CHAPTER 5

SUMMARY AND CONCLUSION
5.1 **Summary of the work done**

Revelations that the DNA is not a passive ingredient for regulation of gene expression by providing a platform for factors to bind, but rather an active member which changes its conformation to alter various processes related to expression regulation, intrigued us to probe into the dynamic nature of DNA. One of such factors, the non–canonical DNA structures, provides a precise mechanism for the DNA to generate plasticity of structure and thus modulate genomic functions.

Our study deals with the G-quadruplex motifs which are composed of stacked G-quartets. Specifically we study the variations within the G-quadruplex to assess their importance in regulation of gene expression and population dynamics. Following points elaborate our major findings:

5.1.1 **Promoter G-quadruplex motifs have low SNP content**

Analysis of SNPs has shown their uneven distribution across the genomic regions. Thus an interesting way would be to concentrate on variants located in important sequence regions such as gene regulatory regions to see their effect on gene regulation. G-quadruplex, one of the well known regulatory motifs, has been reported to be prevalent in promoters. Thus we did our SNP analysis in the putative promoter regions in whole genome. We considered ±1kb of transcription start site (TSS) as the putative promoters.

Promoter sequence of RefSeq genes was downloaded from UCSC (build 36) and the start coordinate of the gene given in UCSC was taken as TSS. Based on a PERL program we fetched all the putative G4 (PG4) motifs within these regions. We used a list of reported SNPs from database for our studies (SNP128) which were validated, in order to rule out any false positives. These were mapped in the promoter and promoter PG4 motifs. We did an initial bioinformatics study within promoters of RefSeq genes to calculate the promoter SNP density. Surprisingly, expected number of SNPs in the quadruplex regions (Quad-SNP) was significantly greater than the actual number of SNPs within Quad-SNP. Comparison of this SNP density with other fragments from the same region, by generating random fragments from same region, revealed similar depletion of SNP density in quadruplex regions. Furthermore, on probing the distribution of the Quad-SNP among the PG4 motifs, we found that only ~4% of PG4 motifs harbour at least one SNP. More than 96% PG4 motifs are totally
devoid of SNPs, thereby indicating resistance of PG4 motif sequences to any change within them. Thus we concluded that SNP density is significantly lower in PG4 motifs.

5.1.2 Quad-SNPs tend to conserve their ancestral allele and stems are more conserved than loops

The depletion of SNP in PG4 motifs indicates important functionality of the motifs. Flexibility in genomic sequence and structure is responsible for inter-individual or inter species-variation. Therefore the otherwise constant sequences of regulatory nature are punctuated by polymorphisms. When the polymorphisms, (SNP in our case), lead to gain of advantage, then positive selection acts thereby stabilising the derived allele, whereas otherwise, balancing selection acts on the loci to conserve the ancestral allele. Since PG4 motifs tend to deter SNP presence and mutations at critical positions in the PG4 motif is known to alter its structure and stability, hence the SNPs, which may alter the structure and stability of the motif, would be expected to be under selection pressure. Based on this concept we questioned whether Quad-SNPs tend to retain their ancestral allele or is flipping to derived allele readily allowed within the motif.

The frequency data for the Quad-SNPs was obtained for four HapMap populations, namely, CEU (Caucasians of European origin), CHB (Chinese), JPT (Japanese) and YRI (Yoruba), consisting of 270 individuals. The chimpanzee genome was used as an extant out-group to distinguish alleles into ancestral (when similar to chimpanzee) or derived (when different from chimpanzee). Among the 1184 validated Quad-SNPs found in this region, allele frequencies from HapMap and ancestral allele information were available for 237 Quad-SNPs. Interestingly we found 195 Quad-SNPs (82.2%) from this list maintained ancestral allele as the major allelic form while in 42 (17.7%) flipping was observed.

From previous studies we know that stems are considered to be more critical in the PG4 structure formation as compared to loops. Thus we hypothesised stem SNPs should be more resistant to higher frequencies of derived alleles in the population. To test this hypothesis we did a population wise analysis of derived allele frequency of all the Quad-SNPs with available allele frequencies in HapMap database. We divided derived allele frequency in low (0 – 0.1), medium (0.1 – 0.5) and high (0.5 – 1) bins. Expectedly we found that in the low derived allele frequency (DAF) bin the proportion of stem SNP was higher than the proportion of loop SNP across all the populations (P = 0.002). Whereas in the higher DAF,
bin the trend was reversed (not significantly), proportion of loop SNPs were higher than stem SNPs. Thus we conclude the PG4 motifs are resistant to change from ancestral allele and conservation is more in stems as compared to loops.

5.1.3 Quad-SNPs play a role in cancer etiology

Role of quadruplex in cancer therapeutics has been a well studied topic. G-quadruplex has been found in the transcriptional regulatory region of many important oncogenes like c-myc, bcl2, VEGFA, k-ras and has been shown to influence gene expression. It has been seen to be prevalent in promoters of oncogenes and depleted in promoters of tumour suppressor genes. So we studied the regulatory effect of Quad-SNPs in the promoter of known oncogenes. The hypothesis was to probe if allelic variation which affected structure was leading to altered regulation in gene activity, thereby making a genotype/allele more common in one phenotype as compared to the other. We included the genes from Cancer Genome Project and search for PG4 motifs in their putative promoter (±1kb) sequences. We then mapped the reported validated SNPs from database (UCSC).

We scanned the genotype of these Quad-SNPs in cancer case and control samples using Illumina golden gate assay. Case here refers to DNA from the cancerous tissue and control refers to the surrounding non-affected tissue. 171 Quad-SNPs within promoter of cancer genes were genotyped, in 121 case-control matched head and neck cancer samples. The genotype calls from Illumina bead studio were analyzed and only SNPs working in 80% samples were considered for further tests. PLINK toolset analysis found seven SNPs in which the minor allele was completely absent from the unaffected population and existed (though in very small frequency) in the affected population.

To re-confirm our results we genotyped the samples containing rare genotypes by SNaPshot. We were able to validate a Quad-SNP on FZD 1 which is a very important oncogene residing in the Wnt/β-catenin pathway. To elucidate its role in gene regulation, we cloned the effective promoter of FZD1 upstream of luciferase gene. Transfection of laryngeal cancer cell line HEP2 showed a significant difference in gene expression between the promoters with the two types of allelic form. To see if this change was restricted to head and neck cancer, or could be seen in other cancers also, we did luciferase assays in HT1080 cell line. Here too we observed a difference of about 24-fold. Consistency of this observation was also checked for MDAMB cell line which showed similar fold change between the two allelic forms.
5.1.4 Promoter Quad-SNP can alter gene expression

Taking into account the differences in gene expression in above experiment we hypothesized, if PG4 motifs are regulatory in nature then alterations in motifs residing in regulatory region would possibly lead to gene expression changes. This hypothesis is supported by previous observations showing mutation in promoter PG4 motifs altering the gene expression.

We analysed the available genotype-gene expression correlation data from HapMap individuals, which highlights the significantly correlated SNPs (Stranger et. al. 2007). We filtered 54 Quad-SNPs included among the correlated SNPs. These Quad-SNPs were found to cause alterations in motif structure with variations in allelic form thereby supporting the assumption that SNPs altering promoter PG4 motifs alter gene expression. Further, to functionally validate this observation, we selected an ubiquitously expressing gene, Ribosomal Protein 3, (RPS3) which plays a critical role in initiation of translation. The functional promoter of this gene harbours a G-quadruplex motif with a validated SNP (G→C). Using CD and melting studies the ‘C’ allele was found to destabilise the PG4 motif. Next we cloned the promoter of RPS3 into basic pGL3, upstream of luciferase. We also created the promoter with other allele by SDM. Promoter activity of ‘C’ allele in the human fibro-sarcoma HT1080 cells showed a conspicuous decrease in excess of 80-fold with respect to the promoter with ‘G’ allele. A decreased fold change of 5-fold in promoter activity on assaying the promoter clone with ‘C’ allele as compared to the ‘G’ allele, re-confirmed our findings in breast cancer cell line MDAMB.

5.1.5 Regulation of Gene expression by PG4 motif

With the previous observations in place quoting polymorphisms present in promoter PG4 motifs alter gene expression we hypothesized that if the motif was of regulatory nature then de novo insertion at a place not having PG4 motif previously, may be able to manipulate the gene expression. Thus we set out to insert a characterized PG4 motif upstream of a constitutively expressing promoter that can provide downstream gene expression quantification. We chose the psiCHECK 2 vector for this purpose. It contains SV40 promoter inserted upstream of the luciferase gene. We inserted a synthetic PG4 motif ‘5'-GGGTGGGTGGGTGGG-3'” upstream of the promoter. Assaying luciferase counts post transfection in A549 cells shows 3.5-fold increase in gene expression. Interestingly if we
The previously debated concept of genetic variations underlying the mechanism of generation of evolutionary variations has recently gained importance for study. This analysis requires data for both genetic and transcriptomic regions. Thus we compared the genomes of human and its ancestor chimpanzee to select region that have been conserved, indicating their functional importance. The promoter conserved quadruplexes containing SNP were selected for the study which hypothesized that the allelic difference between the two species which cause structural change would affect gene expression. We did this for two genes whose expression was known to differ between the species. In these promoters the sequence was variant only at the SNP position; rest of the sequence was completely same. As expected the Quad-SNPs were seen to change structural integrity by CD and melting studies. We cloned the corresponding gene promoters, harboring one allele at a time, in 5′ to luciferase gene. Luciferase assays reported a significant change in gene expression between the two promoters of different alleles. The deviation in magnitude and direction of expression change when compared to gene expression studies could be due to multiple other factors affecting the gene promoters which are species specific.

The figure below (Figure 5.1) summarizes our results and realizes the importance of Quad-SNPs.

5.2 Conclusion and future direction

Current studies focus within the genome for a quest to elucidate the sources of phenotypic variations in humans. This includes inter-individual differences within a population, inter-population differences and variation to disease susceptibility. Also genetic differences in gene expression regulation can lead to evolutionary adaptive differences between species. Though the regulatory roles of G-quadruplexes have been known, our study provides a new perspective to its mechanism of regulation. We elaborate the role of SNPs within the G-
quadruplex motif in its structural alteration, which in turn regulates gene expression. This further enhances the complexity of genome in generating variation in phenotypes at inter-individual level, inter-population level and inter-species level.

Figure 5.1  Figure shows a schematic representation of the various events occurring due to Quad-SNP. It also highlights the type of analysis (bioinformatics, meta-data analysis or experimental) used to infer the results.

With all our current findings in this study we realise that the regulatory potential of the Quad-SNPs has only started unfolding. We have seen that only a fraction of quadruplexes contain SNP. In needs yet to be determined what is the functional relevance of these motifs which contain SNP compared to those that do not contain? Which form of allele give a thermodynamically favourable state? How is the retention of G-quadruplex at one region and its removal from another region determined and regulated? We have shown the changes in gene expression due to Quad-SNP between human and chimps. It would be indeed interesting
to note these changes within the cells of both human and chimp, since the minor allele is retained in lower frequencies within human. Specifically, interaction of the motif with SNP to various chimp and human factors involved in transcription will enhance our understanding.

Thus to conclude, Quad-SNP can be used as an important mechanism of genome regulation by the G-quadruplex motifs.