ELSEVIER

Critical Reviews in Oncology/Hematology 71 (2009) 1–11

CRITICAL REVIEWS IN
Oncology
Hematology
Incorporating Geriatric Oncology

www.elsevier.com/locate/critrevonc

Retroviral integration sites (RIS) mark *cis*-regulatory elements

Cherry Ee Lin Ng [a], Yoshiaki Ito [a,b], Motomi Osato [a,b,*]

[a] *Institute of Molecular and Cell Biology, Singapore*
[b] *Oncology Research Institute, National University of Singapore, Singapore*

Accepted 15 October 2008

Contents

Abstract

Transcription in multicellular eukaryotic organisms involves an elaborate orchestration of the core promoter and *cis*-regulatory elements to drive spatiotemporally and quantitatively correct gene expression. Unlike promoters found immediately upstream of protein-coding genes, the positions of distally located *cis*-regulatory elements relative to a gene of interest are difficult to define. As such, the identification and characterization of these regulatory elements has proved to be challenging. To this end, we propose a combinatorial *in silico* approach involving retroviral integration sites (RIS) mapping together with predicted matrix attachment regions (MARs) mapping and an already well-established comparative genomics approach, to enhance the prediction of potential *cis*-regulatory elements. Predicted elements can be validated by further investigations to ascertain their functions. In view of the abundance of electronically available RIS information, RIS mapping has an unrealized potential to aid in the discovery of novel *cis*-regulatory elements.

Keywords: Retroviral insertional mutagenesis; Gene therapy; Enhancer; Runx; Pu.1; Hox; Wnt-1

1. Introduction

Transcription is a highly complex process in multicellular eukaryotic organisms, involving an elaborate orchestration of the core promoter and *cis*-regulatory elements which include

* Corresponding author at: Institute of Molecular and Cell Biology, 61 Biopolis Drive, Singapore 138673, Singapore. Tel.: +65 6586 9638; fax: +65 6779 1117.
E-mail address: motomi@imcb.a-star.edu.sg (M. Osato).

1040-8428/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.critrevonc.2008.10.008

enhancers, matrix attachment regions (MARs), locus control regions (LCRs), insulators and silencers [1], to drive spatiotemporally and quantitatively correct gene expression. Such *cis*-regulatory elements may reside in introns or within several kilobases (kb) up- and downstream of the transcription unit; in some extreme cases, they may even reside within neighbouring gene(s) one to several megabases (Mb) away from the locus that they regulate. In addition to these intrachromosomal interactions, interchromosomal interactions between promoter and regulatory regions have also been implicated in coordinating gene expression [2]. Evidently, the scattered location of critical regulatory elements makes it difficult to define the appropriate region in which to perform a search for novel regulatory elements.

The identification of *cis*-regulatory elements is central to the understanding of gene transcription and might explain mechanisms of human diseases. Defects in long-range regulatory elements have recently emerged as previously underestimated factors in the genesis of human congenital disorders [3,4]. The deletion of a 39.5-kb LCR-containing region upstream of the β-globin gene is shown to cause Hispanic-type thalassemia, while common types of β-thalassemia are caused by mutations in the coding region [5]. Besides such obvious chromosomal changes, we are now aware of mere point mutations in non-coding regulatory regions being implicated, as exemplified by the case of human preaxial polydactyly (PPD) where dysregulation of sonic hedgehog (*shh*) resulting from a single point mutation within an enhancer located 1 Mb away produces a severe genetic defect [6,7]. Furthermore, in the case of cancer, somatic mutations accumulating in the 5′ non-coding regulatory region of BCL-6 proto-oncogene are reported to be associated with lymphoma progression, namely the transformation of follicular lymphomas to more aggressive large cell lymphomas [8,9]. It is foreseeable that the rapid identification of growing numbers of functional regulatory elements will pave the way for screening of these distally located elements for disease-associated mutations.

2. Identification of *cis*-regulatory elements

Several experimental approaches are currently available to identify distally located *cis*-regulatory elements. Of these, transgenenic approaches in animals is widely believed to be the best method. Large genomic constructs such as yeast artificial chromosome (YAC), bacterial artificial chromosome (BAC), or P1-based artificial chromosome (PAC) clones, which hold the promise of containing most, if not all, of the regulatory elements required for the complex orchestration of gene expression, have been successfully used for reproducing accurate expression of mammalian genes as well as delineating critical distal elements for gene regulation in transgenic mouse models. Subsequent identification and characterization of critical elements within these large constructs frequently involves the use of deletion and site directed mutant constructs. To ascertain the physiological function of *cis*-regulatory elements, further analyses by knocking out the element of interest would also be helpful. Besides mouse, the current line-up of model animals includes chick, zebrafish, medaka and xenopus, which could be cheaper alternatives for *in vivo* models. Nevertheless, transgenic animal experiments are still laborious and time-consuming.

A comparative genomics approach, which relies on evolutionary conservation as a criterion for determining potentially functional elements, can also be taken as an initial strategy for the identification of distal regulatory elements. This approach has become increasingly popular owing to the electronic availability of genomic sequence from numerous vertebrates as well as the concurrent development of genomic alignment, visualization, and analytical bioinformatics tools. Such an *in silico* approach exponentially increases our ability to generate biological hypotheses which aid in the prioritization of putative functional sequences for subsequent experimental testing involving *in vivo* transgenic animal study and *in vitro* experiments: luciferase reporter assays of transfected cells, DNase I hypersensitivity assays and electrophoretic mobility shift assays (EMSA). Additionally, chromatin immunoprecipitation (ChIP) assays also can be employed to confirm the *in vivo* association of proteins with promoters and other regulatory regions. Recent advances in ChIP technologies, such as the ChIP-on-chip, ChIP-PET (paired-end-ditags) and ChIP-Solexa 1G sequencing technology, contribute to the study of mammalian transcriptional regulation by serving as powerful tools in genome-wide location analyses of transcription factor binding regions [10–13].

Mapping of DNase I hypersensitive (HS) sites has traditionally been recognised as an inceptive experiment for the identification of regulatory elements. The presence of a HS site is thought to indicate an open chromatin state and to suggest the binding of a transcription regulatory complex. Genome-wide mapping of HS sites revealed that they are enriched near promoters at the 5′ end of genes, transcription factor binding sites, CpG islands, and active genes, reinforcing the idea that they are markers for regulatory regions [14,15]. Traditional methods involving southern blotting techniques to identify HS sites are labour-intensive and thus their application is limited to small-scale studies covering at most several hundred kbs. In contrast, several protocols have been developed in recent years aimed at large-scale mapping of HS sites for regions of Mb order [14–19] using high-throughput approaches which require either cloning and sequencing [14,18], real-time PCR [20] or microarray [21] analyses. However, these techniques may not be attainable in many laboratories. Interestingly, retroviral integration sites (RIS), in particular that of the gammaretrovirus, murine leukemia virus (MuLV), map within 500–1000 base pairs (bp) of HS sites [22–25]. The preference of MuLV integration near HS sites has been long hypothesized [24,25], and it is with the recent high-throughput and large-scale studies that conclusive evidence is provided for such an integration preference [22,23,26].