Journal of Biotechnology 261 (2017) 63–69

Contents lists available at ScienceDirect

Journal of Biotechnology

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

Phenotype databases for genetic screens in human cells

Benedikt Rauscher, Erica Valentini, Ulrike Hardeland, Michael Boutros*

Division of Signaling and Functional Genomics, German Cancer Research Center (DKFZ) and Heidelberg University, 69120 Heidelberg, Germany

ARTICLE INFO

Keywords:
Functional genomics
Database
Phenotype
High-throughput biology
de.NBI
ELIXIR

ABSTRACT

Genetic screens are powerful tools to identify components that make up biological systems. Perturbations introduced by methods such as RNA interference (RNAi) or CRISPR/Cas9-mediated genome editing lead to biological phenotypes that can be examined to understand the molecular function of genes in the cell. Over the years, many of such experiments have been conducted providing a wealth of knowledge about genotype-to-phenotype relationships. These data are a rich source of information and it is in a common interest to make them available in a simplified and integrated format. Thus, an important challenge is that genetic screening data can be stored in databases in standardized ways, allowing users to gain new biological insights through data mining and integrated analyses. Here, we provide an overview of available phenotype databases for human cells. We review in detail two databases for high-throughput screens, GenomeRNAi and GenomeCRISPR, and describe how these resources are integrated into the German Network for Bioinformatics Infrastructure de.NBI as part of the European infrastructure for life-science information ELIXIR.

1. Introduction

Databases are important bioinformatics resources that make data broadly available and drive research. While databases for sequence-based genomics data, such as whole genome and expression data, are widely available and have become a crucial resource for biomedical research (Aken et al., 2017; Pruitt et al., 2004; Kitts et al., 2016) databases for functional data from genetic screens have yet to achieve a similar level of visibility.

Genetic screens are powerful tools that help to gain a global understanding of the genetic composition of pathways in a cell. Gene perturbation screens are conducted in which many genes are depleted systematically to observe resulting phenotypes. These experiments can be conducted using different methods, such as chemical mutagenesis (Singer and Kusmierek, 1982), RNA interference (RNAi) (Armknecht et al., 2005; Kamath et al., 2003; Maeda et al., 2001) to knockdown genes or CRISPR/Cas9 genome editing (Shalem et al., 2014; Wang et al., 2014) for specific gene knockout. Phenotypes generated by these perturbation methods can be measured and quantified in many different ways. Phenotypic readouts can be simple, such as cell viability to monitor cell fitness, or reporter assays to determine the activity of a pathway of interest (Boutros and Ahringer, 2008), or complex, image-based phenotypes. In perturbation screens with pooled reagents and cells, high-throughput sequencing can be used to quantify the abundance of perturbation reagents in a population of cells (Luo et al., 2008). Furthermore, automated microscopy can be applied to generate hundreds of phenotypes at the same time using image analysis techniques (Zanella et al., 2010; Boutros et al., 2015).

This broad spectrum of possible phenotypes that can be covered by genetic screens can provide a wealth of information to generate new biological insights. However, these phenotypes are not always readily comparable and false assumptions can easily misguide hypothesis generation. Therefore, there is a strong need for databases to store these data in standardized ways to make them broadly accessible and usable for data mining and integrated 'multi-omics' analyses. In this article we first provide an overview of existing resources that tackle these questions focusing on human cell models. We then discuss in detail the databases GenomeRNAi (Schmidt et al., 2013) and GenomeCRISPR (Rauscher et al., 2017) − two repositories that hold data generated in high-throughput screens using RNAi (Fire et al., 1998; Elbashir et al., 2001) and CRISPR/Cas9 (Jinek et al., 2012; Mali et al., 2013), respectively. These databases are supported by the German Network for Bioinformatics Infrastructure (de.NBI) as part of ELIXIR. The vision of de.NBI is to establish a sustainable bioinformatics service infrastructure that can efficiently support developments in biomedical research. In this context, we describe how GenomeRNAi and GenomeCRISPR are integrated with standardized pipelines for the analysis of image based screening data maintained within the Heidelberg Center for Human Bioinformatics (HD-HuB) − one of many de.NBI service centers.

High-throughput screening approaches have been first developed in model organisms (e.g. RNAi in *Caenorhabditis elegans*; Fire et al., 1998; Elbashir et al., 2001) and afterwards conducted in human cell lines as

* Corresponding author.
E-mail address: m.boutros@dkfz.de (M. Boutros).

http://dx.doi.org/10.1016/j.jbiotec.2017.06.008
Received 24 February 2017; Received in revised form 9 June 2017; Accepted 9 June 2017
Available online 16 June 2017

Table 1
Selection of databases storing high-throughput screening experiments results in human cells.

Name	Topic	Numbers	Website
CPD (Cellular Phenotype Database)	Phenotypic data derived from high-throughput RNAi screening.	10 studies currently stored.	www.ebi.ac.uk/fg/sym
GDSC (Genomics of Drug Sensitivity in Cancer)	Link between molecular features of cancers and response to anti-cancer drugs.	1074 cell lines 265 drugs	www.cancerrxgene.org
GenomeRNAi	Manually curated RNAi screenings extracted from the literature.	467 human RNAi screens > 2.7 million number of gene-phenotype associations	www.genomernai.org
GenomeCRISPR	Database for high-throughput CRISPR/Cas9 screening experiments on human cell lines.	> 110 experiments, 63 cell lines	genomecrispr.org

well (Berns et al., 2004). Already in 2012, Shamu et al. urged the need for developing repositories dedicated to screening experiments in mammal cells because such essential resources were only available for few model organisms. As today, the number of repositories of human cell lines screens is limited to a handful, summarized in Table 1 [ref. Table 1]. On the contrary, there are several resources dedicated to the different model organisms like FlyBase for *Drosophila* (Marygold et al., 2013), WormBase for *C. elegans* (Harris et al., 2014) and SGD (Saccharomyces Genome Database; Cherry et al., 2012) which, despite including also results of high-throughput screens, are not the focus of this review.

The Cellular Phenotype Database (CPD; Kirsanova et al., 2015; www.ebi.ac.uk/fg/sym) has been developed as part of the Systems Microscopy Network of Excellence project. It currently stores results of ten RNAi studies in both *Homo sapiens* and *Drosophila melanogaster*. In order to be included in CPD, the studies need to comply with a set of requirements defined in a study description file. In addition, the repository follows a standardized ontology named the Cellular Microscopy Phenotype Ontology (CMPO; Jupp et al., 2016; explained below) and describes genotype-to-phenotype relationships by using heat maps to facilitate data exploration.

Phenotypes in human cell lines and their relationships with specific genotypes and drugs can be found in the Genomics of Drug Sensitivity in Cancer repository (GDSC; Yang et al., 2013; www.cancerrxgene.org). In GDSC large-scale drug screens on over 1000 genetically characterized cell lines are stored with the aim of improving the development of anti-cancer therapeutics. On the website, data can be browsed according to compounds, features (like variant information) and cell line names. In addition, interactive graphical representations allow the user to explore the relations between drugs and genetic modifications.

GenomeRNAi (http://www.genomernai.org) and GenomeCRISPR (http://genomecrispr.dkfz.de) are two examples of databases storing high-throughput screening experiments in human cells curated from literature. They will be described in more detail in the following paragraphs of this review.

2. Ontology resources

All databases storing information about high-throughput screens often need to deal with a high degree of data heterogeneity, which represents a hurdle for the process of database design, for the exchange of information and for the implementation of bioinformatics software. Therefore, in the last years, several resources have been developed to standardize such information by defining minimal information requirements (The Minimum Information About a Cellular Assay, MIACA, miaca.sourceforge.net/) or by establishing ontologies of terms. Of special interest for the topic of high-throughput screens in human cells are the Cellular Microscopy Phenotype Ontology (CMPO; Jupp et al., 2016; www.ebi.ac.uk/cmpo/) and the Human Phenotype Ontology (HPO; Köhler et al., 2017; www.human-phenotype-ontology.github.io/), both summarized in Table 2 (ref. Table 2).

CMPO is species-neutral and dedicated to the annotation of phenotypic observations related to: the cell as a whole, cell processes, cell components, single cells and cell populations. It was specifically developed to describe high content screens and cellular imaging experiments. The ontology comprises a number of functional features for bioinformatics: it is compatible with different tools to annotate free-text in a semi-automated manner and it offer APIs to allow a programmatic access to the data. In addition, users have the possibility to request new CMPO terms, whenever they consider that a relevant term is missing.

HPO contains both a phenotype vocabulary and a disease-phenotype annotation system for *Homo sapiens*. It was established in 2008 at the Charité in Berlin and currently includes 11,000 curated terms as well as a list of 5000 terms in “plain-language”, easier to understand for patients and nonprofessionals (e.g. “color-blind” as synonym of “Dyschromatopsia”). The website features the “HPO web browser” which allow to find terms with associated diseases and genes and, eventually, export the results in CSV and i.d Excel table formats. Several tools for clinical diagnostic employing HPO are also linked in the website.

3. GenomeRNAi – a resource for RNAi phenotypes

RNAi has been a key method for systematic gene perturbation ever since its discovery almost 20 years ago (Fire et al., 1998). Over the years, the method has been applied to perform many experiments. In this context, GenomeRNAi was established in 2007 (Horn et al., 2007) as a resource for high-throughput RNAi screens and has since become the largest database of its kind with several hundred users every month. It currently holds data from 674 screens performed both in *Homo sapiens* and *Drosophila melanogaster* stored in unified formats. For each of these screens, the database connects observed phenotypes with annotations of targeted genes and information about the RNAi reagent used for the perturbation experiment in addition to descriptions of the screen itself. GenomeRNAi further displays phenotypes in the context of genomic information, a feature that is important to uncover potential reagent limitations such as incomplete coverage of annotated splice variants or design biases (e.g. towards UTR regions). GenomeRNAi is agnostic to the type of essay used to measure phenotypes and includes data derived from many different types of screens ranging from reporter-based screens to high-content microscopy approaches. The data are curated manually from peer-reviewed literature ensuring high quality of data. This curation process follows clear guidelines, relying on controlled vocabulary wherever possible. This is crucial as over the years RNAi screens have become increasingly complex and sophisticated. For example, reporter-based screens have been conducted methodically to understand regulatory mechanisms in signaling pathways (Pothof et al., 2003; Boutros and Ahringer 2008). Adding additional layers of complexity, combinatorial knockdown of genes has been used in conjunction with high-throughput microscopy to systematically map genetic interactions in *Drosophila* and human cell lines (Horn et al., 2011; Laufer et al., 2013; Fischer et al., 2015).

GenomeRNAi users can query the database in many different ways (Fig. 1A). First of all, information for a gene of interest can be found