Hermes Materials and Methods

These detailed methods are extracted from the PhD thesis of Leanne Grech.
Contact: Daniel.Jeffares@york.ac.uk or j.bahler@ucl.ac.uk

1.1 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^{o}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^{o}C on selective agar plates to ensure plasmid retention. Strains without plasmids were stored at 4^{o}C on YES agar plates. For long term storage, strains were frozen at -80^{o}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).

1.2 *Hermes* Cell Libraries

1.2.1 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^{o}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^{o}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^{o}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^{o}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^{o}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^{o}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^{o}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.

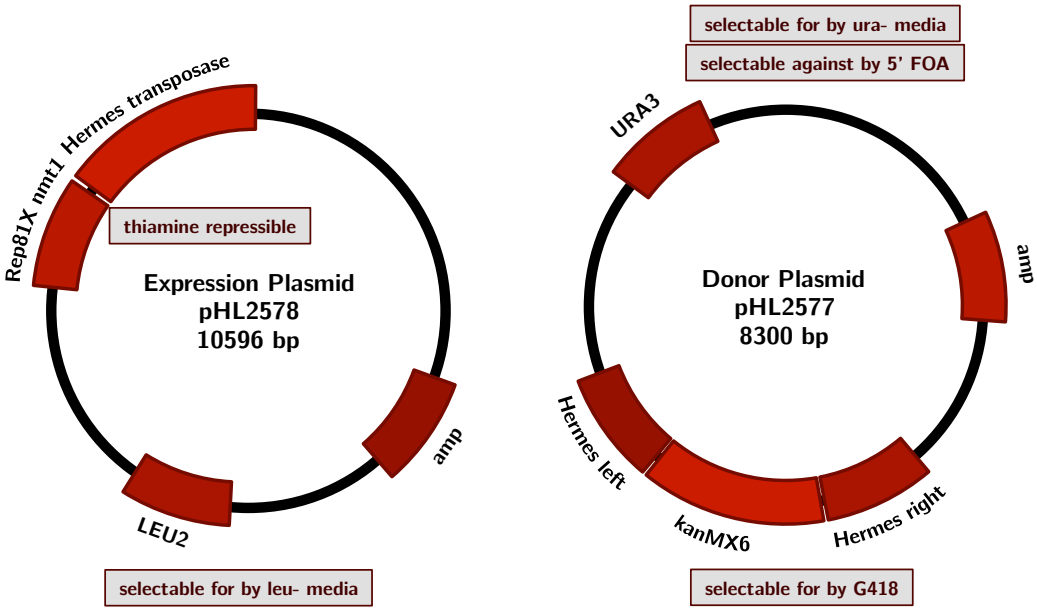


Figure 0.1. Donor and Expression Plasmids for *Hermes* Transposition in Fission Yeast. pHL2577 donor plasmid provides the source of transposon DNA flanked by the left and right terminal inverted repeats (TIRs). Here, the kanMX6 cassette between the TIRs gives cells with insertions resistance to the drug 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)).

1.2.2 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 $OD_{600} \approx 0.05$ (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 $OD_{600} \approx 2$ 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 $OD_{600} \approx 0.05$, with the actual OD_{600} value recorded as $OD_{initial(2)}$. Again, this culture was grown to a final $OD_{600} \approx 2$ 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 $OD_{final} \approx 2$ 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^{o}C.

1.3 *Hermes* DNA Libraries

1.3.1 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^{o}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^{o}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^{o}C for 2 minutes, 58^{o}C for 30 seconds, 72^{o}C for 30 seconds, 15 cycles of 94^{o}C for 30 seconds, 58^{o}C for 30 seconds and 72^{o}C for 30 seconds, followed by a final extension at 72^{o}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^{o}C for 4 minutes, 94^{o}C for 20 seconds, 56^{o}C for 20 seconds, and 72^{o}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 0.1. *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.

1.3.2 Linkers and Primers Design

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

This method amplifies Hermes-flaking DNA from a genome.

A. Linker Ligation

i) DNA Fragments with Hermes Insertions

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

1.4 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]**

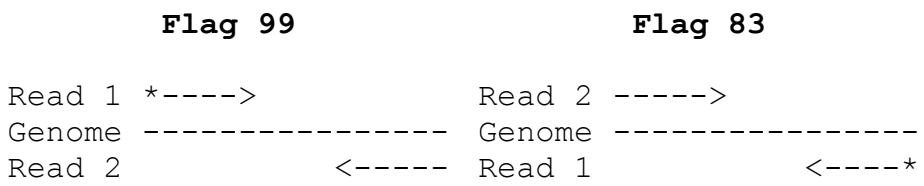
To process Read 2, a Perl script was written by Dr Daniel Jeffares to detect and exclude duplicate reads based on the [10mer] and the first 5 nt of the [Genome]. Then, the output was tallied with Read 1 and the [10mer][Linker] was trimmed using fastx_trimmer.

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.