EMBO
reports

Genomes: the truth is in there

Yves Van de Peer

Genomes are the blueprints of all living organisms. From a meagre 500,000 base pairs in the bacterium *Mycoplasma* or a respectable 3.3 billion in humans, up to a dazzling 149 billion base pairs in the Japanese flower *Paris japonica*; genomes underpin the mystery of life. Eager to better grapple with that mystery, most biologists will spend time browsing journals for papers that present new genome sequences. We want to know the size of the genome and the number of predicted genes—although with every genome published we increasingly learn that these numbers do not matter much (Pertea & Salzberg, 2010).

In some cases, these aberrations in gene numbers can be ascribed to important evolutionary events, such as genome duplications. The discovery that most eukaryotic genome sequences contain remnants of such events is fascinating. We do not know why the ancestral vertebrate survived two genome duplications, or why fish have experienced an additional duplication on top of that. Would complex vertebrates, including us, have evolved if these two rounds of genome duplication had not occurred at the dawn of vertebrate evolution? Would our fish tanks look boring because many of the 25,000 species of teleosts would not have evolved if the fish-specific genome duplication had not occurred? Can we explain Darwin's 'abominable mystery' on the origin and diversification of flowering plants through genome-wide duplications that occurred early in their evolution? Although we cannot be certain of the answers to these questions, there is significant circumstantial evidence that both biological complexity and speciation events might have been compromised if genome-wide duplication events had not occurred.

In order to interpret correctly the timing and evolutionary consequences of events such as whole-genome duplications, it has become clear that we need adequate sampling of genome sequences. For example, soon after the sequencing of large parts of the *Arabidopsis* genome it became obvious that the genome had been duplicated. However, the timing and exact number of genome-wide duplications that had taken place could only be unveiled by comparison with other plant genomes. Comparison also revealed that the developmental genetic 'tool kit' of organisms is generally much older than expected, and this seems to be the case for both plants and animals. For instance, the common ancestors of land plants seem to have contained most of the gene families known to be important for angiosperm development, whereas the metazoan ancestor had a developmental tool kit similar to that of modern complex bilaterians. Interestingly, many developmental genes also seem to have evolved particularly through large-scale gene-duplication events.

The availability of several complete genome sequences and their comparative analyses have already improved our understanding of their composition, structure and evolution. However, the challenge of understanding the genotype–phenotype connection remains. We are getting better at determining the 'parts list' encoded in a genome sequence, including both protein coding and non-coding genes, which is a crucial first step (Gravely *et al*, 2010). For example, finding a complete set of meiotic genes will tell us that the organism is probably able to reproduce sexually, although this might not have been observed *in vivo* (Grimsley et al, 2010). However, we cannot yet predict from a genome sequence which genes are expressed when, where and in which quantities. Yet, elucidating the regulatory wiring of organisms is crucial to understanding how they manage to tune and coordinate their cellular functions in response to a range of diverse internal and external cues (modENCODE *et al*, 2010).

We know that organisms use *cis*-regulatory motifs in a combinatorial manner to generate different expression patterns, and there have been several attempts to predict gene expression from the type, number and organization of regulatory elements in promoters. Beer & Tavazoie (2004) used a probabilistic approach to discern the regulatory motifs and combinatorial rules underlying gene expression in *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. Nguyen & D'haeseleer (2006) developed a computational strategy to analyse systematically how genomic context influences the responsiveness of transcription-factor-binding sites. Others have used thermodynamic models to predict expression patterns, calculating the probabilities of all possible configurations of *trans* factors (including nucleosomes) on the *cis*-regulatory sequence, and summing their contributions to gene expression. Although these and other approaches have been met with criticism, we are getting better at predicting gene expression from the genomic sequence. Further computational dissection of the genome, informed by -omics data such as transcriptomics, ChIP-sequencing data and epigenomics can be expected to go a long way towards unravelling what makes an organism tick.

REFERENCES

Beer MA, Tavazoie S (2004) *Cell* **117:** 185–198

Gravely BR *et al* (2010) Nature [Epub 22 Dec] doi:10.1038/nature09715

Grimsley N *et al* (2010) *Mol Biol Evol* **27:** 47–54

Pertea M, Salzberg SL (2010) *Genome Biol* **11:** 206

Nguyen DH, D'haeseleer P (2006) *Mol Syst Biol* doi:10.1038/msb4100054

modENCODE *et al* (2010) *Science* **330:** 1787–1797

Yves Van de Peer is Professor of Bioinformatics and Genome Biology at Ghent University, Ghent, Belgium. E-mail: yves.vandepeer@psb.vib-ugent.be

EMBO *reports* (2011) **12,** 93.
doi:10.1038/embor.2011.1