

# Contributions of GC on gene expression: recognizing the roles of GC

**Arhondakis Stilianos**

*Laboratory of Molecular Evolution, Stazione Zoologica A. Dohrn, Naples, Italy*

How base composition affects mammalian gene expression is an issue of prime practical and evolutionary interest. In the past few years several groups have addressed the influence of base composition on transcription levels in mammalian genomes observed via genome-wide technologies (Affymetrix, SAGE, ESTs/cDNA libraries). Despite some variability among the reports, especially where they estimate a magnitude for this influence, a persisting trend has emerged: GC-rich genes tend to be expressed at higher levels than GC-poor genes.

Results from our laboratory using publicly available collections of EST data from cDNA libraries representing a wide range of tissues, demonstrated significantly higher counts for GC<sub>3</sub>-richer genes than for GC<sub>3</sub>-poorer genes in human. We also documented a large variability, which is partly a result of differences in experimental protocols that cDNA libraries are prepared, leading to strong GC biases. Conversely, it became clear that GC distributions can also be used as a quality control, to recognize biased libraries.

We also analyzed reported expression levels and genic GC using Affymetrix data from different human tissues, laboratories and array generations, to study technical and experimental effects that may propagate into observed correlation coefficients. Our examples show how correlations are affected by different (i) detection criteria, (ii) array versions, (iii) GC levels of probes (affinity contributions), and (iv) experimental procedures. Although such factors influence results, and are not always easy to discount with precision, our analysis confirms that the observed correlations between GC and gene expression remains robust, and cannot be explained by artefacts. Finally we were able to propose a conservative lower compositional border of the human transcriptomes', with mean GC<sub>3</sub> of coding transcripts' detected within a tissue typically above 55%.

Our results, and observations on technology-specific GC biases, could have important consequences for reproducibility and cross-comparison studies, as well as for differential gene expression studies, i.e., for the main use of genome-wide technologies in the biomedical community.