1    **Hemap: An interactive online resource for characterizing molecular phenotypes**

2    **across hematologic malignancies**

3    Petri Pölönen[1]+, Juha Mehtonen[1]+, Jake Lin[2,3]+, Thomas Liuksiala[2,4]+, Sergei Häyrynen[2],

4    Susanna Teppo[4], Artturi Mäkinen[4,5], Ashwini Kumar[3], Disha Malani[3], Virva Pohjolainen[5],

5    Kimmo Porkka[6], Caroline A. Heckman[3], Patrick May[7], Ville Hautamäki[8], Kirsi J. Granberg[2],

6    Olli Lohi[4], Matti Nykter[2]*, and Merja Heinäniemi[1]*

7

8    [1]Institute of Biomedicine, School of Medicine, University of Eastern Finland, Kuopio, Finland

9    [2]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

10   [3]Institute for Molecular Medicine Finland FIMM, Helsinki Institute of Life Science, University of
11   Helsinki, Helsinki, Finland

12   [4]Tampere Center for Child Health Research, Tampere University and Tampere University Hospital,
13   Tampere, Finland

14   [5]Tampere University Hospital, Fimlab Laboratories, Tampere, Finland

15   [6]Hematology Research Unit Helsinki, University of Helsinki and Helsinki University Central Hospital
16   Cancer Center, Department of Hematology, Helsinki, Finland

17   [7]Luxembourg Center for Systems Biomedicine, University of Luxembourg, Esch-Belval, Luxembourg

18   [8]School of Computing, University of Eastern Finland, Kuopio, Finland

19   + equal contribution * co-corresponding authors

20

21   Conflict of interest statement: CH has research funding from Celgene, Novartis and Orion

22   Pharma unrelated to this study.

23

24   Word count: 5849, Number of figures: 5, Number of Tables: 1. The supplementary material

25   consists of supplemental methods and Supplementary display items; 5 figures and 6 tables

26   with their legends.

27

28   Running title: Organizing molecular diversity of hematologic malignances

29

30   Corresponding authors:
31   Associate Prof. Merja Heinäniemi
32   School of Medicine, University of Eastern Finland, Kuopio, Finland
33   Tel. +358 41 4305724
34   Email: merja.heinaniemi@uef.fi
35
36   Prof. Matti Nykter                          Doc. Olli Lohi, MD
37   Faculty of Medicine and Health Technology   Faculty of Medicine and Health Technology
38   Tampere University, Tampere, Finland        Tampere University, Tampere, Finland
39   Tel. +358 40 5267884                        Tel. +358 50 318 6254
40   Email: matti.nykter@uta.fi                  Email: olli.lohi@uta.fi

41

42

**Abstract**

43

44

45    Large collections of genome-wide data can facilitate the characterization of disease states

46    and subtypes, permitting pan-cancer analysis of molecular phenotypes and evaluation of

47    disease context for new therapeutic approaches. We analyzed 9,544 transcriptomes from

48    more than 30 hematologic malignancies, normal blood cell types, and cell lines, and showed

49    that disease types could be stratified in a data-driven manner. We then identified cluster-

50    specific pathway activity, new biomarkers and in silico drug target prioritization through

51    interrogation of drug target databases. Using known vulnerabilities and available drug

52    screens, we highlighted the importance of integrating molecular phenotype with drug target

53    expression for in silico prediction of drug responsiveness. Our analysis implicated *BCL2*

54    expression level as an important indicator of venetoclax responsiveness and provided a

55    rationale for its targeting in specific leukemia subtypes and multiple myeloma, linked several

56    polycomb group proteins that could be targeted by small molecules (SFMBT1, CBX7 and

57    EZH1) with CLL, and supported *CDK6* as a disease-specific target in AML. Through

58    integration with proteomics data, we characterized target protein expression for pre-B

59    leukemia immunotherapy candidates, including DPEP1. These molecular data can be

60    explored using our publicly available interactive resource, Hemap, for expediting therapeutic

61    innovations in hematologic malignancies.

62

63                                                  **Significance**

64

65    This study describes a data resource for researching derailed cellular pathways and

66    candidate drug targets across hematological malignancies.

67

68                                                  **Introduction**

69

70    Gene expression profiles facilitate genome-wide analyses that can stratify patient subtypes

71    and identify the activity patterns of various cellular pathways under different biological

72    conditions (1-2). Even though a large number of studies have characterized hematologic

73    malignancies and normal blood cell types at genome-wide level since the introduction of

74    microarray technology, most include only tens to hundreds of samples and focus on one

75    disease. Thus, understanding the complete heterogeneity and similarity of diseases states

76    and their subtypes remains an open challenge. Moreover, many hematologic malignancies

77    are rare on the population level, necessitating collecting data across study cohorts.

78

79    Hematological malignancies include acute and chronic leukemias of myeloid and lymphoid

80    lineage, B-, T- and NK cell lymphomas, and multiple myeloma (MM), and a number of

81    premalignant conditions such as myelodysplastic syndrome (MDS), and myeloproliferative

82    neoplasms (MPN). These diseases have highly variable genetic features, unique clinical

83    courses, and varying therapeutic approaches. There is also a marked difference in

84    prevalence, genetic background and prognosis between adult and pediatric blood cancers.

85    In children, acute lymphoblastic leukemia (ALL) is the most common hematological

86    malignancy, while in adults, non-Hodgkin lymphomas (NHL), followed by MM, chronic

87    lymphocytic leukemia (CLL), and acute myeloid leukemia (AML) are the most common.

88    Treatment is moving towards increased utilization of targeted therapies in combination with

89    traditional chemotherapies. Targeted therapies include tyrosine kinase inhibitors such as

90    those developed against BCR-ABL fusion found in CML and some ALL cases, or antibody

91    therapies including CD38-targeting in MM, and engineered CAR-T cells recognizing cell

92    surface CD19 or CD22 antigens in relapsed ALL and NHL (3-5). Yet, current therapies to

93    treat hematologic malignancies rely heavily on drugs that target DNA metabolism in actively

94    proliferating cells or intracellular signaling events that are involved in proliferation (6).

95    Although these drugs have markedly improved progression-free survival, redundancy in

96    signaling and the failure to eradicate quiescent cells (7) can facilitate the rapid development

97    of therapy resistance. Testing a wider portfolio of new drug targets, or repurposing drugs

98    with established clinical indications represent promising strategies (6-7). Molecular profile

99    guided approaches hold promise to improve the efficiency of this process (8).

100

101    We present here a resource that organizes samples from cancer patients, healthy donors

102    and those at pre-malignant stages for comparative analysis based on both curated

103    annotations and data-driven clustering of molecular phenotypes. This hematologic pan-

104    cancer analysis permits the identification of clinically relevant molecular features and the

105    exploration of new drug targeting approaches across the disease hierarchy. The data and

106    analysis tools are made available as an interactive online resource, Hemap,

107    http://hemap.uta.fi/ that synthesizes the curated genome-wide data across different disease

108    subtypes.

109

110

## Materials and Methods

112

**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**).

123

**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.

137

**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

4

142  well the visualization in 2D preserved their relative placement in the original coordinate
143  space. We tested Gaussian Process Latent Variable Model (GPLVM) (13), Locally Linear
144  Embedding (LLE) (14), Principal Component Analysis (PCA) (15), Probabilistic Principal
145  Component Analysis (PPCA) (16), Sammon Mapping (SM) (17) and t-Distributed Stochastic
146  Neighbor Embedding (t-SNE) (18) (see Methods Supplement for parameters). Comparison
147  of the different methods encouraged the selection of t-SNE maps, specifically the Barnes-
148  Hut approximated version of t-SNE implementation (BH-SNE) (19). In final analysis 15%
149  genes with highest variance were used in construction of t-SNE maps (see (20) for
150  justification).

151

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

161

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

5

178

### Gene set analysis

180 The pathway and gene set enrichment analysis available in the Hemap resource was
181 generated based on gene sets retrieved from MsigDB v5.0 (22) (molecular signatures),
182 Wikipathways (06.2015) (23), Recon 1 (24) (metabolic pathways), Pathway Commons 7 (25)
183 and DSigDB v1.0 (26) (drug targets). Gene sets were limited to contain between 5 to 500
184 expressed genes (as defined above) per gene set, resulting in 19,680 gene sets that were
185 evaluated across the dataset. The gene set variation analysis (GSVA) (27), GSVA package
186 1.13.0 in R, was used to calculate gene set enrichment scores (positive for increased and
187 negative for decreased expression) for each sample (parameters mx.diff=F, tau=0.25,
188 rnaseq=F). Significance was evaluated based on empirical *P*-values calculated using 1000
189 random permutations of genes within the gene set, separately for gene set sizes 5-20, 25,
190 30, 40, 50, 75, 100, 200, 300, 400, and 500 to correct for differences in gene set sizes.
191 Hypergeometric test was used to calculate enrichment of significant scores in a specific
192 cluster.

193

### Data sources used for evaluating drug targeting approaches

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

213

**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,

7

250  sample order was randomly shuffled for one category while the original order was preserved
251  for the other categories. Therefore, if the shuffled category features were important for the
252  model fit, model overall fit should decrease as the other features are unchanged. This
253  process was repeated 100 times and median of RMSE and $R^2$ values were computed. Only
254  drugs with good fit when using all the features were kept, having $R^2$ over 0.25 and RMSE
255  less than 0.9.

256

257  **Drug sensitivity testing using patient and healthy donor samples**
258  Bone marrow (BM) aspirates or peripheral blood samples were obtained from AML patients
259  ($N$=52) and healthy donors ($N$=15) after informed written consent using protocols approved
260  by a local Institutional Review Board and in accordance with the Declaration of Helsinki.
261  Mononuclear cells (MNCs) were isolated by density gradient separation (Ficoll-Paque
262  PREMIUM; GE Healthcare, Little Chalfont Bucks, UK) and immediately used for drug testing.
263  Cells were maintained in Mononuclear Cell Medium (MCM; Promocell) or in a 25% HS-5
264  conditioned medium plus 75% RPMI 1640 medium mix (CM). Palbociclib and idarubicin
265  (from Selleck, Houston, TX) were solvated in dimethyl sulfoxide and plated in 5 different
266  concentrations in 10-fold dilutions on 384-well plates using an Echo acoustic dispenser
267  (Labcyte, Sunnyvale, CA, USA), 1-10 000 nM for Palbociclib; 0.1-1000 nM for Idarubicin. 10
268  000 cells were added per well and incubated with the drugs for 3 days at 37°C, 5% $CO_2$.
269  Viability was measured using the CellTiter-Glo reagent (Promega, Madison, WI, USA)
270  according to the manufacturer's instructions and using the PHERAstar (BMG LABTECH,
271  Ortenberg, Germany) or SpectraMax Paradigm (Molecular Devices, Sunnyvale, CA, USA)
272  plate readers. Sensitivity to the drugs was quantified using a drug sensitivity score (DSS),
273  which is a modified area under the curve based metric described previously (36). A selective
274  DSS value was calculated by subtracting the mean DSS of the healthy BM controls from the
275  DSS of individual AML samples.

276

277  **Immunohistochemistry**
278  Anti-DPEP1 antibody (Atlas antibodies, rabbit polyclonal IgG against human renal
279  dipeptidase 1, product number: HPA009426, lot number: A57960) was used with the dilution
280  1:2500 to stain formalin fixed and paraffin embedded bone marrow trephine biopsy samples
281  from pediatric pre-B-ALL patients from the Pirkanmaa ERVA area between the years 2000
282  and 2017. 126 diagnostic samples (including also one Burkitt's lymphoma and a T-
283  lymphoblastic leukaemia/lymphoma case) were stained with a Ventana Benchmark GX
284  using UltraView Universal DAB Detection Kit. Cytoplasmic and membranous staining was
285  graded negative if less than 20 percent of the leukemic blasts were stained, positive if 20

8

286 percent or less than 50 percent of the blasts were positively stained and strong positive if 50
287 percent or a greater proportion of the blasts were positive. The analysis was performed by
288 two pathologists without the knowledge of the patient data or the interpretation of the other
289 analyst. The samples and clinical data were studied with the approval of the Tampere
290 University Hospital Ethical committee (#R16054 and #R13109) and in accordance with the
291 Declaration of Helsinki.

292

293 **Interactive web resource for data analysis**
294 The interactive online resource and the accompanying user guide for the Hemap resource
295 are described in more detail in the Methods Supplement and available at http://hemap.uta.fi//

296

297 <center>**Results**</center>

298

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

309

310 To enable discovery and statistical comparison of previously known and novel molecular
311 phenotypes alongside the annotated disease classes, we utilized a data-driven approach
312 that allows discovery of sample groups and visualizes these for interpretation. First, we
313 compared dimensionality reduction methods that allow visualization of complex data in two
314 dimensional space. The data representation accuracy was quantitatively assessed using the
315 metrics of continuity, trustworthiness and k-nearest neighbor (k-NN) classifier error (see
316 **Methods, Fig. S2**). As a result, t-Distributed Stochastic Neighbor Embedding (t-SNE) (18)
317 and its approximation, Barnes-Hut-SNE (BH-SNE) (19), was selected, as it performed
318 robustly (continuity and trustworthiness, 0.9860 and 0.9943, respectively) in two dimensions

9

319    and still preserved the neighborhood structure (k-NN error 0.0668) (**Figure S2**). The t-SNE

320    map was then utilized for density-based clustering to assign each sample to a cluster (**Fig.**

321    **1B**, see **Methods** for details) and the results were compared to annotated disease classes

322    (**Fig. 1C**). We conclude that both quantitative and biological assessments confirm that our

323    approach faithfully organizes the samples in an unsupervised manner based on their

324    molecular phenotype and disease type. We denote the resulting data-driven sample

325    grouping as the Hemap "cancer-map" in the following text.

326

327    **Comparative analysis associates clinical annotations and pathway activity to the**

328    **molecular disease stratification**

329    The 2D cancer-map revealed a clinically relevant sub-structure (**Table S2**), as exemplified

330    by the different B-cell lymphomas and pre-B-ALL cytogenetic subtypes (colored in **Figs. 2A**

331    and **B**, respectively), providing biological validation for separation of relevant phenotypes on

332    the cancer-map. A detailed comparison to annotations is presented in **Table S2**. Next,

333    statistically significant associations of clusters with gene expression levels, clinical

334    annotations and pathway enrichment scores across different databases were calculated (see

335    **Methods)**. These results can be interactively tabulated and visualized using the online

336    Hemap resource. We selected five most significant pathways at disease cluster level, or

337    those matching pre-B-ALL subtype clusters (**Fig. 2C**) for visualization in a heatmap (see also

338    Table S5). In AML, the pathways for hematopoietic stem cell differentiation, pentose

339    phosphate pathway, renin-angiotensin system, IL-8/CXCR1-mediated signaling events and

340    C-MYB transcription factor networks were most significantly enriched. These reflect well the

341    known progenitor-like phenotype of AML cells. Pentose phosphate pathway, on the other

342    hand, represents a recently uncovered vulnerability (37,38) that is important for AML growth.

343    Similar disease-relevant pathways were also uncovered from T-ALL (TCR pathway), CLL

344    (BCR signaling pathway), lymphomas (cell adhesion molecules (39)) and multiple myeloma

345    (N-glycan biosynthesis (40,41)). In pre-B-ALL clusters, processes related to transcriptional

346    regulation were highly significant (including histone modification, CTCF pathway, and RNA

347    processing). WNT signaling (42,43) was found as a cluster 29-specific (t1;19) enriched

348    pathway, which matches its known relevance in these TCF3-PBX1 fusion carrying cases.

349    Samples expressing a gene or pathway of interest can be visualized as shown in **Fig. 2D**,

350    distinguishing the progenitor-like MLL subtype of pre-B-ALL based on the lack of expression

351    of the differentiation marker *MME* (also known as *CD10*) that is used in clinical diagnostics

10

352　(**Fig. 2E**). Similarly, most significant associations between disease clusters and drug

353　signatures can be examined by e-staining their significance (in red), as illustrated by

354　association of PI3K inhibitor BEZ235 gene set signature from DsigDB to pre-B-ALL (**Fig.**

355　**2F**), which validates a known association between a drug and a disease subtype. Further

356　analysis on the BEZ235 gene set and several case studies on how to generate novel

357　hypothesis are presented in the accompanying **User guide** to demonstrate different analysis

358　(refer to "Explore" and "e-staining" examples).

359

360　**Pan-cancer analysis to recognize vulnerabilities across disease contexts**

361　Parallel to molecular stratification, the diversity of patient profiles in Hemap has the potential

362　to support the development of new therapeutic strategies by leveraging the information

363　about the expression profiles across hematologic malignancies. We analyzed the specificity

364　of drug target expression states across patient groups in a hierarchical manner (**Methods**),

365　as illustrated in **Fig. 3A** (see also **Table S3** for a list of drugs and their targets and **Table S4**

366　for significant associations listed by disease hierarchy). The corresponding significance

367　ranking for targets of approved drugs is shown as heatmaps in **Fig. 3A-B**, where the

368　columns represent different disease contexts and gene targets (in rows) are sorted

369　according to their most significant association. The clinical indication for the drug(s) that

370　could be used to target each gene is indicated in the panel on the right, while e-staining

371　results for example drug targets are shown in **Fig. 3C** (see also **Fig. S3**). Proteasome

372　targeting drugs Bortezomib and Carfilzomib are in use for lymphomas and multiple myeloma.

373　Accordingly, 10/20 genes encoding the proteasome subunits are associated to this disease

374　hierarchy level, or to the pan-cancer category, with highest significance (**Fig. 3A**). In

375　comparison, for precision drugs such as the antibody drugs Elotuzumab (*SLAMF7*, P-val <

376　1e-315) or Daratumumab (*CD38*, P-val 1e-196) approved for MM, or Rituximab (*MS4A1*, P-

377　val < 1e-315 in LY+CLL) used in lymphomas and CLL (**Fig. 3A**) the specific gene targets

378　can be examined. Among all known vulnerabilities (drugs in clinical use / trial) a gene-level

379　analysis detected 84% of targets expressed and 69% were associated with highest

380　specificity score (-log10 P-value) to the respective disease context (see **Table S3**). This is

381　exemplified by the comparison of genes with significant association to lymphoid leukemias

382　(**Fig. 3B)**. *BCL2* targeted by venetoclax is shown as an example of an approved target in

383　CLL that our analysis associates with this disease context and with potential for targeting in

384　MM. The genes marked with asterisk, including *IL2RA* indicate targets of drugs approved for

11

385    other hematologic malignances. Our analysis associated these with re-purposing potential in

386    CLL and/or ALL. *FLT3* is a recently approved target with disease cluster-specific expression

387    in B-lymphoid and myeloid leukemias.

388

389    **Utility of molecular disease stratification for evaluating drug screen results**

390    Next, we examined leukemias at disease subtype level from two *ex vivo* drug screening

391    datasets (12,33). Venetoclax had lower efficacy in T-ALL vs. B-ALL and lowest efficacy was

392    in t1;19 samples in the ALL drug screen (33) which agrees with Venetoclax target *BCL2*

393    gene expression in Hemap (**Fig. 3C)**. Topotecan and dasatinib had the opposite profile, also

394    in agreement with subtype-specific expression of their targets *TOP1MT* and *LCK* (**Fig. S3**).

395    Taken together, out of 15 drugs from this ALL screen tested with our hierarchical analysis,

396    14 (93 %) had a candidate target expressed and 12 (80%) received highest target indication

397    in ALL (**Table S4)**. Using the larger beatAML dataset (12), we set out to examine in an

398    unbiased manner what matters more in predicting drug responsiveness: target expression,

399    genetic lesions traditionally used to stratify patients, or the molecular phenotype as defined

400    by clustering of transcriptome states. We implemented models using elastic nets, where a

401    model for each drug (75 in total) was fit using these three categories of features. To test their

402    importance for model fitting, sample order was randomly shuffled for one category while the

403    original order was preserved for the other categories. The results for 11 drugs that achieved

404    the best overall model fit ($R^2$>0.25) are shown in **Fig. 4A,** including Venetoclax,

405    Panobinostat (HDAC inhibitor), Palbociclib (CDK4/6 inhibitor), 7 kinase inhibitors (many

406    targeting FLT3) and an ALK inhibitor. The average $R^2$ value from 100 tests is colored in the

407    heatmap and summarized as a boxplot next to it. If the shuffled feature was important for the

408    model fit, a decrease in $R^2$ is expected (shift from darker red to dimmer or blue colors) as the

409    other features are unchanged. For venetoclax, this analysis implicated target gene

410    expression as the main predictor (**Fig. S4**). For FLT3-targeting compounds, *FLT3* mutation

411    status was implicated as the top predictive feature (**Fig. S4**). However, overall, the lack of

412    cluster features in the model resulted in lowest predictive power. The disease clusters were

413    the best predictors for Palbociclib and Panobinostat, whereas mutation status had no effect

414    on their model fit. Panobinostat and Palbociclib showed opposite drug responses in clusters

415    13, 2, 6 compared to cluster 1 (**Fig. 4B**). Hemap clusters 17, 5, and 6 corresponded to these

416    clusters (**Fig. S4**) and were similarly enriched for NPM1 and FLT3 mutations or PML-RARA

417    fusion in both data sets. Comparison of clinical phenotypes revealed that blast morphology

12

418    was different between the clusters, linking maturation level to the differential drug response

419    (**Fig. 4C and Fig. S4**).

420

421    Classical targets involved in DNA metabolism (*TOP2A* and *B*) and clinically interesting

422    targets, including *CDK6, BCL2, MDM2* and *VEGFR2* from clinical trials, ranked highly in our

423    disease hierarchy analysis, as shown in **Figs. 3** and **S3**. However, when compared to

424    normal cell types, only 7% of the targets had higher expression in disease than in normal

425    cells (**Table S3**). Palbociclib target *CDK6* is highly expressed in all acute leukemias

426    compared to normal blood cell types, while *TOP2A* has high mRNA levels also in normal

427    blood cells (**Fig. 4D**). To evaluate drug sensitivity that is specific to cancer cells, an

428    experimental *ex vivo* screening approach is exemplified in **Fig. 4D** by comparing in AML

429    patient cell responses to the CDK4/6 inhibitor Palbociclib and Idarubicin targeting *TOP2A*

430    (see **Methods**). Drug sensitivity and selective drug sensitivity scores (DSS and sDSS,

431    respectively, see **Methods**) (36) are compared in box plots (**Fig. 4E**). Overall, the AML

432    patient bone marrow *ex vivo* cultures were more responsive to Idarubicin (refer to **Fig. S4** for

433    AML cell line data). However, a negative score indicating higher effect on normal bone

434    marrow cell viability was observed for Idarubicin in a larger fraction of AML cases compared

435    to Palbociclib. This observation of non-specific response, implied by negative sDSS score, is

436    consistent with our predictions from Hemap data. Therefore, the normal samples included in

437    Hemap could provide valuable additional information for drug target selection. Comparison

438    of BCL2 and BCL2L1 (also known as BCL-XL) levels are presented as another example in

439    **Fig. S4**, relevant to Venetoclax vs Navitoclax toxicity in targeting apoptosis. The Advanced

440    Use Case in the Hemap **User guide** extends this analysis using pathway activities and drug

441    chemical screen data.

442

443    **Evaluating new therapeutic strategies in a pan-hematologic cancer context**

444    Epigenetic regulation has emerged as an important mechanism that can corrupt the gene

445    regulatory network (44), motivating novel therapeutic approaches. Utilizing the disease

446    spectrum in Hemap, we performed a pan-hematologic cancer analysis of epigenetic

447    modifiers (**Table S5**), focusing on genes encoding proteins that are validated targets of small

448    molecule drugs (available from ChEMBL (28)). We found elevated expression of this set of

449    genes significantly enriched in CLL, T-ALL and clusters 28 (pre-B-ALL) and 32 (AML) (**Fig.**

450    **5A**, hypergeometric test adjusted *P*-values 0.0003, 0.0074, 0.0127, 0.0174, respectively, see

451 also **Table S5** for additional mutation frequency information (45) for the genes shown). The

452 expression state for six most significant genes from CLL are shown on the Hemap cancer-

453 map (**Fig. 5A**) and from independent validation RNA-seq data (46) (**Fig. 5B**).

454

455 A second promising new strategy, immunotherapy, can kill cancer cells by targeting surface

456 proteins with antibodies (47) or chimeric antigen receptors (48). However, side effects due to

457 targeting normal blood cells along with development of resistance occur (49). To provide a

458 rational basis for extending the target repertoire, we used disease hierarchy analysis to rank

459 996 candidates available in the Cell Surface Protein Atlas (32) (**Table S6**) resulting in broad,

460 disease and subtype-specific candidates. The top ranked candidate genes in our analysis

461 correspond to those that are uniformly high expressed within the specified disease context.

462 The stem cell antigen CD33, actively pursued for treatment of AML (50), is among highly

463 ranked surface targets in clinical trials shown in **Fig. S3**. Next, we obtained proteomics

464 profiles from 19 B-ALL patients (51) to compare our ranked list for pre-B-ALL (refer to **Table**

465 **S6**) to protein-level expression. The trend between *in silico* drug target rank and protein

466 detection rate is plotted in **Fig. 5C.** Validation rate for top candidates was above 75%. The

467 highly ranked surface targets *CLEC14A, DPEP1, CELSR2, MME, SDK2, INSR, GPM6B,*

468 *ELFN2, FLT3, SLC22A16, FLT4* and *APCDD1* correspond to those with higher expression in

469 pre-B-ALL patients compared to normal blood cells (see also **Fig. S5**). The high gene

470 expression state of *DPEP1* (**Fig. 5D**) in pre-B-ALL was further validated at protein level

471 based on immunohistochemistry of diagnostic bone marrow biopsies. The grading from 117

472 ALL bone marrows and 16 samples representing other lymphoid malignancies or normal

473 lymphoid tissues is presented in **Table 1** and illustrated in **Fig. 5E**.

474

475 To further facilitate the utilization of the data, pre-calculated results are accessible via our

476 interactive web resource (http://hemap.uta.fi/) including the expression state for 4,277 drug

477 target gene sets and 1,094 drug response signatures, which can be further investigated in

478 the context of the 12,433 pathways and molecular signatures (see **Methods** and **User guide**

479 examples). Disease hierarchy analysis for the curated list of drug to target gene associations

480 (11,373 drugs; 1,182 genes) from the Therapeutic Target Database (TTD) (29), DGIdb (30)

481 and targets of FDA approved drugs across disease (31) is available in **Table S4**. In this

482 manner, *in silico* drug target selection based on Hemap can leverage gene and pathway

483 expression, as evaluated across cancer types and normal blood cell types.

14

484

485 **Discussion**

486

487 The integration of available genome-wide data from patients allows uncovering shared

488 disease mechanisms and new therapeutic options. Recent work has highlighted that

489 molecular and genetic data that helps stratify patients can dramatically increase the

490 likelihood of success during clinical development (8,52). However, in several cancer types,

491 including those of hematopoietic and lymphoid tissues, the majority of data have been

492 collected by separate studies concentrating on certain cancer types, which hinders the

493 identification of cancer type specific features. We present an interactive online resource,

494 Hemap (http://hemap.uta.fi/) for analysis across multi-center gene expression datasets to

495 investigate disease subgroups and compare molecular phenotypes across 9,544 samples

496 from hematologic malignancies.

497

498 In practice, the samples included to Hemap are inaccessible to most clinical researchers.

499 The Hemap resource serves to re-purpose data from public repositories for clinical

500 interpretation in an intuitive manner that does not require data analysis expertise. In future

501 versions of Hemap, we plan to include also RNA-seq studies. Presently, the resource

502 already contains the TCGA AML dataset and the User Guide includes examples using this

503 data. Alongside curated disease assignment, we present a data-driven approach that

504 organizes and integrates heterogeneous sample collections in an unbiased manner. To

505 facilitate this, we demonstrated how unsupervised clustering and dimensionality reduction

506 methods, here by the t-SNE method, can be used for organizing the molecular profiles for

507 further downstream analysis. The high level of performance of t-SNE has been shown in

508 context of various data types (18,53-54). In this manner, genes characterizing the patient

509 clusters can be identified for further delineation of their functional role. In CLL, our analysis

510 implicates high expression of several polycomb group proteins (SFMBT1, CBX7 and EZH1)

511 in CLL that could be targeted by small molecules, in line with chromatin state data (46), and

512 their mutation (45) frequencies, highlighting the importance to consider the spectrum of

513 genetic and epigenetic changes in these malignancies. Earlier studies have implicated

514 epigenetic plasticity as a key driver of CLL evolution during treatment (55). Specifically, CLL

515 cases had little to no genetic subclonal evolution, while significant recurrent DNA methylation

516 changes were enriched for regions near Polycomb targets (55). To further elucidate the

517 mechanisms, inclusion of post-treatment data and integrating methylation, chromatin marker

518 and mutation profiles represent important future directions in developing the Hemap

519 resource.

520

521 From a therapeutic perspective, approaches for the development of treatment strategies with

522 a broad disease focus and molecular subtype resolution are urgently needed. We used

523 Hemap to provide a roadmap for candidate drug therapies that allows prioritizing new

524 candidates based on disease-specificity. Our analysis recapitulated known vulnerabilities,

525 providing additional confirmation for targets in current clinical trials: Several compounds

526 targeting Bcl2 have been developed and have shown promise in treating both CLL and Non-

527 Hodgkin's lymphoma (56-57). However, navitoclax that also targets *BCL2L1* (also known as

528 BCL-XL) displays platelet toxicity. This potential for off-target effects was visible as high

529 gene expression level in the erythroid lineage, supporting the choice of the more selective

530 venetoclax. The prevalent high expression also in MM and pre-B-ALL found in our study

531 provides a rationale for the expansion of the testing of these compounds in lymphoid

532 malignancies. This suggestion is additionally supported by a recent study showing that these

533 compounds have promise in MLL-rearranged ALL (58), a pre-B-ALL subtype corresponding

534 to cluster 29 in our dataset. However, Hemap analysis predicts insensitivity in T-ALL and

535 t1;19 subtype, matching recent ALL drug screen data (33). Similarly, the elevated expression

536 of the p53-regulating MDM2 in pre-B-ALL fits with recent data on successful application of

537 antagonists in clinical trials (59), and mechanisms for its high expression ETV6-RUNX1-

538 positive leukemias (60).

539

540 Presently no drug screens have been carried out in primary patient cells across the

541 spectrum of hematologic malignancies in Hemap. The utility of Hemap for drug repurposing

542 was demonstrated in our recent study that identified dasatinib as a targeted therapy for a

543 subgroup of T-ALL patients (61). Here, we examined drug screen datasets to examine how

544 differential drug responsiveness could be linked to disease sub-clusters and drug targets

545 identified from the cancer maps. Using the beatAML dataset, we systematically compared

546 the importance of mutations, clusters and drug target gene expression in predicting drug

547 responses. Clusters were the best predictors of drug response for drugs with best overall

548 model fit. However, the importance of each predictor was largely influenced by drug type.

549 Best predictor for Venetoclax response was *BCL2* expression level. *FLT3* mutation status

16

550 and other mutations were the best predictors for kinase inhibitors. In contrast, disease
551 clusters were the best predictors for Palbociclib and Panobinostat responses to which
552 mutation status had no effect on model fit. Comparison of clusters in which Panobinostat and
553 Palbociclib showed opposite drug responses revealed that blast morphology was different,
554 linking maturation level to differential drug response. Furthermore, their drug targets were
555 differentially expressed in these clusters, pointing out the importance of integrating context
556 and drug target expression for *in-silico* drug screening. Surprisingly, the HDAC expression
557 pattern revealed cytosolic members (*HDAC6* and *HDAC10* (62)) as resistance markers,
558 while nuclear *HDAC4* and *HDAC9* correlated with sensitivity. Our analysis also supported
559 *CDK4/6* as disease specific targets that are known to act as critical activators of normal and
560 leukemic HSC (63). Here, Palbociclib compared favorably to Idarubicin regarding patient
561 blast sensitivity against normal bone marrow cells, reflected in mRNA data from Hemap. The
562 selectivity over normal cells may improve further using combination therapy (63) that allows
563 decreasing the dose. However, additional parameters such as drug target protein level, drug
564 metabolism and cell proliferation rate further contribute to sensitivity and therefore not all
565 patients matching a molecular subtype or expressing the target mRNA can be expected to
566 respond favorably.

567

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

582 cancers (64,65). In future, increased availability of protein-level data from different
583 hematologic malignancies will allow confirming additional targets.

584

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

589

590

605

606

607 **References**
608
609 1.   Orr MS, Scherf U. Large-scale gene expression analysis in molecular target
610      discovery. *Leukemia.* **2002**;16:473–7.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

642  17.  Sammon JW. A nonlinear mapping for data structure analysis. *IEEE T. Comput.*
643       **1969**;5:401-9.

644  18.  van der Maaten L, Hinton G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.*
645       **2008**;9:2579–2605.

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

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

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

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

19

656    23.    Kutmon M, Riutta A, Nunes N, Hanspers K, Willighagen EL, Bohler A, *et al.*
657           WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.*
658           **2015**;44:488-94.

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

662    25.    Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, *et al.* Pathway
663           Commons, a web resource for biological pathway data. *Nucleic Acids Res.*
664           **2011**;39:685-90.

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

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

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

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

674    30.    Cotto KC, Wagner AH, Feng YY, Kiwala S, Coffman AC, Spies G, *et al.* DGIdb 3.0: a
675           redesign and expansion of the drug–gene interaction database. *Nucl Acids Res.*
676           **2018**;46:D1068–73.

677    31.    Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, *et al.* A
678           comprehensive map of molecular drug targets. *Nature Rev Drug Discov.* **2017**;16:19-
679           34.

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

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

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

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

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

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

20

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

698   39.   Drillenburg P, Pals ST. Cell adhesion receptors in lymphoma dissemination. *Blood*.
699         **2000**;95:1900-10.

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

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

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

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

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

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

717   46.   Rendeiro AF, Schmidl C, Strefford JC, Walewska R, Davis Z, Farlik M, *et al*.
718         Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-
719         specific epigenome signatures and transcription regulatory networks. *Nat Commun*.
720         **2016**;7:11938.

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

723   48.   Maude SL, Frey N, Shaw PA, Aplenc R, Barrett DM, Bunin NJ, *et al*. Chimeric Antigen
724         Receptor T Cells for Sustained Remissions in Leukemia. *N. Engl. J. Med*.
725         **2014**;371:1507–17.

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

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

731   51.   Mirkowska P, Hofmann A, Sedek L, Slamova L, Mejstrikova E, Szczepanski T, *et al*.
732         Leukemia surfaceome analysis reveals new disease-associated features. *Blood*.
733         **2013**;121:e149-59.

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

21

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

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

742    55.   Smith EN, Ghia EM, DeBoever CM, Rassenti LZ, Jepsen K, Yoon KA, *et al.* Genetic
743          and epigenetic profiling of CLL disease progression reveals limited somatic evolution
744          and suggests a relationship to memory-cell development. *Blood Cancer J.*
745          **2015**;5:e303.

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

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

751    58.   Benito JM, Godfrey L, Kojima K, Hogdal L, Wunderlich M, Geng H, *et al.* MLL-
752          Rearranged Acute Lymphoblastic Leukemias Activate BCL-2 through H3K79
753          Methylation and Are Sensitive to the BCL-2-Specific Antagonist ABT-199. *Cell Rep.*
754          **2015**;13:2715-27.

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

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

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

764    62.   Boyault C, Sadoul K, Pabion M, Khochbin S. HDAC6, at the crossroads between
765          cytoskeleton and cell signaling by acetylation and ubiquitination. *Oncogene.* **2007**;26:
766          5468-76.

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

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

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

777

778

779 **Tables**

780

781 **Table 1**. **DPEP1 protein expression in bone marrow biopsies based on**

782 **immunohistochemistry grading**.

783

| | 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 |

23

| | | | | |
|---|---|---|---|---|
| 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 |

784

785

786

24

**Figures and Legends**

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

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

25

820

821 **Figure 3. Pan-cancer analysis associates disease contexts with therapeutic strategies**.

822 **A**. The *in silico* drug target analysis across disease hierarchy groups is illustrated

823 schematically and using proteasome and surface protein targeting drugs as examples. On

824 the left, the heatmap columns are organized by disease, and drug targets (in rows) are

825 sorted based on their most significant disease context association (red color tones indicate

826 significant P-value in hypergeometric test, -log(P)). The adjacent heatmap shows the

827 disease indications for drugs known to target the gene in question. Notice that majority of

828 drugs target multiple genes, as illustrated by Bortezomib/Carfilzomib, and only some

829 correspond to precision drugs as exemplified by antibody targets (*SLAMF7*, *CD19*, *MS4A1*

830 and *CD38*). **B.** Comparison of targets of approved drugs with significant association to

831 lymphoid leukemias are shown as in **A**. The disease indications in dimmer red tone reveal

832 potential for re-purposing of drugs approved or in clinical trials in other disease indications

833 (notice that LE includes ALL and CLL). **C**. Example genes highlighted in the heatmaps are e-

834 stained on the t-SNE map as in **Fig. 2C**.

835

836

837 **Figure 4. Evaluation of cluster and disease specificity of drug responses**. **A**. A

838 heatmap comparing how well the drug response data fits different elastic net regression

839 models is shown (color indicates $R^2$ values, drugs with $R^2>0.25$ are shown). The values are

840 summarized as boxplot on the right. Full model included clusters (Clust), gene expression

841 (Gexp) and mutations (Mut), while one category is omitted in the other models. **B.**

842 Palbociclib and Panobinostat drug response AUC values are shown as boxplots for all AML

843 cases and for clusters correlated to differential drug response identified for Palbociclib and

844 Panobinostat. High AUC values mean drug resistance and low drug sensitivity. **C**. Heatmap

845 of FAB morphology markers, cluster specific genetic aberrations and drug target genes

846 (CDK6 for Palbociclib and HDAC4,6,10,2 for Panobinostat) are shown for same clusters as

847 in B. **D**. The gene expression data for *TOP2A* and *CDK6* are e-stained on cancermap.

848 Comparison to normal blood cell types is shown as boxplots of the log2 gene expression

849 signal (T: T-lymphoid, B: B-lymphoid, E: erythroid and M: myeloid). For *CDK6* clusters

850 corresponding to beatAML clusters (as in C) are shown **E.** Drug sensitivity in an AML patient

851 cohort based on DSS and sDSS scores (*N* = 52) are shown as boxplots for Palbociclib

852 (CDK4/6 inhibitor) and the approved AML drug (idarubicin). High difference between DSS

26

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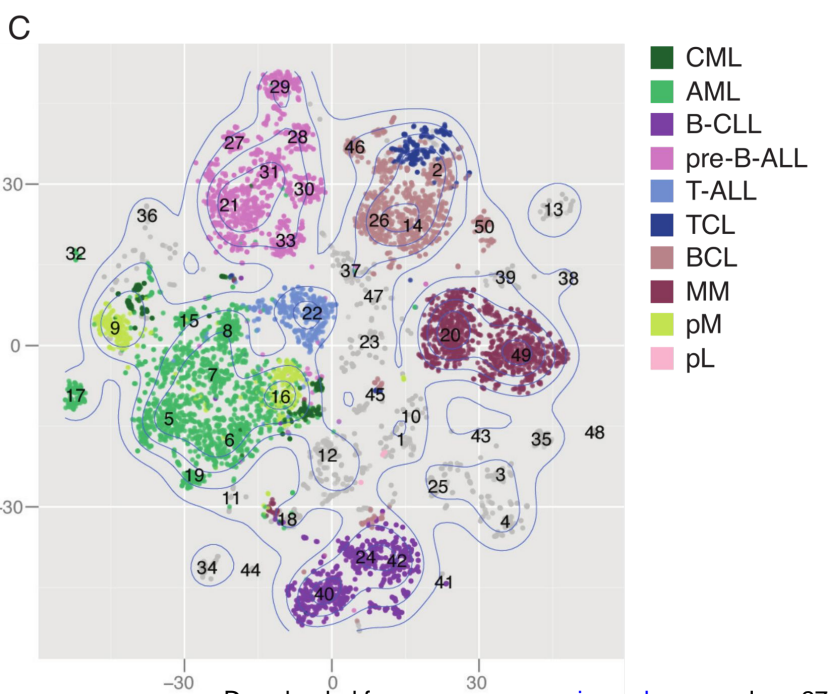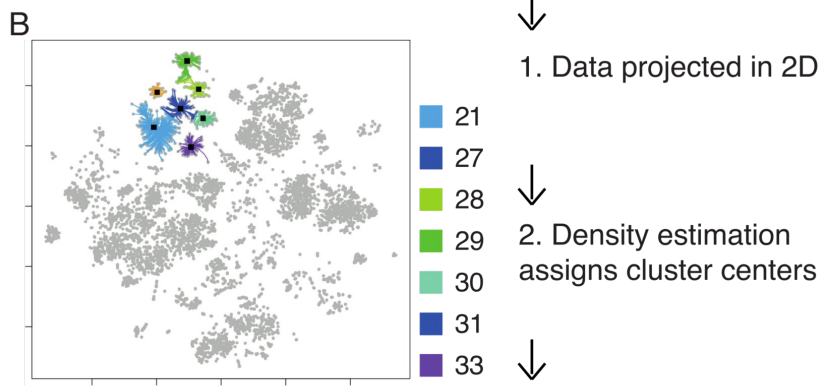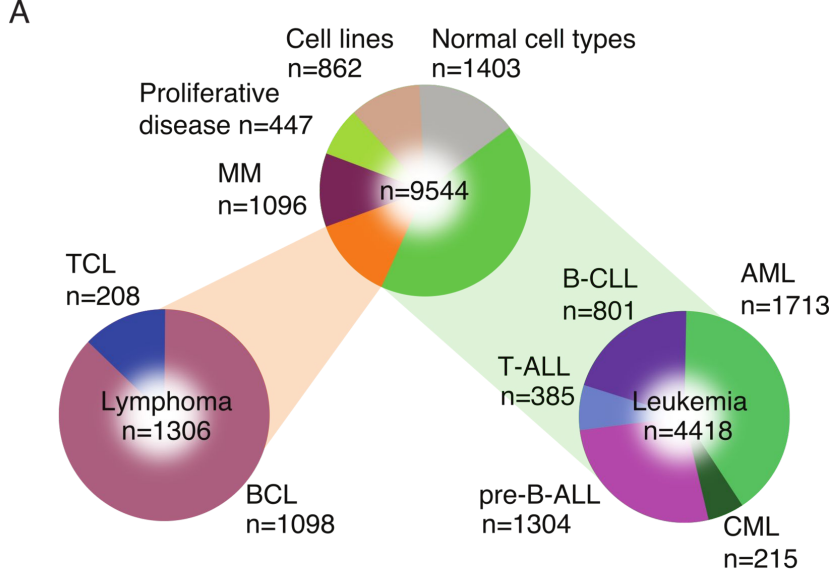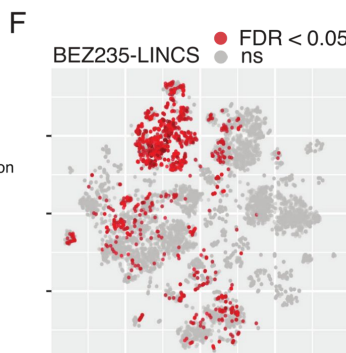
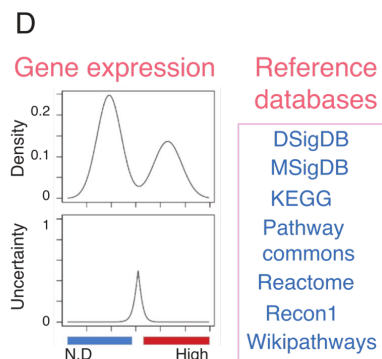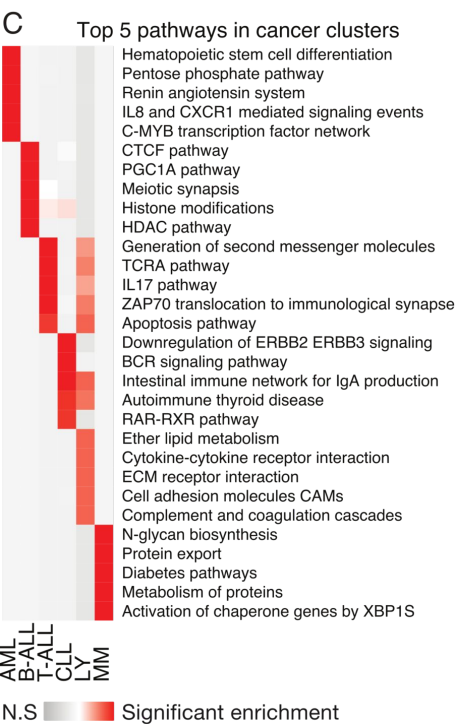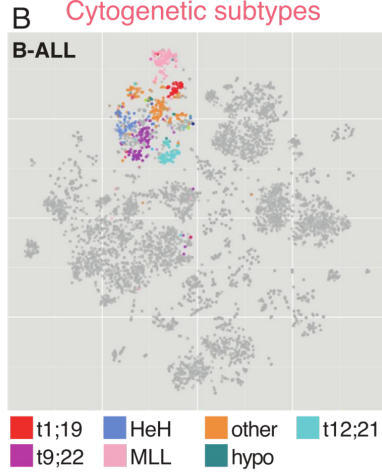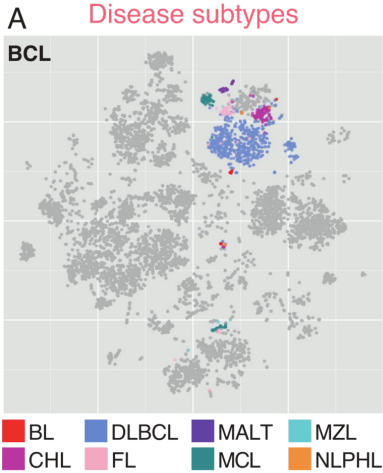854 low difference indicate selectivity in AML cells.

855

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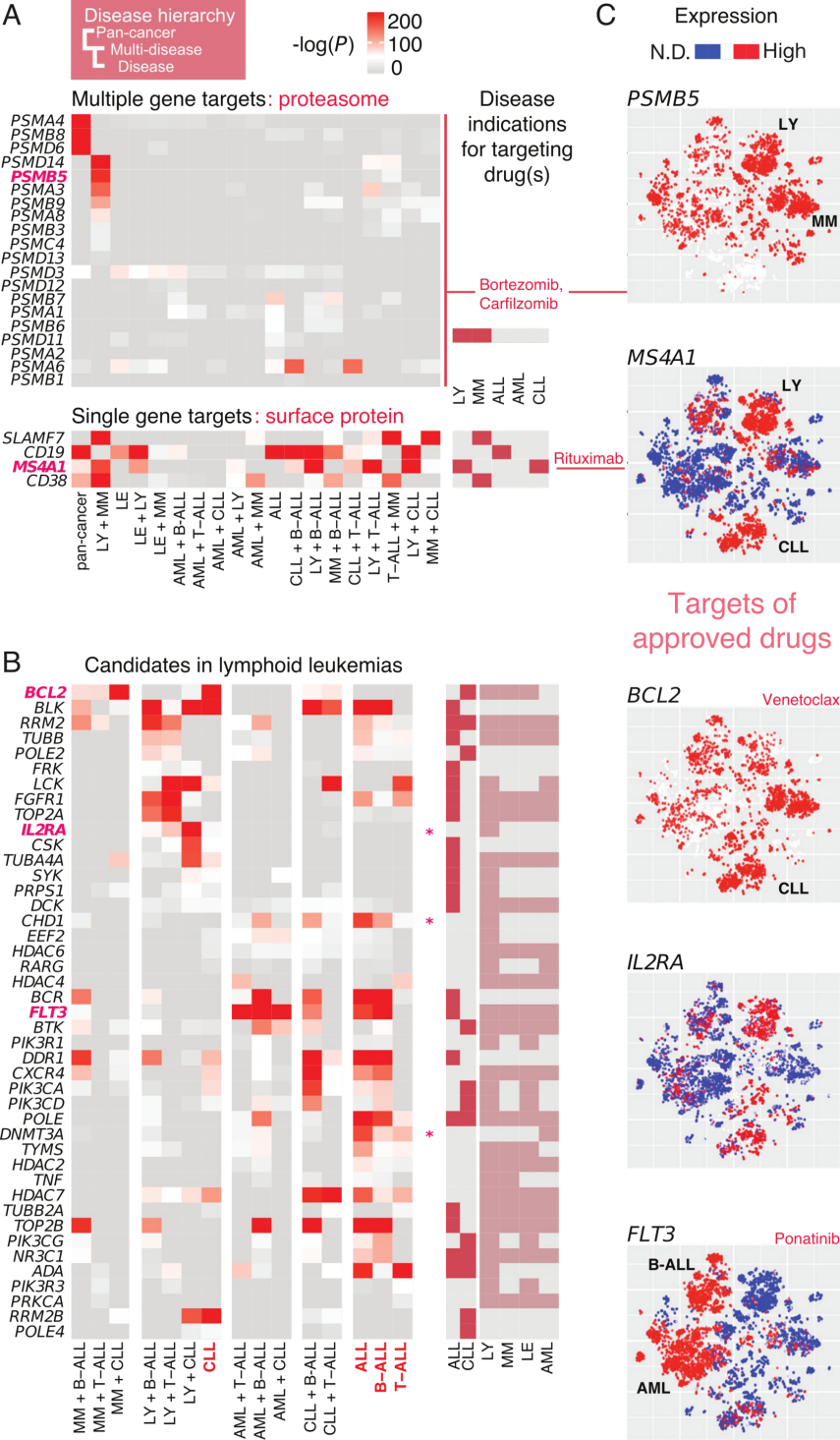
857 **profiles**. **A**.

858 Significantly enriched disease clusters are colored on the map based on pan-cancer analysis

859 of epigenetic modifiers (purple: CLL; blue: T-ALL, green: AML cluster 32; pink: pre-B-ALL

860 cluster 28). The expression states of the most significant drug candidates for CLL (*SFMTB1*,

861 *CBX7*, *EZH1*, *EHMT1*, *KMT2B* and *BAZ2A*) are shown (as in **Fig. 2C**) on the right. **B.** The

862 expression level (log10 cpm) and standard deviation (log2 s.d) of the genes shown in A is

863 indicated on the scatter plot representing independent RNA-seq data[52] (GSE81274, N=10).

864 **C.** Significance ranking of surface target candidates for pre-B-ALL (x-axis, -log10 *P*-value)

865 are plotted against protein level detection rate. Top candidates (P-value < $10^{-250}$) are

866 indicated next to the plot. **D.** *DPEP1* e-staining is shown as in **A**. **E.** DPEP1

867 immunohistochemistry. The sample on the left was interpreted as negative. In the samples in

868 the middle and on the right over 50 percent of the leukemic blasts are showing membranous

869 and cytoplasmic positivity and the staining of the sample was graded as strong positive

870 (FFPE, 40x magnification, Leica DM 3000 microscope, Leica MC190 HD microscope

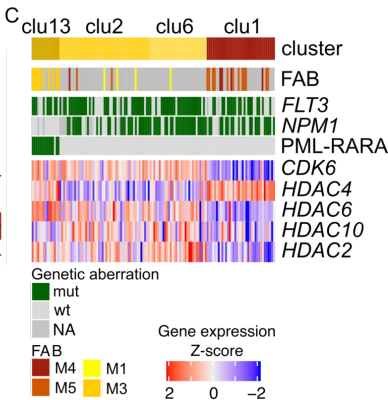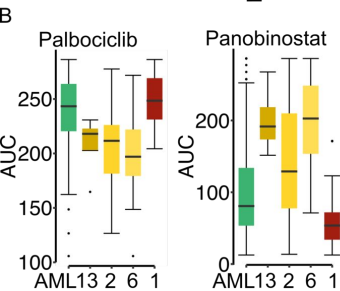871 camera, Leica Application Suite software).

872

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

T-ALL
n=385

AML
n=1713

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

Fig. 1

**A** Disease subtypes

BCL

**B** Cytogenetic subtypes

B-ALL

BL    DLBCL    MALT    MZL
CHL    FL    MCL    NLPHL

t1;19    HeH    other    t12;21
t9;22    MLL    hypo

**C** Top 5 pathways in cancer clusters

Hematopoietic stem cell differentiation
Pentose phosphate pathway
Renin angiotensin system
IL8 and CXCR1 mediated signaling events
C-MYB transcription factor network
CTCF pathway
PGC1A pathway
Meiotic synapsis
Histone modifications
HDAC pathway
Generation of second messenger molecules
TCRA pathway
IL17 pathway
ZAP70 translocation to immunological synapse
Apoptosis pathway
Downregulation of ERBB2 ERBB3 signaling
BCR signaling pathway
Intestinal immune network for IgA production
Autoimmune thyroid disease
RAR-RXR pathway
Ether lipid metabolism
Cytokine-cytokine receptor interaction
ECM receptor interaction
Cell adhesion molecules CAMs
Complement and coagulation cascades
N-glycan biosynthesis
Protein export
Diabetes pathways
Metabolism of proteins
Activation of chaperone genes by XBP1S

AML  B-ALL  T-ALL  CLL  LY  MM

N.S    Significant enrichment

Top 5 pathways in pre-B-ALL clusters

Meiotic synapsis
PGC1A pathway
HDAC pathway
Histone modifications
CTCF pathway
IRAK1 recruits IKK complex
Circadian rhythm pathway
mRNA processing
Tamoxifen metabolism
Degradation of beta catenin
Loss of NLP from mitotic centrosomes
WNT signaling pathway
Sumoylation - transcriptional repression
P53 pathway
mRNA splicing
SRF and MIRs in smooth muscle differentiation
Signaling events mediated by HDAC class III
Transcription
Myogenesis
mTOR pathway
ETS pathway
PAR1 pathway
ACH pathway
Signaling pathways in glioblastoma
NRAGE signals death through JNK
Coregulation of androgen receptor activity

21  27  28  29  30  31  33

**D** Gene expression    Reference databases

Density

Uncertainty

N.D    High

DSigDB
MSigDB
KEGG
Pathway
commons
Reactome
Recon1
Wikipathways

e-staining

**E** *MME*

N.D.    High

**F** BEZ235-LINCS    ● FDR < 0.05    ● ns

Fig.2
Research

**A** Multiple gene targets: proteasome / Disease indications for targeting drug(s)

Single gene targets: surface protein

**B** Candidates in lymphoid leukemias

**C** Expression — N.D. / High

Targets of approved drugs

Fig.3
Research.

**A** Cluster specificity

**B** Palbociclib, Panobinostat

**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

**D** Cancer specificity

TOP2A, CDK6 N.D High

**E** Drug sensitivity — DSS, sDSS — Idarubicin, Palbociclib

Fig. 4

**A** Epigenetic modifiers

*SFMBT1* *CBX7*

N.D. Green High

*EHMT1* *EZH1*

**B** RNA-seq (cpm)

*CBX7*
*EHMT1*
*KMT2B*
*BAZ2A*
*SFBMT1*

s.d (log2)
mean (log10)

*KMT2B* *BAZ2A*

**C** Surface proteins

pre−B−ALL surface targets

% validated
-log10 *P*-value

*CLEC14A*
**DPEP1**
*CELSR2*
*MME*
*SDK2*
*INSR*
*GPM6B*
*ELFN2*
*FLT3*
*SLC22A16*
*FLT4*
*APCDD1*

**D** *DPEP1*

N.D. High

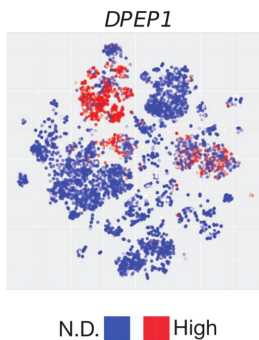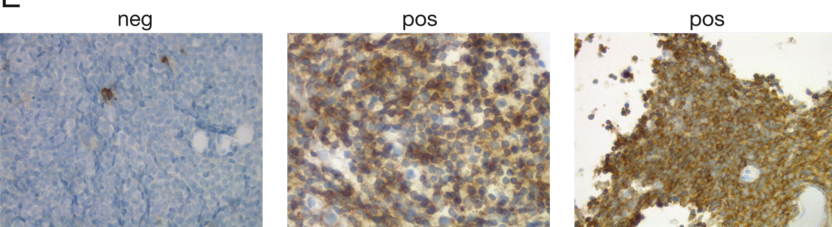**E** neg pos pos

Fig. 5

**AACR** American Association for Cancer Research

# Cancer Research

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

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

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

| | |
|---|---|
| **Updated version** | Access the most recent version of this article at:<br>doi:10.1158/0008-5472.CAN-18-2970 |
| **Supplementary Material** | Access the most recent supplemental material at:<br>http://cancerres.aacrjournals.org/content/suppl/2019/04/02/0008-5472.CAN-18-2970.DC1 |
| **Author Manuscript** | Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited. |

| | |
|---|---|
| **E-mail alerts** | Sign up to receive free email-alerts related to this article or journal. |
| **Reprints and Subscriptions** | To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org. |
| **Permissions** | To request permission to re-use all or part of this article, use this link<br>http://cancerres.aacrjournals.org/content/early/2019/04/02/0008-5472.CAN-18-2970.<br>Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site. |