A new approach to investigating neural transposition challenges current insertion mapping strategies and provides an alternative explanation for differential transposon expression



Christoph D. Treiber

Trinity College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2016

Für Hedi

Abstract

Prior work has suggested the provocative, and somewhat controversial idea, that transposable element mobilization introduces genomic heterogeneity in certain regions of the mammalian brain. Findings in the fruit fly suggested that this might be a conserved phenomenon. Perrat et al. (2013) reported that transposons are overexpressed in the αβ-Kenyon Cells (αβ-KCs) of the *Drosophila* mushroom bodies and whole genome sequencing suggested that *de-novo* insertions map to neural genes within these cells.

In this study I revisited these topics and investigated what impact elevated transposon expression has on the heterogeneity of the genomes of αβ-KCs. Since post-mitotic transposon insertions occur in single DNA molecules, they are rare and difficult to detect, and are particularly hard to distinguish from experimental artefacts. I developed a new protocol to analyse transposon insertions in individual flies, instead of groups. This approach significantly increased the theoretical power to detect rare somatic insertion events, and it also enabled comparison to somatic genomes from cells isolated from the same animal. This advance allowed me to distinguish true somatic events from rare germline polymorphism in the fly population. Furthermore, the new analysis pipeline utilises longer overlapping sequencing reads, which permits the identification of the precise break points of transposon insertion events.

I confirmed that some transposons are more highly expressed in αβ-KCs, but I did not find evidence that this increase correlates with an increased number of transposon insertions. In addition, I did not detect an increase in the number of transposon insertions with age, challenging the idea that damage accumulates with age. A set of *in-silico* simulation experiments revealed that the rate of putative somatic insertion events is comparable to the rate of presently unavoidable experimental artefacts. Lastly, my analyses suggest that elevated and cell-restricted expression of the LINE-like retrotransposon *Ivk* in αβ-KCs might result from its co-expression with the αβ-KC specific gene *CG17684*, within which a germline insertion of *Ivk* resides.

Table of contents

Acknowledgements 8

1. Introduction 11

1.1. Transposons provide flux in the eukaryotic genome 11

1.2. Two major classes of transposons exist in fruit flies 14

1.3. Transposon control mechanisms 15

1.4. Transposon activity in the germline 17

1.5. Transposons activity in somatic cells 18

1.6. Detecting transposable elements 19

1.7. Genomic mosaicism in neurons 25

1.8. The *Drosophila melanogaster* olfactory system 26

2. Results 28

2.1. Analysis of gene expression in αβ-KCs 28

2.2 Analysis of somatic transposon insertions αβ-KCs 50

2.3. Development of MRTemp, a novel approach for transposon mapping 72

2.4. Filtered MRTemp hits 83

2.5. A simulated experiment to test for exon mobilisation 96

2.6. Co-expression of transposons with neighbouring genes 104

3. Discussion 109

3.1. Local transposon hopping had previously been reported in flies 110

3.2. Somatic transposition is unlikely to occur at high frequency in adult fly brains 111

3.3. Co-expression of genes and transposons could present novel gene regulation mechanisms 113

3.4. Single-fly analysis will enable analyses on germline transposon dynamics 114

3.5. The dynamics of chimera formation are similar in bacteria and fruit flies 117

3.6. New *in-silico* simulated control experiments make it possible to test WGS data for chimeric artefacts 119

3.7. New and improved approaches are necessary to test for rare somatic transposon insertions 120

4. Methods 122

5. Supplementary Material 135

6. References 145

7. Abbreviations 155

8. Appendices 158

Appendix 1 – Unix codes 158

Appendix 2 – Shell script to append _1 165

Appendix 3 – Shell script to append _2 166

Appendix 4 – Shell script to extract reads 167

Appendix 5 – Shell script BLAST to genome 168

Appendix 6 – Shell script to extract coverage 169

Appendix 7 – FastQC reports 170

Appendix 8 – UPL primer sequences 196

Acknowledgements

First and foremost, I would like to thank Scott Waddell for giving me the opportunity to do what I love during the last three years. His trust and support were absolutely essential for the research presented in this study. When I started out in Scott's lab, I thought that I was going to discover exciting new concepts and principles that underlie individuality. But as the reader will be able to tell from the title of this thesis, I ended up heading towards a slightly different direction. Scott has taught me not to be discouraged by an unexpected result, and he provided an absolutely unique environment in which I was able to simply get on with (sometimes quite pricey sequencing) experiments. My favourite phrase during the last three years was Scott's reaction to almost all of my ideas for experiments: 'Do it!'. And despite his very busy schedule, Scott's door was never shut. Very importantly, I would like to thank Scott for his comments and edits on this manuscript.

I also want to acknowledge all members of the Waddell lab, for being a scientifically critical and fun group of people. Thanks to Vincent Croset for sharing the frustrations of failed FACS sessions with me, to Johannes Felsenberg for proof-reading this manuscript, and to Paola Cognigni for invaluable advice on fonts.

The facilities that I was able to use made my experiments a lot easier, and I am very thankful to Ruth Brain, Shanice Bailey, Lee Phillimore, Paul Scott, Adam Westhorpe, Fiona Wood, and Sinead Wright at the CNCB, Nigel Rust at the FACS facility, Robert Esnouf at the Research Computing Core, and Sharon Cure at the Trinity College library for all their help.

This work would have not been possible without the generous support of the Wellcome Trust, and I am grateful to Jonathan Flint for his help and support.

I am grateful to David Keays, who has taught me that results mean nothing without a control experiment. 'What if"s never make a lot of sense, but I am certain that I would not be where I am without Dave's support, advice and encouragement to work hard.

The list of people who have played a key role in the process leading up to this manuscript is long, and I suspect that some of the individuals will not be aware of their contribution. To name a few, I learned most of my computational skills from hour-long trouble-shooting sessions with Evan Harrel. Andy Dahl significantly shaped the way I approach scientific problems. Dave Deriso showed me how to get on with things. Paul Fairchild helped me put failed experiments

into perspective. I owe my gratitude to all of them. In addition, it has been an absolute privilege to be DPhil student at Trinity College, Oxford. I was very lucky to have met so many clever and inspiring people at Trinity, and a lot of these people have contributed to my work and life during the last few years.

I am forever grateful to my parents, for making it so very easy for me to follow my dreams, and for always unconditionally supporting me, to my brother Nik, for being my best friend and favourite phone chat buddy, and to Bernd Lettkemann, for always reminding where I came from.

And finally, thank you Clare, for being my companion, my friend (sit next to each other and hold hands), my favourite housemate, my proof reader (and ruthless comma killer), my cornet muse, my English teacher (the culture, not the language), the best team-mate (Go Team Clarly), and so much more…

1. Introduction

1.1. Transposons provide flux in the eukaryotic genome

Transposons are mobile genetic elements that exist in the genomic DNA (gDNA) of almost all eukaryotes, and they constitute around 50% of many eukaryotic genomes[1-5]. The high abundance of transposons in eukaryotes implies that they are an integral part of the genetic information that is stored in and inherited by cells throughout embryonic development, and across generations in all higher organisms. The balance between the conservation of information on one side, and the flexibility to change it on the other, determines the genetic integrity of a species, and also enables organisms to adapt to changes in the environment[6-8]. Therefore, if we want to further our understanding of inheritance and evolution, it will be essential to find out more about the frequency and prevalence of transposon mobilisation.

The fruit fly *Drosophila melanogaster* has been an invaluable system for studying chromosome dynamics and inheritance for more than 100 years. One reason why flies have been a popular model organism for geneticists and evolutionary biologists is the short reproductive life cycle, and the fact that fly strains are relatively easy to maintain[9,10]. Thomas Hunt Morgan, one of the founding fathers of genetics, discovered that genes are carried on chromosomes on the basis of his fruit fly experiments[11]. In a comment by Morgan, published in *Science* in 1911, he

writes: 'All that my hypothesis pretends to account for is that groups of factors that enter together tend to remain together'[12]. This intriguingly simple statement correctly predicted one of the most fundamental principles of inheritance: the notion that genes are located in specific regions on chromosomes, and that the closer two genes are located to each other, the more likely they are to be passed on together to the next generation. The fruit fly has also been instrumental in the outstanding work of Christiane Nüsslein-Volhard and Eric Wieschaus, who identified some of the fundamental principles and key molecules that are important during the development of an organism from the fertilised egg cell into an adult[13]. In their landmark paper, published in 1980, Nüsslein-Volhard and Wieschaus described 15 loci that are each required for the establishment of segmental patterns of fruit fly larvae during embryonal development.

Modern, high throughput genome sequencing techniques now make it possible to sequence the genome extracted from very small amounts of tissue, such as cells isolated from different segments of fly larvae. One of the main findings of these somatic sequencing efforts is the notion that genomes not only differ between, but also within individuals[14]. Therefore, the two areas of inheritance and embryonic development are actually more closely linked than previously thought. The transfer of genetic information is subject to modifications within the somatic tissue of an organism, in a similar way as in a population of individuals. Hence it is conceivable that individual somatic cells with different

genetic modifications within an organism compete with each other during embryonic development. This could lead to a selection of positive traits by a mechanism that is similar to the way natural selection drives evolutionary changes of the genome during sexual reproduction. In addition, genomic heterogeneity within an organism could have a positive effect on the fitness of the host, analogous to the beneficial impact of genomic variation in a population on its ability to adapt to changes in the environment. Random somatic changes to the genome can trigger the development of cancer[15], but changes to the genome have also been found in somatic cells of healthy adults, leading to the idea that organisms could benefit from somatic genome variations[16,17].

Mobile genetic elements were first observed as polymorphisms in the position and number of repetitive DNA sequences on the chromosomes of different fruit fly strains[18-20]. In the early 1970s, Yuri Viktorovich Ilyin and Georgii Pavlovich Georgiev developed a new technique that enabled them to hybridise RNA probes to DNA molecules. Using shotgun cloning, they created a probe for a randomly chosen fruit fly gene, and found that this probe labelled multiple locations on the salivary gland polytene chromosomes of their flies. Astonishingly, the locations of the genes were different between the paternal and the maternal chromosome. They later realised that their randomly cloned fragment contained a mobile genomic element, or transposon. The number and location of the element was highly variable in different fly strains. These findings were the first description of

a transposon in the fruit fly genome. We now know that the fruit fly genome harbours approximately 80 mobile element families[21]. The mobile elements can be categorised into two classes, based on their mechanism of mobilisation.

1.2. Two major classes of transposons exist in fruit flies

1.2.1. Retrotransposons

Retrotransposons replicate using a *copy-paste* mechanism, which comprises of the generation of an RNA intermediate, followed by reverse transcription and the insertion of a newly synthesised DNA fragment[7]. As a result, the copy number of each individual transposon in the host genome increases with every replication cycle. Retrotransposition relies on the RNA polymerase of the host cell, which transcribes the retrotransposon into RNA. These RNA molecules encode for several genes, including a reverse transcriptase (RT). RT binds to the transposon RNA molecule, and creates a cDNA copy of the fragment. This cDNA fragment can then re-insert into the genome at a new location, using one strand of the gDNA as a primer to synthesise a new double-stranded DNA copy of the mobile element. During this process, the nucleotide sequences that flank the target site are sometimes polymerised on both ends of the transposon, leading to so called target site duplications (TSD). Retrotransposons consist of two subtypes of elements, long terminal repeat (LTR) and non-LTR transposons. Phylogenetic analyses of the protein sequence of RT in transposons show that LTR and non-LTR transposons represent two separate classes of mobile elements[22].

1.2.2. DNA transposons

DNA transposons mobilise via a *cut-paste* mechanism[7]. They are excised from the host genome by transposase, an enzyme that is encoded within the transposon sequence. The excised fragment has the potential to re-insert into a new location - a process that is, again, mediated by transposase. This mechanism is typically non-replicative, with the number of copies of a particular DNA transposon only increasing when an element transposes during the DNA synthesis that precedes mitosis. To increase its copy number, the transposon must be cut out from an already duplicated part of the genome, and inserts into a section of the DNA that had not yet been replicated. Insertions of DNA transposons can also lead to the formation of TSDs, resulting from gap-repair at the site of re-insertion. DNA transposons are thought to be inactive in mammals. In fruit flies, the *P* element is a DNA transposon which has swept through all wild type strains of *Drosophila melanogaster*, but is absent from most lab strains[23]. Selective reintroduction of the *P* element mobilisation process has therefore been extensively utilised for genome engineering and transgenesis in laboratory *Drosophila melanogaster*.

1.3. Transposon control mechanisms

The disruptive potential of transposons challenges the integrity of gDNA. Hence, organisms have developed strategies to inhibit transposon mobilisation,

especially in germline cells. The defence system that protects gDNA from transposon mobilisation is sometimes referred to as the immune system of the genome[24-26]. The main challenges for the transposon defence are to identify transposon sequences across generations, and also to be able to adapt to the occurrence of new, active mobile elements. In animals, transposons are down-regulated on a transcriptional level by targeted DNA silencing[27], and post-transcriptionally via a phenomenon called RNA interference (RNAi)[28]. Both of these mechanisms are mediated by RNA induced silencing complexes (RISCs). RISCS are made up of a small noncoding RNA molecule and a member of the Argonaute protein family[29]. The protein components have various functions, and the RNA provides sequence specificity via complementary base-pairing with the target molecule. Three types of RISCs are currently known: (1) Endogenous small interfering RNAs (endo-siRNA) are transcribed from endogenous loci in the fly genome and form short hairpin structures, which are then cleaved into ~21 nucleotide long single stranded RNA molecules by the Dicer2 protein[30]. These mature endo-siRNAs are then bound by the Argonaute protein Ago2 to form a RISC, which subsequently cleaves complementary transposon RNA molecules in somatic cells. (2) In fruit flies, microRNAs (miRNA) are processed by Dicer1, and bind to Ago1 to then translationally repress target genes[31,32]. miRNA bind complementary mRNAs and trigger their cleavage or destabilization[33]. More recently, miRNA activity has also been linked with transposon control in fruit flies[34]. (3) PIWI-interacting RNAs (piRNA) are encoded in clusters on the genome[35]. These loci are a collection of fragments of transposon sequences and

represent the 'immunological memory' of the cell. PiRNAs are expressed in the germline of fruit flies, and they bind to three Argonaute proteins: Piwi, Aubergine (Aub) and Ago3. Piwi-RISCs bind to transposon loci in the nucleus and silence them transcriptionally. Aub and Ago3 RISCs reside in the cytoplasm, where they degrade transposon RNA molecules and also self-amplify via the so called ping-pong replication cycle. This self-proliferation is triggered if active transposon target sequences are encountered, which provides the potential to adapt to more active mobile elements. The piRNA pathway was first identified in the germline of animals, but we now know that Piwi, Aub, Ago3 are also expressed in somatic tissue[36-38]. In higher vertebrates, transposons have also been shown to be transcriptionally silenced by zinc-finger proteins[39,40].

1.4. Transposon activity in the germline

The replication process of both classes of transposons can induce changes to the gDNA of an organism. For example, transposons can insert into genes in the host cell and disrupt their expression[41,42]. In addition, they can insert close to endogenous genes and alter the expression patterns of those genes[43,44]. More generally, transposons have also been shown to influence the overall stability of gDNA, and they are hotspots for DNA double-strand breaks and recombination-driven chromosome rearrangements[45]. The distribution of transposon insertions in the germline appears to be distinct for each fly line, and most transposons are private to a single strain[46]. Likewise, patterns of epigenetic silencing of

transposons exhibit a similar species specificity. These findings suggest that transposons might play a significant role in speciation[47]. In addition, Jacobs et al. have recently shown that KRAB zinc-finger genes, which had previously been linked to transposon silencing, are amongst the fastest evolving genes in the mouse genome. This suggests the existence of an arms race between transposons and their regulatory elements. Furthermore, subsequent mutations of transposon to evade zinc-finger protein control might trigger speciation events of organisms[48].

1.5. Transposons activity in somatic cells

The germline is a cell niche that is very tightly controlled, since every stable modification of the DNA of those cells will propagate to all cells of the offspring. In contrast, modifications of gDNA in somatic cells of organisms are not inherited in the next generation. They can propagate in cell populations within an organism during embryonic development, thereby rendering the individual a genetic mosaic[49]. McConnell et al. sequenced single cells extracted from human brain samples and found that mosaic variations in the copy number of gDNA regions are abundant in human neurons[50]. Evrony et al. reported that somatic transposon insertions, and in addition point mutations in *de-novo* inserted transposons, occur during the embryonic development of the human brain[51]. They found that these events can be used to trace the origin of different cell lineages in the adult brain. These findings have led to the hypothesis that somatic

transposon insertions might be one of several molecular mechanisms that could diversify clonally related groups of neurons in animals[14,16]. In theory, the functional diversity of neurons could result in idiosyncratic behaviour of individuals. This idiosyncrasy would persist throughout the lifetime of an organism, but it would not be passed on to the offspring. As a consequence, such behavioural variation in a single generation could potentially be beneficial for animals to adapt to temporary changes in their environment. Conversely, somatic transposition events might also be detrimental to neuronal function, and they might contribute to age-dependent neuronal decline. In humans the only autonomously active transposons in somatic cells are part of the L1 family and each L1 copy has a very similar, and largely indistinguishable sequence[49]. In fruit flies, around 80 different transposon families are known and most of them are still considered to be active in somatic cells[52,53].

1.6. Detecting transposable elements

There are three ways to monitor transposon activity. First, reporter constructs can be used to test whether cells have the potential to allow transposon activity. Second, the RNA levels of transposon genes can be assessed by real-time quantitative PCR (RT-qPCR), microarrays, or RNA-seq analyses. And third, *de-novo* insertions in the gDNA can be detected by PCR based approaches or by whole-genome sequencing (WGS).

1.6.1. Transposition reporter approaches

In 2005, Muotri et al. found that neuronal precursor cells can host the transposition of a reporter construct[54]. The authors based their conclusions on experiments using an L1 transposition construct. This construct consisted of both open reading frames of the human LINE-1 element (L1) and in addition a gene that encodes for the green fluorescent protein (GFP). The *GFP* gene was interrupted by an intron in the opposite transcriptional direction to that of the rest of construct. As a consequence, GFP can only be expressed from this construct when the L1 construct is transcribed, spliced, reverse transcribed, and re-inserted into the genome at a new location[54-58]. According to the authors, seven days after transfection, HeLa cells, rat hippocampal neural stem cells, and hippocampal-derived neural progenitor cells expressed GFP, indicating that retrotransposition was possible in those cells. In follow up studies, the authors also reported GFP expression in transfected human fetal brain stem cells[59,60].

Whilst these results suggest that adult neurons in humans and mice have the potential to harbour somatic transposition events, it is important to note that all the findings were based on the expression of an exogenous reporter construct. Whether or not these artificial situations reflect what is endogenously happening in adult neurons requires further analysis.

In a separate approach, Li et al. developed a different type of transposon reporter construct in fruit flies[53]. Their *gypsy*-TRAP reporter consisted of a constitutively active promoter (*Tub*p), which drove the expression of Gal80, an inhibitor of the Gal4 transcription factor from the Gal4/UAS binary expression system[61]. The promoter is separated from the transcription site with a fragment from the *ovo* gene that is known to be a relative hotspot target site for insertion of *gypsy*, a fruit fly retrotransposon[62,63]. Flies harbouring this construct were crossed with Gal4/UAS-GFP flies. According to the authors, when *gypsy* inserted into the target site, the promoter was disrupted, leading to loss of expression of the inhibitory Gal80. Gal4-driven GFP expression was therefore released. Li et al. reported more GFP positive cells in aged flies, suggesting that *gypsy* is active in adult fruit fly brains. In a more recent study, Wood and colleagues have reported GFP positive cells in fat body cells of aged fruit flies in the presence of the *gypsy*-TRAP reporter[64].

1.6.2. RT-qPCR-based approaches

Another approach to monitor transposon activity is to assess the expression levels of RNA from transposon sequences. Numerous studies have shown that transposon RNA is abundant in adult neurons[52-54]. However, it should be emphasised that the link between expression and *de-novo* insertion remains elusive, and it is not entirely clear whether increased transposon expression levels lead to increased insertion rates. In addition, primer design for RT-qPCR

experiments on transposons are more challenging than for classical genes for two reasons. Firstly, transposons contain a large number of repeats, which make it harder to identify primers that are unique for a specific target. Secondly, transposons are not spliced, which means that primer pairs cannot be designed to span introns. As a consequence, it is not possible to distinguish mRNA signals from gDNA contamination. The latter issue makes it necessary to always include a DNAse digestion step during the RNA extraction. In addition, samples, for which no cDNA was formed (-RT samples) must be tested alongside cDNA samples to monitor potential gDNA contamination.

1.6.3. gDNA based approaches

The third, and most direct way to monitor transposon insertions is by analysing the gDNA of cells. The challenge this approach faces is the fact that rare somatic transposon insertions can occur in a single DNA molecule, and are therefore difficult to detect in a reliable and reproducible fashion. Several strategies have been developed that try to overcome this issue. Muotri et al. have developed an indirect approach, which is based on qPCR analysis of gDNA, called quantitative multiplex polymerase chain reaction[54]. The assay monitors the amplification of a particular target sequence during PCR, to determine its copy number. Probes that target L1 are used to assess the relative copy number of the retrotransposon in a gDNA sample. Immobile sequences with high genome copy numbers have to be used for normalisation. This approach suggested an increase of *de-novo* L1

insertions in the hippocampus of human brain samples. Another approach for analysing rare somatic transposon insertions is to sequence the entire genome of a tissue sample[65]. This results in an inherent sensitivity issue, as WGS approaches typically have a coverage of around 10x. This means that on average, 10 individual reads exist for each nucleotide in the genome. A tissue sample normally contains millions of individual cells, and hence an event that is unique to one of those cells will be difficult to detect. Again, several approaches aim to address this issue. For example, Baillie et al. have developed an approach that allows them to enrich a DNA sample for L1 sequences, a method known as retrotransposon capture sequencing (RC-seq)[66]. This method increased the power to detect rare somatic insertion events. Alternatively, the sensitivity issue can be approached by reducing the number of cells that are analysed.

Generally, current WGS approaches are based on the formation of DNA fragments, followed by the sequence determination of both ends of these fragments. This analysis results in paired-end reads, which are subsequently mapped onto a reference genome. If a DNA sample contains a section of DNA, which is absent in the reference genome then the read pairs that span this fragment on each side are classified as discordant read pairs. The most common technique to find evidence for transposon insertions in WGS data is to identify those discordant read pairs. One such analysis pipeline was developed by Weng and colleagues, and is called TEMP[67]. TEMP uses the set of discordant read pairs

from a biological sample and maps them to a set of transposon reference sequences. If a read maps onto one of these reference sequences, then this read pair is considered to be one piece of evidence for a transposon insertion. TEMP then calculates a theoretical frequency for the insertion event, based on these reads. Furthermore, it screens the data for so called clipped reads at each potential insertion locus, in order to determine the exact break point of the insertion event. This approach is powerful when characterising frequent insertion events (i.e. events that occurred in the germline of an organism) and events in areas with high sequencing depth, but falls short for rare somatic insertions.

The evidence for most somatic insertion events has, up until now, been based on single, or very few paired-end reads[52,68,69]. These assumptions come with a set of substantial limitations. As the DNA sequence between two sequencing reads is unknown, TEMP (and other similar algorithms) cannot predict the exact break point from this data, and relies instead on a prediction based on the assumed fragment size of the biological sample. There is reason to believe that this is an inappropriate assumption, and this study will present evidence for a broad distribution of DNA fragment sizes. Another caveat is the fact that gDNA must be amplified before it can undergo WGS. It has previously been shown that all amplification methods that have traditionally been used to assess transposon insertions in gDNA can lead to the formation of chimeric DNA molecules[70].

These chimeric fragments would be indistinguishable from true somatic insertion events.

1.7. Genomic mosaicism in neurons

Despite the potential caveats mentioned above, several studies report the existence of somatic transposon insertions[14,71]. Upton et al. recently published a study in which they performed the RC-seq method on single cells from human samples[69]. The authors claim to find rare somatic insertion events in cells from the hippocampus with a frequency of 13.7 insertions per neuron. However, a re-analysis of the same data set by Evrony et al. strongly suggests that the majority of putative insertion events are actually artefacts, and that the true rate is rather closer to 0.2 events per cell or lower[72].

In a separate study, Evrony et al. performed genome-wide L1 profiling on single neurons from human brain samples by enriching the amplified gDNA of cells using an L1-targeted PCR-based approach. This approach specifically amplifies DNA at 3′L1 junctions[73]. The authors reported a rate of *de-novo* L1 insertions with intact 3′ends of less than 0.04 events per cell, when corrected for putative confounding errors. Of the 300 neurons they analysed, they were only able to confirm one single full-length *de-novo* somatic L1 insertion.

1.8. The *Drosophila melanogaster* olfactory system

Fruit flies detect odours with olfactory receptor neurons in the antenna and the maxillary palp[74]. These sensory neurons send projections to individual glomeruli in an area of the fly brain called the antennal lobe. Every type of olfactory receptor neuron that expresses a given olfactory receptor gene projects to the same glomerulus. This means that every chemical compound that is detected by the receptors activates the same subset of glomeruli in every fly. As a result, unique patterns of the, approximately, 50 glomeruli fire when a fly encounters a particular odour. The odour information is then relayed by a set of projection neurons to the mushroom body, a bilateral structure in the fly brain made up of ~2000 Kenyon cells (KCs) per hemisphere. Individual odorants are coded as activity in a specific sparse subpopulation of KCs in the overall ensemble[75,76]. The KCs can broadly be subdivided by their anatomy and function in associative olfactory memory into around 800 αβ-KCs, as well as α′β′, and γ-neurons[77]. For example, whilst α′β′-KC output is critical during memory consolidation, αβ-KCs output is required during appetitive memory retrieval[78,79]. In addition, KC subdivision can be visualised with the expression of several genes that exclusively label different subtypes[80]. In 2013, Perrat et al. reported that the expression of several types of transposons are elevated in the αβ-KCs in comparison to other cell types[52]. This makes αβ-KCs a perfect system for studying the relationship between transposon expression levels and somatic insertion events.

For this study, I developed a protocol that allowed me to analyse the genomes of αβ-KCs in individual flies. The fact that my experimental setup did not require me to sequence the genome of single cells meant that my overall coverage was significantly better than any publicly available data from single cells. To my knowledge, this study is the first where the genomes of neurons from single fruit flies are sequenced. A significant advantage of the fruit fly as a model organism for studying transposition is that flies have multiple transposon types, which have unique sequences. This contrasts with rodents and humans, where actively expressed transposons are almost indistinguishable from each other because they are members of the L1 or Alu families. In this thesis, the advantageous presence of multiple classes of elements made it possible to correlate the expression of a specific transposon type with all its neighbouring genes.

2. Results

A previous microarray expression analysis study from our lab reported that transposons are upregulated in the αβ-KCs of the mushroom body[52]. Subsequent sequencing of gDNA from αβ-KCs suggested the presence of somatic transposon insertions in these cells. In this thesis I revisited these conclusions with newly available tools, and with the most rigorous analysis possible.

I developed MRTemp (Merged Reads TEMP), a new approach that significantly improves the power of detecting rare transposon insertions in gDNA. In addition, I designed and performed a series of *in-silico* simulated experiments to test the prevalence of potential confounding factors of the current and previous transposon insertion mapping strategy.

2.1. Analysis of gene expression in αβ-KCs

I first verified that transposon expression is upregulated in αβ-KCs versus their expression in the rest of the brain. An inherent issue with analysing gene expression between different cell types is that investigators often analyse mRNA that is isolated from labelled cells, purified from different fly strains. The reality of this approach is that differences in the expression patterns of particular genes might result from subtle differences in genetic background. This is especially concerning if transposon expression levels are assessed because the prevalence

and genomic location of mobile genetic elements differ between fly strains[19]. In the prior study Perrat et al.[52] compared gene expression levels by collecting GFP labelled cells from four separate strains. Three of these strains expressed GFP in different subpopulations of Kenyon cells, and another one in the entire brain, except for the mushroom body. To achieve this labelling precision they used a clever combination of intersectional strategies that made use of the Gal4, lexA and flippase-FRT (flippase recombinant target) system. Although this was the best way at the time to specifically label αβ-KCs, the caveat of this design is that any potential differences in the genetic background may mask genuinely differentially expressed genes. Moreover, genes that are differentially expressed in the different fly strains are difficult to distinguish from genuine αβ-KC-specific genes.

2.1.1. Transposons expression varies between fly strains with different genetic backgrounds

I first tested whether transposon expression varied between the fly strains used in Perrat et al.[52] with Taq-man® assays and RT-qPCR. RNA was extracted from whole fly heads of the three Perrat strains that labelled αβ-, α′β′- and γ-KCs. The RNA levels were measured for four of the transposons that were reported to be upregulated in αβ-KCs, compared to α′β′-KCs. In analogy to Perrat et al.[52], I normalised the expression levels to those from flies that had been used to collect α′β′-KCs. The analyses revealed that certain transposons are expressed at

different levels between the strains (Figure 1). For example, the expression of the terminally inverted repeat (TIR) DNA transposon *Bari1* is higher in heads of flies from the αβ-KC expressing strain. As a consequence, the reported difference in *Bari1* expression in αβ-KCs may have actually arisen from elevated expression in the αβ-KC strain, rather than genuine differential expression in αβ-KCs versus α′β′-KCs.

Figure 1. Transposon expression varies with genomic background.

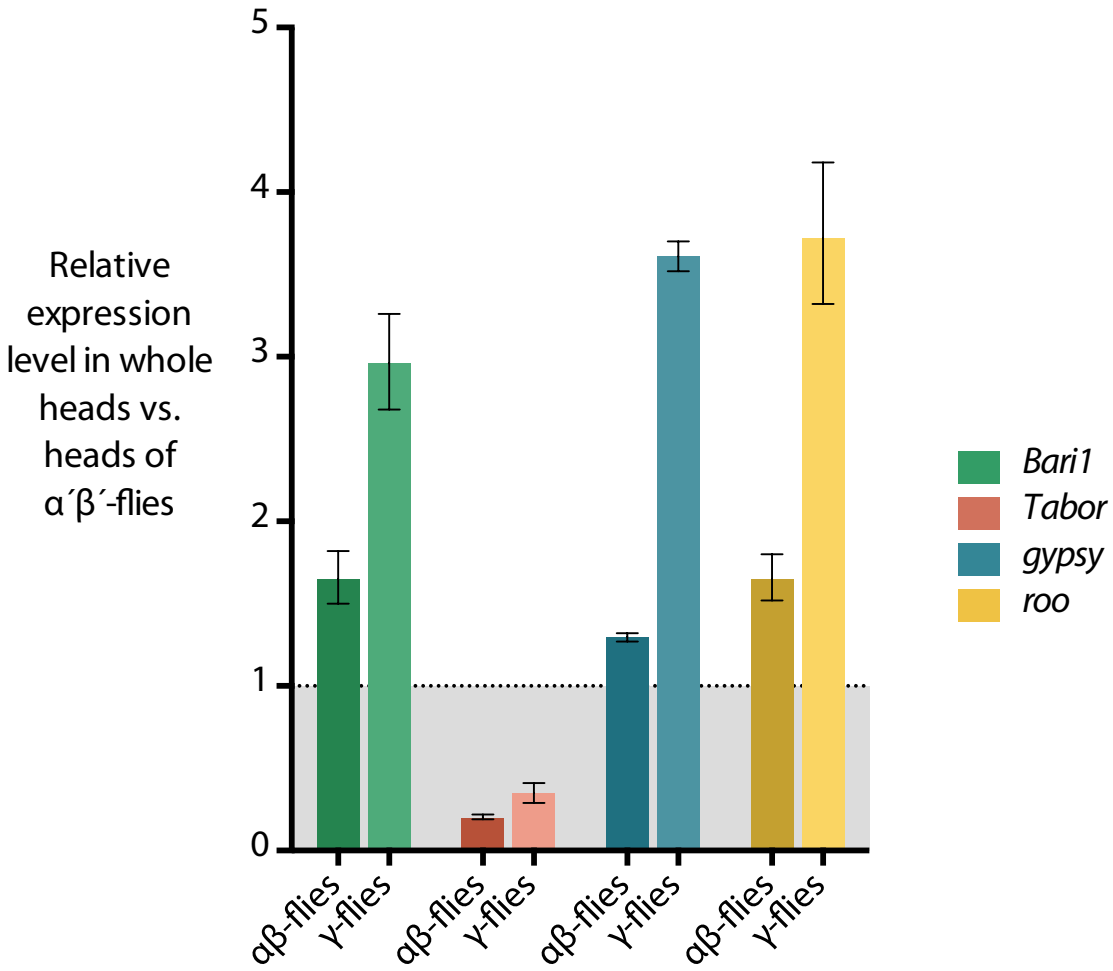


Figure 1. Transposon expression varies with genomic background.

Graph showing the background transposon expression levels in whole fly heads of three different *Drosophila melanogaster* driver lines, used in Perrat et al. Three transposons, *Baril*, *gypsy* and *roo* are expressed at higher levels in the fly strain that was used by Perrat and colleagues to select αβ-KCs and γ-KCs, compared to the flies that were used to collect α′β′-KCs. Conversely, *Tabor* seems to be expressed at significantly higher levels in α'β'-flies. These results highlight the issues that arise with the Perrat method of comparing FAC-sorted cells from different fly strains, as the difference in background transposon expression in all cells could be misinterpreted as expression difference in specific cell populations. Relative expression levels were assessed with Taq-man® assays and RT-qPCR. Ct values were normalised to *GAPDH*. Error bars indicate SEM.

These issues are avoided if tissue samples are collected from the same flies. However, the intersectionally labelled flies used by Perrat et al.[52] are not best suited to such an approach, because the flippase-based strategy is inherently stochastic, with different individual flies having a random set of fluorescently labelled neurons. This means that if GFP positive (αβ-KCs) and GFP negative cells (rest of the brain, RoB) are collected from the same flies, then a randomly unlabelled population of αβ-KCs will contaminate the RoB fraction.

To circumvent these issues with stochastic labelling, I made use of a new *Drosophila* resource, the Janelia split Gal4 project. Split Gal4 is also an intersectional strategy, but does not rely on FRT-recombination. Instead, the expression of each half of the Gal4 transcription factor is driven by a promotor with a distinct expression pattern in the fly brain. Only in the cells in which the two expression patterns overlap is a functional Gal4 reconstituted and Gal4-directed gene expression occurs. The split Gal4 line Mb008b very specifically and consistently labels αβ-KCs. When Mb008b flies are crossed to *UAS-mCherry* flies, the offspring express the fluorophore mCherry in αβ-KCs (Figure 2). Visual inspection of more than 40 fly brains revealed that the expression pattern is consistent between individual flies (data not shown). The offspring of male Mb008b flies and virgin female *UAS-mCherry* flies was used for all subsequent experiments presented in this study.

Figure 2. Mb008b drives *mCherry* expression specifically in αβ-KCs.

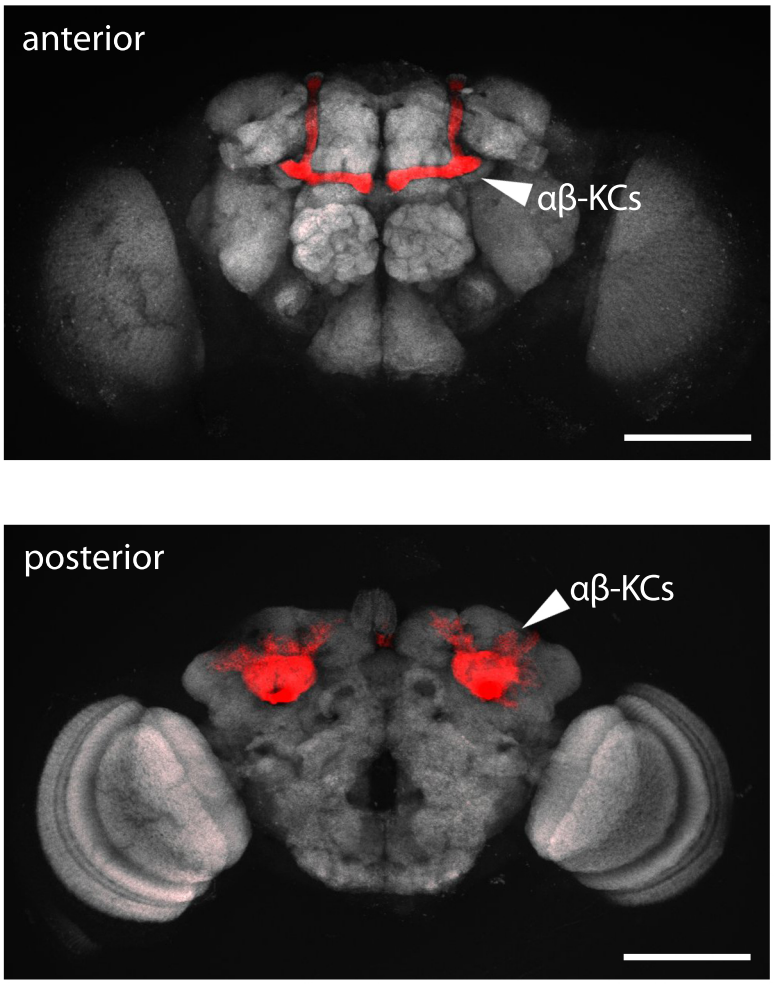


Figure 2. Mb008b drives mCherry expression specifically in αβ-KCs.

Anterior (top) and posterior (bottom) view of single sections from confocal stack images, taken from brains of the split Gal4 line Mb008b, which was used for the majority of experiments in this study. The general neuropil is stained (white) with the anti-bruchpilot antibody NC82 and red indicates the αβ-KC expression pattern of *mCherry*. The scale bar is 100 μm.

To isolate αβ-KCs, and also cells from the rest of the brain, I dissociated the brains dissected from Mb008b/*UAS-mCherry* flies and performed fluorescently labelled cell sorting (FACS). I collected fractions of cells with a high red fluorescence and low green fluorescence signal (Figure 3). This was done to avoid contaminating the mCherry positive fraction with mCherry negative, autofluorescent cells. Autofluorescence is the natural emission of light by a cell. It is independent of the excitation wavelength of the incoming light, and hence an autofluorescent cell will emit a high red- and green fluorescent signal, regardless of whether or not it expresses a fluorophore. Cells from the rest of brain were isolated by selecting a set of mCherry negative FACS cells, with the aim of extracting a similar amount of tissue for each sample. The FACS analysis resulted in a similar number of mCherry positive and negative cells for each sample, suggesting that the split-Gal4 system robustly labels a reproducible and comparable cell population in each fly. RNA was subsequently extracted from the separate cell fractions to test the expression levels of transposons and other genes.

Figure 3. FACS separates αβ-KCs from the rest of the brain.

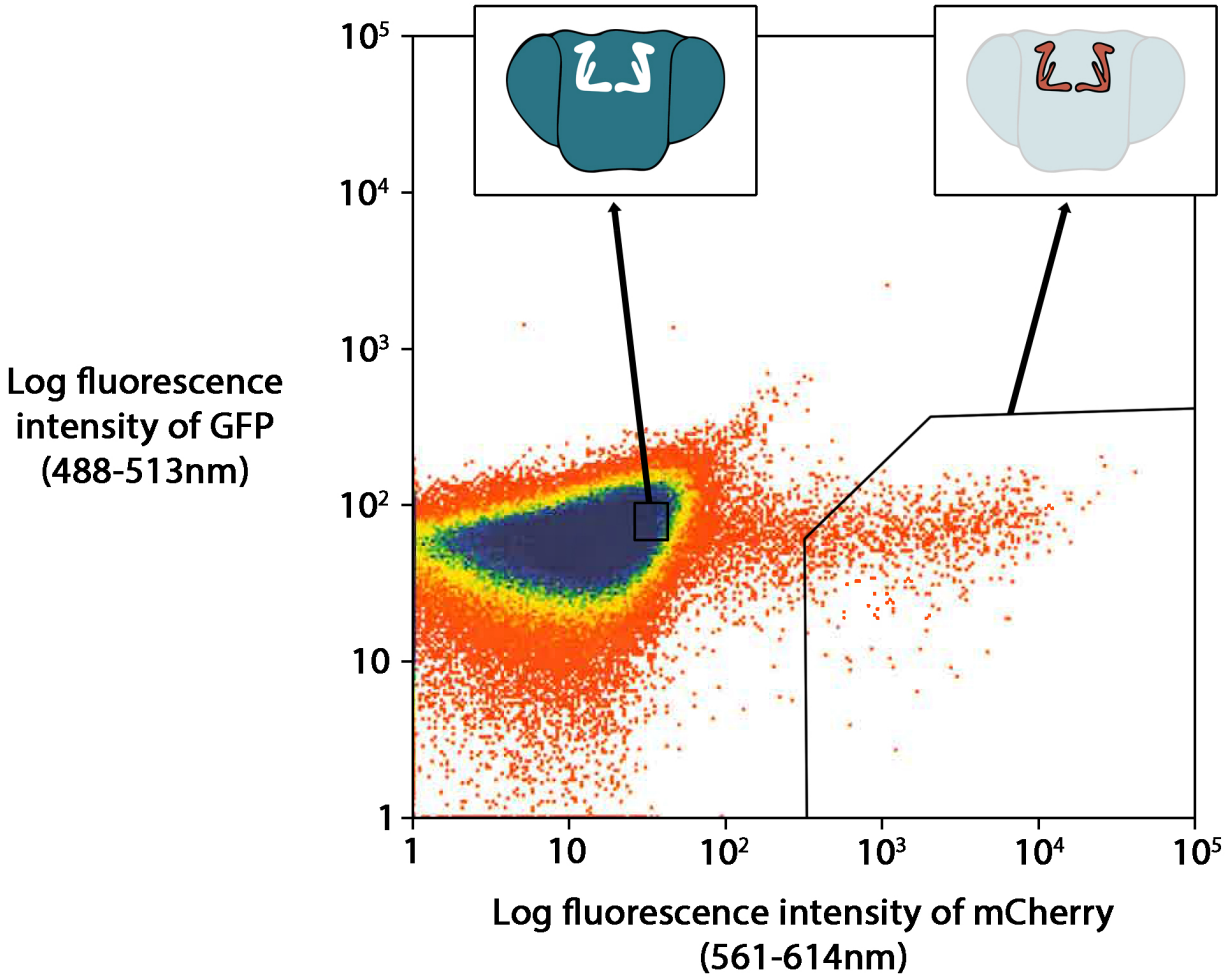


Figure 3. FACS separates αβ-KCs from the rest of the brain.

Sample plot of a FACS run on a single fly brain shows the selection strategy that was used to sort *mCherry*-positive cells and *mCherry*-negative cells from dissociated brain tissue. Log fluorescence values for *mCherry* (561-614nm) are plotted against those for GFP (488-513nm). Signals that arise because of cellular autofluorescence occur at similar intensities in both channels, whilst *mCherry* specific fluorescence is restricted to the channel plotted on the x-axis. Gates for sorting were hand-drawn, with the aim of selecting a pure proportion of *mCherry* cells (right inset) and a second population of cells of similar size, which were *mCherry* negative (left inset). 811 *mCherry* positive, and 8321 *mCherry* negative cells were collected from the sample shown here. Single cells are represented as red points and areas of high-density are coloured blue. The same gates were used for all samples presented in this study.

2.1.2. Quantification of mRNA levels is linear across a wide range of input concentrations

Gene expression levels were again assessed using RT-qPCR. Since the total amount of RNA extracted from FAC-sorted cells is very low, the mRNA needs to be amplified before testing. Multiplexed pre-amplification selectively amplifies target and reference genes. This method includes the combined pre-amplification of multiple genes (the target genes and house-keeping reference genes) and is therefore incompatible with Taq-man® assays. Instead, I used the more customizable UPL® (Universal Probe Library, see methods) system, which required designing primers and choosing specific probes for each gene and transposon that was tested.

When comparing RNA samples, it is essential that the qPCR signal is linear and stable across a wide range of concentrations. I therefore verified the linearity of my assay by measuring the expression levels of two genes, *Ago3*, which encodes for a member of the Argonaute protein family, and *SdHA*, which encodes for succinate dehydrogenase complex flavoprotein subunit A, normalised to the housekeeping gene glyceraldehyde-3-phosphate dehydrogenase gene *GAPDH* in a dilution series made from total RNA extracts from whole flies. A dilution range from 1:1 to 1:1,000,000 revealed that the amplification strategy was reliably linear across a wide range of mRNA levels (Figure 4). This shows that multiplexed RNA

amplification is a reliable method to quantify and compare gene expression levels when the input mRNA amounts are low.

Figure 4. mRNA levels are reliably detected at total RNA amounts ranging from 0.1µg to 0.1pg.

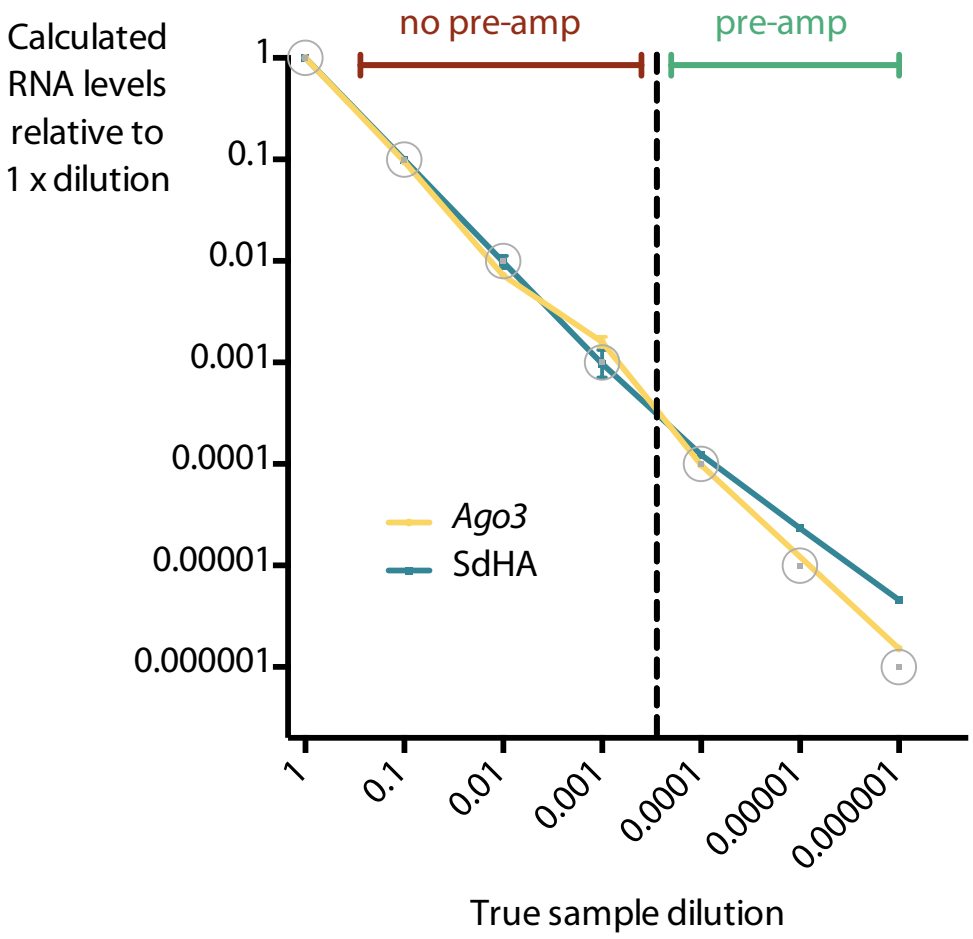


Figure 4. mRNA levels are reliably detected at total RNA amounts ranging from 0.1 µg to 0.1 pg.

Graph showing the linearity of the pre-amplification method used in this study. The calculated RNA levels of *Ago3* and *SdHA* in six different dilutions of total RNA, relative to the initial sample (1 equates to 100ng of total RNA per qPCR well) are plotted. Grey circles indicate the predicted values if the process is totally linear. The first four concentrations were measured with a standard RT-qPCR protocol (no pre-amplification) and the final three concentrations were pre-amplified (see Methods for details). Relative RNA values were calculated with the Ct values of technical triplicates. For pre-amplified samples, the number of pre amplification cycles was added to the Ct values. *GAPDH* was used as a reference gene.

2.1.3. FACS results in pure and homogeneous cell samples

To verify that the FACS approach resulted in pure fractions of mCherry positive and negative cells, I tested both fractions for the expression levels of *mCherry*, and of *Fas2*, which encodes for fasciclin 2, a cell adhesion molecule that has previously been found to be over-expressed in αβ-KCs by immunostaining[81]. The levels of *mCherry* expression were over a thousand times higher in the mCherry-positive fraction, when compared to the mCherry-negative cells. In addition, *Fas2* mRNA levels were 5 times higher in αβ-KCs versus the rest of the brain (Figure 5). These results show that FACS efficiently separates fluorescently labelled cells from the rest of the brain.

Figure 5. mCherry positive FACS fractions are highly enriched for αβ-KCs.

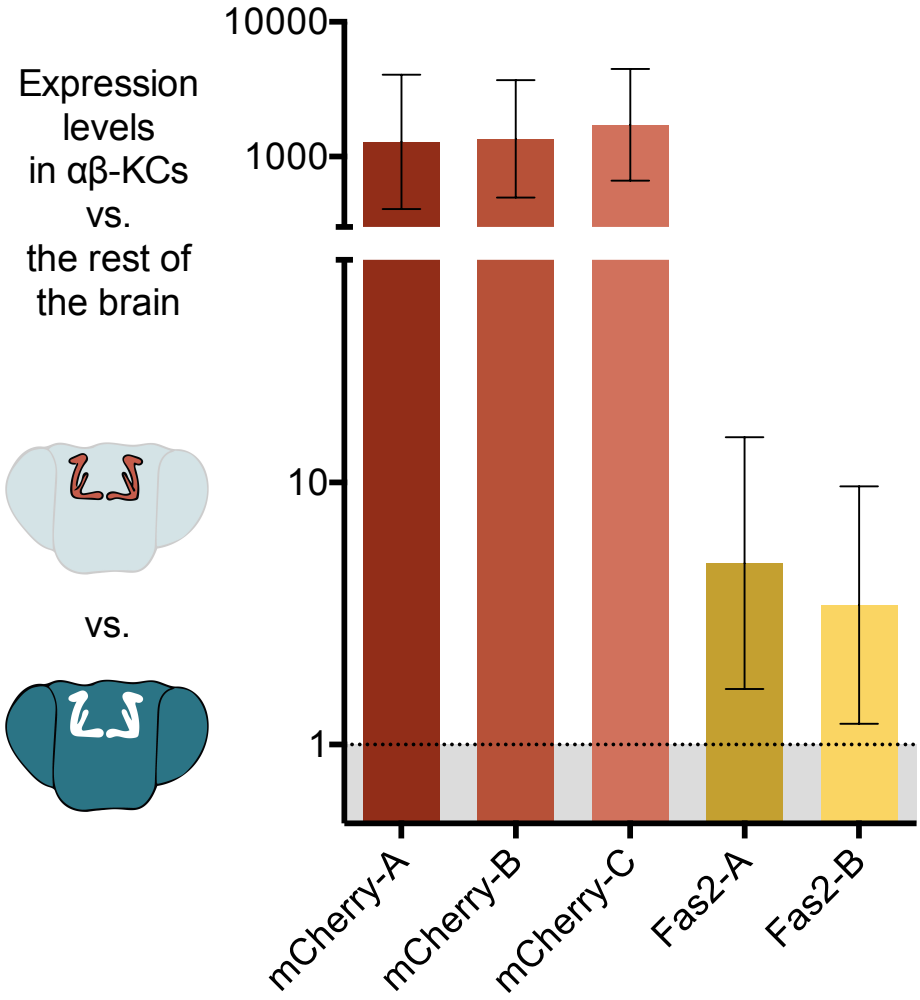


Figure 5. mCherry positive FACS fractions are highly enriched for αβ-KCs.

Graph showing relative gene expression levels in FAC-sorted cells. *mCherry* expression levels are highly enriched in *mCherry* positive fraction of cells. Three independent probes were used to test for *mCherry* expression. *Fas2* is also enriched in the *mCherry* positive cell fraction. Error bars represent combined SEM of 2 separate experimental runs, with 50 flies and 3 technical replicates per run.

I also tested whether the cells collected from the rest of the brain represented a sufficiently randomly sampled population of cells without any obvious bias in gene expression. Two separate populations of cells that were both negative for the fluorophore were collected from four independent samples and tested for levels of expression of the gene *Spock*, which encodes for an ATPase, and the LINE-like retrotransposon *Ivk* (again normalised to *GAPDH*). The analysis showed that the expression levels of both genes did not vary substantially in the two fractions or between multiple biological replicates, suggesting that the strategy to collect *mCherry* negative cells was a reliable way to obtain a sample that was representative of the rest of the brain (Supplementary Figure 1).

2.1.4. Transposons are upregulated in αβ-KCs

I next reassessed the transposon expression results by Perrat et al.[52]. Primers and probes were designed for five of the transposons that were reported to be upregulated in αβ-KCs. mRNA levels were measured with RT-qPCR, performed on groups of 50 flies (25 female and 25 male flies), with four entirely independent repeats. This rigour was necessary, because RT-qPCR experiments are prone to deliver erroneous and irreproducible results if the number of biological replicates are too low[82].

In agreement with Perrat et al.[52], I found that *Ivk*, and the Long Terminal Repeat (LTR) retrotransposon *Tabor* are upregulated in αβ-KCs (Figure 6). However, my analysis revealed a less dramatic increase in expression of a 3 instead of a 15-fold increase for *Ivk*, and a 2.5 instead of a 5-fold increase for *Tabor*. Surprisingly, I did not find any significant differences in the expression levels of the LTR retrotransposons *gypsy* (17-fold increase in Perrat et al.) and *blood* (15-fold increase), or of the LINE-like retrotransposon *Doc3* (15-fold increase). The lack of *gypsy* upregulation seemed particularly noteworthy, because *gypsy* was previously reported to be a key driver of age-dependent memory decline due to the suggested accumulation of somatic insertions in KCs[53].

Figure 6. Transposons *Ivk* and *Tabor* are upregulated in αβ-KCs.

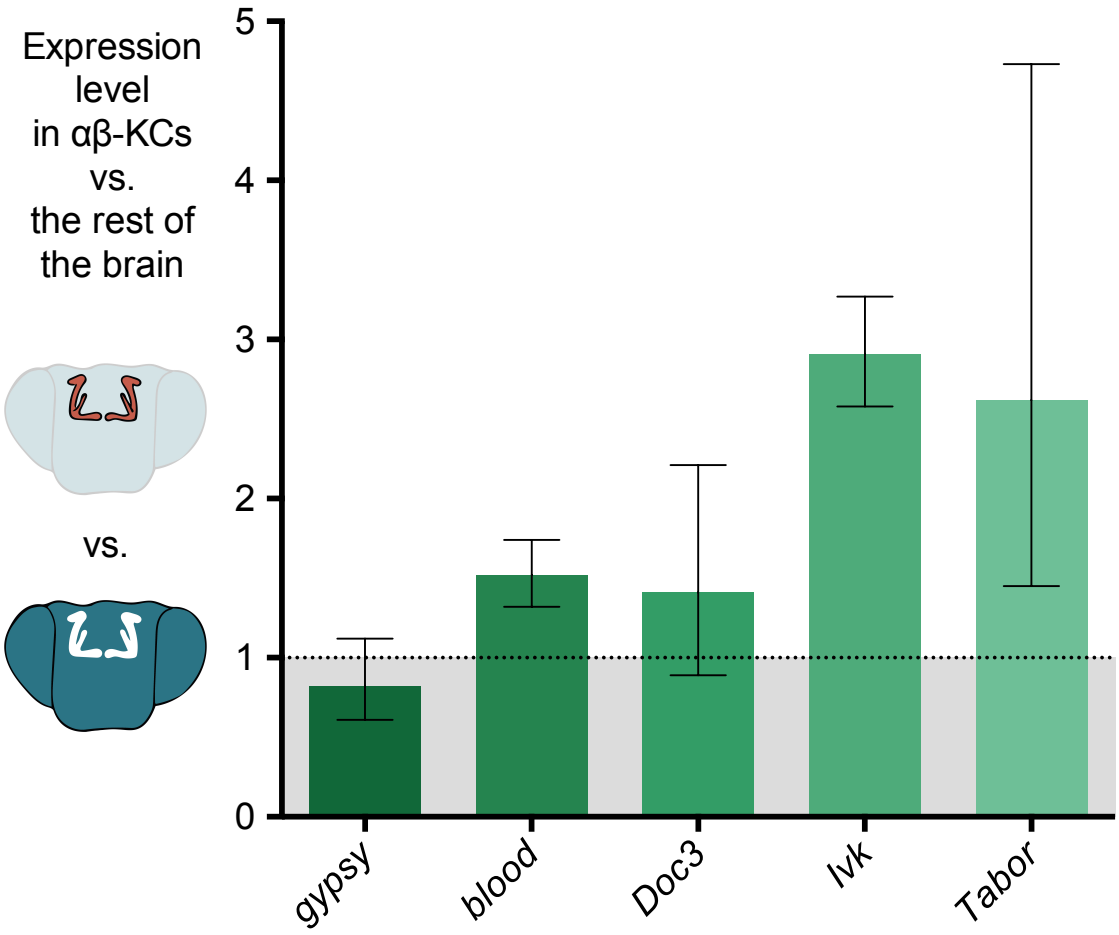


Figure 6. Transposons* Ivk *and* Tabor *are upregulated in αβ-KCs.

Graph showing the relative expression levels of 5 selected transposons in αβ-KCs, compared to the expression level in the rest of the brain. *Ivk* and *Tabor* expression is increased 3-fold in αβ-KCs. Error bars represent combined SEM of 4 separate experimental runs, with 50 flies and 3 technical replicates per run.

2.2 Analysis of somatic transposon insertions αβ-KCs

I next tested whether increased expression of transposons in αβ-KCs had an impact on the frequency of somatic transposon insertion events in these cells.

2.2.1 Multiple displacement amplification results in high coverage sequence data

Perrat et al.[52] studied the frequency and prevalence of somatic transposon insertions by dissecting the brains from 50-60 flies and FAC-sorting them to collect αβ-KCs. gDNA was then extracted from those cells, randomly sheared and a sequencing library was prepared from the DNA fragments. These fragments were then amplified by PCR before being sequenced. This approach unavoidably results in the formation of PCR duplicates, exponentially amplified single gDNA fragments that result in hundreds of identical copies. These duplicates can be easily identified because random DNA fragments map to unique locations in the *Drosophila melanogaster* reference genome, whereas PCR duplicates start and end with the exact same nucleotide sequence. To estimate the number of duplicates in the sequencing data published by Perrat et al., I filtered their original data and found that approximately 90% of the reads that could be mapped to the reference genome were PCR duplicates (Figure 7, red trace). Hence, the average coverage of the prior study without amplification artefacts removed was approximately 2.5x (i.e. every genomic locus was, on average, sampled 2.5 times).

To avoid the formation of PCR duplicates, I amplified the gDNA using a different strategy. Multiple displacement amplification (MDA) uses a bacterial DNA polymerase that amplifies gDNA in an unbiased and linear fashion. This polymerase can displace the DNA strand that is annealed to the template, and therefore the amplification process does not require a cyclic procedure (hence there is no exponential enrichment of certain stretches of DNA). Importantly, this method does not produce PCR duplicates.

Whole-genome sequencing (WGS) analysis of material amplified with MDA resulted in paired-end reads, which mapped to the reference genome with a coverage of 25x for each sample (Table 1, first line). Moreover, a visual inspection of the alignments to the reference genome suggested that the coverage was much more evenly distributed than the alignment of the Perrat et al. data. In addition, filtering the data for PCR amplification artefacts showed that the majority of reads were unique gDNA fragments (Figure 7, blue trace). In summary, MDA is the more reliable and efficient method for DNA amplification and was therefore used for the rest of the experiments presented in this study.

Figure 7. Removal of PCR amplification artefacts reveals low genome coverage in the Perrat et al. sequencing data.

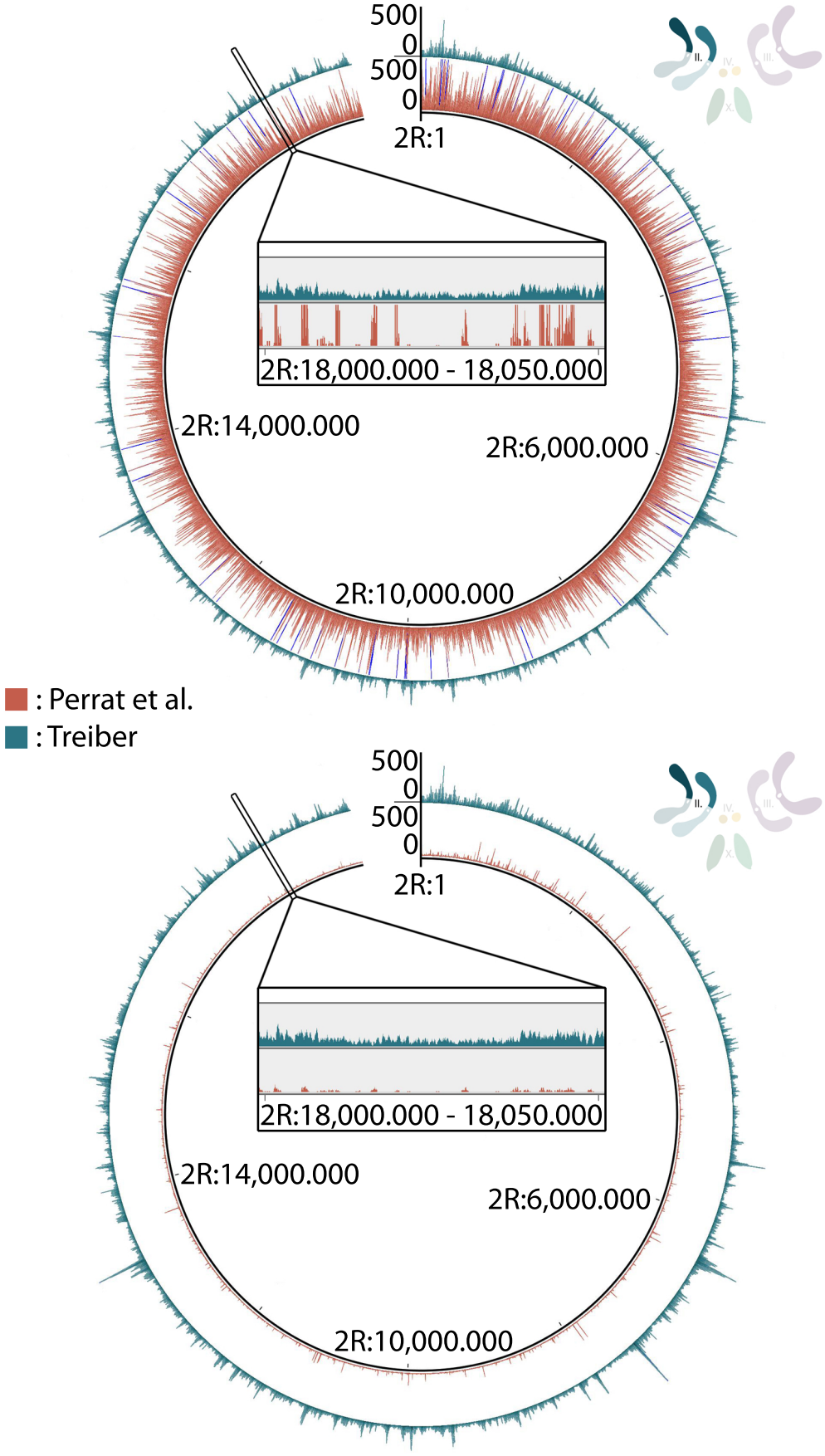


Figure 7. Removal of PCR amplification artefacts reveals low genome coverage of sequencing reads in the Perrat et al. data.

The sequencing coverage on chromosome 2R of data previously published by Perrat et al. (red trace) is compared with sequencing data from groups of flies, produced for the current study (blue trace) on a circular plot. The top panel shows the raw output after the alignment of reads and the bottom panel shows the coverage after removal of PCR duplicates (i.e. reads which share the exact same starting position, see methods and discussion for details).

2.2.2. Putative transposon insertions are detected using TEMP

The TEMP algorithm was used to test the WGS data sets for the presence of transposon insertions. TEMP uses paired-end sequencing reads to identify transposon insertions in gDNA that are absent in the reference genome[67]. Firstly, TEMP filters all discordant read pairs to isolate the pairs where one read uniquely maps to the reference genome (the anchor read) and the other to one of the known *Drosophila melanogaster* transposon reference sequences. Each of these hits counts as one potential piece of evidence for a putative transposon insertion. Germline transposon insertions should be present in all gDNA molecules of a sample and will therefore result in a high number of indicative read pairs. However, rare transposon insertions that occurred in somatic tissue could potentially only be represented by a single discordant read pair.

For my analysis, I catalogued all transposon insertions that were detected in αβ-KCs (2,978) and in the sample from the rest of the brain (4445). If transposon insertion events occurred somatically in αβ-KCs, then they should be absent in the sample of the rest of the brain. Therefore, I intersected the TEMP hits from both samples and selected those that were present in αβ-KCs, and absent in the rest of the brain. This left 2,306 hits, which at first glance suggested a much higher rate of somatic transposon insertions than previously reported[52] (Table 1, first line).

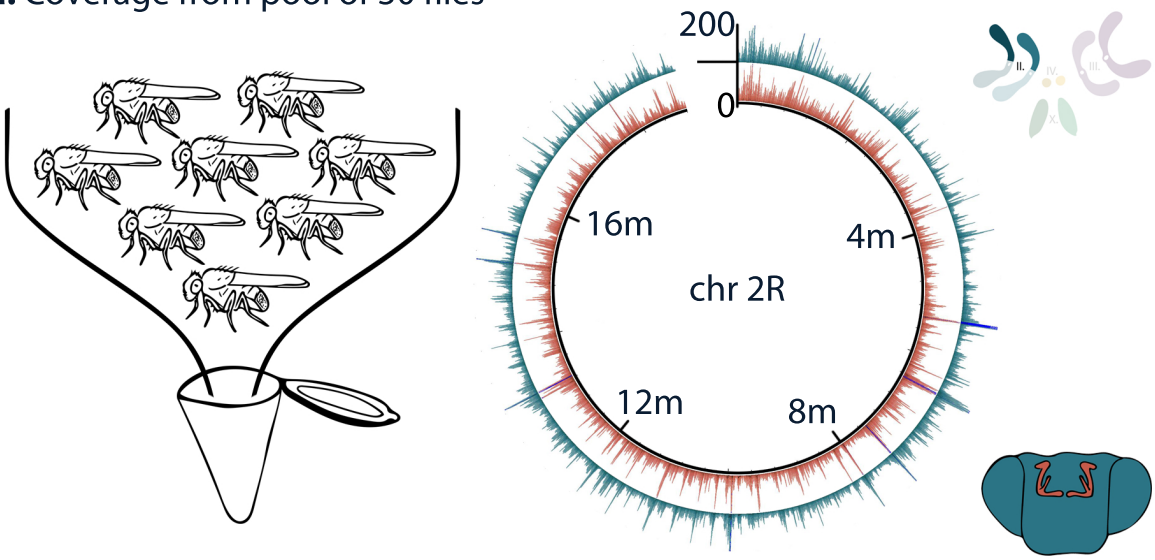
2.2.3. Analyses on groups of flies have limited power

In retrospect we now consider that the analysis of transposon insertions in gDNA prepared from groups of flies inherently lacks the power to reliably detect rare somatic insertion events. To illustrate this, imagine that there are 800 labelled αβ-KCs in each fly brain (see Figure 2), with each genetic locus occurring twice in each cell. This means that based on the experimental approach with groups of 30 flies, each sample contains 48,000 copies of each genetic locus. At a sequencing depth of 25x, in theory only 1 in 1,920 events that occurred in one locus of one cell would be detected, and 1 in 240 events at loci with a local coverage of 200x (Figure 8a). With these numbers in mind the chances of finding a rare *de-novo* event are extremely low using the approach on groups of flies.

Another caveat of using groups of flies is the possibility that low frequency germline transposon insertions in the population would be indistinguishable from genuine *de-novo* insertion events. For example, a germline insertion that was present in all cells of one of the 30 flies, and that happened to be picked up in the αβ-KC sample, and not in the flow-through (a scenario that seems likely, given that every locus is only sampled 12 times on average, and 1 in 60 DNA fragments would harbour the germline insertion), would fall into the exact same category as a true somatic insertion event, being represented by one read pair in αβ-KCs, and by none in the rest of the brain.

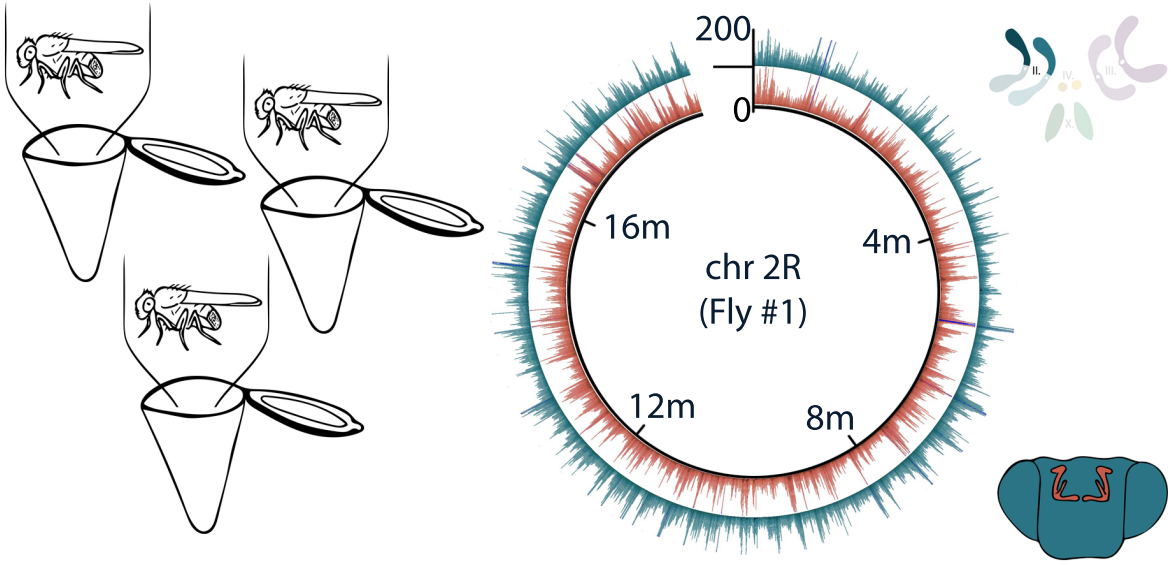
Figure 8. The single-fly approach significantly increases power to detect rare somatic transposon insertion events.

a. Coverage from pool of 30 flies



Theoretical power to detect unique event: ~ 1 in 2000

b. Coverage from single flies



Theoretical power to detect unique event: ~ 1 in 40

Figure 8. The single-fly approach significantly increases power to detect rare somatic transposon insertion events.

a. Schematic showing the original experimental approach, which is comparable to that of previously published studies. Groups of flies are collected, and fluorescently labelled cells are separated from the rest of the brains. As the gDNA from multiple animals is pooled, the theoretical power to detect somatic events is low, and the analysis does not permit a distinction between rare germline insertion event in one of the animals and true somatic insertion events in the cells which are tested. The circular plot shows the gDNA sequencing coverage of *mCherry* positive αβ-KCs (red trace) and *mCherry* negative cells from the rest of the brain (blue trace), on chromosome 2R. Sequencing read alignments on other regions of the gDNA exhibit a similar coverage. ***b.*** The new approach, which has been developed for this study is based on the analysis of single flies. This method increases the theoretical power to detect rare insertion events, and makes it possible to distinguish between germline- and somatic insertion events. The circular plot shows the coverage of both FACS fractions from a single fly brain.

2.2.4. Single-fly analysis substantially increases the power to detect rare somatic insertions

Given the apparent weakness of the 'goup of flies' approach, I developed a method to analyse the genomes of αβ-KCs from single flies with a sequencing coverage of 40x. This methodological advance means that our theoretical discovery rate increased from 1:2,000 to 1:40, and to 1:8 for those loci that have a coverage of 200x (Figure 8b). Even more importantly, this new approach allows us to clearly distinguish germline events with a rare population frequency from true somatic insertion events, as TEMP analyses can be compared within the same individual fly.

The initial raw TEMP output from these experiments suggested that the total number of putative somatic transposon insertion events was significantly higher in single flies than in groups of flies. On average, the TEMP algorithm identified a total of 21,743 hits in αβ-KCs per fly (N=6, SD= 3,026, see Table 1). To distinguish between germline and *de-novo* somatic transposon insertions, I separately intersected the TEMP hits from αβ-KCs with those from the rest of the brain for each fly. This intersectional approach resulted in an average of 10,881 putative αβ-KC specific insertions per fly (N=6, SD=1,960, see Table 1).

2.2.5. New transposon insertions appear to hover around germline insertions

Mapping positive insertion events in the single fly data revealed a recurring and very particular pattern where *de-novo* events appear to cluster around pre-existing germline locations of transposons (Figure 9). These "events" clearly accumulate approximately 10kb up- and down-stream of the germline location of a transposon with the same sequence, a phenomenon that I term 'hovering'. In the example shown in Figure 9, a germline insertion of the transposon *Doc* was found to be inserted in the intronic region of the *turtle* gene. *turtle* encodes for a protein that is known to be involved in axon guidance and it is conceivable that a transposon inserted in a *turtle* exon would impair its normal function or expression pattern. The phenomenon of transposon hovering seemed to occur both in αβ-KC samples (Figure 9, dark green traces), as well as in samples from the rest of the brain (Figure 9, light green traces).

Figure 9. Transposons appear to hover around germline insertion sites.

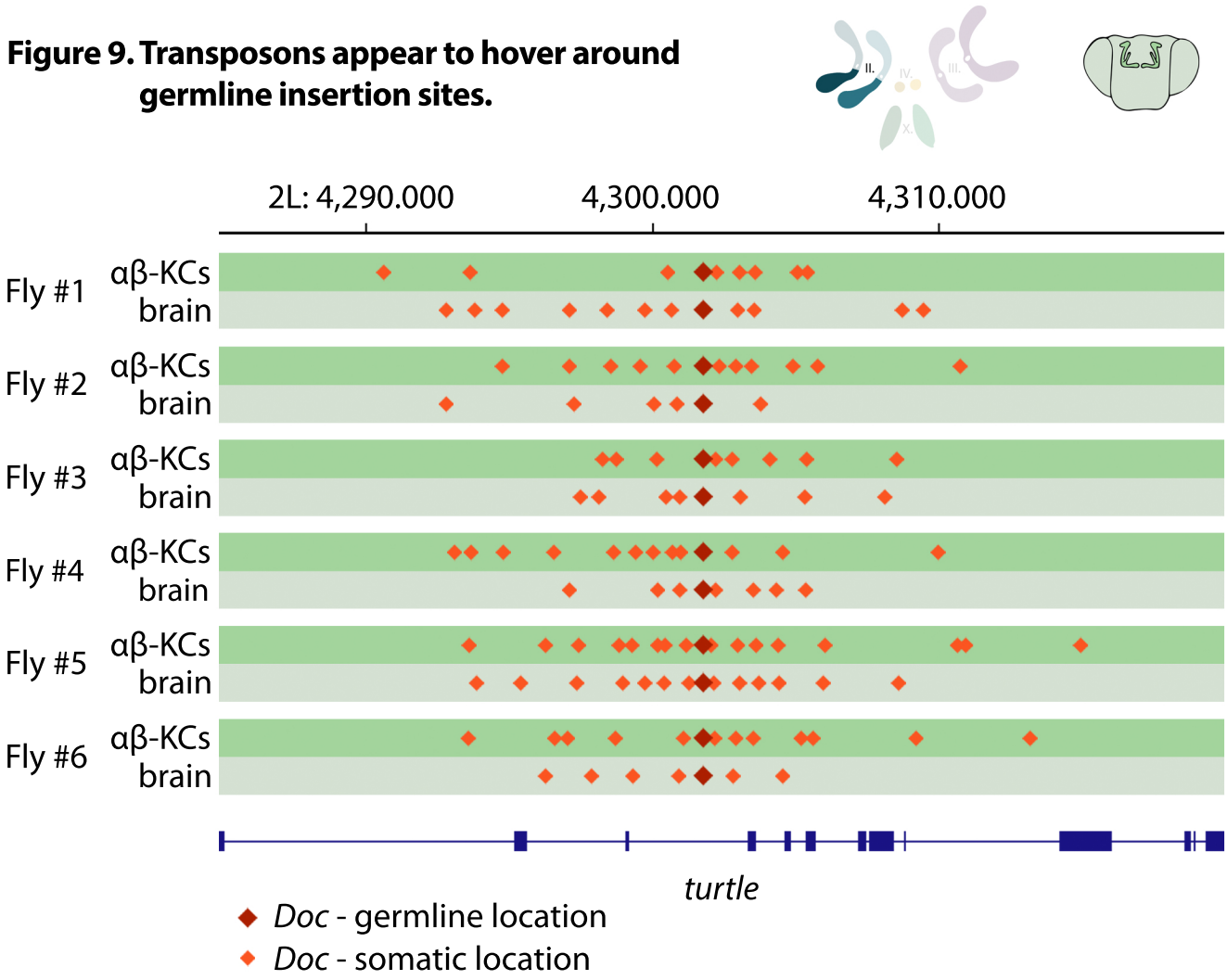


Figure 9. Transposons appear to hover around germline insertion sites.

Graph showing a representative example of a germline transposon insertion that was found in each of the 6 individual flies and that is absent in the *Drosophila melanogaster* reference genome (Release 5.57). Using criteria that are currently employed to identify germline transposon insertion events, putative somatic insertion events of the same transposon are identified at loci, which are approximately 10kb up- and downstream of a germline insertion site of the same transposon type. Dark red diamonds represent a germline insertion of the transposon *Doc*, which was found in each of the 12 samples (αβ-KCs and the rest of the brain), and light red diamonds represent putative somatic *Doc* insertions. The genomic location of the *turtle* gene is shown in blue. Blue boxes indicate exons and blue lines intronic regions of the gene.

2.2.6. Single animal sequencing experiments reveal amplification bias that is conserved across all samples

To test whether the putative TEMP hits were meaningful, I first checked whether our sequencing data is an unbiased and evenly distributed representation of all genetic loci. In theory, every locus should be represented by the same number of sequencing reads. My analysis revealed that the sequencing coverage was not equally distributed across the reference genome. Instead, sequencing reads were clustered across the genome. Surpsigingly, this pattern was conserved across all samples that were analysed in this study (Figure 10). This finding was unexpected because the input material for the WGS reaction was gDNA and not complementary DNA (cDNA), which would be expected to contain over-representations of certain genome regions.

Figure 10. Patchy sequencing coverages results from regionally biased DNA amplification.

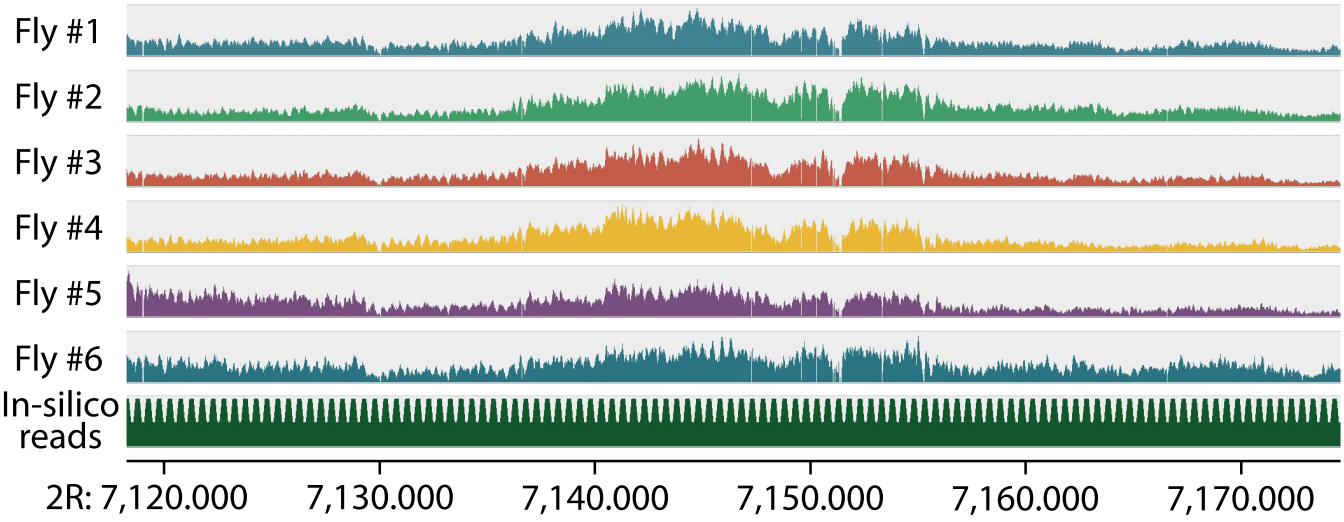


Figure 10. Patchy sequencing coverage results from regionally biased DNA amplification.

Comparison of 6 different DNA sequencing traces showing a stereotypical amplification pattern. Each of the top 6 lines shows the local coverage of sequencing reads at a location on chromosome 2R, which is representative of the majority of the genome. Each sample was amplified using MDA (see Methods). There is an obvious similarity in the distribution of sequencing reads, suggesting that the pattern represents the amplification bias that is introduced by the MDA polymerase. The bottom trace shows the alignment of a set of read pairs, which were created by *in-silico* fragmenting the reference genome sequence, creating reads without any regional amplification bias. The equal distribution of these reads shows that the stereotypic pattern of biological samples is due to true biological amplification dynamics, rather than variations in the mapping efficiency.

This pattern of coverage could either arise because the alignment algorithm that was used for this study (Burrows-Wheeler Alignment[83], BWA, see methods) maps DNA reads with a particular sequence bias, or it could result from non-random variations in the amplification of gDNA. I investigated the source by creating a set of read pairs that were perfectly distributed across the entire reference genome. This idealised *in-silico* data set did not exhibit the same local bias as the biological samples, suggesting that it is not an alignment artefact. Instead, these findings suggest that the uneven sequencing coverage results from intrinsic regional biases of the DNA amplification process. Most importantly, this observation suggests that the quantification of gDNA insertions and deletions at different loci is substantially confounded by regional amplification bias.

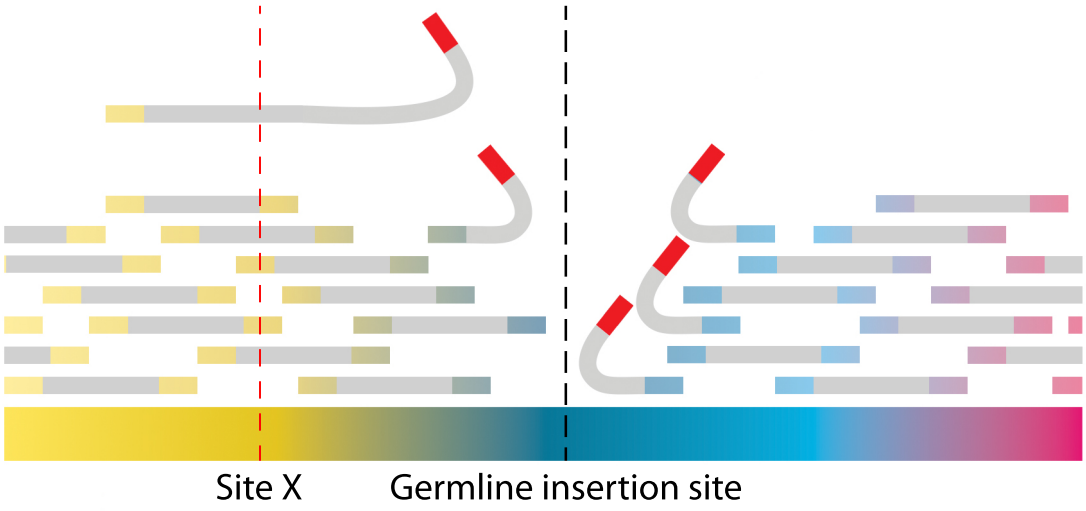
2.2.7. Some sequencing read pairs map further apart from each other than the average DNA fragment size would suggest

The TEMP algorithm infers the location of a putative somatic transposon insertion from the position of the read that uniquely maps onto the genome (the anchor read) and the assumed fragment size of the sequenced DNA fragment. This means that if a DNA fragment that spans the edge of a germline transposon insertion was longer than the assumed fragment length, then the TEMP algorithm would mistakenly categorise this read pair as evidence for a somatic transposon insertion near the germline site (Figure 11a). For example, assume that the average DNA fragment size is 500 basepairs and the divergent fragment is

5000 basepairs. If one end of this large fragment maps to the edge of a transposon, then the other end would map to a region that is located 5000 nucleotides up- or downstream of the germline insertion site. TEMP assumes a fragment size of 500 basepairs and will therefore identify this read pair as evidence for a separate, somatic transposon insertion 4500 nucleotides up- or downstream of the genuine germline insertion site.

Figure 11. Insertion artefacts could arise from variation of insert sizes or from chimeric DNA fragments.

a. Longer DNA fragments result in additional TEMP hits.



b. DNA chimera result in additional TEMP hits.

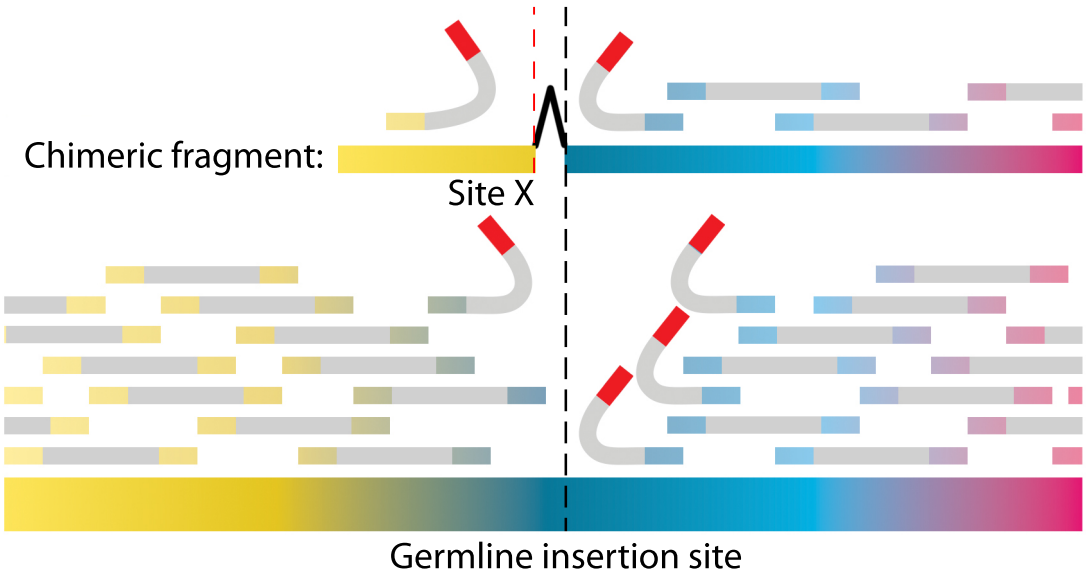


Figure 11. Insertion artefacts could arise from variation of insert sizes or from chimeric DNA fragments.

a. Schematic showing how DNA fragments, which diverge from the average insert size and which have one read maping onto a transposon at a germline insertion site, could potentially be misinterpreted as pieces of evidence for rare somatic insertion events (for example here at site X). ***b.*** Schematic showing how DNA chimeras, which were formed before DNA fragmentation and before the creation of paired-end reads, could potentially lead to read pairs that would be misinterpreted as evidence for a rare somatic insertion events (at site X).

I tested whether variation in DNA fragment size was an issue by looking more closely at the rest of my sequencing data. The sizes of DNA fragments cannot be directly measured from traditional WGS experiments. However, if a set of paired-end reads is aligned to the reference genome, then the distance between the two read pairs on the reference genome can be used to estimate the size of the corresponding DNA fragment. I therefore plotted the predicted fragment sizes of all mapped reads from each sample (Figure 12). Strikingly, the predicted fragment sizes of a substantial portion of DNA fragments diverge from the average fragment size of 500 basepairs. Most importantly, the distribution of fragment sizes is consistent with the spread of putative somatic insertion events that appear to hover around germline transposon insertions, because the size range is similar to the range of hovering events (see Figure 9). We therefore suspect that rather than being evidence of local transposon hopping, hovering is an artefact of the variable DNA fragment size in the WGS library.

Figure 12. Some fragments appear to be up to 10,000 basepairs long.

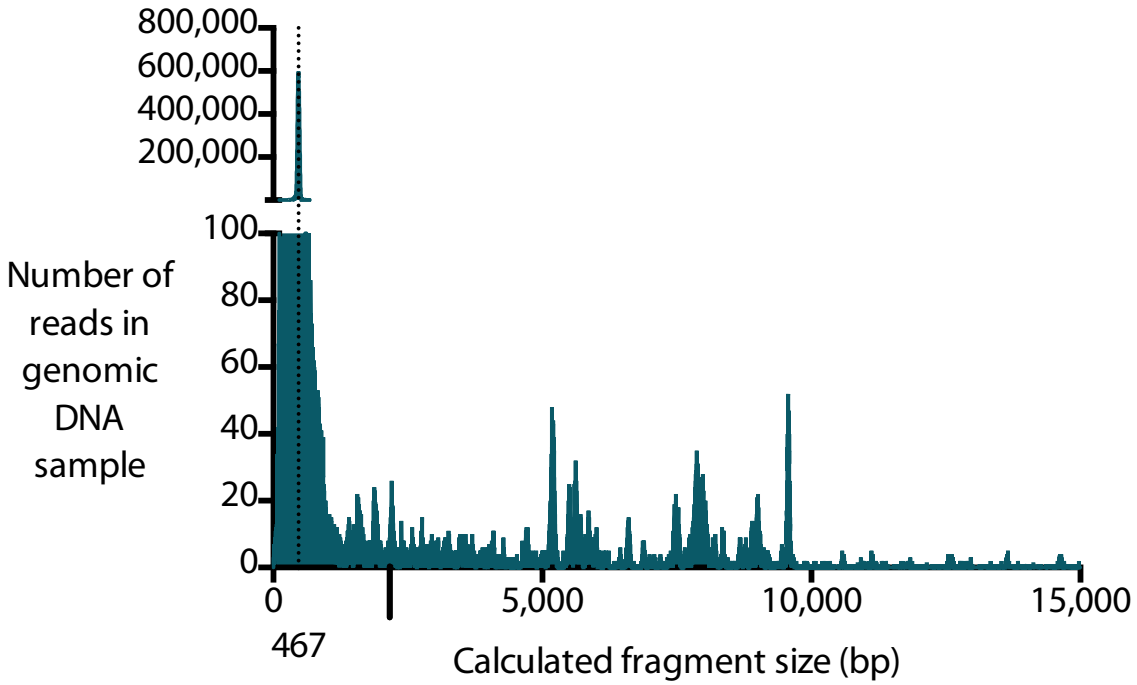


Figure 12. Some fragments appear to be up to 10,000 basepairs long.

An analysis of the calculated fragment sizes, based on the results of BWA read mappings, shows that some fragments are substantially larger than the theoretical fragment size of 500 basepairs. The median size is 467 basepairs. It should be noted that this graph is cropped at 15,000 basepairs and that some fragments appear to have a size up to the length of the longest chromosome in the reference genome sequence. Values of one representative sample are shown.

2.3. Development of MRTemp, a novel approach for transposon mapping

Given the problems associated with classical paired-end read sequencing and TEMP mapping, I developed MRTemp, a new method designed to eradicate DNA fragment size variation. I first generated 350 basepair DNA fragments from amplified gDNA from αβ-KCs and from the rest of the brain from single flies. Next, 250 nucleotides were sequenced from each end of these fragments, so that the entire sequence of the DNA fragment was retrieved. Having the complete sequence of the fragment allows control over the fragment length (Figure 13).

Figure 13. MRTemp reveals break-point sequence and controls for variation of fragment sizes.

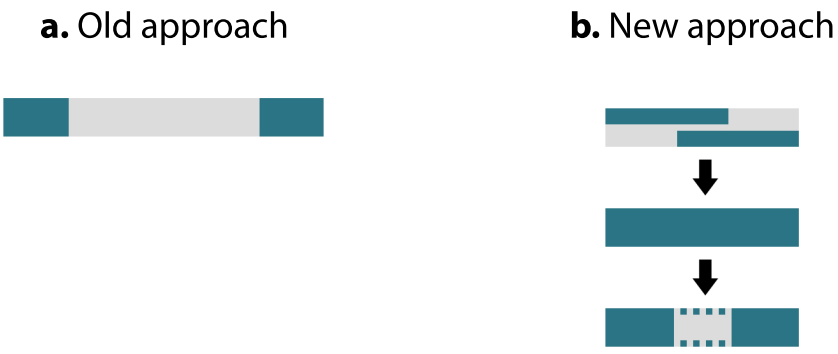


Figure 13. MRTemp reveals breakpoint sequence and controls for variation of fragment sizes.

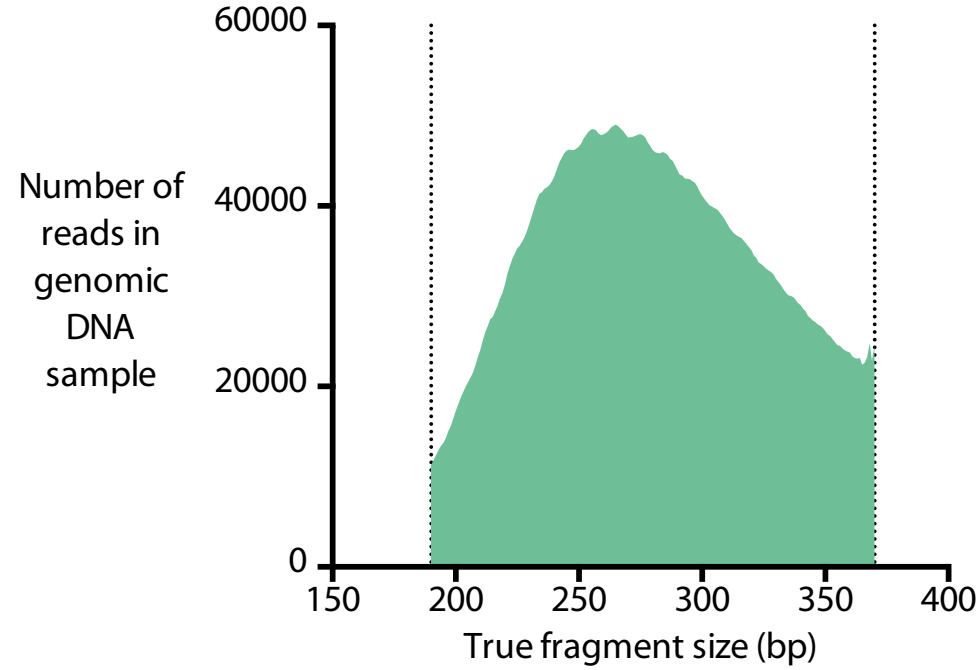
a. Schematic showing the previously published Perrat approach. gDNA was fragmented into 500 basepair long fragments, and then 100 nucleotide long reads at each end of the fragment were sequenced. b. MRTemp, which was developed for this study, is based on the creation of shorter DNA fragments (350 basepairs), followed by long read sequencing from both ends of the fragments, with the aim of obtaining overlapping read pairs. The overlapping reads are then merged, and paired end reads are created based on these merged reads. This new approach removes variation of fragment size and also reveals the exact breakpoint sequence of potential transposon insertion events.

In more detail, each 250 nucleotide read pair is merged to create ~350 basepair long single fragments. A minimum overlap of 10 nucleotides is required to merge each pair. In addition to providing full control over the length of all fragments that are used for the downstream analyses, this also provides insight into the genuine DNA fragment lengths, which in our case vary between 190 and 370 nucleotides (Figure 14a). Next, paired-end *in-silico* reads with a read length of 100 nucleotides and fragment sizes of exactly 250 basepairs are created from the longer merged reads (>249bp). Since all of these read pairs are based on wholly sequenced DNA fragments, the size of each fragment is fixed, and cannot vary. These *in-silico* read pairs are then aligned to a reference genome and discordant read pairs are screened with similar criteria to those in the classic TEMP approach. Once a putative somatic transposon insertion is detected, MRTemp refers back to the 250 basepair long merged DNA fragment, in order to retrieve the precise break point of the insertion.

MRTemp therefore provides three crucial advantages to previous approaches. Firstly, it provides full control over the fragment size. Secondly, it reveals the precise sequence around the breakpoint of potential transposon insertions, which allows unambiguous determination of whether two or more events originate from the same insertion. Lastly, it also reveals the breakpoint within the transposon.

Figure 14. MRTemp shows that the apparent variation of fragment sizes occurs due to the existence of chimeric DNA fragments.

a. True fragment sizes



b. Fragment sizes based on BWA mapping

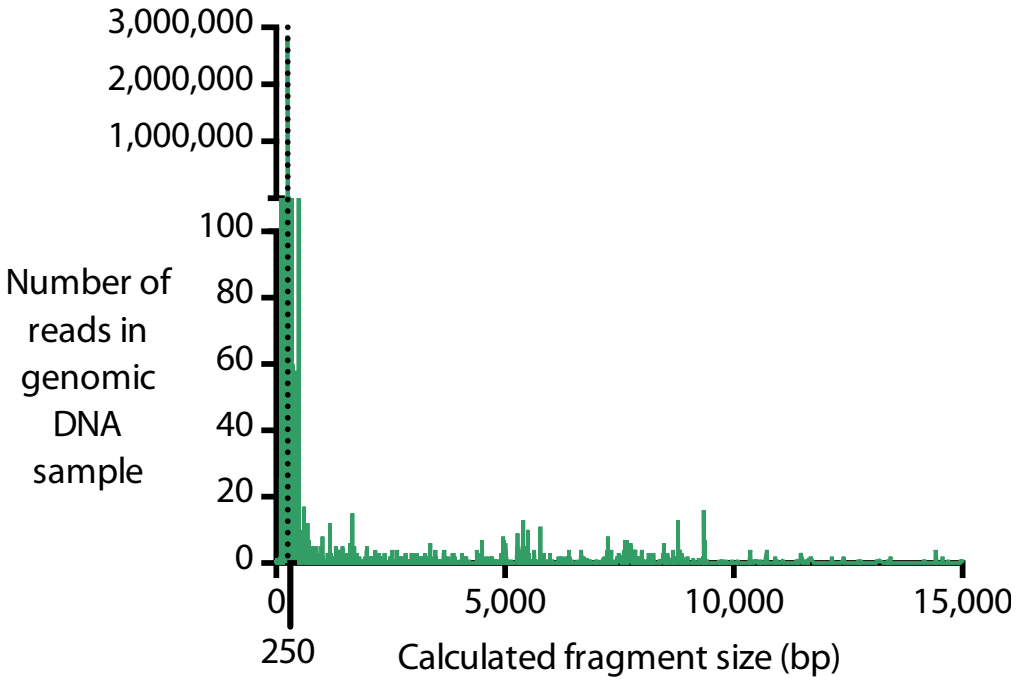


Figure 14. MRTemp shows that the apparent variation of fragment sizes occurs due to the existence of chimeric DNA fragments.

a. True fragment sizes, which are not based on mapping results of read pairs, but rather on merged overlapping reads, reveal a genuine experimental fragment size distribution. The minimum fragment size is 190 basepairs (the size of a single read) and the maximum is 370 basepairs (the size of two reads, with 10 nucleotides overlap). ***b.*** Distribution of calculated fragment sizes, based on the mapping of *in-silico* generated read pairs from MRTemp, which are all exactly 250 basepairs shows that some DNA fragments must be based on chimeric DNA templates. Overall, approximately 4% of fragments appear to have a size that is not 250 basepairs.

2.3.1. Predicted fragment sizes are different from true fragment sizes

Unexpectedly, when the MRTemp read pairs from 6 separate flies were aligned, the theoretical fragment sizes on the basis of this alignment had a similar distribution of predicted fragment lengths to that seen with previous methods (Figure 14b). This result is remarkable because the input read pairs for the alignment originated from DNA fragments that were definitely 250 basepairs long. To test whether this divergence between theoretical and actual fragment sizes might occur due to errors of BWA, I next plotted the theoretical fragment sizes of the set of perfectly distributed *in-silico* read pairs (see chapter 2.2.6). This analysis showed that no fragments mapped further apart than 500 nucleotides (the original fragment size), indicating that predicted fragment sizes above the size of the input fragments did not arise from issues with the mapping of the reads (Supplementary Figure 2).

Taken together, these results suggest that the variation of DNA fragment sizes is not the reason why some read pairs map to the reference genome at locations that are further apart than the assumed fragment size. Therefore, we concluded that variation in fragment size is not responsible for the hovering effect.

2.3.1. DNA chimeras are formed during DNA amplification

The most parsimonious explanation for the finding that the theoretical fragment sizes of merged, overlapping read pairs diverge from the actual fragment sizes is that DNA chimeras were formed during one of the preparatory steps that precede WGS. Chimeric DNA fragments are very problematic because they will result in paired-end reads that could also be misinterpreted as rare somatic insertions (Figure 11b). The input fragments for the MRTemp method are all 250 basepairs long. The predicted fragment lengths from these alignments can therefore be used as a measure for the prevalence and distribution of chimera formation. By assessing the predicted fragment lengths, I estimated that approximately 4% of all mapped read pairs arise from a chimeric DNA template (e.g. 1,267,509 of 30,362,126 read pairs of fly #1 mapped at locations further apart than the expected fragment size).

Lasken et al. have studied and documented the mechanism of chimera formation during the amplification of bacterial DNA[70]. They found that MDA leads to an incredible array of chimeric annealing artefacts. The DNA that is produced during MDA is highly branched, and the DNA strands can be displaced by the MDA polymerase. The freed DNA strands can then prime at a nearby complementary sequence, leading to the formation of a chimeric DNA fragment. This priming will result in a sequence overlap between the two stretches of DNA that form the chimera. The lengths of these overlaps represent the balance

between sufficiently long priming and the probability of a complementary sequence to occur close to the original displacement site. For the single *E.coli* cell that was analysed by Lasken et al., the distribution of these complementary sequences ranged from 2 to 21, with a peak at 5 basepairs. Astonishingly, when I plotted the lengths of chimeric sequences from my sequencing data, I observed a remarkably similar distribution (Figure 15).

Figure 15. Chimeric amplicons contain complementary sequences of various lengths.

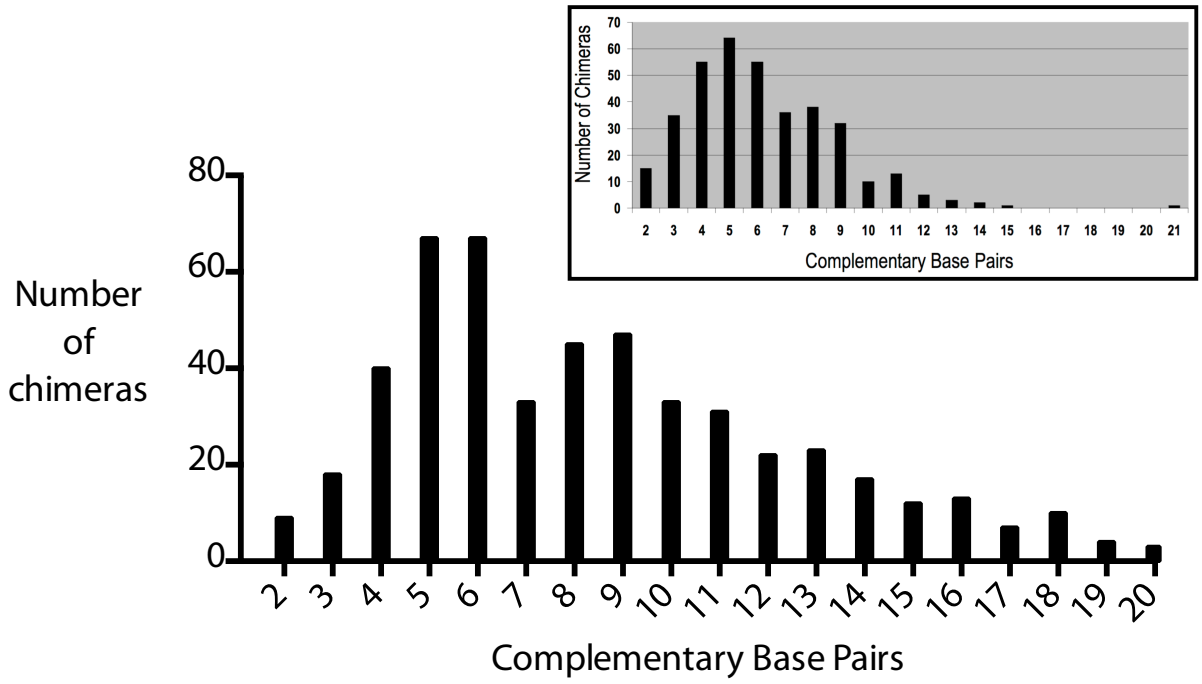


Figure 15. Chimeric amplicons contain complementary sequences of various lengths.

The frequency of specific sizes of palindromic sequences spanning the junction of chimeric amplicons (only those above 1) are shown in this graph. The distribution of chimera sizes of an amplified *Escherichia coli* DNA molecule, as reported by Lasken and colleagues, is shown in the inset (inset was taken from [70]).

In summary, my analysis revealed that chimera formation during DNA amplification leads to artefacts when screening WGS data for somatic transposon insertions. Furthermore, chimeras appear to be preferentially formed between genomic loci that are within 5-10kb of other, which then manifest as hovering events when the reads are analysed using TEMP.

2.4. Filtered MRTemp hits

I next tested whether the reads that predict putative somatic transposon insertions and are not clearly identifiable as amplification artefacts might contain reads that represent genuine transposon insertions. To do this, I filtered the data by removing the putative somatic insertions that occurred within 5000 nucleotides of a transposon germline insertion of the same transposon family. It should be noted, though, that whilst this filter is useful for removing hovering artefacts that arise from regionally restricted chimeras, it remains possible that chimeras form between two more distant DNA stretches. This means that any TEMP hit might be an artefact rather than a rare somatic insertion. To assess whether this was the case, I tested if the presence, absence or number of somatic insertions in αβ-KCs correlated with age or transposon expression levels.

2.4.1. No correlation between transposon expression levels and insertion events

The broad and general assumption seems to be that transposon expression translates to transposon mobility[53]. Hence it would seem reasonable that more highly expressed transposons might produce more somatic insertions. I therefore exploited my prior finding that some transposons were more highly expressed in αβ-KCs than in the rest of the brain (see Figure 6) and correlated the relative expression levels of each transposon type with the number of putative MRTemp hits, which had been filtered to remove hovering artefacts (Figure 16). However, no significant correlation was apparent. I also did not observe a significant correlation if the data was not filtered for hovering (data not shown). Retrotransposition is replicative, which means that each new retrotransposition event results in an additional copy of the mobile transposon. I therefore tested whether the sequence of any of the transposon families was more abundant in the gDNA of αβ-KCs, when compared to gDNA from the rest of the brain from the same individual fly. To do this, I aligned the WGS data sets of each individual fly to a modified reference genome, which I termed DMclean. All transposon sequences in DMclean are removed and one single copy of each transposon sequence is then concatenated to form an additional simulated chromosome. As a consequence, the transposon reads from each WGS sample can only map to one location. I used this alignment to quantify the copy number of transposon sequences taking advantage of the Cufflinks analysis package, which is frequently used to quantify RNA molecules in RNA sequencing experiments[84]. Quantitative

analysis of sequencing reads for each transposon family did not reveal significantly increased copy numbers in αβ-KCs when compared to the rest of the brain, despite the fact that my previous results suggest that some transposons are more highly expressed in αβ-KCs (Figure 17, Supplementary Figure 3).

Figure 16. Expression levels of transposons do not correlate with somatic insertion rates.

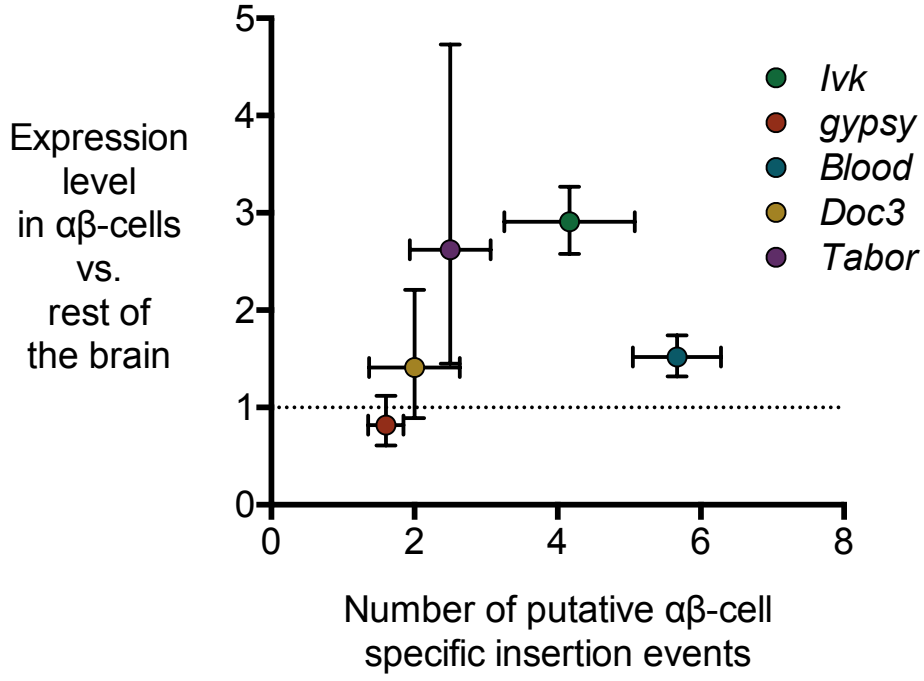


Figure 16. Expression levels of transposons do not correlate with somatic insertion rates.

Plot representing 5 different transposons and showing that there is no linear correlation between the expression levels of transposons in αβ-KCs and the number of putative somatic transposon insertion events in the same cells. The list of putative somatic events was filtered in order to remove unambiguously false positive hits (see Methods). Error bars represent SEM.

Figure 17. Copy number of transposon sequences is the same in αβ-KCs and the rest of the brain.

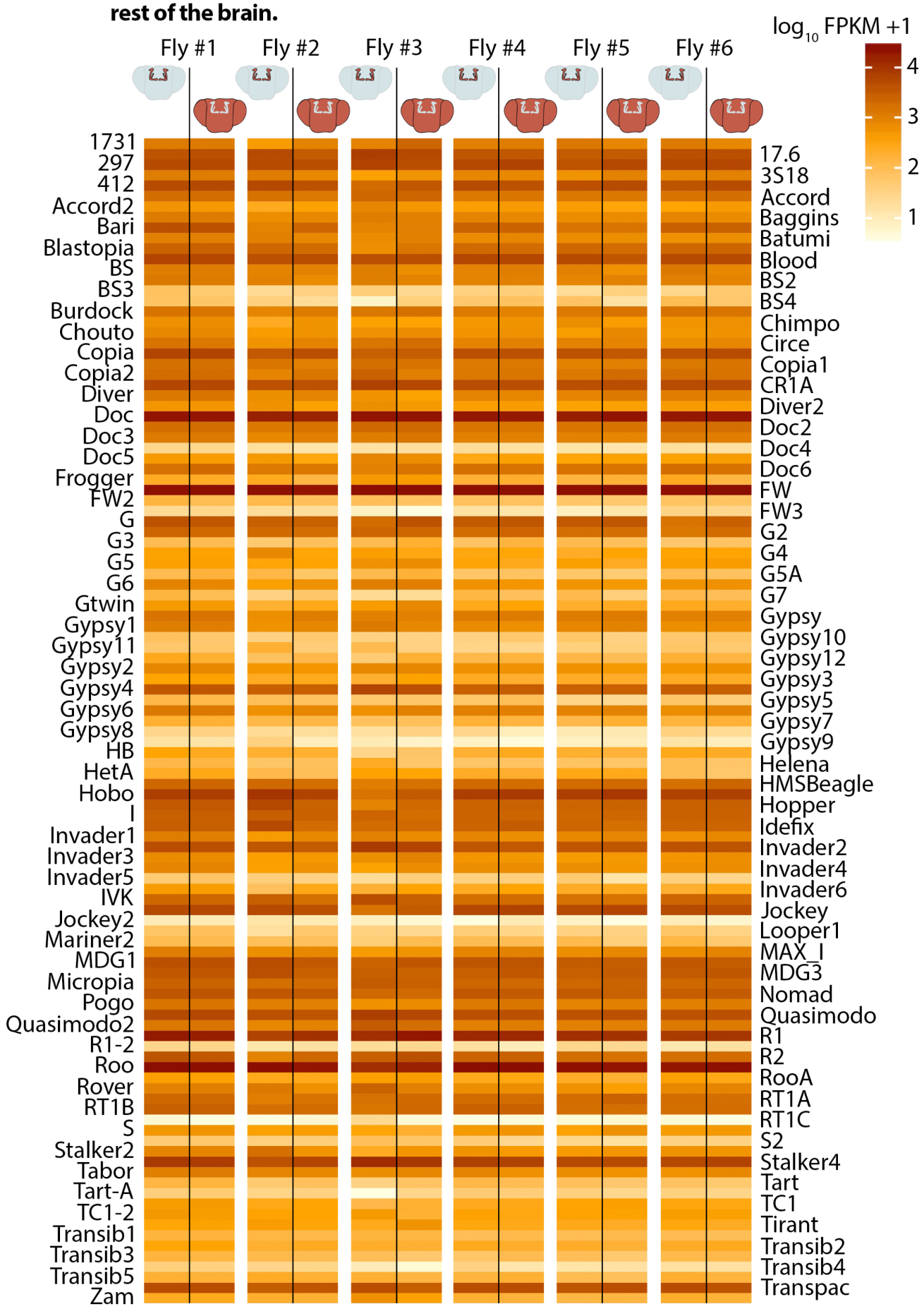


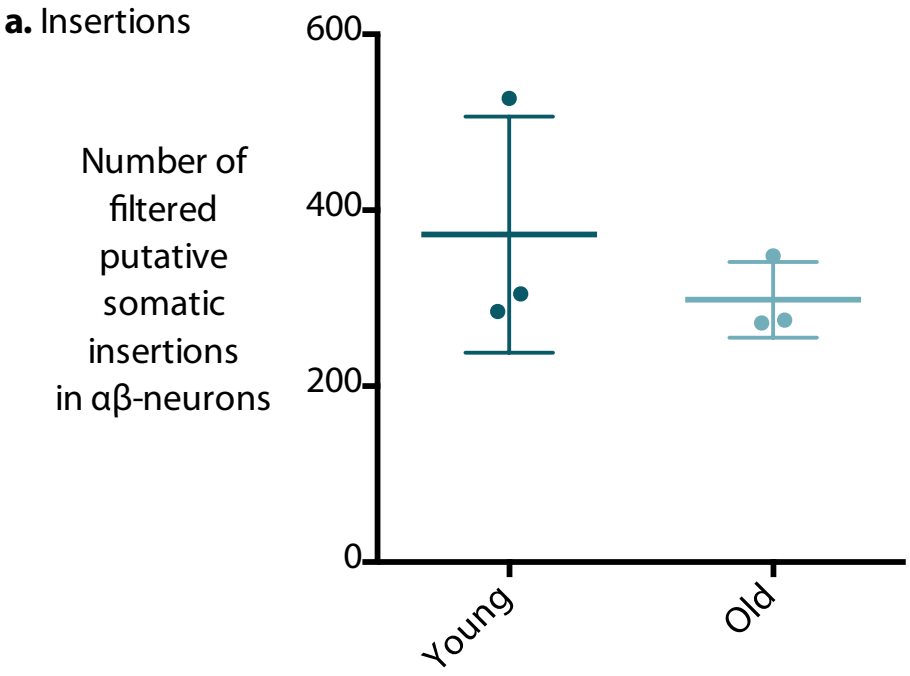
Figure 17. Copy number of transposon sequences is the same in αβ-KCs and the rest of the brain.

Heatmap showing the normalised number of sequencing reads that map onto each of the 111 reference transposon sequences that were analysed in this study. Note that there is no increased amount of transposon sequences in the gDNA from αβ-KCs when compared to the rest of the brain of the same individual, suggesting that there is no significant increase in retrotransposon activity in those cells. Individuals #1 - #3 are young flies (3-4 days) and #4 - #6 are old flies (30 days).

2.4.2. Insertions do not increase with age

If retrotransposons are indeed active in somatic cells, then the number of new insertions might be expected to accumulate throughout the life of the fly. To test whether this was the case for the putative insertion events that I identified in this study, I analysed the gDNA of young (3-4 days) and old (30 days) flies. MRTemp analysis, followed by a filtering step to remove hovering artefacts, showed that there was no significant difference in the number of putative somatic insertion events between the two ages (Figure 18a). I also tested the unfiltered hits and again did not find any significant differences (data not shown). In addition, I used my DMclean genome and the Cufflinks approach to check whether the total amount of transposon sequence within the gDNA increased throughout the lifetime of flies. This analysis also showed no significant difference in the total transposon number in the gDNA of old versus young flies (Figure 18b, Supplementary Figure 4).

Figure 18. The number of putative insertion events does not change with age.



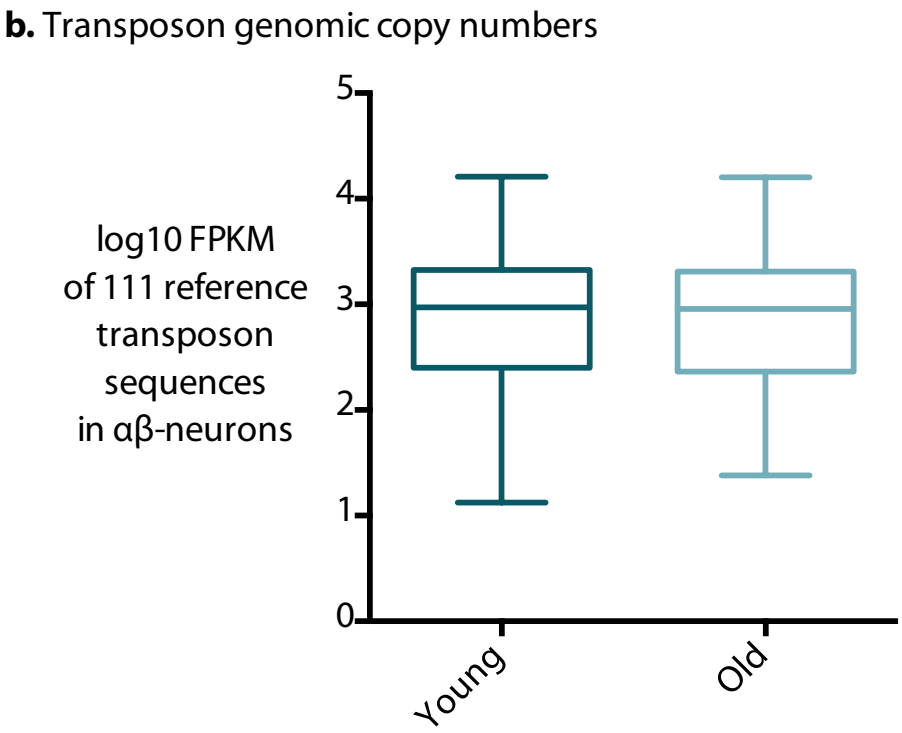


Figure 18. The number of putative insertion events does not change with age.

a. Graph showing that the number of putative filtered somatic insertion events does not differ between young (3-4 days) and old (30 days) flies. Error bars represent SEM. Hits that were present in a sample and absent in the rest of the brain and the reference genome were counted. ***b.*** Boxplot showing the normalised number of reads that map onto each of the 111 reference transposon sequences per αβ-neuron sample of young (3-4 days) and old (30 days) flies. FPKM stands for fragments per kilobase of transposon sequence per million fragments mapped. Whiskers represent Min and Max and the box represents the first and third interquartile interval.

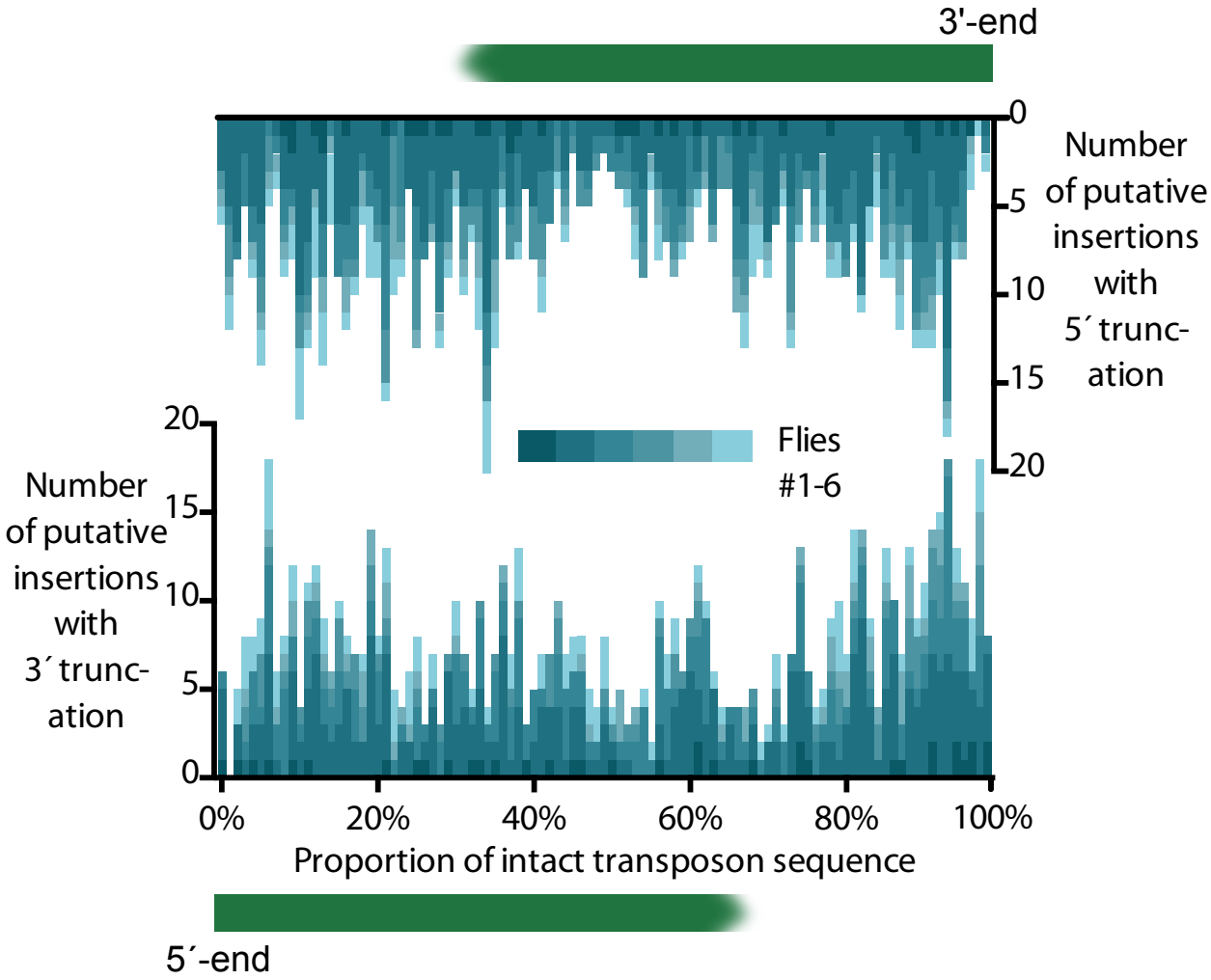
2.4.3. Break points occur throughout the length of LINE-like retrotransposons

During the mobilisation of retrotransposons, a new DNA copy of the mobile element is created at the site of insertion by the priming of a single-stranded RNA molecule from the original element. Therefore, it is reasonable to assume that if transposons are truncated during the process of a genuine somatic insertion event, the position of the break points for those truncation will preferentially occur at the ends of the elements.

Since MRTemp resolves the exact break points in each transposon, I mapped the ends of each putative somatic insertion in 6 individual flies to check where within the transposon the truncations occurred. This analysis showed that there was no bias for either end of the elements (Figure 19a). In contrast, breakpoints were randomly distributed along the entire length of the transposons. The majority of LINE-like retrotransposons in the fruit fly genome are 5′ truncated[85]. To test whether this pattern persisted in the putative MRTemp hits, I also plotted the locations of all LINE-like retrotransposon insertions (Figure 19b). However, this analysis revealed that the breakpoints of the putative *de-novo* insertions were not biased towards the 5′ end. We conclude that the distribution of fragments is more likely to be consistent with the theory that the hits result from random noise in the sequencing data, rather than genuine *de-novo* transposon insertions.

Figure 19. Transposon fragmentation appears to occur at random locations throughout the length of all transposons.

a. Size distribution of all transposon fragments



b. 5' truncation of LINE-like retrotransposons

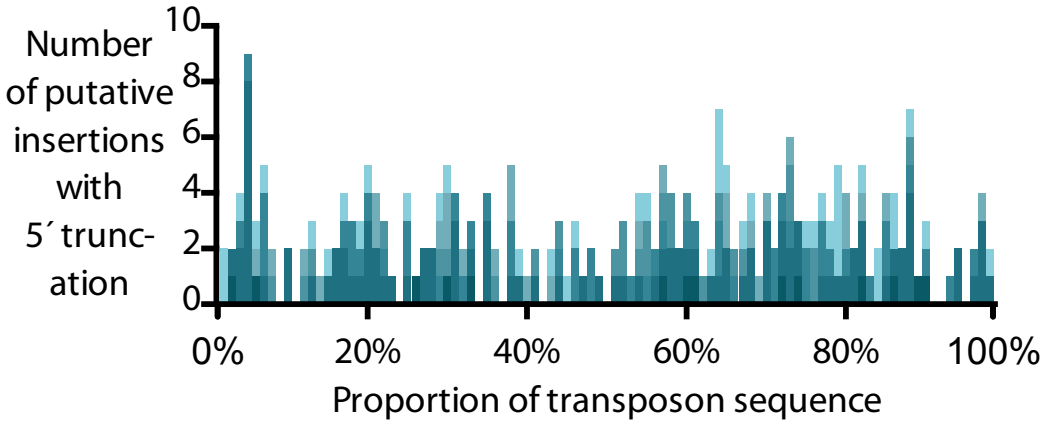


Figure 19. Transposon fragmentation appears to occur at random locations throughout the length of all transposons.

a. Graph showing the frequency of theoretical sizes of transposons in αβ-KCs of 6 individual flies. The top half of the graph shows transposons with 5′ truncations and the bottom half those with 3′ truncations. Note that the breakpoints are randomly distributed across all transposons. The data shown is based on MRTemp analysis. All putative transposon insertions with a predicted frequency of less than 10% are plotted. ***b.*** Graph showing the position of breakpoints in all LINE-like retrotransposons in αβ-KCs of 6 individual flies.

2.5. A simulated experiment to test for exon mobilisation

If most identified rare transposon insertions are actually artefacts, then we should be able to detect “mobilisation” of any stretch of DNA with TEMP. I therefore developed a set of analyses to test whether this was the case. I selected a list of fly genes and used exons from these genes to create a set of immobile genetic elements (IGEs, based on the assumption that exons are not usually mobile). The total number of nucleotides representing those IGEs was the same as in the transposon reference sequences. I also created a modified reference genome, DMsim, in which the IGE sequences were excised from their original location and added on to the end of one of the chromosomes. For example, imagine the genome sequence consisted of region A, followed by region B and then C. If region B were one of the chosen IGEs, then this stretch of DNA sequence would be A-C in DMsim. As the IGEs are most likely present in all flies, WGS data should contain DNA fragments that span the junctions of the IGE on both ends (A-B and B-C). If these sequencing reads are aligned to DMsim, then the A-B and B-C spanning fragments would map as discordant read pairs in this particular locus. The discordant read pairs are then screened for "insertions", using the IGE sequences as a reference, and hence region B should be identified as a germline "insertion" between A and C. Therefore, every genomic locus where an IGE is missing resembles a new, germline transposon insertion. The percentage of the simulated insertions that are detected by TEMP on each individual WGS data set is a reliable predictor of the sensitivity to detect germline insertions in this

particular sample. Most importantly, any additionally predicted mobilisation of an IGE signifies a library artefact and therefore the total number of additional hits is indicative of the false discovery rate for each sample. Running these IGE experiments represents a new analysis for any fruit fly WGS data.

2.5.1. Immobile genetic elements also appear to mobilise

Although one would assume that IGEs are not mobile, my analyses detected events that would qualify as IGE mobilisations in all samples (Table 1). Most importantly, IGEs appear to mobilise in similar patterns to transposons and hover around simulated IGE germline insertion sites (Figure 20). These results show that DNA chimeras are formed throughout the genome and that amplification artefacts present a substantial obstacle for the detection of rare somatic transposon insertions.

Figure 20. Simulated IGEs exhibit similar putative insertion patterns as transposons.



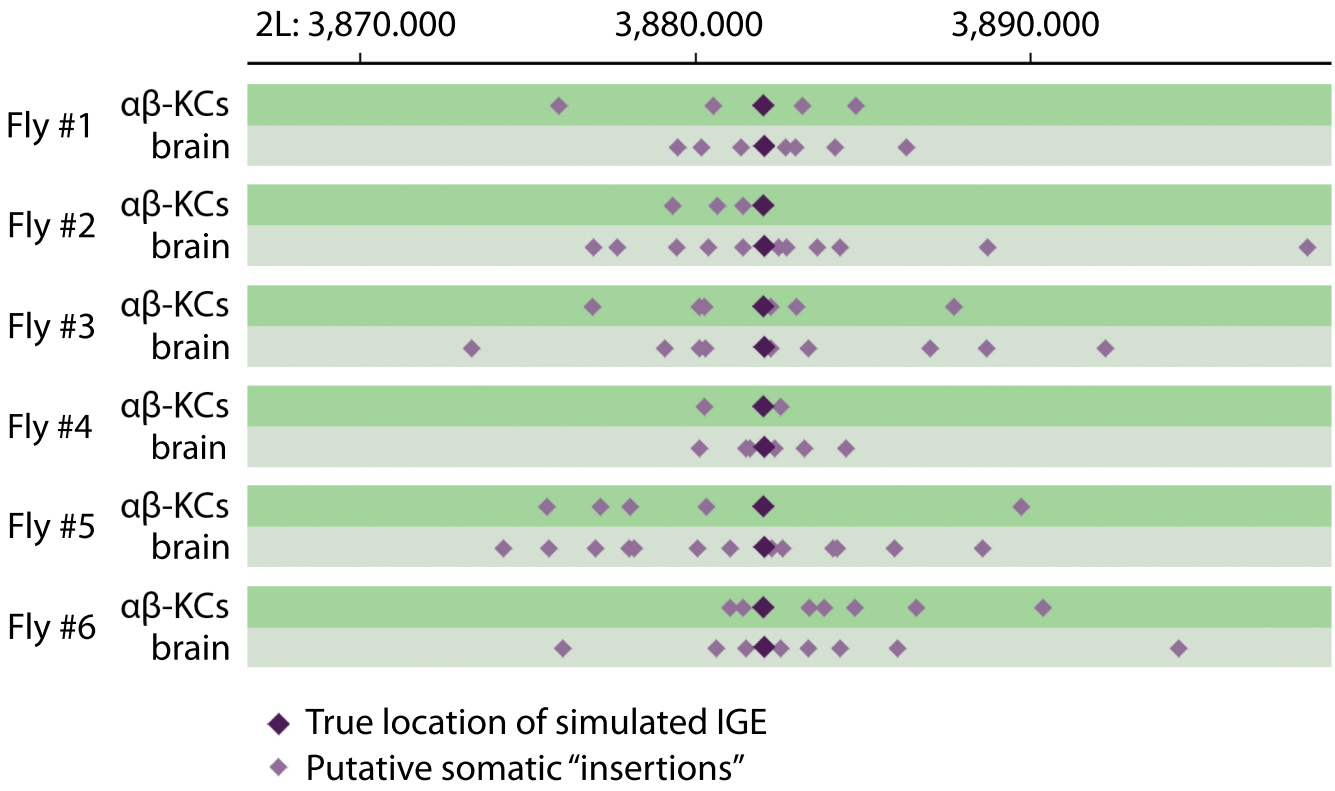


Figure 20. Simulated IGEs exhibit similar putative insertion patterns as transposons.

Representative example of a simulated germline insertion event that was found in each of the 6 individual flies and that is absent in the simulated *Drosophila melanogaster* reference genome (DMsim). Based on the same criteria as in Figure 10, putative somatic insertion events of the same simulated element seemed to occur at loci which are approximately 10kb up- and downstream of the simulated germline insertion site. Dark purple diamonds represent the simulated germline insertion site, which was found in each of the 12 samples (αβ-KCs and the rest of the brain). Light purple diamonds represent putative somatic insertions in each of the samples.

Interestingly, somatic IGE "insertions" occur approximately 10 times less frequently than putative transposon mobilisations. This might suggest that a proportion of the transposon hits represent genuine somatic insertion events. However, as indicated above, although the total number of nucleotides was the same in the collection of reference transposon sequences and IGEs, the total number of times a sequence occurs differs between the two collections. The reason for the difference is the fact that the fly genome contains multiple copies of each transposon family, but one of each IGE. Since transposon sequences covered 30,202,607 nucleotides (17.6%), and the selected IGEs 3,540,381 nucleotides (2.1%) of the reference genome, the rate of IGE mobilisation was effectively the same as the rate of putative somatic transposon insertions, when corrections were made for the fraction of the genome that was tested.

2.5.2. Previously published data contains similar chimeric artefacts

To determine whether the identified problems with the analysis of my transposon mapping data was a general issue that also applied to other studies, I ran the IGE control experiments on WGS data from one published and another unpublished data set from two independent groups.

It is well established that the offspring of male flies carrying the *P* element transposon (P) and female flies without the element (M) are initially sterile and

then regain fertility as they age, a phenomenon termed P-M hybrid dysgenesis[23]. It has been shown that the sterility is the product of *P* element mobilisation in young flies. Khurana et al. reported that P-M hybrid dysgenesis also triggers the activation of resident transposons and that this effect decreases with age, as well as in the subsequent generation[68]. I have tested the authors WGS data and detected IGE "mobilisation", suggesting that they are subject to the same artefact issues (Figure 21). Even more strikingly, I found that the total number of IGE hits varied between different samples and was higher in the sample that was reported to contain an increased number of new transposon insertions.

Figure 21. Sequencing data from other labs is also prone to amplification artefacts.

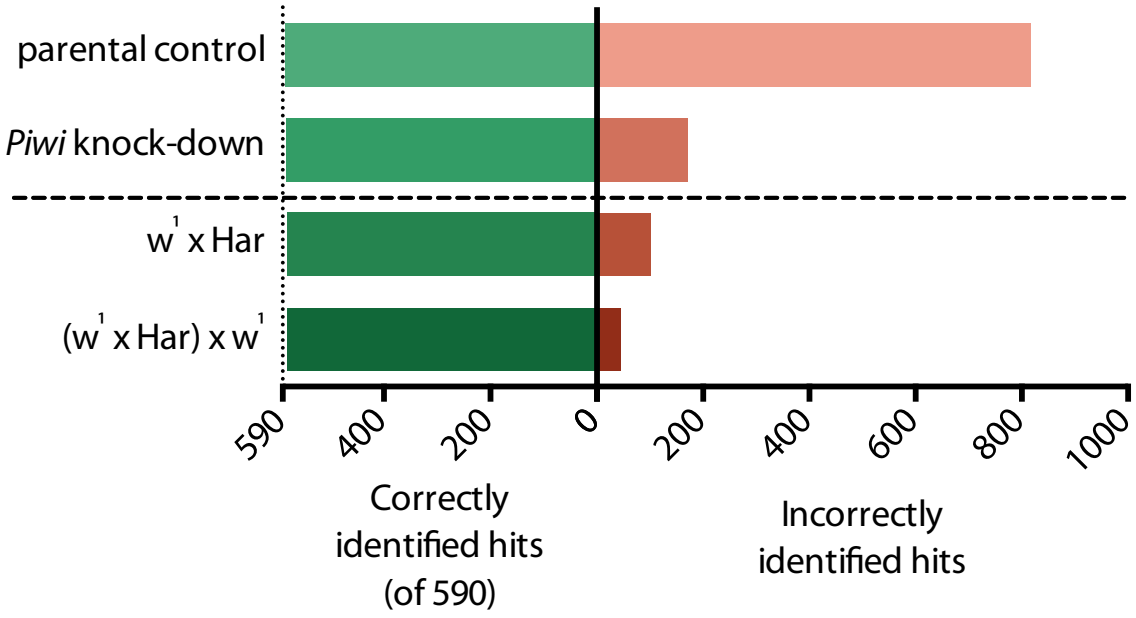


Figure 21. Sequencing data from other labs is also prone to amplification artefacts.

Graph showing number of IGE mobilisations of data from a previously published study[68], and also data from an unpublished study (permission to use this data was generously provided by Dr. Alain Pelisson). Green bars indicate the number of simulated IGE deletions, which were detected with each of the data sets (of 590 in total). Red bars show all putative IGE "mobilisations". Khurana et al. reported an apparent decrease of *de-novo* transposon insertions when w^1 x Har flies were outcrossed to w^1 flies. Note that the number of IGEs also decreases in the outcrossed flies, suggesting that both differences are the result of different rates of artefact formation. Similarly, Pelisson initially reported that parental control flies appeared to harbour more somatic transposon insertions than flies with down-regulated Piwi proteins. The in-silico analysis reveals that the rate of IGEs differs between the two samples.

The unpublished data that I tested was kindly provided by Professor Alain Pelisson, and consisted of WGS data from flies of two different genetic backgrounds. One of the samples was obtained from fly embryos in which Piwi, a protein that acts via piRNA to silence transposons, was knocked down and the other was one of the parental controls with normal Piwi levels. The initial conclusion from the data was unclear because the TEMP approach suggested that somatic transposon insertions decreased when Piwi was knocked down. However, the IGE experiment shed more light on this and showed that the Piwi sample also exhibited a decreased level of amplification artefacts.

It seems likely that in both of these studies, differences in the amount of amplification chimera formation were misinterpreted as differential levels of somatic transposon insertions. In summary, we conclude that errors due to DNA amplification artefacts are a common and unavoidable problem for the analysis of rare somatic transposon insertions.

2.6. Co-expression of transposons with neighbouring genes

My data suggests that although αβ-KCs overexpress some transposons, these elements do not produce a detectable level of new insertions. I therefore queried whether transposon expression might result from activity of one of the loci in which they reside. To test this, all transposon loci of the fly strain to be

investigated must be known, as any of them could account for the particular expression pattern. Multiple copies of each transposon type exist within the fly genome and the germline locations of transposons vary between different fly strains[3]. Hence, the reference genome, which was based on the fly strain *iso1*, will very likely not include the same pattern of germline transposon insertions as the flies that were used for this study. However, the high coverage WGS data on my flies enabled me to identify the endogenous locations of transposons. I pooled the sequencing data from all long-read samples of the rest of the brains, which resulted in a combined genome coverage of 15x. To check the sensitivity of the analysis of germline transposons of this data set, I aligned the combined sequencing reads to DMsim (see chapter 2.5). DMsim contained 590 simulated germline transposon deletions, 541 of which were detected by MRTemp, suggesting that on average 92% of all germline transposon insertions should be detected. This means that I was able to detect the majority of all genomic loci for each transposon type. For example, I found that in addition to the known locations in the reference genome, three additional copies of the LTR retrotransposon *Tabor* were each inserted in the intronic regions of the genes for CG12484, CG10576 and CG15831. None of these germline locations have previously been identified by any of the publicly available transposon databases.

To test whether transposons are co-expressed with neighbouring genes, I exploited the fact that the LINE-like retrotransposon *Ivk* was consistently

overexpressed in αβ-KCs. I started by cataloguing all endogenous germline locations of *Ivk* and found 16 germline insertions, three of which were in introns of annotated fly genes. The genes encoded for an axon guidance protein called Trim9, the ATPase Spock and a peptidase named CG17684. Next, I designed UPL® primers and probes for each of these genes and tested whether any of these three genes was also upregulated in αβ-KCs.

These experiments showed that Trim9 is not expressed in the brain (or is expressed at a level that is below the detection levels of my approach) and Spock is expressed at similar levels in αβ-KCs as in the rest of the brain (Figure 22). Conversely, CG17684 expression was upregulated in αβ-KCs at a similar rate to *Ivk*. This result suggests that the apparent upregulation of *Ivk* could be due to the co-expression with CG17684. *In-silico* analysis revealed two potential cryptic splice sites at the end of *Ivk* (data not shown), which means that *CG17684* may potentially form a chimeric transcript with *Ivk*. In addition, according the reference genome, the gene *CG17684* also contained a full length copy of *Tabor*, another transposon which, according to my prior data, is upregulated in αβ-KCs (Figure 6).

Figure 22. *Ivk* and the neighbouring gene *CG17684* are both upregulated in αβ-KCs.

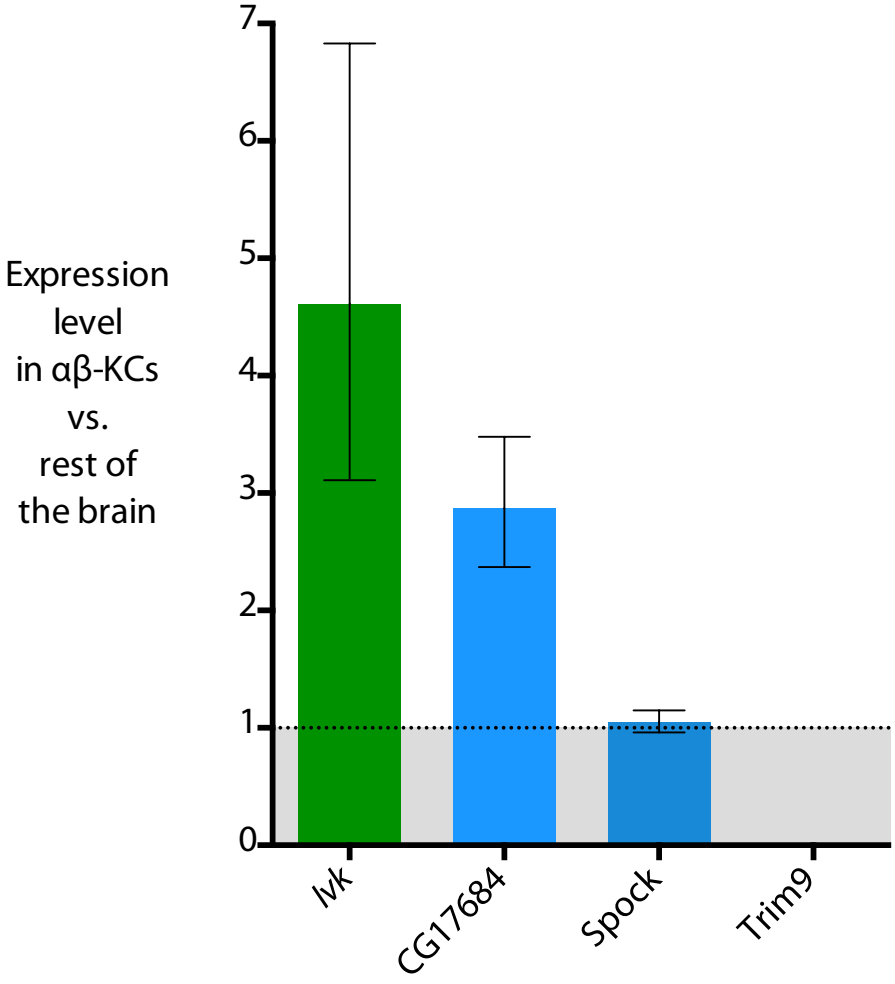


Figure 22. **Ivk *and the neighbouring gene* CG17684 *are both upregulated in αβ-KCs.***

Plot showing the expression levels of *Ivk*, and each of the genes that are neighbours to a full-length germline copy of *Ivk*, between αβ-KCs and the rest of the brain. Error bars represent combined SEM of 4 separate experimental runs, with 50 flies and 3 technical replicates per run. *Trim9* expression was not detectable in any of the samples (although the assay was positively tested on whole RNA extracts).

3. Discussion

In the current study, I present a new method to probe for somatic transposon insertions in the brains of single fruit flies, with a theoretical cellular resolution to identify rare somatic insertion events. My analysis reveals that the amplification of gDNA results in the formation of DNA chimeras, which lead to WGS read pairs that are indistinguishable from reads that would indicate the presence of genuine somatic insertions. I show that despite efforts to filter sequencing reads for putative DNA chimeras, no correlation is evident between the number of *de-novo* insertions and the expression levels of transposons, or the age of the flies. Figure 16 displays the relative expression levels of 5 different transposon types in αβ-KCs, compared to the rest of the brain, plotted on the y-axis and the number of apparent "somatic insertions" (according to the MRTemp analysis), plotted on the x-axis. It is conceivable that a higher expression level of a specific transposon type in αβ-KCs would result in a higher number of somatic insertions of that transposon type in those cells. However, this does not seem to be the case for the "insertions" reported by MRTemp. Similarly, Figure 18a shows a plot of all MRTemp "insertions" of young vs. old flies. Genuine somatic insertions that occur in adult flies would accumulate over time. Yet, I was not able to observe an accumulation using MRTemp. In addition, this accumulation of somatic insertions would result in a higher copy number of transposon sequences in the gDNA of old flies, compared to young flies. Figure 17 and Figure 18b show that this is not the case. In summary, this indicates that despite careful filtering and

stringent criteria, MRTemp "insertions" are most likely not genuine insertions, but rather represent DNA chimera that were formed during DNA amplification. Whilst chimera formation clusters locally (see Figures 9, 12, 14b, 15 and 20), some rare chimera are not formed locally, but distally. It is impossible to identify these distant chimera and hence they represent a fundamental issue for the detection of rare somatic transposon insertions. Furthermore, the observation that transposon sequences do not accumulate in old flies strongly suggests that the rate of somatic transposition cannot be very high. I also present a series of novel *in-silico* simulated control experiments to estimate the error rate of somatic transposition analyses. These experiments confirmed the high prevalence of artefacts due to DNA amplification chimeras in my own data, as well as data from two independent research groups. Taken together, my data suggests that the rate of somatic transposon insertions in αβ-KCs is substantially lower than that of previously reported studies in flies. Finally, I propose an explanation for differential expression of transposons in the fly brain, where transposons are co-expressed from a locus that also includes a cell-specifically expressed neural gene.

3.1. Local transposon hopping had previously been reported in flies

My initial results on somatic transposition in αβ-KCs indicated that transposons preferentially insert in locations within about 10kb up- and downstream of endogenous germline locations in the genome, a phenomenon that I termed hovering. These preliminary results appeared to be promising, as they were

reminiscent of findings on the mobilisation pattern of the DNA transposon Ac/Ds in *Zea mays* and *P* element mobilisation in fruit flies. Both of these mobile elements had previously been reported to preferentially occur at sites near to the site in which they resided[86,87]. If this local mobilisation preference was true for endogenous fruit fly transposons, then germline transposon insertions close to genes could have been an intriguing way to render those genes vulnerable to somatic gene disruption when the control of transposon mobilisation is released in somatic cells. In hindsight, however, it is now evident that the hovering effect that was observed in this study is not a genuine feature of fly cells, but rather a consequence of chimera formation during DNA amplification.

3.2. Somatic transposition is unlikely to occur at high frequency in adult fly brains

The FACS approach, presented in this study works particularly well for the extraction of αβ-KCs from fruit fly brains. This may be because the somata of KCs are located at the posterior end of the brain above the calyx. For FACS, brains are first dissected from fly heads, then digested with a peptidase and physically sheared, in order to separate the cells from each other. It is reasonable to assume that this process is more efficient for the dissociation of cells that are close to the surface of the brain. In addition, KCs have a single neurite that is connected to the cell soma. During trituration neurites are removed and this process is likely to be less destructive if cells contain fewer processes. The yield of intact cells per

brain after FACS was therefore high. The theoretical sensitivity of my strategy to test a small subset of cells in the brains of single flies with a high WGS coverage was substantially higher than the sensitivity of previous efforts in flies. It seems highly unlikely that the rate of somatic transposon insertions in fruit fly brains is particularly high, because despite the moderate level of noise in the data, highly prevalent events would have still been detectable. I tested the data for potential correlations with several parameters, and none of them resulted in any notable interactions with the number of putative transposon insertions.

The relation between transposon expression levels and their rate of somatic insertions is worth assessing more closely. Conceptually, a somatic transposon insertion in a cell requires a single transposon copy for the insertion. However, my analyses (and the data from others) revealed that transposon expression levels are generally similar to those of moderately expressed housekeeping genes. This means that thousands of transposon RNAs appear to be present in the cytoplasm of a cell at any point in time. This, in turn, suggests that either the majority of the RNA copies, which are detected by RT-qPCR are not insertion competent (and are potentially bound to ribosomes, which translate the sequences into enzymes, such as RT), or alternatively, that gDNA is very efficiently protected from new transposon insertions. This could be mediated by endo-siRNAs or piRNAs complementary to transposon sequences. The piRNA pathway was first identified in germline cells, but more recently it has been found to be active in somatic

cells[88]. For example, *Hiwi*, a human PIWI protein, was found to be ubiquitously expressed in somatic tissue. This silencing mechanism could be one explanation for the discrepancy between high expression levels and low rates of somatic insertions. To test whether this might be the case, the relation between transposon expression levels and somatic insertions could be analysed in flies that harbour mutations in Argonaute proteins. If transposon mRNA levels would remain unchanged, and somatic insertions would increase, then this would suggest that transposon are silenced by small RNAs in wild type flies.

3.3. Co-expression of genes and transposons could present novel gene regulation mechanisms

This and previous studies convincingly show that transposons are not expressed uniformly throughout the fruit fly brain[52]. So what could be the reason for this differential expression pattern? One answer to this question is presented in this study. The fact that one of the three genes that contain intact copies of *Ivk* is upregulated in the same cells as *Ivk* inspires an interesting hypothesis. The apparent co-expression of *Ivk* and CG17684 could mean that cells contain chimeric RNA molecules, in which the transposon sequence has become part of the mRNA. If this is the case, then it is also conceivable that the piRNA molecules that are complementary to *Ivk* also silence the expression of the gene that shares the mRNA. This could present an intriguing novel epigenetic way of regulating gene expression in cells. Expression of piRNA against *Ivk* in fly heads has been

reported by Xu et al. [89,90]. If either the piRNA, or Argonaute proteins were down-regulated in αβ-KCs, then this could lead to an increased expression level of CG17684 because the chimeric mRNA molecule is no longer degraded via the small RNA pathway.

3.4. Single-fly analysis will enable analyses on germline transposon dynamics

The fact that I was able to investigate transposon insertions in two tissue samples from the same individual fly was crucial, as it allowed me to distinguish rare germline insertions that occur at a low frequency in the population being analysed for somatic insertions. Historically, flies from the same strain tended to be considered as isogenic (i.e. sharing the same genetic background). My analysis on six individual flies shows that there is a substantial number of private unique germline transposon insertions in individual flies, which are absent in the other flies with a similar genetic background (i.e. they originate from the same parental strains). As pointed out in section 2.2.3., this can very easily be misinterpreted as rare somatic insertions in a particular cell population, because even if several independent groups of flies are analysed, it is still conceivable that rare germline insertions are picked up by accident. This is particularly the case if the analysis lacks the power to identify these rare events in the samples that are used as reference tissue (e.g. the rest of the brain). It will be very interesting to apply the single fly assay to study the population frequency of germline transposon

insertions in isogenic fly strains, and also to compare it to outbred strains. Such experiments will be instrumental to check whether there is a selection for, or against, a particular germline transposon insertion; whether there is any genomic regional bias for or against germline insertions; and whether the germline insertions have any impact on the behaviour of the individuals that carry them. In summary, the fly, in combination with the assay that was established in this study, is a powerful model organism to study these questions, because flies can be kept in isolated, isogenic populations, and yet single individuals can be analysed. Furthermore, flies can also be reared as perfect clones with the use of balancer chromosomes. Crossing schemes exist that lead to the foundation of a new strain, in which all flies share the same founding allele of all genes and chromosome regions. This system could be used to study the evolutionary development of new germline transposon insertions.

The finding that WGS reads from gDNA were not equally distributed across the genome (see section 2.2.6. and Figure 10) revealed a caveat when sequencing data is used to quantify genomic events. It should be noted that in addition to the shared read patterns, such as the ones shown in Figure 10, I also observed cases where specific regions exhibited an over- or under-representation of reads in a single sample, when compared to the rest. This observation might be of significance for RNA-seq experiments. For RNA-seq, cDNA is amplified and sequenced, and the number of reads that map onto a specific gene are used as a

measure of the expression level of this gene. In my experiment, the relative input for each (non-repetitive) genome sequence is one, as every cell contains the same number of gDNA molecules. Nevertheless, if my data had been analysed in the same way as RNA-seq data, then some "genes" would have appeared to be up- or down-regulated in each sample, due to variations in the number of reads at these loci. Hence, the amplification bias that is reported in this study should be taken into account for RNA-seq experiments.

MRTemp, the new method that is presented in this study, led to several important improvements compared to previous assays. Most importantly, MRTemp reveals the exact breakpoints of potential new insertions or of chromosome rearrangements. Knowledge about the exact break points allows us to answer a set of very interesting questions. Firstly, it makes it possible to unambiguously group putative insertion events that are detected in multiple samples because events can be lined up with single nucleotide resolution. Secondly, reads that span both ends of a germline transposon insertion reveal the target site duplication (TSD) at this locus. This is an important piece of information, as it can be used to test whether certain transposon types are associated with specific TSD sizes or sequences. and whether any of the transposons that are detected are inserted in regions with specific sequence motifs.

3.5. The dynamics of chimera formation are similar in bacteria and fruit flies

The notion that MDA leads to the formation of DNA chimeras reveals a fundamental issue when looking at somatic transposon insertions. The less cells being analysed, the more powerful the analysis to detect rare events will be. However, fewer cells also mean that the DNA has to be amplified to a greater degree, and hence the rate of chimera formation will increase as well. The model that Lasken et al. have put forward to explain chimera formation was based on the amplification of a single *E.coli* DNA molecule. To my knowledge, so far no other study has confirmed the findings of Lasken et al. The current study presents the first evidence that DNA chimeras from eukaryotic DNA are formed with the same dynamics as those in *E.coli*. The distribution of overlapping gDNA sequences that are presented in Figure 15 are almost identical to those found by Lasken et al. (see insert in Figure 15). The peak of this distribution represents the balance between the increased affinity for base-pairings of a longer sequence overlap, versus the likelihood that this sequence reoccurs at a certain distance from the original site. In other words, the longer the overlapping sequence, the more efficiently this sequence will prime at the complementary locus, but the further apart this sequence will occur from the original location. The highest theoretical power to detect rare somatic insertion events in samples would be achieved by WGS on gDNA of single-cells. Whilst this has been done on mammalian cells before[50,69,73], an analysis of the raw data from these studies

reveals that the coverage of the genome is extremely patchy and sparse. Single-cell expression profiling has been done exhaustively for more than 10 years and despite some challenges related to the linearity of cDNA amplification[91,92], it has been instrumental for the characterisation of tissue samples and biological processes[93]. However, single-cell DNA sequencing is substantially more challenging, as each cell only contains two molecules of gDNA.

In a recent report, Evrony et al. pointed out that amplification chimeras might lead to artefacts if WGS is used to assess somatic transposon insertions. However, in their study, the authors did not perform any experiments, but simply re-analysed data from the previous study by Upton et al., who had used RC-seq on single neurons from human samples, and reported an estimated LINE1 somatic insertion rate of 13.7 events per cell. Evrony et al. concluded that the data on single neurons contained a plethora of artefacts, and that the rate of somatic insertions is actually about 50-fold lower than assumed. The study presented here is the first experimental evidence for the hypothesis that the presence of amplification artefacts in WGS data causes false positive somatic insertion hits.

3.6. New *in-silico* simulated control experiments make it possible to test WGS data for chimeric artefacts

Previous data simulations that were designed to test transposon detection algorithms focus on simulating WGS reads. This strategy is useful for assessing the efficiency of the algorithm, but it cannot be used to assess the quality of WGS data. I developed an approach that allowed me to test the quality of WGS data, and its suitability for analysing somatic transposon insertions. My experiments with simulated IGEs confirmed that the majority of putative transposon insertions identified by TEMP are really chimeric artefacts. Firstly, the fact that hovering appeared to occur with IGEs shows that local DNA chimera formation is confounding the analysis. Furthermore, IGE mobilisation occurred throughout the entire genome, showing that these chimeras could also be formed distally. Interestingly, the IGE experiments also shed light on the problems that occur when looking for germline transposon insertions in WGS data. The simulated deletions in the reference genome should occur at a population frequency of 100% in any fly sample. Whilst almost all 590 simulated IGE deletions were detected in each single fly short-read sequencing data (samples 3-14, see Table 1, last column), this was not the case for long-read samples, with lower coverage (samples 15-26). On average, only 431 (N=12 samples, SD=91.05) of 590 deletions were detected by those samples. In addition, the deletions that were missed differed between the samples. These findings, and also similar results on WGS data from other labs (see Figure 21, green bars) highlight the issue that germline

insertions with a rare population frequency are easily missed out by WGS analysis with a low sequencing coverage.

3.7. New and improved approaches are necessary to test for rare somatic transposon insertions

In the light of my results, and the conclusions by Evrony et al., it becomes evident that new approaches will need to be developed, to identify and study genuine somatic transposon insertions. Due to the advances in DNA sequencing technology, somatic genomics are now becoming more important for our efforts to understand cellular phenomena in higher organisms, like the fruit fly, and in other areas such as embryonic development and cancer genomics. This study points out issues that will have to be taken into account when somatic tissue is being sequenced. DNA amplification is an essential step for the analysis of somatic genomes. However, my data shows that more effort will be required to develop new methods and protocols that address the issue of DNA chimera formation. The two examples of data from other labs, presented in this study, highlight how artefacts can be misinterpreted if the amount of DNA chimeras varies between samples. This, and other studies show that it is absolutely essential to analyse WGS data rigorously and to perform the appropriate control experiments. My data also challenges the hypothesis that transposition drives behavioural idiosyncrasy in fruit flies. I showed that the genome of fly neurons does not contain hotspots for transposon insertions and that if somatic insertions

do occur at all, then at least their rate of occurrence is too low to have a noticeable impact on the behaviour of a fly.

To my knowledge, this study presents the first direct evidence for an apparent co-expression of a transposon with a neighbouring gene. One reason why this may not have been reported earlier in other organisms could be the fact that these kinds of experiments can only be performed if two criteria are met: The first criterion is that the host genome has to contain different transposon types, so that they can be distinguished based on their DNA sequence. Secondly, all germline transposon insertions of each transposon type must be known, so that all potential neighbouring genes that could drive the expression of a certain transposon can be identified, including at least one that is differentially expressed in the tissue samples that are being analysed.

4. Methods

Fly strains

The flies were raised on standard food at 25°C, and 40-50% humidity. Two separate fly strains were used in order to fluorescently label αβ-KCs. For the first group, w-; c739-Gal4; MB247-lexA males were crossed with w-; *lexAop-FLP*; *UAS>STOP>CD8::GFP* virgins, as previously described[52]. For the second group, split-Gal4 fly strains were used[94]. Mb008b males were crossed with w-; +; *UAS-mCherry* males. The first line was used in order to try and reproduce the results of Perrat and colleagues, and the latter line was used for all other experiments.

Fly brain dissection and immunostaining

Fly brains for confocal imaging were prepared using a standard immunostaining protocol[95]. In summary, fly brains were dissected in phosphate-buffered saline (PBS, 137mM NaCl, 2.7mM KCl, 10mM Na2HPO4, 1.8mM KH2PO4) and fixed with freshly prepared 4% Paraformaldehyde in PBT (PBS, containing 0.3% Triton X-100) for 20 minutes. After fixation, brains were washed 3 x 20 minutes in PBT. Next, brains were incubated in 5% normal goat serum (NGS) in PBT for 30 minutes, and then incubated with the primary antibody NC82 (Mouse, 1:20 dilution in 5% NGS in PBT) at 4° C for 1 week. After incubation, the brains were washed 3 x 20 minutes in PBT, and then incubated with the secondary antibody (Alexa Fluor 488 Goat Anti-mouse, 1:100) at 4°C for 1 week. Finally, brains were

washed 3 x 20 minutes in PBT, and 3 x 20 minutes in PBS, and mounted on microscopy slides (Superfrost PLUS), by carefully placing them between two supporting spacers, and immersing them in antifade mounting media (Vectashield).

Confocal imaging

Brain images were taken on a Leica TCS SP5 confocal LASER microscope, with a HCX IRAPO L 25.0x0.95 objective. Image stacks were merged and processed with ImageJ. Only global changes to brightness and contrast were performed.

Gene expression analysis on whole heads

For the analysis of transposon expression levels in whole fly heads, ~200 flies per sample were collected in Eppendorf® tubes, plunge-frozen in liquid nitrogen, and vortexed, to dislocate the heads from the rest of the body. The sample was then applied onto pre-cooled sieves with mesh sizes of 850, 710 and 355µm. Fly heads were isolated on the 710µm mesh, and further processed. RNA was extracted from fly heads using Trizol, as previously described[96]. Following DNAseI treatment, RNA was next in vitro transcribed into cDNA using the Invitrogen Superscript® III First-Strand Synthesis kit, and oligo-$dT_{(20)}$ as primers to selectively create complementary DNA from mRNA molecules. cDNA from fly

heads was analysed on a Roche LightCycler® 480 II, using the same TaqMan® hydrolysis probes as Perrat et al., for comparison[52].

Fly brain dissociation

The fly dissection protocol was adapted from Nagoshi et al[97]. For expression analysis, groups of 50 sex-matched flies (25 females, 25 males) were dissected on ice in dissecting saline (9.9 mM HEPES-KOH buffer, 137 mM NaCl, 5.4 mM KCl, 0.17mM NaH2PO4, 0.22mM KH2PO4, 3.3 mM glucose, 43.8 mM sucrose, 50µM d(-)-2-amino-5-phosphonovaleric acid, 20µM 6,7-dinitroquinoxaline-2,3-dione, 0.1µM tetrodotoxin), and immediately transferred into SMactive medium (active Schneider's medium) on ice. The brains were then spun down by centrifugation for 1 minute at 2000rpm, then washed with dissecting saline, and incubated in l-cystein-activated papain (50U/ml papain, 1.1mM EDTA, 0.067mM 2-mercaptoethanol, 5.5mM cysteine-HCl) at room temperature for 30 minutes. The digestion was stopped by the addition of 5 volumes of SMactive, followed by two washes with fresh SMactive. Finally, samples were triturated in SMactive, by repeatedly pipetting the solution up and down a flamed P200 pipette tip (pre-wet). The solution was strained through a 20µm cell strainer, and subsequently loaded onto a MoFlo Astrios Cell Sorter.

For gDNA analysis of groups of flies, 30 females were dissected. For single flies, the same protocol was used, except that after every step up until the trituration,

each sample was visually inspected under a dissecting microscope to check whether the brain was still present.

All procedures were performed in Eppendorf® LoBind Protein microcentrifuge tubes, in order to minimise the loss of cells by adhesion to the walls of the tube.

FACS

Fluorescence Activated Cell Sorting (FACS) was performed on a MoFlo Astrios Cell Sorter, at standard settings, and cells were sorted based on 3 criteria. Firstly, the height of the side scatter (SSC) signal at 488nm was plotted against the height of the forward scatter (FSC) 488nm signal, and an area was carefully hand-drawn, in order to select droplets with intact cells, and to avoid droplets with cell debris. Secondly, the 488nm-FSC height was plotted against the width of the 488nm-FSC signal, and again, a selection gate was drawn on this plot. This criterion is used to exclude droplets that contain more than one cell. And thirdly, 488nm-513nm signal (area-log, green fluorescence) was plotted against the 561-579nm signal (area-log, red fluorescence). The rational for this plot was to be able to detect mCherry positive cells (red), and to distinguish them from autofluorescent cells, which would be positive for both wavelengths. On the final plot, three areas of interest were drawn, one to include all mCherry positive cells (Pos), and two separate ones in mCherry-negative cells (Neg-A, Neg-B), which represent the flow-through, and hence the rest of the brains. Gates Pos and Neg-A were used to collect cells for gDNA extraction, and gates Pos, Neg-A and Neg-B

for RNA extractions. Cells were sorted into Eppendorf® LoBind DNA microcentrifuge tubes, for further processing. Samples were visually inspected in a fluorescence stereo microscope (Leica TCS SP5) to confirm purity.

RNA isolation

For gene expression assays, mRNA was extracted from FACS fractions with the Arcturus PicoPure® RNA isolation kit, using the accompanying protocols. In summary, cells were incubated in extraction buffer at 42°C for 30 minutes, and cell debris were removed by centrifugation (3,000 x g, 2 minutes). RNA was isolated on RNA purification columns, treated with DNAseI, to remove DNA contamination, and finally eluted with 20µl elution buffer. Next, cDNA was created as mentioned above. For each sample, a control without RT (-RT) was also generated, in order to test of potential gDNA contamination.

RT-qPCR on FAC-sorted cells

cDNA levels from FAC-sorted cells were analysed on a Roche LightCycler® 480 II, using the Roche Universal ProbeLibrary® (UPL) assay. First, primers and probes were generated using the online UPL assay design center. The primer sequences and probes that were used are listed in appendix 8. All primer pairs were tested on an RNA dilution series from whole flies before they were used, and only primers with linear Ct values were used. The cDNA was always amplified in

batches of 8 primer pairs (2 primer pairs for housekeeping reference genes, and 6 for target genes). First, all primers were pooled, and template cDNA was amplified by PCR, using the HotStart ReadyMix (Kappa biosystem) PCR mix, and 18 PCR cycles. Next, primers were removed using a PCR purification kit (Qiagen). The sample was then split up into three technical replicates, and 8 separate assays per replicate. Finally, the primer pairs and UPL probes for each target- and reference gene was added, together with the UPL mastermix, and analysed in the LightCycler. -RT samples were treated the same way, and samples with gDNA contamination were discarded.

RT-qPCR data analysis

Ct-values, obtained from RT-qPCR assays, were used to test the relative levels of cDNA in each sample, using the $2^{-\Delta\Delta CT}$ method [98,99]. Results from technical triplicates were pooled, and normalised to the geometric mean Ct values of the two housekeeping genes (HG) *GAPDH* and *SdHA* (ΔCt of the target – ΔCt of HGs)[100]. (Note that for Figure 4, only *GAPDH* was used for normalisation.) Then, the average and standard deviation (SD) of ΔCts from each FACS fraction, and each independent biological replicate were calculated. Next, a ΔΔCt value was calculated for each gene or transposon, by subtracting the ΔCt form the rest of the brain from the αβ-KC ΔCts. A combined standard error of the mean (SEM) was calculated by first converting the individual SDs from each sample into SEM values, and then by taking the square root of the sum of squared individual SEMs.

Next, upper and lower limits were calculated by adding and subtracting the SEM from the average. Finally, the mean, upper and lower limit ΔΔCt were converted into fold changes using the formula: 2^(-1 x ΔΔCt).

Genomic DNA extraction and amplification

For gDNA, cells were sorted from groups of 50 flies, or single individuals. 2µl of PBS was added to each of the sorted cells, if the resulting volumes were less than 1µl. Next, gDNA was amplified with a Qiagen® REPLI-g Midi Kit, using the standard protocol[101]. gDNA was extracted using a protocol which was originally provided by Qiagen® online, but which has since become unavailable. The REPLI-g reaction was first equilibrated to room temperature, and 150µl of 96% ethanol were added. The sample was mixed carefully, and then centrifuged at maximum speed for 2 minutes. The supernatant was carefully removed, and the pellet was washed with 100µl of 70% ethanol (avoiding re-suspension of the pellet). The samples were centrifuged again for 2 minutes, and the supernatant was removed, making sure that no ethanol remained in the tube. After 5 minutes of incubation, the pellet was re-suspended in 50µl of TE (10mM Tris, brought to pH 8.0 with HCl, 1mM EDTA) buffer.

In order to avoid DNA contamination of the sample, all solutions, tubes, forceps and pipettes were UV-incubated for 10 minutes before they were used, and appropriate PPE was worn.

Whole-genome sequencing

WGS of short reads was performed by BGI. 500bp short-insert PCR-free libraries were constructed from the amplified gDNA, and 100bp paired-end reads were sequenced on an Illumina HiSeq2000 platform, with a theoretical coverage of 30x.

WGS of long reads was performed by Macrogen. TruSeq DNA PCR-free libraries with 350bp fragment sizes were prepared, and 250bp paired-end reads were sequenced on an Illumina HiSeq2500 platform, with a theoretical coverage of 30x (not taking into account the overlap of the two read pairs).

Data analysis - quality control

Each fastq data set was first tested with FastQC (see the individual reports on read quality in appendix 7). 250bp paired-end reads showed a significant loss of sequencing quality towards the end of the reads, and hence all reads were trimmed to 190bp fragments. On average, 40.243.741, 60.886,136 and 9.022,768 reads were sequenced for each sample in groups of flies, single flies with short reads, and single flies with long reads accordingly.

Formation of *in-silico* read pairs

In order to create 250bp fragments *in-silico*, long, overlapping reads were first assembled into contiguous DNA fragments with FLASH (Version 1.2.11[102]). Next, theses contigs were filtered in order to remove fragments that were smaller than 250bp (28.8%). Of the remaining reads, two 100bp sections, with a gap of 50bp in between, were separated, and merged to form 100bp paired-end reads of 250bp inserts. These reads were mapped to the three reference genomes (DMref, DMsim and DMclean).

DMsim

A set of IGEs was assembled to replace the set of transposon sequences. The IGEs were chosen by first putting together a list of 1,344 exons from the genes that are most highly expressed in adult *Drosophila melanogaster* testes (from FlyAtlas[103], microarray, adult Canton-S testis). The exons were ranked according to their size, and the largest 590 exons were chosen. Next, these 590 exons were removed from their original location in the reference genome, and attached to the end of chromosome 2L. This new reference genome (DMsim), and the list of 590 exons (the IGEs) was used for the *in-silico* simulated control experiments.

DMclean

A modified reference genome was produced by first masking all sequences that mapped to a portion of the set of transposon reference sequences, with RepeatMasker (version 4.06[104]), using the Repbase transposon database[105]. Next, the set of transposon sequences was assembled into a new, theoretical chromosome called 'TE'. Only one copy of each transposon family was used. Finally, the location of each transposon was added to the .gtf files, which contains the locations of all exons.

Alignments

Paired-end reads were aligned using both the Burrows-Wheeler Aligner (BWA)[83], and also TopHat[84]. The parameters that were used for both approaches can be found in appendix 1. *Drosophila melanogaster* reference genome release 5.57 was downloaded from flybase used for all results that are named DMref[106]. The same release was used as the basis for the creation of DMclean and DMsim. On average, 91% of all reads were mapped.

Reference transposon sequences

The set of transposon reference sequences was based on the *Drosophila melanogaster* RepBase transposon library[105]. The same set of transposon reference sequences was used throughout all analyses in this study.

TEMP

The TEMP algorithm has previously been described[67]. Briefly, TEMP first identifies all discordant read pairs in a sequence alignment file, and then maps them to a set of reference transposon sequences. The reads that map a section of these reference sequences are considered to be the first piece of evidence for an insertion at this particular location. Next, the putative insertion location is screened for additional reads that support the presence, or absence, of an insertion at this particular location. For rare, somatic insertion events, usually only one read pair exists, making it impossible to identify the precise breakpoint.

MRTemp

The MRTemp protocol is explained in detail in appendix 1-6. In summary, first putative insertions are identified with TEMP. Next, the corresponding merged long reads, which also contained the sequence of the 50bp gap between the two 100bp reads (and hence the precise breakpoint), are selected and mapped to a set of reference sequences (DMclean for transposons, DMsim for IGEs) with BLAST, in order to identify the exact location of the break point in the genome. In addition, the reads are also mapped to the reference transposon/IGE sequences, in order to determine the breakpoint on the element.

TEMP hovering filter

In order to identify potential somatic transposon insertions, raw TEMP hits were filtered for amplification artefacts on the basis of three criteria. First, the putative insertion should not be present in the sample of the rest of the brain from the same fly. Second, the hit should be absent from the reference genome. And third, the hit should not be within 5kb of another insertion of the same transposon type.

Bedtools

Processing of genome sequencing data was partly done using the bedtools suite[107].

Cuffdiff

In order to assess the copy numbers of reads which map to each transposon, each sample was aligned to the DMclean reference genome, with TopHat. The number of reads per gene (and transposon) was assessed with CuffDiff. More details, including the parameters that were used, can be found in appendix 1-6.

Fragment sizes

The distribution of fragment sizes, based on alignment results, was performed on the Galaxy web platform (version 1.136)[108], using the CollectInsertSizeMetrics function, with the following parameters: MINIMUM_PCT=0.05, DEVIATIONS=10. Fragment sizes of long, overlapping reads was done by measuring the fragment sizes from FLASH-merged reads before they were mapped.

5. Supplementary Material

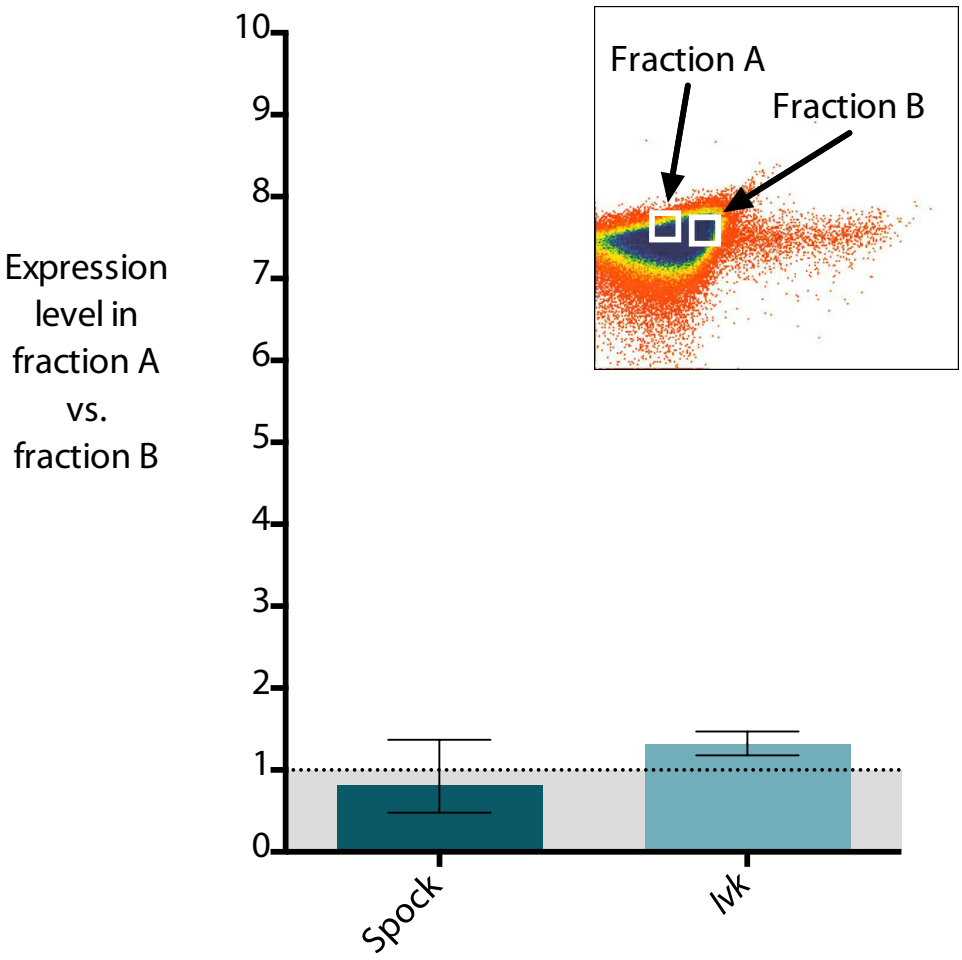
TABLE 1. Whole-genome sequencing data

Sample size	Age	Sample number	Tissue sample	Read lengths	Total reads	% of reads mapped	Mean coverage	Range of insert sizes of 90%	TEMP hits DMel5.57	TEMP hits DMelSIM	**TEMP DMel. 5.57 only αβ**	TEMP DMel SIM only αβ	# true IGEs detected
30 flies	YOUNG	1	αβ-KCs	100	41704908	95.11%	25.52	139-221nt	2978	155	**2306**	139	263/590
		2	r.o.b.	100	38782574	95.41%	24.01	141-237nt	4445	243			382/590
single fly	YOUNG	3	αβ-KCs	100	62628148	96.96%	39.43	434-500nt	23301	1901	**12827**	1899	590/590
		4	r.o.b.	100	62632266	94.96%	38.54	434-496nt	22162	1789			590/590
single fly	YOUNG	5	αβ-KCs	100	62687688	93.21%	38.02	432-496nt	22581	1858	**12783**	1846	590/590
		6	r.o.b.	100	62609858	96.01%	39.21	434-504nt	17224	1931			589/590
single fly	YOUNG	7	αβ-KCs	100	62490412	96.57%	39.13	438-508nt	23974	1817	**10977**	1808	590/590
		8	r.o.b.	100	60824768	93.45%	37.20	452-510nt	24887	1771			590/590
single fly	OLD	9	αβ-KCs	100	58143492	96.58%	36.99	450-506nt	17092	1345	**8133**	1339	590/590
		10	r.o.b.	100	57823896	97.90%	37.25	458-518nt	16692	1321			589/590
single fly	OLD	11	αβ-KCs	100	57628236	97.14%	36.97	458-514nt	18922	1524	**8954**	1516	590/590
		12	r.o.b.	100	58162210	96.96%	37.05	450-510nt	19254	1426			589/590
single fly	OLD	13	αβ-KCs	100	62618772	96.29%	38.72	437-497nt	24587	1675	**11616**	1671	590/590
		14	r.o.b.	100	62383890	90.60%	36.50	435-509nt	21849	1762			590/590
single fly	YOUNG	15	αβ-KCs	250	7304448	91.82%	4.70	250nt	1530	148	**759**	143	484/590
		16	r.o.b.	250	8369114	89.30%	5.17	250nt	1687	156			503/590
single fly	YOUNG	17	αβ-KCs	250	16459322	89.68%	10.45	250nt	2498	322	**1671**	316	454/590
		18	r.o.b.	250	8688532	87.18%	5.27	250nt	1828	186			501/590
single fly	YOUNG	19	αβ-KCs	250	8444438	78.12%	6.41	250nt	2189	61	**1544**	53	216/590
		20	r.o.b.	250	10393246	64.74%	5.72	250nt	1908	79			282/590
single fly	OLD	21	αβ-KCs	250	8846114	89.78%	5.60	250nt	1732	162	**1008**	158	457/590
		22	r.o.b.	250	8504752	91.85%	5.43	250nt	1814	187			506/590
single fly	OLD	23	αβ-KCs	250	7692272	90.51%	5.00	250nt	1627	173	**1017**	163	485/590
		24	r.o.b.	250	7378712	90.26%	5.01	250nt	1575	175			438/590
single fly	OLD	25	αβ-KCs	250	8298164	90.73%	5.39	250nt	1901	151	**1204**	139	433/590
		26	r.o.b.	250	7894110	84.21%	4.83	250nt	1661	143			422/590

Table 1. Whole-genome sequencing data and TEMP results.

Summary table of each DNA sequencing sample that was used for this study, including the summaries of TEMP analyses on transposons, as well as simulated immobile elements. Note that there is an approximate ratio of 10:1 between the putative hits of transposons and of simulated immobile elements, which mirrors the rate of total sequence length in the reference genome between the two groups of elements.

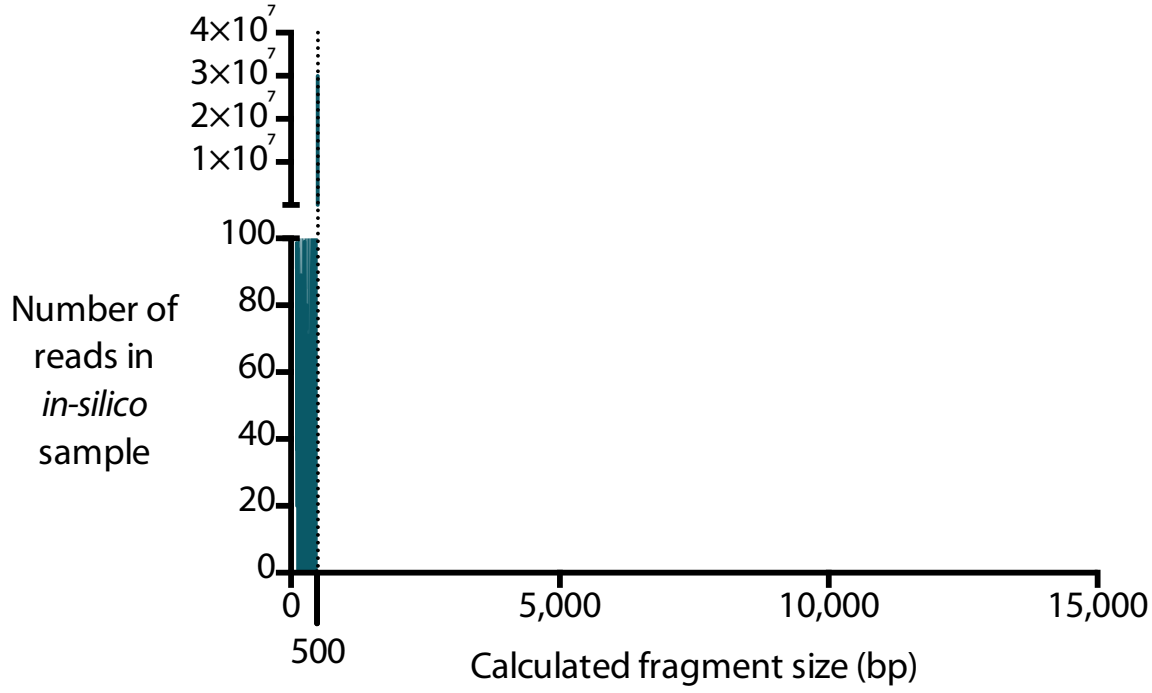
Supplementary Figure 1. Two flow through FACS fractions have similar expression patterns.



Supplementary Figure 1. Two flow through FACS fractions have similar gene expression patterns.

Plot showing the relative gene expression levels of Spock and *Ivk* in two fractions of *mCherry* negative cells. Error bars represent combined SEM of 2 separate experimental runs, with 50 flies and 3 technical replicates per run. Inset shows a schematic of how the FACS gates for the two fractions were chosen.

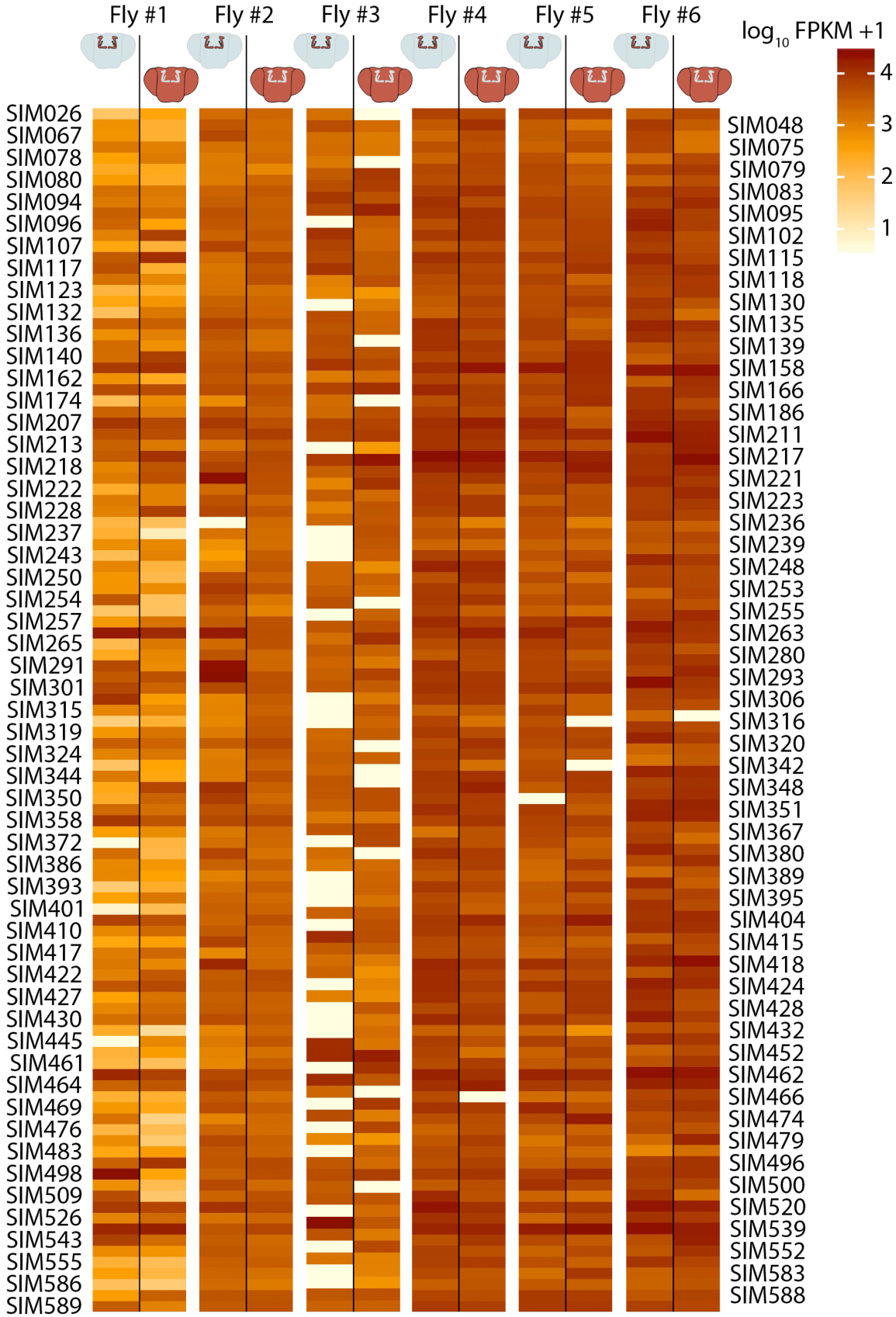
Supplementary Figure 2. Perfect reads do not result in predicted fragment sizes which are larger than the input size.



Supplementary Figure 2. Perfect reads do not result in predicted fragment sizes which are larger than the input size.

Plot showing the predicted input sizes, based on the distance of the two reads in each read pair. Note that none of the reads was found to be further apart than 500 nucleotides, the size of all input fragments.

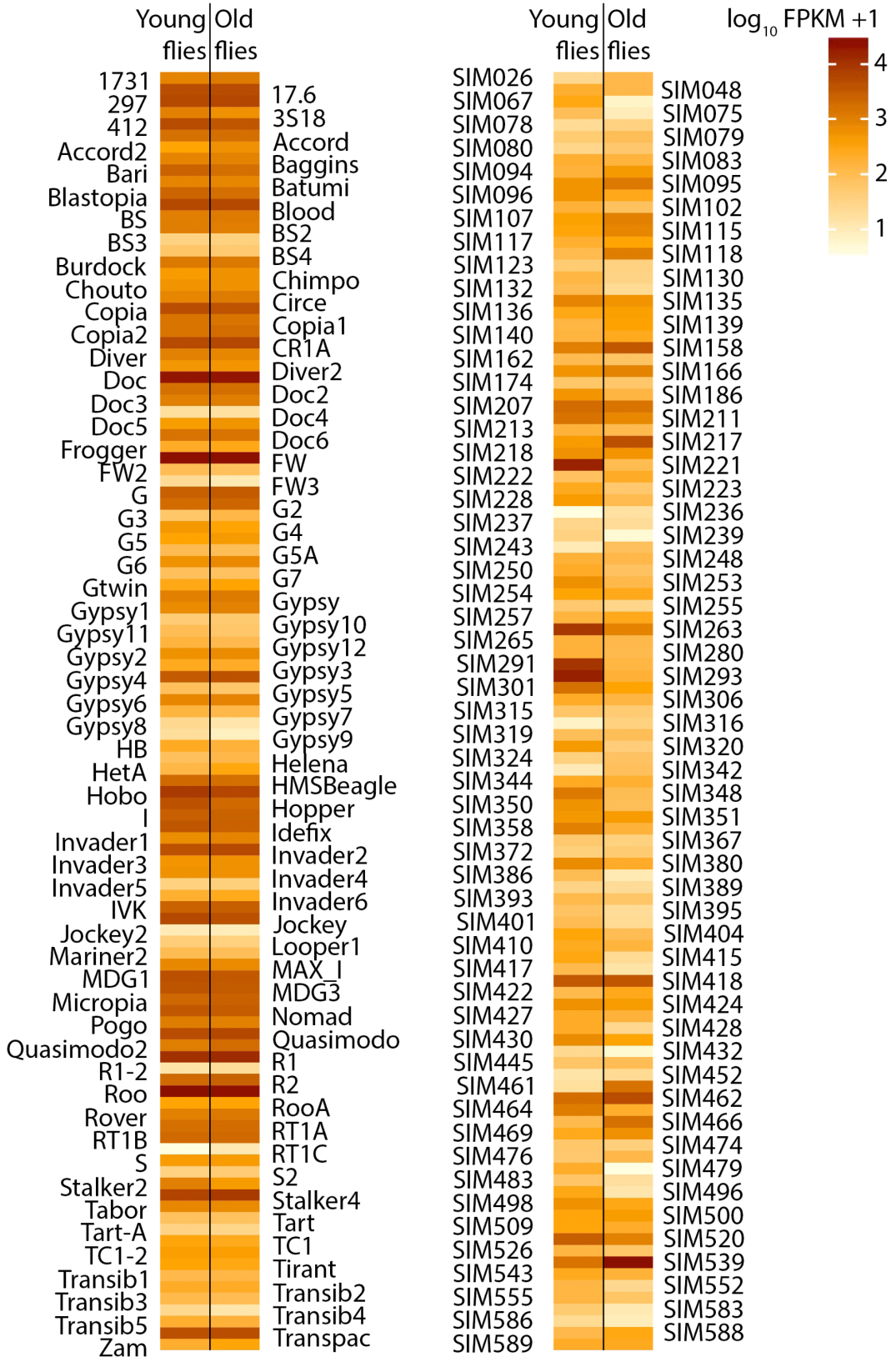
Supplementary Figure 3. Copy number of reads which map onto gene sequences varies between samples due to amplification biases.



Supplementary Figure 3. Copy number of reads which map onto gene sequences varies between samples due to amplification biases.

Heatmap showing the normalised number of sequencing reads that map onto one of 109 reference exon sequences that were used for control analyses in this study. Note that while each exon occurs twice per cell, a basal level of apparent copy number variation remains, possibly due to uneven sequencing coverage. Individuals #1 - #3 are young flies (3-4 days) and #4 - #6 are old flies (30 days).

Supplementary Figure 4. Number of reads which map onto transposon and control gene sequences do not change in old and young flies.



Supplementary Figure 4. Number of reads which map onto transposon and control gene sequences do not change in old and young flies.

Heatmap showing the average normalised number of reads for each transposon family in young and old flies.

6. References

1. McClintock, B. Induction of Instability at Selected Loci in Maize. Genetics 38, 579–599 (1953).
2. Britten, R. J. & Kohne, D. E. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science 161, 529–540 (1968).
3. Adams, M. D. et al. The genome sequence of Drosophila melanogaster. Science 287, 2185–2195 (2000).
4. Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).
5. Mouse Genome Sequencing Consortium et al. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562 (2002).
6. Kazazian, H. H. Mobile Elements: Drivers of Genome Evolution. Science 303, 1626–1632 (2004).
7. Levin, H. L. & Moran, J. V. Dynamic interactions between transposable elements and their hosts. Nature Reviews Genetics 12, 615–627 (2011).
8. Fedoroff, N. V. Presidential address. Transposable elements, epigenetics, and genome evolution. 338, 758–767 (2012).
9. J, G. R. Fly pushing: the theory and practice of Drosophila genetics[;L200 a]. (1997).
10. Arias, A. M. Drosophila melanogaster and the development of biology in the 20th century. Drosophila: Methods and Protocols (2008).
11. Morgan, T. H. The origin of nine wing mutations in Drosophila. Science 33, 496–499 (1911).
12. Morgan, T. H. Chromosomes and associative inheritance. Science 34, 636–638 (1911).
13. Nüsslein-Volhard, C. & Wieschaus, E. Mutations affecting segment number and polarity in Drosophila. Nature 287, 795–801 (1980).
14. Erwin, J. A., Marchetto, M. C. & Gage, F. H. Mobile DNA elements in

thegeneration of diversity and complexityin the brain. Nat. Rev. Neurosci. 15, 497–506 (2014).

15. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. Cell 100, 57–70 (2000).
16. Waddell, S., Barnstedt, O. & Treiber, C. Neural transposition in the Drosophila brain: is it all bad news? Adv. Genet. 86, 65–92 (2014).
17. McConnell, M. J. et al. Mosaic copy number variation in human neurons. Science 342, 632–637 (2013).
18. Ilyin, Y. V., Tchurikov, N. A. & Georgiev. Selection and some properties of recombinant clones of lambda bacteriophage containing genes of Drosophila melanogaster. Nucleic Acids Res. 3, 2115–2127 (1976).
19. Georgiev, G. P. et al. Isolation of eukaryotic DNA fragments containing structural genes and the adjacent sequences. Science 195, 394–397 (1977).
20. Strobel, E., Dunsmuir, P. & Rubin, G. M. Polymorphisms in the chromosomal locations of elements of the 412, copia and 297 dispersed repeated gene families in Drosophila. Cell 17, 429–439 (1979).
21. Kaminker, J. S. et al. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biology 3, RESEARCH0084 (2002).
22. Xiong, Y. & Eickbush, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. The EMBO journal (1990).
23. Bingham, P. M., Kidwell, M. G. & Rubin, G. M. The molecular basis of PM hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. Cell (1982).
24. Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318, 761–764 (2007).
25. Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: the vanguard of genome defence. Nature Publishing Group 12, 246–258 (2011).

26. Haase, A. D. Review ArticleA Small RNA-Based Immune System Defends Germ Cells against Mobile Genetic Elements. Stem Cells International 1–9 (2015). doi:10.1155/2016/7595791
27. Alves, G., Tatro, A. & Fanning, T. Differential methylation of human LINE-1 retrotransposons in malignant cells. Gene 176, 39–44 (1996).
28. Fire, A. et al. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391, 806–811 (1998).
29. Tolia, N. H. & Joshua-Tor, L. Slicer and the Argonautes. Nat Chem Biol 3, 36–43 (2007).
30. Czech, B. et al. An endogenous small interfering RNA pathway in Drosophila. Nature 453, 798–802 (2008).
31. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science 294, 858–862 (2001).
32. Förstemann, K., Horwich, M. D., Wee, L., Tomari, Y. & Zamore, P. D. Drosophila microRNAs Are Sorted into Functionally Distinct Argonaute Complexes after Production by Dicer-1. Cell 130, 287–297 (2007).
33. Ambros, V. The functions of animal microRNAs. Nature 431, 350–355 (2004).
34. Mugat, B. et al. MicroRNA-Dependent Transcriptional Silencing of Transposable Elements in Drosophila Follicle Cells. PLoS Genet 11, e1005194 (2015).
35. Brennecke, J. et al. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. Cell 128, 1089–1103 (2007).
36. Malone, C. D. et al. Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary. Cell 137, 522–535 (2009).
37. Gonzalez, J., Qi, H., Liu, N. & Lin, H. Piwi Is a Key Regulator of Both Somatic and Germline Stem Cells in the Drosophila Testis. CellReports 12, 150–161 (2015).

38. Ross, R. J., Weiner, M. M. & Lin, H. PIWI proteins and PIWI-interacting RNAs in the soma. Nature 505, 353–359 (2014).
39. Rowe, H. M. et al. KAP1 controls endogenous retroviruses in embryonicstem cells. Nature 463, 237–240 (2010).
40. Wolf, G. et al. The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. Genes & Development 29, 538–554 (2015).
41. Han, J. S., Szak, S. T. & Boeke, J. D. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429, 268–274 (2004).
42. Aminetzach, Y. T. Pesticide Resistance via Transposition-Mediated Adaptive Gene Truncation in Drosophila. Science 309, 764–767 (2005).
43. Narita, N. et al. Insertion of a 5'truncated L1 element into the 3'end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. Journal of Clinical Investigation 91, 1862 (1993).
44. Moran, J. V. Exon Shuffling by L1 Retrotransposition. Science 283, 1530–1534 (1999).
45. Xing, J. et al. Mobile elements create structural variation: Analysis of a complete human genome. Genome Research 19, 1516–1526 (2009).
46. Cridland, J. M., Macdonald, S. J., Long, A. D. & Thornton, K. R. Abundance and Distribution of Transposable Elements in Two Drosophila QTL Mapping Resources. Mol. Biol. Evol. (2013). doi:10.1093/molbev/mst129
47. Rebollo, R. et al. A Snapshot of Histone Modifications within Transposable Elements in Drosophila Wild Type Strains. PLoS ONE 7, e44253 (2012).
48. Jacobs, F. M. J. et al. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature 516, 242–245 (2014).

49. Kazazian, H. H. Mobile DNA transposition in somatic cells. BMC Biol 9, 62 (2011).
50. McConnell, M. J. et al. Mosaic copy number variation in human neurons. Science 342, 632–637 (2013).
51. Evrony, G. D. et al. Cell Lineage Analysis in Human Brain Using Endogenous Retroelements. Neuron 85, 49–59 (2015).
52. Perrat, P. N. et al. Transposition-Driven Genomic Heterogeneity in the Drosophila Brain. Science 340, 91–95 (2013).
53. Li, W. et al. Activation of transposable elements during aging and neuronal decline in Drosophila. Nat. Neurosci. 16, 529–531 (2013).
54. Muotri, A. R. et al. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435, 903–910 (2005).
55. Freeman, J. D., Goodchild, N. L. & Mager, D. L. A modified indicator gene for selection of retrotransposition events in mammalian cells. BioTechniques 17, 46–52 (1994).
56. Moran, J. V. et al. High frequency retrotransposition in cultured mammalian cells. Cell 87, 917–927 (1996).
57. Ostertag, E. M., Prak, E. T., DeBerardinis, R. J., Moran, J. V. & Kazazian, H. H. Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Res. 28, 1418–1423 (2000).
58. Muotri, A. R., Zhao, C., Marchetto, M. C. N. & Gage, F. H. Environmental influence on L1 retrotransposons in the adult hippocampus. Hippocampus 19, 1002–1007 (2009).
59. Coufal, N. G. et al. L1 retrotransposition in human neural progenitor cells. Nature 460, 1127–1131 (2009).
60. Muotri, A. R. et al. L1 retrotransposition in neurons is modulated by MeCP2. Nature 468, 443–446 (2011).
61. Pfeiffer, B. D. et al. Tools for neuroanatomy and neurogenetics in Drosophila. PNAS 105, 9715–9720 (2008).
62. Freund, R. & Meselson, M. Long terminal repeat nucleotide sequence and

specific insertion of the gypsy transposon. Proc. Natl. Acad. Sci. U.S.A. 81, 4462–4464 (1984).

63. Dej, K. J., Gerasimova, T., Corces, V. G. & Boeke, J. D. A hotspot for the Drosophila gypsy retroelement in the ovo locus. Nucleic Acids Res. 26, 4019–4024 (1998).

64. Wood, J. G. et al. Chromatin-modifying genetic interventions suppress age-associated transposable element activation and extend life span in Drosophila. PNAS 201604621 (2016). doi:10.1073/pnas.1604621113

65. Bergman, C. M. & Quesneville, H. Discovering and detecting transposable elements in genome sequences. Briefings in Bioinformatics 8, 382–392 (2007).

66. Baillie, J. K. et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479, 534–537 (2011).

67. Zhuang, J., Wang, J., Theurkauf, W. & Weng, Z. TEMP: a computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Res. 42, 6826–6838 (2014).

68. Khurana, J. S. et al. Adaptation to P Element Transposon Invasion in Drosophila melanogaster. Cell 147, 1551–1563 (2011).

69. Upton, K. R. et al. Ubiquitous L1 Mosaicism in Hippocampal Neurons. Cell 161, 228–239 (2015).

70. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol 7, 19 (2007).

71. Singer, T., McConnell, M. J., Marchetto, M. C. N., Coufal, N. G. & Gage, F. H. LINE-1 retrotransposons: mediatorsof somatic variation in neuronalgenomes? Trends in Neurosciences 33, 345–354 (2010).

72. Evrony, G. D., Lee, E., Park, P. J. & Walsh, C. A. Resolving rates of mutation in the brain using single-neuron genomics. Elife (2016). doi:10.7554/eLife.12966.001

73. Evrony, G. D. et al. Single-Neuron Sequencing Analysis of L1

Retrotransposition and Somatic Mutation in the Human Brain. Cell 151, 483–496 (2012).

74. Couto, A., Alenius, M. & Dickson, B. J. Molecular, Anatomical, and Functional Organization of the Drosophila Olfactory System. Current Biology 15, 1535–1547 (2005).

75. Keene, A. C. & Waddell, S. Drosophila olfactory memory: single genes to complex neural circuits. Nat. Rev. Neurosci. 8, 341–354 (2007).

76. Perez-Orive, J. et al. Oscillations and sparsening of odor representations in the mushroom body. Science 297, 359–365 (2002).

77. Tanaka, N. K., Tanimoto, H. & Ito, K. Neuronal assemblies of theDrosophila mushroom body. J. Comp. Neurol. 508, 711–755 (2008).

78. Krashes, M. J., Keene, A. C., Leung, B., Armstrong, J. D. & Waddell, S. Sequential Use of Mushroom Body Neuron Subsets during Drosophila Odor Memory Processing. Neuron 53, 103–115 (2007).

79. Krashes, M. J. & Waddell, S. Rapid Consolidation to a radish and Protein Synthesis-Dependent Long-Term Memory after Single-Session Appetitive Olfactory Conditioning in Drosophila. Journal of Neuroscience 28, 3103–3113 (2008).

80. Aso, Y. et al. The neuronal architecture of the mushroom body provides a logic for associative learning. Elife 3, e04577 (2014).

81. Crittenden, J. R., Skoulakis, E. M., Han, K. A., Kalderon, D. & Davis, R. L. Tripartite mushroom body architecture revealed by antigenic markers. Learn. Mem. 5, 38–51 (1998).

82. Bustin, S. A. et al. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clinical chemistry 55, 611–622 (2009).

83. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).

84. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578

(2012).

85. Eickbush, T. H. & Jamburuthugoda, V. K. The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Research 134, 221–234 (2008).

86. Van Schaik, N. W. & Brink, R. A. Transpositions of modulator, a component of the variegated pericarp allele in maize. Genetics (1959).

87. Tower, J., Karpen, G. H., Craig, N. & Spradling, A. C. Preferential transposition of Drosophila P elements to nearby chromosomal sites. Genetics 133, 347–359 (1993).

88. Keam, S. P. et al. The human Piwi protein Hiwi2 associates with tRNA-derived piRNAs in somatic cells. Nucleic Acids Res. 42, 8984–8995 (2014).

89. Xu, J., Kittler, E., Zapp, M. L., Weng, Z. & Zamore, P. D. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. ... (2008).

90. Yan, Z. et al. Widespread expression of piRNA-like molecules in somatic tissues. Nucleic Acids Res. 39, 6596–6607 (2011).

91. Subkhankulova, T. & Livesey, F. J. Comparative evaluation of linear and exponential amplification techniques for expression profiling at the single-cell level. (Genome biology, 2006).

92. Subkhankulova, T., Gilchrist, M. J. & Livesey, F. J. Modelling and measuring single cell RNA expression levels find considerable transcriptional differences among phenotypically identical cells. BMC Genomics 9, 268 (2008).

93. Subkhankulova. Grouping and classifying electrophysiologically-defined classes of neocortical neurons by single cell, whole-genome expression profiling. Front. Mol.Neurosci. (2010). doi:10.3389/fnmol.2010.00010

94. Pfeiffer, B. D. et al. Refinement of Tools for Targeted Gene Expression in Drosophila. Genetics 186, 735–755 (2010).

95. Wu, J. S. & Luo, L. A protocol for dissecting Drosophila melanogaster

brains for live imaging or immunostaining. Nat Protoc 1, 2110–2115 (2006).

96. Rio, D. C., Ares, M., Hannon, G. J. & Nilsen, T. W. Purification of RNA Using TRIzol (TRI Reagent). Cold Spring Harbor Protocols 2010, pdb.prot5439–pdb.prot5439 (2010).

97. Nagoshi, E. et al. Dissecting differential gene expression within the circadian neuronal circuit of Drosophila. Nat. Neurosci. 13, 60–68 (2009).

98. Livak, K. J. & Schmittgen, T. D. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2–ΔΔCT Method. Methods 25, 402–408 (2001).

99. Nolan, T., Hands, R. E. & Bustin, S. A. Quantification of mRNA using real-time RT-PCR. Nat Protoc 1, 1559–1582 (2006).

100. Barber, R. D. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. Physiological Genomics 21, 389–395 (2005).

101. Dean, F. B. et al. Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl. Acad. Sci. U.S.A. 99, 5261–5266 (2002).

102. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957–2963 (2011).

103. Chintapalli, V. R., Wang, J. & Dow, J. A. T. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat. Genet. 39, 715–720 (2007).

104. Bedell, J. A., Korf, I. & Gish, W. MaskerAid: a performance enhancement to RepeatMasker. Bioinformatics 16, 1040–1041 (2000).

105. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. Trends in Genetics 16, 418–420 (2000).

106. Flybase Consortium. FlyBase: the Drosophila database. Nucleic Acids Res. 24, 53–56 (1996).

107. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature

Analysis. Current protocols in bioinformatics 11.12.1–11.12.34 (John Wiley & Sons, Inc., 2002). doi:10.1002/0471250953.bi1112s47

108. Afgan, E. et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 44, W3–W10 (2016).

7. Abbreviations

αβ-KCs	alpha-beta Kenyon cells
bp	basepairs
BWA	Burrows-Wheeler alignment
cDNA	complementary DNA
DNA	deoxyribonucleic acid
E.coli	*Escherichia coli*
EGFP	enhanced GFP
endo-siRNA	endogenous small interfering RNA
FACS	fluorescently labelled cell sorting
Fas2	fasciclin 2
FRT	flippase recombinant target
FSC	forward scatter
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
gDNA	genomic DNA
GFP	green fluorescent protein
HG	housekeeping gene
IGE	immobile genetic element

kb	kilobases
L1	LINE-1
LINE	long interspersed element
LTR	long terminal repeats
MDA	multiple displacement amplification
miRNA	micro RNA
mRNA	messenger RNA
MRTemp	merged reads TEMP
NGS	normal goat serum
nt	nucleotides
PBS	phosphate buffered saline
PBT	PBS + Triton X-100
PCR	polymerase chain reaction
piRNA	PIWI interacting RNA
PIWI	P element induced wimpy testis
RC-seq	retrotransposon capture sequencing
RISC	RNA induced silencing complex
RNA	ribonucleic acid

RNA-seq	RNA sequencing
RNAi	RNA interference
RoB	rest of the brain
RT	reverse transcriptase
RT-qPCR	real-time quantitative polymerase chain reaction
SD	standard deviation
SdHA	succinate dehydrogenase complex flavoprotein subunit A
SEM	standard error of the mean
SSC	side scatter
TE	transposable element
TEMP	transposable element movement in population
TIR	terminally inverted repeat
TSD	target site duplication
UPL	universal probe library
WGS	whole genome sequencing

8. Appendices

Appendix 1 – Unix codes

BWA alignment

```
>bwa aln -n 3 -l 100 -R DMel.557.fa reads_1.fastq >
reads_1.sai

>bwa aln -n 3 -l 100 -R DMel.557.fa reads_2.fastq >
reads_2.sai

>bwa sampe DMel.557.fa reads_1.sai reads_2.sai
reads_1.fastq reads_2.fastq > aln.sam

>samtools view -Sb aln.sam > aln.bam

>samtools sort aln.bam > aln.sorted.bam

>samtools index aln.sorted.bam

>samtools flagstat aln.sorted.bam
```

TEMP algorithm

```
>TEMP_Insertion.sh -i aln.sorted.bam -s /PATH/TO/scripts -r
TE.fa -t TE.locations.bed -m 3 -f 500
```

Cuffdiff

```
>tophat2 -o tophat_out_A DMelCLEAN.fa reads_A_1.fastq
reads_A_2.fastq

>tophat2 -o tophat_out_B DMelCLEAN.fa reads_B_1.fastq
reads_B_2.fastq


>cuffdiff -o cuffdiff_out DMelCLEAN.allTEs.gtf
tophat_out_A/accepted_hits.bam DMelCLEAN.allTEs.gtf
tophat_out_B/accepted_hits.bam
```

For multiple samples

```
>samtools view -h -o tophat_A1.sam
tophat_out_A1/accepted_hits.bam

>samtools view -h -o tophat_A2.sam
tophat_out_A2/accepted_hits.bam

>samtools view -h -o tophat_B1.sam
tophat_out_B1/accepted_hits.bam

>samtools view -h -o tophat_B2.sam
tophat_out_B2/accepted_hits.bam


>cuffdiff -o cuffdiff_out DMelCLEAN.allTEs.gtf
tophat_A1.sam,tophat_A2.sam tophat_B1.sam,tophat_B2.sam
```

MRTemp

```
1. Crop reads

> on Galaxy:

Tool ID: timmer

Parameters:

Trim this column only: Not available.

Trim from the beginning up to this position: 1

Remove everything from this position to the end: 190

Is input dataset in fastq format?: Yes

Ignore lines beginning with these characters: Nothing
selected.


2. Merge reads

> flash -z -o sample.merged reads.cropped_1.fastq
reads.cropped_2.fastq
```

3. Filter to get all merged reads that are larger than 249 nucleotides

> on Galaxy:

Tool ID: toolshed.g2.bx.psu.edu/repos/devteam/fastq_filter/fastq_filter/1.0.0

Parameters:

Minimum Size: 250

Maximum Size: Not available.

Minimum Quality: Not available.

Maximum Quality: Not available.

Maximum number of bases allowed outside of quality range: Not available.

This is paired end data: Not available.

4. Create _1 file

> on Galaxy:

Tool ID: trimmer

Parameters:

Trim this column only: Not available.

Trim from the beginning up to this position: 1

Remove everything from this position to the end: 100

Is input dataset in fastq format?: Yes

Ignore lines beginning with these characters: Nothing selected.

5. Create _2 file

> on Galaxy:

Tool ID: trimmer

Parameters:

Trim this column only: Not available.

Trim from the beginning up to this position: 151

Remove everything from this position to the end: 250

Is input dataset in fastq format?: Yes

Ignore lines beginning with these characters: Nothing selected.

&

Tool ID: toolshed.g2.bx.psu.edu/repos/devteam/fastx_reverse_complement/cshl_fastx_reverse_complement/1.0.0

to create reverse-complement reads.

6. Append "_1"

> run mybin/add_01.sh -i reads_1.fastq (see appendix 2)

7. Append "_2"

> run mybin/add_02.sh -i reads_2.fastq (see appendix 3)

8. BWA (as above)

output is sample.merged.reads.bwa.sorted.bam

9. TEMP (as above)

output is sample.unpair.uniq.transposons.bed

10. Extract TEMP hits

In Microsoft Excel

> Import sample.unpair.uniq.transposons.bed

> Sort by columns A, B, C

> Copy column D to Textwrangler, add new line at the end,

save as sample.READS.txt

11. Extract reads

> ../READ_extractor.sh -i sample.READS.txt -s /PATH/TO/merged_reads.fastq -r DMclean.fa
(merged_reads.fastq come from step 2; for READ_extractor see appendix 4)

-> output is sample.READS.sequences.txt -> import to Excel file in column G-M

12. Complete Excel file

> add "=IF((H1<10),L1,K1)" to column N

-> this is the breakpoint

> add "=J1&CHAR(9)&(N1-10)&CHAR(9)&(N1+10)&CHAR(9)&E1" to column O

13. TE BLAST merged reads

> creat BLAST database from TE reference sequences

-> this creates folder with TE.DB.fa, TE.DB.fa.nhr, TE.DB.fa.nin and TE.DB.fa.nsq

> run TEBLAST_queries.sh -i sample.READS.txt -s TE.DB.fa
(see appendix 5)

> import output in Excel tab "READS", column O

> in W, enter "=IF((P1<10),T1,S1)"

> in X, enter "=U1-W1"

> in Y, enter "=IF((H1<10),(IF(V1="minus","downstream bit","upstream bit")),(IF(V1="minus","upstream bit","downstream bit")))"

> in Z, enter "=IF((Y1="downstream bit"), 1, W1)"

> in AA, enter "-"

> in AB, enter "=IF((Y1="downstream bit"), W1, U1)"

> in AC, enter “=IF(Y1="downstream bit", "d", "u")”

> in AD, enter “=IF((Y1="downstream bit"), 0, W1/U1)”

> in AE, enter “-“

> in AF, enter “=IF((Y1="downstream bit"), W1/U1, 1)”

> in AB, enter “=J1&CHAR(9)&(N1-10)&CHAR(9)&(N1+10)&CHAR(9)&R1&"|"&AC1&"|"&AD1&"|"&AF1”

> filter column O for “no hits”

> copy AB (until “no hit” lines) to Textwrangler, remove all “, save as sample.breakpoints.bed

14. Merge repeated hits

> bedtools merge -nms -delim “|” -i sample.breakpoints.bed > sample.breakpoints.merged.bed

-> import this file to new tab in Excel sheet

15. Edit merged hits

in this new tab in Excel named “MRTemp”

> before column D, insert column and enter “=C1-10”

> then before E, insert column and enter “=A1&”:”&D1”-“&D1

> then before F, insert 7 columns

> in column F enter “=COUNTA(G1:ZZ1)/4

16. Find coverage

> copy column E to Textwrangler and save as sample.breakpoints.forcov.bed

> run Extract_coverage.sh -i merged.reads.bwa.sorted.bam -r sample.brekpoints.forcov.bed (see appendix 6)

> rename file statistic.txt to sample.breakpoints.coverage.txt

17. Finish Excel file

> insert empty line as new first line

> import sample.breakpoints.coverage.txt to G

> in column H enter “=F2/(F2+G2)”

> in column I enter "
=MIN(O2,S2,W2,AA2,AE2,AI2,AM2,AQ2,AU2,AY2,BC2,BG2,BJ2,BN2,B
R2,BV2,BZ2,CD2,CH2,CL2,CP2,CT2,CX2,DB2,DF2,DJ2,DN2,DR2,DV2,
DZ2,ED2,EH2,EL2,EP2,ET2,EX2,FB2,FF2,FJ2)"

> in column J enter "
=MAX(P2,T2,X2,AB2,AF2,AJ2,AN2,AR2,AV2,AZ2,BD2,BH2,BK2,BO2,B
S2,BW2,CA2,CE2,CI2,CM2,CQ2,CU2,CY2,DC2,DG2,DK2,DO2,DS2,DW2,
EA2,EE2,EI2,EM2,EQ2,EU2,EY2,FC2,FG2,FK2)"

> in column K enter "=(J2-I2)*100"

> in column L enter "
=A2&CHAR(9)&B2&CHAR(9)&C2&CHAR(9)&M2&"|"&F2&"|"&G2&"|"&I2&"
|"&J2"

Appendix 2 – Shell script to append _1

```
#!/bin/bash

usage() {
echo Usage: $0 -i input_file.fastq
}
while getopts i:h o
do
        case $o in
                i)
                    FILE1=$OPTARG ;;
                [?])
                    usage && exit 1;;
        esac
done

if [ -z "$FILE1" ]
then
    usage
    exit
fi

mkdir ~/Desktop/TEMPFOLDER1
pwd > ~/Desktop/TEMPFOLDER1/oripath1
cd ~/Desktop/TEMPFOLDER1
basename "$FILE1" > name1
a=$( cat oripath1 )/$( cat name1 )

cp $a global1
x=0
while read line
do
    ((x++))
    if [ "$x" -eq 1 ]; then
        echo -e "$line" > templine1
        while read g1 g2
            do
            echo $g1"#/1" >> output_1.fastq
            done < ./templine1
    elif [ "$x" -eq 2 ]; then
        echo -e "$line" >> output_1.fastq
    elif [ "$x" -eq 3 ]; then
        echo -e "$line" >> output_1.fastq
    elif [ "$x" -eq 4 ]; then
        echo -e "$line" >> output_1.fastq
        x=0
    fi
done < global1
rm global1
rm ./templine1

cat name1 | rev | cut -c 9- | rev > name1b
b=$( cat oripath1 )/$( cat name1b ).renamed_1.fastq
mv output_1.fastq $b
rm name1
rm name1b
rm oripath1
rm -rd ~/Desktop/TEMPFOLDER1
```

Appendix 3 – Shell script to append _2

```
#!/bin/bash

usage() {
echo Usage: $0 -i input_file.fastq
}
while getopts i:h o
do
        case $o in
                i)
                    FILE2=$OPTARG ;;
                [?])
                    usage && exit 1;;
        esac
done

if [ -z "$FILE2" ]
then
    usage
    exit
fi

mkdir ~/Desktop/TEMPFOLDER2
pwd > ~/Desktop/TEMPFOLDER2/oripath
cd ~/Desktop/TEMPFOLDER2
basename "$FILE2" > name2
a=$( cat oripath )/$( cat name2 )

cp $a global2
x=0
while read line
do
    ((x++))
    if [ "$x" -eq 1 ]; then
        echo -e "$line" > templine2
        while read g1 g2
            do
            echo $g1"#/2" >> output_2.fastq
            done < ./templine2
    elif [ "$x" -eq 2 ]; then
        echo -e "$line" >> output_2.fastq
    elif [ "$x" -eq 3 ]; then
        echo -e "$line" >> output_2.fastq
    elif [ "$x" -eq 4 ]; then
        echo -e "$line" >> output_2.fastq
        x=0
    fi
done < global2
rm global2
rm ./templine2

cat name2 | rev | cut -c 9- | rev > name2b
b=$( cat oripath )/$( cat name2b ).renamed_2.fastq
mv output_2.fastq $b
rm name2
rm name2b
rm oripath
rm -rd ~/Desktop/TEMPFOLDER2
```

Appendix 4 – Shell script to extract reads

```
#!/bin/bash

usage() {
echo Usage: $0 -i list.of.read.names -s source.file.fastq -r masked.reference.genome
}
while getopts i:s:r:t: o
do
        case $o in
                i)
                    FILE1=$OPTARG ;;
                s)
                    FILE2=$OPTARG ;;
                r)
                    FILE3=$OPTARG ;;
                t)
                    FILE4=$OPTARG ;;
                [?])
                    usage && exit 1;;
        esac
done

if [ -z "$FILE1" ] || [ -z "$FILE2" ] || [ -z "$FILE3" ] || [ -z "$FILE4" ]
then
    usage
    exit
fi

name_input=$(basename "$FILE1")
name_inputb=$(echo $name_input | rev | cut -c 5- | rev)
name_source=$(basename "$FILE2")
name_TEs=$(basename "$FILE4")
x=$(echo Tempfolder_)$(echo "$name_inputb")
oripath=$(pwd)
a_input=$(echo "$oripath")/$(echo "$name_input")
c=$(echo "$oripath")/$(echo "$name_inputb").sequences.txt

mkdir $(echo ~/Desktop/)$(echo "$x")
cd $(echo ~/Desktop/)$(echo "$x")

cp $a_input global1
cp $FILE2 sourcefile

while read line
do
        spa=$(echo -e "$line" | rev | cut -c 3- | rev)
        grep -A 1 $spa ./sourcefile | tail -n +2 > localvar
        cat localvar | blastn -db $FILE3 -outfmt "6 qstart qend sseqid sstart send s
        tmp=$(cat ./var1)
        null=$(echo "")
        if [ "$tmp" == "$null" ]; then
        echo "no-hit" > var1
        fi

        cat localvar | blastn -db $FILE4 -outfmt "6 qstart qend sseqid sstart send s

        tmp2=$(cat ./var2)
        if [ "$tmp2" == "$null" ]; then
        echo "no-hit" > var2
        fi
        paste localvar var1 var2 >> global2
```

Appendix 5 – Shell script BLAST to genome

```
#!/bin/bash

usage() {
echo Usage: $0 -i list.of.sequences -s masked.reference.genome
}
while getopts i:s: o
do
        case $o in
                i)
                    FILE1=$OPTARG ;;
                s)
                    FILE2=$OPTARG ;;
                [?])
                    usage && exit 1;;
        esac
done

if [ -z "$FILE1" ] || [ -z "$FILE2" ]
then
    usage
    exit
fi

name_input=$(basename "$FILE1")
name_inputb=$(echo $name_input | rev | cut -c 5- | rev)
name_source=$(basename "$FILE2")
x=$(echo Tempfolder_BLAST_)$(echo "$name_inputb")
oripath=$(pwd)
a_input=$(echo "$oripath")/$(echo "$name_input")
c=$(echo "$oripath")/$(echo "$name_inputb").blastresults.txt

mkdir $(echo ~/Desktop/)$(echo "$x")
cd $(echo ~/Desktop/)$(echo "$x")

cp $a_input global1

while read line
do
        echo $line | blastn -db $FILE2 -outfmt "6 qstart qend sseqid sstart send sle
        tmp=$(cat ./var)
        null=$(echo "")
        if [ "$tmp" == "$null" ]; then
        echo "no-hit" > var
        fi
        cat var >> global2
done < global1
rm global1

mv global2 $c
rm -rd $(echo ~/Desktop/)$(echo "$x")

```

Appendix 6 – Shell script to extract coverage

```
#!/bin/bash

usage() {
echo Usage: $0 -i input_file_aln.sorted.bam -r regions_list.txt
}

while getopts i:r:h o
do
        case $o in
                i)
                    FILE=$OPTARG ;;
                r)
                    REGION=$OPTARG ;;
                [?])
                    usage && exit 1;;
        esac
done

if [ -z "$FILE" ] || [[ -z "$REGION" ]]
then
    usage
    exit
fi

####
# This is creating the text file that will include the stats
echo -e "Coverage" > statistics.txt
####

####
# while loop extracts f1 = chromosome:Start-End
while read f1;
####

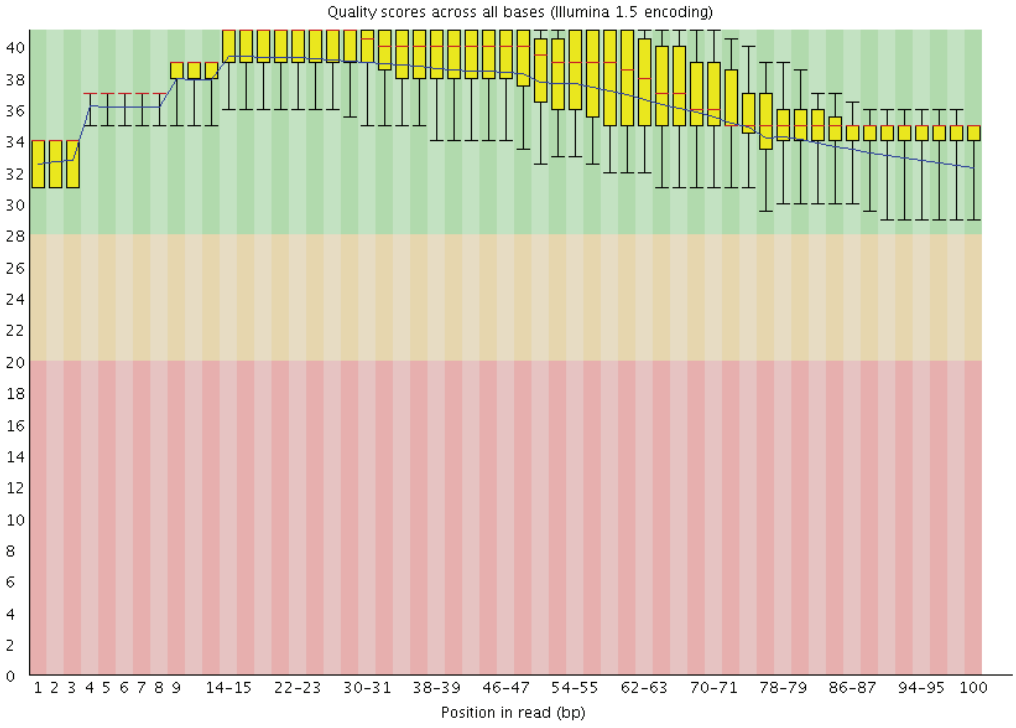
do

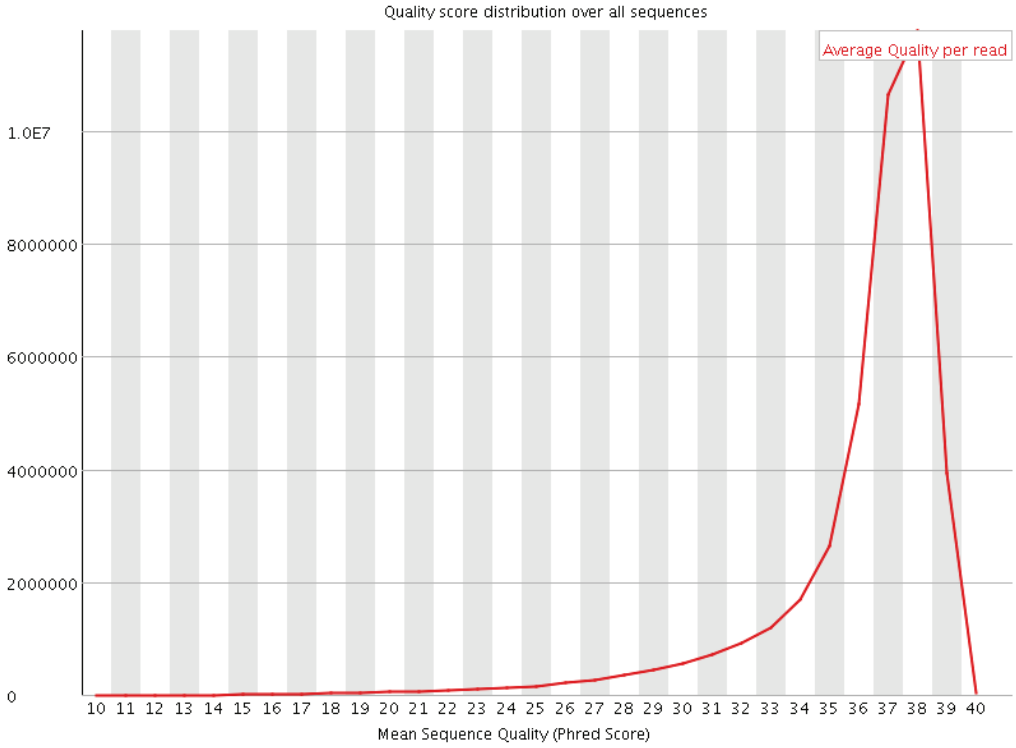
####
# This tests how many reads are in the region that is being analysed
samtools view $FILE $f1 | wc -l >> statistics.txt
####

done < $REGION
```

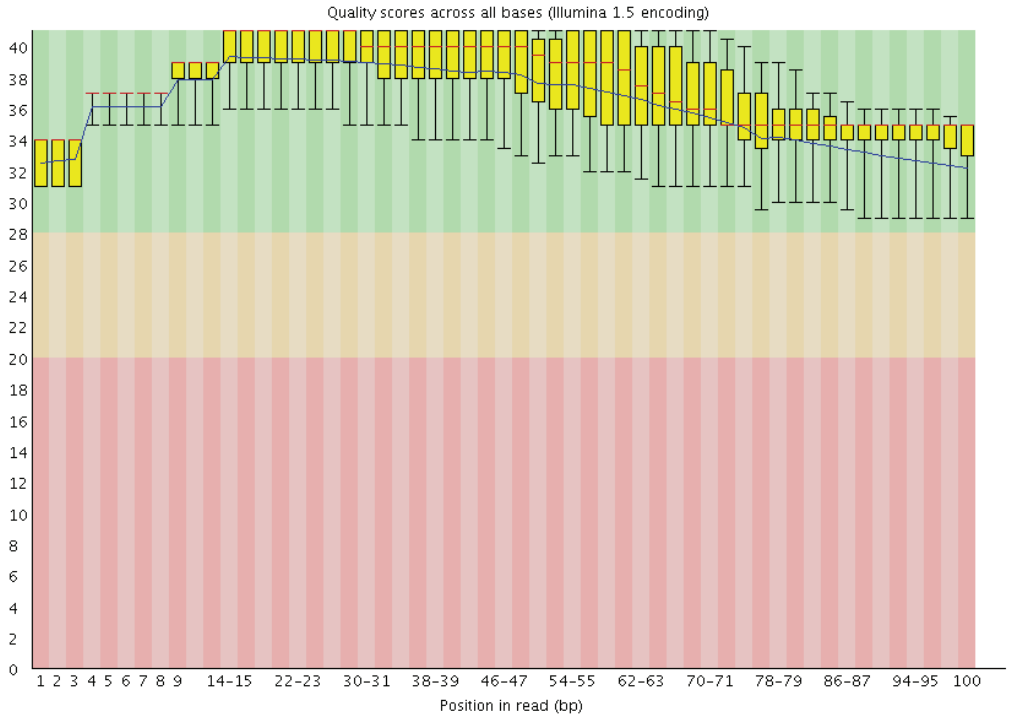
Appendix 7 – FastQC reports

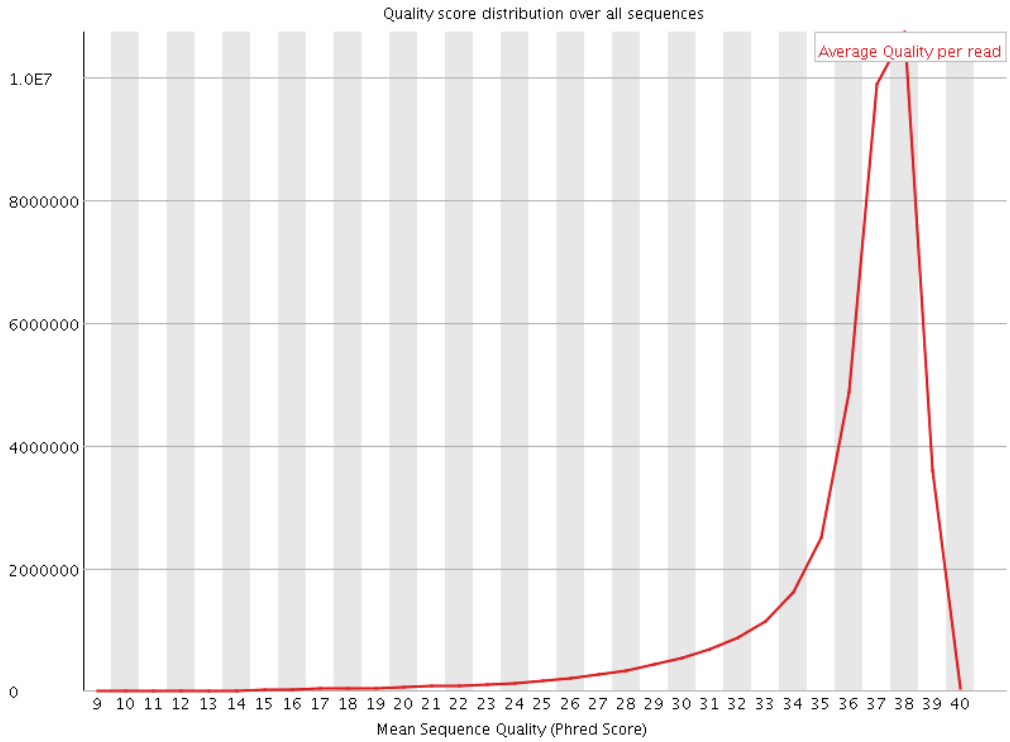
Sample 01 - group of 30 young flies - αβ-KCs - 41,704,908 reads



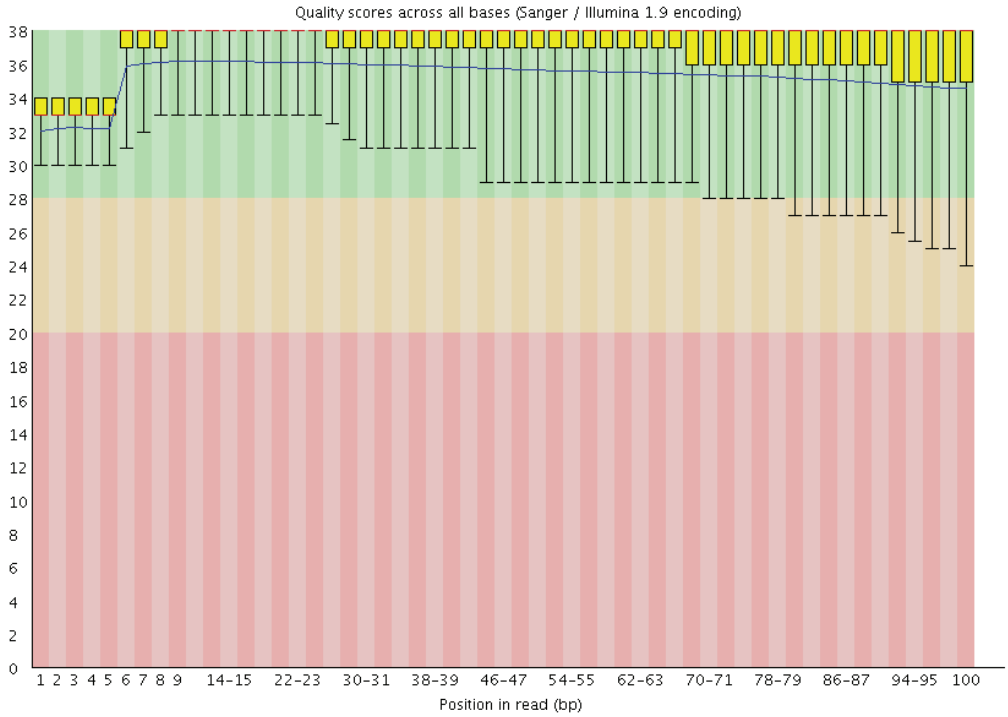


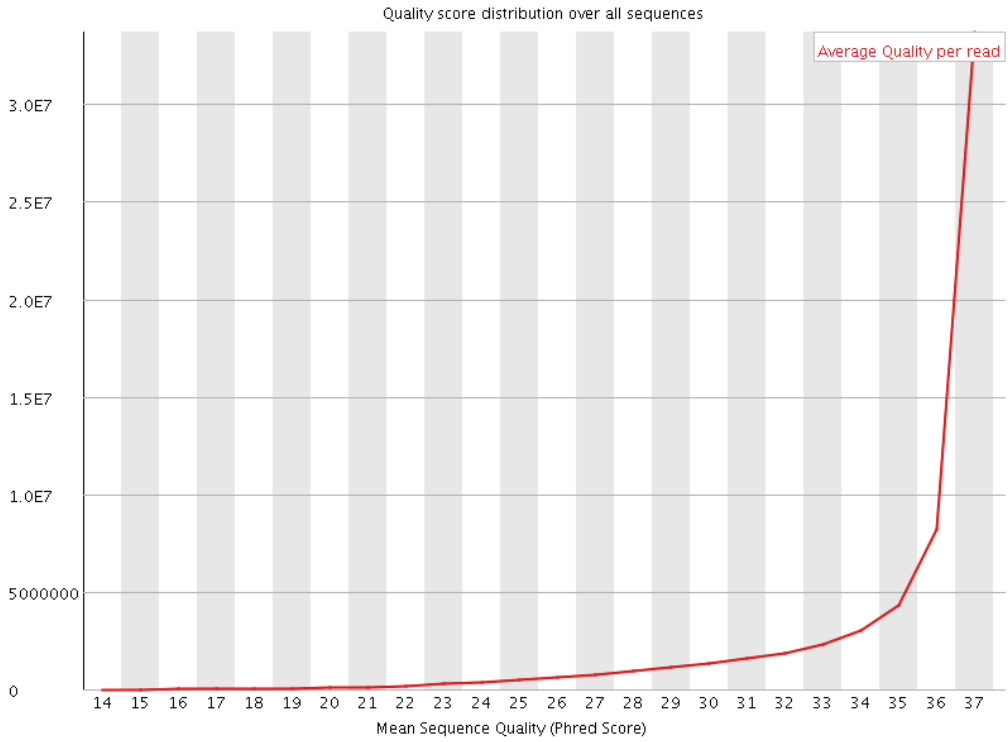
Sample 02 - group of 30 young flies - RoB - 38,782,574 reads



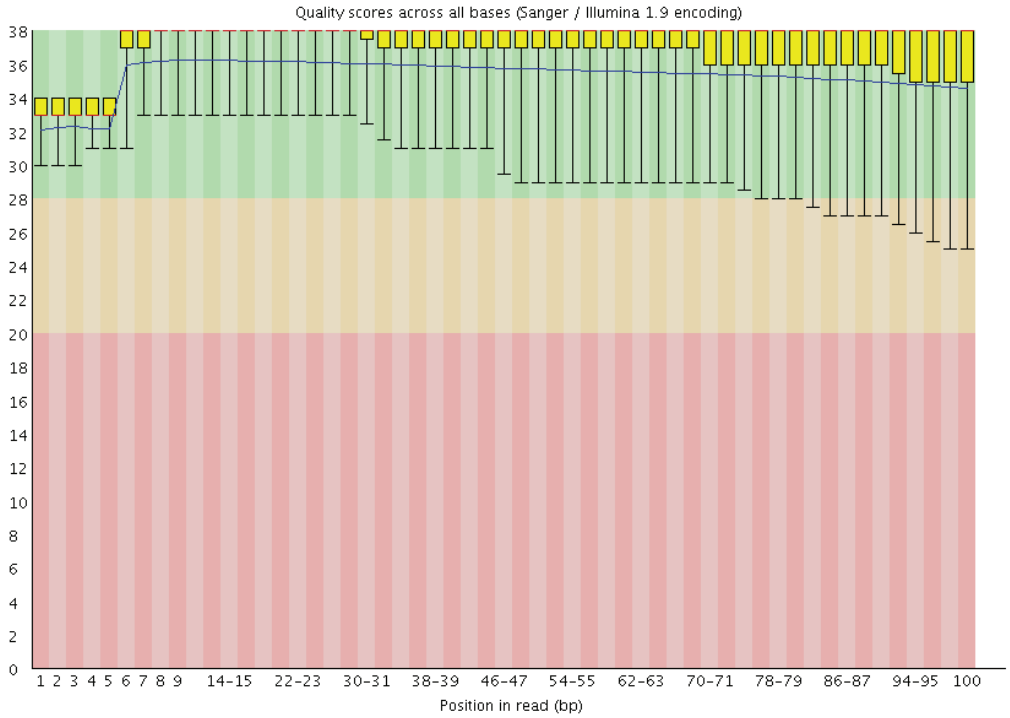


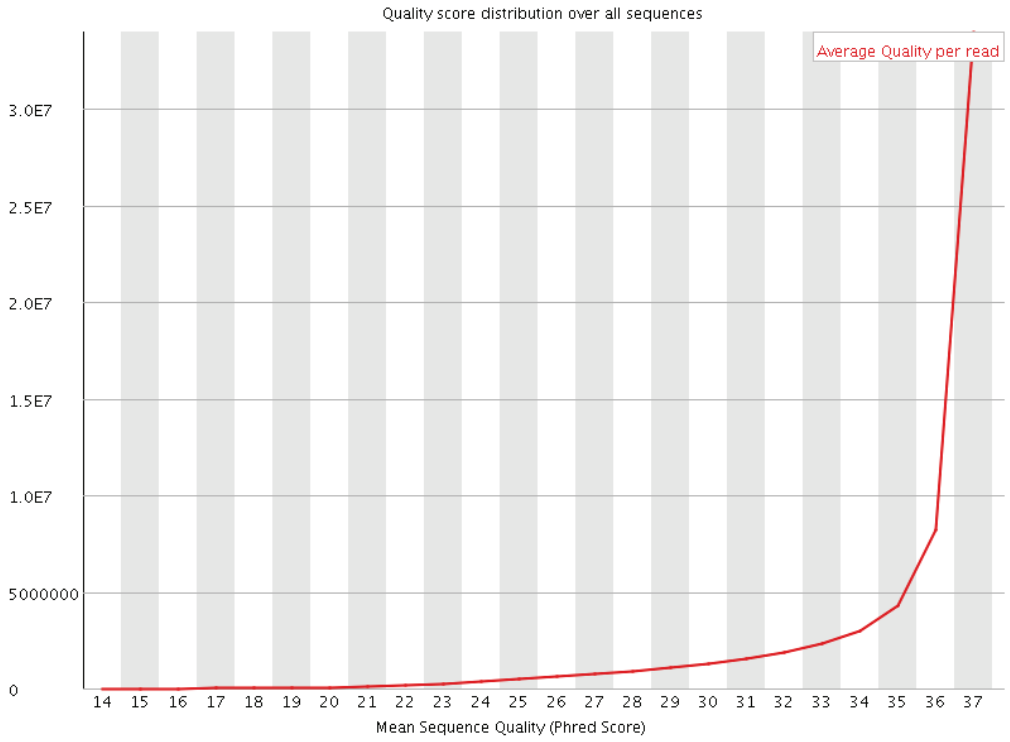
Sample 03 - single young fly - αβ-KCs - 62,628,148 reads



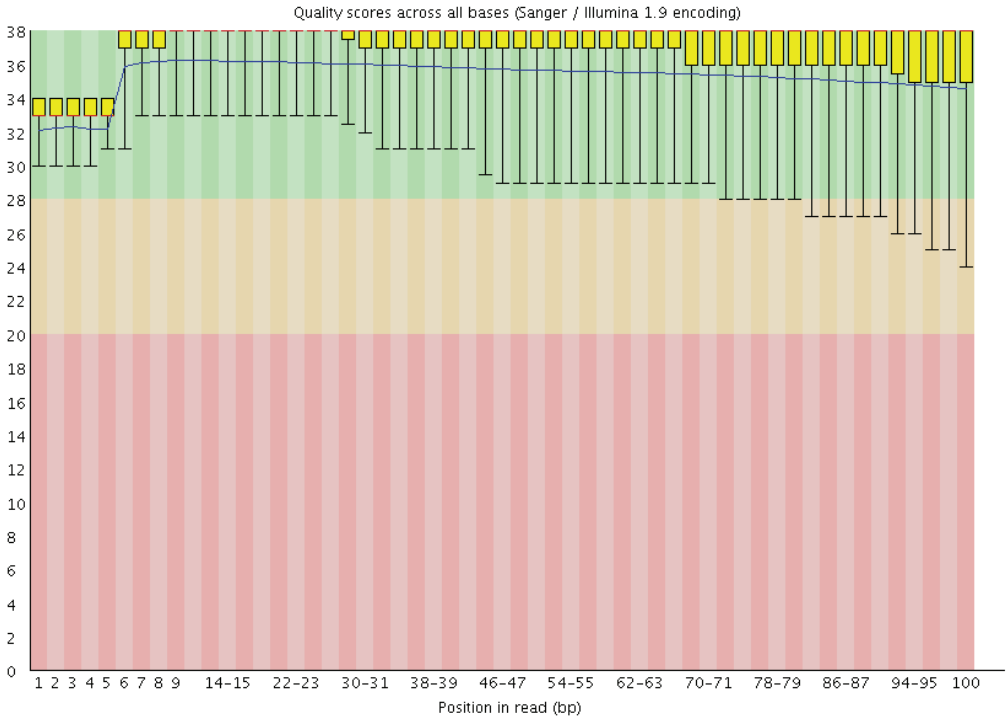


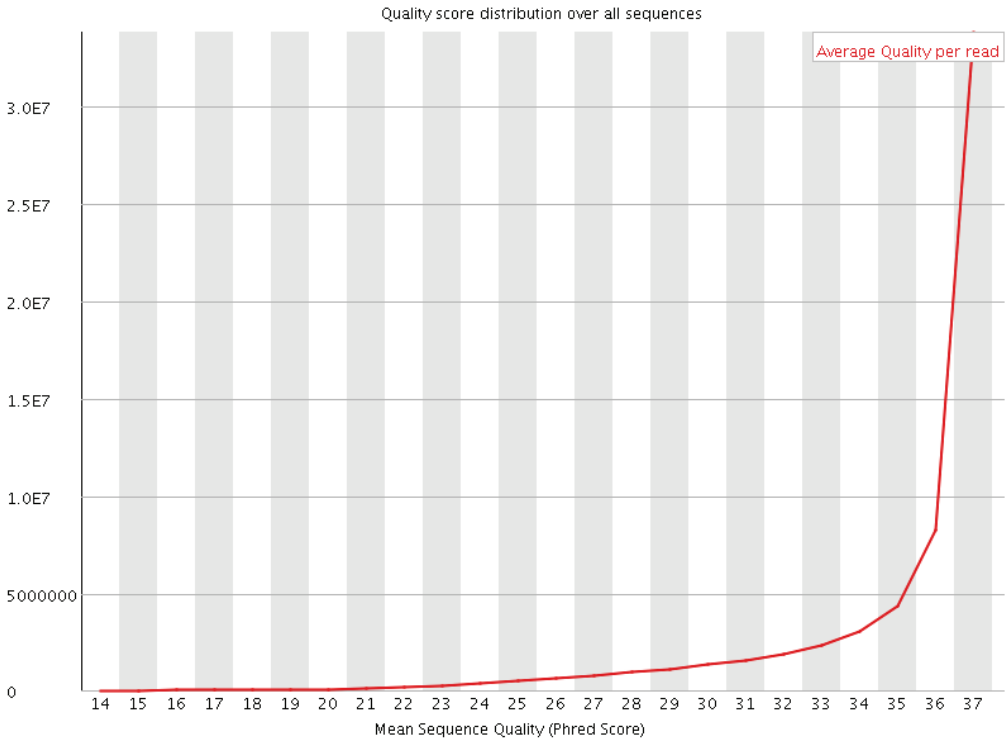
Sample 04 - single young fly - RoB - 62,632,266 reads



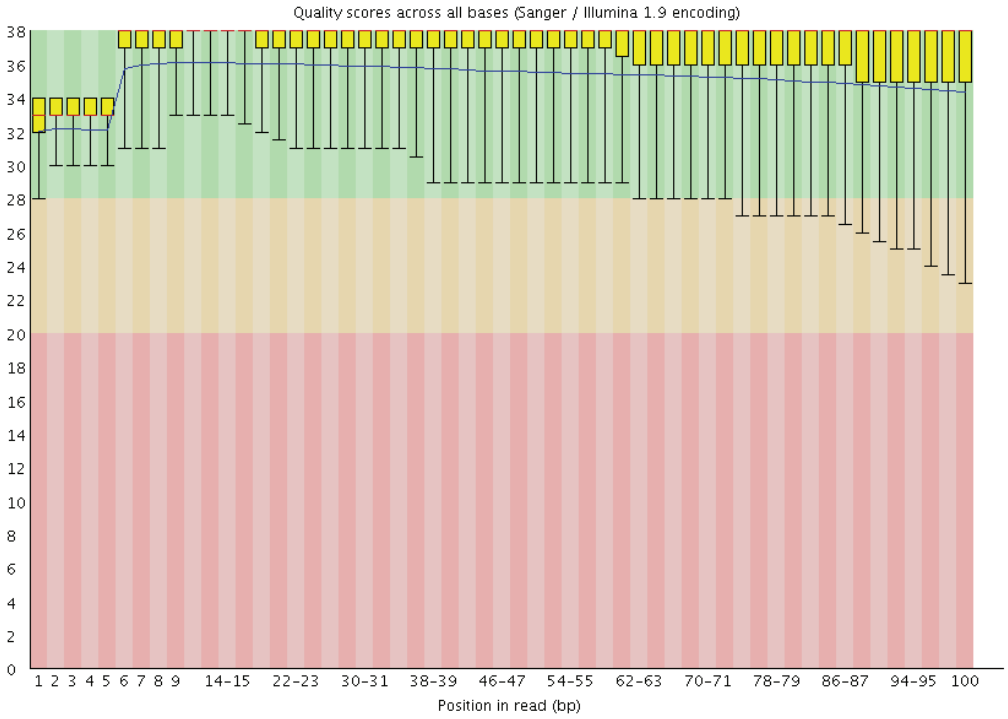


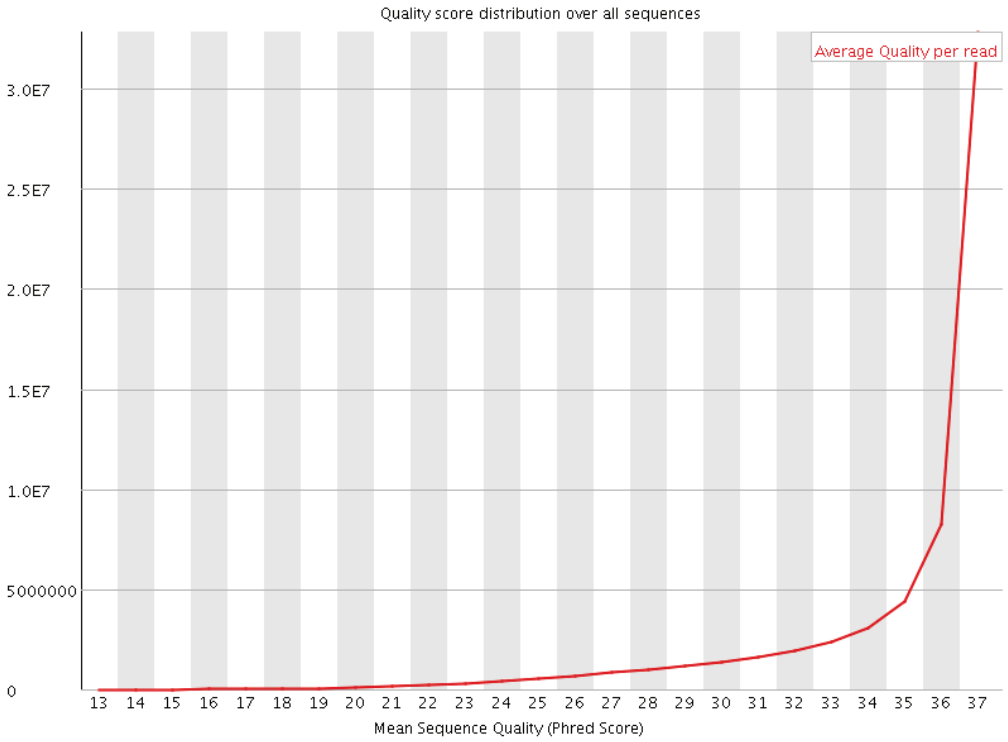
Sample 05 - single young fly - αβ-KCs - 62,687,688 reads



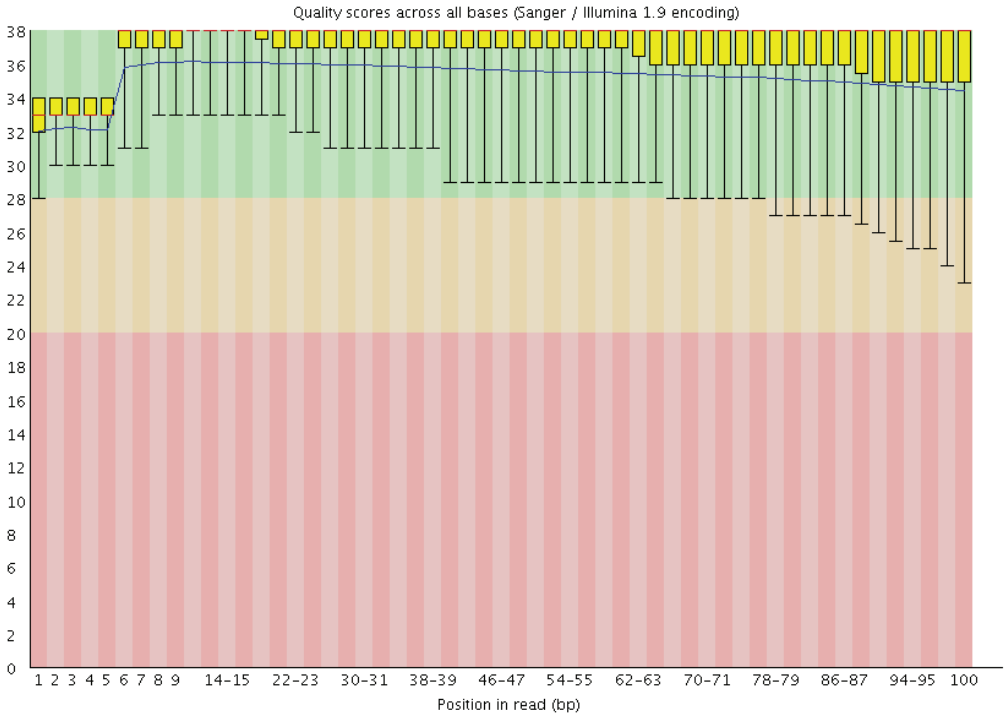


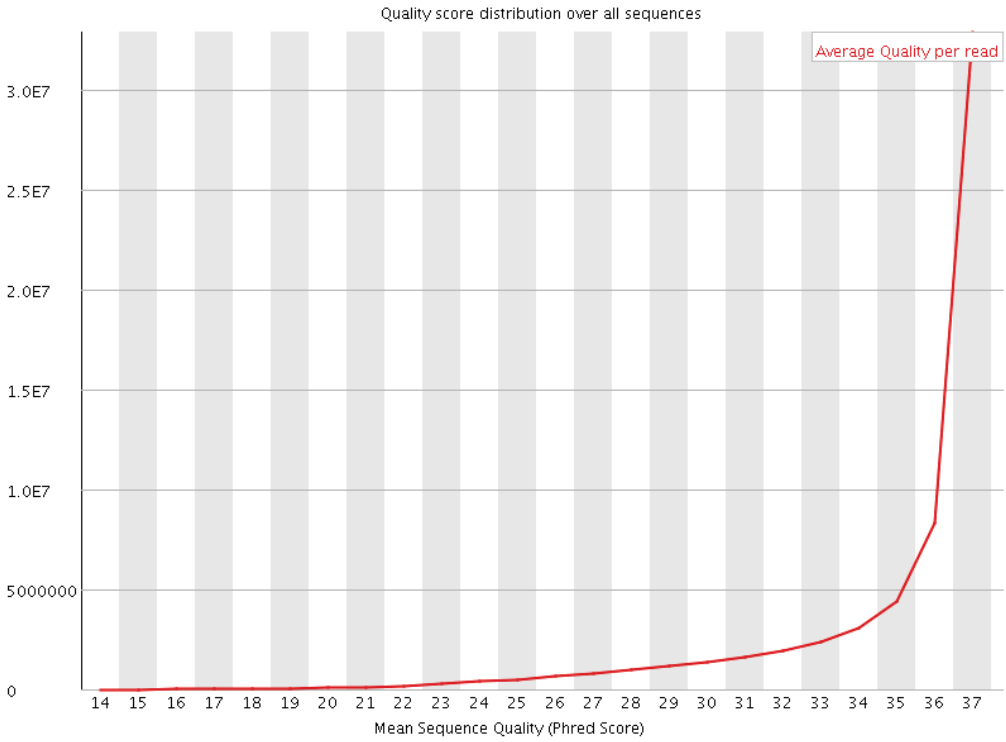
Sample 06 - single young fly - RoB - 62,609,858 reads



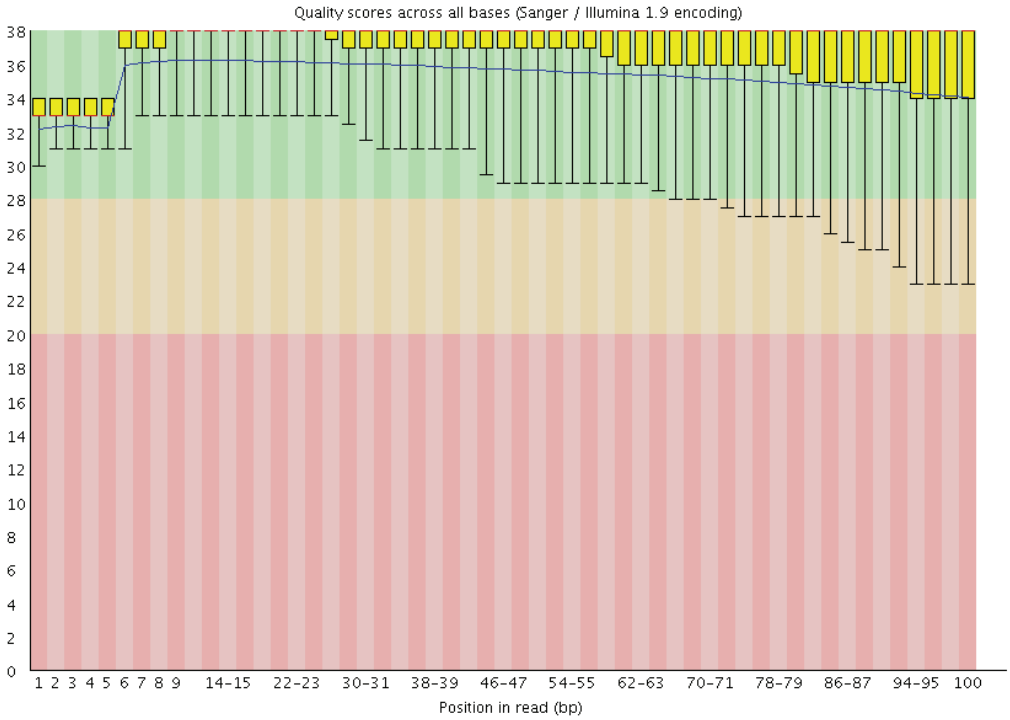


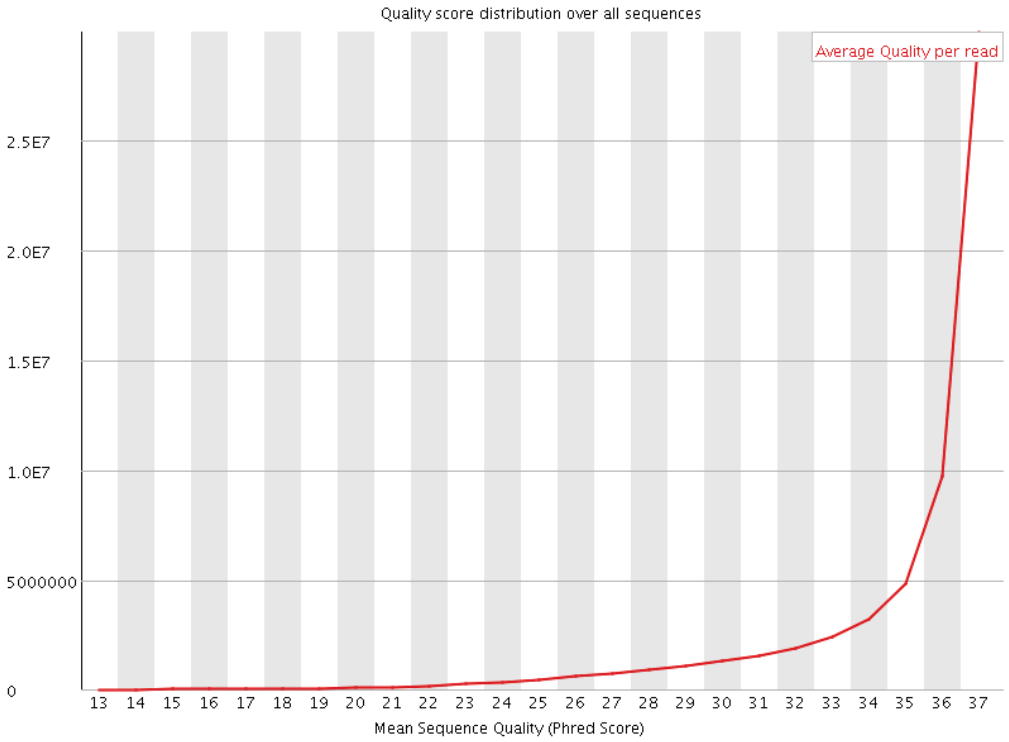
Sample 07 - single young fly - αβ-KCs - 62,490,412 reads



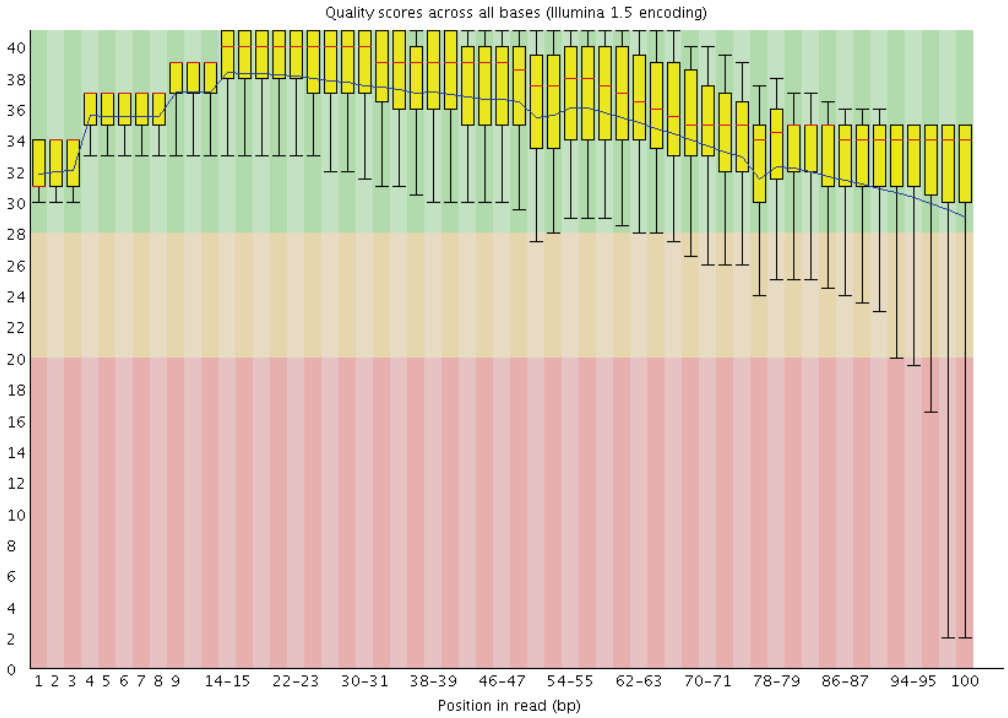


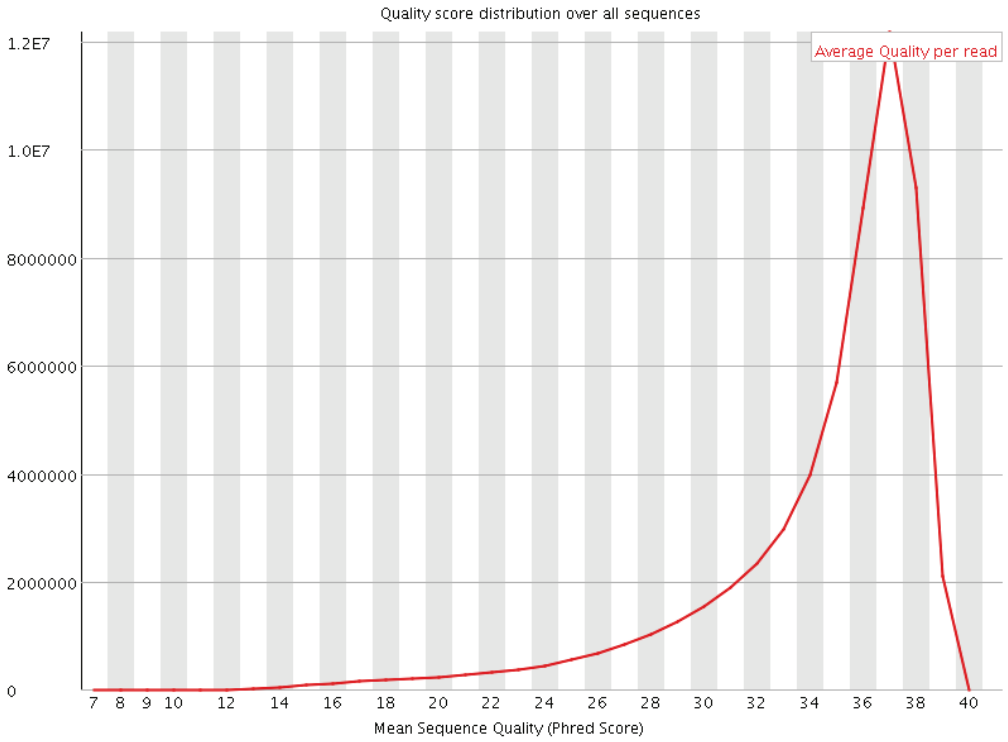
Sample 08 - single young fly - RoB - 60,824,768 reads



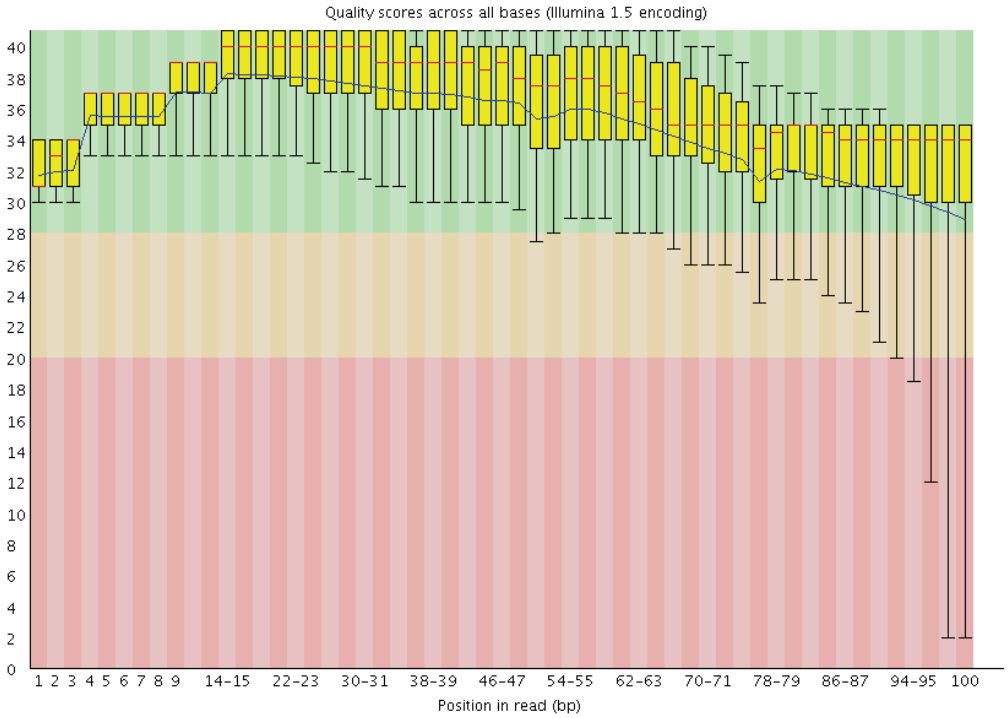


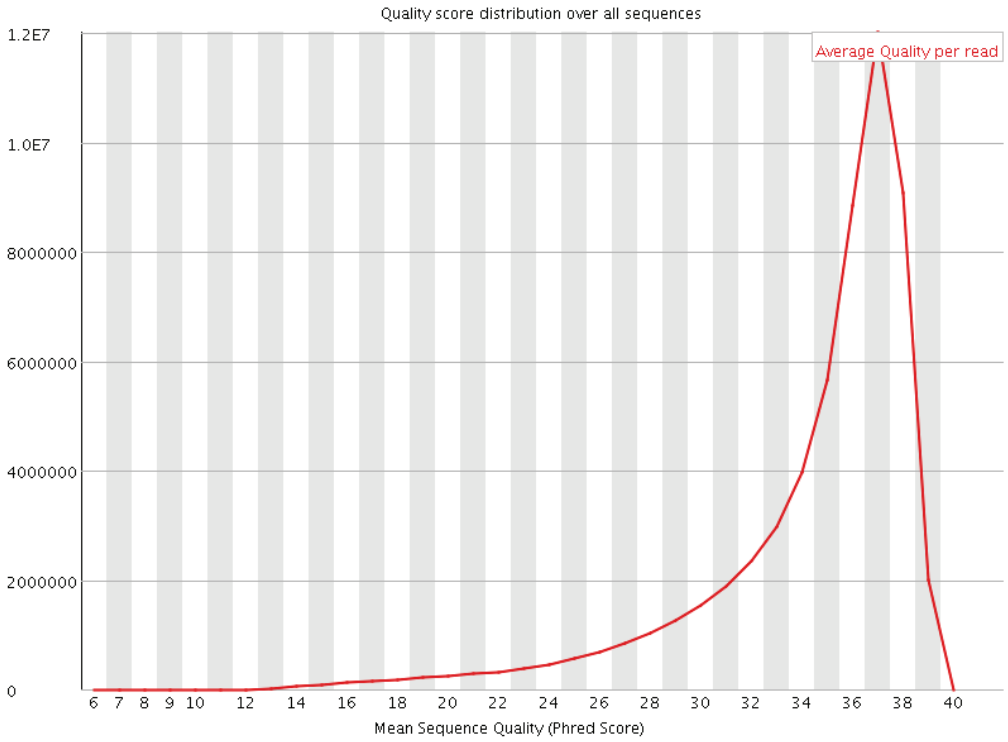
Sample 09 - single old fly - αβ-KCs - 58,143,492 reads



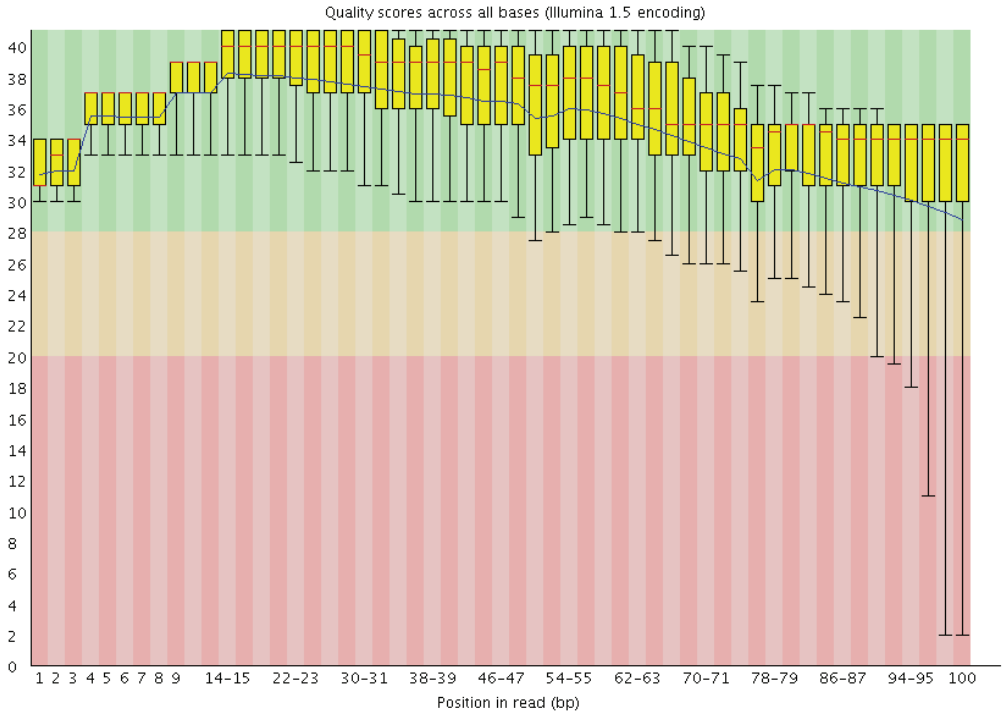


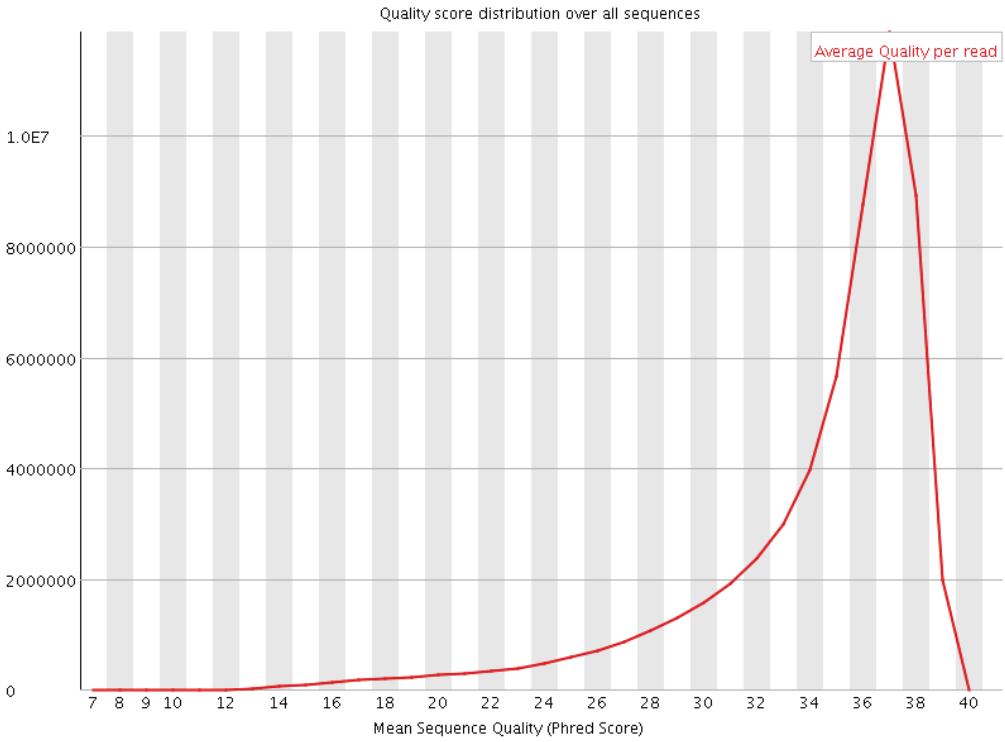
Sample 10 - single old fly - RoB - 57,823,896 reads



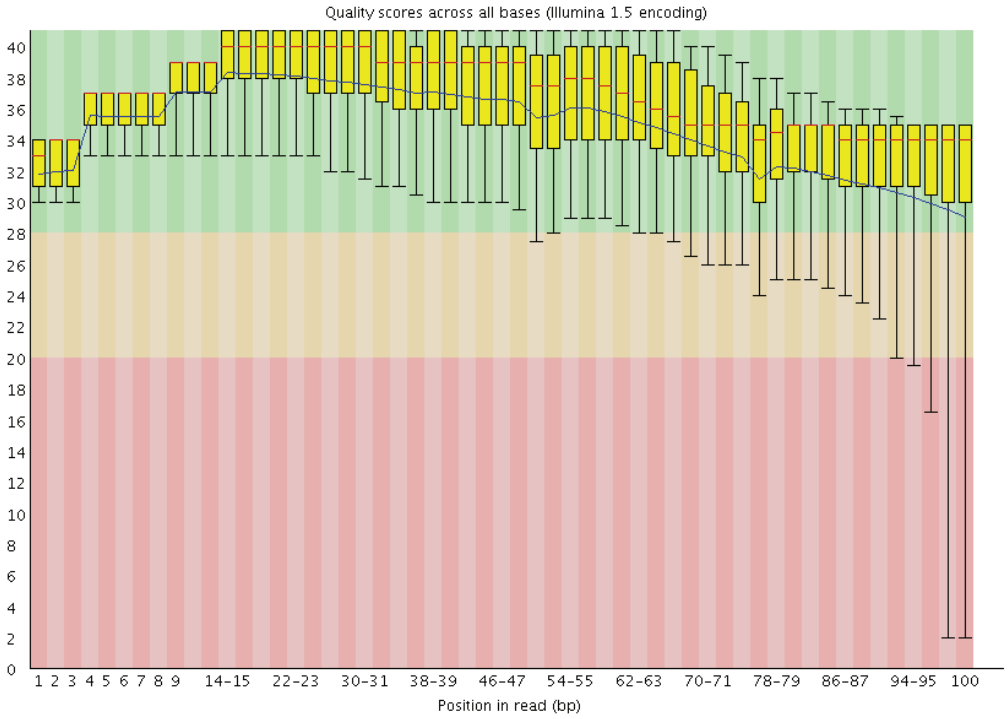


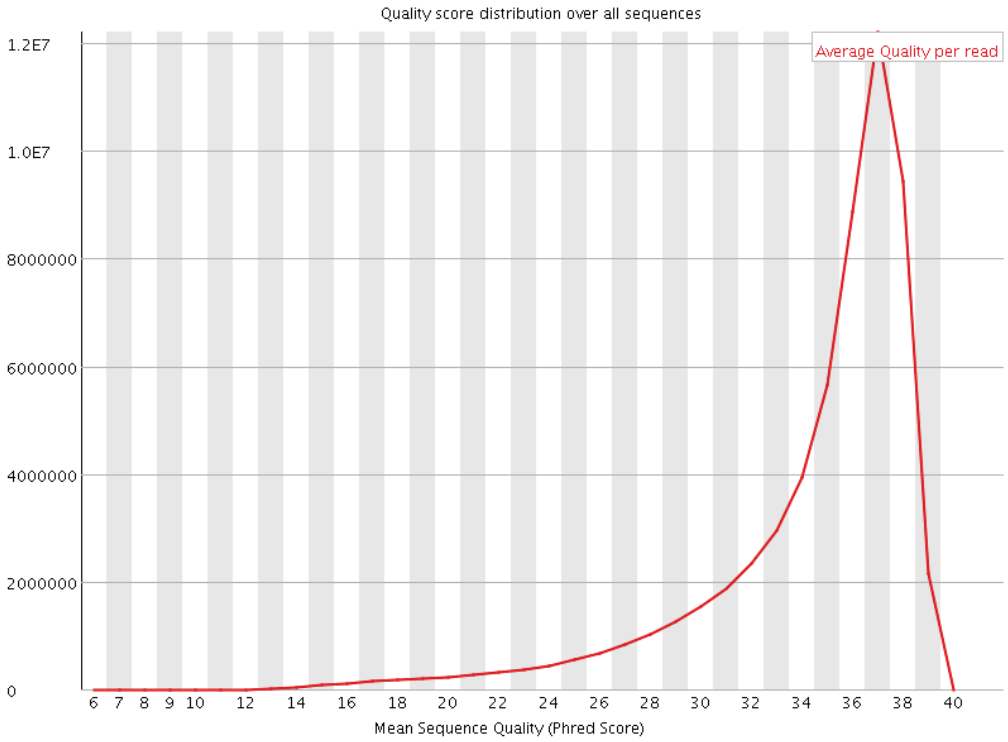
Sample 11 - single old fly - αβ-KCs - 57,628,236 reads



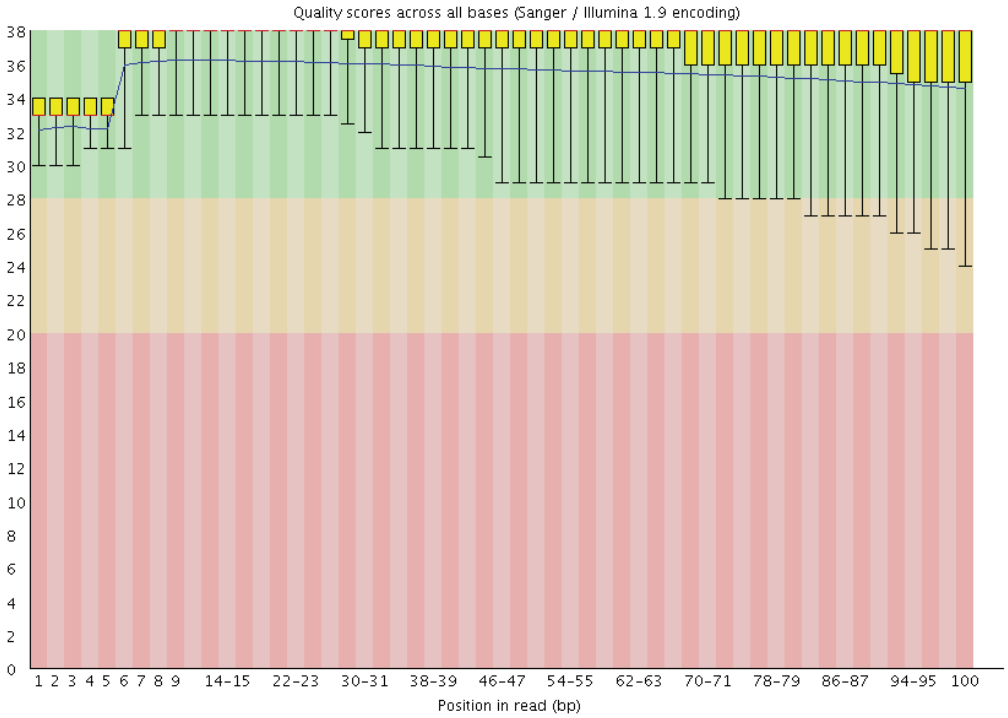


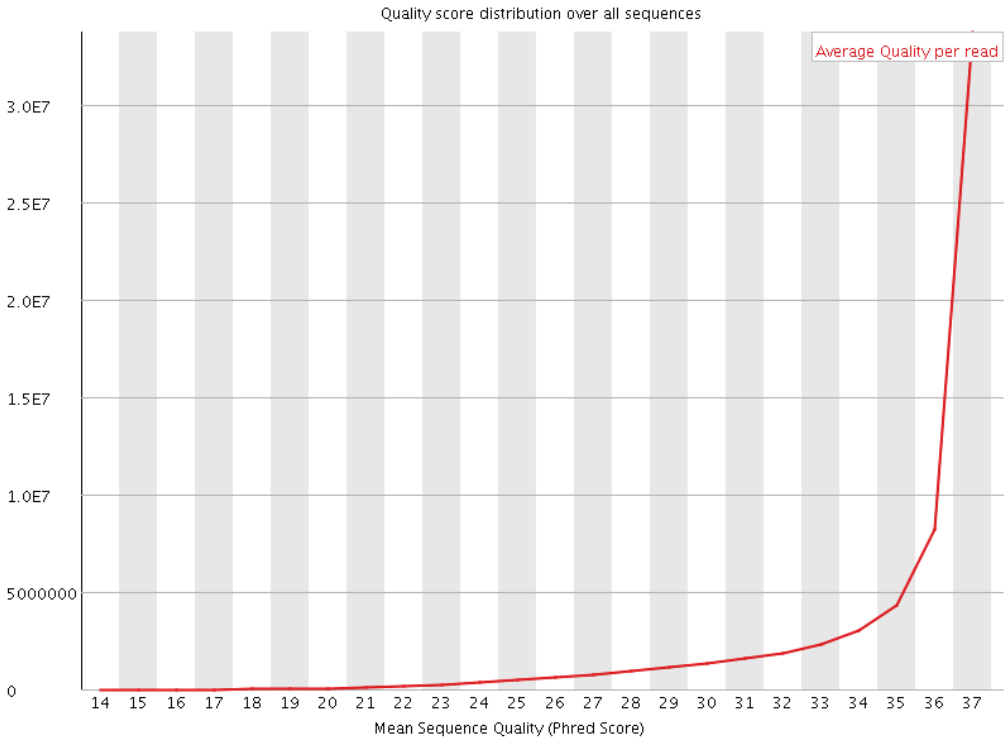
Sample 12 - single old fly - RoB - 58,162,210 reads



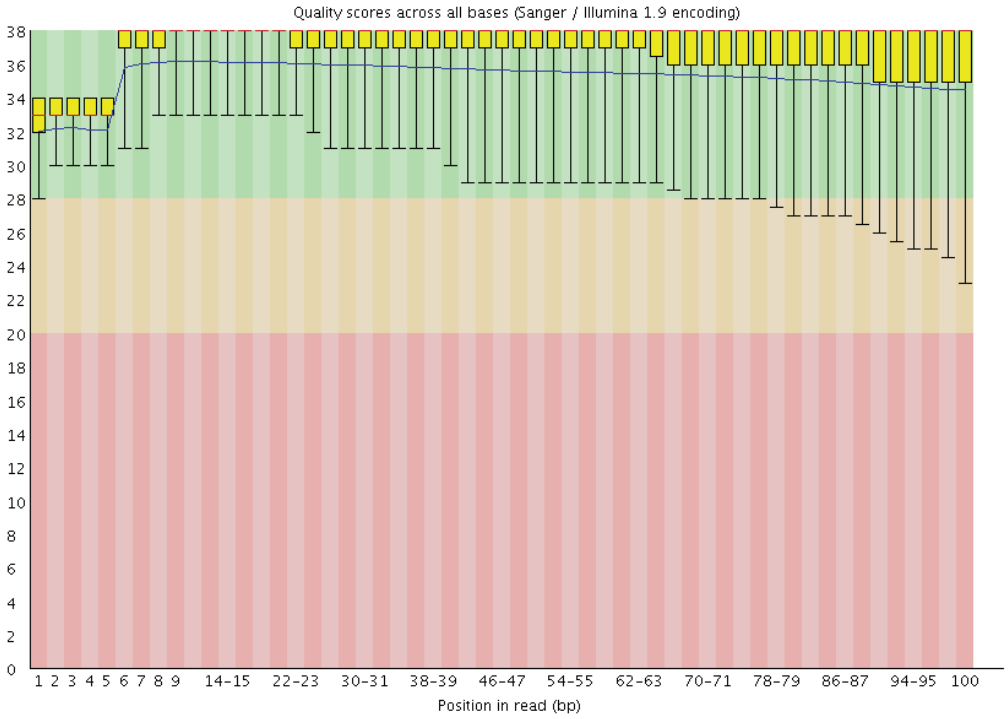


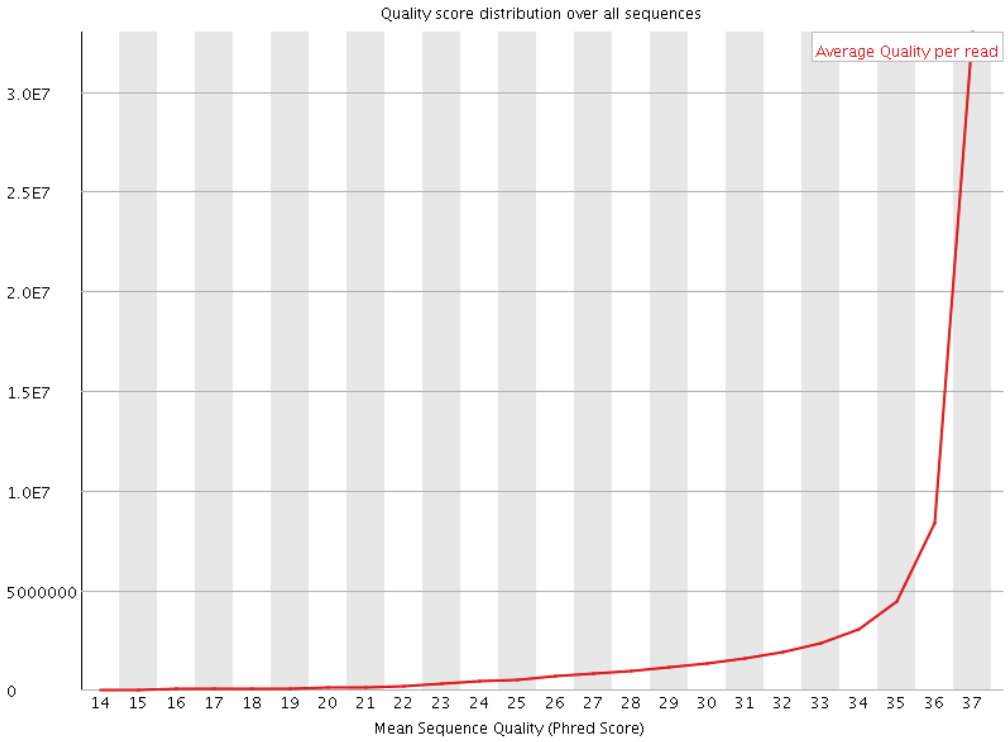
Sample 13 - single old fly - αβ-KCs - 62,618,772 reads



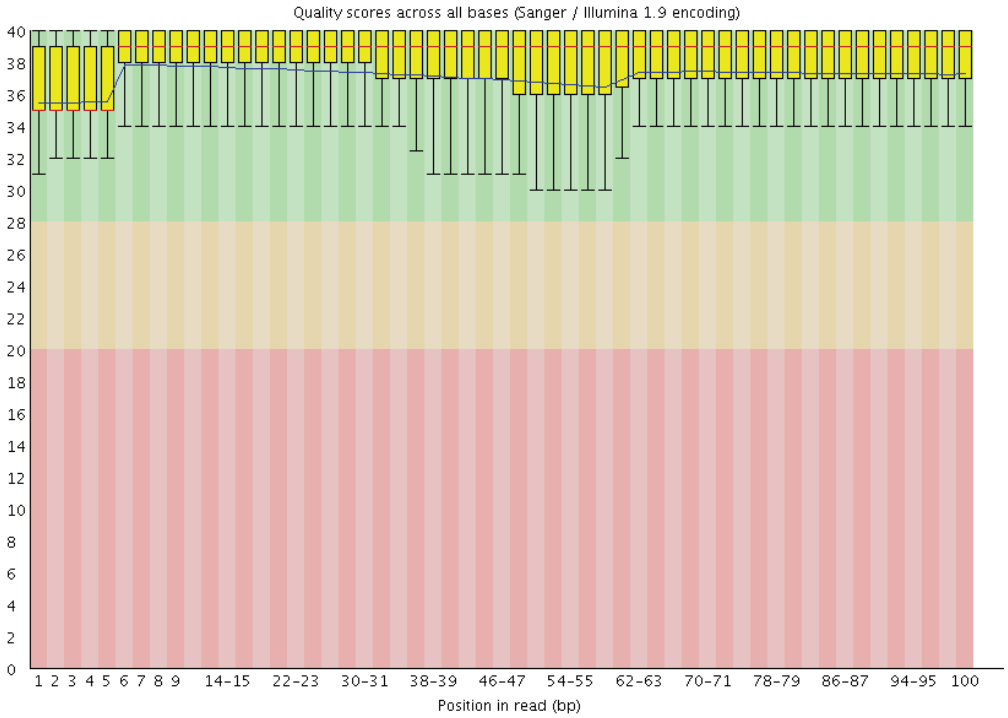


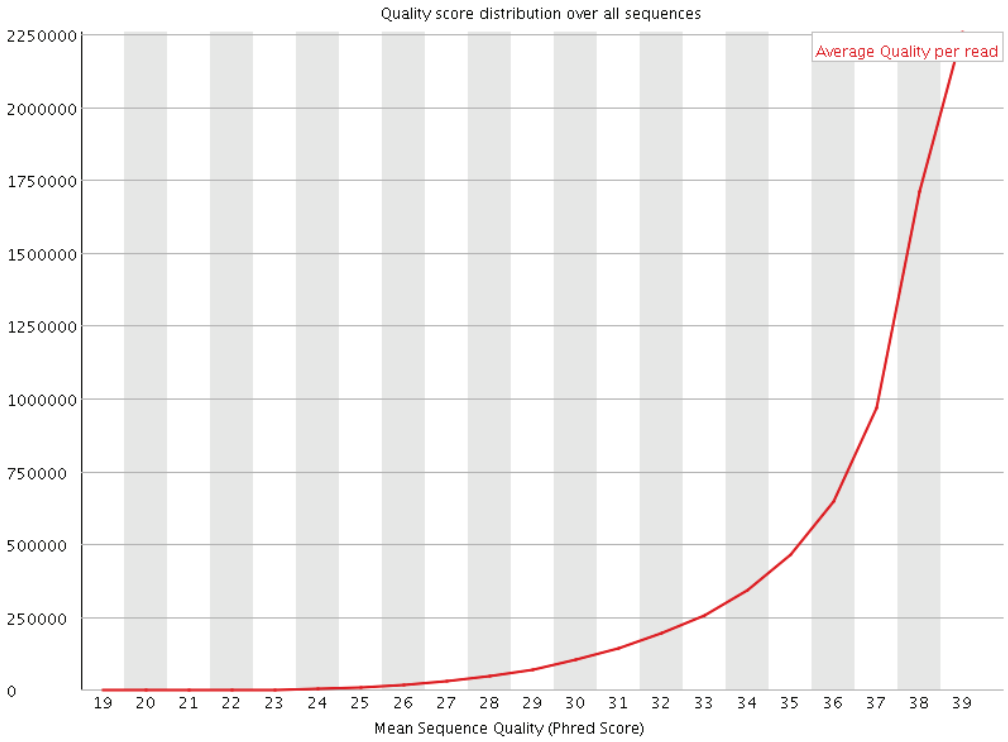
Sample 14 - single old fly - RoB - 62,383,890 reads



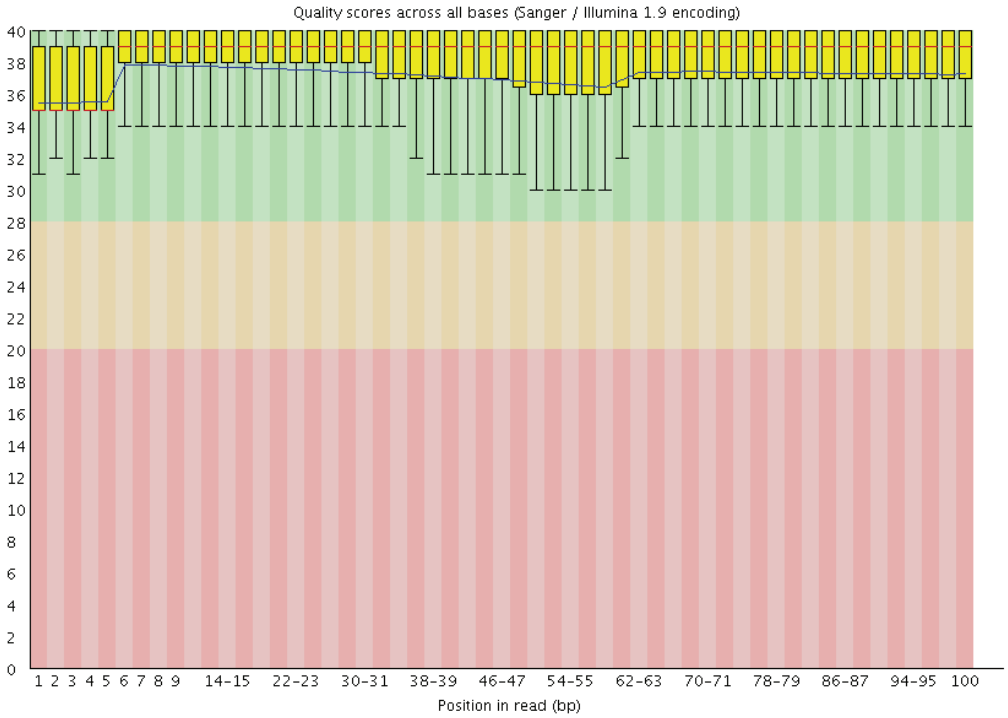


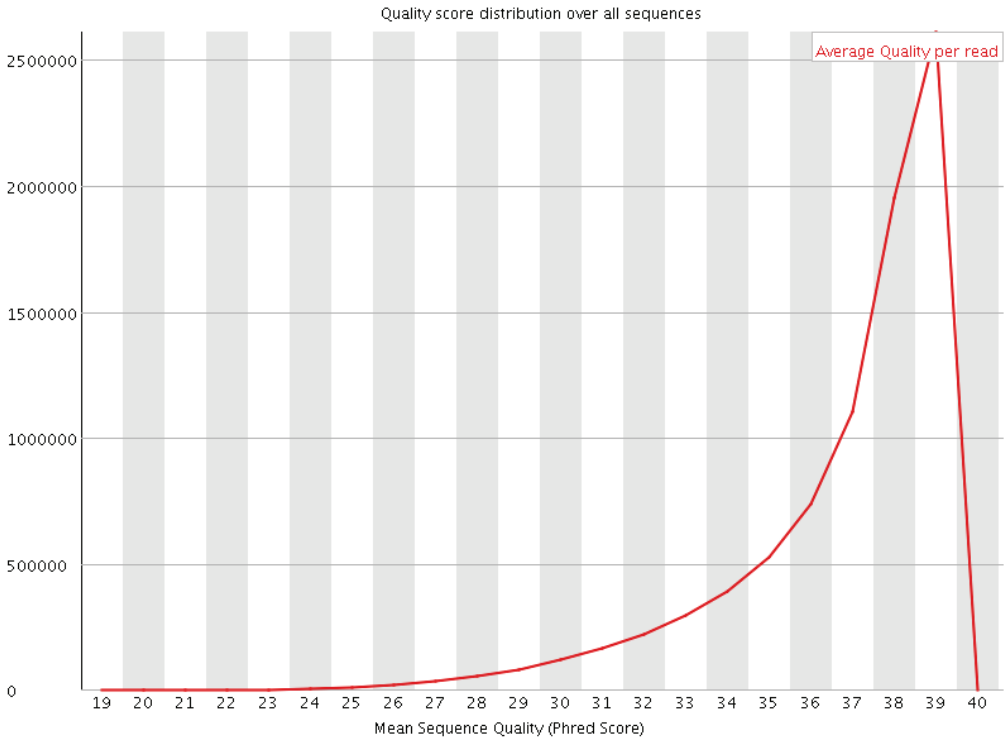
Sample 15 - single young fly - αβ-KCs - 7,304,448 reads



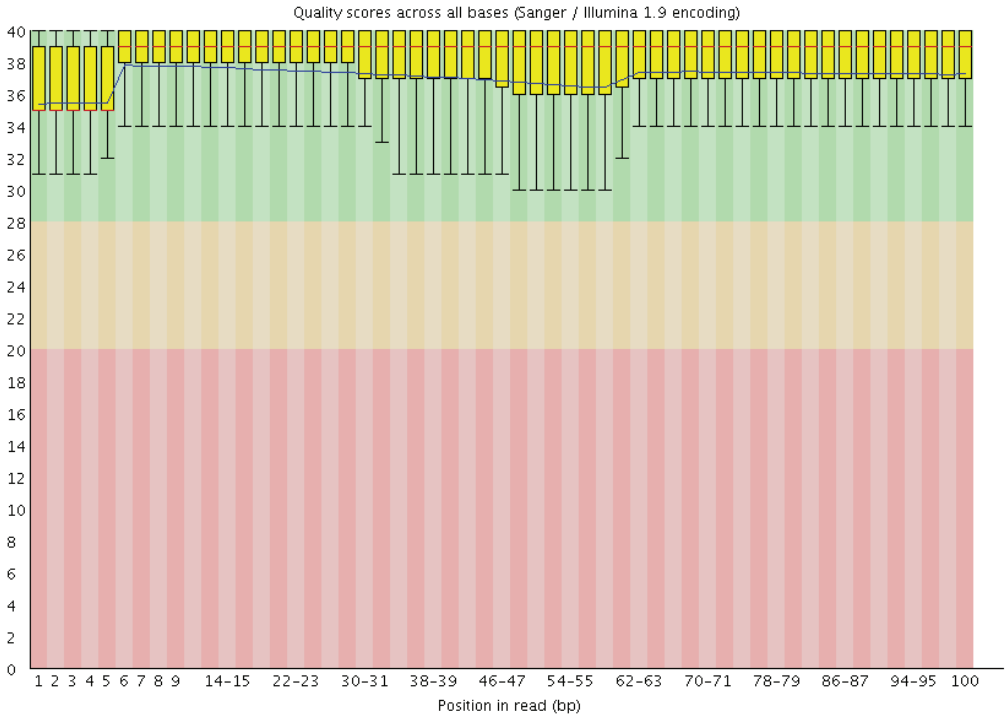


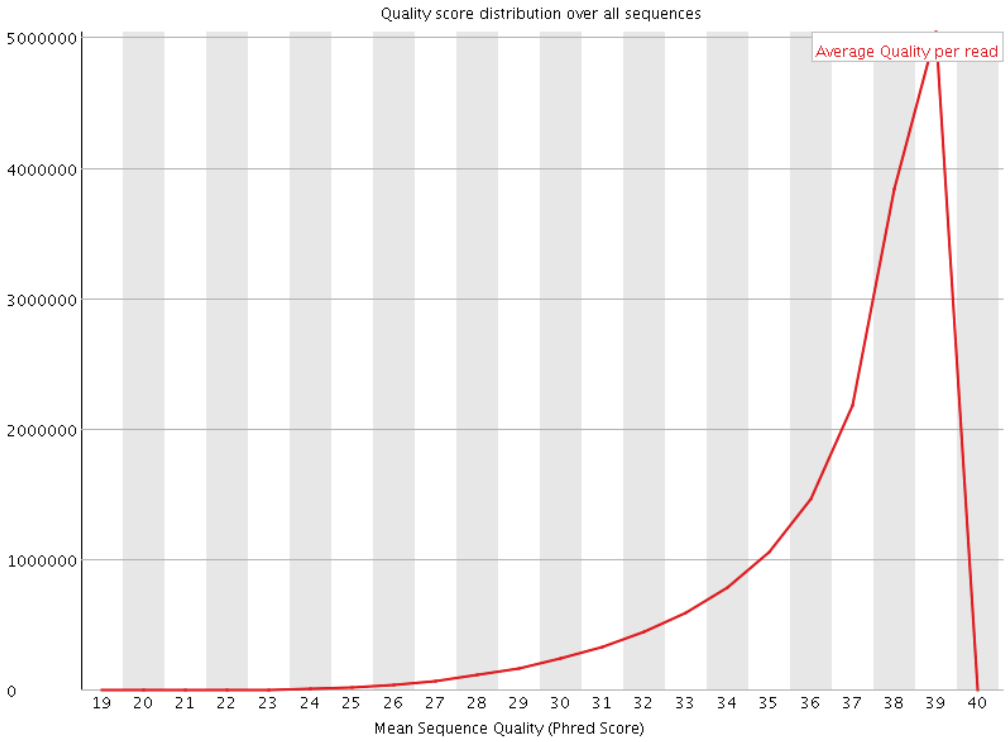
Sample 16 - single young fly - RoB - 8,369,114 reads



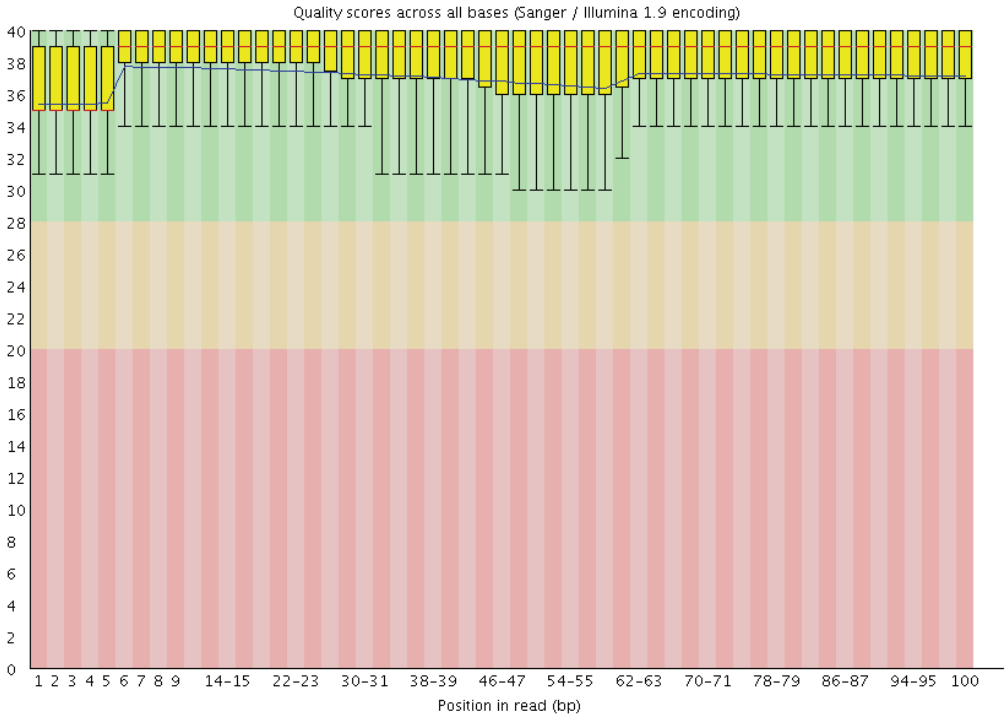


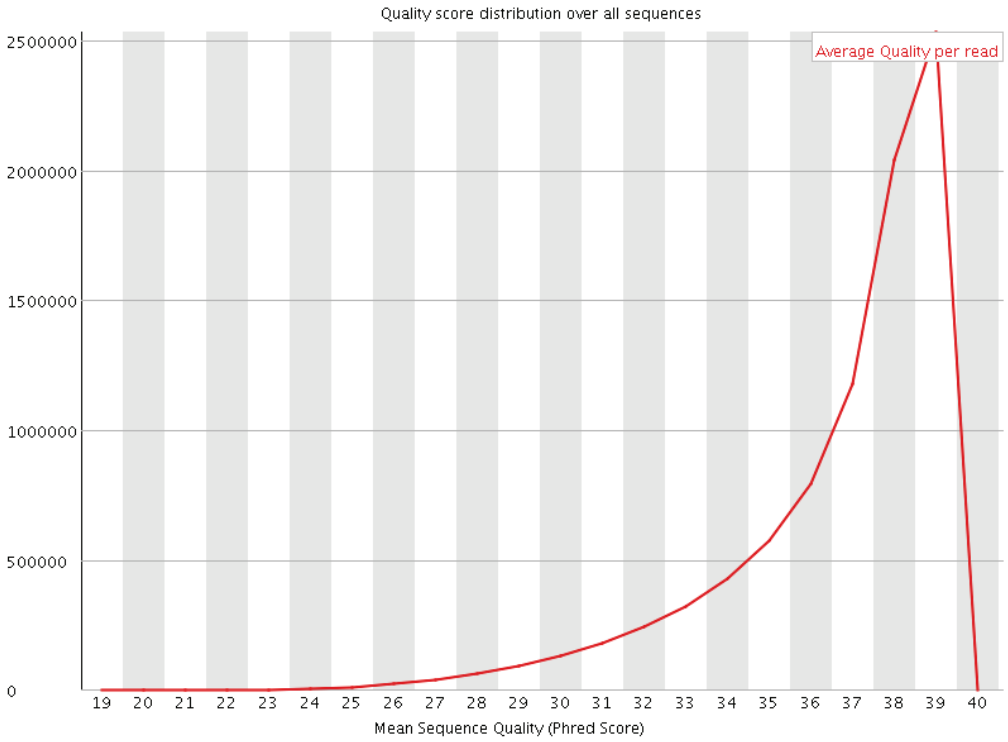
Sample 17 - single young fly - αβ-KCs - 16,459,322 reads



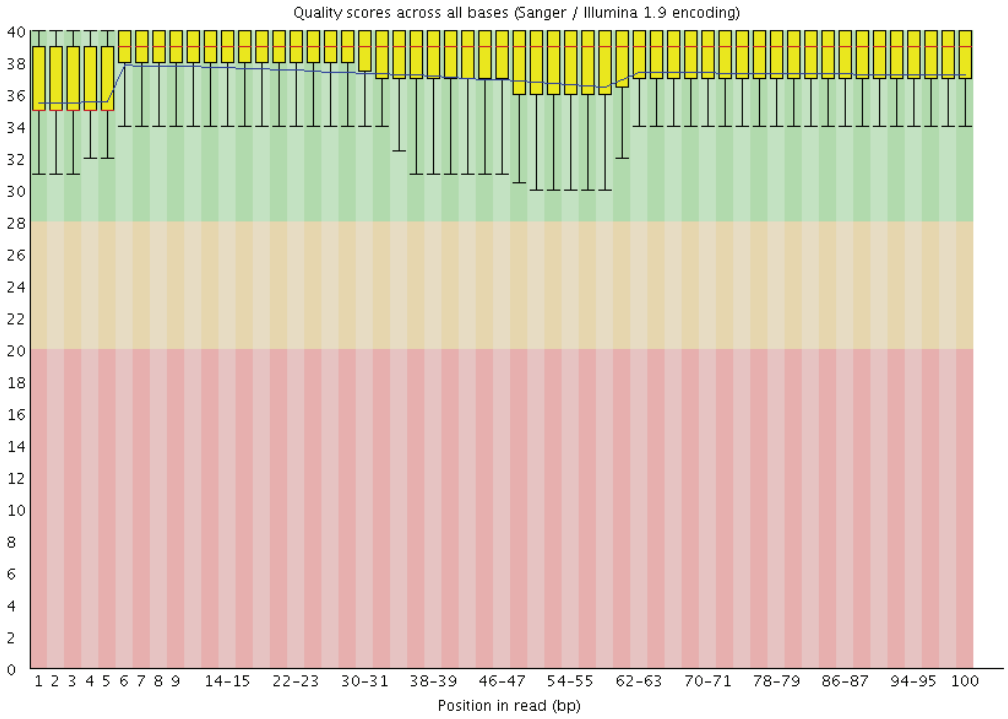


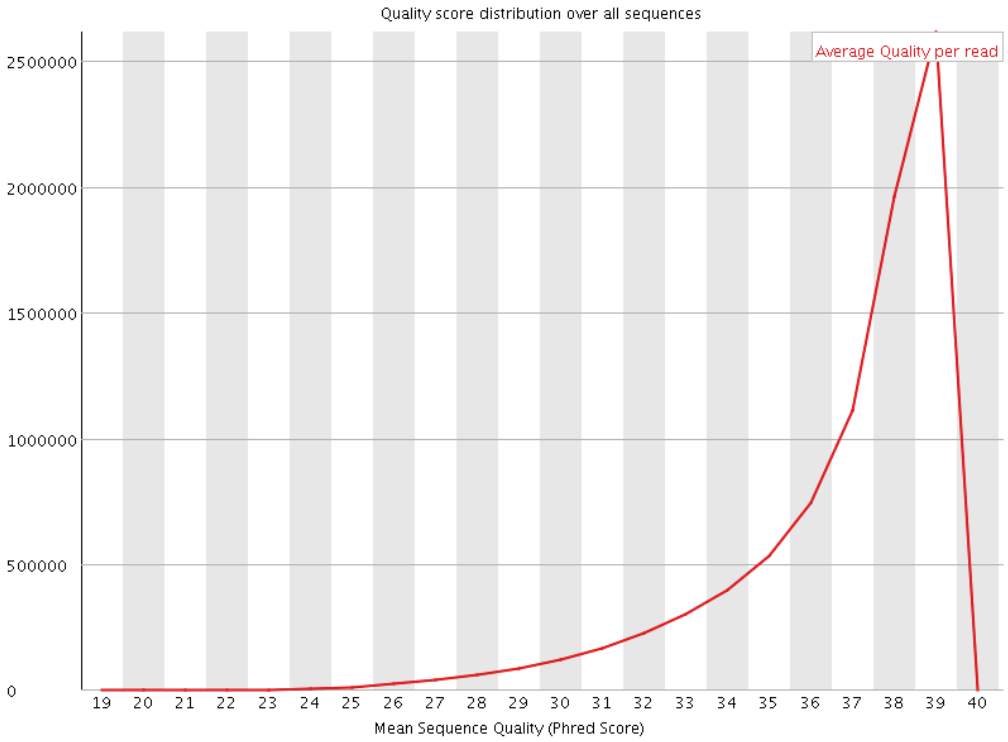
Sample 18 - single young fly - RoB - 8,688,532 reads



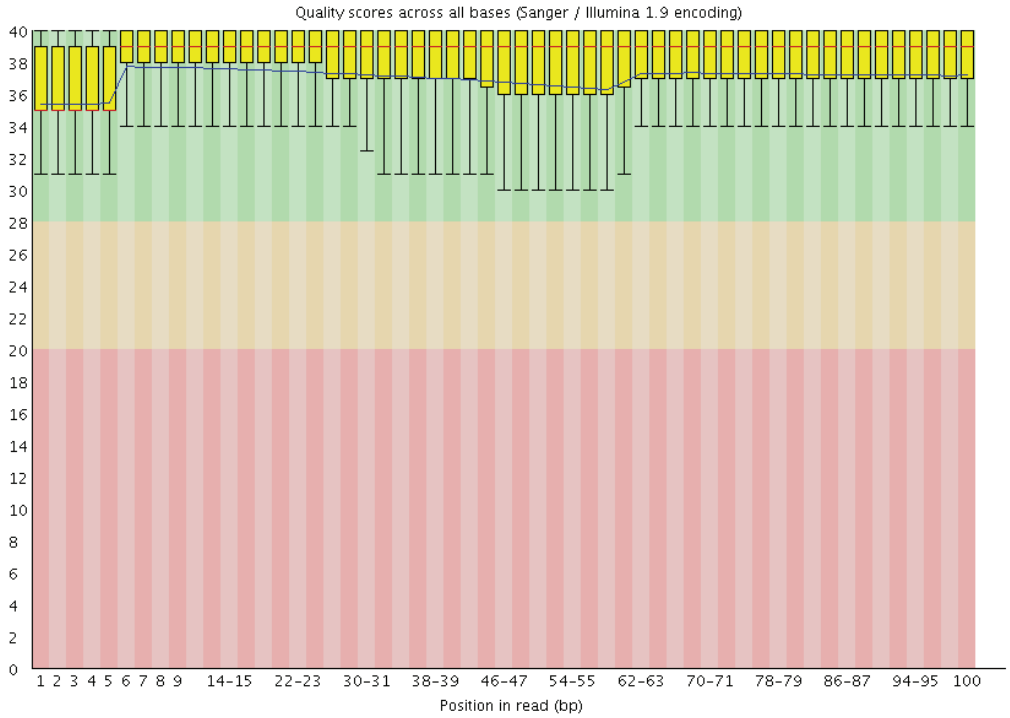


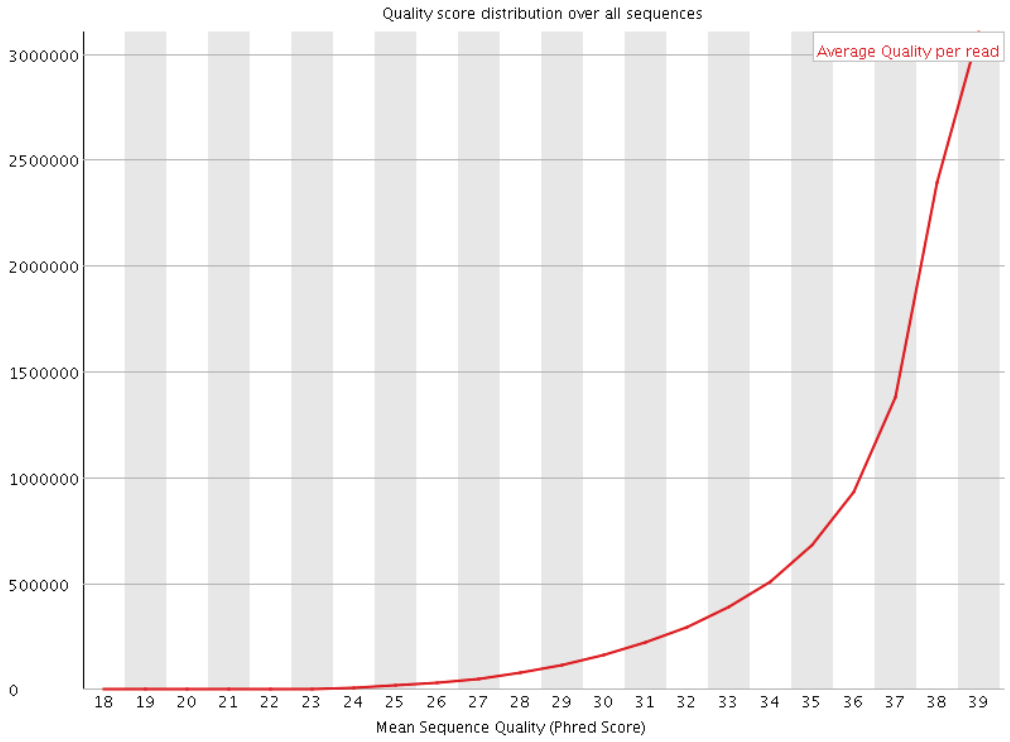
Sample 19 - single young fly - αβ-KCs - 8,444,438 reads



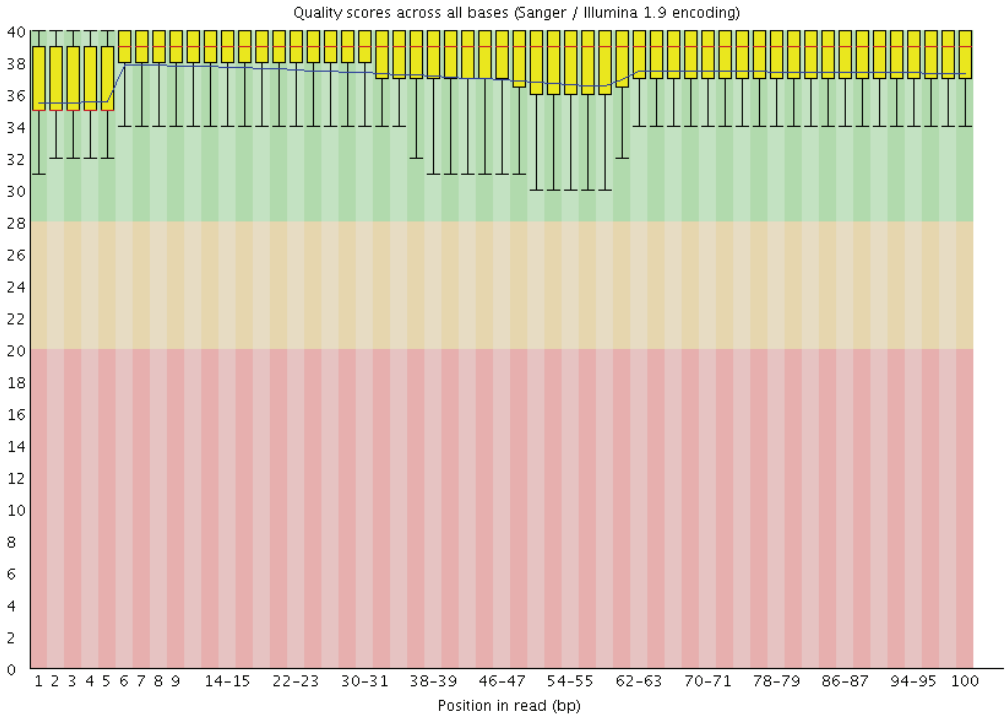


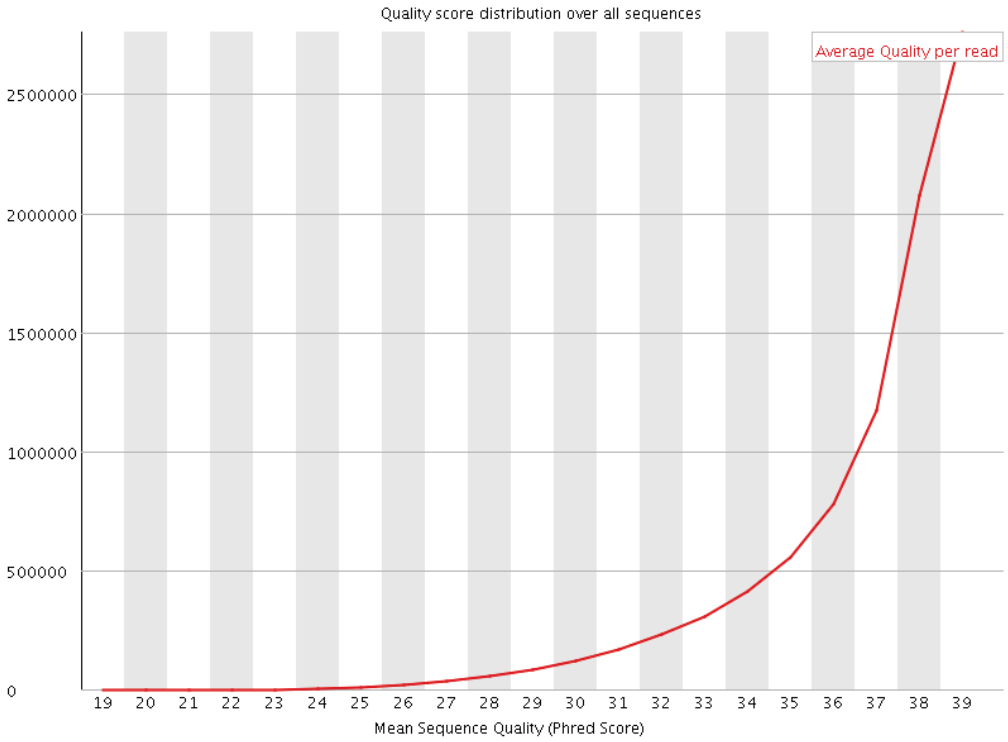
Sample 20 - single young fly - RoB - 10,393,246 reads



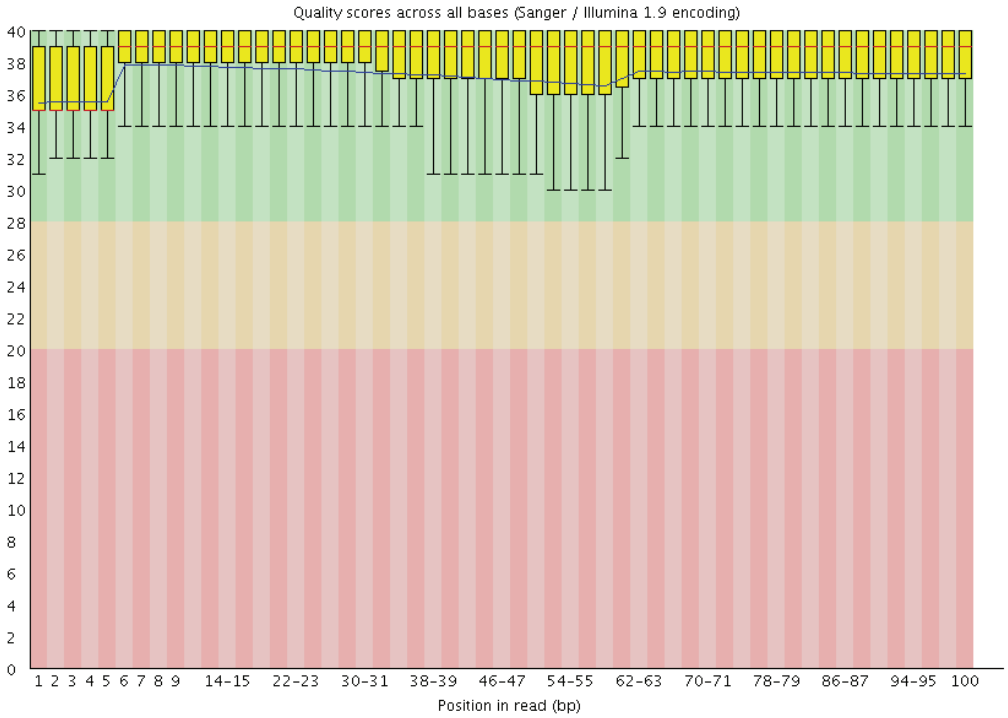


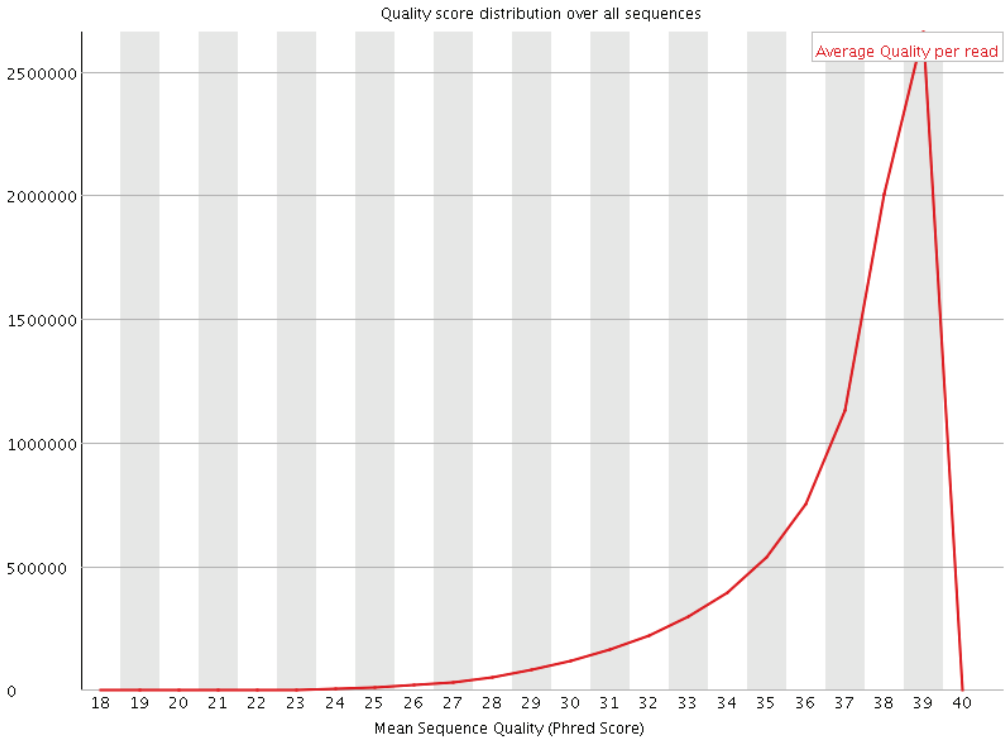
Sample 21 - single old fly - αβ-KCs - 8,846,114 reads



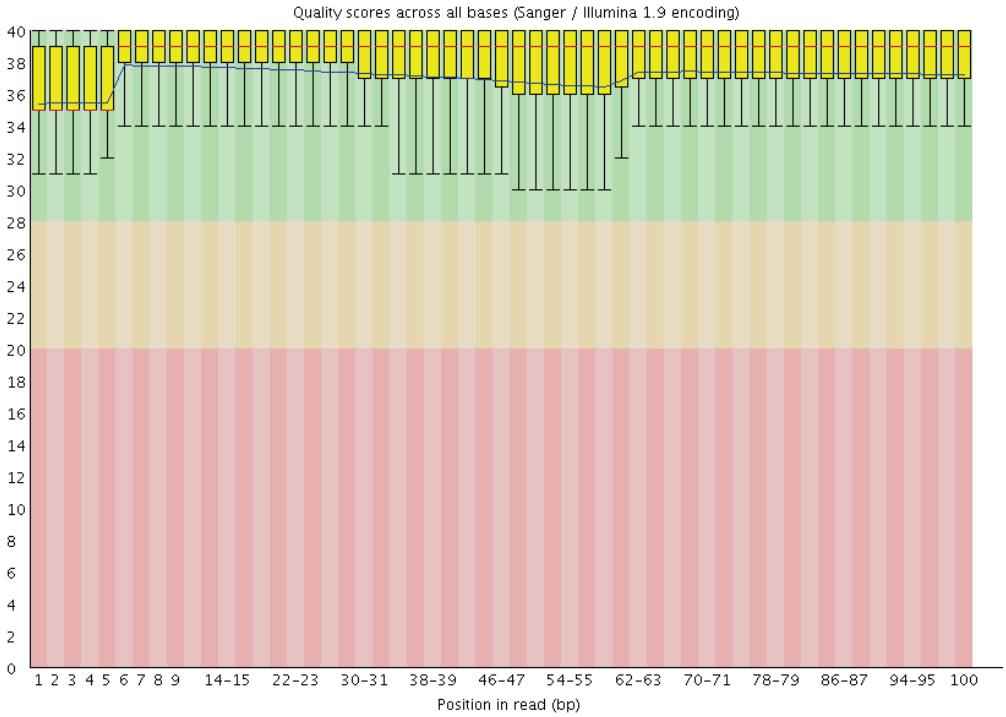


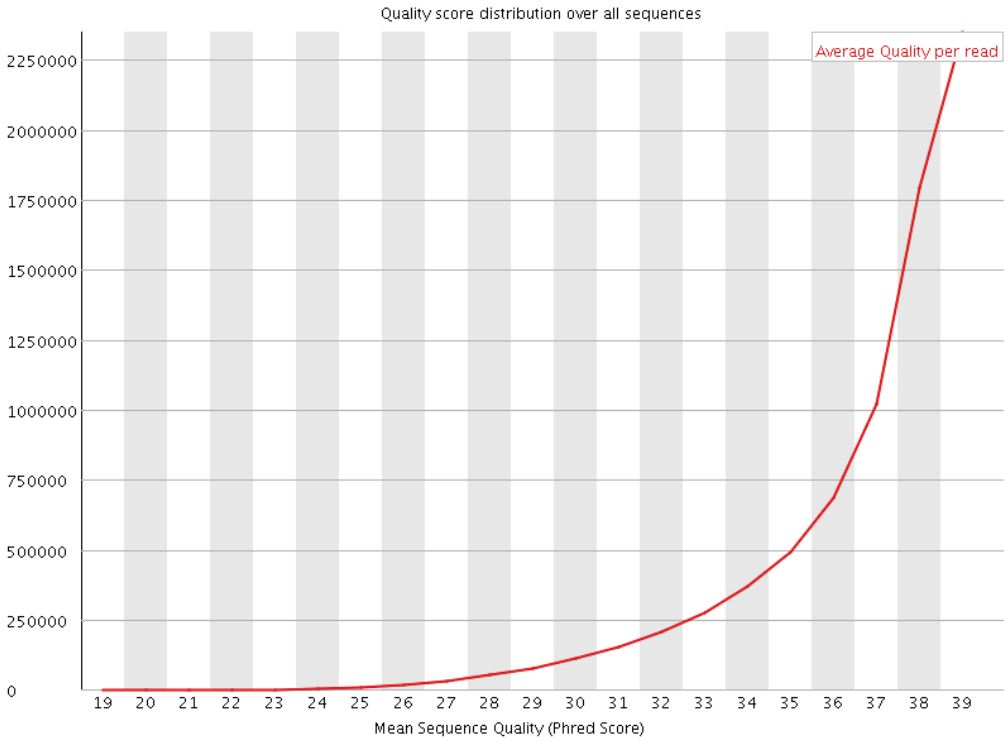
Sample 22 - single old fly - RoB - 8,504,752 reads



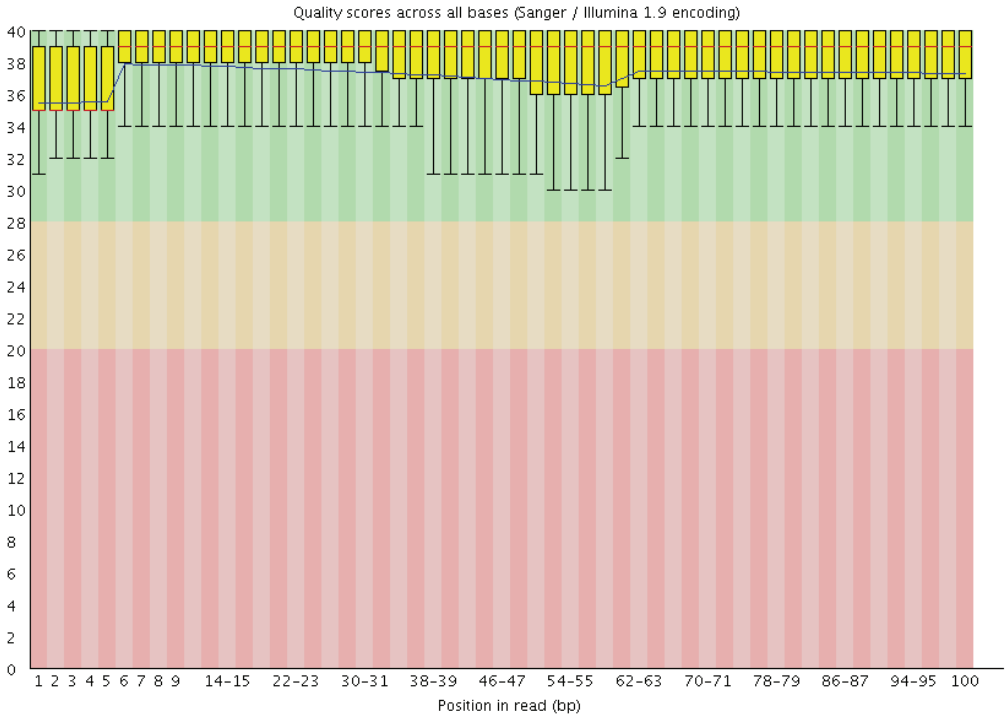


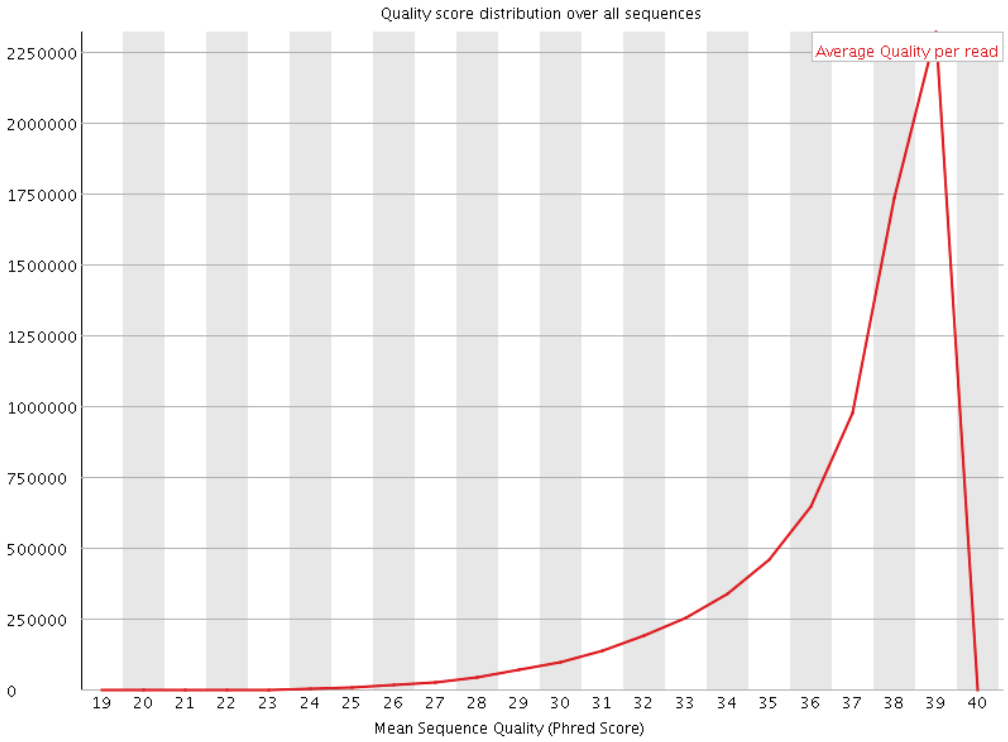
Sample 23 - single old fly - αβ-KCs - 7,692,272 reads



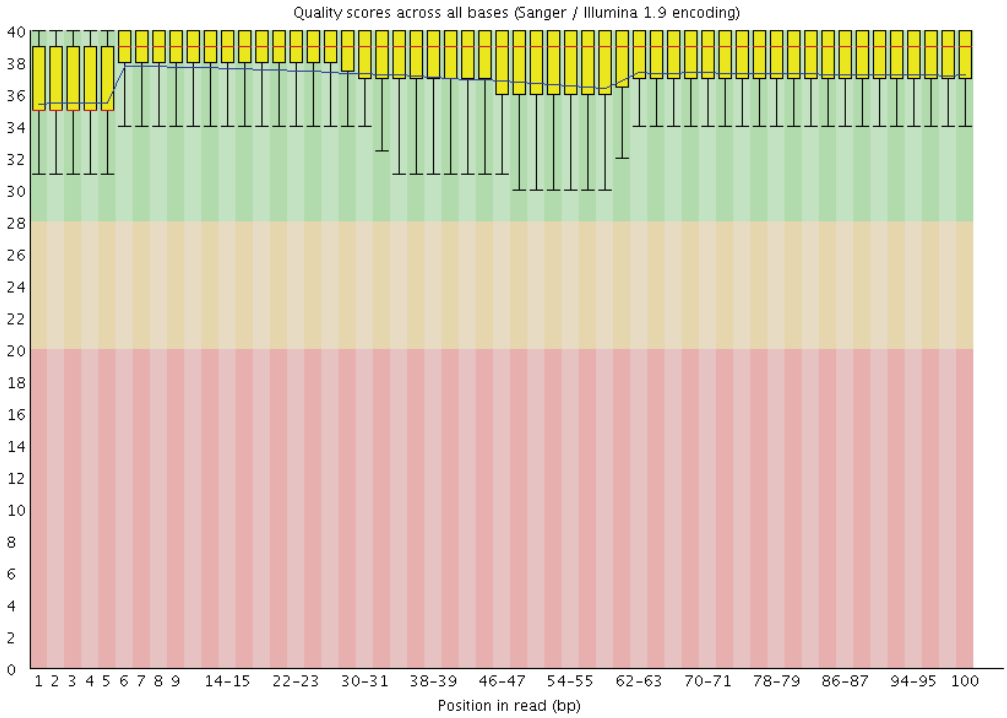


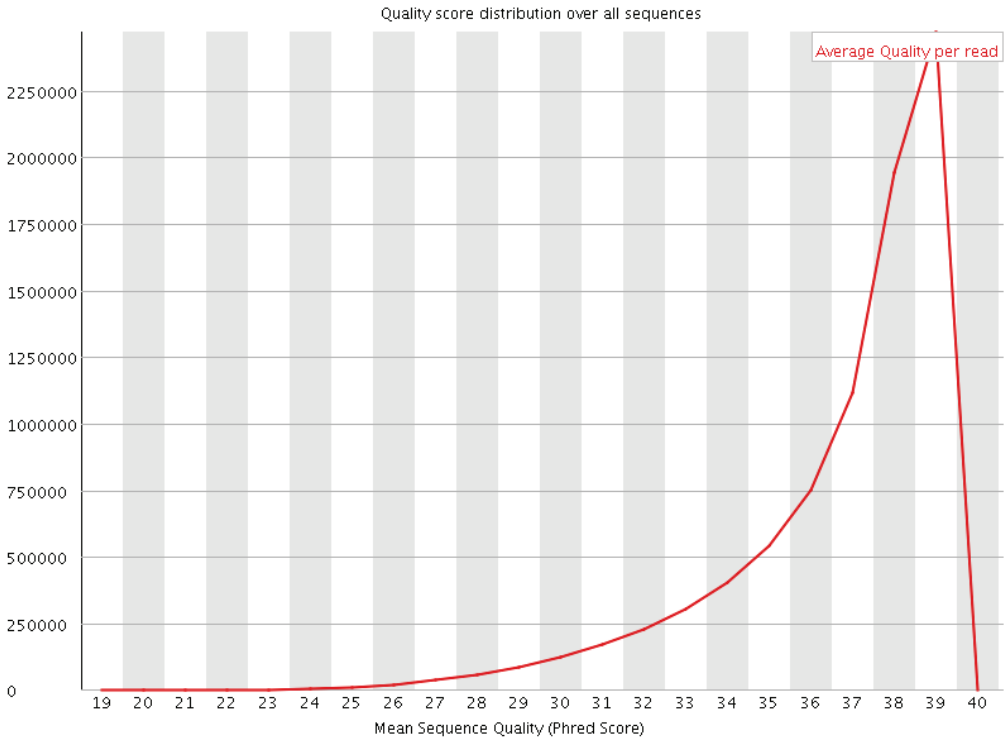
Sample 24 - single old fly - RoB - 7,378,712 reads



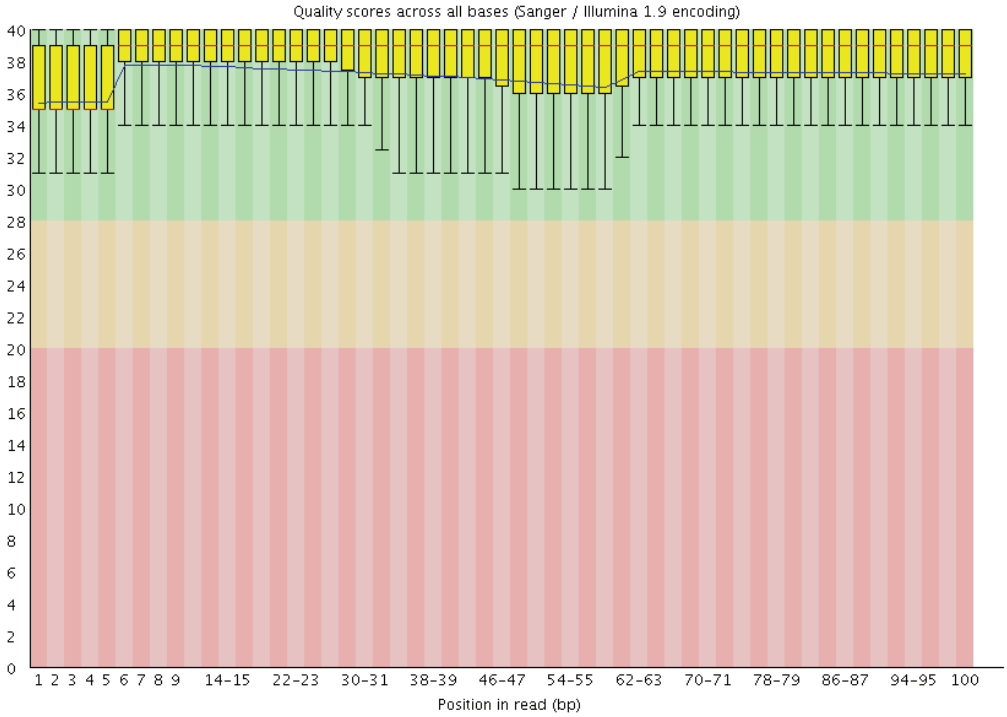


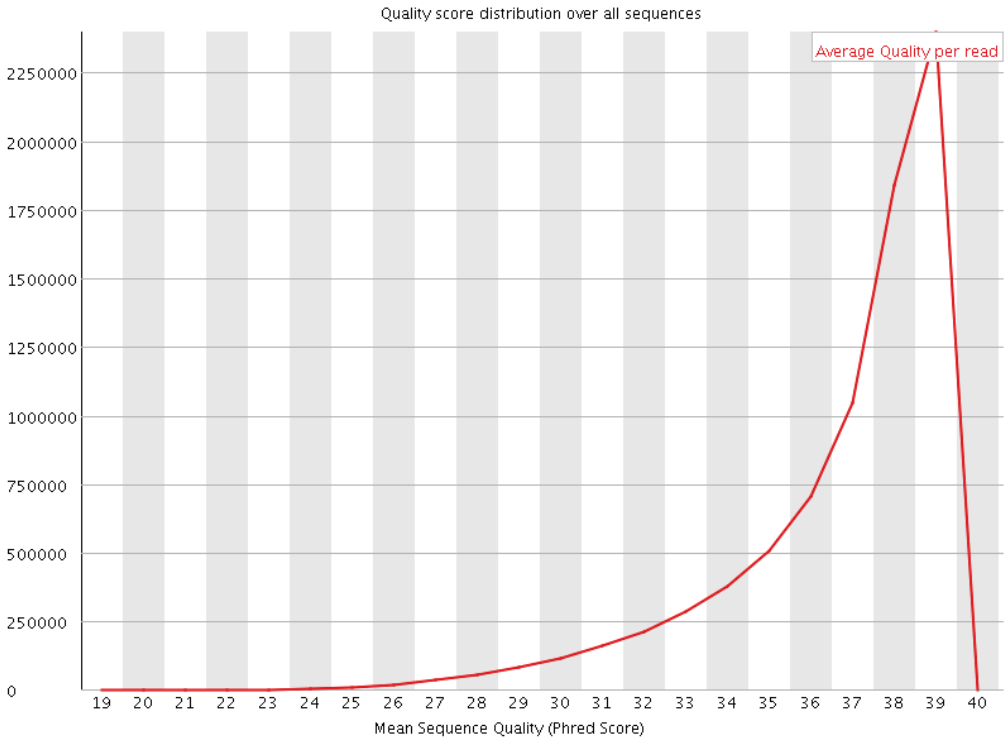
Sample 25 - single old fly - αβ-KCs - 8,298,164 reads





Sample 26 - single old fly - RoB - 7,894,110 reads





Appendix 8 – UPL primer sequences

GAPDH

Forward: AAAAAGCTCCGGGAAAAGG

Reverse: AATTCCGATCTTCGACATGG

UPL probe number: 18

SdHA

Forward: TTCGCGAGGATGAATACGAT

Reverse: CACGAGAGCGTGTGCTTG

UPL probe number: 165

Ago3

Forward: TTATATTACAATTGGGCTGGTACAGT

Reverse: AAATAAGCCAATTTGTGAGCGTA

UPL probe number: 151

mCherry-A

Forward: AGGGCGAGATCAAGCAGAG

Reverse: TTGACCTCAGCGTCGTAGTG

UPL probe number: 161

mCherry-B

Forward: TGACCGTGACCCAGGACT

Reverse: GTCGGAGGGGAAGTTGGT

UPL probe number: 152

mCherry-C

Forward: CCTGTCCCCTCAGTTCATGT

Reverse: CAAGTAGTCGGGGATGTCG

UPL probe number: 125

Fas2-A

Forward: CGCCATCGGCTATTCATC

Reverse: TCAGCCACTTGGATGACG

UPL probe number: 38

Fas2-B

Forward: ATTCAGTCGGCGTCTTTTTG

Reverse: TCTCCAGGATGGGGGACT

UPL probe number: 68

gypsy

Forward: TCGAATACGAAAGCGAGGAT

Reverse: TAACGGCCGACCTAATGCT

UPL probe number: 52

Blood

Forward: TCAATTGGTGGCATATATTGGT

Reverse: TTTCCTCAGTAGTCCTCAGATCG

UPL probe number: 15

Doc3

Forward: AATCCATCTGCGTCGTTTG

Reverse: CGCCACAGTTGGTGCATA

UPL probe number: 105

Ivk

Forward: ACGTTCGACCGTTTCACTCT

Reverse: TTGCAGTGCTTTGATGTGTGT

UPL probe number: 72

Tabor

Forward: GCGATCATGGAAAACTGAAAG

Reverse: CTCAGTGCTGCAATGGATGT

UPL probe number: 34

CG17684

Forward: CGGCATTTCATGGAGCTT

Reverse: TGACGTCATATTAGTGTGCATTTC

UPL probe number: 61

Spock

Forward: TGCGTGTTAAAATGGTGACTG

Reverse: GCGTCTGGTGATGGATTGTA

UPL probe number: 105

Trim9

Forward: CTGCAAGGCTCAAAAGACG

Reverse: GAACTCGGTCGTGGAACG

UPL probe number: 133