BMC Genomics

BioMed Central

Research article

Open Access

Comprehensive analysis of the base composition around the transcription start site in Metazoa

Stein Aerts*[1], Gert Thijs[1], Michal Dabrowski[2], Yves Moreau[1,3] and Bart De Moor[1]

Address: [1]Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Belgium, [2]Laboratory of Transcription Regulation, Nencki Institute, Warsaw, Poland and [3]On leave at Center for Biological Sequence Analysis, BioCentrum, Technical University of Denmark, Lyngby, Denmark

Email: Stein Aerts* - stein.aerts@esat.kuleuven.ac.be; Gert Thijs - gert.thijs@esat.kuleuven.ac.be; Michal Dabrowski - m.dabrowski@nencki.gov.pl; Yves Moreau - yves.moreau@esat.kuleuven.ac.be; Bart De Moor - bart.demoor@esat.kuleuven.ac.be

* Corresponding author

Published: 01 June 2004

BMC Genomics 2004, **5**:34

Received: 27 January 2004
Accepted: 01 June 2004

This article is available from: http://www.biomedcentral.com/1471-2164/5/34

Abstract

Background: The transcription start site of a metazoan gene remains poorly understood, mostly because there is no clear signal present in all genes. Now that several sequenced metazoan genomes have been annotated, we have been able to compare the base composition around the transcription start site for all annotated genes across multiple genomes.

Results: The most prominent feature in the base compositions is a significant local variation in G+C content over a large region around the transcription start site. The change is present in all animal phyla but the extent of variation is different between distinct classes of vertebrates, and the shape of the variation is completely different between vertebrates and arthropods. Furthermore, the height of the variation correlates with CpG frequencies in vertebrates but not in invertebrates and it also correlates with gene expression, especially in mammals. We also detect GC and AT skews in all clades (where %G is not equal to %C or %A is not equal to %T respectively) but these occur in a more confined region around the transcription start site and in the coding region.

Conclusions: The dramatic changes in nucleotide composition in humans are a consequence of CpG nucleotide frequencies and of gene expression, the changes in Fugu could point to primordial CpG islands, and the changes in the fly are of a totally different kind and unrelated to dinucleotide frequencies.

Background

Genomic DNA sequences display compositional heterogeneity on several scales–for example, long-range variations in G+C content (large blocks of DNA of homogeneous composition are often referred to as "isochores" [1]), CpG suppression in vertebrate genomes [2], or skews caused by mutation biases intrinsic to mutation and repair mechanisms [3]. Both neutralist hypotheses and selectionist hypotheses have been made to explain the various compositional variations [4,5]). Until recently it was difficult to investigate more local variations in base composition (for example, at one position relative to some genomic signal). Although there are currently many efforts to understand metazoan gene regulation and transcriptional control, we have only a limited knowledge of the exact start of transcription. In this study we re-evaluate

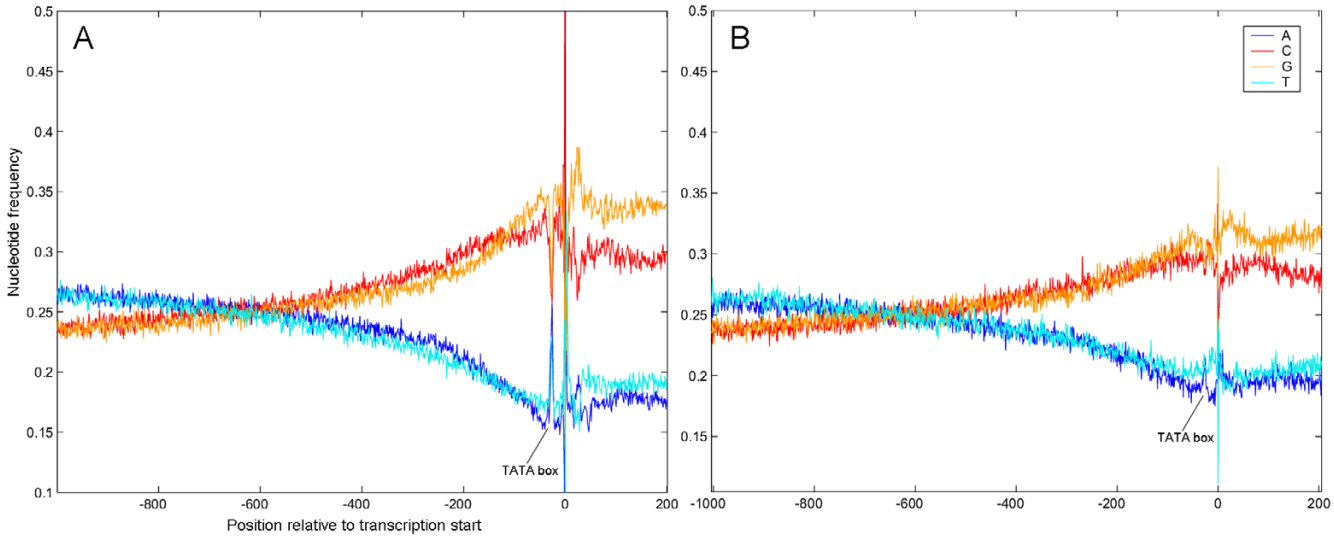

Figure 1
Nucleotide frequencies around the experimentally determined transcription start site (A) and around the annotated gene start in Ensembl of all genes in DBTSS (A) and 5000 randomly selected genes from Ensembl (B).

the average base composition around the transcription start site (TSS) of animal genes. We could both confirm several aspects regarding nucleotide composition and we were able to discover new aspects, especially in invertebrates. It is most obvious from our results that the average nucleotide composition around the transcription start site across the genome is significantly different from the composition in the intergenic and coding regions and some aspects of these composition variations are furthermore different among the investigated species.

Results and discussion

Comparing Ensembl and DBTSS human gene start annotations

From the extraordinary shapes of the composition profiles calculated using the gene start annotations of Ensembl (Figure 1B and Figure 2) it can already be postulated that a significant degree of correct start annotation must be present in Ensembl to get such high resolution. To double check this statement (for human only) we have downloaded all human promoter sequences from the Database of Transcriptional Start Sites (DBTSS). DBTSS contains exact information of the genomic positions of the transcriptional start sites and the adjacent promoters for several thousands of human genes [6]. It can be seen from Figure 1 that the Ensembl data (using 5000 randomly selected genes with at least 100 bp 5'UTR) is noisier but that most of the composition characteristics (as discussed below) are also present in the profiles generated from the Ensembl data. The TATA box is less clear and GC rise is lower for the Ensembl data than for the DBTSS data. We have also checked the quality of the *Drosophila* start points by comparing the nucleotide frequencies around Ensembl (i.e., annotation from FlyBase) gene starts with a data set of experimentally determined TSSs of [7], and they were highly similar [see Additional file 1].

Variations in base composition in different phyla

Figure 2 shows the nucleotide frequencies around TSS for human, fly, and Fugu. A characteristic that is shared among all investigated species is that the A/T content (W) is greater than the G/C content (S) in the intergenic region, for example, at -2000 bp upstream of the TSS. This is the result of the fact that in general the G:C→A:T base pair transition frequency is significantly higher than that of the reverse T:A→C:G transition. Thus accumulation of neutral substitutions results in a generally GC-poor composition of mammalian genomes [8], and apparently also of other vertebrate and also invertebrate genomes. We will further denote this composition as the intergenic background composition (IBC), and we will denote a difference between the A+T content and the G+C content as $\Delta WS = [(A+T)-(G+C)]/(A+T+G+C)$.

The most notable features of the composition profiles are the dramatic changes in ΔWS in the region [-1000,+1000] around the TSS. In human for example, ΔWS changes from ~10% in the IBC to ~-20% at the TSS. A similar polarity switch of ΔWS can be seen in the other vertebrates: mouse, rat, Fugu, and zebrafish (see Figure 2C for

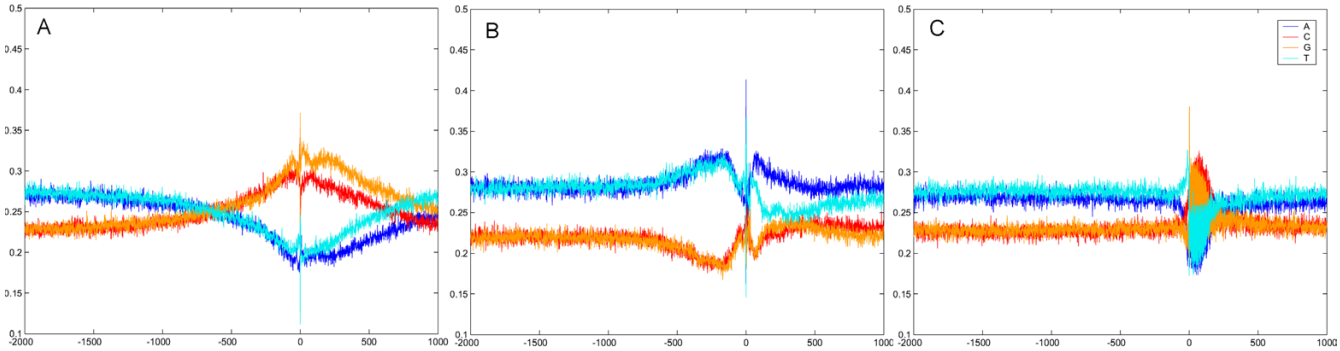

Figure 2
Nucleotide frequencies around the annotated gene start in Ensembl, calculated from 5000 randomly selected genes in human (A), *Drosophila* (B), and Fugu (C).

Fugu and see Additional file 1 for mouse, rat, and zebrafish). The mouse patterns are similar to human. The Fugu and zebrafish patterns also have the same shape with a polarity switch but the composition starts to change later than in mammals and is restored faster as well. The cause of the fast drop in G+C content might be that the 5'UTRs in fish are much shorter than in human so the coding region (where codon usage largely determines base composition) starts immediately after the TSS. A common explanation for the G+C rise that is seen here in the mammalian profile in the proximity of the TSS is the presence of CpG islands, which is related to DNA methylation, or more precisely to a lack of DNA methylation (see below). *Drosophila* (Fig. 2B) also shows a significant change in ΔWS, but without a polarity switch: it increases from ~12% in the IBC to ~26% at the TSS. The maximal difference between ΔWS_{IBC} and ΔWS_{TSS} is not reached at the TSS itself as in vertebrates, but about 150 base pairs before the TSS. The *Drosophila* patterns, showing almost an opposite behavior to that of vertebrates, seems odd at first sight, but because of the absence of DNA methylation in *Drosophila*, a rise in G+C caused by an over-representation of CpG dinucleotides would not be expected anyway (although DNA methylation in insects has been the subject of some debate [9]).

Interestingly, because *Drosophila* does have a change in ΔWS, namely an opposite change to that of vertebrates, there are perhaps factors other than DNA methylation that influence the base composition in this species. One factor could be the general presence of more AT-rich binding sites for transcription factors or histone modification factors [10]. An alternative hypothesis could be that another type of DNA modification than CpG methylation would be involved in a genome-wide marking of promoter regions in *Drosophila*.

Nucleotide composition and CpG islands

Above we have made the remark that the G+C rise in mammals and maybe generally in vertebrates is probably caused by the higher number of CpG dinucleotides in the promoter region. Normally CpGs are present at a frequency of only ~1.5% instead of their expected frequency of ~5% based on the individual frequencies of C and G (0.225 × 0.225). Indeed, most CpGs in the genome are methylated at the cytosine [11] and these methylated cytosines frequently mutate to thymines [12].

To investigate the relationship between CpG frequency and the observed composition profiles, we compared the base compositions between genes with and without a CpG island around the TSS. We did not use a CpG prediction algorithm however to separate CpG-related genes from non-CpG-related genes because CpG island prediction is done using an arbitrary threshold on the number of CpG doublets as compared to the genome frequency. Instead we have taken another approach by simply counting the CpG doublets in the [-400,+400] region around TSS. The same technique was used by Ioshikes and Zhang [13]. A histogram of CpG numbers for 5000 randomly selected genes is bimodal for human (Fig. 3A), but not for fly nor fish (see Figure 3). For human, the first peak represents the genes with CpG numbers that correspond more or less to the genome frequency and the second peak represents genes with more than expected numbers of CpGs.

The histogram of CpG scores for fish (Fig. 3C) shows almost no distinct second peak, but the distribution is slightly broader than the first peak of the human distribution. This could mean that there is some DNA methylation and some CpG over-representation around TSS but not as much as in human. Auf der Maur et al. [14] have suggested that CpG islands of fish may represent a

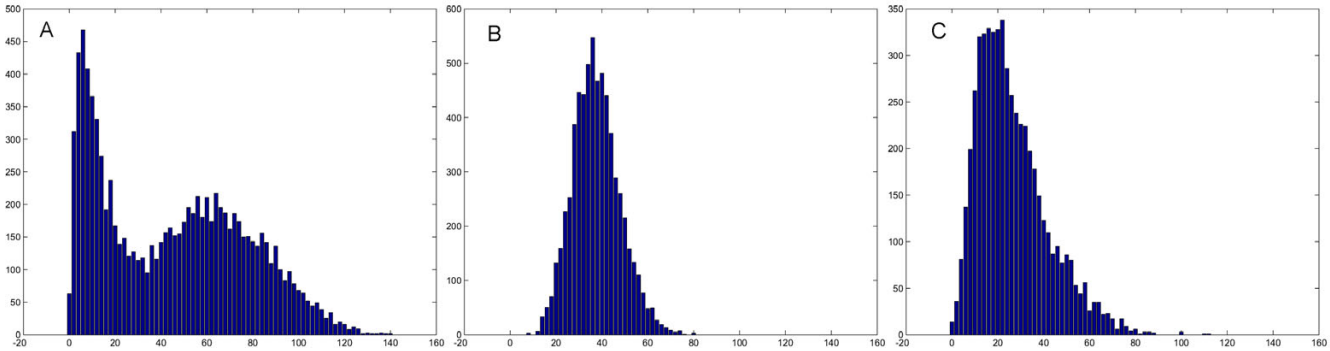

Figure 3
Frequency distributions of the CpG dinucleotide in the [-400,400] region around the TSS in human (A), fly (B), and Fugu (C).

primordial stage of CpG island evolution. This could indeed be a plausible explanation for the Fugu distribution.

The distribution of CpG frequencies in *Drosophila* is a normal distribution (Fig. 3B), which means that there is nothing special about CpG doublets in *Drosophila* and agrees with the absence of DNA methylation in *Drosophila*. To test whether another dinucleotide than CpG is over-represented around TSS in fly, we have performed the same analysis for the fifteen other possible dinucleotides and looked for a distribution like the CpGs for human or fish, but all dinucleotide frequencies were normally distributed and similar to the CpG distribution, although the WpW dinucleotides (AA, AT, TA, TT) had a slightly broader distribution and a higher mean [see Additional file 1].

To see the effect of the CpG concentration on the overall nucleotide composition we have plotted the base composition profiles separately for the 15% lowest scoring and the 15% highest scoring genes (see Figure 4). In human (Fig. 4A,4B), this shows that ΔWS can be completely attributed to CpG over-representation. The results for Fugu (Fig. 4C,4D) show that some genes could have CpG islands (Fig. 4D) since for those the nucleotide composition is similar to the mammalian profiles. This again can be in agreement with the hypothesis of primordial CpG island evolution in other vertebrates than mammals, although other tests are needed to check for a possible functional consequence of the differences between the extremes of the CpG frequency distribution. If we look at the two ends of the distribution of AT dinucleotides in *Drosophila*, we can see a similar breaking apart of the composition profiles into genes with a small ΔWS and genes with a large ΔWS (Fig. 4E,4F). The question remains whether these gene classes in Fugu and fly also have a functional meaning like in human, or that these visualizations are artefacts due to plotting the extremes of the distributions. Below we will test the dependency of the composition profiles on gene expression.

Nucleotide composition and gene expression

It is generally known that the presence of a CpG island around the TSS is related to the expression pattern of the gene. Unmethylated DNA can have an open chromatin structure that facilitates the interaction of transcription factors with the promoter region [15]. Housekeeping genes (HK genes), which are transcribed in all somatic cells and under all circumstances (and thus should be easily activated) frequently have a CpG island in their promoter region [16,17]. Ponger et al. [17] showed that early embryo genes (both housekeeping and tissue specific genes) that are active at the totipotent cell stage or in the blastocyst are associated with CpG islands [17]. We have shown above that our composition profiles are caused by CpG islands, so we can expect to see differences in base composition between genes with different expression patterns. We identified sets of widely and narrowly expressed genes using microarray data using a similar analysis as Eisenberg and colleagues in [18]. We used microarray expression data from 101 different samples taken from 47 different human tissues and cell lines under normal physiological state [19]. The experiments measuring replicates of the same biological condition were averaged to reduce the measurement noise, resulting in 47 data points per probe. We have selected three probe sets with an average reading above 200 standard Affymetrix difference units [18] in the following conditions: (1) in all tissues, these are widely expressed genes; (2) in 20 to 29 out of 47 tissues (medium expression); and (3) in only 1 tissue (narrow expression). Then we mapped the Affymetrix probe identifiers to HUGO gene names using MatchMiner [20] and used these lists to retrieve the corresponding

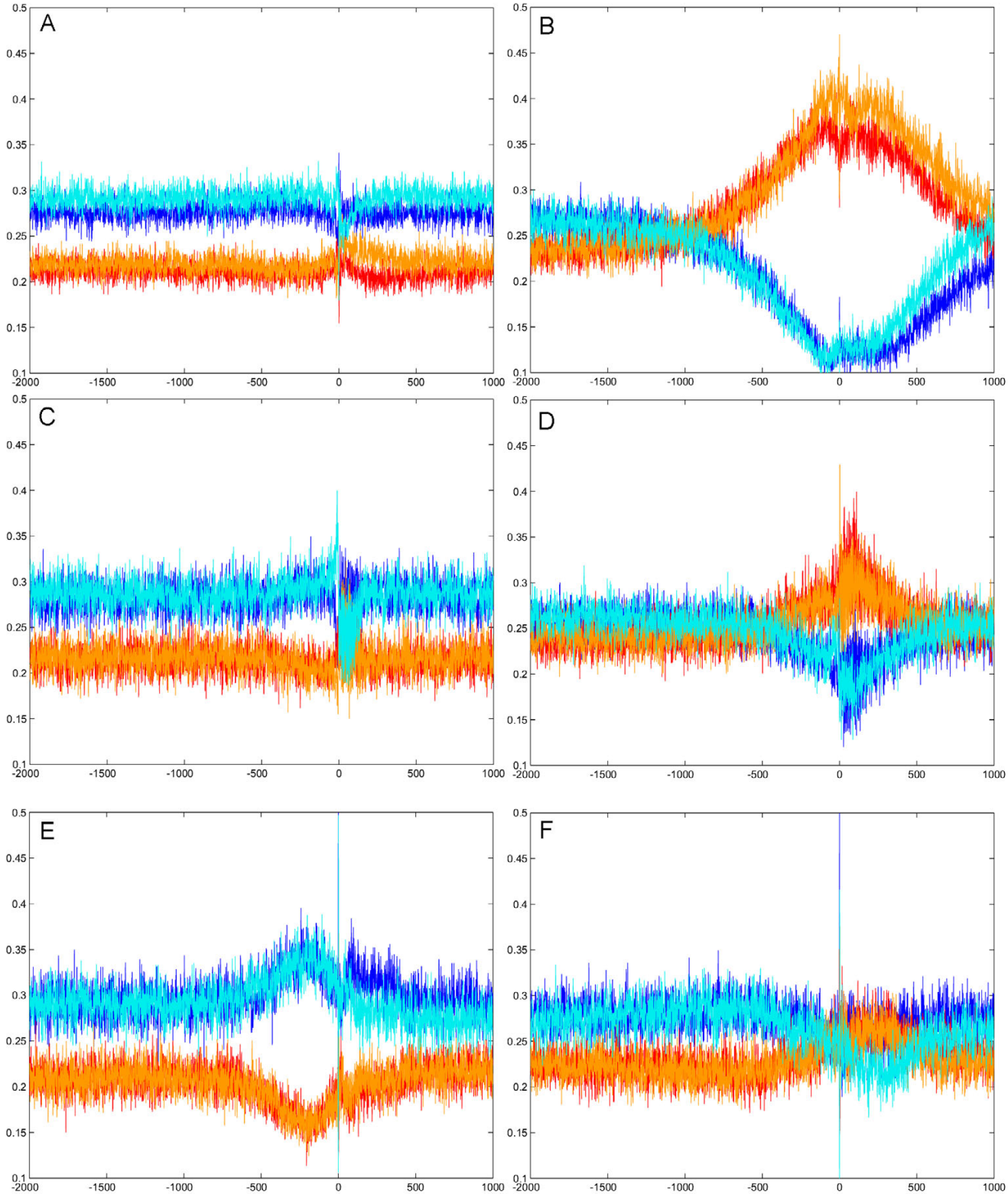

Figure 4
Nucleotide frequencies of several gene classes, separated according to the concentration of a dinucleotide in the [-400,400] region around the TSS. A. Human genes with few CpG doublets. B. Human genes with many CpG doublets. C. Fugu genes with few CpGs. D. Fugu genes with many CpGs. E. Fly genes with many ApTs. F. Fly genes with few ApTs.

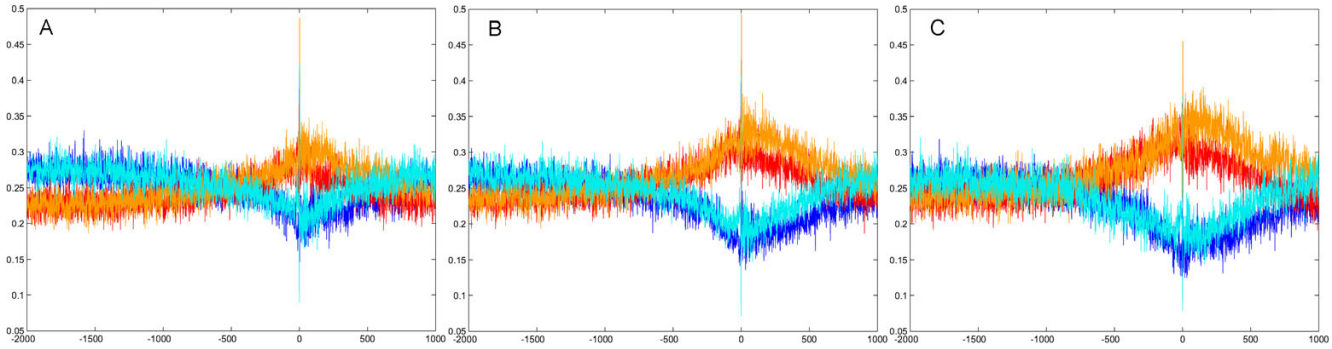

Figure 5
Nucleotide frequencies of three human gene groups: genes with a narrow expression pattern (A), a medium pattern (B), and a wide pattern(C).

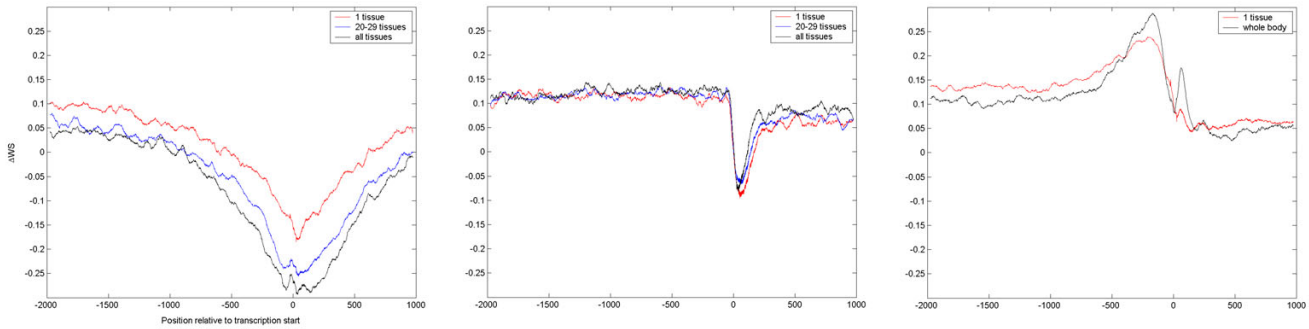

Figure 6
ΔWS profiles. ΔWS = [(A+T)-(G+C)]/(A+T+G+C) is plotted on the *y* axis, at each position *x*. (A) Differences between the ΔWS profiles for human gene groups with narrow, medium, and wide expression can be observed. The significance thereof is assessed (see text and Figure 7). (B) For the orthologous genes in Fugu, there are no observable differences. (C) For narrow and wide expression groups in *Drosophila*, only small differences are present.

sequences using EnsMart. The size of the sets are respectively 647, 886, and 783 genes. Figure 5 shows the average base composition graphs for the three sets. It can be seen that the more widely the genes are expressed, the more pronounced are the variations in ΔWS. A more direct comparison of the ΔWS values for these three gene sets is shown in Figure 6A where ΔWS is plotted along the sequence. From this plot two observations can be made: (1) ΔWS differs between the groups in the background (e.g., 0.1 versus 0.05 for 1 tissue respectively all tissues); and (2) these differences increase around the TSS (-0.17 for 1 tissue versus -0.29 for all tissues, an increase by factor two). The first observation can be linked with literature evidence on GC isochores. Namely, the location of widely expressed genes tends to be correlated with GC-rich isochores [21], and the GC content of exons, introns, and of GC3 positions in GC-rich isochores is higher than in GC-poor isochores.

To assess whether the second observation, which is the focus of this work, is statistically significant, we have calculated the Average Log-Likelihood Ratio (ALLR) between the nucleotide distributions of different expression groups (the formula is given in the Methods section). ALLR was proposed by Wang and Stormo [22] to distinguish probability distributions from each other, as well as from the background. As background profile we use 5000 randomly selected genes. The ALLR has positive values in the regions where the two compared distributions are similar. An illustration of how an ALLR profile behaves for

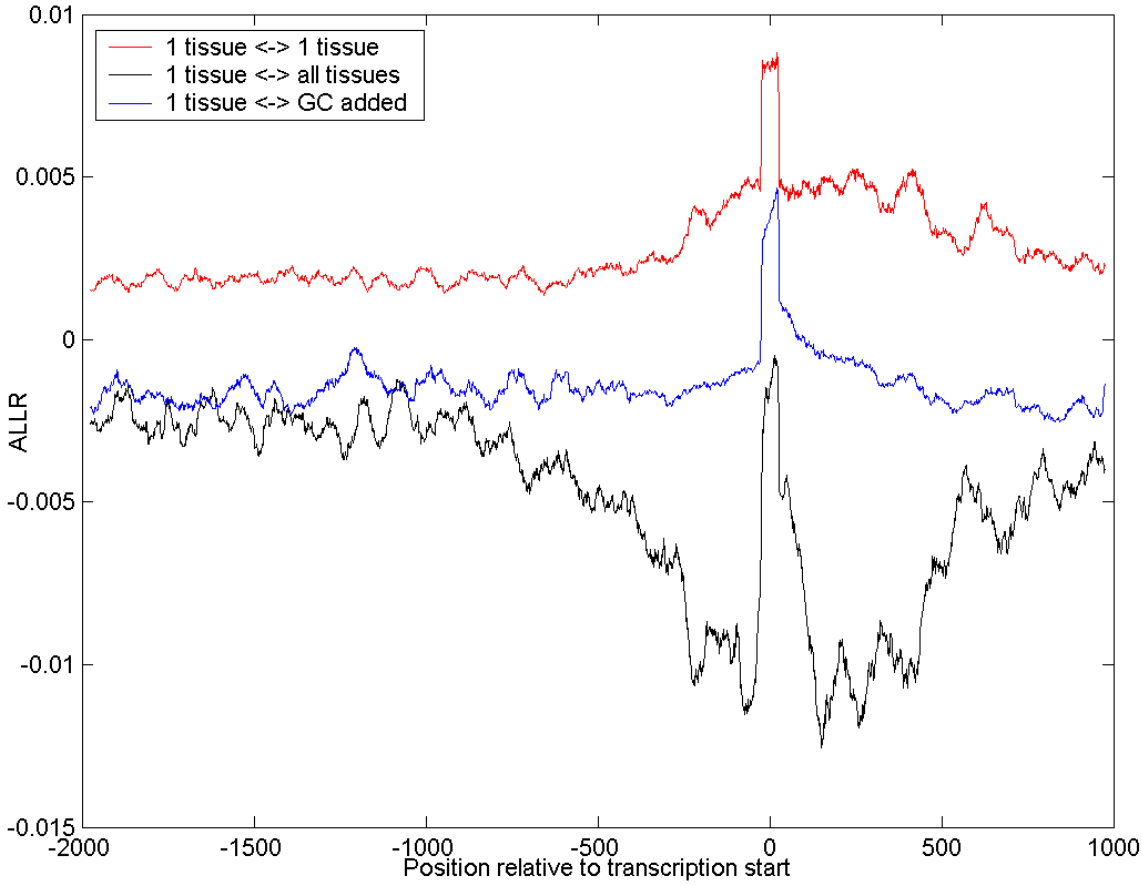

Figure 7
ALLR (Average Log Likelihood Ratio) at each position along the aligned DNA sequences, comparing different distributions. For the red (upper) curve the two compared distributions are the same. The curve lies above zero. Around TSS, the ALLR increases due to a higher similarity with the background profile (5000 random genes). The blue (middle) curve again represents the ALLR values comparing the "1 tissue" expression group with itself, but now the G+C content of one distribution was artificially increased to test the effect caused by GC isochores. The black curve (bottom) compares the "1 tissue" with the "all tissues" expression groups. It represents two effects: one of the GC isochores (where it coincides with the blue curve), and one of the CpG island effect (where it deviates from the blue curve around TSS).

similar distributions is shown in Figure 7 (red, upper curve) where the base composition of the gene group with narrow expression is used for both distributions. What would the ALLR profile look like if the differences in ΔWS between gene groups can be attributed completely to different representations of GC isochore families? To test this, we calculated the ALLR between the gene group with narrow expression and the same data with artificially increased G+C content (middle, blue curve in Figure 7). This curve lies below zero, so the distributions differ. However, the ALLR between the narrow and wide expression groups (lower, black curve) deviates from this curve around the TSS. Here, severe differences between the groups cause the ALLR to decrease further. The CpG doublet counts start to play a more dominant role. Note that this does not rule out the possibility that both phenomena (GC isochores and CpG islands) are outings of the same selection pressure.

These results prove that the composition changes and the effect of methylation on gene expression are functionally conserved (and largely independent of isochore location),

and that the nucleotide compositions are not the result of some kind of mutational bias. The fact that widely expressed genes, regardless of the level of expression, would need a promoter that is easily accessible (e.g., by an open chromatin structure) would make sense in an evolutionary perspective. For these genes one could expect that their regulation depends less on specific *cis*-regulatory modules than on the accessibility of the proximal promoter. To test whether the nucleotide composition also depends on gene expression in Fugu, we have worked under the assumption that Fugu genes that are orthologous to human genes that are widely (or narrowly) expressed, are also widely (or narrowly) expressed. For each of the three human gene sets above, the Fugu orthologous genes were retrieved from the Ensembl data base, the nucleotide frequencies were calculated, and ΔWS was plotted (see Figure 6B). As opposed to human, almost no variation between the groups is observed, maybe because the control of methylation (i.e., keeping promoters unmethylated; for human this is reflected in the second peak of the bimodal CpG distribution) is not or only slightly present in Fugu. It cannot be ruled out however that the absence of a clear trend could be due to the fact that the expression patterns among orthologous genes are not well preserved or again that the gene start annotations are of too low quality.

Drosophila shows a completely different behavior in composition changes so we were interested to see whether these changes also vary with the level of gene expression. Unfortunately we could not find a similar microarray experiment in *Drosophila* that compares different tissues under normal circumstances, and the mapping of human genes to *Drosophila* orthologs results in too few genes. As an alternative we have selected gene sets with different EST expression patterns from the Unigene database, namely (1) Unigene clusters with only one expression site (and leaving out the clusters with *whole body* expression) (narrow expression) and (2) the 2000 clusters with the most expression sites (wide expression). ΔWS for these two sets is displayed in Figure 6C. The difference between the profiles of wide and narrow expression is minimal. The A+T maximum and the G+C minimum (as in Figure 2B) are more or less the same, only the rise in A+T is a little bit steeper in the widely expressed genes (not shown, but reflected in the ΔWS plot of Figure 7C). This finding however might be caused by the quality of the data set and since there is a small observable difference we would not rule out the possibility that differences could be seen in the future when more appropriate data sets are available.

GC and AT skews around the TSS

Chargaff's second parity rule states that the number of As equals the number of Ts, and the number of Cs equals the number of Gs in a *single* strand over windows of sufficient size, often in the order of 1000 bp [23]. In our composition profiles, at least in the intergenic regions, the number of As also equals the number of Ts (and %G=%C), but this is measured at one position across 5000 genes. An 'ergodic' version of Chargaff's second parity rule seems to hold. This variant rule is broken in the [-60,+60] region around the TSS, and also further downstream of the TSS in most species. In vertebrates %A > %T and %G > %C and in invertebrates %T > %A and %C > %G. Such differences are called AT and GC skews and they are measured as (A-T)/(A+T) and (G-C)/(G+C) respectively. The same observation was also made by [24]. The transcription process is asymmetric and might bias mutation patterns between the transcribed and nontranscribed strands by exposing the nontranscribed strand to DNA damage [25]. Both transcription-coupled repair and deamination have been shown experimentally to produce an excess of C→T mutations on the nontranscribed strand in *E. coli* [26,27]. Green and colleagues have shown that A→G transitions can occur significantly more than T→C transitions in transcribed than in non-transcribed regions (in mammals), which can explain the GC skew (G > C) that is present in the whole region after the TSS in vertebrates [3] (we have used the nontranscribed or synonymous strand in all the analyses). In general, they show that transcripts have a significant G+T compositional excess, and we also see that T > A after TSS. Majewski performed a genome-wide study in human and reported the same mutational asymmetry and he further established a correlation between this symmetry and gene expression [8]. All of this however seems only to make sense for the vertebrate skews. Since A > T after the TSS in *Drosophila* (while the opposite is true in human), either the transcriptional machinery that causes the mutational bias differs between these organisms, or else the skews are functionally conserved with a different function in the two phyla. A last observation regarding skews is the sudden AT skew (where the A and T profiles separate in the plots) that occurs right before the TSS in vertebrates and right after the TSS in arthropods. A similar although less pronounced sudden GC skew can be seen right after the TSS in vertebrates, but not in arthropods. For these observations we have no explanation.

Conclusions

In human there is a continuum in gene expression (low-medium-high or narrow-medium-broad) that goes hand in hand with a continuum of CpG doublet concentration around TSS and both are reflected in a continuum of nucleotide frequencies (small-medium-large ΔWS). In other words, genes can differ in their CpG content (and thus in their nucleotide composition) and this difference has a functional meaning (large ΔWS is needed for an 'easy' expression, early in the embryo or in many tissues) and is therefore evolutionary conserved. For CpGs in Fugu these relations are not so clear, perhaps because CpG

islands in fish seem primordial. The changes in fly are of a totally different kind. A possible explanation for the A/T rise in the base composition in *Drosophila* could then be that fly genes differ in their AT-content because of differences in the concentration of AT-rich transcription factor binding sites around the TSS.

Methods

For each sequenced organism that is available in the Ensembl database Release 14 (*Homo sapiens, Mus musculus, Rattus norvegicus, Fugu rubripes, Danio rerio, Drosophila melanogaster* [28], *Anopheles gambiae, Caenorhabditis elegans* [29], and *Caenorhabditis briggsae* [29]), we have randomly selected 5000 stable gene identifiers [30]. These lists were used to retrieve 2000 base pairs (bp) of single stranded DNA from the synonymous strand upstream of the annotated starting point (s) of each gene and 1000 bp downstream. This was done using the EnsMart data mining tool [31]. The analysis and plotting of the average base pair composition of these sequences is done as follows. For each position in the 3000 bp long sequences the percentage of A, T, C, and G over the 5000 genes is calculated and this value is represented on the y axis in Figure 1. The x axis shows the position along the sequence and $x = 0$ corresponds to position +1, the start of the annotated gene or the putative transcription start site. This way of representing the nucleotide composition at an aligned genomic position across many genes – as opposed to a classical average base composition calculated over a window along the DNA strands as in [32] – has been used before for purposes like the study of GC skews in *Arabidopsis* [33], the base composition of complete genes (introns, exons, etc.) [34,24], and promoter prediction [35,36]. Many genes have multiple alternative transcripts with a different TSS. Using DNA regions around each possible TSS of a gene or only around the furthest 5' reaching TSS did not influence the composition profiles (see Additional file 1; this analysis uses the latter). For the sake of brevity we only discuss human, fly, and Fugu profiles. Mouse and rat were very similar to human, profiles of mosquito were very noisy and difficult to interpret, and *C. elegans* and *C. briggsae* are omitted because the interpretation would be too difficult due to trans-splicing at the 5' end of the genes [37,38]. The figures for the undiscussed composition profiles can be found in Additional file 1.

To compare different nucleotide composition profiles, we used the Average Log-Likelihood Ratio (ALLR) [22], which is calculated at each position with the following formula:

$$ALLR = \frac{\sum_{b=A..T} n_{b2} \ln\left(f_{b1} / p_b \right) + \sum_{b=A..T} n_{b1} \ln\left(f_{b2} / p_b \right)}{\sum_{b=A..T} n_{b1} + n_{b2}},$$

where n_{b1} is the number of nucleotides b in profile 1, at a certain position, f_{b1} is the frequency of nucleotide b at this position in profile 1, and p_b is the frequency of b at this position in the background profile.

Authors' contributions

SA designed the experiments, created the data sets, and wrote the manuscript. GT performed all computational analyses and visualizations. MD, YM, and BDM contributed to the manuscript.

Additional material

Additional File 1

This file contains additional figures with nucleotide compositional profiles around the transcription start site for mouse, rat, mosquito, and nematode worms, similar to the profiles shown in Figure 2. Also, it contains a figure that compares the nucleotide composition around transcription start site, either using all transcripts for every gene in random gene set, or using only the 5' furthest reaching transcript. Another figure in this file compares the composition around gene starts in Drosophila *selected in Ensembl with the composition around experimentally determined gene starts. A last section contains all 16 dinucleotide distributions around the transcription start site for human, Fugu, and* Drosophila.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-5-34-S1.pdf]

Acknowledgements

This work is partially supported by Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen (IWT) projects STWW-00162, STWW-Genprom, GBOU-SQUAD-20160; Research Council KULeuven GOA Mefisto-666, GOA-Ambiorics, IDO genetic networks; and FWO: Fund for Scientific Research-Flanders (Belgium) projects G.0115.01 and G.0413.03; IUAP V-22 (2002–2006).

References

1. Bernardi G: **The human genome: organization and evolutionary history.** *Annu Rev Genet* 1995, **29:**445-476.
2. Bird AP: **CpG-rich islands and the function of DNA methylation.** *Nature* 1986, **321(6067):**209-213.
3. Green P, Ewing B, Miller W, Thomas P, Green E: **Transcription-associated mutational asymmetry in mammalian evolution.** *Nat Genet* 2003, **33(4):**514-517.
4. Eyre-Walker A: **Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA.** *Genetics* 1999, **152(2):**675-683.
5. Frank AC, Lobry JR: **Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms.** *Gene* 1999, **238:**65-77.
6. Suzuki Y, Yamashita R, Sugano S, Nakai K: **DBTSS, DataBase of Transcriptional Start Sites: progress report 2004.** *Nucleic Acids Res* 2004, **32:**D78-81.
7. Ohler U, Liao Gc, Niemann H, Rubin GM: **Computational analysis of core promoters in the *Drosophila* genome.** *Genome Biol* 2002, **3(12):**RESEARCH0087.
8. Majewski J: **Dependence of mutational asymmetry on gene-expression levels in the human genome.** *Am J Hum Genet* 2003, **73(3):**688-692.
9. Hendrich B, Tweedie S: **The methyl-CpG binding domain and the evolving role of DNA methylation in animals.** *Trends Genet* 2003, **19(5):**269-277.

10. Sitzler S, Oldenburg I, Petersen G, Bautz EK: **Analysis of the promoter region of the housekeeping gene DmRP140 by sequence comparison of *Drosophila melanogaster* and *Drosophila virilis*.** *Gene* 1991, **100:**155-162.
11. Bird AP: **DNA methylation and the frequency of CpG in animal DNA.** *Nucleic Acids Res* 1980, **8(7):**1499-1504.
12. Coulondre C, Miller JH, Farabaugh PJ, Gilbert W: **Molecular basis of base substitution hotspots in *Escherichia coli*.** *Nature* 1978, **274(5673):**775-780.
13. Ioshikhes IP, Zhang MQ: **Large-scale human promoter mapping using CpG islands.** *Nat Genet* 2000, **26:**61-63.
14. Auf der Maur A, Belser T, Elgar G, Georgiev O, Schaffner W: **Characterization of the transcription factor MTF-1 from the Japanese pufferfish (Fugu rubripes) reveals evolutionary conservation of heavy.** *Biol Chem* 1999, **380(2):**175-85.
15. Bird A: **DNA methylation de novo.** *Science* 1999, **286(5448):**2287-2288.
16. Larsen F, Gundersen G, Lopez R, Prydz H: **CpG islands as gene markers in the human genome.** *Genomics* 1992, **13(4):**1095-1107.
17. Ponger L, Mouchiroud D: **CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences.** *Bioinformatics* 2002, **18(4):**631-633.
18. Eisenberg E, Levanon E: **Human housekeeping genes are compact.** *Trends Genet* 2003, **19(7):**362-365.
19. Su A, Cooke M, Ching K, Hakak Y, Walker J, Wiltshire T, Orth A, Vega R, Sapinoso L, Moqrich A, Patapoutian A, Hampton G, Schultz P, Hogenesch J: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci U S A* 2002, **99(7):**4465-4470.
20. Bussey K, Kane D, Sunshine M, Narasimhan S, Nishizuka S, Reinhold WC, Zeeberg B, Ajay W, Weinstein J: **MatchMiner: a tool for batch navigation among gene and gene product identifiers.** *Genome Biol* 2003, **4(4):**R27.
21. Vinogradov A: **Isochores and tissue-specificity.** *Nucleic Acids Res* 2003, **31(17):**5212-20.
22. Wang T, Stormo G: **Combining phylogenetic data with co-regulated genes to identify regulatory motifs.** *Bioinformatics* 2003, **19(18):**2369-80.
23. Chargaff E: **Structure and function of nucleic acids as cell constituents.** *Fed Proc* 1951, **10:**654-659.
24. Louie E, Ott J, Majewski J: **Nucleotide frequency variation across human genes.** *Genome Res* 2003, **13(12):**2594-601.
25. Francino MP, Ochman H: **Strand asymmetries in DNA evolution.** *Trends Genet* 1997, **13(6):**240-245.
26. Beletskii A, Bhagwat AS: **Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*.** *Proc Natl Acad Sci U S A* 1996, **93(24):**13919-13924.
27. Oller AR, Fijalkowska IJ, Dunn RL, Schaaper RM: **Transcription-repair coupling determines the strandedness of ultraviolet mutagenesis in *Escherichia coli*.** *Proc Natl Acad Sci U S A* 1992, **89(22):**11036-11040.
28. FlyBase Consortium: **The FlyBase database of the Drosophila genome projects and community literature.** *Nucleic Acids Res* 2003, **31:**172-5.
29. Harris T, Chen N, Cunningham F, Tello-Ruiz M, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Chan J, Chen C, Chen W, Davis P, Kenny E, Kishore R, Lawson D, Lee R, Muller H, Nakamura C, Ozersky P, Petcherski A, Rogers A, Sabo A, Schwarz E, Van Auken K, Wang Q, Durbin R, Spieth J, Sternberg P, Stein L: **WormBase: a multi-species resource for nematode biology and genomics.** *Nucleic Acids Res* 2004, **32:**D411-7.
30. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30:**38-41.
31. [http://www.ensembl.org/EnsMart].
32. Shimizu T, Takahashi K, Tomita M: **CpG distribution patterns in methylated and non-methylated species.** *Gene* 1997, **205(1–2):**103-7.
33. Tatarinova T, Brover V, Troukhan M, Alexandrov N: **Skew in CG content near the transcription start site in Arabidopsis thaliana.** *Bioinformatics* 2003, **19(Suppl 1):**I313-I314.
34. Majewski J, Ott J: **Distribution and characterization of regulatory elements in the human genome.** *Genome Res* 2002, **12(12):**1827-36.
35. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Predicting gene regulatory elements in silico on a genomic scale.** *Genome Res* 1998, **8(11):**1202-1215.
36. Ohler U, Niemann H, Liao G, Rubin G: **Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition.** *Bioinformatics* 2001, **17(Suppl 1):**S199-206.
37. Blumenthal T, Evans D, Link C, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu W, Duke K, Kiraly M, Kim S: **A global analysis of Caenorhabditis elegans operons.** *Nature* 2002, **417(6891):**851-4.
38. von Mering C, Bork P: **Teamed up for transcription.** *Nature* 2002, **417(6891):**797-8.