Nucleic Acids Research, 2006, Vol. 34, No. 9 ***2663–2675***
doi:10.1093/nar/gkl354

Long homopurine•homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region

Albino Bacolla, Jack R. Collins[1], Bert Gold[2], Nadia Chuzhanova[3,4], Ming Yi[1], Robert M. Stephens[1], Stefan Stefanov[2], Adam Olsh[2], John P. Jakupciak[5], Michael Dean[2], Richard A. Lempicki[6], David N. Cooper[4] and Robert D. Wells*

Institute of Biosciences and Technology, Center for Genome Research, Texas A&M University System Health Science Center, Texas Medical Center, 2121 West Holcombe Blvd, Houston, TX 77030, USA, [1]Advanced Biomedical Computing Center and [2]Laboratory of Genomic Diversity, NCI-Frederick, Frederick, MD 21702, USA, [3]Biostatistics and Bioinformatics Unit, Cardiff University, Cardiff CF14 4XN, UK, [4]Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK, [5]National Institute of Standards and Technology, DNA Technologies Group, Biotechnology Division, Gaithersburg, MD 20899, USA and [6]Laboratory of Immunopathogenesis and Bioinformatics, SAIC-Frederick, Inc., Frederick, MD 21702, USA

Received February 23, 2006; Revised March 13, 2006; Accepted April 20, 2006

ABSTRACT

Homo(purine•pyrimidine) sequences (R•Y tracts) with mirror repeat symmetries form stable triplexes that block replication and transcription and promote genetic rearrangements. A systematic search was conducted to map the location of the longest R•Y tracts in the human genome in order to assess their potential function(s). The 814 R•Y tracts with ≥250 uninterrupted base pairs were preferentially clustered in the pseudoautosomal region of the sex chromosomes and located in the introns of 228 annotated genes whose protein products were associated with functions at the cell membrane. These genes were highly expressed in the brain and particularly in genes associated with susceptibility to mental disorders, such as schizophrenia. The set of 1957 genes harboring the 2886 R•Y tracts with ≥100 uninterrupted base pairs was additionally enriched in proteins associated with phosphorylation, signal transduction, development and morphogenesis. Comparisons of the ≥250 bp R•Y tracts in the mouse and chimpanzee genomes indicated that these sequences have mutated faster than the surrounding regions and are longer in humans than in chimpanzees. These results support a role for long R•Y tracts in promoting recombination and genome diversity during evolution through destabilization of chromosomal DNA, thereby inducing repair and mutation.

INTRODUCTION

Chromosomal DNA exists principally as a right-handed double helix (B-DNA). However, other conformations, such as triplexes, tetraplexes, slipped structures with hairpin loops, left-handed Z-DNA and cruciforms are also known [reviewed in (1–7)]. These alternative (non-B DNA) conformations are formed at specific sequence motifs and are therefore possible only at discrete chromosomal locations. More than 15 genomic disorders, including neurofibromatosis type I, chronic myeloid leukemia, spermatogenetic failure and recurrent constitutional translocations, have recently been associated with rearrangements mediated by recombination between blocks of repetitive DNA (from a few hundred base pairs to several hundred kilobase pairs in length) almost exclusively composed of direct repeats (DR) and inverted repeats (IR) [reviewed in (8)]. Since double-strand breaks (DSBs) are localized at hotspots within these blocks in most cases, factors other than the primary DNA sequence must be involved, and indeed the locations of the rearrangement fusions are generally found at sequences known to adopt non-B DNA conformations (8–16). This conclusion is further supported by the finding that translocation frequencies correlate with genetic variations in the general population affecting the stability of the putative non-B DNA conformations at breakpoints (16).

*To whom correspondence should be addressed. Tel: +1 713 677 7651; Fax: +1 713 677 7689; Email: rwells@ibt.tamhsc.edu

Triplex DNA requires homo(purine•pyrimidine) sequences (R•Y tracts) and is stabilized by Hoogsteen hydrogen bonds between the purines in the Watson–Crick duplex and a third strand in the major groove, which may be composed of either pyrimidines bound in the parallel orientation (YRY triplexes) or purines (RRY triplexes) bound in the antiparallel orientation (1,3,6,17–19). Specific interactions consist of T-A•T and C^+-G•C triads for YRY triplexes, and G-G•C and A-A•T triads for RRY triplexes; hence, mirror repeat symmetries within R•Y tracts yield fully paired and stable triplexes. The YRY triplexes are additionally stabilized by cytosine protonation at N3, a process that requires low pH in nucleotides, but which occurs cooperatively and at neutral pH in clustered cytosines in a long polynucleotide chain (20–22). Finally, the third strand may be provided by the folding back of a single R•Y tract with mirror repeat symmetry (intramolecular triplex), by the interaction between two R•Y tracts separated by some distance on the same or on two different DNA molecules (intermolecular triplex), or by a single-stranded oligonucleotide composed of either DNA or RNA (6,19,23).

R•Y tracts are genetically unstable in experimental model systems. Whereas numerous biological functions have been attributed to triplexes, including the blockage of DNA replication (24–26) and the interference with transcription (26,27), recent studies have provided revealing insights. First, studies conducted in live mice, mammalian cell cultures and *Escherichia coli* indicate that R•Y tracts are highly mutagenic (28–30), induce DSBs (30,31) and are frequently found at the breakpoint sites of gross deletions and other rearrangements (9,31). Indeed, three types of biochemical and genetic studies have shown that these genomic instabilities were due to the non-B conformations adopted by the R•Y tracts and not to the tracts in their orthodox right-handed B-form (32). Second, long GAA•TTC repeats in the first intron of the frataxin (*FXN*) gene adopt several structures including triplex and sticky DNA, which inhibit transcription of the gene thus reducing the expression of frataxin (26,33,34). Hence, triplexes may be involved in the etiology of Friedreich's ataxia. Third, triplexes may play critical roles in the *bcl-2* major breakpoint region with respect to the RAG-dependent t(14;18) translocation associated with follicular lymphomas (35–37). Fourth, R•Y tracts elicit a biological response in the context of mammalian chromatin also in the absence of perfect mirror repeat symmetry and at moderate-to-short lengths (~20 bp or less) (35–38), indicating that the energetic barrier associated with the duplex-to-triplex structural transitions may be easily overcome.

Long (several hundred base pairs) R•Y tracts in the human genome have been known for >20 years (6,17,39,40), and previous limited studies suggested their abundance in mammalian genomes (41–43). To date, no genome-wide queries, which might be informative with respect to the potential function(s) of these sequences, have been reported. Knowledge about the size and location of non-B DNA conformations in vertebrate genomes would be expected to give critical clues as to their biological functions. The human genome has so far only been surveyed for IR sequences (44), which can form cruciforms. Warburton *et al*. (44) reported that the large IR may have a role in testis gene expression and genome integrity.

Herein, we have applied a data-mining approach to conduct the first systematic search for the longest uninterrupted R•Y tracts in the human genome and tested specific hypotheses by employing experimental methodologies *in silico*. We show that these sequences cluster specifically in the pseudoautosomal region of the sex chromosomes and are found predominantly in genes that are highly expressed in the brain. These determinations, along with comparative analyses of the mouse and chimpanzee genomes, indicate that long R•Y tracts constitute mutational hotspots and are likely to have played a key role in genome plasticity and evolution.

MATERIALS AND METHODS

Computer searches

All R•Y tracts were found using the program PTRfinder (45). The algorithm first maps the DNA sequence to R or Y for purine and pyrimidine, respectively. Then the program locates tandem repeats with this reduced alphabet and given the minimum lengths input by the user. The results were then imported into the Genomic Resource Information Database (GRID, http://grid.abcc.ncifcrf.gov/) (J. R. Collins, R. M. Stephens and J. Shan, manuscript in preparation) for web access and for generating queries to correlate R•Y tracts with gene location and overlap. Genomic build parameters describing tracts of pure R or Y were extracted from the GRID database using the dataset from the Human Genome Browser at the University of California, Santa Cruz (UCSC), which was based on the hg17, NCBI Build 35, May 2004 assembly. The following SQL query was used to retrieve all tracts of type R or Y with length ⩾250 bp: 'SELECT * from 'pupy' WHERE 'len'>=250 AND length('type')=1 ORDER BY 'chrom', 'len''. This query retrieved 818 records, 814 with fully determined coordinates, which were imported into MSExcel spreadsheets. An additional table column with 814 links to the UCSC Human Genome Browser was created using a URL pattern to link each record to the exact genomic position. By opening the links, each record was visually checked for being within or between gene sequences. Thus, 241 of the 814 tracts were found to occur within annotated genes. Of the 228 non-redundant R•Y tract-containing genes, information was gathered for 156 genes, either through OMIM or PubMed (http://www.ncbi.nlm.nih.gov).

In order to compare the 814 tracts and their surrounding sequences in human with their orthologous counterparts in chimpanzee (*Pan troglodytes*), five PERL (Practical Extraction and Reporting Language) scripts were written. A standalone BLAT (Blast-like Alignment Tool) program was downloaded from http://www.soe.ucsc.edu/~kent/src/ and installed on a computer with Linux FC3. The human and chimpanzee (NCBI Build 1 version 1, November 2003) genomes were downloaded from the UCSC website. In addition, the free package CLUSTAL W 1.83 and the EMBOSS application DOTMATCHER (Ian Longden, Sanger Institute, Cambridge, UK), which is under GNU license (http://www.gnu.org), were downloaded and installed.

The first PERL script '1_queryTOfasta.pl' processed the 814-record table containing the repeat data. Coordinates were transformed so as to include 2500 bp of flanking sequences from both sides of each tract. For each record, the respective DNA sequence was extracted from the local human genome files and added to a file in FASTA format containing all records. A second PERL script '2_saBLATfastaVSchromosomes.pl' automated the BLAST search of every record against the local chimpanzee genome. To speed up this process, a previously made chimpanzee '11.ooc' R or Y tracts file was used (for details, see BLAT FAQs and tutorials at the website previously cited). The output was a directory with BLAT-hits results in chimpanzee in PLSX table format, one file for each record. A third PERL script 'selectFROMfile.pl' processed the PLSX-files directory picking up the best score from each file and merging all the best-scores into a single SELECT file, which also contained 814 records. A fourth PERL script '4_plsxTOfastaCHR.pl' processed the 814-record SELECT table in the same way that the first script processed the GRID-database table. It extracted from the local chimpanzee genome each corresponding segment of DNA sequence and added them to a FASTA file. The product of running the first through fourth scripts are two multiple entry DNA FASTA files; one with human queries and the other with chimpanzee matches. A fifth script '5_html_u.pl' compared all query and match sequences from the two files above. It also ran CLUSTAL W and DOTMATCHER for each of the pairs of records. Subsequently, it created an HTML document presenting the results. For ease in troubleshooting, the job was performed in several steps using separated scripts, even though it could readily be integrated into one. The results for the R•Y tracts present within genes have been posted at the following website: http://home.ncifcrf.gov/ccr/lgd/dean_lab/pure_R_or_Y_Comparison.

Similarly, 5 kb human DNA fragments containing the R•Y tracts centrally were used as queries in BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat) searches on the mouse (*Mus musculus*, Mm5, NCBI build 33, May 2004), macaque (*Macaca mulatta*, NCBI, Mmul_0.1, January 2005) and dog (*Canis familiaris*, NCBI, v1.0, July 2004) genomes. Genomic interspersed repeats were identified by RepeatMasker (http://www.repeatmasker.org).

Tissue expression of R•Y tract-containing genes

Tissue gene expression data were downloaded from the Genomic Institute of the Novartis Research Foundation (GNF, http://web.gnf.org/index.shtml), and comprised the Affymetrix U133A chip (~22 000 probe sets) plus a custom GNF Affymetrix chip with ~11 000 probe sets analyzed on 79 human tissues and cell lines. Approximately 17 000 known human genes were mapped to these probe sets and their relative expression levels were analyzed in the 79 tissue types to identify genes preferentially expressed in each tissue. Preferential expressed genes are defined as follows: the expression levels of a given gene across all tissues were transformed to a *Z*-score and genes with a *Z*-score > 1 (i.e. >1 SD above the mean) in a given tissue were classified as highly expressed in that tissue. The *Z*-score was averaged when more than one probe set hybridized to the same gene and for duplicate tissues. A total of ~1 390 000 *Z*-score values for all genes and all tissues were obtained, ~10% of which qualified as highly expressed. *P*-values were obtained by comparing the fraction of the ~17 000 genes highly expressed in a given tissue with the fraction observed for the set of genes containing either the pure R•Y tracts ≥250 bp or the pure R•Y tracts ≥100 bp in length, assuming a binomial distribution. *P*-values were corrected (Bonferroni) for multiple comparison of 79 tissues.

Functional category enrichment analysis and creation of gene-term association networks (GTANs)

Functional category enrichment analyses were performed using a software tool, WholePathwayScope (WPS) (46), developed at the ABCC (NCI-Frederick, MD). Briefly, this analysis is based on Fisher Exact Test for 2 × 2 contingency tables (gene list versus functional category) to estimate and rank the statistical significance of the enrichment of functional categories within a given system (GO terms, BioCarta pathways, KEGG pathways, gene-disease associations, protein interaction partners and protein families) for genes harboring ≥250 pure R•Y tracts (RY250) or ≥100 pure R•Y tracts (RY100). The databases used, which were collected in the WPS database, were as follows: the GO terms and gene-GO term association tables were downloaded from the Gene Ontology Consortium (http://www.geneontology.org); the BioCarta pathways were kindly provided by the CGAP group (Cancer Genome Anatomy Project: http://cgap.nci.nih.gov/), which originated from the Bio Carta pathway collections (http://www.biocarta.com/genes/allPathways.asp); the KEGG pathways were downloaded from the KEGG pathway collections (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.ad.jp/kegg); the gene disease associations were downloaded from the Genetic Association Database (http://geneticassociationdb.nih.gov/); information on Protein Interaction Partners and Protein families (Pfam) information was kindly provided by the DAVID (Database for Annotation, Visualization and Integrated Discovery) group (http://david.niaid.nih.gov). Gene-Term Association Networks (GTANs) were created within WPS, such that for any given gene in a gene list (RY100 or RY250), its associated term(s) was sought in the WPS database. The pairwise gene-term relationships were represented as a graphical layout of a Gene-Term Association Network within WPS, in which any gene and its associated term were linked with an edge to indicate the gene-term association relationship.

Gene size

The sizes for annotated genes were obtained from the lengths of pre-mRNA transcripts, which were downloaded from the National Center for Biotechnology Information (NCBI) website (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/RNA/). The set of 1377 genes used to determine the size distribution of genes expressed in brain was obtained from the tissue distribution analysis, with genes with a *Z*-score > 1 in brain tissues being selected. In most cases, the SigmaPlot 2002 program, version 8.02 (SPSS, Chicago, IL) was used to represent the results graphically and to determine the best fit to the data.

Clustering of R•Y tracts

The order statistics, *r*-scans, as described in Karlin and Macken (47), were used to detect significant clustering of the R•Y tracts observed along each chromosome, by comparing their distribution with that of a uniform Poisson distribution. We assume that $\{X_i\}$ are the distances between adjacent R•Y tracts (n tracts in total) along an individual chromosome that are not necessarily independently and identically distributed (*iid*). Note that all points are mapped into a [0,1] interval. Denote by

$$Y_i^{(r)} = \sum_{j=1}^{i+r-1} X_j, \quad i = 1, \ldots, n-r+1, r \ll n$$

the distance from the *j*-th R•Y tract to its *r*-th nearest neighbor. There is an order statistic associated with the sequence of partial sums

$$Y_1^{(r)}, Y_2^{(r)}, \ldots, Y_{n-r+1}^{(r)}$$

and a minimum defined as

$$m^{(r)} = Y_1^* = \min\{Y_1^{(r)}, Y_2^{(r)}, \ldots, Y_{n-r+1}^{(r)}\}$$

If we assume that distances $\{X_i\}$ are *iid*, then according to Karlin and Macken (47),

$$\Pr\left\{m^{(r)} < \frac{x}{n^{1+(1/r)}}\right\} \approx 1 - \exp(-\lambda), \quad \lambda = \frac{x^r}{r!}, r \geqslant 1$$

We declare significant clustering when the observed minimum for the limit distribution has <0.01 (or <0.05) probability of occurring by chance and therefore

$$m^{(r)} < \frac{1}{n}\left\{\frac{r!(-\ln 0.99)}{n}\right\}^{1/r}$$

Several R•Y tract clusters were detected (after correcting for adjacent tracts separated by single nucleotide interruptions) by testing the first minimum of *r*-scans (where $r = 1, \ldots, 16$).

Complexity analysis

Complexity analysis, as devised by Gusev *et al.* (48), was employed in a search for DR, IR and mirror repeats (MR) in the pseudoautosomal region (PAR1) and the downstream 4.4 Mb region (After_PAR), in the 5 kb sequences flanking the R•Y tracts in the *PRKCB1* and *ADAM18* genes, in the 3′-untranslated region (3′-UTR) of the *DISC1* gene and in the comparative analyses of the R•Y tracts in the mammalian *TIAM1* gene.

RESULTS

R•Y tracts within genes

The human genome was screened for the largest R•Y tracts with the expectation that these could shed light on their potential biological role(s). The total number of tracts equaling or exceeding an arbitrary length cutoff of 250 pure R (adenine, A and guanine, G) or Y (cytosine, C and thymine, T) bases was 814. Approximately 30% (241/814) were located within annotated genes (228, some genes contained more than one tract, Supplementary Table 1), all within introns. The longest tract (1303 bp) was present in the *CENTA1* gene on chromosome 7, which was the only pure intragenic sequence to exceed 1 kb. Tracts were then analyzed to determine the R•Y tract length if occasional interruptions (mostly single nt changes) were ignored. The 2.5 kb *PKD1* R•Y tract (49), which has been instrumental in the characterization of the mutational properties of R•Y sequences (9,25,32), was the third longest, preceded only by a 3.1 kb tract in the *DKFZp434G0625* gene and a tract of nearly 4 kb in the *LOC348094* gene (both encoding products of unknown function). Hence, the *PKD1* gene possesses the longest gene-associated R•Y tract in the human genome for genes of known function. A double logarithmic plot of these tracts arranged in order of size exhibited a linear relationship ($y = 3.499 - 0.458x$, $r^2 = 0.99$, data not shown), indicating an exponential decay distribution. This distribution suggests that selection against or in favor of any R•Y tract length is not exercised to a detectable level within the existing envelope.

Complexity analysis was employed to identify MR separated by any distance within these R•Y tracts. The average fractions of non-redundant MR elements (with no interruptions) per R•Y tract were 16, 4 and 2 for MR lengths of 10, 20 and 30 bp, respectively. In addition, all R•Y tracts had at least one 10 bp-long MR element. Hence, based on the work conducted on the *PKD1* and other R•Y tracts *in vivo* (9,28,30,32,35,37,50,51), we conclude that most, if not all, R•Y tracts with ⩾250 uninterrupted base pairs have the potential to form intramolecular and/or intermolecular triplexes *in vivo*. However, future work will be required to experimentally evaluate triplex formation by these R•Y tracts.

To determine whether there was preferential association within certain categories of genes, we compared the set of R•Y tract-containing genes with that of the human (reference) dataset in proteomic databases (Supplementary Table 2). For the Gene Ontology (GO) Molecular Function, eight terms exceeded a *P*-value of 10^{-4} and included seven terms for channel activities and one term for glutamate receptor activity, consistent with strong enrichment for genes encoding transmembrane proteins in the brain (Supplementary Table 2A). The GO Biological Process analysis revealed 13 terms which exceeded *P*-values of 10^{-4}, and included three terms for cell adhesion and cell communication, four for neuronal function and three for ion transport (Supplementary Table 2B), indicative of a preferential distribution within genes involved in specialized functions at the cell membrane. The GO Cellular Component analysis (Supplementary Table 2C) yielded four terms that exceeded *P*-values of 10^{-4} and which were associated with localization to the cell membrane.

The fibronectin type III, C-cadherin and α-catenin domains, present in many cell surface receptors and cell adhesion molecules, represented the most highly enriched categories (*P*-values $\sim 10^{-4}$) in the Protein Families (Pfam) and Protein Information Resource (PIR) databases (Supplementary Tables 2D and 2E), whereas eight terms including the large neuroactive ligand interaction pathway, which comprises an array of neuronal receptors implicated in neuropeptide and small molecule signaling pathways, were found to be enriched in the more limited BioCarta and Kyoto Encyclopedia of Genes and Genomes databases (Supplementary

Tables 2F and 2G). Analysis of the Protein Interaction Partners database (Supplementary Table 2H) indicated 34/91 enrichments led by DLG4 (psd-95, $P \sim 8 \times 10^{-4}$), a component of the post-synaptic density structure involved in receptor clustering. Finally, analysis of the Disease Association database indicated that the most significant enrichment ($P \sim 5 \times 10^{-3}$) was for candidate genes for schizophrenia susceptibility (Supplementary Table 2I). In summary, the longest R•Y tracts were non-randomly distributed, and were disproportionately associated with genes whose products are localized to the plasma membrane and which perform cell communication and transport functions.

Distinct gene categories

The ⩾250 bp R•Y tract-containing genes represented ~1% of the annotated human gene dataset. We therefore sought to determine whether lowering the R•Y length threshold would yield a larger set of genes with the same non-random distribution profile. A search for genes with known functions containing pure R•Y tracts ⩾100 bp in length yielded a total of 2886 hits and a non-redundant gene set of 1957. Comparison of the distribution of these R•Y tract-containing genes with that of the reference dataset (Supplementary Table 3) revealed that the terms most highly enriched were common to those identified for the genes containing the ⩾250 R•Y tracts (Table 1). This correlation was particularly striking for gene products involved in signal transduction pathways at synapses, which also showed strong associations with susceptibility to schizophrenia (Supplementary Figures 1 and 2, Supplementary Table 4 and Supplementary Text).

Since P-values are sensitive to sample size, we next determined whether the terms were more enriched in ⩾250 or ⩾100 bp R•Y tract-containing genes based on 'fold-enrichment'. Consistently greater enrichments were noted for the ⩾250 bp R•Y tract-containing genes (Table 1). Hence, as the R•Y tract length increases, so does the probability of their association with genes involved in specific functions at the plasma membrane.

A second set of terms was found to be highly enriched in the ⩾100 bp R•Y tract-containing genes but not in the ⩾250 bp R•Y tract-containing genes. These included terms containing transferase and kinase activities (GO_MF) and protein/receptor phosphorylation of genes implicated in development and morphogenesis (GO_BP), implying an enrichment in genes involved in signal transduction pathways (Supplementary Figure 3 and Supplementary Text).

Hence, we conclude that the longest R•Y tracts in the human genome are distributed between two main pools of genes in a length-dependent fashion. The first group, comprising shorter length tracts (100 bp ⩽ R•Y ⩽ 250 bp), co-localized preferentially with genes involved in signal transduction pathways associated with development, whereas a second group, comprising longer tracts (R•Y ⩾ 250 bp), co-localized with genes mostly involved in ion transport, cell adhesion and neurogenesis. Since several terms exceeded

Table 1. Major enriched categories for genes with ⩾ 250 bp and ⩾ 100 bp R•Y

Combined Rank	Category (term—pathway)	P-value		Fold enrichment	
		⩾250 R•Y	⩾100 R•Y	⩾250 R•Y	⩾100 R•Y
GO molecuar function					
10	Ion channel activity	1.95E−05	5.92E−09	4.4	2.2
12	Protein binding	3.14E−03	6.25E−15	1.7	1.6
20	Glutamate receptor activity	6.11E−04	1.92E−07	10.3	4.3
GO biological process					
5	Cell adhesion	1.11E−04	3.36E−12	3.4	2.2
7	Cell communication	2.19E−04	5.24E−15	1.6	1.4
15	Transmission of nerve impulse	1.83E−04	5.24E−08	5.1	2.5
GO cellular component					
2	Membrane	2.78E−06	2.46E−14	1.6	1.3
15	Synapse	2.18E−02	7.69E−05	5.0	2.8
19	Extracellular matrix	4.96E−02	1.33E−08	2.3	2.2
Pfam family					
2	Fibronectin type III	2.84E−05	1.43E−12	5.1	2.8
15	C-cadherin	7.76E−05	3.14E−05	17.0	4.3
17	C2	1.17E−04	3.95E−05	5.4	2.1
PIR family					
3	α-catenin	3.55E−04	1.20E−03	60.4	9.4
11	Cadherin	1.81E−02	4.04E−04	9.5	4.0
KEGG pathway					
5	Neuroactive ligand–receptor interaction	7.06E−03	2.52E−03	2.9	1.5
6	Phosphatidylinositol signaling system	2.30E−02	1.87E−05	4.8	2.6
7	Cholera—infection	1.26E−02	3.39E−03	6.0	2.2
Protein interaction partners					
22	DLG4	7.98E−04	2.06E−03	45.1	6.1
23	RAC1	8.34E−03	9.78E−04	7.1	2.4
Disease association					
4	Schizophrenia	5.23E−03	2.33E−03	7.5	2.6

Combined Rank, sum of the two categories ranks from Supplementary Tables 2 and 3; categories with one entry were excluded; when more than one category with largely redundant gene entries was present, only one was chosen.

P-values of 10^{-10} and the different databases (particularly the large GO_MF, GO_BP, GO_CC, and Pfam Family) were in agreement in identifying terms that contained the same genes, the conclusions are strongly supported by the data.

Tissue gene expression

To determine whether the ⩾250 bp R•Y tract-containing genes (250-set) and the ⩾100 bp R•Y tract-containing genes (100-set) expressed a greater proportion of transcripts in specific tissues as compared with the reference dataset (All-set), we examined the Affymetrix U133A tissue expression dataset, derived from 79 tissues, from the Genomic Institute of the Novartis Research Foundation (GNF). For a given tissue, the fraction of highly expressed genes in the 250-set (and 100-set) was then compared with the fraction of highly expressed genes in the reference dataset (All-set).

The fractions of highly expressed genes containing R•Y tracts (250-set and 100-set) were significantly greater than the corresponding fractions of the All-set in all brain tissues examined, in the atrioventricular node of the fetus and the uterus (Figure 1; *P*-values ranged from 4×10^{-2} to $<1 \times 10^{-13}$ for the 100-set and from 2×10^{-2} to 6×10^{-5} for the 250-set). These fractions were also consistently more enriched in genes with longer R•Y tracts (250-set) than with shorter ones (100-set), regardless of the *P*-values. In summary, these data are compatible either with the view that R•Y tracts with lengths ⩾100 bp (including ⩾250) colocalize preferentially with genes highly expressed in the brain, or that these tracts may perform a role in mediating gene transcription in brain tissues.

Gene size distributions

Some of the genes expressed in human brain tissues, such as contactin-associated protein-like 2 (*CNTNAP2*), dystrophin (*DMD*), and ataxin-2 binding protein (*A2BP1*), are among the largest known, each spanning >1.5 Mb of genomic DNA. We therefore posed the question as to whether the preferential finding of R•Y tracts in brain-expressed genes might be associated with larger gene size. The size distribution for the annotated human gene population followed a Gaussian distribution when the logarithms of gene lengths were analyzed (Figure 2, gray). According to this distribution, the average gene length is ∼18 kb (Figure 2—mode and mean coincide for Gaussian distributions). In contrast, the set of 1377 genes highly expressed in the brain yielded a Gaussian distribution peaking at ∼43 kb (black). Thus, human genes predominantly expressed in brain tissues are on average more than twice as long as those from the total gene population when their log-normal distributions are compared. This notwithstanding, the size distributions of the R•Y tract-containing genes were of disproportionately greater size, with the ⩾100 bp R•Y tract-containing gene set peaking at 108 kb (green) and the ⩾250 bp R•Y tract-containing gene set peaking at 192 kb (red). In addition, the ⩾250 bp R•Y tract-containing genes displayed a pronounced negative skewness, and hence a non-Gaussian behavior. We therefore conclude that larger human genes also tend to host longer R•Y tracts.

Next, we investigated whether larger genes also contained a greater number of R•Y tracts. By analyzing the ⩾100 bp R•Y tract-containing gene set for all pure R•Y tracts ⩾50

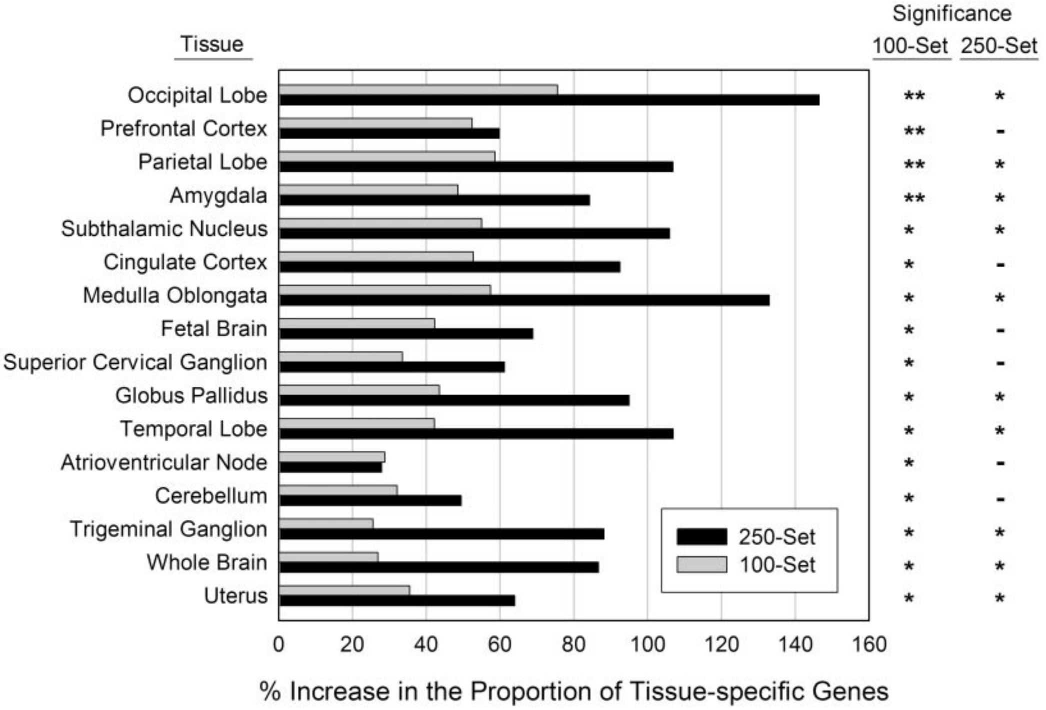

Figure 1. Tissue distribution of R•Y tract-containing genes. The percentage increase in the fraction of R•Y-containing genes (100-set and 250-set) highly expressed (*Z*-score >1) in a given tissue relative to the fraction of all genes highly expressed in the same tissue is displayed on the *x*-axis. *Asterisks*, significantly enriched tissues as determined by binomial probability after a Bonferroni multiple comparison correction. Significance: **, $P < 10^{-7}$; *, $10^{-7} \leqslant P < 0.05$, -, not significant.

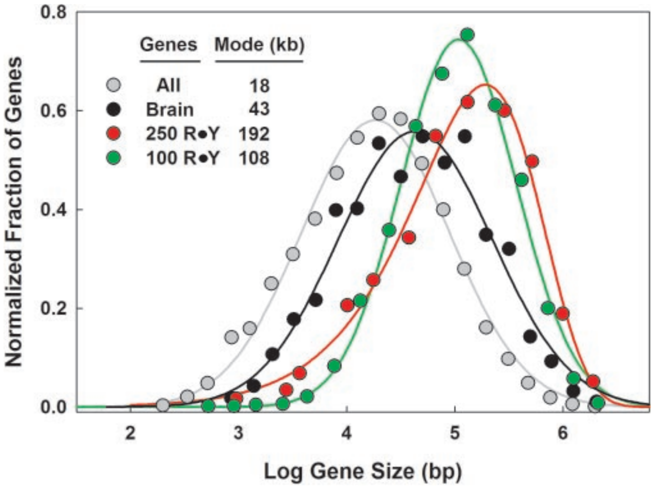

Figure 2. Relationship between gene size and the presence of R•Y tracts. The fractions of genes with log gene size falling within 0.2 (0.25 for ⩾100 bp R•Y tract-containing genes and 0.3 for ⩾250 bp R•Y tract-containing genes) log-intervals plotted as a function of their mean values. Gray, 22,799 human genes; black, 1377 brain-specific human genes; green, 1891 ⩾100 bp R•Y tract-containing human genes; red, 200 ⩾250 bp R•Y tract-containing human genes.

bp in length, 33 genes were found to harbor between 20 and 57 R•Y tracts (Supplementary Table 5). The size of these genes ranged from 38 kb to 2.3 Mb, with a median of ∼800 kb and an average of ∼910 kb, indicating that larger genes also tend to host more R•Y tracts, as predicted.

The gene size distributions of Figure 2 also raise the question as to whether the percent increases in the proportions of highly expressed R•Y tract-containing genes in the brain (Figure 1) could be accounted for by the overall greater length of the genes expressed in this organ. Considering that the distributions P_{All}, P_{Brain}, $P_{\mathrm{RY}\ >\ 100}$ and $P_{\mathrm{RY}\ >\ 250}$ (Figure 2, gray, black, green and red lines, respectively) are interdependent and log-normal (with the exception of $P_{\mathrm{RY}\ >\ 250}$), the predicted ratio increase may be calculated from the following relationship:

$$\frac{\int (P_{RY} \times P_{Brain})dx}{\int (P_{RY} \times P_{All})dx},$$

where P_{RY} is either $P_{\mathrm{RY}>100}$ or $P_{\mathrm{RY}>250}$, using the means and standard deviations that defined the respective curves. The values obtained for the ⩾100 and ⩾250 bp R•Y tract-containing gene sets were 1.30 and 1.32, respectively, for the integrals bound by the log gene sizes of 0 and 8 (which extends beyond all gene sizes), yielding a predicted percent increase of ∼30%. For the 250-set, all percentage increases were greater than this value of 30% for the 13 brain tissues examined (Figure 1). About 2/3 of the 100-set percentage increases in brain tissues were >30% (up to ∼75%), whereas 1/3 were close to 30%. In summary, these data indicate that the number of R•Y tract-containing genes (but not the number of R•Y tracts per kilobase pair) expressed in the brain is greater than expected based on their longer lengths.

The pseudoautosomal region

The total number of pure R•Y tracts ⩾250 bp in the human genome was 814. We evaluated whether these tracts were evenly dispersed following a Poisson-like distribution throughout chromosomes, or whether instead they were clustered in specific regions. By testing each chromosome separately, four small clusters were noted on chromosomes 5, 6 and 12, each involving 2–5 tracts (Supplementary Figure 4). In contrast, two large clusters were identified on the X and Y chromosomes, of 17 and 11 tracts respectively, starting from the telomeric p-arm and extending for 6.2 and 2.6 Mb, respectively. The sequence of the first 2.6 Mb (pseudoautosomal region, or PAR1) is identical on both sex chromosomes, and is essential for homologous recombination during male meiosis and chromosome pairing (52,53). This prompted speculation that the R•Y tracts might play a role in PAR1 function by virtue of their structural properties, perhaps in concert with other non-B DNA-forming sequences, such as IR, DR and MR.

A search for pairs of DR, IR and MR with an arbitrary length cut-off of ⩾62 bp in the first 7.0 Mb of the X chromosome revealed a significant (χ^2 test) over-representation of DR of both ⩾62 bp and ⩾250 bp and IR of ⩾62 bp in the PAR1 region by comparison with the 4.4 Mb region that follows (After_PAR, P < 0.0001). In contrast, MR of ⩾62 bp and IR of ⩾250 bp were not significantly over-represented (Figure 3 and Supplementary Text). In addition, the distribution of MR was characterized by a preponderance of (GAAA•TTTC)$_n$ motifs in PAR1 (10/11 in PAR1 and 8/15 in After-PAR). We surveyed the sequence composition of all DR and IR of ⩾180 bp to determine whether they were uniquely represented in the PAR1 region or whether they corresponded to interspersed elements distributed genome-wide. No significant homology was found to any of the 201 repeats in PAR1. In contrast, 11/21 repeats in the After-PAR region exhibited extensive homologies with other chromosomal sites; indeed 9/11 tracts were identified as LINE1 (7/9), LINE1/LTR (1/9) or LTR (1/9) elements. Additional large segments of repetitive DNA were also identified in PAR1. Seven (>1 kb) were closely examined and found to comprise orderly arrays of tandem repeat blocks (TRB) containing multiple sequence motifs with few interruptions (Supplementary Figure 6). In summary, these data indicated that the clustered R•Y tracts ⩾250 bp form part of a larger family of repetitive elements, mostly DR of unique composition and short IR that densely populate the PAR1 region.

The (GAAA•TTTC)$_n$ repeats

The recurrence of the (GAAA•TTTC)$_n$ repeat within PAR1 was intriguing since this sequence possesses the R•Y mirror symmetry that facilitates triplex-formation (1,3,6,17,18). In addition, its similarity to the (GAA•TTC)$_n$ motifs, which form extraordinarily stable non-B structures (26,34,54,55) raised the question as to whether these structural properties might have been functionally exploited throughout the genome. We addressed this issue by searching for R•Y tracts with mirror symmetry ⩾30 bp in length (sufficient to yield stable triplexes) and comparing their frequencies with those of all microsatellites of comparable repeat units and total lengths. The search for microsatellites with unit lengths from 1 to 29 nt yielded a bimodal distribution with a first peak comprising the mono- to penta-nucleotide repeats (Supplementary Figure 5) and a second peak composed of

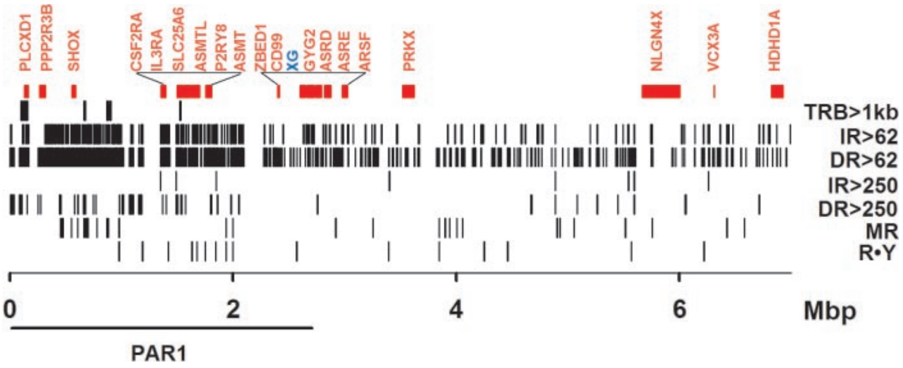

Figure 3. Repetitive elements in the pseudoautosomal region. The vertical lines represent the locations of repetitive elements in the first 7 Mbp of the human X chromosome containing the PAR1 region. R•Y, clustered R•Y repeats from Supplementary Figure 4; MR, mirror repeats ⩾62 bp; DR>250, direct repeats ⩾250 bp; IR>250, inverted repeats ⩾250 bp; DR>62, direct repeats ⩾62 bp but <250 bp; IR>62, inverted repeats ⩾62 bp but <250 bp; TRB>1kb, sequences containing tandem repeat blocks >94.8% pure with a total length >1kb; red, annotated genes; XG (in blue) gene spanning the PAR1 boundary.

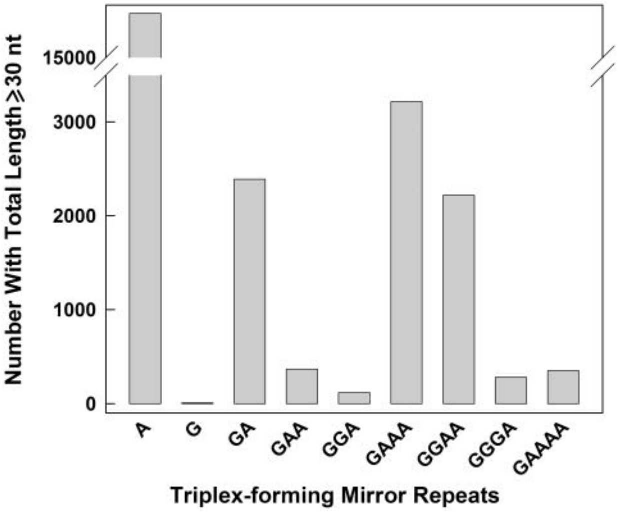

Figure 4. Number of Triplex-forming sequences with mirror symmetry in the human genome. Gray bars, total numbers of uninterrupted R•Y tracts with mirror symmetry ⩾30 bp in length.

⩾15 nt unit lengths, most of which could be decomposed into shorter repeat units displaying an even/odd length asymmetry, as previously noted for the short (1–6mer) repeats (56). Dinucleotides were the most abundant (55 196 copies), followed by tetra- (29 590 copies), mono- (16 686 copies), penta- (8699) and tri-nucleotides (6711). The distribution of R•Y tracts revealed an abundance of $(A\bullet T)_n$ (16 679 copies) over $(G\bullet C)_n$ (seven copies, Figure 4). At present, we do not understand this paucity of $(G\bullet C)_n$ runs. The distributions were followed by $(GAAA\bullet TTTC)_n$ (3217 copies), $(GA\bullet TC)_n$ (2390 copies) and $(GGAA\bullet TTCC)_n$ (2200 copies). All other combinations of G+A repeats contained <400 copies. Therefore, given that poly(A) sequences have the unique property of decreasing triplex stability with increasing length (57–59), these analyses (Figure 4 and Supplementary Figure 5) indicated that the $(GAAA\bullet TTTC)_n$ repeats were over-represented with respect to the total microsatellite population, and were especially abundant among the triplex-forming motifs with mirror symmetry. However, the role, if any, of such tracts and the underlying ability to form intra- or intermolecular triplexes remains to be determined.

Evolutionary comparisons

To determine whether long pure R•Y tracts (⩾250 bp) are unique to human, we queried the chimpanzee, dog, mouse, rat and chicken genomes. Such R•Y tracts lengths were found to be common for all the mammalian/avian species examined (Supplementary Table 6). In addition, the number of pure sequences ⩾50 bp varied from ∼8000 in chicken to >100 000 in murids, attesting to their abundance. A search of the mouse genome for genes with ⩾250 bp R•Y tracts returned a set of 827 non-redundant entries and indicated that the main functional enrichment categories (Supplementary Table 7) largely overlapped those of the ⩾100 bp R•Y tract-containing human gene set (Supplementary Table 3). This indicates that R•Y tracts have been retained within the same gene families since human–mouse divergence ∼80 million years ago (Mya), and also that human membrane-associated genes (Supplementary Table 2) highly expressed in brain have maintained longer R•Y tract lengths than other gene families.

This notwithstanding, little or no homology was found between human and mouse orthologous genes when 5 kb fragments flanking the ⩾250 bp R•Y tracts of the 228 human genes (Supplementary Table 1) were compared. Similarly, analysis of the human and mouse orthologous genes which contained ⩾250 bp R•Y tracts in both species (24 total) revealed little conservation in terms of either the number or location of the tracts (Supplementary Table 8). Thus, whereas R•Y tracts have been retained within gene families, their lengths and locations within individual genes have diverged substantially.

Human–chimpanzee orthologous R•Y tracts

To determine more accurately the extent of sequence conservation, the human intragenic R•Y tracts ⩾250 bp, together with ±2.5 kb of flanking sequence, were compared with their orthologous sites in the chimpanzee genome. These species diverged ∼5–8 Mya from their last common ancestor (LCA) and still share ∼98% sequence identity. Orthologous sequences were identified for most of the 239 tracts (Supplementary Text). A typical dot-plot depicting a comparison of 5 kb of the human *CD99L2* gene with its chimpanzee counterpart is displayed in Figure 5. Sequence identity is evident throughout the region with the exception of the R tract which, although present in both species, is somewhat shorter in chimpanzee. Of all the 142 R•Y sequence pairs analyzed, most displayed the expected ∼98% homology in the R•Y tract-flanking sequences, but none showed sequence conservation within the tracts. We conclude that the R•Y tracts have mutated at a much faster rate than the surrounding sequences. Subsequent analyses (Supplementary Figure 7) provided evidence for a combination of slipped mispairing, recombination-mediated duplications, nucleotide substitutions and possibly also gene conversion, in mediating R•Y tract divergence.

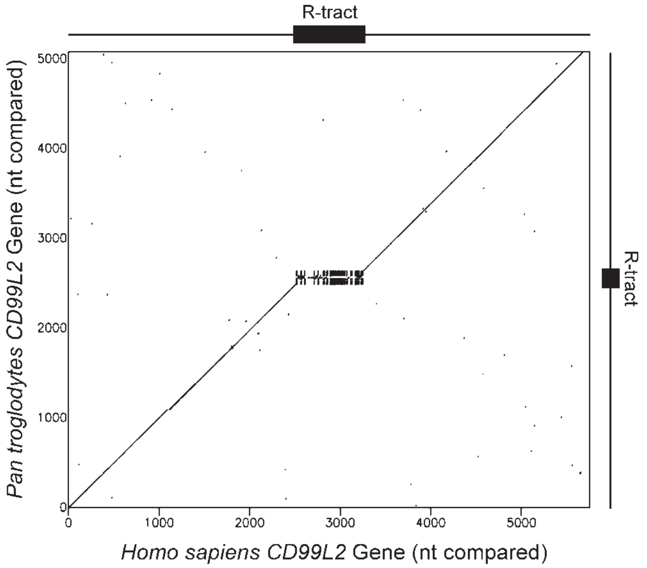

Figure 5. Comparative genomics of R•Y tracts. Dot-plot of the R-tract and 5 kb of flanking sequence in the human *CD99L2* gene with the orthologous gene from chimpanzee.

To determine whether R•Y tracts manifest a length bias in hominids, we compared the total length of all pure tracts ⩾250 bp in both species with their orthologous counterparts (Supplementary Text); 80% of the 582 tracts analyzed were found to be longer in human than in the chimpanzee (Figure 6). This bias was unlikely to be due to the lower accuracy of assembly of the chimpanzee genome (∼3.6× coverage versus 6–10× for the human assembly). Not only did many long chimpanzee R•Y tracts with sequencing gaps match long tracts in the human assembly, but a pairwise plot of all tracts also displayed exponential decay when arranged by size. However, these sizes diverged from the curve fit for ∼150/582 tracts in the chimpanzee, but only for ∼30/582 tracts in the human (inset in Figure 6). These results suggest that long R•Y tracts may have been either more readily acquired or maintained in the human lineage than in the chimpanzee.

Finally, the human–chimpanzee orthologous searches revealed two instances (namely, the *PRKCB1* and *ADAM18* genes) in which the R•Y tract and flanking sequences were

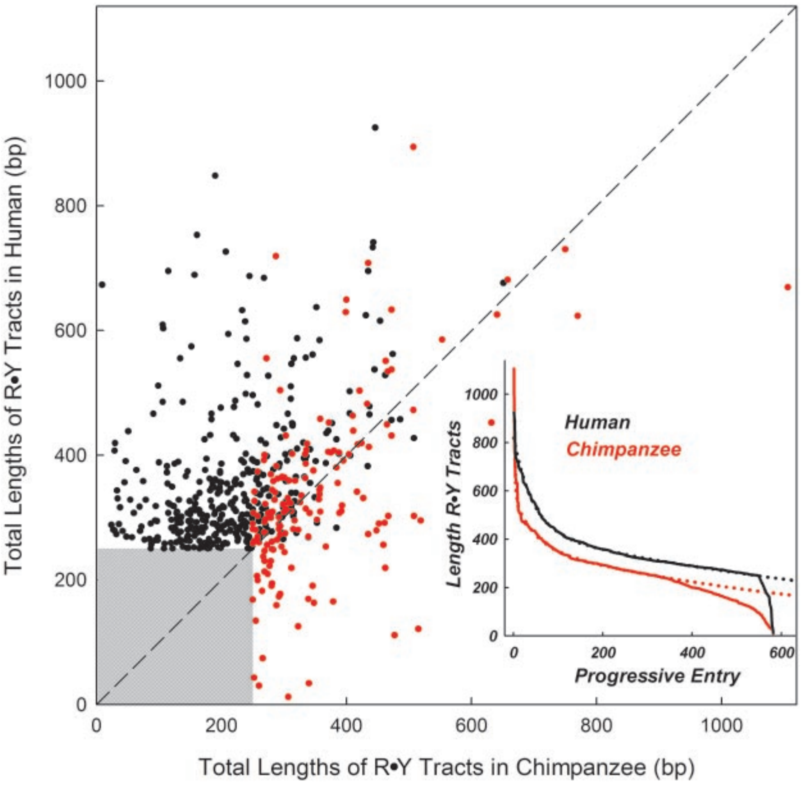

Figure 6. Length comparison of R•Y tracts between human and chimpanzee. The lengths of the tracts in chimpanzee are plotted against the lengths of the orthologous tracts in human. The total R•Y tract lengths are given, including interruptions. Black dots, human pure R•Y tracts ⩾250 bp in length versus orthologous tracts in chimpanzee (436 total); red dots, chimpanzee pure R•Y tracts ⩾250 bp in length versus orthologous tracts in human (166 total). Inset, the 582 unique human/chimpanzee pairs of R•Y tracts from the main panel were ranked by length. Black line, human R•Y tracts (average length = 359 ± 114 bp); red line, chimpanzee R•Y tracts (average length = 260 ± 127 bp); dotted lines, curves fitted to the distributions of R•Y tracts.

present in only one of the two species. In an attempt to understand the possible mutational mechanisms involved, we queried the macaque genome (LCA ~25 Mya) and analyzed the sequence composition and breakpoint junctions. The results of these analyses (Supplementary Figures 8 and 9 and Supplementary Text) were consistent with the R•Y tracts having been inserted and deleted as part of larger (~3.5–8.1 kb) DNA fragments through non-homologous end joining reactions, most of which occurred at IR, suggesting that cruciform structures may have been involved in mediating these genomic rearrangements.

DISCUSSION

This report describes the distribution of the most prominent homopurine•homopyrimidine sequences in the human genome. Their preferred distribution within genes expressed in the brain and encoding membrane-bound proteins with channel and receptor activities contrasts with that of the largest IRs (44), which are mostly associated with testis-expressed genes on the X and Y chromosomes. Whereas IRs are known to form cruciform structures, these long R•Y tracts contain segments that fulfill the requirements for mirror symmetry to foster stable triplex structures, although other non-B DNA conformations, such as slipped structures and occasional tetraplexes, are also possible. Specific types of non-B DNA-forming sequences may therefore be associated with distinct gene families. The R•Y tract enrichment of the PAR1 region also differs from the distribution of the largest IR, which are absent from this portion of the sex chromosomes and are concentrated instead in downstream regions (44). This suggests that different types of non-B DNA conformations may be also functionally distinct. Whereas cruciforms have been proposed to mediate gene conversion thereby contributing to genomic integrity (44), R•Y tracts may be involved in stimulating genomic diversity and recombination by virtue of their ability to fold into triplexes and other conformations.

Length polymorphisms have been noted in R•Y tract-containing alleles in mammals (Supplementary Text). The human–mouse and human–chimpanzee comparisons are consistent with the rapid mutation of long R•Y tracts through length and sequence changes, driven by dynamic mutational mechanisms that include slippage, recombination/repair and nucleotide substitution (Supplementary Figure 7). The total length of all pairs of human–chimpanzee R•Y tracts considered amounts to 213 020 bp for human and 159 020 bp for chimpanzee, i.e. an average length divergence of ~0.039 bases per site per My for the past 6.5 My. This value is more than an order of magnitude higher than the average rate of point mutations between these two species (~0.0019) (60), and is likely to exceed the value of ~0.075 reported for subtelomeric segmental duplication/transfer rates during primate evolution (61) if sequence changes and nucleotide substitutions are also included.

Large-scale analyses indicate a positive correlation between point mutation frequency and repeat sequence density (62), implying a role for dynamic mutation in genomic plasticity and hence genome divergence. Similarly, repetitive sequences increase the frequency of gross rearrangements (9,10,31,63) by engaging distant sites, whereas triplex-forming oligonucleotides increase nucleotode substitution rates (30). Genome-wide surveys have established a positive correlation between R•Y tracts and both recombination rates and nt diversity (64–67). Taken together, these findings support our contention that R•Y tracts, perhaps by virtue of their ability to fold into unconventional DNA conformations, represent hotspots for the generation of DSBs, which may then provide nucleation sites for chromatin remodeling pathways involving DNA repair and recombination (67). Studies of the relationship between genomic instabilities and the presence of repetitive sequences at breakpoint junctions suggest an active role for non-B DNA conformations (including slipped structures, cruciforms and triplexes) in promoting DSBs leading to rearrangements both in the context of human disease [reviewed in (8) and (16,35,68)] and evolution (69,70).

Genes involved in ion transport, synaptic transmission and brain-related functions display accelerated rates of evolution in hominids when compared with murids (60). Concomitantly, human, chimpanzee and other non-human primates exhibit accelerated changes in gene expression in brain tissues, with increased expression specifically in the human lineage (71–73). Such acceleration has been suggested to be 'caused by positive selection that changed the functions of genes expressed in the brains of humans more than in the brains of chimpanzees' (71). The presence of R•Y tracts within large brain-expressed genes may have synergized with increased mutation rates, thereby contributing to accelerated sequence divergence; these tracts may also have potentiated the acquisition of novel transcriptional activities.

The number of RNA transcripts in the mammalian genome is estimated to be at least one order of magnitude greater than the number of annotated genes (currently ~20 000), implying that the majority of the mammalian genome is transcribed (74). Transcription on both strands, generating sense and antisense transcripts with a role in RNA interference and transcriptional regulation, is over-represented in genes encoding cytoplasmic proteins but under-represented in genes encoding membrane and extracellular proteins (75). Expansion of a $(GAA•TTC)_n$ triplex-forming motif is associated with gene silencing in Friedreich's ataxia as a result of strong secondary structure formation (2). It is therefore conceivable that R•Y tracts, particularly those with mirror symmetry such as the highly recurrent $(GAAA•TTTC)_n$ repeats and other MR within the tracts, may also contribute to sense/antisense transcriptional regulation (76).

Several R•Y tract-containing genes encode proteins that function at convergent nodes in signaling pathways at synapses and also represent susceptibility genes for schizophrenia (77,78). This association and the realization that non-B DNA conformations induce genetic instabilities may underscore the delicate balance that exists between the benefits of accelerated evolution and the risks associated with acquiring deleterious mutations.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This research was supported by grants from the National Institutes of Health (NS37554 and ES11347), the Robert A. Welch Foundation, the Friedreich's Ataxia Research Alliance, the Muscular Dystrophy Foundation (Seek-a-Miracle Foundation) to R.D.W., and in part by the Intramural Research Program of the NIH, NCI and Federal funds from the NCI, NIH (contract number N01-CO-12400). D.N.C. acknowledges the financial support of BIOBASE GmbH. Certain commercial equipment, instruments, materials, or companies are identified in this paper to specify adequately the experimental procedure. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are the best available for the purpose. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government. Funding to pay the Open Access publication charges for this article was provided by the NIH and the Robert A. Welch Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Sinden,R.R. (1994) *DNA Structure and Function*. Academic Press, San Diego, CA.
2. Wells,R.D., Dere,R., Hebert,M.L., Napierala,M. and Son,L.S. (2005) Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res.*, **33**, 3785–3798.
3. Mirkin,S.M. and Frank-Kamenetskii,M.D. (1994) H-DNA and related structures. *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 541–576.
4. Rich,A. and Zhang,S. (2003) Timeline: Z-DNA: the long road to biological function. *Nature Rev. Genet.*, **4**, 566–572.
5. Neidle,S. and Parkinson,G.N. (2003) The structure of telomeric DNA. *Curr. Opin. Struct. Biol.*, **13**, 275–283.
6. Soyfer,V.N. and Potaman,V.N. (1996) *Triple-Helical Nucleic Acids*. Springer-Verlag, New York.
7. Hurley,L.H. (2002) DNA and its associated processes as targets for cancer therapy. *Nature Rev. Cancer*, **2**, 188–200.
8. Bacolla,A. and Wells,R.D. (2004) Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.*, **279**, 47411–47414.
9. Bacolla,A., Jaworski,A., Larson,J.E., Jakupciak,J.P., Chuzhanova,N., Abeysinghe,S.S., O'Connell,C.D., Cooper,D.N. and Wells,R.D. (2004) Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc. Natl Acad. Sci. USA*, **101**, 14162–14167.
10. Wojciechowska,M., Bacolla,A., Larson,J.E. and Wells,R.D. (2005) The myotonic dystrophy type 1 triplet repeat sequence induces gross deletions and inversions. *J. Biol. Chem.*, **280**, 941–952.
11. Kuroda-Kawaguchi,T., Skaletsky,H., Brown,L.G., Minx,P.J., Cordum,H.S., Waterston,R.H., Wilson,R.K., Silber,S., Oates,R., Rozen,S. *et al*. (2001) The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nature Genet.*, **29**, 279–286.
12. Repping,S., Skaletsky,H., Lange,J., Silber,S., Van Der Veen,F., Oates,R.D., Page,D.C. and Rozen,S. (2002) Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. *Am. J. Hum. Genet.*, **71**, 906–922.
13. Gotter,A.L., Shaikh,T.H., Budarf,M.L., Rhodes,C.H. and Emanuel,B.S. (2004) A palindrome-mediated mechanism distinguishes translocations involving LCR-B of chromosome 22q11.2. *Hum. Mol. Genet.*, **13**, 103–115.
14. Abeysinghe,S.S., Chuzhanova,N., Krawczak,M., Ball,E.V. and Cooper,D.N. (2003) Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. *Hum. Mutat.*, **22**, 229–244.
15. Chuzhanova,N., Abeysinghe,S.S., Krawczak,M. and Cooper,D.N. (2003) Translocation and gross deletion breakpoints in human inherited disease and cancer II: Potential involvement of repetitive sequence elements in secondary structure formation between DNA ends. *Hum. Mutat.*, **22**, 245–251.
16. Kato,T., Inagaki,H., Yamada,K., Kogo,H., Ohye,T., Kowa,H., Nagaoka,K., Taniguchi,M., Emanuel,B.S. and Kurahashi,H. (2006) Genetic variation affects *de novo* translocation frequency. *Science*, **311**, 971.
17. Wells,R.D., Collier,D.A., Hanvey,J.C., Shimizu,M. and Wohlrab,F. (1988) The chemistry and biology of unusual DNA structures adopted by oligopurine.oligopyrimidine sequences. *FASEB J.*, **2**, 2939–2949.
18. Frank-Kamenetskii,M.D. and Mirkin,S.M. (1995) Triplex DNA structures. *Annu. Rev. Biochem.*, **64**, 65–95.
19. Chan,P.P. and Glazer,P.M. (1997) Triplex DNA: fundamentals, advances, and potential applications for gene therapy. *J. Mol. Med.*, **75**, 267–282.
20. Inman,R.B. (1964) Transitions of DNA Homopolymers. *J. Mol. Biol.*, **9**, 624–637.
21. Wells,R.D. and Larson,J.E. (1972) Buoyant density studies on natural and synthetic deoxyribonucleic acids in neutral and alkaline solutions. *J. Biol. Chem.*, **247**, 3405–3409.
22. Jaishree,T.N. and Wang,A.H. (1993) NMR studies of pH-dependent conformational polymorphism of alternating (C-T)*n* sequences. *Nucleic Acids Res.*, **21**, 3839–3844.
23. Morgan,A.R. and Wells,R.D. (1968) Specificity of the three-stranded complex formation between double-stranded DNA and single-stranded RNA containing repeating nucleotide sequences. *J. Mol. Biol.*, **37**, 63–80.
24. Krasilnikov,A.S., Panyutin,I.G., Samadashwily,G.M., Cox,R., Lazurkin,Y.S. and Mirkin,S.M. (1997) Mechanisms of triplex-caused polymerization arrest. *Nucleic Acids Res.*, **25**, 1339–1346.
25. Patel,H.P., Lu,L., Blaszak,R.T. and Bissler,J.J. (2004) *PKD1* intron 21: triplex DNA formation and effect on replication. *Nucleic Acids Res.*, **32**, 1460–1468.
26. Ohshima,K., Montermini,L., Wells,R.D. and Pandolfo,M. (1998) Inhibitory effects of expanded GAA•TTC triplet repeats from intron I of the Friedreich ataxia gene on transcription and replication *in vivo*. *J. Biol. Chem.*, **275**, 14588–14595.
27. Kohwi,Y. and Panchenko,Y. (1993) Transcription-dependent recombination induced by triple-helix formation. *Genes Dev.*, **7**, 1766–1778.
28. Faruqi,A.F., Datta,H.J., Carroll,D., Seidman,M.M. and Glazer,P.M. (2000) Triple-helix formation induces recombination in mammalian cells via a nucleotide excision repair-dependent pathway. *Mol. Cell. Biol.*, **20**, 990–1000.
29. Luo,Z., Macris,M.A., Faruqi,A.F. and Glazer,P.M. (2000) High-frequency intrachromosomal gene conversion induced by triplex-forming oligonucleotides microinjected into mouse cells. *Proc. Natl Acad. Sci. USA*, **97**, 9003–9008.
30. Vasquez,K.M., Narayanan,L. and Glazer,P.M. (2000) Specific mutations induced by triplex-forming oligonucleotides in mice. *Science*, **290**, 530–533.
31. Wang,G. and Vasquez,K.M. (2004) Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc. Natl Acad. Sci. USA*, **101**, 13448–13453.
32. Bacolla,A., Jaworski,A., Connors,T.D. and Wells,R.D. (2001) *PKD1* unusual DNA conformations are recognized by nucleotide excision repair. *J. Biol. Chem.*, **276**, 18597–18604.
33. Pandolfo,M. and Koenig,M. (1998) Freidreich's Ataxia. In Wells,R.D. and Warren,S.T. (eds), *Genetic Instabilities and Hereditary Neurological Diseases*. Academic Press, San Diego, CA, pp. 373–398.
34. Vetcher,A.A., Napierala,M., Iyer,R.R., Chastain,P.D., Griffith,J.D. and Wells,R.D. (2002) Sticky DNA, a long GAA.GAA.TTC triplex that is formed intramolecularly, in the sequence of intron 1 of the frataxin gene. *J. Biol. Chem.*, **277**, 39217–39227.
35. Raghavan,S.C., Chastain,P., Lee,J.S., Hegde,B.G., Houston,S., Langen,R., Hsieh,C.L., Haworth,I.S. and Lieber,M.R. (2005) Evidence for a triplex DNA conformation at the bcl-2 major breakpoint region of the t(14;18) translocation. *J. Biol. Chem.*, **280**, 22749–22760.

36. Raghavan,S.C. and Lieber,M.R. (2004) Chromosomal translocations and non-B DNA structures in the human genome. *Cell Cycle*, **3**, 762–768.
37. Raghavan,S.C., Swanson,P.C., Wu,X., Hsieh,C.L. and Lieber,M.R. (2004) A non-B-DNA structure at the *Bcl*-2 major breakpoint region is cleaved by the RAG complex. *Nature*, **428**, 88–93.
38. Knauert,M.P., Lloyd,J.A., Rogers,F.A., Datta,H.J., Bennett,M.L., Weeks,D.L. and Glazer,P.M. (2005) Distance and affinity dependence of triplex-induced recombination. *Biochemistry*, **44**, 3856–3864.
39. Hoffman-Liebermann,B., Liebermann,D., Troutt,A., Kedes,L.H. and Cohen,S.N. (1986) Human homologs of TU transposon sequences: polypurine/polypyrimidine sequence elements that can alter DNA conformation *in vitro* and *in vivo*. *Mol. Cell. Biol*., **6**, 3632–3642.
40. Christophe,D., Cabrer,B., Bacolla,A., Targovnik,H., Pohl,V. and Vassart,G. (1985) An unusually long poly(purine)-poly(pyrimidine) sequence is located upstream from the human thyroglobulin gene. *Nucleic Acids Res*., **13**, 5127–5144.
41. Behe,M.J. (1987) The DNA sequence of the human beta-globin region is strongly biased in favor of long strings of contiguous purine or pyrimidine residues. *Biochemistry*, **26**, 7870–7875.
42. Schroth,G.P. and Ho,P.S. (1995) Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. *Nucleic Acids Res*., **23**, 1977–1983.
43. Ussery,D., Soumpasis,D.M., Brunak,S., Staerfeldt,H.H., Worning,P. and Krogh,A. (2002) Bias of purine stretches in sequenced chromosomes. *Comput. Chem*., **26**, 531–541.
44. Warburton,P.E., Giordano,J., Cheung,F., Gelfand,Y. and Benson,G. (2004) Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res*., **14**, 1861–1869.
45. Collins,J.R., Stephens,R.M., Gold,B., Long,B., Dean,M. and Burt,S.K. (2003) An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics*, **82**, 10–19.
46. Ming,Y., Horton,D., Cohen,J.C., Hobbs,H.H. and Stephens,R.M. (2006) WholePathwayScope: a comprehensive pathway-based analysis tool for high-throughput data. *BMC Bioinformatics*, **7**, 30.
47. Karlin,S. and Macken,C. (1991) Some statistical problems in the assessment of inhomogenesis of DNA sequence data. *J. Am. Statist. Assoc*., **86**, 27–35.
48. Gusev,V.D., Nemytikova,L.A. and Chuzhanova,N.A. (1999) On the complexity measures of genetic sequences. *Bioinformatics*, **15**, 994–999.
49. Van Raay,T.J., Burn,T.C., Connors,T.D., Petri,L.R., Germino,G.G., Klinger,K.W. and Landes,G.M. (1996) A 2.5 kb polypyrimidine tract in the *PKD1* gene contains at least 23 H-DNA-forming sequences. *Microb. Comp. Genomics*,, **1**, 317–327.
50. Vasquez,K.M., Wang,G., Havre,P.A. and Glazer,P.M. (1999) Chromosomal mutations induced by triplex-forming oligonicleotides in mammalian cells. *Nucleic Acids Res*., **27**, 1176–1181.
51. Wang,G., Seidman,M.M. and Glazer,P.M. (1996) Mutagenesis in mammalian cells induced by triple helix formation and transcription-coupled repair. *Science*, **271**, 802–805.
52. Filatov,D.A. and Gerrard,D.T. (2003) High mutation rates in human and ape pseudoautosomal genes. *Gene*, **317**, 67–77.
53. Perry,J., Palmer,S., Gabriel,A. and Ashworth,A. (2001) A short pseudoautosomal region in laboratory mice. *Genome Res*., **11**, 1826–1832.
54. Napierala,M., Dere,R., Vetcher,A. and Wells,R.D. (2004) Structure-dependent recombination hot spot activity of GAA.TTC sequences from intron 1 of the Friedreich's ataxia gene. *J. Biol. Chem*., **279**, 6444–6454.
55. Sakamoto,N., Chastain,P.D., Parniewski,P., Ohshima,K., Pandolfo,M., Griffith,J.D. and Wells,R.D. (1999) Sticky DNA: self-association properties of long GAAaTTC repeats in R R Y triplex structures from Friedreich's ataxia. *Molecular Cell*, **3**, 465–475.
56. Subramanian,S., Mishra,R.K. and Singh,L. (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol*., **4**, R13.
57. Sandstrom,K., Warmlander,S., Graslund,A. and Leijon,M. (2002) A-tract DNA disfavours triplex formation. *J. Mol. Biol*., **315**, 737–748.
58. Roberts,R.W. and Crothers,D.M. (1996) Prediction of the stability of DNA triplexes. *Proc. Natl Acad. Sci. USA*, **93**, 4320–4325.
59. James,P.L., Brown,T. and Fox,K.R. (2003) Thermodynamic and kinetic stability of intermolecular triple helices containing different proportions of C+*GC and T*AT triplets. *Nucleic Acids Res*., **31**, 5598–5606.
60. Mikkelsen,T.S., Hillier,L.W., Eichler,E.E., Zody,M.C., Jaffe,D.B., Yang,S.-P., Enard,W., Hellmann,I., Lindblad-Toh,K., Altheide,T.K. *et al*. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
61. Linardopoulou,E.V., Williams,E.M., Fan,Y., Friedman,C., Young,J.M. and Trask,B.J. (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature*, **437**, 94–100.
62. Chiaromonte,F., Yang,S., Elnitski,L., Yap,V.B., Miller,W. and Hardison,R.C. (2001) Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. *Proc. Natl Acad. Sci. USA*, **98**, 14503–14508.
63. Meservy,J.L., Sargent,R.G., Iyer,R.R., Chan,F., McKenzie,G.J., Wells,R.D. and Wilson,J.H. (2003) Long CTG tracts from the myotonic dystrophy gene induce deletions and rearrangements during recombination at the APRT locus in CHO cells. *Mol. Cell. Biol*., **23**, 3152–3162.
64. Kong,A., Gudbjartsson,D.F., Sainz,J., Jonsdottir,G.M., Gudjonsson,S.A., Richardsson,B., Sigurdardottir,S., Barnard,J., Hallbeck,B., Masson,G. *et al*. (2002) A high-resolution recombination map of the human genome. *Nature Genet*., **31**, 241–247.
65. Jensen-Seaman,M.I., Furey,T.S., Payseur,B.A., Lu,Y., Roskin,K.M., Chen,C.F., Thomas,M.A., Haussler,D. and Jacob,H.J. (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res*., **14**, 528–538.
66. Hellmann,I., Prufer,K., Ji,H., Zody,M.C., Paabo,S. and Ptak,S.E. (2005) Why do human diversity levels vary at a megabase scale? *Genome Res*., **15**, 1222–1231.
67. Myers,S., Bottolo,L., Freeman,C., McVean,G. and Donnelly,P. (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science*, **310**, 321–324.
68. Wells,R.D. and Warren,S.T. (1998) *Genetic Instabilities and Hereditary Neurological Diseases*. Academic Press, San Diego, CA.
69. Kehrer-Sawatzki,H., Sandig,C., Chuzhanova,N., Goidts,V., Szamalek,J.M., Tanzer,S., Muller,S., Platzer,M., Cooper,D.N. and Hameister,H. (2005) Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (*Pan troglodytes*). *Hum. Mutat*., **25**, 45–55.
70. Szamalek,J.M., Goidts,V., Chuzhanova,N., Hameister,H., Cooper,D.N. and Kehrer-Sawatzki,H. (2005) Molecular characterisation of the pericentric inversion that distinguishes human chromosome 5 from the homologous chimpanzee chromosome. *Hum. Genet*., **117**, 168–176.
71. Khaitovich,P., Hellmann,I., Enard,W., Nowick,K., Leinweber,M., Franz,H., Weiss,G., Lachmann,M. and Paabo,S. (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, **309**, 1850–1854.
72. Preuss,T.M., Caceres,M., Oldham,M.C. and Geschwind,D.H. (2004) Human brain evolution: insights from microarrays. *Nature Rev. Genet*., **5**, 850–860.
73. Uddin,M., Wildman,D.E., Liu,G., Xu,W., Johnson,R.M., Hof,P.R., Kapatos,G., Grossman,L.I. and Goodman,M. (2004) Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. *Proc. Natl Acad. Sci. USA*, **101**, 2957–2962.
74. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al*. (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
75. Katayama,S., Tomaru,Y., Kasukawa,T., Waki,K., Nakanishi,M., Nakamura,M., Nishida,H., Yap,C.C., Suzuki,M., Kawai,J. *et al*. (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564–1566.

76. Fabregat,I., Koch,K.S., Aoki,T., Atkinson,A.E., Dang,H., Amosova,O., Fresco,J.R., Schildkraut,C.L. and Leffert,H.L. (2001) Functional pleiotropy of an intramolecular triplex-forming fragment from the 3′-UTR of the rat Pigr gene. *Physiol. Genomics*, **5**, 53–65.
77. Harrison,P.J. and Weinberger,D.R. (2005) Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. *Mol. Psychiatry*, **10**, 40–68.
78. Lewis,D.A., Hashimoto,T. and Volk,D.W. (2005) Cortical inhibitory neurons and schizophrenia. *Nature Rev. Neurosci.*, **6**, 312–324.