BloodSpot 3.0
a database of gene and protein expression data in normal and malignant haematopoiesis
Gíslason, Magnús H; Demircan, Gül Sude; Prachar, Marek; Furtwängler, Benjamin; Schwaller, Juerg; Schoof, Erwin m; Porse, Bo torben; Rapin, Nicolas; Bagger, Frederik otzen

Published in:
Nucleic acids symposium series

DOI:
10.1093/nar/gkad993

Publication date:
2024

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC

Citation for published version (APA):
BloodSpot 3.0: a database of gene and protein expression data in normal and malignant haematopoiesis

Magnús H. Gislason¹, Gül Sude Demircan¹, Marek Prchar¹,², Benjamin Furtwängler¹,³,⁴,⁵, Juerg Schwaller⁶,⁷, Erwin M. School⁵, Bo Torben Porse³,⁴,⁸, Nicolas Rapin⁹ and Frederik Otzen Bagger O¹,∗

¹Center for Genomic Medicine, Rigshospitalet Copenhagen University Hospital, Copenhagen DK-2200, Denmark
²Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen DK-2200, Denmark
³The Finsen Laboratory, Copenhagen University Hospital–Rigshospitalet, Copenhagen DK-2200, Denmark
⁴Biotech Research and Innovation Center, Faculty of Health Sciences, University of Copenhagen, Copenhagen DK-2200, Denmark
⁵Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs. Lyngby DK-2800, Denmark
⁶University Children’s Hospital Basel, University of Basel, Basel, Switzerland
⁷Department of Biomedicine, University of Basel, Basel, Switzerland
⁸Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark
⁹Danish National Genome Center, Copenhagen DK-2300, Denmark

To whom correspondence should be addressed. Tel: +45 3545 4668; Email: frederik.otzen.bagger@regionh.dk

Abstract

BloodSpot is a specialised database integrating gene expression data from acute myeloid leukaemia (AML) patients related to blood cell development and maturation. The database and interface has helped numerous researchers and clinicians to quickly get an overview of gene expression patterns in healthy and malignant haematopoiesis. Here, we present an update to our framework that includes protein expression data of sorted single cells. With this update we also introduce datasets broadly spanning age groups, which many users have requested, with particular interest for researchers studying paediatric leukaemias. The backend of the database has been rewritten and migrated to a cloud-based environment to accommodate the growth, and provide a better user-experience for our many international users. Users can now enjoy faster transfer speeds and a more responsive interface. In conclusion, the continuing popularity of the database and emergence of new data modalities has prompted us to rewrite and futureproof the back-end, including paediatric centric views, as well as single cell protein data, allowing us to keep the database updated and relevant for the years to come. The database is freely available at www.bloodspot.eu.

Graphical abstract

Data Curation

Introduction

BloodSpot (1–4) is a resource that enables convenient studying of gene expression dynamics in blood and bone marrow cells, specifically under conditions of hematopoietic differentiation, blood cell maturation and leukaemia, with a focus on acute myeloid leukaemia (AML). A growing number of carefully
curated datasets have enabled researchers to gain insights into the complex gene expression dynamics guiding the differentiation and maturation of blood cells, thereby fostering hypothesis and acting as orthogonal validation to experiments designed for gaining understanding into both normal and pathological haematologic processes (5,6). The value added to the publicly available raw data has been data curation, quality filtering and, importantly, batch correction and data integration. This work has enabled users to analyse at a glance integrated datasets consisting of samples from different studies, which together can provide new and important insights. Popular mention is datasets with leukaemia blast populations side-by-side with sorted normal populations (7).

Advances in single-cell technologies have unlocked unprecedented resolution in profiling individual cellular states, adding new dimensions to our understanding of cellular heterogeneity within complex tissues. A concurrent realisation with the emergence of single cell data has been the fact that profiles of well-defined cellular populations, as sorted by fluorescence-activated cell sorting (FACS) are important references to understand cellular dynamics, which then also acts as a scaffold to understand function identity and state of unsorted cellular clusters. This has prompted us to maintain focus on both sorted and unsorted cells, whereas other projects solely embrace the unlabelled nature of bulk single cells sequencing data with predicted or inferred cell types (8) typically from droplet-based protocols (9). Recently, single cell proteomics has provided a nuanced understanding of functional protein changes associated with genetic shifts (10), effectively bridging the gap between genotype and phenotype, which we have introduced and will expand as more data becomes available. Moreover, paediatric leukemias, that present themselves with unique molecular signatures and aberrant differentiation patterns, distinct from their adult counterparts, have been added to the database to serve the childhood and paediatric AML and AEL communities (11,12). We will continue to grow this area of the database, aiding in therapeutic discoveries and stratifications, which has been a requested addition by our users.

Materials and methods
Server infrastructure
The newest release of BloodSpot includes a comprehensive update of the backend, in order to accommodate new data and provide a distributed system, which will give a more responsive experience for international users, who have long suffered the physical distance to our servers in Copenhagen. This meant migration to a cloud infrastructure, which warranted revamping HTML and PHP code and MySQL database. The MySQL is now migrated to a version 10.6.15 Ubuntu Maria DB database hotel, and PHP code is updated to a modern version 8.2.

Protein data
Protein data has been processed as detailed in Schoof and Furtwängler et al. (10), briefly, cells from the OCI-AML8227 culture model (13) were FACS sorted and processed for subsequent single-cell proteomics mass spectrometry acquisition.
on a ThermoFisher Orbitrap Exploris 480 mass spectrometer. Data was processed using the SCepitre python package (10). Missing values are not plotted in the web interface.

**Gene expression data**

All gene expression datasets compiled from multiple laboratories have been assessed for batch and quality, and batch corrected using ComBat (14). Metadata from Illaria et al. (12) was compiled into a super-group of age_group and revised World Health Organization (WHO) and to allow for separation on both age and disease classification information.

**Visualisation**

Main plots are SinaPlot (15) implemented in D3, where maximum width is given by total number of samples. This behaviour is no longer default in SinaPlot implementations, but allows for easier comparison of distribution of samples across groups. Both SinaPlots and Tree plots are dynamically produced, whereas all survival plots are pre-computed from TCGA AML data (16). Tree plots are based on differentiation patterns, where applicable, or sample group correlation otherwise.

**Results**

BloodSpot (1–4) has served as quick portal into data of sorted normal cells, AML leukemic blasts by WHO classification (17), and lately also single cell datasets (18,19). The database contains gene expression profiles from microarray and RNA-Seq data of various blood cell types across different stages of haematopoesis. This includes stem cells, progenitor cells and mature blood cells of various lineages, for the investigation of differentiation and maturation. Users can search for genes that correlate with the query gene, making it possible to discover co-regulatory mechanisms and find robust marker sets. With this update we introduce additional AML and AEL datasets with a broad range of age groups that we have used as grouping-criteria, together with disease classification. This will allow the study of age effects, to the benefit of the many users who over the years have requested the capability to analyse transcriptional patterns across paediatric and adult leukaemias. In addition, we open space for new data in the proteomics space, where we have already included a single cell proteome dataset ready for browsing and we look forward to expanding this section as more data becomes available. An overview of datasets added since last release can be found in Supplementary Table S1. We have decommissioned the pathway search functionality, which allowed users to search for gene signatures from MSigDB (20) rather than gene names, as it was not being used.

Collectively, we introduce important expansions of the BloodSpot database. This includes the integration of single-cell proteomics data, that paves the way for a richer perspective of blood cell states, and of paediatric leukaemia profiles, augmenting the potential of BloodSpot to continue to serve as a reference for both basic research and clinical investigators within molecular haematology.

To elucidate age related effects we queried TP53 in the newly added dataset from Iacobucci et al. (12), because it is known to be commonly mutated in adult AML, but virtually never in paediatric cases (11,12). We found an increase in variation and an increase in number of cases with very low or no expression of TP53 co-occurring with increased age (Figure 1).

In order to investigate the protein manifestation of a transiently expressed genes we submitted the query gene MPO, a myeloid differentiation marker (21), to three datasets: the
default dataset with AML and normal sorted hematopoietic cells (2), clustered Lin-CD34+ single cells from Velten et al. (19) and single cell sorted proteome cells (10). MPO clearly has distinct expression in MLL and complex AML from three other karyotypes, and displays transient expression with low expression in the stem compartment and high expression in progenitors (Figure 2A). In the single cell expression data there are also clear signs of transient expression (Figure 2B), with cluster 3 notably showing higher expression. Interestingly, the proteome data (Figure 2C) appears not to show accumulation of the protein from stem cells (LSC) to progenitors (PROG) and blasts, which could be explained by a shorter half-life and effect of RNA compared to protein, or due to the nature of the AML model OCI-AML8227 (13).

Conclusions
The BloodSpot 3.0 update represents a significant improvement in the database’s capabilities, incorporating single-cell proteomics data and expanding its coverage to include pediatric leukemia profiles. These enhancements will empower researchers and clinicians in their studies of hematopoiesis, leukemia and related fields. The database is freely available with no registration at www.bloodspot.eu.

Data availability
BloodSpot 3.0 is available at www.bloodspot.eu. All data is publicly available and linked to relevant repositories as they appear in the web interface and supplementary table 1.

Supplementary data
Supplementary Data are available at NAR Online.

Acknowledgements
We thank warmly Ilaria Iacobucci and Charles G. Mullighan, for early access to data and for kindly assisting us with processed data and metadata.

Funding
Work in the E.M.S. lab is supported by grants from the Novo Nordisk Foundation [NNF21OC0071016]; Independent Research Fund Denmark [case no. 2067-00053B]; Lundbeck Foundation [R413-2022-869]; B.F. is the recipient of a fellowship from the Novo Nordisk Foundation as part of the Copenhagen Bioscience PhD Programme [NNF19SA0035442]. Funding for open access charge: Hospital budget.

Conflict of interest statement
None declared.

References
