BIOINFORMATICS DISCOVERY NOTE
Vol. 19 no. 5 2003, pages 549–552
DOI: 10.1093/bioinformatics/btg029

Triplet repeats in human genome: distribution and their association with genes and other genomic regions

Subbaya Subramanian[1], Vamsi M. Madgula[2], Ranjan George[2], Rakesh K. Mishra[1], Madhusudhan W. Pandit[1], Chanderashekar S. Kumar[2] and Lalji Singh[1,*]

[1]Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad 500 007, India and [2]Ingenovis, ilabs ltd. 97, Road No. 3, Banjara Hills, Hyderabad, 500 034, India

Received on August 20, 2002; revised on October 11, 2002; accepted on November 6, 2002

ABSTRACT

Motivation: Simple sequence repeats (SSRs) or microsatellite repeats are found abundantly in many prokaryotic and eukaryotic genomes. Among SSRs, triplet repeats are of special significance because some of them have been linked to various genetic disorders. The objective of the study is to analyze the triplet repeats of complete human genome and to identify the genes that contain the triplet repeats in their coding region. The analysis will help us to identify the candidate genes that have potential for repeat expansion.

Results: We have analyzed triplet repeats in the complete human genome from the publicly available sequences. Our analysis revealed that AGC and CCG repeat were predominantly present in the coding regions of the genome while UTRs and the upstream sequences contained CCG repeats in relative abundance. Analysis of density of triplet repeats (bp/Mb) revealed that AAT and AAC were the abundant repeats whereas ACT and ACG were the rare repeats found in human genome. We could identify about 2135 known or predicted genes that were associated with at least one of the triplet repeat types. A large proportion of putative transcripts that were identified by gene finding programs were found to be associated with triplet repeats. These transcripts will be the candidate genes for analysis of triplet repeat expansion and a possible association with disease phenotypes. Identification of 171 genes which contain a minimum of ten repeat units will be of particular interest in future in correlating their association with any disease phenotype due to the expansion potential of repeats present in them. The list of genes and other details of analysis are given in the online supplementary data (http://www.ingenovis.com/tripletrepeats).

Contact: lalji@ccmb.res.in

*To whom correspondence should be addressed.

INTRODUCTION

Simple sequence repeats (SSRs) or microsatellite repeats are abundantly found in many prokaryotic (Gur-Arie *et al.*, 2000) and eukaryotic (Toth *et al.*, 2000) genomes. Their significance in the genome, however, is poorly understood. Among SSRs, triplet repeats are of special significance because some of them have been linked to various genetic disorders (Pearson and Sinden, 1998). Since the microsatellites are unstable, their number of repeats can grow from one generation to the next (Richard and Paques, 2000). This expansion of repeats in the coding region of the gene or their very presence in the regulatory regions, which may alter the gene expression, are found to be associated with disease phenotype (Cummings and Zoghbi, 2000).

In humans, microsatellites in the form of trinucleotide repeats can be found in genes that are associated with several neurological diseases such as fragile X syndrome (Jin and Warren, 2000), Huntington's disease and several forms of ataxia (Sermon *et al.*, 2001), and Myotonic dystrophy (Timchenko *et al.*, 2001). The expansion of triplet repeats is generally thought to be due to DNA replication slippage and unequal recombination in multiplying cells (Khajavi *et al.*, 2001). The recent completion of human genome sequencing allows us to study the triplet repeats in detail. We have analyzed triplet repeats in the complete human genome for their occurrence, distribution and their association with genes. We have also extensively analyzed the distribution of these repeats and their association with coding and non-coding regions of the genome.

METHODS

The DNA sequences from GenBank (ftp://ftp.ncbi.nlm.nih.gov/genomes/h_sapiens; build number 28) have been used for the purpose of generation of the triplet repeats data. The reference sequences of GenBank are used for

the analysis. All theoretically possible non-overlapping triplet repeat types, viz. AAC, AAG, AAT, ACC, ACG, ACT, AGC, AGG, ATC, CCG were searched in the human genome. We have analyzed the distribution of perfect repeats of the length $\geqslant$12 base pairs. A Java based program has been developed and used to scan the entire genome to find the abundance and distribution of these repeats in coding and non-coding regions of the genome. The occurrence of repeats in the genomic regions such as exons, introns and UTRs has been identified based on the annotation of the human genome sequence in the GenBank database. We calculated the repeat density for each repeat type by analyzing the occurrence of the repeat in each chromosome and measuring the length in base pairs contributed by each repeat type. The repeat density (bp/Mb) of each chromosome was calculated by dividing the total length in terms of base pairs of sequence contributed by triplet repeats by the chromosome size (Mb).

RESULTS

Density and distribution of triplet repeats

The number of occurrences of triplets in various genomic regions is given in online supplementary data: (Chromosome-wise occurrence of triplet repeats in various genomic regions of human genome). As expected, chromosome 1 contained the maximum number of occurrences of triplet repeats and the Y chromosome, the minimum. The density of AAT repeats was highest across the genome (online supplementary data: Density of triplet repeats (length of repeat in bp/Mb). The average length of this repeat was 282 bp/Mb of sequence. This was followed by AAC (avg. 193 bp/Mb), AAG and AGG (both avg. 77 bp/Mb). ACG repeats had the lowest density in most of the chromosomes; the average bp/Mb ranged from 0 to 1. The next lowest density was with ACT repeats (avg. 9 bp/Mb). Analysis of the distribution pattern revealed that the smaller repeat numbers were predominant in the genome. As the repeat number increases the occurrence decreases (online supplementary data: Distribution of repeat number of triplet repeat types). AAT (26 037), AAC (18 707) followed by AGG (10 245) had the maximum number of occurrences for 4 tandem repeats. However, even shorter lengths of 4 tandem repeats of ACG and ACT are not predominantly found in the human genome.

Distribution of triplet repeats in various genomic regions

When triplet repeat types were analyzed in terms of their occurrence in various genomic regions, it was observed that, AAT repeats were abundantly present in the intronic and intergenic sequences as compared with their occurrences in exonic regions and UTRs, AAC being the next abundant triplet repeat (online supplementary data: Occurrence of triplet repeat types in introns and intergenic regions). Repeats like ACG and ACT showed relatively poor presence in all the regions analyzed. This suggests that the occurrence of these repeats is generally less abundant when compared with other triplet repeats irrespective of the genomic regions analyzed. In exonic regions triplet repeats such as AGC, AGG, CCG were found to be most abundant, in that order (online supplementary data: Occurrence of triplet repeat types in exons and UTRs), whereas in the UTRs, the triplets such as CCG and AAT followed by AAC showed predominant occurrence although their values are relatively low when compared with their occurrence in exonic regions. Chromosome-wise analysis of triplet repeat occurrence with respect to various genomic regions is given in online supplementary data.

Association of triplet repeats with genes

From our analysis we could identify about 2135 genes that contained at least one triplet repeat type in their exons with at least four tandem repeats (online supplementary data: Association of triplet repeats with genes). From the list of 2135 genes we have identified 171 genes that contained repeat length of at least 30 bp (supplementary data: Genes containing 10 or more repeats of at least one triplet repeat type). The analysis revealed that the genes maximally contained AGC and CCG repeats. Repeats such as AAC, AAT, AAG, ACC and ATC are not predominantly found in genes; and ACT as well as ACG triplets were completely absent. There are a large number of uncharacterized/putative transcripts containing triplet repeats that can potentially correspond to disease-linked genes. Further analysis revealed that AGC and CCG contributed to the extent of $\sim$67% to these repeats.

Association of repeats with upstream regions of the genes

Analysis of the 500 bp sequences, upstream of the gene, revealed that there are preferential associations of certain repeat types with these sequences. Most of the upstream sequences contained tandem repeats of CCG (online supplementary data: Association of the triplet repeats in the upstream regions of genes). The next most abundant repeat type in the upstream sequences is AGG followed by AAT and AAC. The abundance of AAT and AAC could be explained on the basis of their overall predominance in the genome; but this is not true in the case of CCG and AGG. In such circumstances it is not very easy to explain their predominance in the upstream sequences; however as the upstream regions of the genes contain CpG islands, it is meaningful that most of the upstream sequences contain CCG repeats. Out of 269 genes that contained at least one triplet repeat in the upstream region, none of the upstream sequences contained ACT repeats.

DISCUSSION

A large proportion of the non-coding genome is repetitive DNA, which can be of many types based on the length of the repeating elements and whether they are occurring in tandem. Among different types of repeats, transposable elements are the most abundant and make up to ~45% of the human genome and SSRs occupy ~3% of the genome (International Human Genome Sequencing Consortium, 2001). Among various repeat types triplet repeats are studied to a great extent as their expansion beyond a critical level is known to lead to a disease phenotype (Sinden, 2001).

Triplet repeat distribution

Triplets are found to be present in both coding and non-coding regions of the genome. The exons and UTRs contain many triplet repeats of which the AGC repeats are most abundant in the exonic region. The next most abundant triplet repeats are AGG and CCG. Repeats such as ACG and ACT are rarely present in the exonic regions. The ACT repeats can also act as a stop codon TGA, which may explain their lower occurrences in exons. However, these repeats are poorly represented in the intergenic and intronic regions also, suggesting that existence of such repeated structures may be detrimental to DNA and affect its function. The AAT repeats are most abundant in the intronic regions of human; this is in agreement with Toth *et al.* (2000). They have also reported that AAC repeats are the most abundant repeats in the intronic region of other mammals. As compared with other triplet repeats CCG repeats are predominantly present in the UTRs as well as in the upstream regions of the genes. As UTRs and upstream regions of many genes contain the regulatory elements and CpG islands, it is possible that these repeats might have a regulatory role in genes which contain such repeats in their upstream sequences.

Searching of databases for new putative disease-causing repeat classes has led to the identification of many repetitive elements (Kleiderlein *et al.*, 1998). There are different classes of repeats that are thought to be involved in the disease phenotypes, e.g. unstable repeats of $(AT)_{21}$ (Hewett *et al.*, 1998). Expansion of repeats which is a consequence of replication slippage may also influence the packaging of the DNA and may even have regulatory implications in some cases (Filippova *et al.*, 2001).

Association of triplets with genes

Our search for triplets with repeat number 10 and above revealed that among the repeats that are associated with genes, AGC and CCG repeats are most frequent, followed by AGG repeats. However, triplets such as ACT and ACG are present very rarely in the exonic regions. These observations suggest that repeats of certain amino acids may not occur in the protein sequences; it is quite likely that occurrences of such repeats of amino acids may affect the stability of the protein or it may even interfere with the folding process of the protein. Earlier studies have reported that asparagine repeats are less abundant in mammalian proteins (Kreil and Kreil, 2000). Therefore it is also possible that selective elimination of such repeat types from the coding regions favours events such as speciation and evolutionary advancements. In case of frequently occurring amino acids, the expansion beyond a critical length of amino acid repeats can lead to structural perturbations with a consequence of disease state. The amino acids stretches formed by AGC and CCG repeats are probably allowed in protein sequences, to the extent that they may not influence the stability /folding / function of the protein with their normal state of repeat numbers. Expansion of such repeats beyond a critical length leads to disease and also can influence the chromatin structure and gene expression (Frisch *et al.*, 2001). Further, the analysis of UTRs for the triplet repeats revealed that CCG repeat type is the most abundant repeat type. This could be a reflection of the CpG islands which are likely to be present in the upstream sequences of genes.

Our analysis of triplet repeats in the complete human genome identifies genes, which are associated with these repeats. This analysis will be helpful in future in predicting the genes that are predisposed for disease phenotype. The analysis can also help in identification of candidate genes for population studies. Finally, in our analysis, the triplet association in the genomic regions is based on the annotated gene sequences in the database. It is possible that there are certain exons and introns in the genome, which are not predicted or annotated precisely. Therefore, the ratio of triplet repeat number within genes to intergenic regions is tentative. In future, when the human genome is available in a completely finished form, more number of genes associated with the triplet disorders could be predicted.

ACKNOWLEDGEMENTS

The authors are thankful to Sreedhar, Kavitha and Saritha for their support and technical assistance. We are grateful to Ramesh Aggarwal and K.Thangaraj for providing helpful discussions. Fellowship to S.S. from CSIR is duly acknowledged.

REFERENCES

Cummings,C.J. and Zoghbi,H.Y. (2000) Fourteen and counting: unraveling trinucleotide repeat diseases. *Hum. Mol. Genet.*, **9**, 909–916.

Filippova,G.N., Thienes,C.P., Penn,B.H., Cho,D.H., Hu,Y.J., Moore,J.M., Klesert,T.R., Lobanenkov,V.V. and Tapscott,S.J. (2001) CTCF-binding sites flank CTG/AGC repeats and form a methylation-sensitive insulator at the DM1 locus. *Nature Genet.*, **28**, 335–343.

Frisch,R., Singleton,K.R., Moses,P.A., Gonzalez,I.L., Carango,P., Marks,H.G. and Funanage,V.L. (2001) Effect of triplet repeat expansion on chromatin structure and expression of dmpk and neighboring genes, six5 and dmwd, in myotonic dystrophy. *Mol. Genet. Metab.*, **74**, 281–291.

Gur-Arie,R., Cohen,C.J., Eitan,Y., Shelef,L., Hallerman,E.M. and Kashi,Y. (2000) Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. *Genome Res.*, **10**, 62–71.

Hewett,D.R., Handt,O., Hobson,L., Mangelsdorf,M., Eyer,H.J., Baker,E., Sutherland,G.R., Schuffenhauner,S., Mao,J.J. and Richards,R.I. (1998) FRA10B Structure reveals common elements in repeat expansion and chromosomal fragile sites genesis. *Mol. Cell*, **1**, 773–781.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Jin,P. and Warren,S.T. (2000) Understanding the molecular basis of fragile X syndrome. *Hum. Mol. Genet.*, **9**, 901–908.

Khajavi,M., Tari,A.M., Patel,N.B., Tsuji,K., Siwak,D.R., Meistrich,M.L., Terry,N.H. and Ashizawa,T. (2001) 'Mitotic drive' of expanded CTG repeats in myotonic dystrophy type 1 (DM1). *Hum. Mol. Genet.*, **10**, 855–863.

Kleiderlein,J.J., Nisson,P.E., Jessee,J., Li,W., Becker,K.G., Derby,M.L., Ross,C.A. and Margolis,R.L. (1998) CCG repeats in cDNA s from human brain. *Hum. Genet.*, **103**, 666–673.

Kreil,D.P. and Kreil,G. (2000) Asparagine repeats are rare in mammalian proteins. *Trends Biochem. Sci.*, **25**, 270–271.

Pearson,C.E. and Sinden,R.R. (1998) Trinucleotide repeat DNA structures: dynamic mutations from dynanic DNA. *Curr. Opin. Struct. Biol.*, **8**, 321–330.

Richard,G.F. and Paques,F. (2000) Mini- and microsatellite expansions: the recombination connection. *EMBO Rep.*, **1**, 122–126.

Sermon,K., Seneca,S., De Rycke,M., Goossens,V., Van de Velde,H., De Vos,A., Platteau,P., Lissens,W., Van Steirteghem,A. and Liebaers,I. (2001) PGD in the lab for triplet repeat diseases - myotonic dystrophy, Huntington's disease and Fragile-X syndrome. *Mol. Cell Endocrinol.*, **183** Suppl. 1, S77–85.

Sinden,R.R. (2001) Neurodegenerative diseases: Origins of instability. *Nature*, **411**, 757–758.

Timchenko,N.A., Iakova,P., Cai,Z.J., Smith,J.R. and Timchenko,L.T. (2001) Molecular basis for impaired muscle differentiation in myotonic dystrophy. *Mol. Cell Biol.*, **21**, 6927–6938.

Toth,G., Gaspari,Z. and Jurka,J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967–981.