Functional analysis of regulatory single-nucleotide polymorphisms

Sandra Pampín and José C. Rodríguez-Rey

Purpose of review

The identification of regulatory polymorphisms has become a key problem in human genetics. In the past few years there has been a conceptual change in the way in which regulatory single-nucleotide polymorphisms are studied. We revise the new approaches and discuss how gene expression studies can contribute to a better knowledge of the genetics of common diseases.

Recent findings

New techniques for the association of single-nucleotide polymorphisms with changes in gene expression have been recently developed. This, together with a more comprehensive use of the old in-vitro methods, has produced a great amount of genetic information. When added to current databases, it will help to design better tools for the detection of regulatory single-nucleotide polymorphisms.

Summary

The identification of functional regulatory single-nucleotide polymorphisms cannot be done by the simple inspection of DNA sequence. In-vivo techniques, based on primer-extension, and the more recently developed 'haploChIP' allow the association of gene variants to changes in gene expression. Gene expression analysis by conventional in-vitro techniques is the only way to identify the functional consequences of regulatory single-nucleotide polymorphisms. The amount of information produced in the last few years will help to refine the tools for the future analysis of regulatory gene variants.

Keywords

common diseases, gene expression analysis, promoter, regulatory SNPs, SNPs

Curr Opin Lipidol 18:194–198. © 2007 Lippincott Williams & Wilkins.

Department of Molecular Biology, Faculty of Medicine, University of Cantabria, Santander, Spain

Correspondence to José C. Rodríguez-Rey, Departamento de Biología Molecular, Facultad de Medicina, Avda. Cardenal Herrera Oria s/n. 39011, Santander, Spain
Tel: +34 942 201953; fax: +34 942 201945; e-mail: rodriguj@unican.es

Current Opinion in Lipidology 2007, 18:194–198

Abbreviations

EMSA	electrophoretic mobility shift assay
rSNP	regulatory single nucleotide polymorphism
SNP	single nucleotide polymorphism
TFBS	transcription factor binding site

0957-9672

Introduction

During the past decades genetics has contributed to the discovery of many gene variants associated with human disease. In October 2006 65 251 of these variants had been annotated in the *Cardiff Human Genome Mutation Database* (website: http://archive.uwcm.ac.uk/uwcm/mg/docs/hohoho.html). Most are responsible for monogenic disorders, which are typically caused by a group of rare mutations in the same gene. Familial hypercholesterolemia is a good example. In the year 2002 more than 890 mutations in the LDLR gene had been described. More than 90% are point mutations and the great majority map in the coding region of the gene [1]. In spite of this impressive effort, the understanding of the genetic basis of diseases as common as cardiovascular disease or diabetes remains elusive.

The common disease/common variant hypothesis predicts that the genetic risk for common diseases will be caused by susceptibility alleles present with high frequencies within the population [2,3]. On the other hand, the variants leading to increased susceptibility to common diseases usually produce a mild effect in the phenotype. Some of these mutations have been found in coding sequences. The three well known apoE alleles are a good example. Alleles ε2, ε3 and ε4 encode proteins with different biochemical properties. Having one of these alleles is not enough for producing a disease phenotype, but they are one of the major genetic contributors to the determination of plasma cholesterol levels [4]. Also the Pro12Ala allele of the *PPARG* gene has been found to be associated to type II diabetes [5]. In general, however, these two characteristics, high frequency and mild phenotype, are hardly seen together in coding mutations.

If structural changes in proteins do not suffice to explain common phenotypes maybe their abundance does. The importance of variation in noncoding cis-regulatory regions in the evolution of primate phenotypes had been suggested a long time ago [6]. A recent survey of 140 polymorphisms, previously validated by in-vitro techniques and involved in the regulation of 107 human genes (at the time of the study more than 1% of the named human genes), revealed that variation affecting gene expression is widespread in the human genome. In fact, humans are more polymorphic at functional regulatory sites than they are at coding sequences [7]. In

addition, regulatory single nucleotide polymorphisms (rSNPs) are more likely to produce mild phenotypes. Again, the *LDLR* gene constitutes a good example. Several rSNPs have been published. All of them are present in familial hypercholesterolemia families with no other known change in the *LDLR* gene [8–12]. Most interestingly, some of them are associated with a mild familial hypercholesterolemia phenotype [11,12].

The mild phenotype associated with rSNPs is not the only reason they are so difficult to identify. Promoters are located just upstream of the transcription initiation site and thus they are very easy to spot. Two regions very distant in a linear chromosome can be in contact, however, because of the changes introduced by the compacting of DNA into chromatin. It is not unusual to find control elements very distant from the coding sequence like the locus control region of the globin gene cluster or the region controlling the tissue-specific expression of the *apoE* gene [13,14].

Methods for the quantificaton of allele-specific gene expression

Today a large number of rSNPs have been identified by in-vitro methods. In order to better understand the role of these variants it would be very interesting to find associations between them and changes in gene expression *in vivo*. For that purpose several methods have been designed for the in-vivo analysis of allele-specific gene expression. The single nucleotide primer extension (SNuPE), originally developed for the detection of mutant alleles [15], was subsequently validated for the study of the most extreme case of allele-specific gene expression: imprinted genes. Briefly, the transcripts of both alleles are amplified with a pair of primers flanking the single nucleotide polymorphism (SNP). For determining the relative amount of each allele an oligonucleotide, whose 3′ end is located just before the SNP, is extended in two separated reactions, each one with the radioactive dNTP corresponding to each allele. The ratio between the incorporated radioactivity of both reactions represents the ratio of allele-specific expression [16]. A good example of the application of this method is the study of 13 genes in 96 lymphoblastoid cells. Originally the allele quantification was done with the ABI Prism SNAPshot Multiplex (Applied Biosystems, Foster City, California, USA) [17], a method initially developed for rapid genotyping of pooled samples [18]. A subsequent high-throughput coupled SNuPE with hybridization with Affymetrix HuSNP arrays (Santa Clara, California, USA) [19]. The method has great potential and permitted the acquisition of a great volume of data useful to perform complex genetic analysis. It was successfully applied for the study of the expression of 3554 genes in lymphoblastoid cells from 14 large families. The analysis revealed that the patterns of gene expression are inherited thus underscoring the role of rSNPs in the determination of common phenotypes [20].

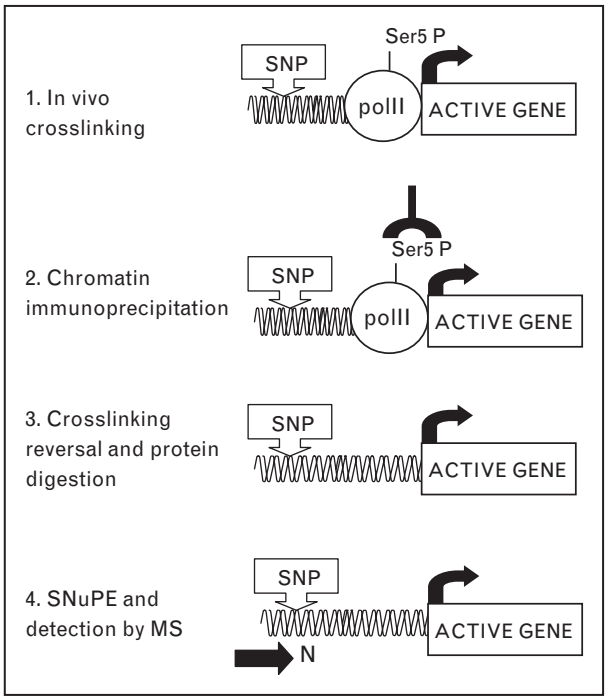
The methods of analysis of allele-specific gene expression based on single nucleotide primer extension perform the analysis on the RNA and therefore cannot discriminate between alleles of those SNPs not located within the transcript. This is a very important limitation if we consider that it excludes any sequence 5′ upstream of the transcription initiation site, that is, most regulatory sequences.

In order to overcome this problem, a very elegant method for the analysis of allele-specific gene expression (haploChIP) was described by Knight *et al.* [21]. The haploChIP method makes use of one of the changes, which accompanies gene expression. In the transition from inactive to actively transcribed genes, the RNA polymerase II (the enzyme which transcribes protein-coding genes) leaves the site where the preinitiation complex was previously formed and moves along the DNA. The event is accompanied by structural changes in the polymerase molecule. One of these changes affects the carboxy-terminal domain of the enzyme. The carboxy terminal domain is a serine-rich domain, which is phosphorylated when the polymerase leaves the initiation site. The phosphorylation is specific for some serines, Ser5 among them, and specific antibodies against phospho-Ser5 recognize only the polymerase bound to the genes being actively transcribed. The method is summarized in Fig. 1. Briefly, proteins are cross-linked to DNA and chromatin is broken down to small pieces and immunoprecipitated with an antiphospho-Ser5 antibody. After reversal of the crosslink, the allele-specific gene expression is analysed by primer extension and mass spectrometry. The method has the advantage of the study of the genes in a natural environment and allows the functional association of DNA variants in the promoter with changes in gene expression. Unfortunately it is not possible to attribute these differences to a particular SNP because it is not uncommon for more than one SNP to exist in a particular regulatory region and even in those cases in which only one SNP is present, the effect of a distant SNP cannot be ruled out. Confirmatory experiments in which isolated SNPs can be studied (see in-vitro assays) are needed [22••].

Functional in-vitro assays for the study of gene expression

The reporter gene assay has been the most widely used method for the study of promoter strength. Briefly, the promoter is cloned directly upstream of the reporter gene in a promoterless plasmid vector. The plasmid is then introduced into cultured cells. Most cells do not integrate the gene into the chromosome but the reporter gene is expressed in the extra chromosomal state. Quantification of protein activity (or amount) is done before 72 h after

Figure 1 In-vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading



The method described by Knight *et al.* [21] allows in-vivo measurement of the differences in allele-specific gene expression. Protein bound to DNA is crosslinked with formaldehyde and chromatin fragments are produced by sonication. The DNA–protein complex is immunoprecipitated with an antibody specific for a phosphorylated serine in the carboxy terminal domain of RNA PolII. Subsequent primer extension is used for determining the levels of allele-specific expression. For that purpose biotinylated oligos located immediately upstream of the single nucleotide polymorphism (SNP) are used. Quantification is done by mass spectrometry. SNuPE, single nucleotide primer extension.

transfection and gives an accurate estimation of the activity of the promoter/regulatory region. In order to avoid interferences from endogenous genes, reporter genes isolated from different organisms are used. The bacterial gene encoding the enzyme chloramphenicol acetyl transferase, an enzyme responsible for resistance to the antibiotic chloramphenicol, was the standard only a few years ago [23]. Now it is being replaced by the luciferase gene of the firefly (*Photinus pyralis*) [24], which can be quantified using a luminometer over a broader linear range (typically five to six orders of magnitude).

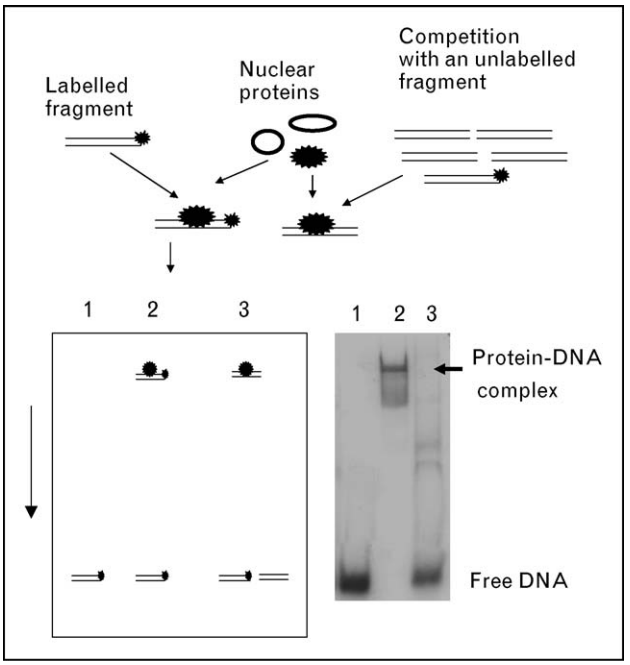
Achieving a good transfection efficiency is of great importance in order to get reproducible results. Initial methods, calcium phosphate [25] and DEAE-dextran [26] rely on endocytosis for the uptake of DNA by the cells. Both are still used for adherent cells. For cells in suspension, electroporation (i.e. the entrance of DNA through pores in the cell membrane induced by a high voltage electroshock) is also a good choice [27,28]. Lipofection makes use of the fusion of DNA cationic lipids complexes to the cell membrane [29]. Different lipids from several manufacturers are marketed and not all of them work equally well with different cell lines, so efficiency tests need to be carried out in advance.

As differences in transfection efficiency result in changes in the activity of the reporter gene, the results must be normalized by the introduction of an internal control of transfection. This is achieved by the co-transfection of another plasmid in which a different reporter gene is cloned downstream of a strong promoter (CMV immediate–early and pSV40 early promoters are the most commonly used). For the control any reporter gene (as long as it is different from the one in the test plasmid) can be used but the dual luciferase assay [30] has become the standard. The test promoter is cloned upstream of the *Photinus pyralis* luciferase. Another luciferase gene from the sea pansy (*Renilla reniformis*) [31] is used as transfection control. Because the substrates of *Renilla* and *Photinus* luciferases are different, the quantification of luminescence due to each luciferase can be performed without the need for dividing samples.

The gene reporter method is very sensitive and the results are typically very consistent. Differences as low as 20% in promoter activity have been reported [32]. Its principal virtue is that the effect of isolated SNPs can be assayed. Because the promoter is not in its natural chromatin environment, however, and because of the different behaviour of cells in culture, the results are sometimes difficult to correlate to in-vivo observations. Since the introduction of the method in the 1980s, the gene reporter assay has been successfully applied to the study of a large number of SNPs in promoters. Only recently, however, has a large number of SNPs been studied simultaneously. Hoogendoorn *et al.* [33] described the search for SNPs in the proximal 500 bp of 170 promoters selected from the *Eukaryotic Promoter Database*. Thirty-five percent of the promoters contained at least one SNP. Subsequent gene reporter assays revealed that around a third of these variants might significantly alter gene expression. Another study by the same group screened for polymorphisms 56 genes previously reported to be differentially expressed in the brains of schizophrenics. Of a total of 54 sequence variants represented in the haplotypes, 12 (about 22%) resulted in functional changes [34]. Most interestingly, the functional mutations are not randomly located in the promoter. As shown in a study of 247 gene promoters 50% of them are clustered in the proximal 100 bp. Incidentally, only 33% of the functional variants were located in a consensus transcription factor binding site (TFBS) [35•].

Altering the affinity transcription factors to DNA by mutations in their binding sites (TFBS) is the most common way in which SNPs can alter gene expression

Figure 2 Electrophoretic mobility shift assay



A short double-stranded oligonucleotide is radioactively labelled and mixed with a nuclear extract containing the transcription factors, which specifically bind to the sequence. The complex is stable in a nondenaturing PAGE, allowing resolution of protein–DNA complexes. Typically, a labelled band with no protein extract (lane 1) is included in order to know the position of noncomplexed DNA. The extra bands in lane 2 correspond to protein–DNA complexes. When an excess of unlabelled oligonucleotide is added to the reaction, the labelled oligo is displaced from the complexes and the extra bands cannot be seen (lane 3).

(for a discussion on how a SNP can affect gene expression by changing DNA topology see Buckland [36••]). Ideally bioinformatics tools should be able to identify TFBS and to discern when a change abolishes the binding. Apart from TFBS sequence, however, there must be some unknown factors that contribute to the binding as transcription factors tolerate a relatively high degree of variation in the TFBS. Besides, only a small part of the TFBS are known, as indicated by the experiments mentioned above [35•]. Therefore the application of in-silico tools is still very limited and accompanying laboratory assays are still needed.

The electrophoretic mobility shift assay (EMSA) is the major method for detecting binding of proteins to DNA. The method is depicted in Fig. 2. Briefly, a labelled double-stranded oligonucleotide (20–25 bp in size) is mixed with a nuclear extract prepared from cells that express the transcription factors. In the presence of transcription factors a complex DNA–protein is formed. Low salt conditions and the cage effect of the gel matrix help to stabilize the complex during electrophoresis. The formation of the protein–DNA complex results in a retardation of mobility and in a separation from the free probe. The specificity of the binding is enhanced by the addition of synthetic polymers such as poly dI–dC and parallel competition assays are carried out as controls of the specificity.

The EMSA is a simple assay and very powerful in combination with bioinformatics. A good example is the study of the promoters of 176 genes coding for G-protein coupled receptors (GPCRs). As a first step, the proximal 5 kb regions were screened for SNPs. The result was the finding of approximately 800 SNPs. Assuming that regions conserved between species most likely mediate biological functions, a second round of selection was carried out using human–mouse conservation as a major selection criteria. Out of the remaining 200 SNPs, 36 were predicted to result in altered binding. Ten of them were selected for EMSA and seven resulted in changes of electrophoretic mobility [37•]. A similar experiment carried out in our laboratory indicated that approximately 80% of the SNPs in which a change in the binding has been predicted *in silico* actually produce changes in the mobility (Pampin *et al.*, in preparation). The combination of EMSA and bioinformatics can be used as a previous filter for selecting SNPs for the more time and labour-consuming reporter gene assays.

Conclusion

The identification of regulatory polymorphisms has become a key problem in human genetics. Coding polymorphisms can be identified *in silico* by sequence inspection and are thus amenable to high-throughput strategies. Conversely, the identification of functional rSNPs is a very laborious task. Considering the existence of several million SNPs it is clear that the assignation of functional significance can only be done by in-silico methods. At present, the application of bioinformatics tools to the identification of rSNPs often gives poor results. By the application of the techniques described in this article many functional polymorphisms might be identified during the next few years. Hopefully the identification of these rSNPs and their addition to the already existing databases will help to improve the bioinformatic tools which in turn will help to elucidate the genetic basis of common diseases.

Acknowledgements

This work has been supported by a grant of the Spanish Fondo de Investigaciones Sanitarias (PI061068 to JCR).

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

Additional references related to this topic can also be found in the Current World Literature section in this issue (pp. 202–203).

1 Villeger L, Abifadel M, Allard D, *et al.* The UMD-LDLR database: additions to the software and 490 new entries to the database. Hum Mutat 2002; 20:81–87.

2 Lander ES. Genomics: launching a revolution in medicine. J Law Med Ethics 2000; 28:3–14.

3 Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet 2001; 17:502–510.

4 Davignon J, Gregg RE, Sing CF. Apolipoprotein E polymorphism and atherosclerosis. Arteriosclerosis 1988; 8:1–21.

5 Altshuler D, Hirschhorn JN, Klannemark M, *et al.* The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 2000; 26:76–80.

6 King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science 1975; 188:107–116.

7 Rockman MV, Wray GA. Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol 2002; 19:1991–2004.

8 Dedoussis GV, Pitsavos C, Kelberman D, *et al.* FH-Pyrgos: a novel mutation in the promoter (−45delT) of the low-density lipoprotein receptor gene associated with familial hypercholesterolemia. Clin Genet 2003; 64:414–419.

9 Francova H, Trbusek M, Zapletalova P, Kuhrova V. New promoter mutations in the low-density lipoprotein receptor gene which induce familial hypercholesterolaemia phenotype: molecular and functional analysis. J Inherit Metab Dis 2004; 27:523–528.

10 Koivisto UM, Palvimo JJ, Janne OA, Kontula K. A single-base substitution in the proximal Sp1 site of the human low density lipoprotein receptor promoter as a cause of heterozygous familial hypercholesterolemia. Proc Natl Acad Sci U S A 1994; 91:10526–10530.

11 Mozas P, Galetto R, Albajar M, *et al.* A mutation (−49C>T) in the promoter of the low density lipoprotein receptor gene associated with familial hypercholesterolemia. J Lipid Res 2002; 43:13–18.

12 Sun XM, Neuwirth C, Wade DP, *et al.* A mutation (T-45C) in the promoter region of the low-density-lipoprotein (LDL)-receptor gene is associated with a mild clinical phenotype in a patient with heterozygous familial hypercholesterolaemia (FH). Hum Mol Genet 1995; 4:2125–2129.

13 Li Q, Peterson KR, Fang X, Stamatoyannopoulos G. Locus control regions. Blood 2002; 100:3077–3086.

14 Dang Q, Walker D, Taylor S, *et al.* Structure of the hepatic control region of the human apolipoprotein E/C-I gene locus. J Biol Chem 1995; 270:22577–22585.

15 Kuppuswamy MN, Hoffmann JW, Kasper CK, *et al.* Single nucleotide primer extension to detect genetic diseases: experimental application to hemophilia B (factor IX) and cystic fibrosis genes. Proc Natl Acad Sci U S A 1991; 88:1143–1147.

16 Singer-Sam J, Chapman V, LeBon JM, Riggs AD. Parental imprinting studied by allele-specific primer extension after PCR: paternal X chromosome-linked genes are transcribed prior to preferential paternal X chromosome inactivation. Proc Natl Acad Sci U S A 1992; 89:10469–10473.

17 Cheung VG, Conlin LK, Weber TM, *et al.* Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet 2003; 33:422–425.

18 Norton N, Williams NM, Williams HJ, *et al.* Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. Hum Genet 2002; 110:471–478.

19 Lo HS, Wang Z, Hu Y, *et al.* Allelic variation in gene expression is common in the human genome. Genome Res 2003; 13:1855–1862.

20 Morley M, Molony CM, Weber TM, *et al.* Genetic analysis of genome-wide variation in human gene expression. Nature 2004; 430:743–747.

21 Knight JC, Keating BJ, Rockett KA, Kwiatkowski DP. In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. Nat Genet 2003; 33:469–475.

22 •• De Gobbi M, Viprakasit V, Hughes JR, *et al.* A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science 2006; 312:1215–1217.
This is a fine example of the combination of chromatin immunoprecipitation and classical functional analysis of gene expression.

23 Gorman CM, Moffat LF, Howard BH. Recombinant genomes which express chloramphenicol acetyltransferase in mammalian cells. Mol Cell Biol 1982; 2:1044–1051.

24 de Wet JR, Wood KV, DeLuca M, *et al.* Firefly luciferase gene: structure and expression in mammalian cells. Mol Cell Biol 1987; 7:725–737.

25 Graham FL, van der Eb AJ. Transformation of rat cells by DNA of human adenovirus 5. Virology 1973; 54:536–539.

26 McCutchan JH, Pagano JS. Enchancement of the infectivity of simian virus 40 deoxyribonucleic acid with diethylaminoethyl-dextran. J Natl Cancer Inst 1968; 41:351–357.

27 Neumann E, Schaefer-Ridder M, Wang Y, Hofschneider PH. Gene transfer into mouse lyoma cells by electroporation in high electric fields. Embo J 1982; 1:841–845.

28 Zimmermann U. Electric field-mediated fusion and related electrical phenomena. Biochim Biophys Acta 1982; 694:227–277.

29 Felgner PL, Gadek TR, Holm M, *et al.* Lipofection: a highly efficient, lipid-mediated DNA-transfection procedure. Proc Natl Acad Sci U S A 1987; 84: 7413–7417.

30 Sherf B, Navarro S, Hannah R, Wood K. Dual-Luciferase™ reporter assay: an advanced co-reporter technology integrating firefly and renilla luciferase assays. Promega Notes 1996; (57):2–9.

31 Matthews JC, Hori K, Cormier MJ. Purification and properties of Renilla reniformis luciferase. Biochemistry 1977; 16:85–91.

32 Yang WS, Nevin DN, Iwasaki L, *et al.* Regulatory mutations in the human lipoprotein lipase gene in patients with familial combined hyperlipidemia and coronary artery disease. J Lipid Res 1996; 37:2627–2637.

33 Hoogendoorn B, Coleman SL, Guy CA, *et al.* Functional analysis of human promoter polymorphisms. Hum Mol Genet 2003; 12:2249–2254.

34 Buckland PR, Hoogendoorn B, Guy CA, *et al.* A high proportion of polymorphisms in the promoters of brain expressed genes influences transcriptional activity. Biochim Biophys Acta 2004; 1690:238–249.

35 • Buckland PR, Hoogendoorn B, Coleman SL, *et al.* Strong bias in the location of functional promoter polymorphisms. Hum Mutat 2005; 26:214–223.
This is very important when designing future searches of regulatory polymorphisms.

36 •• Buckland PR. The importance and identification of regulatory polymorphisms and their mechanisms of action. Biochim Biophys Acta 2006; 1762:17–28.
This is a very good review on the basics of regulatory polymorphisms.

37 • Mottagui-Tabar S, Faghihi MA, Mizuno Y, *et al.* Identification of functional SNPs in the 5-prime flanking sequences of human genes. BMC Genomics 2005; 6:18.
This describes a simple, but efficient method for the bulk analysis of rSNPs.