Check for updates

SOFTWARE TOOL ARTICLE

REVISED DSMZCellDive: Diving into high-throughput cell line data [version 2; peer review: 2 approved]

Julia Koblitz, Wilhelm G. Dirks, Sonja Eberth, Stefan Nagel, Laura Steenpass, Claudia Pommerenke

Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig, 38124, Germany

v2 **First published:** 13 Apr 2022, **11**:420
https://doi.org/10.12688/f1000research.111175.1

Latest published: 20 Jul 2022, **11**:420
https://doi.org/10.12688/f1000research.111175.2

Abstract

Human and animal cell lines serve as model systems in a wide range of life sciences such as cancer and infection research or drug screening. Reproducible data are highly dependent on authenticated, contaminant-free cell lines, no better delivered than by the official and certified biorepositories. Offering a web portal to high-throughput information on these model systems will facilitate working with and comparing to these references by data otherwise dispersed at different sources.

We here provide DSMZCellDive to access a comprehensive data source on human and animal cell lines, freely available at celldive.dsmz.de. A wide variety of data sources are generated such as RNA-seq transcriptome data and STR (short tandem repeats) profiles. Several starting points ease entering the database via browsing, searching or visualising. This web tool is designed for further expansion on meta and high-throughput data to be generated in future. Explicated examples for the power of this novel tool include analysis of B-cell differentiation markers, homeo-oncogene expression, and measurement of genomic loss of heterozygosities by an enlarged STR panel of 17 loci.

Sharing the data on cell lines by the biorepository itself will be of benefit to the scientific community since it (1) supports the selection of appropriate model cell lines, (2) ensures reliability, (3) avoids misleading data, (4) saves on additional experimentals, and (5) serves as reference for genomic and gene expression data.

Keywords

DSMZ, human and animal cell lines, RNA-seq, STR, HLA, omics data, LL-100, leukemia, lymphoma, homeobox

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 20 Jul 2022		✓ view
		↑
version 1 13 Apr 2022	✓ view	? view

1. **Beate Rinner**, Medical University of Graz, Graz, Austria
   BioTechMed-Graz, 8010 Graz, Austria
2. **Arihiro Kohara**, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Cell & Molecular Biology gateway.

This article is included in the Bioinformatics gateway.

Corresponding authors: Julia Koblitz (julia.koblitz@dsmz.de), Laura Steenpass (laura.steenpass@dsmz.de)

Author roles: Koblitz J: Conceptualization, Data Curation, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Dirks WG**: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Eberth S**: Writing – Original Draft Preparation, Writing – Review & Editing; **Nagel S**: Writing – Original Draft Preparation, Writing – Review & Editing; **Steenpass L**: Conceptualization, Investigation, Writing – Review & Editing; **Pommerenke C**: Formal Analysis, Investigation, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: All authors are employed at the Leibniz-Institute DSMZ, a non-profit institute, which distributes the cell lines on which the omics data sets are based on.

Grant information: The author(s) declared that no grants were involved in supporting this work. JK is funded by the Leibniz SAW project DiASPora, Funding No. K280/2019.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Koblitz J, Dirks WG, Eberth S *et al.* **DSMZCellDive: Diving into high-throughput cell line data [version 2; peer review: 2 approved]** F1000Research 2022, **11**:420 https://doi.org/10.12688/f1000research.111175.2

First published: 13 Apr 2022, **11**:420 https://doi.org/10.12688/f1000research.111175.1

REVISED Amendments from Version 1

In this second version of the manuscript, we included an improvement for the heat map visualisation and an updated Figure 4b capturing all colors being mentioned in the legend and all 17 STR loci data in the described database. The legend of Figure 4b was adapted accordingly. Furthermore, we included a comment, why describing the MSI phenomenon based on MOLT-4 would exceed the scope of this manuscript.

Any further responses from the reviewers can be found at the end of the article

Introduction

For more than 70 years, cell lines have become indispensable for life sciences, especially for biomedical research. Human and animal cell lines represent cost-effective model systems of almost unlimited availability. Furthermore, they are relatively easy to manipulate. The vast majority of cell lines are derived from spontaneously immortalised tumour cells carrying specific characteristics. They have become essential models not only for cancer research such as investigating mechanisms of tumorigenesis, targets for therapy, or drug efficacies but also for other areas of biological sciences. Cell lines are made available to the scientific community by cell lines banks, who also control and guarantee for cell line authenticity. In general both, diagnosis and clinical parameters of the donor as well as molecular characteristics of the cell line itself, are required for the selection of an appropriate model cell line for a specific research question.

Usually, the required information about cell lines has to be gathered from different sources, often including the study of literature. In addition to the information provided in the cell line data sheets of the biorepositories themselves, several online databases independent of the biorepositories can be consulted. The largest available online platform listing key information and references from continuous cell lines is Cellosaurus[1]. However, so far the mentioned sources contain only limited information about genetic or transcriptional aberrations characterising a cell line. Today, molecular characteristics obtained from high-throughput sequencing data from cell lines become more and more relevant for the selection of appropriate models. In this context, the cell lines project database of COSMIC (Catalogue of Somatic Mutations in Cancer) provides mutation profiles and copy number variations of over 1,000 cancer cell lines[2]. Further publicly accessible datasets for download are multiple high-throughput sequencing data from the CCLE (Cancer Cell Lines Encyclopedia) panel *via* the DepMap Portal which also offers interactive data visualisation[3]. In order to support the search for appropriate cell line models Jeong *et al.* developed the online database GEMiCCL which enables the comparison of genomic, transcriptomic and copy number data generated in different projects, including CCLE, COSMIC and the NCI-60 cell lines panel[4]. Importantly, the comparison of the data from different sources points out that *e.g.* mutation data can strongly vary for a cell line depending on the data source. This observation is not surprising, as selective culture conditions can foster evolution of cell lines, impacting the genetic and transcriptional diversity of the same cell line between two laboratories[5]. Thus, the major drawbacks during the selection process for an appropriate model cell line are either that the molecular data cannot easily be traced back to the source of the cell line and culture conditions used for the generation of the data, and that bioinformatics skills are required to re-analyse publicly available omics data after downloading.

So far, most molecular data including high-throughput sequencing data are available from different sources but not from the cell line repository itself. This fact ultimately confronts the selection of cell lines with the question from which resource the cell line with the appropriate molecular characteristics is actually available. To overcome this limitation we developed DSMZCellDive, a novel web portal that offers access to evaluated RNA-seq data and STR profiles generated from material of human and animal cell lines that are provided to the scientific community via the DSMZ catalogue.

Methods

Implementation

All data were provided by internal sources at the DSMZ, extracted from various formats using tailored scripts, and finally transferred into an SQL database. The data were integrated by using cell culture identifiers, such as the name of the cell line and DSMZ ACC-No. The SQL database is based on MariaDB Version 10.3. PHP 7 is used to generate the web pages and to query an internal SQL database. The JavaScript libraries jQuery and Plotly.js are used to do asynchronous server requests and draw charts, respectively. Heat maps are created using the R library heatmaply. The PHP library Parsedown is used to display the COI barcoding report.

The importance of machine-readability is increasing steadily. So far, web search engines use a standardised vocabulary (schema.org) to do basic queries. The Bioschemas initiative (bioschemas.org) aims to extend these vocabularies for life sciences. We have integrated Bioschemas profiles and added a machine-readable markup for all cell lines. The Bioschemas representations enhance interoperability and standardisation of DSMZCellDive.

Operation

DSMZCellDive can be accessed at celldive.dsmz.de with every modern browser and is free of charge.

Results

Data sources

DSMZCellDive is designed to provide a data portal to high-throughput and meta data of cell lines hosting diverse data sources. Gene expression data beside HLA (human leukocyte antigen) information, STR (short tandem repeat) profiles, and COI (cytochrome c oxidase I) DNA barcodes are bundled at start as listed in Table 1 and there is more to come. In the following current data sources are described briefly.

The vast majority of data entries in DSMZCellDive is composed of RNA-seq data. We started with our published data on 100 human leukemia and lymphoma cell lines (LL-100 panel)[6]. In contrast to the cited paper, the non-malignant cell line NC-NC, a B lymphoblastoid cell line, is included and RNA-seq data were quantified *via* Salmon[7] and normalised *via* DESeq2[8] in order to keep pace with state-of-the-art data analysis.

HLA genes encode proteins in the major histocompatibility complex (MHC) which play a central role in discriminating self and non-self[9]. Although the HLA gene cluster on chromosome 6 is highly polymorphic, it is not suitable for cell line authentication due to a low exclusion rate and instability of gene expression. Furthermore, HLA typing is important for cancer research since determination of tissue compatibility by tumour neoantigen binding to HLA surface proteins and rejection of specific HLA alleles play a role. Here, HLA typing was determined on LL-100 RNA-seq data *via* arcasHLA, an alignment-based tool[9].

Table 1. Current data in DSMZCellDive. *100 leukemia and lymphoma cell lines + NC-NC, a B lymphoblastoid cell line.

Data source	type	# cell lines	Reference
LL-100	RNA-seq	101*	6
LL-100	HLA	101*	6
STR	STR profiles	4565	10
COI	COI barcodes	197	-

The applied genotyping system (Promega Powerplex 18D) uses STR microsatellite repeats which are located at 17 specific genomic loci that are highly polymorphic in human populations including gender determination *via* Amelogenin. STR typing is serving as a reference technique for identity control of human cell lines at biological resource centers and available as global standard (ANSI/ATCC ASN-0002-2021 (2021). Authentication Of Human Cell Lines: Standardization Of Short Tandem Repeat (STR) Profiling. ANSI eStandards Store). Data sources were kindly provided by ATCC, JCRB, and amended by DSMZ[10].

DNA barcoding for animal samples is frequently based on the Cytochrome c oxidase subunit 1 (COI or COX1) DNA sequence known to exert specific differences between species - prerequisite for species identification. DSMZCellDive harbours COI barcodes for all animal cell lines available at the DSMZ.

Data structure

All data were extracted from various data formats and integrated into a relational SQL database. The database structure contains seven tables in total and can be extended with more data types easily (Figure 1). All data types that belong to DSMZ cell lines are connected to the `celllines` table *via* one-to-many connections to its primary key `cell_id`. Since one cell line can have multiple STR profiles, a meta table is used to connect profiles to cell lines. This should not be confused with the fact that one cell line may carry multiple STR alleles per STR loci, a phenomenon called microsatellite instability (MSI). The whole COI barcoding DNA sequence as well as the report is saved as markdown in a single text field as currently

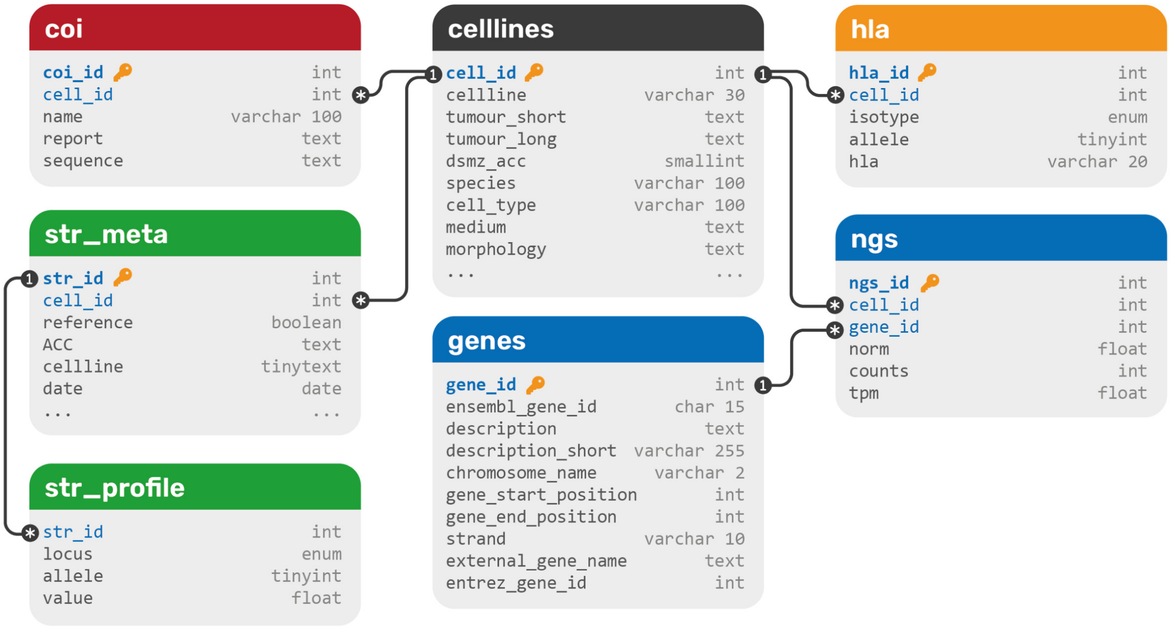


Figure 1. The database structure of DSMZCellDive. Primary keys are blue and marked with a key symbol, foreign keys are blue. Each color represents a different data type: blue for RNA-seq data, green for STR profiles, yellow for HLA typing data, and red for COI barcoding data. All data sets are integrated via a central cell line table (black).

no search or comparison operations are planned on these data, since species-specificity only not genetic individuality is resolved by this technique.

Website data portal

The DSMZCellDive website was designed to be easily extensible by more data types. For this reason, the web layout contains a sidebar instead of a classical navbar, as it allows more content. The header contains breadcrumbs to meet the requirements of the hierarchical structure of the page. The starting page gives an overview on all available data types and provides links to all tables and tools. Each data source has its own overview page, describing the data, available literature and providing further links.

Each RNA-seq data set has its own entry page and supports visualisation as bar chart or heat map. An interactive web interface enables the user to enter genes (up to five for bar charts and up to 20 for heat maps), select tumour entities or cell lines directly, and choose whether normalised data, raw counts, or TPM (transcripts per kilobase million) values should be used. Currently, only RNA-seq data from the LL-100 panel are available; however, it is planned to add more data sets in the near future. For this reason, an integrated view on all RNA-seq data from the different projects was added early on. Since data are normalised within each project, normalised values between projects should not be compared, only TPM data and bar charts are available here.

The STR data sets come with a browser and a search tool. While the browser allows access to unique comprehensive STR data sets of DSMZ and other major cell lines banks, the search tool allows users to compare own STR profiles for similarity matching with regard to authenticity. As a result of the search closest matches according to the following equation are displayed and the ratio to the reference is given in percent distance as described previously[11].

$$\frac{2 \cdot A_{shared}}{A_{query} + A_{ref}} \qquad (1)$$

The result page offers further information on the cell lines and is cross-linked to the respective cell line detail page, which also enables access to the cell line data sheet of the DSMZ catalogue.

All data are integrated on cell line level. Each of the currently almost 900 cell lines has its own detail page, including meta data on the cell line, *i.e.* the species, cell type, morphology, and culture media. Depending on whether it is a human or animal cell line, STR profiles or COI DNA barcode reports are displayed, respectively. If the cell line belongs to any RNA-seq project, a histogram of all genes is shown, as well as HLA typing data.

Use cases

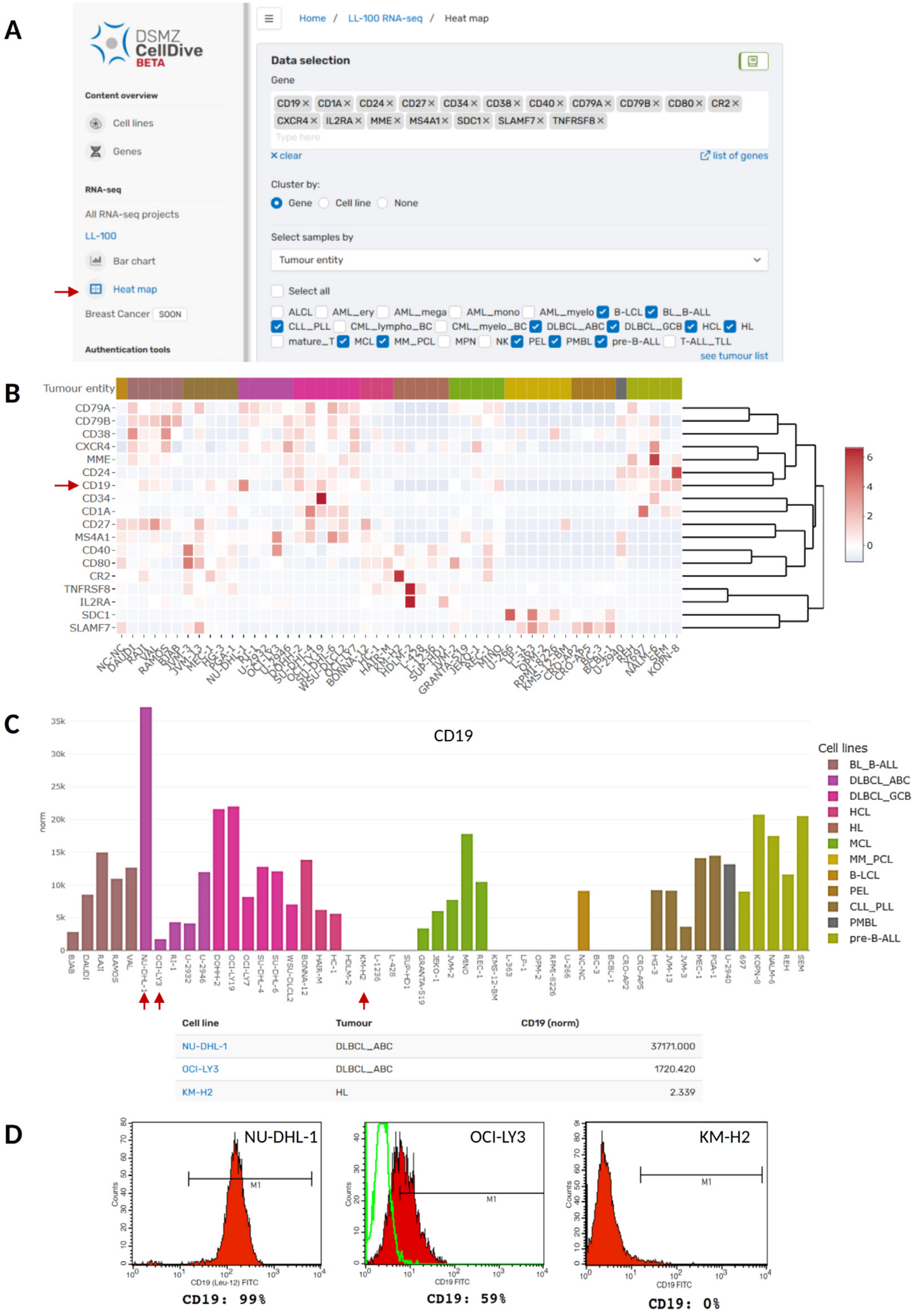
Analysis of B-cell differentiation marker genes

Our new web tool can be applied to analyse and visualise the expression of B-cell development and differentiation markers in B-cell derived cancer cell lines (Figure 2). B-cells originate from multipotent hematopoietic stem cells in the bone marrow and undergo a series of antigen-independent and antigen–dependent differentiation and maturation steps until they finally become memory or plasma B-cells that secrete antibodies. These steps are associated with the expression of well-studied marker genes, many of them encode for cell surface molecules commonly used for immunophenotyping. Importantly, B-cell derived cancer cells still mirror the differentiation and maturation phase of their normal B-cell counterpart, which is also reflected in the expression profile of B-cell marker genes and is crucial for the diagnosis of B-cell neoplasms[12,13]. Figure 2B depicts a heat map illustrating the expression of a customised selection of 18 B-cell differentiation marker genes (*CD1A, CD19, CD24, CD27, CD34, CD38, CD40, CD79A, CD79B, CD80, CR2, CXCR4, IL2RA, MME, MS4A1, SDC1, SLAMF7*, and *TNFRSF8*) across the B-cell derived cell lines included in the LL-100 panel. Intriguingly, specific B-cell markers get lost in the tumour cells. For example, although CD19 is usually expressed in all B-cells from pre-B-cells until the terminal differentiation to plasma cells[14], its expression is typically absent in Hodgkin lymphoma (HL), plasma cell leukemia (PCL), primary effusion lymphoma (PEL) and in a fraction of Diffuse large B-cell lymphoma (DLBCL)[15,16]. Accordingly, *CD19* loss is also seen in cell lines representing HL, PCL, and PEL (Figure 2C). The DLBCL cell line OCI-LY3 is an example for weak CD19 expression on mRNA and protein level compared to DLBCL cell line NU-DHL-1 harboring the highest CD19 expression across all analyzed cell lines (Figure 2C and 2D). Thus, DSMZCellDive can assist the selection of model cell lines *e.g.* with varying levels of *CD19*.

NKL homeobox gene analysis

Homeobox genes encode transcription factors sharing a special helix-turn-helix 3D-structure which mediates interaction with DNA, cofactors and chromatin. This homeodomain is formed by 60 amino acid residues and represents a platform performing gene regulation. Homeobox genes control fundamental processes in development and differentiation during embryogenesis and in the adulthood[17]. Therefore, deregulation of their activity is a common theme in cancer including hematopoietic malignancies.

According to their conserved homeobox sequences, these genes are arranged in eleven classes and several subclasses[18]. The NKL subclass (NK-like) belongs to the ANTP class (according to the *Drosophila antennapedia* gene) and consists of 48 members in humans. Their physiological expression pattern in the hematopoietic compartment has been termed “NKL-code“ comprising eleven genes[19]. This code is a useful tool to evaluate deregulated NKL homeobox genes in myeloid and lymphoid leukemia/lymphoma patients. To date, 24 aberrantly activated NKL homeobox genes are described in T-cell acute lymphoblastic leukemia (T-ALL), representing the strongest group of oncogenes in this malignancy[20]. TLX1 (formerly HOX11) and TLX3 (HOX11L2) are the most frequently deregulated NKL homeobox genes in T-ALL while NKX2-5 is only rarely expressed[19].



DSMZ CellDive BETA

Content overview
- Cell lines
- Genes

RNA-seq
- All RNA-seq projects
- LL-100
- Bar chart
- Heat map
- Breast Cancer SOON

Authentication tools

Home / LL-100 RNA-seq / Heat map

Data selection

Gene

CD19 × CD1A × CD24 × CD27 × CD34 × CD38 × CD40 × CD79A × CD79B × CD80 × CR2 × CXCR4 × IL2RA × MME × MS4A1 × SDC1 × SLAMF7 × TNFRSF8 ×

Type here

× clear | list of genes

Cluster by:
● Gene ○ Cell line ○ None

Select samples by: Tumour entity

□ Select all

□ ALCL □ AML_ery □ AML_mega □ AML_mono □ AML_myelo ☑ B-LCL ☑ BL_B-ALL
☑ CLL_PLL □ CML_lympho_BC □ CML_myelo_BC ☑ DLBCL_ABC ☑ DLBCL_GCB ☑ HCL ☑ HL
□ mature_T ☑ MCL ☑ MM_PCL □ MPN □ NK ☑ PEL ☑ PMBL ☑ pre-B-ALL □ T-ALL_TLL

see tumour list

Cell line	Tumour	CD19 (norm)
NU-DHL-1	DLBCL_ABC	37171.000
OCI-LY3	DLBCL_ABC	1720.420
KM-H2	HL	2.339

Figure 2. Analysis of B-cell marker genes in B-cell derived cancer cell lines of the LL-100 panel using DSMZCellDive. (**A**) For customised analysis of RNA-seq data of the LL-100 project via the new web tool 18 genes and 12 tumour entities were selected to visualise expression pattern in a heat map (red arrow). (**B**) The resulting heat map depicts normalised expression of the genes in rows within the selected cell lines depicted in columns. Red arrow: *CD19*. (**C**) Alternatively, expression of single genes can be visualised in a bar chart. As an example normalised expression of *CD19* (indicated by an arrow in **B**) is shown in the same tumour entities selected in **A** and **B**. Additionally, a table listing the values for normalised *CD19* expression per cell line appears underneath the bar chart. (**D**) Protein expression of CD19 in three of the cell lines depicted in **C** (red arrows). Flow cytometry data were taken from the corresponding cell line data sheets of the DSMZ catalogue to which users are guided by clicking on the blue cell line names in the table depicted in **C**.

Cell lines are useful models to investigate regulation and function of oncogenes including homeobox genes. To identify a leukemia/lymphoma cell line expressing a particular homeo-oncogene, our published LL-100 dataset and the here presented online tool DSMZCellDive may assist to find a suitable cell line model[6]. Figure 3A illustrates generated results, showing gene expression data for *TLX3* and *NKX2-5* as barplots. Both genes are expressed in particular T-ALL cell lines while silenced in cell lines derived from other leukemia/lymphoma entities[21,22]. NKL homeobox gene VENTX is physiologically expressed in lymphoid T-cell progenitors and myeloid conventional dendritic cells (cDC)[23]. MUTZ-3 is an AML (acute myeloid leukemia) cell line derived from a cDC progenitor. This cell line expresses VENTX and represents a useful model to investigate differentiation of dendritic cells and their derived malignancies[23]. Figure 3B shows NKL homeobox gene activities for *VENTX, TLX3* and *NKX2-5* in selected cell lines as heat map, demonstrating strikingly high *VENTX* expression in MUTZ-3. Thus, DSMZcellDive is a useful tool to identify and illustrate normal and aberrant (homeobox) gene expression in leukemia/lymphoma cell lines.

Importance of drifted or lost STR alleles in cell lines

Although *in vitro* evolution of tumour cell lines is well known[5], the underlying genomic alterations often remain obscure. Despite elaborate quality control via STR genotyping of lot charges, crucial genetic changes of a tumour model may remain hidden to the applying scientist.

The importance of MSI and LOH (loss of heterozygosity) can be demonstrated in one of the most commonly used models for AML research, the cytokine-dependent cell line THP-1.

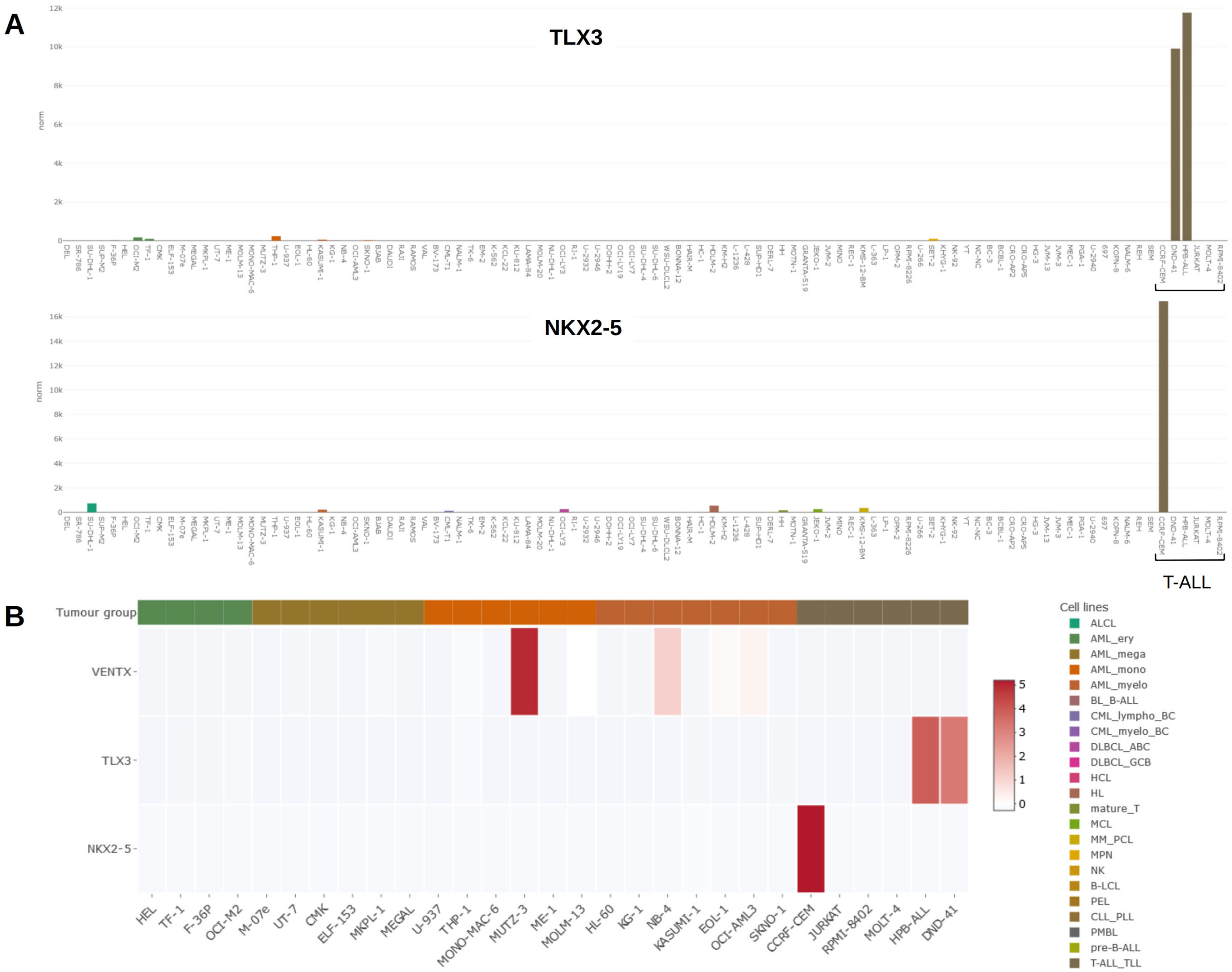
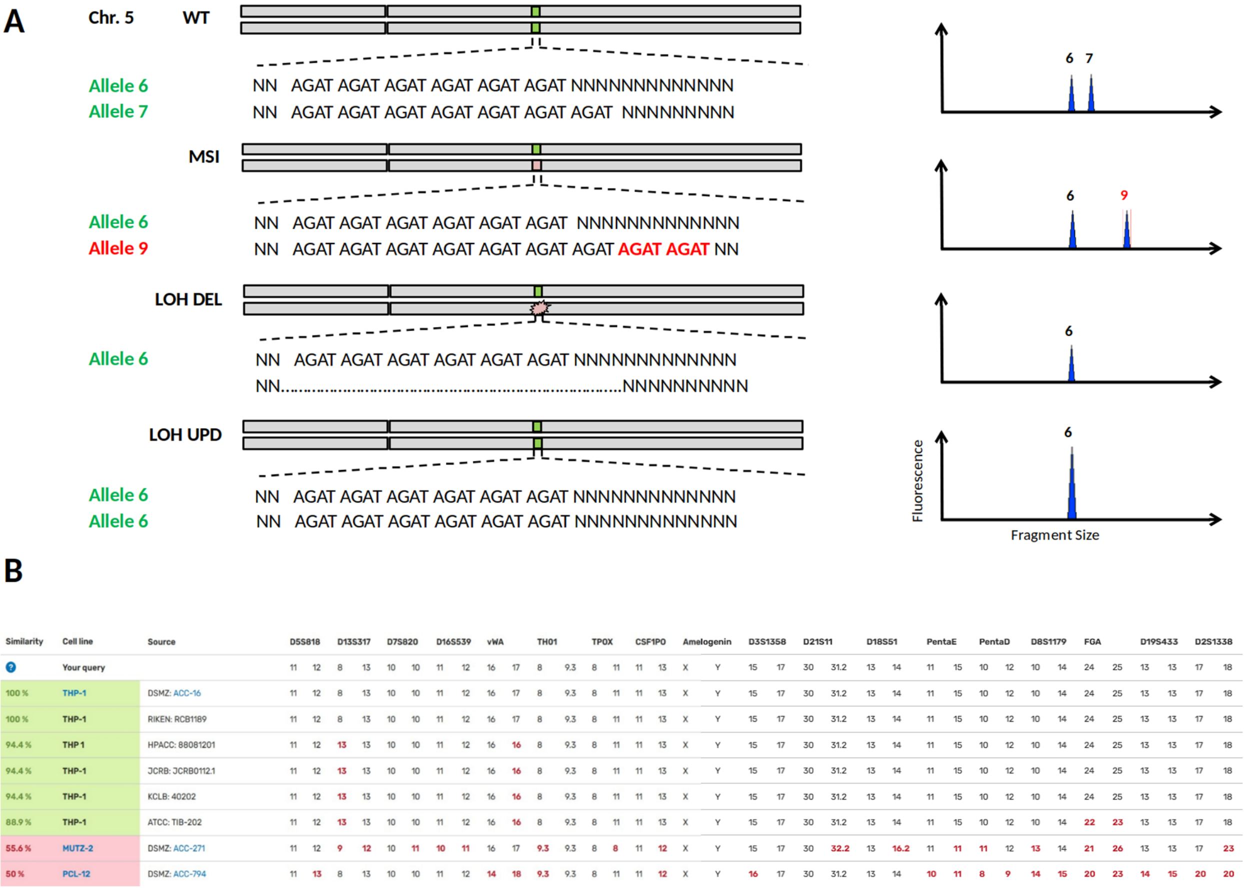


Figure 3. (**A**) Bar plots showing activities for NKL homeobox genes *TLX3* (above) and *NKX2-5* (below) in the LL-100 panel. Both genes are exclusively expressed in T-ALL cell lines DND-41, HPB-ALL (*TLX3*) and CCRF-CEM (*NKX2-5*). T-ALL cell lines are indicated by a bracket. (**B**) Heat map showing activities for NKL homeobox genes *VENTX, TLX3* and *NKX2-5* in AML and T-ALL cell lines (left). *VENTX* is prominently expressed in myelomonocytic cell line MUTZ-3, representing a model for conventional dendritic cells. A color-code assigns tumour entities to the cell lines as indicated in **A** and **B** (right).

A recent publication by Noronha *et al.* describes a large genetic divergence between THP-1 cells from a European and a US biorepository[24]. Although globally standardised STR genotyping in biorepositories is primarily used to prevent the spread of misidentified cell lines, STR typing can distinguish subclones from parental cell lines when minor changes in the STR profile have occurred due to MSI or LOH (Figure 4A). Specifically, it is LOH changes that may qualitatively reveal the loss of a heterozygous chromosomal region but do not allow quantitative conclusions to be drawn. For the divergent THP-1 cell lines, the STR profiles search *via* DSMZCellDive using 17 STR loci yield similarities of 94.4% and 88.9%, respectively (Figure 4B), indicating a close genetic relationship between THP-1 cell lines of different repositories. Despite these high similarities, critical genetic targets of MLL (mixed lineage leukemia) differed substantially between cell bank-specific THP-1 cells[24]. Thus, not only is cross-contamination of cell lines a serious problem for the reproducibility of scientific data, but also silent LOH events within a scientific tumour model.

Thus, DSMZCellDive enables scientists to verify the authenticity of cell lines by providing an extended STR-17 panel, which measures the degree of correspondence between the original cell line and already slightly modified cell lines. Since all commonly measured STR loci for authentication are combined in the STR-17 panel, this is independent of the STR typing kits used. By presenting the diploid STR datasets in two columns, LOH and UPD events can be deduced, which in turn immediately shows users how often and at what point their own STR data has deviated from the STR reference profile. The described tools will increase reliability of *in vitro* data towards trusted scientific conclusions.


A
Chr. 5
WT
Allele 6
Allele 7
NN AGAT AGAT AGAT AGAT AGAT AGAT NNNNNNNNNNNN
NN AGAT AGAT AGAT AGAT AGAT AGAT AGAT NNNNNNNN
6 7
MSI
Allele 6
Allele 9
NN AGAT AGAT AGAT AGAT AGAT AGAT NNNNNNNNNNNN
NN AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT NN
6 9
LOH DEL
Allele 6
NN AGAT AGAT AGAT AGAT AGAT AGAT NNNNNNNNNNNN
NN...NNNNNNNNNN
6
LOH UPD
Allele 6
Allele 6
NN AGAT AGAT AGAT AGAT AGAT AGAT NNNNNNNNNNNN
NN AGAT AGAT AGAT AGAT AGAT AGAT NNNNNNNNNNNN
6
Fluorescence
Fragment Size
B

Similarity	Cell line	Source	D5S818		D13S317		D7S820		D16S539		vWA		TH01		TPOX		CSF1PO		Amelogenin		D3S1358		D21S11		D18S51		PentaE		PentaD		D8S1179		FGA		D19S433		D2S1338	
?	Your query		11	12	8	13	10	10	11	12	16	17	8	9.3	8	11	11	13	X	Y	15	17	30	31.2	13	14	11	15	10	12	10	14	24	25	13	13	17	18
100 %	THP-1	DSMZ: ACC-16	11	12	8	13	10	10	11	12	16	17	8	9.3	8	11	11	13	X	Y	15	17	30	31.2	13	14	11	15	10	12	10	14	24	25	13	13	17	18
100 %	THP-1	RIKEN: RCB1189	11	12	8	13	10	10	11	12	16	17	8	9.3	8	11	11	13	X	Y	15	17	30	31.2	13	14	11	15	10	12	10	14	24	25	13	13	17	18
94.4 %	THP 1	HPACC: 88081201	11	12	13	13	10	10	11	12	16	16	8	9.3	8	11	11	13	X	Y	15	17	30	31.2	13	14	11	15	10	12	10	14	24	25	13	13	17	18
94.4 %	THP-1	JCRB: JCRB0112.1	11	12	13	13	10	10	11	12	16	16	8	9.3	8	11	11	13	X	Y	15	17	30	31.2	13	14	11	15	10	12	10	14	24	25	13	13	17	18
94.4 %	THP-1	KCLB: 40202	11	12	13	13	10	10	11	12	16	16	8	9.3	8	11	11	13	X	Y	15	17	30	31.2	13	14	11	15	10	12	10	14	24	25	13	13	17	18
88.9 %	THP-1	ATCC: TIB-202	11	12	13	13	10	10	11	12	16	16	8	9.3	8	11	11	13	X	Y	15	17	30	31.2	13	14	11	15	10	12	10	14	22	23	13	13	17	18
55.6 %	MUTZ-2	DSMZ: ACC-271	11	12	9	12	10	11	10	11	16	17	9.3	9.3	8	8	11	12	X	Y	15	17	30	32.2	13	16.2	11	11	11	12	13	14	21	26	13	13	17	23
50 %	PCL-12	DSMZ: ACC-794	11	13	8	13	10	10	11	12	14	18	9.3	9.3	8	11	11	12	X	Y	16	17	30	31.2	13	14	10	11	8	9	14	15	20	23	14	15	20	20

Figure 4. Loss of heterozygosity (LOH) or Microsatellite Instability (MSI) in STR profiles of cell lines. (**A**) In the upper half the heterozygous wild type status of STR locus D5S818 on chromosome 5 is shown, corresponding STR electropherograms are depicted on the right. MSI is caused by a replication error during DNA synthesis and may result in a new allele 9 by allelic drift. LOH can occur by a deletion mutation (LOH DEL) or by recombination between paternal chromosomes known as uniparental disomy (UPD). In both LOH cases, genomic DNA is lost and may lead to undesired losses of particular physiological properties at worst. (**B**) Searching for the reference STR profile of the cell line THP-1 via DSMZCellDive delivers similarity scores to other THP-1 profiles and deviating/missing alleles are shown as red numbers. Cell lines highlighted in green indicate authenticity without doubt, while hits below 60% matching similarity marked in pink are unrelated and thought to be definitely genetically different.

Conclusion

With the interactive web portal DSMZCellDive at hand, access to cell lines omics and meta data is bundled at one site. DSMZCellDive presents its own RNA-seq data exclusively from cell lines of the DSMZ. The novel web portal allows researchers to browse and visualize authentic NGS data generated according to high quality standards *e.g.* to support the selection of appropriate model cell lines, while cell lines are acquired. Furthermore, its concept allows implementation of future data readily and, more importantly, provides reliable data of cell lines available at DSMZ, which can serve as reference data for industry and science.

Data and software availability

Underlying data

All data underlying the results are available as part of DSMZCellDive and no additional source data are required.

Software availability

Website: https://celldive.dsmz.de

Source code at github to download and process:

Website: https://github.com/JKoblitz/DSMZCellDive

RNA-seq: https://github.com/claupomm/RNA-seq_ll100

HLA analysis: https://github.com/claupomm/HLA-analysis

Zenodo for archived source code at time of publication:

Website: https://doi.org/10.5281/zenodo.6404422

RNA-seq pipeline: https://doi.org/10.5281/zenodo.6401600

HLA analysis: https://doi.org/10.5281/zenodo.6401594

License: MIT License

Author contributions

JK designed the concept of the web portal and database, wrote all code and the manuscript. WGD provided STR profiling and COI barcoding data and wrote the manuscript. SE and SN provided expertise and wrote the manuscript. LS initiated and designed the basic concept of the web portal. CP provided preliminary work to the web tool, supervised the project, contributed RNA-seq and HLA data, and wrote the manuscript. All authors edited the manuscript.

Acknowledgements

We thank all our colleagues at the Leibniz-Institute DSMZ who contributed to improve the usability of DSMZCellDive. In particular we thank Hilmar Quentmeier for initiating this project, Silke Fähnrich for enhancing the STR part, furthermore Boyke Bunk and Lorenz Reimer for organisational support.

References

1. Bairoch A: **The Cellosaurus, a cell-line knowledge resource.** *J Biomol Tech.* 2018; **29**(2): 25–38.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Iorio F, Knijnenburg TA, Vis DJ, *et al.*: **A landscape of pharmacogenomic interactions in cancer.** *Cell.* 2016; **166**(3): 740–754.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Ghandi M, Huang FW, Jané-Valbuena J, *et al.*: **Next-generation characterization of the Cancer Cell Line Encyclopedia.** *Nature.* 2019; **569**(7757): 503–508.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Jeong I, Yu N, Jang I, *et al.*: **GEMiCCL: mining genotype and expression data of cancer cell lines with elaborate visualization.** *Database (Oxford).* 2018; **2018**: bay041.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Ben-David U, Siranosian B, Ha G, *et al.*: **Genetic and transcriptional evolution alters cancer cell line drug response.** *Nature.* 2018; **560**(7718): 325–330.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Quentmeier H, Pommerenke C, Dirks WG, *et al.*: **The LL-100 panel: 100 cell lines for blood cancer studies.** *Sci Rep.* 2019; **9**(1): 8218.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Patro R, Duggal G, Love MI, *et al.*: ***Salmon* provides fast and bias-aware quantification of transcript expression.** *Nat Methods.* 2017; **14**(4): 417–419.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol.* 2014; **15**(12): 550.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Orenbuch R, Filip I, Comito D, *et al.*: **arcasHLA: high-resolution HLA typing from RNAseq.** *Bioinformatics.* 2020; **36**(1): 33–40.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Dirks WG, MacLeod RA, Nakamura Y, *et al.*: **Cell line cross-contamination initiative: an interactive reference database of STR profiles covering common cancer cell lines.** *Int J Cancer.* 2010; **126**(1): 303–4.
PubMed Abstract | Publisher Full Text
11. Tanabe H, Takada Y, Minegishi D, *et al.*: **Cell line individualization by STR multiplex system in the cell bank found cross-contamination between ECV304 and EJ-1/T24.** *Tissue culture research communications.* 1999; **18**(4): 329–338.
Publisher Full Text
12. Küppers R: **Mechanisms of B-cell lymphoma pathogenesis.** *Nat Rev Cancer.* 2005; **5**(4): 251–62.
PubMed Abstract | Publisher Full Text
13. de Leval L, Jaffe ES: **Lymphoma Classification.** *Cancer J.* 2020; **26**(3): 176–185.
PubMed Abstract | Publisher Full Text
14. Wang K, Wei G, Liu D: **CD19: a biomarker for B cell development, lymphoma diagnosis and therapy.** *Exp Hematol Oncol.* 2012; **1**(1): 36.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Gathers DA, Galloway E, Kelemen K, *et al.*: **Primary Effusion Lymphoma: A Clinicopathologic Perspective.** *Cancers (Basel).* 2022; **14**(3): 722.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Masir N, Marafioti T, Jones M, *et al.*: **Loss of CD19 expression in B-cell neoplasms.** *Histopathology.* 2006; **48**(3): 239–46.
PubMed Abstract | Publisher Full Text
17. Bürglin TR, Affolter M: **Homeodomain proteins: an update.** *Chromosoma.* 2016; **125**(3): 497–521.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Holland PW, Booth HA, Bruford EA: **Classification and nomenclature of all human homeobox genes.** *BMC Biol.* 2007; **5**: 47.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Nagel S: **NKL-Code in Normal and Aberrant Hematopoiesis.** *Cancers (Basel).* 2021; **13**(8): 1961.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Nagel S, Pommerenke C, Scherr M, *et al.*: **NKL homeobox gene activities in**

hematopoietic stem cells, T-cell development and T-cell leukemia. *PLoS One.* 2017; **12**(2): e0171164.
PubMed Abstract | Publisher Full Text | Free Full Text

21. MacLeod RA, Nagel S, Kaufmann M, *et al.*: **Activation of *HOX11L2* by juxtaposition with 3'-*BCL11B* in an acute lymphoblastic leukemia cell line (HPB-ALL) with t(5;14)(q35;q32.2).** *Genes Chromosomes Cancer.* 2003; **37**(1): 84–91.
PubMed Abstract | Publisher Full Text

22. Nagel S, Kaufmann M, Drexler HG, *et al.*: **The cardiac homeobox gene NKX2-5 is deregulated by juxtaposition with BCL11B in pediatric T-ALL cell lines via a novel t(5;14)(q35.1;q32.2).** *Cancer Res.* 2003; **63**(17): 5329–34.
PubMed Abstract

23. Nagel S, Pommerenke C, Meyer C, *et al.*: **NKL Homeobox Gene VENTX Is Part of a Regulatory Network in Human Conventional Dendritic Cells.** *Int J Mol Sci.* 2021; **22**(11): 5902.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Noronha N, Ehx G, Meunier MC, *et al.*: **Major multilevel molecular divergence between THP-1 cells from different biorepositories.** *Int J Cancer.* 2020; **147**(7): 2000–2006.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 28 July 2022

https://doi.org/10.5256/f1000research.136217.r144763

✓

Arihiro Kohara
Laboratory of Cell Cultures, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan

I confirmed that I had read this submission and found that the author had properly modified the manuscript.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: cell biology, genetic mutation

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 06 July 2022

https://doi.org/10.5256/f1000research.122868.r135173

?

Arihiro Kohara
Laboratory of Cell Cultures, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan

In this manuscript, the authors are concerned with the creation of a new web portal that creates a database of cell metadata and characteristic analysis information in cell banks and biorepositories, and will be able to provide useful information for researchers using cell lines. This manuscript can be highly evaluated as an initiative. We also hope that information such as RNA-seq will be expanded in the future, and that fusion gene search based on sequence information and data sharing with other banks will also be expected.

However, I think it will be a better manuscript if you reconsider the following points, so please consider modifying it.

Major

- In the RNA-seq heat map analysis, the clustering results performed based on the similarity of the expression profile of each gene are displayed, but the clustering result based on the similarity of the expression profile of each cell line should also be displayed.

Minor

- The legend in Fig. 4B is not appropriate. There are no cell lines highlighted in yellow, while there is no indication of the color of cells highlighted in red.

- Although the data in Fig. 4B shows only 9 loci data as an example of LOH, it is considered that 17 loci data is collected in this database, so please describe the verification using all the data.

- Also, why not explain the example of MSI-H using MOLT-4 etc.?

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: cell biology, genetic mutation

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 14 Jul 2022

Claudia Pommerenke,

We thank the reviewer for their valuable comments. Please find our responses here:

Major

- *"In the RNA-seq heat map analysis, the clustering results performed based on the similarity of the expression profile of each gene are displayed, but the clustering result based on the similarity of the expression profile of each cell line should also be displayed."*
    - Thank you for your note on the heat map display. Clustering by cell line has been offered as option in the past, albeit by cell line solely without gene clustering. We now enabled clustering by cell line line and gene simultaneously (option “both”).

Minor

- *"The legend in Fig. 4B is not appropriate. There are no cell lines highlighted in yellow, while there is no indication of the color of cells highlighted in red."*
    - Many thanks to the reviewer for the hint. We have now displayed the colors correctly in Fig. 4B and described the cell lines highlighted in red as genetically unbiased.
- *"Although the data in Fig. 4B shows only 9 loci data as an example of LOH, it is considered that 17 loci data is collected in this database, so please describe the verification using all the data."*
    - Figure 4B has been redone so that all STR locations are now captured and the matches across STR 17 are completely shown.
- *"Also, why not explain the example of MSI-H using MOLT-4 etc.?"*
    - An explanation of the MSI phenomenon based on MOLT-4 would unfortunately go beyond the scope of this paper. For a satisfactory explanation, the failure of DNA MMR and the origin of DNA replication errors would have to be presented, so that the scope in this manuscript would be exceeded. In addition, it would need to be discussed much more deeply that the rule established by Amanda Capes-Davies for relatedness of cell lines when there is at least 80% similarity of the STR profile does not apply to MSI cell lines (Capes-Davis, Amanda, et al. "Match criteria for human cell line authentication: where do we draw the line?" International journal of cancer 132.11 (2013): 2510-2519). Finally, MSI lines need additional tests as described in the STR standard of 2021 (ANSI/ATCC ASN-0002-2021 (2021) Authentication Of human cell lines: standardization of Short Tandem Repeat (STR) profiling. ANSI eStandards Store).

Competing Interests: No competing interests were disclosed.

Reviewer Report 24 May 2022

https://doi.org/10.5256/f1000research.122868.r135181

✓ **Beate Rinner**
[1] Department for Biomedical Research, Medical University of Graz, Graz, Austria
[2] BioTechMed-Graz, 8010 Graz, Austria

In the submitted article of Koblitz et al., DSMZCellDive, a comprehensive data source for human and animal cells, will be presented, one aim is to provide high-throughput information through a web portal that aggregates data and references that are otherwise spread across multiple sources.

A variety of data sources are generated, such as RNA-seq transcriptome data and STR (Short Tandem Repeats) profiles, which are essential for working with cell lines.

The new tool is explained and described in detail, e.g. how to access the database via browsing, searching or visualization. The database can be extended with meta- and high-throughput data that will be generated in the future to finally allows a comprehensive summary and detailed characterizations of a cell line.

To illustrate the power of the tool, analysis of B-cell differentiation markers, homeo-oncogenes, homeo-oncogene expression, and measurement of genomic loss of heterozygotes through an expanded STR panel of 17 loci were described.

DSMZCellDive offers enormous added value to the scientific community by supporting the selection of suitable model cell lines and thus ensuring the reliability and reproducibility of experiments.

Optimal in vitro conditions also save resources in the long term. The comprehensive data acquisition and provision of the tool leads to a massive upgrading of in vitro models, which enables the reduction of animal experiments and thus sustainably implements the 3R principles in research.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cell line establishment, Cell line characterization, cell line quality check

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 14 Jul 2022

Claudia Pommerenke,

We thank the reviewer for the evaluation of our work and appreciate the positive comments.

Competing Interests: No competing interests were disclosed.