Hemap: An interactive online resource for characterizing molecular phenotypes across hematologic malignancies

Petri Pölönen[1]+, Juha Mehtonen[1]+, Jake Lin[2,3]+, Thomas Liuksiala[2,4]+, Sergei Häyrynen[2], Susanna Teppo[4], Artturi Mäkinen[4,5], Ashwini Kumar[3], Disha Malani[3], Virva Pohjolainen[5], Kimmo Porkka[6], Caroline A. Heckman[3], Patrick May[7], Ville Hautamäki[8], Kirsi J. Granberg[2], Olli Lohi[4], Matti Nykter[2]*, and Merja Heinäniemi[1]*

[1]Institute of Biomedicine, School of Medicine, University of Eastern Finland, Kuopio, Finland

[2]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

[3]Institute for Molecular Medicine Finland FIMM, Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland

[4]Tampere Center for Child Health Research, Tampere University and Tampere University Hospital, Tampere, Finland

[5]Tampere University Hospital, Fimlab Laboratories, Tampere, Finland

[6]Hematology Research Unit Helsinki, University of Helsinki and Helsinki University Central Hospital Cancer Center, Department of Hematology, Helsinki, Finland

[7]Luxembourg Center for Systems Biomedicine, University of Luxembourg, Esch-Belval, Luxembourg

[8]School of Computing, University of Eastern Finland, Kuopio, Finland

+ equal contribution * co-corresponding authors

Conflict of interest statement: CH has research funding from Celgene, Novartis and Orion Pharma unrelated to this study.

Word count: 5849, Number of figures: 5, Number of Tables: 1. The supplementary material consists of supplemental methods and Supplementary display items; 5 figures and 6 tables with their legends.

Running title: Organizing molecular diversity of hematologic malignancies

Corresponding authors:
Associate Prof. Merja Heinäniemi
School of Medicine, University of Eastern Finland, Kuopio, Finland
Tel. +358 41 4305724
Email: merja.heinaniemi@uef.fi

Prof. Matti Nykter
Faculty of Medicine and Health Technology
Tampere University, Tampere, Finland
Tel. +358 40 5267884
Email: matti.nykter@uta.fi

Doc. Olli Lohi, MD
Faculty of Medicine and Health Technology
Tampere University, Tampere, Finland
Tel. +358 50 318 6254
Email: olli.lohi@uta.fi

Abstract

Large collections of genome-wide data can facilitate the characterization of disease states and subtypes, permitting pan-cancer analysis of molecular phenotypes and evaluation of disease context for new therapeutic approaches. We analyzed 9,544 transcriptomes from more than 30 hematologic malignancies, normal blood cell types, and cell lines, and showed that disease types could be stratified in a data-driven manner. We then identified cluster-specific pathway activity, new biomarkers and in silico drug target prioritization through interrogation of drug target databases. Using known vulnerabilities and available drug screens, we highlighted the importance of integrating molecular phenotype with drug target expression for in silico prediction of drug responsiveness. Our analysis implicated *BCL2* expression level as an important indicator of venetoclax responsiveness and provided a rationale for its targeting in specific leukemia subtypes and multiple myeloma, linked several polycomb group proteins that could be targeted by small molecules (SFMBT1, CBX7 and EZH1) with CLL, and supported *CDK6* as a disease-specific target in AML. Through integration with proteomics data, we characterized target protein expression for pre-B leukemia immunotherapy candidates, including DPEP1. These molecular data can be explored using our publicly available interactive resource, Hemap, for expediting therapeutic innovations in hematologic malignancies.

Significance

This study describes a data resource for researching derailed cellular pathways and candidate drug targets across hematological malignancies.

Introduction

Gene expression profiles facilitate genome-wide analyses that can stratify patient subtypes and identify the activity patterns of various cellular pathways under different biological conditions (1-2). Even though a large number of studies have characterized hematologic malignancies and normal blood cell types at genome-wide level since the introduction of microarray technology, most include only tens to hundreds of samples and focus on one disease. Thus, understanding the complete heterogeneity and similarity of diseases states and their subtypes remains an open challenge. Moreover, many hematologic malignancies are rare on the population level, necessitating collecting data across study cohorts.

Hematological malignancies include acute and chronic leukemias of myeloid and lymphoid lineage, B-, T- and NK cell lymphomas, and multiple myeloma (MM), and a number of premalignant conditions such as myelodysplastic syndrome (MDS), and myeloproliferative neoplasms (MPN). These diseases have highly variable genetic features, unique clinical courses, and varying therapeutic approaches. There is also a marked difference in prevalence, genetic background and prognosis between adult and pediatric blood cancers. In children, acute lymphoblastic leukemia (ALL) is the most common hematological malignancy, while in adults, non-Hodgkin lymphomas (NHL), followed by MM, chronic lymphocytic leukemia (CLL), and acute myeloid leukemia (AML) are the most common. Treatment is moving towards increased utilization of targeted therapies in combination with traditional chemotherapies. Targeted therapies include tyrosine kinase inhibitors such as those developed against BCR-ABL fusion found in CML and some ALL cases, or antibody therapies including CD38-targeting in MM, and engineered CAR-T cells recognizing cell surface CD19 or CD22 antigens in relapsed ALL and NHL (3-5). Yet, current therapies to treat hematologic malignancies rely heavily on drugs that target DNA metabolism in actively proliferating cells or intracellular signaling events that are involved in proliferation (6). Although these drugs have markedly improved progression-free survival, redundancy in signaling and the failure to eradicate quiescent cells (7) can facilitate the rapid development of therapy resistance. Testing a wider portfolio of new drug targets, or repurposing drugs with established clinical indications represent promising strategies (6-7). Molecular profile guided approaches hold promise to improve the efficiency of this process (8).

We present here a resource that organizes samples from cancer patients, healthy donors and those at pre-malignant stages for comparative analysis based on both curated annotations and data-driven clustering of molecular phenotypes. This hematologic pan-cancer analysis permits the identification of clinically relevant molecular features and the exploration of new drug targeting approaches across the disease hierarchy. The data and analysis tools are made available as an interactive online resource, Hemap, http://hemap.uta.fi/ that synthesizes the curated genome-wide data across different disease subtypes.

Materials and Methods

Dataset retrieval and extraction of sample annotation data

Transcriptome datasets for Hemap were retrieved from the NCBI Gene Expression Omnibus (GEO) database (9) and represent samples hybridized to hgu133Plus2 genome-wide microarrays. The meta-data were retrieved based on matching disease ontology terms for hematologic malignancies against sample annotations (R/Bioconductor GEOmetadb package, “gsm” and “gse” tables), followed by manual curation, resulting in 10,470 samples. Refer to Methods Supplement for details. Eight leukemia types, 8 B-cell lymphoma types, 7 T/NK lymphomas, multiple myeloma and 4 proliferative disorders are represented by primary patient samples, while in total 36 disease types are included considering also their sub-categories and cell line data (**Tables S1** and **S2**).

Data preprocessing and quality control

Samples with a median of raw probe intensity distribution deviating more than 1.5 in log2-scale from the median of all medians were deemed outliers and filtered out as well as those with an interquartile range (IQR) deviating more than 0.75 from the median of IQRs. Finally, duplicate samples, as well as all disease types with less than 3 samples (and samples assigned to these), were removed, resulting in 9,544 samples that were processed using the RMA probe summarization algorithm (10) with probe mapping to Entrez Gene IDs (from BrainArray version 18.0.0, ENTREZG). Finally, we employed the bias-correction method (11) to correct for any remaining technical differences (**Fig. S1**). BeatAML (12) clinical and mutation data was obtained from source data file 41586_2018_623_MOESM3_ESM.xlsx. RNAseq counts were obtained from the authors. Genes with over 1 cpm expression in over 1 % of the samples were kept. Data was normalized using limma voom and quantile normalization.

Dimensionality reduction

Dimensionality reduction methods are unsupervised methods that use measures of (dis)similarity and an optimization strategy to return as output sample coordinates in a lower dimension. Metrics of continuity, trustworthiness and k-NN error were used to assess how well the visualization in 2D preserved their relative placement in the original coordinate space. We tested Gaussian Process Latent Variable Model (GPLVM) (13), Locally Linear Embedding (LLE) (14), Principal Component Analysis (PCA) (15), Probabilistic Principal Component Analysis (PPCA) (16), Sammon Mapping (SM) (17) and t-Distributed Stochastic Neighbor Embedding (t-SNE) (18) (see Methods Supplement for parameters). Comparison of the different methods encouraged the selection of t-SNE maps, specifically the Barnes-Hut approximated version of t-SNE implementation (BH-SNE) (19). In final analysis 15% genes with highest variance were used in construction of t-SNE maps (see (20) for justification).

Assignment of cluster centers on t-SNE maps

Kernel density-based clustering algorithm (mean-shift clustering with bandwidth parameter set to 2.5, LPCM-package in R), was used to cluster the data following the dimensionality reduction. This method allows the discovery of sample sets which share similar features without having to pre-specify the number of clusters. The term “cluster” is used in the text to refer to this computational clustering result, and the term “group” is used in context of visual examination. Pairwise statistical analysis between different sample features and clusters was performed as in (21), based on Spearman correlation and the Bonferroni method for multiple hypothesis testing correction (see Methods Supplement for details).

Discretizing gene expression with mixture models

Microarray hybridization generates background signal, which we would like to distinguish from real expression signal. The large sample size of Hemap was leveraged for fitting gene-specific models to cluster the gene expression in two components (expressed and not detected, denoted by 1 and -1, respectively). Gaussian finite mixture models were fitted by expectation-maximization algorithm (R package mclust version 4.3). If the uncertainty value from the model was more than 0.1, we assigned a value of 0 to denote low level. Additionally, each log2 expression value lower than 4 was assigned a value -1 and values higher than 10 a value of 1. This was done to assure minimal amount of misclassifications of data samples to wrong components. The model was chosen by fitting both equal and variable variance models and ultimately choosing the model which achieved a higher Bayesian Information Criterion (BIC) to avoid overfitting. For drug target analysis, we utilized an adjustment for genes where background distribution was not found (gene is always expressed), or if over 90% of the samples had uncertain expression based on the model classification. Expressed state was assigned when >60% of the uncertain samples had expression above 6. Not detected status was re-evaluated similarly (60% at level below 6).

Gene set analysis

The pathway and gene set enrichment analysis available in the Hemap resource was generated based on gene sets retrieved from MsigDB v5.0 (22) (molecular signatures), Wikipathways (06.2015) (23), Recon 1 (24) (metabolic pathways), Pathway Commons 7 (25) and DSigDB v1.0 (26) (drug targets). Gene sets were limited to contain between 5 to 500 expressed genes (as defined above) per gene set, resulting in 19,680 gene sets that were evaluated across the dataset. The gene set variation analysis (GSVA) (27), GSVA package 1.13.0 in R, was used to calculate gene set enrichment scores (positive for increased and negative for decreased expression) for each sample (parameters mx.diff=F, tau=0.25, rnaseq=F). Significance was evaluated based on empirical *P*-values calculated using 1000 random permutations of genes within the gene set, separately for gene set sizes 5-20, 25, 30, 40, 50, 75, 100, 200, 300, 400, and 500 to correct for differences in gene set sizes. Hypergeometric test was used to calculate enrichment of significant scores in a specific cluster.

Data sources used for evaluating drug targeting approaches

Drugs in clinical trials for leukemias, lymphomas or multiple myeloma were obtained from ClinicalTrials.gov (accessed March 7th, 2018) maintained by the U.S. National Institutes of Health, including ongoing and terminated trials. Leukemia clinical trials were further sorted to those with clinical indication associated with AML, pre-B-ALL, CML, CLL or multiple leukemia types. Drugs in use based on approved status in Finland were provided by the Finnish Pharmaceutical Information Centre Ltd and drugs approved by the Food and Drug Administration (FDA) for leukemia, lymphoma and myeloma were queried from FDA website (fda.gov – Drugs – Information on Drugs) (**Table S3**). A table of gene level details for each drug was obtained from DSigDB (26) (DSigDBv1.0 Detailed.txt) and integrated to Hemap *in silico* drug screening analysis. The list of drugs targeting epigenetic modifiers is based on the gene list with 124 genes available from ChEMBL_20 Target Classification Hierarchy (28) (**Table S4**). Analysis using TTD (Therapeutics Targets Database) (29), DGIdb3.0 (30) for FDA approved drugs across a wider disease context (31) as a source database was based on a total of 11,373 unique drugs and 1270 unique genes. Drugs in use and in clinical trial included high confidence targets that were reported in several databases or had an associated Pubmed identifier. A surface marker gene list with total of 996 genes was obtained from Cell Surface Protein Atlas (32) to evaluate putative novel immunotherapy targets.

Drug target *in silico* analysis in hierarchical manner

A disease hierarchy: 1. All disease samples; 2. disease combinations; 3. leukemia, lymphoma, myeloma; 4. AML, pre-B-ALL, T-ALL, CLL; 5. disease clusters; was used to evaluate disease or subtype specific drug target expression. Statistical significance of binary feature enrichment (e.g. high expression state) in a particular sample group was first evaluated using the hypergeometric enrichment test, followed by Bonferroni adjustment of *P*-values. If >90 % of the samples had high expression for a gene in the disease context, Inf score was assigned instead of -log10 *P*-value (hypergeometric test would not be meaningful if the sample size was close to the whole population). Each significant gene was uniquely assigned to the disease group with the lowest *P*-value. In the case of equal *P*-values, a broader disease category was prioritized using the disease hierarchy. As a second filtering level, the Wilcoxon test was used to test whether the drug target gene is expressed at higher level in cancer compared to normal erythroid, myeloid, B-lymphoid or T-lymphoid samples. One normal sample group comparison was accepted for downstream analysis (with the respective comparison annotated as failed). *In silico* drug analysis was benchmarked using two case studies: drugs from Frismantas et. al. (33) and secondly known vulnerabilities (in clinical use/trial). Success rate was reported for drug target gene expression in disease, specificity for disease/subtypes and higher expression relative to normal cells.

BeatAML drug analysis

Spearman’s correlation was computed for each drug area under curve (AUC) values and cancer-map clusters, drug target genes or target gene mutations. Furthermore, mutations with at least 5 observations and significant correlation adj. P<0.05 to drug AUC values or significant fisher exact test adj. P<0.05 in cancer-map clusters were added as features that could improve drug sensitivity prediction.

From total of 122 drugs 47 were excluded based on three criteria. First, 25 drugs with IC50 lower quartile below 10 nm were excluded as these drugs have limited efficacy. Second, 9 drugs with less than 80 samples with measured drug responses were excluded. Third, only drugs with drug target information were kept, resulting in total of 75 drugs. The elastic net implemented in glmnet (34) was trained using tenfold cross-validation using caret (35) trainControl and repeatedcv method. Caret function train and its functionality tuneGrid was used to optimize alpha parameter denoting the L1 and L2 regularization term proportions for elastic net. Each drug had three categories of features to fit the model: clusters, drug target gene expression, or mutations. To test the importance of each category in model fitting, sample order was randomly shuffled for one category while the original order was preserved for the other categories. Therefore, if the shuffled category features were important for the model fit, model overall fit should decrease as the other features are unchanged. This process was repeated 100 times and median of RMSE and R^2 values were computed. Only drugs with good fit when using all the features were kept, having R^2 over 0.25 and RMSE less than 0.9.

Drug sensitivity testing using patient and healthy donor samples

Bone marrow (BM) aspirates or peripheral blood samples were obtained from AML patients (*N*=52) and healthy donors (*N*=15) after informed written consent using protocols approved by a local Institutional Review Board and in accordance with the Declaration of Helsinki. Mononuclear cells (MNCs) were isolated by density gradient separation (Ficoll-Paque PREMIUM; GE Healthcare, Little Chalfont Bucks, UK) and immediately used for drug testing. Cells were maintained in Mononuclear Cell Medium (MCM; Promocell) or in a 25% HS-5 conditioned medium plus 75% RPMI 1640 medium mix (CM). Palbociclib and idarubicin (from Selleck, Houston, TX) were solvated in dimethyl sulfoxide and plated in 5 different concentrations in 10-fold dilutions on 384-well plates using an Echo acoustic dispenser (Labcyte, Sunnyvale, CA, USA), 1-10 000 nM for Palbociclib; 0.1-1000 nM for Idarubicin. 10 000 cells were added per well and incubated with the drugs for 3 days at 37°C, 5% CO2. Viability was measured using the CellTiter-Glo reagent (Promega, Madison, WI, USA) according to the manufacturer's instructions and using the PHERAstar (BMG LABTECH, Ortenberg, Germany) or SpectraMax Paradigm (Molecular Devices, Sunnyvale, CA, USA) plate readers. Sensitivity to the drugs was quantified using a drug sensitivity score (DSS), which is a modified area under the curve based metric described previously (36). A selective DSS value was calculated by subtracting the mean DSS of the healthy BM controls from the DSS of individual AML samples.

Immunohistochemistry

Anti-DPEP1 antibody (Atlas antibodies, rabbit polyclonal IgG against human renal dipeptidase 1, product number: HPA009426, lot number: A57960) was used with the dilution 1:2500 to stain formalin fixed and paraffin embedded bone marrow trephine biopsy samples from pediatric pre-B-ALL patients from the Pirkanmaa ERVA area between the years 2000 and 2017. 126 diagnostic samples (including also one Burkitt's lymphoma and a T-lymphoblastic leukaemia/lymphoma case) were stained with a Ventana Benchmark GX using UltraView Universal DAB Detection Kit. Cytoplasmic and membranous staining was graded negative if less than 20 percent of the leukemic blasts were stained, positive if 20 percent or less than 50 percent of the blasts were positively stained and strong positive if 50 percent or a greater proportion of the blasts were positive. The analysis was performed by two pathologists without the knowledge of the patient data or the interpretation of the other analyst. The samples and clinical data were studied with the approval of the Tampere University Hospital Ethical committee (#R16054 and #R13109) and in accordance with the Declaration of Helsinki.

Interactive web resource for data analysis

The interactive online resource and the accompanying user guide for the Hemap resource are described in more detail in the Methods Supplement and available at http://hemap.uta.fi//

Results

Integrating transcriptomes to characterize molecular states across hematologic malignancies

For the comparative analysis of hematologic malignancies on molecular level, we assembled gene expression profiles from the NCBI GEO database (9), comprising patient samples representing different cancers and proliferative disorders, and including cell lines and normal blood cell types as controls. Sample annotations were curated, and each sample was assigned a disease category. After data quality control and bias correction (see **Methods**, **Fig. S1**), 9,544 samples comprise the final dataset (denoted “Hemap” samples) for subsequent analysis, including 7,279 patient samples (mainly diagnostic) from hematologic malignancies (**Fig. 1A**, **Tables S1** and **S2**).

To enable discovery and statistical comparison of previously known and novel molecular phenotypes alongside the annotated disease classes, we utilized a data-driven approach that allows discovery of sample groups and visualizes these for interpretation. First, we compared dimensionality reduction methods that allow visualization of complex data in two dimensional space. The data representation accuracy was quantitatively assessed using the metrics of continuity, trustworthiness and k-nearest neighbor (k-NN) classifier error (see **Methods, Fig. S2**). As a result, t-Distributed Stochastic Neighbor Embedding (t-SNE) (18) and its approximation, Barnes-Hut-SNE (BH-SNE) (19), was selected, as it performed robustly (continuity and trustworthiness, 0.9860 and 0.9943, respectively) in two dimensions and still preserved the neighborhood structure (k-NN error 0.0668) (**Figure S2**). The t-SNE map was then utilized for density-based clustering to assign each sample to a cluster (**Fig. 1B**, see **Methods** for details) and the results were compared to annotated disease classes (**Fig. 1C**). We conclude that both quantitative and biological assessments confirm that our approach faithfully organizes the samples in an unsupervised manner based on their molecular phenotype and disease type. We denote the resulting data-driven sample grouping as the Hemap "cancer-map" in the following text.

Comparative analysis associates clinical annotations and pathway activity to the molecular disease stratification

The 2D cancer-map revealed a clinically relevant sub-structure (**Table S2**), as exemplified by the different B-cell lymphomas and pre-B-ALL cytogenetic subtypes (colored in **Figs. 2A** and **B**, respectively), providing biological validation for separation of relevant phenotypes on the cancer-map. A detailed comparison to annotations is presented in **Table S2**. Next, statistically significant associations of clusters with gene expression levels, clinical annotations and pathway enrichment scores across different databases were calculated (see **Methods)**. These results can be interactively tabulated and visualized using the online Hemap resource. We selected five most significant pathways at disease cluster level, or those matching pre-B-ALL subtype clusters (**Fig. 2C**) for visualization in a heatmap (see also Table S5). In AML, the pathways for hematopoietic stem cell differentiation, pentose phosphate pathway, renin-angiotensin system, IL-8/CXCR1-mediated signaling events and C-MYB transcription factor networks were most significantly enriched. These reflect well the known progenitor-like phenotype of AML cells. Pentose phosphate pathway, on the other hand, represents a recently uncovered vulnerability (37,38) that is important for AML growth. Similar disease-relevant pathways were also uncovered from T-ALL (TCR pathway), CLL (BCR signaling pathway), lymphomas (cell adhesion molecules (39)) and multiple myeloma (N-glycan biosynthesis (40,41)). In pre-B-ALL clusters, processes related to transcriptional regulation were highly significant (including histone modification, CTCF pathway, and RNA processing). WNT signaling (42,43) was found as a cluster 29-specific (t1;19) enriched pathway, which matches its known relevance in these TCF3-PBX1 fusion carrying cases. Samples expressing a gene or pathway of interest can be visualized as shown in **Fig. 2D**, distinguishing the progenitor-like MLL subtype of pre-B-ALL based on the lack of expression of the differentiation marker *MME* (also known as *CD10*) that is used in clinical diagnostics

(**Fig. 2E**). Similarly, most significant associations between disease clusters and drug signatures can be examined by e-staining their significance (in red), as illustrated by association of PI3K inhibitor BEZ235 gene set signature from DsigDB to pre-B-ALL (**Fig. 2F**), which validates a known association between a drug and a disease subtype. Further analysis on the BEZ235 gene set and several case studies on how to generate novel hypothesis are presented in the accompanying **User guide** to demonstrate different analysis (refer to “Explore” and “e-staining” examples).

Pan-cancer analysis to recognize vulnerabilities across disease contexts

Parallel to molecular stratification, the diversity of patient profiles in Hemap has the potential to support the development of new therapeutic strategies by leveraging the information about the expression profiles across hematologic malignancies. We analyzed the specificity of drug target expression states across patient groups in a hierarchical manner (**Methods**), as illustrated in **Fig. 3A** (see also **Table S3** for a list of drugs and their targets and **Table S4** for significant associations listed by disease hierarchy). The corresponding significance ranking for targets of approved drugs is shown as heatmaps in **Fig. 3A-B**, where the columns represent different disease contexts and gene targets (in rows) are sorted according to their most significant association. The clinical indication for the drug(s) that could be used to target each gene is indicated in the panel on the right, while e-staining results for example drug targets are shown in **Fig. 3C** (see also **Fig. S3**). Proteasome targeting drugs Bortezomib and Carfilzomib are in use for lymphomas and multiple myeloma. Accordingly, 10/20 genes encoding the proteasome subunits are associated to this disease hierarchy level, or to the pan-cancer category, with highest significance (**Fig. 3A**). In comparison, for precision drugs such as the antibody drugs Elotuzumab (*SLAMF7*, P-val < 1e-315) or Daratumumab (*CD38*, P-val 1e-196) approved for MM, or Rituximab (*MS4A1*, P-val < 1e-315 in LY+CLL) used in lymphomas and CLL (**Fig. 3A**) the specific gene targets can be examined. Among all known vulnerabilities (drugs in clinical use / trial) a gene-level analysis detected 84% of targets expressed and 69% were associated with highest specificity score (-log10 P-value) to the respective disease context (see **Table S3**). This is exemplified by the comparison of genes with significant association to lymphoid leukemias (**Fig. 3B)**. *BCL2* targeted by venetoclax is shown as an example of an approved target in CLL that our analysis associates with this disease context and with potential for targeting in MM. The genes marked with asterisk, including *IL2RA* indicate targets of drugs approved for other hematologic malignancies. Our analysis associated these with re-purposing potential in CLL and/or ALL. *FLT3* is a recently approved target with disease cluster-specific expression in B-lymphoid and myeloid leukemias.

Utility of molecular disease stratification for evaluating drug screen results

Next, we examined leukemias at disease subtype level from two *ex vivo* drug screening datasets (12,33). Venetoclax had lower efficacy in T-ALL vs. B-ALL and lowest efficacy was in t1;19 samples in the ALL drug screen (33) which agrees with Venetoclax target *BCL2* gene expression in Hemap (**Fig. 3C)**. Topotecan and dasatinib had the opposite profile, also in agreement with subtype-specific expression of their targets *TOP1MT* and *LCK* (**Fig. S3**). Taken together, out of 15 drugs from this ALL screen tested with our hierarchical analysis, 14 (93 %) had a candidate target expressed and 12 (80%) received highest target indication in ALL (**Table S4)**. Using the larger beatAML dataset (12), we set out to examine in an unbiased manner what matters more in predicting drug responsiveness: target expression, genetic lesions traditionally used to stratify patients, or the molecular phenotype as defined by clustering of transcriptome states. We implemented models using elastic nets, where a model for each drug (75 in total) was fit using these three categories of features. To test their importance for model fitting, sample order was randomly shuffled for one category while the original order was preserved for the other categories. The results for 11 drugs that achieved the best overall model fit (R^2>0.25) are shown in **Fig. 4A,** including Venetoclax, Panobinostat (HDAC inhibitor), Palbociclib (CDK4/6 inhibitor), 7 kinase inhibitors (many targeting FLT3) and an ALK inhibitor. The average R^2 value from 100 tests is colored in the heatmap and summarized as a boxplot next to it. If the shuffled feature was important for the model fit, a decrease in R^2 is expected (shift from darker red to dimmer or blue colors) as the other features are unchanged. For venetoclax, this analysis implicated target gene expression as the main predictor (**Fig. S4**). For FLT3-targeting compounds, *FLT3* mutation status was implicated as the top predictive feature (**Fig. S4**). However, overall, the lack of cluster features in the model resulted in lowest predictive power. The disease clusters were the best predictors for Palbociclib and Panobinostat, whereas mutation status had no effect on their model fit. Panobinostat and Palbociclib showed opposite drug responses in clusters 13, 2, 6 compared to cluster 1 (**Fig. 4B**). Hemap clusters 17, 5, and 6 corresponded to these clusters (**Fig. S4**) and were similarly enriched for NPM1 and FLT3 mutations or PML-RARA fusion in both data sets. Comparison of clinical phenotypes revealed that blast morphology was different between the clusters, linking maturation level to the differential drug response (**Fig. 4C and Fig. S4**).

Classical targets involved in DNA metabolism (*TOP2A* and *B*) and clinically interesting targets, including *CDK6, BCL2, MDM2* and *VEGFR2* from clinical trials, ranked highly in our disease hierarchy analysis, as shown in **Figs. 3** and **S3**. However, when compared to normal cell types, only 7% of the targets had higher expression in disease than in normal cells (**Table S3**). Palbociclib target *CDK6* is highly expressed in all acute leukemias compared to normal blood cell types, while *TOP2A* has high mRNA levels also in normal blood cells (**Fig. 4D**). To evaluate drug sensitivity that is specific to cancer cells, an experimental *ex vivo* screening approach is exemplified in **Fig. 4D** by comparing in AML patient cell responses to the CDK4/6 inhibitor Palbociclib and Idarubicin targeting *TOP2A* (see **Methods**). Drug sensitivity and selective drug sensitivity scores (DSS and sDSS, respectively, see **Methods**) (36) are compared in box plots (**Fig. 4E**). Overall, the AML patient bone marrow *ex vivo* cultures were more responsive to Idarubicin (refer to **Fig. S4** for AML cell line data). However, a negative score indicating higher effect on normal bone marrow cell viability was observed for Idarubicin in a larger fraction of AML cases compared to Palbociclib. This observation of non-specific response, implied by negative sDSS score, is consistent with our predictions from Hemap data. Therefore, the normal samples included in Hemap could provide valuable additional information for drug target selection. Comparison of BCL2 and BCL2L1 (also known as BCL-XL) levels are presented as another example in **Fig. S4**, relevant to Venetoclax vs Navitoclax toxicity in targeting apoptosis. The Advanced Use Case in the Hemap **User guide** extends this analysis using pathway activities and drug chemical screen data.

Evaluating new therapeutic strategies in a pan-hematologic cancer context

Epigenetic regulation has emerged as an important mechanism that can corrupt the gene regulatory network (44), motivating novel therapeutic approaches. Utilizing the disease spectrum in Hemap, we performed a pan-hematologic cancer analysis of epigenetic modifiers (**Table S5**), focusing on genes encoding proteins that are validated targets of small molecule drugs (available from ChEMBL (28)). We found elevated expression of this set of genes significantly enriched in CLL, T-ALL and clusters 28 (pre-B-ALL) and 32 (AML) (**Fig. 5A**, hypergeometric test adjusted *P*-values 0.0003, 0.0074, 0.0127, 0.0174, respectively, see also **Table S5** for additional mutation frequency information (45) for the genes shown). The expression state for six most significant genes from CLL are shown on the Hemap cancer-map (**Fig. 5A**) and from independent validation RNA-seq data (46) (**Fig. 5B**).

A second promising new strategy, immunotherapy, can kill cancer cells by targeting surface proteins with antibodies (47) or chimeric antigen receptors (48). However, side effects due to targeting normal blood cells along with development of resistance occur (49). To provide a rational basis for extending the target repertoire, we used disease hierarchy analysis to rank 996 candidates available in the Cell Surface Protein Atlas (32) (**Table S6**) resulting in broad, disease and subtype-specific candidates. The top ranked candidate genes in our analysis correspond to those that are uniformly high expressed within the specified disease context. The stem cell antigen CD33, actively pursued for treatment of AML (50), is among highly ranked surface targets in clinical trials shown in **Fig. S3**. Next, we obtained proteomics profiles from 19 B-ALL patients (51) to compare our ranked list for pre-B-ALL (refer to **Table S6**) to protein-level expression. The trend between *in silico* drug target rank and protein detection rate is plotted in **Fig. 5C.** Validation rate for top candidates was above 75%. The highly ranked surface targets *CLEC14A, DPEP1, CELSR2, MME, SDK2, INSR, GPM6B, ELFN2, FLT3, SLC22A16, FLT4* and *APCDD1* correspond to those with higher expression in pre-B-ALL patients compared to normal blood cells (see also **Fig. S5**). The high gene expression state of *DPEP1* (**Fig. 5D**) in pre-B-ALL was further validated at protein level based on immunohistochemistry of diagnostic bone marrow biopsies. The grading from 117 ALL bone marrows and 16 samples representing other lymphoid malignancies or normal lymphoid tissues is presented in **Table 1** and illustrated in **Fig. 5E**.

To further facilitate the utilization of the data, pre-calculated results are accessible via our interactive web resource (http://hemap.uta.fi/) including the expression state for 4,277 drug target gene sets and 1,094 drug response signatures, which can be further investigated in the context of the 12,433 pathways and molecular signatures (see **Methods** and **User guide** examples). Disease hierarchy analysis for the curated list of drug to target gene associations (11,373 drugs; 1,182 genes) from the Therapeutic Target Database (TTD) (29), DGIdb (30) and targets of FDA approved drugs across disease (31) is available in **Table S4**. In this manner, *in silico* drug target selection based on Hemap can leverage gene and pathway expression, as evaluated across cancer types and normal blood cell types.

Discussion

The integration of available genome-wide data from patients allows uncovering shared disease mechanisms and new therapeutic options. Recent work has highlighted that molecular and genetic data that helps stratify patients can dramatically increase the likelihood of success during clinical development (8,52). However, in several cancer types, including those of hematopoietic and lymphoid tissues, the majority of data have been collected by separate studies concentrating on certain cancer types, which hinders the identification of cancer type specific features. We present an interactive online resource, Hemap (http://hemap.uta.fi/) for analysis across multi-center gene expression datasets to investigate disease subgroups and compare molecular phenotypes across 9,544 samples from hematologic malignancies.

In practice, the samples included to Hemap are inaccessible to most clinical researchers. The Hemap resource serves to re-purpose data from public repositories for clinical interpretation in an intuitive manner that does not require data analysis expertise. In future versions of Hemap, we plan to include also RNA-seq studies. Presently, the resource already contains the TCGA AML dataset and the User Guide includes examples using this data. Alongside curated disease assignment, we present a data-driven approach that organizes and integrates heterogeneous sample collections in an unbiased manner. To facilitate this, we demonstrated how unsupervised clustering and dimensionality reduction methods, here by the t-SNE method, can be used for organizing the molecular profiles for further downstream analysis. The high level of performance of t-SNE has been shown in context of various data types (18,53-54). In this manner, genes characterizing the patient clusters can be identified for further delineation of their functional role. In CLL, our analysis implicates high expression of several polycomb group proteins (SFMBT1, CBX7 and EZH1) in CLL that could be targeted by small molecules, in line with chromatin state data (46), and their mutation (45) frequencies, highlighting the importance to consider the spectrum of genetic and epigenetic changes in these malignancies. Earlier studies have implicated epigenetic plasticity as a key driver of CLL evolution during treatment (55). Specifically, CLL cases had little to no genetic subclonal evolution, while significant recurrent DNA methylation changes were enriched for regions near Polycomb targets (55). To further elucidate the mechanisms, inclusion of post-treatment data and integrating methylation, chromatin marker and mutation profiles represent important future directions in developing the Hemap resource.

From a therapeutic perspective, approaches for the development of treatment strategies with a broad disease focus and molecular subtype resolution are urgently needed. We used Hemap to provide a roadmap for candidate drug therapies that allows prioritizing new candidates based on disease-specificity. Our analysis recapitulated known vulnerabilities, providing additional confirmation for targets in current clinical trials: Several compounds targeting Bcl2 have been developed and have shown promise in treating both CLL and Non-Hodgkin's lymphoma (56-57). However, navitoclax that also targets *BCL2L1* (also known as BCL-XL) displays platelet toxicity. This potential for off-target effects was visible as high gene expression level in the erythroid lineage, supporting the choice of the more selective venetoclax. The prevalent high expression also in MM and pre-B-ALL found in our study provides a rationale for the expansion of the testing of these compounds in lymphoid malignancies. This suggestion is additionally supported by a recent study showing that these compounds have promise in MLL-rearranged ALL (58), a pre-B-ALL subtype corresponding to cluster 29 in our dataset. However, Hemap analysis predicts insensitivity in T-ALL and t1;19 subtype, matching recent ALL drug screen data (33). Similarly, the elevated expression of the p53-regulating MDM2 in pre-B-ALL fits with recent data on successful application of antagonists in clinical trials (59), and mechanisms for its high expression ETV6-RUNX1-positive leukemias (60).

Presently no drug screens have been carried out in primary patient cells across the spectrum of hematologic malignancies in Hemap. The utility of Hemap for drug repurposing was demonstrated in our recent study that identified dasatinib as a targeted therapy for a subgroup of T-ALL patients (61). Here, we examined drug screen datasets to examine how differential drug responsiveness could be linked to disease sub-clusters and drug targets identified from the cancer maps. Using the beatAML dataset, we systematically compared the importance of mutations, clusters and drug target gene expression in predicting drug responses. Clusters were the best predictors of drug response for drugs with best overall model fit. However, the importance of each predictor was largely influenced by drug type. Best predictor for Venetoclax response was *BCL2* expression level. *FLT3* mutation status and other mutations were the best predictors for kinase inhibitors. In contrast, disease clusters were the best predictors for Palbociclib and Panobinostat responses to which mutation status had no effect on model fit. Comparison of clusters in which Panobinostat and Palbociclib showed opposite drug responses revealed that blast morphology was different, linking maturation level to differential drug response. Furthermore, their drug targets were differentially expressed in these clusters, pointing out the importance of integrating context and drug target expression for *in-silico* drug screening. Surprisingly, the HDAC expression pattern revealed cytosolic members (*HDAC6* and *HDAC10* (62)) as resistance markers, while nuclear *HDAC4* and *HDAC9* correlated with sensitivity. Our analysis also supported *CDK4/6* as disease specific targets that are known to act as critical activators of normal and leukemic HSC (63). Here, Palbociclib compared favorably to Idarubicin regarding patient blast sensitivity against normal bone marrow cells, reflected in mRNA data from Hemap. The selectivity over normal cells may improve further using combination therapy (63) that allows decreasing the dose. However, additional parameters such as drug target protein level, drug metabolism and cell proliferation rate further contribute to sensitivity and therefore not all patients matching a molecular subtype or expressing the target mRNA can be expected to respond favorably.

Cancer cells display remarkable plasticity: resistance to recently approved CD19-targeting CAR-T therapy has been shown to occur via mutations or splicing defects at the CD19 locus or lineage-switching (49). To combat the diversity of resistance mechanisms, there is a demand to diversify the target repertoire. In pre-B-ALL, we identified promising surface protein candidates, prioritizing targets with consistently high levels within the given disease context and low levels in normal blood cell types. Over 75% of the highly ranked candidates were confirmed using proteomics (51), and additional literature confirmation was found for five candidates. Moreover, we validated DPEP1 as a potential surface target in pre-B-ALL by immunohistochemical staining of diagnostic bone marrow biopsies. Positive staining was found in each subtype for majority of cases, except in MLL where both the Hemap gene expression data and protein staining indicated low or undetectable levels. The validation cohort consisted of pediatric cases, while Hemap analysis included also adult samples. DPEP1 is a zinc-dependent metalloproteinase that is expressed aberrantly in several cancers, and has been implicated as a potential therapeutic target in colon and pancreatic cancers (64,65). In future, increased availability of protein-level data from different hematologic malignancies will allow confirming additional targets.

In conclusion, the interactive Hemap resource facilitates comparative analyses across multiple hematologic malignancies. We envision that the mechanistic insight gained by concomitant identification of molecular subtypes, genetic aberrations and derailed cellular pathways will expedite therapeutic innovations and clinical utility.

Acknowledgements

We would like to thank the Finnish Pharmaceutical Information Centre Ltd for information provided regarding approved drugs in hematologic malignancies, Dr. Aik Choon Tan for the advice with drug signature gene sets, staff of the FIMM Technology Center High-Throughput Biomedicine Unit and Sequencing Lab. Members of the T. Aittokallio, C.A. Heckman, O. Kallioniemi, S. Mustjoki, K. Porkka and K. Wennerberg groups for their help to analyze drug response profiles, and Dr. Sheila Reynolds and Roger Kramer for assistance with feature matrix analysis. The work was supported by grants from the Academy of Finland (project no. 312043 (MN), no. 310829 (MN), no. 259038 (KG), no. 276634 (MH), no. 277816 (OL)), The Finnish Cultural Foundation (Interdisciplinary Science Workshops, MH), Sigrid Juselius Foundation (MN), and Cancer Society of Finland (MN, MH, OL, KG), Paulo Foundation (OL), Foundation for Pediatric Research (OL), Jane and Aatos Erkko Foundation (OL), Nokia Foundation (VH) and University of Eastern Finland (MH). We thank CSC – IT Center for Science and UEF bioinformatics center for providing computational resources.

References

1. Orr MS, Scherf U. Large-scale gene expression analysis in molecular target discovery. *Leukemia.* **2002**;16:473–7.

2. Ylipää A, Yli-Harja O, Zhang W, Nykter M. Characterization of aberrant pathways across human cancers. *BMC Syst. Biol.* **2013**;7:S1.

3. Kumar SK, Rajkumar V, Kyle RA, van Duin M, Sonneveld P, Mateos MV, *et al*. Multiple myeloma. *Nat Rev Dis Primers*. **2017**;3:17046.

4. Nangalia J, Green AR. Myeloproliferative neoplasms: from origins to outcomes. *Blood*. **2017**;130:2475-83.

5. June CH, O'Connor RS, Kawalekar OU, Ghassemi S, Milone MC. CAR T cell immunotherapy for human cancer. *Science*. **2018**;359:1361-65.

6. McCabe B, Liberante F, Mills KI. Repurposing medicinal compounds for blood cancer treatment. *Ann. Hematol.* **2015**;94:1267–76.

7. Corces-Zimmerman MR, Hong WJ, Weissman IL, Medeiros BC, Majeti R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc. Natl. Acad. Sci. U. S. A.* **2014**;111:2548–53.

8. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*. **2016**;166:740-54.

9. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. **2002**;30:207-10.

10. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **2003**;31:e15.

11. Eklund AC, Szallasi Z. Correction of technical bias in clinical microarray data improves concordance with known biological information. *Genome Biol.* **2008**;9:R26.

12. Tyner JW, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL, *et al*. Functional genomic landscape of acute myeloid leukaemia. *Nature*. **2018**,562:526-31.

13. Lawrence ND. Gaussian process latent variable models for visualization of high dimensional data. *Adv. Neural Inf. Process. Syst.* **2004**;16.3:329-36.

14. Roweis ST, Lawrence KS. Nonlinear dimensionality reduction by locally linear embedding. *Science.* **2000**;290:2323-6.

15. Hotelling H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**;24:417.

16. Tipping ME, Bishop CM. Probabilistic Principal Component Analysis. *J. Roy. Stat. Soc. B.* **1999**;61:611-22.

17. Sammon JW. A nonlinear mapping for data structure analysis. *IEEE T. Comput.* **1969**;5:401-9.

18. van der Maaten L, Hinton G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**;9:2579–2605.

19. van der Maaten L. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn Res.* **2014**;15:1-21.

20. Mehtonen J, Pölönen P, Häyrynen S, Lin J, Liuksiala T, Granberg K, *et al*. Data-driven characterization of molecular phenotypes across heterogenous sample collections. *bioRxiv*. **2018**. https://doi.org/10.1101/248096.

21. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature.* **2014**;513:202-9.

22. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2005**;102:15545-50.

23. Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, *et al.* WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.* **2015**;44:488-94.

24. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.* **2007**;104:1777-82.

25. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **2011**;39:685-90.

26. Yoo M, Shin J, Kim J, Ryall KA, Lee K, Lee S, *et al.* DSigDB: drug signatures database for gene set analysis. *Bioinformatics.* **2015**;31:3069-71.

27. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics.* **2013**;14:7.

28. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2014**;42:D1083–90.

29. Yang H, Qin C, Li YH, Tao L, Zhou J, Yu CY, *et al.* Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res.* **2016**;44:D1069-74.

30. Cotto KC, Wagner AH, Feng YY, Kiwala S, Coffman AC, Spies G, *et al*. DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucl Acids Res*. **2018**;46:D1068–73.

31. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, *et al*. A comprehensive map of molecular drug targets. *Nature Rev Drug Discov*. **2017**;16:19-34.

32. Bausch-Fluck D, Hofmann A, Bock T, Frei AP, Cerciello F, Jacobs A, *et al.* A mass spectrometric-derived cell surface protein atlas. *PLoS One.* **2015**;10:e0121314.

33. Frismantas V*,* Dobay MP, Rinaldi A, Tchinda J, Dunn SH, Kunz J, *et al*. Ex vivo drug response profiling detects recurrent sensitivity patterns in drug-resistant acute lymphoblastic leukemia. *Blood*. **2017**;129:e26-37.

34. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*. **2010**;33:1-22.

35. Kuhn M. Building predictive models in R using the caret package. *Journal of statistical software*. **2008**;28:1-26.

36. Yadav B, Pemovska T, Szwajda A, Kulesskiy E, Kontro M, Karjalainen R, *et al*. Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. *Sci. Rep.* **2014**;4:5193.

37. Bhanot H, Weisberg EL, Reddy MM, Nonami A, Neuberg D, Stone RM, *et al*. Acute myeloid leukemia cells require 6-phosphogluconate dehydrogenase for cell growth and NADPH-dependent metabolic reprogramming. *Oncotarget*. **2017**;8:67639-50.

38. Mizuno H, Kagoya Y, Koya J, Masamoto Y, Kurokawa M. Activated Pentose Phosphate Pathway Mediated By Fbp-1 Upregulation Supports Progression of Acute Myeloid Leukemia with High EVI-1 Expression. *Blood*. **2018**;132:757.

39. Drillenburg P, Pals ST. Cell adhesion receptors in lymphoma dissemination. *Blood*. **2000**;95:1900-10.

40. Mittermayr S, Lê GN, Clarke C, Millán Martín S, Larkin, AM, O'Gorman P, *et al*. Polyclonal immunoglobulin GN-glycosylation in the pathogenesis of plasma cell disorders. *Journal of proteome research*. **2016**;16:748-62.

41. Pang X, Li H, Guan F, Li X. Multiple Roles of Glycans in Hematological Malignancies. *Frontiers in oncology*. **2018**;8:364.

42. Diakos C, Xiao Y, Zheng S, Kager L, Dworzak M, Wiemels JL. Direct and indirect targets of the E2A-PBX1 leukemia-specific fusion protein. *PloS one*. **2014**;9:e87602.

43. Karvonen H, Perttilä R, Niininen W, Hautanen V, Barker H, Murumägi A, *et al*. Wnt5a and ROR1 activate non-canonical Wnt signaling via RhoA in TCF3-PBX1 acute lymphoblastic leukemia and highlight new treatment strategies via Bcl-2 co-targeting. *Oncogene*. **2019**;DOI:10.1038/s41388-018-0670-9.

44. Huether R, Dong L, Chen X, Wu G, Parker M, Wei L, *et al*. The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. *Nat Commun.* **2014**;5:3630.

45. Ramsay AJ, Martínez-Trillos A, Jares P, Rodríguez D, Kwarciak A, Quesada V. Next-generation sequencing reveals the secrets of the chronic lymphocytic leukemia genome. *Clin Transl Oncol*. **2013**;15:3-8.

46. Rendeiro AF, Schmidl C, Strefford JC, Walewska R, Davis Z, Farlik M, *et al.* Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat Commun*. **2016**;7:11938.

47. Robak T, Blonski JZ, Robak P. Antibody therapy alone and in combination with targeted drugs in chronic lymphocytic leukemia. *Semin Oncol*. **2016**;43:280-90.

48. Maude SL, Frey N, Shaw PA, Aplenc R, Barrett DM, Bunin NJ, *et al.* Chimeric Antigen Receptor T Cells for Sustained Remissions in Leukemia. *N. Engl. J. Med.* **2014**;371:1507–17.

49. Jacoby E, Nguyen SM, Fountaine TJ, Welp K, Gryder B, Qin H, *et al.* CD19 CAR immune pressure induces B-precursor acute lymphoblastic leukaemia lineage switch exposing inherent leukaemic plasticity. *Nat Commun*. **2016**;7:12320.

50. Laszlo GS, Estey EH, Walter RB. The past and future of CD33 as therapeutic target in acute myeloid leukemia. *Blood Rev.* **2014**;28:143–53.

51. Mirkowska P, Hofmann A, Sedek L, Slamova L, Mejstrikova E, Szczepanski T, *et al*. Leukemia surfaceome analysis reveals new disease-associated features. *Blood.* **2013**;121:e149-59.

52. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, *et al*. The support of human genetic evidence for approved drug indications. *Nat genet.* **2015**;47:856-60.

53. Amir el-AD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, *et al*. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol.* **2013**;31:545-52.

54. Shekhar K, Brodin P, Davis MM, Chakraborty AK. Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE). *Proc. Natl. Acad. Sci. U. S. A.* **2014**;111:202-7.

55. Smith EN, Ghia EM, DeBoever CM, Rassenti LZ, Jepsen K, Yoon KA, *et al*. Genetic and epigenetic profiling of CLL disease progression reveals limited somatic evolution and suggests a relationship to memory-cell development. *Blood Cancer J*. **2015**;5:e303.

56. Anderson MA, Huang D, Roberts A. Targeting BCL2 for the treatment of lymphoid malignancies. *Semin. Hematol.* **2014**;51:219–27.

57. Roberts AW, Davids MS, Pagel JM, Kahl BS, Puvvada SD, Gerecitano JF, *et al.* Targeting BCL2 with Venetoclax in Relapsed Chronic Lymphocytic Leukemia. 2015. *N. Engl. J. Med.* **2015**;374:311-22.

58. Benito JM, Godfrey L, Kojima K, Hogdal L, Wunderlich M, Geng H, *et al.* MLL-Rearranged Acute Lymphoblastic Leukemias Activate BCL-2 through H3K79 Methylation and Are Sensitive to the BCL-2-Specific Antagonist ABT-199. *Cell Rep.* **2015**;13:2715-27.

59. Andreeff M, Kelly KR, Yee K, Assouline S, Strair R, Popplewell L, *et al.* Results of the Phase I Trial of RG7112, a Small-Molecule MDM2 Antagonist in Leukemia. *Clin. Cancer Res*. **2015**;22:868-76.

60. Kaindl U, Morak M, Portsmouth C, Mecklenbräuker A, Kauer M, Zeginigg M, *et al*. Blocking ETV6/RUNX1-induced MDM2 overexpression by Nutlin-3 reactivates p53 signaling in childhood leukemia. *Leukemia*. **2014**;28:600–8.

61. Laukkanen S, Grönroos T, Pölönen P, Kuusanmäki H, Mehtonen J, Cloos J, *et al. In silico* and preclinical drug screening identifies dasatinib as a targeted therapy for T-ALL. *Blood Cancer J*. **2017**;7:e604.

62. Boyault C, Sadoul K, Pabion M, Khochbin S. HDAC6, at the crossroads between cytoskeleton and cell signaling by acetylation and ubiquitination. *Oncogene*. **2007**;26: 5468-76.

63. Yang C, Boyson CA, Di Liberto M, Huang X, Hannah J, Dorn DC, *et al.* CDK4/6 Inhibitor PD 0332991 Sensitizes Acute Myeloid Leukemia to Cytarabine-Mediated Cytotoxicity. *Cancer Res.* **2015**;75:1838–45.

64. Zhang G, Schetter A, He P, Funamizu N, Gaedcke J, Ghadimi BM, *et al*. DPEP1 inhibits tumor cell invasiveness, enhances chemosensitivity and predicts clinical outcome in pancreatic ductal adenocarcinoma. *PloS one*. **2012**;7:e31507.

65. Eisenach PA, Soeth E, Röder C, Klöppel G, Tepel J, Kalthoff H, *et al*. Dipeptidase 1 (DPEP1) is a marker for the transition from low-grade to high-grade intraepithelial neoplasia and an adverse prognostic factor in colorectal cancer. *British journal of cancer*. **2013**;109:694-703.

Tables

Table 1. DPEP1 protein expression in bone marrow biopsies based on immunohistochemistry grading.

	DPEP1 Immunohistochemistry			Total
	Negative	Positive	Strong positive	
pre-B-ALL				
BCR-ABL1	0	2	0	2
ETV6-RUNX1	7	16	10	33
Hyperdiploid	13	16	1	30
Hypodiploid	1	0	0	1
MLL rearranged	5	0	0	5
TCF3-PBX1	4	0	0	4
other	18	16	8	42
Total	48	50	19	117
Other disease/tissues				
BL	2	0	0	2
T-lymphoblastic leukaemia/lymphoma	6	1	0	7

MCL	1	0	0	1
CLL	1	0	0	1
PTCL	1	0	0	1
CHL (NSCHL)	1	0	0	1
Tonsils	1	0	0	1
Thymus	1	0	0	1
Spleen	1	0	0	1
Total	15	1	0	16

Figures and Legends

Figure 1. A molecular stratification of hematologic malignancies and normal blood cell types is captured in a t-SNE visualization. **A**. Composition of the hematologic transcriptome dataset. Of the 9,544 samples, 6,820 represent hematologic malignancies (leukemia, lymphoma or myeloma), and the rest consist of cancer cell lines, proliferative diseases (myeloid denoted pM and lymphoid denoted pL), normal blood cells (healthy donor or patient). See also **Table S2**. **B**. The transcriptome data projected in 2D using t-SNE is shown. Each dot represents one of the 9,544 samples. Cluster assignment based on density estimation is shown in color for seven distinct clusters visible on the cancer map. **C**. The separation between annotated disease types (indicated by color) is shown: the lymphoid malignancies separate into acute lymphoid leukemias (pre-B-ALL in pink and T-ALL in blue), lymphomas (top right), multiple myeloma (adjacent to B-cell lymphomas) and chronic lymphoblastic leukemia (CLL, below). The myeloid diseases (AML, CML and myeloproliferative disease) are grouped closely. Samples representing normal cell types or cell lines are in grey color. Numbers refer to data driven cluster assignment (see Table S2).

Figure 2. Comparison of molecular phenotypes based on the cancer-map. Sample attribute visualizations are exemplified that allow characterizing the molecular phenotypes. Different BCL types (in **A**) and pre-B-ALL subtypes (in **B**) are colored based on sample annotations (refer to **Table S2** for abbreviations). **C.** The five most significant pathways per disease cluster (above) or pre-B-ALL cluster (below) are shown as a heatmap (tones of red indicate significant enrichment to cluster (hypergeometric test, scaled P-value). The pre-B-ALL cluster number and color (as in B) are indicated below the heatmap. **D.** The bimodal log2 gene expression signal distribution can be used to separate samples with low or non-detectable expression (N.D., in blue) from samples expressing the gene (in red). Alternatively significance of enrichment for gene sets and pathways from different databases can be selected for visualization (e-staining) on the cancer map. **E.** The corresponding gene expression state is shown on the cancer-map for the B-lymphoid differentiation marker *MME*, where the color tones correspond to scaled log2 expression values (red: high, white low; blue: not detected). **F.** Gene set enrichment for BEZ235 targets is e-stained, with empirical P-value < 0.05 shown in red.

Figure 3. Pan-cancer analysis associates disease contexts with therapeutic strategies. **A**. The *in silico* drug target analysis across disease hierarchy groups is illustrated schematically and using proteasome and surface protein targeting drugs as examples. On the left, the heatmap columns are organized by disease, and drug targets (in rows) are sorted based on their most significant disease context association (red color tones indicate significant P-value in hypergeometric test, -log(P)). The adjacent heatmap shows the disease indications for drugs known to target the gene in question. Notice that majority of drugs target multiple genes, as illustrated by Bortezomib/Carfilzomib, and only some correspond to precision drugs as exemplified by antibody targets (*SLAMF7*, *CD19*, *MS4A1* and *CD38*). **B.** Comparison of targets of approved drugs with significant association to lymphoid leukemias are shown as in **A**. The disease indications in dimmer red tone reveal potential for re-purposing of drugs approved or in clinical trials in other disease indications (notice that LE includes ALL and CLL). **C**. Example genes highlighted in the heatmaps are e-stained on the t-SNE map as in **Fig. 2C**.

Figure 4. Evaluation of cluster and disease specificity of drug responses. **A**. A heatmap comparing how well the drug response data fits different elastic net regression models is shown (color indicates R^2 values, drugs with R^2>0.25 are shown). The values are summarized as boxplot on the right. Full model included clusters (Clust), gene expression (Gexp) and mutations (Mut), while one category is omitted in the other models. **B**. Palbociclib and Panobinostat drug response AUC values are shown as boxplots for all AML cases and for clusters correlated to differential drug response identified for Palbociclib and Panobinostat. High AUC values mean drug resistance and low drug sensitivity. **C**. Heatmap of FAB morphology markers, cluster specific genetic aberrations and drug target genes (CDK6 for Palbociclib and HDAC4,6,10,2 for Panobinostat) are shown for same clusters as in B. **D**. The gene expression data for *TOP2A* and *CDK6* are e-stained on cancermap. Comparison to normal blood cell types is shown as boxplots of the log2 gene expression signal (T: T-lymphoid, B: B-lymphoid, E: erythroid and M: myeloid). For *CDK6* clusters corresponding to beatAML clusters (as in C) are shown **E.** Drug sensitivity in an AML patient cohort based on DSS and sDSS scores (*N* = 52) are shown as boxplots for Palbociclib (CDK4/6 inhibitor) and the approved AML drug (idarubicin). High difference between DSS

and sDSS values indicate response in the bone marrow normal mononuclear cells, whereas low difference indicate selectivity in AML cells.

Figure 5. Connecting the map of patient gene expression states to drug target profiles. **A**.

Significantly enriched disease clusters are colored on the map based on pan-cancer analysis of epigenetic modifiers (purple: CLL; blue: T-ALL, green: AML cluster 32; pink: pre-B-ALL cluster 28). The expression states of the most significant drug candidates for CLL (*SFMTB1*, *CBX7*, *EZH1*, *EHMT1*, *KMT2B* and *BAZ2A*) are shown (as in **Fig. 2C**) on the right. **B.** The expression level (log10 cpm) and standard deviation (log2 s.d) of the genes shown in A is indicated on the scatter plot representing independent RNA-seq data[52] (GSE81274, N=10). **C.** Significance ranking of surface target candidates for pre-B-ALL (x-axis, -log10 *P*-value) are plotted against protein level detection rate. Top candidates (P-value < 10^{-250}) are indicated next to the plot. **D.** *DPEP1* e-staining is shown as in **A**. **E.** DPEP1 immunohistochemistry. The sample on the left was interpreted as negative. In the samples in the middle and on the right over 50 percent of the leukemic blasts are showing membranous and cytoplasmic positivity and the staining of the sample was graded as strong positive (FFPE, 40x magnification, Leica DM 3000 microscope, Leica MC190 HD microscope camera, Leica Application Suite software).

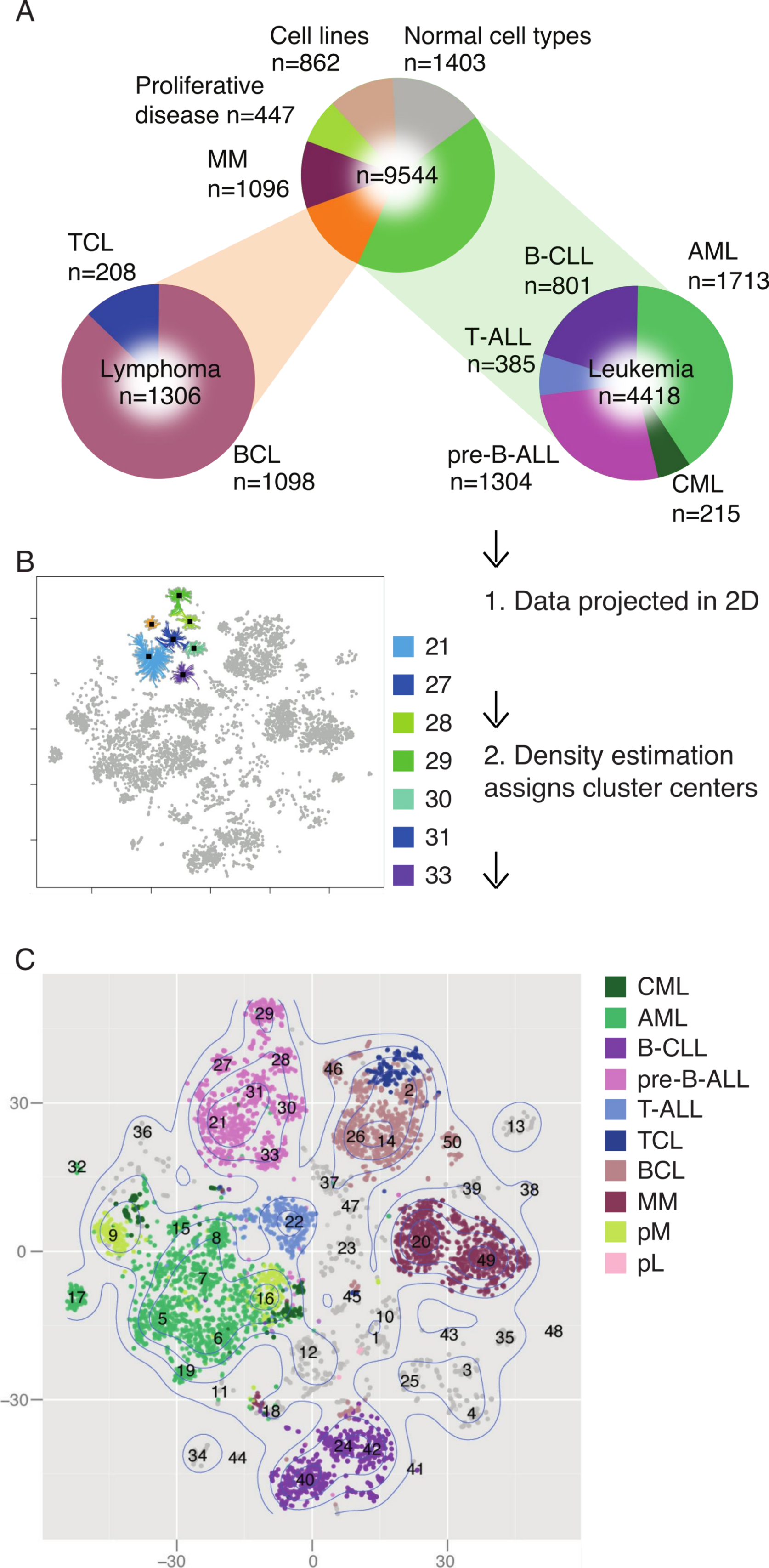
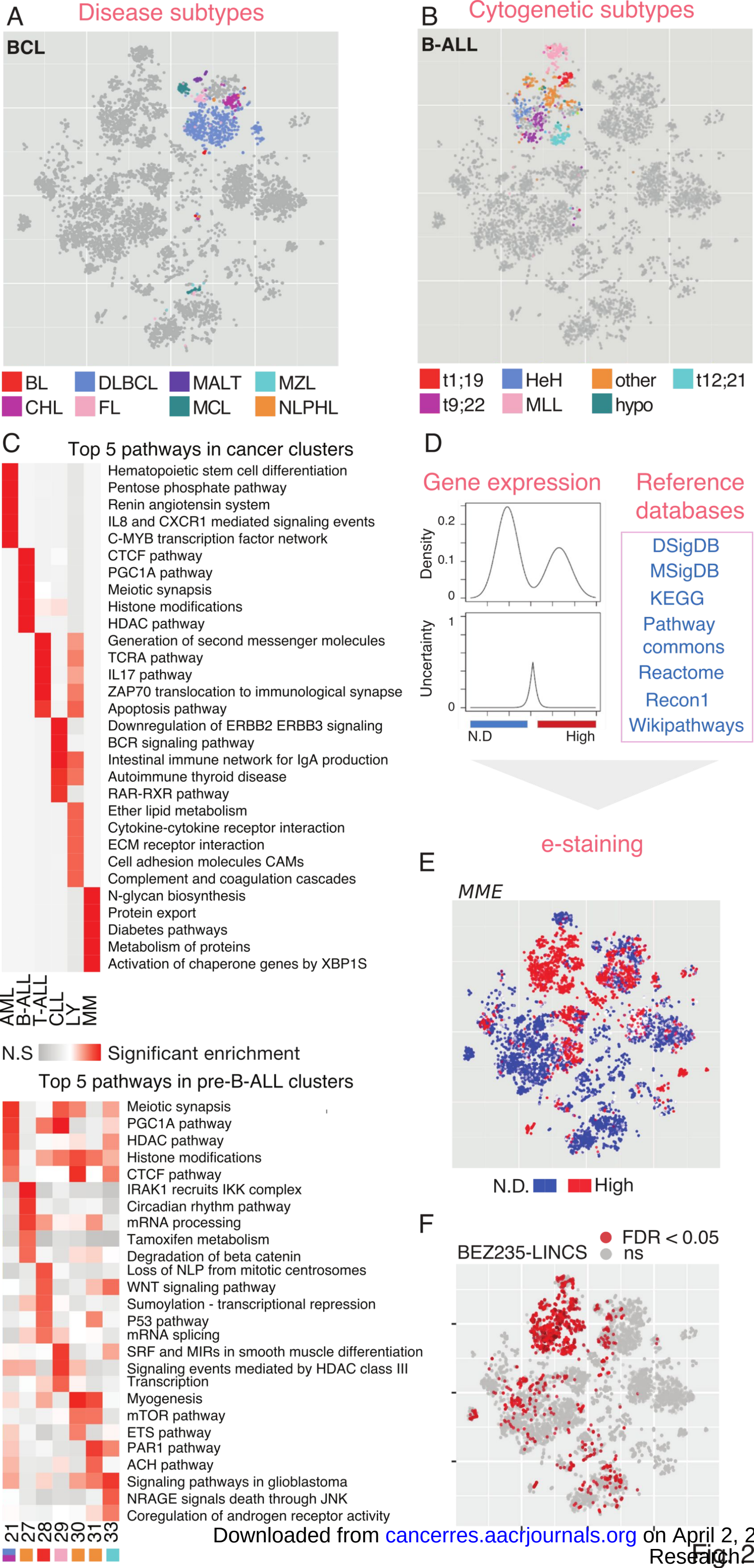

A
Cell lines n=862
Normal cell types n=1403
Proliferative disease n=447
MM n=1096
n=9544
TCL n=208
Lymphoma n=1306
BCL n=1098
B-CLL n=801
AML n=1713
T-ALL n=385
Leukemia n=4418
pre-B-ALL n=1304
CML n=215
B
21
27
28
29
30
31
33
1. Data projected in 2D
2. Density estimation assigns cluster centers
C
CML
AML
B-CLL
pre-B-ALL
T-ALL
TCL
BCL
MM
pM
pL
30
0
−30
−30
0
30
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

Fig. 1



Research.

Fig.2

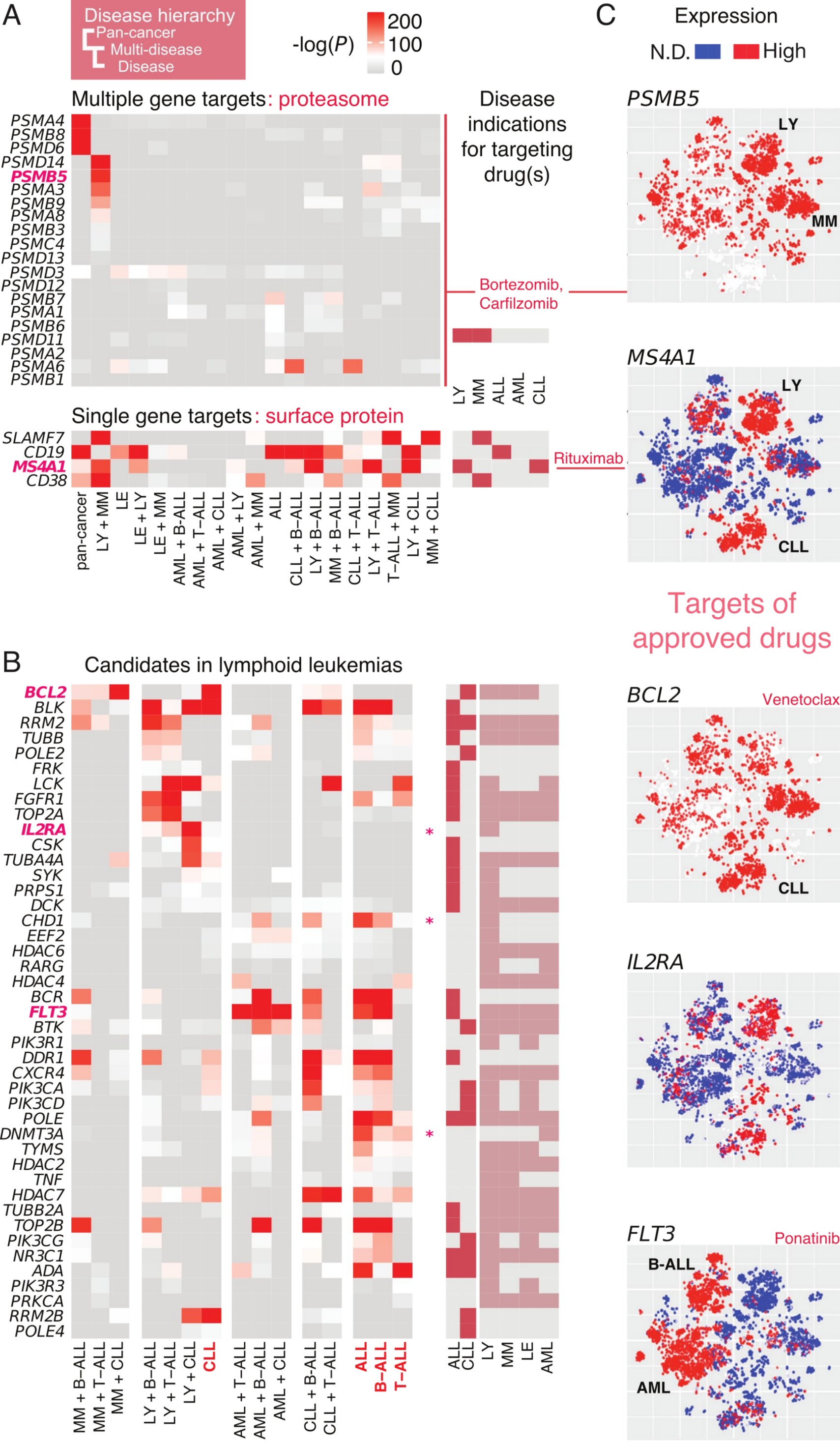

A
Disease hierarchy
Pan-cancer
Multi-disease
Disease
-log(P)
200
100
0
Multiple gene targets: proteasome
PSMA4
PSMB8
PSMD6
PSMD14
PSMB5
PSMA3
PSMB9
PSMA8
PSMB3
PSMC4
PSMD13
PSMD3
PSMD12
PSMB7
PSMA1
PSMB6
PSMD11
PSMA2
PSMA6
PSMB1
Disease indications for targeting drug(s)
Bortezomib, Carfilzomib
LY
MM
ALL
AML
CLL
Single gene targets: surface protein
SLAMF7
CD19
MS4A1
CD38
Rituximab
pan-cancer
LY + MM
LE
LE + LY
LE + MM
AML + B-ALL
AML + T-ALL
AML + CLL
AML + LY
AML + MM
ALL
CLL + B-ALL
LY + B-ALL
MM + B-ALL
CLL + T-ALL
LY + T-ALL
T-ALL + MM
LY + CLL
MM + CLL
B
Candidates in lymphoid leukemias
BCL2
BLK
RRM2
TUBB
POLE2
FRK
LCK
FGFR1
TOP2A
IL2RA
CSK
TUBA4A
SYK
PRPS1
DCK
CHD1
EEF2
HDAC6
RARG
HDAC4
BCR
FLT3
BTK
PIK3R1
DDR1
CXCR4
PIK3CA
PIK3CD
POLE
DNMT3A
TYMS
HDAC2
TNF
HDAC7
TUBB2A
TOP2B
PIK3CG
NR3C1
ADA
PIK3R3
PRKCA
RRM2B
POLE4
*
*
*
MM + B-ALL
MM + T-ALL
MM + CLL
LY + B-ALL
LY + T-ALL
LY + CLL
CLL
AML + T-ALL
AML + B-ALL
AML + CLL
CLL + B-ALL
CLL + T-ALL
ALL
B-ALL
T-ALL
ALL
CLL
LY
MM
LE
AML
C
Expression
N.D.
High
PSMB5
LY
MM
MS4A1
LY
CLL
Targets of approved drugs
BCL2
Venetoclax
CLL
IL2RA
FLT3
Ponatinib
B-ALL
AML

Fig. 3

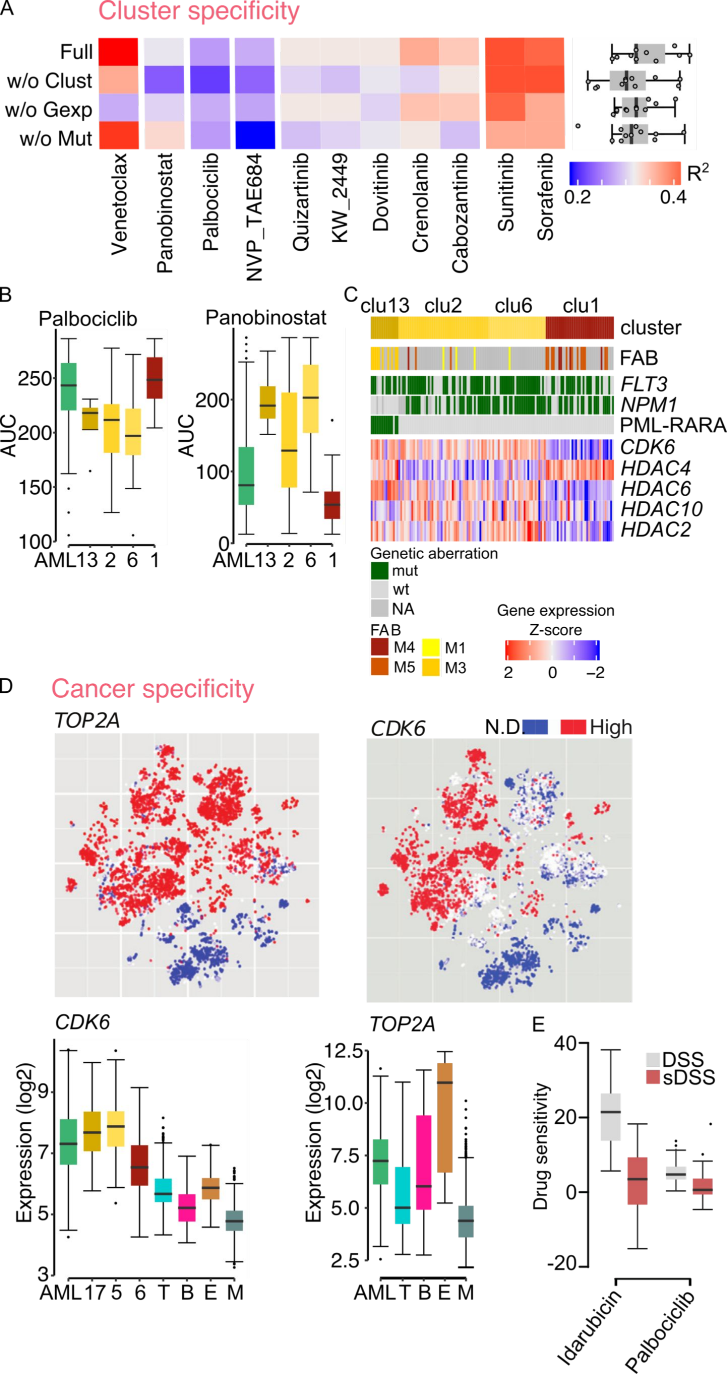

A
Cluster specificity
Full
w/o Clust
w/o Gexp
w/o Mut
Venetoclax
Panobinostat
Palbociclib
NVP_TAE684
Quizartinib
KW_2449
Dovitinib
Crenolanib
Cabozantinib
Sunitinib
Sorafenib
R2
0.2
0.4
B
Palbociclib
Panobinostat
AUC
250
200
150
100
AUC
200
100
0
AML13 2 6 1
AML13 2 6 1
C
clu13
clu2
clu6
clu1
cluster
FAB
FLT3
NPM1
PML-RARA
CDK6
HDAC4
HDAC6
HDAC10
HDAC2
Genetic aberration
mut
wt
NA
FAB
M4
M5
M1
M3
Gene expression
Z-score
2
0
−2
D
Cancer specificity
TOP2A
CDK6
N.D.
High
CDK6
Expression (log2)
3
5
7
9
AML17 5 6 T B E M
TOP2A
Expression (log2)
2.5
5.0
7.5
10.0
12.5
AML T B E M
E
Drug sensitivity
40
20
0
-20
DSS
sDSS
Idarubicin
Palbociclib

Fig. 4

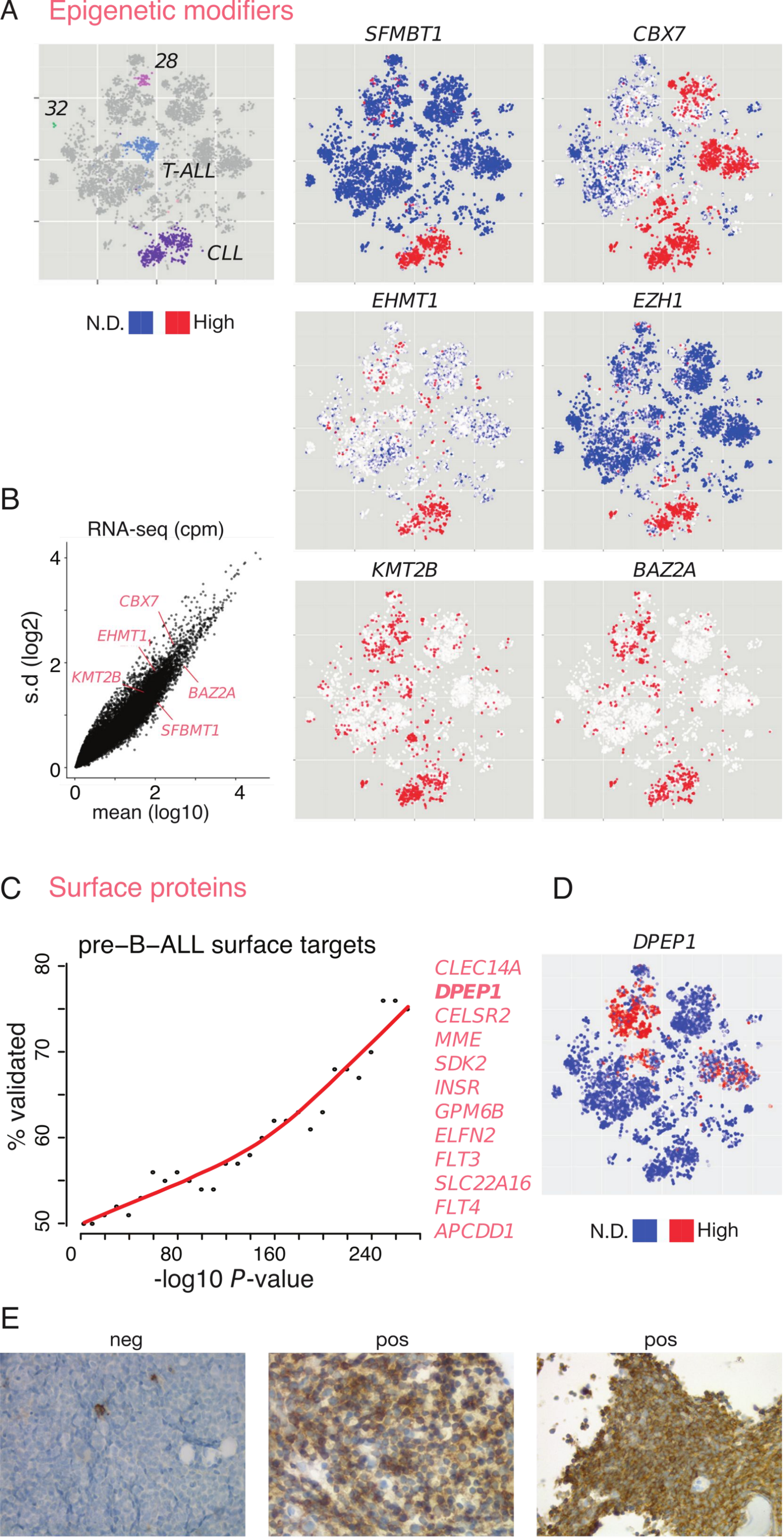

A
Epigenetic modifiers
28
32
T-ALL
CLL
N.D.
High
SFMBT1
CBX7
EHMT1
EZH1
KMT2B
BAZ2A
B
RNA-seq (cpm)
s.d (log2)
mean (log10)
0
2
4
CBX7
EHMT1
KMT2B
BAZ2A
SFBMT1
C
Surface proteins
pre−B−ALL surface targets
% validated
50
60
70
80
0
80
160
240
-log10 P-value
CLEC14A
DPEP1
CELSR2
MME
SDK2
INSR
GPM6B
ELFN2
FLT3
SLC22A16
FLT4
APCDD1
D
DPEP1
N.D.
High
E
neg
pos
pos

Fig. 5

Author Manuscript Published OnlineFirst on April 2, 2019; DOI: 10.1158/0008-5472.CAN-18-2970
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

Cancer Research
The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

AACR American Association for Cancer Research

Hemap: An interactive online resource for characterizing molecular phenotypes across hematologic malignancies

Petri Pölönen, Juha Mehtonen, Jake Lin, et al.

Cancer Res Published OnlineFirst April 2, 2019.