Published online 4 July 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 18* ***8005–8016***
doi:10.1093/nar/gkr536

Zinc-finger transcription factors are associated with guanine quadruplex motifs in human, chimpanzee, mouse and rat promoters genome-wide

Pankaj Kumar[1], Vinod Kumar Yadav[1], Aradhita Baral[2], Parveen Kumar[1], Dhurjhoti Saha[2] and Shantanu Chowdhury[1,2,*]

[1]G.N.R. Knowledge Centre for Genome Informatics, [2]Proteomics and Structural Biology Unit, Institute of Genomics and Integrative Biology, CSIR, Mall Road, Delhi 110 007, India

Received May 2, 2011; Revised June 6, 2011; Accepted June 12, 2011

ABSTRACT

Function of non-B DNA structures are poorly understood though several bioinformatics studies predict role of the G-quadruplex DNA structure in transcription. Earlier, using transcriptome profiling we found evidence of widespread G-quadruplex-mediated gene regulation. Herein, we asked whether potential G-quadruplex (PG4) motifs associate with transcription factors (TF). This was analyzed using 220 position weight matrices [designated as transcription factor binding sites (TFBS)], representing 187 unique TF, in >75 000 genes in human, chimpanzee, mouse and rat. Results show binding sites of nine TFs, including that of AP-2, SP1, MAZ and VDR, occurred significantly within 100 bases of the PG4 motif ($P < 1.24E-10$). PG4–TFBS combinations were conserved in 'orthologously' related promoters across all four organisms and were associated with >850 genes in each genome. Remarkably, seven of the nine TFs were zinc-finger binding proteins indicating a novel characteristic of PG4 motifs. To test these findings, transcriptome profiles from human cell lines treated with G-quadruplex-specific molecules were used; 66 genes were significantly differentially expressed across both cell-types, which also harbored conserved PG4 motifs along with one/more of the nine TFBS. In addition, genes regulated by PG4–TFBS combinations were found to be co-regulated in human tissues, further emphasizing the regulatory significance of the associations.

INTRODUCTION

The regulation of gene expression in eukaryotes is highly complex and often occurs through the coordinated action of multiple transcription factors (TF). A simplistic model posits specific DNA sequence motifs or *cis*-regulatory elements dictate binding of TF leading to activation or repression of genes. Emerging evidence suggests the possibility that a subset of such *cis*-regulatory elements may adopt distinct conformation(s) that additionally specify TF-DNA interactions. In this context, it is interesting to consider the DNA secondary structures adopted by guanine-rich sequences called G-quadruplexes (or G4 DNA)—a unique self-arrangement of Hoogsten base-paired, intramolecular or intermolecular, association of DNA strands in parallel/antiparallel orientation stabilized by charge coordination with monovalent cations (especially K^+) (1–4).

A large volume of evidence from genome-wide computational studies suggest prevalence of potential G4 (PG4) motifs in promoters of a wide range of species. Initially observed in a genome-wide study comprising 18 bacterial species where PG4 motifs were found to be enriched within regulatory regions (5); this was also found to be the case when >140 bacteria were tested (6). Further studies showed enrichment of PG4 motifs in promoters of human (7,8), chimpanzee (8), mouse (8), rat (8) and chicken (9) genomes; moreover, occurrence of numerous human promoter PG4 motifs were found to be conserved within corresponding mouse and rat promoters (8). In addition, emerging evidence also suggests role of G-quadruplexes in chromatin packaging (10–12), recombination (13) and CpG methylation (14).

In vitro evidence for functional role of the G-quadruplex structure in transcription has been shown for few genes. *c-MYC* was the first case, where a G-quadruplex-forming

*To whom correspondence should be addressed. Tel: +91 11 2766 6157; Fax: +91 11 2766 7471; Email: shantanuc@igib.res.in

sequence in the nuclease hypersensitive element upstream of the P1 promoter was shown to affect *c-MYC* transcription (15). Similarly, transcription was influenced by G-quadruplex-forming sequence motifs within the core promoter of human c-*KIT* (16) and *k-RAS* oncogenes (17). Promoter-G-quadruplex were also reported for a number of other genes such as *VEGF*, *PDGF*, *HIF1α*, *BCL-2*, *RB* and *RET* (18) in addition to *thymidine kinase 1*, where a non-canonical G-quadruplex motif formed from repeats constituting two guanines instead of three was found to be functionally active (19). In line with these studies, transcriptome profiling performed in human cancer cells indicated changes in gene expression in presence of established intracellular G-quadruplex binding ligands suggesting a genome-wide role of G-quadruplex motifs in transcription (20).

Encouraged by these findings, attempts were made to probe involvement of potential *trans* factors in recognition of G-quadruplex motifs. Using chromatin immunoprecipitation (ChIP) assays we recently demonstrated that the non-metastatic factor NM23-H2 binds to the c-*MYC* promoter via a G-quadruplex element (21). In line with this, interactions of recombinant hnRNP A1/Up1 with the *KRAS* promoter G-quadruplex (22), Myc-associated zinc-finger protein (MAZ)/poly(ADP-ribose) polymerase 1 (PARP-1) binding to the G-quadruplex element in the murine *KRAS* promoter (23) and binding of nucleolin/hnRNP proteins to the G-quadruplex forming sequences of the *VEGF* promoter was shown (24). Moreover, G-quadruplex motifs in the promoter of three muscle-specific genes, human sarcomeric mitochondrial creatine kinase, muscle creatine kinase and integrin α-7 of mouse were shown to bind the homodimeric form of the TF MyoD *in vitro* (25). Although these studies suggest G-quadruplex–TF interactions as possible regulatory mechanisms, focus on individual promoters and TF has not tested the fuller scope of such structure specific interactions.

We hypothesized that functionally active quadruplex motifs must associate with one or more TF and reasoned that given the large number of PG4 motifs found near transcription start sites (TSS) the ones that are most likely to be functional, as a first approximation, would be conserved across species. With this in mind using the strategy shown in Figure 1 we sought to find out PG4-transcription factor binding site (TFBS) associations in a genome-wide context in human, chimpanzee, mouse and rat. Findings were tested using genome-wide transcriptome profiling data generated in two cell lines after treatment with a molecule that binds quadruplex motifs inside cells. Further validation was obtained from tissue-specific expression of genes harboring PG4–TFBS combinations.

MATERIALS AND METHODS

Sequence retrieval and analysis

The ±2-kb region centered at annotated TSS of 20 664 human, 20 601 chimpanzee, 19 656 mouse and 15 162 rat non-redundant promoter sequences were retrieved from UCSC build hg18 for human, PanTro2 for chimpanzee, mm9 for mouse and rn4 for rat. PG4 motif forming sequences with stem size three were searched within these promoters with a customized algorithm as described earlier (5). Briefly, we adopted a general pattern G_n-N_{L1}-G_n-N_{L2}-G_n-N_{L3}-G_n, where G is guanine; N is any nucleotide including G; $n = 3–5$, maintaining a constant n within a single motif while the number of nucleotide with loops (L1, L2 and L3) could vary from 1 to 7. The program was rerun with cytosine (C) instead of guanine (G) to identify motifs on the complimentary strand and appropriately corrected for strand orientation. We restricted our program to a stem size of 3 and loop length of 1–7 considering that most *in vitro* characterizations and experiments have used these guidelines for PG4 motifs, though recent work shows that non-canonical motifs are also possible with varying loop and stem sizes (5,19,26).

Analysis of TFBS

Analysis of conservation of PG4 motifs in orthologous promoters of human, chimpanzee, mouse and rat were carried out using algorithms previously published by us (8). Herein, we extended our previous study to include chimpanzee and used NCBI HomoloGene for ortholog information. Using human genes having at least one PG4 motif within ±2 kb of TSS, we searched for the corresponding promoter region in chimpanzee, mouse and rat to retrieve 13 437 human-chimpanzee, 14 940 human–mouse and 13 764 human–rat promoter pairs. For each promoter pair, PG4 motif(s) was searched within 200 bases with respect to the human PG4 motif position in the corresponding chimpanzee, mouse and rat promoter (Figure 2). These promoter-pairs were considered for further analysis and designated as $PG4_{CP\text{-}H}$ (PG4 conserved promoter set human), $PG4_{CP\text{-}C}$ (PG4 conserved promoter set chimpanzee), $PG4_{CP\text{-}M}$ (PG4 conserved promoter set mouse), and $PG4_{CP\text{-}R}$ (PG4 conserved promoter set rat).

We considered 220 PWMs, which represented 187 unique TFs as potential TFBS. $PG4_{CP\text{-}H}$, $PG4_{CP\text{-}C}$, $PG4_{CP\text{-}M}$ and $PG4_{CP\text{-}R}$ were analyzed for presence of these TFBS using MATCH™ (TRANSFAC® professional 12.1) (27). In order to analyze the enrichment of TFBS elements on conserved-set promoters we considered the rest of the promoters (i.e. excluding the conserved set) as a control set. The total occurrence of any given TFBS on each conserved-set promoter was considered as the observed frequency. Similarly, the occurrence of a TFBS in control set promoters, gave the randomly expected frequency. The discrepancy between observed and expected frequency was evaluated by determining the statistically variable chi-square (χ^2), independently for human, chimpanzee, mouse and rat.

PG4–TFBS inter-distance analysis

Using the positions of conserved PG4 motif and TFBS that were found to be significantly enriched on $PG4_{CP\text{-}H}$, $PG4_{CP\text{-}C}$, $PG4_{CP\text{-}M}$ and $PG4_{CP\text{-}R}$ sets, promoter wise n*n combinations of PG4–TFBS were generated and their respective inter-distance (distance between conserved PG4

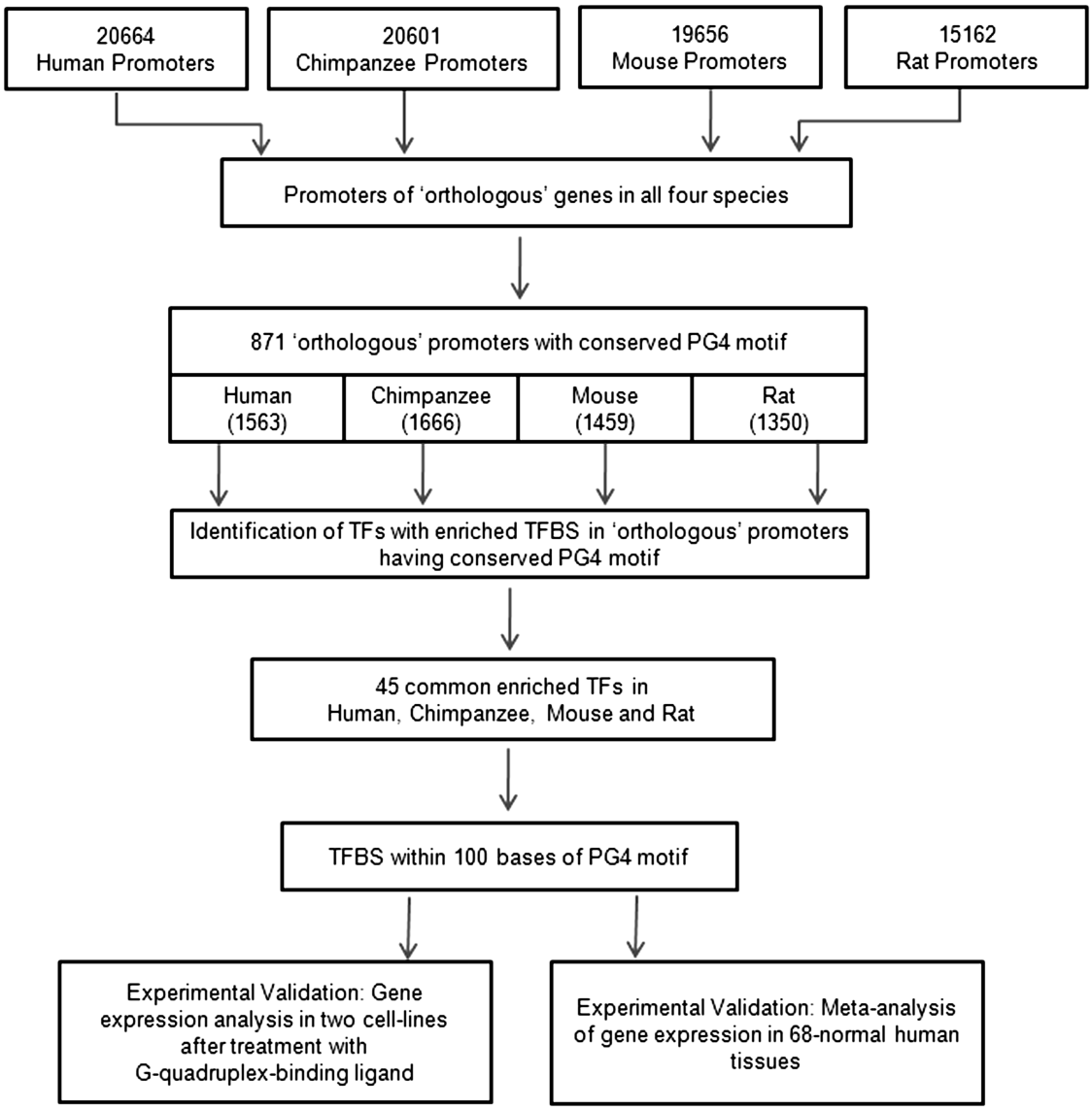

Figure 1. Flowchart summarizing the approach adopted in this study. Schema of the strategy followed to test genome wide association of TF with PG4 motif(s).

motif and TFBS) were calculated. The inter-distance values were then grouped in bins of 100 and their respective percentage frequency within each bin was calculated.

Analysis of PG4 motif co-occurrence with TFBS elements

To analyze the co-occurrence significance of TFBS with PG4 motif individually on $PG4_{CP\text{-}H}$, $PG4_{CP\text{-}C}$, $PG4_{CP\text{-}M}$ and $PG4_{CP\text{-}R}$ sets, we first evaluated the randomly expected co-occurrence frequency of individual TFBS with PG4 motif. The actual promoter-wise co-occurrence of individual TFBS element with PG4 motif was then compared with random expectation of co-occurrence frequency to analyze significance. This is based on a previously published method (28). Briefly, F ($f1$, $f2$) the frequency of co-occurrence of individual TFBS with PG4 motif within m-base pairs (window size) in any n-base pair long sequence is given by

$$F(f1,f2) = \frac{F(f1)F(f2)((2n-m)(m+1)-n)}{n*n}$$

Where F ($f1$) is the promoter-wise expected frequency of PG4 motif, F ($f2$) is the promoter-wise expected frequency of individual TFBS; m is 200 bases and n is 4000 bases (in our case).

The actual co-occurrence frequency of PG4 motif and individual TFBS site within 200 bases in $PG4_{CP\text{-}H}$, $PG4_{CP\text{-}C}$, $PG4_{CP\text{-}M}$ and $PG4_{CP\text{-}R}$ set sequences were obtained by querying the promoter-wise TFBS position and PG4 motif conservation files using in house Perl scripts. In order to calculate the statistical significance of co-occurrence, χ^2-test was performed for individual TFBSs. For example, given '$n-1$' degrees of freedom

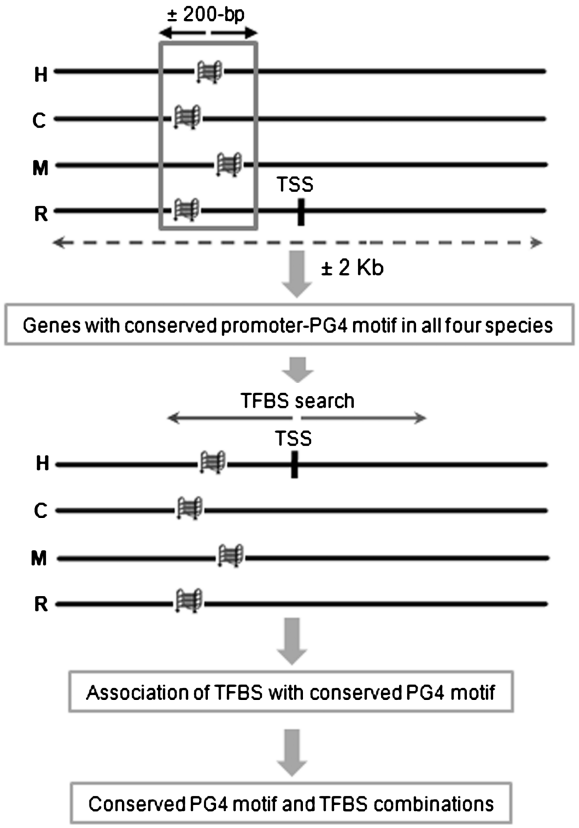

Figure 2. PG4 motif positional conservation across orthologously related promoters. Scheme showing identification of conserved PG4 motif within ±200 bases in orthologously related promoters (±2 kb of TSS) of human, chimpanzee, mouse and rat and search for associated TFBS. H represents a human gene and C, M and R represent their orthologous in chimpanzee, mouse and rat, respectively.

(e.g. $n-1 = 699$ for SP1 in human), to exclude false positives with a simple Bonferroni correction, a reasonable significance level would be $P = 0.005/699 = 7.15 \times 10\text{E-}06$, which corresponds to $\chi^2 = 873.26$.

Analysis of PG4–TFBS enrichment on differentially expressed genes

Genome-wide expression data for HeLa S3 and A549 cells after treatment with TMPyP4, previously published from our laboratory (20), were used for this analysis. The 1161 differentially expressed genes (863 up and 298 down at ≤20% FDR) were compared with human conserved-set genes ($PG4_{CP\text{-}H}$). The ±2-kb sequence (centered at TSS) of genes found to be common with TMPyP4-treated differentially expressed genes were analyzed for enrichment of nine TF (earlier shown to be enriched within 100 bases of conserved PG4 motif). The total occurrence of individual TFBS on each of these genes were considered as observed frequency. The expected frequencies for these nine TFBS were found as described earlier for TFBS enrichment analysis. χ^2-test was performed for individual TF to get statistical significance.

Comparison with experimentally determined ChIP-seq/ ChIP-on-chip TFBS

The *in-silico* binding positions predicted using TRANSFC for human conserved-set promoter were compared with experimentally determined and publically available ChIP-on-Chip (ChIP followed by microarrays) data for SP1 (29), NF-Y (29) and ChIP-seq (ChIP followed by parallel sequencing) data for STAT1 (30), at the time of this study. The sequences of ChIP-on-chip and ChIP-seq binding coordinate intervals for common promoters were fetched from UCSC (hg18) and searched for the SP1 and NF-Y binding positions using TRANSFAC consensus motifs (PWM). For STAT1 we used binding consensus motif proposed by authors (30). The observed TFBS positions were mapped with respect to TSS and PG4–TFBS inter-distance values were calculated, which were finally grouped in bins of 100 to calculate respective frequencies. To calculate the level of similarity for SP1, NF-Y and STAT1 between TRANSFAC-predicted and ChIP-on-chip/ChIP-seq binding sites and their respective inter-distance from conserved PG4 motifs we first plotted the frequency distribution of the TFBSs with respect to PG4 motif (5′ base) position (Supplementary Figure S1). We also calculated the correlation coefficient of the two data sets for statistical significance.

Co-expression and significance analysis

Tissue-specificity of genes harboring PG4 motifs and a TFBS within a 100 base window on the conserved set promoters was checked in 68 human tissues (31). Analysis was largely based on a previously described method (21). The expression data of each gene across all tissues was first normalized to be mean 0 and variance 1 before ranking them as per their normalized expression level in each tissue, hence generating 68 tissue-specific ranked gene lists. We generated two distinct sets of genes namely set A and set B for TFBS that significantly co-occurred with PG4. Set A corresponds to genes with TFBS within 100 bases of conserved PG4 motif. Set A includes 785, 320, 456, 420, 372, 369, 384, 296 and 253 genes for Kid3, KROX, AP-2, SP1, ETF, MAZ, VDR, ZF5 and WT1, respectively. Set B corresponds to genes where respective TFBSs were found beyond ±100 bases of the conserved PG4 motif. Set B includes 86, 228, 332, 280, 324, 344, 385, 296 and 262 genes for Kid3, KROX, AP-2, SP1, ETF, MAZ, VDR, ZF5 and WT1, respectively.

Enrichment of expression of a given gene set $\boldsymbol{S}$ in a particular tissue and its significance was analyzed from the whole ranked list of genes $\boldsymbol{T}$ for the tissue after evaluating the non randomness of ranks of $\boldsymbol{S}$ within $\boldsymbol{T}$, using the Mann–Whitney rank sum statics. After summing the ranks of $\boldsymbol{S}$ in list $\boldsymbol{T}$, we tested the significance of this rank sum against the rank sum of control set (10 random sets of same cardinality from all genes in $\boldsymbol{T}$, excluding $\boldsymbol{S}$). If μ and $\sigma 2$ are the mean and variance of the control set, then enrichment (z-score) of $\boldsymbol{S}$ is given by $(\mu - S)/\ \sigma 2$, which measures enrichment in terms of number of standard deviations away from the mean of the control sets. A z-score of ≥4.0 was considered to be significant in the present study.

RESULTS

More than 40 TFBS–PG4 motif associations are conserved across human, chimpanzee, mouse and rat promoters genome-wide

We found 5005 human–chimpanzee, 4929 human–mouse and 2263 human–rat promoter pairs with PG4 motifs. Out of these 871 promoters harbored at least one conserved PG4 motif (that is present in all the four organisms) and in total 1563, 1666, 1459 and 1350 PG4 motifs in human, chimpanzee, mouse and rat, respectively (Table 1). We reasoned that the 871 promoters harboring one/more conserved PG4 motifs had the maximum likelihood of being functionally relevant in the context of PG4 motif-mediated transcription. KEGG pathway analysis was performed using web-based tool GeneCodis (32) to check for potential importance of genes harboring conserved PG4 motif(s). Significant over-representation ($P < 6.8E\text{-}05$; after correction for multiple hypothesis testing) was found in MAPK signaling, regulation of actin cytoskeleton, focal adhesion, TGF-β signaling, Wnt signaling and apoptosis (Supplementary Table S1).

Next, we asked which TFBS were predominant within the promoters harboring conserved PG4 motifs. In order to statistically analyze the TFBS enrichment we considered 871 $PG4_{CP\text{-}H}$, $PG4_{CP\text{-}C}$, $PG4_{CP\text{-}M}$ and $PG4_{CP\text{-}R}$ along with control sets of 19 793 human, 19 730 chimpanzee, 18 785 mouse and 14 292 rat promoter sequences, where the control sequences were devoid of any conserved PG4 motif (Figure 3). This revealed 120 622, 112 351, 112 855 and 108 669 binding sites for 184, 184, 180 and 181 different TFs on the $PG4_{CP\text{-}H}$, $PG4_{CP\text{-}C}$, $PG4_{CP\text{-}M}$ and $PG4_{CP\text{-}R}$ sets, respectively. Considering a significance level of $P < 0.005$, we obtained target sites for 63 TFs in human, 60 TFs in chimpanzee, 63 TFs in mouse and 60 TFs in rat. Out of these 45 TFs were found to be common to all four species (Supplementary Table S2) indicating that many TF target sites were significantly enriched in association with PG4 motifs.

Target sites of seven zinc-finger TF significantly co-occur with PG4 motifs

Next we checked whether association of PG4 motifs with TFBS had any particular distribution with respect to their relative positioning within a promoter. Inter-distance between all conserved PG4 motifs and TFBS of each of the 45 TFs was mapped within the 871 conserved-set promoters independently for human, chimpanzee, mouse and rat and represented as percentage frequency (fraction of all associations per TFBS) for each PG4–TFBS combination in a window of 100 bases (Figure 4). Interestingly, we noted that for any particular PG4–TFBS combination, the inter-distance distribution was largely distinct, and moreover, the respective distributions were very similar in all the four species. Interestingly, many TFBS either overlapped or were within ±100 bases of the conserved PG4 motif. Considering the potentially important implication of this, we analyzed statistical significance of the co-occurrence for PG4–TFBS pairs which were within an inter-distance of 100 bases using a previously published method (28). This gave target sites of 21, 16, 12 and 11 TF–PG4 combinations, in human, chimpanzee, mouse and rat, respectively. Of these, TFBS for nine factors were found to be common within all the four species (Table 2). We noted with interest that seven out of the nine factors [SP1, MAZ, WT1, KROX (EGR-2), Kid3 (ZNF354C), ZF5 (ZFP161) and VDR] whose target sites were found within 100 bases of the PG4 motif had the zinc-finger motif, particularly the cysteine2–histidine2 (C2H2) domain (Table 3). This was also true for many of the TFs that co-occur within PG4-harboring promoters (Supplementary Table S3). Consistent with this finding one earlier study found that a large number of upstream PG4 motifs are enriched with target sites of SP1 (33). Zinc-finger factors, particularly the Cys2-His2 type represent a significant number among all TF. Though, keeping this in mind, rigorous methods for statistical corrections were devised (Figure 3 and see 'Materials and Methods' section), we further pondered on the likelihood of associations that could be artifacts merely because of high numbers. Out of 187 unique TFs studied here, 33 (0.18) were Cys2-His2 zinc-fingers. We found six Cys2-His2 type zinc fingers out of nine to be associated with PG4 motifs, constituting a fraction of 0.66 ($P < 0.001$; two-tail fisher exact test) suggesting an enrichment that is more than expected by chance. On the other hand, by a similar analogy other TFs with high numbers would be expected to have more association with PG4 motifs. This was not the case; 21 out 187 TFs were leucine zipper factors, however none of these were found to be associated with PG4 motifs in our analysis.

Experimentally determined gene expression reveals role of PG4-zinc-finger associations

To test the physiological significance of the above findings, we resorted to gene expression analysis of human cells

Table 1. Distribution of PG4 motifs near TSS

	ORFs studied	Total no. of PG4 motif in promoters[a]	Promoters with at least one PG4 motif	Conserved PG4 motifs in 871 orthologously[b] related promoters
Human	20 664	50 939	14 836	1563
Chimpanzee	20 601	41 811	14 184	1666
Mouse	19 656	33 738	13 738	1459
Rat	15 163	20 148	9470	1350

[a] ±2 kb centered at TSS.
[b] Human, chimpanzee, mouse and rat.

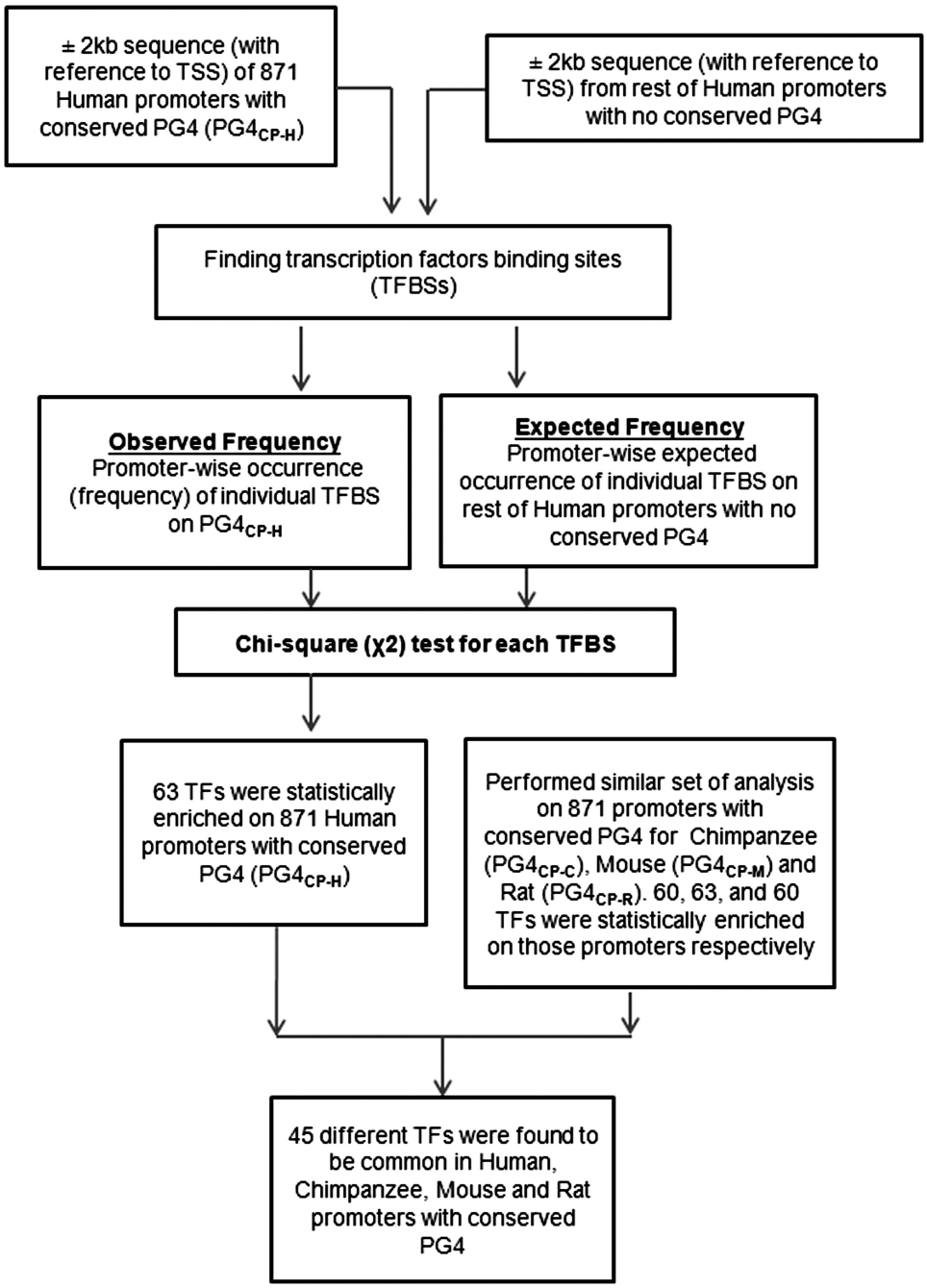

Figure 3. Strategy followed for genome wide comparative analysis to identify enriched presence of TFBS within promoters harboring conserved PG4 motifs in human, chimpanzee, mouse and rat.

treated with the cationic porphyrin ligand (TMPyP4) that binds selectively to G-quadruplex motifs inside cells (15). In a previous study we demonstrated that the effect of TMPyP4 and other ligands that selectively bind to G-quadruplexes show similar genome-wide expression changes largely consistent with presence of the quadruplex motif in promoters (20), though there could be other secondary mechanisms that influence the transcriptome. The gene expression datasets in lung adenocarcinoma (A549) and cervical carcinoma (HeLaS3) cells were analyzed to determine whether genes having PG4–TFBS associations show significant change in expression.

We found 66 genes harboring conserved PG4 motifs within ± 2 kb of TSS that gave significant differential expression (FDR cutoff $\leq 20\%$) consistently across four replicates in both cell lines, after treatment with TMPyP4 (Figure 5). Next we asked if target sites for any of the nine TFs, including the seven zinc-finger factors, were present along with the conserved PG4 motifs in the 66 genes. Interestingly, we found significantly

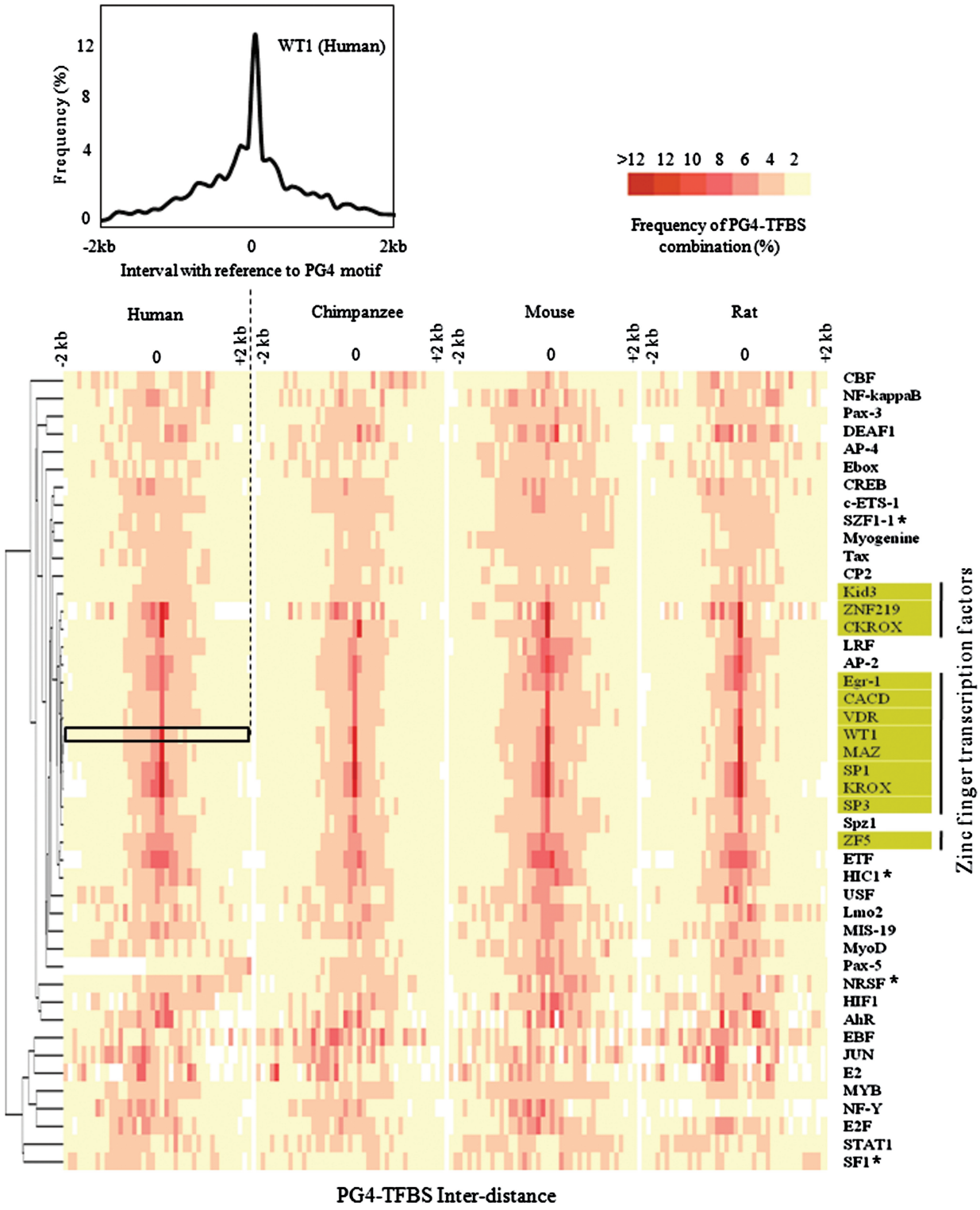

Figure 4. Zinc finger TFBS are closely associated with PG4 motifs. Distribution of TFBS with respect to PG4 motif (PG4–TFBS inter-distance) in promoters of human, chimpanzee, mouse and rat that harbor conserved PG4 motifs. Pseudocolor represents percentage frequency of PG4–TFBS inter-distance values in bins of 100 bases relative to nearest PG4 motif. Asterisk represents additional zinc finger TF found in the study.

enriched occurrence in each of the 66 genes relative to the randomly expected chance of occurrence of each target site ($P < 1.02E\text{-}05$); number of differentially expressed genes that harbor significant PG4–TFBS combinations are given in Table 4. Figure 5 shows the expression arrays representing all the 66 differentially expressed genes after replicate treatments with TMPyP4 along with the corresponding promoters where the relative positions

of the TFBS and conserved PG4 motif are shown. Furthermore, we noted that in each of the 66 genes, target sites of one or more of the nine TFs were present within ±100 bases of the conserved PG4 motif. Together this suggests wide spread functional role of PG4 motifs in gene expression, however, it may be noted that TMPyP4 selectivity towards G-quadruplex DNA vis-à-vis duplex DNA is modest. Therefore, though the gene expression results reported earlier (20) were additionally validated using more selective G-quadrulex binding ligands like the carbazole derivative, BMVC (20), and also a second cationic porphyrin (TpPy) (20), these findings will require to be tested further for individual genes.

PG4–TFBS combinations from *in vivo* genome-wide ChIP-seq and ChIP-on-chip data

The PG4–TFBS associations were found by us using TRANSFAC motifs (PWMs) that are built based on both functional and predicted target sites and constitute all possible genomic occurrences. Therefore, we attempted to validate our predictions using genome-wide experimentally determined TFBS for three TF SP1, NF-Y and STAT1 out of 45 TFBS enriched within conserved-set promoters. In case of SP1 we found 402 combinations on 63 promoters where a conserved PG4 motif and the experimentally determined SP1 site occurred within ±2 kb of TSS. Interestingly, as shown in Supplementary Figure S1a, the frequency of SP1 distribution with respect to the PG4 motif in the case of ChIP-chip was very similar to the ones observed with TFBS determined from TRANSFAC (6520 PG4-SP1 combinations on 700 promoters; $r = 0.93$; $P < 0.0001$), though functional SP1 sites were only a fraction of those found in TRANSFAC. A similar analysis using the TF NF-Y gave 113 conserved PG4-NF-Y combinations in 30 ChIP-chip identified promoters. Distribution of the occurrence of NF-Y sites (with respect to PG4 motifs) when compared to 479 PG4-NF-Y combinations on 118 promoters from TRANSFAC analysis indicated significant similarity in the two distributions ($r = 0.76$; $P < 0.0001$; Supplementary Figure S1b). In case of STAT1 we observed 325 PG4-STAT1 combinations on 82 promoters reported by ChIP-seq experiments. Comparing with 987 PG4-STAT1 combinations on 210 promoters by TRANSFAC once again gave a frequency distribution which was very similar ($r = 0.86$; $P < 0.0001$; Supplementary Figure S1c). These results further indicated that the PG4–TFBS co-occurrences observed using TRANSFAC data are likely to be true in functional cases, though the functional set in most cases is expected to be limited for a variety of reasons, including chromatin compaction (that limits presentation of all available TFBS) and co-factor requirements for TF binding (which is expected to be context dependent).

Table 2. Significance of PG4–TFBS co-occurrence within ±100 bp in promoters with conserved PG4 motifs in human, chimpanzee, mouse and rat

	TF name	Human *P*-value	Chimpanzee *P*-value	Mouse *P*-value	Rat *P*-value
1	SP1	<E-300[a]	2.43E-143	1.20E-76	1.36E-41
2	WT1	<E-300[a]	5.31E-41	7.70E-11	1.24E-10
3	KROX	<E-300[a]	2.20E-31	1.35E-48	1.56E-20
4	MAZ	<E-300[a]	1.59E-105	3.56E-57	2.35E-18
5	VDR	1.24E-262	1.40E-269	1.46E-19	1.44E-11
6	Kid3	<E-300[a]	<E-300[a]	<E-300[a]	<E-300[a]
7	ZF5	<E-300	<E-300[a]	1.48E-190	8.27E-139
8	ETF	<E-300[a]	6.33E-268	1.03E-123	4.06E-81
9	AP-2	<E-300[a]	<E-300[a]	4.04E-69	2.84E-74

[a]Indicates value <E-310.

Table 3. Functional annotation of TFBS significantly co-occurring with conserved PG4 motifs (within 100 bases)

TF Name	Classification of TF	Involvement in Biological processes/ Pathways[a]	Key regulated genes by TF
SP1	Zinc-coordinating DNA binding domains, C2H2 zinc-finger domain, Ubiquitous factors	Cell cycle; MAPK signaling; TGF-β signaling	*CDK1, CDK2, CDK4, CCND2, IL-10, c-MET*
VDR	Zinc-coordinating DNA binding domains, Cys4 zinc finger of nuclear receptor type, Thyroid hormone receptor-like factors	Cell-cycle progression, proliferation and growth, Osteoblastic differentiation	*TCTP, p73, BRCA1,*
KROX	Zinc-coordinating DNA binding domains, C2H2 zinc-finger domain, cell-cycle regulators	Cell cycle, apoptosis	*BNIP3L, BAK, EFNA1, SFN,*
WT1	Zinc-coordinating DNA binding domains, C2H2 zinc-finger domain, cell-cycle regulators, GLI-like	Cell cycle, MAPK signaling, apoptosis	*CCNA1, p21, BCL-2*
MAZ	Zinc-coordinating DNA binding domains, C2H2 zinc-finger domain	Cell cycle, apoptosis, lymphocyte development, neural differentiation	*c-MYC, PPARgamma1, BCL-2, RAG-2, DCC*
Kid3	Zinc-coordinating DNA binding domains, C2H2 zinc-finger domain, Krueppel-like	Kidney and brain development	*HP1α, MOD1, MOD2*
ZF5	Zinc-coordinating DNA binding domains, C2H2 zinc-finger domain, Krueppel-like	Cell cycle, cell proliferation, induction of programmed cell death	*c-MYC, TK1*
AP-2	Basic Domains, bHSH	Cell cycle TGF-β signaling, MAPK signaling	*ESDN, EREG, CXCL2, CDKN1A, COX-2*
ETF	Helix-turn-helix, TEA domain	Cell cycle	*P53*

[a]Relevant references showing involvement of particular TF in biological processes/pathways are given in Supplementary Table S2.

Table 4. TFBS enriched on promoters harboring conserved PG4 motif that are differentially expressed in A549 and HeLaS3 cells after treatment with G-quadruplex binding ligand

TFBS associated with conserved PG4 motif (within 100 bases)	Differentially expressed genes with PG4–TFBS in promoters	*P*-value of PG4–TFBS enrichment in promoters
AP2	34	2.05E-40
ETF	28	4.61E-53
Kid3	58	1.64E-09
KROX	19	3.29E-13
MAZ	24	5.55E-13
SP1	27	1.88E-25
VDR	18	1.02E-05
WT1	20	5.31E-11
ZF5	41	2.68E-44

Genes having PG4-zinc-finger associations are co-expressed

In order to further test the regulatory significance of the associations, we analyzed the transcriptome profile of normal human tissues for genes with TFBS–PG4 pairs in promoters. This was based on the reasoning that regulatory control by any TFBS in association with the PG4 motif for a group of genes is likely to result in significantly enriched (or altered) expression response (either up or downregulation) within specific tissues relative to other randomly picked genes. Two groups of genes were analyzed for each of the nine PG4–TFBS associations found above; genes harboring PG4–TFBS associations either within ± 100 bases (set A) or beyond ± 100 bases (set B) of conserved PG4 motif (see 'Materials and Methods' section). Using gene expression data from 68 normal human tissues we observed significantly enriched expression-response (z-score >4.0; see 'Materials and Methods' section for details of statistical analysis) for set A in all cases in most tissues (Supplementary Figure S2). This was also true in many cases for the genes in set B. Interestingly, the TF Kid3 (ZNF354C), KROX (EGR2), SP1 and AP-2 showed largely distinct expression in set A relative to set B, indicating the likelihood that close proximity of the TFBS with PG4 may be functionally relevant. On the other hand, in case of MAZ, ETF, ZF5 (ZFP161), WT1 and VDR z-scores appeared similar in set A and set B underscoring the possibility that occurrence of the PG4 along with the TFBS within the promoter was important for gene expression in addition to proximal positioning of PG4–TFBS.

DISCUSSION

We found target sites of 45 TF out of 187 analyzed are enriched in promoters harboring PG4 motifs. The functional importance of this is implied by the fact that binding sites of all the 45 TF and PG4 motif occurrences were maintained across four organisms in orthologously related promoters. Remarkably, target sites of nine TFs, including seven zinc-finger factors, were found to be predominantly occurring within 100 bases of a PG4 motif; again, we noted, this was found across the four vertebrate lineages. These observations were confirmed by analyzing transcriptome data generated using a ligand that binds to G-quadruplex motifs inside cells. More than 60 genes, which significantly changed expression on ligand treatment in two cell lines of different origin harbored closely associated PG4 motif and zinc-finger target sites. Finally, genes with PG4–TFBS associations in promoters showed significant co-regulation in transcriptome profiles of 68 human tissues, implicating functional relevance of the G-quadruplex–TFBS associations. Taken together, these findings give strong indications of a genome-wide regulatory role for PG4–TFBS associations and suggest the importance of close proximity in specific cases, implicating a broader role of transcriptional regulation by G-quadruplex elements.

Recently Cogoi *et al.* (23) reported interaction between the Myc associated zinc-finger protein MAZ and the G-quadruplex motif present in the promoter of murine *KRAS* resulting in activation of *KRAS* expression. Using pull down and ChIP assays, they demonstrated *in vivo* binding of MAZ to the quadruplex-forming element in the murine *KRAS* promoter (23). In this context, the approach by Isalan *et al.* (34) who used a phage display-based technique to search for protein factors that could bind to the telomeric quadruplex motif is notable, which identified an engineered Cys2-His2 zinc-finger protein that was both sequence and structure-specific for the telomeric quadruplex motif (35). These independent studies involving particular cases of quadruplex-zinc-finger interactions support the findings reported here from an unbiased genome scale study.

TF PG4 motif associations as putative regulators of cell cycle-related genes

Interestingly, all the nine TFs with target sites significantly co-occurring with conserved PG4 motifs on the promoters of human, chimpanzee, mouse and rat, have been implicated in progression through cell cycle. For example, Tapias *et al.* have shown that the TF SP1 regulates cell-cycle progression through *CDK4* and *CDKN1A*/*p21* interaction (36). In addition, we noted several instances where a cell-cycle gene could be potentially regulated by presence of PG4–TFBS combinations in promoters (see Supplementary Data for details). Based on these it is tempting to speculate that PG4 motifs in association with TF may influence cell cycle related cellular function. Though, this appears to be in line with observations made in a recent study [*vide infra* (37)], further work will be required to directly test this possibility.

G-quadruplex ligand interactions affect non-telomeric functions

Reduced cell proliferation, particularly in tumors, has been reported using various G-quadruplex binding ligands like the cationic porphyrin TMPyP4 [tetra(*N*-methyl-4-pyridyl)-porphyrin chloride], papaverine-derived ligands 6a,12a-diazadibenzo-[a,g]fluorenylium

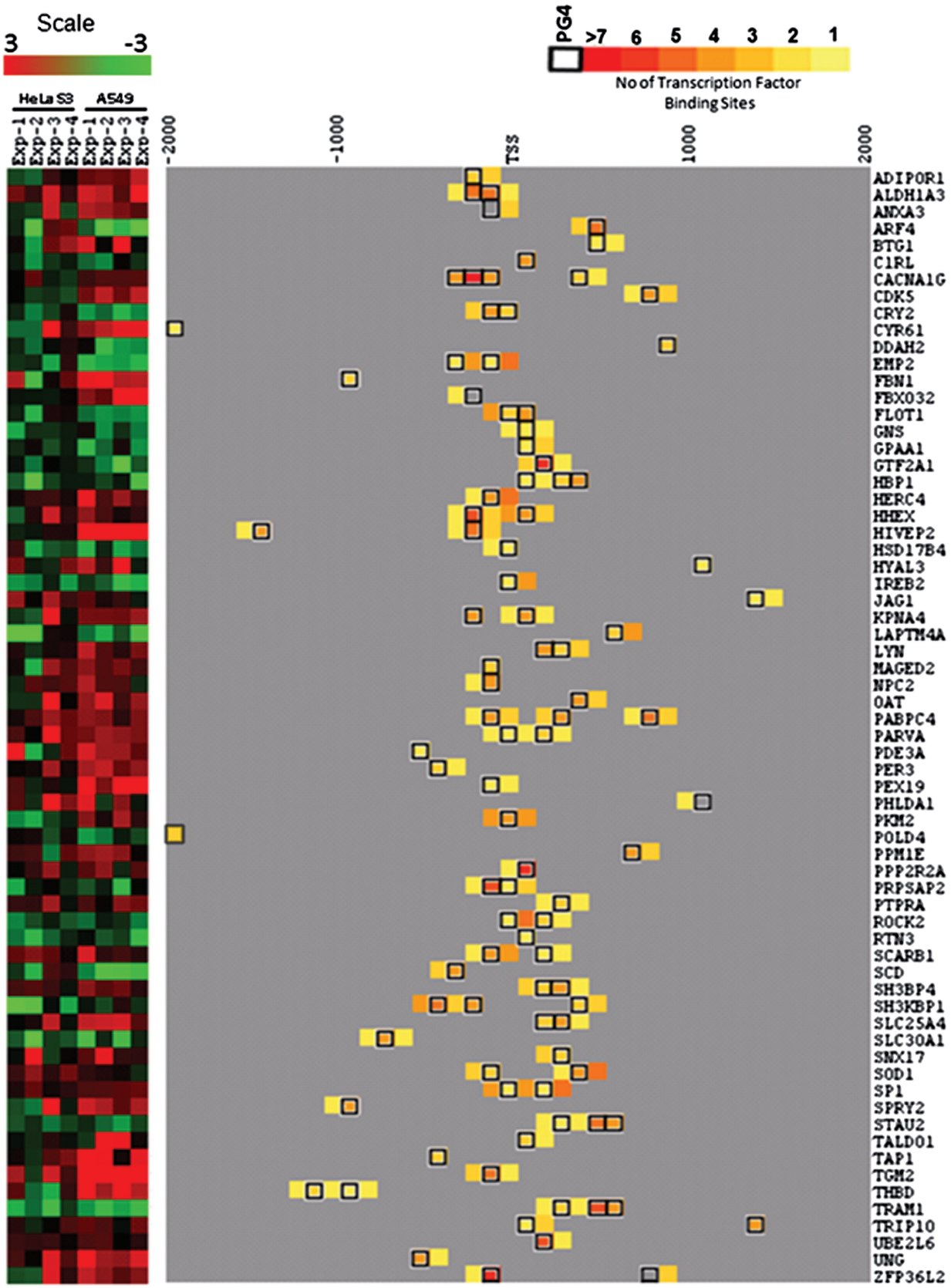

Figure 5. Genes harboring conserved PG4–TFBS associations are differentially expressed in presence of G-quadruplex binding ligand. Left panel: expression profile of genes with conserved PG4 motif that have significant differential expression in both cell lines on treatment with ligand. Pseudocolor representing their relative expression values in HeLaS3 and A549 cells. Right panel: association of PG4 motif with TFBS in promoters of differentially expressed genes shown in left panel; black box represents PG4 motif and associated pseudocolor shows number of TFBS within 100-base windows relative to TSS.

and 2,3,9,10-tetramethoxy-12-oxo-12H-indolo[2,1-a]-isoquinolinium chloride, BRACO-19 (3,6,9-trisubstuted acridine ligand), RHSP4 [3,11-difluoro-6,8,13-trimethyl-8H-quino(4,3,2-k1)acridinium methosulphate and telomestatin (38,39)]. A study by Grand *et al*. (40) showed reduced tumor growth due to decreased expression of *c-MYC* in presence of TMPyP4 which reduces hTERT and various others genes that together regulate telomere length and thereby enhance proliferative capacity of the cell. Based on our current results it is possible that the effect on cell proliferation/cell-cycle regulation observed in presence of the above ligands may be due to interaction with G-quadruplex motifs present in promoters, which thereby alter regulatory mechanisms involving TF, in addition to inhibition of telomerase or telomeric DNA amplification (41) by binding to telomeric G-quadruplex motifs. A recent study using a synthetic analog of telomestatin, HXDV (a hexaoxzole macrocycle) showed anti-proliferative activity and inhibition of cell-cycle progression leading to M-phase cell-cycle arrest due to specific G-quadruplex binding affinity of HXDV inside cells (37). Interestingly, the M-phase cell-cycle arrest was found to be independent of the telomerase status of cells (found also in telomerase-negative cells). This is consistent with our findings suggesting disruption of promoter G-quadruplexes and associated TF interactions lead to arrest in cell-cycle progression.

G-quadruplex DNA and zinc-finger proteins as binding pairs

Versatility of the zinc-finger binding pocket has been widely studied. The modular nature of the pocket and the variety of DNA (and RNA) elements, within a given generic code, that zinc-finger factors recognize is intriguing (42,43). Interestingly, this has led to the discovery of tailor-made nucleases that use the specificity of a given DNA sequence and the best-fit zinc-finger binding domain (44). On the other hand, G-quadruplex research increasingly points out the possibility of a vast number of structural motifs. Moreover, emerging reports suggest biological roles where variety within the G-quadruplex structural domain is evident (38). In this context, it is interesting to consider the implications of G-quadruplex–zinc-finger interactions as a pair, where the respective modular variations possible in both DNA and protein domains can be exploited. Keeping these in mind, it is tempting to speculate that *cis* and *trans* aspects of the G-quadruplex and zinc-finger interactions, respectively, have co-evolved to provide answers regarding how domain variations in DNA and the cognate protein binding domain are best utilized. Secondly, the perplexing number of G-quadruplexes in the genome is intriguing, raising considerable doubt regarding how many would be functional. Contextual use of G-quadruplex motifs is a plausible answer, presumably through *trans*-associations with protein factors that contextually extrude or bind G-quadruplexes in cell-type/state-specific fashion. Zinc-finger factors appear to be well-suited for this purpose, both, due to their contextual presence as well as domain-variations within a general theme, emphasizing the implications of the PG4 motif-zinc-finger associations found in this study.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Authors acknowledge Ram Krishna Thakur and Munia Ganguli for helpful discussions, careful reading, editing and comments on the manuscript

FUNDING

Senior research fellowship by Indian Council of Medical Research (to P.K.); Council of Scientific and Industrial Research (to V.Y. and A.B.); Department of Science and Technology, Govt. of India (LS-03/2006-07 to D.S.); European community seventh framework program (FP7/2007-2013) under grant agreement number (200754) the GEN2PHEN project (to Pr.K); research grant from CSIR (SIP006 to S.C.). Funding for open access charge: Research project SIP006 of Council of Scientific and Industrial Research.

Conflict of interest statement. None declared.

REFERENCES

1. Balagurumoorthy,P. and Brahmachari,S.K. (1994) Structure and stability of human telomeric sequence. *J. Biol. Chem*., **269**, 21858–21869.
2. Gellert,M., Lipsett,M.N. and Davies,D.R. (1962) Helix formation by guanylic acid. *Proc. Natl Acad. Sci. USA*, **48**, 2013–2018.
3. Sen,D. and Gilbert,W. (1988) Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*, **334**, 364–366.
4. Sundquist,W.I. and Klug,A. (1989) Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature*, **342**, 825–829.
5. Rawal,P., Kummarasetti,V.B., Ravindran,J., Kumar,N., Halder,K., Sharma,R., Mukerji,M., Das,S.K. and Chowdhury,S. (2006) Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. *Genome Res*., **16**, 644–655.
6. Yadav,V.K., Abraham,J.K., Mani,P., Kulshrestha,R. and Chowdhury,S. (2008) QuadBase: genome-wide database of G4 DNA–occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res*., **36**, D381–D385.
7. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res*., **35**, 406–413.
8. Verma,A., Halder,K., Halder,R., Yadav,V.K., Rawal,P., Thakur,R.K., Mohd,F., Sharma,A. and Chowdhury,S. (2008) Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species. *J. Med. Chem*., **51**, 5641–5649.
9. Du,Z., Kong,P., Gao,Y. and Li,N. (2007) Enrichment of G4 DNA motif in transcriptional regulatory region of chicken genome. *Biochem. Biophys. Res. Commun*., **354**, 1067–1070.
10. Halder,K., Halder,R. and Chowdhury,S. (2009) Genome-wide analysis predicts DNA structural motifs as nucleosome exclusion signals. *Mol. Biosyst*., **5**, 1703–1712.
11. Hershman,S.G., Chen,Q., Lee,J.Y., Kozak,M.L., Yue,P., Wang,L.S. and Johnson,F.B. (2008) Genomic distribution and

functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **36**, 144–156.
12. Wong,H.M. and Huppert,J.L. (2009) Stable G-quadruplexes are found outside nucleosome-bound regions. *Mol. Biosyst.*, **5**, 1713–1719.
13. Mani,P., Yadav,V.K., Das,S.K. and Chowdhury,S. (2009) Genome-wide analyses of recombination prone regions predict role of DNA structural motif in recombination. *PLoS ONE*, **4**, e4399.
14. Halder,R., Halder,K., Sharma,P., Garg,G., Sengupta,S. and Chowdhury,S. (2010) Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. *Mol. Biosyst.*, **6**, 2439–2447.
15. Siddiqui-Jain,A., Grand,C.L., Bearss,D.J. and Hurley,L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl Acad. Sci. USA*, **99**, 11593–11598.
16. Bejugam,M., Sewitz,S., Shirude,P.S., Rodriguez,R., Shahid,R. and Balasubramanian,S. (2007) Trisubstituted isoalloxazines as a new class of G-quadruplex binding ligands: small molecule regulation of c-kit oncogene expression. *J. Am. Chem. Soc.*, **129**, 12926–12927.
17. Cogoi,S. and Xodo,L.E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res.*, **34**, 2536–2549.
18. Patel,D.J., Phan,A.T. and Kuryavyi,V. (2007) Human telomere, oncogenic promoter and 5′-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic Acids Res.*, **35**, 7429–7455.
19. Basundra,R., Kumar,A., Amrane,S., Verma,A., Phan,A.T. and Chowdhury,S. (2010) A novel G-quadruplex motif modulates promoter activity of human thymidine kinase 1. *FEBS J.*, **277**, 4254–4264.
20. Verma,A., Yadav,V.K., Basundra,R., Kumar,A. and Chowdhury,S. (2009) Evidence of genome-wide G4 DNA-mediated gene expression in human cancer cells. *Nucleic Acids Res.*, **37**, 4194–4204.
21. Thakur,R.K., Kumar,P., Halder,K., Verma,A., Kar,A., Parent,J.L., Basundra,R., Kumar,A. and Chowdhury,S. (2009) Metastases suppressor NM23-H2 interaction with G-quadruplex DNA within c-MYC promoter nuclease hypersensitive element induces c-MYC expression. *Nucleic Acids Res.*, **37**, 172–183.
22. Paramasivam,M., Membrino,A., Cogoi,S., Fukuda,H., Nakagama,H. and Xodo,L.E. (2009) Protein hnRNP A1 and its derivative Up1 unfold quadruplex DNA in the human KRAS promoter: implications for transcription. *Nucleic Acids Res.*, **37**, 2841–2853.
23. Cogoi,S., Paramasivam,M., Membrino,A., Yokoyama,K.K. and Xodo,L.E. (2010) The KRAS promoter responds to Myc-associated zinc finger and poly(ADP-ribose) polymerase 1 proteins, which recognize a critical quadruplex-forming GA-element. *J. Biol. Chem.*, **285**, 22003–22016.
24. Uribe,D.J., Guo,K., Shin,Y.J. and Sun,D. (2011) Heterogeneous Nuclear Ribonucleoprotein K and Nucleolin as Transcriptional Activators of the Vascular Endothelial Growth Factor Promoter through Interaction with Secondary DNA Structures. *Biochemistry*, **50**, 3796–3806.
25. Yafe,A., Shklover,J., Weisman-Shomer,P., Bengal,E. and Fry,M. (2008) Differential binding of quadruplex structures of muscle-specific genes regulatory sequences by MyoD, MRF4 and myogenin. *Nucleic Acids Res.*, **36**, 3916–3925.
26. Guedin,A., Gros,J., Alberti,P. and Mergny,J.L. (2010) How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res.*, **38**, 7858–7868.
27. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
28. Qiu,P., Ding,W., Jiang,Y., Greene,J.R. and Wang,L. (2002) Computational analysis of composite regulatory elements. *Mamm. Genome*, **13**, 327–332.
29. Reed,B.D., Charos,A.E., Szekely,A.M., Weissman,S.M. and Snyder,M. (2008) Genome-wide occupancy of SREBP1 and its partners NFY and SP1 reveals novel functional roles and combinatorial regulation of distinct classes of genes. *PLoS Genet.*, **4**, e1000133.
30. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
31. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
32. Carmona-Saez,P., Chagoyen,M., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.*, **8**, R3.
33. Todd,A.K. and Neidle,S. (2008) The relationship of potential G-quadruplex sequences in cis-upstream regions of the human genome to SP1-binding elements. *Nucleic Acids Res.*, **36**, 2700–2704.
34. Isalan,M., Patel,S.D., Balasubramanian,S. and Choo,Y. (2001) Selection of zinc fingers that bind single-stranded telomeric DNA in the G-quadruplex conformation. *Biochemistry*, **40**, 830–836.
35. Patel,S.D., Isalan,M., Gavory,G., Ladame,S., Choo,Y. and Balasubramanian,S. (2004) Inhibition of human telomerase activity by an engineered zinc finger protein that binds G-quadruplexes. *Biochemistry*, **43**, 13452–13458.
36. Tapias,A., Ciudad,C.J., Roninson,I.B. and Noe,V. (2008) Regulation of Sp1 by cell cycle related proteins. *Cell Cycle*, **7**, 2856–2867.
37. Tsai,Y.C., Qi,H., Lin,C.P., Lin,R.K., Kerrigan,J.E., Rzuczek,S.G., LaVoie,E.J., Rice,J.E., Pilch,D.S., Lyu,Y.L. *et al.* (2009) A G-quadruplex stabilizer induces M-phase cell cycle arrest. *J. Biol. Chem.*, **284**, 22535–22543.
38. Balasubramanian,S., Hurley,L.H. and Neidle,S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug Discov.*, **10**, 261–275.
39. Yang,D. and Okamoto,K. (2010) Structural insights into G-quadruplexes: towards new anticancer drugs. *Future. Med. Chem.*, **2**, 619–646.
40. Grand,C.L., Han,H., Munoz,R.M., Weitman,S., Von Hoff,D.D., Hurley,L.H. and Bearss,D.J. (2002) The cationic porphyrin TMPyP4 down-regulates c-MYC and human telomerase reverse transcriptase expression and inhibits tumor growth in vivo. *Mol. Cancer Ther.*, **1**, 565–573.
41. De,C.A., Cristofari,G., Reichenbach,P., De,L.E., Monchaud,D., Teulade-Fichou,M.P., Shin-Ya,K., Lacroix,L., Lingner,J. and Mergny,J.L. (2007) Reevaluation of telomerase inhibition by quadruplex ligands and their mechanisms of action. *Proc. Natl Acad. Sci. USA*, **104**, 17347–17352.
42. Lu,D., Searles,M.A. and Klug,A. (2003) Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition. *Nature*, **426**, 96–100.
43. Wolfe,S.A., Nekludova,L. and Pabo,C.O. (2000) DNA recognition by Cys2His2 zinc finger proteins. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 183–212.
44. Klug,A. (2010) The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu. Rev. Biochem.*, **79**, 213–231.