Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes

Johanna Eddy and Nancy Maizels

Nucl. Acids Res. 2008 36: 1321-1333

- JANE: suggesting journals, finding experts
- SOAP: short oligonucleotide alignment program
- A comparison of common programming languages used in bioinformatics

JANE: Journal/Author Name Estimator

- http://biosemantics.org/jane/

- Web-tool for determining :
  - ✓ which journal is most appropriate for publishing your results
  - ✓ which other scientists can be called upon to review your work

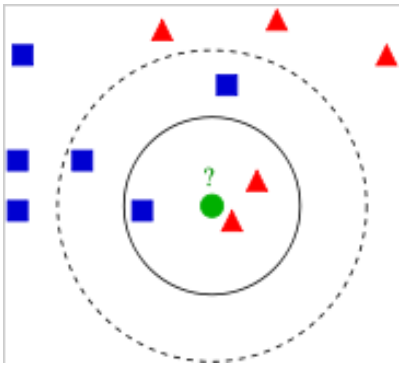
- 4 171 368 Medline articles from 4513 journals

- JANE includes articles that:
  - ✓ contain abstract
  - ✓ are published in the last 10 years
  - ✓ belong to categories: standard research, review and application articles
  - ✓ belonged to a journal with at least 25 publications in the last 10 years and at least one publication in the last 12 months

Schuemie M.J., Kors J.A. "**JANE: suggesting journals, finding experts.**" *Bioinformatics* 2008 Mar 1;24(5):727-8.

JANE: Implementation

- Lucene search engine (MoreLikeThis algorithm) creates the ordered list of most similar articles
- Weighted *k*-nearest neighbor classification to determine the journal/author list
- Confidence scores (normalized Lucene similarity scores) from 0 – 100%
- Best performance with *k* = 50



Example of *k*-NN classification. The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If *k = 3* it is classified to the second class because there are 2 triangles and only 1 square inside the inner circle. If *k = 5* it is classified to first class (3 squares vs. 2 triangles inside the outer circle).

Schuemie M.J., Kors J.A. "**JANE: suggesting journals, finding experts."** *Bioinformatics* 2008 Mar 1;24(5):727-8.

JANE: Comparion with other tools

- Pubmed: only Medline records can be used as queries
- GoPubMed: use only boolean keyword-based query

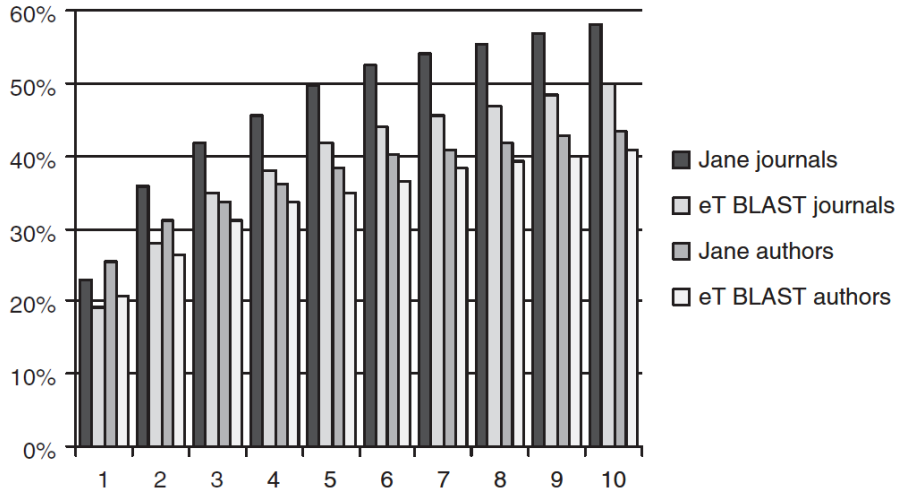


Fig. 2. Cumulative histogram of the rank of the correct journal and the highest ranking correct author in the result lists of eTBLAST and Jane for a test set of 1000 abstracts (e.g. for Jane, the correct journal appeared at the top of the list for 23% of the abstracts, it appeared in the top 2 for 36% of the abstracts, etc.).

eT BLAST	**Jane**
20 CPU cluster	2 CPU
114 sec	0.6 sec

Schuemie M.J., Kors J.A. "**JANE: suggesting journals, finding experts.**" *Bioinformatics* 2008 Mar 1;24(5):727-8.

SOAP: Short Oligonuc. Alignment Program

- http://soap.genomics.org.cn/
- Efficient application for aligning gapped and ungapped short oligonucleotides onto reference sequences
- Useful with Illumina-Solexa sequencing systems: 25-50 nt. length reads
- BLAST and BLAT: unable to cope with huge amount of reads
- SSAHA: optimized for long reads
- ELAND: ungapped alignments up to 32 nt.
- Maq: ungapped alignments using sequence quality information

Li R., Li Y., Kristiansen K., Wang J. "**SOAP: short oligonucleotide alignment program.**" *Bioinformatics* 2008 Mar 1;24(5):713-4.

SOAP: Implementation

- Store the reference sequences in RAM. Two bits for each base, so one byte can store 4 bps
- Split reads into 4 parts - a,b,c,d, two mismatches will be distributed on at most two of the 4 parts at the same time
- Use look up table to judge how many mismatches between reference and read. To have best efficiency, the table used 3 bytes to check a fragment of 12-bp on a time. The table occupied 2^24=16Mb RAM
- Search for identical hits first, if no hits, then 1-mismatch hits will be picked up, then 2-mismatch hits, then gapped hits.

$$RAM = \frac{L}{3} + (4 * 3 + 8 * 6) * 4^S + (4 + 1) * 3 * \frac{L}{4} + 4 * 2^{24}$$

where L is the total length of the reference sequences; S is seed size. For small reference like yeast, $L = 12$ Mb and selected seed size $S = 10$ bp, about 200 Mb RAM is needed; but for the whole human genome, $L = 3$ Gb and a selected seed size $S = 12$ bp, about 14 Gb RAM in total will be needed.

Li R., Li Y., Kristiansen K., Wang J. "**SOAP: short oligonucleotide alignment program.**" *Bioinformatics* 2008 Mar 1;24(5):713-4.

SOAP: Performance

Table 1. Comparison of performance and sensitivity among short oligonucleotide alignment programs

Program	Time consumed (s)	Reads aligned (%)
blastn (−F F −W 11)	165 780	85.47
blastn (−F F −W 15)	150 660	84.66
Blat (−tileSize = 8)	22 032	85.07
Eland	166	88.53
Maq	458	88.39
Soap	134	88.46
Soap iterative	161	90.9
Soap iterative + gapped	486	91.15

We used a query dataset containing 9 914 527 single-end reads (length 32 bp) generated by Illumina-Solexa Genome Analyzer. The DNA sample is a mixture of long PCR products of a 5 Mb human genome region. For blastn we tried 11(better sensitivity) and 15(faster) bp word size, and disabled DUST masking of low-complexity sequence (−F F). For blat, tileSize parameter was set at 8. SOAP used 12 bp seed, SOAP iterative will iteratively trim off 2 bp at 3′-end of read and redo alignment until hits were detected or the remaining sequence is shorter than 27 bp. Sensitivity is calculated under the same threshold by allowing at most 2 mismatches. SOAP gapped will allow one continuous insertion or deletion with size between 1 to 3 bp. After checking sequencing quality, we found the remaining unmappable reads are in low sequencing quality.

Li R., Li Y., Kristiansen K., Wang J. "**SOAP: short oligonucleotide alignment program.**" *Bioinformatics* 2008 Mar 1;24(5):713-4.

Language battle

- Comparison of the memory usage and speed of execution for three standard bioinformatics methods, implemented in programs using one of six different programming languages:

  C, **C++, C#**, **Java**, **Perl** and **Python**

- Focused on bioinformatic methods

- Windows vs. Unix

  Linux: Fedora core 7, kernel 2.6.21-1.3228
  Windows: Windows XP professional, Version 2002, service pack 2
  Intel(R) Core(TM)2 CPU 6400 @ 2.13GHz
  4GB DDR2 memory
  250GB hard drive

Fourment M., Gillings M.R. "**A comparison of common programming languages used in bioinformatics.**" *BMC Bioinformatics* 2008 Feb 5;9(1):82.

Tested algorithms

- **Sellers algorithm** (Sellers P.H.) - simple global sequence alignment method using a dynamic programming approach with a gap penalty

  For sequences of sizes n and m, the running time of the algorithm is O(nm) and the amount of memory used is in O(nm)

- **Neighbor-Joining method** (Saitou and Nei, 1987) - distance-based algorithm for constructing phylogenetic trees and is probably the most widely used distance based method.

  The running time of the algorithm is O(n3) and the amount of memory used is in O(n2)

- **BLAST parsing** - BLAST results can be as large as several gigabytes and a program is usually needed to parse the interesting parts or to feed another program.

Fourment M., Gillings M.R. "**A comparison of common programming languages used in bioinformatics.**" *BMC Bioinformatics* 2008 Feb 5;9(1):82.

Performance (1)

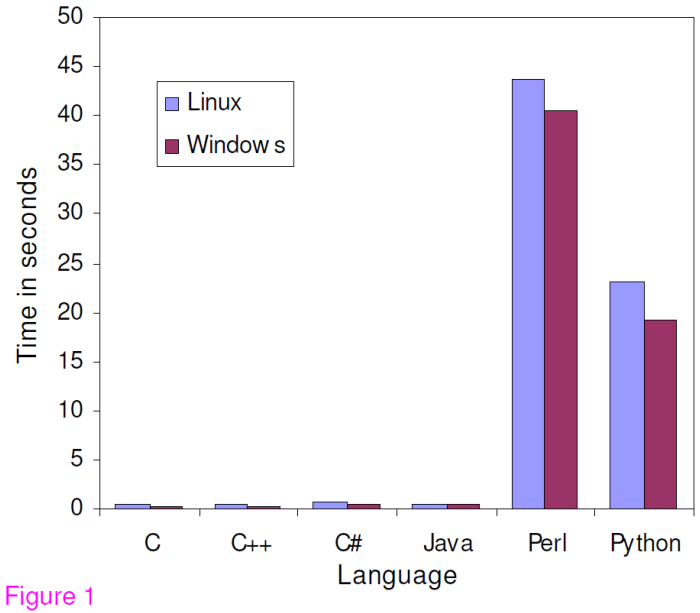


Figure 1

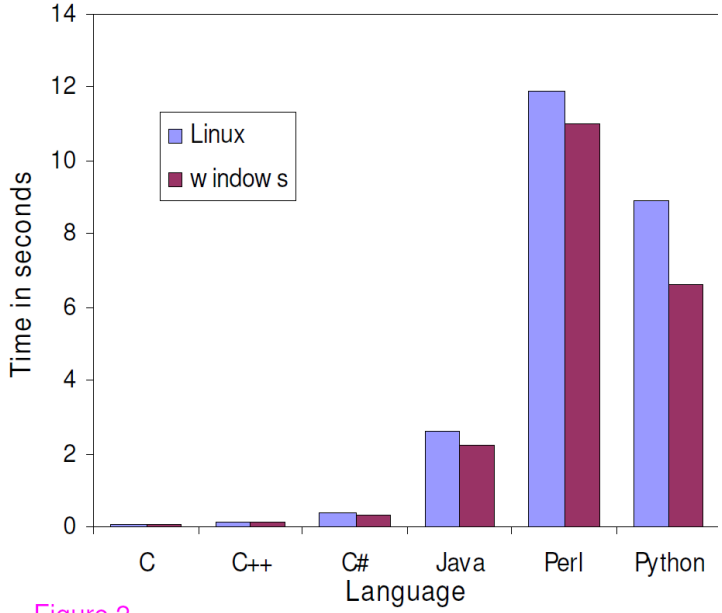


Figure 2

Speed comparison of the global alignment program

Speed comparison of the global alignment algorithm using a gap penalty of 10 implemented in C, C++, C#, Java, Perl and Python. The programs were run on Linux and Windows platforms. Two DNA sequences of 3216bp and 3217bp were used.

Speed comparison of the Neighbor-Joining program

Speed comparison of the Neighbor-Joining algorithm using the Jukes-Cantor evolutionary model implemented in C, C++, C#, Java, Perl and Python. The programs were run on Linux and Windows platforms. The input file was an alignment of 76 DNA sequences.

Fourment M., Gillings M.R. "**A comparison of common programming languages used in bioinformatics.**" *BMC Bioinformatics* 2008 Feb 5;9(1):82.

Performance (2)

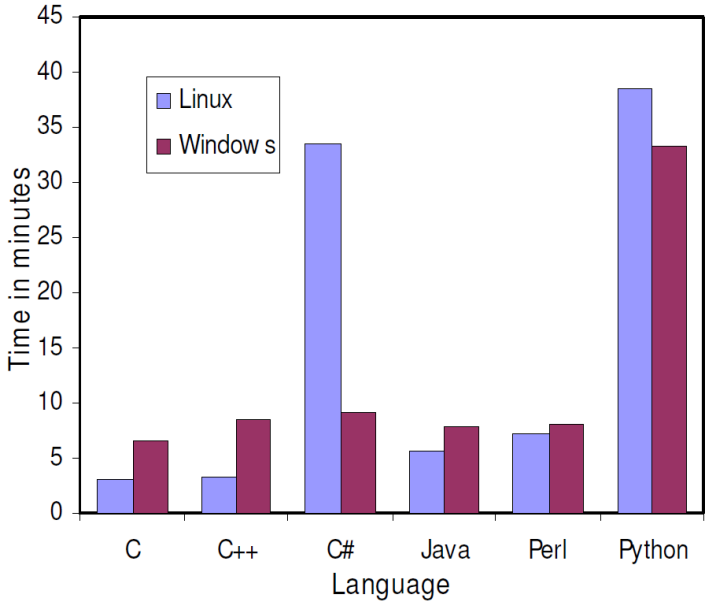


Figure 3

Speed comparison of the BLAST parsing program

Speed comparison of the BLAST parsing program implemented in C, C++, C#, Java, Perl and Python. The programs were run on Linux and Windows platforms. The input file was a 9.8 Gb file from a BLASTP run.

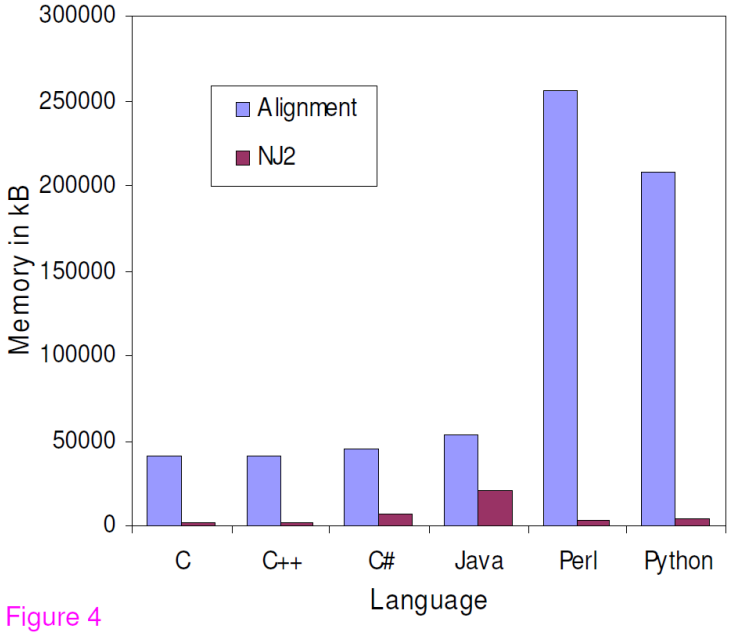


Figure 4

Memory usage comparison of the Neighbor-Joining and global alignment programs

Memory usage comparison for the Neighbor-Joining and global alignment programs implemented in C, C++, C#, Java, Perl and Python. The programs were run on a Linux platform.

Fourment M., Gillings M.R. "**A comparison of common programming languages used in bioinformatics.**" *BMC Bioinformatics* 2008 Feb 5;9(1):82.

Performance (3)

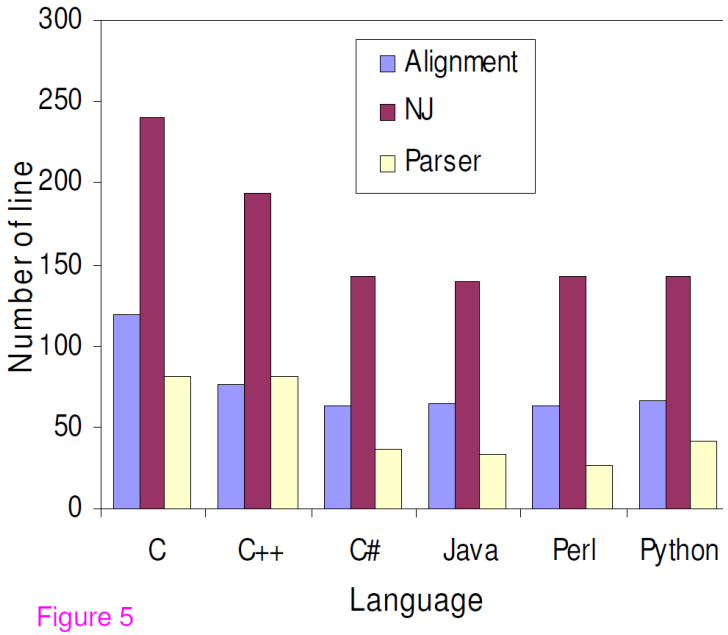


Figure 5

Number of lines for each program

Number of lines for the global alignment, BLAST parser and Neighbor-Joining programs implemented in C, C++, C#, Java, Perl and Python.

Tables

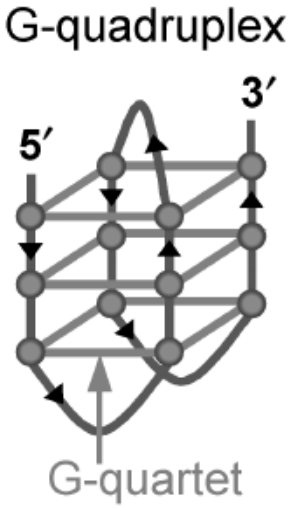
Table 1: Language list with respective compiler or interpreter name and version

	Linux		Windows	
Language	Compiler/interpreter	version	Compiler/interpreter	version
C	GNU gcc	4.1.1	gcc	3.4.2
C++	GNU g++	4.1.1	g++	3.4.2
C#	gmcs/mono	1.1.17.1	.NET csc	2.0.50727
Java	Sun JDK javac/java	1.5.0_09	Sun JDK javac/java	1.5.0_12
Perl	Perl	5.8.8	Active state perl	5.8.8
Python	Python	2.4.4	python	2.5.1

Fourment M., Gillings M.R. "**A comparison of common programming languages used in bioinformatics.**" *BMC Bioinformatics* 2008 Feb 5;9(1):82.

G-quadruplex formation

- **Does the G-richness in gene regulatory areas influence the regulation of gene expression?**
- G-rich sequences form potentially G4 DNA -> (TTAGGG)4 in telomeric regions
- >300 000 sites in human genome to from G-quadruplex structures
- G4 DNA/RNA are very stable structures



- G4 DNA formation in genes correlates with their function: protooncogenes, tumor supressor genes
- G4 DNA is recognized by conserved proteins: RecQ helicases, MutSα

Eddy J, Maizels N. "**Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes.**" *Nucl. Acids Res*. 2008 36: 1321-1333.

Data

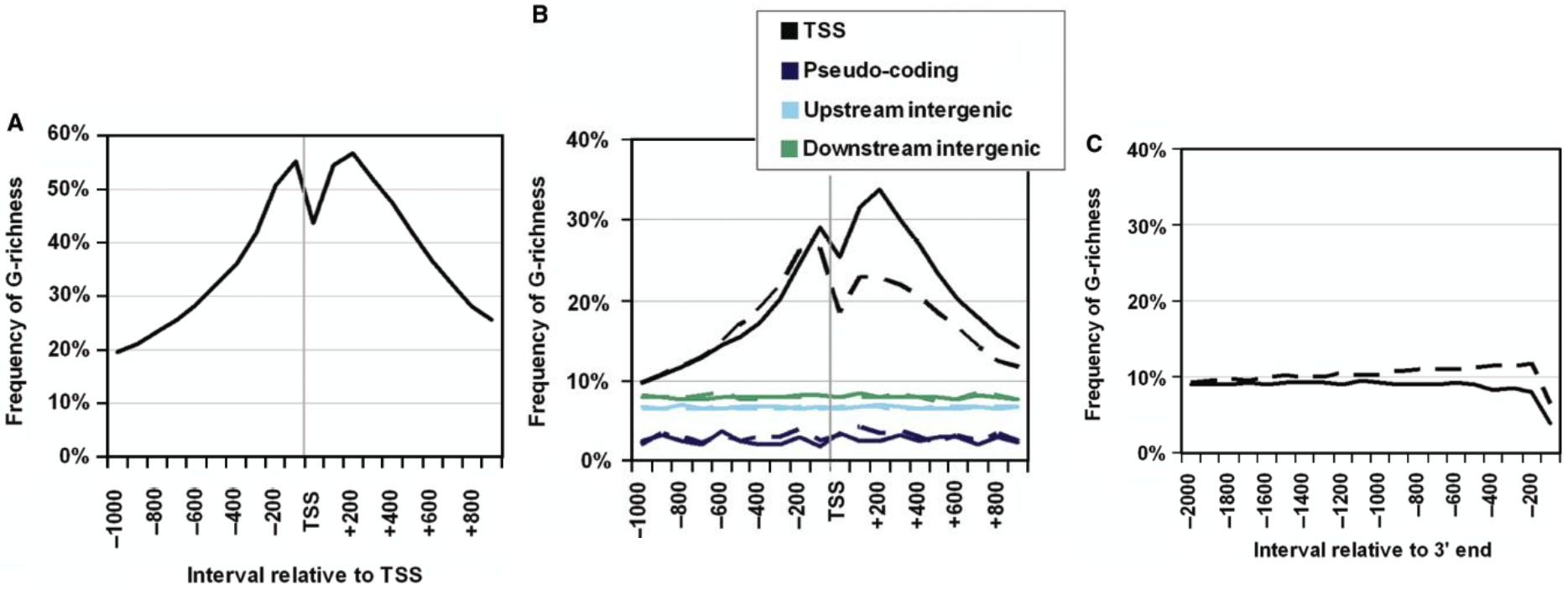
- RefSeq genes: human (18 217), mouse (18 543), chiken (4 782), frog (5 530) and zebrafish (10 578)

- 2kb flanking TSS (+/- 1000 bp), 5 kb upstream and downsteam of gene sequences

- DNA regulatory motifs:
  - ✓ SP1 – RGGCGKR
  - ✓ KLF – GGGGTGGGG
  - ✓ EKLF – AGGGTGKGG
  - ✓ MAZ – GGGAGGG
  - ✓ EGR-1 – GCGTGGGCG
  - ✓ AP-2 – CGCCNGSGGG

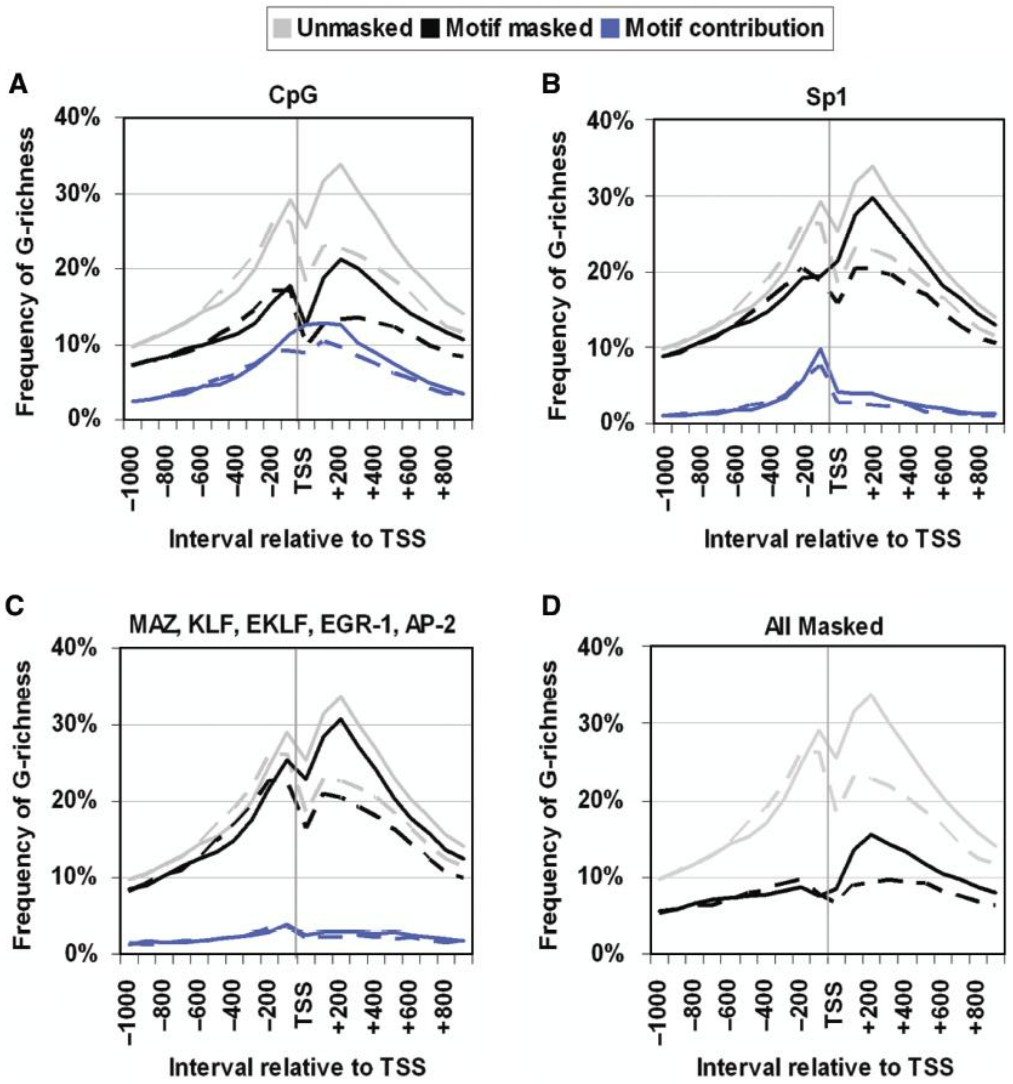
- RNA regulatory motifs:
  - ✓ hnRNP A family – UAGGGU/A
  - ✓ Hn RNP H family – GGGA

Eddy J, Maizels N. "**Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes.**" *Nucl. Acids Res*. 2008 36: 1321-1333.



Strand-biased G-richness in human genes.
Percentage of genes with four or more G-runs per 100 bp interval was calculated for the indicated regions: **(A)** G-richness of duplex DNA within the 2 kb window spanning the TSS; analysis includes 18 217 human RefSeq genes. **(B)** Strand bias of G-richness. Nontemplate strands (solid lines) and template strands (dashed lines) of human RefSeq genes (black); 1000 random pseudo-coding sequences (blue); intergenic sequences 3 kb upstream of the TSS (cyan); and intergenic sequences 3 kb downstream of the 3' ends of the genes (green). G-richness of nontemplate and template strands is indistinguishable within intergenic sequences. **(C)** Strand bias of G-richness within 2 kb of the 3' ends of genes. Nontemplate strands (solid line) and template strands (dashed line).

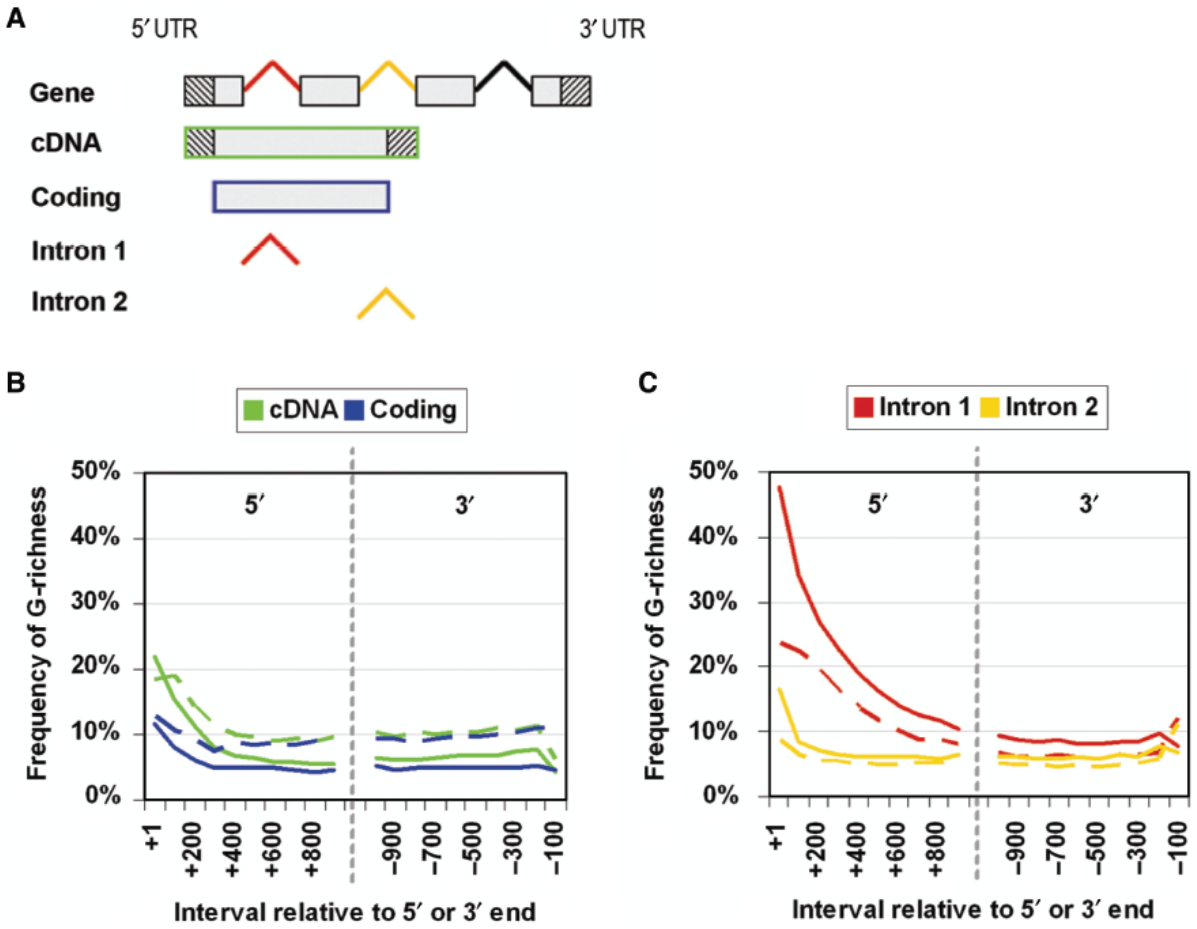
Eddy J, Maizels N. "**Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes."** *Nucl. Acids Res*. 2008 36: 1321-1333.



G-richness upstream but not downstream of the TSS can be attributed to canonical regulatory motifs in duplex DNA.

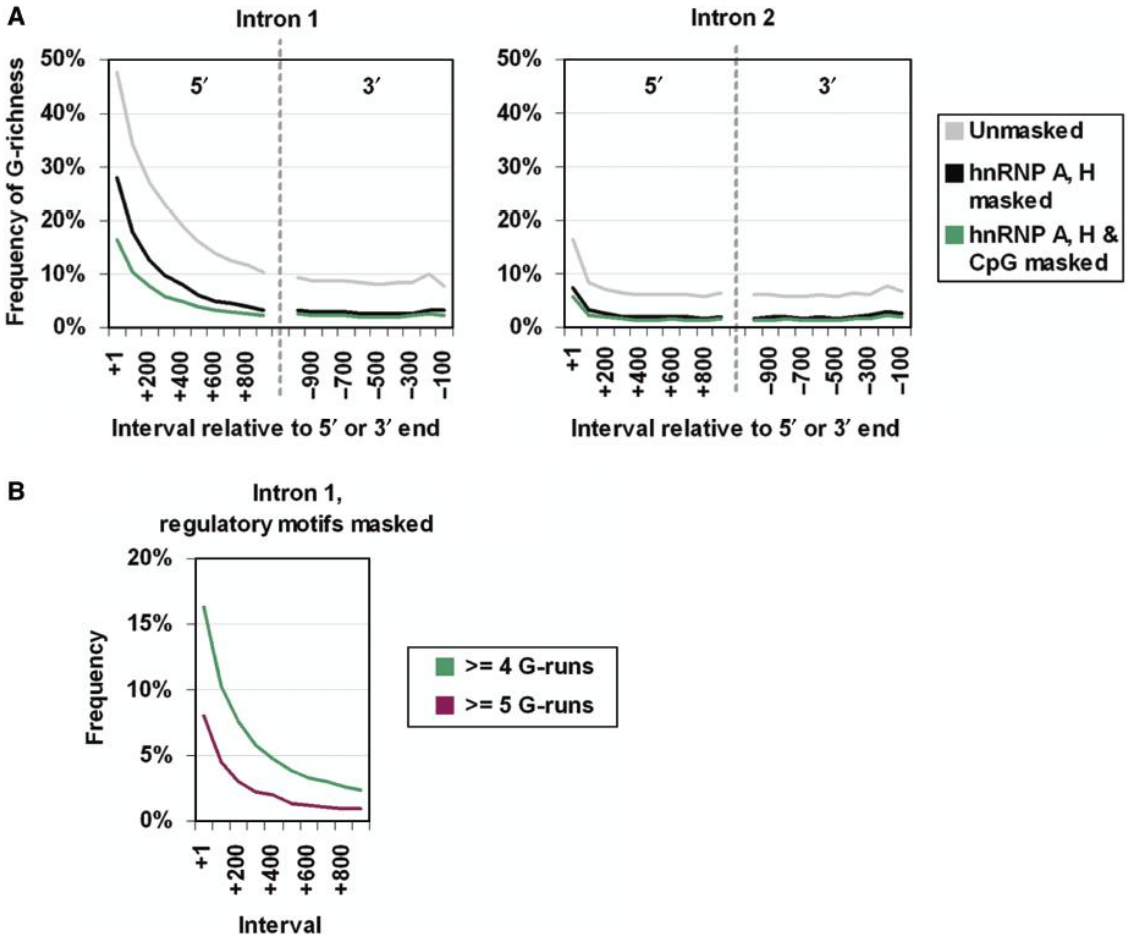
Percentage of genes in which G-richness of nontemplate (solid lines) and template (dashed lines) strands was contributed by specific motifs was analyzed for all 18 217 human RefSeq genes within the 2 kb window spanning the TSS. In each panel G-richness of unmasked sequences is shown for comparison (gray). Motifs tested were: **(A)** G-richness contributed solely by CpG dinucleotides (blue); G-richness calculated with CpG dinucleotides masked (black). Gaussian fit (data not shown) for nontemplate strand G-richness contributed by CpG dinucleotides only, represented by the solid blue line (R2=0.99). **(B)** G-richness contributed solely by SP1 motifs (blue); G-richness calculated with SP1 motifs masked (black). Gaussian fits (data not shown) for nontemplate strand, SP1 motifs only (R2=0.80); and for SP1 motifs masked (R2=0.95). **(C)** G-richness contributed by motifs for 5 transcription factors, MAZ, KLF, EKLF, EGR-1, and AP-2 (blue); G-richness with these 5 transcription factor motifs masked (black). **(D)** G-richness with CpG dinucleotides and motifs for transcription factors SP1, MAZ, KLF, EKLF, EGR-1 and AP-2 masked (black).

Eddy J, Maizels N. "**Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes.**" *Nucl. Acids Res*. 2008 36: 1321-1333.



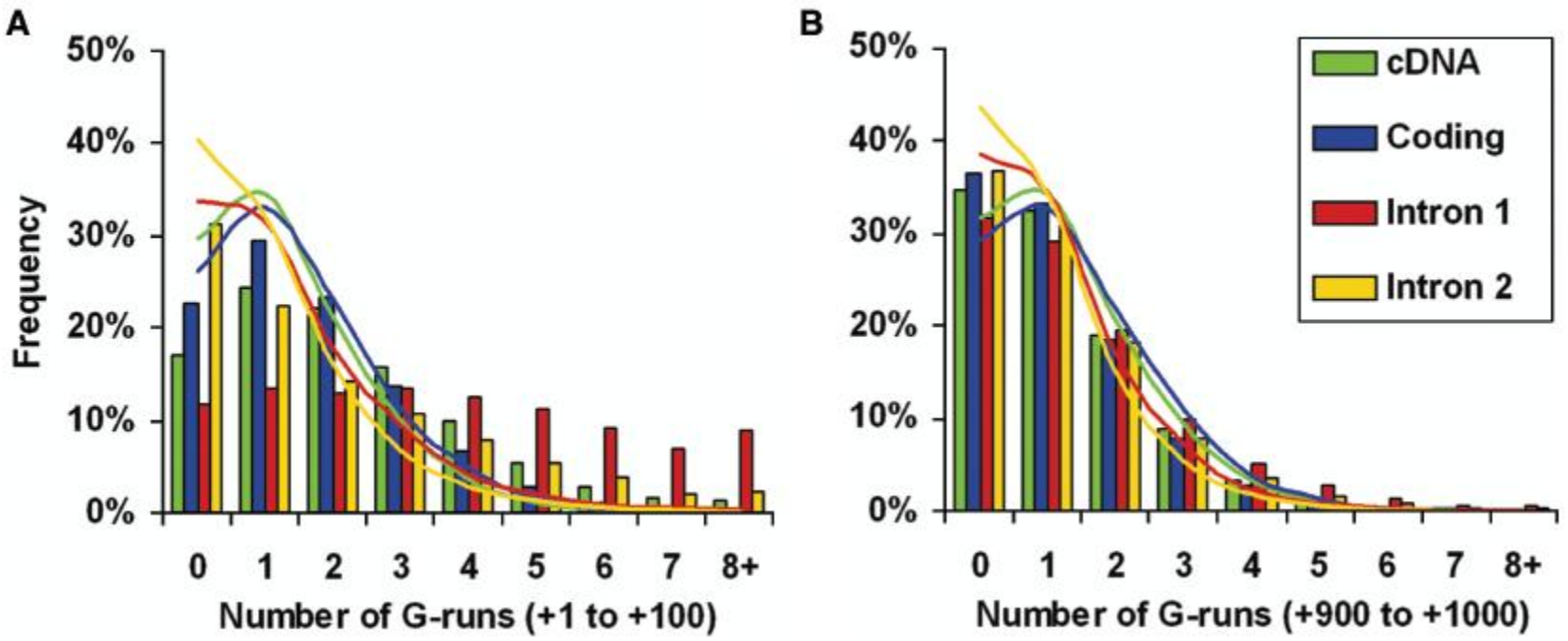
G-richness mapped to functional regions within human genes.
(A) Diagram of a prototype gene with 5′ UTR (reverse hatched boxes), coding exons (gray boxes), introns (carats), and 3′ UTR (forward hatched boxes) indicated. **(B)** G-richness of 19 056 unique cDNA sequences (green), and 13 640 unique coding sequences (blue). G-richness was calculated for the first 1 kb of each sequence relative to the 5′ end, and the last 1 kb of each sequence relative to the 3′ end, for specific elements of a typical gene, for all sequences greater than 1 kb in length, and distinguishing nontemplate (solid lines) and template (dashed lines) strands. Vertical lines separate analyses of 5′ and 3′ regions. **(C)** G-richness of 13 433 unique first intron sequences (red), and 11 540 unique second intron sequences (gold). Analyses and notations as in **(C)**.

Eddy J, Maizels N. "**Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes.**" *Nucl. Acids Res*. 2008 36: 1321-1333.



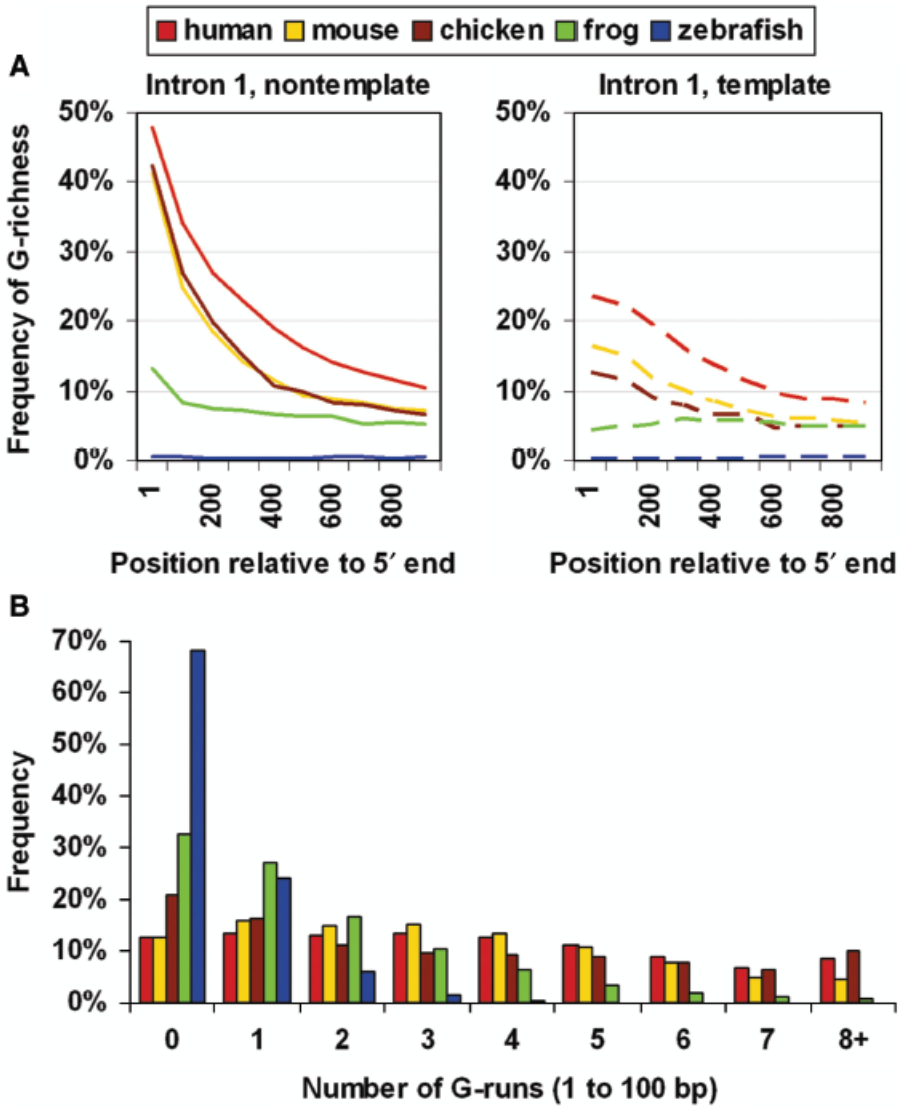
hnRNP A and hnRNP H motifs and CpG dinucleotides contribute to but do not account for G-richness of human first introns. **(A)** Percentage of 13 433 unique first intron sequences (left) or 11 540 unique second intron sequences (right) in which G-richness of nontemplate (solid lines) strands was contributed by specific motifs, within the first 1 kb of each sequence relative to the 5′ end, and the last 1 kb of each sequence relative to the 3′ end, for all sequences that are greater than 1 kb in length. G-richness of unmasked sequences (gray) is shown for comparison with G-richness with motifs for hnRNP A and hnRNP H masked (black), and hnRNP A and hnRNP H plus CpG dinucleotides masked (green). Vertical lines separate the 5′ and 3′ analyses. **(B)** Multiplicity of G-runs in first intron sequences with motifs for hnRNP A and hnRNP H and CpG dinucleotides masked. G-richness with four or more G-runs (green) as in **(A)**, and G-richness redefined as five or more G-runs (plum).

Eddy J, Maizels N. "**Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes.**" *Nucl. Acids Res*. 2008 36: 1321-1333.



The G-rich element at the 5' end of first introns has high potential to form polymorphic G-quadruplex structures.
Numbers of G-runs were enumerated in 100 nt intervals within the nontemplate strand for each specific element of a typical gene (Figure 3A), including cDNA (green), coding (blue), first intron (red) and second intron (gold), for all sequences greater than 100 bp in length. The distribution of numbers of G-runs is shown for two intervals, comparing the observed value of each genomic region (bars) to the value predicted based upon analysis of the same sequences randomly shuffled (lines). Intervals analyzed were: (A) 100 nt interval from +1 to +100 relative to the 50 end. (B) 100 nt interval from +900 to +1000 relative to the 50 end.

Eddy J, Maizels N. "**Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes.**" *Nucl. Acids Res*. 2008 36: 1321-1333.



The G-rich element at the 5’ end of first introns is conserved.
Comparison of G-richness of the first intron sequences of human (red), mouse (gold), chicken (brown), frog (green) and zebrafish (blue). **(A)** G-richness was calculated for first intron sequences of mouse (11 816), chicken (3399), frog (4193), zebrafish (5787), and compared with human (13 433). Regions analyzed were the 100 nt interval from +1 to +100 relative to the 5’ end, for all unique first introns greater than 1 kb in length, for the nontemplate strand (left, solid lines), and template strand (right, dashed lines). **(B)** Distribution of numbers of G-runs in the first 100 nt of the nontemplate strand of the first intron, for all unique intron sequences greater than 100 bp.

Eddy J, Maizels N. “**Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes.”** *Nucl. Acids Res*. 2008 36: 1321-1333.

Conclusions

- Upstream of the TSS known motifs that account the G-richness, downstream of the TSS new motifs
- G-richness is high in nontemplate strand downstream (+200 - +300 bp) of the TSS
- 5’ end of the first intron’ nontemplate strand of many human genes is G-rich
- Proximity to the promoter may enable the G-rich elements to regulate gene expression either in transcription or splicing process

Eddy J, Maizels N. “**Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes.**” *Nucl. Acids Res*. 2008 36: 1321-1333.