BMC Genomics

BioMed Central

Database

Expoldb: expression linked polymorphism database with inbuilt tools for analysis of expression and simple repeats

Vineet K Sharma[1], Anu Sharma[2], Naveen Kumar[1], Mamta Khandelwal[1], Kiran Kumar Mandapati[2], Shirley Horn-Saban[3], Liora Strichman-Almashanu[4], Doron Lancet[4], Samir K Brahmachari[1] and Srinivasan Ramachandran*[1]

Address: [1]G.N. Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, Mall Road, Delhi 110 007, India, [2]Functional Genomics Unit, Institute of Genomics and Integrative Biology, Mall Road, Delhi 110 007, India, [3]Microarray facility, Department of Biological Services, Weizmann Institute of Science, Rehovot 76100, Israel and [4]Department of Molecular Genetics and Crown Human Genome Center, Weizmann Institute of Science, Rehovot 76100, Israel

Email: Vineet K Sharma - vineetks@gmail.com; Anu Sharma - anukaush@gmail.com; Naveen Kumar - naveenkumar@igib.res.in; Mamta Khandelwal - kmamta@hcltech.com; Kiran Kumar Mandapati - kirankumar341@rediffmail.com; Shirley Horn-Saban - shirley.saban@weizmann.ac.il; Liora Strichman-Almashanu - liora.almashanu@weizmann.ac.il; Doron Lancet - doron.lancet@weizmann.ac.il; Samir K Brahmachari - skb@igib.res.in; Srinivasan Ramachandran* - ramuigib@gmail.com

* Corresponding author

Published: 13 October 2006

BMC Genomics 2006, **7**:258 doi:10.1186/1471-2164-7-258

Received: 06 July 2006
Accepted: 13 October 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/258

Abstract

Background: Quantitative variation in gene expression has been proposed to underlie phenotypic variation among human individuals. A facilitating step towards understanding the basis for gene expression variability is associating genome wide transcription patterns with potential *cis* modifiers of gene expression.

Description: EXPOLDB, a novel Database, is a new effort addressing this need by providing information on gene expression levels variability across individuals, as well as the presence and features of potentially polymorphic $(TG/CA)_n$ repeats. EXPOLDB thus enables associating transcription levels with the presence and length of $(TG/CA)_n$ repeats. One of the unique features of this database is the display of expression data for 5 pairs of monozygotic twins, which allows identification of genes whose variability in expression, are influenced by non-genetic factors including environment. In addition to queries by gene name, EXPOLDB allows for queries by a pathway name. Users can also upload their list of HGNC (HUGO (The Human Genome Organisation) Gene Nomenclature Committee) symbols for interrogating expression patterns. The online application 'SimRep' can be used to find simple repeats in a given nucleotide sequence. To help illustrate primary applications, case examples of Housekeeping genes and the *RUNX* gene family, as well as one example of glycolytic pathway genes are provided.

Conclusion: The uniqueness of EXPOLDB is in facilitating the association of genome wide transcription variations with the presence and type of polymorphic repeats while offering the feature for identifying genes whose expression variability are influenced by non genetic factors including environment. In addition, the database allows comprehensive querying including functional information on biochemical pathways of the human genes.

EXPOLDB can be accessed at http://expoldb.igib.res.in/expol

Background

Functional genomics in the post human genome sequencing era is greatly facilitated by correlating expression data with sequences of potential regulatory elements. The primary repositories of gene expression data such as Gene Expression Omnibus (GEO) [1], UniGene [1], Gene Expression Database (GXD) [2], and Gene Expression Atlas (GNF) [3] provide useful information on gene expression obtained from microarrays and other techniques, however, they provide limited information on the role of genetic elements that can potentially modulate gene expression. Thus, there is a need for databases integrating gene expression information with sequence information of potential genetic regulators with propensity for exhibiting sequence variability.

One such potential regulator is the dinucleotide repeat $(TG/CA)_n$. The $(TG/CA)_n$ repeats are widely distributed, considered to be *cis* regulators of transcription, and above 12 repeat units tend to be polymorphic [4-6]. Segments of DNA consisting of $(TG/CA)_n$ repeats (with $n \geq 23$), display, under conditions close to physiological, the propensity to adopt a Z-form [7-9], a conformation which affects the movement of the RNA polymerase [10], and binding of transcriptional factors [11]. In addition, these repeats have been observed to be associated with recombination sites [12] and mRNA splicing [13], which elect them as functional elements in humans [14].

The $(TG/CA)_n$ repeats can be divided into three length categories, based on their biological properties [14]. Type I repeats ($6 \leq n < 12$) have very low propensity for polymorphism [15]. Type II repeats ($12 \leq n < 23$) are likely polymorphic, as more than 93% of the $(CA)_n$ repeats of $n \geq 12$ units were found to display length polymorphism and act as *cis* regulators of transcription [4]. Type III repeats ($n \geq 23$) were shown to have a propensity to adopt conformations such as Z DNA [7-10], and to be associated with recombination sites [12]. In general, $(TG/CA)_n$ repeats of $n \geq 12$ units exert a down regulatory effect on transcription, which is positively correlated with the length of repeats [16]. A few examples of genes, whose transcription levels were shown to be modulated by $(TG/CA)_n$ repeats, are human *HSD11B2* [16], *MMP-9* [18], *IFN-γ*[19], *EGFR* [20], and housekeeping genes [22] and others such as rat *α-lactalbumin* [9], *prolactin* [17], and *nucleolin* [11], and tilapia *prolactin-1*[21]. These repeats also exhibit preference for binding to nuclear factors in some instances [24] and stimulate mRNA splicing [13].

In this work, we report the construction of EXPOLDB (EXpression linked POLymorphism DataBase), a novel database focusing on the effect of $(TG/CA)_n$ repeats on transcription level and variation between individuals. In this first release, EXPOLDB was constructed using expression information from our GeneChip experiments [25] using novel analysis tools.

Construction and content

Gene expression data

All data in EXPOLDB resides locally and are retrievable conforming to 'open access'. In this first release, we used GeneChip (HG-U95Av2 arrays, Affymetrix) expression data of blood leukocytes obtained from 13 normal healthy human individuals including 5 pairs of monozygotic twins (GEO series accession No. GSE928) [25].

Mean expression value for each gene was computed from the $\log_{10}$ transformed 'signal values' with 'P' (Present) calls. The coefficient of variation (CV) was computed as SD/Mean where SD is the standard deviation. 'Signal log ratio' is the difference in expression level for a transcript between two experiments, and is computed by Affymetrix Microarray Suite Software MAS 5.0. Differentially expressed genes in pair-wise comparisons were identified as those with a signal log ratio > 1.585, after considering the experimental noise [25].

Sequence retrieval and mapping of $(TG/CA)_n$ repeats

Genomic sequences of human genes (Build 35) were retrieved in GenBank format from Entrez Gene [1] and parsed to obtain the exon and intron information and the gene sequence. Identification of uninterrupted Type I, Type II and Type III intragenic $(TG/CA)_n$ repeats was carried out using the Perl script 'SimRep' [14,22]. The genomic region between the most upstream transcript start and the most downstream transcript end of a gene, with the addition of 1 kb 5' of upstream flanking region, was scanned for repeats. The positions of the repeats were noted and displayed with respect to gene structure (exons and introns). For genes with reported alternative spliced variants, the repeats distribution is displayed with respect to the exon-intron structure of the gene corresponding to each individual splice variant. Information on known polymorphic repeats was obtained from CEPH database [26] and mapped to genes using UniSTS [1]. *Alu* repeats were mapped using RepeatMasker [27].

Expression and functional information from other sources

Information about genes expressed in Blood was derived from Expressed Sequence Tags (ESTs), retrieved from the UniGene database (Build 160) [1]. Genes were classified as highly expressed according to a previously defined criteria (H; >0.0363% of the total detected transcription), moderately expressed (M; from 0.0363% to 0.0121%), or weakly expressed (W; <0.0121%) genes [28]. The dataset of human housekeeping genes was obtained from Eisenberg and Levanon [29]. Pathway information on genes was obtained from NetAffx (version 23rd June 2004) [30].

Statistical Analysis

Statistical tests were carried out using the 'Statistical Pages' [31]. Differences in expression levels between genes with intragenic $(TG/CA)_n$ repeats (only Type II and Type III were considered – based on their current experimental evidence as *cis* regulators of transcription with propensity for polymorphism) and genes without repeats ($n < 6$ units) were evaluated using *t-test*.

Database and Web interface

The back end data was prepared in MS Access 2000 (Microsoft Corporation Inc., USA). Server side scripting was prepared using ASP (Active Server Pages, Version 3.0), PHP (PHP: Hypertext Preprocessor, version 5.0) and Perl (Practical Extraction and Report Language, version 5.8.1). The client side scripting was prepared using JavaScript and HTML (Hyper Text Markup Language, version 4.0). Internet Information Server (IIS, version 6.0) was used as Web server.

Utility

The tetrapodic layout of EXPOLDB is shown in Figure 1. All attributes of a gene, are singularly linked to its official HGNC symbol serving as the primary key. A brief description of the potential uses of EXPOLDB is presented below.

Examining gene expression and variability

EXPOLDB houses gene expression information from monozygotic twins and other unrelated individuals. Queries enable retrieving information about genes expressed in blood, or genes exhibiting inter-individual expression variability. These query pages can be accessed through 'Query EXPOLDB'. Wild cards (*) and multiple keywords can be used with Boolean operators. Queries can be limited by different attributes, such as Chromosome number, HGNC gene symbol, gene function, UniGene ID, accession number of known polymorphic repeats, biochemical pathway, and the range of expression variability. Two indices are provided to assess variation in gene expression: 'coefficient of variation' (CV), and 'signal log ratio'. CV is provided with the 'Expression in Blood" query page, and indicates the overall variability in a set of individuals. We

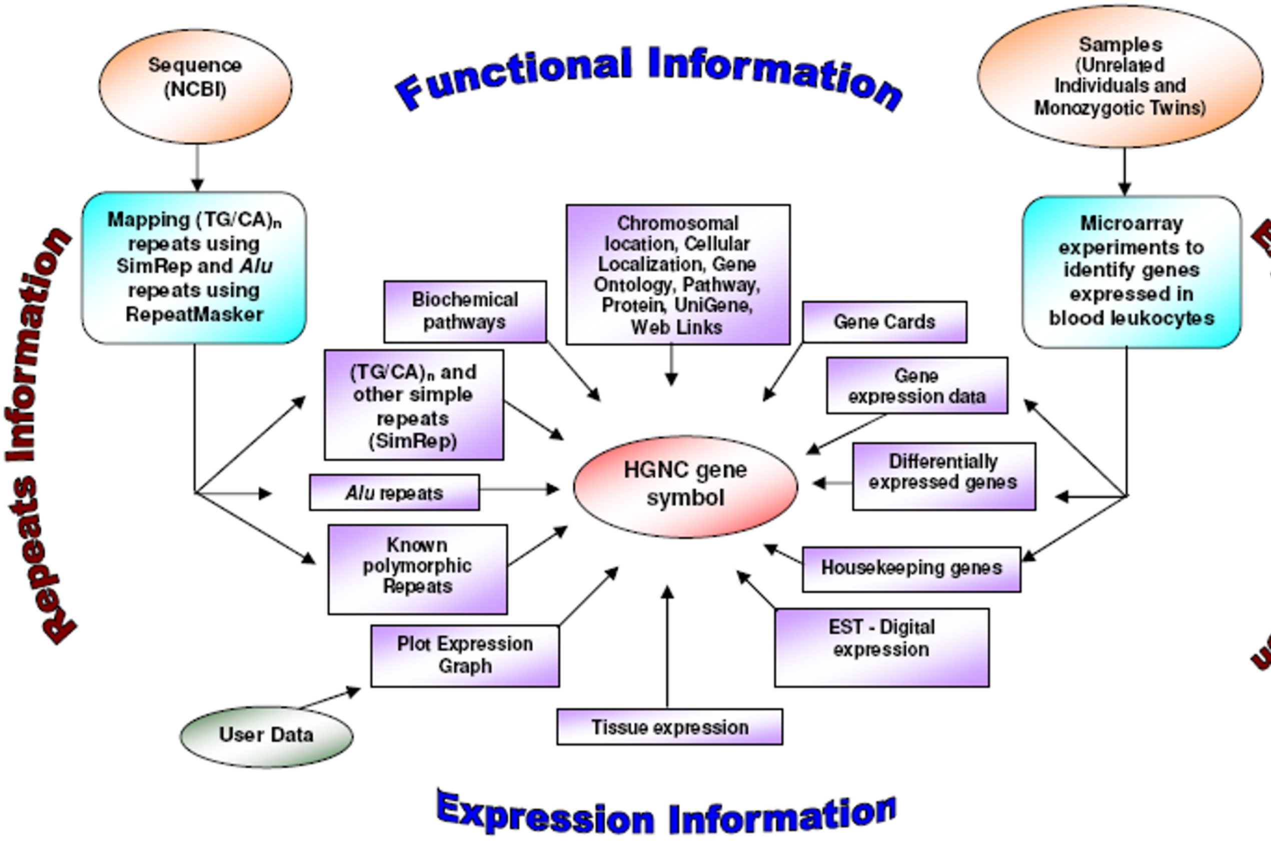

Functional Information
Sequence (NCBI)
Samples (Unrelated Individuals and Monozygotic Twins)
Repeats Information
Experimental Information
Mapping (TG/CA)n repeats using SimRep and Alu repeats using RepeatMasker
Microarray experiments to identify genes expressed in blood leukocytes
Chromosomal location, Cellular Localization, Gene Ontology, Pathway, Protein, UniGene, Web Links
Biochemical pathways
Gene Cards
(TG/CA)n and other simple repeats (SimRep)
Gene expression data
Alu repeats
HGNC gene symbol
Differentially expressed genes
Known polymorphic Repeats
Housekeeping genes
Plot Expression Graph
EST - Digital expression
User Data
Tissue expression
Expression Information

Figure 1

Tetrapodic layout of EXPOLDB: The 4 domains of layout are shown in bold face type. Note that all attributes of a gene are singularly linked to its official HGNC symbol serving as the primary key have presented one application of CV, to identify 'control' or 'reference' housekeeping genes whose expression is most constant across all individuals regardless of their genetic relationship [32]. On the other hand, 'signal log ratio' is provided with the "Differential Expression" query page, and since it is generally used for pair-wise comparisons it is more suitable to assess differential expression between monozygotic twins. Discordance in expression in this case could indicate the lack of genetic effect, and potential involvement of environmental factors. The mathematical relationship between the two measures has not been worked out yet in the literature and therefore, users are advised to exercise caution while using the two measures. In principle, the use of a given metric is guided by the biological question at hand.

The query for differentially expressed genes offers three limited sets, in addition to searching all genes: (1) Genes differentially expressed in monozygotic twins, which are less likely to be influenced by genetic factors, (2) Genes differentially expressed in unrelated, age matched (20–23 yr) female individuals but not in monozygotic twins (including all possible pair-wise comparisons), these are more likely to be affected by genetic factors, and (3) Differentially expressed housekeeping genes (in all pair-wise comparisons including monozygotic twins).

Querying

Submission of a query produces a 'Results' page listing all the resultant gene matches. The expression of genes in different individuals can be examined visually either singly or collectively as bar charts in the graphic display by selecting the appropriate square boxes placed in front of the listed genes. Detailed information of genes (EXPOLDB profile) can be obtained by clicking on the HGNC gene symbol displayed on the 'Results' page.

EXPOLDB Profile

The EXPOLDB profile of a gene summarizes the information on the function, expression and repeat content of the gene. The table 'Expression in Blood' provides mean expression, coefficient of variation (CV) and EST based expression status ('H'/'M'/'W' for High/Medium/Weak). The expression levels of a gene in unrelated age and gender matched individuals and in monozygotic twins can be examined visually in the form of bar charts by clicking on the 'Show Expression Graph' button. The table 'Differentially expressed genes' displays information in the two categories 'Unrelated age and gender matched female individuals' and 'Monozygotic Twins'.

Repeat Table

The table '$(TG/CA)_n$ Repeats' provides information on $(TG/CA)_n$ repeats categorized into three types (I, II and III). The lengths and positions of repeats within the gene structure (exons and introns) and in 1 kb upstream flanking region are displayed. Because stretches of repeats separated by short intervals may act in concert to modulate transcription [33], the table also reports $(TG/CA)_n$ repeats within a range of 50 bp flanking each repeat. The table 'Polymorphic Repeats' provides information on the presence of known polymorphic repeats with their heterozygosity index and the number of alleles in CEPH families [26]. All this information can also be retrieved collectively for multiple genes by using the 'Advanced Query for Retrieving Data for Multiple Genes' option available on the Results page.

Information on other intragenic simple repeats can be probed in the table 'Other simple repeats' by specifying the repeat type, minimum cut-off length to score a repeat and clicking on the button 'Run SimRep'. The table on '*Alu* Repeats' provides information on the total content of intragenic *Alu* repeats. The details specifically for the young and active *Alu* Y repeats can be pulled out by clicking on the button '*Alu* Y'. Other related information on gene function and expression can be accessed using the links provided to the publicly available databases such as KEGG [34], GeneCards [35], GDB [36], UCSC Golden Path [37], Ensembl [38], PubMed [1], GXD [2], HuGEIndex [39] and GNF [3].

SimRep – An online application to identify simple repeats

'SimRep' is an online application to identify dinucleotide and other microsatellite repeats in a given nucleotide sequence including all available human gene sequences. Users can search either for a dinucleotide repeat by selecting it from the pull down menu or for a specific microsatellite repeat of their choice by entering the pattern to be searched. Patterns can be specified using the standard four base symbols A, T, G and C, as well as other symbols recommended by IUPAC (International Union of Pure and Applied Chemistry). The minimum length for scoring a repeat can be specified in the field 'Enter Cut-Off'. SimRep reports the length and location of the repeats or patterns in the given sequence in the form of a table. The positions of repeats either in forward strand (+) or in reverse strand (-) are reported with respect to the forward strand only as per the convention followed by genome sequence annotation groups. In the case of palindromic dinucleotide repeats such as GC, AT only one strand is reported.

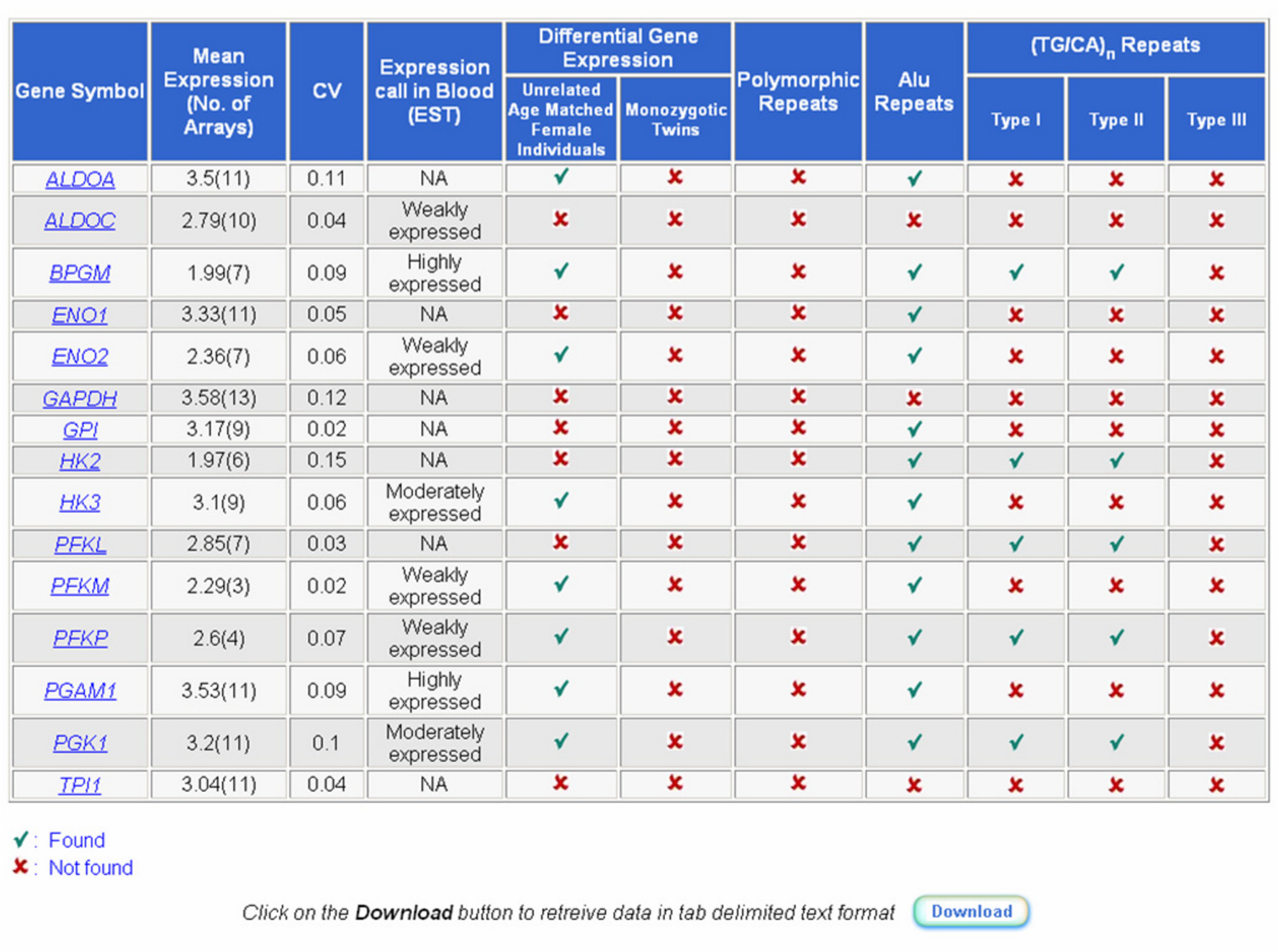
Examining gene expression and variability in Biochemical Pathways

With the present focus of biology shifting towards adopting a systemic approach to understand the complexity of human biology, biochemical pathways have become a focus of investigations. EXPOLDB offers this facility by providing information on gene expression and its variability in 134 biochemical pathways (from the KEGG and

GenMAPP databases) as organized by NetAffx [30]. Information on expression patterns including expression status, variability between monozygotic twins and between unrelated individuals (age and gender matched), repeats, polymorphic markers and functions can be queried for the genes involved in a defined pathway. For example, in the glycolytic pathway (Figure 2), none of the genes were differentially expressed in the monozygotic twin pairs, indicating that if expression variability is found for genes in this pathway it is likely not determined by environmental factors. Only 5 genes (*BPGM*, *HK2*, *PFKL*, *PFKP*, and *PGK1*) out of 15 genes (including isoforms) contained Type I and II repeats. 9 out of 15 genes (including isoforms) had low CVs (≤ 0.08) across all individuals and 8 were differentially expressed among unrelated individuals (age and gender matched). Among the 5 genes with Type I and II repeats, 3 (*BPGM*, *HK2*, *PGK1*) had CVs ≥ 0.8 and 3 (*BPGM*, *PFKP*, *PGK1*) were differentially expressed. These observations support the traditional practice of using the genes of the glycolytic pathway as 'controls' or as 'reference' genes in gene expression studies. Three genes *ALDOC*, *ENO1* and *GPI* showed no variation in expression between any two pairs of individuals including monozygotic twins, contained no repeats and had low CVs (<0.08) and therefore are devoid of potential factors that causes expression variation. If verified independently, these genes could be used as 'controls' or 'reference' genes in mRNA quantitation experiments.

Correlating genome wide expression with the incidence of $(TG/CA)_n$ repeats

Eukaryotic transcription is inherently complex and involves interaction of large numbers of proteins [40]. Therefore, examining the role of regulatory elements in gene expression regulation requires a set of genes with either a common organization of promoter, upstream ele-



Gene Symbol	Mean Expression (No. of Arrays)	CV	Expression call in Blood (EST)	Differential Gene Expression		Polymorphic Repeats	Alu Repeats	$(TG/CA)_n$ Repeats		
				Unrelated Age Matched Female Individuals	Monozygotic Twins			Type I	Type II	Type III
ALDOA	3.5(11)	0.11	NA	✓	✗	✗	✓	✗	✗	✗
ALDOC	2.79(10)	0.04	Weakly expressed	✗	✗	✗	✗	✗	✗	✗
BPGM	1.99(7)	0.09	Highly expressed	✓	✗	✗	✓	✓	✓	✗
ENO1	3.33(11)	0.05	NA	✗	✗	✗	✓	✗	✗	✗
ENO2	2.36(7)	0.06	Weakly expressed	✓	✗	✗	✓	✗	✗	✗
GAPDH	3.58(13)	0.12	NA	✗	✗	✗	✗	✗	✗	✗
GPI	3.17(9)	0.02	NA	✗	✗	✗	✓	✗	✗	✗
HK2	1.97(6)	0.15	NA	✗	✗	✗	✓	✓	✓	✗
HK3	3.1(9)	0.06	Moderately expressed	✓	✗	✗	✓	✗	✗	✗
PFKL	2.85(7)	0.03	NA	✗	✗	✗	✓	✓	✓	✗
PFKM	2.29(3)	0.02	Weakly expressed	✓	✗	✗	✓	✗	✗	✗
PFKP	2.6(4)	0.07	Weakly expressed	✓	✗	✗	✓	✓	✓	✗
PGAM1	3.53(11)	0.09	Highly expressed	✓	✗	✗	✓	✗	✗	✗
PGK1	3.2(11)	0.1	Moderately expressed	✓	✗	✗	✓	✓	✓	✗
TPI1	3.04(11)	0.04	NA	✗	✗	✗	✗	✗	✗	✗

✓ : Found
✗ : Not found

Click on the ***Download*** *button to retreive data in tab delimited text format*

Figure 2
A summary of the expression patterns and repeat distribution in genes of the glycolytic pathway.

ments, CpG islands etc., or with common expression profiles emanating from their clustered localization in coordinately regulated genomic regions [41]. Examples for gene sets conforming to these specifications are housekeeping genes, and genes belonging to the same family with similar architecture of regulatory elements such as the *RUNX* gene family [42].

Housekeeping genes

Housekeeping genes are expressed constitutively in all tissues to maintain cellular functions, and their expression pattern is less likely to be affected by variations in tissue specific factors, polymorphism in chromatin structure across different individuals, or experimental artifacts such as the number of different cell types in blood samples (if similar quantities of total RNA is taken) [39,41]. In the human housekeeping genes, the mean expression of genes without repeats (n < 6 units) was observed to be significantly higher than the mean expression of genes containing Type II and Type III $(TG/CA)_n$ repeats (*t-test*, df = 455, $P < 0.006$) suggesting the down modulatory role of these repeats to be in conformity with previous observations [9,14-23].

The RUNX family

Similar analysis was carried out with *RUNX* family. The mammalian RUNX genes comprise a small family of three genes *RUNX1*, *RUNX2* and *RUNX3* containing the 'runt domain' (RD), that act as master regulators of gene expression in major developmental pathways [42,43]. Sequence analysis suggests that *RUNX3* is the evolutionary founder of the mammalian *RUNX* family [43], and there exists extensive structural similarities between the three mammalian *RUNX* genes. Thus, the *RUNX* family provides a set of genes with similar architecture to investigate the effects of $(TG/CA)_n$ repeats in expression.

The records for *RUNX1* and *RUNX3* were retrieved (*RUNX2* was not found to be expressed in blood). The gene *RUNX1* has several Type II $(TG/CA)_n$ repeats: $(CA)_{17}$, $(CA)_{22}$, $(TG)_{12}$, $(TG)_{13}$, $(TG)_{14}$, $(TG)_{21}$, $(TG)_{23}$, $(TG)_{24}$ and $(CA)_{22}$ in introns and one interrupted $(TG)_7$-CG-$(TG)_9$ repeat in exon 8, whereas *RUNX3* is devoid of these repeats (n < 6 units). The mean expression of *RUNX3* is significantly higher than the mean expression of *RUNX1* (*t-test*, df = 13, $P < 0.0002$). These results corroborate with the common trend of down regulatory role of repeats as observed earlier [9,14-23].

Literature Resource

We have compiled a useful list of publications of several studies from the perspective of $(TG/CA)_n$ repeats as *cis* modulators of gene expression and other upcoming multifaceted roles of these repeats [23]. This list is likely to grow with the availability of more information and at present provides a useful wealth of information on the multifaceted roles of these repeats.

Discussion

Similar Databases

To our knowledge, EXPOLDB is the first systematic attempt to correlate gene expression and its variability with the presence and type of $(TG/CA)_n$ repeats. Other available databases focus singly on either gene expression or repetitive sequences.

Unique features of EXPOLDB

EXPOLDB is constructed for facilitating examination of the effect of repetitive elements in *cis* on expression variability. In addition, it allows distinguishing between genetic factors and other factors influencing expression levels (such as environment), by comparing expression between monozygotic twins. In particular, this data could be used as a sieve while identifying genes whose expression varies primarily due to genetic factors. Further, the variability in expression and the repeat content can be examined and correlated using either a gene centric or a pathway centric approach. Graphic display of expression values in the form of bar charts aids visual comparisons and stimulates novel questions. The tool SimRep can be used to identify other dinucleotide and user specified simple repeats in recent build (build 35) of human genes sequences and in a given nucleotide sequence.

Limitations

At present, as per global status, there is limited data on monozygotic twins and on the polymorphic status of several repeats and on gene expression data from different populations. Our efforts in constructing EXPOLDB are likely to stimulate and facilitate investigations on this aspect of variability in gene expression. We envisage that the emerging role of $(TG/CA)_n$ repeats as 'functional elements' [4-23] and the global efforts on generating expression data are likely to result in the growth and use of this database.

Conclusion

We envision that our effort to organize the gene expression data and the variability contained in it from the perspective of simple repeats as *cis* regulators of transcription could enhance other efforts in this subject and could serve as a seed database by offering genome wide expression data with facility to correlate genetic information for Systems Biology projects.

Availability and requirements

EXPOLDB is accessible at http://expoldb.igib.res.in/expol requires Explorer version 5.5 or above, FireFox version 1.0.4 or above.

List of abbreviations

EXPOLDB: Expression Linked Polymorphism database, CV: Coefficient of Variation, SD: Standard deviation, HGNC: HUGO Gene Nomenclature Committee

Authors' contributions

This project symbolizes Indo-Israel friendship with mutually benefiting interests. All authors have contributed together towards this goal. VKS carried out a major part of the work including writing of computer programs, planning artistic GUI, downloading data from NCBI, analysis of microarray and sequence data and wrote the manuscript, AS helped in carrying out the microarray experiments and analysis of microarray data, NK and MK helped in software coding web enablement, SHS helped in carrying out the microarray experiments, LSA and DL offered constructive scientific criticisms with focus on providing benefits from user view point, SKB provided scientific suggestions particularly on twins during the work. SR is the group leader, working in many arms of the project including experiments, provided scientific suggestions and criticisms for improving the database, guided in statistical analysis, conforming to ethical principles, critical examination, presentation and manuscript preparation.

Acknowledgements

VKS and AS are recipients of Senior Research Fellowships from CSIR, India. We thank Dipayan Das Gupta and Sushant K. Das Sutar for assistance in program writing, the Department of Biotechnology, Govt. of India and the Ministry of Science, Israel for grants under the Indo-Israel cooperation, and the CSIR for a grant under the Task force 'In Silico biology for drug target development' (CMM0017).

References

1. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Helmberg W, Kapustin Y, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2006:D173-D180.
2. Hill DP, Begley DA, Finger JH, Hayamizu TF, McCright IJ, Smith CM, Beal JS, Corbani LE, Blake JA, Eppig JT, Kadin JA, Richardson JE, Ringwald M: **The mouse Gene Expression Database (GXD): updates and enhancements.** *Nucleic Acids Res* 2004:D568-D571.
3. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci USA* 2002, **99(7):**4465-4470.
4. Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J: **A comprehensive genetic map of the human genome based on 5,264 microsatellites.** *Nature* 1996, **380(6570):**152-154.
5. Brahmachari SK, Meera G, Sarkar PS, Balagurumoorthy P, Tripathi J, Raghavan S, Shaligram U, Pataskar S: **Simple repetitive sequences in the genome: structure and functional significance.** *Electrophoresis* 1995, **16(9):**1705-1714.
6. Ellegren H: **Microsatellites: simple sequences with complex evolution.** *Nat Rev Genet* 2004, **5(6):**435-445.
7. Haniford DB, Pulleyblank DE: **The in-vivo occurrence of Z DNA.** *J Biomol Struct Dyn* 1983, **1(3):**593-609.
8. Nordheim A, Rich A: **The sequence (dC-dA)n X (dG-dT)n forms left-handed Z-DNA in negatively supercoiled plasmids.** *Proc Natl Acad Sci USA* 1983, **80(7):**1821-1825.
9. Meera G, Ramesh N, Brahmachari SK: **Zintrons in rat alpha-lactalbumin gene.** *FEBS Lett* 1989, **251(1–2):**245-249.
10. Peck LJ, Wang JC: **Transcriptional block caused by a negative supercoiling induced structural change in an alternating CG sequence.** *Cell* 1985, **40(1):**129-137.
11. Rothenburg S, Koch-Nolte F, Rich A, Haag F: **A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity.** *Proc Natl Acad Sci USA* 2001, **98(16):**8985-8990.
12. Majewski J, Ott J: **GT Repeats are associated with recombination on human chromosome 22.** *Genome Res* 2000, **10:**1108-1114.
13. Hui J, Stangl K, Lane WS, Bindereif A: **HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats.** *Nat Struct Biol* 2003, **10(1):**33-37.
14. Sharma VK, Brahmachari SK, Ramachandran S: **(TG/CA)n repeats in human gene families: abundance and selective patterns of distribution according to function and gene length.** *BMC Genomics* 2005, **6(1):**83.
15. Rockman MV, Wray GA: **Abundant raw material for *Cis*-regulatory evolution in humans.** *Mol Biol Evol* 2002, **19:**1991-2004.
16. Agarwal AK, Giacchetti G, Lavery G, Nikkila H, Palermo M, Ricketts M, McTernan C, Bianchi G, Manunta P, Strazzullo P, Mantero F, White PC, Stewart PM: **CA-Repeat polymorphism in intron 1 of *HSD11B2* : effects on gene expression and salt sensitivity.** *Hypertension* 2000, **36:**187-194.
17. Naylor LH, Clark EM: **d(TG)n.d(CA)n sequences upstream of the rat prolactin gene form Z-DNA and inhibit gene transcription.** *Nucleic Acids Res* 1990, **18(6):**1595-1601.
18. Shimajiri S, Arima N, Tanimoto A, Murata Y, Hamada T, Wang KY, Sasaguri Y: **Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene.** *FEBS Lett* 1999, **455(1–2):**70-74.
19. Pravica V, Asderakis A, Perrey C, Hajeer A, Sinnott PJ, Hutchinson IV: **In vitro production of IFN-gamma correlates with CA repeat polymorphism in the human *IFN-gamma* gene.** *Eur J Immunogenet* 1999, **26:**1-3.
20. Gebhardt F, Zanker KS, Brandt B: **Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1.** *J Biol Chem* 1999, **274:**13176-13180.
21. Streelman JT, Kocher TD: **Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia.** *Physiol Genomics* 2002, **9(1):**1-4.
22. Sharma VK, B-Rao C, Sharma A, Brahmachari SK, Ramachandran S: **(TG:CA)(n) repeats in human housekeeping genes.** *J Biomol Struct Dyn* 2003, **21(2):**303-310.
23. Sharma VK, Sharma A, Kumar N, Khandelwal M, Mandapati KK, Horn-Saban S, Strichman-Almashanu L, Lancet D, Brahmachari SK, Ramachandran S: **EXPOLDB Literature Link.** [http://expoldb.igib.res.in/expol/literaturelinks.html].
24. Epplen JT, Kyas A, Maueler W: **Genomic simple repetitive DNAs are targets for differential binding of nuclear proteins.** *FEBS Lett* 1996, **389(1):**92-95.
25. Sharma A, Sharma VK, Horn-Saban S, Lancet D, Ramachandran S, Brahmachari SK: **Assessing natural variations in gene expression in humans by comparing with monozygotic twins.** *Physiol Genomics* 2005, **21(1):**117-123.
26. The Foundation Jean Dausset-Centre d'Etude du Polymorphisme Humain (CEPH): **CEPH database.** 1984 [ftp://ftp.cephb.fr/ceph_genotype_db/ceph_db/Ver_9/mkr/].
27. Bedell JA, Korf I, Gish W: **MaskerAid: a performace enhancement to RepeatMasker.** *Bioinformatics* 2000, **16(11):**1040-1041.
28. Bortoluzzi S, d'Alessi F, Romualdi C, Danieli GA: **The human adult skeletal muscle transcriptional profile reconstructed by a novel computational approach.** *Genome Res* 2000, **10:**344-349.
29. Eisenberg E, Levanon EY: **Human housekeeping genes are compact.** *Trends Genet* 2003, **19(7):**362-365.
30. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose MA: **NetAffx: Affymetrix probesets and annotations.** *Nucleic Acids Res* 2003, **31(1):**82-86.
31. Pezzullo JC: [http://statpages.org/]. Departments of Pharmacology and Biostatistics at Georgetown University, in Washington, DC, USA

32. Sharma VK, Sharma A, Kumar N, Khandelwal M, Mandapati KK, Horn-Saban S, Strichman-Almashanu L, Lancet D, Brahmachari SK, Ramachandran S: **Most constant housekeeping genes.** [http://expoldb.igib.res.in/expol/mostconstantgenes.html].
33. Agarwal AK, White PC: **Structure of the VPATPD Gene Encoding Subunit D of the Human Vacuolar Proton ATPase.** *Biochem Biophys Res Commun* 2000, **279:**543-547.
34. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006:D354-D357.
35. Safran M, Chalifa-Caspi V, Shmueli O, Olender T, Lapidot M, Rosen N, Shmoish M, Peter Y, Glusman G, Feldmesser E, Adato A, Peter I, Khen M, Atarot T, Groner Y, Lancet D: **Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE.** *Nucleic Acids Res* 2003, **31(1):**142-146.
36. Cuticchia AJ: **Future vision of the GDB human genome database.** *Hum Mutat* 2000, **15(1):**62-67.
37. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ: **The UCSC Genome Browser Database: update 2006.** *Nucleic Acids Res* 2006:D590-598.
38. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Flicek P, Graf S, Hammond M, Herrero J, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Kokocinski F, Kulesha E, London D, Longden I, Melsopp C, Meidl P, Overduin B, Parker A, Proctor G, Prlic A, Rae M, Rios D, Redmond S, Schuster M, Sealy I, Searle S, Severin J, Slater G, Smedley D, Smith J, Stabenau A, Stalker J, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Hubbard TJ: **Ensembl 2006.** *Nucleic Acids Res* 2006:D556-561.
39. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, Weng Z, Mutter GL, Frosch MP, Macdonald ME, Milford EL, Crum CP, Bueno R, Pratt RE, Mahadevappa M, Warrington JA, Stephanopoulos G, Stephanopoulos G, Gullans SR: **A compendium of gene expression in normal human tissues.** *Physiol Genomics* 2001, **7:**97-104.
40. Coulson RM, Ouzounis CA: **The phylogenetic diversity of eukaryotic transcription.** *Nucleic Acids Res* 2003, **31(2):**653-660.
41. Lercher MJ, Urrutia AO, Hurst LD: **Clustering of housekeeping genes provides a unified model of gene order in the human genome.** *Nat Genet* 2002, **31(2):**180-183.
42. Levanon D, Glusman G, Bettoun D, Ben-Asher E, Negreanu V, Bernstein Y, Harris-Cerruti C, Brenner O, Eilam R, Lotem J, Fainaru O, Goldenberg D, Pozner A, Woolf E, Xiao C, Yarmus M, Groner Y: **Phylogenesis and regulated expression of the RUNT domain transcription factors *RUNX1* and *RUNX3*.** *Blood Cells Mol Dis* 2003, **30:**161-163.
43. Bangsow C, Rubins N, Glusman G, Bernstein Y, Negreanu V, Goldenberg D, Lotem J, Ben-Asher E, Lancet D, Levanon D, Groner Y: **The *RUNX3* gene – sequence, structure and regulated expression.** *Gene* 2001, **279:**221-232.
44. Bustin SA: **Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays.** *J Mol Endocrinol* 2000, **25:**169-193.