Chromatin architecture aberrations in prostate cancer and leukemia

by

James Hawley

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Graduate Department of Medical Biophysics
University of Toronto

Contents

Chapter 1

Epigenetic dynamics underlying B cell acute lymphoblastic leukemia relapse

J.R.H, L.G.-P., A.M., J.E.D., and M.L. conceptualized the study. S.M.D., L.G.-P., R.J.V., E.W, J.M., O.I.G., I.G., S.Z.X., M.H., S.R.O., G.N., S.M.C., J.E., C.J.G., J.S.D., M.D.M., C.G.M., and J.E.D. were involved with primary data acquisition. J.R.H., L.G.-P., A.M., and M.C.-S.-Y., J.E.D., and M.L. were involved with the statistical and computational data analysis and biological interpretation. J.R.H. performed all analyses with the DNA methylation (DNAme) data, M.C.-S.-Y. with the RNA sequencing (RNA-seq) data, and A.M. with the assay for transposase-accessible chromatin sequencing (ATAC-seq) data and integration. J.R.H, L.G.-P., and A.M. designed the figures. J.E.D. and M.L. oversaw the study.

1.1 Abstract

1. Relapse of B-cell acute lymphoblastic leukemia (B-ALL) is common in both pediatric and adult patients

2. Previous investigation of the origins of B-ALL relapse identifies growth of subclones present at diagnosis as critical factor

3. Here, we use patient matched diagnosis and relapse samples and multiple PDXs to investigate

epigenetic dynamics that occur over B-ALL relapse

4. Find most differentially methylated regions (DMRs) are unique to each patient and are not recurrent

5. The few recurrent DMRs occur in the promoter regions of genes associated with differentiation and stem cell characteristics

6. Suggests that relapse selects for clones with stem cell characteristics, both genetically and epigenetically

1.2 Introduction

1. Relapse of B-ALL is common in both pediatric and adult patients

2. Previous investigation of the origins of B-ALL relapse identifies growth of subclones present at diagnosis as critical factor

3. Importance of DNAme, CTCF binding, and chromatin organization in hematopoietic stem and progenitor cells (HSPCs) function raises the possibility of these roles being important in B-ALL as well

4. Previous work has identified stem-associated phenotypes and chromatin remodelling pathways as important in these subclones

5. Here, we use patient matched diagnosis and relapse samples and multiple PDXs to investigate epigenetic dynamics that occur over B-ALL relapse

1.3 Results

1.3.1 Multi-omic integration of B-ALL relapse patients links DNA methylation to relapse status

1.3.2 Widespread loss of DNA methylation over normal B-cell differentiation

1.3.3 Widespread gain of DNA methylation over B-ALL relapse

1.3.4 Recurrent DNA methylation changes identify stem cell pathways in relapse

1.4 Discussion

1. summary of our work
   (a) Relapse is a major barrier to treating B-ALL
   (b) Requires rigorous molecular and multi-omic investigation to understand the origins of relapse and whether it can be predicted at diagnosis
   (c) Multi-omic studies here identify DNAme as an important biomarker of B-ALL relapse
   (d) DNAme and RNA-seq changes indicate presence of stem-like cells at diagnosis that become dominant at relapse
2. context of current work in regard to relapse treatment and other studies
3. future directions for investigation
   (a) effect of targeted DNAme of promoter regions for important stem genes on engraftment
   (b) effect of demethylating agents on relapsed B-ALL patients
   (c) combination therapy of demethylating agents with chemotherapy to reduce the potential outgrowth of relapse-fated subclones

1.5 Methods

1.5.1 Patient selection and sample collection

Patient samples were obtained at diagnosis and relapse from patients with B-ALL as previously described [1]. All samples were frozen viably and stored long term at -150 °C. Samples were selected retrospectively based on paired-sample availability.

Human cord blood samples were obtained with informed consent from Trillium and Credit Valley Hospital according to procedures approved by the University Health Network Research Ethics Board, as previously described [1]. Cells were stained with the following antibodies (all from BD Biosciences, unless otherwise stated):

- FITC anti-CD45RA (1:50, 555488)
- PE anti-CD90 (1:50, 555596)
- PE-Cy5 anti-CD49f (1:50, 551129)
- V450 anti-CD7 (1:33.3, 642916)
- PE-Cy7 anti-CD38 (1:100, 335790)
- APC anti-CD10 (1:50, 340923)
- APC-Cy7 anti-CD34 (1:200, custom made by BD Biosciences)

Cells were sorted from cord blood cells on the basis of markers listed in Table 1.1, as previously described [2], on a FACSAria III (Becton Dickinson), consistently yielding > 95 % purity.

Table 1.1: **Cell surface markers used to isolate cell populations from cord blood pools.**

Cell type(s)	Surface markers
HSCs & MPPs	CD34+ CD38- CD45RA-
CMPs, GMPs, & MEPs	CD34+ CD38+ CD10- CD19+
LMPPs & MLPs	CD34+ CD38- CD45RA+
EarlyProBs, PreProBs, & ProBs	CD34+ CD38+ CD10+ CD19+
B	CD34- CD38+ CD19+ CD33- CD3- CD56-

1.5.2 Patient-derived xenograft generation and limiting dilution assays

Patient-derived xenografts (PDXs) were generated as previously described [1]. Clinical samples were stained with the following antibodies:

- anti-CD19 PE (BD Biosciences, clone 4G7)
- anti-CD3 FITC (BS Biosciences, clone SK7) or anti-CD3 APC (Beckman Coulter, clone UCHT11)
- anti-CD45 APC (BD Biosciences, clone 2D1) or anti-CD45 FITC (BD Biosciences, clone 2D1)
- anti-CD34 APC-Cy7 (BD Biosciences, clone 581)

Each sample was sorted on a FACSAria III (BD Biosciences) for leukemic blasts ($CD19^{+}CD45^{dim/-}$) and T cells ($CD3^{+}CD45^{hi}$). NOD scid gamma (NSG) mice were bred according to protocols established and approved by the Animal Care Committee at the University Health Network. 8-to-12-week-old mice were sublethally irradiated at 225 cGy 24 hours prior to transplants. Only female mice were used. Intra-femoral injections of 10 to 250 000 sorted leukemic blasts were performed as previously described [3]. Mice were sacrificed 20-to-30 weeks post-transplant or at the onset of disease symptoms. Human cell engraftment in the injected femur, bone marrow (non-injected bones, left tibia, right tibia, left femur), spleen, and central nervous system were assessed using human-specific antibodies for CD45 (PE-Cy7, BD Biosciences, clone HI30; v500 BD Biosciences, clone HI30), CD44 (PE, BD Biosciences, clone 515; FITC, BD Biosciences, clone L178), CD3 (APC, BD Biosciences, clone UCHT1), CD19 (PE-Cy5, Beckman Coulter, clone J3-119), CD33 (PE-Cy7, BD Biosciences, clone P67-6; APC, BD Biosciences, clone P67-6), and CD34 (APC-Cy7, BD Biosciences, clone 581)

analyzed on an LSRII (BD Biosciences). Mice were considered to be engrafted when > 0.1 % of cells in the injected femur were positive for one or more human B-ALL-specific cell surface marker (CD45, CD44, CD19, and CD34). Confidence intervals for the frequency of leukemia initiating cells was calculated using ELDA [4].

1.5.3 Human cell isolation from patient-derived xenografts

Cells from the injected femur, bone marrow, and spleen, were frozen viably after sacrifice. Injected femur and bone marrow of mice engrafted with > 10 % human cells were combined. These cells were depleted of mouse cells using the Miltenyi Mouse Cell Depletion Kit (Miltenyi Biotec; samples with > 20 % engraftment) or by cell sorting with human CD45 and human CD19 and/or CD34 cell surface antibodies to a purity of > 90 %, as determined by post-processing flow cytometry. Central nervous system cells from mice with > 60 % engraftment were used directly for DNA isolation. DNA was isolated using the QIAamp DNA Blood Mini or Micro Kit (Qiagen).

1.5.4 Primary and patient-derived xenograft sample sequencing

RNA sequencing

RNA-seq was performed as previously described [1]. Briefly, amplified complementary DNA (cDNA) was sequenced as paired-end libraries on an Illumina HiSeq2000. The libraries were sequenced as 2×75 bp for the adult and 2×100 bp for the pediatric samples.

DNA methylation capture sequencing

DNA methylation capture sequencing (MeCapSeq) was performed using the SeqCapEpi CpGiant kit (Roche NimbleGen). Briefly, the DNA library is prepared and bisulfite converted, amplified, and enriched using capture probes for targeted bisulfite-converted DNA fragments, then sequenced on a short-read sequencing machine. More specifically, library preparation for MeCapSeq was performed with the KAPA Library Preparation Kits, bisulfite conversion of genomic DNA was performed with the Zymo EZ DNA Methylation Lightning kit, bisulfite-converted DNA libraries were amplified using the KAPA HiFi HotStart Uracil+ ReadyMix kit, and finally hybridized to probes from the SeqCap Epi Enrichment Kit. Captured DNA fragments were sequenced on an Illumina HiSeq 2500 as 2×125 bp to a target depth of 70×10^6 read pairs per sample.

Assay for transposase-accessible chromatin sequencing

Library preparation for ATAC-seq was performed with the Nextera DNA Sample Preparation Kit (FC-121-1030, Illumina), according to a previously reported protocol [5]. ATAC-seq libraries were sequenced with an Illumina HiSeq 2500 sequencer to generate single-end 50 bp reads.

1.5.5 Sequencing data analysis

Differential gene expression analysis

The methods are described in [REF 1]. Briefly, RNA-seq reads were aligned against the GRCh38 reference human genome with STAR (v2.5.2b) [6] and annotated with the Ensembl reference (v90). Default parameter were used with the following exceptions: chimeric segments were screened with a minimum size of 12 bp, junction overlap of 12 bp, and maximum segment reads gap of 3 bp; splice junction overlap of 10 bp; maximum gap between aligned mates of 100 000 bp; maximum aligned intron of 100 000; and alignSJstitchMismatchNmax of 5 1 5 5. Transcript counts were obtained with HTSeq (v0.7.2) [7]. Data was library size normalized using the RLE method, followed by a variance stabilizing transformation using DESeq2 (v1.22.1) [8]. Principal component analysis plots were generated on a per sample basis using the top 1 000 variable genes. For downstream analysis, the mean expression of each sample clone condition was used. For per-patient analyses, differentially expressed genes were identified between disease stage and clone status using DESeq2. Genes with an false discovery rate (FDR) < 0.05 and absolute $\log_2$(fold change) > 1 were considered significant.

Identification of accessible chromatin peaks

ATAC-seq reads were aligned against the GRCh38 reference human genome with Bowtie2 (v2.0.5) [9] with default parameters. Accessible peaks were identified with MACS2 (v2.0.10) [10] with the following command:

```
macs2 callpeak -f BED -g hs --keep-dup all -B --SPMR --nomodel --
    shift -75 --extsize 150 -p 0.01 --call-summits -n {sample_name}
     -t {input_bam}
```

A catalogue of peaks from all samples was collected with a custom R script. ATAC-seq signal was mapped from each sample to this catalogue using Bedtools [11] for downstream analysis.

Bisulfite sequencing pre-processing

Sequencing read qualities were assessed with FastQC (v0.11.8) [12]. Low quality bases were trimmed with Trim Galore! (v0.6.3) [13] with the following command:

```
trim_galore --gzip -q 30 --fastqc_args '-o TrimGalore' {
   sample_mate1} {sample_mate2}
```

Trimmed reads were aligned to the GRCh38 reference human genome with Bismark (v0.22.1) [14] with default parameters. Duplicates were removed from the resulting alignment file with the following command:

```
deduplicate_bismark -p --bam {input_bam}
```

The deduplicated BAM file was sorted by position with sambamba (v0.7.0) [15]. M-biases were calculated with MethylDackel (v0.4.0) [16], and methylation β values were extracted from the BAM files with the following command:

```
MethylDackel extract --mergeContext --OT 3,124,3,124 --OB
   3,124,3,124 {ref_genome} {dedup_sorted_bam}
```

Both M and β values were for each CG dinucleotide (CpG) were used in downstream analyses.

Similarity network fusion

Preprocessed data from each sample was collected with the following features: normalized gene expression abundance for all genes, chromatin accessibility signal within previously identified accessible peaks, and mean β value for all CpGs listed in the manifest for targeted bisulfite sequencing kit. These features and sample labels were processed with the SNFtool R package [17] to perform the similarity network fusion analysis. Graphs were constructed for all samples deriving from a single patient where each node is a sample and each edge is weighted according to the determined similarity between the samples. Edges whose weights were below specific thresholds were removed from the graph. The threshold weight for the fused graph was 0.05. Similar graphs were constructed using the individual components for each sample (e.g. using just the similarity in RNA-seq data), and the component graphs were compared to the fused graph, to compare the importance of each feature. Threshold weights for these individual graphs were determined to be 6×10^{-5} for DNAme, 4×10^{-4} for gene expression, and 2×10^{-4} for chromatin accessibility.

Differentially methylated region identification

DMRs were identified using the dmrseq R package (v1.3.8) [18] with an absolute filtering cutoff value of 0.05 and using the sequencing batch as an adjustment covariate. Normal samples from all donors were compared pairwise based on their sorted cell type. B-ALL samples were compared by their designated disease stage (Dx, DRI, or Rel), and were compared both across all patients (e.g. all Dx samples against all Rel samples), or within a single patient (e.g. all Dx samples from Patient 1 against all Rel samples from Patient 1). A multiple testing correction with the FDR method was performed [19]. Regions with an FDR < 0.1 were determined to be significant.

Gene ontology enrichment analysis

Gene ontology enrichment analysis was performed using the PANTHER classification system (database version 2019-10-08) [20]. Gene symbols for the genes whose promoter regions contained the recurrently hyper-methylated regions in all B-ALL patient samples were supplied, with the entire human genome as the background. An over-representation Fisher test for biological processes was performed with an FDR correction. Biological processes at the top of the hierarchy with an FDR < 0.05 were determined to be significant.

Glossary

3C chromatin conformation capture

ANOVA Analysis of Variance

AR androgen receptor

ATAC-seq assay for transposase-accessible chromatin sequencing

B-ALL B-cell acute lymphoblastic leukemia

cDNA complementary DNA

ChIP-seq chromatin immunoprecipitation sequencing

CMP common myeloid progenitor

CPC-GENE Canadian Prostate Cancer Genome Network

CpG CG dinucleotide

crRNA CRISPR RNA

CRE *cis*-regulatory element

DEPMAP Cancer Dependency Map

DHS DNase I hypersensitive sites

DMR differentially methylated region

DNAme DNA methylation

EarlyProB early progenitor B cell

FDR false discovery rate

FN false negative

FP false positive

FOX forkhead box

GLM generalized linear model

GMP granulocyte-macrophage progenitor

gRNA guide RNA

HSC hematopoietic stem cell

HSPC hematopoietic stem and progenitor cell

IID independent and identically distributed

JS James-Stein

kbp kilobase

KO knockout

LDA limiting dilution assay

LMPP lymphoid-primed multi-potent progenitor

MeCapSeq DNA methylation capture sequencing

MEP megakaryocyte-erythrocyte progenitor

MSE mean square error

mCRPC metastatic castration-resistant prostate cancer

MLP monocyte-lymphoid progenitor

MPP multi-potent progenitor

NSG NOD scid gamma

OLS ordinary least squares

mRNA messenger RNA

PCa prostate cancer

PDX patient-derived xenograft

PreProB pre-progenitor B cell

ProB progenitor B cell

RNAi RNA interference

RNA-seq RNA sequencing

shRNA small hairpin RNA

siRNA small interfering RNA

SNV single nucleotide variants

SRA Seqeuence Read Archive

SV structural variant

TAD topologically associated domain

TCGA The Cancer Genome Atlas

TN true negative

TP true positive

TF transcription factor

tracrRNA trans-activating CPRISR RNA

UTR untranslated region

WGS whole genome sequencing

WT wild-type