

Cranfield University

Greg Deakin

**The Functional Role of Methylated Short Tandem Repeats in
Early Mouse Development**

School: Cranfield Health
Course Title: Applied Bioinformatics
Award: M.Sc.
Year: 2011
Supervisor: Conrad Bessant

25th August 2011

This thesis is submitted in partial fulfilment of the requirements for the Degree of Master of Science.

© Cranfield University, date 2011. All rights reserved. No part of this publication may be reproduced without the written permission of the copyright holder.

1. Abstract

Short tandem repeats, or microsatellites are ubiquitous throughout all genomes that have been explored. In common with other sequences, the DNA in microsatellites has DNA marks in the form of chromatin methylation. Regulation of DNA methylation and changes in their pattern is critical for the establishment of unique cell states throughout development in mammals. DNA methylation is extensively reprogrammed during the early phases of mammalian development to establish unique developmental patterning. Whether microsatellites are also reprogrammed with developmental patterns is unknown. In this thesis, we assessed the characteristics of di- and trinucleotide microsatellites in the NCBIM37 *Mus musculus* assembly and observed a marked difference in quantity and length of microsatellites of differing motif, not explained by any known mechanism. Secondly we assessed the quantities of di-, tri- and tetranucleotide microsatellites in experimentally determined methylomes of *Mus musculus* at various stages in development. Our results indicate that at least one tetranucleotide microsatellite motif and more tentatively a second trinucleotide microsatellite follow a pattern of methylation consistent with reprogramming. Finally we show that the genes containing these specific microsatellites in the NCBIM37 genome have strong links to known developmental processes.

Table of Contents

1. Abstract.....	2
2. Introduction.....	5
2.1. Tandem Repeats.....	5
2.2. Fuzzy Repeats	7
2.3. Epigenetic Reprogramming during Development.....	8
2.3.1 Early Embryo Reprogramming.....	10
2.3.2 Germ Cell Reprogramming.....	10
3. Project Objectives.....	13
3.1. Basic analysis of the Mus musculus assembly.....	13
3.2. Analysis of microsatellites	13
3.3. Analysis of Experimental Datasets.....	14
3.4. Assessing the Functional Significance.....	15
3.4.1 Finding genes.....	15
3.4.2 Finding Functional Relationships.....	15
4. Methodology.....	16
4.1. Basic analysis of the Mus musculus genome.....	16
4.1.1 Calculating the expected levels of di- and trinucleotides from base composition.....	17
4.1.2 Calculating the expected number of tandem repeats.....	20
4.2. Analysis of microsatellites	21
4.3. Analysis of Experimental Datasets.....	23
4.3.1 Experimental Data.....	23
4.3.2 Calculating the expected number of methylated tandem repeats.....	24
4.3.3 Calculating Fold Change.....	26
4.4. Assessing the functional significance	27
4.4.1 Finding Genes.....	27
4.4.2 Generating Random Gene Lists.....	27
4.4.3 Gene Function.....	28
5. Results and Discussion.....	30
5.1. Basic analysis of the Mus musculus genome.....	30
5.2. Analysis of microsatellites	34
5.2.1 Expected number of tandem repeats.....	34
5.3. Analysis of experimental datasets.....	44
5.4. Assessing the Functional Significance	53
5.4.1 Finding Genes.....	53
5.4.2 Determining gene function.....	54
6. Conclusions.....	61
6.1.1 Comments and Criticisms.....	62
7. References.....	64

Index of Tables

Table 1: Experimental Data Description.....	26
Table 2: Dinucleotide Microsatellite Exponents.....	38
Table 3: Microsatellites Found in Genes.....	57
Table 4: Number of gene list related terms returned by DAVID Functional Annotation Chart.....	58
Table 5: GGA/AGG Developmental Annotation Clusters.....	59
Table 6: CTTC/AAGG Developmental Annotation Clusters (2).....	60
Table 7: CTTC/AAGG Regulatory Annotation Clusters.....	61
Table 8: GGA/AGG Developmental Annotation Clusters.....	62
Table 9: Control Annotation Cluster.....	63

Index of Diagrams

Diagram 1: Dinucleotide Microsatellite Methylation Fold Changes.....	45
Diagram 2: Trinucleotide Microsatellite Methylation Fold Changes.....	47
Diagram 3: Trinucleotide Microsatellite Methylation Fold Changes.....	48
Diagram 4: CTTC, GGA and their Cyclic Permutations' Methylation Fold Changes... ..	52

Index of Figures

Figure 1: Nucleotide Frequencies.....	30
Figure 2: Proportional Chromosome Length / GC Frequency.....	31
Figure 3: Dinucleotide Frequencies.....	32
Figure 4: Trinucleotide Frequencies.....	33
Figure 5: Dinucleotide Microsatellite Frequencies.....	36
Figure 6: Percentage of Dinucleotides in Microsatellites.....	36
Figure 7: Total Nucleotide in Dinucleotide Microsatellites.....	37
Figure 8: Mean Dinucleotide Exponents.....	38
Figure 9: Trinucleotide Microsatellite Quantity.....	39
Figure 10: Percentage of Trinucleotides in Repetitive Regions.....	40
Figure 11: Total Nucleotides in Trinucleotide Microsatellites.....	42
Figure 12: Trinucleotide Average Exponent.....	43

List of Abbreviations and Definitions

SSR Short Sequence Repeat

STR Short Tandem Repeat

PGC Primordial Germ Cell

E Embryonic Day

MeDIP-seq Methylated DNA Immunoprecipitation Sequencing

DAVID Database for Annotation, Visualization and Integration Discovery

Motif Nucleotide sequence of tandem repeating unit

2. Introduction

The principal aim of this study is to determine if there is any regulation of methylation in short repetitive DNA sequences during early embryogenesis in the mouse species *Mus musculus*. If this can be shown it may indicate a functional role for these short tandem repeats.

2.1. Tandem Repeats

The genomes of all eukaryotes so far examined contain many types of repetitive DNA. From long, short and micro repetitive regions, to many as yet unclassified sequences of varying lengths and repetitiveness (1). The advent of high throughput technology has only enhanced our knowledge of the underlying genomic complexity.

Repetitive DNA falls naturally into several distinct groups, regions of DNA that are interspersed with other genomic sequences, retroviral and retrotransposon repeats which have characteristic properties and consecutively repeating sequences, more commonly known as tandem repeats (2). This thesis will focus on the latter and more specifically at the shorter repeats often termed microsatellites, though also known interchangeably in the literature as short sequence repeats (SSR) and short tandem repeats (STR). The term microsatellite will be used throughout to refer to tandem repeats with a repetitive unit (period) of no more than six nucleotides.

Microsatellites are generally thought to arise by a specific mutational mechanism known as DNA slippage (3). The process starts by an initial seeding of short, randomly produced "proto" microsatellites. Then, due to the short sequence length and repetitive nature of both the proto and subsequent microsatellite, during replication the two strands of the DNA molecule can slip

against each other and come out of alignment. The result of this slippage, as replication continues, can be a gain or loss of one or more repeat units. Most, though not all of these so called indel (insertion/deletion) mutations are corrected by the mismatch repair system. Microsatellite repeat number (exponent) is therefore dependent on the primary slippage rate and the efficiency of the mismatch repair system. For any length of DNA it is possible to calculate the predicted number of repetitive regions of a given exponent and period, proportional to the base composition of the DNA (4). However the mechanism of formation leads to much higher mutation rates than would be predicted from base composition alone (5, 6).

The process of microsatellite elongation is therefore an intrinsic property of double stranded DNA and requires no specific enzymes, which has been further confirmed by *in vitro* experiments and studies on cloned microsatellites (7, 8). Indeed some authors, given the mechanism of formation and that conservation between species of both sequence and occurrence is low (9) propose microsatellites to be selectively neutral, randomly or almost randomly distributed over the genome (10, 11). Bachrog *et al.* detected a significant positive correlation between genome-wide AT content and microsatellite density (12), which fits with the slippage mechanism of microsatellite genesis (from a proto microsatellite) and elongation as a random process. Then again as others have suggested some microsatellites are found in regulatory or protein-coding sequences indicating they may be targets of natural selection (13-17). Of course once formed the microsatellites are subjected to the same evolutionary pressures as other genomic regions, and if found selectively useful will be adaptively co-opted. Alternatively it is not difficult to see how they could be negatively selected, especially if found in coding regions. Metzgar *et al.* reported that microsatellites occur much less frequently in coding regions in *Mus* (amongst other genus) (18) . Further, there is some evidence of negative selection for di- and tetranucleotide microsatellites which occur even less frequently in coding regions than trinucleotides (19), presumably as they will

lead to frame-shift mutations.

You-chun Li *et al.* list three core areas where evidence exists for a functional role for microsatellites, chromatin organisation, regulation of DNA metabolic process, and regulation of gene activity (19). For instance many microsatellites are capable of forming unusual yet stable DNA structures, which can produce a mechanism for control of transcription, especially as on unwinding they produce unique protein recognition sites (20). Deletion of specific di-, tri- and tetranucleotide microsatellites has been shown to change transcriptional activity (21, 22). Varying the distance a microsatellite is from a promoter region can alter expression levels of a gene (23), as can changes in the microsatellite exponent (24-27). Hamada *et al* also found that TG microsatellites found in the human genome are mostly between 20 and 60 bp which is also the exponent range for maximum activity (24). Changes in exponent can also lead to changes to, and regulation of translation (28, 29). When found within introns microsatellites have also been shown to affect transcription (30-32). Microsatellites found short distances upstream of genes are known to form binding sites for regulatory proteins (9), and downstream of the start codon have been shown to affect translation levels (33).

In summary microsatellites found within regulatory regions, introns, exons, upstream and downstream, and of varying exponent length have been shown to affect transcription and translation. However, perfect microsatellites do not always reflect biological reality.

2.2. Fuzzy Repeats

Any repeat sequence as well as being subject to slippage mutations is subject to both point and standard indel mutations. This is especially important if we are looking for repetitive sequences that have been co-opted into a functional role.

In some regards evolutionary biology can be seen as a historical science; any microsatellite which has a functional role is likely to be around for a while, allowing mutations time to build up. Perfect repeats are either recent additions to a genome or under high selection for specific sequences. These so called fuzzy tandem repeats are of particular interest.

Longer fuzzy tandem repeats have been found in regulatory regions of eukaryotic genes (34). They have also been shown to form cooperative arrays of binding sites and interact with transcription factors (35-38). Much less is known about the role of shorter repetitive sequences (fuzzy microsatellites), especially when highly mismatched, though they are known to be abundant in both exons and transcription regulating regions (39). A close examination of the distribution of fuzzy microsatellites in relation to the known position of genomic features, such as transcription regulation regions, and frequency within introns, exons and untranslated regions may hint at a functional role.

When found in transcribed regions fuzzy microsatellites can reflect sequence periodicities in protein sequence, or even structural features such as hydrophobic helices (40, 41). In complex regulatory regions, such as enhancers and silencers, fuzzy microsatellites appear to be linked with some types of binding sites for transcription factors (36, 38, 42) and other evidence suggests that a fuzzy microsatellite with a period similar to a binding site can modulate the exact response to the concentration of regulators (43, 44).

2.3. Epigenetic Reprogramming during Development

Epigenetics is concerned with chemical modifications of DNA and of its associated chromatin proteins. While these modification do not directly alter the primary sequence of DNA, epigenetic marks do contain heritable information (both within an organism and between generations) and play key roles in

regulating genome function and in development. The regulation of mammalian development is controlled by sequence-specific transcription factors, but the epigenetic modifications also allow differential cellular states to be established (45). Of these, cytosine methylation is one of the best studied epigenetic modifications of DNA and is known to play an essential role in normal embryonic development (46,47).

DNA methylation can occur on any cytosine, but is *proportionally* most common on when the nucleotide is found in a CG dinucleotide (*i.e.* more likely to be methylated in a CG dinucleotide than an CA, CC or CT). In both mammals and plants this methylation is maintained by DNA methyltransferases. There are several of these enzymes, but the specific action of Dnmt1 (it prefers hemimethylated substrate) suggests a mechanism for the maintenance of specific methylation patterns (48). It is thought that patterns inscribed on the genome at defined developmental time points in precursor cells could be maintained by Dnmt1. These patterns could then lead to predetermined programs of gene expression during development in descendants of the precursor cells (49,50). This would provide a mechanism to explain how patterns of differentiation could be maintained by populations of cells. In addition, specific demethylation events in differentiated tissues could then lead to further changes in gene expression as and when required.

While the epigenetic landscape of the genome is generally stable in somatic cells of multicellular organisms, they are extensively reprogrammed during early development. Initially the methylation levels of mature sperm and egg are similar to those in somatic cells, though the pattern of methylation is not necessarily the same (48). Post fertilisation two, separate genome-wide methylation reprogramming events take place, in primordial germ cells (PGCs) and in the early zygote beginning immediately after fertilisation (51-53).

2.3.1 Early Embryo Reprogramming

From the moment of fertilisation a remarkable transformation of the once sperm genome takes place. The specialised histone molecules needed to compact the sperm genome are replaced by their standard counterparts. On completion genome-wide demethylation occurs in the male genome. This was originally thought to be finished before DNA replication starts (54-56), but recent evidence points to two phases, one before and one soon after DNA replication commences, in the S and G₂ phases (56). The female germ line sequences are protected from these rounds of active demethylation, however they subsequently become demethylated in the next few cell cycles due to the absence of methylation maintenance by Dnmt1 which is excluded from the nucleus (57-60), a more passive method of demethylation.

There are some sequences which escape these demethylation events. For instance, paternally expressed germ line imprints such as H19 and Rasgrf1, this also indicates they must have evolved a demethylation protection mechanism (61). In addition, all differentially methylated regions (DMRs) with a germ line imprint are resistant to passive demethylation (62). How this resistance is brought about is not yet clear.

Subsequent to the end of the demethylation at around embryonic day (E) 3.5, the morula stage just before the time of embryo implantation (52), specific methylation patterns are written to the genome. However, sequences undergoing dynamic DNA methylation during this early phase of embryogenesis remain unknown. Some DMRs are protected from this *de novo* methylation in the embryo. Again the mechanism remains unclear how this is brought about.

2.3.2 Germ Cell Reprogramming

In the mouse PGCs are initially highly methylated and show normal patterns of imprinting (63-67). However, early in their development, they undergo a demethylation event which is complete by E 13 to 14. Recent evidence points to this reprogramming event occurring in two distinct phases. The first phase, starting near E 8.5 sees the demethylation of specific genes, while others are potentially upregulated. After entering the developing gonads, a second phase of reprogramming between E11.5 – E13.5 sees the erasure of imprints and demethylation of many other sequences (reviewed in 78).

Once the genomes of the male and female PGCs have been demethylated, the cells enter mitotic and meiotic arrest for male and female cells respectively. Then, at a later stage the cells are remethylated with presumably developmental patterns. The timing of this process differs between male and female germ cells, with male cells begin remethylation from E 16, but female cells not until after birth.

The genome-wide demethylation event occurs in both male and female germ cells. It is known that during this phase 97% of CpG nucleotides are demethylated, compared to the 70 to 80% methylation found in embryonic stem cells and somatic cells. The demethylation is truly global, with most promoters, genic and intergenic sequences and transposons (68), and importantly all or almost all DMRs becoming hypomethylated.

Feng *et al.* (69) have made two important observations; demethylation in PGCs occurs between E10.5 to E13.5, however some loci are demethylated at earlier stages and therefore demethylation is not necessarily coordinated time-wise throughout the genome (70). Also all current knowledge with regards to demethylation relates to CpG, nothing is known about the erasure of non-CG

methylation in PGCs. The full mechanism for genome wide DNA methylation modification, and the exact timing have yet to be unravelled (51).

As to methylation of microsatellite little is known beyond when things go wrong. For instance, in fragile X syndrome, hypermethylation of expanded CGG or CCG microsatellites in the FMR1 gene leads to hypermethylation of the adjacent CpG rich promoter (71) . As to a possible function, the literature is silent.

Previous analysis of methylated DNA from mouse embryos has suggested that different microsatellites appear in different quantities, but has never been fully studied. This thesis will use bioinformatic approaches to analyse the publicly available *Mus musculus* genome and large datasets of methylated DNA sequences. The information contained within the data may help in addressing this deficiency.

3. Project Objectives

The design for this project can be divided into several distinct sections:

1. Basic analysis of the *Mus musculus* NCBI37 assembly
2. Analysis of microsatellites in the genome
3. Analysis of experimentally determined methylomes
4. Assessing the functional significance

The first three sections are exploratory in nature and will consist of developing visual methods to check for unexpected biases. Section four will use the results from the experimental analysis to ask more detailed questions of different types of microsatellites and ultimately to find a link between microsatellites, their methylation patterns and development.

3.1. Basic analysis of the *Mus musculus* assembly

There are a number of very simple questions to answer before commencing with the main parts of the project, such as the size and nucleotide content of the genome, and the proportions of specific short sequences (motifs) present in the genome. Methods will be developed to predict the expected numbers of each motif in the genome based on the proportion of each nucleotide.

The principal objective of section 1 is exploring the genome and will allow the results from analysing the microsatellite content and the experimental datasets to be put into context. Additionally, to find if there are any underlying genomic differences that will need to be accounted for in subsequent sections.

3.2. Analysis of microsatellites

Locating perfect tandem repeats of any sequence and any length in a genomic sequence is a trivial exercise. However, as discussed in the introduction perfect

repeats are not a true reflection of biology. Allow a certain amount of “fuzziness” in a sequence is more challenging. One of the key objectives of this project is to search for fuzzy tandem repeats. A perfect repeat finder could well discover several microsatellites in what would better be described as a single fuzzy microsatellite and completely miss other repetitive sequences. Though the interesting question investigating the amount of fuzziness in functional microsatellites won't be explored here.

The fuzzy repeat finder must be able report on the consensus pattern and period, location, and exponent for all discovered microsatellites. The consensus pattern is the “best fit” motif which fits data in the tandem repeat. For instance if the basic pattern size is 4, the repeat could potentially be reported as 4, 8, 12 and etc. The consensus for fuzzy repeats might not always be immediately obvious. Then, using the results from the previous section will allow expected microsatellite values to be calculated. The second objective for this section will be to identify any differences between actual and expected values.

3.3. Analysis of Experimental Datasets

The third section will use the fuzzy repeat finder developed in section 2 to find microsatellites within experimental datasets. The datasets are whole genome sequences of methylated DNA derived from mouse cells at various stages of development. Included are both male and female primordial germ cells, embryonic stem cells, and a number of knock-out or knock-down cell lines.

The main project objective will be to look how methylation of microsatellites varies for specific motifs and across the datasets. Hopefully, we will see examples of methylation reprogramming as set out in the introduction. It is not clear what patterns will be found, but ideally datasets from similar cell types at similar developmental times will show similar levels of methylation. As an

example, E 14 PGCs from male and female should have the same methylation pattern, however as the male germ cells are progressively remethylated from day E 16 and the female not, the methylation patterns *could* be different .

3.4. Assessing the Functional Significance

This section will depend on the results of section 3, and whether it seems any methylation levels for microsatellites are being regulated. The experimental analysis results will be used in conjunction with the assembled genome to try and find any positional relationship between the interesting microsatellite(s) and genes.

3.4.1 Finding genes

Using the algorithms developed in section 2, or developing further algorithms it will be possible to locate microsatellites within genes, or within a specified number of base pairs, both upstream and/or downstream. Ideally a generalised algorithm should be developed so if the chromosomal position of a microsatellite is known its proximity to any structure with a known position can be found. As made clear in the introduction, proximity to genes seems to be key in determining whether a microsatellite potentially has a functional role.

3.4.2 Finding Functional Relationships

After identifying genes the next step will be to see if there is any functional similarity between them. In view that the aim of this project is to determine if methylated microsatellites play a role in development, finding genes or functional networks related to development would be ideal.

4. Methodology

Scripts written in Perl and R have been used extensively throughout the methodology. Perl was chosen for its string handling capabilities and R for its superior handling and manipulation of large datasets.

4.1. Basic analysis of the *Mus musculus* genome

Perl scripts have been written (available in supplementary material) to calculate the following characteristics from the publicly available NCBI m37 *Mus musculus* genome:

1. length of each chromosome
2. base composition
3. dinucleotide prevalence
4. trinucleotide prevalence

The values for the base composition and nucleotide prevalence need to be corrected to take account of the complementarity of DNA. This is best shown if you consider the union of a DNA sequence S and its reverse complement S^I into a single sequence $S+S^I=\bar{S}$. Then, following Burge *et al.* (72), let f_x denote the discovered frequency of nucleotide X (A, T, C or G), f_{xy} the frequency of a dinucleotide and f_{xyz} the frequency of a trinucleotide.

Therefore the frequency of $f_{\bar{a}}$ is equal to $f_{\bar{t}}$ is equal to $\frac{1}{2}(f_a + f_t)$, with the same calculation possible for $f_{\bar{c}}$ and $f_{\bar{g}}$. Similar corrections are made for the frequencies $f_{\bar{xy}}$ and $f_{\bar{xyz}}$. e.g. $f_{\bar{GA}}$ is equal to $f_{\bar{TC}}$ is equal to

$$\frac{1}{2}(f_{GA} + f_{TC})$$

4.1.1 Calculating the expected levels of di- and trinucleotides from base composition

Using the same notation as above the following equation is used to calculate the relative frequency of a dinucleotide:

$$P(\bar{x}\bar{y}) = \frac{f_{\bar{x}\bar{y}}}{f_{\bar{x}} \cdot f_{\bar{y}}}$$

e.g. again for the tandem repeat with motif GA;

$$P(\bar{G}\bar{A}) \text{ is equal to } \frac{1}{2}(f_{GA} + f_{TC}) / \frac{1}{4}(f_G + f_C) \cdot (f_A + f_T)$$

It follows from this equation that the expected values for the sixteen possible dinucleotides can be calculated using the formula:

$$E(\bar{x}\bar{y}) = \frac{\sum [\bar{x}\bar{y}] \cdot f_{\bar{x}} \cdot f_{\bar{y}}}{16}$$

Where $\sum [\bar{x}\bar{y}]$ = the sum of the corrected values for all dinucleotides.

Calculating the expected values for trinucleotides follows the same process. However a third order correction must be made to account for the frequencies of the constituent dinucleotides f_{xy} , f_{yz} and f_{xnz} (where n can be any nucleotide):

$$P(\bar{x}\bar{y}\bar{z}) = \frac{f_{\bar{x}\bar{y}\bar{z}} \cdot f_{\bar{x}} \cdot f_{\bar{y}} \cdot f_{\bar{z}}}{f_{\bar{x}\bar{y}} \cdot f_{\bar{y}\bar{z}} \cdot f_{\bar{x}\bar{z}}}$$

Again the expected values for the sixty four different trinucleotides can now be calculated using the above equation to give:

$$E(x\bar{y}z) = \frac{\sum [x\bar{y}z] \cdot f_{\bar{x}} \cdot f_{\bar{y}} \cdot f_{\bar{z}} \cdot f_{x\bar{y}} \cdot f_{\bar{y}z} \cdot f_{x\bar{z}}}{64}$$

Where $\sum [x\bar{y}z]$ = the sum of the corrected values for all trinucleotides

Below is a worked example for calculating the expected value for the motif GGA.

The following required data is taken from tables S1, S2 and S3:

Nucleotide	Count	Nucleotide	Count
GGA	41977221	TCC	41995168
CC	134073230	GG	134087829
GA	159020818	TC	159111635
GAA	52048862	TTC	52094358
GCA	35875336	TGC	35898020
GTA	29433240	TAC	29399473
A	744681828	T	745397519
C	534146040	G	534300392
Total nucleotides	2558525779		

Nucleotide counts for given motifs. Taken from tables S1, S2 & S3

The first step is to calculate the corrected values to account for reverse complements for all the above table (with the exception of total nucleotides). This is straightforward, for GGA and TCC:

$$x\bar{y}z = (41977221 + 41995168) / 2 = 41986194.5$$

The complete set of corrected values are given in the table below. All further calculations will use these corrected figures.

Base complements	Corrected count
GGA/TCC	41986194.5
CC/GG	134080529.5
GA/TC	159066226.5
GAA/TTC	52071610
GCA/TGC	35886678
GTA/TAC	29416356.5
A/T	745039673.5
C/G	534223216

Values corrected for base complementarity

The next step is to calculate the frequency of G and A:

$$f_{\bar{x}} = f_{\bar{y}} = (\text{Total G} / \text{Total Nucleotides}) * 4$$

$$= (534223216 / 2558525779) * 4 = 0.84$$

$$f_{\bar{z}} = (\text{Total A} / \text{Total Nucleotides}) * 4$$

$$= (745039673.5 / 2558525779) * 4 = 1.12$$

Next, the frequency of the constituent dinucleotides is required. These are GG, GA and GnA, where n is any nucleotide. GnA is derived from the data in the corrected values table:

$$\text{GnA} = (\text{GAA} + \text{GCA} + \text{GGA} + \text{GTA}) / 4 = 159360839$$

To calculate the frequency of a GA dinucleotide

$$f_{\bar{y}\bar{z}} = (\text{GA total} / (\text{G total} * \text{A total})) * \text{total nucleotides}$$

$$= (159066226.5 / (534223216 * 745039673.5)) * 2558525779 = 1.02$$

$$f_{\bar{x}\bar{y}} = (134080529.5 / (534223216 * 534223216)) * 2558525779 =$$

$$1.20$$

$$f_{x\bar{y}z} = (159360839 / (534223216 * 745039673.5)) * 2558525779 = 1.02$$

The sum of all trinucleotides $\sum [x\bar{y}z] = 2558524482$

Therefore the expected value for GGA =

$$E(x\bar{y}z) = (2558524482 * 0.84 * 0.84 * 1.12 * 1.02 * 1.20 * 1.02) / 64 = 40899086$$

(Note that the sum of nucleotides does not equal total nucleotides - 2 as might be expected, due in part to the data coming from individual FASTA files for each chromosome, but more so due to there being a few unknown regions in the genome. For each unknown region you lose 2 from the trinucleotide count in comparison to the expected value given the nucleotide count.)

4.1.2 Calculating the expected number of tandem repeats

Using the formula given by deWachter (4) it is possible to calculate the expected number of tandem repeats of a specific motif and exponent in any length of DNA, given the probability of the motif, the exponent number and the number of nucleotides in the DNA.

$$E(M_e) = fM^e \cdot (1 - fM)^2 \cdot (2p + (N - ep - 2p + 1))$$

Where fM is the probability (corrected as above) of a motif, e is the exponent, p is the period and N is the number of nucleotides in the sequence.

As a worked example, using the data from tables S1 and S3, for the motif GGA and an exponent length of 4.

GGA Count	41977221
TCC Count	41995168
Total Nucleotides	2558525779

To calculate the corrected GGA count:

$$\begin{aligned} \overline{GGA} &= (\text{GGA count} + \text{TCC count}) / 2 \\ &= (41977221 + 41995168) / 2 = 41986194.5 \end{aligned}$$

To calculate the motif probability:

$$\begin{aligned} fM &= \text{corrected count} / (\text{nucleotide count} - \text{period} + 1) \\ &= 41986194.5 / (2558525779 - 3 + 1) = 0.01641031 \end{aligned}$$

Therefore:

$$\begin{aligned} &fM^e \cdot (1 - fM)^2 \cdot (2p + (N - ep - 2p + 1)) \\ &= 0.01641031^4 \cdot (1 - 0.01641031)^2 \cdot (2 \cdot 3 + (2558525779 - 4 \cdot 3 - 2 \cdot 3 + 1)) \\ &= 179.5084 \end{aligned}$$

The expected number of GGA repeats produced by chance in the *Mus musculus* genome with an exponent of 4 will be 180.

An alternative method for calculating the expected number of tandem repeats is by simulation using the data. tandem_sim2.pl (supplementary material) is an example of such a simulation.

4.2. Analysis of microsatellites

Gary Benson's Tandem Repeats Finder program (73) was used to locate all tandem repeats with a period of five or fewer nucleotides. This program returns a list of all repeats of the specified period and lower, their chromosomal locations, the consensus repeat motif and the actual repeat sequence (amongst other data). For sequences which can be described equally well with more than one period, the program limits the redundancy in the output to at most three. In this case as microsatellites of five or fewer repeats are used some repeats may be reported at both period 2 and 4.

The following parameter values were used for all runs of Tandem Repeat Finder:

Mismatch penalty	5
Indel penalty	5
Minimum score	50
Maximum period	5

The minimum score was the default setting. Mismatch penalty and indel penalty set the permissiveness of these types of sequence disparities, and are both set to the middle of three settings. Middle range settings were chosen as the method was used to examine proportional quantities of microsatellites of varying motif. Extreme values could potentially favour particular motifs.

4.3. Analysis of Experimental Datasets

4.3.1 Experimental Data

No.	Sample	Description	Group
1	J1 1 50ng (2-181 2B)	Embryonic stem cells	1
2	J1 1 400ng (2-172 B)	Embryonic stem cells	1
3	PGC 11.5 1 (3-16(1))	Primordial germ cells	2
4	P3MEF methyl	Embryonic Fibroblast	10
5	TKO methyl	Triple methylase knockout	5
6	N95 methyl	Maintenance methylation knockout	3
7	Sperm methyl	Mature sperm	9
8	J1 2 (3-170 5)	Embryonic stem cells	1
9	Tet methyl	Tet knockdown (RNAi)	4
10	PGC 13.5 male AID (3-182/197 – 5B)	AID knockout	(7)
11	PGC 13.5 female AID (3-182 – 6B)	AID knockout	(7)
12	PGC 16.5 male 1 (3-182 1B)	Primordial germ cells	6
13	PGC 16.5 female 1 (3-182 2B)	Primordial germ cells	7
14	PGC 13.5 male 1 (3-182 3B)	Primordial germ cells	7
15	PGC 13.5 female 1 (3-182 4B)	Primordial germ cells	7
16	PGC 11.5 AID (3-182 7B)	AID knockout	(2)
17	E14-ES-methyl	Embryonic stem cells	1
18	E14-EmBod-methyl	Embryoid body	8
19	J1-ES-p14 MeDIP	Embryonic stem cells	1
20	E14-ES-Rua-MeDIP	Embryonic stem cells	1
21	EB-E14-D13-Rua-MeDIP	Embryoid body	8
22	NP95 -/- ES-MeDIP	Maintenance methylation knockout	3
23	Tet1/2-KD L2 MeDIP	Tet knockdown (RNAi)	4
24	Tet1 KD Single L1 MeDIP	Tet knockdown (RNAi)	4
25	PGC 11.5 2 (4-127 1B)	Primordial germcells	2
26	PGC 16.5 male 2 (4-127 4B)	Primordial germcells	6
27	PGC 13.5 male 2 (4-127 2B)	Primordial germcells	7
28	PGC 13.5 female 2 (4-127 3B)	Primordial germcells	7
29	PGC 16.5 female 2 (4-127 5B)	Primordial germcells	7

Table 1: Experimental Data Description. The experimental datasets, a description and grouping based on expected similar methylation patterns.

MeDIP-seq datasets consisting of full genome sequences of methylated DNA were made available in FASTQ format, as set out in Table 1. MeDIP-seq (Methylated DNA immunoprecipitation sequencing) is a technique for isolating methylated DNA fragments, followed by sequencing of these fragments. Sequencing in this case was via the Illumina platform with reads of 40bp. Each dataset has been added to groups which are expected to have similar

methylation patterns.

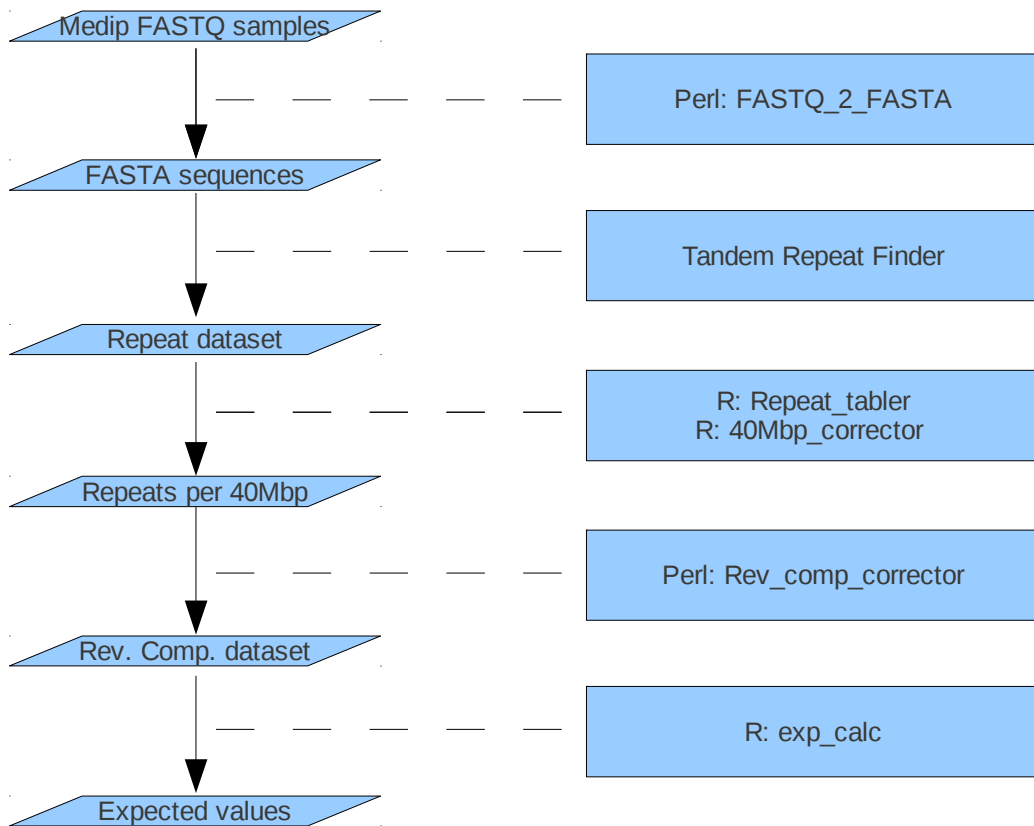
The data consists mainly of two types of cells, ES and PGCs. The ES cells are expected to show similar methylation patterns. The PGCs are taken from various stages of methylation reprogramming and it is reasonable to assume they will show variation in methylation pattern. E 11.5 is close to the end of demethylation, E 13.5 is after demethylation, but before remethylation commences. By day E16.5 males germ cells will have started remethylation, but female germ cells will still be demethylated.

The embryoid body includes cells at various stages of differentiation and presumably different methylation patterns, while mouse embryonic fibroblast are no longer pluripotent and sperm fully differentiated and consequently varying methylation patterns.

AID deficient samples are expected to show a deficiency in the demethylation of CpGs (78). Tet1 and Tet2 are 5mC hydroxylases, cells with these enzymes reduced might show a reduction in demethylation. N95 is a methylation maintenance enzyme, removal or reduction in this enzyme could lead to a reduction of methylation levels.

4.3.2 Calculating the expected number of methylated tandem repeats

A limitation of Tandem Repeat Finder is it accepts input in FASTA file format only. As the experimental data was in FASTQ format it required conversion to FASTA. Workflow 1 was used for calculating expected levels of repeats from the FASTQ experimental datasets.



Workflow 1: *FASTQ_2_FASTA* converts the sequence portion of a FASTQ file to FASTA. *Tandem Repeat Finder* finds fuzzy microsatellites in the FASTA formatted sequence produced by *FATSQ_2_FASTA* and returns a data set of all repeats, their motifs, exponent, and consensus period. *Repeat_tabler* takes the repeat dataset and calculates numbers of microsatellites per motifs and returns them as a table. *40Mbp_corrector* converts the microsatellites per motif to microsatellites per motif per 40Mbp (allows comparison across samples). *Rev_comp_corrector* corrects the values for base pare complementarity (as shown previously). *Exp_calc* calculates the expected values from the Rev. Comp. dataset using the formula: $\frac{\text{row means}}{\sum(\text{row means})} \times \sum(\text{dataset values})$ where row means is the mean value for a motif across all datasets.

4.3.3 Calculating Fold Change

To reduce the number of results a selection criterion was applied.

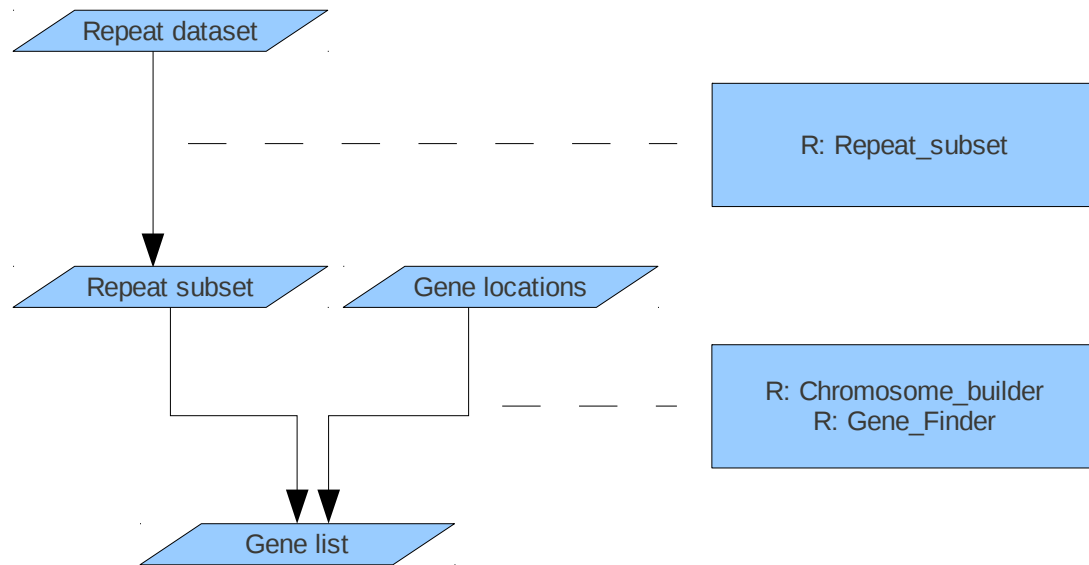
All motifs where the mean number of microsatellites per 40Mbp across the datasets was 2 or less were excluded from analysis. A mean of 2 was chosen as anything lower than this would potentially have less than 30 microsatellites in some datasets (*i.e.* some datasets were less than 600Mbp). For tetra nucleotides a second selection criterion was also applied to reduce the number of motifs to 20. The selection criterion consisted of multiplying the variance across datasets for each motif by the observed / expected variance across datasets for each motif. The 20 highest scoring motifs were selected. This method was chosen to reduce the large biases towards sample mean (towards either high or low) if either variance measure was used alone.

Observed / expected results for motifs passing the selection criteria were calculated to examine fold change. For scoring a positive or negative fold change two different thresholds were used depending on the mean value for a motif across all datasets, above 30 +/- 1.5, below 30 +/- 2. Though this dividing point was fairly arbitrary, two thresholds were chosen to reduce (the higher probability of) small sample values passing a fold change threshold by chance.

4.4. Assessing the functional significance

4.4.1 Finding Genes

Gene list containing tandem repeats or located close to tandem repeats were discovered using Workflow 2.



Workflow 2: The Repeat dataset was created as per Workflow 1. The Gene locations dataset contains the start and end points, and Ensembl gene ID for all genes (or any other available genome structure) in the *Mus musculus* assembly. Repeat_subset divides the repeat dataset by motif. Chromosome_builder builds an array containing the location for each motif in the Repeat subset dataset and the location for each gene. Gene_Finder takes the Chromosome_builder array and returns a list of genes, given an input motif, which are either located within a gene, within a specified number of bases (upstream or downstream), or within any other structure contained in the gene locations file, e.g. with exons or introns.

4.4.2 Generating Random Gene Lists

To generate a random selection of genes biased for size an iterative method was developed in Perl. Using a list of all *Mus musculus* genes in the NCBI37 assembly, including their size, a random gene was selected and the size was stored. Iteration continued to randomly select genes until the sum of the size

for any gene was larger than a threshold value, this particular gene was then marked as chosen. Again iteration continued until the required number of genes had been chosen. The Ensembl ID of the chosen genes was then returned.

4.4.3 Gene Function

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v.6.7 (74, 75) was used to find developmental processes associated with the gene lists generated by the Gene_Finder algorithm.

The Functional Annotation Chart function was used to produce a list of related terms to the input gene-lists. Terms were investigated further depending on passing two criteria. First the term must have had a greater than two fold enrichment score (computed by DAVID) and secondly the term did not occur in any control groups, or had a minimum fold change at least twice that for all control groups. Two control gene-list were included consisting of 1000 and 2000 genes created using the method given in 4.4.2. The threshold value was set to 5000000 based on the size of the largest gene (4434883 bp), guaranteeing that any gene would need to be randomly chosen twice before being selected. These were used to control for gene size as it was assumed that microsatellites had a higher chance of occurring in larger genes.

Genes associated with the terms that passed the criteria were then used with the DAVID Functional Annotation Clustering Tool to return clusters of related terms. The Functional Annotation Clustering Tool was used with the classification stringency set to low to allow weakly related terms to form a single cluster. This was to increase the chance of developmental processes being classified in the same cluster rather than several separate clusters. For control purposes a further 400 genes randomly selected for size bias as per 4.4.3 were used with the Functional Annotation Clustering tool to compare clusters.

5. Results and Discussion

5.1. Basic analysis of the *Mus musculus* genome

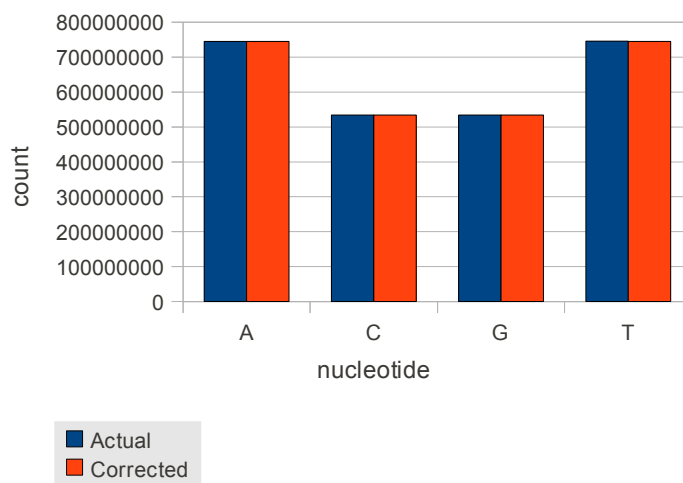


Figure 1: Nucleotide Frequencies. Total count of nucleotides found in *Mus musculus* NCBIM37 assembly and their corrected values to allow for base complementarity. Corrected values are calculated such that $(C+G)/2$ for C and G, and $(A+T)/2$ for A and T

Figures 1 and 2 summarise the raw nucleotide data (available in table S1).

Figure 1 is a graphical representation of nucleotide quantities, showing both the actual quantities calculated from the genome and corrected values to take account of the complementarity of DNA. Methods section 4.1 explains in detail how the corrected values have been calculated and why high concordance is expected.

Figure 2 illustrates the effect of chromosome length on GC content, by showing the GC frequency for each chromosome as a proportion of the total genome length. Apart from the shortest chromosome (actually the mitochondrial genome), they are all clustered at or near the mean value of 41.8% GC. This is important as it is known that the nucleotide proportions influence the rate of formation of new microsatellites (12). Any difference particularly on longer chromosomes could unduly skew the proportion of certain microsatellites.

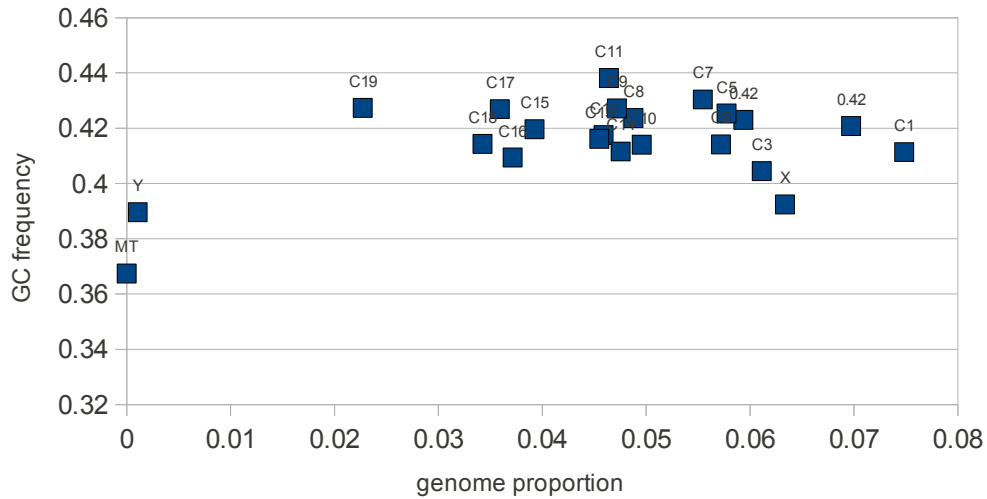


Figure 2: Proportional Chromosome Length / GC Frequency. *Mus musculus* NCBIM37 assembly GC frequency for each chromosome as a proportion of the total genome length.

Dinucleotide and trinucleotide values were calculated as given in the methods section 4.1, and expected values were calculated from base composition using the equations given in methods 4.1.1. Figure 3 and Figure 4 depict the actual and expected di- and trinucleotide quantities respectively (original data available in Tables S2 and S3). To remove duplicate data only a single value for each complementary pair has been graphed. For consistency, for each complementary motif the first alphabetically has been used throughout all graphs.

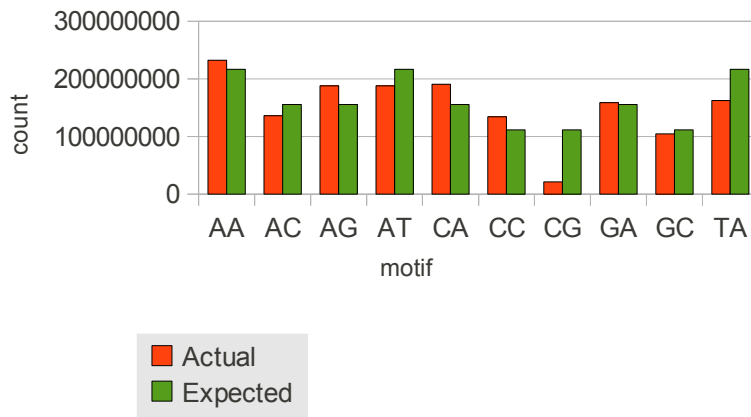


Figure 3: Dinucleotide Frequencies. Actual and expected count of dinucleotides found in the *Mus musculus* NCBIM37 assembly. The expected values were calculated using the formula $E(\bar{x}\bar{y}) = \frac{\sum [\bar{x}\bar{y}] \cdot f_{\bar{x}} \cdot f_{\bar{y}}}{16}$ where $\sum [\bar{x}\bar{y}]$ is the sum of dinucleotides found (corrected for their reverse complement) $f_{\bar{x}}$ and $f_{\bar{y}}$ are the corrected frequency of the constituent nucleotides.

The dinucleotide graph shows very clearly the under representation of CG and at levels much lower than expected from nucleotide content. This result is expected and is due to the well known phenomenon of deamination of methylated CpG (75). However, as the expected results for trinucleotides contain second order corrections for their constituent dinucleotides, this pattern is not seen in the trinucleotide graph.

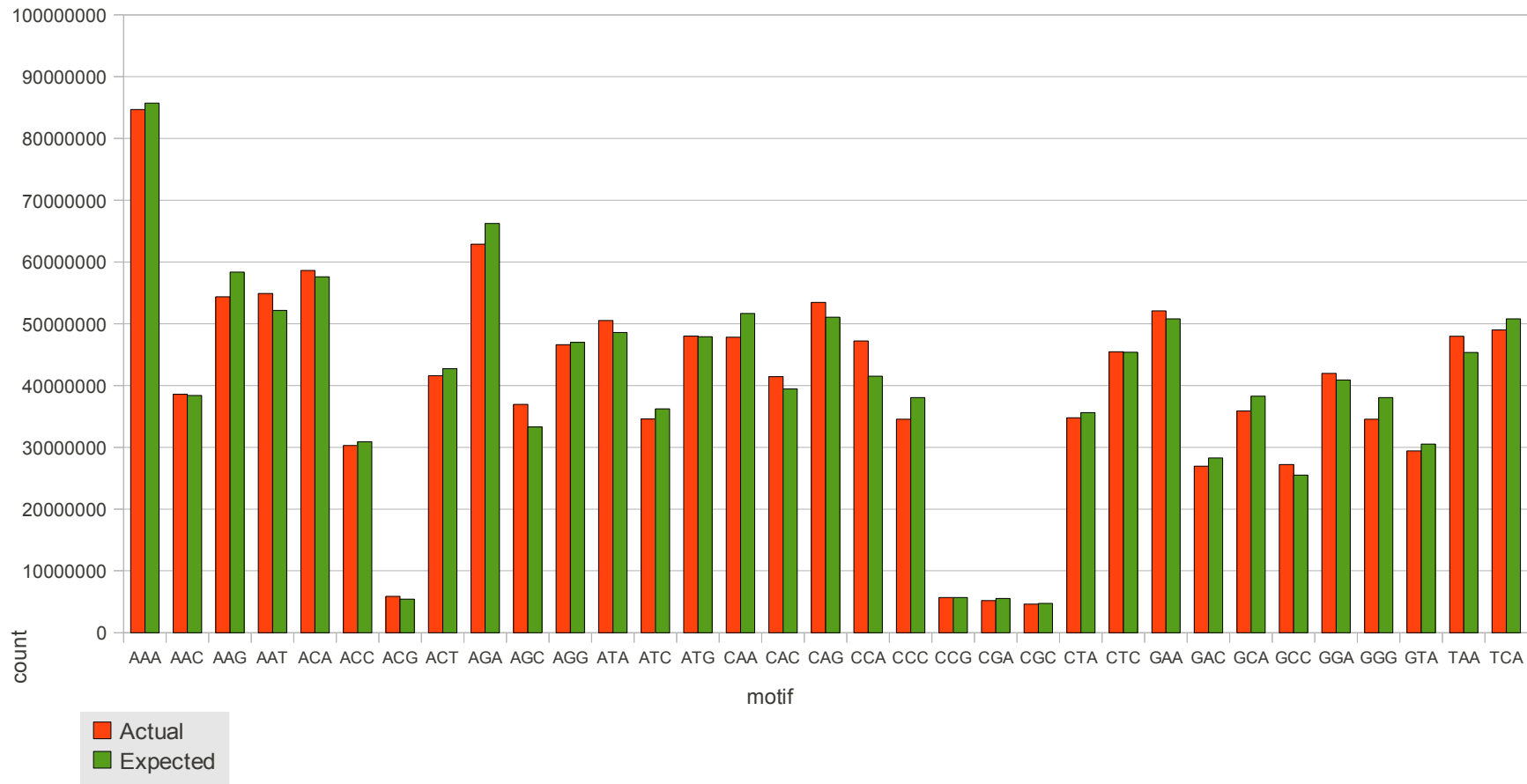


Figure 4: Trinucleotide Frequencies. Actual and expected count of trinucleotides found in the *Mus musculus* NCBIM37 assembly. The expected values were calculated using the formula $E(x\bar{y}z) = \frac{\sum [x\bar{y}z] \cdot f_{\bar{x}} \cdot f_{\bar{y}} \cdot f_{\bar{z}} \cdot f_{x\bar{y}} \cdot f_{\bar{y}z} \cdot f_{x\bar{z}}}{64}$ where $\sum [x\bar{y}z]$ is the sum of trinucleotides found (corrected for their reverse complement), $f_{x\bar{y}}$, $f_{\bar{y}z}$ and $f_{x\bar{z}}$ the corrected frequency of trinucleotides (with n representing any nucleotide), and $f_{\bar{x}}$, $f_{\bar{y}}$ and $f_{\bar{z}}$ the corrected frequency of the constituent nucleotides.

5.2. Analysis of microsatellites

The microsatellite composition of the *Mus musculus* genome was discovered using the workflow 4.2 . Microsatellites with a period of two and three nucleotides have been analysed.

5.2.1 Expected number of tandem repeats

The Tandem Repeat Finder program will find repeats of a minimum exponent which is dependent on the settings used. Table 2 shows the *minimum* exponent found for dinucleotides and the expected values due to chance, calculated using the formula given in the methods. It's clear, as predicted by theory that the dinucleotide microsatellites in *Mus musculus* differ markedly from those predicted by chance association of nucleotides. A similar pattern is shown for trinucleotides (table S6).

Motif	Minimum Exponent	Predicted Count	Average Exponent	Predicted Average
AA	51	1.55E-44	72.75	1.10
AC	12.5	1.48E-05	29.28	1.06
AG	12.5	1.48E-05	32.36	1.08
AT	12.5	1.76E-05	38	1.08
CA	12.5	2.25E-07	32	1.08
CC	NA	NA	NA	1.06
CG	12.5	2.61E-17	14.24	1.01
CT	12.5	1.48E-05	32.36	1.08
GA	12.5	1.87E-06	31.9	1.07
GC	12.5	1.05E-08	14.71	1.04
GG	NA	NA	NA	1.06
GT	12.5	2.79E-07	29.28	1.06
TA	12.5	2.47E-06	37.92	1.07
TC	12.5	1.87E-06	31.9	1.07
TG	12.5	1.76E-05	32	1.08
TT	13.5	1.86E-05	13	1.10

Table 2: Dinucleotide Microsatellite Exponents. *Dinucleotide microsatellites found with Tandem Repeat Finder, the shortest and average microsatellite exponent, the expected number of microsatellites of the given minimum exponent based on nucleotide prevalence and the expected average microsatellite exponent.*

Graphical representations for di- and trinucleotide repeats are presented below. Figures have been provided showing the actual and expected values for the

quantity of di- and trinucleotide microsatellites (figures 5 and 9), total nucleotides in repetitive regions (figures 6 and 10), exponent length and percentage of nucleotides in repetitive regions (figures 7 and 11). All graphs represent similar ideas, but taken as a whole give a clearer overview of what is happening. It's worth noting that the expected values are relative to all actual values found, *i.e.* comparison is being made between the motifs.

There are some consistent patterns across all graphs, first there are two points to note. The almost complete lack of mononucleotide repeats (AA, GG, TTT, CCC, and *etc.*) is due to a limitation of the repeat finding algorithm, in almost all circumstances mononucleotides are classified as single period repeats (there are 60246 single period repeats in the data, almost exclusively poly A and poly T). For this reason they have been removed from all calculations and are not shown on the graphs. Secondly, again consistent with the results from the basic analysis is the almost complete lack of any repetitive regions containing CG (note these are repetitive regions so include repeats such as GCC and GAC). The results of repeats containing CG have been left in the calculations as they represent actual biology and are not an artefact of repeat finding.

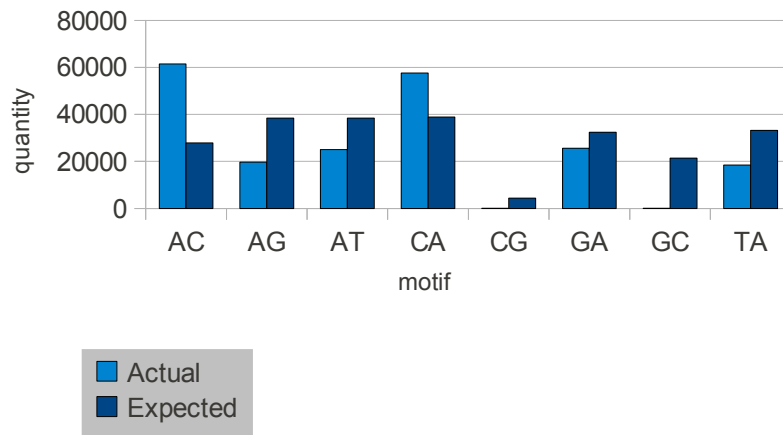


Figure 5: Dinucleotide Microsatellite Frequencies. Actual and expected count of dinucleotide microsatellites found in the *Mus musculus* NCBIM37 assembly. The expected values have been calculated using the formula $E(M_{xy}^q) = \frac{\sum [M_{xy}^q] \cdot P(\bar{xy})}{16}$ where $\sum [M_{xy}^q]$ is the sum of all dinucleotide microsatellites (corrected for their reverse compliments) and $f \bar{xy}$ is the corrected frequency of the specific dinucleotide.

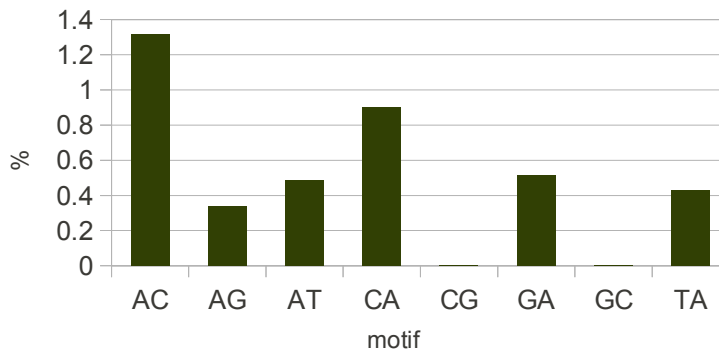


Figure 6: Percentage of Dinucleotides in Microsatellites. Dinucleotide microsatellites as a percentage of the *Mus musculus* NCBIM37 assembly.

Examining the dinucleotide graphs, there is a clear over representation, both in total nucleotide quantity and percentage of dinucleotides in microsatellites for AC. However, the mean exponent graph indicates that these microsatellites do not differ in length in comparison to other microsatellites. This is consistent with the findings of Kruglyak *et al* in yeast (77) and Bachtrog *et al.* for drosophila melanogaster (12).

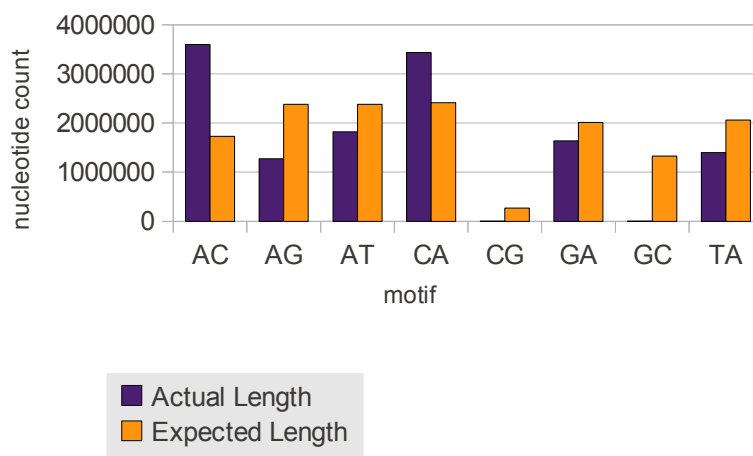


Figure 7: Total Nucleotide in Dinucleotide Microsatellites. Total and expected number of nucleotides found in dinucleotide microsatellites. The expected number of nucleotides has been calculated using the formula $E(M_{xy}^l) = \sum [M_{xy}^l] \cdot f_{\bar{x}\bar{y}}$ where $\sum [M_{xy}^l]$ is the sum of nucleotides in all dinucleotide microsatellites (corrected for their reverse compliments) and $f_{\bar{x}\bar{y}}$ is the corrected frequency of the specific dinucleotide.

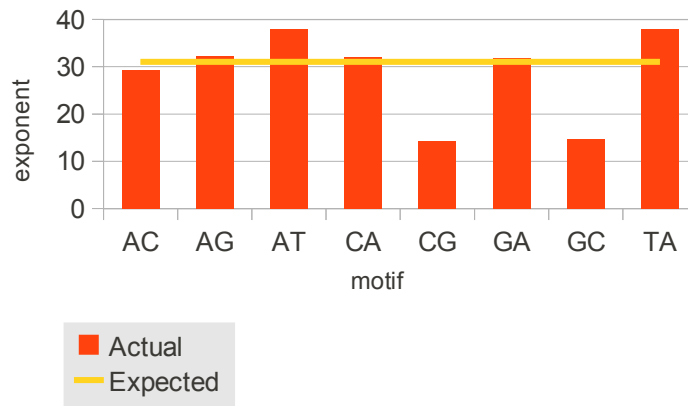


Figure 8: Mean Dinucleotide Exponents. Mean dinucleotide exponents and the expected mean. Expected mean has been calculated from the expected total microsatellite length and expected microsatellite quantity.

The trinucleotide graphs show a similar, though slightly more complex pattern. Again, there are clear differences in the proportions of trinucleotides in microsatellites, and the quantity of microsatellites of specific motifs. However, unlike dinucleotides some of these motifs also have a longer average exponent. A closer examination of the exponent length graph reveals an unexpected pattern. The six motifs which are over represented are AAG, AGA, GAA, AGG, GAG and GGA. These are exclusively motifs containing both and only A and G nucleotides. With the exception of AGA, microsatellites of these motifs are also found in excess. This may lead to the suggestion that the process of microsatellite synthesis favours these particular motifs. However, Schlotter and Tautz (7) provide evidence that the mechanism of microsatellite formation might favour motifs with higher AT content. GGA and AGA (they don't present data for the other four motifs) come in the middle of their range of trinucleotide microsatellites in terms of synthesis rate and total length.

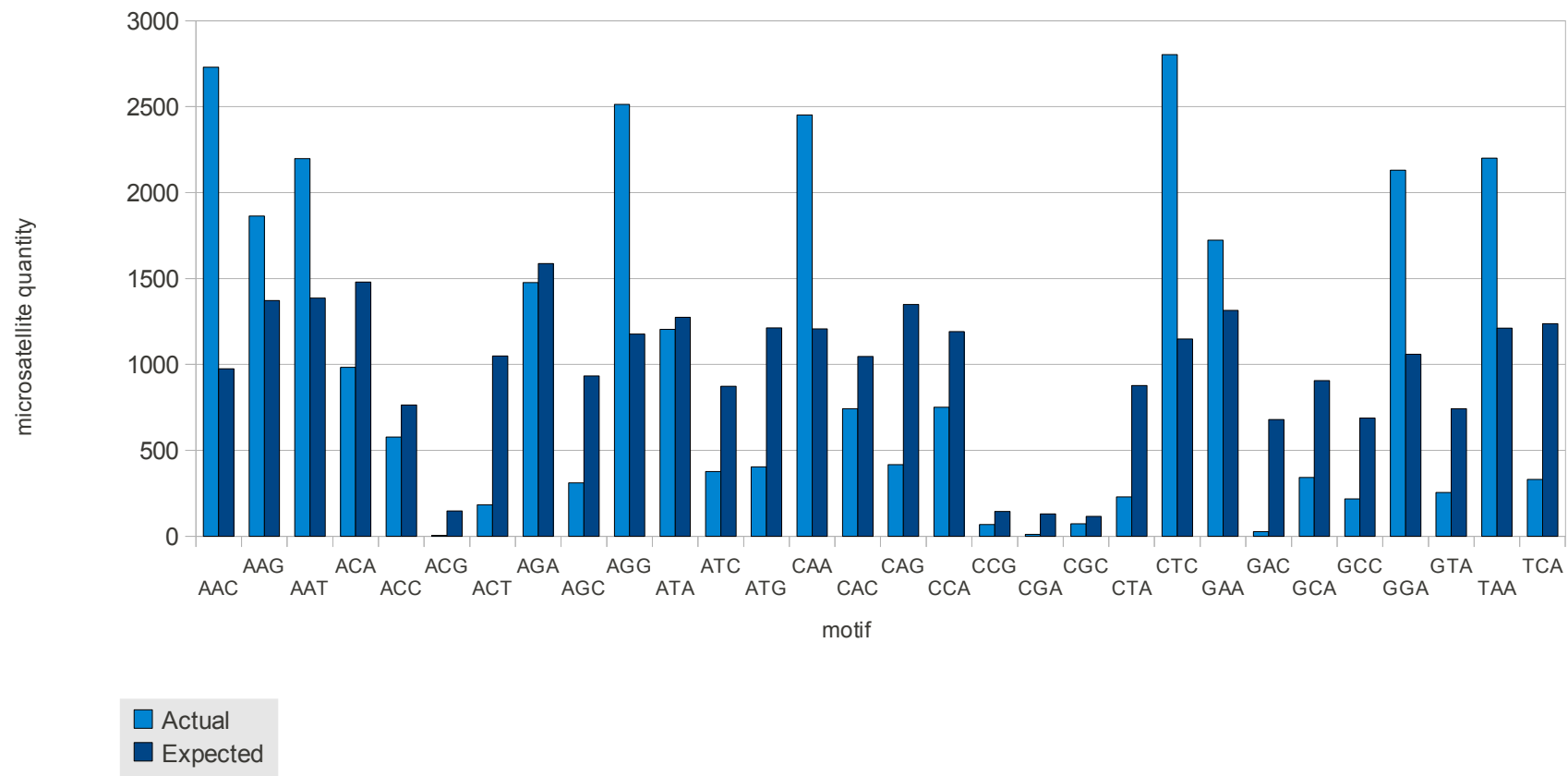


Figure 9: Trinucleotide Microsatellite Quantity. Actual and expected count of trinucleotide microsatellites found in the genome. The expected values have been calculated using the formula $E(M_{xyz}^q) = \frac{x\bar{y}z}{\sum [x\bar{y}z]} \cdot \sum [M_{xyz}^q]$ where $\sum [M_{xyz}^q]$ is the sum of all trinucleotide microsatellites (corrected for their reverse compliments), $x\bar{y}z$ is the corrected number of the specific trinucleotide and $\sum [x\bar{y}z]$ is the sum of all trinucleotides.

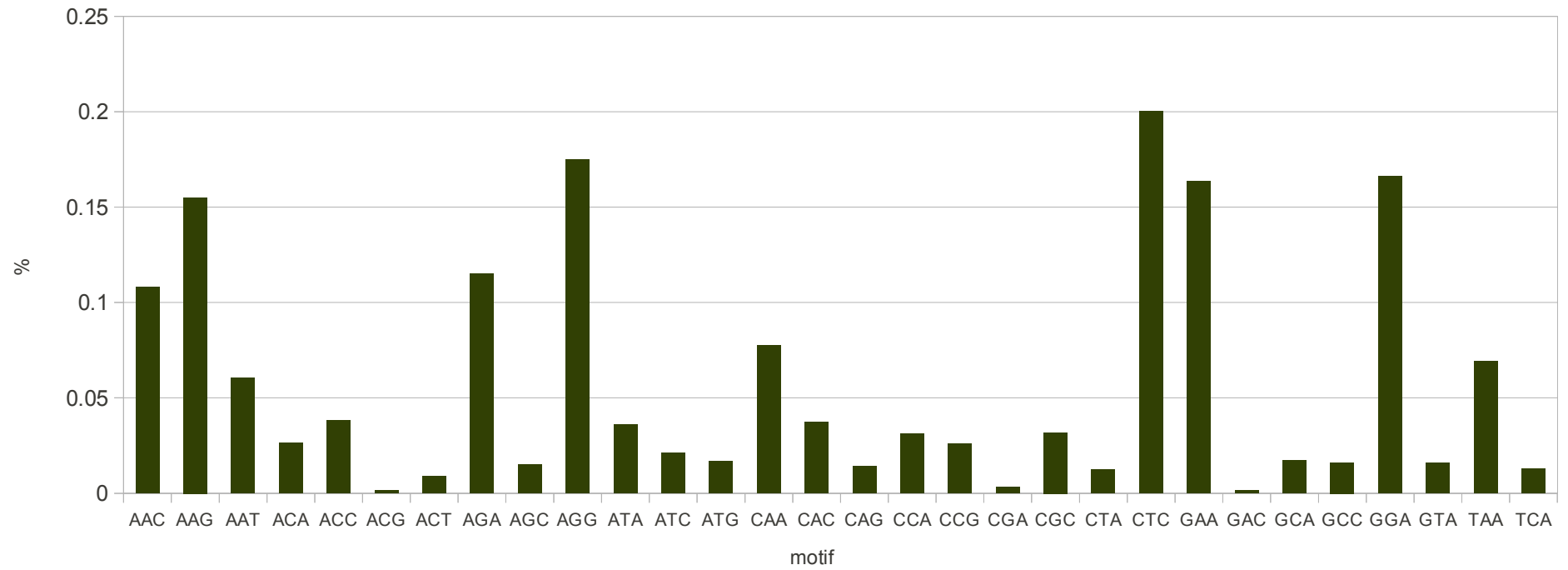


Figure 10: Percentage of Trinucleotides in Repetitive Regions. Trinucleotide microsatellites as a percentage of the *Mus musculus* NCBIM37 assembly.

A second unexpected result is found when examining motifs containing two C nucleotides. As would be expected, when coupled with G they are under represented compared to the expected values in all graphs. What is not expected is they are also all under represented when coupled with A, but they are all over represented when with T. Thirdly and perhaps the most unexpected finding is a very clear difference between motifs consisting of two different nucleotides and those with three. All motifs with three nucleotides are under represented in all graphs.

There are interesting questions to be asked about these observations, what microsatellites are over represented, do they follow a pattern and why are they over represented? Equally important are the reverse condition, why are some microsatellites under represented?

The answer to the first two questions is in the data. The over represented have already been stated, and an obvious pattern from looking at the graphs is that not only are they over represented they are also the most prevalent. This would certainly not be an *a priori* necessary condition for over representation. As to the why, could they be performing particular functions, are they being favoured by selection? Again for the under represented microsatellites, are they being selected against because of particular structural features or do they have a general regulatory function and are negatively selected if they're in the wrong place? These are not exclusive possibilities, though it should be noted that the method used to calculate the expected results means they are relative to each other.

The other questions concerning the unexpected patterns within the trinucleotide data, though not pursued here would form an interesting starting point for a future study.

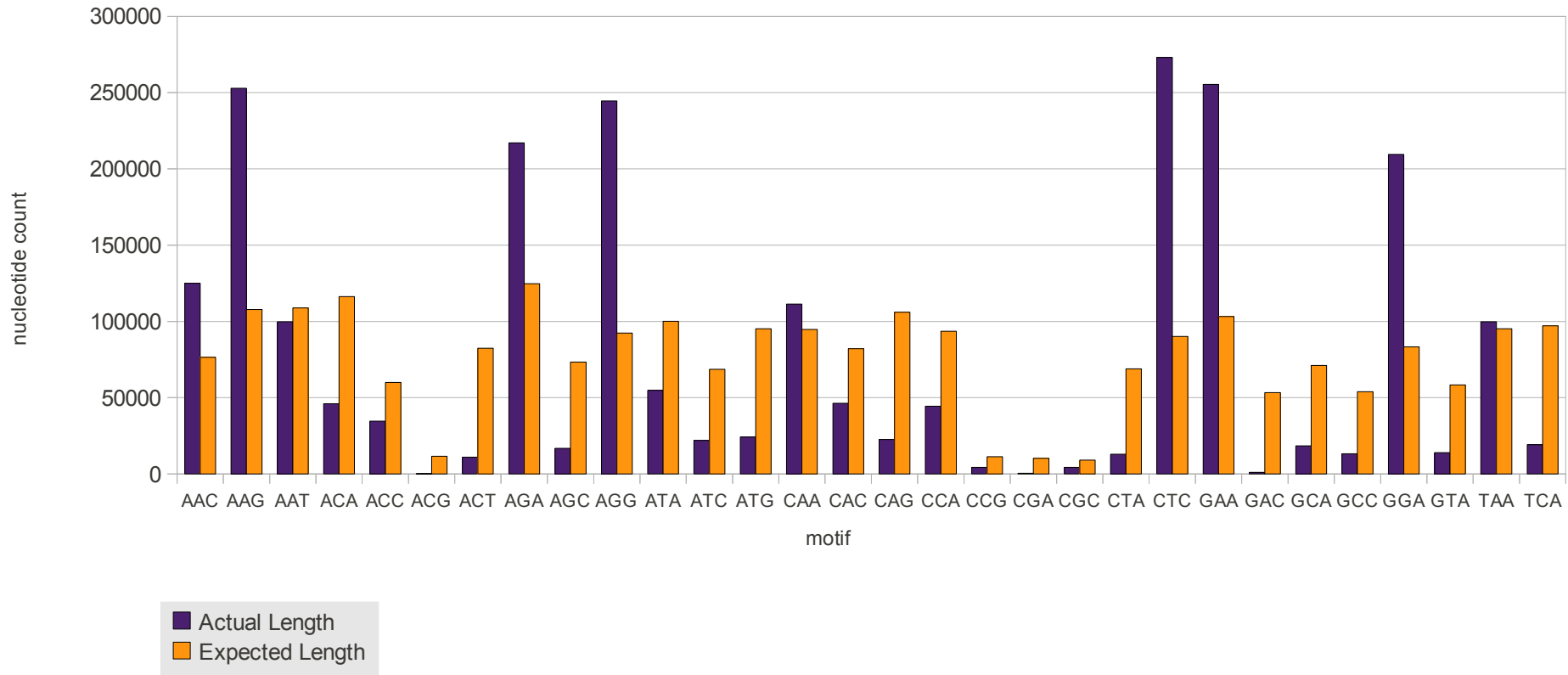


Figure 11: Total Nucleotides in Trinucleotide Microsatellites. Total and expected number of nucleotides found in trinucleotide microsatellites. The expected number of nucleotides has been calculated using the formula $E(M_{xyz}^l) = \frac{x\bar{y}z}{\sum [x\bar{y}z]} \cdot \sum [M_{xyz}^l]$ where $\sum [M_{xyz}^l]$ is the sum of all nucleotides in trinucleotide microsatellites (corrected for their reverse compliments), $x\bar{y}z$ is the corrected number of the specific trinucleotide and $\sum [x\bar{y}z]$ is the sum of all corrected trinucleotides.

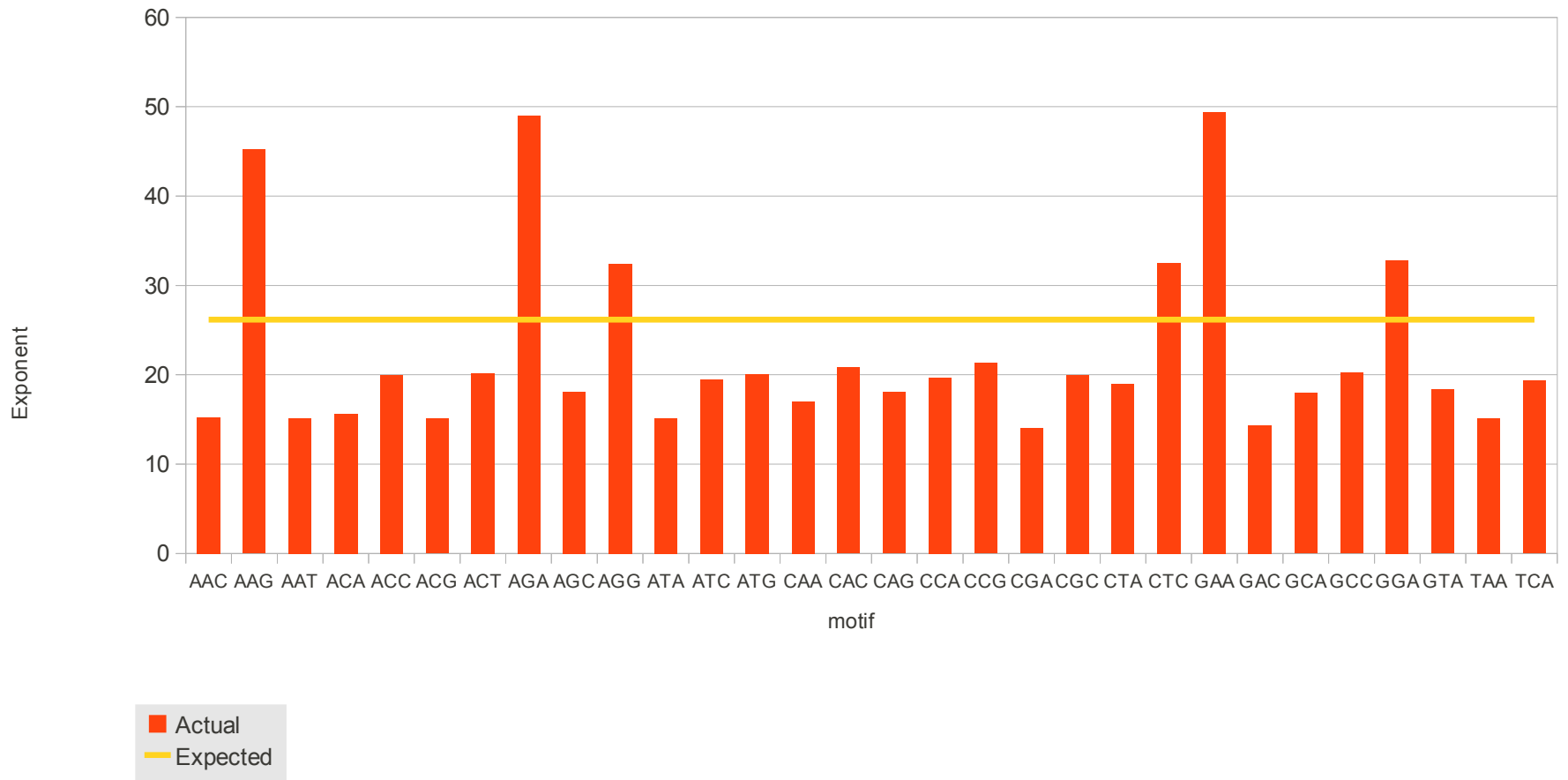


Figure 12: Trinucleotide Average Exponent. Mean trinucleotide exponents and the expected mean. Expected mean has been calculated from the expected total microsatellite length and expected microsatellite quantity.

5.3. Analysis of experimental datasets

The microsatellite content of the experimental datasets (table 1) was found using the method in section 4.2. Fold changes for microsatellite quantities of specific motifs found in the data were calculated using methods 4.3.2 (to calculate expected results) and 4.3.3.

Observer/expected graphs, passing the criteria are shown in Graphs S1, S2 and S3 for di-, tri- and tetra nucleotides. Values passing the +ve and -ve fold change threshold are shown diagrammatically as diagrams 1, 2 and 3, for di-, tri- and tetranucleotides respectively. The diagrams were visually examined to search for patterns within the data that match the groupings as given in Table 1 (using a scoring method is discussed in the conclusion). The PGCs which are expected to show clear sign of reprogramming have been highlighted in the group column.

No.	Dataset	Group	AC	CA	AG	GA
1	J1 1 50ng (2-181 2B)	1				
2	J1 1 400ng (2-172 B)	1				
3	PGC 11.5 1 (3-16(1))	2				
4	P3MEF methyl	10				
5	TKO methyl	5				
6	N95 methyl	3				
7	Sperm methyl	9				
8	J1 2 (3-170 5)	1				
9	Tet methyl	4				
10	PGC 13.5 male AID (3-182/197 – 5B)	(7)				
11	PGC 13.5 female AID (3-182 – 6B)	(7)				
12	PGC 16.5 male 1 (3-182 1B)	6				
13	PGC 16.5 female 1 (3-182 2B)	7				
14	PGC 13.5 male 1 (3-182 3B)	7				
15	PGC 13.5 female 1 (3-182 4B)	7				
16	PGC 11.5 AID (3-182 7B)	(2)				
17	E14-ES-methyl	1				
18	E14-EmBod-methyl	8				
19	J1-ES-p14 MeDIP	1				
20	E14-ES-Rua-MeDIP	1				
21	EB-E14-D13-Rua-MeDIP	8				
22	NP95 -/- ES-MeDIP	3				
23	Tet1/2-KD L2 MeDIP	4				
24	Tet1 KD Single L1 MeDIP	4				
25	PGC 11.5 2 (4-127 1B)	2				
26	PGC 16.5 male 2 (4-127 4B)	6				
27	PGC 13.5 male 2 (4-127 2B)	7				
28	PGC 13.5 female 2 (4-127 3B)	7				
29	PGC 16.5 female 2 (4-127 5B)	7				




-ve fold change: 
+ve fold change: 
No change: 

Diagram 1: Dinucleotide Microsatellite Methylation Fold Changes. Fold changes for each dataset, calculated from the observed / expected quantities of microsatellites, that passed the threshold are shown in green and red (+ve and -ve fold change). Datasets where no overall fold change is discernible in the given motif are shown as grey. The group column highlights datasets where reprogramming of methylation is expected.

The first observation about the three fold change diagrams (diagram 1, 2, 3) is that there is a trend towards similar fold change patterns for motifs which are cyclic permutations, e.g. the four motifs AGAA, AAGA, AAAG and GAAA. The reason for this trend would seem obvious considering they have essentially the same sequence for a given length microsatellite. However, it is intriguing that they should have similar +ve and -ve fold change patterns, if there is nothing interesting going on is there any reason to assume this would happen? These are effectively independent trials and are strongly suggestive of the maintenance of different methylation patterns for different microsatellites.

For matching the experimental data to the possible regulation patterns, the

dinucleotide microsatellites (diagram 1) are uninformative. However, both the trinucleotide and the tetranucleotide microsatellites are more interesting. Using the information from section 5.2.1 to inform the analysis of the methylated data, the motif GGA is of particular interest.

No.	Sample	Group	GGA	CTC	AGG	AAG	AGA	GAA	ACC	CAC	CCA	AAC	ACA	CAA	AGC	CAG	GCA	ACT	CTA	ATC	TCA	GTA	ATG	
1	J1 1 50ng (2-181 2B)	1																						
2	J1 1 400ng (2-172 B)	1																						
3	PGC 11.5 1 (3-16(1))	2																						
4	P3MEF methyl	10																						
5	TKO methyl	5																						
6	N95 methyl	3																						
7	Sperm methyl	9																						
8	J1 2 (3-170 5)	1																						
9	Tet methyl	4																						
10	PGC 13.5 male AID (3-182/197 – 5B)	(7)																						
11	PGC 13.5 female AID (3-182 – 6B)	(7)																						
12	PGC 16.5 male 1 (3-182 1B)	6																						
13	PGC 16.5 female 1 (3-182 2B)	7																						
14	PGC 13.5 male 1 (3-182 3B)	7																						
15	PGC 13.5 female 1 (3-182 4B)	7																						
16	PGC 11.5 AID (3-182 7B)	(2)																						
17	E14-ES-methyl	1																						
18	E14-EmBod-methyl	8																						
19	J1-ES-p14 MeDIP	1																						
20	E14-ES-Rua-MeDIP	1																						
21	EB-E14-D13-Rua-MeDIP	8																						
22	NP95 +/- ES-MeDIP	3																						
23	Tet1/2-KD L2 MeDIP	4																						
24	Tet1 KD Single L1 MeDIP	4																						
25	PGC 11.5 2 (4-127 1B)	2																						
26	PGC 16.5 male 2 (4-127 4B)	6																						
27	PGC 13.5 male 2 (4-127 2B)	7																						
28	PGC 13.5 female 2 (4-127 3B)	7																						
29	PGC 16.5 female 2 (4-127 5B)	7																						

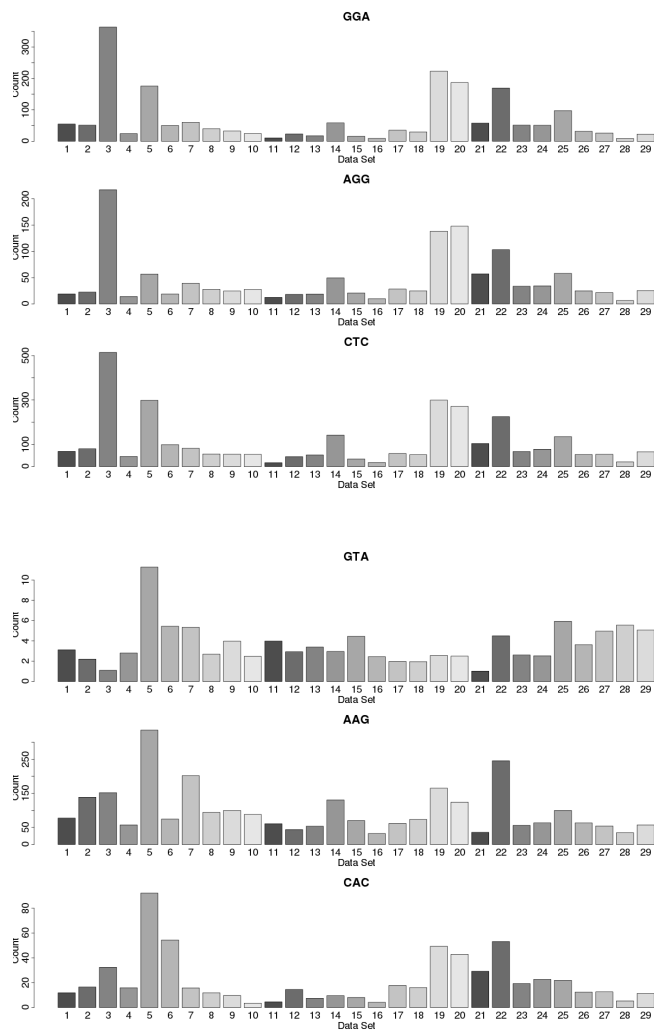
-ve fold change: ■
+ve fold change: ■
No change: ■

Diagram 2: Trinucleotide Microsatellite Methylation Fold Changes. Fold changes for each dataset, calculated from the observed / expected quantities of microsatellites, that passed the threshold are shown in green and red (+ve and -ve fold change). Datasets where no overall fold change is discernible in the given motif are shown as grey. The group column highlights datasets where reprogramming of methylation is expected.

No.	Sample	Group	CTTC	GGAA	AGGA	AAGG	AGAA	AAGA	AAAG	GAAA	AGAT	GATA	TAGA	ATCC	CATC	TCCA	ATGG	CACA	ACAC	ACCT	GGTA	ACAT
1	J1 1 50ng (2-181 2B)	1																				
2	J1 1 400ng (2-172 B)	1																				
3	PGC 11.5 1 (3-16(1))	2																				
4	P3MEF methyl	10																				
5	TKO methyl	5																				
6	N95 methyl	3																				
7	Sperm methyl	9																				
8	J1 2 (3-170 5)	1																				
9	Tet methyl	4																				
10	PGC 13.5 male AID (3-182/197 – 5B)	(7)																				
11	PGC 13.5 female AID (3-182 – 6B)	(7)																				
12	PGC 16.5 male 1 (3-182 1B)	6																				
13	PGC 16.5 female 1 (3-182 2B)	7																				
14	PGC 13.5 male 1 (3-182 3B)	7																				
15	PGC 13.5 female 1 (3-182 4B)	7																				
16	PGC 11.5 AID (3-182 7B)	(2)																				
17	E14-ES-methyl	1																				
18	E14-EmBod-methyl	8																				
19	J1-ES-p14 MeDIP	1																				
20	E14-ES-Rua-MeDIP	1																				
21	EB-E14-D13-Rua-MeDIP	8																				
22	NP95 +/- ES-MeDIP	3																				
23	Tet1/2-KD L2 MeDIP	4																				
24	Tet1 KD Single L1 MeDIP	4																				
25	PGC 11.5 2 (4-127 1B)	2																				
26	PGC 16.5 male 2 (4-127 4B)	6																				
27	PGC 13.5 male 2 (4-127 2B)	7																				
28	PGC 13.5 female 2 (4-127 3B)	7																				
29	PGC 16.5 female 2 (4-127 5B)	7																				

-ve fold change: ■
+ve fold change: ■

Diagram 3: Trinucleotide Microsatellite Methylation Fold Changes. Fold changes for each dataset, calculated from the observed / expected quantities of microsatellites, that passed the threshold are shown in green and red (+ve and -ve fold change). Datasets where no overall fold change is discernible in the given motif are shown as grey. The group column highlights datasets where reprogramming of methylation is expected.



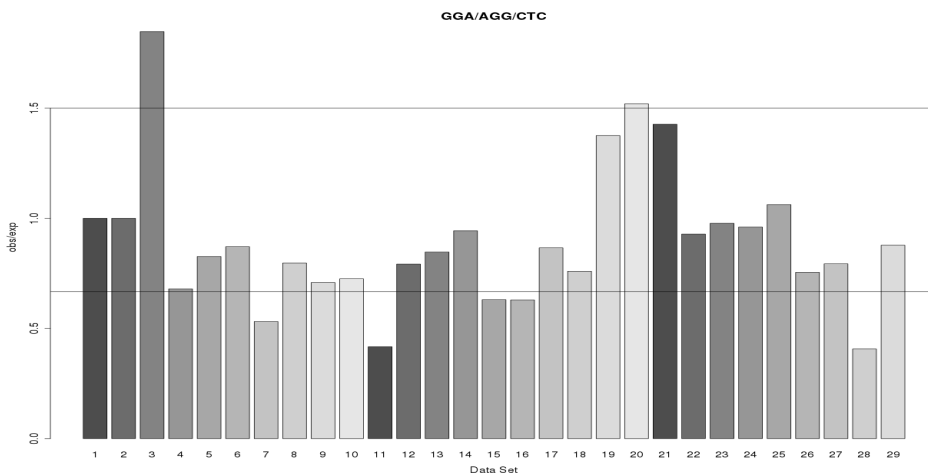
Graph 1: GGA/AGG/CTC Fold Changes. Graphs of the microsatellite quantity found in each dataset for the cyclic permutations of GGA, and three randomly selected motifs.

The diagram for trinucleotide fold change (diagram 2) points to several lines of evidence supporting the regulation of GGA. It matches 7 of the 8 highlighted datasets expected to be informative for regulation during PGC reprogramming. There is a difference between male and female PGCs at E 16.5, which is indicative of a regulatory change. Additionally it was one of the motifs highlighted in section 5.2.1 that was in excess in the genome.

There are several contradictory results, not least because it does not appear to

be that good a match for its cyclic permutations, especially AGG. However graphs of the absolute values (Graph 1) indicate that they are closely matched and, the controls indicate more so to each other than towards other motifs.

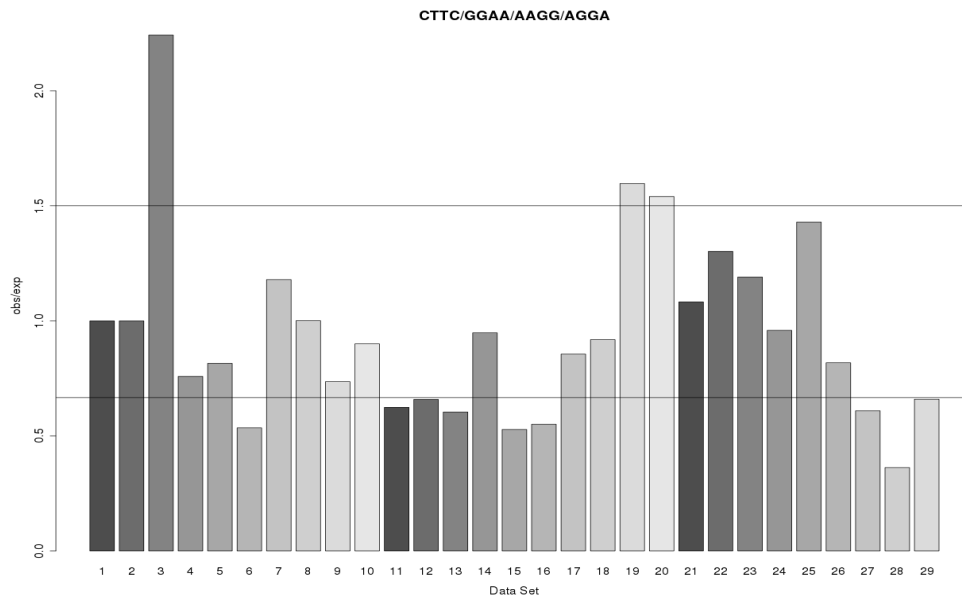
Most damning however, when the results are combined they combine to greatly reduce the evidence for a pattern indicating regulation (Graph 2 and diagram 3).



Graph 2: GGA and Cyclic Permutations Fold Change: Observed / expected microsatellite quantities found in datasets for GGA and its cyclic permutations. The upper bar represent the +ve and the lower the -ve fold change thresholds.

This does not completely rule out these motifs from being regulated, but it does considerably weaken the argument and more contributory evidence is required.

There is the alternative, that specifically GGA microsatellites are being regulated, but without a possible mechanism of action this remains highly speculative.



Graph 3: CTTC and Cyclic Permutations Fold Change. Observed / expected microsatellite quantities found in datasets for CTTC and its cyclic permutations. The upper bar represent the +ve and the lower the -ve fold change thresholds.

In the tetranucleotide data, CTTC and two of its three cyclic permutations show a similar pattern to GGA, with 7 out of 8 matches with the highlighted cells, and thus may also be good candidates for regulation. In this case combining the results from the cyclic permutations does not weaken the patterns (Graph 3 and Diagram 4) with 6 out of 8 still matching. For the two none matching datasets, dataset 12 (PGC E16.5 Male) is very close to the -ve fold-change threshold (value 0.65, threshold 0.66). Additionally E16.5 is near the start of remethylation, it is not unreasonable to assume that some of the microsatellites in question have yet to be remethylated in this dataset. The second contradictory result (Dataset 14, E 13.5 Male) is anomalous and difficult to explain. The fold change patterns for the other datasets are generally in line with expectation, *i.e.* they have the same fold change for datasets in the same group. There are a few exceptions which need to be stated. The PGC E 11.5 datasets (3 and 25) show a different fold change. Given that the latest research points to two distinct phases of demethylation, finishing and starting at about E

11.5 respectively and that the timings are not fully known (51), again it might not be unreasonable suggesting that the difference in relative quantities of microsatellites found in these datasets reflects the ambiguity in the timing of demethylation events. Of the remaining datasets, 4 out of 6 ES cells have a matching fold change pattern, all three Tet datasets are consistent and do not point to any demethylation and the EB datasets are consistent.

No.	Sample	Group	CTTC	GGA
1	J1 1 50ng (2-181 2B)	1		
2	J1 1 400ng (2-172 B)	1		
3	PGC 11.5 1 (3-16(1))	2		
4	P3MEF methyl	10		
5	TKO methyl	5		
6	N95 methyl	3		
7	Sperm methyl	9		
8	J1 2 (3-170 5)	1		
9	Tet methyl	4		
10	PGC 13.5 male AID (3-182/197 – 5B)	(7)		
11	PGC 13.5 female AID (3-182 – 6B)	(7)		
12	PGC 16.5 male 1 (3-182 1B)	6		
13	PGC 16.5 female 1 (3-182 2B)	7		
14	PGC 13.5 male 1 (3-182 3B)	7		
15	PGC 13.5 female 1 (3-182 4B)	7		
16	PGC 11.5 AID (3-182 7B)	(2)		
17	E14-ES-methyl	1		
18	E14-EmBod-methyl	8		
19	J1-ES-p14 MeDIP	1		
20	E14-ES-Rua-MeDIP	1		
21	EB-E14-D13-Rua-MeDIP	8		
22	NP95 -/- ES-MeDIP	3		
23	Tet1/2-KD L2 MeDIP	4		
24	Tet1 KD Single L1 MeDIP	4		
25	PGC 11.5 2 (4-127 1B)	2		
26	PGC 16.5 male 2 (4-127 4B)	6		
27	PGC 13.5 male 2 (4-127 2B)	7		
28	PGC 13.5 female 2 (4-127 3B)	7		
29	PGC 16.5 female 2 (4-127 5B)	7		

-ve fold change: 
+ve fold change: 
No change: 

Diagram 4: CTTC, GGA and their Cyclic Permutations' Methylation Fold Changes. Fold changes calculated from the observed / expected quantities of the combined cyclic permutations for the given microsatellites. Datasets that passed the threshold are shown in green and red (+ve and -ve fold change). Datasets where no overall fold change is discernible in the given motif are shown as grey. The group column highlights datasets where reprogramming of methylation is expected.

Other motifs do not show such clear signs of regulation as CTTC and its cyclic

permutations (or GGA). Some do show consistency across the groups, such as AAAG, AGAT and their cyclic permutations, but as they don't show clear signs of reprogramming it is a case of not ruling them in rather than ruling them out.

5.4. Assessing the Functional Significance

5.4.1 Finding Genes

Lists of genes within the NCBIM37 assembly containing microsatellites with the motifs GGA, AGG, CTTC, AAGG, CAA and AAGA were discovered using Workflow 1 in section 4.4.1 (Table 3). GGA and AGG, CTTC and AAGG, CAA and AAGA were combined to form three gene list given as GGA/AGG, CTTC/AAGA and CAA/AAGA in table 3. There is a slight discrepancy between the numbers of genes in the original list and the number reported by DAVID, due to the genes not being listed within the DAVID database. CAA and AAGA were included for control purposes, to eliminate terms with high concordance with microsatellites in general.

Motif	Quantity	In Genes	DAVID
GGA	4263	1204	1046
AGG	5030	1292	1128
CTTC	5214	1302	1106
AAGG	5655	1439	1289
CAA	4904	1130	1007
AAGA	5797	1386	1180
GGA/AGG			1966
CTTC/AAGG			2173
CAA/AAGA			1986

Table 3: Microsatellites Found in Genes. The total quantity of microsatellite for each motif and the numbers found within genes in the NCBIM37 *Mus musculus* assembly. DAVID lists the number of genes containing microsatellites of the given motif found in the DAVID database.

5.4.2 Determining gene function

The gene-lists GGA/AGG and CTTC/AAGG were then used with method 4.4.2. The gene-list CAA/AAGA and the two random lists were used as the controls for this method.

Gene List	Terms	Unique Terms	Related Genes
GGA/AGG	1123	182	410
CTTC/AAGG	1101	141	391

Table 4: Number of gene list related terms returned by DAVID Functional Annotation Chart. Unique Terms is the number of terms found after filtering for gene size. The Related Genes is the number of genes related to the Unique Terms.

Table 4 lists the total terms, the number of terms that passed the criteria and the total number of genes (found in the gene-list) related to the unique terms (genes listed in tables S8 and S9). The annotation clusters returned using method 4.4.2 are telling, especially for CTTC/AAGG.

Tables 5-8 show a number of the more significant clusters found in the gene list for CTTC/AAGG and GGA/AGG. The enrichment score is the geometric mean (in -log scale) of member's p-values in a corresponding annotation cluster. Lower numbered clusters are typically of more interest, but is dependent on the biology that is being investigated. For reference there were 177 clusters found for CTTC/AAGG and 139 found for GGA/AAG. The P-Value rates how significant the term is, to quote from the DAVID website:

“a modified Fisher Exact P-Value, for gene-enrichment analysis. It ranges from 0 to 1. Fisher Exact P-Value = 0 represents perfect enrichment. Usually P-Value is equal or smaller than 0.05 to be considered strongly enriched in the annotation categories”

Cluster	Enrichment Score	Term	P-Value	Fold Change
Annotation Cluster 9	4.34	GO:0016477~cell migration GO:0051674~localization of cell GO:0048870~cell motility GO:0048762~mesenchymal cell differentiation GO:0060485~mesenchyme development GO:0014031~mesenchymal cell development GO:0014033~neural crest cell differentiation GO:0014032~neural crest cell development GO:0001755~neural crest cell migration GO:0001667~ameboidal cell migration GO:0001764~neuron migration GO:0007507~heart development	7.62E-08 4.22E-07 4.22E-07 4.09E-06 4.87E-06 2.54E-05 1.83E-04 1.83E-04 2.67E-04 1.04E-03 9.69E-03 1.32E-01	3.83 3.37 3.37 7.81 7.66 7.33 8.12 8.12 9.99 5.95 3.83 1.72
Annotation Cluster 12	3.98	GO:0006928~cell motion GO:0030030~cell projection organization GO:0030182~neuron differentiation GO:0048666~neuron development GO:0000904~cell morphogenesis involved in differentiation GO:0000902~cell morphogenesis GO:0032989~cellular component morphogenesis GO:0031175~neuron projection development GO:0048858~cell projection morphogenesis GO:0048667~cell morphogenesis involved in neuron differentiation GO:0007411~axon guidance GO:0007409~axonogenesis GO:0032990~cell part morphogenesis GO:0048812~neuron projection morphogenesis GO:0009953~dorsal/ventral pattern formation GO:0007389~pattern specification process GO:0003002~regionalization	2.11E-09 2.77E-07 6.23E-07 2.44E-06 1.13E-05 2.02E-05 4.84E-05 6.05E-04 8.52E-04 9.71E-04 1.02E-03 1.16E-03 1.36E-03 2.23E-03 6.31E-03 7.84E-03 2.50E-02	3.44 3.24 2.88 3.15 3.43 2.85 2.62 2.81 2.84 2.94 3.91 3.05 2.71 2.83 4.19 2.16 2.15
Annotation Cluster 31	2.84	GO:0048729~tissue morphogenesis GO:0035295~tube development GO:0060429~epithelium development GO:0035239~tube morphogenesis GO:0001763~morphogenesis of a branching structure GO:0007423~sensory organ development GO:0060562~epithelial tube morphogenesis GO:0002009~morphogenesis of an epithelium GO:0048732~gland development GO:0030855~epithelial cell differentiation GO:0030879~mammary gland development GO:0048754~branching morphogenesis of a tube GO:0003006~reproductive developmental process GO:0050678~regulation of epithelial cell proliferation GO:0022612~gland morphogenesis GO:0001657~ureteric bud development GO:0008283~cell proliferation GO:0060603~mammary gland duct morphogenesis GO:0060675~ureteric bud morphogenesis GO:0001658~branching involved in ureteric bud morphogenesis GO:0001655~urogenital system development GO:0060443~mammary gland morphogenesis GO:0001822~kidney development GO:0001656~metanephros development	4.26E-06 5.82E-06 8.74E-06 1.02E-05 2.26E-05 1.41E-04 1.48E-04 1.75E-04 2.05E-04 1.40E-03 1.78E-03 2.92E-03 3.99E-03 5.84E-03 6.24E-03 2.33E-02 2.90E-02 2.93E-02 3.55E-02 3.55E-02 3.73E-02 5.38E-02 6.08E-02 6.41E-02	3.38 3.19 3.11 3.81 4.29 2.83 4.14 3.32 3.11 3.42 4.01 3.70 2.32 4.25 3.65 4.56 2.01 5.89 5.47 5.47 2.36 4.64 2.50 3.30
Annotation Cluster 37	2.42	GO:0048729~tissue morphogenesis GO:0035239~tube morphogenesis GO:0001704~formation of primary germ layer GO:0048598~embryonic morphogenesis GO:0001707~mesoderm formation GO:0048562~embryonic organ morphogenesis GO:0048332~mesoderm morphogenesis GO:0007369~gastrulation GO:0007498~mesoderm development GO:0048568~embryonic organ development GO:0007389~pattern specification process GO:0048589~developmental growth GO:0007178~transmembrane receptor protein serine/threonine kinase signaling pathway GO:0001503~ossification GO:0040007~growth GO:0060348~bone development GO:0007179~transforming growth factor beta receptor signaling pathway GO:0009880~embryonic pattern specification mmu04350:TGF-beta signaling pathway GO:0016202~regulation of striated muscle tissue development GO:0048634~regulation of muscle development GO:0001701~in utero embryonic development GO:0009952~anterior/posterior pattern formation	4.26E-06 1.02E-05 6.82E-05 6.84E-05 3.01E-04 3.03E-04 4.09E-04 6.61E-04 8.17E-04 1.68E-03 7.84E-03 1.55E-02 1.60E-02 2.08E-02 3.03E-02 3.46E-02 3.59E-02 6.22E-02 8.36E-02 9.57E-02 1.01E-01 1.63E-01 2.10E-01	3.38 3.81 7.66 2.56 7.44 3.33 7.05 4.66 5.19 2.54 2.16 3.06 3.44 2.89 2.18 2.59 3.99 4.37 2.55 3.65 3.56 1.58 1.75

Table 5: GGA/AGG Developmental Annotation Clusters. Annotation clusters produced using the DAVID Annotation Clustering tool with a gene list of 391 genes (filtered for size) derived from genes containing GGA or AGG motif microsatellites. P-Values are modified Fisher Exact probabilities and the enrichment score is the geometric mean (in -log scale) of the p-values in the cluster.

Cluster	Enrichment Score	Term	P-Value	Fold Change
Annotation Cluster 11	4.32	GO:0006928~cell motion GO:0043005~neuron projection GO:0030182~neuron differentiation GO:0000902~cell morphogenesis GO:0031175~neuron projection development GO:0040007~growth GO:0030030~cell projection organization GO:0032989~cellular component morphogenesis GO:0042995~cell projection GO:0030424~axon GO:0000904~cell morphogenesis involved in differentiation GO:0048858~cell projection morphogenesis GO:0048666~neuron development GO:0032990~cell part morphogenesis GO:0048812~neuron projection morphogenesis GO:0048667~cell morphogenesis involved in neuron differentiation GO:0007409~axonogenesis GO:0007411~axon guidance GO:0031290~retinal ganglion cell axon guidance	2.02E-09 1.98E-07 5.86E-07 4.69E-06 9.21E-06 2.37E-05 2.47E-05 3.89E-05 4.16E-05 6.99E-05 7.80E-05 1.47E-04 1.67E-04 2.53E-04 3.57E-04 1.56E-03 1.81E-03 5.50E-03 5.50E-02	3.37 3.76 2.82 2.94 3.33 3.39 2.73 2.59 2.23 4.49 3.08 3.06 2.61 2.91 3.10 2.79 2.90 3.34 7.79
Annotation Cluster 12	4.26	GO:0007178~transmembrane receptor protein serine/threonine kinase signaling pathway GO:0004675~transmembrane receptor protein serine/threonine kinase activity GO:0005024~transforming growth factor beta receptor activity GO:0007389~pattern specification process mmu04350:TGF-beta signaling pathway GO:0007369~gastrulation IPR000472:TGF-beta receptor/activin receptor GO:0030509~BMP signaling pathway GO:0001707~mesoderm formation GO:0001701~in utero embryonic development GO:0048332~mesoderm morphogenesis GO:0001704~formation of primary germ layer GO:0007498~mesoderm development	2.59E-09 9.33E-07 9.33E-07 1.07E-06 2.47E-05 3.24E-05 1.22E+00 4.19E-04 2.84E-03 2.94E-03 3.62E-03 4.53E-03 5.44E-03	7.45 18.46 18.46 3.20 4.49 5.40 0.00 9.08 6.06 2.31 5.74 5.45 4.31
Annotation Cluster 14	4.21	GO:0016202~regulation of striated muscle tissue development GO:0048634~regulation of muscle development GO:0060284~regulation of cell development GO:0051153~regulation of striated muscle cell differentiation GO:0051147~regulation of muscle cell differentiation GO:0048641~regulation of skeletal muscle tissue development GO:0048742~regulation of skeletal muscle fiber development compositionally biased region:Poly-Gln	1.15E-08 1.51E-08 3.20E-05 2.72E-04 7.44E-04 1.91E-03 8.81E-03 1.18E-02	10.38 10.14 3.66 9.91 8.07 9.08 9.08 2.93
Annotation Cluster 22	3.58	GO:0016202~regulation of striated muscle tissue development GO:0048634~regulation of muscle development GO:0046620~regulation of organ growth GO:0050678~regulation of epithelial cell proliferation GO:0030323~respiratory tube development GO:0060541~respiratory system development GO:0050679~positive regulation of epithelial cell proliferation GO:0060420~regulation of heart growth GO:0030324~lung development GO:0002053~positive regulation of mesenchymal cell proliferation GO:0010464~regulation of mesenchymal cell proliferation GO:0040008~regulation of growth GO:0060043~regulation of cardiac muscle cell proliferation GO:0055024~regulation of cardiac muscle tissue development GO:0055021~regulation of cardiac muscle growth GO:0048638~regulation of developmental growth GO:0045843~negative regulation of striated muscle development GO:0048635~negative regulation of muscle development GO:0001525~angiogenesis	1.15E-08 1.51E-08 2.56E-05 5.13E-05 6.26E-05 1.54E-04 1.69E-04 2.32E-04 2.34E-04 2.72E-04 3.39E-04 7.22E-04 2.87E-03 2.87E-03 2.87E-03 3.21E-03 2.92E-02 3.51E-02 7.38E-02	10.38 10.14 8.81 5.77 4.18 3.81 8.20 15.14 3.93 9.91 9.48 2.55 13.21 13.21 13.21 5.89 10.90 9.91 2.19

Table 6: CTTC/AAGG Developmental Annotation Clusters (2). Annotation clusters produced using the DAVID Annotation Clustering tool with a gene list of 410 genes (filtered for size) derived from genes containing CTTC or AAGG motif microsatellites. P-Values are modified Fisher Exact probabilities and the enrichment score is the geometric mean (in -log scale) of the p-values in the cluster.

Cluster	Enrichment Score	Term	P-Value	Fold Change		
Annotation Cluster 1	10.31	GO:0007167~enzyme linked receptor protein signaling pathway	1.31E-13	4.66		
		mmu05200:Pathways in cancer	1.17E-11	3.53		
		GO:0008284~positive regulation of cell proliferation	7.78E-08	3.45		
Annotation Cluster 2	8.05	GO:0042981~regulation of apoptosis	3.06E-16	3.61		
		GO:0043067~regulation of programmed cell death	4.45E-16	3.57		
		GO:0010941~regulation of cell death	5.31E-16	3.55		
		GO:0043066~negative regulation of apoptosis	9.86E-06	3.19		
		GO:0043069~negative regulation of programmed cell death	1.34E-05	3.13		
		GO:0060548~negative regulation of cell death	1.42E-05	3.11		
		GO:0043523~regulation of neuron apoptosis	6.33E-03	3.63		
		GO:0043524~negative regulation of neuron apoptosis	4.79E-02	3.63		
Annotation Cluster 3	6.44	GO:0042981~regulation of apoptosis	3.06E-16	3.61		
		GO:0043067~regulation of programmed cell death	4.45E-16	3.57		
		GO:0010941~regulation of cell death	5.31E-16	3.55		
		GO:0043085~positive regulation of catalytic activity	1.98E-13	4.73		
		GO:0051336~regulation of hydrolase activity	2.62E-11	5.00		
		GO:0051345~positive regulation of hydrolase activity	4.99E-10	7.63		
		GO:0043065~positive regulation of apoptosis	3.52E-07	3.52		
		GO:0043068~positive regulation of programmed cell death	4.07E-07	3.49		
		GO:0010942~positive regulation of cell death	4.68E-07	3.46		
		Apoptosis	1.92E-06	3.21		
		GO:0006919~activation of caspase activity	2.28E-06	9.91		
		GO:0012501~programmed cell death	2.47E-06	2.53		
		GO:0008219~cell death	3.92E-06	2.44		
		GO:0043280~positive regulation of caspase activity	4.62E-06	9.08		
		GO:0010952~positive regulation of peptidase activity	4.62E-06	9.08		
		GO:0006915~apoptosis	4.80E-06	2.50		
		GO:0016265~death	6.42E-06	2.38		
		GO:0043281~regulation of caspase activity	7.44E-06	7.27		
		GO:0052548~regulation of endopeptidase activity	7.44E-06	7.27		
		GO:0052547~regulation of peptidase activity	8.82E-06	7.12		
		IPR001315:Caspase Recruitment	1.97E-05	11.93		
		IPR007111:NACHT nucleoside triphosphatase	6.93E-05	9.69		
		SM00114:CARD	1.67E-04	8.09		
		GO:0012502~induction of programmed cell death	2.08E-04	3.26		
		GO:0006917~induction of apoptosis	2.08E-04	3.26		
		domain:CARD	9.47E-03	8.92		
		mmu04621:NOD-like receptor signaling pathway	1.49E-01	2.42		
		GO:0009617~response to bacterium	2.63E-01	1.62		
		Annotation Cluster 6	5.48	GO:0009967~positive regulation of signal transduction	3.22E-10	5.07
				GO:0010647~positive regulation of cell communication	3.89E-10	4.81
GO:0010627~regulation of protein kinase cascade	5.57E-06			3.98		
GO:0010740~positive regulation of protein kinase cascade	2.77E-05			4.95		
GO:0008543~fibroblast growth factor receptor signaling pathway	1.40E-04			8.48		
GO:0043408~regulation of MAPKKK cascade	1.01E-03			3.91		
GO:0043410~positive regulation of MAPKKK cascade	1.71E-03			5.41		

Table 7: CTTC/AAGG Regulatory Annotation Clusters. Annotation clusters produced using the DAVID Annotation Clustering tool with a gene list of 410 genes (filtered for size) derived from genes containing CTTC or AAGG motif microsatellites. P-Values are modified Fisher Exact probabilities and the enrichment score is the geometric mean (in -log scale) of the p-values in the cluster.

Cluster	Enrichment Score	Term	P-Value	Fold Change
Annotation Cluster 9	4.34	GO:0016477~cell migration GO:0051674~localization of cell GO:0048870~cell motility GO:0048762~mesenchymal cell differentiation GO:0060485~mesenchyme development GO:0014031~mesenchymal cell development GO:0014033~neural crest cell differentiation GO:0014032~neural crest cell development GO:0001755~neural crest cell migration GO:0001667~ameboidal cell migration GO:0001764~neuron migration GO:0007507~heart development	7.62E-08 4.22E-07 4.22E-07 4.09E-06 4.87E-06 2.54E-05 1.83E-04 1.83E-04 2.67E-04 1.04E-03 9.69E-03 1.32E-01	3.83 3.37 3.37 7.81 7.66 7.33 8.12 8.12 9.99 5.95 3.83 1.72
Annotation Cluster 12	3.98	GO:0006928~cell motion GO:0030030~cell projection organization GO:0030182~neuron differentiation GO:0048666~neuron development GO:0000904~cell morphogenesis involved in differentiation GO:0000902~cell morphogenesis GO:0032989~cellular component morphogenesis GO:0031175~neuron projection development GO:0048858~cell projection morphogenesis GO:0048667~cell morphogenesis involved in neuron differentiation GO:0007411~axon guidance GO:0007409~axonogenesis GO:0032990~cell part morphogenesis GO:0048812~neuron projection morphogenesis GO:0009953~dorsal/ventral pattern formation GO:0007389~pattern specification process GO:0003002~regionalization	2.11E-09 2.77E-07 6.23E-07 2.44E-06 1.13E-05 2.02E-05 4.84E-05 6.05E-04 8.52E-04 9.71E-04 1.02E-03 1.16E-03 1.36E-03 2.23E-03 6.31E-03 7.84E-03 2.50E-02	3.44 3.24 2.88 3.15 3.43 2.85 2.62 2.81 2.84 2.94 3.91 3.05 2.71 2.83 4.19 2.16 2.15
Annotation Cluster 31	2.84	GO:0048729~tissue morphogenesis GO:0035295~tube development GO:0060429~epithelium development GO:0035239~tube morphogenesis GO:0001763~morphogenesis of a branching structure GO:0007423~sensory organ development GO:0060562~epithelial tube morphogenesis GO:0002009~morphogenesis of an epithelium GO:0048732~gland development GO:0030855~epithelial cell differentiation GO:0030879~mammary gland development GO:0048754~branching morphogenesis of a tube GO:0003006~reproductive developmental process GO:0050678~regulation of epithelial cell proliferation GO:0022612~gland morphogenesis GO:0001657~ureteric bud development GO:0008283~cell proliferation GO:0060603~mammary gland duct morphogenesis GO:0060675~ureteric bud morphogenesis GO:0001658~branching involved in ureteric bud morphogenesis GO:0001655~urogenital system development GO:0060443~mammary gland morphogenesis GO:0001822~kidney development GO:0001656~metanephros development	4.26E-06 5.82E-06 8.74E-06 1.02E-05 2.26E-05 1.41E-04 1.48E-04 1.75E-04 2.05E-04 1.40E-03 1.78E-03 2.92E-03 3.99E-03 5.84E-03 6.24E-03 2.33E-02 2.90E-02 2.93E-02 3.55E-02 3.55E-02 3.73E-02 5.38E-02 6.08E-02 6.41E-02	3.38 3.19 3.11 3.81 4.29 2.83 4.14 3.32 3.11 3.42 4.01 3.70 2.32 4.25 3.65 4.56 2.01 5.89 5.47 5.47 2.36 4.64 2.50 3.30
Annotation Cluster 37	2.42	GO:0048729~tissue morphogenesis GO:0035239~tube morphogenesis GO:0001704~formation of primary germ layer GO:0048598~embryonic morphogenesis GO:0001707~mesoderm formation GO:0048562~embryonic organ morphogenesis GO:0048332~mesoderm morphogenesis GO:0007369~gastrulation GO:0007498~mesoderm development GO:0048568~embryonic organ development GO:0007389~pattern specification process GO:0048589~developmental growth GO:0007178~transmembrane receptor protein serine/threonine kinase signaling pathway GO:0001503~ossification GO:0040007~growth GO:0060348~bone development GO:0007179~transforming growth factor beta receptor signaling pathway GO:0009880~embryonic pattern specification mmu04350:TGF-beta signaling pathway GO:0016202~regulation of striated muscle tissue development GO:0048634~regulation of muscle development GO:0001701~in utero embryonic development GO:0009952~anterior/posterior pattern formation	4.26E-06 1.02E-05 6.82E-05 6.84E-05 3.01E-04 3.03E-04 4.09E-04 6.61E-04 8.17E-04 1.68E-03 7.84E-03 1.55E-02 1.60E-02 2.08E-02 3.03E-02 3.46E-02 3.59E-02 6.22E-02 8.36E-02 9.57E-02 1.01E-01 1.63E-01 2.10E-01	3.38 3.81 7.66 2.56 7.44 3.33 7.05 4.66 5.19 2.54 2.16 3.06 3.44 2.89 2.18 2.59 3.99 4.37 2.55 3.65 3.56 1.58 1.75

Table 8: GGA/AGG Developmental Annotation Clusters. Annotation clusters produced using the DAVID Annotation Clustering tool with a gene list of 391 genes (filtered for size) derived from genes containing GGA or AGG motif microsatellites. P-Values are modified Fisher Exact probabilities and the enrichment score is the geometric mean (in -log scale) of the p-values in the cluster.

The clusters found using the CTTC/AAGG gene list are highly significant for a large number of development processes (tables 5 and 6). For example, embryonic morphogenesis (P-Value 1.15E-09), kidney development (P-Value 3.36E-08), heart development (P-Value 8.21E-14), limb development (P-Value 8.84E-07) and many more examples. In addition there are very high enrichment scores for a number of clusters related to regulation (table 7), in particularly clusters 2 and 3 (enrichment scores 8.05 and 6.44) which are related to the regulation of apoptosis, itself an important developmental process.

The results for GGA/AAG are not nearly as convincing (table 8), but still significant. For example, gastrulation (P-Value 6.61E-04), tissue morphogenesis (P-Value 5.82E-06), formation of primary germ layer (P-Value 6.82E-05), and again many more examples.

With the exception of cluster 14 (table 9), the control gene-list produced as per method 4.4.2 (table s10) did not produce clusters resembling those for CTTC/AAGG or GGA/AAG. This cluster is similar to cluster 12 for CTTC &c. and cluster 11 for GGA &c., suggesting that this cluster at least is due to a bias for microsatellites in larger genes.

Cluster	Enrichment Score	Term	P-Value	Fold Change
Annotation Cluster 14	3.27	neuron differentiation	2.0E-6	3.1E0
		neuron development	7.8E-6	3.4E0
		cell projection organization	7.9E-6	3.2E0
		cell morphogenesis	1.7E-5	3.2E0
		neuron projection development	3.8E-5	3.6E0
		cell projection morphogenesis	6.4E-5	3.7E0
		cell morphogenesis involved in neuron differentiation	8.5E-5	3.8E0
		cellular component morphogenesis	9.9E-5	2.8E0
		cell part morphogenesis	1.1E-4	3.5E0
		neuron projection morphogenesis	2.5E-4	3.6E0
		cell morphogenesis involved in differentiation	3.9E-4	3.3E0
		axon guidance	8.3E-4	4.5E0
		locomotory behavior	1.2E-3	2.9E0
		axonogenesis	1.8E-3	3.3E0
		cell motion	7.9E-3	2.1E0
		neuron migration	5.3E-2	3.5E0
		cell migration	2.2E-1	1.6E0
		cell motility	3.6E-1	1.4E0
		localization of cell	3.6E-1	1.4E0

Table 9: Control Annotation Cluster. Annotation clusters produced using the DAVID Annotation Clustering tool with 400 random genes biased for larger size. P-Values are modified Fisher Exact probabilities and the enrichment score is the geometric mean (in -log scale) of the p-values in the cluster.

The results presented from this analysis are highly suggestive of a role for the microsatellites CTTC and its cyclic permutations and to a lesser extent GGA and its cyclic permutations in regulating aspects of development.

6. Conclusions

The analysis of microsatellites presented here has demonstrated that different motifs confer different properties. It has been shown that the quantity of di- and trinucleotide microsatellites is not evenly distributed in comparison to the proportion of their constituent di- and trinucleotides in the genome. Additionally trinucleotide microsatellites that are found in excess generally have been shown to have higher exponents. Further trends show that trinucleotide microsatellites containing two or three different nucleotides have characteristic proportional quantities. Analysis of methylation patterns on di-, tri and tetranucleotide microsatellites has provided some evidence that at least certain motifs show characteristic methylation patterns through PGC reprogramming and on into ES cells. Finally, genes found to contain the microsatellites with methylation patterns that indicate they undergo reprogramming in PGCs have been shown to be related to known developmental processes.

The results are consistent with demonstrating a link between methylation of microsatellites and mouse development. This is the first step and is still a long way from proving conclusively that these microsatellites have a functional role. There are several areas of interest that come out of this study that warrant further research, and could help to further investigate the functional role of microsatellites.

Extending the bioinformatic analysis as given here to include the larger period motifs is trivial, but time consuming. The algorithms developed for section 5.4

allowed searching for microsatellites within exons, introns or any other genomic feature with a known location, or upstream or downstream of these features, though was not implemented in this study. Both of these extensions would be of benefit in searching for similar trends in other microsatellites and refining the search for a functional role. Additional analysis to determine whether microsatellite location is correlated with gene location (or other genomic structures) would be useful in showing a direct link between genes and microsatellites.

To help understand why microsatellites with certain motifs are found in excess may require experimental analysis. As already mentioned it does seem reasonable to suggest that this is due to physical properties, but this is contrary to the current evidence. The same is true for why trinucleotides with two or three different nucleotides are distinctly different, experimental confirmation is required to see if this is a function of slippage or the result of some other force.

Ultimately the microsatellites suggested here would need to be experimentally verified for function. A first step, perhaps to see if variation in microsatellite exponent for a few of the genes listed here affect gene expression.

6.1.1 Comments and Criticisms

Though the results and conclusions presented here are valid, are interesting and perhaps point to a novel method of control in early mouse development, there are various methods used within this thesis that could be improved.

In no particular order, on reflection it might have been useful to combine all cyclic permutations in sections 5.2 and 5.3 and analyse these results. The method used here has the benefit of keeping the motifs as independent test and

allows checking for consistency across the cyclic permutations. Combining the results would have made it easier to check for patterns in the data, and allowed more (or all) of the tetranucleotide data to be analysed.

The creation of a biased random gene list in section 5.4.1 was based on the assumption that the probability of a microsatellite being present in a larger gene is a linear relationship, this is not necessarily true. Further analysis on microsatellite distribution could be undertaken to confirm (or deny) whether this was a good assumption.

Not all cyclic permutations for CTTC and GGA were used in section 5.4. This was due to a limitation in the DAVID software not being able to handle gene-lists longer than 3000 genes. However, it would have been possible to combine the cyclic permutation gene lists and select a random sample of 3000 genes from each. Both the method used and that suggested are based on the reasonable suggestion that there is no difference between cyclic permutations.

It would have been expedient to produce a scoring method for rating motifs against expected methylation patterns (section 5.3) before examining the results. It proved difficult to produce an unbiased scoring system once the results were known.

Following these additional methods, while unlikely to affect the final conclusions would have provided additional support.

7. References

1. Singer M. & Berg T. *Genes and Genomes*. University Science Books, Mill Valley, California. (1991).
2. Brown, T. A. *Genomes*. BIOS Scientific Publishers, Oxford, UK. (1999).
3. Tautz, D. & Schlötterer, C. Simple Sequences. *Curr Opin. Gen. Dev.* **4**, 832–837 (1994).
4. deWachter, R. The number of repeats expected in random nucleic acid sequences and found in genes. *J. Theor. Biol.* **91**, 71–98 (1981).
5. Epplen, C. et al. On the essence of ‘meaningless’ simple repetitive DNA in eukaryote genomes. In: *DNA Fingerprinting: State of the Science* (eds Pena, S. D. J., Chakraborty, R., Epplen, J. T. & Jeffreys A. J.), pp. 29–45. Birkhäuser Verlag, Basel, Switzerland (1993) .
6. Richards, R. I., Holman, K., Yu, S. & Sutherland G. R. Fragile X syndrome unstable element, p(CCG)_n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. *Hum. Mol. Genet.* **2**, 1429–1435 (1993).
7. Schlötterer, C. & Tautz D. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* **20**, 211–215 (1992).

8. Hentschel, C. C. Homocopolymer sequences in the spacer of a sea urchin histone gene repeat are sensitive to S1 nuclease. *Nature* **295**, 714–716 (1982).
9. Csink, A. K. & Henikoff, S. Something from nothing: the evolution and utility of satellite repeats. *Trends Genet.* **14**, 200–204 (1998).
10. Schlötterer, C. Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**, 365–371 (2000).
11. Schlötterer, C., Wiehe, T. Microsatellites, a Neutral Marker to Infer Selective Sweeps. In: *Microsatellites: Evolution and Applications* (eds Goldstein, D. B., Schlötterer, C.), pp. 238–47. Oxford University Press, Oxford (1999).
12. Bachtrog, D., Weiss, S., Zangerl, B., Brem, G. & Schlötterer, C. Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome. *Mol. Biol. Evol.* **16**, 602–610 (1999).
13. Rockman, M. V. & Wray, G. A. Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.* **19**, 1991–2004 (2002).
14. Fondon, J. W., Garner, H. R. Molecular origins of rapid and continuous morphological evolution. *Proc. Natd. Acad. Sci. USA* **101**, 18058–8063 (2004).
15. Hammock, E. A. & Young, L. J. Microsatellite instability generates

- diversity in brain and sociobehavioral traits. *Science* **308**, 1630–1634 (2005).
16. Molla, M., Delcher, A., Sunyaev, S., Cantor, C. & Kasif, S. Triplet repeat length bias and variation in the human transcriptome. *Proc. Natd. Acad. Sci. USA* **106**, 17095–17100 (2009).
17. Kozlowski, P., de Mezer, M. & Krzyzosiak, W. J. Trinucleotide repeats in human genome and exome. *Nucl. Acids Res* **38**, 1–13 (2010).
18. Metzgar, D., Bytof, J. & Wills, C. Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.* **10**, 72–80 (2000).
19. You-chun, L. *et al.* Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.* **11**, 2453–2465 (2002).
20. Catasti, P., Chen, X., Mariappan, S. V., Bradbury, E. M. & Gupta, G. DNA repeats in the human genome. *Genetica* **106**, 15–36 (1999).
21. Hoffman, E. K., Trusko, S. P., Murphy, M. & George, D. L. An S1 nuclease-sensitive homopurine/homopyrimidine domain in the c-Ki-ras promoter interacts with a nuclearfactor. *Proc. Natd. Acad. Sci. USA* **87**, 2705–2709 (1990).
22. Lafyatis, R., Denhez, F., Williams, T., Sporn, M. & Roberts, A. Sequence specific protein binding to and activation of the TGF-beta3 promoter

- through a repeated TCCC motif. *Nucleic Acids Res.* **19**, 6419–6425 (1991).
23. Stallings, R. L., Ford, A. F., Nelson, D., Torney, D. C., Hildebrand, C. E. & Moyzis, R. K. Evolution and distribution of (GT)_n repetitive sequences in mammalian genomes. *Genomics* **10**, 807–815 (1991).
24. Miret, J. J., Pessoa-Brandao, L. & Lahue, R. S. Orientation-dependent and sequence-specific expansions of CTG/CAG trinucleotide repeats in *Saccharomyces cerevisiae*. *Proc. Natd. Acad. Sci. USA* **95**, 12438–12444 (1998).
25. Hamada, H., Petrino, M. G., Kakunaga, T., Seidman, M. & Stollar, B. D. Characterization of genomic poly (dT-dG) ·poly (dC- dA) sequences: structure, organization, and conformation. *Mol. Cell. Biol.* **4**, 2610–2621 (1984a).
26. Hamada, H., Seidman, M., Howard, B. H. & Gorman, C. M. Enhanced gene expression by the poly (dT-dG) ·poly (dC-dA) sequence. *Mol. Cell. Biol.* **4**, 2622–2630 (1984b).
27. Okladnova, O. et al. A promoter associated polymorphic repeat modulates PAX-6 expression in human brain. *Biochem. Bioph. Res. C.* **248**, 402–405 (1998).
28. Sandberg, G. & Schalling, M. Effect of in vitro promoter methylation and CGG repeat expansion on FMR-1 expression. *Nucleic Acids Res.* **14**,

2883 –2887 (1997).

29. Timchenko, N. A., Welm, A. L., Lu, X. & Timchenko, L. T. CUG repeat binding protein (CUGBP1) interacts with the 5' regions of C/EBPbeta mRNA and regulates translation of C/EBPbeta isoforms. *Nucleic Acids Res.* **27**, 4517– 4525 (1999).

30. Meloni, R., Albanese, V., Ravassard, P., Treilhou, F. & Mallet, J. A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element *in vitro*. *Hum. Mol. Genet.* **7**, 423–428 (1998).

31. Gebhardt, F., Zanker, K. S. & Brandt, B. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J. Biol. Chem.* **274**, 13176–13180 (1999).

32. Gebhardt, F., Burger, H. & Brandt, B. Modulation of EGFR gene transcription by a polymorphic repetitive sequence — a link between genetics and epigenetics. *Int. J. Biolmarker* **15**, 105 –110 (2000).

33. Martin-Farmer, J. & Janssen, G. R. A downstream CA repeat sequence increases translation from leadered and unleadered mRNA in *Escherichia coli*. *Mol. Microbiol.* **31**, 1025–1038 (1999).

34. Shi, X.M. et al. Tandem repeat of C/EBP binding sites mediates PPARgamma2 gene transcription in glucocorticoid-induced adipocyte differentiation. *J. Cell Biochem.* **76**, 518–527 (2000).

35. Gao, Q. & Finkelstein, R. Targeting gene expression to the head: the *Drosophila* orthodenticle gene is a direct target of the Bicoid morphogen. *Development* **125**, 4185–4193 (1998).
36. Ott, R. W. & Hansen, L. K. Repeated sequences from the *Arabidopsis thaliana* genome function as enhancers in transgenic tobacco. *Mol. Gen. Genet.* **252**, 563–571 (1996).
37. Meloni, R. et al. A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro. *Hum. Mol. Genet.* **7**, 423–428 (1998).
38. Ramchandran, R. et al. A (GATA)(7) motif located in the 5' boundary area of the human beta-globin locus control region exhibits silencer activity in erythroid cells. *Am. J. Hematol.* **65**, 14–24. (2000).
39. Nakamura, Y. et al. VNTR (variable number of tandem repeat) sequences as transcriptional, translational, or functional regulators. *J. Hum. Genet.* **43**, 149–152 (1998).
40. Katti, M. V., Sami-Subbu, R., Ranjekar, P. K. & Gupta, V. S. Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci* **9**, 1203-1209 (2000).
41. Li, L. et al. Pseudo-periodic partitions of biological sequences. *Bioinformatics* **20**, 295–306 (2004).

42. Antoniewski, C. et al. Direct repeats bind the EcR/USP receptor and mediate ecdysteroid responses in *Drosophila melanogaster*. *Mol. Cell. Biol.* **16**, 2977–2986 (1996).
43. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. From DNA to Diversity. *Molecular Genetics and the Evolution of Animal Design*. Blackwell Science, Malden, MA (2001).
44. Davidson, H. et al. Genomic sequence analysis of *Fugu rubripes* CFTR and flanking genes in a 60 kb region conserving synteny with 800 kb of human chromosome 7. *Genome Res.* **10**, 1194–1203 (2000).
45. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**, 425–432 (2007).
46. Lei, H. et al. De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development* **122**, 3195–3205 (1996).
47. Okano, M., Bell, D.W., Haber, D.A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for *de novo* methylation and mammalian development. *Cell* **99**, 247–257 (1999).
48. Bestor, T. H. The DNA methyltransferases of mammals. *Hum. Mol. Genet.* **9**, 2395–2402 (2000).
49. Holliday, R. & Pugh, J. E. DNA Modification Mechanisms and Gene

- Activity during Development . *Science* **187**, 226-232 (1975).
50. Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell. Genet.* **14**, 9-25 (1975).
51. Hemberger, M., Dean, W. & Reik, W., Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nat. Rev. Mol. Cell Biol.* **10**, 526-537 (2009).
52. Surani, M. A., Hayashi, K. & Hajkova, P. Genetic and epigenetic regulators of pluripotency. *Cell* **128**, 747-762 (2007).
53. Sasaki, H. & Matsui, Y. Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nat. Rev. Genet.* **9**, 129-140 (2008).
54. Oswald, J. *et al.* Active demethylation of the paternal genome in the mouse zygote. *Curr. Biol.* **10**, 475-478 (2000).
55. Mayer, W., Niveleau, A., Walter, J., Fundele R. & Haaf, T. Embryogenesis: demethylation of the zygotic paternal genome. *Nature* **403**, 501-502 (2000).
56. Wossidlo, M. *et al.* Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. *EMBO J.* **29**, 1877-1888 (2010).

57. Mertineit, C. *et al.* Sex-specific exons control DNA methyltransferase in mammalian germ cells. *Development* **125**, 889-897 (1998).
58. Howell, C. Y. *et al.* Genomic imprinting disrupted by a maternal effect mutation in the Dnmt1 gene. *Cell* **104**, 829-838 (2001).
59. Howlett, S. K. & Reik W. Methylation levels of maternal and paternal genomes during preimplantation development. *Development* **113**, 119-127 (1991).
60. Rougier, N. *et al.* Chromosome methylation patterns during mammalian preimplantation development. *Genes Dev.* **12**, 2108-2113 (1998).
61. Reik, W. & Walter, J. Evolution of imprinting mechanisms: the battle of the sexes begins in the zygote. *Nature Genet.* **27**, 255-256 (2001).
62. Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development, *Science* **293**, 1089-1093 (2001).
63. Monk, M., Boubelik, M. & Lehnert, S. Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. *Development* **99**, 371-382 (1987).
64. Surani, A. Imprinting and the initiation of gene silencing in the germ line. *Cell* **93**, 309-312 (1998).

65. Kafri, T. *et al.* Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. *Genes Dev.* **6**, 705-714 (1992).
66. Brandeis, M. *et al.* The ontogeny of allele-specific methylation associated with imprinted genes in the mouse. *EMBO J.* **12**, 3669-3677 (1993).
67. Tada, T. *et al.* Epigenotype switching of imprintable loci in embryonic germ cells
Dev. Genes Evol. **207**, 551-561 (1998).
68. Popp, C. *et al.* Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463**, 1101-1105 (2010).
69. Feng, S., Jacobsen, S.E. & Reik W. Epigenetic reprogramming in plant and animal development. *Science* **330**, 622-627 (2010).
70. Seki, Y. *et al.* Cellular dynamics associated with the genome-wide epigenetic reprogramming in migrating primordial germ cells in mice. *Development* **134**, 2627-2638 (2007).
71. Grabczyk, E., Kumari D. & Usdin K. Fragile X syndrome and Friedreich's ataxia: two different paradigms for repeat induced transcript insufficiency, *Brain Res. Bul.* **56** (3-4), 367-373 (2001).
72. Burge, C., Campbell, A., & Karlin, S. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natd. Acad. Sci. USA*

- 89**, 1358-1362 (1992).
73. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucl. Aci. Res.* **27**, 573–580 (1999).
74. Huang, D. W., Sherman B. T. & Lempicki R. A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc.* **4**(1), 44-57 (2009).
75. Huang, D. W., Sherman B. T. & Lempicki R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1-13 (2009)
76. Scarano, E., Iaccarino, M., Grippo, P. & Parisi, E. The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos. *Proc. Natl. Acad. Sci. USA* **57** (5), 1394–400 (1967).
77. Kruglyak, S., Durrett, R., Schug, M. D. & Aquadro C. F. Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. *Mol. Biol. Evol.* **17**(8), 1210–1219 (2000).
78. Morgan, H. D., Dean, W., Coker, H. A., Reik, W. & Petersen-Mahrt, S. K. Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. *J. Biol. Chem.* **279**, 52353–52360 (2004).