SYNOPSIS

BACKGROUND

Genomic analysis of various organisms have shown approximately constant pattern of coding DNA content during evolution. Also, it has been observed that complexity of highly evolved genome are directly proportional to its non-coding content in comparison to least complex as well as less evolved organisms. Analysis of non-coding DNA of human and other closest relatives (viz. chimpanzee and macaque), has also revealed that approximately 45% is contributed by repetitive sequences. In contrast to earlier hypothesis where repetitive DNA sequences were considered as junks in the genome, these are increasingly being implicated in genome organization and function as well as in diseases. Repetitive DNA sequences are broadly classified into tandem repeats (microsatellite, mini-satellite and macro-satellite) and interspersed repeats (LTR and non-LTR (LINE and SINE)). Whereas majority of the tandem repeats (except microsatellite) are located at centromere and telomeric regions, the interspersed repeats are widely distributed across the genome and play critical role in genome reorganization. The latter therefore would be a source for genomic differences both within and between the species. Interspersed repeats are primarily transposable elements popularly called jumping genes and proliferate in the genome by cut and paste (eg. DNA transposons) as well as copy and paste (eg. retrotransposons) mechanisms. Depending on the site of insertion to genes, they exert different effects on gene’s structure, function and expression. For example when inserted at promoter proximal regions of a gene they could affect regulation of gene expression, whereas in the UTR regions they may affect efficiency of translation through interaction with miRNA and anti-sense transcripts etc.; when present in introns they could create exon-intron splice sites and thus be involved in generating protein diversity. The other class of repeats called microsatellites are simple repeats about <150 base pair in length with repeat unit <13 base pair. These repeats exhibit extensive length polymorphisms which in some cases be involved in genetic syndromes. These repetitive regions are also highly prone to substitutions and contribute to inter-individual and interspecies variation.
MOTIVATION

In the post sequencing era of genomics, when, we have an opportunity to discover the key genetic variations which are responsible for phenotypic and physiological differences between species through comparative genomics. At the one end, conserved DNA sequence regions between distantly related species provide indications of functionally critical region in the genome; sequence variability between closely related species could suggest regions for variability between them. Distinguishing differences that could be responsible for functional divergence between species is still a challenge and require functional validation. Microarray based analysis from different tissues between human and chimpanzee have revealed a significant number of genes that differ in expression despite similarity at the protein sequence level indicate difference between these two closely related species could be due to differences in regulation. In this study we attempted a genome wide study the contribution of non-coding and repetitive sequence in contributing to divergence between human and chimpanzee at 5kb promoter proximal region. We considered macaque as an out-group. We also tried to identify biological processes that have been impacted divergence and those that are conserved between species. Further we also looked at gain and depletion of TFBS as a consequence of this divergence particularly those in repetitive regions. We also attempted to explore the functional efficiency of TFBS in the repetitive elements especially Alu repeat taking a comparison of human and mouse promoter sequence.

OBJECTIVES

1. To carry out genome wide comparative analysis of 5kb upstream regions of genes in human, chimpanzee and rhesus monkey and identify genes which show non-coding promoter sequence divergence specially in the human lineage.
2. To carry out identification and analysis involving Gene-Ontology and Pathways of the highly diverged genes to choose candidates for further analysis.
3. To identify functional elements in the diverged sequences especially from repeat using bioinformatics approaches.
4. To test the functional consequence of putative regulatory non-coding sequences in zebrafish or other suitable expression system.
Divergence study at 5kb promoter proximal region

We retrieved 5kb promoter region from human genes as the reference point to look for JC-divergence (Jukes and Cantor 1969) between its orthologous sequences from chimpanzee and macaque. Changes in the promoter region of a gene could potentially affect expression profiles and lead to differences in pathway and biological processes ultimately to phenotypic or physiological changes. We considered macaque as an outgroup to identify human and chimpanzee specific divergence events. We restricted our study to those events of that could be contributed by changes in 5kb promoter proximal regions. Evidence from microarray analysis that have shown gene expression differences as more critical between primates rather than coding sequence difference was one of the motivation to consider these regions. This could help to identify key genes that could be prioritized for human and chimpanzee specific traits. Analysis of the JC-divergence between these three species revealed some outlier genes that are highly diverged either between human and chimpanzee or diverged from macaque to both human and chimpanzee lineages. The former was observed to be enriched in response to stimulus whereas latter in cellular macromolecule metabolic processes. This suggests divergence between macaque and human-chimp lineage at metabolic process level and between human and chimpanzee at the immunity level.

We analyzed the upper and lower 25% of the total diverged genes to identify enriched biological processes that are highly or least diverged between macaque, chimpanzee and human. This analysis revealed interesting outcomes about neuronal and immune system related biological processes which were earlier shown to be highly evolved between human and chimpanzee but at the coding level. We observed promoter region of genes related to system development and regulation of metabolic process to be highly conserved between all the three species (macaque, chimpanzee and human), suggesting that these genes could be highly critical for basic functions of the individual. In the highly diverged set we observed enrichment of neurological system process and G-protein coupled receptor protein signaling pathway between human and chimpanzee suggesting association with expression changes of their downstream genes that may contribute to species specific traits.
Contribution of repeat sequences towards the promoter sequence divergence

Repetitive sequences, which cover around 45% of the human genome are highly mutable and have also been shown to harbor various regulatory elements, we therefore also explored the contribution of repeats in divergence of promoter proximal regions of the genes between macaque, chimpanzee and human. We considered Alu-repeat primarily because it is primate specific, most abundant, harbors a large number of TFBS and in earlier studies in the lab has shown a non-random distribution in the genome. We measured the percentage contribution of repetitive DNA towards the promoter region divergence of each gene. We observed varying degree of repetitive DNA contribution towards the JC-divergence for individual genes. Analyzing the upper and lower 25% of the genes where repeats contribute to divergence or conserved respectively, our study revealed that processes related to regulation of immune system process and synaptic transmission the repeats contribute substantially to divergence between human and chimpanzee in comparison to non-repetitive regions. When we studied contribution of repeats to divergence between macaque and human, embryological skeletal system morphogenesis, sensory perception, anterior/posterior pattern formation and cognition biological processes were found to be enriched. Earlier analysis also shown the biasness in the genome wide distribution of repetitive sequences and which suggest that there has been a positive and crucial role of repetitive sequences towards divergence of regulation of specific genes and pathways which could lead to phenotypic and physiological differences to the individuals. Regulation of immune system and synaptic transmission are unique to the observed phenotypic and physiological differences between human and chimpanzee we thus postulate that repetitive sequences may have played a critical and important role in conferring differences between human and chimpanzee. In contrast we observed 63 genes wherein repeats were found almost conserved across two species. These mapped to genes relating to embryonic development, central nervous system development, skeletal system development and regulation of cell-differentiation. This study thus helps prioritize repetitive sequences that could contribute to divergence as well as functional conservation for primate specific characteristics.
Transcription factor binding site divergence (gain and depletion)

Encouraged from the results of above analyses that repetitive sequences contribute significantly to the promoter region sequence divergence of some crucial biological processes, we next explored the gain and depletion of regulatory sites (TFBS) which could act as enhancers or suppressors for their downstream genes. We selected genes from four biological processes (regulation of cell death, cognition, cell communication and response to stimulus) to study for the gain and depletion of TFBS in human and chimpanzee and considered macaque as an outlier. Taking macaque as an out-group is essential to identify human and chimpanzee specific events. If TFBS is present in human but absent in both macaque and chimpanzee it could be considered as human specific event; on the other hand, if any TFBS absent in human but present in macaque and chimpanzee it could be considered as depletion in human. Putative TFBS within the promoter sequences of the genes from selected biological processes were identified using p-match tool from TRANSFAC database with high quality matrix. In-house tools and algorithms have been designed for analysis of the conservation as well as gain and depletion of TFBS between orthologous promoter sequences of macaque, chimpanzee and human genes. Several TFBS were observed to be significantly gained and depleted in human and chimpanzee lineages in the given biological processes, some of them are discussed in the context of their enrichment in specific biological processes. We observed 35 TFBS that were significantly gained or depleted in chimpanzee and human from the mentioned four biological processes. The details are given in the following tables:

Regulation of cell death

Regulation of cell-death process includes events which modulate the rate or frequency of cell death. So we looked for the TFBS which are gained in human and tried to correlate the function of binding TF whether they relate to apoptosis or cell death processes and if it is, then it’s an indication for differential regulation and selection for apoptosis related processes in human lineage as suggested in a recent articles. We observed CKROX, CREB, FAC1, Sp3, USF2 and WT1 transcription factor related enrichment of TFBS in human lineage for this biological process and also these TFs have been found involved in T-cell lineage commitment, apoptosis, antiproliferative,
tumor association and transcriptional repression activities. On the other hand sites for MAF and RFX transcription factors have been found depleted in chimpanzee lineage and these TFs have also been shown to be involved in apoptotic pathway and MHC class II genes regulation.

**Cell-signaling**

*Cell signaling* involves any process that mediates the transfer of information from one cell to another. In this process we found sites significantly gained for CACD, FACL, KROX, Sp1, tal-1beta, WT1 and ZNF217 transcription factors in the chimpanzee lineage compared to human lineage and sites for Pax-4, Apz1 and TFE have been found depleted in human lineage and these TFs were found to be involved in various signaling pathways and apoptosis.

**Cognition**

*Cognition* is operation of the mind by which an organism becomes aware of objects of thought or perception; it includes the mental activities associated with thinking, learning, and memory. Sites of transcription factor FACL, Ikaros which are significantly gained in human lineage and their association with expression in fetal brain and Alzheimer disease as well as in pituitary gland are highly correlated with increment of cognitive processes in human lineage. On contrary there has been a significant depletion of sites of transcription factor FACL and NF1 in chimpanzee.

**Response to stimulus**

*Response to stimulus* is process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a stimulus. The process begins with detection of the stimulus and ends with a change in state or activity or the cell or organism. A large number of TFBS have been found significantly gained/depleted in human and gained but no depletion in chimpanzee among which sites for Egr-1, PLZF, RFX and Tal-1beta have been gained in chimpanzee which shown to be expressed in various cells of immune system. Sites for Crx, DBP, Ikaros, MYB and Sp1 have been gained in human and found to be associated with photoreceptor specific genes, coordination with circadian rhythm, lymphoid lineage specification and response to chemotherapeutic drugs. Sites for Cart-
1, CIZ, Kid3, RBP-Jkappa, Sp3, Spz1 and WT1 have been significantly depleted in human lineage and shown to associated with forebrain mesenchyme survival and initiation of neural tube, nucleocytoplasmic shuttling activity, development of brain and kidney, mediating Notch signaling and T-cell vs. B-cell fate decisions in lymphoid progenitors, Sp3 as opposing role for Sp1, regulating spermatogenesis and expression of WT1 in T-cells.

Except a few cases, above results have shown a positive correlation with the TFBS function and the biological processes where they have been gained or depleted in the human or chimpanzee lineages and indicates that human and chimpanzee divergence are not random. As a general observation, most of the TFBS which are gained or depleted in human and chimpanzee are involved in cell-signaling or apoptosis related pathways and are correlated with the biological processes. This could be further explored for the candidate genes and should be looked for further validation through functional genomics approach.

**Contribution of repetitive sequences in novel gain of TFBS in human lineage**

When we considered novel TFBS which are significantly gained in human from of the four selected biological processes and looked for the contribution of repetitive sequences in novel TFBS, we observed that ~40% of the novel TFBS are contributed by repetitive DNA sequences in which largest contribution from SINE (~24%) followed by LINE (~11%) then simple repeats (~7%) followed by DNA-elements (~3%) and other repeats (~0.1%). Though the percentage contribution of SINE and Simple repeats in human genome is less than LINE elements, their comparative contribution to novel TFBS gain in human is relatively higher and suggest their positive selection and functional role in the divergence between human and chimpanzee.

**Functional validation of TFBS gained in human in Alu-repetitive DNA**

Through the computational study, as we observed that Alu-repetitive DNA harbor ~1/4th of the novel TFBS gain in human compared to chimpanzee. We tried to validate the functionality of putative TFBS in Alu-element in human lineage (primates) in comparison to mouse (murine). Mouse has been selected for the comparison as to prove
hypothesis that during the evolutionary process Alu-repeats have gained novel TFBS through mutations in them to maintain the normal expression of the downstream gene. Because promoter region of neurophysiological processes have been observed to be highly diverged in our one of the preliminary study and also in this analysis we found neurological system processes, we have selected cell-lines from the neuronal origin as the expression system for analysis of repeat harbored putative TFBS. Retinoic acid receptor pathway has been selected for the study of regulation of gene expression on two basis, one it has very less number of steps are between stimulus and the effect and another its presence in Alu-repeats. So this would be interesting to validate transcriptional inducibility of retinoic acid binding site (RABS), harbored by Alu-repeat in human. In search for a gene suitable for the analysis of its promoter region, we have analyzed the differentially regulated genes from a microarray data (retinoic acid (30μM concentration) treated vs. untreated) of human primary melanocyte and mouse melanocyte cell-line (B16) for 24hrs. We have sorted out up-regulated gene in both human and mouse cell-line in treaded ones compared to untreated. Then we have selected out the KIF1B gene’s promoter based upon the criteria that it harbors three putative RABS in the Alu-repeat and two in the non-repetitive region in human while all (five putative RABS) in the non-repetitive region of mouse to find out whether RABS harbored by Alu-repeat is functionally active or not in human cell-line. We have validated the up-regulation of KIF1B gene by qPCR-analysis upon treatment of retinoic acid in both human (WM266.1) and mouse (B16) cell-lines. These putative RABS have been identified by using PROMO program developed in our lab previously. We have cloned 2.5 kb region of promoter sequence harboring putative RABS from human and mouse into the firefly luciferase plasmid vector (pGL4.23) having minimal promoter at the upstream of firefly luciferase gene then transfected this vector into human and mouse melanocyte cell-line (WM266.1 and B16 respectively) to measure the expression of luciferase enzyme (dual luciferase assay) and compared with the control to obtain the fold change. We observed that both human and mouse clones have given > 2.5 and >3.5 fold upregulation of the luciferase gene respectively upon treatment with retinoic acid. Next we had to confirm that whether putative RABS harbored by repetitive Alu-element is functional or not, to verify this we mutated each of the four sites in the human clone insert (two inside the Alu-repeat and two outside the repeat) using site
directed mutagenesis procedure and transfected in human WM266.1 cell-line to check for basal expression of luciferase gene before and after treatment with retinoic acid in comparison to non-mutated clones. We observed that mutation-1 (M-1) inside the RABS from repeat and mutation-5 (M-5) outside the repetitive DNA RABS exerts major effects towards down-regulation of the luciferase gene expression. Our analysis demonstrate that putative TFBS contributed by repetitive sequences which gained or depleted in the promoter proximal region, could contribute to the functional diversity in orthologous genes of individuals between and within species.

SUMMARY AND FUTURE DIRECTIONS

The genome wide comprehensive analysis of 5kb promoter proximal region in respect of divergence and regulatory aspect between human, chimpanzee and macaque has given insight about the peculiar biological processes some of which have also been shown to be highly diverged or evolved during the divergence time between these species. Our analysis has been novel towards finding genes and the regulatory elements which have been uniquely gained or depleted in these lineages and create a base ground from where the specific genes could be picked up and could be validated for the particular TFBS whether they are functionally relevant to the particular category which are diverged between the species. Our result also demonstrate about half of the regulatory elements contributed by repetitive sequences (majorly Alu-repeats) and also that these repeats are primate specific it is obvious that these must have played a crucial role to the evolution of primate lineage as we validated the functional relevance of them by harboring various regulatory elements.