Exploring the Functional Landscape of the Fission Yeast Genome via *Hermes* Transposon Mutagenesis

Leanne Grech

Submitted for the Degree of Doctor of Philosophy (PhD)

University College London

Primary Supervisor: Professor Jürg Bähler (UCL)

January 2018

Declaration

"I, Leanne Grech, confirm that the work presented in this thesis is my own, and where information has been derived from other sources, I confirm that this has been indicated in the thesis."

Abstract

In general, the non-protein-coding dark matter of eukaryotic genomes remains poorly understood. Neither diversity studies, comparative genomics, nor the biochemical outputs allow fine-scale descriptions of the genomic elements that are required for an organism to grow. Transposon mutagenesis offers an alternative approach to locate functional regions. In principle, insertion mutants are created in large pools, and mutants harbouring harmful insertions are quantitatively removed from the population. Subsequent sequencing of mutant libraries should reveal functionality in regions with fewer insertions. Transposon mutagenesis works well in bacteria.

We applied the *Hermes* transposon system to locate functional regions in the *Schizosaccharomyces pombe*, or fission yeast genome. We created multiple dense insertion libraries, during log phase growth and chronological ageing, achieving a saturating (or near-saturating) insertion density of 1 insertion per 13 nucleotides of the genome for log phase samples. To account for the complexity and stochastic nature of the data, we applied a five-state hidden Markov model (HMM) that includes generalised linear models to account for systematic insertion biases (e.g. nucleosomes).

The HMM state provides a semi-quantitative estimate of the functional significance of the genome at single nucleotide-level resolution. The HMM state values are strikingly consistent (but more detailed than) genome annotations. Here, we show that transposon insertions have functional consequences in 90% of the genome, including 80% of the non-protein-coding regions. Specifically, we discover 85 essential ncRNAs during vegetative growth, and 218 during ageing. We also find 54 pro-ageing and 136 anti-ageing genes. Overall, this functional annotation map distinguishes sub gene-level genomic segments that have differential effects on cell survival, and so will have extensive utility to the community.

Contents

Acknowledgements

It has been a metamorphic roller coaster ride,
and I am fortunate to have had you all by my side.
Jürg, it was the humorous lighthearted satire, e.g.
what is the best thing about Switzerland? Discuss.
I don't know, but their flag is a big plus.

Dan, for teaching me the ins and outs of R.
I now know about gilding the lilies; bizarre.
S. pombe is an elephant, we are blind men;
we'll figure it out, it's a question of when.

María y Bala, por ser amable y paciente.
"la vida es hermosa, muy muy caliente"
One Day, we'll write poetry on a typewriter,
and the future'll be a hundred times brighter.

Mimi, and the friends in the Bähler clan,
united forever through a chronological lifespan.
I'll remember our lunches, sharing life stories.
I've done a PhD in one of the best laboratories!

Andy, Julie, Kath and Steve, thanks ever so much,
for giving us both a comfortable roof over our head,
and letting us win once at Ticket to Ride (too bad).
I look forward to the memories ahead.

Dad, ma, Nadine, Sakura, jien dik li jien
għax dejjem ħabbejtuni mingħajr waqfien.
Għallimtuni li f'din il-ħajja, il-kuntentizza prezzjuża,
u fejn hemm l-imħabba, l-għana mhiex bżonnjuża.
Grazzi ta' kollox. Inħobbkom b'qalbi kollha.

Chris, qalbi, the serendipitous micropalaeontologist.
You pushed me up through the deepest trough,
and for that I cannot thank you enough.
This PhD was the start of something greater,
and now here we are, four years later.

List of Acronyms and Abbreviations

bc	barcode
BIC	Bayesian information criterion
bp	base pair
BWA	Burrows-Wheeler Aligner
canonRNA	canonical RNA
CDS	coding DNA sequence
CEN	centromere
CFU	colony forming unit
CLS	chronological life span
CMH	Cochran–Mantel–Haenszel
DAPI	4',6-diamidino-2-phenylindole
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dsDNA	double stranded DNA
eCDS	essential coding DNA sequence
EDTA	ethylenediaminetetraacetic acid
EMM	Edinburgh minimal medium
FDR	false discovery rate
FOA	5-fluoroorotic acid
FYPO	fission yeast phenotype ontology
G418	G418 disulfate salt solution
gRNA	guide RNA
HMM	Hidden Markov Model
IGF	insulin-like growth factor

IPKM	insertions per kilobase per million insertions
kb	kilobase
kDa	kilodaltons
leu	leucine
LiAc	lithium acetate
lincRNA	long intergenic non-coding RNA
lncRNA	long non-coding RNA
LTR	long terminal repeat
Mbp	megabase pair
MT	mitochondria
NaCl	sodium chloride
nCDS	non-essential coding DNA sequence
ncRNA	non-coding RNA
NEB	New England Biolabs
NFR	nucleosome-free region
NGS	next generation sequencing
NHEJ	non-homologous end joining
nm	nanometre
nmt1	no-message-in-thiamine promoter
NOA	no annotation
nt	nucleotide
NUMT	nuclear mitochondrial DNA segment
OD_{600}	optical density at 600 nm
ORF	open reading frame
PCR	polymerase chain reaction

PEG	polyethylene glycol
pg	picogram
pH	potential of hydrogen
RNA	ribonucleic acid
rpm	revolutions per minute
rRNA	ribosomal RNA
S. pombe	*Schizosaccharomyces pombe*
S.O.C.	super optimal broth with catabolite repression
SAM	sequence alignment map
SBS	sequencing by synthesis
snoRNP	small nucleolar ribonucleoprotein
snRNP	small nuclear ribonucleoprotein
TE	tris-EDTA
TEL	telomere
TIR	terminal inverted repeat
Tn-seq	transposon sequencing
TORC1	target of rapamycin complex 1
tRNA	transfer RNA
UCL	University College London
UIS	unique insertion sites
UMI	unique molecular identifier
ura	uracil
UTR	untranslated region
YES	yeast extract with supplements

List of Figures

List of Tables

Chapter 1 BACKGROUND

1.1 *Schizosaccharomyces pombe* as a Model Organism

Schizosaccharomyces pombe, or fission yeast, is a unicellular haploid eukaryote characterised by rod-shaped cells, first developed as an experimental genetic model in the 1950s by Urs Leupold. Since then, it has emerged as a popular model organism of great scientific importance. In particular, it is significant in the field of cellular and molecular biology, when studying cell cycle control (Nurse 1990), centromere structure (Allshire and Karpen 2008), cytokinesis (Goyal *et al.* 2011), DNA repair and recombination (Phadnis *et al.* 2011), heterochromatin assembly mediated by RNA interference (Goto and Nakayama 2012), and mitosis and meiosis (Harigaya and Yamamoto 2007). For the most part, this can be attributed to the fact that *S. pombe* is amenable to genetic manipulations, as well as having a rapid division cycle with a generation time of two to four hours. Moreover, its haploid life cycle facilitates the recovery of what would be recessive alleles in diploid organisms, while its compact genome reduces the likelihood that a gene knockout will be masked by redundant genes.

Importantly, *S. pombe* has a fully sequenced (Wood *et al.* 2002), well-annotated (Wood *et al.* 2012) genome; ~12.6 Mbp in size, and distributed across three chromosomes. In general, several mechanisms and pathways have been found to be conserved from fission yeast to higher eukaryotes such as humans. In addition, conserved genes that are essential for eukaryotic cell organisation were discovered, and these are believed to have originated with the appearance of eukaryotic life. Similarly, conserved genes that are important for multicellular organisation were identified, suggesting that the transition from prokaryotes to eukaryotes required more novel genes than did the transition from unicellular to multicellular organisation (Wood *et al.* 2002).

S. pombe can also be seen as an informative predictor of human gene function, which means that it could be used to model complex disease processes. Indeed, over 20 protein-coding genes have been implicated in cancer, specifically, in the cell cycle, checkpoint controls, DNA damage and repair, and in other processes known to maintain genomic stability. *S. pombe* has likewise been exploited to understand the neurodegenerative lysosomal storage disorder Batten disease (Haines *et al.* 2009). In addition, systematic screens for mutants resistant to TORC1 inhibition have revealed genes involved in cellular ageing and growth (Rallis *et al.* 2014). Overall, *S. pombe* has proven a valuable model organism to systematically interrogate the genome in a manner that is not possible, or at least more limited in higher eukaryotes.

1.2 Functional Elements in the Fission Yeast Genome

In order to understand the human genome, we must first discover and interpret all functional elements within its sequence. For this reason, it is important to recognise what constitutes function and what sets the boundaries of an element. Since there is no universal definition of function, only an intuitive one, each scientific discipline has to rely on different lines of evidence to define function, making it a very controversial field to pursue (Germain *et al.* 2014). In fact, while the biochemical approach quantifies evidence of molecular activities, the evolutionary method measures selective constraint, and the genetic process evaluates phenotypic consequences of mutations. Together, these three different approaches can provide information on the biological significance of an element. Indeed, groups of functional genomic elements identified through each method can be quantitatively enriched for each other (Kellis *et al.* 2014). Operationally, functional elements have been defined as discrete genomic segments that encode a defined product (such as protein or non-coding RNA) or display a reproducible biochemical signature (such as protein binding or a specific chromatin structure) (Dunham *et al.* 2012).

Having said this, it is now established that only ~1.5% of the human genome codes for protein sequence (Lander *et al.* 2001). However, comparative studies with mammalian genomes (Waterston *et al.* 2002, Gibbs *et al.* 2004, Lindblad-Toh *et al.* 2005) have shown that at least 5% is under selective constraint and therefore perhaps functional, of which ~3.5% consists of non-coding sequences with apparent roles (Lindblad-Toh *et al.* 2011). In general, this created an enigmatic aura, leading to the label of 'dark matter', similar to the 'dark matter' of the universe, which we cannot easily detect or understand, but that nevertheless exists and is exposed to experimental queries. Overall, existing research on these non-coding regions, which form a part of this once proverbial genomic 'dark matter', suggests that these regions have important biological roles in cellular homeostasis, development, differentiation, and metabolism. In fact, their aberrant expression or regulation (dysregulation) is being found in numerous human diseases, including cancer, cardiovascular, developmental, and neurological diseases. In consequence, translational research is examining the potential use of these non-coding elements as biomarkers and molecular targets in medical theranostics.

However, considering the sheer volume of research investigations carried out to date, this 'dark matter' remains, for the most part, poorly understood, as are most of the other functional elements of the human genome. With this in mind, scientists often turn to the genomes of model eukaryotic organisms such as insects, worms, and yeasts. One of the benefits of studying functional elements in model organisms is the ability to biologically validate the elements discovered using methods that cannot always be applied in humans. For instance, Siepel *et al.* (2005) carried out a comprehensive search for conserved elements in four insect species (three species of *Drosophila melanogaster* and *Anopheles gambiae*), two species of *Caenorhabditis elegans*, and seven species of *Saccharomyces cerevisiae*. Here, Siepel *et al.* (2005) found that from yeasts to vertebrates, increasing fractions of conserved bases lie outside of exons of protein-coding genes, therefore

reflecting the importance of non-protein-coding and other regulatory sequences in higher eukaryotes.

Studies within the fission yeast community have also helped discover new, conserved, functional elements. For example, Wilhelm *et al.* (2008) interrogated the *S. pombe* transcriptome under multiple conditions and detected widespread transcription in over 90% of the genome. In doing so, they provided information on novel, mostly non-coding transcripts and untranslated regions (UTRs), thus improving the existing genome annotation. Studies by Fawcett *et al.* (2014) and Jeffares *et al.* (2015) independently showed that in addition to introns and UTRs, intergenic regions exhibit lower levels of nucleotide diversity, suggesting that a considerable amount of non-coding transcripts are under selective constraint and thus likely to be functional. Having said this, it is important to note that such population genomics studies (data from within species) do not have the power to locate specific functional elements, only classes that are conserved. Similarly, comparative genomics studies (data from between species) do not have the power to locate functional elements that are not conserved over a long period of time (such as genomic elements specific to *S. pombe*). Overall, this shows that even a fully sequenced and well-annotated genome, such as the *S. pombe* genome, has genomic complexities far beyond current annotation. Importantly, it demonstrates that the hunt for functional elements within the fission yeast genome is ongoing and needs to be further pursued.

1.3 Mutagenising the Fission Yeast Genome

Over the past decade, the advent of high throughput sequencing, or next generation sequencing (NGS), has greatly accelerated the rate at which genes are discovered. In itself, this has challenged the existing methods for defining the functions of genes. In fact, several methods have now been developed to gain further insight into gene function. Perhaps one of the most novel approaches, especially in haploid cells, is that generating systematic deletions and disruptions

within the genome. Indeed, the first genome-wide deletion library for fission yeast is now commercially available by the Bioneer Corporation. Here, Kim *et al.* (2010) determined the essentiality of 4,836 protein-coding genes, and found that 26.1% of *S. pombe* genes are essential and 73.9% are non-essential for viability of haploid cells in the growth conditions used. In this deletion library, each mutant contains a pair of unique molecular barcodes that can be monitored either by microarrays or by NGS techniques, thus making it quicker to identify the genes that are responsible for a particular phenotype (Han *et al.* 2010, Kim *et al.* 2010). While it is still a powerful reverse genetics tool, this deletion library is not without its limitations. One of its major drawbacks is that it contains mutants with deletions only in the open reading frames (ORFs). Therefore, it does not specifically target intergenic and non-coding elements (Guo *et al.* 2013), and the question as to how much genetic information is actually transacted by non-coding regions still remains unanswered.

1.3.1 Transposon Saturation Mutagenesis

To this end, other different approaches have to be taken into account, such as saturation mutagenesis, which makes it possible to create libraries of mutants containing all possible mutations in a gene sequence (Zheng *et al.* 2004). Generally, for transposon mutagenesis libraries, this involves the use of ubiquitous transposable elements; specialised sequences that are capable of moving within a host genome in a non-replicative manner. In an experimental setting, a transposon can be modified to carry almost any sequence cargo, however, in order to do this, it needs to be flanked by sequences termed terminal inverted repeats (TIRs). TIRs are specific to the transposon, and are recognised by a transposase enzyme that is necessary in mediating the movement of the transposon. Since transposons are also inherently mutagenic, numerous molecular applications based on transposons have been established in multiple model organisms.

In fission yeast, three types of transposons have been analysed in a genome-wide context: the class I retrotransposon *Tf1* and the two class II cut-and-paste DNA transposons (a) *piggyBac*, isolated from the cabbage looper moth *Trichoplusia ni*, and (b) *Hermes*, isolated from the housefly *Musca domestica*. First attempts to use *Tf1* as a mutagen were not very successful because most of the insertions occurred in clusters in a window 500 nt upstream of ORFs (Behrens *et al.* 2000, Singleton and Levin 2002, Bowen *et al.* 2003, Guo and Levin 2010). However, because *Tf1* was reported to harbour a chromodomain, and since it is known that chromodomains interact with heterochromatic regions, Cherry *et al.* (2014) anticipated possible *Tf1* targeting to heterochromatic regions of *S. pombe*. Indeed, their results revealed that *Tf1* can insert within silent regions of the genome.

In regard to the *piggyBac* transposon, Li *et al.* (2011) used an *S. pombe* strain with a selectable transposon excision marker and an integrated transposase gene, to show that most insertion sites occur within intergenic regions and TTAA sites, with limited local hopping effect and little chromosomal bias. In general, the preference for intergenic regions was assumed to be due to selective pressure on insertions in ORFs that resulted in reduced fitness. One of the main problems of the *piggyBac* transposon is that there are too few potential insertion sites within the genome. Nonetheless, using this system, they managed to obtain different types of alleles, including null, hypomorphic and hypermorphic alleles, which were all broadly distributed among the three yeast chromosomes.

Overall, these studies show how useful transposon mutagenesis could be in the exploration of the *S. pombe* genome, in particular when compared to the Bioneer deletion library, which in general created null alleles of non-essential genes. In addition, with the deletion library, it is more arduous to perform a screen in a genetic background that is different from that of the library, because to do so would involve crossing the desired background into each deletion strain within the

library. It is indeed much easier to introduce transposable elements into the preferred genetic background. One other advantage of transposon mutagenesis is that transposons often insert at individual locations as single, unmodified copies, creating defined boundaries between genomic and transposon DNA, which therefore makes the insertion sites easier to map with high throughput sequencing (Dos Santos *et al.* 2012).

1.3.2 Hermes Transposon Mutagenesis

Hermes was first developed into a transposon mutagenesis tool for fission yeast by Evertts *et al.* (2007). In this work, they describe a system that contains the *Hermes* transposase, driven by a repressible promoter, and the transposon, composed of a drug resistance marker flanked by the TIRs of *Hermes*. Upon promoter induction, and subsequent transposase expression, the transposon is excised at its TIRs and then integrated into chromosomal DNA. One advantage of this approach is that the transposase and the transposon are encoded on separate plasmids which means that transformations can be carried out in different genetic backgrounds. Moreover, the risk of local hopping is reduced; local hopping, a shared feature of cut-and-paste transposons, is the phenomenon in which transposons preferentially land into cis-linked sites in the vicinity of the donor locus (Ivics and Izsvák 2010).

One other advantage of the *Hermes* transposon system in *S. pombe* is that there is only a slight preference in the insertion sites. In fact, Guo *et al.* (2013) observed that *Hermes* efficiently disrupts intergenic regions and ORFs with bias towards nucleosome-free sites. Similar studies carried out in other organisms, such as *Drosophila melanogaster* (Guimond *et al.* 2003) and *Saccharomyces cerevisiae* (Gangadharan *et al.* 2010), also showed that *Hermes* insertions are biased towards nucleosome-free and TA-rich regions. In addition, Gangadharan *et al.* (2010), show that *Hermes* can insert in either orientation once it recognises an 8 bp nTnnnnAn target site (Figure 1.1). Overall, the *Hermes* transposon system

appears to have a clear advantage over the other transposon systems, which makes it better suited in creating transposon mutagenesis libraries, and therefore, in potentially discovering functional elements within the *S. pombe* genome.

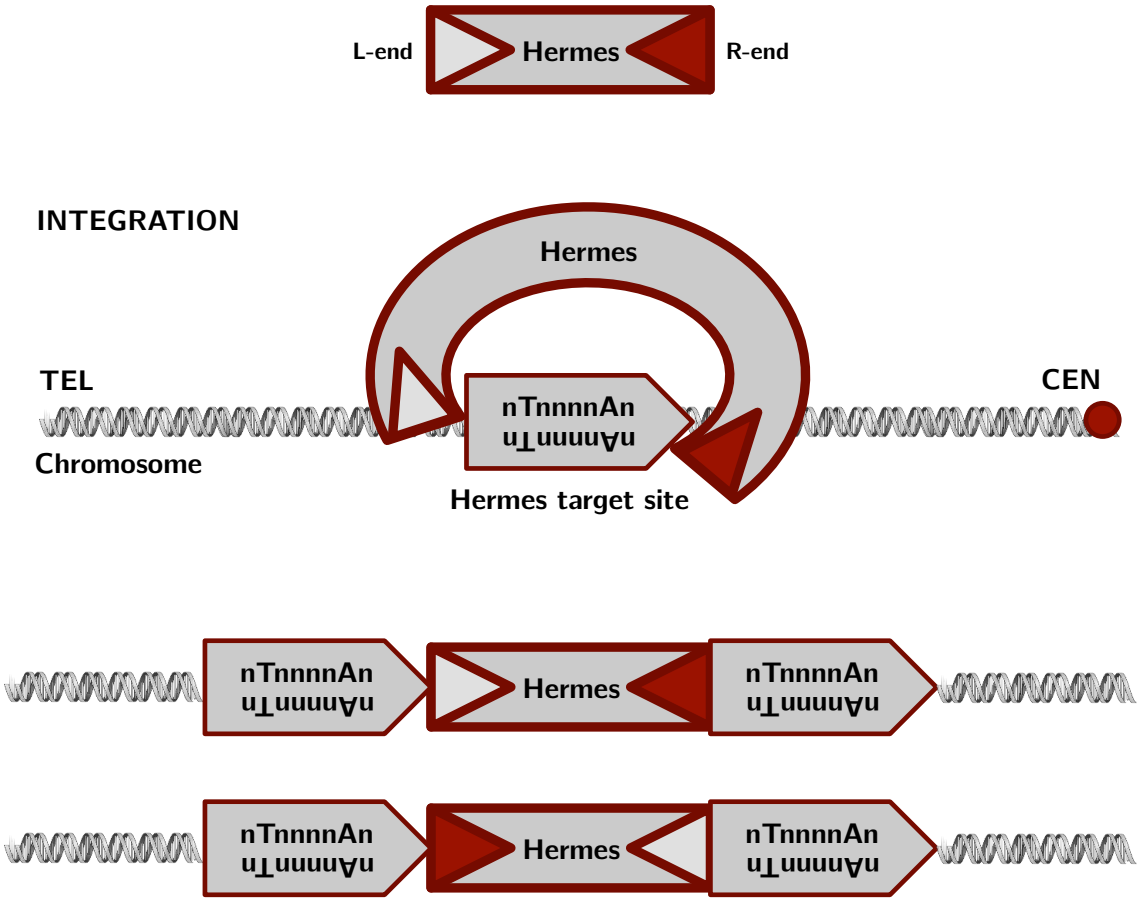


Figure 1.1. *Hermes* Transposition Mechanism. *Hermes* inserts in either orientation once it recognises an 8 bp nTnnnnAn target site (adapted from Gangadharan *et al.* (2010)).

Indeed, the approach of coupling transposon mutagenesis with NGS technologies is quite powerful. It is a method that has been well exemplified in multiple organisms, including prokaryotes (van Opijnen and Camilli 2013), yeasts (Oh *et al.* 2010, Li *et al.* 2011, Guo *et al.* 2013, Michel *et al.* 2017), and haploid mammalian cells (Pettitt *et al.* 2013).

Here, this technique will be used to examine, at a fine-scale resolution, the fitness consequences of transposon insertions during mitotic proliferation as well as

chronological ageing. In general, the biological mechanisms at the heart of the ageing process remain unsolved. However, research on the role of genes involved in ageing and longevity is ongoing. *S. pombe* has emerged as a model organism to unravel chronological ageing. In this field, transposon mutagenesis, together with high throughput sequencing, can help to discover genes with changing importance as a function of ageing.

1.4 Aims: Rationale and Significance

In spite of extensive efforts, a lot remains to be discovered within the *S. pombe* genome. While the dispensability, or essentiality, of most protein-coding genes under the standard growth condition is known, not much is known with regards to the non-protein-coding genes. In order to predict genomic function, one option would be to knockout genes. However, another approach could also be used in parallel: during transposon mutagenesis, transposons are randomly inserted in as many positions in the genome as possible, creating mutations that can be easily mapped and sequenced. Subsequently, essential elements can be distinguished from non-essential ones, in that genomic regions lacking insertion sites will be considered as essential whereas those harbouring a lot of insertions will not (Figure 1.2).

In brief, if the transposon is inserted in the middle of an essential regulatory element, the effects on the cell will be drastic, whereas the opposite is true for non-essential elements. Taking all of this into consideration, this study will explore the functional landscape of the *S. pombe* genome through *Hermes* transposon mutagenesis. In particular, this work aims to answer questions such as: How much of the non-protein-coding genes are functional during log phase growth and chronological ageing? How well does transcriptional activity predict an essential region? How much of the non-conserved genes (between species) are functional, and are there any conserved genes that do not seem to be functional in *S. pombe*?

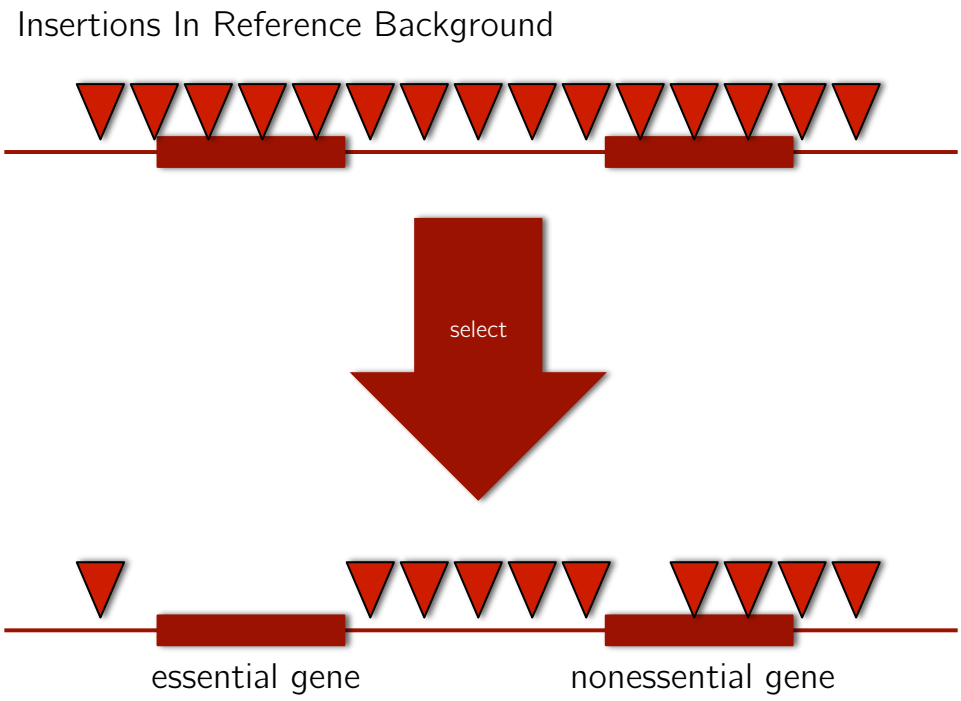


Figure 1.2. Essential vs. Non-Essential Genes. Following selection and sequencing, essential genes can be distinguished from non-essential genes, since essential genes accumulate far fewer insertions than non-essential ones.

Chapter 2 MATERIALS AND METHODS

2.1 Strains and Media

EMM (Edinburgh Minimal Medium) and YES Broth (yeast extract, glucose, and amino acid supplement), both purchased from Formedium™ (Norfolk, UK), were used for the cultivation of fission yeast. EMM is a selective medium. YES is used for non-selective, vegetative growth. 2% agar was added for solid-phase growth on plates. For liquid growth, cells were grown in an incubator shaker (Infors, Surrey, UK) at 32°C and 170 rpm. Cell growth was approximated by a cell density meter (Biochrom Ltd., Cambridge, UK), measuring optical density (OD) at a wavelength of 600 nm, where an OD of 0.1 was taken to correspond to 2 x 10^6 cells/ml. Strains carrying plasmids were stored at 4°C on selective agar plates to ensure plasmid retention. Strains without plasmids were stored at 4°C on YES agar plates. For long term storage, strains were frozen at -80°C in YES, or EMM, and 50% glycerol.

For the *Hermes* cell libraries, thiamine was added at a final concentration of 5 µg/ml to repress expression from the no-message-in-thiamine (*nmt1*) promoter (Maundrell 1990). 5-FOA (5-fluoroorotic acid) (Zymo Research Corporation, Irvine, California) was used at a final concentration of 2 mg/ml and G418 (G418 disulfate salt solution) (Formedium™, Norfolk, UK) was used at a final concentration of 50 mg/ml. For the *Hermes* DNA libraries, all oligonucleotides were purchased through Invitrogen™ (Paisley, UK).

2.2 *Hermes* Cell Libraries

2.2.1 Fission Yeast Transformation

A. Preparing The Plasmids

So as to propagate the plasmids, One Shot TOP10® Chemically Competent *Escherichia coli* were used as these allow stable replication of high-copy number

plasmids. For each transformation, one vial of One Shot TOP10® Chemically Competent *E. coli* was thawed on ice. 10 pg to 100 ng of the plasmid DNA were added to the vial, mixed gently by swirling or tapping, and then the vial was incubated on ice for 30 minutes. Cells were heat shocked for 30 seconds at 42°C without shaking and then placed on ice for 2 minutes. Subsequently, 250 µl of pre-warmed S.O.C. Medium (Invitrogen™, Paisley, UK) were aseptically added to each vial. This was then capped tightly and shaken horizontally at 225 rpm for 1 hour at 37°C in a shaking incubator. From each transformation, 20 to 200 µl were spread on a pre-warmed selective plate (lysogeny broth agar plate containing 100 µg/ml ampicillin) which was then incubated overnight at 37°C.

Using a sterile pipette tip, a single colony was picked from the freshly streaked selective plate and then used to inoculate a culture of 1 to 5 ml lysogeny broth medium containing 100 µg/ml ampicillin. This bacterial culture was incubated overnight (12 to 16 hours) at 32°C in a shaking incubator. Growth was characterised by a cloudy haze in the medium. Bacterial cells were then harvested by centrifugation at > 4000 rpm (6800 x g) in a conventional, table-top microcentrifuge for 6 minutes at room temperature (15 to 25°C). To purify plasmid DNA, the 'Plasmid DNA Purification using the QIAprep Spin Miniprep Kit and a Microcentrifuge' protocol (page 22 on the QIAprep® Miniprep Handbook) was employed. To quantify the purified plasmid DNA, the NanoDrop 2000 UV/Vis Spectrophotometer was used, where a 260/280 ratio of ~1.8 assessed the purity of the DNA.

For long term storage of the bacteria, a glycerol stock was created by gently mixing 500 µl of the overnight culture with 500 µl of 50% glycerol in a 2 ml cryovial and then freezing the glycerol stock tube at -80°C.

B. Transforming The Plasmids

First, an *S. pombe* strain was selected taking into account the configuration of both the donor and the expression plasmids (Figure 2.1). Owing to its leu– and ura– genotype and its suitability for detecting the desired phenotype, the JB980 (ura4–D18 leu1–32 h–) strain was grown in 50 ml YES until an OD_{600} of 0.8 to 1.0. For each transformation, 20 ml of cells were pelleted in a falcon tube by centrifuging at 2500 rpm for 5 minutes at room temperature and discarding the supernatant. Pelleted cells were washed in 50 ml sterile water, centrifuged again, and the supernatant removed. Following that, the cells were transferred to a microcentrifuge tube in 1 ml sterile water, centrifuged, and the supernatant discarded. Cells were then washed in 1 ml LiAc-TE, centrifuged, and the supernatant removed. 100 µl LiAc-TE were added to resuspend the cells and the resulting cell suspension was incubated for 10 minutes at 32°C.

Subsequently, 100 µl of cells were gently mixed with 2 µl of carrier DNA (at 10 mg/ml) and up to 5 µl of donor plasmid DNA, and incubated for 10 minutes at room temperature. For the negative control, sterile water was used instead of the donor plasmid DNA. 260 µl of fresh and sterile 40% PEG/LiAc-TE were gently mixed to the cell suspension followed by 30 to 60 minutes incubation at 29°C to 30°C. 43 µl of pre-warmed DMSO were also gently mixed in. Cells were heat shocked for 5 minutes at 42°C, put on ice for 2 minutes, and then centrifuged. Following the removal of the supernatant, 100 µl of sterile water were added and plated on EMM + N + leucine plates to select for the donor plasmid. To confirm the positive colonies, single colonies were re-streaked on EMM + N + leucine plates.

So as to help avoid recombination between plasmids, the donor plasmid (pHL2577) was transformed first followed by a second transformation to introduce the expression plasmid (pHL2578). During the second transformation, to select for both the donor and the expression plasmids, cells were plated on

EMM + N + thiamine plates. For the negative control, an empty expression vector (Rep81X) was transformed in place of the expression plasmid.

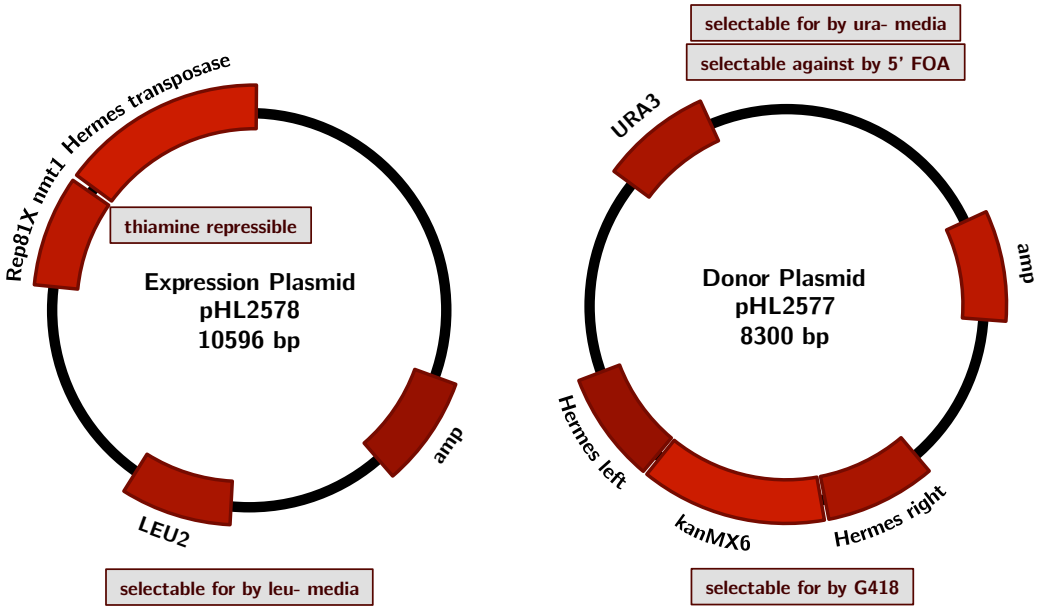


Figure 2.1. Donor and Expression Plasmids for *Hermes* Transposition in Fission Yeast. pHL2577 donor plasmid provides the source of transposon DNA flanked by the left and right terminal inverted repeats (TIRs). Here, the kanMX6 cassette between the TIRs gives cells with insertions resistance to the drug G418 (Geneticin). With regards to the pHL2578 expression plasmid, this contains the transposase gene which is under the control of the Rep81X *nmt1* promoter; removal of thiamine allows the expression of the transposase (adapted from Evertts *et al.* (2007)).

2.2.2 Cell Library Construction

Following the sequential introduction of the donor and the expression plasmids into the JB980 strain, one colony was picked to inoculate 50 ml EMM – Leu – Ura + 15 µM Thiamine. When this starter culture reached stationary phase, at an $OD_{600} \approx 2$ to 5, 5 ml of cells were pelleted at 2000 rpm for 5 minutes and then washed for four times in 25 ml EMM – Leu – Ura – Thiamine to remove thiamine. Removal of thiamine allowed the expression of the transposase from the *nmt1* promoter. Following that, an aliquot was taken to inoculate 50 ml EMM – Leu –

Ura – Thiamine to an OD_{600} ≈ 0.05 (i.e. 1 x 10^6 cells/ml), with the actual OD_{600} value recorded as $OD_{initial(1)}$ and dubbed as the cell number at generation zero. This culture was grown to a final OD_{600} ≈ 2 to 5, with the actual OD_{600} value recorded as $OD_{final(1)}$. Using these two values and the equation below, the number of cell generations was calculated. Subsequently, when the OD_{final} was reached, cells were plated to monitor the transposition frequency.

$$n = \frac{\left[\ln\left(\frac{OD_{final}}{OD_{initial}}\right)\right]}{0.693}$$ where *n* is the number of cell generations and *ln* is the natural log.

Next, after the first serial passage, an aliquot was taken to inoculate 50 ml EMM – Leu – Ura – Thiamine to an OD_{600} ≈ 0.05, with the actual OD_{600} value recorded as $OD_{initial(2)}$. Again, this culture was grown to a final OD_{600} ≈ 2 to 5, with the actual OD_{600} value recorded as $OD_{final(2)}$. In accordance with Park *et al.* (2009), this was repeated for about 25 generations, that is approximately 6 serial passages. Importantly, after each serial passage, at each OD_{final}, cells were plated to monitor the transposition frequency, as explained below.

Once each serial passage reached an OD_{final} ≈ 2 to 5, five ten-fold serial dilutions (e.g. undiluted, 1:10, 1:100 and 1:1000) were prepared. The three least dilute cultures were plated onto FOA and G418 and YES plates. Colonies growing on these plates represented the cells that had lost the donor plasmid but retained the transposon. On the other hand, the three most dilute cultures were plated onto YES plates. Colonies growing on these plates represented all of the cells. After approximately 3 days of growth, the number of colonies on each plate was counted, and, using the dilution factor, the number of resistant cells in the original culture was determined. To calculate the transposition frequency, the number of colonies on the FOA and G418 plates was divided by the number of colonies on the YES plates. The transposition frequency was then expressed relative to the generation number.

So as to select against cells carrying the donor plasmid, cells from the final 50 ml cultures were used to inoculate a 500 ml culture of EMM + Leu + Ura + FOA + Thiamine. Finally, after approximately 24 hours of growth, a 500 ml culture of YES + FOA + G418 was inoculated to an initial OD_{600} of 0.5 and grown to an approximate final OD_{600} of 5. Overall, this selected for insertional mutations. Ten 50 ml aliquots were pelleted by centrifugation, resuspended in 50% glycerol in YES, and stored at -80°C.

2.3 *Hermes* DNA Libraries

2.3.1 DNA Library Construction

In order to create the *Hermes* DNA libraries, frozen aliquots of the cell libraries were first streaked. Genomic DNA was then extracted using the phenol/chloroform extraction method as described by Sambrook *et al.* (1989) and quantified using the Qubit dsDNA Broad Range Assay Kit (Invitrogen™, Paisley, UK).

Once extracted, DNA was sheared to an average size of 200 bp using a Covaris S2 ultrasonicator (Covaris, Woburn, Massachusetts) in a final volume of 120 µl TE (Quail *et al.* 2008). For each cycle, the parameters were:- Intensity: 5%, Duty Cycle: 10%, Cycles Per Burst: 200, Treatment Time: 60 seconds, and Power Mode: Frequency Sweeping. This was repeated for a total of 6 cycles. 1 µg of the sheared DNA was end repaired using the NEB End Repair Module (NEB, Hitchin, UK) according to manufacturer's instructions. End repaired DNA was purified with 1.8x Agencourt AMPure XP beads (Beckman Coulter, Danvers, Massachusetts) and resuspended in 50 µl sterile water.

Subsequently, forked linkers were annealed to a final concentration of 10 µM in an annealing buffer made up of 1 mM EDTA, 10 mM Tris pH 7.5, and 50 mM NaCl. This was done by heating to 90°C then cooling slowly to room temperature over 1 hour. Using the NEBNext Quick Ligation Module (NEB, Hitchin, UK), according

to manufacturer's instructions, 10 µl of annealed linkers were ligated to 25 µl of purified DNA in a 50 µl reaction for 15 minutes at 20°C. One of the linkers contained a random 10 bp sequence which acted as a unique molecular identifier (UMI) in that it was able to distinguish biologically unique insertions over PCR-derived amplifications (Kivioja *et al.* 2012). Linker ligated DNA was purified with 1.8x Agencourt AMPure XP beads and resuspended in 25 µl sterile water.

Following that, linker ligated DNA was digested with 20 units of KpnI-HF (NEB, Hitchin, UK) in a final volume of 50 µl for 2 hours. This was carried out so as to remove any of the pHL2577 donor plasmid, containing the *Hermes* transposon, which could have still been present during genomic preparations. This is because KpnI-HF cuts 21 bp away from the transposon sequence thus making it impossible for the plasmid to be amplified during the first PCR. Besides, it is a rare cutter of the fission yeast genome, and therefore it does not introduce any significant biases when detecting the insertions in the later stages of the procedure. DNA was then purified with 1.8x Agencourt AMPure XP beads and resuspended in 25 µl sterile water.

So as to enrich for fragments containing the *Hermes* transposon, DNA was then amplified with the BIOTAQ™ DNA polymerase (Bioline, Essex, UK) and pre-designed primers. Specifically, the forward primer was designed to bind to the ligated linker, whereas the reverse primer was designed to bind within the right TIR sequence. Cycle parameters for amplification were as follows: 5 cycles of 94°C for 2 minutes, 58°C for 30 seconds, 72°C for 30 seconds, 15 cycles of 94°C for 30 seconds, 58°C for 30 seconds and 72°C for 30 seconds, followed by a final extension at 72°C for 10 minutes.

Following this first round of PCR, DNA was purified with 1.8x Agencourt AMPure XP beads and resuspended into 25 µl sterile water. 2 µl were then used in a second round of PCR to attach the multiplex oligonucleotides for Illumina MiSeq sequencing (NEB, Hitchin, UK). Cycle parameters for amplification were as

follows: 15 cycles of 94°C for 4 minutes, 94°C for 20 seconds, 56°C for 20 seconds, and 72°C for 20 seconds. PCR products were first purified with 1.8x Agencourt AMPure XP beads, then size selected with 0.9x Agencourt AMPure XP beads. DNA was eluted from the beads in 25 µl sterile water. Finally, the molarity and size of the libraries were determined using an Agilent High Sensitivity DNA Chip on the 2100 Bioanalyser platform (Agilent Technologies, Santa Clara, California). Based on the size of the DNA amplicons, the DNA concentration was calculated using:

$$\text{concentration (nM)} = \frac{\text{concentration (ng/µl)}}{660\ \text{g/mol} \times \text{average library size}} \times 10^6$$

Step	Method Summary
1. DNA Extraction	phenol/chloroform extraction
2. DNA Shearing	Covaris ultrasonicator, ≈ 200 bp
3. End Repair	fragmented DNA to blunt ended DNA
4. Linker Ligation	random 10 bp as a unique molecular identifier
5. KpnI-HF Digestion	removes any residual pHL2577 donor plasmid
6. PCR 1	enriches for fragments containing the *Hermes* insert
7. PCR 2	attaches adapters for Illumina MiSeq sequencing
8. Bioanalyser	determines the molarity and size of the libraries

Table 2.1. *Hermes* DNA Libraries Method Summary. It is important to note that purification was carried out after steps 3 to 7 using 1.8x Agencourt AMPure XP beads.

Finally, 2 nM libraries were pooled together for paired-end sequencing. The MS-102-2022 MiSeq reagent kit v2 (300 cycles) (Illumina, Cambridge, UK) was used to sequence the libraries.

2.3.2 Linkers and Primers Design

Key:

- **Linkers**
- ***Hermes* Sequence**
- **Universal Primer and Complements**
- **BIOTAQ Extending <---**

A. Linker Ligation

i) DNA Fragments with Hermes Insertions

In essence, the reason why these mutation libraries are able to be sequenced is based on the fact that the *Hermes* transposon sequence is known, and therefore it is feasible to select for fragments containing the sequence. Consequently, it is possible to identify where the sequence was inserted and which regions of the genome were disrupted. Simply put, the *Hermes* insertions look like this, with the baseline dots representing the *S. pombe* genome:

```
5-.....AGAGAACTTCAACAAGCCACAGGC-[more Hermes sequence].....-3
3-.....TCTCTTGAAGTTGTTCGGTGTCCG-[more Hermes sequence].....-5
```

ii) Linker Sequences

Then, the linkers are attached at both ends:

```
Linker1-Random10mer:  5-TTCAGACGTGTGCTCTTCCGATCT-[NNNNNNNNNN]-CTCCGCTTAAGGGAC-3
Linker2:                                          3-NH2-3AmM-GAGGCGAATTCCCTG-5
```

Or, shown the other way around:

```
Linker1-Random10mer:  3-CAGGGAATTCGCCTC-[NNNNNNNNNN]-TCTAGCCTTCTCGTGTGCAGACTT-5
Linker2:              5-GTCCCTTAAGCGGAG-NH2-3AmM-3
```

In Linker 1, the underlined unpaired 24 nt sequence provides the priming site for the forward primer in the first PCR. In Linker 2, the 5' end has a phosphate group and the 3' end has an amino group which acts as a blocking group thus preventing linker-linker amplification.

With the *Hermes* insertions, the sequences look like this:

```
5-TTCAGACGTGTGCTCTTCCGATCT-[NNNNNNNNNN]-CTCCGCTTAAGGGAC...AGAGAACTTCAACAAGCCACAGG-[Hermes]...GTCCCTTAAGCGGAG-NH2-3AmM-3
                           3-NH2-3AmM-GAGGCGAATTCCCTG...TCTCTTGAAGTTGTTCGGTGTCC-[Hermes]...CAGGGAATTCGCCTC-[NNNNNNNNNN]-TCTAGCCTTCTCGTGTGCAGACTT-5
```

It is important to note that at this point each DNA fragment has a covalently bound random 10mer (from the Linker1-Random10mer). Indeed, it has two, one at each end, but only the top strand is sequenced.

B. PCR 1

PCR 1 uses two primers:

i) 1-Transposon-4NNNN

3-CTCTTGAAGTTGTTCGGTGTCC-[NNNN]-TCTAGCCTTCTCGCAGCAC-5

Or, shown the other way around:

5-CACGACGCTCTTCCGATCT-[NNNN]-CCTGTGGCTTGTTGAAGTTCTC-3

where:

- complementary to the *Hermes* right end
- same as a part of the universal primer in PCR 2

ii) Linker1-Amp

5-TTCAGACGTGTGCTCTTCCGATCT-3

In Linker1-Amp, the 24 nt sequence is the same as the underlined unpaired sequence of the Linker1-Random10mer. It is important to note that complementary sequences are only present after the first PCR 1 cycle (see next page).

PCR 1: First Cycle

```
5-TTCAGACGTGTGCTCTTCCGATCT-[NNNNNNNNNN]-CTCCGCTTAAGGGAC...AGAGAACTTCAACAAGCCACAGG-[Hermes]...GTCCCTTAAGCGGAG-NH2-3AmM-3
                                          <---  3-CTCTTGAAGTTGTTCGGTGTCC-[NNNN]-TCTAGCCTTCTCGCAGCAC-5
```

1-Transposon-4NNNN anneals to and extends the top strand. During the first cycle, there are no sequences that Linker1-Amp can anneal to, and therefore, only fragments containing *Hermes* are extended.

PCR 1: Second Cycle (and all other cycles)

Subsequent to the first cycle and extension by the 1-Transposon-4NNNN primer, the dsDNA looks like this:

```
5-TTCAGACGTGTGCTCTTCCGATCT-[NNNNNNNNNN]-CTCCGCTTAAGGGAC...AGAGAACTTCAACAAGCCACAGG-[Hermes]...GTCCCTTAAGCGGAG-NH2-3AmM-3
3-AAGTCTGCACACGAGAAGGCTAGA-[NNNNNNNNNN]-GAGGCGAATTCCCTG...TCTCTTGAAGTTGTTCGGTGTCC-[NNNN]-TCTAGCCTTCTCGCAGCAC-5
```

Therefore, the top and bottom strands can now be amplified by the 1-Transposon-4NNNN and Linker1-Amp primers respectively.

1-Transposon-4NNNN extends like so (as above):

```
5-TTCAGACGTGTGCTCTTCCGATCT-[NNNNNNNNNN]-CTCCGCTTAAGGGAC...AGAGAACTTCAACAAGCCACAGG-[Hermes]...GTCCCTTAAGCGGAG-NH2-3AmM-3
                                          <---  3-CTCTTGAAGTTGTTCGGTGTCC-[NNNN]-TCTAGCCTTCTCGCAGCAC-5
```

Linker1-Amp extends like so:

```
5-TTCAGACGTGTGCTCTTCCGATCT-3 --->
3-AAGTCTGCACACGAGAAGGCTAGA-[NNNNNNNNNN]-GAGGCGAATTCCCTG...TCTCTTGAAGTTGTTCGGTGTCC-[NNNN]-TCTAGCCTTCTCGCAGCAC-5
```

C. PCR 2

PCR 2 uses two multiplexing primers:

i) Universal Primer

3-**TCTAGCCTTCTCGCAGCACATCCCTTTCTCACATCTAGAGCCACCAGCGGCATAGTAA**-5

where:

- same as a part of the 1-Transposon-4NNNN primer in PCR 1

ii) Primer Index N (e.g. Index 1)

5-**CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTC**-3

where:

- same as a part of Linker1-Amp

PCR 2: Cycles

So, the dsDNA looks like this:

5-**TTCAGACGTGTGCTCTTCCGATCT-[NNNNNNNNNN]-CTCCGCTTAAGGGAC**...**AGAGAACTTCAACAAGCCACAGG-[NNNN]-AGATCGGAAGAGCGTCGTG**-3
3-**AAGTCTGCACACGAGAAGGCTAGA-[NNNNNNNNNN]-GAGGCGAATTCCCTG**...**TCTCTTGAAGTTGTTCGGTGTCC-[NNNN]-TCTAGCCTTCTCGCAGCAC**-5

Universal Primer extends like so:

```
5-TTCAGACGTGTGCTCTTCCGATCT-[NNNNNNNNNN]-CTCCGCTTAAGGGAC...AGAGAACTTCAACAAGCCACAGG-[NNNN]-AGATCGGAAGAGCGTCGTG-3
                                                          <---  3-TCTAGCCTTCTCGCAGCACATCCCTTTCTCACATCTAGAGCCACCAGCGGCATAGTAA-5
```

Primer Index N extends like so:

```
5-CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTC-3  --->
                                  3-AAGTCTGCACACGAGAAGGCTAGA-[NNNNNNNNNN]-GAGGCGAATTCCCTG...TCTCTTGAAGTTGTTCGGTGTCC-[NNNN]-TCTAGCCTTCTCGCAGCAC-5
```

2.4 Bioinformatics Pipeline

When the MiSeq run was complete, FASTQ files were generated by the MiSeq Reporter, which is a pre-installed software on MiSeq sequencers. FASTQ files contained sequence reads and their quality scores, excluding clusters that did not pass filter. So as to analyse the raw FASTQ files, a custom bioinformatics pipeline was created. For each library, the pipeline encompassed the following main steps:

A. Processing Read 1
B. Processing Read 2
C. Mapping
D. Processing SAM Files
E. Determining *Hermes* Insertion Counts

A. Processing Read 1

Read 1 Architecture: **[4mer][*Hermes*][Genome]**

The FASTQ file for Read 1 was first scanned for the read architecture above. Then, the [4mer] was trimmed off by the fastx_trimmer, a command line tool available within the FASTX-Toolkit (Hannon Lab, Cold Spring Harbor Laboratory, New York, USA).

Command: fastx_trimmer [-h] [-f N] [-l N] [-z] [-v] [-i INFILE] [-o OUTFILE]

So as to identify and keep reads with the [*Hermes*] insertions, excluding those within the pHL2577 donor plasmid, the Reaper program was used. Reaper is one of the three standalone tools available within the Kraken suite, with the other two being Tally and Sequence Imp. Reaper is used for demultiplexing, trimming and filtering short read sequencing data. It can handle barcodes, strip low quality bases, and trim adapter sequences. It is fast because it is written in C and it uses very little memory (Davis *et al.* 2013).

Command: reaper -i sample.fastq -meta sample.txt -geom no-bc -5p-sinsert l/e[/g[/o]] --fastqx-out

In Reaper, geometry-dependent read processing is possible, with the three supported geometries being no-bc, 3p-bc, and 5p-bc. Such read processing depends on the absence or presence of barcodes and on the geometry of the read. In this context, the geometry refers to the read design, that is a description of what a read looks like. For this data, the most suitable geometry was deemed to be no-bc (no barcode). Now, if the reads are not barcoded, it is possible to run the program with or without a metadata file. If the metadata file is used, however, as it was in this case, it requires the 3-prime adapter sequence (3p-ad) and the tabu sequence (tabu). For this data, the tabu sequence was set to the first 200 nt of the pHL2577 donor plasmid sequence, and reads contaminated with it were consequently discarded. Finally, the command line was given the --fastqx-out option which resulted in the inclusion of a new field on the identifier line, specifically, the record offset number.

Using the --fastqx-in option, Tally then identifies this number and utilises it to pair up the processed reads. Tally, one of the other standalone tools available within the Kraken suite, removes redundancy from sequence files by collapsing identical reads to a single entry while recording the number of instances of each. However, it can also re-pair reaper-processed files without tallying, as was done in this case.

Command: tally -i out1.gz -j out2.gz -o out1.unique.gz -p out2.unique.gz --fastqx-in --no-tally --with-quality

B. Processing Read 2

Read 2 Architecture: **[10mer][Linker][Genome]**

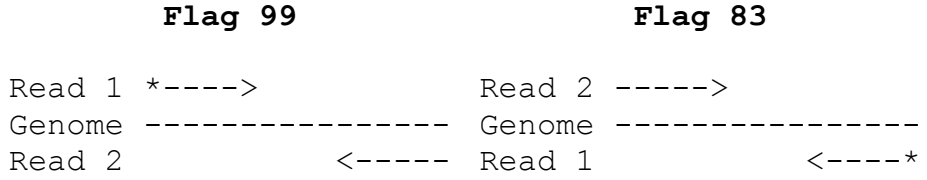
To process Read 2, a Perl script was written by Dr Daniel Jeffares to detect and exclude duplicate reads based on the [10mer] and the first 5 nt of the [Genome]. Then, the output was tallied with Read 1 and the [10mer][Linker] was trimmed using fastx_trimmer.

C. Mapping

Following Tally processing, the collated files were processed with the Burrows Wheeler Alignment (BWA) software package, used for mapping low-divergent sequences against a large reference genome. BWA consists of three algorithms but the one used in this pipeline was the BWA-MEM algorithm. This is because BWA-MEM is fast, more accurate, and can handle longer sequences ranging from 70 bp to 1 Mbp (Li and Durbin 2009). Using the BWA-MEM algorithm, the paired-end reads were aligned to the *S. pombe* reference genome and to the pHL2577 donor plasmid, with the final alignment being outputted in the SAM (Sequence Alignment/Map) format.

D. Processing SAM Files

SAM files were then converted to BAM files which are the binary version of SAM files. BAM files were analysed with SAMtools; an open source suite of utilities used to manipulate alignments, including sorting, merging, indexing and generating alignments in a per-position format (Li *et al.* 2009). Indeed, upon further analysis, the reads were flagged with a number, either 99, 147, 83, or 163, which meant that the reads were mapped in the correct orientation and within the insert size. Based on the flag information, reads with flag 99 and flag 83 were considered to be the only ones relevant to the read architecture.



Next, the BAM files were sorted based on genomic position. SAMtools were used, for flag 99 and flag 83 reads with a mapping score of at least 30, to separate both the chromosome number and the insertion positions. For flag 83 reads, however, a specific Perl script had to be written because a BAM file only states the position at the start of the read and flag 83 reads have the insertion at the rightmost end.

With this in mind, the Perl script was written so as to output the rightmost position of the flag 83 reads. Subsequently, the files for flag 99 and flag 83 reads were modified to include the + and – signs to indicate respectively which insertions came from the forward strand and which insertions came from the reverse strand. In addition, these +/– signs were important in estimating unique insertions, since insertions found on the same chromosome but on different strands were considered to be unique events. Finally, the files for flag 99 and flag 83 reads, containing the chromosome number and the insertion positions, were merged.

E. Determining *Hermes* Insertion Counts

Using a Perl script, merged BAM files were finally processed to determine the total number of unique insertions within the genome. Outputted plain text files were then loaded into the R statistical package to look for any biologically meaningful patterns (see Chapters 4 and 5).

2.5 Chronological Lifespan Assay

In Chapter 2.2.2, detailing the construction of *Hermes* cell libraries, the last step was to select for insertional mutations in a 500 ml culture of YES + FOA + G418. In effect, that was the end of the log phase libraries and the beginning of the chronological lifespan (CLS) assays, as illustrated in Figure 2.2. In CLS assays, the aim is to monitor the loss of cell viability in a culture over time. In this case, the start of the assay (time point 0, or t0) was established when the culture reached stationary phase, specifically, when the optical density remained constant following two cell cycles. Here, time points were taken at 24-hour intervals.

First, at each time point, the number of cells per µl in the culture was measured using a haemocytometer. In individual microcentrifuge tubes, a serial dilution was then carried out to achieve a concentration of 2 cells per µl. 3x 100 µl of this dilution were plated in three separate YES plates. Consistently, 5 sterile glass beads were used to evenly spread the 100 µl (≈ 200 cells) over the entire surface

of each plate. Following two days incubation, CFUs, or colony forming units, were counted using a manual colony counter. Ultimately, taking into account the dilution factor, the average number of colonies per ml was calculated and used as a measure of survival.

Overall, this procedure was repeated for each time point, each time adjusting the dilution to attain a concentration of 20 to 300 colonies per plate. In addition, at each time point, normalisation was performed with respect to time point 0 until the culture reached 0.1% of the initial survival. So as to illustrate this pattern of decline over time, a survival curve was plotted. Importantly, at each time point, aliquots were taken, of which a selected few were processed for Illumina MiSeq sequencing.

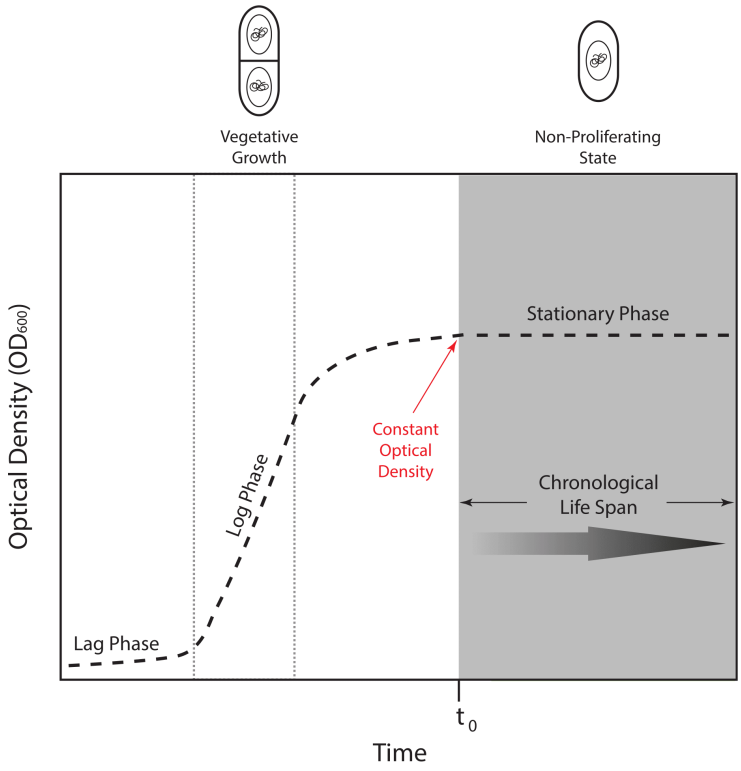


Figure 2.2. Distinguishing Log Phase and Ageing Libraries. For the log phase *Hermes* libraries, cells were harvested before time point 0. t0 for the ageing libraries was established when the cell culture reached stationary phase, that is, when the OD_{600} remained unchanged. (Design: Dr Graeme C. Smith.)

2.6 Fluorescence Microscopy

On a microscope slide, 2 µl of cells with chromosomal insertions were stained with 2 µl of calcofluor, 2 µl of DAPI, and 2 µl of anti-fading mounting media. In fluorescence microscopy, calcofluor stains the cell wall, DAPI stains the nucleus, and the mounting media prevents the fluorophores from dying. Carefully, one drop of immersion oil was then placed onto the coverslip.

Chapter 3 CREATING *HERMES* LIBRARIES

3.1 Method Overview

In this chapter, the focus will be on the creation of the *Hermes* libraries and the design and optimisation processes behind it. Overall, the construction of the libraries can be divided into two stages, namely, cell and DNA libraries. Briefly, the approach for creating *Hermes* libraries starts with sequential lithium acetate transformations to introduce the donor and the expression plasmids into a strain of fission yeast. CFUs, or colony forming units, containing both plasmids are grown in liquid selective medium lacking leucine and uracil to select for the plasmids. So as to produce the cell libraries, cultures are first grown in a series of flasks for a total of approximately 25 cell generations, and then ultimately, cells with chromosomal insertions are selected for. For the cellular ageing screen, a chronological lifespan (CLS) assay ensues. In the second stage of construction, DNA extracted from the cell libraries is sheared with a Covaris high performance ultrasonicator, then end repaired, linker ligated, digested with KpnI-HF to remove any residual pHL2577 donor plasmid, enriched for a region of *Hermes* right TIR and its flanking genomic DNA by a first PCR, and then a second PCR to attach multiplex oligonucleotides for Illumina sequencing (Figure 3.1).

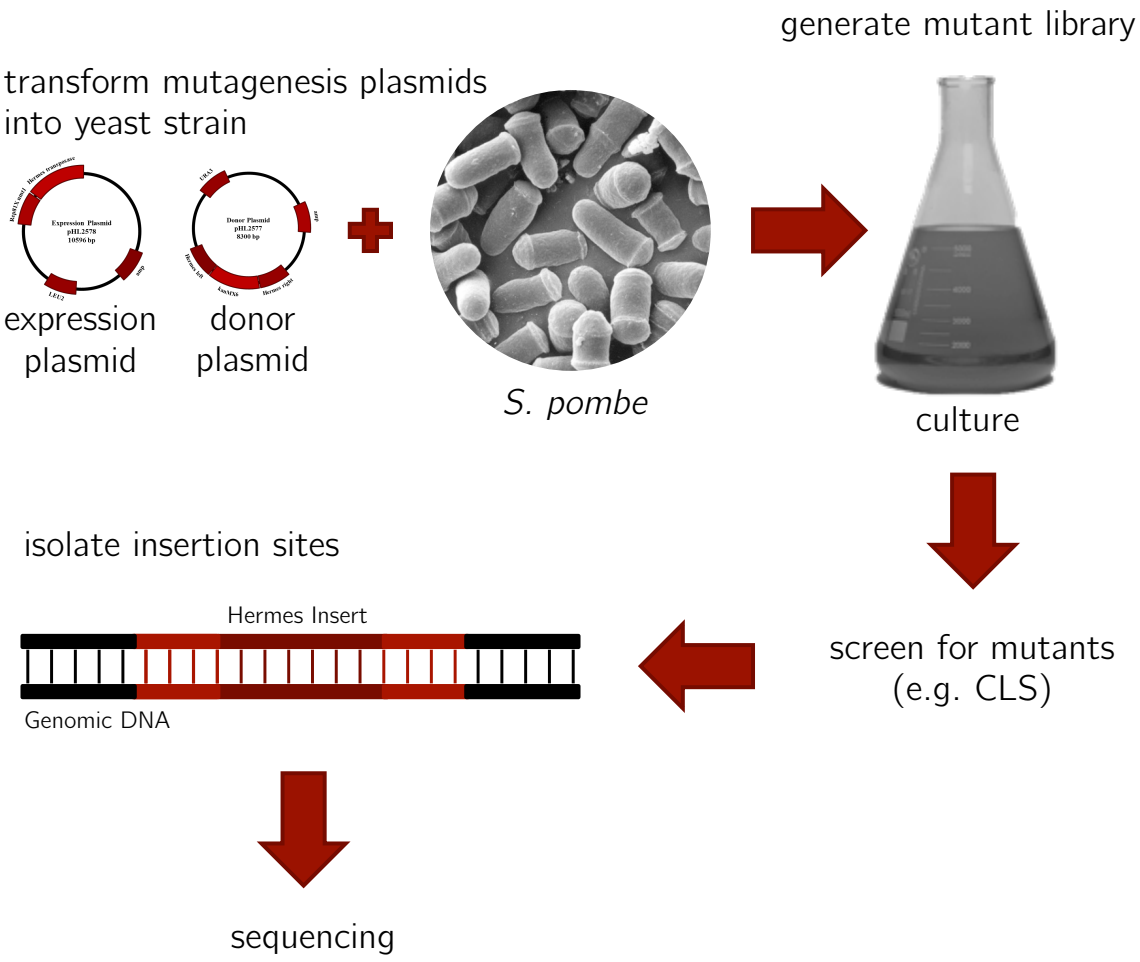


Figure 3.1. Method Summary. *Hermes* libraries construction can be divided into two stages, namely, cell mutant and DNA libraries (see main text for details).

3.2 *Hermes* Cell Libraries Optimisation

In the earlier stages of this research investigation, the aim was to create complex *Hermes* libraries on a large scale and in an efficient manner. To this end, several of the steps in the procedure were attempted multiple times using either a different approach or different techniques. For example, in addition to the standard lithium acetate (LiAc) method used to transform the donor and the expression plasmids (Forsburg and Rhind 2006), two other techniques were tested. Specifically, these involved generating either cryopreserved LiAc competent cells (Suga and Hatakeyama 2005) or protoplasts, which are cells lacking cell walls (Flor-Parra *et al.* 2014). In terms of efficiency, using competent cells proved to be more cost and time effective when compared to both the LiAc and the protoplast

techniques. However, overall, more positive colonies resulted from the standard LiAc method, therefore explaining its use here in this investigation.

S. pombe cells transformed with both the donor and the expression plasmids were then grown for an average of 25 generations, that is about 6 serial passages. Figure 3.2 shows that for the ten cell libraries created within this research, the generation number ranges from 19.3 (for Libraries 1 and 2, or LG1 + LG2) to 28.7 (for Libraries 7 and 8, or LG7 + LG8). In a comparison of the two pools, it appears that the generation number, within these limits, does not have an effect on the average proportion of unique insertion sites. Indeed, this is in accordance with Evertts *et al.* (2007), stating that an average of 25 cell generations is ideal for generating mutation libraries, because within that time, a considerable number of cells with insertions are produced while keeping the occurrence of double insertions to a minimum.

In this procedure, the generation of cells with insertions was monitored by a transposition assay, where serially diluted cells were plated onto two different types of plates, (i) YES + FOA + G418 to determine the number of cells that lost the donor plasmid but had a genomic insertion, and (ii) YES to determine the total number of cells within the culture. Subsequently, the transposition frequency was calculated by dividing the number of CFUs, or colony forming units, on the FOA and G418 plates by the number of CFUs on the YES plates. Here, results show that for the ten cell libraries, the transposition frequency ranges from 0.02% to 0.17%, with an average of 0.06% of cells having a chromosomal insertion. Initially, this seems like a low percentage, however, when compared to the published research from Park *et al.* (2009), where the transposition frequency ranges from 1.50% to 2.75%, this result is not that dissimilar, especially if the approach for the transposition assay is taken into account. Park *et al.* (2009) used EMM + FOA instead of YES plates to calculate the transposition frequency. Plating onto EMM + FOA plates determines the number of cells that lost the

donor plasmid. Plating onto YES plates, on the other hand, determines the total number of cells within the culture. In theory, this number is higher than the number of cells that lost the donor plasmid, which would therefore explain the somewhat lower values in this investigation.

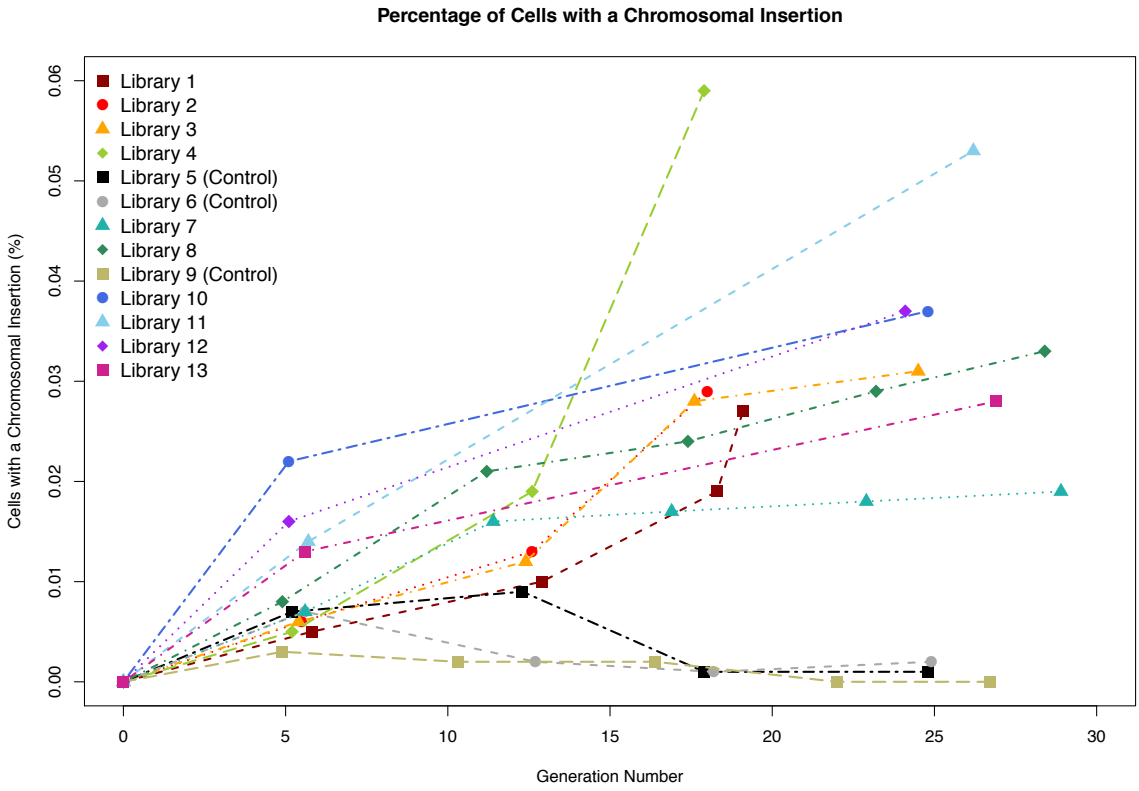


Figure 3.2. Quantitative Transposition Assay. For this assay, the transposition frequencies were calculated after each serial passage (that is, after about 4 to 6 cell generations), except for Libraries 10 to 13, where the frequencies were calculated after the first and last passages only. For a better graphical representation, frequencies > 0.06%, specifically for Libraries 2 (0.160%) and 4 (0.165%), are not shown. Here, cells that do not express the transposase enzyme were used as negative controls. In these controls, no transposition events were observed.

In another attempt to improve complexity, the two flasks for Libraries 7 and 8, containing cells accumulating insertions, were each split into five flasks. Basically, this was done to minimise competition and therefore any bias towards more competitive insertion mutants. However, comparing these two libraries with the

other libraries, it seems that this approach does not influence the average proportion of unique insertion sites at all. In one last effort, fluoroorotic acid, or FOA, used in the final steps to select against cells carrying the donor plasmid, was altogether eliminated. It is well known that plasmids can be passively lost from cells in the absence of continued drug selection. Libraries VT-2 and VT-3 tested this, showing, however, that this strategy does not have an effect either. Indeed, for Libraries VT-2 and VT-3 respectively, only 73% and 68% of sequence reads were mapped to the *S. pombe* genome, which is less than the average (90%) for libraries that utilised FOA.

3.3 *Hermes* DNA Libraries

3.3.1 Design

In general, developing libraries for Illumina sequencing involves fragmenting the DNA (either enzymatically or mechanically), repairing the fragmented DNA, ligating linkers, size selecting, and then carrying out a single PCR step to amplify the library (Quail *et al.* 2008). Now, while this is suitable for most libraries, there are instances where modification is required. In fact, to adapt this workflow to the *Hermes* libraries, additional steps had to be included, such as the KpnI-HF digestion step, which gets rid of any residual donor plasmid containing the transposon sequence. Moreover, a preceding PCR step (dubbed PCR 1) had to be added, with the aim of enriching for fragments containing the *Hermes* insert. Ultimately, this was followed by a second PCR step (dubbed PCR 2) where the Illumina adapters were added to allow multiplex sequencing.

In this study, DNA was first extracted from the genome of the insertion mutants. Extracted DNA was sheared with a Covaris ultrasonicator and end repaired to blunt the heterogeneous ends. In reference to Figure 3.3, linkers were then ligated to the terminal ends of the sheared and end repaired templates. Here, one of the linkers incorporated a random 10 bp sequence (dubbed a 10mer) which acts as a

unique molecular identifier (UMI), in that it helps to distinguish unique insertions from those derived from the subsequent PCR amplifications.

Basically, before the PCR, each DNA molecule is simplistically assumed to have arisen from a single cell. It can thus be inferred that the count of an insertion at position *p* is directly proportional to the number of cells harbouring an insertion at position *p*. So as to determine the true number of original DNA molecules, and therefore, to interpret the read count in a more accurate and representative manner, a molecular barcode was used. In essence, the concept of molecular barcoding is that each original DNA molecule is attached to a unique barcode, such that reads that have different barcodes represent different original molecules, whereas reads that have the same barcode are the result of PCR amplification of one original molecule. In fact, when processing the data after sequencing, reads with the same UMI were assumed to have arisen from a PCR duplication, while reads with different UMIs were assumed to have arisen from single cells and were therefore present in the initial, unamplified sample. It is important to note that molecular barcoding does not prevent PCR duplication from occurring. Rather, it provides a solution to track duplicates and treat them accordingly for downstream analysis (Peng *et al.* 2015).

In reference to Chapter 2.3.2 and Figure 3.3, it is also important to understand that one of the linkers was designed in such a way so as to provide the priming site for the forward primer in the subsequent PCR step (PCR 1). In addition, for PCR 1, the reverse primer was designed to bind to the *Hermes* right TIR, thus allowing the boundary between the transposon and the genomic region disrupted by its insertion to be amplified. Following the first PCR, a second PCR (PCR 2) was carried out so as to attach the multiplex oligonucleotides. Ultimately, before sequencing, the quality, quantity and size of the libraries were checked on the Agilent 2100 Bioanalyser system; electropherograms with clean, sharp peaks in the 200 to 300 bp region indicated that the libraries were suitable for sequencing.

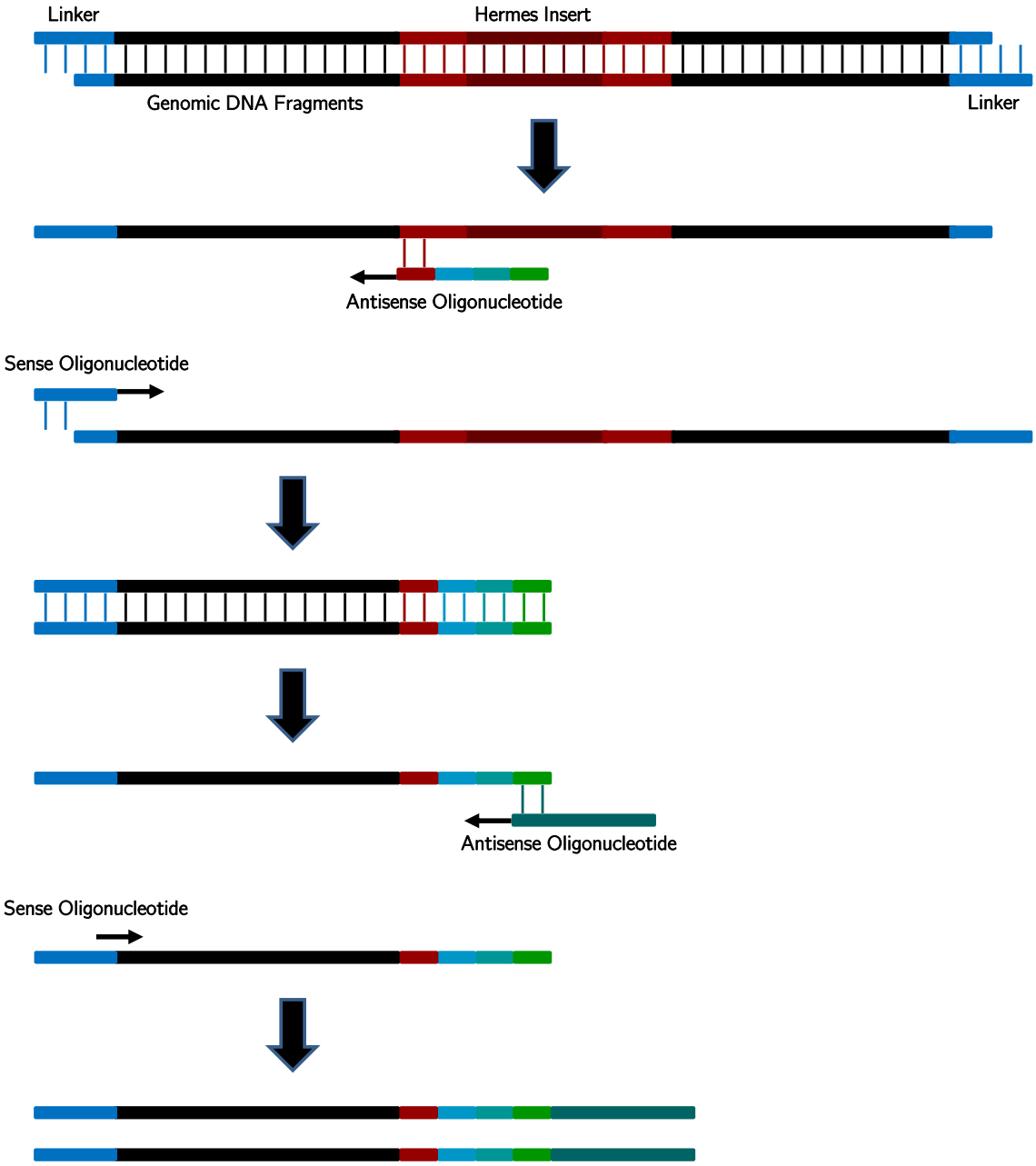


Figure 3.3. Schematic Diagram of the *Hermes* DNA Libraries Workflow. Genomic DNA disrupted by *Hermes* transposon insertions is extracted, sheared, end repaired, and linker ligated at both terminal ends; one of the ligated linkers incorporates a random 10 bp sequence and provides the priming site for the forward primer in PCR 1. In PCR 1, the reverse primer binds to the *Hermes* right terminal inverted repeat (TIR), therefore enriching for fragments containing the transposon. In PCR 2, multiplex oligonucleotides are added so as to allow the fragments to bind to the flow cell during sequencing.

3.3.2 Optimisation

Similar to the *Hermes* cell libraries, the DNA libraries were optimised in an attempt to increase complexity. For DNA extraction, for example, three different methods were tested, namely the phenol/chloroform extraction procedure (Sambrook *et al.* 1989), the MasterPure™ Complete DNA Purification Kit, and an in-house technique exploiting the QIAGEN 100/G genomic tips and glucanex, a yeast lytic enzyme from *Trichoderma harzianum* (Petit *et al.* 1994). Overall, the phenol/chloroform extraction method proved to be the most effective in terms of yield. Next, for Illumina sequencing, mechanical shearing remains the method of choice for achieving high sensitivity and unbiased results. In this investigation, a Covaris S2 ultrasonicator was used to shear the DNA, owing to the fact that it is accurate, quick, reproducible, and the results are without GC or temperature bias. Optimisation was not performed for the end repair and linker ligation steps.

In any scientific study, experimental optimisation also means having reliable negative and positive controls which ensure the validity of the results. In this study, a good negative control was hypothesised to be genomic DNA from the JB980 strain lacking the *Hermes* insertion. In contrast, a suitable positive control was hypothesised to be DNA of expected length and known to contain the *Hermes* insertion. In total, three positive controls were used in this investigation: (a) the pHL2577 donor plasmid cut with the BseYI restriction enzyme, (b) same as (a) but spiked, at different concentrations, into genomic DNA from the JB980 strain, and (c) genomic DNA extracted from single transformed colonies then digested with two restriction enzymes – MseI which is a common genome cutter, followed by KpnI-HF so as to remove the donor plasmid.

In general, these controls were used to validate a number of steps within the workflow, specifically that the linkers anneal to the appropriate regions, that digestion with KpnI-HF eliminates interference from the donor plasmid, and also, that the conditions and primers chosen for PCR 1 amplification are optimal. In

regard to the digestion step, it is important to note that, in theory, KpnI-HF cuts 21 bp away from the transposon sequence thus making it impossible for the plasmid to be amplified during the first PCR. So as to test this, the two positive controls (a) and (b) described above were either treated with or without KpnI-HF. Lanes 7 and 9 in Figure 3.4 show that the untreated controls produced the expected 811 bp band (described below), as opposed to the treated controls, which did not produce any bands at all, thus confirming the role of the KpnI-HF restriction enzyme in removing any residual donor plasmid.

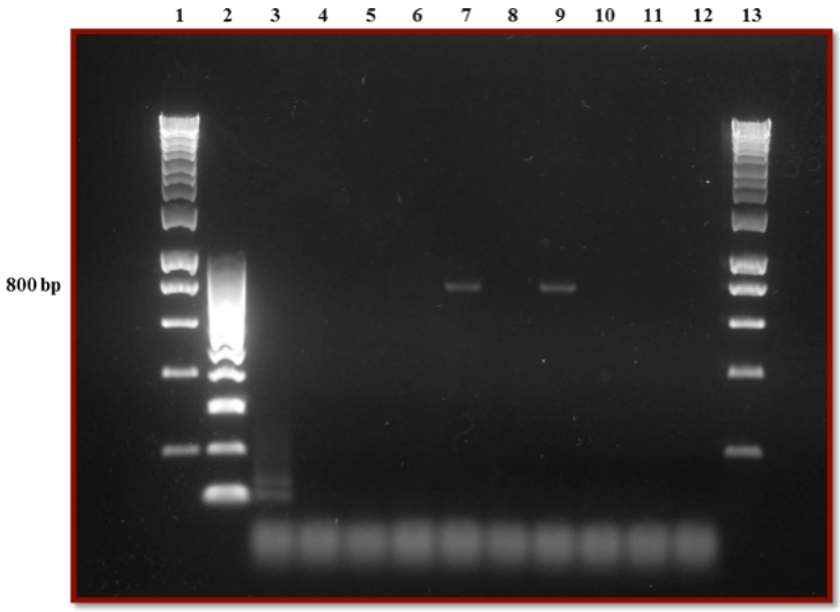


Figure 3.4. Negative and Positive Controls. 2.0% TBE / agarose gel visualised with the nucleic acid stain GelRed™. Lanes 1 and 13: HyperLadder™ 1 kb. Lane 2: HyperLadder™ 100 bp. Lane 3: Two Single Transformants from Libraries 1 and 2. Lanes 4 and 5: JB980. Lane 6: JB980 + spike. Lane 7: JB980 + spike – KpnI. Lane 8: BseYI cut pHL2577. Lane 9: BseYI cut pHL2577 – KpnI. Lane 11: pHL2577. Lanes 10 and 12: Water.

In regard to PCR 1, all four controls proved to be useful. In fact, when the negative control was treated in the same way as the libraries, no band was visible after PCR 1, as can be seen in lanes 5 and 10 in Figure 3.4. In contrast, for the two positive controls (a) and (b) described above, when these were untreated with

KpnI-HF, the expected 811 bp band was observed. In both of these controls, the 8300 bp donor plasmid was cut with the BseYI restriction enzyme. BseYI cuts the plasmid in two positions, between 2776 bp and 2781 bp, and between 7116 bp and 7121 bp (Figure 3.5). During PCR 1, the reverse primer binds to the *Hermes* right TIR, found between 3498 bp and 3514 bp in the plasmid. It was therefore expected for the amplified fragment to be 717 bp (2781 bp to 3498 bp) in size. However, when this is added to the 49 bp linker sequence and the 45 bp sequence derived from the reverse primer, it adds up to 811 bp, which is equal to the bands seen in lanes 7 and 9 in Figure 3.4. Figure 3.6 illustrates the results from control (b) therefore providing an additional layer of confidence. In regard to control (c), this was hypothesised to be the better positive control. In theory, and as shown in lane 3 in Figure 3.4, this control should contain only one insertion, that is only one sequence, and thus, the PCR 1 product should be observed as a single band representing that one transposon sequence. Incidentally, this is in contrast to the DNA extracted from the libraries, which are represented by different length products owing to the random shearing and the random distribution of insertions across the genome.

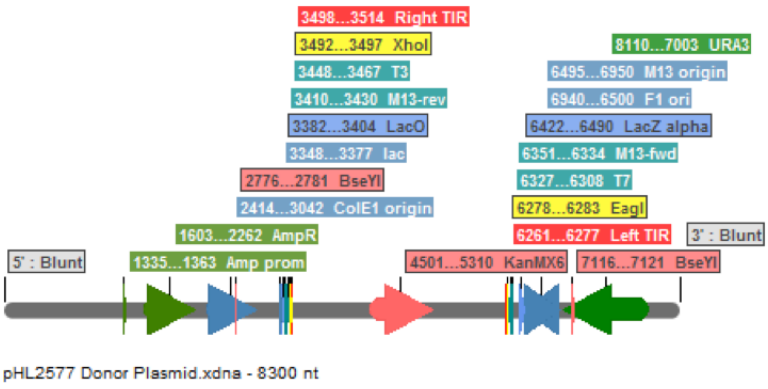


Figure 3.5. Features and Unique Sites of the Linearised pHL2577 Donor Plasmid. BseYI cuts between 2776 bp and 2781 bp and between 7116 bp and 7121 bp.

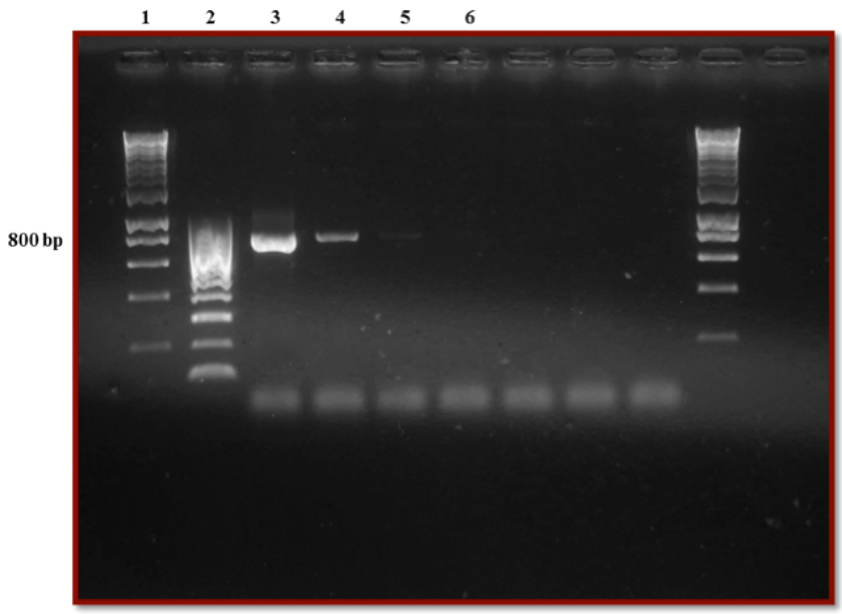


Figure 3.6. Spike Control. 2.0% TBE / agarose gel visualised with the nucleic acid stain GelRed™. Lane 1: HyperLadder™ 1 kb. Lane 2: HyperLadder™ 100 bp. Lane 3: 100 pg pHL2577 per 100 ng genome. Lane 4: 10 pg pHL2577 per 100 ng genome. Lane 5: 1 pg pHL2577 per 100 ng genome. Lane 6: Water.

In a last effort to improve complexity, a multiplexing scheme using 96-well plates was introduced, where every µg of DNA extracted from the cell libraries was sheared, end repaired, linker ligated, digested with KpnI-HF, amplified through a first and second PCR, and then sequenced. In this way, the majority of insertions present in the cell libraries were represented in the sequence reads. Overall, this was the approach that most significantly increased the complexities of the *Hermes* libraries.

3.4 Bioinformatics Pipeline

It has been a decade since the introduction of next generation sequencing (NGS) technology. So as to cater for the rapid growth of demand, there are now over thirty companies offering NGS products and services, however, Illumina is the dominant supplier (Baker 2010). In fact, the Illumina MiSeq was the platform used in this research investigation; the MiSeq desktop sequencer allows multiplexing of different samples and targets small genomes such as the genome of fission yeast.

In brief, Illumina's sequencing by synthesis (SBS) technology uses four fluorescently-labelled nucleotides to sequence DNA templates that are immobilised on the surface of a so-called flow cell; a flow cell is a thick glass slide with channels or lanes on which cluster generation occurs (Quail *et al.* 2012). Ultimately, SBS results in highly accurate base-by-base sequencing that eliminates sequence-context specific errors, enabling robust base calling across the genome. Overall, this offers an attractive approach for localising transposon insertion sites.

With Illumina, and most NGS platforms, it is possible to specify the length of the reads and whether they are single-end or paired-end. Here, read length refers to the number of base pairs that are read at a time, for example, one read might consist of 50 bp, 100 bp, or more. However, longer reads provide more reliable information about the location of specific base pairs. In fact, having longer reads addresses a common challenge that arises during sequencing, that is that the same read sequences can appear in multiple places within the genome. In this study, 2x 75 bp reads (using a 300-cycle MiSeq Reagent Kit) were generated for each of the *Hermes* libraries that were sequenced.

In regard to whether the reads are single-end or paired-end, it is known that paired-end reading is more effective than single-end sequencing in resolving structural rearrangements such as gene insertions. In addition, paired-end reads are more likely to align to a reference genome, therefore improving the quality of the entire dataset. In single-end reading, the sequencer reads a fragment from only one end to the other, whereas in paired-end reading it starts at one read, finishes this direction at the specified read length (75 bp), and then starts another round of reading from the opposite end. In doing so, it improves the ability to ascertain the location of various reads in the genome. In this study, paired-end reading was used for all sequenced libraries.

Once the MiSeq runs were complete, Illumina's BaseSpace Sequence Hub was used to assess and manage the sequencing data. Specifically, it was utilised to

determine the total number of reads, the quality score distribution (% ≥ Q30), and the percentage of clusters passing filter (%PF). In BaseSpace, %PF is an indication of signal purity from each cluster, whereas a quality score is a prediction of the probability of an error in base calling. For example, for base calls with a quality score of Q30, one base call in 1,000 is predicted to be incorrect. BaseSpace was also used to download the raw FASTQ files which were then double checked with the FastQC tool (Babraham Institute, Cambridge, UK). FastQC is a quality control application for high throughput sequence data, which assesses the overall quality of the runs, and spots any potential biases or problems. It calculates, amongst others, the per base sequence quality and the per base and per sequence GC content.

Raw FASTQ files were then processed. So as to identify and keep reads with chromosomal *Hermes* insertions, while excluding those within the donor plasmid, the Reaper program was used on Read 1. For Read 2, a Perl script was written to detect and exclude duplicate reads based on the random 10mer and the first 5 nt of the genome (Figure 3.7). Together, these eliminated a large proportion of reads. Incidentally, in a pilot MiSeq run not included in this work, a comparison was made between two libraries differing only in their KpnI-HF treatment. Here, libraries that were not digested with the restriction enzyme had a higher proportion of reads within the donor plasmid, thus validating the controls in Chapter 3.3.2 and the use of KpnI-HF.

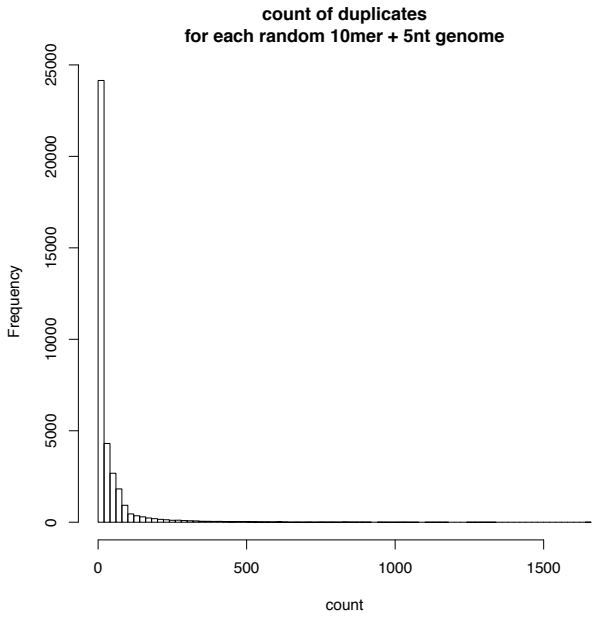


Figure 3.7. Filtering Duplicate Reads. For Read 2, duplicate reads were discarded based on the 10mer introduced during linker ligation and the first 5 nt of the *S. pombe* genome. Here, this frequency histogram illustrates the proportion of reads, for one of the libraries, that were filtered out.

Following read filtering (with Reaper) and re-pairing (with Tally), the BWA-MEM algorithm was used to align the reads to both the reference genome and the plasmid, with the final alignment being outputted in the SAM format. SAM files were converted into the more storage efficient BAM files. SAMtools then allowed the reads to be flagged and those reads with 99 and/or 83 flags were selected for. It is important to note that for a read, the flag is representative of its orientation, that is, it shows whether the read is on the forward or positive strand (Read 1) or whether it is on the reverse or negative strand (Read 2). In addition, flags provide information about the uniqueness of the read, and in fact, at this stage in the pipeline, it was possible to count the number of unique insertion sites.

3.5 Sequenced *Hermes* Libraries

Overall, the approach of coupling transposon mutagenesis with NGS technologies is quite powerful. In fact, in this research investigation, this technique was used to describe all the functional elements in the fission yeast genome, with a special

focus on unknown and/or non-protein-coding regions, such as long non-coding RNAs (more detail in Chapters 4 and 5). Using *Hermes* transposon mutagenesis, the principle was to saturate the genome with insertions. Intuitively, since insertions in functional regions will kill cells or slow their growth, selection for fit cells will result in a lack of insertions in functional regions. In bacteria, but using the *Himar1* mariner transposon (Chao *et al.* 2013), this approach proved to be effective. However, in eukaryotes, it has not been exploited as well yet. Indeed, while there are genome-wide analyses of *Hermes* transposon insertions in budding (Gangadharan *et al.* 2010) and in fission yeast (Guo *et al.* 2013), neither one describes non-protein-coding regions in detail.

So as to address this, fourteen, independent, log phase *Hermes* cell libraries were generated. However, since five of these were pooled, nine log phase libraries were sequenced in total, with some sequenced twice (Table 3.1). Together, these add up to 923,235 unique insertions, that is, 1 insertion per 13 nt of the genome (Table 3.2). Moreover, two of the log phase libraries were aged via a chronological lifespan assay, and so, ageing *Hermes* libraries were also generated and sequenced. For an exploration of both the log phase and the ageing datasets, refer to Chapters 4 and 5 respectively.

	Libraries	Index	Total Reads	Total Counts	Total Sites	Proportion Unique
Log Phase Libraries	LG1 + LG2 †	LG.1.log	1,507,513	966,275	113,152	0.117
		LG.6.log	5,375,975	3,302,294	193,795	0.059
	LG3 + LG4 †	LG.2.log	1,770,789	1,126,182	110,363	0.098
		LG.7.log	6,718,097	4,098,752	197,754	0.048
	LG7 + LG8 †	LG.3.log	897,124	668,780	71,681	0.107
		LG.8.log	3,342,831	2,360,090	130,796	0.055
	LG10 + LG11	LG.11.log	2,074,327	1,322,780	134,930	0.102
		LG.23.log	4,842,477	2,861,102	172,390	0.060
	LG12 + LG13	LG.12.log	1,983,459	1,120,068	154,887	0.138
		LG.24.log	5,233,867	2,640,987	214,128	0.081
	HL-B †	HL.4.log	1,249,659	1,024,078	241,870	0.236
		HL.9.log	4,984,174	3,858,776	470,980	0.122
	HL-D †	HL.5.log	1,275,625	984,723	175,482	0.178
		HL.10.log	5,411,286	3,884,519	348,480	0.090
	VT-2	VT.21.log	2,489,588	494,817	15,754	0.032
	VT-3	VT.22.log	3,210,462	455,007	17,399	0.038
Ageing Libraries	LG1 + LG2 (t0)	LG.13.age0	2,126,418	1,388,330	158,538	0.114
		LG.25.age0	4,769,942	2,915,756	215,197	0.074
	LG1 + LG2 (t2)	LG.14.age2	2,799,595	1,684,623	185,734	0.110
		LG.26.age2	6,551,428	3,399,150	253,426	0.075
	LG1 + LG2 (t4)	LG.15.age4	2,437,930	1,278,106	156,548	0.123
		LG.27.age4	5,931,849	2,363,412	205,611	0.087
	LG1 + LG2 (t6)	LG.16.age6	2,090,620	712,865	126,623	0.178
		LG.28.age6	5,528,095	1,191,659	161,731	0.136
	LG3 + LG4 (t0)	LG.17.age0	1,876,680	1,252,614	143,209	0.114
	LG3 + LG4 (t2)	LG.18.age2	1,854,546	1,148,851	142,657	0.124
	LG3 + LG4 (t4)	LG.19.age4	1,378,278	659,771	103,936	0.158
	LG3 + LG4 (t6)	LG.20.age6	2,033,675	536,334	78,617	0.147

Table 3.1. Summary of the Log Phase and Ageing *Hermes* Libraries. From left to right, the *libraries* are labelled according to the initials of the researcher who created them, LG for Ms Leanne Grech (University College London, UK), HL for Dr Henry L. Levin (NICHD, NIH, Bethesda, USA), and VT for Dr Victor Álvarez Tallada (Universidad Pablo de Olavide, Sevilla, Spain). † marks the libraries that were not multiplexed. In the second column, the *index* provides a guideline for the text files stored on the large.cs.ucl.ac.uk server path /SAN/bahlerlab/hermes/insertion-data. *Total Reads* presents the total number of paired-end sequence reads in the FASTQ files. Next, the *Total Counts* and *Total Sites* calculate the total number of insertion counts and sites, where a site refers to a single insertion, and the count refers to the number of times that insertion is present within the libraries. *Proportion Unique* is the total sites divided by the total counts.

Libraries	Total Insertion Counts	Mean Insertion Counts	Unique Insertion Sites (UIS)	Genome Saturation (UIS/nt)
all log phase	25,798,137	27.94	923,235	13.60
ageing t0	5,495,541	16.67	329,681	38.02
ageing t2	6,159,442	16.95	363,483	34.48
ageing t4	4,251,766	14.91	285,197	43.91
ageing t6	2,417,815	11.16	216,659	57.77
all libraries	44,122,701	40.17	1,098,430	11.44

Table 3.2. Summary of the *Hermes* Insertion Counts and Sites. Here, a *site* refers to a single insertion, and the *count* refers to the number of times that insertion is present within the libraries. In the far right column, *genome saturation* is calculated as the total number of unique insertion sites (UIS) per nucleotide of the genome. In other words, it is calculated as 12,600,000 bp divided by the UIS. 12.6 Mbp is the size of the *S. pombe* genome. For all log phase *Hermes* libraries, for example, this works out as 12,600,000 / 923,235 = 13.6 nt, that is, 1 insertion per 13 nt of the genome.

3.6 The *Hermes* Genome Browser

***Hermes* Genome Browser: http://bahlerweb.cs.ucl.ac.uk/bioda/.**

In order to extend the usefulness of the *Hermes* insertion data, all log phase and ageing datasets, Hidden Markov Model (HMM) states, conservation measures, and nucleosome densities (see Chapter 4) have been visualised in an interactive genome browser powered through Biodalliance 0.13.7 (Down *et al.* 2011). The *Hermes* Genome Browser is a powerful tool in functional genomics, as exemplified in Figure 3.8 illustrating a gene with lots of *Hermes* insertions during log phase but not during ageing. In this browser, users can search using either a genomic location or a gene name. It is recommended to view the *Hermes* Genome Browser on a Firefox or a Safari web browser.

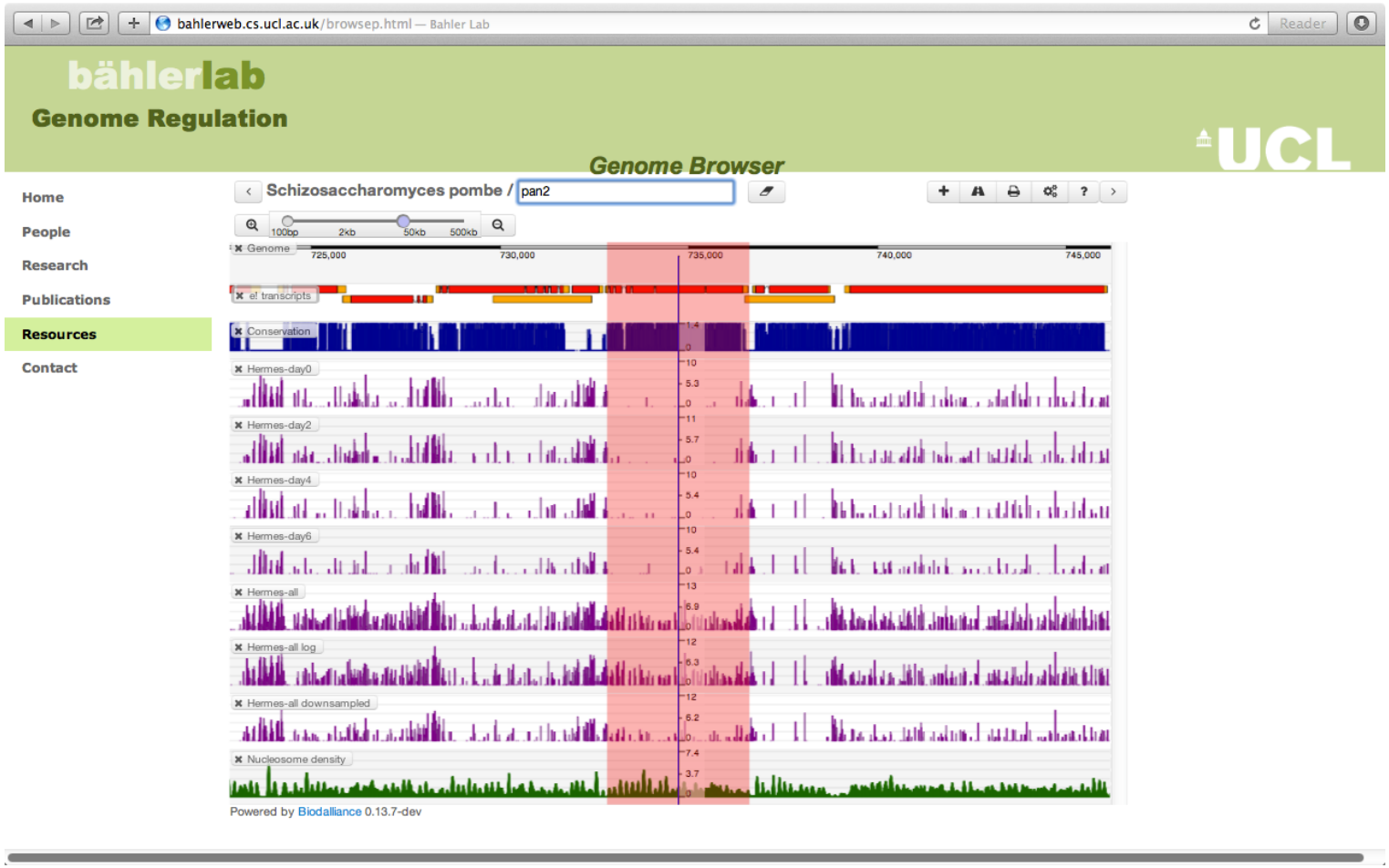


Figure 3.8. Searching for a Gene on the *Hermes* Genome Browser. Highlighted, the *pan2* protein-coding gene, which appears to be essential during ageing (few insertions) but non-essential during log phase (many insertions).

3.7 Summary of the Main Results

Hermes insertion libraries were constructed as described (Park *et al.* 2009), with the exception that the transposition frequency was calculated by dividing the number of colonies on YES 5-FOA+G418 plates by the number of colonies on YES plates. *Hermes* insertion libraries were created in an *S. pombe* strain with the genotype *ura4*–D18 *leu1–32* h– (JB980 in our collection). In general, < 0.2% of cells in libraries contained genomic *Hermes* insertions, so the expectation is that the majority of insertion mutants contain a single insertion.

For each of the libraries, all DNA extracted was processed to enrich for capture of rare insertions. For the same purpose, ligation and PCR reactions were performed in 96-well plates, using a maximum of 1 µg of DNA per well, and then re-pooling before sequencing. So as to distinguish between unique chromosomal insertions and those derived from PCR amplifications, a unique molecular identifier was designed, specifically, a random 10 bp in the adapter sequence. In total, nine log phase *Hermes* libraries were sequenced. For two libraries, ageing was induced through glucose starvation as described (Roux *et al.* 2009). For these two libraries, cells were collected and processed when the culture reached stationary phase (age 0) and 2, 4, and 6 days later.

Overall, 1,098,430 unique insertion sites were sequenced, that is, 1 insertion per 11 bp of the fission yeast genome. In order to extend the usefulness of the *Hermes* insertion data, all log phase and ageing datasets have been visualised in an interactive genome browser (http://bahlerweb.cs.ucl.ac.uk/bioda/).

Chapter 4 DISSECTING THE LOG PHASE DATASET

In this chapter, the aim is to exhaustively explore the log phase dataset, starting with a brief overview and a comparative analysis to existing data, both published and unpublished. Using a Hidden Markov Model (HMM), Chapter 4 next focuses on the clustering of insertion sites into distinct regions of essentiality across the entire genome, similar to the approach taken by DeJesus and Ioerger (2013) for the *Mycobacterium tuberculosis* genome. Figures 4.2 to 4.4, 4.10, 4.12 to 4.21, 4.23, and 6.2 to 6.3 are the work of Dr Daniel Jeffares; I generated all other graphs using the R programming language on a Linux operating system.

4.1 Overview

In analogy to the structure of the Earth, this dataset can be viewed as having both an inner core and an outer crust; getting to the core requires sifting through multiple layers of data. Within the outermost, superficial layer lie three crucially important questions. First, how frequent, and where does the *Hermes* transposon integrate? Second, is there any difference between log phase and ageing libraries? Third, is the insertion data a good predictor of gene essentiality?

4.1.1 How Frequent, and Where Does The Hermes Transposon Integrate?

In order to answer the first question, *Hermes* insertion sites were plotted on the log10 scale for both the nuclear and the mitochondrial genome (Figure 4.1). In agreement with previous studies (Evertts *et al.* 2007, Guo *et al.* 2013), the plot shows that *Hermes* integrates across all three chromosomes with minimal bias. Interestingly, insertions also occur in the mitochondria. 923,235 unique insertions were sequenced in total, which means that the log phase dataset is saturated with *Hermes* insertions once every 13 nucleotides of the genome.

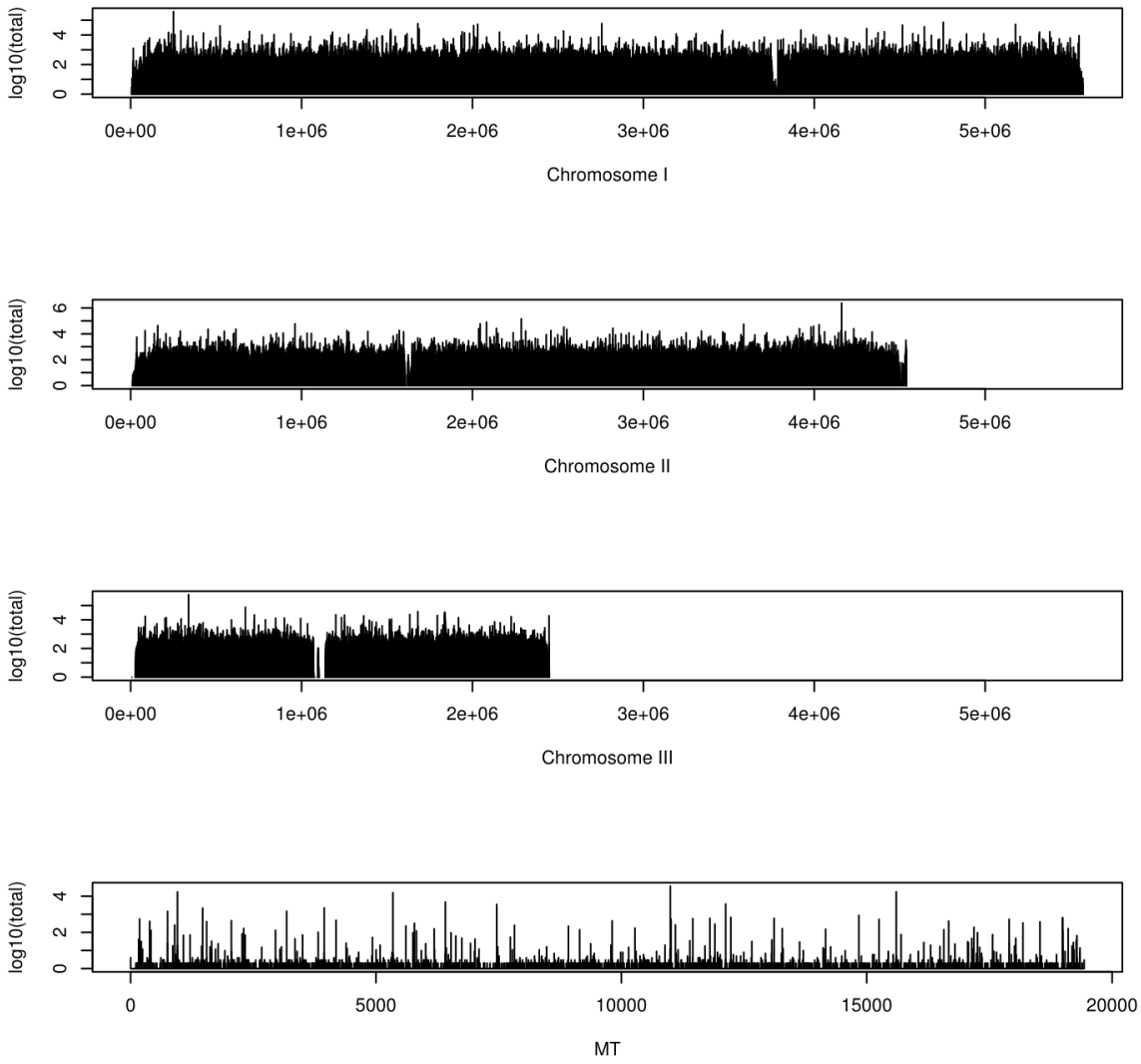


Figure 4.1. Sites of *Hermes* Integration. *Hermes* insertion sites were plotted on the log10 scale for both the nuclear and the mitochondrial genome. Plot shows that *Hermes* integrates across all chromosomes (I, II, and III) of the fission yeast genome. Insertions also occur in the mitochondria (MT).

4.1.2 Is there any Difference between Log Phase and Ageing Libraries?

In order to advance to the next part of the analysis, and to the HMM, it was imperative to evaluate the similarity, if any, between all of the log phase, and their analogous ageing libraries. For this, the IPKM of each library and transcript was calculated as insertion counts per kilobase per million insertions. Figure 4.2 is a heat map showing the Pearson correlation between the IPKM of all libraries. Overall, there are two assumptions for this clustering. First, since insertions are normalised for gene length, gene counts should be similar between libraries.

Second, log phase and ageing libraries should cluster distinctly since both represent two different biological conditions. Interestingly, Figure 4.2 shows that ten of the log phase libraries form three clusters (a, b, and c), and four form two clusters with their corresponding ageing libraries (d and e). Therefore, what this means is that for gene-based analyses (refer to Chapter 5), ageing libraries can still be considered separate from log phase ones, but, for an analysis such as the HMM, both sets of libraries can be pooled together so as to increase the statistical power of the data.

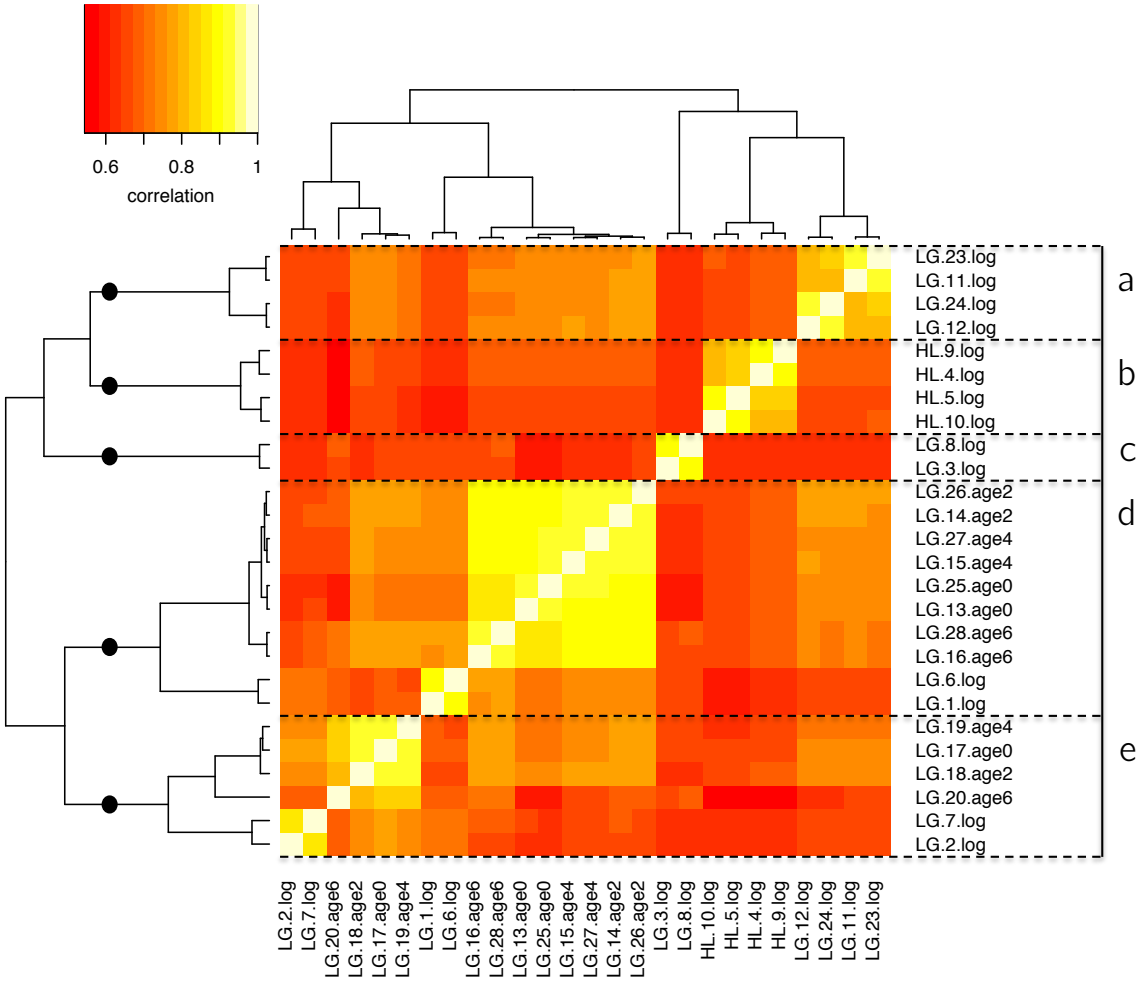


Figure 4.2. Clustering of Log Phase and Ageing Libraries. a = LG10+LG11, LG12+LG13, b = Dr Henry L. Levin's libraries, c = LG7+LG8, d = LG1+LG2, and e = LG3+LG4. Here, a clustered heat map visualising the Pearson's rank correlation coefficients (r) from the correlation analyses between the IPKM of all libraries.

4.1.3 Is the Insertion Data a Good Predictor of Gene Essentiality? A Comparative Analysis.

In search of a pattern, and to determine whether the *Hermes* insertion data is a suitable predictor of gene essentiality, a comparison to (a) genome annotation was made. *Hermes* insertion data was also compared to relevant, published and unpublished data, such as (b) gene expression levels, (c) colony sizes and growth scores, and (d) constraint and genetic diversity.

(a) Genome Annotation

In the beginning, both insert counts and inserts per site (that is, the proportion of insertions per genome length) were correlated to genome annotation (Figure 4.3, Table 4.1). Encouragingly, protein-coding regions of essential genes, which are considered important regions, have fewer insertions. In contrast, long terminal repeats (LTRs), untranslated regions (UTRs), and regions without any annotation (which make up 18% of the genome) have more insertions. Overall, intergenic long ncRNAs (ig/lncRNAs) have the highest proportion of insertions. On a closer look, the most informative measure appears to be the average insert count using all sites, which also includes sites with zero insertions; this measure includes information from counts of sites with one or more insertions, and inserts per site, both of which contain information.

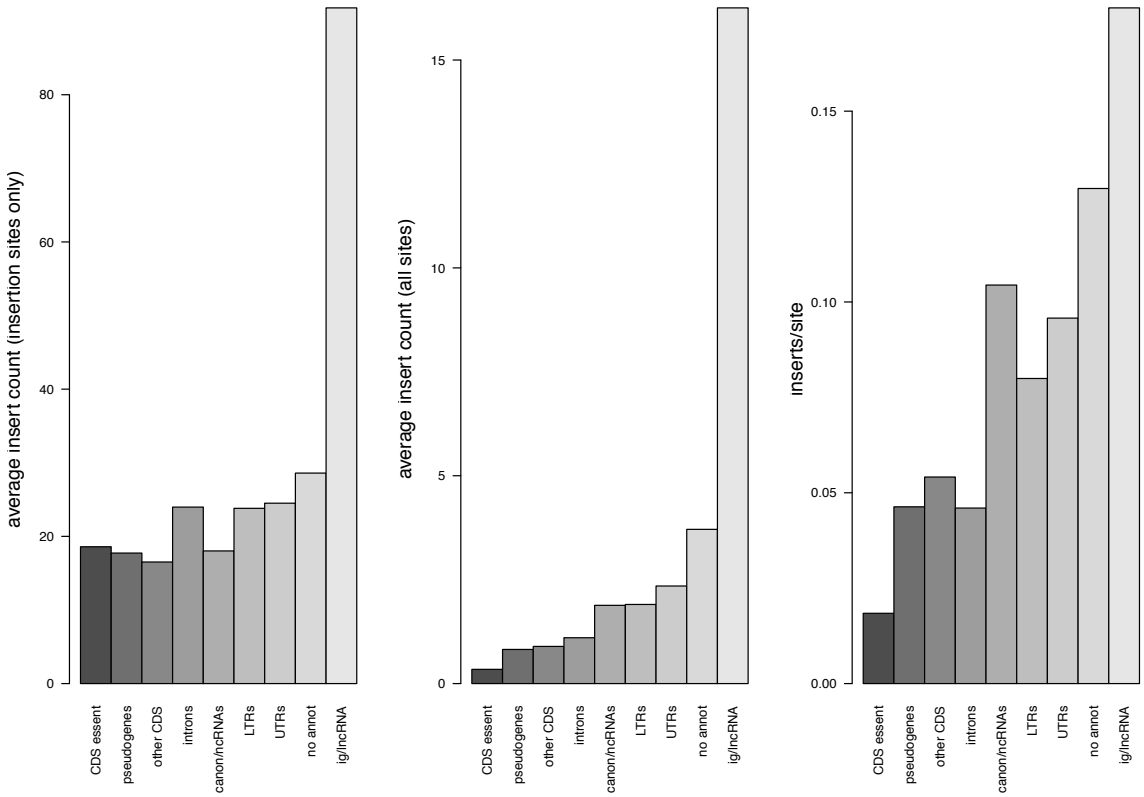


Figure 4.3. Biological Signals of the Log Phase Data. *Left*, shows the average insertion count using only sites with an insertion. *Middle*, shows the average insertion count using all sites including those with zero insertions. *Right*, shows the number of insertions (of any count) per site in the fission yeast genome. From left to right, annotations are coding sequences of essential genes, pseudogenes, coding sequences of non-essential genes, introns, canonical RNAs (rRNAs, sno/snRNAs, and tRNAs), Tf retrotransposons and solo long terminal repeats (LTRs), 5' and 3' untranslated regions (UTRs), regions without any annotation, and intergenic long ncRNAs (ig/lncRNAs). Here, plots use the total of all log phase datasets.

site	average insert count (insertion sites only)	average insert count (all sites)	inserts per site
CDS essential	18.59	0.34	0.02
pseudogenes	17.73	0.82	0.05
other CDS	16.52	0.89	0.05
introns	23.98	1.10	0.05
canonical ncRNAs	18.02	1.88	0.10
LTRs	23.81	1.90	0.08
UTRs	24.51	2.35	0.10
no annotation	28.60	3.71	0.13
ig/ncRNAs	91.80	16.25	0.18

Table 4.1. Biological Signals of the Log Phase Data. Figure 4.3 in table form. In reference to Chapter 4.2, individual values are useful as initial parameters for the Hidden Markov Model (HMM), where coding sequences of essential genes (CDS essential) are used as an expectation for regions where insertions are deleterious, long terminal repeats (LTRs) are used for intermediate regions, and intergenic ncRNAs (ig/ncRNAs) for regions where insertions are advantageous.

In order to provide further supporting evidence for the quality and validity of the data, the genome was fragmented into 20 kb windows and the insertion count was plotted for each position. PomBase annotation for essential and non-essential genes was overlaid on top of the plots. Figure 4.4 shows two examples from chromosome I, but on looking at all the generated plots, the common observation is that the insertion count agrees with genome annotation. However, on a few occasions, there are discrepancies between the two.

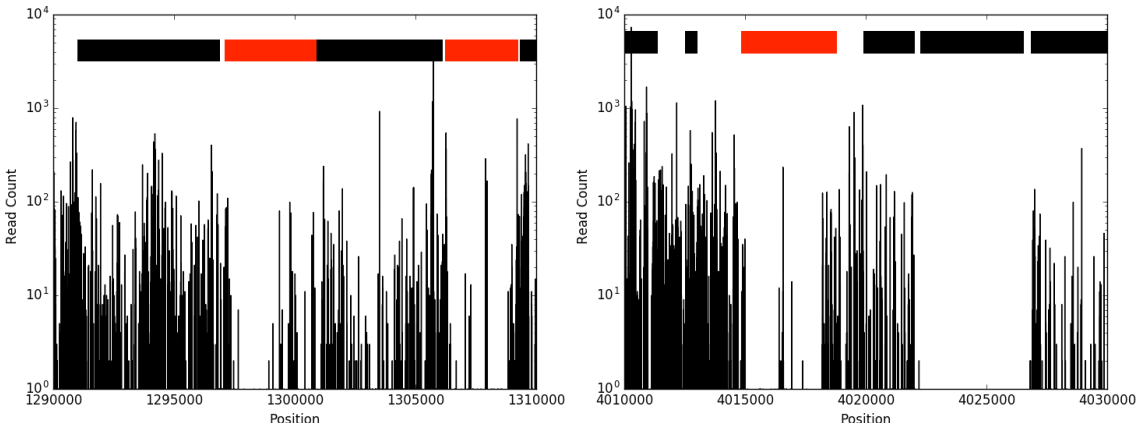


Figure 4.4. *Hermes* Integration vs. PomBase Annotation. Overall, *Hermes* insertion count data agrees with genome annotation (see annotated essential gene at position 1,295,500 to 1,300,000 (*left*) with few insertions). Occasionally, there are inconsistencies between the two (see annotated non-essential gene at position 40,225,000 to 40,275,000 (*right*) with no insertions). It is important to note that even between dense regions, there are positions with no insertions; chromosome I (*shown here*), for example, has an average of 17 zero insertion sites between each insertion. Colour Key: red = essential genes, black = non-essential genes.

So as to delve deeper, a script ("hermes-counts-per-gene.pl") that calculates insertion counts for each gene was written by Dr Daniel Jeffares. It outputs gene i.d. (*gene*), gene length (*len*), number of unique insertion sites in the gene (*sites*), number of unique forward and reverse insertion sites in the gene (*fwd, rev*), minimum (*min*), maximum (*max*), mean and median insertion count number, and finally, a count of unique insertion sites per gene length (*sitespernt*). For the minimum insertion count number, sites with no insertions were assigned a zero value, whereas for the maximum, mean and median, zero values were not counted and only sites that have insertions were included.

The Perl script was used to further test the hypothesis that essential genes accumulate fewer insertions than non-essential genes whether in the log phase or during ageing. Reassuringly, plotting unique insertion sites per gene length for both gene sets, during log phase, early (t0, t2) and late ageing (t4, t6), shows

that non-essential genes do indeed gather more insertions than essential genes (Figure 4.5).

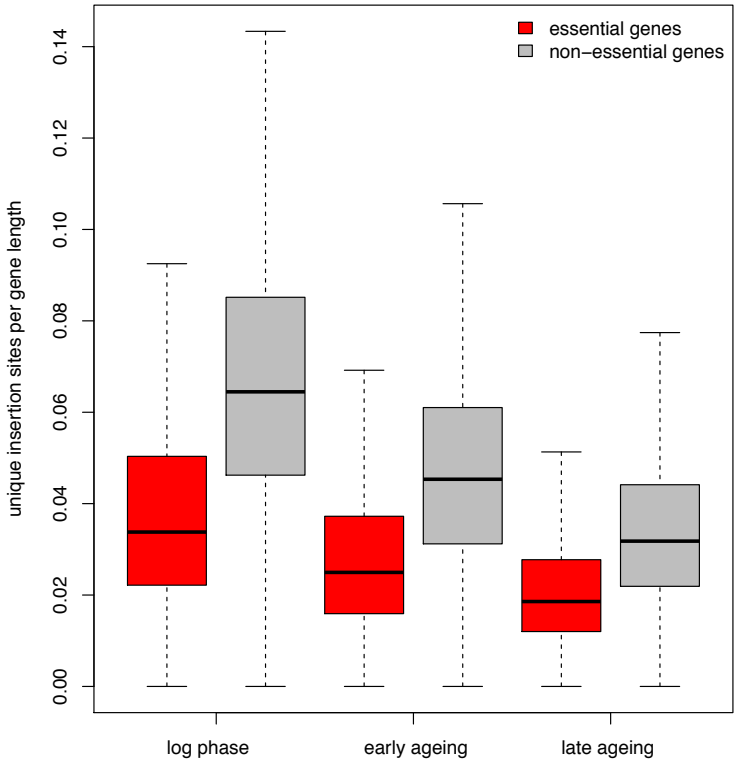


Figure 4.5. *Hermes* Integration as a Marker of Gene Essentiality. Plotting unique insertion sites per gene length (*sitespernt*) for log phase, early ageing and late ageing libraries shows that essential genes accumulate less insertions than non-essential genes. Here, using *sitespernt* instead of *sites* (that is, unique insertion sites in the gene) corrects for gene length bias. PomBase Fission Yeast Phenotype Ontologies (FYPO) 0002061 and 0002060 were respectively used for lists of essential and non-essential genes.

Replotting Figure 4.5, to include the list of long intergenic non-coding RNAs (lincRNAs) being deleted with the CRISPR/Cas9 genome editing system (Rodríguez-López *et al.* 2016), results in Figure 4.6. In brief, the criteria for these ncRNA knockouts include, amongst others, the distance from neighbouring protein-coding genes and relevant RNA-seq information (as published in Marguerat *et al.* (2012)). Plotting the insertion count data for all 122 lincRNAs, as well as the essential and non-essential genes, shows that the lincRNAs have considerably

more insertions; whether this has any bearing on the function of ncRNAs in general is yet to be established.

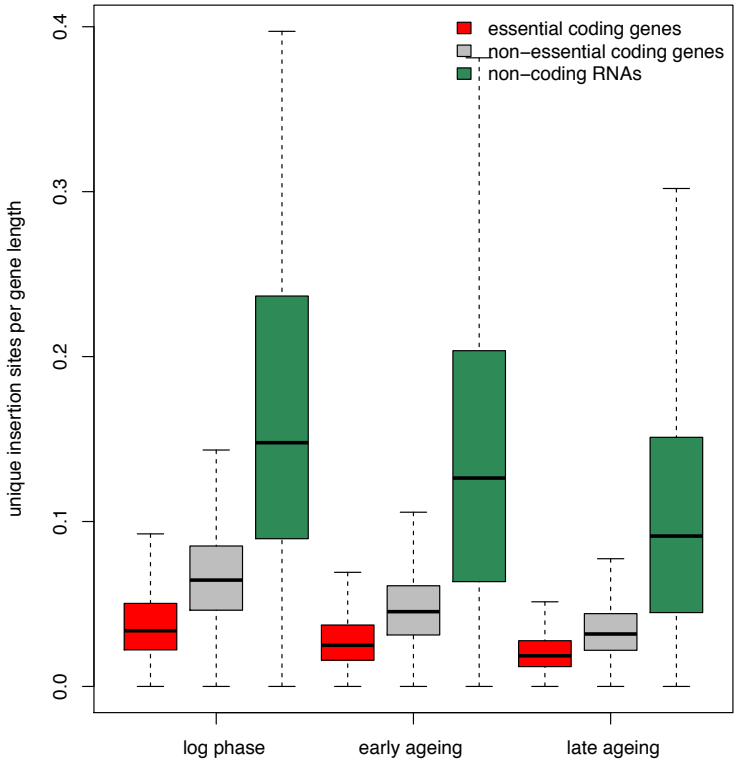


Figure 4.6. *Hermes* Integration and CRISPR/Cas9-Deleted LincRNAs. In the Bähler Lab (University College London), 122 long intergenic non-coding RNAs are being knocked out with the CRISPR/Cas9 genome editing tool. Plotting the *Hermes* insertion count data for these lincRNAs, in addition to the essential and non-essential genes from Figure 4.5, shows that the lincRNAs have substantially more insertions.

(b) Gene Expression Levels

Next, the focus shifted to how transcriptomes adjust to the requirements of the cell, a topic described in detail by Marguerat *et al.* (2012). In this study, RNA-seq and mass spectrometry were combined to analyse how changes in both the cell physiology and volume are reflected in the cellular concentrations of all coding and non-coding RNAs. Marguerat *et al.* analysed both proliferating cells that constantly replenish their RNAs and proteins, and postmitotic cells that do not divide or grow due to a reversible arrest in a quiescent state. It was reported that

the transcriptome is larger in proliferating than in quiescent cells, thus reflecting the higher need for transcription during division and growth. Taking this into account, we compared the expression levels during proliferation to the log phase insertion data, expecting to observe an inverse correlation between count of unique insertion sites and copies per cell, that is, the lower the count, the higher the mRNA copy numbers, and vice versa. Overall, the premise is that essential genes are more highly expressed than non-essential ones, as observed by Mata and Bähler (2003). Figure 4.7 shows a weak correlation therefore implying that expression levels are not particularly suitable for teasing out gene essentiality.

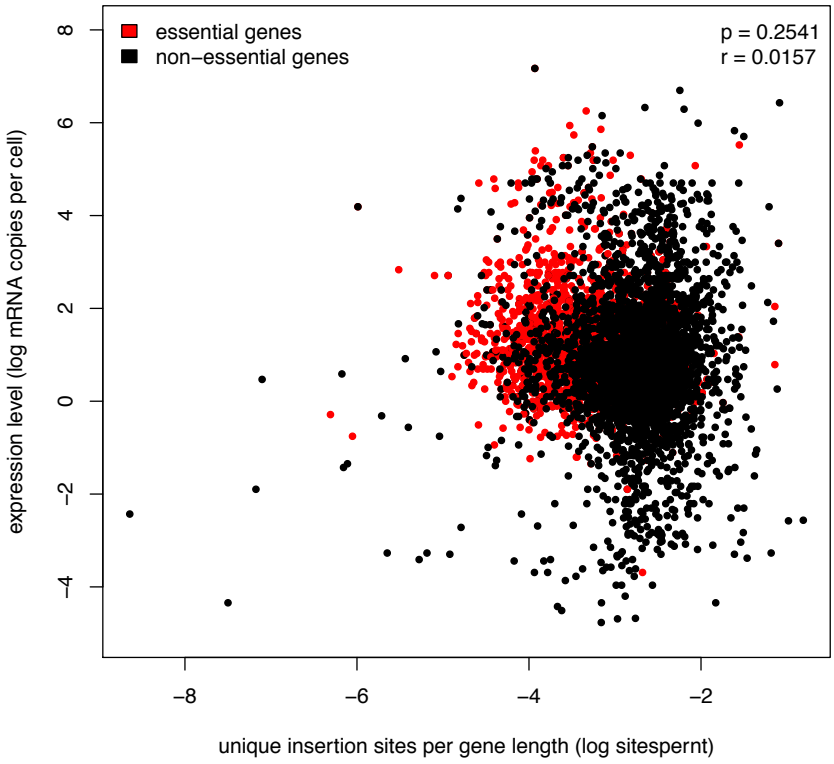


Figure 4.7. Comparison of *Hermes* Integration to Gene Expression Levels. Data for gene expression levels was obtained from Marguerat *et al.* (2012).

(c) Colony Sizes and Growth Scores

Sideri *et al.* (2014) analysed the Bioneer fission yeast deletion library for mutants with differences in proliferation under standard conditions. In detail, parallel phenotyping and quantitative bar-seq were used to compare the proliferation of deletion mutants grown competitively in the same culture. Sideri *et al.* performed two independent biological repeats of the mutant pool grown exponentially in EMM for 9 hours. In the supplementary data (Table S8), growth scores are provided for each mutant, which we use as a measure of fitness and as a comparison to the log phase insertion data. Unfortunately, Figure 4.8 (*top*) shows that there is no significant correlation between count of unique insertion sites and growth scores. It is speculative, but this could be due to the fact that YES, not EMM, was used as growth medium for the proliferation of the *Hermes* mutants. In a somewhat similar approach, Malecki and Bähler (2016) used colony size as a proxy to calculate the fitness of each of the mutants in the deletion library. In this analysis, the expectation is to observe fewer insertions in smaller, slower, less fit mutants, and more in larger, faster, fitter mutants. Figure 4.8 (*bottom*) shows a weak correlation therefore suggesting that colony sizes are also not particularly suited to tease out gene essentiality.

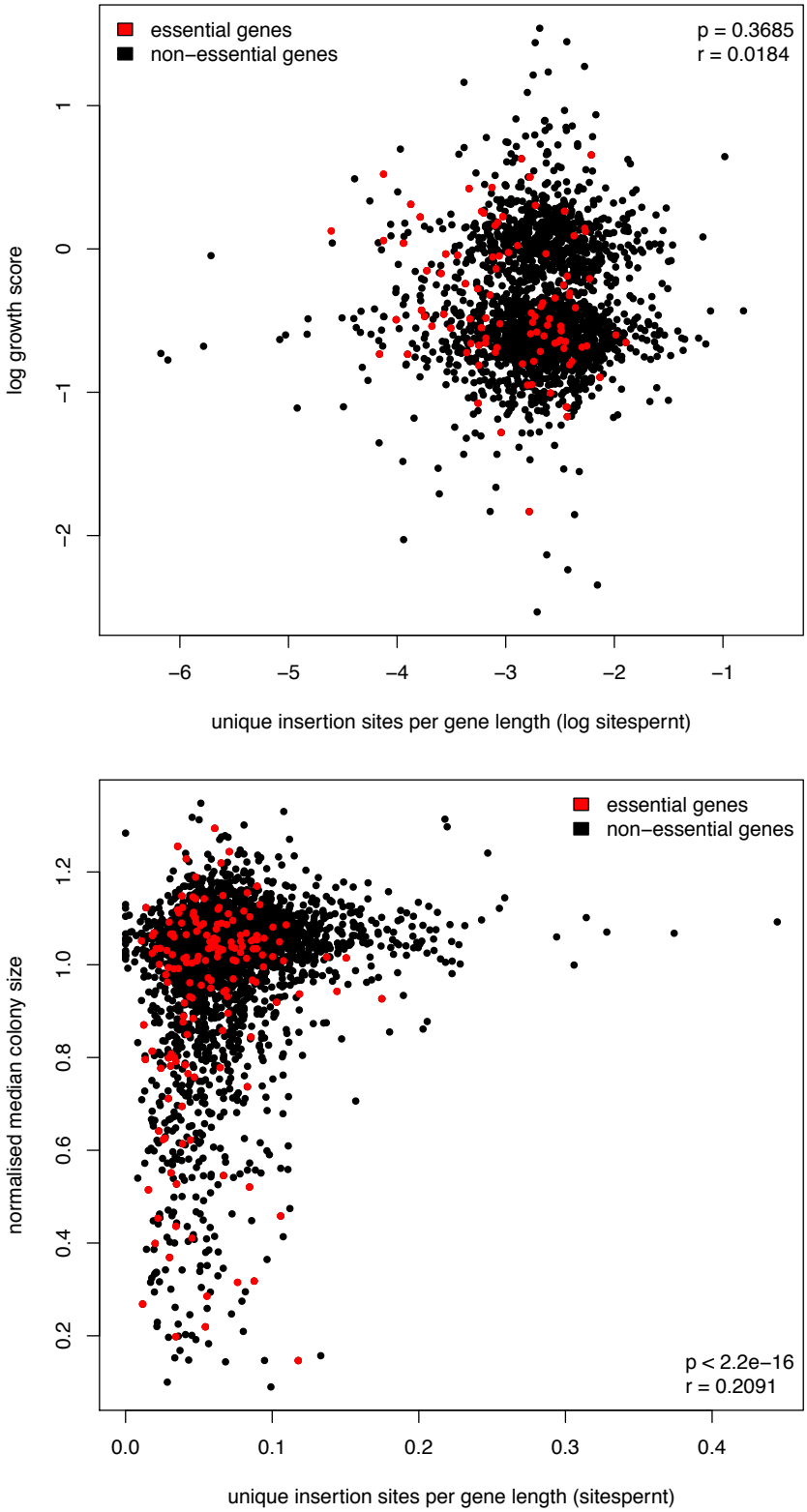


Figure 4.8. Comparison of *Hermes* Integration to Growth Scores (*top*) and Colony Sizes (*bottom*). Exploited as a measure of fitness, data for growth scores and normalised median colony sizes were obtained from Sideri *et al.* (2014) and Malecki and Bähler (2016) respectively.

In an attempt to establish whether the growth scores from Sideri *et al.* and the expression level data from Marguerat *et al.* can also predict gene essentiality, a box plot comparing the respective values for essential and non-essential genes was made. Figure 4.9 shows that while there is no correlation (p = 0.4485, *left*) for growth scores, there is a significant correlation (p < 2.2e-16, *right*) between expression levels for essential and non-essential genes. Nonetheless, the *Hermes* insertion data is still a better predictor, as it bears a stronger correlation between essential and non-essential genes (c.f. Figure 4.5).

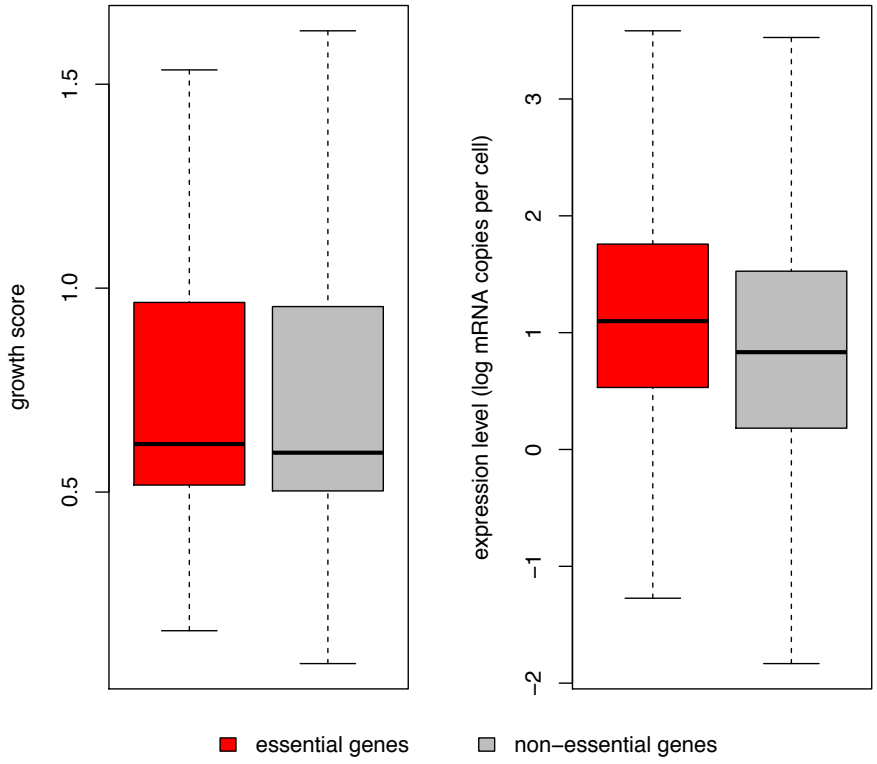


Figure 4.9. Growth Scores (*left*) and Gene Expression Levels (*right*) as Predictors of Gene Essentiality. Used as a fitness proxy, data for growth scores and gene expression levels were obtained from Sideri *et al.* (2014) and Marguerat *et al.* (2012) respectively.

(d) Constraint and Genetic Diversity

Next, we determined whether the *Hermes* insertion data relates to evolutionary data. To this end, an analysis of genetic diversity between coding regions showed that protein-coding regions are more conserved (less genetically diverse) than introns and 5'/3' untranslated regions (UTRs), and both of these are more conserved than regions of the genome with no annotation (Fawcett *et al.* 2014, Jeffares *et al.* 2015). In addition, estimates of constraint between *Schizosaccharomyces* species are consistent with genetic diversity.

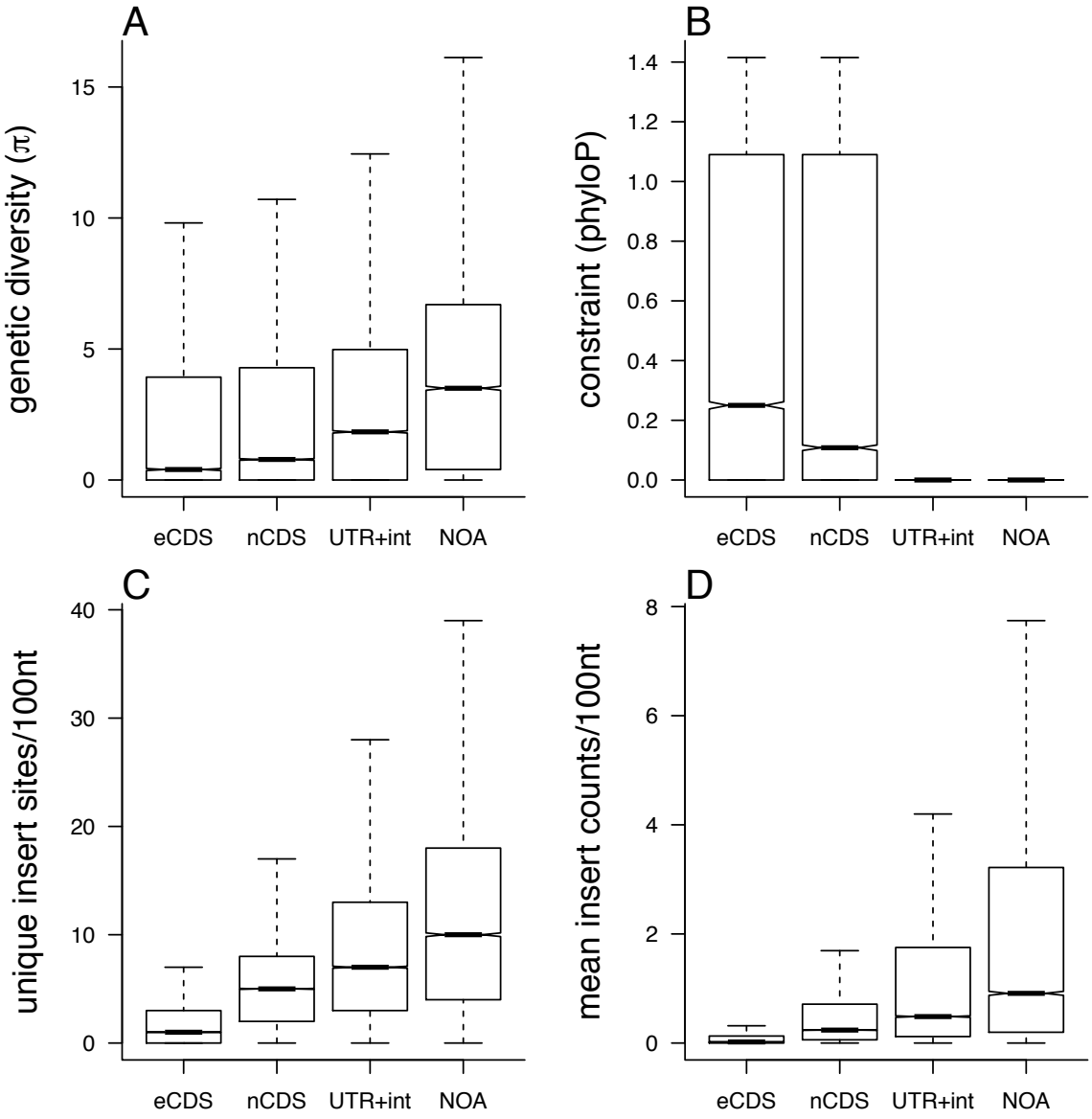


Figure 4.10. *Hermes* Insertion Data relates to Evolutionary Data and Genome Annotation. In summary, for the protein-coding regions of essential

genes (eCDS), protein-coding regions of non-essential genes (nCDS), 5'/3' UTRs and introns (UTR+int), and regions of the genome without any annotation (NOA), we show (A) the genetic diversity from 57 strains of *S. pombe* (Jeffares *et al.* 2015), measured in 100 nt windows, and (B) an estimate of the constraint (conservation) between four *Schizosaccharomyces* species (mean phyloP conservation over 100 nt windows). Similarly, for pooled log phase *Hermes* data, we show (C) the number of unique insertion sites/100 nt, and (D) the mean insertion counts/100 nt.

Overall, this comparative analysis shows that the genome scale patterns of *Hermes* insertions relate well to our broad expectations of the functional elements of the genome, to the genome scale trends of divergence between related *Schizosaccharomyces* species, and to diversity within the species. While divergence and diversity are able to predict generic aspects of the genome with respect to function, they do not have sufficient resolution to pinpoint specific functional elements. Primarily, this is because polymorphic sites are present at low density and their relationship to constraint is affected by recombination rate (Campos *et al.* 2014) and recent events of selection, which can purge diversity in surrounding areas (Cheeseman *et al.* 2012). Therefore, considering the density of our transposon insertion libraries, and the fact that the libraries are generated from multiple samples without evolutionary history, we expect to circumvent these problems and locate functional elements with high accuracy.

4.2 Hidden Markov Model

Tn-seq data can be analysed in different ways, using reads to determine the absence or presence of insertions in a gene, or alternatively, using the read count, that is the number of reads at each site. Intrinsically, both approaches have challenges, depending on the quality of the libraries and the sequencing dataset. DeJesus and Ioerger (2013) described a novel method for examining Tn-seq data using Hidden Markov Models (HMMs). HMMs explore sequential datasets, in which a sequence of observed values is explained by a hidden state sequence, in this case, the essentiality of each site, which cannot be directly observed. HMMs can use information from read counts to infer the state probability distribution and the most likely state sequence. In a simple two-state HMM, for example, the genome of an organism can be viewed as an alternating state sequence of essential (State 1) and non-essential regions (State 2).

So as to characterise the essentiality of the entire fission yeast genome, an initial HMM with three states was assembled by Dr Maarten Speekenbrink at University College London. The HMM is based on Visser and Speekenbrink (2010) with the theory outlined in Section 4.2.1 below. Dr Daniel Jeffares and Mr Christoph Sadée debugged the code, trained the model, altered it to incorporate log2 transformed data, and finally, expanded it to a fifth state.

4.2.1 The HMM Model

First, consider a discrete Markov model, before advancing to the more complex HMM; consider that each position in the 12.6 Mbp fission yeast genome could be in one of several pre-defined states:

$$q_t = S_j \qquad (1)$$

where:

q_t = state of nucleotide t
t = position of nucleotide with $1 \leq t \leq T$
S_j = state j.

In this case, this results in a state sequence Q = $q_1 \ldots q_t \ldots q_T$ with T = 12.6 Mbp if considering the entire genome at once. It is also easy to analyse sub-domains of the genome separately such as individual chromosomes. S_j is part of a pre-defined set of states such as:

$$S = \{S_1, S_2, S_3\} \tag{2}$$

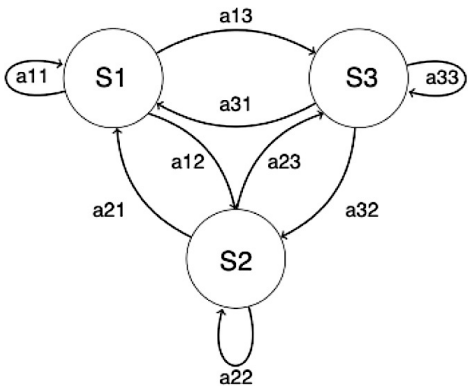
where:

S_1 = essential
S_2 = intermediate
S_3 = non-essential.

For this model, the states correspond to the essentiality for growth, that is, if a nucleotide at one position is essential, intermediate, or non-essential for the growth of the cell. It is, however, challenging to infer whether a nucleotide at one position is essential, and so, for this reason, the essentiality of genomic regions should be taken into account instead. Therefore, consider that the state at position q_t is predicted by some probability depending on the preceeding states $P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \ldots]$, or using the first order Markov assumption, the most adjacent preceding state:

$$a_{ij} = P[q_t = S_j | q_{t-1} = S_i] \tag{3}$$

where: the transition probabilities, a_{ij}, are part of a Markov chain with probabilities for each transition, and can be represented as a transition matrix (Figure 4.11).



$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Figure 4.11. Three State, Ergodic Markov Chain (*left*) and Transition Matrix (*right*). In this Hidden Markov Model, the three hidden states (S1, S2, and S3) are defined as "essential", "intermediate", and "non-essential".

For this model, an ergodic process is considered, meaning that one state can be reached from another $a_{ij} > 0$. In addition, since the state is not directly observable, it has to be inferred by the absence or presence of an insertion site and its count; absence means increased likelihood to be essential, whereas presence means increased likelihood to be intermediate or non-essential. Overall, this calls for a Hidden Markov Model, defined by M, the number of distinct observation symbols per hidden state, which correspond to all count values including sites with zero insertions. In consequence of the biases described in Chapter 1 (nTnnnnAn motif and nucleosome occupancy), a high insertion count can still correspond to an essential state, and therefore, all observation symbols are still possible for each state. Here, the individual observation symbols, v_k, are part of the set V.

$$V = \{v_1, v_2, \ldots, v_M\} \tag{4}$$

Next, the data can be interpreted as an observation sequence $O = O_1 O_2 \ldots O_t \ldots O_T$, where the position is denoted by t and the observation at O_t is v_k. Each state has an observation symbol probability distribution, B, associated to it. Equation 6 is the probability of being in state S_j when the observation at position t is $O_t = v_k$. It is important to note that B is affected by the two biases and is thus not only based on the observation sequence.

$$B = \{b_j(k)\} \tag{5}$$

$$b_j(k) = P[v_k \text{ at } t | q_t = S_j] \tag{6}$$

For a complete description of the HMM, the last parameter to be defined is π, the initial state distribution. π is required as each state depends on the previous state by some probability a_{ij}, and since there is no preceding state for the first position,

there must be an initial state distribution defining the state of the first element at t = 1. In Equation 9, all HMM distributions are grouped together into λ.

$$\pi = \pi_j \quad (7)$$

$$\pi_j = P[q_1 = S_j] \quad (8)$$

$$\lambda = \{A, B, \pi\} \quad (9)$$

So now, the problem is defined by (a) how to best adjust the model parameters λ, given the two biases, in order to maximise the probability of finding the observation sequence of insertion sites and counts, and (b) what is the best state sequence ($Q = q_1 \ldots q_t \ldots q_T$) for the given parameters and observation sequence? In this HMM, the first problem was resolved using the expectation maximisation (EM) algorithm, whereas the second problem was tackled using the Verterbi algorithm. The HMM was implemented through the depmixS4.R package, which has an EM algorithm that models the observation symbol probability distribution, B, as a linear response model (Visser and Speekenbrink 2010); this allows for the addition of the 8 nt motif and nucleosome occupancy biases as covariates.

For each position in the genome, given the two biases, there is a likelihood for transposition to occur. Dr Chris Illingworth (University of Cambridge) used insertion data from Guo *et al.* (2013) to calculate the probability of transposition based on the 8 nt nTnnnnAn motif. In his approach, Dr Illingworth calculated the percentage of each nucleotide at each position and compared this to the percentage composition across the entire genome. Overall, 20 positions were identified for which the composition differed from the genome composition by at least 1%. For each position t, the probability of observing one nucleotide is given by $p_t(\alpha)$: $1 \leq t \leq 20$, where α denotes any of the four possible nucleotides, AGCT. In Equation 10, $p^{gw}(\alpha)$ is the genome-wide probability of observing nucleotide α. In his quest, the next step was to fragment the genome into 20 nt windows, and for each window, calculate a likelihood measure (L).

$$L = \underbrace{\sum_{t} \log p_t(\alpha_t)}_{\text{similarity to motif}} - \underbrace{\sum_{t} \log p^{gw}(\alpha_t)}_{\text{similarity to base comp}} \quad (10)$$

Finally, the state sequence Q was calculated using post model parameter optimisation. The Verterbi algorithm maximises P(Q|O,λ) (same as P(Q,O|λ)), the probability of finding a state sequence for the given model parameters λ and observation sequence O. In detail, this is done via an iterative process, when moving along the observation sequence, by defining δ, the highest probability along a single path.

$$\delta_t(i) = \max P[q_1,q_2,...,q_t = i,O_1O_2...O_t|\lambda] \quad (11)$$

$$\delta_{t+1}(j) = [\max \delta_t(i)a_{ij}] \cdot b_j(O_{t+1}) \quad (12)$$

Therefore, the most likely state sequence is then the sequence of arguments that maximises Equation 12, which can be evaluated sequentially from the first position 1 to the last position T, resulting in an assigned state for each nucleotide position. In a similar approach, a measure for the nucleosome bias was established based on the average nucleosome density (Atkinson *et al.* 2017). Ultimately, following the inclusion of the two biases, the HMM was applied separately to each of the three chromosomes. In a brief overview,

(a) the first runs were carried out on the raw read count data, then

(b) a correlation between read count data and cell growth was uncovered and implemented into a three-state HMM and (c) a four-state HMM, and finally,

(d) a five-state HMM was constructed, assigning an additional state to extremely high read counts.

4.2.2 HMM Optimisation

(a) First Run

In this HMM, the states were classified as "S1 = essential", "S2 = intermediate" and "S3 = non-essential". Unlike DeJesus and Ioerger (2013), who specified appropriate likelihood functions for read counts, the states for this model were

defined based on annotation and the associated insertion counts and frequencies. In reference to Figure 4.3 and Table 4.1, coding sequences of essential genes were used to define State 1, long terminal repeats and untranslated regions to describe State 2, and intergenic non-coding RNAs and non-annotated regions to designate State 3. For this model, an ergodic symmetric initial transition matrix was chosen, with the larger values on the diagonal assuring data smoothing and continuous states (Equation 13). π, the initial distribution, was selected as in Equation 14, with an increased likelihood for State 1 at the first position (π_1 = 0.5) due to a long insertion sequence at the start.

$$A_{i,j} = \begin{pmatrix} 0.998 & 0.001 & 0.001 \\ 0.001 & 0.998 & 0.001 \\ 0.001 & 0.001 & 0.998 \end{pmatrix} \quad (13)$$

$$\pi = \{\pi_1 = 0.5, \pi_2 = 0.25, \pi_3 = 0.25\} \quad (14)$$

Initially, running the HMM with these parameters results in fast alternating, and few continuous states, with S1 being the most abundant (Figure 4.12). So as to put this into perspective, Kim *et al.* (2010) found that in fission yeast, only 26.1% of genes are essential. It is thus apparent that the issue here is that the HMM is too sensitive, rapidly changing to S3 when there is an insertion with a high count, and to S1 when there is a short sequence with no insertions. Overall, the indication is that the transition matrix in Equation 13 does not effectively smooth out the data, hence resulting in few and short continuous states. Figure 4.12 does, however, show predicted regions correlating with annotation and both insertion counts and frequencies.

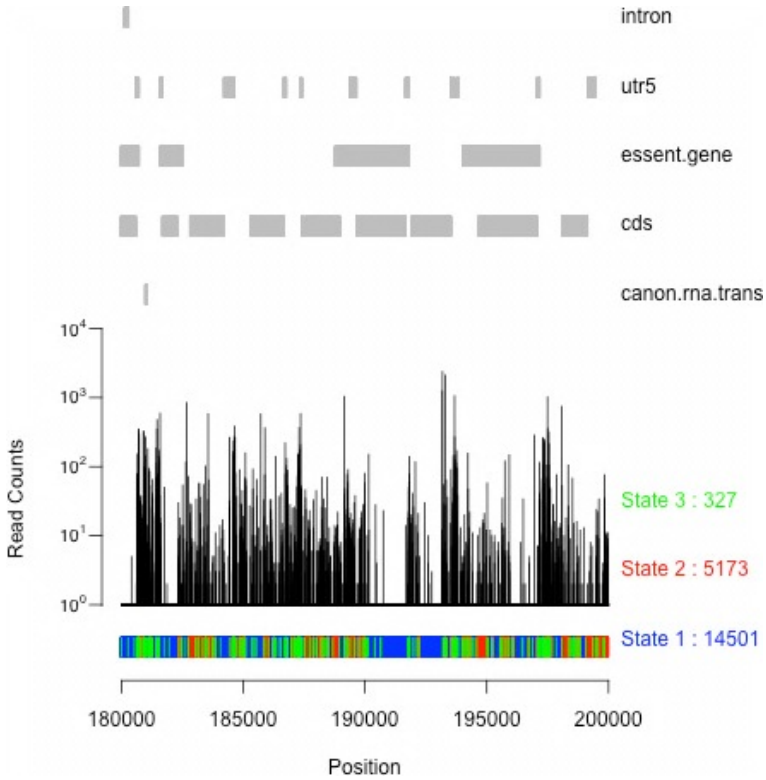


Figure 4.12. HMM, First Run. Illustrated, a 20 kb region within chromosome II with PomBase annotation on top and predicted states at the bottom. Here, the states are classified as "State 1 = essential" (*blue*), "State 2 = intermediate" (*red*) and "State 3 = non-essential" (*green*). Qualitatively, S3 appears to be more abundant than the indicated 327 nt, but this is due to plotting a large number of insertions in a rather small window. Quantitatively, S1 is the most frequent state for this region (14,501 nt) and in general. HMM's predicted regions correspond to annotation and both insertion counts and frequencies, as exemplified by the region to the right of the 190,000 nt position, predicted as essential. However, with the initial parameters, the overall result is unsatisfactory, with fast alternating, and few continuous states.

(b) Read Count Data and Cell Growth

In an effort to construct a fully functioning model, the non-zero insertion count data was analysed further and used to plot log2 read count histograms for each chromosome. Figure 4.13 (*top*) is the histogram for chromosome I, characterised by an exponential decay, with smaller read counts being the most frequent.

Interestingly, a closer look at the histogram (Figure 4.13, *bottom*) reveals a pattern consistent with the geometric series 1, 2, 4, 8, 16, 32. Specifically, at log2 read counts of 3, 4 and 5 (inverse log2 read counts of 8, 16, and 32), the frequency increases, an observation reflected in the other chromosomes as well.

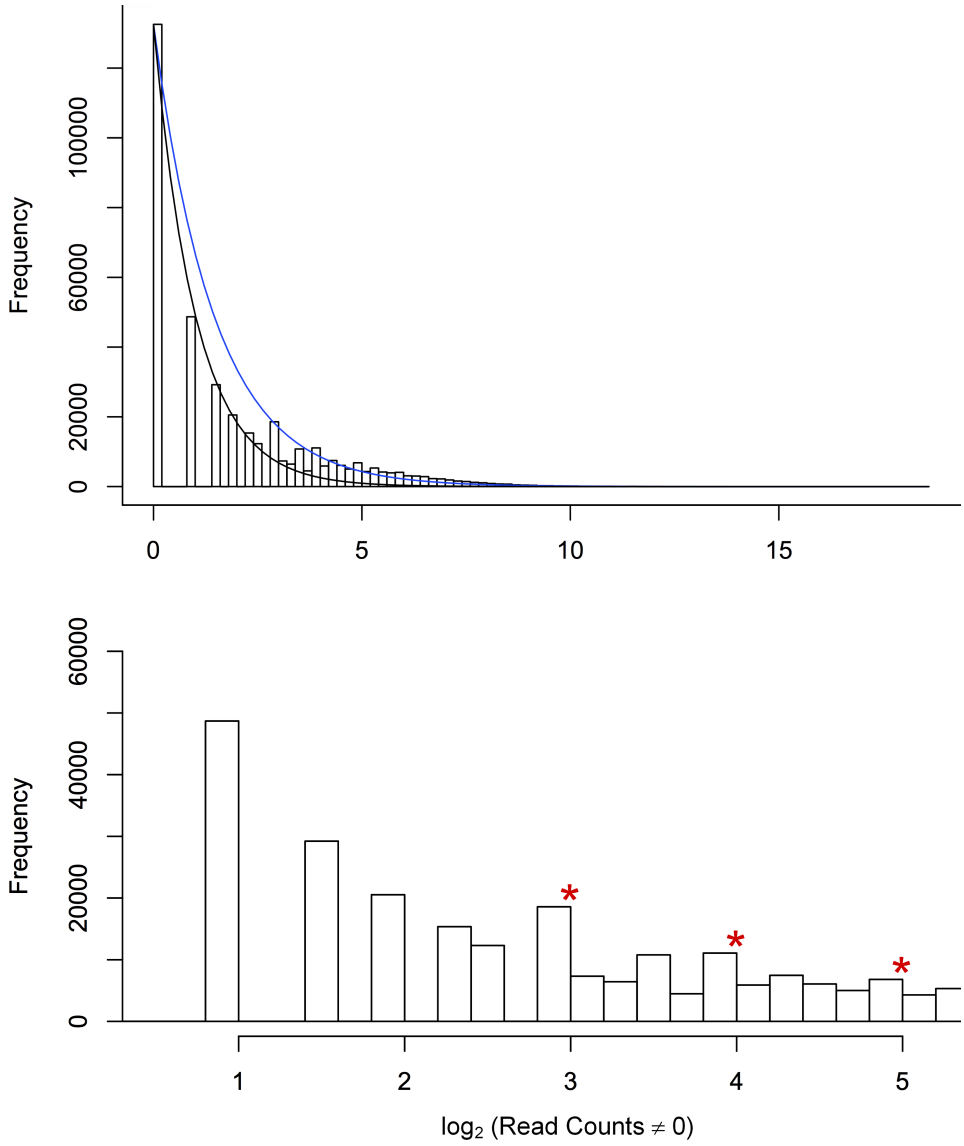


Figure 4.13. Log2 Non-Zero Read Count Histograms. *Top*, the histogram for chromosome I, characterised by two exponential decay functions, $y = Ae^{-\lambda x}$ (black) and $y = A(2^{-\lambda x})$ (blue), fitted to the data with A = maximum frequency and $\lambda = 1$. *Bottom*, a zoomed histogram exposing peaks at log2 read counts of 3, 4 and 5 (marked with red asterisks).

In order to rationalise this result, a correlation was made to the mitotic cell cycle:

$$x(t) = A \cdot 2^{t/\tau} \tag{15}$$

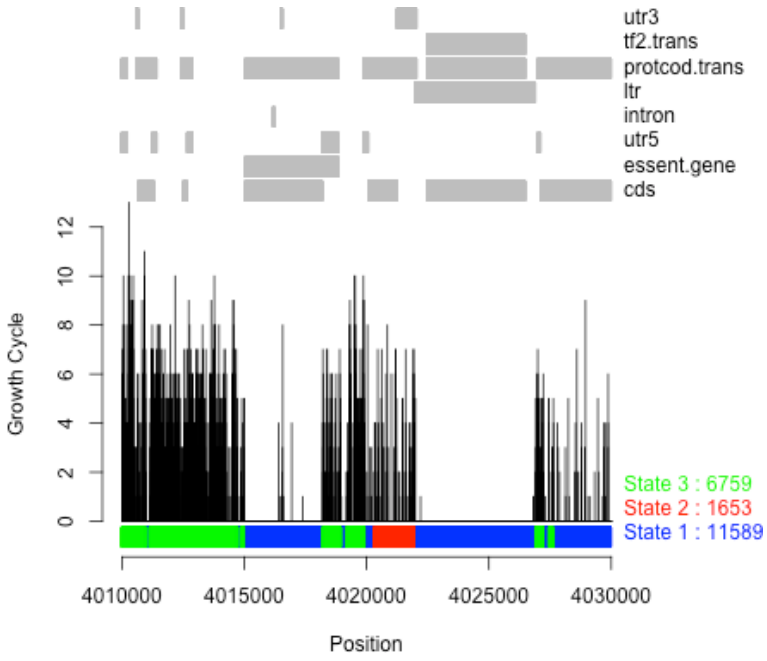
where:

x = cells at time t
t/τ = growth, or mitotic cell cycle
τ = time required for one growth, or mitotic cell cycle
A = x(0) = initial population.

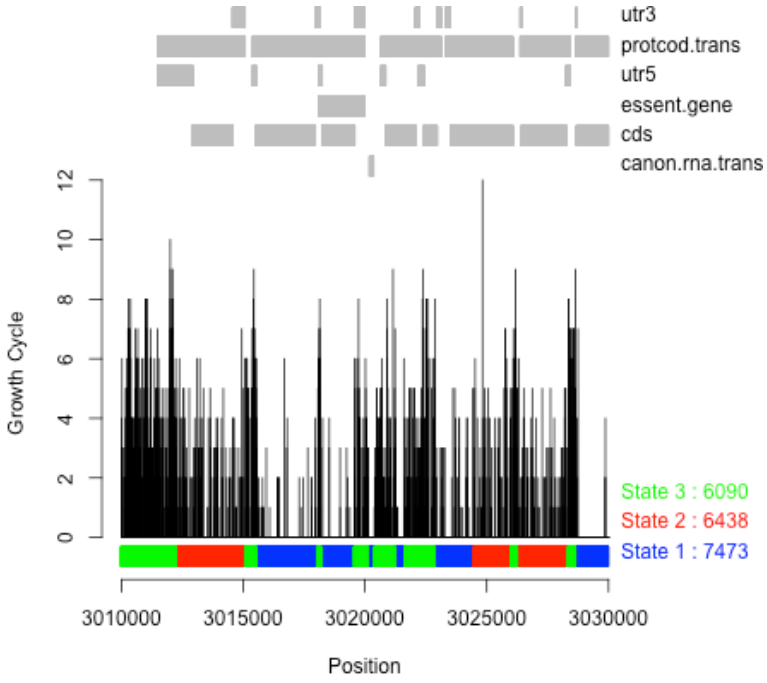
It was previously assumed that the read count for one insertion position *p* reflects transposition occuring at site *p* in different cells within the population. However, a different interpretation, in view of this result, is that transposition occurs at position *p* in one cell, and as that cell undergoes mitosis, it produces two identical daughter cells both with an insertion at position *p*. In turn, those 2 produce 4, then 8, then 16, then 32 cells, all with an insertion at position *p*; this is consistent with the geometric series described above. Therefore, using Equation 16 below, the read count can then be converted to a mitotic, or growth cycle value. It is important to note that the growth cycle is specific to each individual insertion since the exact start of transposition cannot be determined for each individual cell.

$$\log_2(x) = t/\tau = \text{growth cycle} \tag{16}$$

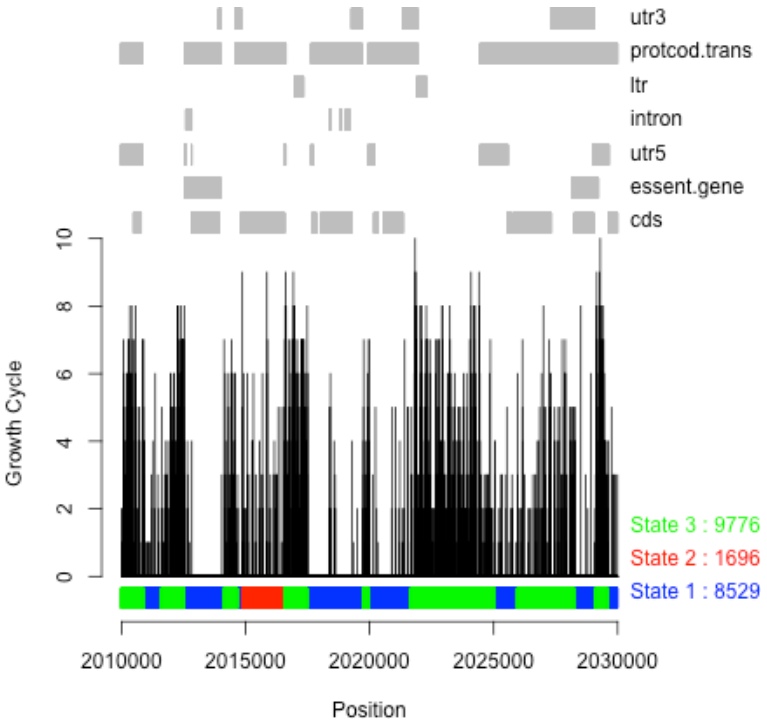
So, with this in mind, for the second run of the HMM, all non-zero read counts were log2 transformed, resulting in a growth cycle value for each insertion site; the same initial state probability, training set, and transition matrix as the first run were used. Figure 4.14 illustrates the outcome, that is, more continuous states.



(a) Chromosome I.



(b) Chromosome II.



(c) Chromosome III.

Figure 4.14. HMM, Second Run. Illustrated, 20 kb regions within each of the three chromosomes with annotation on top and predicted states at the bottom. PomBase annotations are canonical RNAs (canon.rna.trans), coding sequences (cds), essential genes (essent.gene), introns (intron), solo long terminal repeats (ltr), protein-coding transcripts (protcod.trans), Tf retrotransposons (tf2.trans), and 3' and 5' untranslated regions (utr3, utr5). HMM states are classified as "State 1 = essential" (*blue*), "State 2 = intermediate" (*red*) and "State 3 = non-essential" (*green*). In this run, even more so than the first run (c.f. Figure 4.12), the HMM's predicted regions correspond to annotation and both insertion counts and frequencies. In addition, this run generated more continuous states and only a few short states, but the latter could be due to undersampling.

In this HMM, it is encouraging that all states have an average continuous length greater than 150 nt (Table 4.2), which is the average length of DNA wrapped around a nucleosome core, because this implies that the nucleosome bias is eliminated. It also suggests that state prediction is based solely on the fitness of the cells.

	Average Continuous State Lengths (nt)		
	State 1 "essential"	**State 2 "intermediate"**	**State 3 "non-essential"**
Chromosome I	1337	2144	895
Chromosome II	1293	1850	841
Chromosome III	1518	3033	1333

Table 4.2. Average Continuous State Lengths. It is reassuring that all states have an average continuous length greater than 150 nt; the average length of DNA wrapped around a nucleosome core. States < 150 nt imply more accessible, nucleosome-free regions and therefore higher insertion probabilities. In contrast, states > 150 nt suggest less accessible, nucleosome-occupied regions with the absence or presence of insertions based solely on the fitness of the cells.

In addition to the average lengths, the maximum continuous state lengths were also looked at. Indeed, for chromosomes I, II and II, the maximum lengths for S1 for example, were 49,630, 53,300 and 43,530 respectively. Figure 4.15 illustrates the 43,530 nt long region within chromosome III, providing a perfect example of how powerful the *Hermes* insertion data can be to the *S. pombe* community.

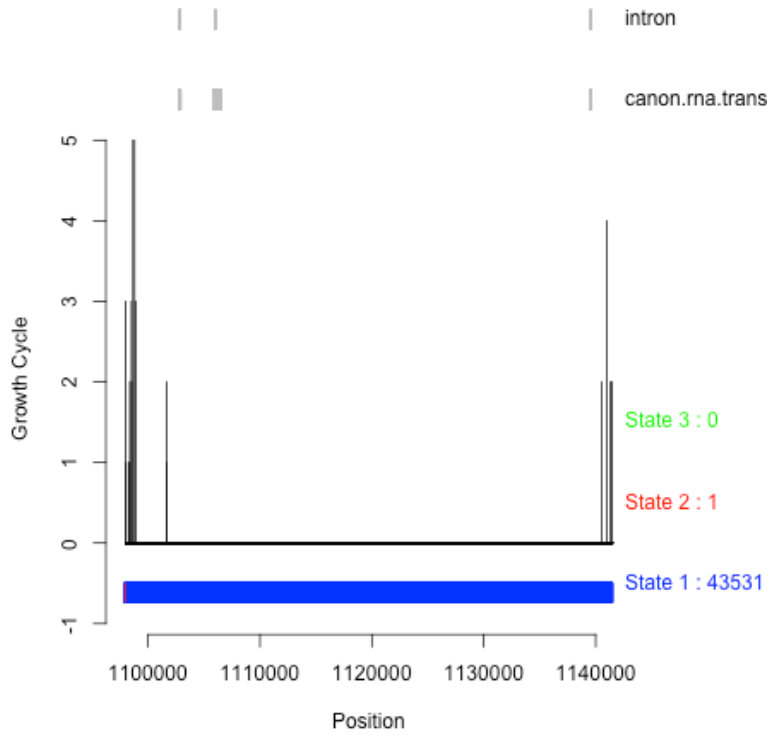


Figure 4.15. HMM vs. PomBase Annotation. Illustrated, a 40 kb region within chromosome III with PomBase annotation on top and predicted states at the bottom. Interestingly, this region is classified as essential by the HMM, but it is not annotated accordingly in the PomBase database.

Ultimately, so as to quantitatively validate the quality of the model, growth cycle values were reversed to read counts. In turn, as in Table 4.3, the total percentage, insertion frequencies and mean read counts of a state within each chromosome, were calculated. Overall, the results indicate that essential regions (S1) are much more conserved than intermediate (S2) and non-essential (S3) regions due to the lower insertion frequencies and mean read counts.

HMM States	Chromosome Number	Total Chromosome %	Insertion Frequencies	Mean Read Counts
State 1 "essential"	I	47	0.020	305.94
	II	51	0.022	314.12
	III	45	0.019	364.08
State 2 "intermediate"	I	23	0.063	1897.85
	II	17	0.071	1852.73
	III	17	0.058	2289.88
State 3 "non-essential"	I	30	0.156	4087.38
	II	32	0.176	5583.15
	III	38	0.126	6820.37

Table 4.3. HMM Quantitative Validation. *Total Chromosome %*: the ratio of nucleotides classified as one state compared to the total amount of nucleotides. *Insertion Frequencies*: the ratio of all insertions within one state compared to the total amount of nucleotides within the same state across the entire chromosome. *Mean Read Counts*: the total of non-zero read counts within one state divided by the non-zero insertions within the same state across the entire chromosome. Overall, S1 regions are more conserved owing to the lower insertion frequencies and mean read counts. In addition, S1 regions account for a larger percentage of each chromosome, however this could be due to undersampling.

(c) Four-State HMM

In order to represent extremely high growth cycles, a fourth state was introduced. In this HMM, the initial state probability was as in Equation 17, the training set for the fourth state included growth cycles above the 99th percentile, and the transition matrix was as in Equation 18.

$$\pi = \{\pi_1 = 0.5, \pi_2 = 0.2, \pi_3 = 0.2, \pi_4 = 0.1\} \tag{17}$$

$$A_{i,j} = \begin{pmatrix} 0.998 & 0.0006 & 0.0006 & 0.0006 \\ 0.0006 & 0.998 & 0.0006 & 0.0006 \\ 0.0006 & 0.0006 & 0.998 & 0.0006 \\ 0.0006 & 0.0006 & 0.0006 & 0.998 \end{pmatrix} \tag{18}$$

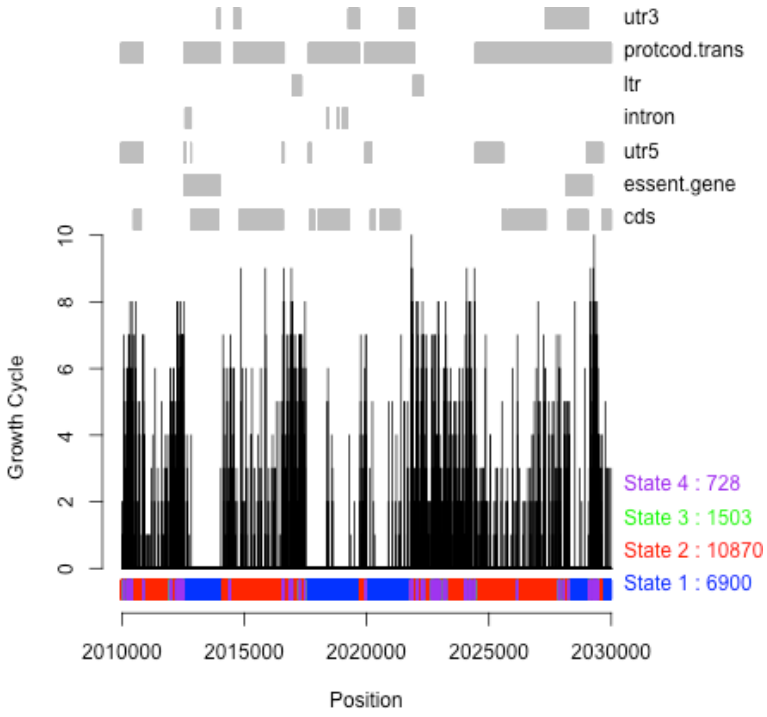


Figure 4.16. Four-State HMM. Illustrated, a 20 kb region within chromosome III with PomBase annotation on top and predicted states at the bottom. Here, the states are classified as "State 1 = essential" (*blue*), "State 2 = very important" (*red*), "State 3 = somewhat important" (*green*), and "State 4 = not important" (*purple*). S2 is on average the longest continuous state within all chromosomes.

(d) Five-State HMM

Overall, in these HMMs, the initial state distributions (π) were $1/n$ where n is the number of states, and the transition probabilities (a_{ij}) were 0.95 for positions remaining in the same state and $0.05/(n\text{-}1)$ for all other transitions. So, for the five-state HMM, the initial state distribution was 1/5 for all states, and the probabilities were 0.95 if remaining in the same state, and 0.01 for all other transitions.

In order to determine the best-fitting model, both a selection and a stopping criterion were applied. For the former, the Bayesian Information Criterion (BIC) was implemented. The BIC, a popular criterion for model selection, provides a measure of the weight of evidence favouring one model over another. Generally, the model with the lowest BIC is considered the better fit (Weakliem 1999). Figure 4.17 shows that, for three datasets, adding more states to the HMM enhances its robustness. Indeed, this was true for all datasets on which the BIC was implemented. Ultimately, a five-state HMM was selected, with states classified as "State 1 = essential", "State 2 = very important", "State 3 = important", "State 4 = somewhat important", and "State 5 = not important". Here, classification was based on coding sequences of essential genes (S1), coding sequences of non-essential genes (S2), auxiliary UTRs and introns (S3), unannotated regions (S4), and the top 10% of insertion sites (S5).

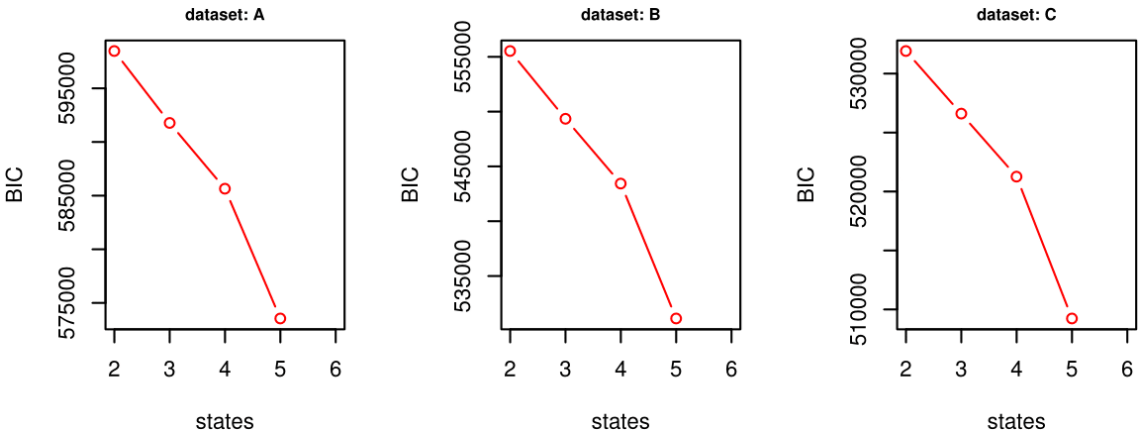


Figure 4.17. Selection of a Five-State HMM based on the Bayesian Information Criterion (BIC). BIC values for three datasets showing that adding more states to the HMM improves its robustness. In general, the model with the lowest BIC is preferred, in this case, the five-state HMM.

In addition to the selection criterion, a stopping criterion was applied. In an iterative algorithm, such as the five-state HMM algorithm, a stopping criterion is needed to determine when to stop the iteration. In this case, the stopping criterion was set as the log-likelihood (*loglik*) of the data. In statistics, the likelihood of a parameter value, θ, given outcomes *x*, is equal to the probability assumed for those observed outcomes given those parameter values. Generally, the higher the log-likelihood, the better the fit to the data. For all datasets tested, the log-likelihood plateaued at 150 iterations (Figure 4.18). Overall, the run time was 20 hours.

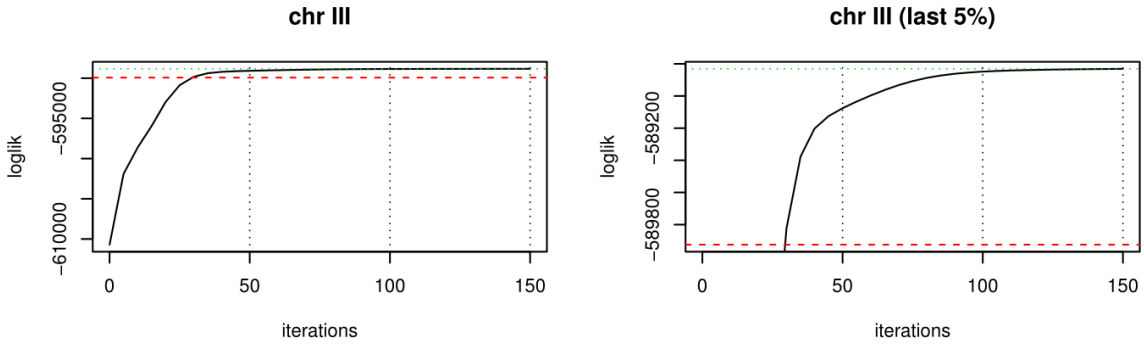


Figure 4.18. The Five-State HMM Iterative Algorithm. Plotted, a graph of log-likelihood (*loglik*) against iterations for chromosome III. In this algorithm, the total number of iterations was set to 150.

Ultimately, so as to ascertain that the model is robust to downsampling, the five-state HMM was run on a subset of both the log phase and the ageing data (Figure 4.19).

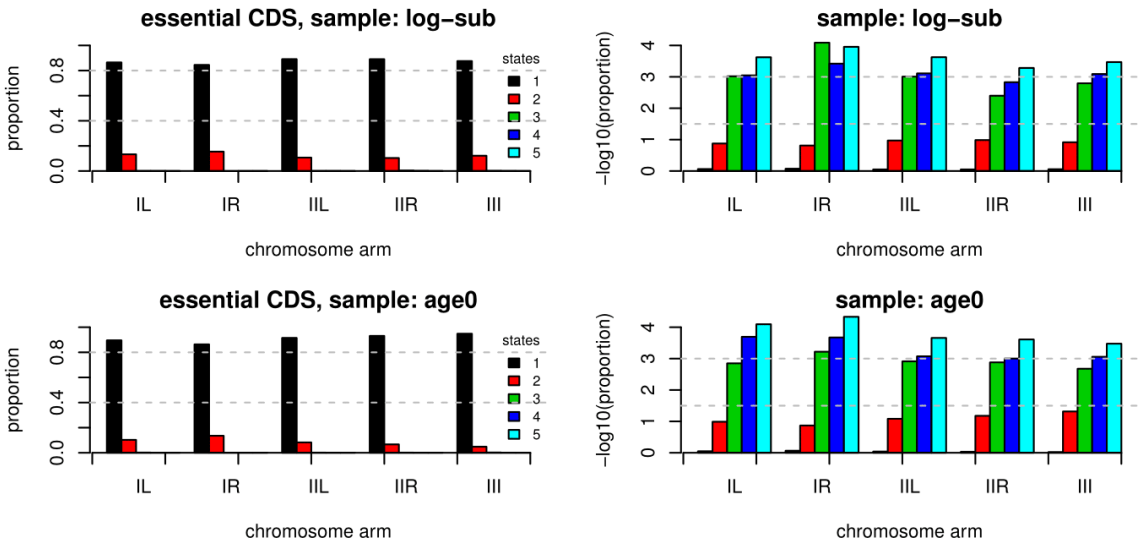


Figure 4.19. The Five-State HMM and its Robustness to Downsampling. Running the five-state HMM on a subset of both the log phase (*top*) and the ageing (*bottom*) data confirmed that the established five-state HMM is robust to downsampling. On the x-axis, IL and IR refer to the left and right arms of chromosome I respectively; IIL and IIR to the left and right arms of chromosome II respectively; and III to chromosome III.

4.2.3 HMM Results

Once the five-state HMM was established and proofed, it was critical to determine whether the five states have any biological importance. Specifically, what can be inferred from the HMM about the functional elements in the genome? Figure 4.20 shows that the HMM has clear biological relevance since 99% of essential coding sequences (eCDS) were assigned to either S1 or S2; eCDS are the most highly conserved regions and contain the lowest within-species diversity. Overall, 91% of the genome was defined as S1 or S2, indicating that insertion mutations of this kind have detectable functional consequences over the majority of the genome. 40% of the fission yeast genome that is not protein-coding also contains function, since 81% of the non-protein-coding regions are S1 or S2. Based on this analysis, UTRs, unannotated regions and non-coding RNAs all contain similar proportions of functional DNA.

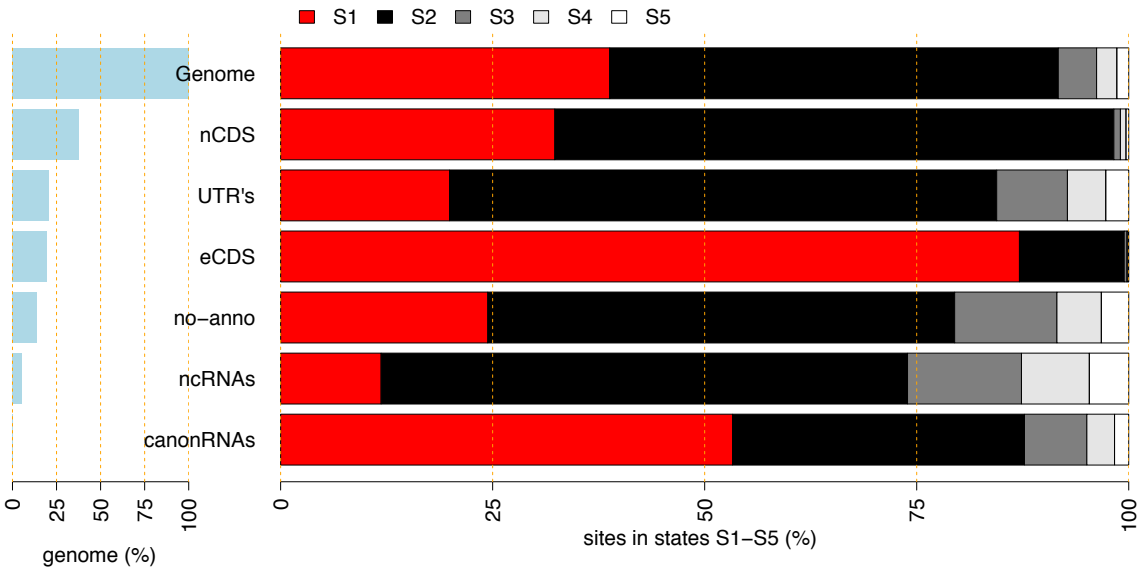


Figure 4.20. Inferring Functional Elements from the Five-State HMM. For the genome, protein-coding regions of non-essential genes (nCDS), untranslated regions (UTRs), protein-coding regions of essential genes (eCDS), regions with no annotation (no–anno), non-coding RNAs (ncRNAs), and canonical RNAs (canonRNAs), we show the percentage of sites that are in states S1 to S5.

(a) Protein-Coding Regions

Consequently, we examined whether the HMM contained additional information about the functional significance of protein-coding genes. To this end, we calculated the mean HMM state for each protein-coding gene, speculating that this is a good measure of the fitness effect of deleting that gene. Figure 4.21 shows that essential protein-coding genes (whose deletions are inviable) have a greater proportion of sites assigned to S1 than non-essential genes, as expected. Also, mean states for essential and non-essential genes form overlapping distributions, showing that the binary classification of essential/non-essential genes could be improved using this mean HMM state.

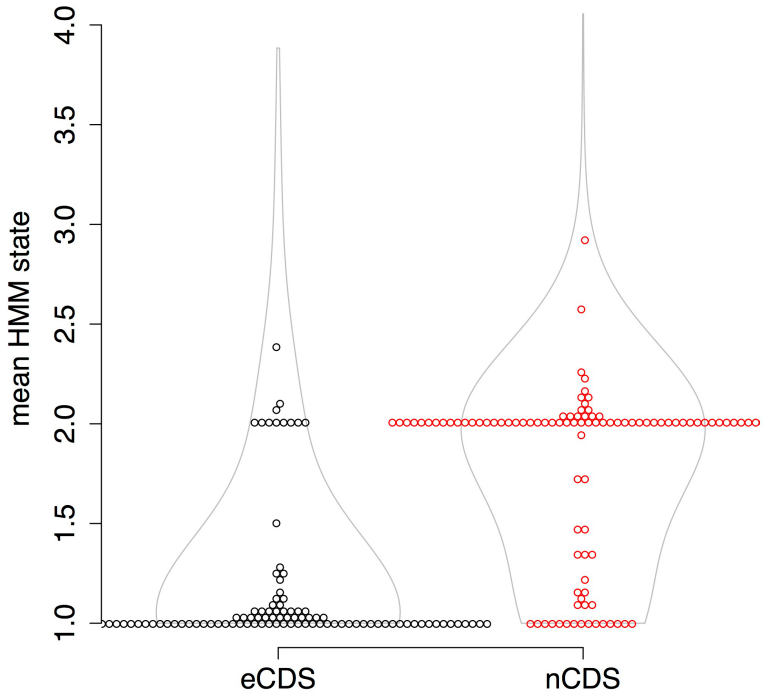


Figure 4.21. Mean HMM States for Essential (eCDS) and Non-Essential (nCDS) Coding Sequences. eCDS have a greater proportion of sites assigned to S1 than nCDS.

So as to address this principle further, the next aim was to determine whether the mean HMM state relates to other measures of fitness, such as growth rates of knockout mutants (Malecki and Bähler 2016). 3419 genes whose deletions are viable were explored, and information about their functional consequences was obtained. Here, the mean HMM state positively correlated with colony size of the deletion mutants (Pearson $r = 0.29$, $p < 10^{-16}$), indicating that this measure is related to the fitness cost of gene deletion. From these viable knockout mutants, 620 have a mean HMM state of 1, showing that transposon insertion mutations have strong effects, even though the deletions are viable. Using the AnGeLi tool (Bitton *et al.* 2015), we discovered that these 620 genes are strongly enriched (Fisher tests $p < 10^{-10}$) for gene ontology terms that are consistent with functional roles during mitotic growth, such as mitotic cell cycle, and with deletion phenotypes, such as abnormal subcellular components and decreased vegetative cell population growth. In addition, this gene set is enriched for 1:1 human orthologs. Collectively, these results indicate that mean HMM states are able to

define functionally important genes beyond the binary classification of essential/non-essential.

(b) Non-Protein-Coding Regions

In general, the *Hermes* insertion data can be used to predict which ncRNAs are required for growth. In view of the HMM, these are projected to be regions with mean HMM state < 1.5, coverage > 50%, and not overlapping coding genes. Overall, 85 transcripts met these criteria (Figure 4.22), with over 75% having 100% coverage. It is assumed that knocking out these ncRNAs will result in growth-arrested or slow-growing cells. Rodríguez-López *et al.* (2016), who deleted ncRNAs using the CRISPR/Cas9 gene editing approach, show the opposite for SPNCRNA.37, SPNCRNA.137, and SPNCRNA.284. It is possible that the fitness cost of a deletion is different to that of a *Hermes* insertion, meaning that while the deletion of certain ncRNAs (with mean HMM state < 1.5) does not inhibit cell growth, *Hermes* insertions within the same ncRNAs are deleterious.

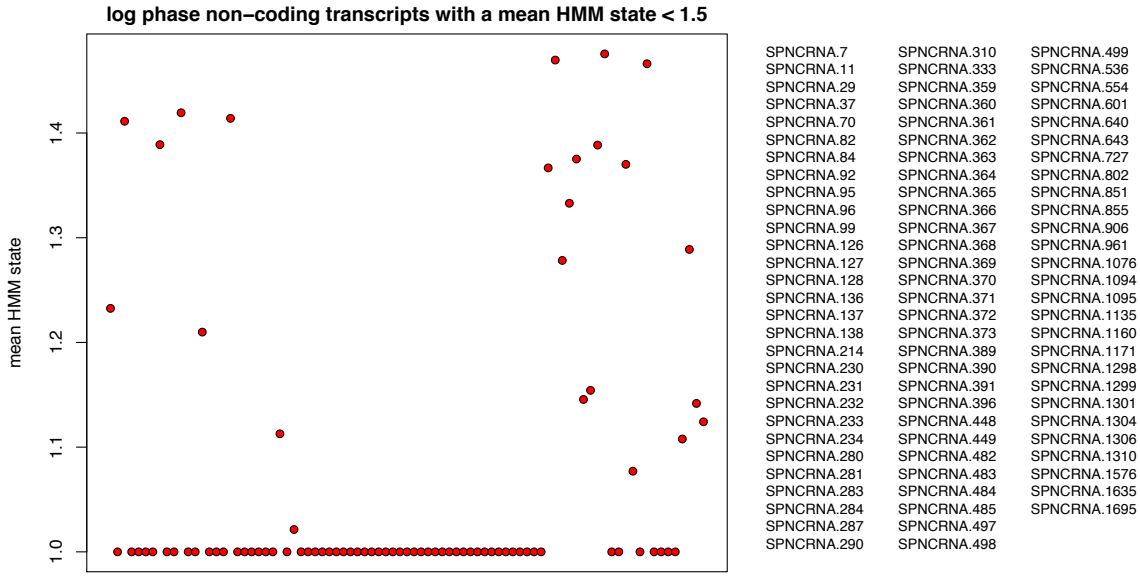


Figure 4.22. Log Phase Non-Coding Transcripts with a Mean HMM State < 1.5. Overall, we find 85 ncRNAs with mean HMM state < 1.5, coverage > 50%, and not overlapping coding genes (Supplementary List A).

4.3 Mitochondrial Insertions

In this work, an interesting aspect of the data that also required a more comprehensive evaluation is the mitochondrial insertions. Using both datasets, we discovered that over 100 nt windows, the mitochondrial genome has far more *Hermes* insertions per site than the nuclear genome. Indeed, the mitochondrial genome has a median of 1 insertion per 4.5 nt whereas the nuclear genome has a median of 1 insertion per 20 nt. In principle, the occurrence of mitochondrial insertions can be attributed to NUMTs, or nuclear mitochondrial DNA segments. Lopez *et al.* (1994) coined NUMTs to describe transposition of mitochondrial DNA into eukaryotic nuclear genomes. In both the lab and in nature, NUMTs have been shown to enter nuclear DNA via non-homologous end joining (NHEJ) at double-strand breaks (Hazkani-Covo *et al.* 2010). NUMT sequences of different length and size have been detected across a range of organisms, from yeasts (Sacerdot *et al.* 2008) to humans (Mourier *et al.* 2001).

NUMTs have also been elucidated in *S. pombe* (Lenglez *et al.* 2010), and therefore, the theoretical likelihood that the observed mtDNA insertions correspond to insertions in NUMT sequences is high. In addition, this is compounded by the fact that NUMTs are located in non-protein-coding regions (Lenglez *et al.* 2010), and we find that the mitochondrial genome has more unique insertions per site than the nuclear genome. However, while the case for NUMTs is strong, it is also possible that what we see are actual insertions in the mitochondria. Regardless, we also discovered that there is little biological signal in the mitochondria (Figure 4.23), perhaps because *Hermes* insertions in one of multiple copies had negligible consequences for the cell. Therefore, we excluded mitochondrial sites from further examination.

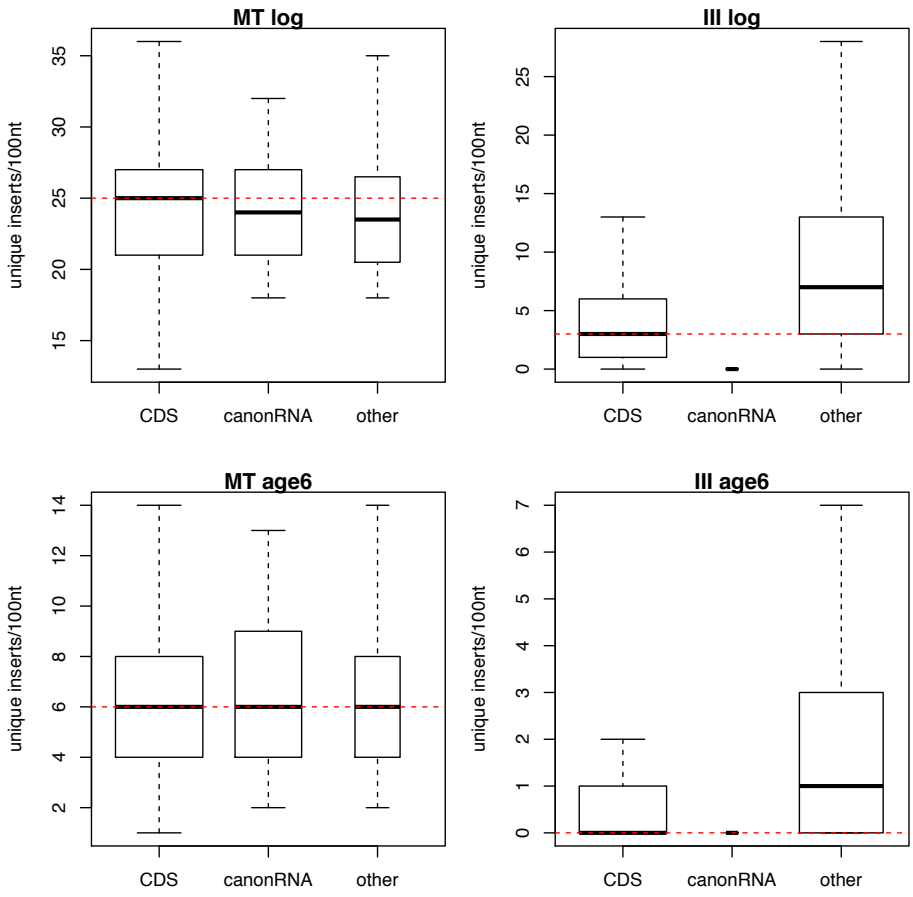


Figure 4.23. Mitochondrial Insertions. Here, we calculated unique insertions per site for coding sequences (CDS), canonical RNAs, and all other regions in the mitochondria (MT), and also chromosome III as a comparison. Overall, results show that there is little biological signal in the mitochondria.

4.4 Summary of the Main Results

Overall, results show that *Hermes* integrates across all three chromosomes of the genome with minimal bias, an observation that is consistent with previous studies. *Hermes* insertion data is a strong predictor of gene essentiality, as determined by comparative analysis to genome annotation, gene expression levels, colony sizes and growth scores, and constraint and genetic diversity.

In order to characterise the essentiality of the entire fission yeast genome, a five-state Hidden Markov Model (HMM) was developed, with states classified as in Table 4.4. For State 5, insertion densities (insertions/100 nt) were used instead of insertion counts.

State No.	Classification	Training Data
1	essential	coding sequences of essential genes
2	very important	coding sequences of non-essential genes
3	important	auxiliary UTRs and introns
4	somewhat important	unannotated regions
5	not important	top 10% of insertion sites

Table 4.4. HMM State Categorisation. In tabular form, a classification of the five HMM states used in this research investigation.

HMM results show that the five-state model has clear biological relevance since 99% of essential coding sequences (eCDS) were assigned to either S1 or S2; eCDS are the most highly conserved regions and contain the lowest within-species diversity. Overall, 91% of the genome was defined as S1 or S2, indicating that *Hermes* insertion mutations have detectable functional consequences over the majority of the genome. 40% of the genome that is not protein-coding also contains function, since 81% of the non-protein-coding regions are S1 or S2. Specifically, we find 85 ncRNAs with mean HMM state < 1.5, coverage > 50%, and not overlapping coding genes.

Chapter 5 EVALUATING THE AGEING DATASET

5.1 Introduction

In both budding and fission yeast, two different types of ageing have been studied: replicative and chronological. In replicative lifespan, the viability of yeast cells decreases with the number of cell generations, whereas in chronological lifespan (CLS), it decreases with the time the cells spend in stationary phase. In other words, in CLS, cells show a decline in viability after entering stationary phase, until, ultimately, all cells in the culture are dead. It has been established that starvation is not the only cause of cell death (Fabrizio and Longo 2003). In fact, it has been proposed that the loss of viability in chronologically ageing cells is mainly due to ethanol production (Fabrizio *et al.* 2005), toxicity induced by reactive oxygen species, and loss of mitochondrial function (Longo and Fabrizio 2002). It is the result of a cocktail of pathways, molecular mechanisms, and genes, only some of which have been identified so far.

In most of the studies that have attempted to understand how CLS works, the model organism used was the budding yeast *Saccharomyces cerevisiae*. However, an assay for the fission yeast *Schizosaccharomyces pombe* has also been described (Roux *et al.* (2006, 2009)). In this screen, in addition to log phase libraries, aged *Hermes* libraries were generated via two, independent CLS assays. In total, seven, twenty-four-hour-apart time points were taken (Figure 5.1) although only the even time points (t0, t2, t4, t6) were processed and analysed. For these time points, fluorescence microscopy images were also captured (Figure 5.2). Consistently, for both repeats, the beginning of the CLS curve (time point 0, or t0) was taken following two cell cycles when the optical density remained unchanged. Progressing from t0 to t6, the aim is to determine which genes change during ageing.

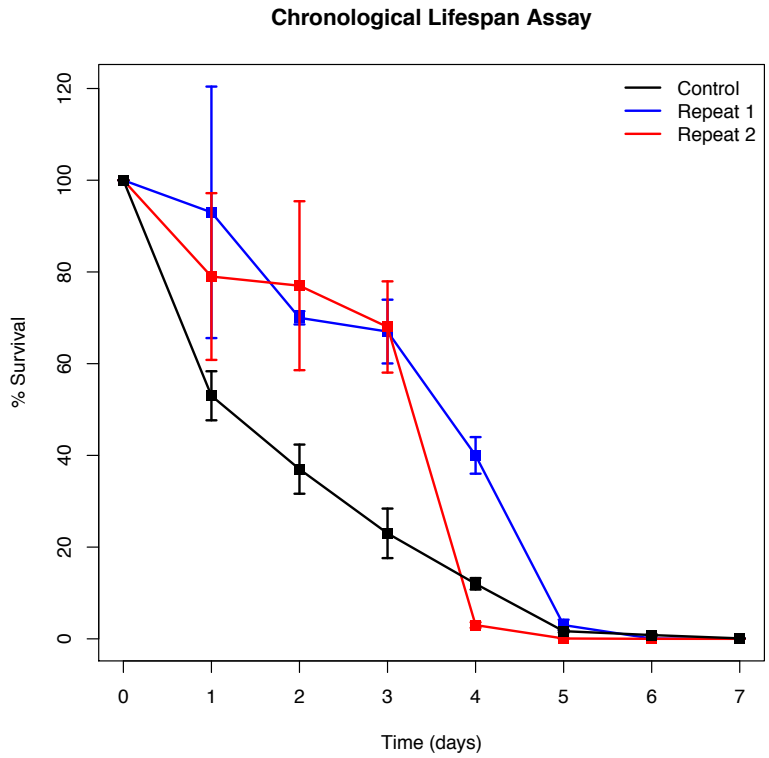


Figure 5.1. Chronological Lifespan (CLS) of the *Hermes* Mutants. CLS assays monitor the decline in viability without detectable regrowth in a cell culture over time. Illustrated, CLS survival curves of two, independent *Hermes* mutant pools, with error bars at each time point representing the range of three technical replicates. Data for wild type control was obtained from Rallis *et al.* (2013).

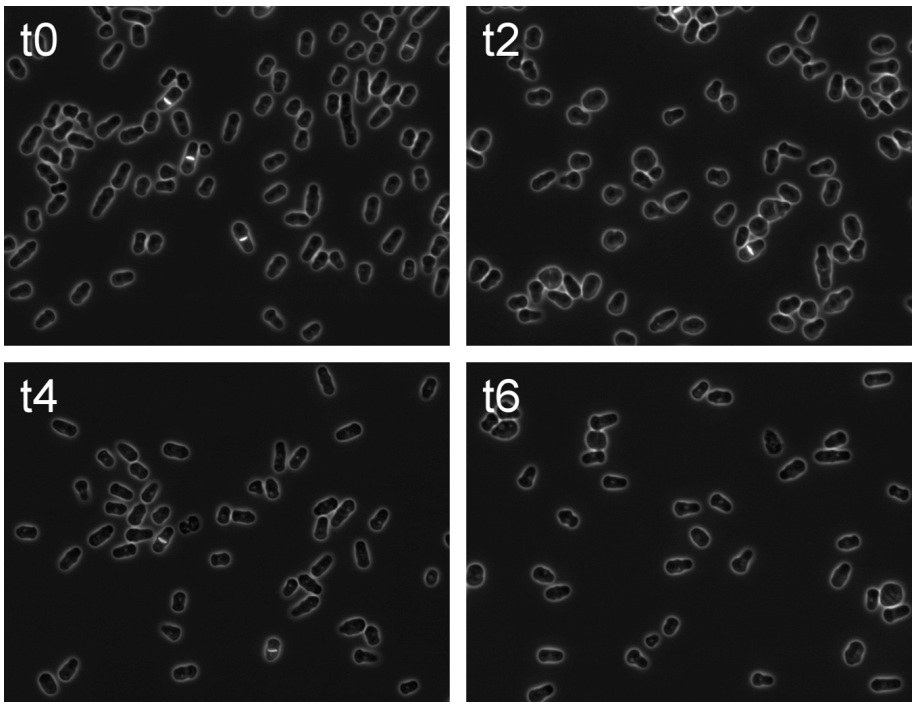


Figure 5.2. Fluorescence Microscopy Images of the *Hermes* Mutants. Overall, *Hermes* mutant cells exhibit an altered morphology when compared to wild type cells. Progressing from t0 to t6, cell death is characterised by a decrease in the number of cells undergoing cell division, a reduction in the cellular volume, and rounding-up of the cells (viewed at 63x magnification).

5.2 In Search of Genes that Change during Ageing

Similar to any other genomic dataset, the aged *Hermes* libraries can be analysed and dissected in a number of different ways. Each approach on its own has its caveats, but the aim ultimately is to uncover a biologically meaningful pattern when all analyses are taken together into consideration. Pivotal in the collective analysis is the difference between unique insertion sites and unique insertion counts, where a *site* refers to a single insertion, and the *count* refers to the number of times that insertion is present within the sequenced libraries. For each statistical test conducted, either one of these measures is used. Overall, the main focus of each approach is to discover genes that change during ageing.

5.2.1 Using Unique Insertion Sites

So as to uncover genes that change with time, a few Cochran–Mantel–Haenszel (CMH) tests were performed. CMH tests, similar to chi-square tests but with one degree of freedom, are used to test for conditional independence between repeats. Here, a CMH test was used to compare the number of unique insertion sites at each time point between the two aged libraries. Over 5000 genomic features, comprising 122 long intergenic non-coding RNAs, essential and non-essential genes, were screened. For this test, the hypothesis was that with time, the number of unique insertion sites changes (decreases or increases) in a gene but remains constant in the region surrounding it. Therefore, two 5 kb regions, one on each side of the genomic feature (Figure 5.4, *top*), were also included in the test. Each gene's p-value was then corrected by the Benjamini-Hochberg method for

multiple testing, with a false discovery rate of < 0.05 as the significance threshold for the adjusted p-values. In total, seven genes that change significantly in both aged libraries were found (Table 5.1).

systematic i.d.	name	p-value	adjusted p-value
SPAC56F8.04c	ppt1	1.37E-08	6.86E-05
SPAC9G1.02	wis4	1.67E-06	2.79E-03
SPBC25H2.13c	cdc20	1.08E-05	1.35E-02
SPBPB7E8.02	-	1.22E-06	2.79E-03
SPCC1281.06c	ole1	5.86E-05	4.19E-02
SPCC1753.04	tol1	1.80E-05	1.81E-02
SPCC417.08	tef3	3.30E-05	2.75E-02

Table 5.1. First Cochran–Mantel–Haenszel (CMH) Test Results. Overall, seven genes with significant, Benjamini-Hochberg adjusted p-values < 0.05 were uncovered.

For each of the seven genes, a ratio between unique insertion sites in the gene (G) and the gene region (R) was calculated at all four time points. In general, only SPCC417.08 (or *tef3*) shows an increase in G/R ratio with time (Figure 5.3). Indeed, compared to the gene region, unique insertion sites in this gene seem to be doubling with time. Kim *et al.* (2010) established that deleting *tef3* results in an inviable cell population. Therefore, according to both current annotation and our data, *tef3* is essential during vegetative growth. Here, we also show that *tef3* is a pro-ageing gene, meaning that it is advantageous to have insertions in this gene during ageing.

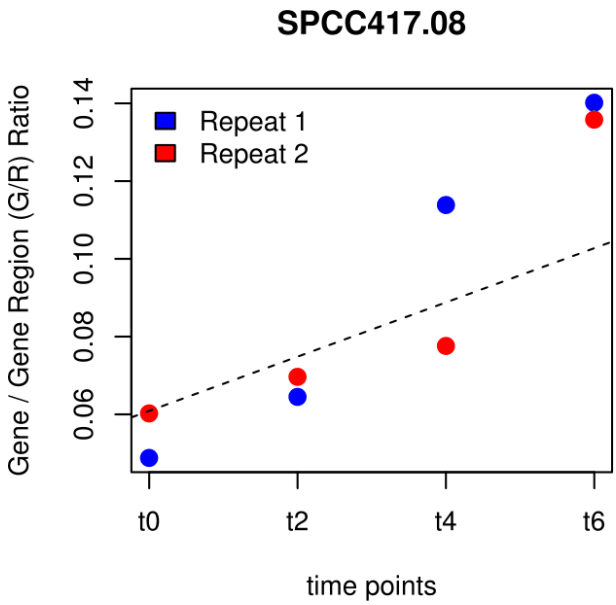


Figure 5.3. Gene / Gene Region Ratios for the *tef3* gene. SPCC417.08, or *tef3*, is an essential gene that shows a marked increase in G/R ratio with time.

While informative, this gene-based analysis has at least two limitations. First, the overall 10 kb gene region for any one gene could be overlapping with neighbouring genes and their gene regions, therefore biasing the G/R ratios. Second, this approach favours long genes with their relatively high number of unique insertions, thus making it more difficult to find true significant p-values. In fact, a quick and simple chi-square test shows that for two genes of different length, if the fold change in the number of unique insertions from t0 to t6 is the same in both genes, the p-value is almost always lower for the longer gene. One way of correcting this artefact is to use a standard length for every gene.

To this end, a modified version (v2.0) of this analysis employs a similar pipeline as above, but instead of gene and 10 kb gene regions, it uses mean and median unique insertion sites of 1, 2 or 3 kb gene windows. Each gene window is centred around the gene midpoint and extends equally in both directions. For example, for a 3 kb gene window, there is 1.5 kb either side of the gene midpoint (Figure 5.4, *bottom*). Similar to the original approach, a CMH test was carried out separately

on each of the three gene windows followed by Benjamini-Hochberg corrections for multiple testing.

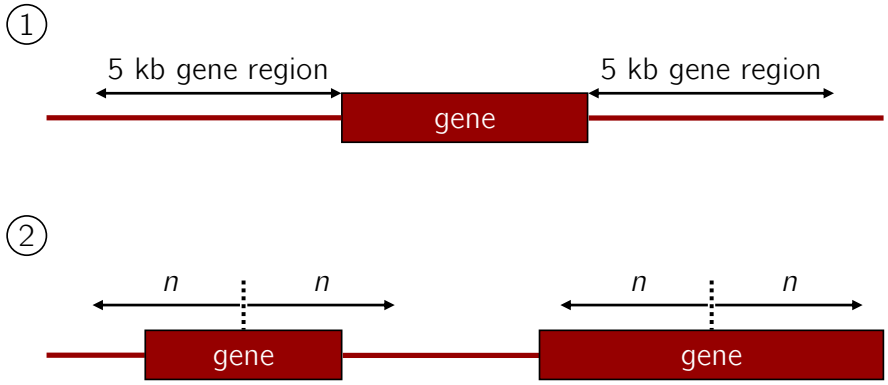


Figure 5.4. Gene-Based Ageing Analyses. *Top*, analysis v1.0 uses the number of unique insertion sites within a gene and the overall 10 kb region surrounding it. *Bottom*, analysis v2.0 uses the mean and median unique insertion sites of *2n* kb gene windows where n = 500 nt, 1 kb and 1.5 kb for 1, 2 and 3 kb gene windows respectively. For both analyses, a gene was defined as a whole, therefore, it was set to include untranslated regions, introns, and exons, rather than just the coding sequences.

In reference to Table 5.2, a cross-comparison of the top 10 genes with the lowest significant p-values within each of the four analyses (i.e. G and 10 kb R (v1.0) and three gene windows (v2.0: 1/2/3 kb)) was performed. Interestingly, out of the 9 resulting genes, 8 overlap in two of the analyses whereas SPBC21.06c (or *cdc7*) is the only gene uncovered in three. Since this cross-comparison only includes genes that overlap in two or more analyses, Table 5.3 also lists the top 10 genes with Benjamini-Hochberg adjusted p-values < 0.05 for at least one of the four analyses. Intersecting the two lists through a Venn diagram results in 5 overlapping genes (Figure 5.5).

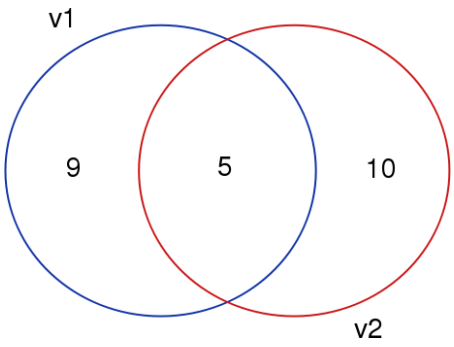


Figure 5.5. Venn Diagram Intersecting Gene Lists in Table 5.2 and Table 5.3. (see main text for details)

For all 14 genes, an IPKM value was calculated and plotted at each time point. IPKM, analogous to the more commonly known RPKM, stands for insertions per kilobase per million insertions. IPKM was computed using:

IPKM = IPK/(total unique insertion sites in a library/1,000,000)

where *IPK = C/(L/1000)* and C is unique insertion sites in a region (say, a gene) and L is the length of the region in nt. Identical to the G/R ratio results, SPCC417.08 (or *tef3*) is the only gene that consistently, in both repeats, shows an increase in IPKM with time (Figure 5.6).

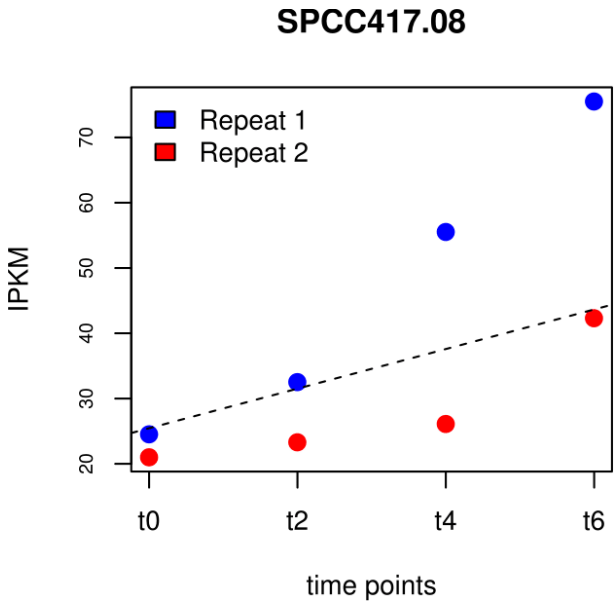


Figure 5.6. IPKM for the *tef3* gene. IPKM stands for insertions per kilobase per million insertions. SPCC417.08, or *tef3*, is an essential gene that shows a marked increase in IPKM with time.

gene			v2.0: 1 kb		v2.0: 2 kb		v2.0: 3 kb		v1.0	
systematic i.d.	*name*	*ontology term*	*p-value*	*adjusted*	*p-value*	*adjusted*	*p-value*	*adjusted*	*p-value*	*adjusted*
SPAC1687.12c	coq4	cofactor metabolic process	**8.14E-04**	**0.50925**	1.69E-01	0.99998	2.53E-01	0.99999	**5.39E-04**	**0.26969**
SPAC56F8.04c	ppt1	cofactor and lipid metabolic process	**4.96E-06**	**0.02483**	1.54E-01	0.99998	3.46E-01	0.99999	**1.37E-08**	**0.00007**
SPBC21.06c	cdc7	mitotic cytokinesis and mitotic cell cycle regulation; signalling	**1.75E-04**	**0.29171**	**1.86E-04**	**0.26117**	**4.29E-06**	**0.02154**	2.73E-01	0.99999
SPBC887.14c	pfh1	DNA recombination, repair and replication; mitochondrion and telomere organisation	**2.94E-04**	**0.36812**	1.35E-02	0.99998	9.53E-02	0.99999	**4.40E-04**	**0.24440**
SPBP4H10.15	aco2	cellular amino acid metabolic process; mitochondrion organisation; precursor metabolites and energy generation	2.87E-02	0.99998	**1.06E-03**	**0.66484**	**2.33E-04**	**0.29307**	1.27E-02	0.99999
SPBPB7E8.02	-	-	4.40E-03	0.99998	**1.92E-04**	**0.26117**	2.09E-03	0.99999	**1.22E-06**	**0.00279**
SPCC162.09c	hmg1	cofactor and lipid metabolic process	7.69E-02	0.99998	**3.09E-04**	**0.30992**	**1.85E-04**	**0.29307**	2.87E-02	0.99999
SPCC1753.04	tol1	lipid metabolic process	**4.27E-04**	**0.40621**	1.15E-01	0.99998	2.69E-01	0.99999	**1.80E-05**	**0.01805**
SPCC417.08	tef3	cytoplasmic translation	1.93E-01	0.99998	1.93E-03	0.80799	**1.34E-04**	**0.29307**	**3.30E-05**	**0.02751**

Table 5.2. CMH Test: Cross-Comparison of the Top Genes with the Lowest Benjamini-Hochberg Adjusted P-Values. Gene-Centric Analysis v1.0 (far right column) performs a Cochran–Mantel–Haenszel (CMH) test on the number of unique insertion sites at each time point for each gene and the corresponding 10 kb gene region. Gene-Centric Analysis v2.0 also performs a CMH test but using mean and median unique insertion sites of 1, 2 or 3 kb gene windows (see main text for details). For both analyses, each p-value was corrected by the Benjamini-Hochberg method for multiple testing, with a false discovery rate of < 0.05 as the significance threshold for the adjusted p-values. Here, the 9 genes listed are the result of a cross-comparison of the top 10 genes with the lowest Benjamini-Hochberg adjusted p-values (highlighted in bold) within each of the four analyses, and therefore, only genes present in two or more analyses are included.

gene			v2.0: 1 kb		v2.0: 2 kb		v2.0: 3 kb		v1.0	
systematic i.d.	*name*	*ontology term*	*p-value*	*adjusted*	*p-value*	*adjusted*	*p-value*	*adjusted*	*p-value*	*adjusted*
SPAC4A8.03c	ptc4	autophagy; membrane organisation; precursor metabolites and energy generation; signalling	**1.55E-05**	**0.03868**	6.65E-01	0.99998	8.34E-01	0.99999	3.49E-02	0.99999
SPAC56F8.04c	ppt1	cofactor and lipid metabolic process	**4.96E-06**	**0.02483**	1.54E-01	0.99998	3.46E-01	0.99999	**1.37E-08**	**0.00007**
SPAC9G1.02	wis4	signalling	3.78E-01	0.99998	2.11E-03	0.81550	2.64E-03	0.99999	**1.67E-06**	**0.00279**
SPBC21.06c	cdc7	mitotic cytokinesis and mitotic cell cycle regulation; signalling	1.75E-04	0.29171	1.86E-04	0.26117	**4.29E-06**	**0.02154**	2.73E-01	0.99999
SPBC25H2.13c	cdc20	chromatin organisation; DNA repair and replication; transcription regulation	1.52E-02	0.99998	9.11E-02	0.99998	3.77E-02	0.99999	**1.08E-05**	**0.01345**
SPBPB7E8.02	-	-	4.40E-03	0.99998	1.92E-04	0.26117	2.09E-03	0.99999	**1.22E-06**	**0.00279**
SPCC1281.06c	ole1	lipid metabolic process	6.48E-03	0.99998	1.27E-02	0.99998	3.55E-01	0.99999	**5.86E-05**	**0.04185**
SPCC1322.01	rpm1	mitochondrion organisation	1.91E-02	0.99998	**8.67E-06**	**0.04354**	1.03E-01	0.99999	2.69E-02	0.99999
SPCC1753.04	tol1	lipid metabolic process	4.27E-04	0.40621	1.15E-01	0.99998	2.69E-01	0.99999	**1.80E-05**	**0.01805**
SPCC417.08	tef3	cytoplasmic translation	1.93E-01	0.99998	1.93E-03	0.80799	1.34E-04	0.29307	**3.30E-05**	**0.02751**

Table 5.3. CMH Test: Genes with Benjamini-Hochberg Adjusted P < 0.05. Gene-Centric Analysis v1.0 (far right column) performs a Cochran–Mantel–Haenszel (CMH) test on the number of unique insertion sites at each time point for each gene and the corresponding 10 kb gene region. Gene-Centric Analysis v2.0 also performs a CMH test but using mean and median unique insertion sites of 1, 2 or 3 kb gene windows (see main text for details). For both analyses, each p-value was corrected by the Benjamini-Hochberg method for multiple testing, with a false discovery rate of < 0.05 as the significance threshold for the adjusted p-values. Here, the 10 genes listed are those with Benjamini-Hochberg adjusted p-values < 0.05 (highlighted in bold) for at least one of the four analyses (c.f. Table 5.2).

5.2.2 Using Unique Insertion Counts

In the next part of the analysis, the same CMH tests as above were carried out, followed by Benjamini and Hochberg's multiple testing correction method. However, in this case, unique insertion counts were used instead of unique insertion sites. Over 1200 genes resulted when the three count datasets (for 1, 2, and 3 kb gene windows) were merged and the genes with adjusted p-values < 0.05 in each of the datasets were subsetted. For these genes, a Spearman rank correlation coefficient test was then performed between the four time points and a ratio of insertion counts and mean insertion counts. In statistics, a Spearman correlation test assesses monotonic relationships; a monotonic relationship exists when an increase in one variable is accompanied by an increase (or decrease) in the other variable. Specifically, the Spearman coefficient determines the direction and strength of the relationship between two variables. Provided that there are no repeated data values, an ideal Spearman coefficient of +1 or −1 occurs when each of the variables is a perfect monotonic function of the other.

In this Spearman correlation test, three output parameters were taken into account: the correlation coefficient (*r*), the correlation p-value (*cor.p*), and the Benjamini-Hochberg adjusted correlation p-value (*adj.cor.p*). Table 5.4 shows that 13 out of over 1200 genes have absolute correlation coefficient values $r > 0.9$, with 5 having a positive correlation and 8 having a negative one. For all 13 genes, graphs of ratio of count and mean count against time points were plotted, with two contrasting examples provided in Figure 5.7. In addition, 4 out of 13 genes have adjusted correlation p-values *adj.cor.p* < 0.05. For each of these 4 genes, insertion counts were converted to insertion counts per million insertion counts in the library, and then, for each time point, graphs of log2 insertion counts per million against chromosome position were plotted, using the same two genes as in Figure 5.7 as examples (Figure 5.8).

genes with a positive (+ve) correlation coefficient		genes with a negative (-ve) correlation coefficient	
systematic i.d.	*r*	*systematic i.d.*	*r*
SPAC343.18	0.98	SPAC1783.06c	−0.98
SPBC6B1.10	0.98	SPBC17G9.08c	−0.98
SPAC20G8.01	0.93	SPCC550.01c	−0.98
SPAPYUG7.06	0.93	SPAC17H9.10c	−0.95
SPBC119.13c	0.91	SPAC22A12.17c	−0.93
		SPBC1347.01c	−0.93
		SPBC21C3.09c	−0.93
		SPAC637.11	−0.91

Table 5.4. Spearman Correlation Test Results. 13 out of over 1200 genes (see main text for details) have absolute correlation coefficient values $r > 0.9$ between count / mean count ratio and time, with 5 having a positive correlation and 8 having a negative relationship.

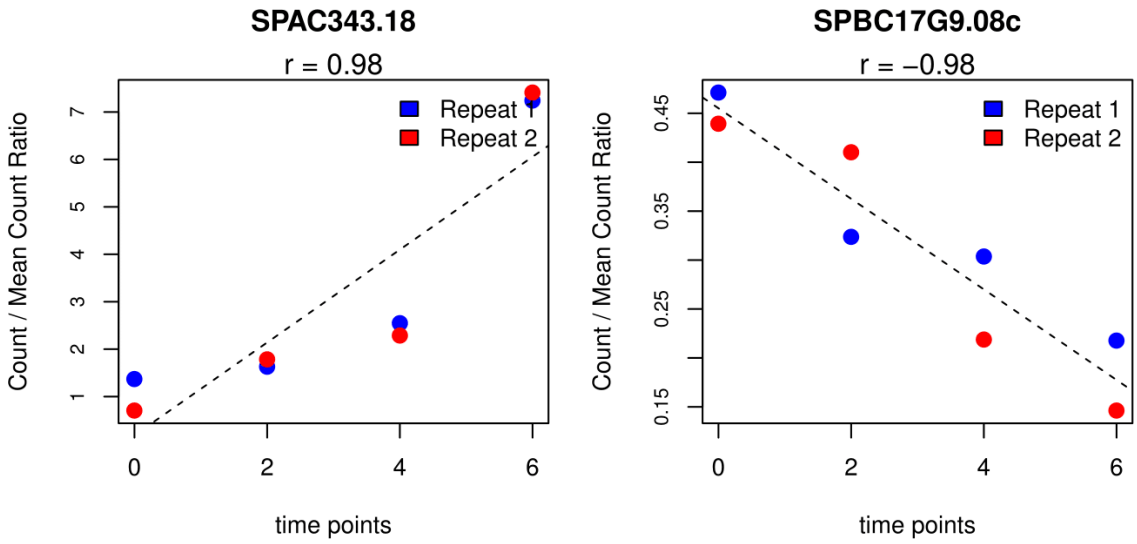


Figure 5.7. Count / Mean Count Ratio Plots. *Left*, for SPAC343.18, or *rfp2*, there is a high positive correlation coefficient between count / mean count ratio and time, indicating that insertions are enriched with time. *Right*, for SPBC17G9.08c, or *cnt5*, there is a high negative correlation coefficient between count / mean count ratio and time, suggesting that insertions are depleted with time.

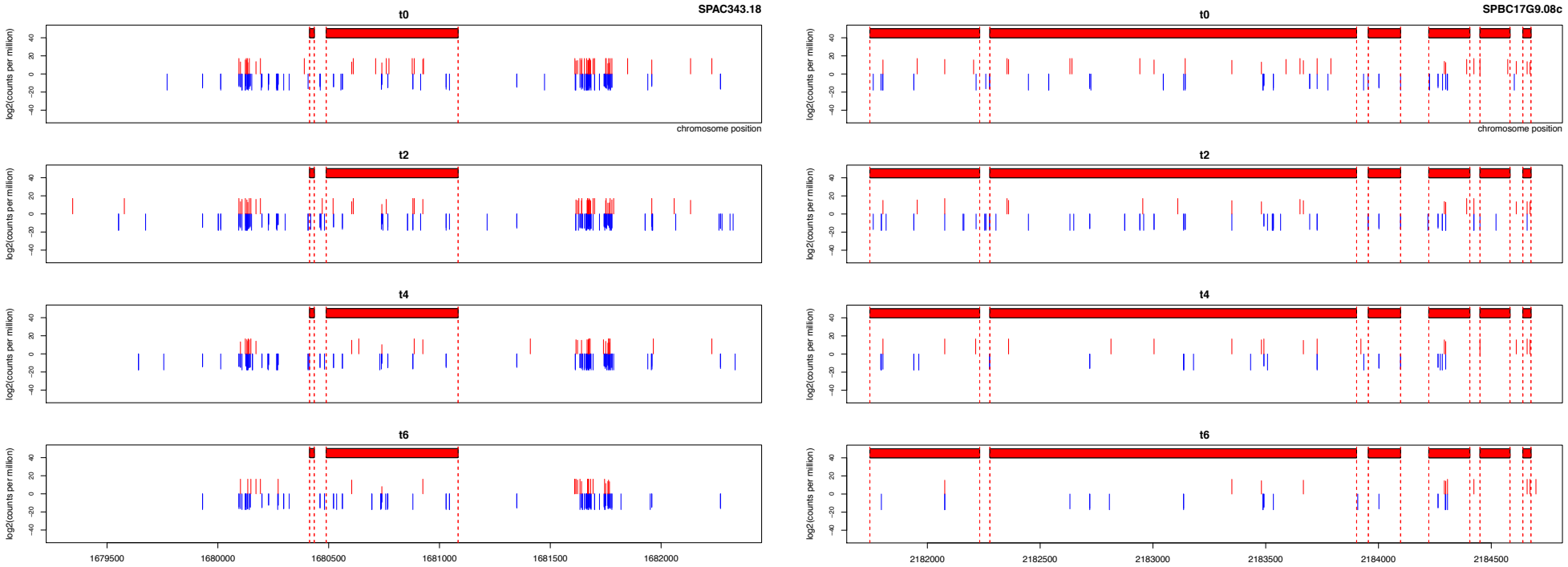


Figure 5.8. Log2 Insertion Counts Per Million vs. Chromosome Position Plots. Illustrated, two genes with Benjamini-Hochberg adjusted correlation p-values < 0.05. *Left*, SPAC343.18, or *rfp2*, is enriched for insertions with time, at least for one of the repeats. *Right*, SPBC17G9.08c, or *cnt5*, is depleted for insertions with time. For each time point, 3 kb windows were plotted, with the red boxes at the top of each plot outlining the coding sequences of the gene. Repeats 1 and 2 are represented by blue (-log2 scale) and red (+log2 scale) lines respectively.

So as to expand the search, 190 genes with a correlation p-value *cor.p* < 0.05 were filtered, with 54 having a positive correlation and 136 having a negative one. Independently, both lists were then analysed with AnGeLi, a tool for the comprehensive and customised interrogation of gene lists from fission yeast (Bitton *et al.* 2015). In this gene list enrichment analysis, Benjamini and Hochberg (false discovery rate (FDR)) was applied as the multiple testing correction method with a < 0.01 cutoff, using all genes as background and biological process as category. Table 5.5 shows the output for both analyses. Interestingly, 29 out of the 54 positively correlated genes were enriched for FYPO:0002061, annotating genes whose deletion results in inviable vegetative cell populations, and are therefore essential. In contrast, 124 out of the 136 negatively correlated genes were enriched for FYPO:0002060, annotating genes whose deletion results in viable vegetative cell populations, and are therefore non-essential.

For the SPAC343.18 (or *rfp2*) gene (Figure 5.7, *left*), for example, characterised by a positive correlation coefficient, this indicates that the gene could be essential during log phase growth (at t0) but not at later time points. On the other hand, for the SPBC17G9.08c (or *cnt5*) gene (Figure 5.7, *right*), characterised by a negative correlation coefficient, this suggests that the gene is perhaps not essential during growth but is switched on as cells age (at t2, t4, and t6).

Genes with a Positive Correlation (Pro-Ageing Genes)				
FYPO	***FYPO Annotation***	***List Frequency***	***Background Frequency***	***Corrected P-Value***
FYPO:0000001	phenotype	100 (54/54)	70.8 (4959/7005)	1.51E-05
FYPO:0000002	cell phenotype	100 (54/54)	70.1 (4908/7005)	1.51E-05
FYPO:0000003	cell population phenotype	100 (54/54)	70.6 (4949/7005)	1.51E-05
FYPO:0000136	cellular physical quality phenotype	100 (54/54)	69.7 (4881/7005)	1.51E-05
FYPO:0002057	cell population viability	100 (54/54)	70.5 (4938/7005)	1.51E-05
FYPO:0004639	abnormal cellular physical quality phenotype during vegetative growth	59.3 (32/54)	24.2 (1696/7005)	4.91E-05
FYPO:0002061	**inviable vegetative cell population**	**53.7 (29/54)**	**20.7 (1452/7005)**	**8.36E-05**
FYPO:0000004	cell viability	98.1 (53/54)	69.2 (4845/7005)	8.62E-05
FYPO:0002059	inviable cell population	53.7 (29/54)	20.9 (1466/7005)	8.62E-05
FYPO:0003037	abnormal cell phenotype	68.5 (37/54)	33.3 (2332/7005)	8.93E-05
Genes with a Negative Correlation (Anti-Ageing Genes)				
FYPO:0000124	viable cell	90.4 (123/136)	52.1 (3651/7005)	1.53E-18
FYPO:0001491	viable vegetative cell	90.4 (123/136)	52.1 (3651/7005)	1.53E-18
FYPO:0000003	cell population phenotype	100 (136/136)	70.6 (4949/7005)	4.26E-18
FYPO:0002057	cell population viability	100 (136/136)	70.5 (4938/7005)	4.26E-18
FYPO:0000001	phenotype	100 (136/136)	70.8 (4959/7005)	6.19E-18
FYPO:0002177	viable vegetative cell with normal cell morphology	83.1 (113/136)	44.3 (3100/7005)	7.13E-18
FYPO:0002058	viable cell population	91.9 (125/136)	56.1 (3928/7005)	1.46E-17
FYPO:0002060	**viable vegetative cell population**	**91.2 (124/136)**	**54.9 (3844/7005)**	**1.46E-17**
FYPO:0000004	cell viability	98.5 (134/136)	69.2 (4845/7005)	2.54E-16
FYPO:0000136	cellular physical quality phenotype	98.5 (134/136)	69.7 (4881/7005)	4.56E-16

Table 5.5. AnGeLi (Analysis of Gene Lists) Results. 190 genes, with a Spearman correlation p-value < 0.05 between count / mean count ratio and time, were filtered, with 54 having a positive correlation (Supplementary List B) and 136 having a negative correlation (Supplementary List C). Included in this table are the top 10 results of both enrichment analyses, sorted in ascending order according to the Benjamini-Hochberg corrected p-values. FYPO stands for Fission Yeast Phenotype Ontology. Highlighted in bold, the FYPOs describing the viability of a vegetative cell population upon deletion of a gene; inviable for 53.7% of genes with a positive correlation and viable for 91.2% of genes with a negative correlation.

5.3 Application of the HMM on the Ageing Dataset

In concurrence, we also applied the five-state Hidden Markov Model (HMM) on the ageing dataset (Figures 5.9 and 5.10).

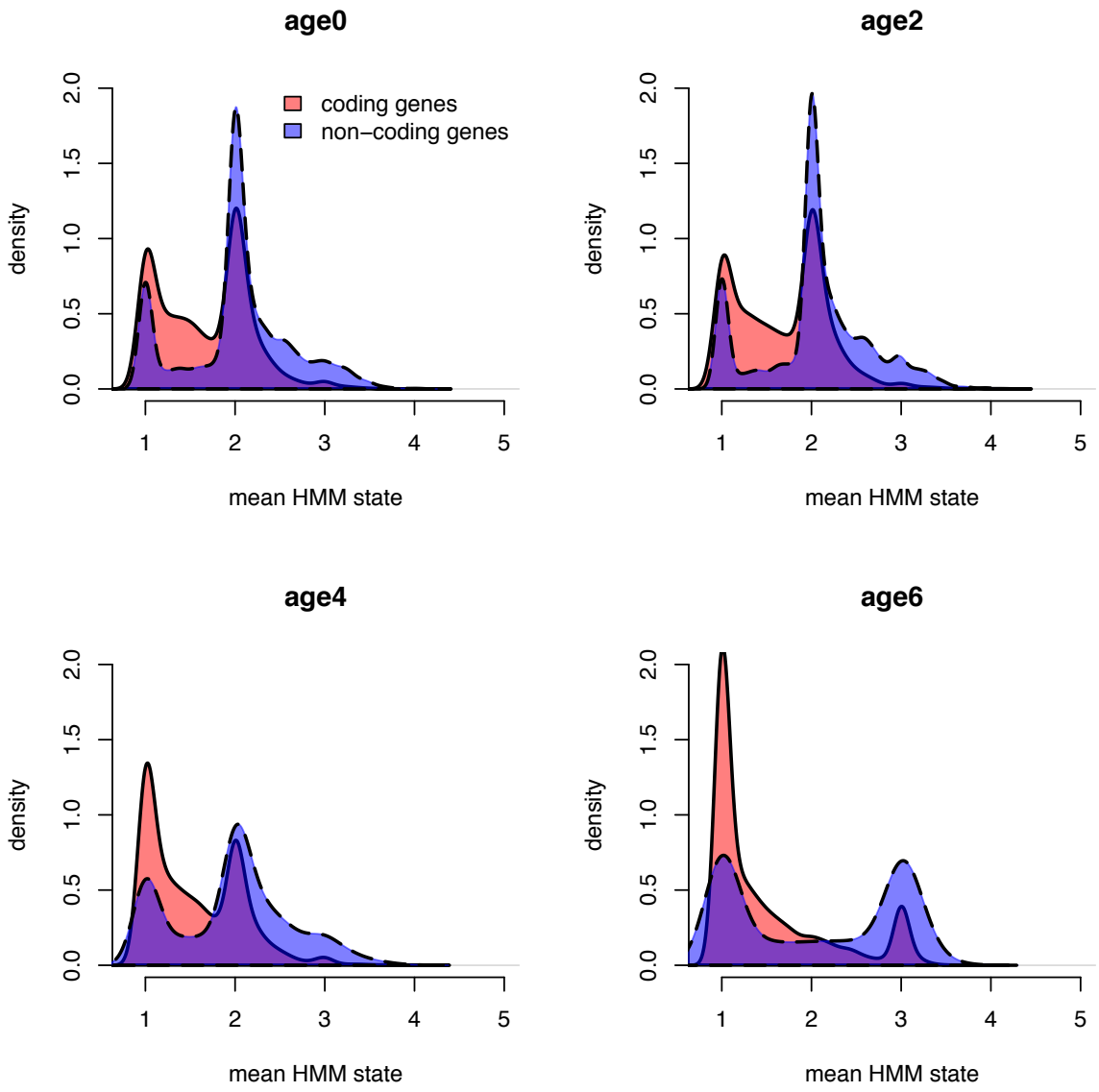


Figure 5.9. The Ageing Landscape. Overall, we demonstrate that progressing from age 0 to age 6 results in an increase of the number of essential protein-coding transcripts (mean HMM state = 1). In regard to the non-protein-coding transcripts, we observe that while most are not essential, a few are still important (see Figure 5.10).

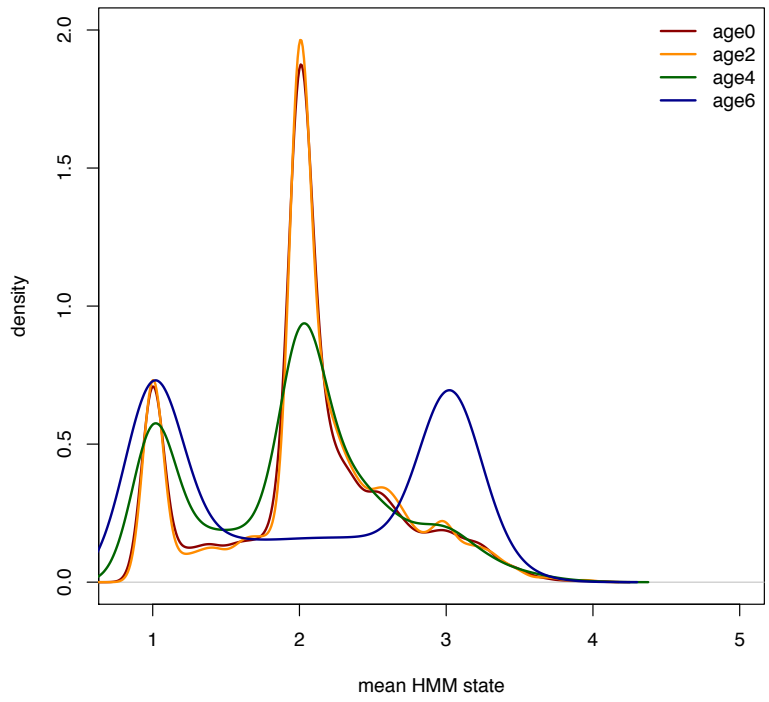


Figure 5.10. The Ageing Landscape of the Non-Protein-Coding Transcripts. Here, we show that most non-coding transcripts are not essential during ageing (mean HMM state = 2 / 3), however, a few are still required for cell survival (mean HMM state = 1). In addition, we observe a bipolar distribution at age 6. Overall, we find 218 ncRNAs with mean HMM state < 1.5, coverage > 50%, and overlapping all four time points (Supplementary List D).

5.4 Summary of the Main Results

Hermes insertion libraries were aged via chronological lifespan (CLS) assays as described (Roux *et al.* 2009). For these libraries, cells were collected and processed when the cultures reached stationary phase (age 0) and 2, 4, and 6 days later.

So as to examine whether genes changed their functional importance during ageing, a gene-based analysis was carried out. Here, the hypothesis was that with time, the number of unique insertion counts changes in a gene but remains constant in the region surrounding it. To this end, Cochran–Mantel–Haenszel tests were independently carried out on the mean (for the observed values) and

the median (for the expected values) unique insertion counts of 1, 2 and 3 kb gene windows, with each gene window centred around the gene midpoint. Over 5000 genomic features were screened. P-values were corrected for multiple testing using the Benjamini–Hochberg method, with a false discovery rate of < 0.05 as the significance threshold for the adjusted p-values.

1200 genes deviated significantly from the expectation after p-value adjustment. For these genes, a Spearman rank correlation coefficient test was performed between the four time points and a ratio of insertion counts and mean insertion counts. 190 out of 1200 genes had a correlation p-value < 0.05, with 54 having a positive correlation and 136 having a negative one. Both gene lists were analysed with AnGeLi (Bitton *et al.* 2015). Here, positively correlated genes, whose relative insertion densities increased with ageing, were enriched for the essential genes phenotype ontology (FYPO:0002061, genes whose deletion results in inviable vegetative cell populations). In contrast, negatively correlated genes, whose relative insertion densities decreased with time, were enriched for the non-essential genes phenotype ontology (FYPO:0002060, genes whose deletion results in viable vegetative cell populations).

In this work, we also applied the five-state Hidden Markov Model on the ageing dataset. Overall, we demonstrated that progressing from age 0 to age 6 results in an increase of the number of state 1 protein-coding transcripts. In addition, we show that most non-protein-coding transcripts are not essential during ageing, however, a few are still important for survival.

Chapter 6 DISCUSSION

6.1 Another Piece of the Puzzle

In this large-scale experiment, we have addressed one of the central aims of biological research: to fully describe the coordinated collection of functions and processes that combine to create living organisms, such as fission yeast. Gradually, a more comprehensive picture of *S. pombe* is coming to light, as novel datasets from both small- and large-scale investigations are merged to refine the current annotation of gene structures and assign function to them. Our *Hermes* insertion dataset is another piece of the puzzle.

Specifically, our goal was to identify functional elements in the *S. pombe* genome during growth and chronological ageing. To this end, we combined high-density transposon mutagenesis with high throughput DNA sequencing. First, we constructed near saturation transposon libraries, which can, in principle, harbour mutations disrupting all non-essential loci in the genome. NGS was then used to map *Hermes* insertion sites *en masse*. In light of the relevant literature and our expectations for the data, we next developed a five-state Hidden Markov Model (HMM) to quantitatively discriminate between loci that are dispensable or required during the two conditions of interest.

In a previous attempt to identify the function of *S. pombe* genes, Guo *et al.* (2013) showed a lack of *Hermes* insertion sites in essential genes and an excess in non-genic regions. In general, this is consistent with the fact that insertion densities are good indicators of the relative functional importance of particular annotation types. So as to formally establish whether our *Hermes* insertion data reflects functional constraint, we compared *Hermes* insertions to both the divergence trends between related *Schizosaccharomyces* species, and the diversity within the species. Overall, there were two messages to take home: first, *Hermes*

insertions are indicative of functional regions, and second, our data is able to locate functional elements with high precision.

In a consistent manner, the relative levels of genetic constraint and diversity between species showed that the coding regions of essential genes (eCDS) were subject to higher constraint than coding regions of non-essential genes (nCDS). In turn, this was followed by 5'/3' untranslated regions (UTRs) and introns, with unannotated regions (generally intergenic regions) being the least constrained. In addition, both *Hermes* insertion densities (unique insertion positions/100 nt) and mean insertion counts were consistent with this ranking, showing that *Hermes* insertions are suitable indicators of functional regions, for both protein-coding and non-coding transcripts.

In general, evolutionary studies are valuable for measuring the fitness effects of mutational changes in genomes (that is, function), however, it is also important to take into account their shortcomings. If sufficient related genomes are available, for instance, a selective constraint analysis can locate regions that are conserved (Lindblad-Toh *et al.* 2011). However, such studies will fail to uncover functional genomic elements that are not conserved over long periods of evolutionary time, because these will not be retained. In addition, there is a paradox here, in that an increase in phylogenetic depth is required to detect smaller or weaker areas of constraint, but that same increase in phylogenetic depth means that species are more diverged, and so have more unshared functional elements (Cooper and Brown 2008). Likewise, patterns of diversity are able to predict generic aspects of the genome with respect to function (Fawcett *et al.* 2014, Jeffares *et al.* 2015), but do not have sufficient resolution to pinpoint specific functional elements. Overall, this is because polymorphic sites are present at low densities. In addition, the relationship of polymorphic sites to constraint is affected by both the recombination rate (Campos *et al.* 2014) and recent events of selection, which can purge diversity in surrounding areas (Cheeseman *et al.* 2012). In contrast to

these evolutionary studies (and their pitfalls), near saturation insertion libraries, generated from independent repeats, are able to define functional operons in bacterial genomes in detail (Zhang *et al.* 2012, DeJesus and Ioerger 2013). Similarly, our *Hermes* insertion data was able to locate functional elements in the *S. pombe* genome with high accuracy.

6.2 The Importance of the HMM

HMMs are at the core of numerous biological applications, including gene finding, multiple sequence alignment, profile searches, and regulatory site identification. Indeed, HMMs are often considered the Lego's of computational sequence analysis (Eddy 2004). In general, HMMs are more advantageous than the other methods used to examine Tn-seq data (Zhang *et al.* 2012, Zomer *et al.* 2012, DeJesus *et al.* 2013). One of the advantages of using an HMM is that it is not limited to annotated gene boundaries, and can therefore pinpoint independent non-coding RNAs, protein domains, and regulatory regions, that are essential under the condition of interest. HMMs also allow for the possibility that only a segment of a gene might be required. In addition, to veer from the basic essential or non-essential classification, more states can be introduced to capture distinct genomic regions. In this work, in fact, a fifth state was added to capture sites with the highest 10% of unique insertions per 100 nt.

Using HMMs, it is also possible to account for insertion biases. In this work, the two biases were the nTnnnnAn motif and the nucleosome occupancies. In order to understand these biases, we must first understand the complex structure that the *Hermes* transposase forms with the DNA. In this regard, it is important to consider how *Hermes* works at the molecular level. It is known that the *Hermes* approach described in Everts *et al.* (2007) is bipartite, in that it contains the *Hermes* transposase, which is driven via a repressible promoter (see Figure 2.1, expression plasmid), and the transposon, which is composed of a drug resistance marker flanked with terminal inverted repeats (TIRs) of *Hermes* (see Figure 2.1,

donor plasmid). Upon promoter induction, and subsequent transposase expression, the transposon is excised at its TIRs and then integrated into chromosomal DNA. In the meantime, we are left to ponder how the transposase is able to locate its transposon ends amidst a sea of chromosomal DNA. Hickman *et al.* (2014) provided an answer, revealing that *Hermes* forms an octameric ring organised as a tetramer of dimers. In their research, Hickman *et al.* (2014) discovered that, with respect to the chemical steps of transposition, there is a difference between isolated dimers and octamers, with the former being active *in vitro* and the latter being active *in vivo*. It was also shown that the octamer provides multiple specific and non-specific DNA binding domains, and it is these multiple sites of interaction that allow the transposase to locate its transposon.

So far, research supports the octamer as being the active species. Its size, 560 kDa, is several orders larger than histones and so this could help to understand the preference for *Hermes* to integrate into nucleosome-free regions (NFRs) (Gangadharan *et al.* 2010, Guo *et al.* 2013). In addition to the fact that NFRs are, at large, more accessible to DNA binding proteins, it would be difficult for such a large protein complex structure to interact with DNA that is wrapped around histones. Moreover, stretches of T and A are higher in NFRs, which could also account for the preference of *Hermes* to integrate into nucleosome-depleted regions. NFRs include transcriptional regulatory regions, such as enhancers, promoters, and terminators, which have low nucleosome occupancies and often contain nucleosome-depleted regions. In contrast, protein-coding regions have high nucleosome occupancies (Yadon *et al.* 2010). In general, all of this was taken into consideration when developing the HMM.

In this experiment, the applied five-state HMM helped to distinguish segments at the subgene level that have differential effects on cell survival. From the HMM results, we concluded that *Hermes* insertions have functional consequences in 90% of this compact genome, including 80% of the non-protein-coding regions.

Specifically, we show that 99% of essential coding sequences (eCDS) were assigned to either S1 or S2. eCDS are the most highly conserved regions and contain the lowest within-species diversity.

6.3 Ageing: Understanding How and Why

Comedian George Carlin once said that life should be lived in reverse; die first to get it out of the way, work for forty years until you are young enough to enjoy your retirement, go to school, play, become a little baby, then spend your last nine months floating in the womb. What if, however, instead of living life backwards, we could just slow down ageing and live longer, healthier lives? What if ageing is just a disease that we could cure?

From caloric restriction to telomerase, this is an exciting new era in the field of ageing research, particularly, the biology of ageing in mammals. In fact, several screens have been carried out in model organisms such as yeast, worms, and flies, leading to the identification of genes involved in ageing. Some of these genes were also confirmed in mammals (Kennedy 2008). It is possible to extrapolate from one eukaryotic species to another because a lot of functions and pathways known to control longevity are conserved within the eukaryotic kingdom (Fontana *et al.* 2010, Kaeberlein 2010). Moreover, an overlap exists between the genetic determinants of longevity across a range of eukaryotic species (Kaeberlein and Kennedy 2005).

In this screen, using *S. pombe* as a eukaryotic model for ageing, we aged *Hermes* insertion libraries via chronological lifespan assays. CLS assays measure the length of time non-dividing cells survive. In this case, the cell cultures were first grown to stationary phase, and then over time, samples were removed to assess the survival of the population. Overall, the strength of the system lies in the ease of manipulating genes, and the evolutionarily conserved nature of the longevity genes. To this end, we performed a gene-based analysis to examine whether (and

if so which) genes changed their functional importance over time. Using such an unbiased approach, our expectation was to discover genes with a link to ageing, and therefore provide a better assessment of the total number of longevity genes.

In *S. pombe*, several pro- and anti-longevity genes have been characterised. Sideri *et al.* (2014), for example, discovered long-lived mutants in quiescent cells deprived of nitrogen. Rallis *et al.* (2014), on the other hand, revealed genes involved during stationary phase under glucose starvation. Here, the screen was carried out for mutants resistant to TORC1 inhibition. TORC1, target of rapamycin complex 1, promotes ageing in multiple organisms. Combined, these two studies generated a list of 116 anti-longevity (or pro-ageing) genes, which increase lifespan when knocked out. In relation to our ageing insertion data, the initial expectation was for these genes to be good positive controls, that is, to exhibit an increase in the number of insertions as a function of age. However, for most (88) of the 116 genes, the opposite was true. Indeed, plotting the count / mean count ratio at each time point for both repeats showed there to be a negative correlation. It is possible that this is due to a difference in experimental conditions; it is known that some mutants are long-lived in one condition but short-lived in another. For example, the *gsk3* mutant, whose gene encodes glycogen-synthase kinase 3, is long-lived during nitrogen starvation (Sideri *et al.* 2014). In contrast, the *sck2* mutant (S6 protein kinase) is long-lived during stationary phase (Roux *et al.* 2006) but not during nitrogen starvation (Sideri *et al.* 2014).

In order to ensure that the information stored within the ageing insertion dataset is mined to its fullest, we pursued a different but similar route. Specifically, we performed relevant statistical tests to uncover significant correlations ($p < 0.05$) between the count / mean count ratio and each of the four time points. Here, the robustness of the data was validated. Overall, we unveiled two lists, one containing 54 positively correlated genes, and another one encompassing 136

negatively correlated genes. In this context, the expectation was for the former list to denote genes that are essential during log phase but not throughout ageing, and for the latter list to represent genes that are required for longevity but not for growth. In the case of most genes, the AnGeLi tool suggested this to be true, in that deleting the positively correlated genes gives rise to inviable cell populations, whereas knocking out the negatively correlated genes, on the other hand, results in viable cell populations. In PomBase, according to the FYPO annotation data, an inviable cell population designates essential genes, while a viable cell population identifies non-essential genes. Overall, this fits well with our results (Figure 6.1), given that in PomBase, null mutants are annotated as inviable or viable during vegetative growth, also referred to as log phase.

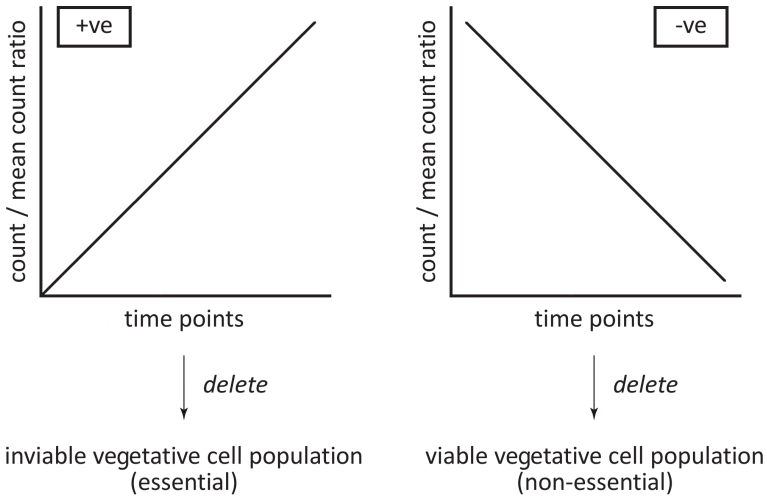


Figure 6.1. Mining the Ageing Insertion Dataset. Overall, we found 54 positively correlated, pro-ageing genes and 136 negatively correlated, anti-ageing genes that are in accordance with the FYPO annotation data in PomBase.

It is speculative, but this could be attributable to antagonistic pleiotropy; a theory of ageing proposed by Williams (1957) as an evolutionary explanation for senescence. In general, Williams states that an allele with a negative impact on performance in late life could be selected for if it has a positive influence on reproduction in early life. In genetics, antagonistic pleiotropy is frequently used to describe a gene that controls multiple traits, where at least one trait is beneficial

to the organism's fitness and at least one is detrimental. It is often an evolutionary trade-off between early- and late-life performance. In budding yeast, for instance, Qian *et al.* (2012) measured the fitness of over 5000 non-essential genes, and found that hundreds of genes harm rather than benefit the organism, citing widespread antagonistic pleiotropy. In a more specific example, Li *et al.* (2014) investigated the *de novo*-originated gene, *MDF1*, which promotes growth but suppresses mating. In fission yeast, Rallis *et al.* (2014) found that slow-growing mutants live longer, whereas fast-growing mutants live shorter. In this context, the trade-off is between growth and ageing. It is therefore plausible that what we observe in our *Hermes* insertion data are pleiotropic genes with a positive effect during growth and a negative role during ageing, and vice versa.

In light of the five-state HMM, we also demonstrated that progressing from age 0 to age 6 results in an increase of the number of state 1 protein-coding genes. In a comprehensive exploration of the two later time points, we observe that over 60% of these protein-coding transcripts are enriched for the cellular metabolic process GO term (Bitton *et al.* 2015). In general, this could be due to the fact that ageing is a consequence of rapid metabolic adaptation essential for survival in changing environmental situations. In addition, we show that most non-protein-coding genes are not essential during ageing. However, a few (14%) are still important across all four time points.

6.4 Non-Coding RNAs: Functional or Junk?

Traditionally, coding sequences have received much of the attention of geneticists, however, the emerging abundance of ncRNAs, which can outnumber coding transcripts, has raised important questions concerning their general significance. For example, how much genetic information is transacted by ncRNAs? Does this 'dark matter' form a network of regulatory information to control gene expression, or is it merely opportunistic transcriptional noise?

In recent years, the ncRNA annotation of the *S. pombe* genome has escalated. In addition to the 307 transfer RNAs (tRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), and ribosomal RNAs (rRNAs), there are now other annotated ncRNAs, for the most part, from RNA-seq data (Watanabe *et al.* 2002, Wilhelm *et al.* 2008, Rhind *et al.* 2011, Eser *et al.* 2016, Atkinson *et al.* 2017). Intriguingly, a large number of ncRNAs seem to be antisense to protein-coding genes, with some displaying meiosis-specific expression (Bitton *et al.* 2011), and others having a regulatory function during sexual differentiation (Watanabe *et al.* 2002). Of the remainder, few had been characterised, up until the recent work of Atkinson *et al.* (2017) which uncovered 5775 novel long non-coding RNAs (lncRNAs); lncRNAs contribute a substantial portion to non-coding transcriptomes. Overall, this hints at a potential for expansion of ncRNA research in *S. pombe*.

In this regard, and due to the nature of both datasets, our aim was to hunt down the essential ncRNAs, and challenge the conception that non-protein-coding genes are junk. Overall, our data illustrates that 40% of non-protein-coding regions have low transposon insertion densities and therefore seem to be functional; a result that has not been documented before. On further examination, we also show that for the log phase dataset, 85 non-protein-coding genes have a mean state < 1.5, coverage > 50%, and not overlapping protein-coding transcripts. In brief, we find 85 ncRNAs that are considered essential according to the HMM. It is important to note that for our five-state HMM, classification for state 1 was trained on coding sequences of essential genes.

6.4.1 The Role of ncRNAs in Ageing

Our next aim was to understand whether ncRNAs in particular affect the ageing process in *S. pombe*. It is well acknowledged that the expression of ncRNAs in eukaryotic genomes changes with age. It is also known that ncRNAs evolved in eukaryotes as epigenetic regulators of gene expression, with small microRNAs

(miRNAs, 20–24 nt) and long non-coding RNAs (lncRNAs, > 200 nt) being the most abundant regulatory ncRNAs (Sierra *et al.* 2015). Intriguingly, both of these have been implicated in ageing.

In 1993, from *C. elegans* studies, came the first evidence of an involvement of miRNA in ageing and lifespan. MiRNAs are highly conserved non-coding RNAs that negatively regulate gene expression by directly targeting cellular mRNAs. In the worm *C. elegans*, the miRNA *lin-4* was shown to target the *lin-14* mRNA, and also the insulin/IGF signalling cascade (Lee *et al.* 1993). In turn, this pathway is a molecular downstream target of regulatory molecules that influence ageing (Grillari and Grillari-Voglauer 2010, Jung and Suh 2012). In addition to *lin-4*, other miRNAs, such as miR-1 and miR-145, have been reported to target the insulin/IGF-I receptor and related signalling molecules (Jung and Suh 2014). Together, these and others have provided support for the role of miRNAs in regulating the lifespan of most model organisms. In relation to how these small inhibitory RNAs contribute to human ageing, a number of miRNAs expression profiles (e.g. Noren Hooten *et al.* (2010), ElSharawy *et al.* (2012)) have demonstrated that changes in miRNA expression also occur with human ageing. Overall, this suggests that miRNAs and their targets have the potential to be used as age indicators. However, notwithstanding this wealth of evidence of miRNA involvement in the ageing of other organisms, to this date no miRNAs have been found in *S. pombe*. Therefore, with this avenue closed, we must look at the larger counterpart of miRNAs, lncRNAs, whose relevance in ageing has just become apparent.

It has been demonstrated that lncRNAs affect six major molecular traits of ageing (Grammatikakis *et al.* 2014). Specifically, lncRNAs have been implicated in –

(a) Cell Proliferation and Senescence: It is known that as organisms age, senescent cells accumulate. One of the hallmarks of cellular senescence is a failure to progress through the cell cycle. In fact, senescent cells arrest growth with a

DNA content characteristic of G1 phase (Campisi and d'Adda di Fagagna 2007). MALAT1, a conserved lncRNA amongst mammals, for example, has been shown to repress senescence. Indeed, depleting MALAT1 in breast (Zhao *et al.* 2014) and cervical (Guo *et al.* 2010) cancer cells induced G1 arrest and reduced both cell growth and cell proliferation.

(b) Communication Among Cells: It is also known that ageing involves alterations at the level of intercellular communication; one of the changes is inflammaging (an inflammatory state that accompanies ageing in mammals), which can result from senescent cells secreting cytokines (López-Otín *et al.* 2013). Intriguingly, lncRNAs associated with inflammation have also been found. Rapicavoli *et al.* (2013), for instance, discovered 54 mouse lncRNAs induced by inflammatory signalling via TNFα, an inflammatory cytokine.

(c) Controlling Telomere Length: In gerontology, cellular senescence has been associated with telomere shortening. Similar to the plastic tips at the ends of shoelaces, telomeres are structures at the ends of chromosomes that protect chromosomal DNA from damage. In the ciliate *Tetrahymena*, Nobel laureates Carol Greider and Elizabeth Blackburn discovered that telomerase counteracts telomere shortening (Greider and Blackburn 1989). LncRNA TERC, one of the components of the telomerase complex, has been implicated in the maintenance of telomere length, and therefore, the prevention of senescence. Indeed, mice deficient in TERC exhibited short telomeres, chromosomal damage, and premature ageing (Samper *et al.* 2001). LncRNA TERRA, on the other hand, was shown to suppress telomere elongation (Redon *et al.* 2010).

(d) Epigenetic Gene Expression: It has been demonstrated that epigenetic changes modulate gene expression during cellular senescence and ageing. Overall, there is a decline in DNA methylation (Johnson *et al.* 2012), disruption of heterochromatin (Tsurumi and Li 2012), and histone modifications. It was shown, for instance, that disruption of histone modifications influences lifespan in model organisms such as flies (Siebold *et al.* 2010) and worms (Greer *et al.* 2010). Several lncRNAs are involved in these epigenetic changes; some are implicated in

histone modifications (e.g. ANRIL), while most others contribute to the regulation of DNA methylation (e.g. *H19* and *Xist*) (Grammatikakis *et al.* 2014).

(e) Proteostasis: It has been shown that the disruption of proteostasis, or protein homeostasis, leads to age-associated diseases such as Alzheimer's, Huntington's, and Parkinson's diseases (Balch *et al.* 2008). In general, proteostasis includes numerous biological processes; lncRNAs have been linked to some of these processes, including autophagy (e.g. 7SL, HULC, and MEG3), protein synthesis and degradation (e.g. AS Uchl1, HOTAIR, and lncRNA-p21), and protein trafficking (e.g. Gadd7, GAS5, and PANDA) (Grammatikakis *et al.* 2014).

(f) Stem Cell Function: It is known that as stem cells age, their renewal capacity deteriorates, and their potential to differentiate into different cell types is altered (Ahmed *et al.* 2017). So far, a few lncRNAs that affect stem cell homeostasis have surfaced. Indeed, some lncRNAs (e.g. linc-RoR (Loewer *et al.* 2010)) were found to regulate stem cell transcription factors, which in turn regulate the expression of these lncRNAs.

In 2002, when the genome sequence of *S. pombe* was elucidated, Wood *et al.* (2002) suggested that *S. pombe* makes for a good model organism to understand human disease gene function. In this regard, just like ncRNAs have been implicated in human ageing, an extrapolation is that age-associated ncRNAs could also be present in *S. pombe*. However, even though genome-wide ageing studies have been carried out, nothing is known about the relationship between ncRNAs and ageing in this organism. To this end, and due to the nature of the dataset, our aim was to hunt down ncRNAs that appear to be essential during ageing. In this work, we demonstrate that most non-protein-coding transcripts are not required during ageing (mean HMM state = 2 / 3), however, a few are still important (mean HMM state = 1). In total, we find 218 age-related ncRNAs that are considered essential according to the HMM.

6.5 Future Work

6.5.1 HMM State Blocks

It is evident, from the results that have been generated thus far, that the *Hermes* insertion data is a powerful tool. Indeed, there is a vast potential for future research. In the current pipeline, for instance, we have state blocks; runs of the genome that are the same HMM state, which will further elucidate the functional elements in the *S. pombe* genome. To this end, we identified runs of genome sites that are all the same state within the log phase dataset. Since most state blocks are short (length = 1 nt), blocks with length < 10 nt were removed. In total, 15,451 blocks remained, which cover 89% of the genome. Figure 6.2 illustrates that all blocks have different distributions, such as the bimodal state 2 blocks. Moreover, Figure 6.3 shows that while there are almost equal numbers of state 1, 2, and 3 blocks, there are much more state 4 and 5 blocks.

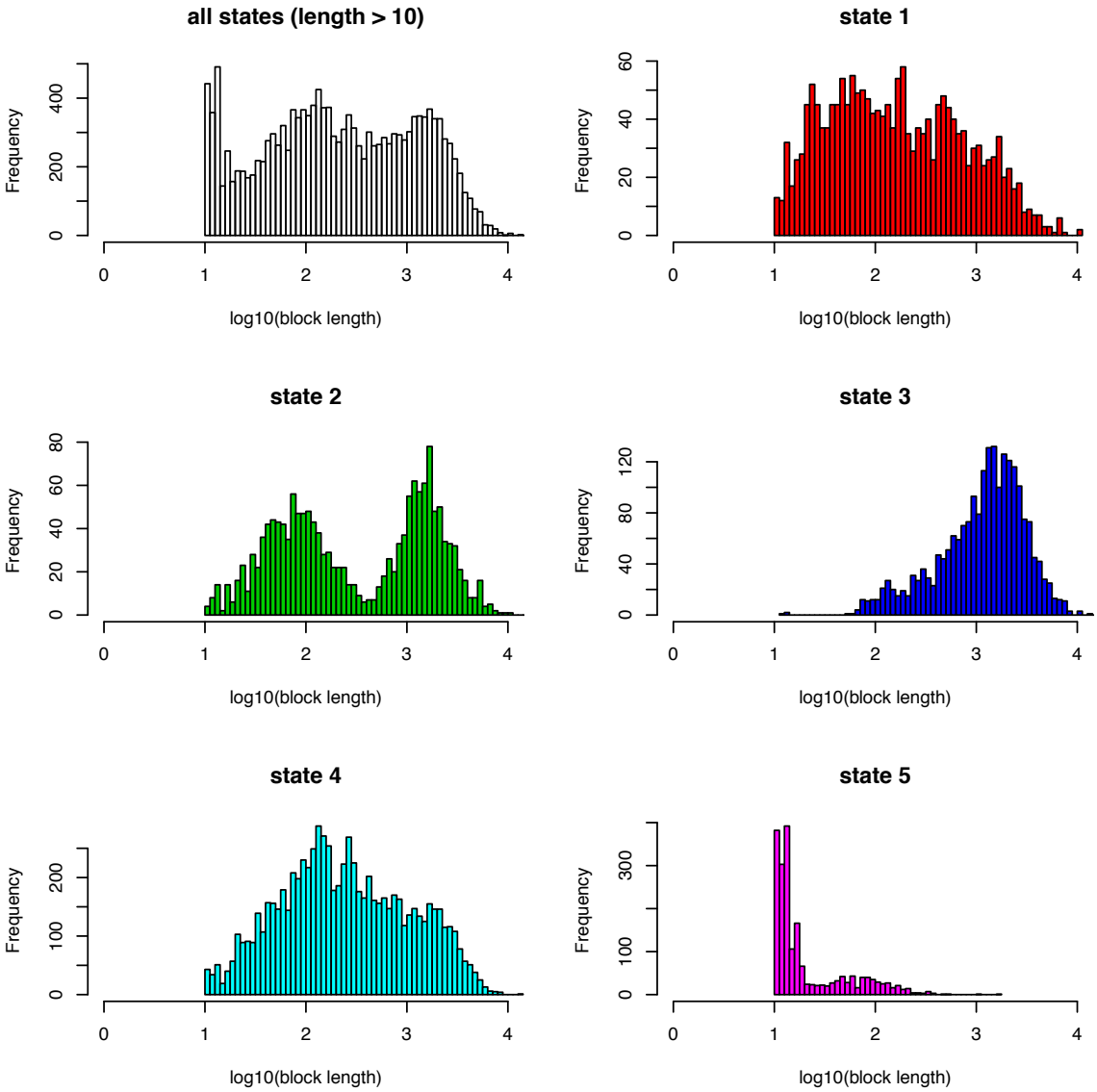


Figure 6.2. Distribution of HMM State Blocks. Illustrates that all state blocks have different distributions. It is intriguing that there are no large regions that are altogether unimportant.

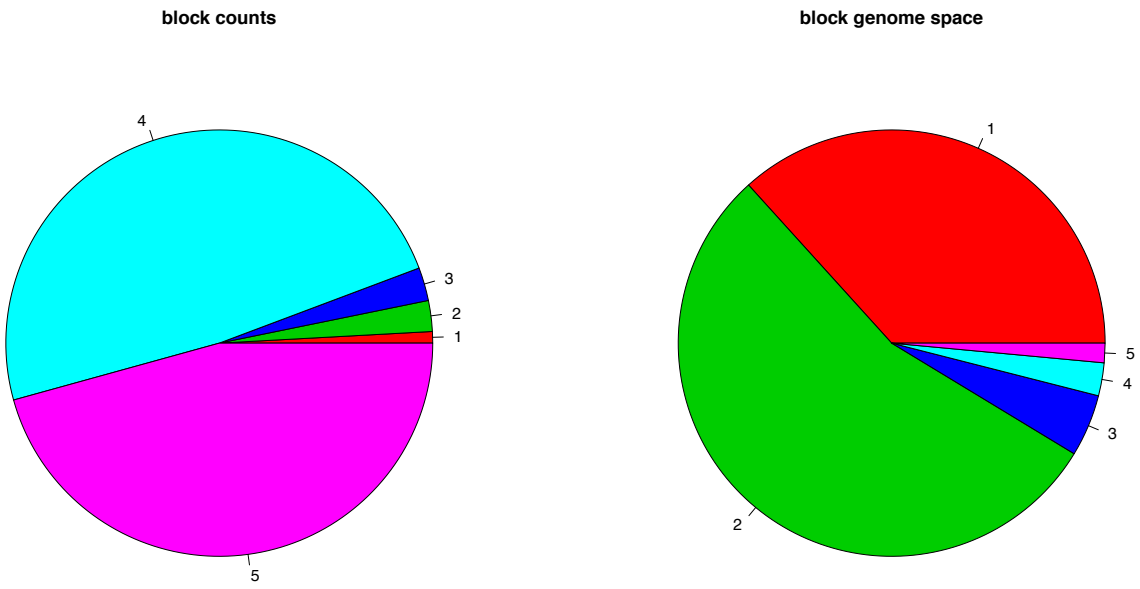


Figure 6.3. Count of HMM State Blocks. Illustrates that while there are almost equal numbers of state 1, 2, and 3 blocks, there are many more state 4 and 5 blocks.

Subsequent to this, our aim is to look at conservation and RNA expression levels within these state blocks, and answer questions such as: How close are state block edges to annotation edges? Could there be state 1 blocks that have low or no RNA-seq expression, and state 3 or 4 blocks that have high RNA-seq expression? So far, we used samples from Atkinson *et al.* (2017); two 100% stationary phase (that is, as soon as stationary phase was reached) and two 50% stationary phase, to compute the normalised coverage of RNA-seq data at the nucleotide level, that is, read coverage per million reads normalised by sample depth. In this regard, the next goal is to interrogate whether an HMM-defined region is transcribed or not.

6.5.2 CRISPR/Cas9 Verification

Rodríguez-López *et al.* (2016) established an efficient and fast PCR-based method for the deletion of DNA sequences in the *S. pombe* genome, through modifications and optimisations of the CRISPR/Cas9 gene editing system. CRISPR/Cas9 is the biggest game changer to hit genetics since PCR. In general, it relies on the Cas9 nuclease complexed with a guide RNA (gRNA) to target

DNA, which then edits the DNA to disrupt genes or insert desired sequences. In their open access research, Rodríguez-López *et al.* (2016) also provide a web tool that users can exploit to design the different primers required for deletion of genomic regions. So far, using this optimised CRISPR/Cas9 technology, over 80 non-coding RNA genes have been deleted in *S. pombe*, most of which have low expression levels. So as to understand the function of these ncRNAs, the authors are also phenotyping the mutants (personal communication; unpublished). In particular, this will include screening under various conditions (including, amongst others, DNA damage, heavy metal and oxidative stress), using the RoToR HDA robot from Singer Instruments. In addition, growth in different carbon and nitrogen sources will also be checked.

Overall, the *Hermes* insertion data generated in this work has augmented our knowledge of functional elements in the *S. pombe* genome. However, verification of function for transcripts assigned as essential by the HMM is required. CRISPR/Cas9 could be used to this end.

6.5.3 Time Points and Experimental Conditions

In the future, an aspect of the data that could also be addressed is the ageing time points. In total, seven time points were collected during the construction of the libraries; however, four (t0, t2, t4, t6) were processed and studied. Overall, in contrast to the log phase libraries, the ageing libraries had less unique insertion sites per nucleotide of the genome. So as to increase this number, and with it the power of the ageing dataset, the remaining time points could be processed as well.

So far, the two experimental conditions tested were ageing and vegetative growth. *Hermes* transposon mutagenesis, when combined with NGS, is a flexible tool that can be adapted to test other conditions. Indeed, it is also suitable to screen for CLS in quiescent cells deprived of nitrogen. In addition, it can be used to explore the functional landscape of the *S. pombe* genome under heat or oxidative stress.

For the heat shock experiment, one approach would be to incubate mid-log phase cultures of the insertion mutants at 50°C for up to one hour (conditions estimated from Roux *et al.* (2006)). For the oxidative stress test, cultures can first be diluted with YES and then exposed to hydrogen peroxide at a final concentration of 0.2 mM, 1.0 mM or 2.0 mM for up to one hour (conditions estimated from Pekmez *et al.* (2008)). So as to measure cell survival after treatment, drop tests can then be carried out on solid medium. DNA libraries can be constructed as per Chapter 2.3.

6.6 Conclusion

Over the past six decades, *S. pombe* has emerged as one of the two main yeast model organisms. Indeed, publications containing gene-specific data for *S. pombe* have flourished since the Swiss geneticist Urs Leupold made its acquaintance during the 1940s. Regardless, a lot remains in the dark, with secret messages dispersed around the three chromosomes. In light of this, the aim of the PhD was to explore the functional landscape of the *S. pombe* genome. To this end, we coupled *Hermes* transposon mutagenesis with next generation sequencing, a combination that proved to be flexible, powerful, and robust. Specifically, we created multiple dense insertion libraries, during log phase growth and chronological ageing, achieving a saturating (or near-saturating) insertion density of 1 insertion per 13 nucleotides of the genome for log phase samples. In general, this was a significant improvement over the previous attempt (Guo *et al.* 2013).

So as to account for the complexity and stochastic nature of the data, we applied a five-state Hidden Markov Model (HMM), where the HMM state provides a semi-quantitative estimate of the functional significance of the genome at single nucleotide-level resolution. HMM state values are consistent but more detailed than genome annotations. So far, amongst the numerous results, we have shown that transposon insertions have functional consequences in 90% of the *S. pombe* genome, including 80% of the non-protein-coding regions. Overall, this functional annotation map (published online) distinguishes sub gene-level genomic segments that have differential effects on cell survival, and is therefore a valuable resource for the research community. It also builds the foundation for other studies, which together will shed light on our understanding of the biology of *S. pombe*. In turn, such knowledge will provide novel insights into the biology of non-model organisms such as ourselves.

References

Ahmed, A. S. I., Sheng, M. H. C., Wasnik, S., Baylink, D. J. and Lau, K.-H. W. (2017). Effect of Aging on Stem Cells. *World Journal of Experimental Medicine* 7 (1): 1-10.

Allshire, R. C. and Karpen, G. H. (2008). Epigenetic Regulation of Centromeric Chromatin: Old Dogs, New Tricks? *Nature Reviews Genetics* 9 (12): 923-937.

Atkinson, S. R., Marguerat, S., Bitton, D. A., Rodríguez-López, M., Rallis, C., *et al.* (2017). Long Non-Coding RNA Repertoire and Regulation by Nuclear Exosome, Cytoplasmic Exonuclease and RNAi in Fission Yeast. bioRxiv: 158477.

Baker, M. (2010). Next-Generation Sequencing: Adjusting to Data Overload. *Nature Methods* 7 (7): 495-499.

Balch, W. E., Morimoto, R. I., Dillin, A. and Kelly, J. W. (2008). Adapting Proteostasis for Disease Intervention. *Science* 319 (5865): 916-919.

Behrens, R., Hayles, J. and Nurse, P. (2000). Fission Yeast Retrotransposon Tf1 Integration is Targeted to 5' Ends of Open Reading Frames. *Nucleic Acids Research* 28 (23): 4709-4716.

Bitton, D. A., Grallert, A., Scutt, P. J., Yates, T., Li, Y., *et al.* (2011). Programmed Fluctuations in Sense/Antisense Transcript Ratios Drive Sexual Differentiation in *S. pombe*. *Molecular Systems Biology* 7 (1): 559.

Bitton, D. A., Schubert, F., Dey, S., Okoniewski, M., Smith, G. C., Khadayate, S., Pancaldi, V., Wood, V. and Bähler, J. (2015). AnGeLi: A Tool for the Analysis of Gene Lists from Fission Yeast. *Frontiers in Genetics* 6: 330.

Bowen, N. J., Jordan, I. K., Epstein, J. A., Wood, V. and Levin, H. L. (2003). Retrotransposons and their Recognition of Pol II Promoters: A Comprehensive Survey of the Transposable Elements from the Complete Genome Sequence of *Schizosaccharomyces pombe*. *Genome Research* 13 (9): 1984-1997.

Campisi, J. and d'Adda di Fagagna, F. (2007). Cellular Senescence: When Bad Things Happen to Good Cells. *Nature Reviews Molecular Cell Biology* 8 (9): 729-740.

Campos, J. L., Halligan, D. L., Haddrill, P. R. and Charlesworth, B. (2014). The Relation between Recombination Rate and Patterns of Molecular Evolution and Variation in *Drosophila melanogaster*. *Molecular Biology and Evolution* 31 (4): 1010-1028.

Chao, M. C., Pritchard, J. R., Zhang, Y. J., Rubin, E. J., Livny, J., Davis, B. M. and Waldor, M. K. (2013). High-Resolution Definition of the *Vibrio cholerae*

Essential Gene Set with Hidden Markov Model-Based Analyses of Transposon-Insertion Sequencing Data. *Nucleic Acids Research* 41 (19): 9033-9048.

Cheeseman, I. H., Miller, B. A., Nair, S., Nkhoma, S., Tan, A., *et al.* (2012). A Major Genome Region underlying Artemisinin Resistance in Malaria. *Science* 336 (6077): 79-82.

Cherry, K. E., Hearn, W. E., Seshie, O. Y. and Singleton, T. L. (2014). Identification of Tf1 Integration Events in *S. pombe* under Nonselective Conditions. *Gene* 542 (2): 221-231.

Cooper, G. M. and Brown, C. D. (2008). Qualifying the Relationship between Sequence Conservation and Molecular Function. *Genome Research* 18 (2): 201-205.

Davis, M. P. A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. and Enright, A. J. (2013). Kraken: A Set of Tools for Quality Control and Analysis of High-Throughput Sequence Data. *Methods* 63 (1): 41-49.

DeJesus, M. A. and Ioerger, T. R. (2013). A Hidden Markov Model for Identifying Essential and Growth-Defect Regions in Bacterial Genomes from Transposon Insertion Sequencing Data. *BMC Bioinformatics* 14: 303.

DeJesus, M. A., Zhang, Y. J., Sassetti, C. M., Rubin, E. J., Sacchettini, J. C. and Ioerger, T. R. (2013). Bayesian Analysis of Gene Essentiality Based on Sequencing of Transposon Insertion Libraries. *Bioinformatics* 29 (6): 695-703.

Dos Santos, P. C., Fang, Z., Mason, S. W., Setubal, J. C. and Dixon, R. (2012). Distribution of Nitrogen Fixation and Nitrogenase-Like Sequences Amongst Microbial Genomes. *BMC Genomics* 13: 162.

Down, T. A., Piipari, M. and Hubbard, T. J. P. (2011). Dalliance: Interactive Genome Viewing on the Web. *Bioinformatics* 27 (6): 889-890.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., *et al.* (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* 489 (7414): 57-74.

Eddy, S. R. (2004). What is a Hidden Markov Model? *Nature Biotechnology* 22 (10): 1315-1316.

ElSharawy, A., Keller, A., Flachsbart, F., Wendschlag, A., Jacobs, G., *et al.* (2012). Genome-Wide miRNA Signatures of Human Longevity. *Aging Cell* 11 (4): 607-616.

Eser, P., Wachutka, L., Maier, K. C., Demel, C., Boroni, M., Iyer, S., Cramer, P. and Gagneur, J. (2016). Determinants of RNA Metabolism in the *Schizosaccharomyces pombe* Genome. *Molecular Systems Biology* 12 (2): 857.

Evertts, A. G., Plymire, C., Craig, N. L. and Levin, H. L. (2007). The *Hermes* Transposon of *Musca domestica* is an Efficient Tool for the Mutagenesis of *Schizosaccharomyces pombe*. *Genetics* 177 (4): 2519-2523.

Fabrizio, P., Gattazzo, C., Battistella, L., Wei, M., Cheng, C., McGrew, K. and Longo, V. D. (2005). Sir2 Blocks Extreme Life-Span Extension. *Cell* 123 (4): 655-667.

Fabrizio, P. and Longo, V. D. (2003). The Chronological Life Span of *Saccharomyces cerevisiae*. *Aging Cell* 2 (2): 73-81.

Fawcett, J. A., Iida, T., Takuno, S., Sugino, R. P., Kado, T., *et al.* (2014). Population Genomics of the Fission Yeast *Schizosaccharomyces pombe*. *PLOS ONE* 9 (8): e104241.

Flor-Parra, I., Zhurinsky, J., Bernal, M., Gallardo, P. and Daga, R. R. (2014). A Lallzyme MMX-Based Rapid Method for Fission Yeast Protoplast Preparation. *Yeast.* 31 (2): 61-66.

Fontana, L., Partridge, L. and Longo, V. D. (2010). Extending Healthy Life Span–from Yeast to Humans. *Science* 328 (5976): 321-326.

Forsburg, S. L. and Rhind, N. (2006). Basic Methods for Fission Yeast. *Yeast.* 23 (3): 173-183.

Gangadharan, S., Mularoni, L., Fain-Thornton, J., Wheelan, S. J. and Craig, N. L. (2010). DNA Transposon *Hermes* Inserts into DNA in Nucleosome-Free Regions *in vivo*. *PNAS* 107 (51): 21966-21972.

Germain, P.-L., Ratti, E. and Boem, F. (2014). Junk or Functional DNA? ENCODE and the Function Controversy. *Biology and Philosophy* 29 (6): 807-831.

Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., *et al.* (2004). Genome Sequence of the Brown Norway Rat Yields Insights into Mammalian Evolution. *Nature* 428 (6982): 493-521.

Goto, D. B. and Nakayama, J. (2012). RNA and Epigenetic Silencing: Insight from Fission Yeast. *Development, Growth and Differentiation* 54 (1): 129-141.

Goyal, A., Takaine, M., Simanis, V. and Nakano, K. (2011). Dividing the Spoils of Growth and the Cell Cycle: The Fission Yeast as a Model for the Study of Cytokinesis. *Cytoskeleton* 68 (2): 69-88.

Grammatikakis, I., Panda, A. C., Abdelmohsen, K. and Gorospe, M. (2014). Long Noncoding RNAs (lncRNAs) and the Molecular Hallmarks of Aging. *Aging* 6 (12): 992-1009.

Greer, E. L., Maures, T. J., Hauswirth, A. G., Green, E. M., Leeman, D. S., *et al.* (2010). Members of the Histone H3 Lysine 4 Trimethylation Complex Regulate Lifespan in a Germline-Dependent Manner in *C. elegans*. *Nature* 466 (7304): 383-387.

Greider, C. W. and Blackburn, E. H. (1989). A Telomeric Sequence in the RNA of *Tetrahymena* Telomerase Required for Telomere Repeat Synthesis. *Nature* 337 (6205): 331-337.

Grillari, J. and Grillari-Voglauer, R. (2010). Novel Modulators of Senescence, Aging, and Longevity: Small Non-Coding RNAs Enter the Stage. *Experimental Gerontology* 45 (4): 302-311.

Guimond, N., Bideshi, D. K., Pinkerton, A. C., Atkinson, P. W. and O'Brochta, D. A. (2003). Patterns of *Hermes* Transposition in *Drosophila melanogaster*. *Molecular Genetics and Genomics* 268 (6): 779-790.

Guo, F., Li, Y., Liu, Y., Wang, J., Li, Y. and Li, G. (2010). Inhibition of Metastasis-Associated Lung Adenocarcinoma Transcript 1 in CaSki Human Cervical Cancer Cells Suppresses Cell Proliferation and Invasion. *Acta Biochimica et Biophysica Sinica* 42 (3): 224-229.

Guo, Y. and Levin, H. L. (2010). High-Throughput Sequencing of Retrotransposon Integration Provides a Saturated Profile of Target Activity in *Schizosaccharomyces pombe*. *Genome Research* 20 (2): 239-248.

Guo, Y., Park, J. M., Cui, B., Humes, E., Gangadharan, S., *et al.* (2013). Integration Profiling of Gene Function with Dense Maps of Transposon Integration. *Genetics* 195 (2): 599-609.

Haines, R. L., Codlin, S. and Mole, S. E. (2009). The Fission Yeast Model for the Lysosomal Storage Disorder Batten Disease Predicts Disease Severity Caused by Mutations in *CLN3*. *Disease Models and Mechanisms* 2 (1-2): 84-92.

Han, T. X., Xu, X.-Y., Zhang, M.-J., Peng, X. and Du, L.-L. (2010). Global Fitness Profiling of Fission Yeast Deletion Strains by Barcode Sequencing. *Genome Biology* 11 (6): R60.

Harigaya, Y. and Yamamoto, M. (2007). Molecular Mechanisms underlying the Mitosis-Meiosis Decision. *Chromosome Research* 15 (5): 523-537.

Hazkani-Covo, E., Zeller, R. M. and Martin, W. (2010). Molecular Poltergeists: Mitochondrial DNA Copies (*numts*) in Sequenced Nuclear Genomes. *PLOS Genetics* 6 (2): e1000834.

Hickman, A. B., Ewis, H. E., Li, X., Knapp, J. A., Laver, T., *et al.* (2014). Structural Basis of *hAT* Transposon End Recognition by *Hermes*, an Octameric DNA Transposase from *Musca domestica*. *Cell* 158 (2): 353-367.

Ivics, Z. and Izsvák, Z. (2010). The Expanding Universe of Transposon Technologies for Gene and Cell Engineering. *Mobile DNA* 1 (1): 25.

Jeffares, D. C., Rallis, C., Rieux, A., Speed, D., Převorovský, M., *et al.* (2015). The Genomic and Phenotypic Diversity of *Schizosaccharomyces pombe*. *Nature Genetics* 47 (3): 235-241.

Johnson, A. A., Akman, K., Calimport, S. R. G., Wuttke, D., Stolzing, A. and de Magalhães, J. P. (2012). The Role of DNA Methylation in Aging, Rejuvenation, and Age-Related Disease. *Rejuvenation Research* 15 (5): 483-494.

Jung, H. J. and Suh, Y. (2012). MicroRNA in Aging: From Discovery to Biology. *Current Genomics* 13 (7): 548-557.

Jung, H. J. and Suh, Y. (2014). Regulation of IGF-1 Signaling by MicroRNAs. *Frontiers in Genetics* 5: 472.

Kaeberlein, M. (2010). Lessons on Longevity from Budding Yeast. *Nature* 464 (7288): 513-519.

Kaeberlein, M. and Kennedy, B. K. (2005). Large-Scale Identification in Yeast of Conserved Ageing Genes. *Mechanisms of Ageing and Development* 126 (1): 17-21.

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., *et al.* (2014). Defining Functional DNA Elements in the Human Genome. *PNAS* 111 (17): 6131-6138.

Kennedy, B. K. (2008). The Genetics of Ageing: Insight from Genome-Wide Approaches in Invertebrate Model Organisms. *Journal of Internal Medicine* 263 (2): 142-152.

Kim, D.-U., Hayles, J., Kim, D., Wood, V., Park, H.-O., *et al.* (2010). Analysis of a Genome-Wide Set of Gene Deletions in the Fission Yeast *Schizosaccharomyces pombe*. *Nature Biotechnology* 28 (6): 617-623.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. and Taipale, J. (2012). Counting Absolute Numbers of Molecules Using Unique Molecular Identifiers. *Nature Methods* 9 (1): 72-74.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., *et al.* (2001). Initial Sequencing and Analysis of the Human Genome. *Nature* 409 (6822): 860-921.

Lee, R. C., Feinbaum, R. L. and Ambros, V. (1993). The *C. elegans* Heterochronic Gene *lin-4* Encodes Small RNAs with Antisense Complementarity to *lin-14*. *Cell* 75 (5): 843-854.

Lenglez, S., Hermand, D. and Decottignies, A. (2010). Genome-Wide Mapping of Nuclear Mitochondrial DNA Sequences Links DNA Replication Origins to Chromosomal Double-Strand Break Formation in *Schizosaccharomyces pombe*. *Genome Research* 20 (9): 1250-1261.

Li, D., Yan, Z., Lu, L., Jiang, H. and Wang, W. (2014). Pleiotropy of the *De Novo*-Originated Gene *MDF1*. *Scientific Reports* 4: 7280.

Li, H. and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14): 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., *et al.* (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25 (16): 2078-2079.

Li, J., Zhang, J.-M., Li, X., Suo, F., Zhang, M.-J., Hou, W., Han, J. and Du, L.-L. (2011). A *PiggyBac* Transposon-Based Mutagenesis System for the Fission Yeast *Schizosaccharomyces pombe*. *Nucleic Acids Research* 39 (6): e40.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., *et al.* (2011). A High-Resolution Map of Human Evolutionary Constraint Using 29 Mammals. *Nature* 478 (7370): 476-482.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., *et al.* (2005). Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog. *Nature* 438 (7069): 803-819.

Loewer, S., Cabili, M. N., Guttman, M., Loh, Y.-H., Thomas, K., *et al.* (2010). Large Intergenic Non-Coding RNA-RoR Modulates Reprogramming of Human Induced Pluripotent Stem Cells. *Nature Genetics* 42 (12): 1113-1117.

Longo, V. D. and Fabrizio, P. (2002). Regulation of Longevity and Stress Resistance: A Molecular Strategy Conserved from Yeast to Humans? *Cellular and Molecular Life Sciences* 59 (6): 903-908.

Lopez, J. V., Yuhki, N., Masuda, R., Modi, W. and O'Brien, S. J. (1994). *Numt*, a Recent Transfer and Tandem Amplification of Mitochondrial DNA to the Nuclear Genome of the Domestic Cat. *Journal of Molecular Evolution* 39 (2): 174-190.

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. and Kroemer, G. (2013). The Hallmarks of Aging. *Cell* 153 (6): 1194-1217.

Malecki, M. and Bähler, J. (2016). Identifying Genes Required for Respiratory Growth of Fission Yeast. *Wellcome Open Research* 1: 12.

Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R. and Bähler, J. (2012). Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell* 151 (3): 671-683.

Mata, J. and Bähler, J. (2003). Correlations between Gene Expression and Gene Conservation in Fission Yeast. *Genome Research* 13 (12): 2686-2690.

Maundrell, K. (1990). *nmt1* of Fission Yeast: A Highly Transcribed Gene Completely Repressed by Thiamine. *Journal of Biological Chemistry* 265 (19): 10857-10864.

Michel, A. H., Hatakeyama, R., Kimmig, P., Arter, M., Peter, M., Matos, J., De Virgilio, C. and Kornmann, B. (2017). Functional Mapping of Yeast Genomes by Saturated Transposition. *eLife* 6: e23570.

Mourier, T., Hansen, A. J., Willerslev, E. and Arctander, P. (2001). The Human Genome Project Reveals a Continuous Transfer of Large Mitochondrial Fragments to the Nucleus. *Molecular Biology and Evolution* 18 (9): 1833-1837.

Noren Hooten, N., Abdelmohsen, K., Gorospe, M., Ejiogu, N., Zonderman, A. B. and Evans, M. K. (2010). MicroRNA Expression Patterns Reveal Differential Expression of Target Genes with Age. *PLOS ONE* 5 (5): e10724.

Nurse, P. (1990). Universal Control Mechanism Regulating Onset of M-Phase. *Nature* 344 (6266): 503-508.

Oh, J., Fung, E., Schlecht, U., Davis, R. W., Giaever, G., St. Onge, R. P., Deutschbauer, A. and Nislow, C. (2010). Gene Annotation and Drug Target Discovery in *Candida albicans* with a Tagged Transposon Mutant Collection. *PLOS Pathogens* 6 (10): e1001140.

Park, J. M., Evertts, A. G. and Levin, H. L. (2009). The *Hermes* Transposon of *Musca domestica* and its Use as a Mutagen of *Schizosaccharomyces pombe*. *Methods* 49 (3): 243-247.

Pekmez, M., Arda, N., Hamad, İ., Kiğ, C. and Temizkan, G. (2008). Hydrogen Peroxide-Induced Oxidative Damages in *Schizosaccharomyces pombe*. *Biologia* 63 (2): 151-155.

Peng, Q., Vijaya Satya, R., Lewis, M., Randad, P. and Wang, Y. (2015). Reducing Amplification Artifacts in High Multiplex Amplicon Sequencing by Using Molecular Barcodes. *BMC Genomics* 16: 589.

Petit, J., Boisseau, P. and Arveiler, B. (1994). Glucanex: A Cost-Effective Yeast Lytic Enzyme. *Trends in Genetics* 10 (1): 4-5.

Pettitt, S. J., Rehman, F. L., Bajrami, I., Brough, R., Wallberg, F., *et al.* (2013). A Genetic Screen Using the *PiggyBac* Transposon in Haploid Cells Identifies *Parp1* as a Mediator of Olaparib Toxicity. *PLOS ONE* 8 (4): e61520.

Phadnis, N., Hyppa, R. W. and Smith, G. R. (2011). New and Old Ways to Control Meiotic Recombination. *Trends in Genetics* 27 (10): 411-421.

Qian, W., Ma, D., Xiao, C., Wang, Z. and Zhang, J. (2012). The Genomic Landscape and Evolutionary Resolution of Antagonistic Pleiotropy in Yeast. *Cell Reports* 2 (5): 1399-1410.

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H. and Turner, D. J. (2008). A Large Genome Centre's Improvements to the Illumina Sequencing System. *Nature Methods* 5 (12): 1005-1010.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. and Gu, Y. (2012). A Tale of Three Next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers. *BMC Genomics* 13: 341.

Rallis, C., Codlin, S. and Bähler, J. (2013). TORC1 Signaling Inhibition by Rapamycin and Caffeine Affect Lifespan, Global Gene Expression, and Cell Proliferation of Fission Yeast. *Aging Cell* 12 (4): 563-573.

Rallis, C., López-Maury, L., Georgescu, T., Pancaldi, V. and Bähler, J. (2014). Systematic Screen for Mutants Resistant to TORC1 Inhibition in Fission Yeast Reveals Genes Involved in Cellular Ageing and Growth. *Biology Open* 3 (2): 161-171.

Rapicavoli, N. A., Qu, K., Zhang, J., Mikhail, M., Laberge, R.-M. and Chang, H. Y. (2013). A Mammalian Pseudogene LncRNA at the Interface of Inflammation and Anti-Inflammatory Therapeutics. *eLife* 2: e00762.

Redon, S., Reichenbach, P. and Lingner, J. (2010). The Non-Coding RNA TERRA is a Natural Ligand and Direct Inhibitor of Human Telomerase. *Nucleic Acids Research* 38 (17): 5797-5806.

Rhind, N., Chen, Z., Yassour, M., Thompson, D. A., Haas, B. J., *et al.* (2011). Comparative Functional Genomics of the Fission Yeasts. *Science* 332 (6032): 930-936.

Rodríguez-López, M., Cotobal, C., Fernández-Sánchez, O., Borbarán Bravo, N., Oktriani, R., *et al.* (2016). A CRISPR/Cas9-Based Method and Primer Design Tool for Seamless Genome Editing in Fission Yeast. *Wellcome Open Research* 1: 19.

Roux, A. E., Leroux, A., Alaamery, M. A., Hoffman, C. S., Chartrand, P., Ferbeyre, G. and Rokeach, L. A. (2009). Pro-Aging Effects of Glucose Signaling through a G Protein-Coupled Glucose Receptor in Fission Yeast. *PLOS Genetics* 5 (3): e1000408.

Roux, A. E., Quissac, A., Chartrand, P., Ferbeyre, G. and Rokeach, L. A. (2006). Regulation of Chronological Aging in *Schizosaccharomyces pombe* by the Protein Kinases Pka1 and Sck2. *Aging Cell* 5 (4): 345-357.

Sacerdot, C., Casaregola, S., Lafontaine, I., Tekaia, F., Dujon, B. and Ozier-Kalogeropoulos, O. (2008). Promiscuous DNA in the Nuclear Genomes of Hemiascomycetous Yeasts. *FEMS Yeast Research* 8 (6): 846-857.

Sambrook, J., Fritsch, E. F. and Maniatis, T. A. (1989). Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, New York.

Samper, E., Flores, J. M. and Blasco, M. A. (2001). Restoration of Telomerase Activity Rescues Chromosomal Instability and Premature Aging in $Terc^{-/-}$ Mice with Short Telomeres. *EMBO Reports* 2 (9): 800-807.

Sideri, T., Rallis, C., Bitton, D. A., Lages, B. M., Suo, F., Rodríguez-López, M., Du, L.-L. and Bähler, J. (2014). Parallel Profiling of Fission Yeast Deletion Mutants for Proliferation and for Lifespan During Long-Term Quiescence. *G3* 5 (1): 145-155.

Siebold, A. P., Banerjee, R., Tie, F., Kiss, D. L., Moskowitz, J. and Harte, P. J. (2010). Polycomb Repressive Complex 2 and Trithorax Modulate *Drosophila* Longevity and Stress Resistance. *PNAS* 107 (1): 169-174.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., *et al.* (2005). Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes. *Genome Research* 15 (8): 1034-1050.

Sierra, M. I., Fernández, A. F. and Fraga, M. F. (2015). Epigenetics of Aging. *Current Genomics* 16 (6): 435-440.

Singleton, T. L. and Levin, H. L. (2002). A Long Terminal Repeat Retrotransposon of Fission Yeast has Strong Preferences for Specific Sites of Insertion. *Eukaryotic Cell* 1 (1): 44-55.

Suga, M. and Hatakeyama, T. (2005). A Rapid and Simple Procedure for High-Efficiency Lithium Acetate Transformation of Cryopreserved *Schizosaccharomyces pombe* Cells. *Yeast.* 22 (10): 799-804.

Tsurumi, A. and Li, W. X. (2012). Global Heterochromatin Loss: A Unifying Theory of Aging? *Epigenetics* 7 (7): 680-688.

van Opijnen, T. and Camilli, A. (2013). Transposon Insertion Sequencing: A New Tool for Systems-Level Analysis of Microorganisms. *Nature Reviews Microbiology* 11 (7): 435-442.

Visser, I. and Speekenbrink, M. (2010). depmixS4: An R Package for Hidden Markov Models. *Journal of Statistical Software* 36 (7): 1-21.

Watanabe, T., Miyashita, K., Saito, T. T., Nabeshima, K. and Nojima, H. (2002). Abundant Poly(A)-Bearing RNAs That Lack Open Reading Frames in *Schizosaccharomyces pombe*. *DNA Research* 9 (6): 209-215.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., *et al.* (2002). Initial Sequencing and Comparative Analysis of the Mouse Genome. *Nature* 420 (6915): 520-562.

Weakliem, D. L. (1999). A Critique of the Bayesian Information Criterion for Model Selection. *Sociological Methods and Research* 27 (3): 359-397.

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J. and Bähler, J. (2008). Dynamic Repertoire of a Eukaryotic Transcriptome Surveyed at Single-Nucleotide Resolution. *Nature* 453 (7199): 1239-1243.

Williams, G. C. (1957). Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution* 11 (4): 398-411.

Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Lyne, R., *et al.* (2002). The Genome Sequence of *Schizosaccharomyces pombe*. *Nature* 415 (6874): 871-880.

Wood, V., Harris, M. A., McDowall, M. D., Rutherford, K., Vaughan, B. W., *et al.* (2012). PomBase: A Comprehensive Online Resource for Fission Yeast. *Nucleic Acids Research* 40 (D1): D695-D699.

Yadon, A. N., Van de Mark, D., Basom, R., Delrow, J., Whitehouse, I. and Tsukiyama, T. (2010). Chromatin Remodeling around Nucleosome-Free Regions leads to Repression of Noncoding RNA Transcription. *Molecular and Cellular Biology* 30 (21): 5110-5122.

Zhang, Y. J., Ioerger, T. R., Huttenhower, C., Long, J. E., Sassetti, C. M., Sacchettini, J. C. and Rubin, E. J. (2012). Global Assessment of Genomic Regions Required for Growth in *Mycobacterium tuberculosis*. *PLOS Pathogens* 8 (9): e1002946.

Zhao, Z., Chen, C., Liu, Y. and Wu, C. (2014). 17β-Estradiol Treatment Inhibits Breast Cell Proliferation, Migration and Invasion by Decreasing MALAT-1 RNA Level. *Biochemical and Biophysical Research Communications* 445 (2): 388-393.

Zheng, L., Baumann, U. and Reymond, J.-L. (2004). An Efficient One-Step Site-Directed and Site-Saturation Mutagenesis Protocol. *Nucleic Acids Research* 32 (14): e115.

Zomer, A., Burghout, P., Bootsma, H. J., Hermans, P. W. M. and van Hijum, S. A. F. T. (2012). ESSENTIALS: Software for Rapid Analysis of High Throughput Transposon Insertion Sequencing Data. *PLOS ONE* 7 (8): e43012.

Supplementary Lists

List A includes 85 non-protein-coding genes with mean HMM state < 1.5, coverage > 50%, and not overlapping coding genes.

SPNCRNA.7	SPNCRNA.310	SPNCRNA.498
SPNCRNA.11	SPNCRNA.333	SPNCRNA.499
SPNCRNA.29	SPNCRNA.359	SPNCRNA.536
SPNCRNA.37	SPNCRNA.360	SPNCRNA.554
SPNCRNA.70	SPNCRNA.361	SPNCRNA.601
SPNCRNA.82	SPNCRNA.362	SPNCRNA.640
SPNCRNA.84	SPNCRNA.363	SPNCRNA.643
SPNCRNA.92	SPNCRNA.364	SPNCRNA.727
SPNCRNA.95	SPNCRNA.365	SPNCRNA.802
SPNCRNA.96	SPNCRNA.366	SPNCRNA.851
SPNCRNA.99	SPNCRNA.367	SPNCRNA.855
SPNCRNA.126	SPNCRNA.368	SPNCRNA.906
SPNCRNA.127	SPNCRNA.369	SPNCRNA.961
SPNCRNA.128	SPNCRNA.370	SPNCRNA.1076
SPNCRNA.136	SPNCRNA.371	SPNCRNA.1094
SPNCRNA.137	SPNCRNA.372	SPNCRNA.1095
SPNCRNA.138	SPNCRNA.373	SPNCRNA.1135
SPNCRNA.214	SPNCRNA.389	SPNCRNA.1160
SPNCRNA.230	SPNCRNA.390	SPNCRNA.1171
SPNCRNA.231	SPNCRNA.391	SPNCRNA.1298
SPNCRNA.232	SPNCRNA.396	SPNCRNA.1299
SPNCRNA.233	SPNCRNA.448	SPNCRNA.1301
SPNCRNA.234	SPNCRNA.449	SPNCRNA.1304
SPNCRNA.280	SPNCRNA.482	SPNCRNA.1306
SPNCRNA.281	SPNCRNA.483	SPNCRNA.1310
SPNCRNA.283	SPNCRNA.484	SPNCRNA.1576
SPNCRNA.284	SPNCRNA.485	SPNCRNA.1635
SPNCRNA.287	SPNCRNA.497	SPNCRNA.1695
SPNCRNA.290		

List B includes 54 pro-ageing genes with a Spearman correlation p-value < 0.05 between count / mean count ratio and time.

Systematic I.D.	Name	Ontology Term
SPAC1071.11	-	FMN binding, oxidoreductase activity
SPAC1142.08	fhl1	regulation of transcription (DNA-templated), transcription (DNA-templated)
SPAC17G8.01c	trl1	tRNA metabolic process
SPAC1F5.10	fal1	ribosome biogenesis
SPAC20G8.01	cdc17	DNA recombination, DNA repair, DNA replication
SPAC20G8.09c	nat10	ribosome biogenesis
SPAC22E12.18	-	human CCNDBP1 ortholog
SPAC23C11.14	zhf1	detoxification, transmembrane transport
SPAC23H4.10c	thi4	cofactor metabolic process, vitamin metabolic process
SPAC25B8.02	sds3	chromatin organisation, regulation of transcription (DNA-templated), transcription (DNA-templated)
SPAC25B8.19c	loz1	regulation of transcription (DNA-templated), transcription (DNA-templated)
SPAC27D7.14c	tpr1	transcription (DNA-templated)
SPAC29B12.08	clr5	chromatin organisation, regulation of transcription (DNA-templated), transcription (DNA-templated)
SPAC2F3.11	-	exopolyphosphatase activity, metal ion binding
SPAC2F7.11	nrd1	conjugation with cellular fusion, regulation of transcription (DNA-templated), transcription (DNA-templated)
SPAC31G5.13	rpn11	mitotic sister chromatid segregation, protein catabolic process, protein modification by small protein conjugation or removal, regulation of mitotic cell cycle phase transition
SPAC343.18	rfp2	DNA repair, protein modification by small protein conjugation or removal
SPAC644.10	med11	transcription (DNA-templated)
SPAC664.10	klp2	conjugation with cellular fusion, establishment or maintenance of cell polarity, microtubule cytoskeleton organisation, mitotic sister chromatid segregation
SPAC8F11.07c	cdc24	DNA recombination, DNA repair, DNA replication

SPAPB1E7.09	ogm2	carbohydrate derivative metabolic process, protein glycosylation
SPAPYUG7.06	sdu1	hydrolase activity
SPBC119.11c	pac1	conjugation with cellular fusion, ribosome biogenesis
SPBC119.13c	prp31	mRNA metabolic process
SPBC11G11.03	mrt4	mRNA metabolic process, ribosome biogenesis
SPBC1289.04c	pob1	vesicle-mediated transport
SPBC12C2.10c	pst1	chromatin organisation, regulation of transcription (DNA-templated), transcription (DNA-templated)
SPBC146.14c	sec26	vesicle-mediated transport
SPBC16A3.14	-	metal ion binding, structural constituent of ribosome, superoxide dismutase activity
SPBC16C6.07c	rpt1	mitotic sister chromatid segregation, protein catabolic process, regulation of mitotic cell cycle phase transition
SPBC1778.01c	zuo1	cytoplasmic translation, protein folding
SPBC17G9.03c	krs1	cellular amino acid metabolic process, cytoplasmic translation, tRNA metabolic process
SPBC29A10.07	pom152	nucleocytoplasmic transport
SPBC2G2.12	hrs1	cellular amino acid metabolic process, cytoplasmic translation, tRNA metabolic process
SPBC30D10.11	gpi1	carbohydrate derivative metabolic process, lipid metabolic process
SPBC30D10.12c	rsm27	mitochondrial translation
SPBC336.11	vps52	vesicle-mediated transport
SPBC354.10	def1	DNA repair, protein catabolic process, telomere organisation
SPBC3B9.04	oms1	S-adenosylmethionine-dependent methyltransferase activity
SPBC3E7.01	fab1	conjugation with cellular fusion, lipid metabolic process, signalling
SPBC428.20c	alp6	establishment or maintenance of cell polarity, microtubule cytoskeleton organisation, protein complex assembly
SPBC4F6.14	nop4	ribosome biogenesis
SPBC651.08c	rpc1	transcription (DNA-templated)
SPBC660.13c	ssb1	DNA recombination, DNA repair, DNA replication, telomere organisation
SPBC6B1.10	prp17	mRNA metabolic process

SPBC776.08c	utp22	ribosome biogenesis
SPCC1393.08	-	regulation of transcription (DNA-templated), transcription (DNA-templated)
SPCC1827.05c	-	ribosome biogenesis
SPCC1902.01	gaf1	conjugation with cellular fusion, regulation of transcription (DNA-templated), transcription (DNA-templated)
SPCC417.08	tef3	cytoplasmic translation
SPCC553.08c	ria1	ribosome biogenesis
SPCC5E4.06	smc6	DNA recombination, DNA repair
SPCC757.09c	rnc1	mRNA 3'-UTR AU-rich region binding, mRNA binding, protein binding
SPCC895.05	for3	actin cytoskeleton organisation, establishment or maintenance of cell polarity, microtubule cytoskeleton organisation, mitotic cytokinesis, protein complex assembly

List C includes 136 anti-ageing genes with a Spearman correlation p-value < 0.05 between count / mean count ratio and time.

Systematic I.D.	Name	Ontology Term
SPAC1039.06	-	cell wall organisation or biogenesis
SPAC11E3.04c	ubc13	DNA repair, protein modification by small protein conjugation or removal
SPAC1296.05c	lcp1	regulation of transcription (DNA-templated), transcription (DNA-templated)
SPAC139.06	hat1	chromatin organisation, transcription (DNA-templated), regulation of transcription (DNA-templated)
SPAC13D6.02c	byr3	cytoplasmic translation
SPAC13F5.03c	gld1	carbohydrate metabolic process
SPAC13G7.09c	-	conserved fungal protein
SPAC140.01	sdh2	carbohydrate derivative metabolic process, generation of precursor metabolites and energy, nucleobase-containing small molecule metabolic process
SPAC1556.06	meu1	*Schizosaccharomyces* specific protein
SPAC16C9.06c	upf1	mRNA metabolic process
SPAC1751.01c	gti1	transmembrane transport
SPAC1783.06c	atg12	autophagy
SPAC17A5.11	rec12	DNA recombination, meiotic nuclear division
SPAC17H9.10c	ddb1	chromatin organisation, DNA repair, protein catabolic process, protein modification by small protein conjugation or removal, regulation of transcription (DNA-templated), transcription (DNA-templated)
SPAC1805.08	dlc1	conjugation with cellular fusion, meiotic nuclear division, mitotic sister chromatid segregation
SPAC1834.03c	hhf1	chromatin organisation
SPAC1834.10c	-	mitochondrion organisation, protein complex assembly
SPAC19A8.13	usp101	mRNA metabolic process
SPAC1A6.09c	lag1	lipid metabolic process
SPAC1B3.08	-	COP9 signalosome complex subunit 12 (predicted)

SPAC1D4.02c	grh1	vesicle-mediated transport
SPAC1F7.09c	dal2	nitrogen cycle metabolic process
SPAC1F8.08	-	*Schizosaccharomyces pombe* specific protein
SPAC20G8.10c	atg6	autophagy, lipid metabolic process
SPAC22A12.05	rpc11	mRNA metabolic process, transcription (DNA-templated), tRNA metabolic process
SPAC22A12.17c	-	short chain dehydrogenase (predicted)
SPAC22G7.03	-	*Schizosaccharomyces* specific protein
SPAC22H12.01c	mug35	*Schizosaccharomyces* specific protein
SPAC22H12.03	-	lipid metabolic process
SPAC23A1.08c	rpl3401	cytoplasmic translation
SPAC23C11.12	hcn1	meiotic nuclear division, mitotic sister chromatid segregation, protein catabolic process, regulation of mitotic cell cycle phase transition
SPAC23G3.04	ies4	chromatin organisation, transcription (DNA-templated), regulation of transcription (DNA-templated)
SPAC24H6.03	cul3	protein catabolic process, protein modification by small protein conjugation or removal
SPAC27E2.12	-	*Schizosaccharomyces pombe* specific protein
SPAC29B12.05c	-	mitochondrial S-adenosylmethionine-dependent methyltransferase (predicted)
SPAC29B12.11c	-	human WW domain binding protein-2 ortholog
SPAC2F3.08	sut1	transmembrane transport
SPAC2F7.17	mrf1	mitochondrial peptide chain release factor (predicted)
SPAC31A2.08	mrp20	mitochondrial ribosomal protein subunit L23 (predicted)
SPAC31A2.11c	cuf1	regulation of transcription (DNA-templated), transcription (DNA-templated)
SPAC3A12.17c	cys12	cellular amino acid metabolic process
SPAC3C7.08c	elf1	nucleocytoplasmic transport, ribosome biogenesis
SPAC3F10.04	gsa1	detoxification
SPAC3G6.02	rpn15	mitotic sister chromatid segregation, nucleocytoplasmic transport, protein catabolic process,

		protein complex assembly, regulation of mitotic cell cycle phase transition
SPAC3G9.04	ssu72	mitotic sister chromatid segregation, transcription (DNA-templated)
SPAC3H1.08c	-	transmembrane transport
SPAC3H1.11	hsr1	regulation of transcription (DNA-templated), transcription (DNA-templated)
SPAC4D7.07c	csi2	microtubule cytoskeleton organisation, mitotic sister chromatid segregation
SPAC4F10.14c	btf3	protein folding, protein targeting
SPAC4G8.02c	sss1	protein targeting, transmembrane transport
SPAC4G9.11c	cmb1	DNA repair
SPAC4G9.20c	ymc1	transmembrane transport
SPAC513.04	-	*Schizosaccharomyces pombe* specific protein
SPAC521.02	wss1	DNA repair
SPAC56F8.04c	ppt1	cofactor metabolic process
SPAC57A10.04	mug10	signalling
SPAC5H10.03	-	phosphoglycerate mutase family
SPAC5H10.04	-	NADPH dehydrogenase (predicted)
SPAC637.11	rpm2	mitochondrion organisation
SPAC637.13c	slm1	actin cytoskeleton organisation, establishment or maintenance of cell polarity, signalling
SPAC683.03	-	*Schizosaccharomyces pombe* specific protein
SPAC823.09c	-	threonine aspartase (predicted)
SPAC823.14	ptf1	Mst2 histone acetytransferase acytyltransferase complex (predicted)
SPAC890.02c	alp7	establishment or maintenance of cell polarity, microtubule cytoskeleton organisation
SPAC8C9.04	-	*Schizosaccharomyces* specific protein
SPAC959.05c	pdi4	protein disulfide isomerase (predicted)
SPAC9G1.03c	rpl3001	cytoplasmic translation, ribosome biogenesis
SPAC9G1.15c	mzt1	microtubule cytoskeleton organisation, protein complex assembly

SPACUNK4.16c	tps3	carbohydrate metabolic process
SPACUNK4.19	mug153	*Schizosaccharomyces pombe* specific protein
SPAP14E8.05c	-	UPF0136 family mitochondrial protein (implicated in heme biosynthesis)
SPAP27G11.16	-	*Schizosaccharomyces pombe* specific protein
SPAPB24D3.04c	mag1	DNA repair
SPBC106.12c	tho4	nucleocytoplasmic transport, regulation of transcription (DNA-templated), transcription (DNA-templated)
SPBC106.13	gid9	carbohydrate metabolic process, protein catabolic process
SPBC115.01c	rrp46	mRNA metabolic process, ribosome biogenesis, tRNA metabolic process
SPBC11C11.06c	-	*Schizosaccharomyces* specific protein
SPBC12D12.02c	cdm1	DNA repair, DNA replication
SPBC1347.01c	rev1	DNA repair, mitochondrion organisation
SPBC14C8.04	ilv6	cellular amino acid metabolic process
SPBC14C8.09c	dbl3	IMPACT domain protein, possible chaperone (predicted)
SPBC15D4.13c	-	human ASCC1 ortholog
SPBC1683.06c	urh1	carbohydrate derivative metabolic process, cofactor metabolic process, nucleobase-containing small molecule metabolic process
SPBC1683.12	-	transmembrane transport
SPBC16A3.03c	ppr7	mitochondrion organisation
SPBC16G5.19	-	*Schizosaccharomyces pombe* specific protein
SPBC17G9.08c	cnt5	signalling
SPBC18H10.19	vps38	autophagy
SPBC21C3.09c	-	cellular amino acid metabolic process
SPBC21C3.10c	-	vitamin metabolic process
SPBC25B2.10	-	Usp (universal stress protein) family protein
SPBC26H8.01	thi2	vitamin metabolic process

SPBC27.03	meu25	*Schizosaccharomyces* specific protein
SPBC28F2.03	ppi1	protein folding
SPBC29A10.13	atp7	carbohydrate derivative metabolic process, nucleobase-containing small molecule metabolic process, transmembrane transport
SPBC2D10.03c	-	DUF866 domain protein
SPBC2D10.09	snr1	cellular amino acid metabolic process
SPBC2G2.04c	mmf1	mitochondrial matrix protein (YjgF family protein Mmf1)
SPBC30D10.05c	-	cofactor metabolic process
SPBC30D10.09c	-	HVA22/TB2/DP1 family protein
SPBC336.13c	mmp2	mitochondrion organisation, protein maturation, protein targeting
SPBC3B8.07c	dsd1	lipid metabolic process
SPBC4F6.08c	mrpl39	mitochondrial ribosomal protein subunit L39 (predicted)
SPBC56F2.14	mrpl44	mitochondrial ribosomal protein subunit l44 (predicted)
SPBC685.04c	aps2	vesicle-mediated transport
SPBC725.03	-	cofactor metabolic process, vitamin metabolic process
SPBC725.11c	php2	carbohydrate metabolic process, regulation of transcription (DNA-templated), transcription (DNA-templated)
SPBC83.16c	-	protein with a role in clearing protein aggregates (predicted)
SPBC8D2.02c	vps68	vesicle-mediated transport
SPBC8D2.09c	msl1	mRNA metabolic process
SPBP19A11.01	gcv3	cellular amino acid metabolic process
SPBP4H10.08	qcr10	carbohydrate derivative metabolic process, generation of precursor metabolites and energy, nucleobase-containing small molecule metabolic process
SPBPB7E8.02	-	PSP1 family protein
SPCC1259.05c	cox9	generation of precursor metabolites and energy
SPCC1259.09c	pdx1	cofactor metabolic process

SPCC1259.12c	gid1	carbohydrate metabolic process, protein catabolic process
SPCC1281.07c	gst4	detoxification
SPCC1281.08	wtf11	wtf element Wtf11
SPCC1322.01	rpm1	mitochondrion organisation
SPCC13B11.01	adh1	carbohydrate derivative metabolic process, carbohydrate metabolic process, cofactor metabolic process, generation of precursor metabolites and energy, nucleobase-containing small molecule metabolic process
SPCC1442.14c	hnt1	adenosine 5'-monophosphoramidase (predicted)
SPCC23B6.05c	ssb3	DNA recombination, DNA repair, DNA replication, telomere organisation
SPCC24B10.03	-	*Schizosaccharomyces* specific protein
SPCC338.05c	mms2	DNA repair, protein modification by small protein conjugation or removal
SPCC338.10c	cox5	carbohydrate derivative metabolic process, generation of precursor metabolites and energy, nucleobase-containing small molecule metabolic process
SPCC338.12	pbi2	membrane organisation
SPCC417.09c	-	regulation of transcription (DNA-templated), transcription (DNA-templated)
SPCC417.11c	-	cofactor metabolic process
SPCC548.06c	ght8	transmembrane transport
SPCC550.01c	coa4	mitochondrion organisation, protein complex assembly
SPCC550.07	-	acetamidase (predicted)
SPCC663.04	rpl39	cytoplasmic translation
SPCC70.02c	inh1	carbohydrate derivative metabolic process, nucleobase-containing small molecule metabolic process, transmembrane transport
SPCC794.02	wtf5	wtf element Wtf5
SPCC965.04c	yme1	mitochondrion organisation, protein catabolic process
SPCP20C8.02c	-	*S. pombe* specific UPF0321 family protein 1

List D includes 218 non-protein-coding genes with mean HMM state < 1.5, coverage > 50%, and overlapping all four time points.

SPNCRNA.7	SPNCRNA.333	SPNCRNA.565
SPNCRNA.11	SPNCRNA.359	SPNCRNA.568
SPNCRNA.16	SPNCRNA.360	SPNCRNA.569
SPNCRNA.22	SPNCRNA.361	SPNCRNA.572
SPNCRNA.29	SPNCRNA.362	SPNCRNA.576
SPNCRNA.37	SPNCRNA.363	SPNCRNA.578
SPNCRNA.47	SPNCRNA.364	SPNCRNA.579
SPNCRNA.61	SPNCRNA.365	SPNCRNA.581
SPNCRNA.70	SPNCRNA.366	SPNCRNA.584
SPNCRNA.84	SPNCRNA.367	SPNCRNA.601
SPNCRNA.92	SPNCRNA.368	SPNCRNA.603
SPNCRNA.95	SPNCRNA.369	SPNCRNA.604
SPNCRNA.96	SPNCRNA.370	SPNCRNA.605
SPNCRNA.99	SPNCRNA.371	SPNCRNA.606
SPNCRNA.117	SPNCRNA.372	SPNCRNA.633
SPNCRNA.119	SPNCRNA.373	SPNCRNA.639
SPNCRNA.128	SPNCRNA.389	SPNCRNA.673
SPNCRNA.129	SPNCRNA.390	SPNCRNA.676
SPNCRNA.136	SPNCRNA.391	SPNCRNA.677
SPNCRNA.137	SPNCRNA.396	SPNCRNA.688
SPNCRNA.138	SPNCRNA.446	SPNCRNA.695
SPNCRNA.139	SPNCRNA.448	SPNCRNA.700
SPNCRNA.154	SPNCRNA.449	SPNCRNA.734
SPNCRNA.201	SPNCRNA.471	SPNCRNA.750
SPNCRNA.230	SPNCRNA.482	SPNCRNA.759
SPNCRNA.231	SPNCRNA.483	SPNCRNA.789
SPNCRNA.232	SPNCRNA.484	SPNCRNA.796
SPNCRNA.233	SPNCRNA.485	SPNCRNA.799
SPNCRNA.234	SPNCRNA.487	SPNCRNA.802
SPNCRNA.280	SPNCRNA.497	SPNCRNA.805
SPNCRNA.281	SPNCRNA.498	SPNCRNA.838
SPNCRNA.283	SPNCRNA.499	SPNCRNA.847
SPNCRNA.284	SPNCRNA.503	SPNCRNA.885
SPNCRNA.287	SPNCRNA.513	SPNCRNA.898
SPNCRNA.289	SPNCRNA.532	SPNCRNA.908
SPNCRNA.290	SPNCRNA.536	SPNCRNA.914
SPNCRNA.291	SPNCRNA.541	SPNCRNA.924
SPNCRNA.292	SPNCRNA.556	SPNCRNA.936
SPNCRNA.293	SPNCRNA.558	SPNCRNA.937
SPNCRNA.310	SPNCRNA.563	SPNCRNA.952

SPNCRNA.961
SPNCRNA.1002
SPNCRNA.1017
SPNCRNA.1033
SPNCRNA.1062
SPNCRNA.1072
SPNCRNA.1076
SPNCRNA.1079
SPNCRNA.1094
SPNCRNA.1095
SPNCRNA.1096
SPNCRNA.1097
SPNCRNA.1104
SPNCRNA.1105
SPNCRNA.1118
SPNCRNA.1126
SPNCRNA.1127
SPNCRNA.1131
SPNCRNA.1135
SPNCRNA.1136
SPNCRNA.1139
SPNCRNA.1148
SPNCRNA.1160
SPNCRNA.1171
SPNCRNA.1173
SPNCRNA.1174
SPNCRNA.1178
SPNCRNA.1180
SPNCRNA.1181
SPNCRNA.1183
SPNCRNA.1191
SPNCRNA.1193
SPNCRNA.1195
SPNCRNA.1204
SPNCRNA.1212
SPNCRNA.1217
SPNCRNA.1224
SPNCRNA.1240
SPNCRNA.1250
SPNCRNA.1263
SPNCRNA.1264
SPNCRNA.1265
SPNCRNA.1266
SPNCRNA.1267
SPNCRNA.1268
SPNCRNA.1276
SPNCRNA.1277
SPNCRNA.1289
SPNCRNA.1294
SPNCRNA.1298
SPNCRNA.1299
SPNCRNA.1301
SPNCRNA.1303
SPNCRNA.1304
SPNCRNA.1306
SPNCRNA.1307
SPNCRNA.1309
SPNCRNA.1310
SPNCRNA.1321
SPNCRNA.1354
SPNCRNA.1356
SPNCRNA.1381
SPNCRNA.1383
SPNCRNA.1405
SPNCRNA.1412
SPNCRNA.1429
SPNCRNA.1454
SPNCRNA.1456
SPNCRNA.1476
SPNCRNA.1481
SPNCRNA.1489
SPNCRNA.1495
SPNCRNA.1508
SPNCRNA.1515
SPNCRNA.1525
SPNCRNA.1529
SPNCRNA.1535
SPNCRNA.1539
SPNCRNA.1570
SPNCRNA.1577
SPNCRNA.1582
SPNCRNA.1590
SPNCRNA.1592
SPNCRNA.1603
SPNCRNA.1613
SPNCRNA.1618
SPNCRNA.1620
SPNCRNA.1634
SPNCRNA.1635
SPNCRNA.1638
SPNCRNA.1639
SPNCRNA.1640
SPNCRNA.1655
SPNCRNA.1661
SPNCRNA.1675
SPNCRNA.1682
SPNCRNA.1695
SPNCRNA.1696