CHAPTER SIX: COMPLEX RELATIONSHIPS BETWEEN SUBSTITUTIONS AND TRANSCRIPTION

6.1 INTRODUCTION

Transcriptional strand bias has been described in reporter gene assays and more recently, in cancer genome sequences, and is believed to reflect the activity of nucleotide excision repair (NER). NER is a non-specific repair process activated by sensing bulky DNA distortion caused by mutagenic biochemical modifications (Nouspikel, 2009). Across the genome, DNA distortion is sensed by the XPC protein complex, which results in the opening of a denaturation bubble via the TFIIH complex. The damaged strand is incised at both the 5' and 3' ends resulting in an oligonucleotide gap which is filled in by DNA polymerase δ or DNA polymerase ε, and the nick is sealed by a DNA ligase. A particular class of NER exists that is coupled to transcription, called transcription-coupled repair (TCR). DNA lesion sensing is likely to involve stalling of RNA polymerase II (RNAPII) but otherwise repair proceeds in the same way as described for global NER (Figure 1.6 chapter 1). A consequence of transcription-coupled repair is that DNA damage on the transcribed strand is repaired more efficiently than damage on the non-transcribed strand. Thus, fewer mutations accumulate on the transcribed strand.

For example, DNA damage induced by short-wavelength ultraviolet light (UVB 290-320 nm and UVC < 290nm) can cause the formation of covalent modifications between two adjacent pyrimidines on the same DNA strand resulting in cyclobutane pyrimidine dimers (CPDs). These DNA distorting chemical modifications are the ideal substrate for TCR. This has been documented via reporter assays in a variety of model systems, such as mouse models (Vreeswijk et al., 1998) and more comprehensively, in a malignant melanoma cell line COLO-829. Here, a comparison of the dominant C>T/G>A transition mutation in transcribed genomic regions uncovered a strand bias, with fewer on the transcribed strand ($P < 0.0001$). This strand bias was attributed to preferential repair of the ultraviolet-induced pyrimidine dimers that underlie C>T /G>A mutations on the transcribed strand (Pleasance et al., 2010a). It was believed that this was consistent with the past operation of transcription-coupled repair on ultraviolet-light-induced DNA damage in COLO-829. Similar analyses have been extensively documented for the by-products of tobacco-smoke. Adducts from B[a]DPE have been mapped to preferential codons in TP53 reporter assays of human bronchial epithelial cells and these have been shown to coincide with C>A/G>T transversion mutations (Denissenko et al., 1996; Pfeifer, 2000; Pfeifer et al., 2002). A strong transcriptional bias with fewer transversions on the transcribed strand has been attributed to the past activity of TCR (Hainaut and Pfeifer, 2001). Whole genome analysis of a lung cancer cell line, NCI-H209, has reiterated this verdict (Pleasance et al., 2010b).

These studies revealed how detailed analyses of comprehensive catalogues of somatic mutations in cancer could help to uncover clues regarding specific repair processes that have been operative. In this chapter, asymmetries in substitution mutation prevalence are sought, integrating the analysis with transcriptomic data where appropriate. Variation is explored between genes, between transcriptional strands in each gene and along the length of each gene, in order to gain further insight into the mutagenic exposure and repair pathways that have fashioned these twenty-one breast cancer genomes.

6.2 DEFINING STRAND BIAS OF SUBSTITUTIONS IN CANCER GENOMES

Base substitutions that fall within a genomic footprint, corresponding to ~40% of the human genome, can be classified according to transcriptional strand (Figure 6.1). The six mutation-types of substitutions can therefore be further sub-classified according to whether they are on the transcribed or non-transcribed strand.

Figure 6.1

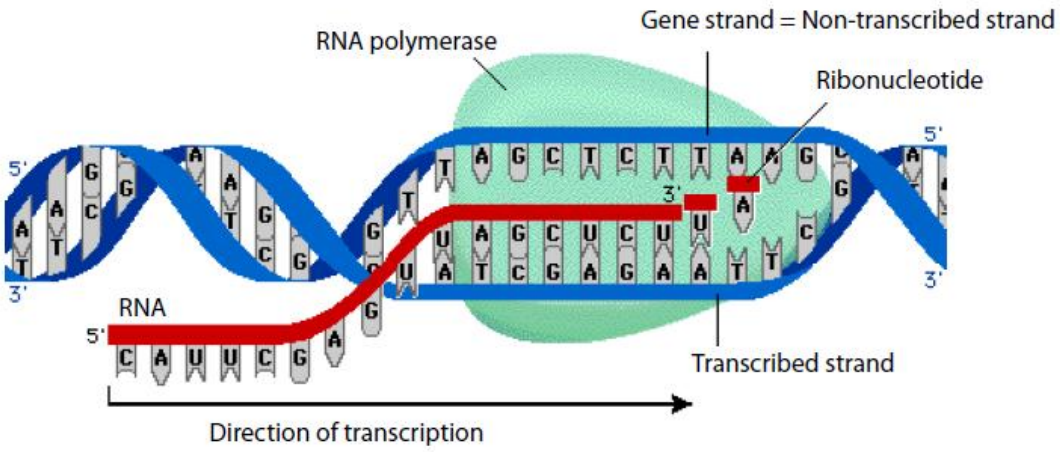

Figure 6.1: Transcriptional strands. The nucleotide sequence of transcribed RNA is identical to the sense/non-template/non-transcribed strand, except that U replaces T, and is complementary to that of the anti-sense/template/transcribed strand.

6.3 GENE EXPRESSION DATA OF THE BREAST CANCER GENOMES

Gene expression data were derived from the Illumina Human HT12 Expression BeadChip array, run in duplicate, with all seventeen samples batched together and normalised. Overall, gene expression data was available for 14,721 genes, with the genomic footprint of these genes encompassing 867,657,063 bases, corresponding to approximately 27.7% of the genome.

Standard hierarchical clustering based on expression array data showed that the seventeen samples for which expression data is available clustered according to histopathological status as would be expected (Figure 6.2).

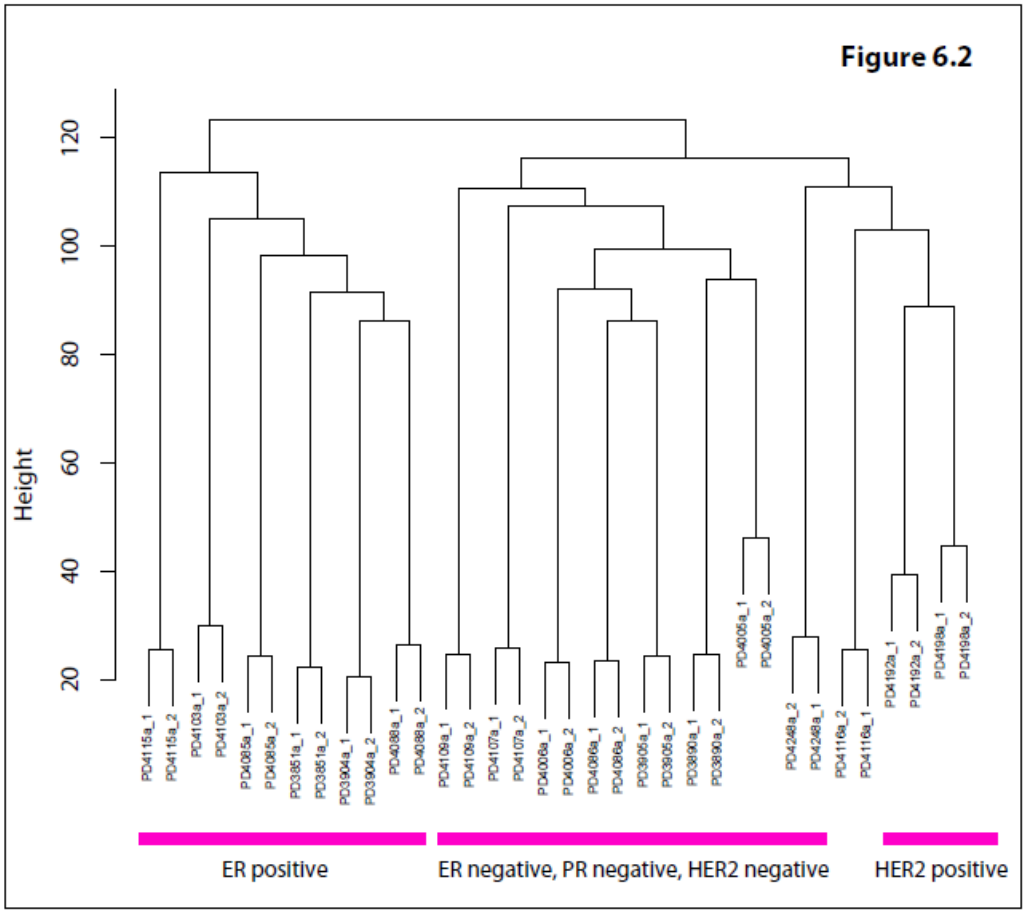

Figure 6.2: Cluster dendrogram of transcriptomic profile of seventeen breast cancer genomes.

In order to examine correlations between mutation prevalence and gene expression as well as to consider transcriptional strand biases, a Poisson regression or generalised linear mixed effects model was used for the analysis (as described in the Experimental Procedures, section 2.6.2). The overall fitted curve for each mutation-type represented the combined effects across all seventeen cases for which there was expression data. The relationships described in the following sections were based on this analysis.

6.3.1 Difference in the prevalence of mutations on the transcribed and non-transcribed strands

The relationship between transcriptional strand and the prevalence of somatic substitutions was examined first. The differences in the prevalence of mutations on the transcribed versus non-transcribed strands (transcriptional strand bias) across all protein coding genes in aggregate were sought. For a given level of gene expression, the effects of transcription-coupled repair (TCR) are revealed by the significant separation of curves for mutations on the transcribed and non-transcribed strands.

A moderate degree of transcriptional strand bias was detectable for C>A/G>T transitions across the 21 breast cancer genomes ($p=1.75 \times 10^{-15}$) and appeared to be present in almost all cases (Figure 6.3). This bias was characterised by fewer C>A mutations on the non-transcribed strand than the transcribed strand.

A strand bias was also observed for T>G/A>C mutations ($p=1.5 \times 10^{-4}$) with fewer T>G mutations on transcribed than non-transcribed strands. No evidence of a transcriptional strand bias was observed for C>G/G>C, C>T/G>A, T>A/A>T or T>C/A>G mutations.

The most widely acknowledged cause of transcriptional strand bias is TCR of nucleotide excision repair (NER) which is believed to remove nucleotides with bulky adducts from the transcribed strands of genes. Assuming that TCR was responsible for the observed strand biases, the presence of fewer C>A mutations on non-transcribed than transcribed strands would suggest that bulky adduct damage to guanine may be the cause of the observed mutations. Similarly, the presence of fewer T>G mutations on transcribed compared to non-transcribed strands would suggest that there may have been bulky adduct damage to thymine. The nature of these ubiquitous mutagenic exposures in breast cancer, and whether they are exogenous or endogenous in origin, is unknown. However, the assumption that TCR is involved is not necessarily correct and it may ultimately transpire that other DNA repair processes, or indeed DNA damage mechanisms, may differentially affect the transcribed and non-transcribed strands of genes.

6.3.2 The relationship between levels of gene expression and the prevalence of somatic mutation

It is known from studies in diverse biological lineages ranging from bacteria (Gouy and Gautier, 1982) to *Caenorhabditis elegans* and *Drosophila* (Duret and Mouchiroud, 1999), that the rate of substitution-related evolutionary change of a gene is often related to its level of expression. Relative substitution rates across genes varied widely but essentially showed a strong negative correlation with the level of gene expression across these biological extractions. Therefore, the relationship between levels of gene expression and the prevalence of somatic mutation was investigated as in a previously studied malignant melanoma and small cell lung cancer respectively (Pleasance et al., 2010a; Pleasance et al., 2010b), the authors demonstrated that levels of gene expression correlated inversely with mutation prevalence. This phenomenon was observed on both the transcribed and non-transcribed strands of genes indicating that it was independent of transcriptional strand. The mechanism underlying this phenomenon is not well-explored.

In the seventeen breast cancers for which gene expression data was available, an inverse correlation of substitution prevalence with gene was observed for C>A/G>T ($p=2.47 \times 10^{-9}$), C>T/G>A ($p=7.5 \times 10^{-3}$), T>A/A>T ($p=1.09 \times 10^{-6}$) and T>C/A>G ($p=1.83 \times 10^{-4}$) mutations for both transcribed and non-transcribed strands (Figure 6.3). This finding reinforces the observation made in a single malignant melanoma and small-cell lung cancer, but extends it to multiple cancer samples of a different tissue-type. No correlation was observed for C>G/G>C or T>G/A>C mutations.

There could be two reasons for this observation. First, for four out of the six classes of mutations, an alternative repair pathway which is related to the degree of expression but that operates on both strands and is at least as numerically important as TCR appears to be at play. Thus, significantly lower mutation prevalence, on both transcribed and non-transcribed strands, was observed in more highly expressed genes. The alternative argument would be that highly expressed genes are under enhanced selective constraint with purifying selection modulating and restricting mutagenic accumulation in highly expressed genes. However, a large proportion of mutations in the genomic footprint are within introns and it remains unclear why purifying selection would be acting on mutations in these regions.

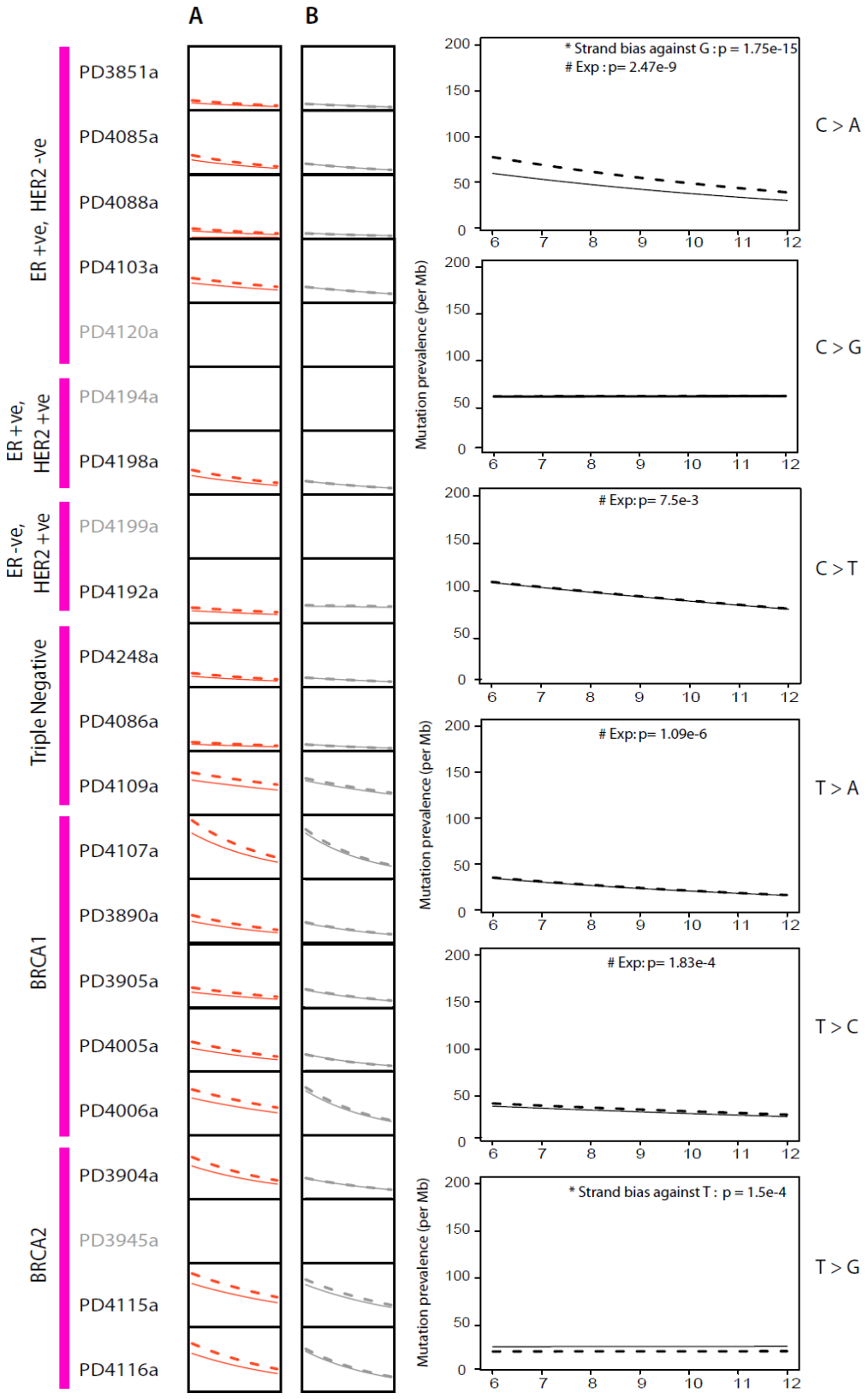

Figure 6.3: The relationship between mutation prevalence and transcription and/or expression. Mutation prevalence is expressed as the number of mutations per Mb from 0 to 2 per Mb on the vertical axis. Log 2 expression levels range from 6 to 12 on the horizontal axis. Lines are fitted curves to the data for A and B. (A) C > A mutations for each cancer genome; (B) T > A mutations for each cancer genome. Breast cancer samples without expression data are shown in gray. (C) Overall effect of transcription and gene expression on mutation prevalence by mutation type. P-values of significance are provided for each mutation-type if a strong effect was seen in either strand bias and/or relationship with expression.

6.4 THE MUTATION PREVALENCE INCREASES AT INCREASING DISTANCE FROM THE TRANSCRIPTION START SITE

Previously, a sharp decline in the mutations at methylated CpG dinucleotides was observed in germline cells in the vicinity of the 5' end of genes (Polak and Arndt, 2008). More recently, an inverse relationship between distance from transcription start site and mutation prevalence was demonstrated in a malignant melanoma cancer genome (Pleasance et al., 2010a). Therefore, the relationship between distance from the transcriptional start site and mutation prevalence in protein coding genes was next examined. Transcription start site coordinates and the genomic footprint of all genes were obtained from the Ensembl v58 API. 1 kb bins beginning from each transcription start site were defined, marching along the length of all genes. The number of genes which completely encompassed each 1kb bin was scored for each bin. The number of mutations present in each bin was also counted. The fraction of genes mutated in each bin is presented in Figure 6.4.

There was evidence of increasing mutation prevalence at increasing distance from the transcription start site (Figure 6.4a), suggesting that the influences of transcription upon mutagenesis described above wane as proximity to the transcription start site decreases. The result confirms the observation previously made on ultraviolet light induced C>T/G>A mutations in a melanoma cell line (Pleasance et al., 2010a), extending it to many more cancer samples of different classes and across many different mutation types.

The effect appears to be particularly pronounced in the first 1kb from the transcription start site (Figure 6.4b). In germline cells, a localised strand asymmetry showing an excess of C>T over G>A substitutions in the non-transcribed strand was confined to the first 1-2 kb downstream of the 5' end of genes (Polak and Arndt, 2008). The authors hypothesised that the exposed non-transcribed strand near the 5' end of genes was more susceptible to cytosine deamination of methylated CpG dinucleotides. To investigate if this was a feature of somatically acquired mutations, strand bias was interrogated in bins of 1kb from the transcription start site and then increasingly larger bins thereafter for all the C>X mutation-types, because the effect was believed to be prominent closer to the transcription start site. Here, all 21 genomes were included in the analysis. Although initially no significant difference between the transcriptional strands were seen, when C>T mutations were

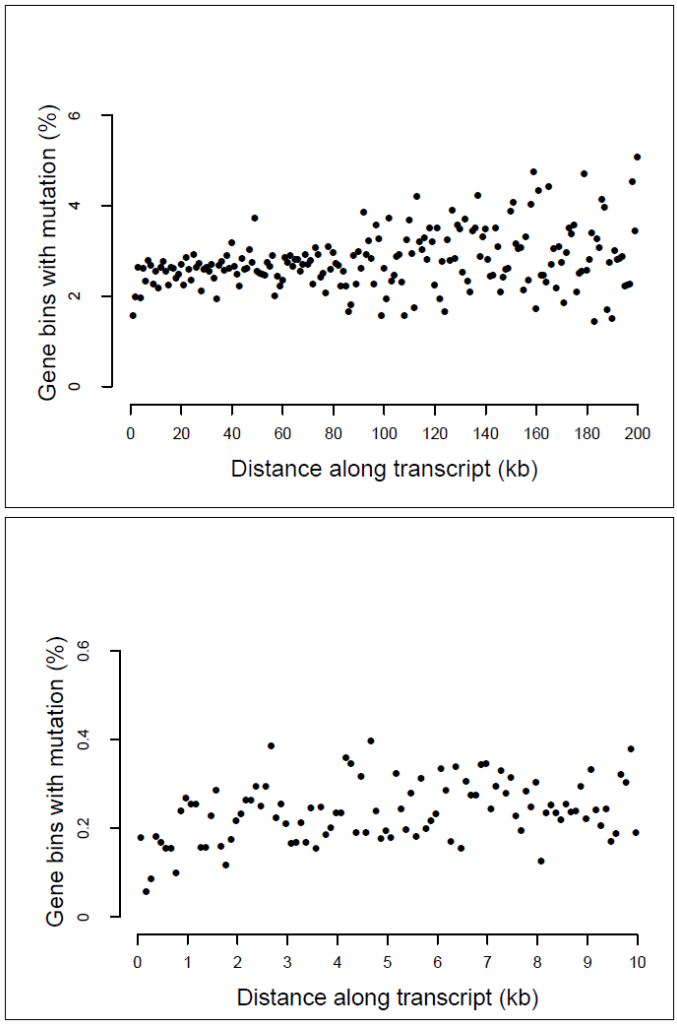

Figure 6.4: The effect of distance from transcription start site on mutation prevalence. (A) Each dot represents a 1kb bin at increasing distances from all transcription start sites (TSS) up to 200kb. The y axis shows the percentage of genes in each bin carrying a somatic mutation. The mutation prevalence increases as distance increases from the TSS. (B) This is particularly marked in the first 1kb after the TSS. Each dot represents a 100bp bin.

separated by whether they occurred at CpG dinucleotides or otherwise, an overall strand asymmetry was seen for the C>T mutations at CpGs, with more mutations seen in the non-transcribed strand. Unlike what was previously documented in the germline, this was not confined to the first 1-2kb (Figure 6.5). When C>Ts at CpGs were treated in isolation, the asymmetry extended to approximately 10kb.

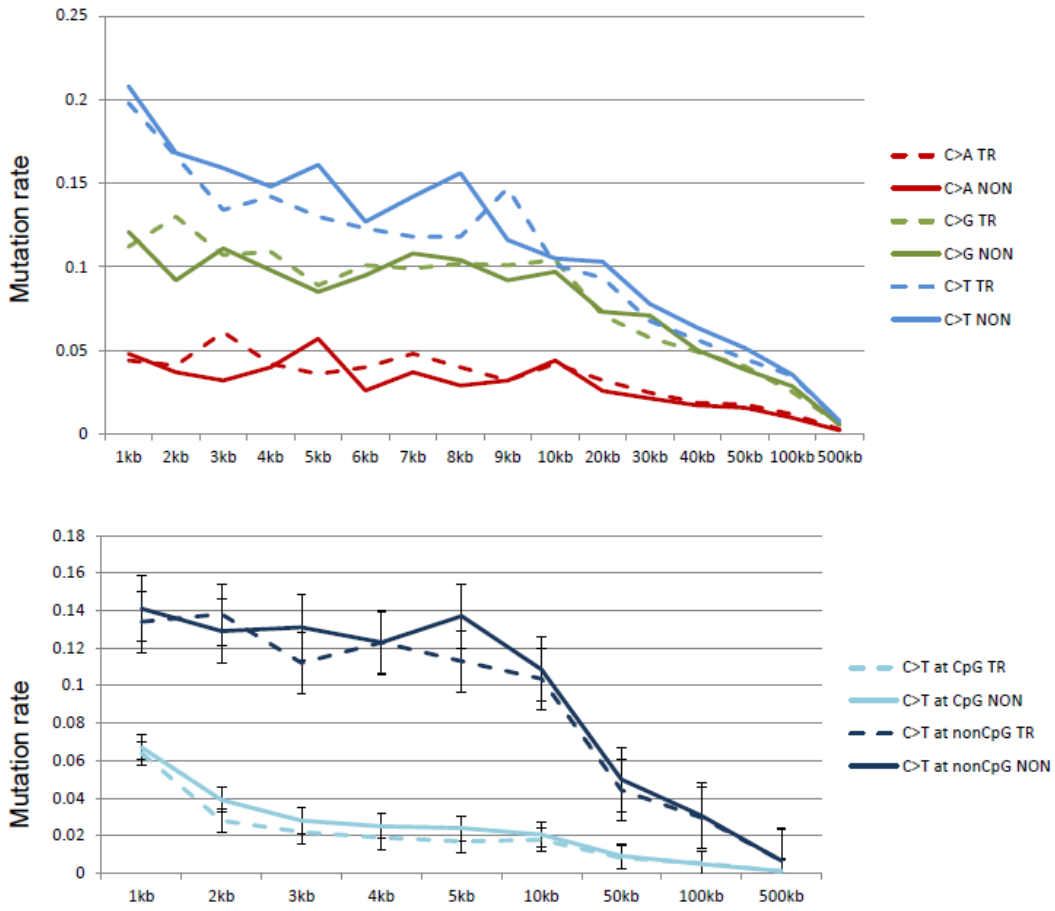

Figure 6.5: Mutation rate with distance from transcription start site, not corrected for the total number of genes involved. Note the strand bias against C>T mutations on the transcribed strand which is not limited to first 1-2kb from the transcription start site, although it is not statistically significant for each genomic bin. The horizontal axis describes the bins of genomic distance from the transcription start site. The vertical axis describes the mutation rate per Mb.

6.5 THE MUTATION PREVALENCE IS HIGHER IN INTRONS THAN IN EXONS

Previously, a reduction in mutation prevalence in exons compared to introns was reported in a single cancer genome and the possibility of the action of negative selection on mutations in the coding sequence was raised, given that preferential targeting of nucleotide excision repair has not been reported towards exons specifically. By comparing the rate of mutation in various parts of the genome, a reduction in mutation prevalence in exons (2.64 per Mb) compared to introns (2.90 per MB) was observed in the twenty-one breast cancers ($p = 1.79X10^{-5}$).

6.6 DISCUSSION

The results from these 21 genomes have yielded further insights into the complex relationship between mutagenesis and transcription in breast cancers. First, transcriptional strand bias was documented with evidence of fewer mutations on the transcribed strand of G>T mutations and for T>G mutations. Second, a marked inverse correlation between gene expression levels and mutation prevalence was seen for C>A/G>T, C>T/G>A, T>A/A>T and T>C/A>G mutations for both transcribed and non-transcribed strands. Thirdly, increasing mutation prevalence was seen with increasing distance from transcription start site. Finally, mutation prevalence was found to be higher in introns than in exons.

6.6.1 Transcriptional strand bias invokes transcription-coupled repair being operative

A transcriptional strand bias was found for C>A/G>T mutations and also for T>G/A>C mutations in most of the cancers but not for the remaining classes of mutations. This suggests that the extent of TCR differs for the various classes of mutation, possibly reflecting differences in the ability of the TCR machinery to recognize and/or repair different adduct lesions.

If TCR is responsible for these strand biases, DNA damage through covalent binding of bulky adducts may be implicated in breast cancer pathogenesis. The exposures generating such covalent modifications could, in principle, be exogenous, and indeed many carcinogens are known to cause adducts on guanine, for example the by-products of tobacco smoke. Alternatively, the exposure could be endogenous in origin, for example due to reactive oxygen species (Hori et al., 2011). Oxidised bases are, however, generally thought to be repaired by base excision repair (BER).

6.6.1.2 Potential atypical substrates for transcription-coupled repair in breast cancers

Here, potential atypical substrates for the action of transcription-coupled repair in breast cancers are speculated. A particularly unique class of oxidative DNA lesion generated by hydroxyl radicals called cyclopurines, are characterised by a covalent bond between the purine and the sugar moiety of the sugar-phosphate backbone making them troublesome for base excision repair (BER) and ideal candidates for nucleotide excision repair (NER) (Bishop and Bell, 1985; Simon et al., 1985). Furthermore, lipid peroxidation has also been known to yield a highly reactive product, malondialdehyde, which can form bulky DNA adducts on guanine (Katzen et al., 1985) again challenging the effectiveness of base excision repair, but posing the perfect substrate for nucleotide excision repair. In fact, malondialdehyde adducts in the transcribed strand of expressed genes were shown to be strong blocks to RNA polymerase II (RNAPII) and are targets for cellular transcription-

coupled repair (TCR) (Shih et al., 1981). Furthermore, viral DNA site-specifically adducted with a malondialdehyde-analogue, exocyclic adduct propanodeoxyguanosine (PdG) and incorporated into NER-deficient and proficient strains demonstrated a 4-fold increase in the frequencies of transversions and transitions in E. coli strains deficient in NER (Johnson et al 1987) (Bishop, 1985).

Oestrogens can cause damage to DNA by generating electrophilic species that can covalently bind to DNA. This is thought to proceed through catechol oestrogen metabolites, which can be oxidised to intermediates that bind to DNA. Therefore, oestrogen could generate DNA lesions particularly to guanines which may become substrates for nucleotide excision repair. First, stable oestrogen adducts which constitute ideal candidates for nucleotide excision repair can be formed through 2,3-quinone oxidative species (Shih et al., 1981). Second, both endogenous and synthetic oestrogens have been shown to induce oxidative DNA damage in addition to specific DNA adducts (Spencer et al., 2011) (Spencer et al., 2012). However, it remains unclear whether such DNA damage would be repaired by nucleotide excision repair or base excision repair. Notwithstanding, there is ample epidemiological evidence linking this endogenously operating and widely exogenously administered mutagen to breast cancer and it should not be overlooked as a potential source for DNA damage.

Commonly occurring depurinating and base deamination events, which are not the usual substrates for transcription-coupled repair, could potentially affect the progression of the transcription complex, thereby providing an opportunity for transcription-related repair. Abasic sites on the transcribed strand were found to block transcription by mammalian RNA polymerase II (RNAPII) in in vitro transcription assays with site-specific lesions whilst not causing any such block when the abasic site was in the non-transcribed strand (Hanawalt and Spivak, 2008). The prevailing dogma is that lesions that block RNAPII will be subject to transcription-coupled repair, and these findings would suggest that an abasic site could be sufficient to initiate transcription-coupled repair (Tornaletti et al., 2006; Wang et al., 2006). This implies that bulky adduct formation is not *always* necessary to stall RNAPII and initiate transcription-coupled repair, at least in vitro.

6.6.1.3 Other forms of transcription-related DNA repair may be operative in breast cancers

Although TCR is the most widely acknowledged transcription-dependent repair pathway, it is possible that a version of base excision repair may exist which is itself coupled to transcription. Alternatively, RNAPII stalling via atypical substrates could initiate other forms of transcription-related repair. If transcription-coupled repair is not involved, the data would suggest that there exist other, currently uncharacterised, forms of transcription-related DNA repair pathways.

6.6.1.4 Strand bias could be due to transcription-related DNA *damage*, not repair

The underlying assumption that mutagenic damage is equivalent in both transcriptional strands and that strand asymmetry arises from partiality of repair to the transcribed strand may of course be incorrect. At the present time, it is hard to identify an example of a mutagen which shows strand bias but the possibility that a mutagen preferentially targets one of two strands cannot be dismissed.

6.6.2 Evidence for an alternative repair pathway associated with levels of gene expression

An inverse relationship between levels of gene expression and mutation prevalence was previously reported in a malignant melanoma and a small cell lung cancer cell line (Pleasance et al., 2010a; Pleasance et al., 2010b). This finding has been extended in this study for some mutation types to include seventeen primary breast cancers. The relationship again seems to be inverse in nature, with more somatic substitutions accumulating in poorly expressed genes. This expression-related phenomenon is independent of transcriptional strand as both strands appear to be similarly affected (Figure 6.3). T>G/A>C mutations exhibited a transcriptional strand bias but no correlation between expression and mutation prevalence. Conversely, C>T/G>A, T>A/A>T, and T>C/A>G mutations showed correlations between gene expression and mutation prevalence but no strand bias (Figure 6.3).

This relationship could be due to less efficient repair in poorly expressed genes. However, the proficient repair of the non-transcribed strand cannot be attributed to transcription-coupled repair which targets the transcribed strand of genes. One possibility is that the genome-wide form of nucleotide excision repair is recruited more effectively to highly transcribed genes, perhaps as a result of differing chromatin configuration.

There have been hints in the past of a type of global nucleotide excision repair called transcription domain-associated repair (DAR) which may account for efficient repair of both strands in expressed

genes. Whilst DAR depends upon transcription it does not depend upon RNA polymerase II stalling due to a lesion. In a series of strand-specific repair studies on HL60 and THP1 cells, repair of both strands continued to occur in parts of the gene that the polymerase never reached and continued despite blocking RNAPII activity with RNAPII inhibitors like α–amanitin (Nouspikel et al., 2006). Furthermore, using siRNA experiments, DAR has been shown to be dependent on XPC, a protein central to global genome repair and not essential for transcription-coupled repair. Conversely, transcription domain associated repair appeared to be independent of transcription coupled repair-specific proteins, CSA and CSB (Barnes et al., 1993). DAR may therefore be a subset of global nucleotide excision repair, perhaps restricted to certain genomic regions by chromatin configuration. It was proposed that genomic domains within which transcription is active are more accessible than the bulk of the genome to the recognition and repair of lesions through the global pathway (Barnes et al., 1993). Since then however, relatively little work has been seen in transcription domain associated repair. The finding here of an inverse relationship between gene expression level and mutation prevalence, acting on both transcribed and non-transcribed strands lends some support to the presence of a repair phenomenon related to expression which is independent of and different to transcription-coupled repair, and could be evidence supporting transcription domain associated repair.

These data also show a trend towards a higher prevalence of somatic substitutions at the 3′ compared to the 5′ ends of genes. This may be due to aborted transcription, such that 3′ ends are overall less transcribed than 5′ ends, with the consequence that expression-related repair processes are deployed less at 3′ ends and hence the mutation prevalence is higher.