Molecular Biology I: Transcriptional Regulation

Dr Nick Plant

Centre for Toxicology, Faculty of Health and Medical Sciences, University of Surrey, Guildford, Surrey, GU2 7XH, UK

Email: N.Plant@Surrey.ac.uk

Nick Plant is a Reader in Molecular Toxicology within the Centre for Toxicology at the University of Surrey, UK. His research interest include the use of systems modelling to understand the role of nuclear receptors in co-ordinating body responses to therapeutic drugs

Words (Total): 2429

Words (Abstract): 202

Figures: 2 (300 word equivalent)

Tables: 0

Keywords

Transcription; PCR; TaqMan; microarray; Quantification; ChIP; EMSA

Conflicts of Interest

None declared

Abstract

Whereas the DNA of the cell may be envisaged as the blueprints for a human, it is the processes of transcription and translation that act as the production line to convert these blueprints into the active protein required to produce the complex biological interactions that allow a cell to function. At its very simplest level the process of transcription is concerned with the recruitment of RNA polymerase II (Pol II) to the transcription start site of the target gene: The faster this process occurs then the more transcripts can be produced in a given amount of time and hence the more mRNA is produced to be converted to protein. This rate of Pol II recruitment can be modified through the action of a number of DNA-binding proteins (transcription factors), such as the nuclear receptors that are key sensors for the presence of therapeutic drugs within the cell

In the current article we will examine the techniques available for examining the DNA:protein interactions within the regulatory regions of target genes, which result in the increased recruitment of RNA polymerase II. In addition, we will examine those techniques designed to measure the next stage, the transcription of gene coding regions to make mRNA transcripts.

General Introduction

The human body is constantly exposed to a milieu of chemicals, including therapeutic drugs, environmental contaminants and food products. As such, it must be able respond to these changes in chemical level, preventing toxicity. In addition, it is important to maintain homeostasis for a large number of endogenous chemicals, such as the steroid hormones, ensuring healthy physiology. A key mechanism for the body's ability to respond to fluctuating chemical levels is the transcriptional control of genes whose protein products are involved in the absorption, distribution, metabolism and excretion (ADME) of both endogenous and foreign chemicals.

At the simplest level, the level of transcription of a gene is controlled by the rate of Pol II recruitment to the transcription start site (Figure 1). This rate of recruitment can be modified by the action of transcription factors (TFs); proteins that interact with specific sequences within the regulatory regions of a gene and act as a molecular bridge to grab Pol II and recruit it to the transcription start site. These TFs may be divided into two broad categories; firstly, the mere presence of *generic TFs* (e.g. Sp1, TBP) is sufficient to encourage recruitment of Pol II. Secondly, *ligand-activated TFs* (e.g. the nuclear receptors; PXR, PPAR, GR) respond to the presence of an activating chemical, and then increase the rate of recruitment of both generic TFs and Pol II. These two systems represent the *generic* and *responsive* modes of gene expression, with the former being important for basal expression of genes, whereas the latter is central to the body's response to altered chemical levels, such as is seen during pathophysiology or therapeutic treatment (1, 2). To be able to understand gene expression it is therefore necessary to be able to measure the protein:DNA interactions that occur within gene regulatory regions, as well as the resultant increase in mRNA transcript production rate.

DNA:Protein Interactions

The prerequisite for DNA transcription is the interaction of proteins with the regulatory regions of the target gene. As described above, this can be as simple as the recruitment of Pol II to the transcription start site, or the complex interaction of multiple proteins within a regulatory region that ultimately results in the efficient recruitment of Pol II. Assays that investigate these protein:DNA interactions range from the extremely simple, but potentially flawed, to highly complex, but biologically relevant, and we shall an example of each in turn.

One of the simplest assays for DNA:Protein interactions is the Electromobility Shift Assay (EMSA), alternatively referred to as the gel retardation assay. There is an obvious difference in both mass and bulk for protein-free DNA versus protein-bound DNA. Hence, when separated by electrophoresis they will migrate at different rates through the resolving gel, and this is the principle behind EMSA. Initially, a short region of DNA (20-50bp), where you think your protein may bind, is labelled for easy detection; common labels include radioactivity, fluorophores or biotinylated colourimetric markers. This DNA is then incubated with protein for 30 minutes to allow interaction, and then separated by polyacrylamide gel electrophoresis. After electrophoresis, the DNA is detected and its movement relative to protein-free DNA observed. Protein:DNA interactions will cause the DNA to be 'retarded' and move less distance through the gel (Figure 2). However, it is important to demonstrate that any observed interaction is actually specific and not the result of a 'sticky-protein', which would bind to any piece of DNA regardless of the biological relevance. To check for specificity a competition assay is carried out, whereby excess unlabelled probe DNA is added to the labelled DNA/protein mix during the incubation step. If the interaction is specific then binding to both the labelled and unlabelled probe DNA will be equally likely; as the unlabelled DNA is in excess this will effectively compete out the labelled DNA and the retarded band will disappear. In comparison, if the interaction is non-specific then the band will not be competed out as efficiently by the unlabelled probe DNA and the retarded DNA will remain. A final check for specificity that is often carried out is the so-called super-shift assay. Here an antibody to the target protein is also added to the labelled DNA and protein in the incubation mix. Obviously, the interaction of all three will result

in an even larger complex that will be even further retarded in the gel, and as the antibody is specific for the target protein this super-shifted band will only appear if the interacting protein is indeed the protein of interest(3).

The major limitation of the EMSA is that it examines protein:DNA interactions when the DNA is in a completely naked state. In vivo, DNA is wound around histones to produce *chromatin*, and this can have a significant impact on transcription levels: When DNA is tightly bunched (heterochromatin), this can prevent TFs from reaching their binding sites, interacting with DNA, and altering transcription rates of target genes. In contrast, euchromatin represent the unwound state of chromatin, which is most permissive to protein:DNA interactions and transcriptional regulation. Therefore, an EMSA will tell you if there is the potential for a protein:DNA interaction, but not if it is likely to occur in vivo; to examine this we need to use a more complex assay such as chromatin immunoprecipitation (ChIP), which studies protein:DNA interaction in the chromatin context.

In ChIP, interacting proteins are first fixed to genomic DNA using formaldehyde and then the genomic DNA broken into small (<500bp) fragments, usually by sonication. Using an antibody to the protein of interest any DNA bound to this protein is isolated through immunoprecipitation, followed by reversal of the formaldehyde cross links to free the DNA. Each fragment of isolated DNA represents a piece of DNA to which the target protein was bound and through sequencing or PCR these regions can be identified, telling you where in the genome your protein bound to the DNA. Carrying out this analysis for several different proteins, including histones, generic and responsive transcription factors, as well as Pol II itself, lets you build up a picture of the interactions occurring in regulatory regions of genes *in vivo* (3).

As mentioned previously, ChIP is a far more complex technique, being more technically demanding, time consuming and expensive than EMSA. However, much work has been carried out to reduce this burden, and the recent development of 'ChIP-on-chip' allows the study of many different protein:DNA interactions, in the chromatin context, at one time (4).

Measurement of Transcription

Both EMSA and ChIP technologies will provide an indication of whether particular proteins can interact with a region of DNA, and under what conditions. However, they do not directly measure if this interaction leads to an alteration in RNA transcription. *Reporter gene assays* allow the study of transcription in a naked DNA format, sharing the same ease-of-use but lack of applicability to in vivo that we saw with EMSA. By contrast, *qRT-PCR* and *DNA microarrays* allow us to study transcription in cell system, thus incorporating chromatin effects.

For a reporter gene assay it is first necessary to attach the regulatory region of interest (up to several kilobases of DNA) upstream of an easily measurable reporter gene, such as fluorescent proteins or enzymes which convert chemicals for subsequent colourimetric/luminescent detection. This construct is contained within a plasmid, a small circular piece of self-replicating DNA, which can be transfected into a mammalian cell line of choice for your assay. The fact that this assay occurs within a mammalian cell line is a major advantage over the EMSA as it puts the experiment into an environment that is much closer to the *in vivo* situation and hence increases confidence in the answer produced. In the basic assay, your reporter construct is transfected into the cell line of interest and then transcription stimulated through the addition of the protein of interest, a stimulating chemical, or a combination of both. Following stimulation for 12 to 72 hours the protein coded for by the reporter gene is measured, with an increase in production being indicative of transcriptional activation (5). This system thus represents a simple method to study the impact of external stimuli of transcription of a target gene, albeit with the caveat that these effects are not observed in the context of chromatin.

To measure transcription occurring from DNA in the chromatin context, two approaches are commonly used. As we will see the first measures the levels of a single mRNA transcript, whereas

the latter allows measurement of thousands of different mRNA transcripts in parallel, allowing information on the transcription rates within a whole cell to be derived (the transcriptome). Polymerase chain reaction (PCR) is a simple technique using primers specific to a target region of DNA to achieve its exponential amplification. If PCR is coupled with an initial conversion of RNA into DNA, through the use of a reverse transcriptase enzyme, then it is capable of measuring the levels of RNA within a cell (RT-PCR). Finally, if this measurement is undertaken in a robust manner and compared to a standard curve of known amounts of target RNA/DNA then this will produce a fully quantitative measurement of transcript levels within the cell: qRT-PCR, often shortened to qPCR. For qPCR two commonly used detection systems exist, Sybr-green and TaqMan. Sybr-green is a fluorescent chemical that intercalates into the major groove of DNA in a stochiometric fashion, meaning that the level of fluorescence is directly related to the level of DNA in the solution. The benefit of this approach is its relative low cost, but as Sybr-green binds to DNA non-specifically then extra care must be taken to ensure that you are only measuring the level of your target transcript and not of anything else. In comparison the 5'nuclease assay (TaqMan) uses a specific probe against the target transcript that contains a fluorophore and quencher. Each amplification round degrades the probe, freeing the flurophore in a manner directly proportional to the level of target transcript. Through the use of a specific probe, TaqMan has the benefit of guaranteed high specificity towards your target transcript, but this increased specificity comes at an increased cost per reaction (1).

Whereas qPCR assays determine the level of only a single, or very few, mRNA transcripts DNA microarrays are a capable of measuring the level of *every* mRNA transcript within the cell simultaneously. Such technology has the power to examine whole transcriptome differences between normal and tumour tissue (6), normal and chemically stimulated tissue (7) or for mapping mRNA transcript levels across a large range of samples/tissues (8, 9). Each microarray is comprised of thousands of spots of DNA, each of which is specific for a target mRNA transcript. RNA is extracted from the cell/tissue of interest and then labelled with a fluorescent dye for later detection. This mixture of labelled mRNAs is then hybridized to the microarray, and brighter fluorescence for a

given spot reflects more of that mRNA transcript being present in the original RNA mixture. Such techniques can be performed as one- or two-colour arrays: In a one-colour array RNA from both normal and stimulated cells are labelled with the same fluorescent marker and then hybridised to different microarrays. Each is quantified separately and the results compared to see differential expression. In a two-colour system, normal and stimulated cells are mixed with different fluorescent markers, usually red and green, and then hybridised to the same microarray. If an mRNA transcript is present at the same level in both normal and stimulated cells then the resultant spot will be orange; if more mRNA is present in the normal sample then the green fluorophore will predominant and the spot will be green; more mRNA present in the stimulated sample will produce a red spot. In essence, both the one- and two-colour systems provide the same answer, the relative expression of a mRNA transcripts in one sample compared to another, but use subtly different procedures to achieve this.

An exciting development that allows data generation exceeding even that of high throughput DNA microarrays are *next-generation sequencers*. With these machines it is possible to sequence entire genomes in a fraction of the time, and cost, of traditional technologies meaning that projects such as the 1000 genome project (http://www.1000genomes.org/), which aims to produce a repository of genetic variation, are possible (3, 10). RNA-Seq is the application of these technologies to the measurement of mRNA levels, and is set to revolutionise our ability to produce transcriptome-level mRNA profiles in the next few years (11).

Conclusions

For the body to function, and for it to be able for it to respond to the alterations away from homeostasis caused by pathogenesis or chemical exposure, it is vital that there is a flexible, responsive system to allow the expression of proteins as and when required by the cell. Control of transcription is the first stage of this response network, whereby information from the body's blueprints (DNA) is turned into the mRNA transcripts required to produce the basic machine parts for body functioning, the proteins. The next stage in this response system is the conversion of these

individual proteins into the complex, multi-protein machines required to make the body run, which will be examined in the next article.

References

1. Plant N. Molecular Toxicology. London: BIOS; 2003.

2. Plant N, Aouabdi S. Nuclear receptors: the controlling force in drug metabolism of the liver? Xenobiotica. 2009;39(8):597-605.

3. Dale JW, von Schantz M, Plant N. From Genes to Genomes: Concepts and Applications of DNA Technology. 3 ed. Oxford: Wiley-Blackwell; 2011.

4. Sugii S, Evans RM. Epigenetic codes of PPAR gamma in metabolic disease. FEBS Letters. [Review]. 2011 Jul;585(13):2121-8.

5. El-Sankary W, Gibson GG, Ayrton A, Plant N. Use of a reporter gene assay to predict and rank the potency and efficacy of CYP3A4 inducers. Drug Met Disp. 2001;29(11):1499-504.

6. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA. 2001;98(26):15149-54.

7. Pogribny IP, Bagnyukova TV, Tryndyak VP, Muskhelishvili L, Rodriguez-Juarez R, Kovalchuk O, et al. Gene expression profiling reveals underlying molecular mechanisms of the early stages of tamoxifen-induced rat hepatocarcinogenesis. Toxicol Appl Pharm. 2007 Nov 15;225(1):61-9. .

8. Plant KE, Everett DM, Gibson GG, Lyon J, Plant NJ. Transcriptomic and phylogenetic analysis of Kpna genes: a family of nuclear import factors modulated in xenobiotic-mediated liver growth. Pharmacogenet Genomics. 2006 Sep;16(9):647-58.

9. Plant N. Can systems toxicology identify common biomarkers of non-genotoxic carcinogenesis? Toxicol. [Journal article]. 2008 Jul 10;254:164-9.

10. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour. [Review]. 2011 Sep;11(5):759-69.

11. Malone JH, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. [Review]. 2011 May;9.

Figure Legend

Figure 1: The Basic Transcriptional Process. Ligand-responsive and generic transcription factors bind to response elements within the regulatory regions of a gene. This increases the recruitment of RNA polymerase II to the transcription start site, allowing Pol II to move along the protein coding region of the gene producing mRNA transcripts. These transcripts are then translated by the ribosome to produce the final protein.

Figure 2: Interpretation of an EMSA. Labelled DNA will migrate faster than DNA bound to protein, allowing protein:DNA interactions to be identified (lane 1+2). These interactions are shown to be specific through the use of unlabelled DNA to 'outcompete' the specific signal, illustrated by the lighter shifted band (lane 3).Finally, addition of an antibody produces a 'supershift' and positively identifies the protein binding to the DNA.

Mol Biol I: Figure 1 + 2