

RESEARCH ARTICLE

# Very long intergenic non-coding RNA transcripts and expression profiles are associated to specific childhood acute lymphoblastic leukemia subtypes

Maxime Caron<sup>1,2</sup>, Pascal St-Onge<sup>1</sup>, Simon Drouin<sup>1</sup>, Chantal Richer<sup>1</sup>, Thomas Sontag<sup>1</sup>, Stephan Busche<sup>2</sup>, Guillaume Bourque<sup>2</sup>, Tomi Pastinen<sup>2</sup>, Daniel Sinnett<sup>1,3\*</sup>

**1** CHU Sainte-Justine Research Center, Montreal, Quebec, Canada, **2** Department of Human Genetics, McGill University, Montreal, Quebec, Canada, **3** Department of Pediatrics, University of Montreal, Montreal, Quebec, Canada

\* [daniel.sinnett@umontreal.ca](mailto:daniel.sinnett@umontreal.ca)



**OPEN ACCESS**

**Citation:** Caron M, St-Onge P, Drouin S, Richer C, Sontag T, Busche S, et al. (2018) Very long intergenic non-coding RNA transcripts and expression profiles are associated to specific childhood acute lymphoblastic leukemia subtypes. PLoS ONE 13(11): e0207250. <https://doi.org/10.1371/journal.pone.0207250>

**Editor:** Obul Reddy Bandapalli, German Cancer Research Center (DKFZ), GERMANY

**Received:** June 7, 2018

**Accepted:** October 26, 2018

**Published:** November 15, 2018

**Copyright:** © 2018 Caron et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Whole transcriptome, ChIP-seq and WGBS datasets are available on the Gene Expression Omnibus (GEO) under accession numbers GSE89071 and GSE120677.

**Funding:** This study was supported by research funds provided by the Terry Fox Research Institute (<http://www.tfri.ca/>) and the Canadian Institutes of Health Research (<http://www.cihr-irsc.gc.ca/e/193.html>). DS holds the François-Karl-Viau Research Chair in Pediatric Oncogenomics. The operation of

## Abstract

Very long intergenic non-coding RNAs (vlincRNAs) are a novel class of long transcripts (~50 kb to 1 Mb) with cell type- or cancer-specific expression. We report the discovery and characterization of 256 vlincRNAs from a cohort of 64 primary childhood pre-B and pre-T acute lymphoblastic leukemia (cALL) samples, of which 61% are novel and specifically expressed in cALL. Validation was performed in 35 pre-B and pre-T cALL primary samples. We show that their expression is cALL immunophenotype and molecular subtype-specific and correlated with epigenetic modifications on their promoters, much like protein-coding genes. While the biological functions of these vlincRNAs are still unknown, our results suggest they could play a role in cALL etiology or progression.

## Introduction

Childhood acute lymphoblastic leukemia (cALL) represents approximately 25% of all pediatric cancer cases. Despite remarkable improvements in survival, with 5 year event-free survival rates of approximately 80%, non-responding or relapsing patients still represent one of the most frequent cause of disease-related death in children [1]. Childhood ALL is a complex disease comprising multiple molecular subtypes with distinctive somatic genetic alterations such as aneuploidy, chromosomal rearrangements, and point mutations [1]. High hyperdiploid cases (HHD) and those harboring the t(12;21) [ETV6/RUNX1] rearrangement, together representing about half of pre-B cALL cases. Both subtypes are associated with a good prognosis [1, 2]. Other less frequent (< 10%) subtypes, such as MLL-rearranged, t(1;19) [TCF3/PBX1], or t(9;22) [BCR/ABL1], are associated with intermediate to poor outcomes [1, 2]. These genetic alterations contribute to leukemogenesis by altering key regulatory processes, subverting normal proliferation control, blocking differentiation, and promoting resistance to death signals [2]. Although ~75% of cALL cases can be currently sub-classified clinically using standard

the Calcul Québec supercomputer is funded by the Canada Foundation for Innovation (<https://www.innovation.ca/>), RMGA (<https://www.rmga.qc.ca/en/>), and the Fonds de recherche du Québec - Nature et technologies (<http://www.frqnt.gouv.qc.ca/en/accueil>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

cytology or molecular diagnostics techniques [1, 2], accurate patient risk stratification is still an ongoing challenge. Interestingly, a recent study showed that expression profiles could classify up to 98% of cases [3]. While these studies primarily focused on the analysis of protein-coding transcripts [3–6], long non-coding RNA (lncRNA) transcripts have also been shown to have pre-B cALL subtype-specific expression and can modulate cell proliferation, apoptosis, migration, and treatment resistance [4–7]. Recently, a new class of lncRNAs known as “very long intergenic non-coding RNAs” (vlincRNAs) has been described. So far only a few thousand vlincRNAs, whose size ranges from 50kb to 1Mb, have been identified. However, it is known that these transcripts show cell type-specific expression patterns and seem to have biological functions [8–10]. In this study, we described vlincRNA populations expressed in cALL primary samples through whole-transcriptome sequencing, assessed their cALL-subtype specificity, and investigated putative expression regulation mechanisms. The insight gained from our results on this new class of transcripts will spur further research on their expression and function not only in cALL but also in other cancer types.

## Results

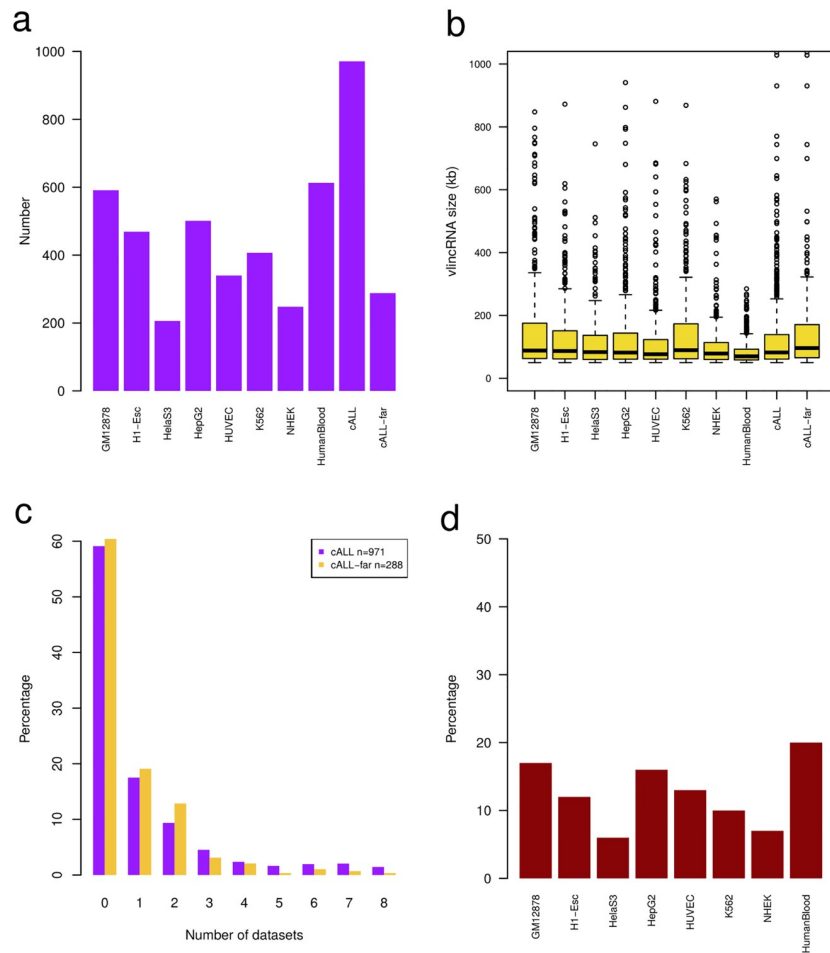
### Identification of vlincRNAs expressed in cALL patients

We sequenced the whole transcriptome of 64 cALL patients (57 pre-B and 7 pre-T) and 4 matched normal cell populations derived from human cord blood (3 CD10<sup>+</sup>CD19<sup>+</sup> pre-B and 1 CD3<sup>+</sup> pre-T) (S1 Table; [6]). In this “discovery cohort” we identified a total of 971 expressed vlincRNAs with a median size of 82.2 kb [50 kb–1.03 Mb], in line with previous estimates and supporting the sensitivity of the discovery method (Fig 1A and 1B; [9]). When compared to available public datasets [8, 9], we found that 59% (574 / 971) of vlincRNAs were unique and specific to our cALL samples (< 25% reciprocal strand specific overlap) while the remaining 41% overlapped with those found in at least one existing dataset (Fig 1C and 1D). We identified a high-confidence subset of 288 vlincRNAs (31.4%; 288 / 917) located at least 10 kb away from protein coding genes on either strand. From these, we selected 256 autosomal vlincRNAs having at least 100 reads per sample in a minimum 3 samples to perform differential expression analyses.

### VlincRNA expression is pre-B cALL molecular subtype-specific

Since it was previously reported that lncRNA expression profiles can accurately classify pre-B cALL molecular subtypes [4, 7], we investigated whether this also held true for vlincRNAs. Principal component analysis on the selected 256 vlincRNAs showed cALL immunophenotype-specific expression patterns highlighted by pre-B and pre-T cALL cases clustering separately (Fig 2A). We observed cALL subtype discrimination between pre-T, HHD, t(12;21) [ETV6/RUNX1], t(9;22) [BCR/ABL1], and unclassified (“Other”) pre-B cALL subtypes when performing hierarchical clustering using vlincRNA normalized expression values (cluster purity = 0.85; Fig 2B). Furthermore, we observed subtype-specific vlincRNA expression patterns when either all 256 vlincRNAs (Fig 2C) or the top 25 vlincRNAs differentially expressed in primary cALL samples (relative to control) are considered (Fig 2D). Interestingly, a sample originally classified as “Other” at diagnosis clustered with t(12;21) samples using vlincRNA expression (Fig 2). We determined by RT-PCR that the t(12;21) [ETV6/RUNX1] fusion was indeed expressed and reassigned this sample to the t(12;21) [ETV6/RUNX1] subtype.

Some vlincRNAs had expression levels orders of magnitude higher in specific subtypes, suggesting a subtype-specific role (e.g., HHD-specific or t(12;21)-specific vlincRNAs shown in Fig 3). To assess the robustness of these findings, we further analyzed the transcriptome of 35 independent primary samples (30 pre-B and 5 pre-T) as a replication cohort (S1 Table) using



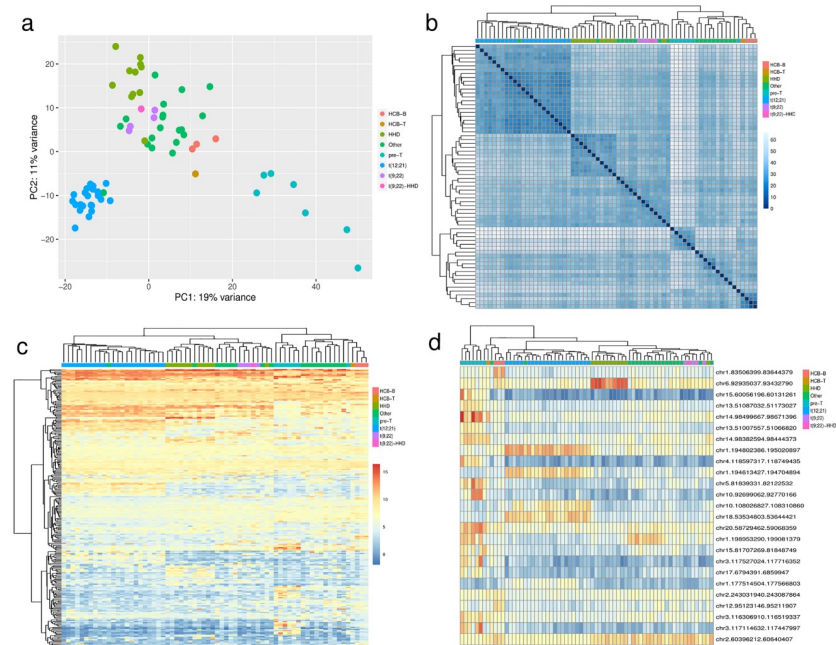
**Fig 1. Characterization of cALL-derived vlincRNAs.** (A) Number of vlincRNAs previously reported for diverse cell types (GM12878, H1-Esc, HeLaS3, HepG2, HUVEC, K562, NHEK, HumanBlood) [9] and discovered in our cALL dataset, including ‘far-from-genes’ cALL vlincRNAs that are at least 10 kb away from protein-coding genes. (B) Boxplot of vlincRNA transcript sizes. (C) Percentage of all (n = 971) and ‘far-from-genes’ (n = 288) cALL vlincRNAs overlapping vlincRNAs a given number of public datasets. (D) Percentage of cALL vlincRNAs overlapping vlincRNAs from public datasets (HumanBlood data is unstranded).

<https://doi.org/10.1371/journal.pone.0207250.g001>

the Illumina HiSeq2500/4000 platforms (see [Methods](#)). PCA analysis and hierarchical clustering using minimally expressed vlincRNAs in the replication cohort (n = 273/288) confirmed the subtype-specific expression patterns observed in the discovery cohort (cluster purity of 0.94), [S2A and S2B Fig](#). Subtype-specific vlincRNA expression between discovery and replication samples also showed strong Pearson correlations ([S2C Fig](#)), confirming that vlincRNA expression patterns are robust reliable discriminators of cALL subtypes. We also looked at the expression levels of t(12;21)-specific vlincRNAs in REH cells and found similar expression patterns in this cell line ([S2D Fig](#)). Interestingly, we observed different vlincRNA expression profiles for pre-B and pre-T cALLs, indicating that the expression of certain vlincRNA is cancer type-specific.

### Epigenetic regulation of cALL-specific vlincRNAs

To delineate the basis of the vlincRNA expression regulation in cALL, we investigated epigenetic regulation by generating ChIP-seq data of the H3K4me1, H3K4me3, and H3K27ac active

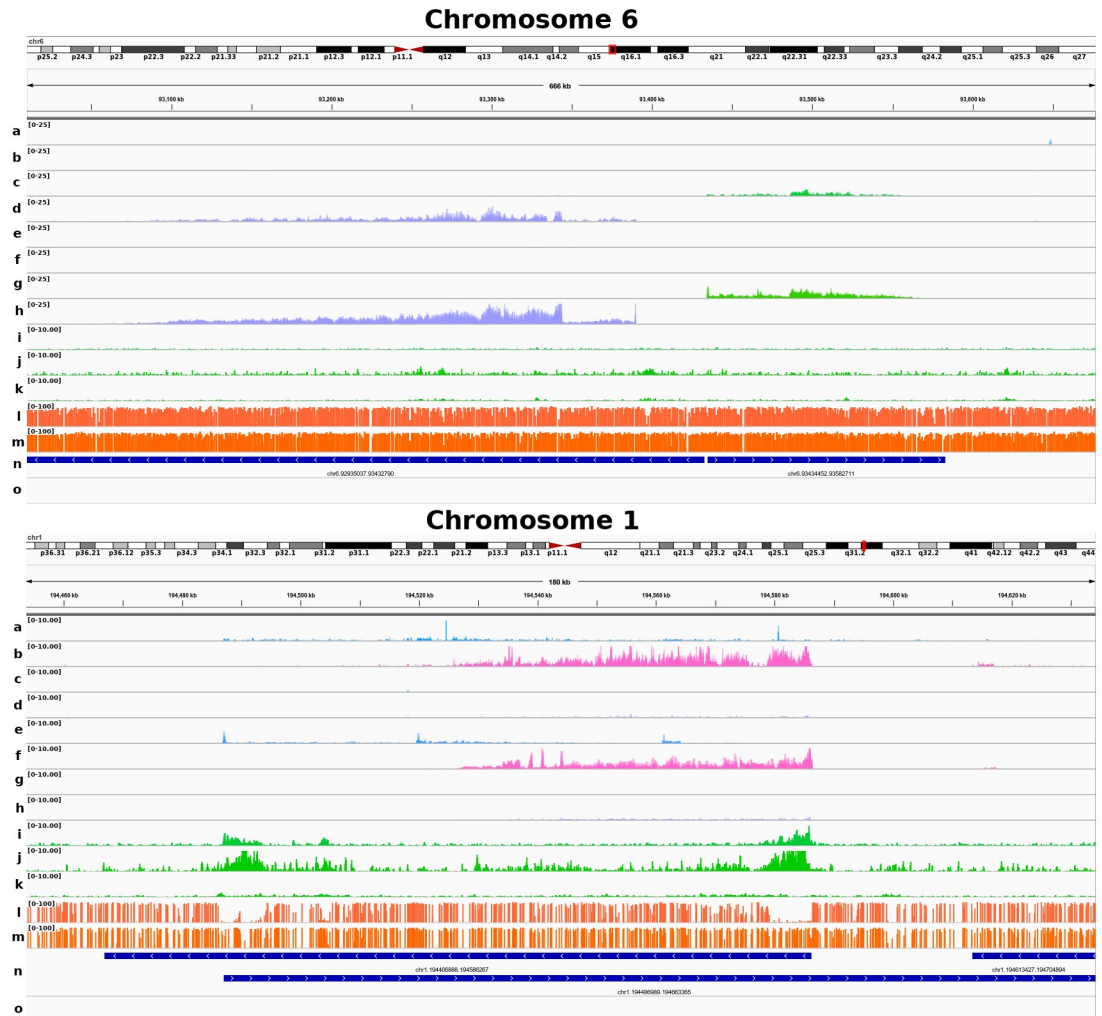


**Fig 2. Subtype-specific classification of cALL using vlincRNA expression profiles.** (A) PCA plot of the discovery samples (n = 68) using the DESeq2 regularized log transform (rld) normalized read counts of minimally expressed cALL vlincRNAs (n = 256). (B) Hierarchical clustering of discovery samples using Euclidean distance on vlincRNA normalized rld expression values. Cluster purity = 0.85 using 8 clusters. (C) Heatmap of vlincRNA normalized rld expression values in the discovery samples. (D) Heatmap of normalized rld expression values of the top 25 significantly differentially expressed cALL vlincRNAs (adj. p-value < 0.05) in the discovery samples using the likelihood ratio test on subtypes with DESeq2. HCB-B: CD19<sup>+</sup>CD10<sup>+</sup> pre-B cells isolated from human cord blood. HCB-T: CD3<sup>+</sup> pre-T cells isolated from human cord blood.

<https://doi.org/10.1371/journal.pone.0207250.g002>

chromatin marks from a pool of two additional t(12;21)[*ETV6/RUNX1*] cases. Nearly a third (30.9%; 79 / 256) of cALL-specific putative vlincRNA promoters ( $\pm 10$  kb) overlapped active chromatin regions, representing a significant enrichment (2.2-fold enrichment; Fisher p-value =  $4.0 \times 10^{-6}$ ). Enriched active chromatin region overlaps was positively correlated with vlincRNA expression in t(12;21) [*ETV6/RUNX1*] samples: the highest expression quartile has a higher fraction of overlap than the lowest quartile (Q1: 56.3% vs. Q4: 7.8%; Fig 4). DNA methylation was also investigated to further characterize the epigenetic landscape of cALL vlincRNAs by performing whole genome bisulfite sequencing (WGBS) of three t(12;21)[*ETV6/RUNX1*] cases and one control CD10<sup>+</sup>CD19<sup>+</sup> pre-B cell sample isolated from human cord blood. This experiment showed that the 79 active vlincRNA promoters described above were hypomethylated compared to control cells. This observation is further supported by the analysis of publicly available WGBS results [11] and 450K DNA methylation data [12, 13] of 242 samples (see Methods; Fig 5 and S4 Fig). Transcription factor motif enrichment analyses of highest expressed quartiles vlincRNA candidate promoters ( $\pm 10$ kb; n = 64) in the t(12;21) and HHD subtypes revealed distinct regulators in both subtypes (S5 Fig). Motifs from the *ETS* transcription factor family, with roles in tissue development and cancer progression, were significantly enriched in the t(12;21) subtype, while binding motifs for AP-1 subunits *JUN* and *FOS*, implicated in apoptosis, proliferation, and differentiation, were significantly enriched in the HHD subtype.

Finally, we investigated whether other hematological cell types exhibited a similar trend by looking at H3K4me3 read density on vlincRNA promoters overlapping active chromatin

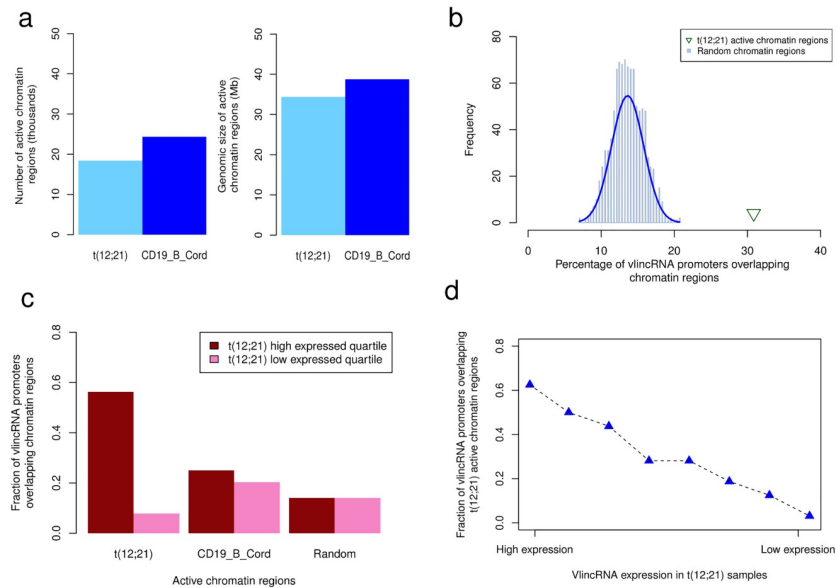


**Fig 3. Genome wide tracks of gene expression, active histone marks and CpG methylation of subtype-specific cALL vlincRNAs.** Epigenetic tracks of vlincRNAs expressed in specific subtypes. Top, HHD-specific vlincRNAs expressed on chromosome 6 in both the SOLiD and Illumina sequencing platforms. Bottom, t(12;21)-specific vlincRNAs expressed on chromosome 1 in both the SOLiD and Illumina sequencing platforms. Splicing patterns are more clearly defined with Illumina RNA-seq data. (A,B) Normalized read coverage of t(12;21) vlincRNA expression (SOLiD platform, + and—strands). (C,D) Normalized read coverage of HHD vlincRNA expression (SOLiD platform, + and—strands). (E,F) Normalized read coverage of t(12;21) vlincRNA expression (Illumina platform, + and—strands). (G,H) Normalized read coverage of HHD expression (Illumina, + and—strands). (I,J,K) Normalized read coverage of the H3K4me3, H3K27ac and H3K4me1 ChIP-seq histone marks in the t(12;21) pool sample. (L,M) WGBS methylation levels of the merged t(12;21) cases (n = 3) and the CD10<sup>+</sup>CD19<sup>+</sup> control cells. (n) Coordinates of vlincRNA transcripts (n = 288) found in the discovery samples (n = 68). (o) RefSeq gene annotations.

<https://doi.org/10.1371/journal.pone.0207250.g003>

regions (identified above) in three ENCODE cell lines: GM12878 (B-lymphocytes), K562 (chronic myeloid leukemia), and Jurkat (pre-T ALL). Interestingly, these cell lines also showed active chromatin enrichment on a fraction of vlincRNA promoters identified in t(12;21) [ETV6/RUNX1] samples, suggesting that vlincRNA expression could be regulated through epigenetic mechanisms (Fig 6; [15]). Given that ~40% of vlincRNAs expressed in cALL are not cALL-specific, it is not surprising that a subset of t(12;21) [ETV6/RUNX1] actively regulated vlincRNAs could be under H3K4me3 activation in other hematological cell lines.





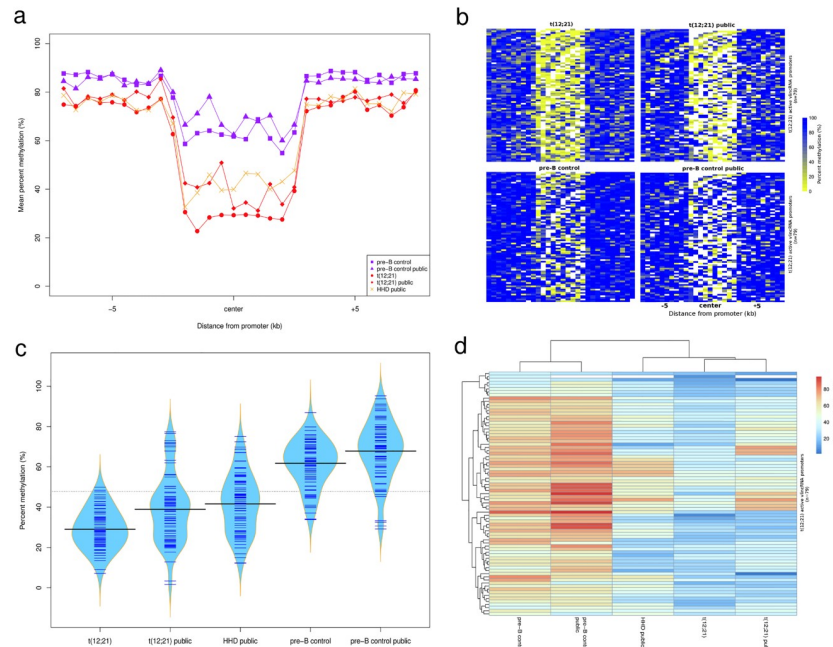
**Fig 4. Epigenetic regulation at cALL vlincRNA promoters.** (a) Number (left) and genomic size (right) of active chromatin regions in two t(12;21) samples and CD19<sup>+</sup> B cord. Active chromatin regions in the t(12;21) samples were generated using ChromHMM on three chromatin marks (H3K4me3, H3K27ac and H3K4me1) and retaining states having the H3K4me3 mark present. Active CD19<sup>+</sup> B cord chromatin regions were obtained from the Roadmap Epigenomics Project [14] by keeping states having the H3K4me3 mark present (states 1,2,3,10, and 11) from the 15-state model. (b) Percentage of vlincRNA promoters (n = 256) overlapping active t(12;21) chromatin regions (30.9%; 79 / 256) and random regions (mean 14.06%; 36 / 256). Random regions were generated a thousand times using shuffleBed from bedtools. (c) Percentage of vlincRNA promoters overlapping active t(12;21) chromatin regions, active CD19<sup>+</sup> B cord chromatin regions and random regions in the t(12;21) high and low expressed vlincRNA quartiles (Q1 and Q4, n = 64). (d) Fraction of vlincRNA promoters overlapping active t(12;21) chromatin regions sorted by t(12;21) vlincRNA expression (8 bins of 32 promoters). HCB-B: CD19<sup>+</sup>CD10<sup>+</sup> pre-B cells isolated from human cord blood.

<https://doi.org/10.1371/journal.pone.0207250.g004>

## Discussion

In this study, we identified and characterized very long intergenic non-coding RNAs (vlincRNA) expression patterns in 64 primary cALL samples (57 pre-B and 7 pre-T) and showed that they are specific to cALL subtypes and that epigenetic modifications correlated with their expression. We found 971 vlincRNA transcripts expressed in cALL primary samples, of which 59% (574 / 971) were unique to our dataset. We further selected 256 high-confidence autosomal vlincRNAs having at least 100 reads per sample in a minimum 3 samples to perform differential expression analyses (156 / 256, or 61% unique to cALL). The median size and range of the vlincRNA transcripts we identified was of 82.2 kb [50 kb—1.03 Mb], which is concordant with previously reported sizes (~50 kb—1 Mb), with a median size of ~80 kb [9, 16]. Although it was reported that vlincRNAs are mostly unspliced [10], we have observed putative splicing patterns in some vlincRNAs (Fig 3), indicating that they could be expressed in both spliced and unspliced forms or less likely be novel pre-mRNAs.

Despite the current cure rate, cALL patients whose disease is refractory to treatment or relapses face a dismal prognosis. In addition, this high cure rate has been achieved through risk-based treatments administered at the expense of considerable toxicity and decreased quality of life. Indeed, more than 70% of young adult survivors will suffer from long-term effects because of their treatments [17, 18]. Thus, the development of improved risk stratification strategies leading to personalized and targeted treatment is thus essential to improve patient outcome and long-term quality of life. Previous studies done using microarray technology [5] or whole-transcriptome sequencing had shown that protein-coding [4, 7] or long non-coding

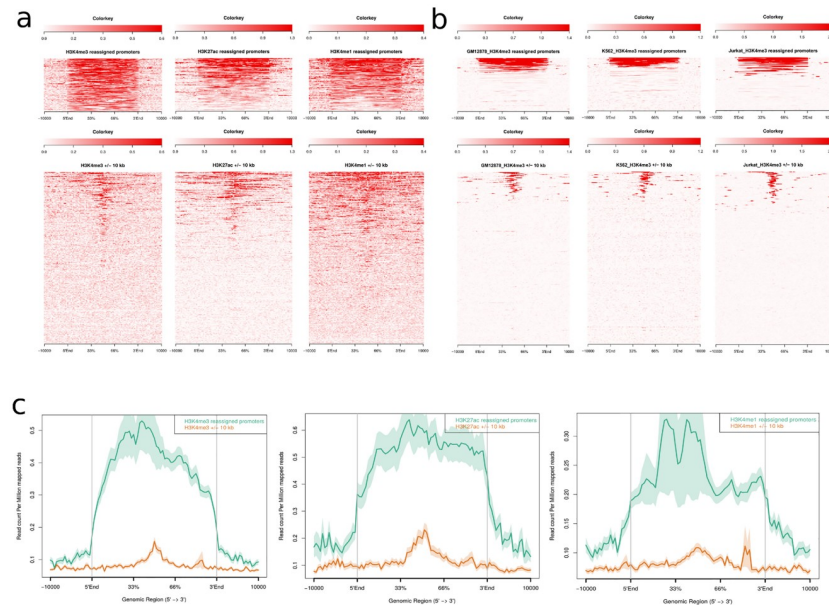


**Fig 5. Methylation regulation of t(12;21) active vlincRNA promoters.** (a) WGBS mean methylation levels of t(12;21) active promoters (n = 79) in t(12;21), HHD and pre-B control CD10<sup>+</sup>CD19<sup>+</sup> datasets from this study and public data [11]. Each data point represents 10% of promoter size and flanking regions size (±10kb). (b) WGBS methylation heatmap of t(12;21) active promoters in t(12;21) and pre-B control CD10<sup>+</sup>CD19<sup>+</sup> datasets from this study and public data (white = no covered CpGs). Each column represents 10% of promoter size and flanking regions size (±10kb). (c) DNA methylation beanplot of t(12;21) active promoters in t(12;21), HHD and pre-B control CD10<sup>+</sup>CD19<sup>+</sup> datasets from this study and public data. (d) Hierarchical clustering of methylation levels of t(12;21) active promoters in t(12;21), HHD and pre-B control CD10<sup>+</sup>CD19<sup>+</sup> datasets showing similar methylation profiles between this study and public data.

<https://doi.org/10.1371/journal.pone.0207250.g005>

RNA transcription profiles [4–7] could be used to discriminate cALL disease subtypes. Here, we have demonstrated that vlincRNA expression profiles can discriminate between cALL subtypes, particularly the t(12;21) [ETV6/RUNX1] and HHD subtypes, which together account for ~60% of pre-B cALL cases. These results, validated in 35 other primary cALL samples suggest that vlincRNA expression patterns can be used as molecular biomarkers for more accurate disease subtype classification. It would be interesting to see if these results scale with increased cohort size to reach the high cALL subtype classification accuracy demonstrated by Lilljebjorn *et al.* [3]. Furthermore, we show that vlincRNA expression is cancer-specific as pre-B and pre-T cALL exhibit different expression patterns. This observation is concordant with that of St-Laurent and colleagues [9] which have reported a wide variation in vlincRNA expression pattern across cancerous and normal cell types. Although these are both acute leukemias, they originate from distinct cell types, play different role, carry specific molecular alterations and thus can be considered as different cancers. These data and the consistency in expression patterns observed across both our discovery and validation cohorts strongly suggest that they are not transcriptional noise but rather disease-specific and even disease subtype-specific.

We further showed that putative vlincRNA promoters are enriched in active chromatin histone marks and had lower DNA methylation levels than healthy cell counterparts. These data support St-Laurent *et al.*'s hypothesis of vlincRNA transcriptional epigenetic regulation [9]. Together these data strongly suggest that subtype-specific vlincRNA expression is regulated by epigenetic changes.



**Fig 6. Histone mark ChIP-seq read density plots of candidate t(12;21) active promoters.** (a) Heatmaps of H3K4me3, H3K27ac and H3K4me1 ChIP-seq normalized read densities for candidate (bottom,  $\pm 10$ kb around the 5' start of vlincRNAs,  $n = 256$ ) and reassigned (top,  $n = 79$ ) t(12;21) active vlincRNA promoters. (b) H3K4me3 ChIP-seq normalized read densities in GM12878, K562 and Jurkat cell lines on candidate and reassigned t(12;21) active vlincRNA promoters. (c) ChIP-seq normalized read density plots of the H3K4me3, H3K27ac and H3K4me1 histone marks for candidate and reassigned t(12;21) vlincRNA promoters.

<https://doi.org/10.1371/journal.pone.0207250.g006>

We demonstrated that vlincRNA expression is cALL subtype-specific, but the question remains about the function of these molecules in normal cells and their role in cancer etiology and progression. Silencing experiments performed on vlincRNAs expressed in K562 chronic myeloid leukemia cells resulted in an increase in apoptosis, particularly for transcripts that were more broadly expressed across cell types, suggesting that cell type-specific vlincRNAs have more specialized roles [9]. Others have shown that vlincRNAs are involved in senescence control [10]. Although further experiments would be required to confirm this, we speculate that cALL subtype-specific vlincRNAs could have distinct biological roles between subtypes. Other cALL vlincRNAs that are more broadly expressed across cALL subtypes or other cancer cell types could also play a role in proliferation or apoptosis; again, additional work is required to assess their roles.

In conclusion, we have identified cALL subtype-specific vlincRNAs transcripts whose expression is controlled by well-known epigenetic mechanisms and modulators. In addition to protein-coding genes and long non-coding RNAs, we have shown that vlincRNA expression profiles can accurately classify cALL subtypes.

## Materials and methods

### Sample cohort

Study samples consist of 102 cALL patients (90 pre-B and 12 pre-T, of which 57 pre-B and 7 pre-T belong to the RNA-seq “discovery cohort”) from the established Quebec Childhood ALL cohort [19] (S1 Table) and 4 control samples (3 pre-B CD10<sup>+</sup>CD19<sup>+</sup> and 1 pre-T CD3<sup>+</sup>) isolated from human cord blood. Patients (50 females and 52 males) aged from 1–17 years (median 5.54 years) were diagnosed in the Division of Hematology-Oncology at the Sainte-



Justine UHC, Montreal, Canada, between 1994 and 2017. They underwent treatment with Dana Farber Cancer Institute (DFCI) ALL Consortium protocols DFCI 91–01, 95–01, 2000–01, 2005–01, or 2011–01 [20]. Mononuclear cells were isolated from bone marrow (BM) aspirates or peripheral blood cells at diagnosis and contained a mean level of 90% leukemic blasts. The Sainte-Justine Institutional Review Board approved the research protocol, and written informed consent was obtained from all participating individuals or their parents.

## RNA sequencing

Total RNA was extracted from white blood cell pellets obtained from bone marrow or peripheral blood tissue, followed by a DNase treatment to remove possible genomic DNA contamination. The discovery and control samples were sequenced on the SOLiD System (Life Technologies) following standard manufacturer protocols. Briefly, ribosomal RNAs were removed using the RiboMinus Eukaryote kit (Invitrogen), cDNA libraries for all samples were prepared using the SOLiD total RNA-seq kit (Applied Biosystems), RNA was fragmented into 100–200bp fragments, purified and ligated to SOLiD adapters, fragments were then reverse-transcribed and size-selected, cDNAs were enriched by PCR and purified, Clonally amplified beads were enriched and subjected to paired-end whole transcriptome sequencing (50bp x 35bp or 75bp x 35bp). Reads were aligned to the human genome (hg19) using the Lifescope Genomic Analysis Software (Applied Biosystems). Replication samples and the Reh cell line were sequenced on the Illumina platform). Briefly, ribosomal RNAs were removed using the Ribo-Zero human/mouse/rat included in the TruSeq Stranded Total RNA LT sample preparation kit (Illumina Cat#RS-122-2201) as per manufacturer's protocol. Final libraries were quality controlled on a Bioanalyzer and underwent 75 or 100bp paired-end sequencing on Illumina HiSeq2500/4000 systems. Reads were aligned to the human genome (hg19) using STAR v2.4.2 [21]. All sequencing experiments were done at the Integrated Centre for Pediatric Clinical Genomics, Sainte-Justine UHC, Montreal, Canada. Whole transcriptome datasets are available on the Gene Expression Omnibus (GEO) under accession number GSE89071.

## VlincRNA identification and expression profiling

The vlincRNA identification procedure was derived from St-Laurent *et al.* [9]. Briefly, 68 SOLiD strand-specific total RNA-seq datasets (64 samples and 4 controls) were pooled, split by chromosome and filtered for mapping quality  $\geq 40$  using Picard (Broad Institute) and SAMtools [22]. Strand-specific base coverage was obtained using BEDTools' genomeCoverageBed v2.25.0 [23]. Covered bases overlapping UCSC or RefSeq protein coding genes from the UCSC Table Browser (April 2016) were removed. Bases having less than Q3 (75%) read density or overlapping blacklisted regions (~170Mb total) were removed. Blacklisted regions consisted of ENCODE's EncodeDacMapabilityConsensus [15], svelter's exclude file [24] and canvas' filter file [25]. Remaining covered bases were merged if less than 500bp apart and the resulting segments were merged if less than 10kb apart. Finally, strand-specific vlincRNAs were defined as segments longer than 50kb ( $n = 971$ ). A subset of these, 'far from genes' ( $n = 288$ ), were defined as being at least 10 kb away from protein coding genes regardless of strand orientation. Expression read counts for the SOLiD datasets ( $n = 68$ ) and the Illumina datasets ( $n = 35$ ) were obtained using htseq-count v0.6.1p1 [26]. Autosomal vlincRNAs having at least 100 reads per sample in minimum 3 samples ( $n = 256$ ) were used to perform differential expression analyses on the SOLiD dataset using the DESeq2 likelihood ratio test [27] on subtypes, with sequencing batch and one sva variable as covariates (Fig 2A and 2B) [28]. Expression cluster purity was calculated on eight clusters for the discovery cohort and five clusters for the replication cohort (cluster numbers returned by mclust [29]).

### Chromatin immunoprecipitation (ChIP) sequencing

ChIP-sequencing was performed on bone marrow aspirates or peripheral blood mononuclear cells of a pool of two t(12;21) cases using the previously described procedure [30]. Briefly, the DNA of 10 to 30 million cells was immunoprecipitated using the following antibodies: H3K4me1 (Abcam; ab8895), H3K4me3 (Diagenode; pAb-003-050), and H3K27ac (Abcam; ab4729). Library preparation for ChIP-seq assays was carried out using the Paired-End DNA Sample Prep Kit V1 (Illumina; PE-102-1001) and sequenced using the HiSeq2000 sequencing system (2 x 100bp) at the McGill University and Genome Quebec Innovation Center. Reads were aligned with bwa v.0.6.1 [31] and peaks were called using macs2 [32] with an input control and the “*-broad*” option.

### VlincRNA promoter epigenetic analyses

Candidate promoters of vlincRNAs were defined as  $\pm 10$ kb around the start of expression. Promoter coordinates were refined in the t(12;21) [ETV6/RUNX1] subtype by taking the corresponding ChIP-seq peaks of the H3K4me3, H3K4me1, and H3K27ac chromatin marks and using them as input in ChromHMM v.1.11 [33] to distribute chromatin into 8 states. Active promoter regions were defined as having the H3K4me3 mark present ( $n = 19313$ , total size = 35.9Mb, median size = 1.4kb; Fig 4A). Regions separated by less than 1kb were merged and those longer or equal to 400bp were retained. The longest active regions overlapping candidate promoter regions were retained and reassigned as being active promoters (Figs S3 and 6A and 6C). Active CD19<sup>+</sup> B cord chromatin regions were obtained from the Roadmap Epigenomics Project [14] by keeping states having the H3K4me3 mark present (states 1,2,3,10, and 11) from the 15-state model. The background distribution of active overlapping regions was obtained by generating a thousand random genomic intervals based on t(12;21) [ETV6/RUNX1] putative promoter-active region sizes (Fig 4B). The relationship between expression and overlap of active promoters was obtained by grouping vlincRNAs into t(12;21) [ETV6/RUNX1] expression quartiles and reporting the percentage of overlap. Read density plots and heatmaps of chromatin marks around putative and active vlincRNA promoters were obtained using ngsplot v2.47 [33]. Motif analyses were done with HOMER [34] using the putative promoters ( $\pm 10$ kb around the start of transcription) of the high expressed vlincRNA quartiles in the t(12;21) [ETV6/RUNX1] and HHD subtypes.

### Whole genome bisulfite sequencing (WGBS)

WGBS was performed on bone marrow aspirates or peripheral blood mononuclear cells of three t(12;21) [ETV6/RUNX1] cases and on one human cord blood CD10<sup>+</sup>CD19<sup>+</sup> control using a previously described procedure [34]. Briefly, WGBS gDNA library preparations were carried out using the TruSeq DNA Sample Prep Kit v2 (Illumina) with an added bisulfite conversion step. Amplified libraries were validated and quantified on Bioanalyzer High Sensitivity DNA Chips and underwent 100bp paired-end sequencing on the Illumina HiSeq2000 system at the McGill University and Genome Quebec Innovation Center. Reads were aligned to a bisulfite-converted reference genome using bwa and methylation calls were obtained using ngtgen-utils [35]. Mean coverage of the three t(12;21) cases were 6X, 13X, and 20X and mean coverage of the pre-B CD10<sup>+</sup>CD19<sup>+</sup> control was 18X (bisulfite conversion rates > 99.5%). The C/T read counts of the three t(12;21) cases were merged to create a single 39X high coverage dataset.

## 450K methylation arrays

IDAT files measuring DNA methylation at single CpG-site resolution from the Illumina Infinium HumanMethylation450 BeadChips Array were generated in-house and obtained from previous studies [12, 13]. Samples are four healthy B cell stages: mpp (n = 5), preB-I (n = 6), preB-II (n = 7) and immature B (n = 5) and three cALL subtypes: t(12;21) (n = 61), HHD (n = 82) and 'Other' (n = 76). IDAT Files were processed using minfi [36] with funnorm normalization [37]. Failed probes (detection p-value > 0.01 in more than 20% of samples), probes on sex chromosomes and probes overlapping SNPs were removed. Beta values were used as estimates of methylation levels.

## Methylation analyses

WGBS methylation values upstream (-10kb), downstream (+10kb) and on active t(12;21) [*ETV6/RUNX1*] promoters were obtained by splitting the regions into ten equal-size bins. Methylation values of covered CpGs overlapping the bins were averaged. Boxplots and hierarchical clustering of promoter methylation levels were obtained by averaging methylation values of covered CpGs across the promoters. For 450K datasets, beta values of probes overlapping active t(12;21) [*ETV6/RUNX1*] promoters (34 promoters, 161 probes) were averaged to obtain one methylation value per promoter per sample.

## Data access

Whole transcriptome, ChIP-seq and WGBS datasets are available on the Gene Expression Omnibus (GEO) under accession numbers GSE89071 and GSE120677.

## Supporting information

**S1 Fig. Discovery of cALL vlincRNAs.** Genome wide tracks illustrating the discovery procedure of vlincRNAs in cALL samples. Blue = forward strand, pink = reverse strand. (a,b) Strand specific RNA-seq read coverage of pooled discovery samples (n = 68, SOLiD platform). (c) RefSeq gene annotations. (d,e) Strand specific protein coding gene coordinates of merged UCSC and RefSeq annotations. (f,g) Strand specific bases covered by at least one read using genomeCoverageBed from bedtools. (h,i) Strand specific covered bases having more than 75% read density and not overlapping protein coding genes or blacklisted regions. (j) Strand specific transcripts longer than or equal to 50 kb after merging covered bases less than 500 bp apart and merging resulting segments less than 10 kb apart.

(TIF)

**S2 Fig. Subtype-specific classification of replication cALL samples using vlincRNA expression profiles.** (a) PCA plot of the replication samples (Illumina platform, n = 35) using the DESeq2 regularized log transform (rld) normalized expression of minimally expressed cALL vlincRNAs (n = 273). (b) Hierarchical clustering of the replication samples using Euclidean distance on vlincRNA normalized rld expression values. Cluster purity = 0.94 using 5 clusters. (c) Subtype-specific Pearson correlations of vlincRNA mean log<sub>2</sub> normalized expression across discovery and replication samples (the t(9;22)-HHD sample was not included in this analysis). (d) Normalized rld expression of the top five t(12;21)-specific vlincRNAs in REH cells. t(12;21)-specific vlincRNAs were determined as having a minimum of 2 fold change higher expression than in the other subtypes and sorting the fold change in descending order.

(TIF)

**S3 Fig. Redefining coordinates of t(12;21) active promoters using histone mark ChIP-seq data.** (a,b) Overlay of strand specific RNA-seq normalized read coverage of the t(12;21) Illumina samples (n = 22, + and—strands respectively). (c,d,e) Normalized read coverage of the H3K4me3, H3K27ac and H3K4me1 ChIP-seq histone marks of the t(12;21) pool sample. (f,g) WGBS methylation levels of the merged t(12;21) cases (n = 3) and the CD10<sup>+</sup>CD19<sup>+</sup> control sample respectively. (h) VlnCRNA transcripts (n = 256) discovered in the cALL discovery samples (n = 68). (i) RefSeq gene annotations. (j) Active chromatin regions containing the H3K4me3 mark from ChromHMM. (k) Candidate promoter regions  $\pm 10$  kb around the 5' start of vlincRNAs. (l) Redefined promoter coordinates by keeping the largest active chromatin regions overlapping the candidate promoters. A third (30.9%; 79 / 256) of the candidate vlincRNA promoters are defined as active.

(TIF)

**S4 Fig. Pearson correlations of 450K methylation levels of t(12;21) active promoters in four normal B cell stages and three cALL subtypes.** Pairwise Pearson correlations of 450K methylation levels of t(12;21) active promoters (n = 34) in four healthy B cell stages (mpp, preB-I, preB-II, immature B) and three cALL subtypes (t(12;21), HHD, 'Other') from both in-house and public datasets (242 samples total). 450K beta values were used as methylation levels. For each promoter, methylation levels were obtained by averaging the values of all overlapping 450K probes. For more than half of t(12;21) active promoters (57%; 45 / 79), no probes overlapped. Circle size increases as Pearson correlation decreases.

(TIF)

**S5 Fig. Transcription factor motif enrichment in candidate promoters of highly expressed t(12;21) and HHD vlincRNAs.** Significantly enriched motifs (q-value < 0.05) in candidate promoters ( $\pm 10$  kb around the 5' start of vlincRNAs) of t(12;21) and HHD high expressed quartiles vlincRNAs (Q4, n = 64).

(TIF)

**S1 Table. Study samples.**

(XLSX)

**S2 Table. List of vlincRNAs discovered in the cALL samples.**

(XLSX)

## Acknowledgments

The authors are indebted to the patients and their parents for participating in this study. Patient tissue samples were provided by the Sainte-Justine UHC Pediatric Cancer Biobank. RNA-seq was performed at the Integrated Clinical Genomics Centre In Pediatrics, CHU Sainte-Justine. Computations were made on the supercomputer Briarée from Université de Montréal, managed by Calcul Québec and Compute Canada. We thank Joseph L. Wiemels and Seung-Tae Lee for sharing their WGBS and 450K methylation data.

## Author Contributions

**Conceptualization:** Maxime Caron, Simon Drouin, Daniel Sinnett.

**Data curation:** Maxime Caron, Pascal St-Onge, Chantal Richer, Thomas Sontag, Stephan Busche, Guillaume Bourque, Tomi Pastinen.

**Formal analysis:** Maxime Caron, Pascal St-Onge.

**Funding acquisition:** Simon Drouin, Daniel Sinnett.

**Investigation:** Maxime Caron, Pascal St-Onge, Daniel Sinnett.

**Methodology:** Maxime Caron, Simon Drouin, Daniel Sinnett.

**Project administration:** Simon Drouin, Daniel Sinnett.

**Resources:** Simon Drouin, Thomas Sontag, Stephan Busche, Guillaume Bourque, Tomi Pastinen, Daniel Sinnett.

**Supervision:** Simon Drouin, Daniel Sinnett.

**Validation:** Maxime Caron, Pascal St-Onge, Chantal Richer.

**Visualization:** Maxime Caron.

**Writing – original draft:** Maxime Caron, Simon Drouin, Daniel Sinnett.

**Writing – review & editing:** Maxime Caron, Pascal St-Onge, Simon Drouin, Chantal Richer, Stephan Busche, Daniel Sinnett.

## References

1. Woo JS, Alberti MO, Tirado CA. Childhood B-acute lymphoblastic leukemia: a genetic update. *Exp Hematol Oncol*. 2014; 3:16. <https://doi.org/10.1186/2162-3619-3-16> PMID: 24949228.
2. Mullighan CG. Molecular genetics of B-precursor acute lymphoblastic leukemia. *J Clin Invest*. 2012; 122(10):3407–15. <https://doi.org/10.1172/JCI61203> PMID: 23023711.
3. Lilljebjorn H, Henningsson R, Hyrenius-Wittsten A, Olsson L, Orsmark-Pietras C, von Palffy S, et al. Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nat Commun*. 2016; 7:11790. <https://doi.org/10.1038/ncomms11790> PMID: 27265895.
4. Casero D, Sandoval S, Seet CS, Scholes J, Zhu Y, Ha VL, et al. Long non-coding RNA profiling of human lymphoid progenitor cells reveals transcriptional divergence of B cell and T cell lineages. *Nat Immunol*. 2015; 16(12):1282–91. <https://doi.org/10.1038/ni.3299> PMID: 26502406.
5. Fernando TR, Rodriguez-Malave NI, Waters EV, Yan W, Casero D, Basso G, et al. LncRNA Expression Discriminates Karyotype and Predicts Survival in B-Lymphoblastic Leukemia. *Mol Cancer Res*. 2015; 13(5):839–51. <https://doi.org/10.1158/1541-7786.MCR-15-0006-T> PMID: 25681502.
6. Ouimet M, Drouin S, Lajoie M, Caron M, St-Onge P, Gioia R, et al. A childhood acute lymphoblastic leukemia-specific lncRNA implicated in prednisolone resistance, cell proliferation, and migration. *Oncotarget*. 2016. <https://doi.org/10.18632/oncotarget.13936> PMID: 27980230.
7. Lajoie M, Drouin S, Caron M, St-Onge P, Ouimet M, Gioia R, et al. Specific expression of novel long non-coding RNAs in high-hyperdiploid childhood acute lymphoblastic leukemia. *PLoS One*. 2017; 12(3):e0174124. <https://doi.org/10.1371/journal.pone.0174124> PMID: 28346506.
8. Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, et al. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol*. 2010; 8:149. <https://doi.org/10.1186/1741-7007-8-149> PMID: 21176148.
9. St Laurent G, Shtokalo D, Dong B, Tackett MR, Fan X, Lazorthes S, et al. VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. *Genome Biol*. 2013; 14(7):R73. <https://doi.org/10.1186/gb-2013-14-7-r73> PMID: 23876380.
10. Lazorthes S, Vallot C, Briois S, Aguirrebengoa M, Thuret JY, St Laurent G, et al. A vlincRNA participates in senescence maintenance by relieving H2AZ-mediated repression at the INK4 locus. *Nat Commun*. 2015; 6:5971. <https://doi.org/10.1038/ncomms6971> PMID: 25601475.
11. Lee ST, Muench MO, Fomin ME, Xiao J, Zhou M, de Smith A, et al. Epigenetic remodeling in B-cell acute lymphoblastic leukemia occurs in two tracks and employs embryonic stem cell-like signatures. *Nucleic Acids Res*. 2015; 43(5):2590–602. <https://doi.org/10.1093/nar/gkv103> PMID: 25690899.
12. Busche S, Ge B, Vidal R, Spinella JF, Saillour V, Richer C, et al. Integration of high-resolution methylation and transcriptome analyses to dissect epigenomic changes in childhood acute lymphoblastic leukemia. *Cancer Res*. 2013; 73(14):4323–36. <https://doi.org/10.1158/0008-5472.CAN-12-4367> PMID: 23722552.



13. Lee ST, Xiao Y, Muench MO, Xiao J, Fomin ME, Wiencke JK, et al. A global DNA methylation and gene expression analysis of early human B-cell development reveals a demethylation signature and transcription factor network. *Nucleic Acids Res.* 2012; 40(22):11339–51. <https://doi.org/10.1093/nar/gks957> PMID: 23074194.
14. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenyk M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015; 518(7539):317–30. <https://doi.org/10.1038/nature14248> PMID: 25693563.
15. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012; 489(7414):57–74. <https://doi.org/10.1038/nature11247> PMID: 22955616.
16. St Laurent G, Wahlestedt C, Kapranov P. The Landscape of long noncoding RNA classification. *Trends Genet.* 2015; 31(5):239–51. <https://doi.org/10.1016/j.tig.2015.03.007> PMID: 25869999.
17. Haddy TB, Mosher RB, Reaman GH. Late effects in long-term survivors after treatment for childhood acute leukemia. *Clinical pediatrics.* 2009; 48(6):601–8. <https://doi.org/10.1177/0009922809332680> PMID: 19264722.
18. Mody R, Li S, Dover DC, Sallan S, Leisenring W, Oeffinger KC, et al. Twenty-five-year follow-up among survivors of childhood acute lymphoblastic leukemia: a report from the Childhood Cancer Survivor Study. *Blood.* 2008; 111(12):5515–23. <https://doi.org/10.1182/blood-2007-10-117150> PMID: 18334672.
19. Healy J, Belanger H, Beaulieu P, Lariviere M, Labuda D, Sinnott D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. *Blood.* 2007; 109(2):683–92. <https://doi.org/10.1182/blood-2006-02-003236> PMID: 17008550.
20. Silverman LB, Stevenson KE, O'Brien JE, Asselin BL, Barr RD, Clavell L, et al. Long-term results of Dana-Farber Cancer Institute ALL Consortium protocols for children with newly diagnosed acute lymphoblastic leukemia (1985–2000). *Leukemia.* 2010; 24(2):320–34. <https://doi.org/10.1038/leu.2009.253> PMID: 20016537.
21. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013; 29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635> PMID: 23104886.
22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009; 25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352> PMID: 19505943.
23. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010; 26(6):841–2. <https://doi.org/10.1093/bioinformatics/btq033> PMID: 20110278.
24. Zhao X, Emery SB, Myers B, Kidd JM, Mills RE. Resolving complex structural genomic rearrangements using a randomized approach. *Genome Biol.* 2016; 17(1):126. <https://doi.org/10.1186/s13059-016-0993-1> PMID: 27287201.
25. Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics.* 2016; 32(15):2375–7. <https://doi.org/10.1093/bioinformatics/btw163> PMID: 27153601.
26. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics.* 2015; 31(2):166–9. <https://doi.org/10.1093/bioinformatics/btu638> PMID: 25260700.
27. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014; 15(12):550. <https://doi.org/10.1186/s13059-014-0550-8> PMID: 25516281.
28. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* 2012; 28(6):882–3. <https://doi.org/10.1093/bioinformatics/bts034> PMID: 22257669.
29. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. *Bioinformatics.* 2001; 17(10):977–87. PMID: 11673243.
30. Adoue V, Schiavi A, Light N, Almlof JC, Lundmark P, Ge B, et al. Allelic expression mapping across cellular lineages to establish impact of non-coding SNPs. *Mol Syst Biol.* 2014; 10:754. <https://doi.org/10.15252/msb.20145114> PMID: 25326100.
31. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25(14):1754–60. <https://doi.org/10.1093/bioinformatics/btp324> PMID: 19451168.
32. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008; 9(9):R137. <https://doi.org/10.1186/gb-2008-9-9-r137> PMID: 18798982.

33. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*. 2012; 9(3):215–6. <https://doi.org/10.1038/nmeth.1906> PMID: 22373907.
34. Busche S, Shao X, Caron M, Kwan T, Allum F, Cheung WA, et al. Population whole-genome bisulfite sequencing across two tissues highlights the environment as the principal source of human methylome variation. *Genome Biol*. 2015; 16:290. <https://doi.org/10.1186/s13059-015-0856-1> PMID: 26699896.
35. Johnson MD, Mueller M, Game L, Aitman TJ. Single nucleotide analysis of cytosine methylation by whole-genome shotgun bisulfite sequencing. *Curr Protoc Mol Biol*. 2012; Chapter 21:Unit21 3. <https://doi.org/10.1002/0471142727.mb2123s99> PMID: 22870857.
36. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014; 30(10):1363–9. <https://doi.org/10.1093/bioinformatics/btu049> PMID: 24478339.
37. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol*. 2014; 15(12):503. <https://doi.org/10.1186/s13059-014-0503-2> PMID: 25599564.