is a novel chromosomal passenger protein that can complement Aurora-B kinase function in mitotic cells. *Cell Motility and the Cytoskeleton*, *59*(4), 249–263. <http://doi.org/10.1002/cm.20039>

- Schleif, R. (1992). DNA looping. *Annual Review of Biochemistry*, *61*, 199–223.  
<http://doi.org/10.1146/annurev.bi.61.070192.001215>
- Schleif, R. (2010). AraC protein, regulation of the l-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action. *FEMS Microbiology Reviews*. <http://doi.org/10.1111/j.1574-6976.2010.00226.x>
- Schmitz, J. (2012). SINEs as driving forces in genome evolution. In *Repetitive DNA* (pp. 92–107). <http://doi.org/10.1159/000337117>
- Schneider, K. L., Xie, Z., Wolfgruber, T. K., & Presting, G. G. (2016). Inbreeding drives maize centromere evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(8), E987–96.  
<http://doi.org/10.1073/pnas.1522008113>
- Seitan, V. C., Banks, P., Laval, S., Majid, N. a., Dorsett, D., Rana, A., ... Strachan, T. (2006). Metazoan Scc4 homologs link sister chromatid cohesion to cell and axon migration guidance. *PLoS Biology*, *4*(8), 1411–1425.  
<http://doi.org/10.1371/journal.pbio.0040242>
- Sessa, F., Mapelli, M., Ciferri, C., Tarricone, C., Areces, L. B., Schneider, T. R., ... Musacchio, A. (2005). Mechanism of Aurora B activation by INCENP and inhibition by hesperadin. *Molecular Cell*, *18*(3), 379–391.  
<http://doi.org/10.1016/j.molcel.2005.03.031>
- Shang, W. H., Hori, T., Martins, N. M. C., Toyoda, A., Misu, S., Monma, N., ... Fukagawa, T. (2013). Chromosome Engineering Allows the Efficient Isolation of Vertebrate Neocentromeres. *Developmental Cell*, *24*(6), 635–648.  
<http://doi.org/10.1016/j.devcel.2013.02.009>
- Shelby, R. D., Monier, K., & Sullivan, K. F. (2000). Chromatin assembly at kinetochores is uncoupled from DNA replication. *Journal of Cell Biology*, *151*(5), 1113–1118. <http://doi.org/10.1083/jcb.151.5.1113>
- Shelby, R. D., Vafa, O., & Sullivan, K. F. (1997). Assembly of CENP-A into centromeric chromatin requires a cooperative array of nucleosomal DNA contact sites. *Journal of Cell Biology*, *136*(3), 501–513.  
<http://doi.org/10.1083/jcb.136.3.501>
- Shi, J., Wolf, S. E., Burke, J. M., Presting, G. G., Ross-Ibarra, J., & Dawe, R. K. (2010). Widespread gene conversion in centromere cores. *PLoS Biology*, *8*(3).  
<http://doi.org/10.1371/journal.pbio.1000327>
- Shuaib, M., Ouararhni, K., Dimitrov, S., & Hamiche, A. (2010). HJURP binds CENP-A via a highly conserved N-terminal domain and mediates its deposition at centromeres. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(4), 1349–54. <http://doi.org/10.1073/pnas.0913709107>
- Singh, S. M., & Panda, A. K. (2005). Solubilization and refolding of bacterial inclusion body proteins. *Journal of Bioscience and Bioengineering*, *99*(4), 303–310. <http://doi.org/10.1263/jbb.99.303>
- Smit, A., Hubley, R., & Green, P. (2013). RepeatMasker Open-4.0. 2013-2015 .  
<Http://www.repeatmasker.org>. Retrieved from <http://repeatmasker.org>
- Steiner, C. C., & Ryder, O. A. (2011). Molecular phylogeny and evolution of the Perissodactyla. *Zoological Journal of the Linnean Society*, *163*(4), 1289–1303.  
<http://doi.org/10.1111/j.1096-3642.2011.00752.x>
- Studier, F. W. (2005). Protein production by auto-induction in high-density shaking cultures. *Protein Expression and Purification*, *41*(1), 207–234.  
<http://doi.org/10.1016/j.pep.2005.01.016>
- Sullivan, L. L., Boivin, C. D., Mravinac, B., Song, I. Y., & Sullivan, B. A. (2011). Genomic size of CENP-A domain is proportional to total alpha satellite array

- size at human centromeres and expands in cancer cells. *Chromosome Research*, 19(4), 457–470. <http://doi.org/10.1007/s10577-011-9208-5>
- Sun, X., Le, H. D., Wahlstrom, J. M., & Karpen, G. H. (2003). Sequence analysis of a functional *Drosophila* centromere. *Genome Research*, 13(2), 182–194. <http://doi.org/10.1101/gr.681703>
- Takeuchi, K., Nishino, T., Mayanagi, K., Horikoshi, N., Osakabe, A., Tachiwana, H., ... Fukagawa, T. (2014). The centromeric nucleosome-like CENP-T-W-S-X complex induces positive supercoils into DNA. *Nucleic Acids Research*, 42(3), 1644–1655. <http://doi.org/10.1093/nar/gkt1124>
- Tamm, I., Wang, Y., Sausville, E., Scudiero, D. A., Vigna, N., Oltersdorf, T., & Reed, J. C. (1998). IAP-family protein Survivin inhibits caspase activity and apoptosis induced by Fas (CD95), bax, caspases, and anticancer drugs. *Cancer Research*, 58(23), 5315–5320.
- Tanaka, K. (2013). Regulatory mechanisms of kinetochore-microtubule interaction in mitosis. *Cellular and Molecular Life Sciences*. <http://doi.org/10.1007/s00018-012-1057-7>
- Tang, Z., Sun, Y., Harley, S. E., Zou, H., & Yu, H. (2004). Human Bub1 protects centromeric sister-chromatid cohesion through Shugoshin during mitosis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(52), 18012–18017. <http://doi.org/10.1073/pnas.0408600102>
- Tanno, Y., Susumu, H., Kawamura, M., Sugimura, H., Honda, T., & Watanabe, Y. (2015). The inner centromere – shugoshin network prevents chromosomal instability. *Science*, 349(6253), 1237–1241. <http://doi.org/10.1126/science.aaa2655>
- Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2), 178–192. <http://doi.org/10.1093/bib/bbs017>
- Tsukahara, T., Tanno, Y., & Watanabe, Y. (2010). Phosphorylation of the CPC by Cdk1 promotes chromosome bi-orientation. *Nature*, 467(7316), 719–723. <http://doi.org/10.1038/nature09390>
- Vader, G., Kauw, J. J. W., Medema, R. H., & Lens, S. M. a. (2006). Survivin mediates targeting of the chromosomal passenger complex to the centromere and midbody. *EMBO Reports*, 7(1), 85–92. <http://doi.org/10.1038/sj.embor.7400562>
- Vassilev, L. T. (2006). Cell cycle synchronization at the G2/M phase border by reversible inhibition of CDK1. *Cell Cycle*. <http://doi.org/10.4161/cc.5.22.3463>
- Ventura, M., Antonacci, F., Cardone, M. F., Stanyon, R., D’Addabbo, P., Cellamare, A., ... Rocchi, M. (2007). Evolutionary formation of new centromeres in macaque. *Science (New York, N.Y.)*, 316(5822), 243–246. <http://doi.org/10.1126/science.1140615>
- Ventura, M., Weigl, S., Carbone, L., Cardone, M. F., Misceo, D., Teti, M., ... Rocchi, M. (2004). Recurrent sites for new centromere seeding. *Genome Research*, 14, 1696–1703. <http://doi.org/10.1101/gr.2608804>
- Verhagen, a M., Ekert, P. G., Pakusch, M., Silke, J., Connolly, L. M., Reid, G. E., ... Vaux, D. L. (2000). Identification of DIABLO, a mammalian protein that promotes apoptosis by binding to and antagonizing IAP proteins. *Cell*, 102(1), 43–53. [http://doi.org/10.1016/S0092-8674\(00\)00009-X](http://doi.org/10.1016/S0092-8674(00)00009-X)
- Vidale, P., Magnani, E., Nergadze, S. G., Santagostino, M., Cristofari, G., Smirnova, A., ... Giulotto, E. (2012). The catalytic and the RNA subunits of human telomerase are required to immortalize equid primary fibroblasts. *Chromosoma*, 121(5), 475–488. <http://doi.org/10.1007/s00412-012-0379-4>

- Vihko, P., Isomaa, V., & Ghosh, D. (2001). Structure and function of 17beta-hydroxysteroid dehydrogenase type 1 and type 2. *Molecular and Cellular Endocrinology*, *171*, 71–76.
- Vong, Q. P., Cao, K., Li, H. Y., Iglesias, P. A., & Zheng, Y. (2016). Chromosome Alignment and Segregation Regulated by Ubiquitination of Survivin Author(s): Queenie P. Vong, Kan Cao, Hoi Y. Li, Pablo A. Iglesias and Yixian Zheng Source:, *310*(5753), 1499–1504.
- Voullaire, L. E., Slater, H. R., Petrovic, V., & Choo, K. H. (1993). A functional marker centromere with no detectable alpha-satellite, satellite III, or CENP-B protein: activation of a latent centromere? *American Journal of Human Genetics*, *52*, 1153–1163.
- Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., ... Lindblad-Toh, K. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science (New York, N.Y.)*, *326*, 865–867. <http://doi.org/10.1126/science.1178158>
- Waizenegger, I. C., Hauf, S., Meinke, a, & Peters, J. M. (2000). Two distinct pathways remove mammalian cohesin from chromosome arms in prophase and from centromeres in anaphase. *Cell*, *103*(3), 399–410. [http://doi.org/10.1016/S0092-8674\(00\)00132-X](http://doi.org/10.1016/S0092-8674(00)00132-X)
- Wang, F., Ulyanova, N. P., Van Der Waal, M. S., Patnaik, D., Lens, S. M. A., & Higgins, J. M. G. (2011). A positive feedback loop involving haspin and aurora B promotes CPC accumulation at centromeres in mitosis. *Current Biology*, *21*(12), 1061–1069. <http://doi.org/10.1016/j.cub.2011.05.016>
- Wang, L. H.-C., Mayer, B., Stemmann, O., & Nigg, E. A. (2010). Centromere DNA decatenation depends on cohesin removal and is required for mammalian cell division. *Journal of Cell Science*, *123*, 806–813. <http://doi.org/10.1242/jcs.058255>
- Warburton, P. E., Cooke, C. A., Bourassa, S., Vafa, O., Sullivan, B. A., Stetten, G., ... Earnshaw, W. C. (1997). Immunolocalization of CENP-A suggests a distinct nucleosome structure at the inner kinetochore plate of active centromeres. *Current Biology : CB*, *7*(11), 901–904. [http://doi.org/10.1016/S0960-9822\(06\)00382-4](http://doi.org/10.1016/S0960-9822(06)00382-4)
- Watanabe, Y., & Kitajima, T. S. (2005). Shugoshin protects cohesin complexes at centromeres. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1455), 515–521, discussion 521. <http://doi.org/10.1098/rstb.2004.1607>
- Weaver, B. A. (2014). How Taxol/paclitaxel kills cancer cells. *Molecular Biology of the Cell*, *25*(18), 2677–81. <http://doi.org/10.1091/mbc.E14-04-0916>
- Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., ... Peters, J.-M. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, *451*(7180), 796–801. <http://doi.org/10.1038/nature06634>
- Wolfgruber, T. K., Sharma, A., Schneider, K. L., Albert, P. S., Koo, D. H., Shi, J., ... Presting, G. G. (2009). Maize centromere structure and evolution: Sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. *PLoS Genetics*, *5*(11). <http://doi.org/10.1371/journal.pgen.1000743>
- Wong, L. H., & Choo, K. H. A. (2004). Evolutionary dynamics of transposable elements at the centromere. *Trends in Genetics*. <http://doi.org/10.1016/j.tig.2004.09.011>
- Xu, Z., Ogawa, H., Vagnarelli, P., Bergmann, J. H., Hudson, D. F., Ruchaud, S., ...

- Samejima, K. (2009). INCENP-aurora B interactions modulate kinase activity and chromosome passenger complex localization. *Journal of Cell Biology*, *187*, 637–653. <http://doi.org/10.1083/jcb.200906053>
- Yamagishi, Y., Honda, T., Tanno, Y., & Watanabe, Y. (2010). Two histone marks establish the inner centromere and chromosome bi-orientation. *Science (New York, N.Y.)*, *330*, 239–243. <http://doi.org/10.1126/science.1194498>
- Yan, H., & Jiang, J. (2007). Rice as a model for centromere and heterochromatin research. *Chromosome Research*, *15*(1), 77–84. <http://doi.org/10.1007/s10577-006-1104-z>
- Yan, M., Wang, L. C., Hymowitz, S. G., Schilbach, S., Lee, J., Goddard, a, ... Dixit, V. M. (2000). Two-amino acid molecular switch in an epithelial morphogen that regulates binding to two distinct receptors. *Science (New York, N.Y.)*, *290*(5491), 523–527. <http://doi.org/10.1126/science.290.5491.523>
- Yang, F., Fu, B., O'Brien, P. C. M., Robinson, T. J., Ryder, O. A., & Ferguson-Smith, M. A. (2003). Karyotypic relationships of horses and zebras: Results of cross-species chromosome painting. *Cytogenetic and Genome Research*, *102*(1–4), 235–243. <http://doi.org/10.1159/000075755>
- Yue, Z., Carvalho, A., Xu, Z., Yuan, X., Cardinale, S., Ribeiro, S., ... Earnshaw, W. C. (2008). Deconstructing Survivin: Comprehensive genetic analysis of Survivin function by conditional knockout in a vertebrate cell line. *Journal of Cell Biology*, *183*(2), 279–296. <http://doi.org/10.1083/jcb.200806118>
- Zeitlin, S. G., Shelby, R. D., & Sullivan, K. F. (2001). CENP-A is phosphorylated by Aurora B kinase and plays an unexpected role in completion of cytokinesis. *Journal of Cell Biology*, *155*(7), 1147–1157. <http://doi.org/10.1083/jcb.200108125>
- Zeng, P. Y., Vakoc, C. R., Chen, Z. C., Blobel, G. A., & Berger, S. L. (2006). In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation. *BioTechniques*, *41*(6), 694–698. <http://doi.org/10.2144/000112297>
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., ... Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, *9*, R137. <http://doi.org/10.1186/gb-2008-9-9-r137>
- Zhou, J., Yao, J., & Joshi, H. C. (2002). Attachment and tension in the spindle assembly checkpoint. *Journal of Cell Science*, *115*(Pt 18), 3547–3555. <http://doi.org/10.1242/jcs.00029>
- Zlotina, A., Galkina, S., Krasikova, A., Crooijmans, R. P. M. A., Groenen, M. A. M., Gaginskaya, E., & Deryusheva, S. (2012). Centromere positions in chicken and Japanese quail chromosomes: De novo centromere formation versus pericentric inversions. *Chromosome Research*, *20*(8), 1017–1032. <http://doi.org/10.1007/s10577-012-9319-7>

## Chapter 8 Appendices

### *Appendix I*

Analysis of domains, which the donkey ChIPSeq reads map to on the EquCab genome.

#### *Repetitive elements across the EAS4 centromere orthologous domain in the horse*

The abundance of SINEs in the horse domain corresponding to the EAS4 centromere was 1.52% less than that observed across the whole genome. LINEs associated with this region, particularly L1 elements higher than whole genome levels. LTR elements remained similar to whole genome levels and an increase in TcMar-Tigger elements was observed at this domain in EquCab, 4% higher than whole genome levels.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	15	2248	1.03	3.53
ALUs	0	0	0	0
MIRs	14	2186	1	3.49
<b>LINEs:</b>	63	59393	27.16	21.59
LINE1	46	52830	24.16	16.25
LINE2	15	6078	2.78	4.66
L3/CR1	2	485	0.22	0.48
<b>LTR elements:</b>	36	15860	7.25	6.29
ERV1	12	4432	2.03	2.11
ERV1-MaLRs	13	4520	2.07	2.62
ERV_classI	10	6812	3.12	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	17	3211	1.47	3.67
hAT-Charlie	8	1025	0.47	1.87
TcMar-Tigger	2	358	0.16	0.9

**Table 8.1** Repetitive elements across the entire horse orthologous region of EAS4 compared with whole genome levels

#### *Repetitive elements across the EAS5 centromere orthologous domain in the horse*

There was a 2.5% decrease in the amount of SINEs present at this loci in the horse compared with the whole genome. The overall abundance of LINEs increased from 21.59% observed across the entire genome to 27.16%, notably L1 elements which increased by 7.91% while L2 elements dropped by 1.88% to 2.78% at this domain. LTR elements were comparable with whole genome levels, with the exception of ERVclassI which was almost 2% higher. hAT-Charlie and TcMar-Tigger levels were less than half that of whole genome levels

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	15	2248	1.03	3.53
ALUs	0	0	0	0
MIRs	14	2186	1	3.49
<b>LINEs:</b>	63	59393	27.16	21.59
LINE1	46	52830	24.16	16.25
LINE2	15	6078	2.78	4.66
L3/CR1	2	485	0.22	0.48
<b>LTR elements:</b>	36	15860	7.25	6.29
ERV_L	12	4432	2.03	2.11
ERV_L-MaLRs	13	4520	2.07	2.62
ERV_classI	10	6812	3.12	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	17	3211	1.47	3.67
hAT-Charlie	8	1025	0.47	1.87
TcMar-Tigger	2	358	0.16	0.9

**Table 8.2** Repetitive elements across the entire horse orthologous region of EAS5 compared with whole genome levels

*Repetitive elements across the EAS7 centromere orthologous domain in the horse*

Examination of repetitive elements in the EAS7 centromere horse orthologous region versus the whole genome showed a 1.56% decrease in SINEs. There was an increase in the overall LINE abundance by 17.33%, particularly L1 elements which increased by 16.69%, more than doubling whole genome levels (16.25%). LTR element abundance remained almost constant with whole genome levels with a minor increase of 0.28% at this domain. hAT-Charlie and TcMar-Tigger levels decreased by 0.27% and 0.22% respectively.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	23	3203	1.97	3.53
ALUs	0	0	0	0
MIRs	22	3157	1.95	3.49
<b>LINEs:</b>	67	63123	38.92	21.59
LINE1	43	53421	32.94	16.25
LINE2	21	8667	5.34	4.66
L3/CR1	3	1035	0.64	0.48
<b>LTR elements:</b>	19	10661	6.57	6.29
ERV_L	8	3043	1.88	2.11
ERV_L-MaLRs	9	3332	2.05	2.62
ERV_classI	2	4286	2.64	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	17	3766	2.32	3.67
hAT-Charlie	12	2600	1.6	1.87
TcMar-Tigger	4	1109	0.68	0.9

**Table 8.3** Repetitive elements across the entire horse orthologous region of EAS7 compared with whole genome levels



*Repetitive elements across the EAS8 centromere orthologous domain in the horse*

Repeatmasker analysis of the domain corresponding to EAS8 centromere in the horse showed a 2% reduction in SINEs compared to whole genome levels. Overall LINE abundance increased by 10.67%, while L2 elements (dropped by 1.47%) were less than whole genome levels, L1 elements increased by 12.82%. There was an overall reduction in LTR element abundance by 2.14% while hAT-Charlie (-1.58%) and TcMar-Tigger (-0.67%) levels were also decreased.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	12	2128	1.53	3.53
ALUs	0	0	0	0
MIRs	12	2128	1.53	3.49
<b>LINEs:</b>	45	44814	32.26	21.59
LINE1	32	40381	29.07	16.25
LINE2	13	4433	3.19	4.66
L3/CR1	0	0	0	0.48
<b>LTR elements:</b>	13	5760	4.15	6.29
ERV1	4	2032	1.46	2.11
ERV1-MaLRs	7	3237	2.33	2.62
ERV_classI	1	436	0.31	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	6	801	0.58	3.67
hAT-Charlie	3	404	0.29	1.87
TcMar-Tigger	2	319	0.23	0.9

**Table 8.4** Repetitive elements across the entire horse orthologous region of EAS8 compared with whole genome levels

*Repetitive elements across the EAS9 centromere orthologous domain in the horse*

SINE abundance was decreased by 2.81% at the horse EAS9 centromeric orthologous domain in comparison to whole genome levels, while LINE levels almost doubled, increasing by 20.22%. This increase was attributed to L1 elements which increased by 25.33%, conversely there was a reduction in L2 elements with 38.8 times less present at this domain compared to whole genome levels. There was also a decrease of 3.96% in LTR elements while there was a 2.07% increase in DNA elements: hAT Charlie levels were decreased by 1.51%, conversely TcMar-Tigger were increased by 4.47%

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	3	321	0.72	3.53
ALUs	0	0	0	0
MIRs	3	321	0.72	3.49
<b>LINEs:</b>	26	18714	41.81	21.59
LINE1	24	18612	41.58	16.25
LINE2	1	54	0.12	4.66
L3/CR1	1	48	0.11	0.48
<b>LTR elements:</b>	4	1042	2.33	6.29
ERV_L	4	1042	2.33	2.11
ERV_L-MaLRs	0	0	0	2.62
ERV_classI	0	0	0	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	4	2567	5.74	3.67
hAT-Charlie	1	162	0.36	1.87
TcMar-Tigger	3	2405	5.37	0.9

**Table 8.5** Repetitive elements across the entire horse orthologous region of EAS9 compared with whole genome levels

*Repetitive elements across the EAS10 centromere orthologous domain in the horse*

There was a 1.08% reduction in SINEs at this domain in the horse compared to levels observed across the entire horse genome. LINE levels were increased by 7.37%, with L1 increasing by 10.64% while L2 levels decreased from whole genome levels of 4.66% to 1.71% at this domain. LTR element levels were comparable with whole genome levels, decreasing slightly by 0.43%. DNA elements were decreased by 2.33%.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	40	6258	2.45	3.53
ALUs	0	0	0	0
MIRs	40	6258	2.45	3.49
<b>LINEs:</b>	106	73989	28.96	21.59
LINE1	81	68709	26.89	16.25
LINE2	20	4372	1.71	4.66
L3/CR1	4	688	0.27	0.48
<b>LTR elements:</b>	43	14963	5.86	6.29
ERV_L	10	4706	1.84	2.11
ERV_L-MaLRs	19	5495	2.15	2.62
ERV_classI	13	4621	1.81	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	18	3416	1.34	3.67
hAT-Charlie	8	1371	0.54	1.87
TcMar-Tigger	2	546	0.21	0.9

**Table 8.6** Repetitive elements across the entire horse orthologous region of EAS10 compared with whole genome levels

*Repetitive elements across the EAS11 centromere orthologous domain in the horse*

There was a decrease of 1.39% in SINEs at this domain in the horse compared with

levels observed across the whole genome. LINE abundance was almost doubled, increasing by 20.54%: L1 and L2 elements increased by 18.52% and 2.28% respectively. There was also an increase in LTR elements (3.07%) while DNA element abundance remained comparable with whole genome levels dropping by 0.33%.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	11	1849	2.14	3.53
ALUs	0	0	0	0
MIRs	11	1849	2.14	3.49
<b>LINEs:</b>	38	36479	42.13	21.59
LINE1	24	30105	34.77	16.25
LINE2	12	6034	6.97	4.66
L3/CR1	0	0	0	0.48
<b>LTR elements:</b>	13	8106	9.36	6.29
ERV1	5	5437	6.28	2.11
ERV1-MaLRs	6	2085	2.41	2.62
ERV_classI	1	444	0.51	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	11	2891	3.34	3.67
hAT-Charlie	7	2260	2.61	1.87
TcMar-Tigger	2	327	0.38	0.9

**Table 8.7** Repetitive elements across the entire horse orthologous region of EAS11 compared with whole genome levels

*Repetitive elements across the EAS12 centromere orthologous domain in the horse*

Repeatmasker analysis of this domain in the horse showed SINE levels decreased to less than half that seen across the entire horse genome. There was an increase in LINE/L1 elements more than doubling from 16.25% to 39.5%, in contrast there was a decrease of 3.13% in L2 elements. LTR elements remained comparable with whole genome levels and hAT-Charlie (1.08%) and TcMar-Tigger (0.2%) abundance was lower than whole genome levels.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	31	4510	1.28	3.53
ALUs	0	0	0	0
MIRs	31	4510	1.28	3.49
<b>LINEs:</b>	127	144308	41.1	21.59
LINE1	103	138687	39.5	16.25
LINE2	23	5377	1.53	4.66
L3/CR1	1	244	0.07	0.48
<b>LTR elements:</b>	54	22911	6.53	6.29
ERV_L	21	11511	3.28	2.11
ERV_L-MaLRs	18	6325	1.8	2.62
ERV_classI	12	4030	1.15	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	28	9603	2.74	3.67
hAT-Charlie	11	2789	0.79	1.87
TcMar-Tigger	3	2115	0.6	0.9

**Table 8.8** Repetitive elements across the entire horse orthologous region of EAS12 compared with whole genome levels

*Repetitive elements across the EAS13 centromere horse orthologous domain*

SINE abundance was reduced at the EAS13 orthologous domain in EquCab (1.25%), when compared with levels observed across the entire horse genome. LINE levels rose 15.14%, with an increase in L1 (17.81%) and a decrease in L2 (2.32%) levels. There was an increase in LTR abundance by 3.7%, with ERVL (2.46%), ERVL-MaLR (0.03%) and ERV classI (1.35%) all higher than whole genome levels. DNA element levels were decreased, dropping from 3.67% across the genome to 1.08% at the horse orthologous domain.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	30	4989	2.28	3.53
ALUs	0	0	0	0
MIRs	30	4989	2.28	3.49
<b>LINEs:</b>	8	80505	36.73	21.59
LINE1	78	74642	34.06	16.25
LINE2	24	5133	2.34	4.66
L3/CR1	6	730	0.33	0.48
<b>LTR elements:</b>	43	21885	9.99	6.29
ERV_L	20	10023	4.57	2.11
ERV_L-MaLRs	18	5813	2.65	2.62
ERV_classI	4	5546	2.53	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	13	2361	1.08	3.67
hAT-Charlie	10	1740	0.79	1.87
TcMar-Tigger	2	473	0.22	0.9

**Table 8.9** Repetitive elements across the entire horse orthologous region of EAS13 compared with whole genome levels

*Repetitive elements across the EAS14 centromere horse orthologous domain*

Repetitive element analysis at the Eas14 horse orthologous domain showed a decrease in SINEs when compared to whole genome levels. LINE levels increased by 11.87%, with L1 levels increasing (14.99%) and L2 levels dropping (2.54%). There was also an increase LTR elements, with ERV ClassI (1.94%) and ERVL-MaLRs (1.3%) levels rising while ERVL levels (0.81%) decreased. Conversely DNA element levels were decreased at this domain (1.11%) when compared to levels observed across the whole genome.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	45	6222	2.1	3.53
ALUs	0	0	0	0
MIRs	43	6105	2.06	3.49
<b>LINEs:</b>	122	99067	33.46	21.59
LINE1	88	92481	31.24	16.25
LINE2	33	6266	2.12	4.66
L3/CR1	0	0	0	0.48
<b>LTR elements:</b>	66	25260	8.53	6.29
ERVL	14	3839	1.3	2.11
ERVL-MaLRs	32	11088	3.75	2.62
ERV_classI	14	9225	3.12	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	48	7583	2.56	3.67
hAT-Charlie	33	5408	1.83	1.87
TcMar-Tigger	2	416	0.14	0.9

**Table 8.10** Repetitive elements across the entire horse orthologous region of EAS14 compared with whole genome levels

*Repetitive elements across the EAS16 centromere orthologous domain in the horse*

Analysis of repetitive elements across this domain showed a decrease in SINE abundance when compared with the rest of the horse genome. LINE levels more than doubled increasing by 23.42% to 45.01%, this increase is due to an L1 element increase of 24.26%, while L2 levels dropped by 16%. There was a drop in LTR abundance, with a complete absence of ERV class I elements at this domain. DNA element levels decreased by 3.12% with both hAT-Charlie and TcMar-Tigger levels reduced.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	4	733	1.72	3.53
ALUs	0	0	0	0
MIRs	4	733	1.72	3.49
<b>LINEs:</b>	17	19225	45.01	21.59
LINE1	11	17301	40.51	16.25
LINE2	6	1924	4.5	4.66
L3/CR1	0	0	0	0.48
<b>LTR elements:</b>	5	1629	3.81	6.29
ERV_L	3	996	2.33	2.11
ERV_L-MaLRs	2	633	1.48	2.62
ERV_classI	0	0	0	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	2	237	0.55	3.67
hAT-Charlie	1	194	0.45	1.87
TcMar-Tigger	1	43	0.1	0.9

**Table 8.11** Repetitive elements across the entire horse orthologous region of EAS16 compared with whole genome levels

*Repetitive elements across the EAS18 centromere orthologous domain in the horse*

SINE abundance at the EAS18 centromere EquCab orthologous region was 2.1% less than levels observed across the genome. A 2.05% increase was observed in LINEs, with a increase in L1 (5.14%) and a decrease in L2 (2.8%). An increase was also observed in LTR elements (3.22%), with a rise in ERVL (1.54%), ERVL-MaLRs (0.95%) and ERV classI (0.91%). DNA elements at this domain decreased 2.68% when compared to level observed across the genome.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	13	1993	1.43	3.53
ALUs	0	0	0	0
MIRs	13	1993	1.43	3.49
<b>LINEs:</b>	40	32831	23.64	21.59
LINE1	29	29707	21.39	16.25
LINE2	10	2580	1.86	4.66
L3/CR1	0	0	0	0.48
<b>LTR elements:</b>	31	13207	9.51	6.29
ERV_L	11	5070	3.65	2.11
ERV_L-MaLRs	13	4962	3.57	2.62
ERV_classI	4	2902	2.09	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	11	1380	0.99	3.67
hAT-Charlie	6	621	0.45	1.87
TcMar-Tigger	3	436	0.31	0.9

**Table 8.12** Repetitive elements across the entire horse orthologous region of EAS18 compared with whole genome levels

*Repetitive elements across the EAS19 centromere orthologous domain in the horse*

Analysis of repetitive elements across the entire EAS19 CENP-A binding domain in the horse shows a drop in SINE abundance (2.5%) while there is an increase in LINE abundance (5.57%) with L1 levels (7.91%) rising and L2 levels (1.88%) dropping. There is also an increase in LTR levels rising by 0.96%, with the abundance of all three subclass rising. DNA element abundance at this domain is more than half that of whole genome levels (decrease of 2.2%)

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	15	2248	1.03	3.53
ALUs	0	0	0	0
MIRs	14	2186	1	3.49
<b>LINEs:</b>	63	59393	27.16	21.59
LINE1	46	52830	24.16	16.25
LINE2	15	6078	2.78	4.66
L3/CR1	2	485	0.22	0.48
<b>LTR elements:</b>	36	15860	7.25	6.29
ERVL	12	4432	2.03	2.11
ERVL-MaLRs	13	4520	2.07	2.62
ERV_classI	10	6812	3.12	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	17	3211	1.47	3.67
hAT-Charlie	8	1025	0.47	1.87
TcMar-Tigger	2	358	0.16	0.9

**Table 8.13** Repetitive elements across the entire horse orthologous region of EAS19 compared with whole genome levels

*Repetitive elements across the EAS27 centromere orthologous domain in the horse*

Analysis of the entire EAS27 centromere associated domain in horse showed a reduction in SINE levels (1.36%) when compared to the whole genome. In contrast LINE abundance increased by 11.98%, with both L1 (11.76%) and L2 (0.66%) levels rising. There was a drop in overall LTR abundance of 1.59%, with a reduction in all element classes. There was also a reduction in DNA elements (0.71%).

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	23	3658	2.17	3.53
ALUs	0	0	0	0
MIRs	23	3658	2.17	3.49
<b>LINEs:</b>	74	56617	33.57	21.59
LINE1	50	47245	28.01	16.25
LINE2	20	8978	5.32	4.66
L3/CR1	3	222	0.13	0.48
<b>LTR elements:</b>	21	7929	4.7	6.29
ERV_L	6	1649	0.98	2.11
ERV_L-MaLRs	10	4174	2.48	2.62
ERV_classI	1	442	0.26	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	20	4992	2.96	3.67
hAT-Charlie	7	1903	1.13	1.87
TcMar-Tigger	7	1504	0.89	0.9

**Table 8.14** Repetitive elements across the entire horse orthologous region of EAS27 compared with whole genome levels

*Repetitive elements across the EAS30 centromere orthologous domain in the horse*  
 Repetitive element analysis at the horse domain corresponding to the EAS30 centromere showed a 2.53% reduction in SINES. Conversely LINE abundance had increased by 2.51%, with L1 levels (3.97%) up and L2 levels (0.78%) down. There was also a rise in overall LTR abundance by 2.62%, with ERVL levels dropping (0.76%) and both ERVL-MaLRs (2.04%) and ERV classI (1.58%) levels increasing. DNA element abundance had also increased, rising by 0.46% as a result of a 0.59% increase in TcMar-Tigger.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	7	990	1	3.53
ALUs	0	0	0	0
MIRs	7	990	1	3.49
<b>LINEs:</b>	31	23862	24.1	21.59
LINE1	21	20017	20.22	16.25
LINE2	10	3845	3.88	4.66
L3/CR1	0	0	0	0.48
<b>LTR elements:</b>	16	8818	8.91	6.29
ERV_L	3	1334	1.35	2.11
ERV_L-MaLRs	8	4613	4.66	2.62
ERV_classI	4	2732	2.76	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	11	4117	4.16	3.67
hAT-Charlie	6	1836	1.85	1.87
TcMar-Tigger	2	1478	1.49	0.9

**Table 8.15** Repetitive elements across the entire horse orthologous region of EAS30 compared with whole genome levels



*Repetitive elements across the EASX centromere orthologous domain in the horse*

Analysis of repetitive elements at this domain in the horse genome showed a reduction in the levels of SINEs (2.97%) when compared to the whole genome. LINE levels were increased by 16.33%, with L1 (19.36%) levels up while L2 (3.1%) levels were decreased. LTR element abundance was increased by 1.94%, with both ERVL (2.5%) and ERVL-MaLRs (0.09%) levels up while there was a drop in ERV class I abundance (0.27%). There was a decrease in DNA elements, with levels dropping by 0.4%.

Elements	No. of elements	Length occupied bp	% of sequence	% of sequence Whole Genome
<b>SINEs:</b>	5	579	0.56	3.53
ALUs	0	0	0	0
MIRs	5	579	0.56	3.49
<b>LINEs:</b>	54	39337	37.92	21.59
LINE1	44	36943	35.61	16.25
LINE2	6	1622	1.56	4.66
L3/CR1	4	772	0.74	0.48
<b>LTR elements:</b>	21	8535	8.23	6.29
ERVL	8	4784	4.61	2.11
ERVL-MaLRs	9	2809	2.71	2.62
ERV_classI	4	942	0.91	1.18
ERV_classII	0	0	0	0
<b>DNA elements:</b>	13	3395	3.27	3.67
hAT-Charlie	7	1863	1.8	1.87
TcMar-Tigger	2	841	0.81	0.9

**Table 8.16** Repetitive elements across the entire horse orthologous region of EASX compared with whole genome levels

## Appendix II

Centromere coordinates in the primary and immortalized donkey fibroblasts.

Cen	Peak	Primary coordinates (nt)	Immortalised coordinates (nt)	Primary domain span (kb)	Immortalised domain span (kb)
Eca5/Eas16	1	74,873,938- 74,940,021	74,873,938- 74,930,101	66.1	56.2
Eca6/Eas19	1	14,180,898- 14,252,252	14,180,898- 14,252,252	71.4	71.4
Eca8/Eas7	1	41,917,682- 42,137,806	41,972,718- 42,121,470	220.1	148.8
Eca9/Eas12	1	31,943,974- 32,066,330	32,135,195- 32,246,892	122.4	78.0
Eca9/Eas12	2	31,950,769- 32,028,722	32,139,967- 32,224,362	111.7	84.4
Eca11/Eas13	1	46,711,805- 46,848,897	46,729,294- 46,827,799	137.1	98.5
Eca13/Eas14	1	7,222,926- 7,298,581	7,223,492- 7,298,413	75.7	74.9
Eca13/Eas14	2	7,350,216- 7,476,466	7,349,587- 7,474,729	126.3	125.1
Eca14/Eas9	1	29,616,607- 29,692,084	29,616,607- 29,692,084	75.5	75.5
Eca17/Eas11	1	16,741,838- 16,859,286	16,754,025- 16,842,919	117.4	88.9
Eca19/Eas5	1	4,917,307- 5,024,460	4,928,659- 5,024,735	107.2	96.1
Eca19/Eas5	2	5,031,390- 5,160,423	5,036,009- 5,135,847	129.0	99.8
Eca20/Eas8	1	26,418,570- 26,510,778	26,418,570- 26,510,778	92.2	92.2
Eca25/Eas10	1	8,576,403- 8,698,559	8,596,732- 8,685,857	122.2	89.1
Eca25/Eas10	2	8,703,672- 8,817,899	8,721,979- 8,836,225	114.2	114.2
Eca26/Eas18	1	22,363,191- 22,473,453	22,369,150- 22,460,799	110.3	91.6
Eca26/Eas18	2	22,495,599- 22,528,174	22,495,599- 22,528,174	32.6	32.6
Eca27/Eas27	1	19,710,802- 19,931,051	19,722,819- 19,807,301	220.2	84.5
Eca27/Eas27	2	-	19,809,999- 19,888,529	-	78.5
Eca28/Eas4	1	12,864,618- 13,058,668	12,894,614- 13,029,445	194.1	134.8
Eca30/Eas30	1	17,709,309- 17,823,588	17,720,873- 17,811,906	114.3	91.0
EcaX/EasX	1	26,926,226- 27,077,732	26,962,334- 27,061,644	151.5	99.3

**Table 8.17 Centromere coordinates and sizes in the immortalized and primary cell line**