Chapter 5

Recurrent patterns of protein expression signatures in pediatric acute lymphoblastic leukemia: recognition and therapeutic guidance

Hoff FW1,2,*, Hu CW3, Qiu Y1, Ligeralde A1, Yoo SY4, Scheurer ME5, de Bont ESJM2, Qutub AA3, Kornblau SM1,2, Horton TM6 *

1. Department of Leukemia, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA
2. Department of Pediatric Oncology/Hematology, Beatrix Children’s Hospital, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
3. Department of Bioengineering, Rice University, Houston, TX, USA
4. Department of Bioinformatics and Computational Biology, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA
5. Department of Pediatrics and Department of Epidemiology, Texas Children’s Cancer and Hematology Centers, Baylor College of Medicine, Houston TX.
6. Department of Pediatrics, Baylor College of Medicine/Dan L. Duncan Cancer Center and Texas Children’s Cancer Center, Houston, TX, USA

*Contributed equally to the work

Published in: Molecular Cancer Research 2018; 16(8):1263-1274
DOI: 10.1158/1541-7786.MCR-17-0730
Abstract

Pediatric acute lymphoblastic leukemia (ALL) is the most common pediatric malignancy, and the second leading cause of pediatric cancer death in developed countries. While the cure rate for newly-diagnosed ALL is excellent, the genetic heterogeneity and chemoresistance of leukemia cells at relapse makes individualized curative treatment plans difficult. We hypothesize that genetic events would coalesce into a finite number of protein signatures that could guide the design of individualized therapy. Custom reverse phase protein arrays were produced from pediatric ALL \( (n = 73) \) and normal CD34\(^+ \) \( (n = 10) \) samples with 194 validated antibodies. Proteins were allocated into 31 Protein Functional Groups (PFG) to analyze them in the context of other proteins, based on known associations from the literature. The optimal number of protein clusters was determined for each PFG. Protein networks showed distinct transition states, revealing “normal-like” and “leukemia-specific” protein patterns. Block clustering identified strong co-correlation between various protein clusters that formed ten “Protein Constellations” (CON). Patients that expressed similar recurrent combinations of CONs compiled seven distinct “Protein Expression Signatures” (SIG), correlating with risk stratification, cytogenetics and laboratory features. Most CONs and SIGs were specific for T-cell ALL or pre-B-cell ALL, however some CONs showed significant overlap. Several SIGs were associated with Hispanic ethnicity, suggesting that ethnic pathophysiological differences likely exist. Additionally, some CONs were enriched for “normal-like” protein clusters, whereas others had exclusively “leukemia-specific” patterns.

Implications: Recognition of proteins that have universally altered expression, together with proteins that are specific for a given SIG, suggests targets for directed combinatorial inhibition or replacement to enable personalized therapy.
Introduction

Pediatric acute lymphoblastic leukemia (ALL) is the most common form of cancer in children accounting for approximately 25% of all childhood malignancies. Despite dramatic improvements in outcome over the past few decades, with 5-year survival rates approaching 90%, relapsed ALL remains one of the leading causes of pediatric cancer mortality and morbidity. To improve therapeutic outcome in high-risk patients and relapsed ALL, we need an improved understanding of individual molecular pathophysiology. Defining what signaling pathways and regulatory network dependencies are crucial to driving the underlying malignancy would facilitate the use of targeted therapies on an individualized basis.

High-throughput next-generation sequencing has led to an advanced understanding of the genetic heterogeneity of pediatric ALL; this in turn has led to a focus on novel therapies that target frequently mutated candidate genes. This research has revealed multiple recurrent genetic alterations, involving genes involved in lymphoid development, cell-cycle regulation, tumor-suppression, apoptosis, lymphoid signaling, and transcriptional regulation. However, with the exception of the BCR-ABL tyrosine kinase inhibitors, most recurrent genetic events identified to date lack therapeutic agents that specifically target the mutated proteins resulting from these genetic mutations. Furthermore, those genetic and epigenetic changes occur in a near infinite number of combinations and the physiological consequences of combinatorial genetic mutations are largely undefined. This genetic heterogeneity makes personalized rational treatment combinations challenging.

Since the molecular consequences of genetic and epigenetic events are predominantly mediated by the altered expression and function of proteins, we hypothesize that genetic heterogeneity coalesces into a more finite number of protein expression patterns, and that these protein expression patterns reveal key protein dependencies that could identify therapeutic targets. Gene expression profiling (GEP) has revealed recurrent patterns of gene expression, but has the limitation that messenger RNA transcript expression correlates with protein abundance for less than 50% of genes. GEP also does not reflect post-translation modifications (PTM) and protein activation states. Since proteins function in networks and functionally related pathways, rather than on individual basis, we further hypothesize that analyzing proteins using a network-based approach should identify crucial recurrent protein expression patterns that define subpopulations of pediatric ALL. We therefore set out to define unique protein expression
patterns across pediatric ALL patients with the goal of informing risk classification and suggesting novel combinational therapy.

Methods

Patient population
Peripheral blood (PB) mononuclear cells were collected from 73 ALL patients (67 newly diagnosed and 6 relapsed pediatric ALL) that were evaluated at the Texas Children’s Hospital (TXCH) between July 2010 and June 2015. Samples were collected prior to induction therapy and in accordance with Institutional Review Board (IRB)-approved policies. Informed written consent was obtained in accordance with the Declaration of Helsinki, and applicable local and state laws. Demographics are described in Table 1. Sixteen patients were diagnosed with T-cell ALL and 57 with pre-B cell ALL. A high percentage were of Hispanic ethnicity (n = 45/73, 62%). Single-nucleotide polymorphisms (SNP) were determined for 54 patients to verify their genetic ancestry. Patients were stratified into risk groups according to the Children’s Oncology Group (COG)\textsuperscript{12} and were treated under a variety of COG protocols (Table S1). All but 6 patients achieved complete remission (CR), and only four relapsed. Sixteen patients underwent stem cell transplantation and 63 (86%) were alive at the end of follow-up (28 to 350 weeks). Mutation analysis was restricted to that performed as part of routine clinical care and included analysis of MLL, CDKN2A, IgH, TCF3, ETV6 and RUNX1. This mutation information was available for all but two patients.

Table 1. Patient characteristics (n = 73)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Number (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cases</td>
<td>73 (100%)</td>
</tr>
<tr>
<td>ALL subtype</td>
<td></td>
</tr>
<tr>
<td>Pre-B ALL</td>
<td>57 (78%)</td>
</tr>
<tr>
<td>T-ALL</td>
<td>16 (22%)</td>
</tr>
<tr>
<td>Age (yr.) Median (range)</td>
<td>7.3 (0.2-18.0)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>40 (55%)</td>
</tr>
<tr>
<td>Declared Ethnicity</td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>61 (84%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>44 (72%)</td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>17 (28%)</td>
</tr>
<tr>
<td>Black American</td>
<td>6 (8%)</td>
</tr>
<tr>
<td>Asian</td>
<td>4 (5%)</td>
</tr>
<tr>
<td>Mixed</td>
<td>2 (3%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>“Single nucleotide polymorphism” ethnicity</td>
<td></td>
</tr>
<tr>
<td>European</td>
<td>9 (12%)</td>
</tr>
</tbody>
</table>
108 | Chapter 5

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>African</td>
<td>6 (8%)</td>
</tr>
<tr>
<td>American Indian</td>
<td>38 (52%)</td>
</tr>
<tr>
<td>Asian</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>Not done</td>
<td>19 (26%)</td>
</tr>
</tbody>
</table>

**Cytogenetics**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Favorable</td>
<td>15 (21%)</td>
</tr>
<tr>
<td>Intermediate</td>
<td>42 (58%)</td>
</tr>
<tr>
<td>Unfavorable</td>
<td>15 (21%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>1 (1%)</td>
</tr>
</tbody>
</table>

**Risk group**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Low risk</td>
<td>4 (5%)</td>
</tr>
<tr>
<td>Standard/intermediate risk</td>
<td>29 (40%)</td>
</tr>
<tr>
<td>High/very high risk</td>
<td>40 (55%)</td>
</tr>
</tbody>
</table>

**CNS status**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CNS-1</td>
<td>46 (63%)</td>
</tr>
<tr>
<td>CNS-2</td>
<td>20 (27%)</td>
</tr>
<tr>
<td>CNS-3</td>
<td>6 (8%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>2 (3%)</td>
</tr>
</tbody>
</table>

**Response**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete remission</td>
<td>67 (92%)</td>
</tr>
<tr>
<td>Resistant</td>
<td>4 (5%)</td>
</tr>
<tr>
<td>Fail</td>
<td>2 (3%)</td>
</tr>
</tbody>
</table>

**Status**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Alive</td>
<td>63 (86%)</td>
</tr>
</tbody>
</table>

**NOTE:** Cytogenetic aberrations were classified into favorable, intermediate, and unfavorable cytogenetics. Favorable: hyperdiploid, diploid and t(12:21) EV76/RUNX1 translocation; unfavorable: 11q23 rearrangement, hypodiploid, t(9:22) BCR/ABL1 translocation, 5q deletion. Patients that were not classified as favorable or unfavorable were defined as having intermediate cytogenetics. Central nervous system (CNS) involvement was categorized into three groups according to the COG standard. CNS-1: no blasts in the cerebrospinal fluid (CSF), CNS-2: <5% blasts in the CSF with or without red blood cells, CNS-3: >5% blasts in CSF. Risk group stratification was done according to the AALL protocols.

**Reverse phase protein arrays (RPPA) methodology**

The antibody based high-throughput RPPA methodology was performed on 73 samples from pediatric patients with ALL, ten cryopreserved CD34+ normal bone marrow (BM) samples (AllCells, Almeda CA) and 95 leukemic cell lines samples. RPPA was performed at MD Anderson Cancer Center (Houston, TX). Cell lines were obtained from the ATCC and different laboratories and were tested for mycoplasma using the mycoplasma PCR detection kit (Applied Biological Materials Inc. Catalog No. G238). Patient samples were processed into RPPA lysates on the day of collection and no samples were prepared from cryopreserved cells. The methodology and validation of the technique are fully described in previous publications.13-15

Briefly, the whole cell lysate protein preparations were made from the mononuclear cell fraction of ficolled PB and normalized to a concentration of 1x10⁴ cells/μL. Samples were printed in five serial (1:2) dilutions onto slides along with normalization and expression controls. Slides were probed with 194 strictly validated primary antibodies and a secondary antibody to amplify the signal, and finally a stable dye was precipitated using the chromogen DAB (Agilent, Santa Clara, CA).16 This included antibodies against 149 different proteins
along with 36 antibodies targeting phosphorylation sites, six targeting cleaved forms of Caspase, NOTCH1 and PARP1, and three targeting Histone methylation sites. A “Rosetta Stone” table of manufacturer, antibody name, and primary and secondary antibody dilution can be found in Table S2. The stained slides were analyzed using Microvigene® software version 3.0 (Vigene Tech, Carlisle, MA) to produce quantified data.

**Nomenclature protein and antibody names**

Since neither the HUGO\textsuperscript{17}, HUPO\textsuperscript{18} or MiMI\textsuperscript{19} naming systems account for PTM, we used a nomenclature in which the HUGO gene symbol is followed by a period, then the type of PTM, “p” for phosphorylated, “cl” for cleaved or “Me” for methylation, followed by the letter code for the affected amino acid and its sequence position. For example, AKT1\_2\_3pThr\textsuperscript{308} is AKT phosphorylated on Threonine at position 308. Placing the PTM after the protein name enables alphabetical sorting and inclusion of the affected site.

**Data normalization and processing**

*SuperCurve* algorithms were used to generate a single value from the five serial dilutions.\textsuperscript{20} Loading controls\textsuperscript{21} and topographical normalization\textsuperscript{22} procedures were performed to account for protein concentration and background staining variations. Since all samples had replicates, the average expression level of the replicates was used as a single expression level. All protein expression levels were shifted relative to the median of the normal CD34\textsuperscript{+} BM samples.

**Computational analysis**

The computational analysis was done using the “MetaGalaxy” analysis (Figure S1), because we had previously seen in adult acute myeloid leukemia (AML) that this approach, which analyzes proteins in the context of functionally related proteins, obtained more clinically interesting patient groups compared to the traditional approaches.\textsuperscript{23} In contrast to the traditional unsupervised hierarchical clustering that ignores all the known relationships between proteins, and has the additional disadvantage of weighing each component equally, we first divided the 194 proteins in 31 functionally related protein groups, defined as a “Protein Functional Group” (PFG). This allocation into functional related groups was done based on their known function or pathway membership from the existing literature or based on strong associations within the dataset (e.g., BRCA2 to the “Cell Cycle” PFG and DDX17 to the “Ribosome” PFG). Because proteins could have multiple functions, proteins could belong to multiple PFG. The proteins involved in each PFG are listed in Table S3.
To identify if subsets of cases with similar (correlated) expression of core member proteins within each PFG did exist, a combination of Progeny Clustering\(^{24}\) (a bootstrapping and stability-based method for selecting cluster numbers) in combination with k-means\(^{25}\) (generation of protein clusters) was used. Subsets of patients were identified based on their relative protein expression similarities (i.e., Euclidean distance), which were defined as a “Protein Cluster” (PC). The optimal number of PCs for each PFG was determined using the clustering solution stability scores. For some PFG an alternative number of clusters was chosen or small clusters were merged into the closest group to make more biologically relevant clusters. Linear discriminant analysis was performed to determine which of the PCs was statistically most similar to the normal CD34\(^+\) samples.\(^{26}\) This PC was then set as cluster 1 and was positioned to the far left. Principal component analysis (PCA)\(^{27}\) was used to visualize the distribution of the protein clusters relative to that of normal CD34\(^+\) BM samples. Associations between PCs and clinical/laboratory features were assessed using the Fisher’s exact test for categorical variables and the Kruskal-Wallis test by ranks for continuous variables. Survival curves were generated using the Kaplan-Meier method. Protein networks were constructed from known protein associations that were obtained from the STRING literature database (combined score > 0.9)\(^{28}\) in combination with computationally reconstructed interactions from RPPA data using graphical lasso\(^{29}\) and StARS\(^{30}\) (for model selection based on stability). Since the STRING database does not consider post-translational modifications, the protein names were used to query literature-based interactions for PTM sites.

Next, we rebuilt the overall picture by combining the individual protein clusters into one binary matrix to assess whether we could identify patterns of PCs from various PFG that recurrently co-occurred together. This matrix indicated the PC-membership for each patient in all PFG; “1” if a patient was a member of that PC, “0” if not a member. Block clustering\(^{31}\) was performed to search for strong recurrent correlations between protein clusters from various PFG that were defined as a “Protein Constellation” (CON). A group of patients that expressed similar patterns of CONs was defined as a “Protein Expression Signature” (SIG). The optimal number of CONs, that formed SIGs, was obtained by selecting the combination that generated the largest sum of the squared difference between the expected and observed values, divided by the expected value. The expected value was defined as the product of PC-membership within that CON, divided by the frequency of patients that fell within a given SIG. Correlations between SIGs and clinical features/outcomes were assessed similarly as for the individual PFG. Lists of proteins that were significantly over or under expressed relative to the normal CD34\(^+\)
cells were generated for each CON and SIG using the Wilcoxon signed-rank test with an FDR-adjusted $p$-value ($p < 0.01$). The most discriminative proteins that allow classification into the SIGs were selected using Random Forest. All the statistical tests and plots were generated in R (Version 0.99.484 – © 2009-2015 RStudio, Inc.). Networks were generated in Cytoscape (Version 3.3.0).

Results

Existence of “normal-like” and “leukemia-specific” protein patterns
To characterize heterogeneity in protein expression between pediatric ALL patients we started our analysis with evaluating proteins in the context of their own PFG. Therefore, the progeny clustering algorithm was applied that enabled identification of an optimal number of PCs: a subset of cases with similar (correlated) expression of core protein components of a PFG. The number of PCs ranged from 3 to 5 clusters for each PFG (Figure 1A). The measure of cluster stability was based on the co-occurrence probability matrix (Figure S2). Overall, clusters showed high stability and reproducibility with scores of 0.6 to 0.9. No confounding variables that affected the clustering analysis were found for the processing time of the samples, or for the date of collection (Figure S3). Next, PCA was performed to determine if the PCs were similar to the normal CD34+ samples, or if it was sufficiently dissimilar to be specific to a leukemic state, based on their graphic distribution on the PCA plot in comparison to the normal CD34+ samples. Most PFG ($n = 23$) had at least one cluster with expression similar to the normal CD34+ samples (Checkered pattern Figure 1A). In contrast, leukemia-specific clusters, lacking overlap with the normal CD34+ cells, were observed for all 31 PFG (Solid fill Figure 1A) with eight PFG (Cell Cycle, Differentiation, MEK, PKC, STP upstream, T-cell, Transcription and WNT-signaling) having only leukemia-specific clusters. For 8 of the PFG we could identify more than one “normal-like” pattern.

Protein functional groups reveal different protein activity states
To visualize interactions between core member proteins of each PFG and other probed proteins in the dataset, protein networks were generated. Networks were built by integrating previously known protein interactions from the literature and strong correlations in the dataset. The median expression for each protein cluster was then calculated relative to the normal CD34+ cells and overlaid onto the networks to reveal the overall differences in expression and activation associated with each protein cluster. For instance, for the “Friend Leukemia Virus
Integration 1” (FL11) that was formed by five core protein members (FLI1, NCL, NPM1, STMN1, and WTAP), we were able to recognize five distinct protein clusters (PC-1, PC-2, PC-3, PC-4 and PC-5) (Figure 2A). By convention, the protein expression levels in PC-1 were statistically the closest to normal and showed most proteins with expression similar to the normal CD34⁺ samples (Figure 2B). The greatest variation between the protein clusters was observed in the expression of key protein STMN1; progressively higher expression in PC-2 and PC-4 and increasingly lower expression in PC-3 and PC-5 (Figure 2C). Another example that showed the concept of different transition states was the “Apoptosis Occurring” PFG. Here we could recognize three different protein clusters (PC-1, PC-2 and PC-3) (Figures available online, http://leukemiaproteinarlas.org/pediatric-all/ApopOccur/). Increased evidence of apoptosis activation, in the form of cleavage of Parp1, Caspase 3 and Caspase 7 was evident in PC-2 and PC-3, representing two apoptosis “on”-states.

**Figure 1.** The optimal number of protein clusters for all protein functional groups. (A) The optimal number of protein clusters that was identified for each of the 31 PFG is illustrated. Protein patterns that showed sufficient overlap with the normal CD34⁺ samples on the PCA plot were assigned as "normal-like" protein clusters and are shown as checkered boxes. Protein clusters that were sufficiently dissimilar from the normal CD34⁺ were assigned as "leukemia-specific" and are shown as solid boxes. (B) Representation of pediatric ALL protein clusters that were mimicked by at least one of the leukemic
cell lines. Green ticks indicate that a protein cluster had a cell line with a protein expression pattern equivalent. The red crosses indicate that none of the cell lines were found to express a comparable protein expression pattern.

**CONs express recurrent patterns of protein expression**

Although traditional approaches that cluster patients directly by taking all proteins together with unsupervised hierarchical clustering could clearly separate pediatric ALL patients from the healthy subjects (Figure S4), we supposed that we would find better, more robust patterning within the pediatric ALL samples, if we created smaller subsets of proteins based on known functional relationships and then built up the overall picture from these individual building blocks (Figure S5). Therefore, we developed a novel approach that defined patient SIGs by looking for recurrences in expression patterns within PFG and from there built higher order structures by performing hierarchical clustering of those smaller patterns.

As described, PCs were first defined within each PFG, which resulted in a total of 114 protein clusters for the 31 PFG. As each patient was represented by one of the PCs of each individual PFG, each patient was a member of 31 out of the 114 protein clusters. Secondly, all PCs were compiled into a single binary matrix, which we called a “MetaGalaxy” (Figure 3A). Block clustering was conducted to search for recurrent associations between various PCs, which were defined as a CON (horizontally in Figure 3A). Patients that showed recurrent patterns of CONs were together defined as SIG (vertically in Figure 3A). An optimization calculation was performed to determine the optimal number of CONs and SIGs. This was determined by selecting the matrix where the squared sum of the difference between the observed and expected values of each combination of CONs and SIGs, divided by the expected value, was maximal. This suggested the presence of ten CONs and seven SIGs. For instance, CON-4 that was horizontally formed by 16 PCs was strongly associated with patients that formed SIG-7. The expected occurrence in this CON was 9% for each SIG, based on the presence of 107 single blue points that indicated PC-membership out of 1168 (CON-4; 107 single blue points vs. a potential of 16 x 73 patients = 1168, 107/1168 = 0.09). Within this CON, SIG-7 showed an observed occurrence significantly above expectation of 94% (64 of the 68 points were blue, vs. an expected number of 0.09 x 68 = 6). In contrast, none of the patients in SIG-2 had a membership for any of the PCs within this CON (0/128 blue points) \( (p < 0.001) \). Likewise, CON-9 that was formed by three PCs clusters had an expected presence rate of 62% (135 blue points vs. a potential of 219 points (3 x 73 patients)). Within this CON, SIG-1 had an observed
Figure 2. Relative protein expression levels for the proteins involved in "Friend Leukemia Virus Integration 1" (FLI1) PFG. (A) This heatmap shows the relative protein expression levels for the five core member proteins of the "FLI1" PFG: STMN1, FLI1, NPM1, WTAP and NCL. The Progeny Clustering algorithm (coupled with k-means) was performed and identified an optimal number of five PCs. The colors reflect the median expression levels relative to the normal CD34+ samples. Proteins expressed greatly below normal are shown as dark blue, and proteins expressed significantly above
normal are shown in dark red (maroon). Proteins within the range of the normal cells are colorized in green. Each column represents a single patient. The annotation bar shows patient membership for the different ALL subtypes [pre-B cell (yellow) and T-cell (magenta)] and for the five defined PCs [PC-1 (red), PC-2 (magenta), PC-3 (yellow), PC-4 (light green) and PC-5 (dark green)]. (B) Principal component analysis (PCA) visualized the global distribution of the patients in their assigned protein cluster relative to the normal CD34+ samples. From the PCA partial similarity between normal CD34+ cells [black*] and PC-1 [red*] was observed, while the leukemia specificity of PC-2 [magenta*], PC-3 [yellow*], PC-4 [light green*] and PC-5 [dark green*] was demonstrated by the lack of overlap with the normal CD34+ cells. Each plotted dot represents one patient. (C) Protein networks show interactions between the five core protein members (large nodes) and associated proteins (small nodes). Colors reflect the relative median protein expression within that PC; ranged from high (maroon) to low (dark blue). Dotted ( . . ) lines indicate known associations from the literature, dashed lines (- - -) indicate interactions based on strong correlation in the dataset and solid lines ( — ) indicate interactions both seen in the literature and our dataset. Arrows show transition from the most normal state PC-1 to the more "on"-states PC-2 and PC-4 and the more "off" states in PC-3 and PC-5 relative to PC-1.

presence rate above expected of 94% (34 vs. 36 (3 x 12) blue points) and SIG-7 had an observed presence rate below expected of 0% (0 vs. a potential of 12 (3 x 4) points) ($p < 0.001$). A list of the protein clusters in each CON is shown in Table S4. An example of the optimization calculation is shown in Figure S6.

Most CONs were associated with a single ALL subtype, with CON-3 and CON-5 only being found in T-cell ALL and CON-2, 4, 6, 8 and 10 being exclusive to pre-B cell ALL. However, CON-1 and CON-9 showed some overlap between pre-B cell ALL and T-cell ALL, suggesting shared protein deregulation. A clear distinction was observed between the T-cell specific SIG-1 ($n = 12/12$, 100%), the pre-B ALL dominant SIG-2, 3 and 4 ($n = 19/23$, 83%) and pre-B ALL exclusive SIG-5, 6 and 7 ($n = 38/38$, 100%). Because the majority of T-cell ALL cases were within SIG-1, we conducted a separate analysis of only T-cell ALL samples. As shown in Figure 3B, we observed three T-cell SIGs based on 6 CONs. A list of the PCs in each CON is shown in Table S5. A similar analysis was performed using only B-cell ALL cases, but this was not different from what was seen in SIG-2 to SIG-8 (Figure S7/ Table S6).

Because we identified PCs that showed sufficient overlap with our healthy CD34+ cells to be defined as normal-like PCs, we were then interested in whether CONs were enriched or depleted for those clusters. Interestingly, we found CONs that showed enrichment for normal-like patterns (CON-1, 8, 9, and 10) and CONs that had exclusively leukemia-specific patterns (CON-2, 3, and 5) ($p = 0.011$).
**Protein patterns correlate with outcomes and clinical and laboratory features**

Typical of pediatric ALL, this cohort was characterized by a high CR rate \((n = 67, 93\%)\) and a low therapy resistance rate \((n = 4, 5\%)\) that in combination with a low relapse rate \((n = 4, 5\%)\) resulted in a high survival \((n = 63, 86\%)\). Given the paucity of events, SIGs do not show statistically significant correlation with overall survival (OS) (Figure 4A) or disease-free survival (DFS) (Figure 4B), which was defined as having an event (relapse or death) post induction that led to CR. However, three out of the four relapse cases were within SIG-6. Univariate Cox Proportional-Hazard analysis showed no other relationships between the survival probability and any of the collected patients features (Table S7).

On the other hand, SIGs were significantly associated with patient demographics and laboratory variables (Table 2). Favorable cytogenetics were overrepresented in SIG-5 and SIG-7, and intermediate cytogenetics were overrepresented in SIG-1 and SIG-4 \((p = 0.017)\). No associations were seen for single cytogenetic abnormalities, such as the frequently harbored 11q23 rearrangement. This lack of association with specific cytogenetic types again highlights the large heterogeneity among ALL patients and is likely due to various combinations of mutations. As expected for the T-cell ALL SIG, CDKN2A was highly mutated \((n = 9/12, 75\%)\) compared to the overall mutation rate \((n = 20/64, 31\%)\) \((p = 0.007)\).\(^{34}\) A low CDKN2A mutation frequency was observed for SIG-2 \((n = 1/7, 14\%)\), SIG-4 \((n = 1/9, 10\%)\) and SIG-7 \((n = 0/2, 0\%)\).

**Protein SIGs are associated with Hispanic ethnicity**

In numerous studies Hispanic patients with pediatric acute leukemia have fared worse than Caucasians,\(^{35-38}\) but whether this arises from different underlying biology, or is related to socioeconomic factors is currently unknown. Our study population, drawn from an area in Southern Texas, was enriched for Hispanic patients. Pre-B cell ALL SIG-3, 5 and 7 showed a similar proportion of Hispanic patients compared to the overall population \((n = 45/73, 62\%)\). However, both SIG-2 \((n = 6/8, 75\%)\) and SIG-6 \((n = 17/20, 85\%)\) were enriched for Hispanic patients, whereas Hispanic patients were underrepresented in SIG-4 \((n = 3/9, 33\%)\) \((p = 0.021, df = 2)\). This imbalance in ethnic compositions was even stronger after verification by genetic ancestry mapping using SNP; two of the non-Hispanic patients in SIG-2 were of African descent, also having inferior disease outcome.\(^{39}\) Two of the self-reported non-Hispanics in SIG-6 were actually Hispanics by SNP and one patient stating Hispanic ancestry was Asian by SNP typing. This brings the total number of Hispanics in SIG-6 to 18 \((n = 18/20, 90\%)\). SIG-4 had
Figure 3. "MetaGalaxy" analysis identifies strong correlation between PCs from various PFGs. (A) Block clustering identified the existence of ten CONs (horizontally) and seven SIGs (vertically). Each column represents a single patient and is positive (blue) for 31 out of the 114 protein clusters. Each row represents a single PFG-PC. The annotation bar shows a clear division in ALL type [pre-T cell (magenta), pre-B cell (yellow)], and shows the patient characteristics including: gender, age, treatment risk group, CNS status, cytogenetics, declared ethnicity, SNP verified ethnicity, CDKN2A mutation status, relapse, achievement of complete remission and the vital status of the patient. (B) Block clustering limited to the T-cell ALL samples enabled recognition of six CONs (horizontally) and three SIGs (vertically). The annotation bar shows patient characteristics for ethnicity and suggests ethnicity-associated CONs.
Figure 4. Survival analysis for the identified ALL SIGs and the protein summary plot for the CONs associated with the Hispanic ethnicity within the T-ALL patients. The Kaplan-Meier curves for overall survival (A) and disease-free survival (DFS) (B) were generated according to the seven ALL SIGs. Line colors match with the colored annotation bar on the "MetaGalaxy". (C) Proteins with significantly higher or significantly lower protein expression levels relative to normal CD34+ cells within T-ALL CON-3 (enriched for Hispanic ethnicity) and CON-5 (enriched for non-Hispanic ethnicity) are shown. Proteins in CON-3 were predominantly involved in PFG "Cell Cycle", "FLI1" and "IAP-Apoptosis" and proteins in CON-5 were involved PFG "BH3 Apoptosis", "Cell Cycle", "FLI1", "IAP-Apoptosis" and "MEK". Colors reflect the relative median expression within that specific CON, ranged from the lowest (dark blue) to relatively normal (cyan-green-yellow) to the highest (maroon) expression.

one additional Hispanic patient by SNP determination, making that SIG less imbalanced (n = 4/9, 44%).

When T-cell ALL cases were considered separately, three SIGs were present with most of the Hispanic cases in SIG1 and SIG-2 (n = 6/8, 75%) and only two cases in SIG-3 (n = 3/8, 38% following SNP analysis). CON-1, 2, 4 and 6 were unassociated with ethnicity, while CON-3 was found in the Hispanic cases and CON-5 was strongly present in the non-Hispanic cases. Notably, both CON-3 and CON-5 contained PCs from the PFG “Cell Cycle”, “FLI1” and “IAP-Apoptosis”. Expression summary plots are shown in Figure 4C. The Hispanic-associated CON-3 lacked the up-regulation of CCND3, DUSP6, RB1, RB1-pSer307,511 and STMN1 seen in the
non-Hispanic CON-5, but had up-regulation of unphosphorylated FKHRL1 (FOXO3). CON-3 showed higher levels of anti-apoptosis proteins, including XIAP and BIRC5 and lacked the suppressed expression of BCL2 and DIABLO.

**Proteomics to predict potential protein leads for targeted therapy**

Because most potential drugs target proteins, we generated lists of potential druggable targets for each SIG and CON (Figure S8, *figures available online*, http://leukemiaproteinatlas.org/pediatric-all/global/). These potential target leads were identified as being significantly over and under expressed relative to normal CD34+ cells. Figure 5A shows an example of all differentially expressed proteins when compared to the controls in SIG-6 that comprised 3 out of the 4 relapse patients. For example, proteins PARP1 and cleaved PARP1 together with LEF1, PIK3CA and BRAF were all higher expressed. From these lists, we could then reveal proteins that were universally changed in the same direction in at least 6 of the 7 SIGs (Figure 5B). Hypothetically, rational combinations of targeted therapies directed against SIG-specific proteins together with targeted therapies directed against universally altered expressed proteins could be used therapeutically in specific subsets of patients alone, or in addition to, standard therapy to overcome treatment resistance. However, this hypothesis needs validation with future experiments.

**Selection of discriminative proteins to aid in risk stratification and determine therapy**

In order to classify patients into one of the 7 SIGs based on a limited number of proteins, Random Forest was utilized to select the proteins with the highest distinctiveness (Figure S9). This resulted in a correct classification rate of 78% (*n* = 57/73), whereby variation in protein expression enabled a higher than overall classification accuracy for SIG-1 (*n* = 12/12, 100%), 5 (*n* = 12/14, 86%), 6 (*n* = 17/20, 85%) and 7 (*n* = 4/4, 100%). For instance, patients in SIG-1 could be separated based on their relatively low CDKN1A in combination with their relatively high GATA1 and NOTCH3, and SIG-7 could be discerned based on their low CASP3 levels. Inferior classification capability was found for SIG-2 (*n* = 4/8, 50%) and 3 (*n* = 1/6, 17%), which may be explained given that neither SIG-2 nor SIG-3 was exclusively associated with any CON and that none of the most discriminative proteins were significantly different compared to the other SIGs.
Leukemic cell lines only partially mimic protein patterns

Leukemic cell lines are frequently used to investigate the pathobiology of leukemia, but immortalization and cryopreservation of those cells likely alter the biology of the cell from their leukemic patient cell of origin. To determine if cell lines express differences or similarities in protein expression patterns compared to the pediatric ALL patient samples, we generated a new RPPA with 95 leukemic cell line samples, including cell lines derived from pediatric and adult ALL (e.g., Jurkat, REH), and AML patients (e.g., Kasumi-1, HL-60, Molm13, Molm14, OCIAML3). Arrays were probed with 235 antibodies of which 163 ($n = 163/194$, 84%) overlapped with the antibodies used on the pediatric patient array. Because the cell line and the pediatric acute leukemia patient array both had cells from healthy donors included, alignment of the control CD34+ samples from both enabled comparison of the arrays.

Overall, unsupervised hierarchical clustering and PCA clearly demonstrated completely distinct proteomic profiles for pediatric ALL patient samples and leukemic cell lines (Figure S10). Individual comparison of PCs showed that only 53 out of the 114 (46.5%) PCs had at least one cell line equivalent (Figure 1B/ Table S8). None of the CONs or SIGs seen in the ALL patients was replicated in the cell lines.

Pediatric leukemia online portal

In addition to the PFG “FLII” that is discussed in this paper, results from all PFG analyses are published online and can be assessed at: http://leukemiaproteinatlas.org/pediatric-all/.

Discussion

Heterogeneity within the genetic and epigenetic landscape of pediatric ALL makes personalized medicine challenging. To assist in the process of both risk stratification and medication management, we have demonstrated that pediatric ALL could be characterized by the “MetaGalaxy” approach into a finite number of recurrent proteins expression patterns that could identify key protein targets based on individual protein expression.

The MetaGalaxy analysis is a two-step approach that starts with the analysis of proteins in the context of other proteins that are known to be functionally related or known to interact with each other, and then globally searches for protein patterns that frequently co-occur. This
Table 2. Demographics and laboratory features for seven identified SIGs

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Variable</th>
<th>Total</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (%)</td>
<td>100%</td>
<td>12%</td>
<td>27%</td>
<td>19%</td>
<td>11%</td>
<td>16%</td>
<td>8%</td>
<td>5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALL type (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-B</td>
<td>78%</td>
<td>0%</td>
<td>75%</td>
<td>83%</td>
<td>89%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>0.000</td>
</tr>
<tr>
<td>T-cell</td>
<td>22%</td>
<td>100%</td>
<td>25%</td>
<td>17%</td>
<td>11%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White blood cell count (x k/μL)</td>
<td>Median</td>
<td>31</td>
<td>397</td>
<td>59</td>
<td>118</td>
<td>12</td>
<td>21</td>
<td>66</td>
<td>7</td>
<td>0.000</td>
</tr>
<tr>
<td>Peripheral blood absolute blast (k/μL)</td>
<td>Median</td>
<td>25</td>
<td>319</td>
<td>41</td>
<td>99</td>
<td>9</td>
<td>12</td>
<td>56</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>Peripheral blast (%)</td>
<td>Median</td>
<td>31</td>
<td>397</td>
<td>59</td>
<td>118</td>
<td>12</td>
<td>21</td>
<td>66</td>
<td>7</td>
<td>0.000</td>
</tr>
<tr>
<td>Lactate dehydrogenase (x U/L)</td>
<td>Median</td>
<td>2069</td>
<td>9914</td>
<td>4702</td>
<td>3609</td>
<td>2049</td>
<td>1195</td>
<td>1500</td>
<td>965</td>
<td>0.000</td>
</tr>
<tr>
<td>Bilirubin (mg/dl)</td>
<td>Median</td>
<td>0.3</td>
<td>0.6</td>
<td>0.3</td>
<td>0.2</td>
<td>0.5</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.004</td>
</tr>
<tr>
<td>Fibrinogen (mg/dl)</td>
<td>Median</td>
<td>332</td>
<td>191</td>
<td>330</td>
<td>267</td>
<td>495</td>
<td>410</td>
<td>360</td>
<td>398</td>
<td>0.001</td>
</tr>
<tr>
<td>Human Leucocyte Antigen - antigen D-related (%)</td>
<td>Median</td>
<td>100%</td>
<td>0%</td>
<td>98.50%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>97.50%</td>
<td>100%</td>
<td>0.000</td>
</tr>
<tr>
<td>Cytogenetics (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Favorable</td>
<td>21%</td>
<td>0%</td>
<td>14%</td>
<td>0%</td>
<td>11%</td>
<td>50%</td>
<td>15%</td>
<td>75%</td>
<td></td>
<td>0.003</td>
</tr>
<tr>
<td>Intermediate</td>
<td>58%</td>
<td>92%</td>
<td>57%</td>
<td>67%</td>
<td>78%</td>
<td>29%</td>
<td>55%</td>
<td>25%</td>
<td></td>
<td>0.021</td>
</tr>
<tr>
<td>Risk Group (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
<td>16%</td>
<td>67%</td>
<td>13%</td>
<td>17%</td>
<td>0%</td>
<td>7%</td>
<td>0%</td>
<td>25%</td>
<td></td>
<td>0.000</td>
</tr>
<tr>
<td>CDKN2A (%)</td>
<td>Yes</td>
<td>21%</td>
<td>75%</td>
<td>14%</td>
<td>25%</td>
<td>11%</td>
<td>33%</td>
<td>22%</td>
<td>0%</td>
<td>0.025</td>
</tr>
</tbody>
</table>

NOTE: Significant patient characteristics and ALL features are shown for the overall patient cohort as well as for each SIG. Other nonsignificant variables that were checked, but which lacked association with the SIGs included: age at diagnosis, sex, race, CNS status, infection, IGH gene rearrangement, TCF3 gene rearrangement, ETV6 mutation, RUNX1 mutation, the percentage of bone marrow blasts, bone marrow and peripheral blood monocytes or promyelocytes, hemoglobin, platelet count, albumin and creatinine. P-values were generated using the Kruskal–Wallis test by ranks for continuous variables and the Fisher exact test for categorical variables.
approach arises from the supposition that traditional unsupervised hierarchical clustering ignores known protein interactions and weights each component equally. We hypothesized that if we created smaller subsets of proteins with known functional relationships (i.e., protein functional groups; PFG) and then built overall interaction networks from individual PCs within the PFG as building blocks, that we would build more robust protein patterns. Furthermore, this analysis provides insight into which protein patterns resemble normal cells, and which represent distinct protein expression patterns and activation states between protein clusters.

The existence of recurrent protein patterns led to our hypothesis that overexpressed proteins could function as candidate druggable targets for inhibition or deactivation, while under expressed proteins could function as targets for replacement or reactivation. This concept of replacement has been successfully demonstrated in acute promyelocytic leukemia, where RARα in the fusion gene cannot reach the nucleus, but all-trans retinoid acid can replace this loss of function.\(^\text{40}\) For proteins that are over expressed or significantly activated, use of small molecular inhibitors to has proved a viable strategy. The paradigm for this is the use of imatinib (Gleevec) and other tyrosine kinase inhibitors (e.g., bosutinib, nelotinib and dasatinib) to suppress the constitutively activated ABL kinase activity seen in Ph+ leukemia patients.\(^\text{41}\) By identifying many targets for each SIG, possible rational combinations of targeted therapy could be identified that could be used alone, or in combination with standard chemotherapy. For instance, reactivation of the universally suppressed GATA1 may be useful in inducing differentiation during hematopoiesis.\(^\text{42}\) Likewise, the universal loss of NR4A1 (Nur77)\(^\text{43,44}\) and TCF4\(^\text{45}\) expression poses an opportunity to restore stem cell regulation by restoring normal expression and/or function. To test this hypothesis, we performed proteomic profiling on leukemia cell lines to find representative cell lines that resemble with the protein expression patterns seen in pediatric ALL patients. However, only half of the PCs in patients showed similarities to cell lines, calling into question the relevance of leukemia cell lines in testing drug combinations in future experiments.

If aberrantly expressed proteins could aid in determining patient’s risk group, then classification based on SIGs could be performed at diagnosis and implemented during risk stratification (i.e., prior to consolidation therapy). This process would first need to be tested and validated in larger data sets with more divergence in therapy outcome. If this methodology were shown to be predictive, development of an ELISA or forward phase protein array kit
could potentially classify patients in real time, making routine determination of SIG-
membership both feasible and potentially useful for post-induction therapy determination.

A highly important observation was the association of the Hispanic ethnicity with SIGs. Numerous studies have reported an inferior outcome for patients with Hispanic ethnicity. It is uncertain whether this arises from a different pathophysiology or socioeconomic factors. We observed a clear skewing of some Hispanic patients to specific SIGs, suggesting that for many Hispanic patients the difference in outcome arises from underlying differences in the pathophysiology of their leukemia. A similar finding was noted by Harvey et al. who observed that, within high-risk pediatric ALL, there was a gene expression signature associated with the Hispanic ethnicity that had a very poor 4-year relapse free survival. In our study, this was most pronounced in the differential expression of two T-ALL-CON. Protein expression summaries were notable for over expression of CCND3, DUSP6, RB1, RB1-pSer807,811 and STMN1, and under expression of BCL2 in non-Hispanic groups, and over expression of FKHRL1 along with decreased expression of XIAP and BIRC5 in the Hispanic enriched SIGs. This suggests that leukemia in Hispanics is associated with a less proliferative “push” in combination with greater resistance to apoptosis due to relatively higher levels of BCL2 and
“IAP-proteins” BIRC5 and XIAP. Since malignancies with higher proliferation rates are more sensitive to cell cycle specific chemotherapy agents, and since cells with reduced anti-apoptosis potential are less likely to survive chemotherapy, the CONs provide plausible explanations as to why some Hispanic ALL patients do worse than their non-Hispanic counterparts. However, this observation first needs further validation in a larger patient cohort.

One of the limitations in our study was the small number of patient samples and the restricted number of antibodies targeting phosphorylation sites represented in the array. Repetition of the analysis in a larger cohort of patients, will enable identification of more SIGs that could more accurately discriminate patients and would likely show heterogeneity in outcome. Moreover, it would be interesting to test how additional mutational analysis using genomic and gene expression sequencing could provide more insight in correlation between mutational events and protein expression. A previous study in pediatric ALL observed correlation between the mutational state of NOTCH1 and/or FBXW7 with aberrant NOTCH1 protein expression. However, they also observed NOTCH1 protein activation in some patients without the presence of a NOTCH1 mutation. Another study showed that although patients with mutations in the PTEN/AKT pathway were found to have decreased expression of PTEN
compared to wild-type controls, there was no difference in phosphorylation of AKT or
downstream AKT targets.  

In conclusion, our findings demonstrate the existence of SIGs and CONs in a cohort of pediatric
ALL patients. Elaboration of this approach could be extended to other diseases as well, to
compare SIGs across diseases and to identify disease specific and universal protein expression
patterns.

Figure 5. Significantly higher and lower expressed proteins relative to the normal CD34⁺ samples in
SIG-6. (A) An example of all the significantly altered expressed proteins compared with the normal
CD34⁺ cells for SIG-6 is shown. Higher expressed proteins (up) suggest targets for inhibition and lower
expressed proteins (down) suggest protein targets for replacement or activation. Blue circles denote
proteins that were universally altered in similar direction in all SIGs and red circles point out SIG-
specific protein targets. Colors indicate the relative median protein expression for that SIG, ranged from
the lowest (dark blue) to the highest (maroon) expression. (B) Proteins that were universally changed
in the same direction (in six of the seven SIGs) compared with normal CD34⁺ samples are shown. Non-
significantly different proteins compared with normal CD34⁺ samples are shown in white (blank).

Acknowledgements

This research was funded by the Hyundai Hope on Wheels research grant to TM Horton, the
Takeda/Millennium R01-CA164024 grant to TM Horton, the St Baldrick’s Consortium Grant
to M.E. Scheurer, and the Ladies Leukemia League to ME Scheurer.

Conflict of interest

The authors declare no competing interests.
Reference list


