Published online 26 October 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D917–D924***
doi: 10.1093/nar/gkv1101

BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis

Frederik Otzen Bagger[1,2,3,4], Damir Sasivarevic[5], Sina Hadi Sohi[5], Linea Gøricke Laursen[1,2,3,4], Sachin Pundhir[1,2,3,4], Casper Kaae Sønderby[3], Ole Winther[2,3,5], Nicolas Rapin[1,2,3,4,*] and Bo T. Porse[1,2,4,*]

[1]The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, University of Copenhagen, Denmark, [2]Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark, [3]The Bioinformatics Centre, Department of Biology, Faculty of Natural Sciences, University of Copenhagen, Denmark, [4]Danish Stem Cell Centre (DanStem) Faculty of Health Sciences, University of Copenhagen, Denmark and [5]DTU Compute, Technical University of Denmark, Lyngby, Denmark

Received September 02, 2015; Revised October 05, 2015; Accepted October 11, 2015

ABSTRACT

Research on human and murine haematopoiesis has resulted in a vast number of gene-expression data sets that can potentially answer questions regarding normal and aberrant blood formation. To researchers and clinicians with limited bioinformatics experience, these data have remained available, yet largely inaccessible. Current databases provide information about gene-expression but fail to answer key questions regarding co-regulation, genetic programs or effect on patient survival. To address these shortcomings, we present BloodSpot (www.bloodspot.eu), which includes and greatly extends our previously released database HemaExplorer, a database of gene expression profiles from FACS sorted healthy and malignant haematopoietic cells. A revised interactive interface simultaneously provides a plot of gene expression along with a Kaplan–Meier analysis and a hierarchical tree depicting the relationship between different cell types in the database. The database now includes 23 high-quality curated data sets relevant to normal and malignant blood formation and, in addition, we have assembled and built a unique integrated data set, BloodPool. Bloodpool contains more than 2000 samples assembled from six independent studies on acute myeloid leukemia. Furthermore, we have devised a robust sample integration procedure that allows for sensitive comparison of user-supplied patient samples in a well-defined haematopoietic cellular space.

INTRODUCTION

A decade of intense studies of the genetic programs underlying normal and malignant haematopoiesis has resulted in a number of gene-expression data sets, which can potentially help answer questions concerning the molecular mechanisms governing normal haematopoiesis and how these are de-regulated in cancer. To researchers and clinicians with limited bioinformatics experience, these data have been available through online databases in the form of raw or semi-processed files but remained largely inaccessible for analysis, let alone comparison with user-supplied in-house data. Recently, a number of web interfaces have been generated to facilitate single gene queries of in-house data (ImmGen Gene Skyline (1), Gene-expression Atlas (2), Leukemia Gene Atlas (3) and Differentiation Map (2)) or curated, compiled and processed data sets (HemaExplorer (3), Gene Expression Commons (4), A HeamAtlas (5), BloodChIP (6), BloodExpress (7) and CODEX (8)). These tools provide information on the expression of single genes, but fail to answer the main questions as to whether these genes influence patient survival or if genes or pathways are regulated in similar or inverse patterns. We have previously published a comprehensive database of mRNA microarray samples from FACS sorted healthy and leukemic bone marrow samples (3) which has proven a useful and popular resource for researchers working within the areas of cellular differentiation, haematopoiesis and leukaemia.

*To whom correspondence should be addressed. Tel: +45 3545 6023; Fax: +45 7262 0285; Email: bo.porse@finsenlab.dk
Correspondence may also be addressed to Nicolas Rapin. Tel: +45 2643 2878; Fax: +45 7262 0285; Email: nicolas.rapin@finsenlab.dk
Present Address: Frederik Otzen Bagger. European Molecular Biology Laboratory—European Bioinformatics Institute (EMBL-EBI), Welcome Trust Genome Campus, Hinxton, Cambridge, UK, and Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge CB2 0PT, UK.

Here, we present a complete overhaul and significantly expanded version of the original database, with a new and interactive interface, all freely available online. The new database redefines current approaches to explorative data integration, presentation and visualisation of gene-expression in the haematopoietic system. Consequently, all these improvements called for a new name: BloodSpot.

The core function of BloodSpot is to provide an expression plot of genes in healthy and cancerous haematopoietic cells at specific differentiation stages. To present these haematopoietic gene profiles, we have developed a novel visualization chart that simply integrates the benefits of strip-charts and violin plots. The server accepts either a unique gene name (gene alias) or a gene signature name from the MSigDB database. Of note, an auto-complete mechanism helps finding the right names for genes and gene signatures. To contextualise the haematopoietic gene expression profile, two additional levels of visualisation are available: an interactive hierarchical tree that shows the relationship between the samples displayed and a Kaplan–Meier plot based on a high-quality Acute Myeloid Leukemia (AML) data set (9). Additionally, we added a large body of curated data sets to the database, which users can query seamlessly. Significantly, we provide a new integrated data set of samples from AML patients along with FACS sorted samples from healthy individuals. This new integrated data set provides the most detailed picture of the gene expression landscape in healthy and malignant haematopoiesis to date. Finally, the database provides the possibility of comparing user-supplied leukaemia samples to healthy cells.

The platform is freely available, and requires no login, at: www.bloodspot.eu

DATA CONTENT UPDATES

Available data sets

BloodSpot is a database of mRNA expression in healthy and malignant haematopoiesis and includes data from both humans and mice. The database is sub-divided into several data sets that are each accessible for browsing through the new interface. Data sets are organised by organism of origin and disease status. The data sets are organised as follows: first, human healthy haematopoietic cells, then human leukaemia and finally healthy mouse haematopoietic cells. BloodSpot contains the data sets from our previous HemaExplorer (3) as well as new published data sets, all manually processed as described in Rapin *et al.* (10). All data sets available in BloodSpot were generated using oligonucleotide microarray chips, except for one mouse data set that was generated using RNA sequencing technology. For completeness, the database also includes the content of other online databases that we deem relevant for the study of haematopoiesis in the framework of BloodSpot. These external databases include the Differentiation Map (DMAP) (2) and the Immunological Genome project (ImmGen) (1).

In total the platform encompasses more than 5000 samples (see Tables 1–3). All data sets were controlled for quality, appropriately normalised and adjusted for batch effects when necessary (11,12).

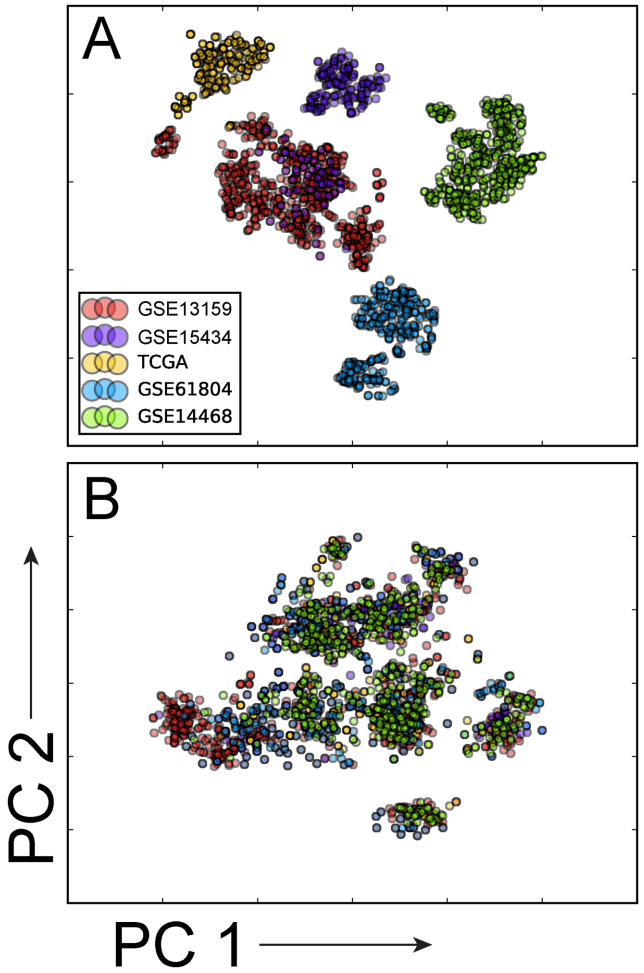

Figure 1. Principal component analysis (PCA) plot of BloodPool samples. (**A**) before batch correction, (**B**) after batch correction. Batches are coloured by study of origin.

BloodPool

One new feature of BloodSpot is BloodPool, an aggregated and integrated data set grouping the results of multiple studies focusing on AML. By means of our batch correction methods this data set can be used to study gene expression (programs) in AML in comparison with healthy corresponding cells (see Figure 1). Using the computational method developed in Rapin *et al.* (10), we have also computed gene expression fold changes relative to their nearest normal counterparts for all AML profiles in BloodPool. BloodPool is available for browsing within BloodSpot and can be selected as any of the other available data sets.

MSigDB and CMAP gene signatures integration

We collected all gene signatures available from the Molecular Signatures Database (MSigDB) (13) (version 4.0) (http://www.broadinstitute.org/gsea/msigdb/) and computed, for each signature, the mean expression values for all samples in all data sets. These mean values summarise the expression of a signature for each sample. Connectivity map (CMAP) (13) signatures were generated with the rank matrix provided by

Table 1. Data sets for normal hematopoiesis

Data set	Organism	Source	Sample numbers	Cell types	Reference
Normal hematopoiesis with AMLs	Human	GSE42519	34	HSC, MPP, CMP, MEP, GMP, early PM, late PM, MY, MM, BC, PMN	Rapin *et al.* (20)
Normal hematopoiesis (HemaExplorer)	Human	GSE17054	2	HSC	Majeti *et al.* (21)
Normal hematopoiesis (HemaExplorer)	Human	GSE19599	4	GMP, MEP	Andersson *et al.* (22)
Normal hematopoiesis (HemaExplorer)	Human	GSE11864	2	Monocytes	Hu *et al.* (23)
Normal hematopoiesis (HemaExplorer)	Human	E-MEXP-1242	2	Monocytes	Wildenberg *et al.* (24)
Normal hematopoiesis (DMAP)	Human	GSE24759	211	Normal Hematopoiesis	Novershtern *et al.* (2)
Mouse normal hematopoietic system	Mouse	GSE14833, GSE6506	67	Normal Hematopoiesis	Di Tullio *et al.* (25), Chambers *et al.* (26)
ImmGen data sets	Mouse	GSE15907	>700	Normal Hematopoiesis	Ref (1,27–29)

Table 2. Data sets for leukemic patients

Data set	Organism	Source	Patient numbers	Cell types	Reference
AML Normal Karyotype data sets	Human AML	GSE15434	251	NK-AML, WBM	Kohlman *et al.* (28)
AML TCGA data sets	Human AML	TCGA	183	Various genetic aberrations, including t(8;21), inv(16), t(15;17), t(11q23), complex karyotype, WBM	TCGA (9)
Leukemia MILE study	Human AML, ALL, CML, CLL and MDS	GSE13159	2096	AML, ALL and preleukemic stages.	Haferlach *et al.* (29,30)
AML versus normal	Human AML	GSE6891, GSE13159	91 251	NK-AML, WBM	de Jonge *et al.* (31,32)
Bloodpool	Human AML	GSE13159, GSE15434, TCGA, GSE61804, GSE14468	2076	Mainly AML, ALL and preleukemic stages.	all references above

the database. For each combination of compound and concentration, we reported the top and bottom 500 genes and produced gene signatures. The data displayed in BloodSpot represent the mean value of all genes in a given signature.

Data normalisation

All data were normalised and batch corrected to eliminate potential lab batch effects. For this we performed Robust Multi-array Average (RMA) (14) normalisation of all microarray .CEL data files partitioned by origin, and next applied ComBat (http://jlab.byu.edu/ComBat/) (12) an empirical Bayes method implemented in the R language. The batches were defined to be the study name/number, while the covariates was assigned to the relevant cell type. The resulting integrated gene expression databases can be visualised directly or compared to external samples provided by the user. See Tables 1–3 for an overview of the data presented in BloodSpot and the normalisation procedure used. All AML data sets available in BloodSpot are normalised according to Rapin *et al.* (10) and further batch corrected using ComBat when necessary. This processing schema ensures that the samples are normalised in the context of normal haematopoiesis and according to state of the art batch correction methods, regardless of the origin of the data.

For RNA-seq data, we used the Blue Collar Bioinformatics RNA-seq pipeline (mapping on mm10 mouse genome with TopHat version 2 (15), (https://bcbio-nextgen.readthedocs.org/)) to obtain normalised count data from raw fastq files from Lara-Astiaso *et al.* (16). We report count data processed using the variance stabilising transformation method from the DESeq2 package (17).

Abbreviations and sample annotations

Abbreviations for all cell types can be found below the plot by clicking the 'Abbreviations' link. Typically, the user can find more detailed information about each cell type such as a longer, more informative name, and for healthy cells data sets the immunophenotype, when available. Links to the raw unprocessed data can also be found here.

Available genes

The server is restricted to genes found in our database of Affymetrix Human 133U plus 2, Affymetrix Human

Table 3. Data set overview

Data set	Features	Samples	Normalisation method
Leukemia MILE study	67191	2095	1
Normal human hematopoiesis with AMLs	67191	296	1,7
Immgen Key populations	47273	256	2
AML versus normal	67191	252	3
AML TCGA data set	67191	244	1
AML TCGA data set versus normal	67191	244	3
AML Normal Karyotype	54675	234	1
AML Normal Karyotype versus normal	67191	234	3
Normal human hematopoiesis (DMAP)	35459	211	4
Immgen abT cells	47273	190	2
Immgen Dentritic cells	47273	151	2
Immgen MFs Monocytes Neutrophils	47273	114	2
Immgen B cells	47273	103	2
Normal human hematopoiesis (HemaExplorer)	57270	77	5
Immgen gdT cells	47273	76	2
Immgen Stem and progenitor cells	47273	76	2
Mouse normal hematopoietic system	57613	67	4
Immgen Activated T cells	47273	55	2
Immgen NK cells	47273	47	2
Immgen Stromal cells	47273	39	2
Mouse normal (RNA seq)	45426	52	6
BloodPool	67191	2120	1,7
BloodPool versus normal	67191	2076	3,7

Normalisation method legend:
1 Each cancer sample is normalised together with a set of samples from sorted normal myeloid populations. All samples where normalised using RMA. Comparison of gene expression values is not possible with other data sets in Bloodspot.
2 All samples from the ImmGen data sets were normalised together with RMA. Samples were subsequently attributed to the different data sets in BloodSpot. This means that comparison of gene expression values is possible across all ImmGen data sets.
3 The data are normalised according to Rapin *et al.* Briefly, each cancer sample is normalised together with a set of samples from sorted normal myeloid populations. Next, using a PCA-based method, the 5 closest normal samples from the cancer sample are averaged and this computed normal sample are next compared to the cancer sample allowing for computation of gen expression fold changes. See Supplementary Methods and Rapin *et al.* (10).
4 All sampleswhere
normalised using RMA. Comparison of gene expression values is not possible with other datasets in Bloodspot.
5
See our previous work (Bagger *et al.* (3)).
6 The data were processed using the bcbio nextgen RNA-seq pipeline. Count data were subsequently processed with DESeq2's variance stabilising transformation method.
7 The data was batch corrected using ComBat, taking study number as batch.

133UA and Affymetrix Human 133UB chips for human, and GeneChip Mouse Genome 430 2.0 and Affymetrix Mouse Gene 1.0 ST Arrays for mouse. For the RNA-seq data set UCSC annotation for the mm10 genome was used.

In order to handle gene aliases, a dictionary of gene aliases was constructed from NCBI ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/ and The HUGO Gene Nomenclature Committee (HGNC) www.genenames.org. Ambiguous gene aliases were not included when constructing the dictionary. The alias conversion is only used when the query is not an official gene symbol or probe name. The end result allows for greater flexibility regarding gene names input and faster browsing.

FUNCTIONALITY UPDATES

Both the back-end and the front-end have been completely redesigned for interactive usage and speed of execution. The interface is built with a range of new functionalities, with a focus on simplicity of use (see Figure 2).

Unified input

BloodSpot takes a single gene name (or unambiguous gene alias) or gene signature name as query. Users can search for keywords such as 'carcinomas' or 'cell cycle' and will be provided with a list of matching gene signature names. When relevant, it is possible to select which probe-set to display from the list in the upper right corner of the main plot. By default, the probe with the overall highest intensity is at the top of the list. The option 'Max probe' will use the one probe with the highest intensity within each population.

Default plot

When visiting the interface the plot at the centre of the screen in the default view. This representation is a novel improved jitter strip chart of gene expression, a swift novel visualisation plot that draws from bar plots and violin plots where the jitter is controlled by the density of samples and normalised over all the columns in the chart. Thus the width of the data cloud shows how many samples have similar values (see Figure 3A and a comparison to existing data plot types in Supplementary Figure S1). For more details on this visualisation method please see (Sidiropoulos, N.,

Figure 2. BloodSpot interface details. After a gene alias is submitted to display its expression pattern, any of the top three panels can be clicked to magnify content. The three panels show, from left to right, a survival plot based on a high-quality AML data set displaying a full Kaplan–Meier analysis for any query gene or gene signature, an improved jitter strip chart of gene-expression plot that draws from bar plots and violin plots and an interactive hierarchical tree that shows the relationship between the samples displayed and allows changing the focus of the display. The Select Population button allows the user to select which populations to display. The Gene Correlations button shows in a table how much other genes or gene signatures correlate with the displayed gene. It is possible to click on the genes in the table to display their expression profile. The Print as PDF button allows the user to export the current plot in PDF format. The T-Test button allows you to perform significance test between pairs of populations (legend is as follows: NS: non significant; $*P < 0.05$; $**P < 0.01$; $***P < 0.001$). The Export Data as Text button allows you to export the raw data as text (CSV format). The Upload your own sample button allows for the upload of an Affymetrix HU133 plus 2.0 .CEL file and for viewing it in the context of normal haematopoiesis. The drop down menu in the upper right corner of the main plot can be used to select a probe representing the gene of interest; by default, the probe with the highest intensity is chosen. At the bottom of the main plot, a list of abbreviations is available that includes immunophenotypes when applicable.

Sohi, S.H., Rapin, N. and Bagger, F.O. (2015) SinaPlot: an enhanced chart for simple and truthful representation of single observations over multiple classes. *bioRxiv,* http://dx.doi.org/10.1101/028191). Both an R-package and a web-server have been implemented for those interested in make use of this plot type that we have named SinaPlot.

Survival plot

The chart shown to the left of the BloodSpot interface is a survival plot based on a high-quality AML data set from The Cancer Genome Atlas (TCGA). It displays a full Kaplan–Meier analysis of survival. The survival plots are only available for human data sets, sharing probes with the microarray platform used by the TCGA (Affymetrix U133 Plus 2) (see Figure 3B).

Tree plot

The chart shown to the right of the BloodSpot interface is an interactive hierarchical tree that shows the relationship between the samples displayed and allows changing the focus of the display. It is possible to mouse over the nodes to get the full name for long names. Nodes can be clicked to collapse a branch of the tree—this will also update the default plot in the middle and remove the same populations there (see Figure 3C).

Correlation of genes and gene signatures

For each gene and signature in every data set, we report the top correlating genes or signatures. Taking the haematopoietic fingerprint (e.g. the expression value of one gene over all haematopoietic cells) of all probe-sets and gene signatures in a given data set, we calculated the correlation matrix (Pearson) and present the highest positive and negative correlating genes/signatures. This feature allows for investigation of new associations between putative co-regulated genes or signatures that exhibit similar or inverse expression patterns over the course of haematopoiesis (see Figure 3D).

Other built-in tools

Cell populations may be removed from the graphs using the 'Select population' button. The current plot displayed can be exported as PDF in publication-ready quality using the 'Print as PDF' button. The 'T-Test' button can be used to add the results from a students t-test for significance between pairs of populations to the plot. The legend is as following: NS: non-significant; $*P < 0.05$; $**P < 0.01$; $***P < 0.001$. The significance marks relies on t statistics for unequal sample sizes but assuming equal variance and the critical values are compared with a two-tailed probability. Finally, raw data can be exported as CSV using the 'Export Data as Text' button.

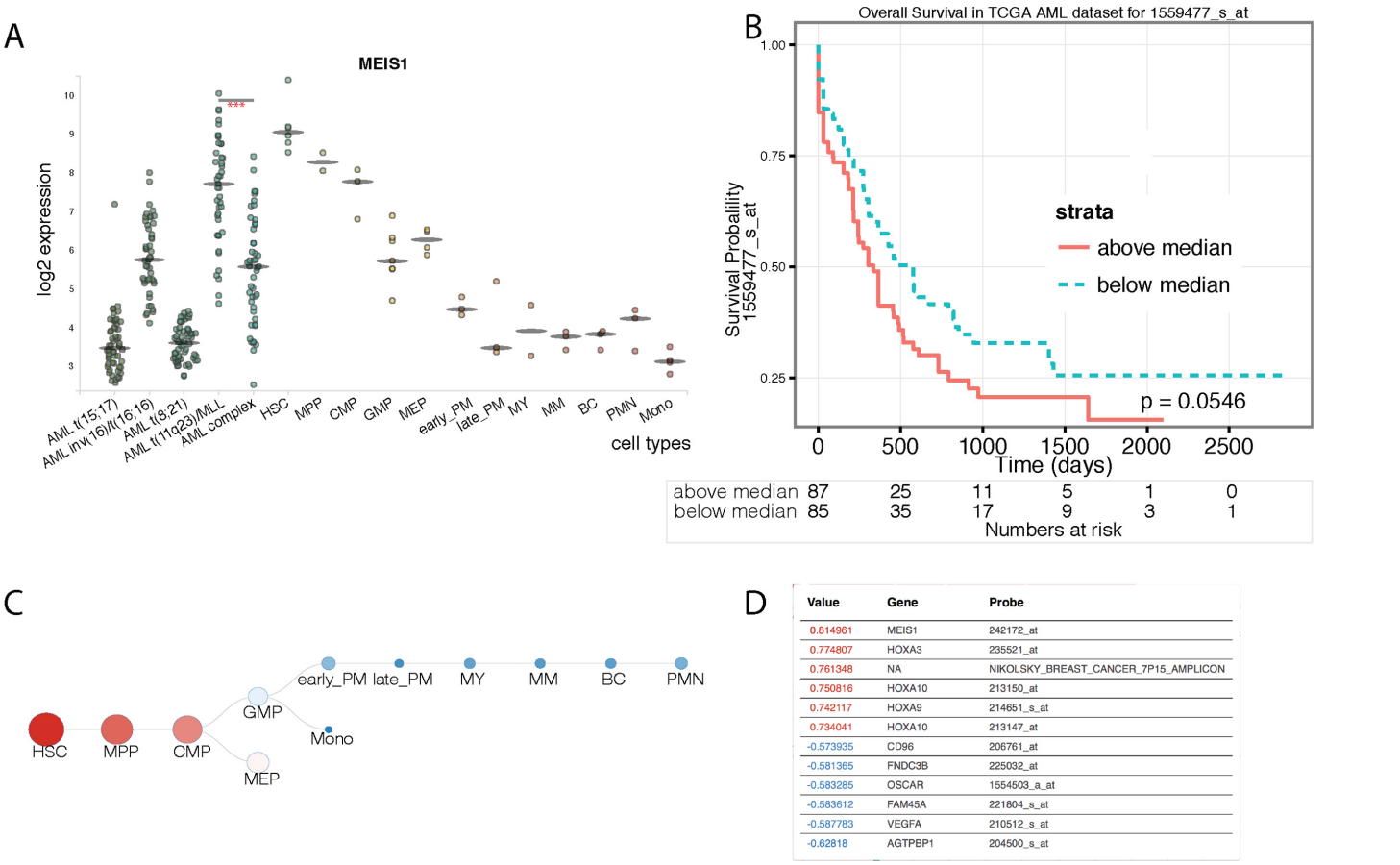

Figure 3. Main plots from BloodSpot for *MEIS1*. (**A**) Default view in BloodSpot. The plot is a novel improved jitter strip chart of gene expression that draws from bar plots and violin plots where the jitter is controlled by the density of samples and normalised over all the columns in the chart. (**B**) Survival plot based on a high-quality AML data set from The Cancer Genome Atlas (TCGA). It displays a full Kaplan–Meier analysis of survival. The survival plots are only available for human data sets, sharing probes with the microarray platform used by the TCGA. (**C**) Interactive hierarchical tree that shows the relationship between the samples displayed. Hovering over the nodes provides the full names of cell populations. Nodes can be clicked to collapse a branch of the tree—this will also update the default plot in the middle and remove the same populations there. The colour in the nodes represents the median expression of the queried gene. To accentuate the display in the trees, node size is also proportional to gene expression. Trees are based on literature (hierarchical differentiation), or overall sample correlation (correlation of samples). (**D**) Example table of genes and gene signatures correlating with MEIS1 expression in the default data set. This table appears when the user clicks on the 'correlation' button.

Upload sample

By clicking the 'Upload sample' button it is possible to analyse user-supplied samples produced on the Affymetrix U133 plus 2 platform. Significantly, doing so allows for the comparison of any myeloid microarray data to normal human haematopoiesis. The resulting analysis is then displayed in a private session in the framework of BloodSpot along with a principal component analysis that shows the location of the uploaded sample in the hematopoietic sample space. The analysis is anonymous and requires no login. The resulting data set, including the uploaded sample, can then be queried along with the default data sets in a private session. All names and array information are stripped from the uploaded file before creating the database for the user session. Hence, the uploaded sample in the private session will appear simply as S_1 in all charts. The private sessions and uploaded data are deleted every day at GMT 1.30 pm.

EXAMPLES OF USE OF BLOODSPOT

To demonstrate the use of BloodSpot, we provide in the following section an example relying on data and analysis provided by the database.

MEIS1 is part of a transcriptional program required for the maintenance of MLL-rearranged AML (18). The expression of this gene is therefore often up-regulated in MLL leukaemias. Using Bloodspot, we investigated the expression pattern of *MEIS1*, and found it to be expressed at high levels in stem cells with decreasing expression as the cells differentiate (Figure 3A and C). Using the correlation function, we find that MEIS1 expression also correlates with the expression patterns of a number of Homeobox genes, including *HOXA3*, *HOXA9* and *HOXA10* which are also typically expressed early during haematopoiesis (19) (Figure 3D). Switching to the BloodPool data set, *MEIS1* is found to be up-regulated in MLL leukaemias (Figure 4). Although the *P*-value in the survival plot does not reach statistical significance (0.055; see Figure 3B), the influence

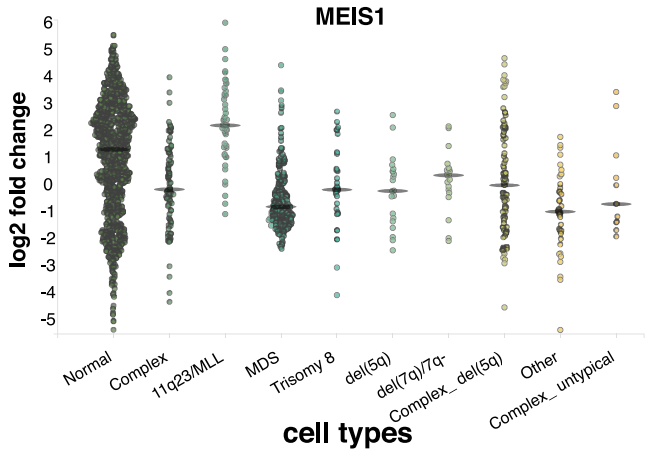

Figure 4. *MEIS1* expression relative to the nearest normal counterpart in different AML subtypes, including MLL-rearranged AML.

of *MEIS1* expression in leukemic patients may be of potential relevance.

DISCUSSION

Here we have presented a web-based database that allows for browsing of haematopoietic gene-expression fingerprints in human, murine and malignant haematopoiesis in a large number of high-quality data set containing several hematopoietic cell types. The tool facilitates the easy assessment of gene-expression data and how this links to patient survival, investigation of gene-expression signatures, as well as analysis of user generated data and export of data and figures. Focusing on simplicity, BloodSpot has features that allow clinicians or biologists to quickly retrieve relevant information on the expression of specific genes/pathways, and further explore co-regulated patterns of gene-expression as well as impact on patient survival. Our statistical framework supports the upload of user-generated patient data for integration and comparison with our database of healthy cells. This will allow assessment of the origin of the blast population in AML patients as well as assessment of well known and novel genetic markers in the context of normal haematopoiesis, both of which could be important for stratification of difficult patient cases.

We have also integrated the largest pool of AML patient microarray samples to date and have computed gene expression fold changes for these profiles, thanks to our cancer versus normal method previously described in (10) and curation and labelling of external data followed by ComBat (12). In conclusion, we have curated and populated a database and developed an analysis platform, which will allow researchers as well as clinicians to access and analyse gene expression data related to both normal and malignant haematopoiesis. We believe that the database should be of interest to all researchers and clinicians interested in haematopoiesis, leukaemia, basic immunology and gene expression in developmental systems.

Additional to information on gene-expression BloodSpot addresses two key questions, namely, how gene-expression patterns of single genes impact on patient survival, and which other genes display similar expression patterns in the haematopoietic system. Thus the platform will help broaden the basis on which to generate hypotheses about potential therapeutic targets and expand the understanding of co-regulated genes and pathways, to support experimental findings from animal model systems.

AVAILABILITY

Bloodspot is accessible at www.bloodspot.eu

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Danish Research Council for Strategic Research, as well as through a centre grant from the NovoNordisk Foundation (The Novo Nordisk Foundation section for Stem Cell biology in Human Disease). Furthermore, F.O.B. was supported by the Lundbeck foundation. We thank Nicolas Hillau for the animated logo of BloodSpot.

FUNDING

Funding for open access charge: Danish Research Council for Strategic Research (09-065157, 10-092798); NovoNordisk Foundation (The Novo Nordisk Foundation section for Stem Cell biology in Human Disease).
Conflict of interest statement. None declared.

REFERENCES

1. Miller,J.C., Brown,B.D., Shay,T., Gautier,E.L., Jojic,V., Cohain,A., Pandey,G., Leboeuf,M., Elpek,K.G., Helft,J. *et al.* (2012) Deciphering the transcriptional network of the dendritic cell lineage. *Nat. Immunol.*, **13**, 888–899.
2. Novershtern,N., Subramanian,A., Lawton,L.N., Mak,R.H., Haining,W.N., McConkey,M.E., Habib,N., Yosef,N., Chang,C.Y., Shay,T. *et al.* (2011) Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, **144**, 296–309.
3. Bagger,F.O., Rapin,N., Theilgaard-Mönch,K., Kaczkowski,B., Thoren,L.A., Jendholm,J., Winther,O. and Porse,B.T. (2013) HemaExplorer: a database of mRNA expression profiles in normal and malignant haematopoiesis. *Nucleic Acids Res.*, **41**, D1034–D1039.
4. Seita,J., Sahoo,D., Rossi,D.J., Bhattacharya,D., Serwold,T., Inlay,M.A., Ehrlich,L.I., Fathman,J.W., Dill,D.L. and Weissman,I.L. (2012) Gene Expression Commons: an open platform for absolute gene expression profiling. *PLoS One*, **7**, e40321.
5. Watkins,N.A., Gusnanto,A., de Bono,B., De,S., Miranda-Saavedra,D., Hardie,D.L., Angenent,W.G.J., Attwood,A.P., Ellis,P.D., Erber,W. *et al.* (2009) A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood*, **113**, e1–e9.
6. Chacon,D., Beck,D., Perera,D., Wong,J.W.H. and Pimanda,J.E. (2014) BloodChIP: a database of comparative genome-wide transcription factor binding profiles in human blood cells. *Nucleic Acids Res.*, **42**, D172–D177.
7. Miranda-Saavedra,D., De,S., Trotter,M.W., Teichmann,S.A. and Göttgens,B. (2009) BloodExpress: a database of gene expression in mouse haematopoiesis. *Nucleic Acids Res.*, **37**, D873–D879.
8. Sánchez-Castillo,M., Ruau,D., Wilkinson,A.C., Ng,F.S.L., Hannah,R., Diamanti,E., Lombard,P., Wilson,N.K. and Göttgens,B. (2015) CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.*, **43**, D1117–D1123.

9. Cancer Genome Atlas Research Network. (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074.
10. Rapin,N., Bagger,F.O., Jendholm,J., Mora-Jensen,H., Krogh,A., Kohlmann,A., Thiede,C., Borregaard,N., Bullinger,L., Winther,O. *et al.* (2014) Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients. *Blood*, **123**, 894–904.
11. Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.
12. Johnson,W.E., Li,C. and Rabinovic,A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.
13. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
14. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
15. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.
16. Lara-Astiaso,D., Weiner,A., Lorenzo-Vivas,E., Zaretsky,I., Jaitin,D.A., David,E., Keren-Shaul,H., Mildner,A., Winter,D., Jung,S. *et al.* (2014) Immunogenetics. Chromatin state dynamics during blood formation. *Science*, **345**, 943–949.
17. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
18. Willer,A., Jakobsen,J.S., Ohlsson,E., Rapin,N., Waage,J., Billing,M., Bullinger,L., Karlsson,S. and Porse,B.T. (2015) TGIF1 is a negative regulator of MLL-rearranged acute myeloid leukemia. *Leukemia*, **29**, 1018–1031.
19. Argiropoulos,B. and Humphries,R.K. (2007) Hox genes in hematopoiesis and leukemogenesis. *Oncogene*, **26**, 6766–6776.
20. Rapin,N., Bagger,F.O., Jendholm,J., Mora-Jensen,H., Krogh,A., Kohlmann,A., Thiede,C., Borregaard,N., Bullinger,L., Winther,O. *et al.* (2014) Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients. *Blood*, **123**, 894–904.
21. Majeti,R., Becker,M.W., Tian,Q., Lee,T.L.M., Yan,X., Liu,R., Chiang,J.H., Hood,L., Clarke,M.F. and Weissman,I.L. (2009) Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 3396–3401.
22. Andersson,A., Edén,P., Olofsson,T. and Fioretos,T. (2010) Gene expression signatures in childhood acute leukemias are largely unique and distinct from those of normal tissues and other malignancies. *BMC Med. Genomics*, **3**, 6.
23. Hu,X., Chung,A.Y., Wu,I., Foldi,J., Chen,J., Ji,J.D., Tateya,T., Kang,Y.J., Han,J., Gessler,M. *et al.* (2008) Integrated regulation of Toll-like receptor responses by Notch and interferon-γ pathways. *Immunity*, **29**, 691–703.
24. Wildenberg,M.E., van Helden-Meeuwsen,C.G., van de Merwe,J.P., Drexhage,H.A. and Versnel,M.A. (2008) Systemic increase in type I interferon activity in Sjögren's syndrome: a putative role for plasmacytoid dendritic cells. *Eur. J. Immunol.*, **38**, 2024–2033.
25. Di Tullio,A., Vu Manh,T.P., Schubert,A., Castellano,G., Månsson,R. and Graf,T. (2011) CCAAT/enhancer binding protein alpha (C/EBP(alpha))-induced transdifferentiation of pre-B cells into macrophages involves no overt retrodifferentiation. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 17016–17021.
26. Chambers,S.M., Boles,N.C., Lin,K.-Y.K., Tierney,M.P., Bowman,T.V., Bradfute,S.B., Chen,A.J., Merchant,A.A., Sirin,O., Weksberg,D.C. *et al.* (2007) Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell*, **1**, 578–591.
27. Painter,M.W., Davis,S., Hardy,R.R., Mathis,D., Benoist,C. and Immunological Genome Project Consortium. (2011) Transcriptomes of the B and T lineages compared by multiplatform microarray profiling. *J. Immunol.*, **186**, 3047–3057.
28. Kohlmann,A., Bullinger,L., Thiede,C., Schaich,M., Schnittger,S., Döhner,K., Dugas,M., Klein,H.-U., Döhner,H., Ehninger,G. *et al.* (2010) Gene expression profiling in AML with normal karyotype can predict mutations for molecular markers and allows novel insights into perturbed biological pathways. *Leukemia*, **24**, 1216–1220.
29. Haferlach,T., Kohlmann,A., Wieczorek,L., Basso,G., Kronnie,G.T., Bene,M.C., De Vos,J., Hernandez,J.M., Hofmann,W.K., Mills,K.I. *et al.* (2010) Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the International Microarray Innovations in Leukemia Study Group. *J. Clin. Oncol.*, **28**, 2529–2537.
30. Kohlmann,A., Kipps,T.J., Rassenti,L.Z., Downing,J.R., Shurtleff,S.A., Mills,K.I., Gilkes,A.F., Hofmann,W.-K., Basso,G., Dell'orto,M.C. *et al.* (2008) An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the Microarray Innovations in LEukemia study prephase. *Br. J. Haematol.*, **142**, 802–807.
31. Verhaak,R.G.W., Wouters,B.J., Erpelinck,C.A.J., Abbas,S., Beverloo,H.B., Lugthart,S., Löwenberg,B., Delwel,R. and Valk,P.J.M. (2009) Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica*, **94**, 131–134.
32. de Jonge,H.J.M., Valk,P.J.M., Veeger,N.J.G.M., ter Elst,A., den Boer,M.L., Cloos,J., de Haas,V., van den Heuvel-Eibrink,M.M., Kaspers,G.J.L., Zwaan,C.M. *et al.* (2010) High VEGFC expression is associated with unique gene expression profiles and predicts adverse prognosis in pediatric and adult acute myeloid leukemia. *Blood*, **116**, 1747–1754.
33. Desch,A.N., Randolph,G.J., Murphy,K., Gautier,E.L., Kedl,R.M., Lahoud,M.H., Caminschi,I., Shortman,K., Henson,P.M. and Jakubzick,C.V. (2011) CD103 +pulmonary dendritic cells preferentially acquire and present apoptotic cell-associated antigen. *J. Exp. Med.*, **208**, 1789–1797.
34. Malhotra,D., Fletcher,A.L., Astarita,J., Lukacs-Kornek,V., Tayalia,P., Gonzalez,S.F., Elpek,K.G., Chang,S.K., Knoblich,K., Hemler,M.E. *et al.* (2012) Transcriptional profiling of stroma from inflamed and resting lymph nodes defines immunological hallmarks. *Nat. Immunol.*, **13**, 499–510.
35. Narayan,K., Sylvia,K.E., Malhotra,N., Yin,C.C., Martens,G., Vallerskog,T., Kornfeld,H., Xiong,N., Cohen,N.R., Brenner,M.B. *et al.* (2012) Intrathymic programming of effector fates in three molecularly distinct [gamma][delta] T cell subtypes. *Nat. Immunol.*, **13**, 511–518.