Methods 96 (2016) 69–74

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

TransgeneOmics – A transgenic platform for protein localization based function exploration

Susanne Hasse, Anthony A. Hyman, Mihail Sarov *

Max Planck Institute of Molecular Cell Biology and Genetics Dresden, Germany

ARTICLE INFO

Article history:
Received 1 August 2015
Received in revised form 9 October 2015
Accepted 10 October 2015
Available online 17 October 2015

Keywords:
BAC transgenesis
Recombineering
Protein localization
Genome engineering
Protein tagging

ABSTRACT

The localization of a protein is intrinsically linked to its role in the structural and functional organization of the cell. Advances in transgenic technology have streamlined the use of protein localization as a function discovery tool. Here we review the use of large genomic DNA constructs such as bacterial artificial chromosomes as a transgenic platform for systematic tag-based protein function exploration.

1. Introduction

For decades, immunostaining with protein specific antibodies has provided a reliable method for protein localization and antibodies are now available for a large number of proteins. Over the last 10 years the Human Protein Atlas program (HPA) has generated a collection of over 50,000 polyclonal antibodies targeting 19,100 (95%) human protein-coding genes [1]. Many of them have already been used to generate validated protein expression profiles and to systematically map the subcellular localization in fixed cells [2–5]. Antibodies are a versatile affinity reagent and can be applied in any cell line or tissue from the target species and even across species when the epitope is conserved. However, antibodies have a number of known limitations that have to be taken into account in every experiment. Although *in vivo* localization approaches based on fluorescently labeled single chain Fab fragments (scFabs) or camelid single chain antibodies (often referred as nanobodies) is possible [6–8], the vast majority of the available antibodies require cell fixation and permeabilization, which can cause cell shrinkage or leakage of endogenous components [9,10]. Antibodies are also prone to cross-reactivity and often cannot discriminate between protein isoforms that have different subcellular localizations [11].

The ability to observe protein localization *in vivo* can often reveal functionally relevant dynamics that cannot be inferred from immunostaining [12–14]. Since the first use of the *Aequorea victoria* green fluorescent protein (GFP) over 20 years ago [15] a large array of fluorescent proteins with various properties and methods for expression of fluorescently tagged proteins (Table 1) have been developed [16,17]. The cloning of cDNA derived open reading frames (ORFs) into standard expression vectors was one of the early approaches for systematic tag-based protein localization mapping [18]. A proteome-scale localization map was generated in budding yeast through the systematic cloning of ORFs into plasmid vectors for inducible overexpression of V5 tagged proteins, followed by immunostaining with an anti-V5 antibody [19] and similar approaches have been used in other systems [20–22]. Due to its simplicity this approach remains very popular and large scale ORF resources are now available for many commonly used model systems, typically in vector formats that allow the easy shuffling to expression vectors containing fluorescent or epitope tags [20,23–26]. However, the heterologous promoters and 3′ regulatory elements often used with these type of vectors do not reflect the endogenous expression levels of most proteins and can disturb cellular functions. Since certain diseases can be caused by abnormally high protein levels [27], cDNA transgenes have been used in systematic screens for such phenotypes, for example in an activator screen for the antioxidant response element in human cells [28]. In addition, overexpression may lead to saturation of the specific binding sites of a protein leading to lead to mislocalization and/or obscuring of the normal localization pattern.

The most reliable way to ensure endogenous regulatory control of the tagged protein expression is the direct targeting of the genomic locus. Two major strategies, the transposon or viral vector mediated random insertion and the homology directed precise

* Corresponding author.

http://dx.doi.org/10.1016/j.ymeth.2015.10.005
1046-2023/

Table 1
Comparison of various methods for system-scale protein localization.

	Cost	Scalability	Signal strength	Isoform selectivity	Major advantages	Major limitations
Immunostaining	High	Low	Varies	Low	Applicable to any sample	Cross-reactivity, cost
cDNA transgenes	Low	Good	High	Good	Ease of cloning	Position effects, overexpression
gDNA transgenes	Low	Good	Near native levels	Good	Comparable with targeting at a lower cost, hypomorphic or lethal protein variants can be studied	Position effects, fragmentation
Gene targeting	High	Low	Native levels	Good	Endogenous expression levels	High costs, the tag can affect gene function

targeting, have been developed in parallel with their own advantages and limitations. The main advantage of the random approach is the ability to rapidly generate and map a large number of insertions. The tag can be either inserted directly or through an exchange of a previously inserted gene trap cassette by recombinase mediated cassette exchange (RMCE) [19,29–35]. Some of the technical limitations of this approach, including a insertion bias towards a particular sequence content, were mostly overcome with later generations of these vectors. However, all random approaches suffer from diminishing returns, where the number of repeatedly hit genes grows and the number of new genes rapidly declines, and only a fraction of the inserts can result in a functional tagged protein. Until recently, homology directed repair (HDR) was only practical in a small number of model systems and has only been used for proteome-scale localization studies in a few species including *Saccharomyces cerevisiae* [36]. The discovery of the CRISPR/Cas technology has dramatically increased the efficiency of gene targeting in systems where such tools were already available and has made genome engineering feasible in species that so far appeared resistant to this approach. However, a number of technical and logistic challenges still need to be addressed before systematic genome-wide HDR targeting based tagging in mammalian systems becomes reality.

Genomic DNA (gDNA) transgenes provide an alternative that combines the ease and efficiency of cDNA transgenics with preservation of the endogenous regulatory expression control comparable with gene targeting. In this review, we explore various applications of the gDNA transgene approach and compare it to other methods for protein localization.

2. gDNA transgenesis as platform for protein function exploration

2.1. Methods for gDNA transgene engineering

The first attempts to use gDNA for large scale GFP tagging was cloning of random fragments of the fission yeast genome into expression vectors, resulting in one of the early protein localization maps [37]. This approach is not practical for metazoan genes, which can be orders of magnitude bigger than in yeast. However, large insert gDNA libraries of bacterial artificial chromosomes (BACs), P1-derived artificial chromosomes (PACs), or yeast artificial chromosomes (YACs) that can cover most genes with their endogenous cis regulatory context are available for many model systems. When used as transgenes, these clones typically exhibit near-physiological levels and patterns of gene expression and are routinely used in mapping mutations by complementation [38]. YACs can be engineered *in vivo* by homologous recombination in yeast [39,40] and the development of efficient methods for homologous recombination mediated engineering in *Escherichia coli* [41,42] made it possible to apply this approach to the easier to handle fosmid/BAC/PAC constructs. Most *E. coli* cloning hosts exhibit very low levels of spontaneous recombination, which ensures a stable maintenance of large and repetitive clones. A simple two-component recombination system, consisting of an exonuclease and a strand annealing protein from the phage lambda or related phages, is sufficient for an efficient and precise homologous recombination in *E. coli* with only 30–50 bp of homology [41,43,44]. As recombineering is independent of the availability of restriction sites and can be used to insert, delete, change or retrieve any sequence of interest to and from any DNA that can be propagated in *E. coli* it was soon adapted to a wide spectrum of applications. BAC recombineering provided an efficient way to engineer point mutations, create translational or transcriptional reporters and targeting constructs for genome editing [45–48] for almost any gene of interest in a range of model organisms (mammals: [49–51]; fly: [52,53]; worm: [54–57]). With recombineering, the tag coding sequence can be inserted into the gene of interest at the N- or the C-terminus or at any internal position of choice, allowing the tagging of specific splice forms. The tag is typically inserted as a cassette with a selectable marker, which can be removed by site-specific excision through recombinases like Cre or Flp. Methods have been developed to allow the easy exchange of the tag in the transgene by either homologous recombination [45] or RMCE [58].

2.2. System-scale gDNA transgenesis

The high recombineering efficiency, near absence of unintended recombination background and the ease of handling *E. coli* in liquid culture lead to the development of high throughput pipelines for multi-step recombineering in a 96 well plate format for parallel engineering of a large number of constructs in *Caenorhabditis elegans* [59], mammals [60] and Drosophila [52,53,61] (Fig. 1). Using this approach a *C. elegans* library of 14637 tagged fosmid transgenes was constructed, which enabled the *in vivo* localization or affinity purification of 73% of the proteome [54]. The fosmid collection was used to create over 580 transgenic *C. elegans* lines. All constructs, lines and associated localization patterns are accessible at https://transgeneome.mpi-cbg.de. Since the release of the *C. elegans* TrangeneOme resource, 2806 constructs have been distributed to labs around the world in a community-wide effort to generate a proteome scale resource of tagged transgenic lines. Using a similar approach [52,53,61] a collection covering approximately 10000 *Drosophila* genes was recently created, from which transgenic lines for 826 genes were already established. The 'tagged FlyTransgeneOme' (fTRG) lines are available at the VDRC stock centre (http://stockcenter.vdrc.at) [61]. Systematic BAC transgenesis in mammalian cell lines [60] has led to the generation of over 6000 BAC-tagged cell lines, including HeLa, U2OS, mouse embryonic stem cells and MDCK (http://hymanlab.mpi-cbg.de/bac). Over 5800 lines have been sent to 160 labs in 26 countries. Their applications spanned from the investigation of small amounts of different lines [62,63] to studies involving several hundred BAC-lines [49–51,64]. Hutchins et al. studied 696 tagged human proteins that carried a C-terminal localization and