Genome acrobatics: understanding complex genomes

‘The fifth base has been ignored for far too long.’

Stephan Beck, Head of Human Sequencing, The Sanger Centre, Cambridge, UK

Genomes are often portrayed as static DNA molecules that simply store genetic information. They are, however, capable of far more than that. Genomic DNA – arranged as chromosomes in humans – is a dynamic biomolecule with wide-ranging control over its own functionality. For instance, DNA can change the way it functions by undergoing conformational changes and/or chemical modifications. Such changes have profound functional implications which, among other aspects, we need to discern if we want to understand the biology of our genome in its full complexity. To do that, we need three major types of information: (1) a human genome reference sequence, (2) an index of human sequence variants and (3) an index of human sequence modifications.

The Human Genome Project has provided an annotated reference sequence, which presents the biggest advance in our understanding of complex genomes[1,2]. The comparison of DNA sequences from different individuals has provided many sites in the genome where the reference sequence shows base differences between individuals[3]. Such sequence variants, also known as single nucleotide polymorphisms (SNPs), are useful genetic markers and, in some cases, are the underlying cause of altered functionality and even disease. As yet, there is little information available on *in vivo* sequence modifications, such as the chemical alteration of bases. Modifications that do not result in a change of the actual DNA sequence are described as epigenetic. The reversible methylation of cytosines is currently the only known endogenous epigenetic modification in our genome and gives rise to a complex epigenome(s). Although there is only one (albeit diploid) genome per individual, which is the same in every normal nucleated human cell, there are many epigenomes per individual depending on tissue and/or cell type and developmental stage. The recently formed Human Epigenome Consortium aims to expand our limited knowledge of epigenetic sequence modifications[4].

Methylation variable positions (MVPs)

Epigenetic modifications, such as methylation, have a major role in genome function and stability. In addition, they show great potential as diagnostic markers for diseases. Cytosine methylation is involved in imprinting, gene regulation, chromatin structure and disease[5,6]. By comparing the global patterns of cytosine methylation, the consortium aims to generate a publicly available map of methylation variable positions (MVPs) of the human genome. MVPs are defined as positions of differential methylation status when compared in different

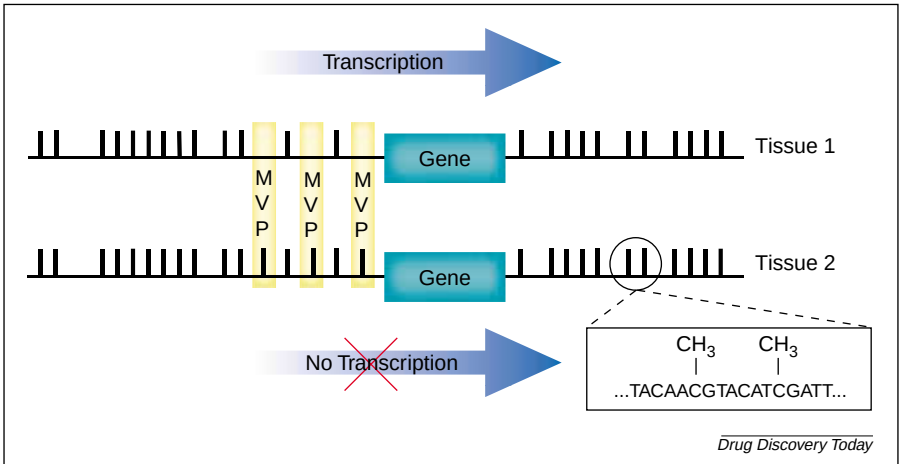

Figure 1. Schematic illustration of methylation variable positions (MVPs) and their possible effect on gene transcription. MVPs are defined as positions where the cytosine methylation status [indicated by the presence or absence of vertical bars (|) or (CH_3) predominantly at CpG dinucleotides] varies between tissues. Transcription can be affected by a single to a few MVPs in the regulatory region upstream of genes.

1359-6446/01/$ – see front matter PII: S1359-6446(01)02036-0

tissues and cell lines, usually of healthy versus diseased origin (Fig. 1).

The primary target sites for MVP discovery are CpG-rich gene promoter regions, which are often CpG islands. As illustrated in Figure 1, hypermethylation of such sites will down-regulate or silence transcription at that locus. If the locus happens to encode a tumour suppressor gene, this can lead to tumourigenesis and the corresponding MVPs are likely to be of diagnostic significance. Conversely, hypomethylation can lead to the inappropriate activation of genes in certain tissues or developmental stages, which can also result in genetic imbalance and disease. MVPs are best identified by DNA sequencing using the bisulfite sequencing method[7] but other techniques, based on chip technology and MS are also being developed. Treatment of genomic DNA with bisulfite deaminates unmethylated cytosines and converts them into uracils, which are scored as thymines (T) after sequencing. Methylated cytosines are not converted by bisulfite and are scored as cytosines (C).

The epigenome pilot project

The Epigenome Consortium reached its first milestone last year. Three of its members, The Sanger Centre (Cambridge, UK), Epigenomics AG (Berlin, Germany) and the Centre National de Genotypage (Evry, France), were jointly awarded a grant for an epigenome pilot study to generate a map of MVPs of the human major histocompatibility complex (MHC; Ref. 8). The MHC is among the most gene-dense regions in the human genome and is associated with numerous diseases, including most, if not all, autoimmune diseases[9]. It is probable that some of these diseases will have an epigenetic component. Together with the completed sequence and SNP maps, these data will provide a unique resource to tackle all MHC-associated diseases in the future.

Now that the sequencing of many reference genomes is either complete or well underway, the time has come to study their epigenomes too. The fifth base has been ignored for far too long already.

References

1 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
2 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
3 The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928–933
4 Beck, S. *et al.* (1999) From genomics to epigenomics: a loftier view of life. *Nat. Biotechnol.* 17, 1144
5 Robertson, K.D. and Wolffe, A.P. (2001) DNA methylation in health and disease. *Nat. Rev. Genet.* 1, 11–19
6 Special issue (2001) Epigenetics. *Science* 293, 1063–1105
7 Frommer, M. *et al.* (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* 89, 1827–1831
8 Hagman, M. (2000) Mapping a subtext in our genetic book. *Science* 288, 945–946
9 The MHC Sequencing Consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. *Nature* 401, 921–923

Stephan Beck
Head of Human Sequencing
The Sanger Centre
Wellcome Trust Genome Campus
Hinxton, Cambridge
UK CB10 1SA
tel: +44 1223 494952
fax: +44 1223 494919
e-mail: beck@sanger.ac.uk