Single-cell developmental classification of B cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse

Zinaida Good[1–4,11], Jolanda Sarno[5,6,11], Astraea Jager[1,2,5], Nikolay Samusik[1,2], Nima Aghaeepour[1,2], Erin F Simonds[1,2,10], Leah White[5], Norman J Lacayo[5], Wendy J Fantl[7], Grazia Fazio[6], Giuseppe Gaipa[6], Andrea Biondi[6], Robert Tibshirani[8,9], Sean C Bendall[3], Garry P Nolan[1,2,12] & Kara L Davis[1,2,5,12]

Insight into the cancer cell populations that are responsible for relapsed disease is needed to improve outcomes. Here we report a single-cell-based study of B cell precursor acute lymphoblastic leukemia at diagnosis that reveals hidden developmentally dependent cell signaling states that are uniquely associated with relapse. By using mass cytometry we simultaneously quantified 35 proteins involved in B cell development in 60 primary diagnostic samples. Each leukemia cell was then matched to its nearest healthy B cell population by a developmental classifier that operated at the single-cell level. Machine learning identified six features of expanded leukemic populations that were sufficient to predict patient relapse at diagnosis. These features implicated the pro-BII subpopulation of B cells with activated mTOR signaling, and the pre-BI subpopulation of B cells with activated and unresponsive pre-B cell receptor signaling, to be associated with relapse. This model, termed 'developmentally dependent predictor of relapse' (DDPR), significantly improves currently established risk stratification methods. DDPR features exist at diagnosis and persist at relapse. By leveraging a data-driven approach, we demonstrate the predictive value of single-cell 'omics' for patient stratification in a translational setting and provide a framework for its application to human cancer.

Despite high rates of initial responses to frontline treatment, cancer mortality largely results from relapse or metastasis. Although there is debate as to whether resistant cancer cells are present at the time of initial diagnosis or whether they emerge under the pressure of therapy, many studies have suggested that it is the former[1–4]. Such cells can be rare and are not accurately represented in animal models or patient-derived xenografts[5,6]. Hence, the identification and study of the cellular species underlying cancer persistence will require both high-throughput single-cell analyses of primary human tissues and new analytical tools to align these rare populations with clinical outcomes.

B cell precursor acute lymphoblastic leukemia (BCP-ALL) is a common childhood malignancy. Despite dramatic improvements in survival using current treatment regimens, relapse is the most frequent cause of cancer-related death among children with BCP-ALL[7]. BCP-ALL is characterized by the clonal proliferation of blast cells that bear the hallmarks of immature B cells in the bone marrow (BM) and/or peripheral blood. Known molecular alterations stall the development of B lymphocytes (B lymphopoiesis) in individuals with BCP-ALL[8–12].

Healthy B lymphopoiesis occurs through sequential developmental stages that are marked by losses and appearances of surface proteins, intracellular mediators of DNA rearrangement, and activation of signaling pathways that regulate decisions of cell fate[13,14]. We previously applied single-cell mass cytometry (also known as 'cytometry by time-of-flight' or CyTOF) analysis to align developing B cells into a unified trajectory, which enabled us to better define human pre-pro-B, pro-B and pre-B cells and their regulatory signaling during early developmental checkpoints[14].

Currently, for children with BCP-ALL, risk prediction strategies integrate clinical, genetic and treatment-response features gathered during the first months of treatment[15]. As in most risk-prediction scenarios, prediction is imperfect. We reasoned that performing deep phenotypic single-cell studies of diagnostic leukemic samples could identify cell populations that are predictive of relapse, and that we could discover novel aspects of resistance to treatment in this disease.

To build on our study of normal early B lymphopoiesis, we performed a mass cytometry analysis of primary diagnostic samples from patients with BCP-ALL. By aligning individual BCP-ALL cells with
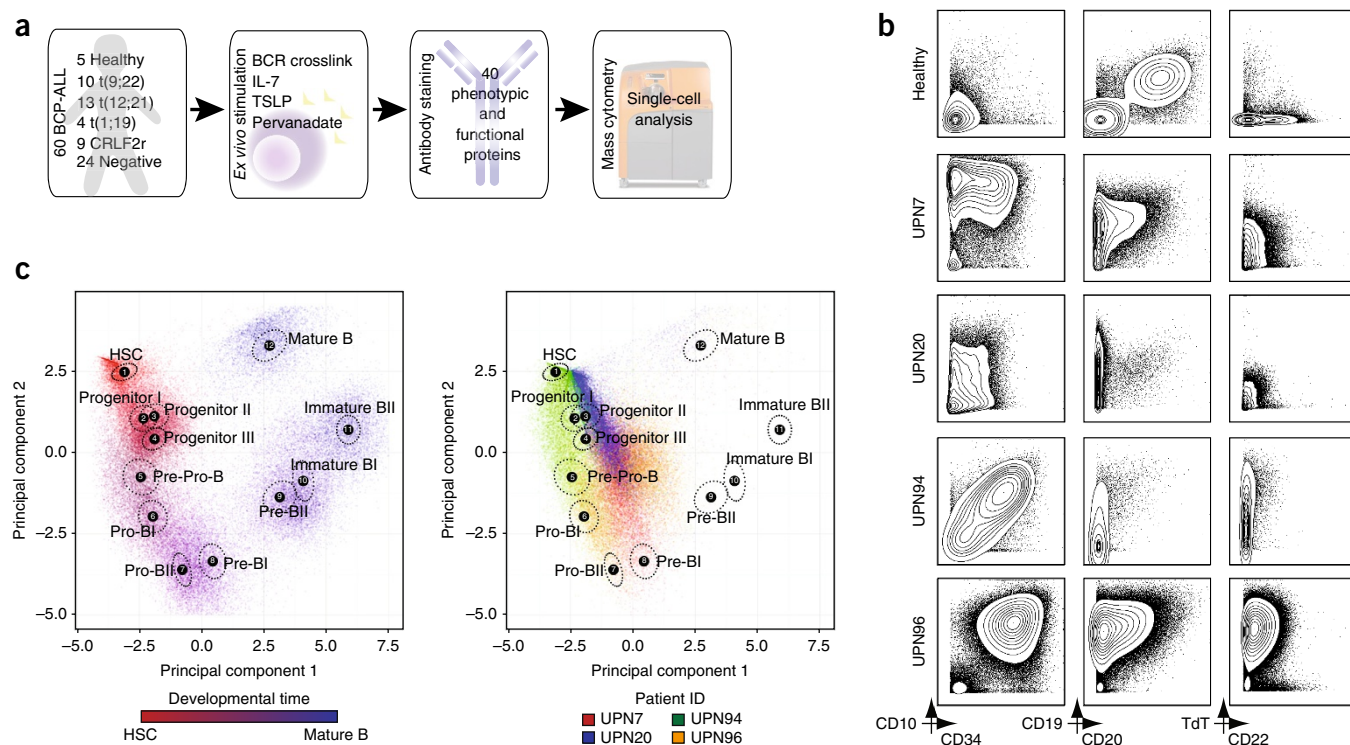
Figure 1 Mass cytometry analysis of samples from patients with BCP-ALL reveals phenotypic heterogeneity of leukemic cells. (**a**) Summary of sample processing steps for mass cytometry analysis of primary samples from patients with BCP-ALL (see **Supplementary Tables 1–3** for patient information, antibody panel, and perturbation conditions, respectively). BM aspirates from 60 patients with BCP-ALL and five healthy donors (control) were included. Prognostic cytogenetic translocations identified at diagnosis, as well as relevant *ex vivo* perturbations used to uncover cell state, are indicated. 'Negative' patients were negative for any of the prognostic cytogenetic translocations analyzed. (**b**) Mass cytometry analysis of commonly used diagnostic antigens expressed by lineage-negative BM cells (see **Supplementary Fig. 1a** for gating) from four representative patients (UPN7, UPN20, UPN94 and UPN96) with BCP-ALL and one healthy donor. TdT, terminal deoxynucleotidyl transferase. (**c**) Left, PCA of 12 manually gated B cell subpopulations in samples from healthy donors (1,000 cells sampled from each of *n* = 5 donors). The first two principal components were constructed using 11 markers that define B cell developmental populations (see **Supplementary Fig. 1b–d** for gating, marker weights, and variance captured by each principal component). The developmental time color scale was defined by setting hematopoietic stem cells as red and mature B cells as blue. Intermediate populations were placed on this red-to-blue color gradient at equal intervals. For each stage, a black dot indicates the population centroid, and the surrounding circle indicates standard error based on five healthy donors. Right, PCA of samples from four patients with BCP-ALL shown projected onto healthy B cell progression. Each sample uniquely occupies the PCA space while overlapping with multiple healthy populations and other patient samples.

developmental states along the normal B cell trajectory, we found that there was an expansion across the pre-pro-B cell to pre-BI cell transition. By applying machine learning to proteomic features extracted from these expanded cell populations, we constructed a predictive model of relapse that was validated in an independent patient cohort. This model revealed six cellular features that implicated a developmental phenotype and behavioral identity of two cell populations in portending relapse. Analysis of pairs of samples from individual patients that were taken at the time of diagnosis and at relapse (hereafter referred to as matched diagnosis–relapse pairs) confirmed the persistence of these predictive features at relapse. Thus, analysis of BCP-ALL samples that distinguishes developmental states at high resolution reveals that a unique and reproducible cellular behavior across patients is a main driver of relapse.

RESULTS
Deep phenotyping reveals developmental heterogeneity in leukemia cells from patients with BCP-ALL
To understand the extent to which childhood BCP-ALL mimics the differentiation of its tissue of origin, we profiled diagnostic primary

BM aspirates from 60 individuals with diverse clinical genetics by single-cell mass cytometry and compared the profiles to those of BM aspirates from five healthy donors (hereafter referred to as healthy BM (**Fig. 1a** and **Supplementary Tables 1–3**). Examination of the expression of proteins that are routinely used in flow-cytometry-based diagnosis of leukemic blasts revealed the expected patterns of expression, with overexpression of the early lymphoid antigen CD10 and the adhesion molecule CD34 as compared to healthy BM (**Fig. 1b**). To visualize similarities between BCP-ALL cells and normally developing B cells, we compared BCP-ALL cells to their healthy BM counterparts using principal component analysis (PCA) (**Fig. 1c** and **Supplementary Fig. 1**). Developing B cells from healthy donors occupied a remarkably clear path in this representation space (**Fig. 1c**, left). Once projected into the same space, BCP-ALL cells from individual patients fell into areas with similarity to multiple healthy populations, with a heavy skewing toward early stages of B lymphopoiesis (**Fig. 1c**, right), as expected[8]. We thus reasoned that aligning individual leukemic cells to their closest developmental state would enable us to view each BCP-ALL sample as a set of aberrant developing B cell populations, which could potentially uncover novel aspects of BCP-ALL biology.
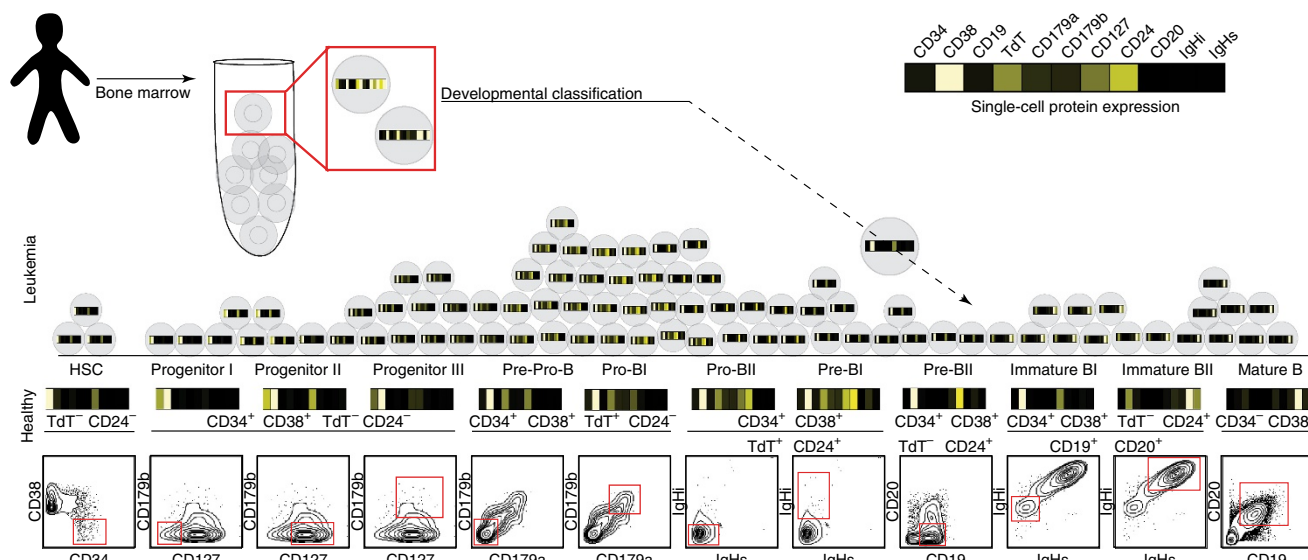
Figure 2 Single-cell developmental classifier for BCP-ALL. BM aspirates from five healthy donors were manually gated into 12 consecutive developmental stages of B lymphopoiesis (final gate is shown as a red box on a contour plot in the profiles at the bottom; the text above indicates prior gate(s) on lineage-negative cells; see **Supplementary Fig. 1a,b** for complete gating strategy). The mean arsinh-transformed expression of 11 proteins with relevance to normal B lymphopoiesis, shown in the heat bar, was determined for each healthy cell population (shown above the contour plots, where black indicates low expression, and white indicates high expression). Single cells from each BCP-ALL sample were then assigned to the most similar population of healthy cells based on the shortest Mahalanobis distance calculated from expression of the same 11 proteins. Cells with a distance greater than the classification threshold to all developmental populations remained unclassified (<1% for each patient). IgHi, intracellular immunoglobulin heavy chain; IgHs, surface immunoglobulin heavy chain.

Organization of BCP-ALL cells with a single-cell developmental classifier

To approximate the developmental state of each leukemic cell, we constructed a single-cell developmental classifier. Building on our recent study[14], we manually partitioned healthy BM into 12 developing B cell populations and three populations of mixed progenitors or mature non-B cells (**Supplementary Fig. 1b**). These 15 populations were defined by the expression of 11 proteins involved in B cell development, which provided a phenotypic maturity 'barcode' (**Fig. 2**). Mahalanobis distance was used to assign each cell to its closest developmental population (**Supplementary Fig. 2a**). We used manually gated healthy populations to test the reliability of this approach. In a ten-fold cross-validation approach, in which we iteratively defined the classifier using nine-tenths of the data and tested on the remaining one-tenth; the classifier had a good predictive performance for each population and a 92% overall accuracy (**Supplementary Fig. 2b–d**). If a cell was misclassified, it was most likely to fall into a neighboring population, which suggested that the biological effect of misclassification was likely to be small (**Supplementary Fig. 2c**). We found that the Mahalanobis distance was superior to other distance metrics (such as the cosine[14,16], Euclidian[17] and Manhattan[18] metrics) in assigning cells to the correct developmental population (**Supplementary Fig. 2d**). Moreover, additional phenotypic markers did not improve classification relative to that using the original 11 proteins (**Supplementary Fig. 2e**).

We used this single-cell classifier to assign each leukemic cell into the most phenotypically similar developmental stage. We emphasize that in this case we selected the closest developmental stage and do not imply that a leukemic cell is equivalent to a normal developing B cell (for example, a leukemic cell classified as a pre-BI cell is not suggested to be a pre-BI cell but rather is considered to be a 'pre-BI-like' cell).

Leveraging prior knowledge of normal cellular differentiation to delineate cancer cells into biologically meaningful populations enabled direct comparisons of these populations in healthy donors and patients.

BCP-ALL expands across the pre-pro-B to pre-BI developmental transition

Once classified, the frequency of cells in each developmental compartment was examined across all of the samples. As compared to healthy controls, we found a significant expansion across the pre-pro-B to pre-BI transitional populations in samples from individuals with leukemia (**Fig. 3a**). Despite the clonal nature of leukemia, phenotypic heterogeneity was observed within individual ALL samples, such that in 100% of cases there was an expansion of cells in more than one developmental population: 5.3 ± 1.9 (mean ± s.d.) of expanded populations/patient (**Supplementary Fig. 3a**).

Because prognosis and risk stratification in BCP-ALL are based partially on recurrent chromosomal rearrangements, we determined whether there was any association between these translocation events and the classification of leukemia cells. All of the known genetic alterations associated with BCP-ALL showed little correlation to overall developmental classification (i.e., the assignment of leukemic cells to their most similar healthy counterpart), although in some cases expansion of particular populations did reach statistical significance (**Fig. 3b**). Specifically, patients with translocation 1;19 (t(1;19), which creates a fusion between the genes transcription factor 3 and PBX homeobox 1 (*TCF3–PBX1*)) had a contraction in immature B cell populations as compared to patients without this translocation, corroborating prior findings[19,20]. In cytokine-receptor-like factor 2 (*CRLF2*)-rearranged ALL, the pro-BII and pre-BII compartments were contracted. In some cases, we had access to matched BM and peripheral blood samples at the time of diagnosis. In general, the overall
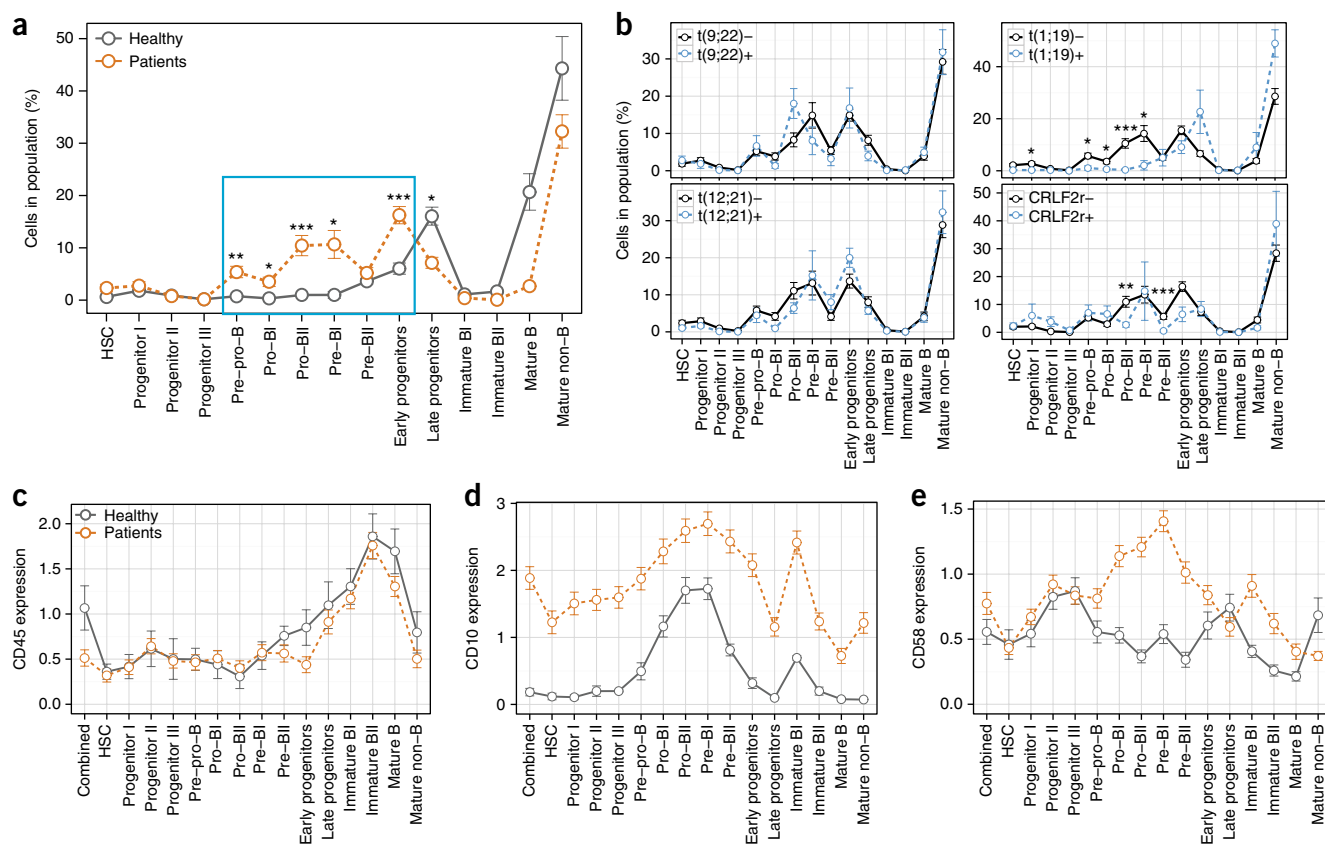
Figure 3 Developmental classification reveals that BCP-ALL results in an expansion of cells across the pre-pro-B to pre-BI transition. (**a**) Percentage of cells from healthy donor BM (n = 5, gray line) or diagnostic BM samples from patients with BCP-ALL (n = 60, orange line) after classification into each developmental subpopulation. Cell populations that were significantly expanded in leukemic samples are shown in the blue box (pre-pro-B, P = 0.0012; pro-BI, P = 0.011; pro-BII, P = 0.00013; pre-BI, P = 0.011; early progenitors, P = 0.00013; late progenitors contracted, P = 0.036), and the remaining populations did not change significantly ($P \geq 0.05$). (**b**) Percentage of cells in each developmental population from patients with BCP-ALL, grouped by diagnostic cytogenetics: (i) translocation t(9;22)(q34;q11) *BCR–ABL1*: t(9;22)– (n = 50) versus t(9;22)+ (n = 10) (top left); (ii) translocation t(1;19)(q23;p13) *TCF3–PBX1*: t(1;19)– (n = 56) versus t(1;19)+ (n = 4): progenitor I, P = 0.015; pre-pro-B, P = 0.037; pro-BI, P = 0.026; pro-BII, P = 6.2×10^{-6}; pre-BI, P = 0.022 (top right); (iii) translocation t(12;21)(p13;q22) *ETV6–RUNX1*: t(12;21)– (n = 47) versus t(12;21)+ (n = 13) (bottom left); and (iv) *CRLF2*-rearranged: CRLF2r– (n = 51) versus CRLF2r+ (n = 9): pro-BII, P = 0.0033; pre-BII, P = 0.00017 (bottom right). (**c–e**) Antigen expression on BM developmental populations from healthy donors (n = 5, gray line) or patients with BCP-ALL (n = 60, orange line): CD45 (**c**), CD10 (**d**) and CD58 (**e**). "Combined" denotes expression in all cells without developmental classification. Throughout, data are mean ± s.e.m. *P < 0.05, **P < 0.01, ***P < 0.001 by an unpaired two-tailed Welch's *t*-test using Bonferroni correction for multiple comparisons.

classification did not substantially change on the basis of the use of BM versus peripheral blood samples; however, some pro-B to pre-B transitional populations were not as frequently seen in the peripheral blood as in the BM (**Supplementary Fig. 3b**). The weak association of single-cell developmental classification with recurrent chromosomal rearrangements emphasizes the general applicability of our approach across BCP-ALL samples with diverse clinical genetics.

BCP-ALL cells maintain the hallmarks and co-expression patterns of early B cells

By comparing the developmental leukemic cell subsets to B cell subsets from healthy individuals, we found three major patterns: (i) an expression pattern similar to that in healthy B cells (e.g., CD45 and IKAROS), (ii) overexpression in all cell populations (e.g., CD10 and PAX5), and (iii) a developmentally inappropriate expression pattern (e.g., CD58, CD123 and CD43) (**Fig. 3c–e** and **Supplementary Fig. 3c–e**). Unexpectedly, despite the aberrant expression of three (of 24 examined) proteins associated with B cell development, BCP-ALL

cells generally maintained the expected developmental progression pattern of protein expression in leukemic populations as compared to that in healthy BM. Notably, PAX5, a master B cell transcription factor encoded by a gene that is frequently mutated in BCP-ALL[21], was highly overexpressed on all leukemic cell populations, even in the case of two patients who harbored heterozygous *PAX5* deletions according to genomic analysis (**Supplementary Fig. 3f** and **Supplementary Table 4**). In contrast to the general maintenance of phenotypic expression of surface and/or intracellular developmental proteins in leukemic cells, there was a higher frequency of BCP-ALL cells with activated basal signaling than in the healthy BM controls. In particular, the frequency of cells with active ribosomal protein S6 (RPS6), which is activated during protein translation and is downstream of mechanistic target of rapamycin (mTOR), or with activated cAMP-response-element-binding protein (CREB) expression was markedly greater in leukemic populations than in healthy B cells (**Supplementary Fig. 3g**). Thus, using healthy B cell progenitor populations to organize the leukemic subpopulations in patients with BCP-ALL provides a refined view

of aberrations in phenotypic and regulatory molecule expression in leukemic cell populations.

Developmentally dependent predictor of relapse, a data-driven model based on the cellular features of BCP-ALL cells at diagnosis

Long-term survival without recurrence of disease is the primary clinical indicator of therapeutic success for patients with BCP-ALL. In patients for whom ≥3 years of follow-up data were available ($n = 54$), 31% relapsed ($n = 17$), a slightly higher rate than the 15–20% expected relapse rate for childhood BCP-ALL[15]. We first determined whether developmental classification alone could be used to stratify patients who would go on to relapse from those who would remain in remission. Cells from patients who relapsed were not enriched in a particular developmental state relative to cells from patients who did not relapse (**Supplementary Fig. 4a**), suggesting that relapse was not strictly connected to a particular phenotypic developmental state.

We used the cellular features at the time of diagnosis for each patient to construct a model to predict clinical outcome (relapse versus continued remission) and identify a short list of leukemic cell features at the time of diagnosis that were sufficient to predict relapse (**Fig. 4a**). To do this, we used a machine-learning approach, termed elastic net, that was designed to identify a small set of predictive features while preserving predictive power[22]. Because we could not exclude the possibility that some patients could relapse after the last documented follow-up appointment, we modeled relapse as a time-to-event outcome in an elastic-net-regularized Cox model[23]. We applied this method to a set of cell features from the expanded leukemic cell populations, including frequency of cells in each population, expression of surface and intracellular proteins, and frequency of cells with activated signaling molecules in the unperturbed state and in response to each perturbation (**Supplementary Table 5**). We also included patient age and white blood cell count at diagnosis, as these features are considered important clinically (**Supplementary Table 1**). This resulted in 352 features per patient across 54 patients, for a total of 19,008 data points.

Using random sampling, we divided the patient cohort into a training cohort (80% of patients, $n = 44$) and a validation cohort (20% of patients, $n = 10$). We applied pre-validation[24,25] to estimate the performance of the relapse prediction model within the training cohort, and then we validated the final model on the independent validation cohort. We therefore assessed the predictive performance of our model twice: within the training cohort and then within the validation cohort. We termed the resulting model 'developmentally dependent predictor of relapse' (DDPR; pronounced 'deeper').

Activated signaling in pro-BII and pre-BI cells at diagnosis predicts relapse

Of the 352 features, hierarchical clustering of six cellular features identified by DDPR almost perfectly separated patients according to their last documented relapse status (**Fig. 4b**). Examining individual features revealed two developmentally dependent patterns: (i) the ability of cells to respond to *ex vivo* stimulation was associated with continuous remission (**Fig. 4b**, yellow box) and (ii) an increased frequency of cells with basally active RPS6 signaling was associated with relapse (**Fig. 4b**, orange box). All six features were confined to pro-BII and pre-BI cell populations. In pro-BII cells of patients who ultimately relapsed, we observed high basal activity of RPS6 and a lack of response to the phosphatase inhibitor pervanadate, as

assessed by phosphorylated eukaryotic translation initiation factor 4E–binding protein 1 (p4EBP1), due to high basal activation of 4EBP1 (**Fig. 4c** (left) and **Supplementary Fig. 4b**). Similarly, in pre-BI cells, the response to pre-B cell receptor (pre-BCR) crosslinking or pervanadate treatment, as assessed by pRPS6 and pCREB levels, and the response to thymic stromal lymphopoietin (TSLP) stimulation, as assessed by phosphorylated spleen tyrosine kinase (pSYK), were significantly blunted in patients who went on to relapse (**Fig. 4c** (right), **Supplementary Fig. 4b** (right, **c**). In essence, the signaling features that were predictive of relapse related to high basal activation of the mTOR pathway in pro-BII cells, as well as high basal activation and a lack of response to stimulation of the pre-BCR pathway in pre-BI cells. Differences in these features were apparent even when applied to a single common genetic subgroup of patients with translocation *ETV6–RUNX1* (t(12;21)(p13;q22); **Supplementary Fig. 4d**), who generally have a favorable prognosis.

To assess DDPR performance, we calculated an integrated cumulative/dynamic area under the curve (iAUC)[26] and a *C*-statistic[27], the measures appropriate for censored time-to-event data. In the training cohort, DDPR had a predicted (cross-validated) iAUC value of 0.92 and a *C*-statistic of 0.87 (**Fig. 4d**, left). Applying the model to the independent validation cohort resulted in an iAUC value of 0.85 and a *C*-statistic of 0.87 (**Fig. 4d**, right), indicating strong model performance. To determine whether predictive cellular features could be detected in bulk leukemia cells without developmental classification, we repeated this analysis using features from all of the cells in the blast cell gate. Using data from unclassified bulk cells, we found that the model performed inferiorly to DDPR (training cohort iAUC$_{DDPR}$ = 0.92 versus iAUC$_{bulk}$ = 0.71; validation cohort iAUC$_{DDPR}$ = 0.85 versus iAUC$_{bulk}$ = 0.66; **Supplementary Fig. 4e**). DDPR performed well as a risk-stratification method at diagnosis in predicting relapse-free survival (RFS) in a retrospective analysis of both cohorts ($P = 2.8 \times 10^{-7}$; **Fig. 4e** and **Supplementary Fig. 4f**). Thus, organizing data from primary diagnostic leukemia samples using single-cell developmental classification was critical for predicting future clinical outcomes.

DDPR synergizes with current risk-stratification methods

Current risk-prediction methodology integrates a combination of clinical (National Cancer Institute (NCI)-Rome criteria[28]) and genetic (cytogenetic) features at diagnosis, as well as an early response to therapy (prednisone response and/or minimal residual risk (MRD)[29,30]). Final risk, which guides clinical decisions, integrates all of these features and is generally determined 3 months following initiation of treatment, although for some patients risk may be known earlier.

As expected, NCI–Rome criteria successfully stratified 53 patients with available risk data according to RFS (log-rank $P = 0.0083$; **Fig. 5a**, top). However, integrating DDPR with NCI–Rome criteria resulted in a significant improvement in risk stratification, as assessed by the integrated discrimination improvement index (IDI), continuous net reclassification improvement (NRI) and median improvement (MI) for time-to-event data[31,32] at 5 years following diagnosis (**Fig. 5a** (bottom) and **Supplementary Fig. 5a**). Within the 45 patients for whom MRD and final risk were available (per AIEOP-BFM 2000 protocol definitions), either MRD ($P = 0.0086$; **Fig. 5b**, top) or final risk ($P = 0.044$; **Fig. 5c**, top) alone performed well in stratifying patients according to RFS. Combining DDPR prediction with either MRD (**Fig. 5b**, bottom) or final risk (**Fig. 5c**, bottom) significantly improved patient stratification at 5 years following diagnosis ($P < 0.05$; **Supplementary Fig. 5a**).
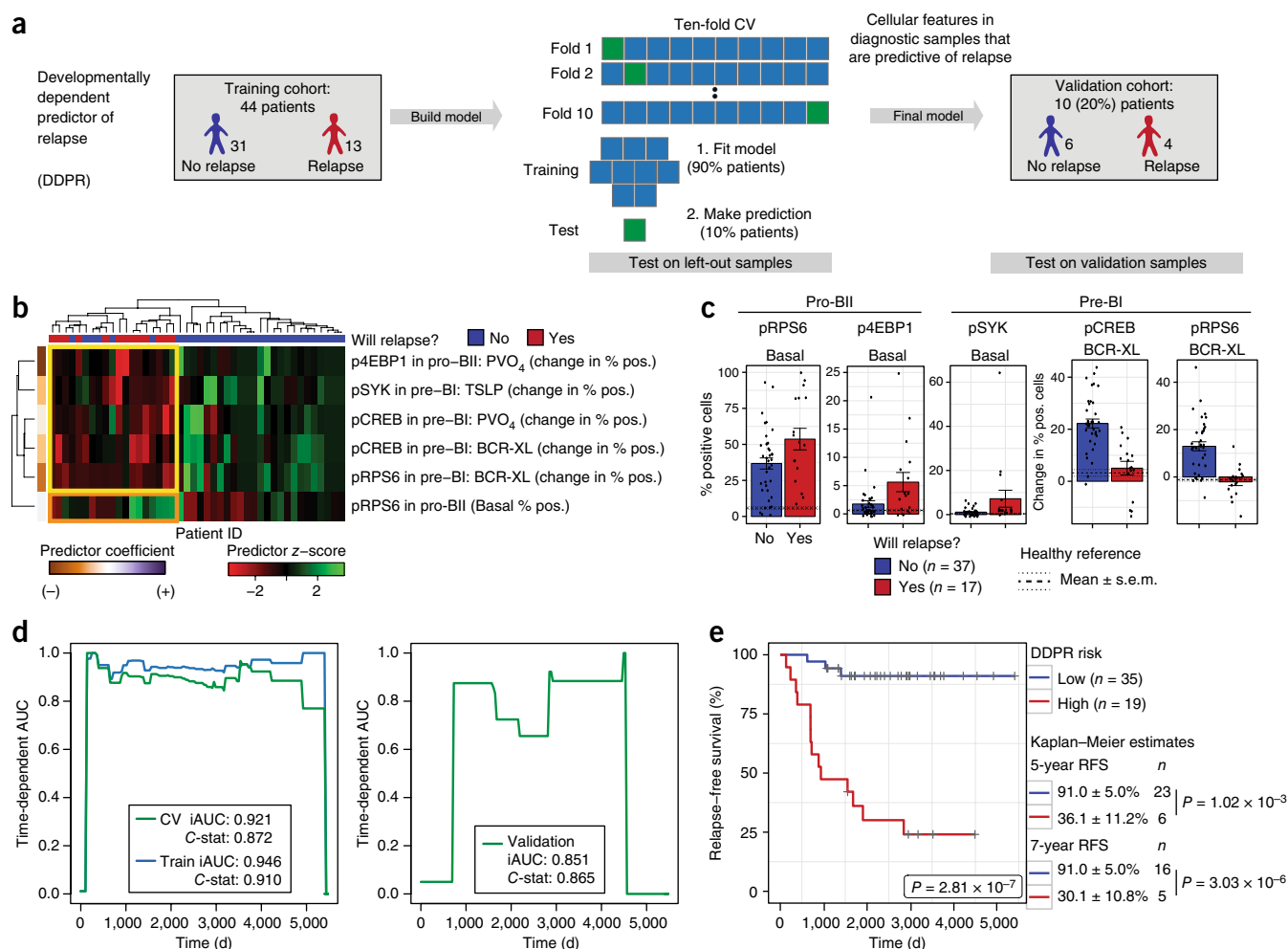
Figure 4 DDPR predicts which patients will go on to relapse based on the features of expanded BCP-ALL populations at the time of diagnosis. (**a**) Construction of the DDPR model that predicts relapse in patients with BCP-ALL. Data from 54 patients with ≥3 years of follow-up data were divided into training ($n = 44$) and validation ($n = 10$) cohorts. Cellular features available to DDPR included signaling in the basal state, changes in signaling state following perturbations, mean arsinh-transformed expression of surface and intracellular proteins, and frequency of cells in the expanded developmental populations. DDPR performance was estimated using ten-fold cross-validation (CV) within the training cohort to yield pre-validated relative risk for each patient. The final DDPR model (elastic-net-regularized Cox model) was then built using all of the training cohort samples. Once constructed, DDPR was applied to predict relative risk for samples in the validation cohort. (**b**) Hierarchical clustering of six predictive features of relapse identified by DDPR within the training cohort. The last documented relapse status is shown above the heat map as relapse (red) or continuous complete remission (blue). Coefficients of predictors are shown on the left of the heat map. Yellow box indicates five features with negative correlation to relapse. Orange box indicates one feature with positive correlation to relapse. BCR-XL, B cell receptor crosslink; PVO_4, pervanadate. (**c**) Percentage of cells positive for key DDPR cellular features in pro-BII and pre-BI cells in all patients ($n = 54$). Data are mean ± s.e.m. P values are not shown, because these features were selected to be different and nonredundant between classes (unpaired two-tailed Welch's t-test from left to right would yield: $P = 0.055$, $P = 0.044$, $P = 0.13$, $P = 5.7 \times 10^{-6}$ and $P = 1.6 \times 10^{-7}$). Dashed lines indicate mean levels in the corresponding developmental populations within healthy BM aspirates of five healthy donors; dotted lines indicate s.e.m. (**d**) Time-dependent AUC curves showing performance for relapse prediction in the training (left) and validation (right) cohorts. iAUC and C-statistic (C-stat) summary measures are shown for each curve built using pre-validated (green, left), overall model fit (blue, left) and predicted (green, right) relative risk of relapse with reference to the sample average. (**e**) Kaplan–Meier analysis of RFS of all patients with ≥3 years of follow-up data ($n = 54$), stratified by DDPR risk group. An estimate for relative risk of relapse was used to assign a risk group to each patient (pre-validated in the training cohort; predicted in the validation cohort; Online Methods). P values were calculated using the log-rank test. Log-rank tests for: training cohort alone, $P = 5.6 \times 10^{-6}$; validation cohort alone, $P = 0.040$. RFS estimates, s.e.m., number of patients at risk and P values for both groups at 5 and 7 years are shown on the right (5 years, $P = 1.02 \times 10^{-3}$; 7 years, $P = 3.03 \times 10^{-6}$).

In clinical practice, relapse risk determines treatment decisions for patients. Key DDPR features were clearly different in patients who went on to relapse from those who remained in continuous remission, regardless of their NCI–Rome or MRD status (**Supplementary Fig. 5b,c**), suggesting that cellular phenotypes identified by DDPR may provide rational drug targets for patients at risk of relapse.

Cells characterized by activated mTOR and deficits in pre-BCR signaling persist from diagnosis to relapse

The analyses described thus far do not establish whether features identified by DDPR are present at relapse. To understand how BCP-ALL is remodeled under the pressure of treatment, seven matched diagnosis–relapse sample pairs were analyzed (**Fig. 6a**). These samples
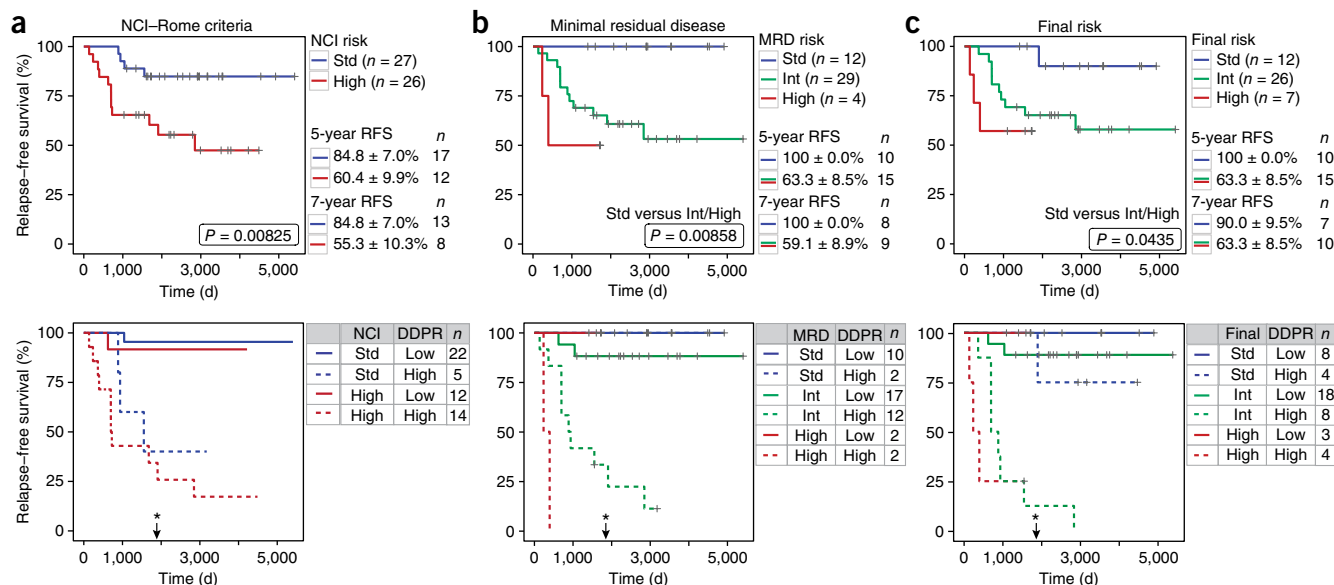
Figure 5 DDPR synergizes with existing risk-stratification methods to improve relapse-free survival prediction for patients with BCP-ALL. (**a**) Kaplan–Meier analysis showing RFS for patients with ≥3 years of follow-up data and known NCI–Rome criteria ($n = 53$), stratified by NCI–Rome criteria alone (top) or by NCI–Rome criteria and DDPR (bottom). (**b,c**) RFS as in **a** for patients with known MRD (**b**) or final risk (**c**), as determined by protocol definitions ($n = 45$), stratified by the clinical risk group alone (top) or risk group and DDPR (bottom). An estimate for relative risk of relapse was used to assign a DDPR risk group to each patient (pre-validated in the training cohort; predicted in the validation cohort; Online Methods). Kaplan–Meier estimates with s.e.m. and the number of people at risk are shown for 5-year and 7-year RFS in the top plots (5-year and 7-year P values (respectively): NCI–Rome criteria: $P = 0.033$ and $P = 0.040$, MRD: $P = 0.157$ and $P = 0.084$, final risk: $P = 0.157$ and $P = 0.169$). All P values were calculated using the log-rank test. P values in **b,c** are between the standard-risk (Std) and the intermediate- or high-risk (Int/High) groups due to low number of patients in the high-risk group. Black arrows with asterisks indicate a significant (*$P < 0.05$) improvement in patient risk stratification at 5 years following diagnosis achieved by adding DDPR to each established criterion: continuous net reclassification improvement (NRI) for NCI–Rome criteria: $P = 0.027$, MRD: $P = 0.033$, final risk: $P = 0.013$ (see **Supplementary Fig. 5a** for NRI estimates and 95% confidence intervals).

comprised diverse underlying prognostic genetics, including those with good risk (*ETV6–RUNX1*; $n = 1$), poor risk (*BCR–ABL1*; $n = 2$) and without known prognostic genetic aberrations ($n = 4$). In contrast to the samples collected at the initial time of diagnosis, which showed an expansion spanning the pre-pro-B to pre-BI transition, at relapse the expansion narrowed almost exclusively to the pre-BI population, but this expansion was also present in the diagnosis specimen (**Fig. 6a**, red box). In these paired samples, DDPR predictive features were present at diagnosis and were either maintained or exacerbated at relapse (**Fig. 6b**). Taken together, these results indicate that the cellular populations and features associated with poor outcome exist at diagnosis and persist at relapse.

Because we are able to examine the concomitant expression of proteins in the same cell, we examined single-cell pairwise correlation between pRPS6 and p4EBP1 in pro-BII cells from the matched specimens. We found that the levels of these proteins were correlated at both the time of diagnosis and after relapse (Spearman's ρ (± s.e.m.) for pRPS6–p4EBP1: at diagnosis, $\rho = 0.31$ (± 0.10); at relapse, $\rho = 0.39$ (± 0.05); **Fig. 6c** (left) and **Supplementary Fig. 6a**). Similarly, pre-BI cells displayed a moderate correlation between their implicated molecules pSYK, pCREB and pRPS6 (for example, pCREB–pRPS6: at diagnosis, $\rho = 0.27$ (± 0.04); at relapse, $\rho = 0.45$ (± 0.05); **Fig. 6c** (right) and **Supplementary Fig. 6b**).

To quantify the strengths of signaling relationships between these proteins in the pro-BII and pre-BI populations, we used 'conditional density resampled estimate of mutual information' (DREMI) to estimate the dependency between each pair of proteins[33]. In leukemic pre-BI cells, pSYK, pCREB and pRPS6 were indeed likely to belong to

cross-correlated signaling networks that persisted from diagnosis to relapse (**Fig. 6d**). These pre-BI cells could not respond to pre-BCR crosslinking by further increasing the levels of pCREB and pRPS6 (**Supplementary Fig. 6b**). Similarly, in the pro-BII population, DREMI confirmed an expected dependency between p4EBP1 and pRPS6 (**Supplementary Fig. 6c**). Notably, for both p4EBP–pRPS6 and pSYK–pCREB dependencies, the strength of the signaling relationship was enhanced at the time relapse, indicating that there was either a modest strengthening of the relationship between these molecules or an increased homogeneity of cell populations. In either case, the evidence points toward the importance of these signaling relationships during progression toward an aggressive tumor state. By comparison, in patients who did not relapse, co-activation of CREB, RPS6 and SYK at baseline was diminished in pre-BI cells, and these cells had the ability to co-activate CREB and RPS6 in response to pre-BCR crosslinking (**Supplementary Fig. 6d**).

To determine whether the predicted cellular signaling phenotypes could be therapeutically targeted, we treated primary diagnostic samples *ex vivo* with chemical inhibitors that target the activated pathways. Treatment of healthy or leukemic cells with BEZ235, a dual phosphatidylinositol-4,5-bisphosphate 3-kinase (PI3K) and mTOR inhibitor (PI3K–mTORi), reduced the frequency of pro-BII cells with activated p4EBP1 (**Fig. 6e**, left). However, phosphorylation of RPS6 was not as strongly inhibited by treatment with BEZ235 in leukemic pro-BII and pre-BI cells, suggesting that there were multiple routes for RPS6 activation (**Supplementary Fig. 6e**). By contrast, treatment with dasatinib, a dual ABL and SRC family kinase inhibitor (ABL–SFKi), reduced the frequency of pre-BI cells
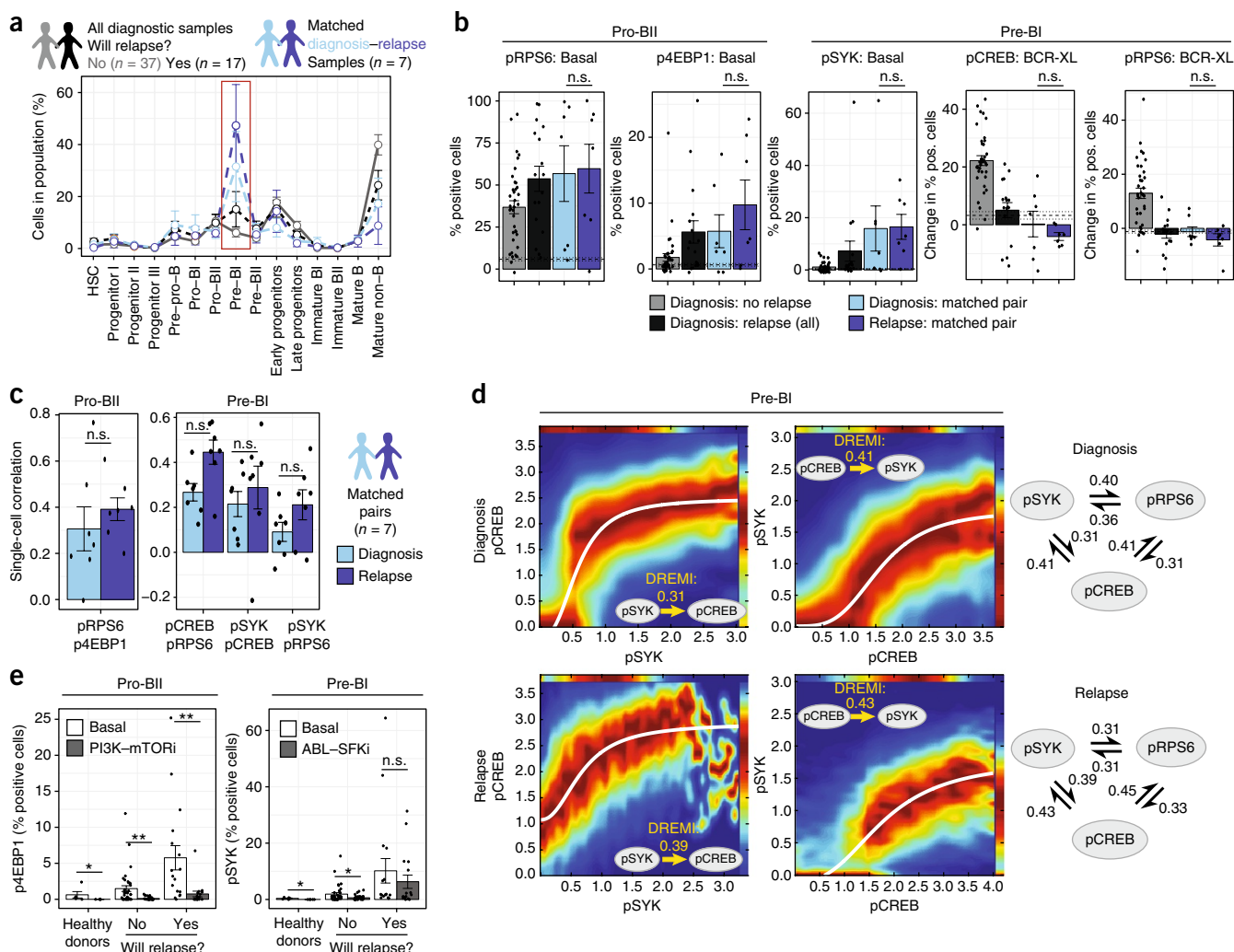
Figure 6 Cells with DDPR features are present at the time of diagnosis and persist at relapse. (**a**) Percentage of cells in each developmental population of all diagnostic BCP-ALL samples from patients who either stayed in continuous remission for ≥3 years (gray, $n = 37$) or went on to relapse (black, $n = 17$), as compared to that in matched diagnosis–relapse pairs (diagnosis, blue; relapse, purple; $n = 7$). Red box highlights a significant ($P = 0.0030$) expansion of the pre-BI population at relapse as compared to that in diagnostic samples from patients who did not relapse. P values were calculated using a two-sided Tukey's honest significance test and were corrected for multiple comparisons using Bonferroni correction. (**b**) Bar plots showing key DDPR features in all diagnostic samples compared to matched diagnosis–relapse pairs, as in **a**: percentage of pro-BII cells with pRPS6 or p4EBP1 in the nonstimulated (basal) state, percentage of pre-BI cells with pSYK in the basal state, and change from the basal state in the percentage of pre-BI cells with pCREB or pRPS6 signaling following BCR cross-linking (BCR-XL). (**c**) Bar plots showing the Spearman's rank correlation coefficient for the key DDPR features listed in **b**, calculated for matched diagnosis–relapse pairs ($n = 7$): single-cell correlation of arsinh-transformed values between pRPS6 and p4EBP1 in pro-BII cells (left), or between pCREB and pRPS6, pSYK and pCREB, or pSYK and pRPS6 in pre-BI cells (right). None of the DDPR features changed significantly from the time of diagnosis to relapse in **b,c** (paired two-tailed Welch's t-test applied to matched diagnosis–relapse pairs only). (**d**) DREMI analysis and conditional density rescaled visualization (DREVI) analysis for DDPR features in pre-BI cells. Up to 5,000 pre-BI cells from matched diagnosis–relapse pairs ($n = 7$) were sampled and pooled before analysis. Left, estimated conditional density functions for pSYK-to-pCREB signaling response (pSYK→pCREB) and for pCREB→pSYK at diagnosis and at relapse; sigmoidal response functions were fitted to each plot. Right, quantification of strengths of pairwise signaling relationships within the network formed by pSYK, pCREB and pRPS6 at diagnosis and at relapse. (**e**) Bar plots showing response of DDPR features (basal p4EBP1 in pro-BII cells and basal pSYK in pre-BI cells) to short-term *ex vivo* treatment (see **Supplementary Table 3**) in samples from healthy donors ($n = 5$) or patients with BCP-ALL (no relapse, $n = 37$; relapse, $n = 17$). Shown are the effects of BEZ235 (a dual PI3K and mTOR inhibitor, PI3K–mTORi) in pro-BII cells (healthy, $P = 0.023$; no relapse, $P = 0.0032$; relapse, $P = 0.0092$) and of dasatinib (a dual BCR–ABL kinase and SRC family kinase inhibitor, ABL–SFKi) in pre-BI cells (healthy, $P = 0.031$; no relapse, $P = 0.048$; relapse, $P = 0.22$). Effects were assessed using a two-tailed Welch's t-test. Throughout, data are mean ± s.e.m. *$P < 0.05$; **$P < 0.01$; n.s, not significant ($P \geq 0.05$).

with activated SYK in patients who remained in remission to levels that approached those in healthy BM; yet, its effect was minor in patients who went on to relapse (**Fig. 6e**, right). Treatment with dasatinib was also able to slightly reduce the frequencies of pro-BII cells with activated RPS6 and 4EBP1, as well as of pre-BI cells with

activated RPS6 and CREB (**Supplementary Fig. 6f**). Taken together, these data indicate that some DDPR features may be therapeutically targetable; however, knowledge of underlying developmental signaling must guide the design of therapeutic approaches for improving patient outcomes.

DISCUSSION

Deep phenotyping of primary cancer samples provides a unique opportunity to examine intratumoral heterogeneity and link patient outcomes to specific cellular populations. Yet, it remains challenging to organize these massive data sets into meaningful models. Data from single-cell analyses of primary tumors must be stratified to understand how cellular diversity of tumors affects disease progression or treatment response.

This study leveraged the phenotypic profiling depth enabled by mass cytometry to perform high-parameter single-cell analyses of primary patient samples, without extended *ex vivo* culture or passage through an immunodeficient mouse, to preserve biology as it exists in the patient. We addressed the intrinsic heterogeneity of the single-cell data by showing that BCP-ALL cells could be reliably aligned to a developmental continuum of normal cell populations. In doing so, we organized diverse data from multiple patients into a 'universal' physiological standard that allowed relevant comparisons. This approach not only allows for the identification of key cell populations and their behaviors in relation to clinical outcome, but also provides mechanistic insight into the persistent disease state.

Direct comparison of BCP-ALL cells to human B cell progenitors enabled identification of developmental states that were most vulnerable to malignant transformation: the pre-pro-B to pre-BI transition. During this transition, normal B cells rearrange their immunoglobulin heavy chain locus with coordinated cell fate decisions. We have previously demonstrated changes in network structure surrounding the differentiation of pro-B cells into pre-B cells, particularly regarding the transition between the interleukin (IL)-7 receptor and pre-BCR signaling pathways[14]. Signaling through the IL-7 receptor in healthy pro-B cells activates the JAK–STAT and PI3K signaling pathways[34,35]. Subsequently, normally developing B cells proceed to further expand and differentiate after receiving a strong pre-BCR signal that informs them of successful heavy chain rearrangement[34]. DDPR identified activation of signaling molecules in the pre-BCR (pSYK and pCREB) and PI3K–mTOR (pRPS6 and p4EBP1) pathways around this developmental transition in individuals who were likely to relapse. Aberrant BCR-related signaling is known to be pathogenic in other B cell malignancies, such as diffuse large B cell lymphoma[36]. Dysregulation of pre-BCR signaling has been described in Philadelphia-chromosome-rearranged (Ph+) BCP-ALL, which is likely mediated by upregulation of inhibitory regulators or mutations in genes that encode them[37–39]. Our results suggest that leukemia may exploit this normal developmental process to maintain pre-BCR signaling at a 'just-right' level. The ABL–SFKi was effective in targeting pre-BCR signaling in pre-BI cells, as assessed by phosphorylation of SYK, but only in patients who would not go on to relapse. Thus, an understanding of normal developmental signaling states, and how leukemic cells maintain or diverge from these states in their developmental context, can guide therapeutic efforts to affect patient outcomes.

As we have demonstrated previously, cellular behavior is a key determinant of resistance to therapy[40–44]. It is not just the outward identity of leukemia cells, as measured by the presence of molecules on the cell surface, that determines therapy resistance, but instead how those cells behave, as measured by intracellular signaling states. This model cannot currently determine the origin of leukemic transformation, but our data suggest that the drive of leukemic cells to differentiate remains strong as they attempt to continue their developmental program. Given the limited size of the patient cohort studied, DDPR must be applied to larger cohorts of patients. It will be of interest to extend this type of analysis to adult BCP-ALL, as it is well understood that an older age at diagnosis carries a higher risk of relapse. Moreover, this developmental model of BCP-ALL would benefit from integration with genomic, epigenetic and transcriptomic investigations of both healthy and leukemic populations. Defining the mechanisms that control survival and proliferation decisions at the pro-BII-to-pre-BI developmental transition will be informative with regard to how BCP-ALL exploits these pathways and how genetic lesions associated with leukemia cooperate with the underlying developmental program.

Altogether, this study shows that aligning transformed cells to their normal developmental path can improve the risk-stratification system and identify, with improved precision, the most relevant cell populations for further study and therapeutic targeting. More broadly, this study highlights the translational value of understanding cancer at the single-cell level and applying machine learning to guide treatment paradigms for patients with cancer.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS

G.P.N. and K.L.D. conceptualized the study; Z.G., S.C.B., E.F.S., L.W., N.S., N.A., G.F., W.J.F., R.T. and K.L.D. developed the methods; Z.G., J.S., A.J. and K.L.D. performed all experiments; Z.G., N.S. and R.T. developed the software; Z.G. and R.T. performed formal data analysis for all of the data generated; N.J.L., G.G., A.B., K.L.D. and G.P.N. provided funding and patient samples supporting the study; and Z.G., J.S., S.C.B., R.T., G.G., K.L.D. and G.P.N. wrote and edited the manuscript.

1. Anderson, K. *et al.* Genetic variegation of clonal architecture and propagating cells in leukemia. *Nature* **469**, 356–361 (2011).
2. Walter, M.J. *et al.* Clonal architecture of secondary acute myeloid leukemia. *N. Engl. J. Med.* **366**, 1090–1098 (2012).

3. Welch, J.S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).

4. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).

5. Klco, J.M. *et al.* Functional heterogeneity of genetically defined subclones in acute myeloid leukemia. *Cancer Cell* **25**, 379–392 (2014).

6. Duque-Afonso, J. & Cleary, M.L. The AML salad bowl. *Cancer Cell* **25**, 265–267 (2014).

7. Bhojwani, D. & Pui, C.H. Relapsed childhood acute lymphoblastic leukemia. *Lancet Oncol.* **14**, e205–e217 (2013).

8. Greaves, M.F. Differentiation-linked leukemogenesis in lymphocytes. *Science* **234**, 697–704 (1986).

9. Bhojwani, D. *et al.* Biologic pathways associated with relapse in childhood acute lymphoblastic leukemia: a Children's Oncology Group study. *Blood* **108**, 711–717 (2006).

10. Mulligan, C.G. *et al.* Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322**, 1377–1380 (2008).

11. Hogan, L.E. *et al.* Integrated genomic analysis of relapsed childhood acute lymphoblastic leukemia reveals therapeutic strategies. *Blood* **118**, 5218–5226 (2011).

12. Mulligan, C.G. *et al.* BCR–ABL1 lymphoblastic leukemia is characterized by the deletion of *Ikaros. Nature* **453**, 110–114 (2008).

13. LeBien, T.W. & Tedder, T.F. B lymphocytes: how they develop and function. *Blood* **112**, 1570–1580 (2008).

14. Bendall, S.C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).

15. Teachey, D.T. & Hunger, S.P. Predicting relapse risk in childhood acute lymphoblastic leukemia. *Br. J. Haematol.* **162**, 606–620 (2013).

16. Spitzer, M.H. *et al.* An interactive reference framework for modeling a dynamic immune system. *Science* **349**, 1259425 (2015).

17. Samusik, N., Good, Z., Spitzer, M.H., Davis, K.L. & Nolan, G.P. Automated mapping of phenotype space with single-cell data. *Nat. Methods* **13**, 493–496 (2016).

18. Zunder, E.R., Lujan, E., Goltsev, Y., Wernig, M. & Nolan, G.P. A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry. *Cell. Stem Cell* **16**, 323–337 (2015).

19. Bicocca, V.T. *et al.* Cross-talk between ROR1 and the pre-B cell receptor promotes survival of t(1;19) acute lymphoblastic leukemia. *Cancer Cell* **22**, 656–667 (2012).

20. Crist, W. *et al.* Prognostic importance of the pre-B cell immunophenotype and other presenting features in B lineage childhood acute lymphoblastic leukemia: a Pediatric Oncology Group study. *Blood* **74**, 1252–1259 (1989).

21. Mulligan, C.G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukemia. *Nature* **446**, 758–764 (2007).

22. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

23. Tibshirani, R. The lasso method for variable selection in the Cox model. *Stat. Med.* **16**, 385–395 (1997).

24. Tibshirani, R.J. & Efron, B. Pre-validation and inference in microarrays. *Stat. Appl. Genet. Mol. Biol.* **1**, Article1 (2002).

25. Höfling, H. & Tibshirani, R. A study of pre-validation. *Ann. Appl. Stat.* **2**, 643–664 (2008).

26. Uno, H., Cai, T., Tian, L. & Wei, L.J. Evaluating prediction rules for *t*-year survivors with censored regression models. *J. Am. Stat. Assoc.* **102**, 527–537 (2007).

27. Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B. & Wei, L.J. On the *C*-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statist. Med.* **121**, 1105–1117 (2011).

28. Smith, M. *et al.* Uniform approach to risk classification and treatment assignment for children with acute lymphoblastic leukemia. *J. Clin. Oncol.* **14**, 18–24 (1996).

29. Basso, G. *et al.* Risk of relapse of childhood acute lymphoblastic leukemia is predicted by flow cytometric measurement of residual disease on day 15 bone marrow. *J. Clin. Oncol.* **27**, 5168–5174 (2009).

30. Borowitz, M.J. *et al.* Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia and its relationship to other prognostic factors: a Children's Oncology Group study. *Blood* **111**, 5477–5485 (2008).

31. Pencina, M.J., D'Agostino, R.B. Sr. & Steyerberg, E.W. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat. Med.* **30**, 11–21 (2011).

32. Uno, H., Tian, L., Cai, T., Kohane, I.S. & Wei, L.J. A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Stat. Med.* **32**, 2430–2442 (2013).

33. Krishnaswamy, S. *et al.* Conditional density-based analysis of T cell signaling in single-cell data. *Science* **346**, 1250689 (2014).

34. Clark, M.R., Mandal, M., Ochiai, K. & Singh, H. Orchestrating B cell lymphopoiesis through interplay of IL7 receptor and preB cell receptor signaling. *Nat. Rev. Immunol.* **14**, 69–80 (2014).

35. O'Reilly, L.A. *et al.* MEK/ERK-mediated phosphorylation of Bim is required to ensure survival of T and B lymphocytes during mitogenic stimulation. *J. Immunol.* **183**, 261–269 (2009).

36. Alizadeh, A.A. *et al.* Distinct types of diffuse large B cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).

37. Trageser, D. *et al.* Pre-B cell receptor–mediated cell cycle arrest in Philadelphia chromosome–positive acute lymphoblastic leukemia requires IKAROS function. *J. Exp. Med.* **206**, 1739–1753 (2009).

38. Feldhahn, N. *et al.* Mimicry of a constitutively active pre-B cell receptor in acute lymphoblastic leukemia cells. *J. Exp. Med.* **201**, 1837–1852 (2005).

39. Shojaee, S. *et al.* PTEN opposes negative selection and enables oncogenic transformation of pre-B cells. *Nat. Med.* **22**, 379–387 (2016).

40. Kotecha, N. *et al.* Single-cell profiling identifies aberrant STAT5 activation in myeloid malignancies with specific clinical and biologic correlates. *Cancer Cell* **14**, 335–343 (2008).

41. Gibbs, K.D. Jr. *et al.* Decoupling of tumor-initiating activity from stable immunophenotype in HoxA9-Meis1-driven AML. *Cell. Stem Cell* **10**, 210–217 (2012).

42. Irish, J.M. *et al.* B cell signaling networks reveal a negative prognostic human lymphoma cell subset that emerges during tumor progression. *Proc. Natl. Acad. Sci. USA* **107**, 12747–12754 (2010).

43. Irish, J.M. *et al.* Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* **118**, 217–228 (2004).

44. Levine, J.H. *et al.* Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).

ONLINE METHODS

Bone marrow samples from patients and healthy donors. Reagents and methods relevant to this work can also be found in the **Life Sciences Reporting Summary**. Fresh human BM was obtained from healthy adult donors ($n = 5$ (three females, ages 20, 27 and 44 years old, and two males, ages 26 and 28 years old); average age, 29 years; AllCells, Alameda, CA, USA). De-identified BM samples from pediatric patients with BCP-ALL were obtained under informed consent from the Lucile Packard Children's Hospital at Stanford (Stanford, CA, USA; Ph+ samples, $n = 9$) and the Pediatric Clinic University of Milan Bicocca (Monza, Italy; $n = 51$) for a total of 60 primary diagnostic patient samples. Use of these samples was approved by the Institutional Review Boards at both institutions. All relevant ethical regulations were followed in this study. No a priori power calculation was performed, as this is the first study of its kind. Patients were included based on availability of sufficient cells for a mass-cytometry-based study and on the availability of at least 3 years of follow-up data from the date of diagnosis. Infants with MLL-rearranged BCP-ALL were excluded from this study. Clinical data were available for these samples, including minimal residual disease (MRD) risk group and final risk assignment as per the AIEOP-BFM ALL 2000 protocol (https://clinicaltrials.gov identifier: NCT00613457)[45], diagnostic cytogenetics, age at diagnosis, gender, white blood cell count at diagnosis, date of diagnosis, date of relapse and date of the last follow-up (**Supplementary Table 1**). Median follow-up time was 5.5 years. In agreement with the AIEOP-BFM ALL 2000 study[45], median time to relapse was 2.0 years. Median follow-up time for patients in continuous complete remission (CCR) was 7.6 years. Gene copy number and mutational analysis for the genes *IKZF1*, *P2RY8–CRLF2*, *CDKN2A*, *CDKN2B*, *PAX5*, *ETV6*, *BTG1*, *RB1* and *ERG* was performed using multiplex-ligation-dependent probe amplification (MLPA) for 20 patients; *IGH2–CRLF2* rearrangement was tested by fluorescent *in situ* hybridization (FISH) analysis for 20 patients; and *JAK2* mutations were identified by high-resolution melt (HRM) analysis for 16 patients (**Supplementary Table 4**).

Clinical protocol definitions. Treatment protocols for each patient are indicated in **Supplementary Table 1**. Therapy for all of the patients was based on the Berlin–Frankfurt–Munster (BFM) backbone[46]. For Italian patients (the majority in this study), MRD risk and final risk were assigned as per the AIEOP-BFM ALL 2000 clinical protocol[45]. Briefly, after 7 d of monotherapy with prednisone and one intrathecally administered dose of methotrexate, treatment was complemented by administration of corticosteroid, vincristine, daunorubicin and L-asparaginase. Remission induction was followed by intravenous administration of cyclophosphamide and cytarabine, intrathecal administration of methotrexate and oral administration of mercaptopurine. Risk group assignment resulted from a combination of presenting features and response to therapy, as measured by microscopic examination of peripheral blood for leukemic cells at day 8 and MRD at time point (TP) 1 (day 33) and TP2 (day 78), as assessed by PCR analysis of immunoglobulin gene rearrangements. Patients were determined to be: standard risk (SR) if MRD was negative at both TP1 and TP2 and no adverse clinical features were observed; intermediate risk (IR) if MRD was positive at TP1 or TP2 but $<10^{-3}$; or MRD high risk (HR) if MRD $> 10^{-3}$ at TP2. The HR group included patients with any of the following criteria: Ph+ ALL, t(4;11) or *MLL–AF4*; 'prednisone-poor response' (PPR) >1,000 blasts/ml on day 8; failure to achieve CR on TP1; or high degree of MRD at TP2 ($>1 \times 10^{-3}$)[45]. Clinical remission was defined by the presence of <5% blasts in the BM by morphologic inspection.

Mass cytometry. Samples were processed as previously described[47]. Briefly, viably preserved BM cells were thawed and resuspended in 90% RPMI medium (ThermoFisher Scientific, Waltham, MA, USA)r?] with 10% FCS supplemented with 20 U/ml sodium heparin (Sigma-Aldrich, St. Louis, MO, USA), 0.025 U/ml Benzonase (Sigma-Aldrich), 1× L-glutamine and 1× penicillin–streptomycin (Invitrogen, Carlsbad, CA, USA). Cells were rested at 37 °C for 30 min and stained for viability with cisplatin as described[48]. Following viability staining, cells were perturbed under the following conditions: treatment with pervanadate, IL-7, thymic stromal lymphopoietin (TSLP), dasatinib, BEZ-235 or tofacitinib, or by BCR crosslinking; sources, concentrations and time points are listed in **Supplementary Table 3**. Cells were then fixed with paraformaldehyde (PFA;

Electron Microscopy Sciences, Hatfield, PA, USA) to a final concentration of 1.6% for 10 min at room temperature. Cells were barcoded using 20-plex palladium barcoding plates prepared in-house as described[49]. To control for batch effects, we included at least one healthy BM reference sample within each barcoding plate. A total of 36 barcode plates were used in this study. Following barcoding, cells were pelleted and washed once with cell-staining medium (CSM; PBS with 0.5% BSA and 0.02% sodium azide) to remove residual PFA. Blocking was performed with Purified Human Fc Receptor Binding Inhibitor (eBioscience, San Diego, CA, USA), following the manufacturer's instructions. Antibodies to surface markers were added, yielding 50- or 100-μl final reaction volumes, and samples were incubated at room temperature for 30 min (**Supplementary Table 2**). Cells were pelleted and washed with CSM before permeabilization with 4 °C methanol for 10 min at 4 °C and optional storage at −80 °C. Cells were washed with CSM and stained with intracellular-marker- and/or phospho-specific antibodies in 50 μl for 30 min at room temperature (**Supplementary Table 2**). Cells were washed once in CSM, then stained with 1:5,000 ^{191}Ir/^{193}Ir DNA intercalator (Fluidigm, South San Francisco, CA, USA) in PBS with 1.6% PFA for 20 min at room temperature. Cells were washed once with CSM, washed twice with double-distilled water, filtered to remove aggregates and resuspended in ^{139}La/^{142}Pr/^{159}Tb/^{169}Tm/^{175}Lu normalization beads[50] immediately before analysis using a CyTOF1 mass cytometer (Fluidigm). Throughout the analysis, cells were maintained at 4 °C and introduced at a constant rate of ~300 cells/s.

Magnetic-bead-based depletion of non-B cell subpopulations from healthy samples. To enrich Ficoll-treated BM from healthy donors for rare hematopoietic and B lymphocyte progenitors, cells were incubated with biotin-conjugated antibodies (**Supplementary Table 6**) for 30 min at a concentration of 5 million cells per 100 μl. Cells were washed with CSM twice and then incubated with BD Streptavidin Particles Plus (BD Biosciences, Franklin Lakes, NJ, USA) at the manufacturer's recommended concentration for 30 min at room temperature. Particle-labeled cells were resuspended in CSM to approximately 2×10^7 to 8×10^7 cells/ml and placed in a magnetic holder for 7 min. The supernatant was transferred to a new tube, and the beads or cells were washed, resuspended and placed back in the magnetic holder for an additional round of depletion and supernatant recovery. This washing procedure was repeated. Cells from the supernatant were then pelleted by centrifugation at 250*g* for 5 min. Depleted healthy cells were then stimulated and fixed before being aliquoted for use as controls on each barcode plate. These samples were then stained and analyzed alongside the leukemia samples.

Processing of mass cytometry data. Data were normalized together using bead normalization[50], and files were debarcoded as described[49]. After debarcoding, we obtained ≥600,000 single-cell events per patient. Single-cell protein expression data were extracted using Bioconductor software (http://www.bioconductor.org) and transformed using the inverse hyperbolic sine (arsinh) function with a cofactor of 5. To control for batch effects among barcoding plates, we performed percentile normalization using the healthy reference BM sample(s) that were included within each plate (normalization values across barcoding plates were in the range of 0.97 to 1.18, indicating that batch effects before normalization were small). Expression of proteins in each population of interest was determined by calculating the mean level of expression after arsinh transformation. Calculation of the percentage of positive cells for each phosphorylated protein was based on a mass cytometry cutoff of ≥10 counts. For single-cell correlation between two antigens, we calculated Spearman's rank correlation coefficient using arsinh-transformed data for each patient separately. To perform DREMI and DREVI analyses, we sampled up to 5,000 cells from each patient and assessed strengths of pairwise interaction on pooled samples using a MATLAB-based software, simpledremi (http://www.c2b2.columbia.edu/danapeerlab/html/dremi.html)[33].

Manual gating. Single cells were gated using Cytobank software (https://www.cytobank.org) based on event length and ^{191}Ir/^{193}Ir DNA content (to avoid debris and doublets) as described[47]. Following single-cell gating, live non-apoptotic cells were gated based on cleaved poly(ADP-ribose) polymerase (cPARP) and ^{195}Pt content[48]. Platelets and erythrocytes were excluded by gating on CD61 and CD235a, respectively. The remaining fraction was gated to exclude T cells

and myeloid cells on the basis of CD3e and of CD33 and CD16 expression, respectively. After further exclusion of CD38[high] plasma cells, the remaining fraction was defined as lineage-negative blasts (Lin− B[+]; see **Supplementary Fig. 1a** for gating). Further analysis was applied to this Lin− B[+] fraction unless otherwise noted.

Single-cell developmental classification. Lin− B[+] fraction from healthy human BM was gated into 15 developmental populations of normal B lymphopoiesis, mixed progenitors and mature non-B cell fractions, as shown in **Supplementary Figure 1b**. The distribution of each population was based on the expression of 11 B cell developmental proteins that were used for manual gating: CD34, CD38, CD127, CD24, terminal deoxynucleotidyl transferase (TdT), CD179a, CD179b, intracellular immunoglobulin heavy chain (IgHi), surface IgH (IgHs), CD19 and CD20. Prior to classification, each leukemia sample was normalized to control for batch effects (see subsection "Processing of mass cytometry data"). Lin− B[+] cells from each leukemia sample were then assigned to the most similar healthy fraction based on the shortest Mahalanobis distance among distances to all healthy developmental populations in these 11 dimensions. For stability in calculating the Mahalanobis distance, the covariance matrix Eigen values were set to ≥0.2. A cell was designated 'unclassified' if none of the distances were below the classification threshold (Mahalanobis distance = 11 based on the number of dimensions). These single-cell classification parameters were optimized via ten-fold cross-validation using cells with known population assignment from healthy BM donors. Optimization parameters included additional protein markers as well as cosine, Manhattan or Euclidian distance metrics. By using BM from adult healthy donors as a reference, we assumed that B lymphopoiesis proceeded through the same stages of development in both children and adults. The percentages of B cell developmental subpopulations in healthy samples are not relevant to the performance of the single-cell developmental classifier. As such, the gradual decline in the B cell output with age does not confound single-cell developmental classification under this assumption.

Relapse predictive modeling. To construct a predictive model of relapse, termed the DDPR, we allocated 54 patients with BCP-ALL, each with ≥3 years of follow-up data, into training (80% of patients, $n = 44$) and validation (20% of patients, $n = 10$) cohorts using a random assignment function that preserved proportions of cases and controls. A complete set of mass cytometry features available to DDPR is listed in **Supplementary Table 5**. It includes frequency of cells in each of five expanded populations (pre-pro-B, pro-BI, pro-BII, pre-BI and early progenitors), average expression of 24 proteins in these populations (CD10, CD19, CD20, CD22, CD24, CD34, CD38, CD43, CD45, CD58, CD79b, CD123, CD127, CD179a, CD179b, HLA-DR, IgHi, IgHs, IKAROS, Ki-67, PAX5, RAG1, TdT and TSLPr) and frequency of cells with each of nine activated signaling molecules (p4EBP1, pSTAT5, pPLC-γ2, pAKT, pSYK, pRPS6, pERK1–ERK2, pCREB and pIKAROS) in the unperturbed state and in response to each of four perturbations (BCR crosslinking, IL-7, TSLP or pervanadate treatment; as per **Supplementary Table 3**). Two clinical features available to DDPR (age and white blood cell count at diagnosis) are in **Supplementary Table 1**. Missing values were imputed as a median of all values. All features were scaled using their mean and s.d. within the training cohort, and these scaling parameters were also applied to the validation cohort. DDPR was built using elastic net[22], a regularized machine-learning approach that utilizes both L_1 and L_2 penalty types to prevent overfitting. To take advantage of the follow-up data in our cohort, relapse was modeled as a time-to-event outcome in a regularized Cox proportional-hazard model[23]. To estimate DDPR predictive performance, we

used pre-validation[24,25]. In a ten-fold cross-validation approach, we sequentially constructed an elastic-net-regularized Cox model using 90% of the training cohort samples and tested that model on the remaining 10% of the training cohort samples. We then used all of the training cohort samples to construct the final model. Feature-scaling parameters and final DDPR coefficients are listed in **Supplementary Table 7**. To test DDPR predictive performance, we applied the final model to the validation cohort. To assess performance, we reported an iAUC value[6] and a C-statistic[27] for censored time-to event data calculated using predicted relative risk (with reference to the sample average; RR) values within each cohort. To stratify patients into low- and high-DDPR risk groups, we selected a RR threshold value (0.9967) that was based on the optimal log-rank P value calculated from the fitted RR values within the training cohort. We then applied this threshold value to pre-validated (training cohort) and predicted (validation cohort) RR values to yield the final DDPR risk group assignments for all of the patients (**Supplementary Table 1**).

Statistical analysis. Data analysis was performed using R statistical software (http://www.r-project.org). To test statistical significance between two groups, we applied a two-tailed unpaired Student's t-test. When more than two groups were compared, we used Tukey's honest significance difference test; we included a Bonferroni correction for multiple comparisons. When equal variance assumption was not met based on an F test, a Welch's t-test was used instead of Student's t-test; normality was assessed using the Shapiro–Wilk test. A paired Welch's t-test was used for paired samples. The Kaplan–Meier method was used to estimate RFS rates; differences between groups were assessed using the log-rank test. DDPR RFS curves were built using predicted risk group assignment for each patient (i.e., based on pre-validated RR for the training cohort and predicted RR for the validation cohort). To test for synergy between current risk-stratification methods and DDPR, we calculated integrated discrimination improvement index (IDI), continuous net reclassification improvement (NRI) and median improvement (MI) for censored time-to-event data and their statistical significance at 5 years[31,32].

Life Sciences Reporting Summary. Further information on experimental design is available in the **Life Sciences Reporting Summary**.

Data availability. The mass cytometry data are available at https://github.com/kara-davis-lab/DDPR/releases.

45. Conter, V. *et al.* Molecular response to treatment redefines all prognostic factors in children and adolescents with B cell precursor acute lymphoblastic leukemia: results in 3,184 patients of the AIEOP-BFM ALL 2000 study. *Blood* **115**, 3206–3214 (2010).
46. WHO Expert Committee. Acute lymphoblastic leukemia in children: a brief review of the internationally available protocols. (Second Meeting of the Subcommittee of the Expert Committee on the Selection and Use of Essential Medicines) 1–11 (World Health Organization, Geneva, 2008).
47. Bendall, S.C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
48. Fienberg, H.G., Simonds, E.F., Fantl, W.J., Nolan, G.P. & Bodenmiller, B. A platinum-based covalent viability reagent for single-cell mass cytometry. *Cytometry A* **81**, 467–475 (2012).
49. Zunder, E.R. *et al.* Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. *Nat. Protoc.* **10**, 316–333 (2015).
50. Finck, R. *et al.* Normalization of mass cytometry data with bead standards. *Cytometry A* **83**, 483–494 (2013).

nature research

Corresponding author(s):   Nolan & Davis; NMED-A82533D

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.

For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

▶ Experimental design

1.   Sample size

Describe how sample size was determined.

> We did not predetermine a samples size as and a power calculation could not be performed, as this is the first study of its kind. Sample size was determined based on availability of samples with sufficient cellular material and annotated clinical metadata. For relapse prediction models, samples included in this analysis required at least 3 years of follow up from initial diagnosis.

Data exclusions

Describe any data exclusions.

> Patient samples that did not have at least 3 years of follow up data from the time of diagnosis were excluded from the DDPR relapse prediction modeling. These comprised UPN 60-69, as shown in Table S1 . These samples were included in the developmental classification.

Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

> The patient samples utilized in this work were analyzed in a series of batches. Each sample was divided into 5 treatment conditions as described in Table S3. Thus, each sample was replicated 5 times and responses across patients were reproducible as well as across batches of samples to enable generalization regarding per patient responses and across patient responses.

Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

> During the cross-validation construction of the DDPR model, samples were randomly assigned to training and validation cohorts using a random assignment in R that preserves proportions of cases vs. controls.

Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

> Investigators were not blinded to group allocation.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

> All analyses were performed using Cytobank (cytobank.org) and R. Single-cell protein expression data were extracted using Bioconductor software (www.bioconductor.org).DREMI is available for download at www.c2b2.columbia.edu/danapeerlab/html/dremi.html. The developmental classifier and DDPR code is available to editors and reviewers and will be made available upon request. FCS files are available on Github (https://github.com/kara-davis-lab/DDPR/releases).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

> Materials, outside of patient samples, utilized in this study are not unique and are available for distribution by a third party. All antibodies are commercially available.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

All antibodies were validated in human cells (cell lines or primary cells) known to our laboratory to be positive or negative controls for a given antibody target. Mass cytometry antibody reagents listed below in the following order:
Protein/ Clone/ Lot Number/ Manufacturer/ Metal Isotope/ Surface (S) or Intracellular(I) Stain

CD10/ HI10a/ 6155527/ Biolegend/ Gd156/ S
CD123/ 6H6/ B199259/ Biolegend/ Eu151/ S
CD127/ HCD127/ B173990/ Biolegend/ Dy162/ S
CD16/ 3G8/ B175991/ Biolegend/ Yb171/ S
CD179a/ HSL96/ B129864/ Biolegend/ Sm149/ I
CD179b/ HSL11/ B179047/ Biolegend/ Gd158/ I
CD19/ H1B19/ B157781/ Biolegend/ Nd142/ S
CD20/ 2H7/ B164952/ Biolegend/ Sm147/ S
CD22/ HIB22/ B165323/ Biolegend/ Nd143/ S
CD235/ HIR2/ B132247/ Biolegend/ In113/ S
CD24/ ML5/ B167884/ Biolegend/ Gd160/ S
CD3/ HIT3a/ B151232/ Biolegend/ Er170/ S
CD33/ HIM3-4/ B183522/ Biolegend/ Yb171/ S
CD34/ 8G12/ B163230/ Biolegend/ Nd148/ S
CD38/ HIT2/ B170151/ Biolegend/Er168/ S
CD43/ CD43-10G7/ B149905/ Biolegend/ Er167/ S
CD45/ HI30/ B159992/ Biolegend/ In115/ S
CD58/ TS2-9/ B145718/ Biolegend/ Tm169/ S
CD61/ VI-PL2/ B176028/ BD Biosciences/ In113/ S
CD79b/ CB3-1/ 4203934/ Biolegend/ Nd146/ S
HLA-DR/ L243 B161762/ Biolegend/ Yb174/ S
IgHi/ polyclonal/ 10689/ Invitrogen/ Eu153/ I
IgHs/ polyclonal/ 10689/ Invitrogen/ Lu175/ S
IKAROS/ D10E5 2/ Cell Signaling Technology/ Nd145/ I
IgL kappa/ MHK-49/ B162243/ Biolegend/ Sm154/ I
IgL lambda/ MHL-38/ B171739/ Biolegend/ SM154/ I
PAX5/ 1H9/ B178991/ eBioscience/ Ho165/ I
RAG1/ D36B3/ 3968BF/ Cell Signaling Technology/ Dy163/ I
TdT/ E17-1519/ 21361/ BD Biosciences/ Dy164/ I
CRLF2/ 1B4/ E028811/ eBioscience/ Dy161/ S
AKT (pS473)/ 193H12/ 4060BF/ Cell Signaling Technology/ Tb159/ I
4EBP1(pT37/T46)/ 236B4/ 18/ Cell Signaling Technology/ Nd144/ I
cPARP/ F21-852/ 5089576/ BD Biosciences/ La139/ I
CREB (pS133)/ 87G3/ 9198BF/ Cell Signaling Technology/ Yb176/ I
ERK1/2 (pT202/pY204)/ D13/ 4370BF/ Cell Signaling Technology/ Yb173/ I
Ki-67/ B56/ 3305519/ BD Biosciences/ Sm152/ I
IKAROS (pS63)/ STA9/ 1/ Epitomics (made for Nolan lab)/ Gd155/ I
PLCg2 (pY759)/ K86-689.37/ 26057/ BD Biosciences/ Pr141/ I
rpS6 (pS235/pS236)/ N7-548/ 4044686/ BD Biosciences/ Yb172/ I
STAT5 (pY694)/ 47/ 4044688/ BD Biosciences/ Nd150/ I
ZAP70/SYK (pY319/pY352)/ 17a/ 85582/ BD Biosciences/ Er166/ I

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used in the study.

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used in the study.

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used in the study.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No eukaryotic cell lines were used in the study.

▶ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

No animals were used in the study.

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Clinical data were available for the samples used in this study, including minimal residual disease (MRD) risk group and final risk assignment as per AIEOP-BFM ALL 2000 protocol (ClinicalTrials.gov identifier: NCT00613457), diagnostic cytogenetics, age at diagnosis, gender, white blood cell count at diagnosis, date of diagnosis, date of relapse, and date of the last follow-up.