Identification of epigenetic modifications that contribute to pathogenesis in therapy-related AML: Effective integration of genome-wide histone modification with transcriptional profiles

Xinan (Holly) Yang*, Bin Wang, John M Cunningham

From The 4th Translational Bioinformatics Conference and the 8th International Conference on Systems Biology (TBC/ISB 2014)
Qingdao, China. 24-27 October 2014

Abstract

Background: Therapy-related, secondary acute myeloid leukemia (t-AML) is an increasingly frequent complication of intensive chemotherapy. This malignancy is often characterized by abnormalities of chromosome 7, including large deletions or chromosomal loss. A variety of studies suggest that decreased expression of the EZH2 gene located at 7q36.1 is critical in disease pathogenesis. This histone methyltransferase has been implicated in transcriptional repression through modifying histone H3 on lysine 27 (H3k27). However, the critical target genes of EZH2 and their regulatory roles remain unclear.

Method: To characterize the subset of EZH2 target genes that might contribute to t-AML pathogenesis, we developed a novel computational analysis to integrate tissue-specific histone modifications and genome-wide transcriptional regulation. Initial integrative analysis utilized a novel "seq2gene" strategy to explore largely the target genes of chromatin immunoprecipitation sequencing (ChIP-seq) enriched regions. By combining seq2gene with our Phenotype-Genotype-Network (PGNet) algorithm, we enriched genes with similar expression profiles and genomic or functional characteristics into "biomodules".

Results: Initial studies identified SEMA3A (semaphoring 3A) as a novel oncogenic candidate that is regulated by EZH2-silencing, using data derived from both normal and leukemic cell lines as well as murine cells deficient in EZH2. A microsatellite marker at the SEMA3A promoter has been associated with chemosensitivity and radiosensitivity. Notably, our subsequent studies in primary t-AML demonstrate an expected up-regulation of SEMA3A that is EZH2-modulated. Furthermore, we have identified three biomodules that are co-expressed with SEMA3A and up-regulated in t-AML, one of which consists of previously characterized EZH2-repressed gene targets. The other two biomodules include MAPK8 and TATA box targets. Together, our studies suggest an important role for EZH2 targets in t-AML pathogenesis that warrants further study.

Conclusion: These developed computational algorithms and systems biology strategies will enhance the knowledge discovery and hypothesis-driven analysis of multiple next generation sequencing data, for t-AML and other complex diseases.

* Correspondence: xyang2@uchicago.edu
Section of Hematology/Oncology, Dept. of Pediatrics, Comer Children’s Hospital University of Chicago, Chicago, USA

© 2015 Yang et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Introduction

The significance of non-coding DNA regulators in human disease has drawn increasing attention. For example, human non-coding regions contain collections of transcription factor binding sites and other regulatory elements called “cis-regulatory” regions. These cis-regulatory elements are sufficient to activate transcription in a defined spatial and temporal expression domain [1]. Cis-regulation can occur on either side of a target transcript and regulators can reside far from their regulatory targets [2]. However, identifying cis-regulatory elements and their domain-specific targets remains a major challenge for current computational biology. To address the challenge, we here perform a “sequence-regulator-network” study to integrate information from histone modification and transcriptional regulation. This method both generates and validates genomic hypotheses, and could have a broad impact in studying regulatory mechanisms of gene expression in systems biology. Here, we select therapy-related acute myeloid leukemia (t-AML) as a clinically significant context to apply the method.

T-AML, including therapy-related myelodysplastic syndrome, accounts for approximately 10 to 20 percent of myeloid malignancies [3]. T-AML complicates conventional chemoradiotherapies that are used to treat a variety of primary malignancies and is associated with a uniformly poor prognosis, with a median survival of six months [3]. Complete loss of chromosome 7 (-7) and 5 or partial deletion involving the long arm of chromosome 7 (del7q) are highly recurrent chromosomal aberrations in AML and t-AML.

Specific interest has focused on the link between common chromosome 7 abnormalities and the location of the EZH2 gene, the histone methyltransferase enhancer of Zeste homologue 2 present at position 7q36.1 [4]. Not surprisingly, EZH2 expression is significantly reduced in -7/del7q patients with myeloid disorders when compared with healthy controls [5]. What remains an enigma is the recently reported dual role of EZH2 in malignant cell development. EZH2 is a component of the polycomb group complex, which is vital for hematopoietic cell development. In normal cells, EZH2 suppresses its targets through depositing the histone modification mark H3K27me3 (trimethylation on lysine 27 of histone H3) [6]. In several epithelial cancers, overexpression of wild-type EZH2 has been found to promote tumor progression or metastasis [7,8]. However, inactive mutated EZH2 or its low expression in myeloid malignancy contributes to tumorigenesis by suppressing differentiation, thus directing cells toward a leukemic stem cell state [5,9,10]. Conflictingly, Xu et al. reported that EZH2 overexpression was associated with poor patient outcome in myeloid disorders and chemotherapy reduced expression of EZH2 [11].

These conflicting observations about EZH2 suggest a context-specific regulatory mechanism, which may be explained by plastic epigenetic modification [12]. Histone methylation is an important epigenetic modification in chromatin. The histone modification mark H3K27me3 reflects EZH2-involved Polycomb-mediated repression, and the deposition of H3K27me3 is development-stage- and tissue-specific [13]. Thus, we studied the presence of H3K27me3 to understand the conflicting roles of EZH2 – as either an oncogene or a tumor suppressor in different tumors. However, abnormal EZH2-associated regulation to particular target genes remains unclear in leukemia, specifically in t-AML.

Using ChIP-seq (Chromatin Immunoprecipitation Sequencing) of histone marks and other regulatory proteins, researchers can perform genome-wide searches for intergenic functional elements (including promoters and enhancers), but might also identify non-enhancers with similar signatures [14]. Therefore in this study, we evaluated whether the selective regions control the developmental expression of the target genes using transcriptomic measurements. To identify EZH2 target genes and their functional regions in t-AML, we developed a novel computational integrative analysis with histone modification of H3K27me3 and gene expression.

In the proposed integrative analysis, there are three distinguishing features: 1) a novel “seq2gene” strategy links genomic regions to more neighboring coding genes on both sides, 2) selective transcriptional and epigenetic data mining between cells, and 3) the prediction of “biomodules”. The seq2gene strategy links genomic regions to a broad range of neighboring genes rather than the nearest one. The rationale is that enhancers can target long-range DNA targets, and often multiple enhancers (five or more) target the same genes [14]. The strategy then incorporates epigenetic regulatory patterns that differ between cell lines with disease-specific transcript alterations in t-AML. Finally, the selected target gene (seed) is associated with a group of genes, the “biomodule” that share similar expression patterns and genomic or functional characteristics, using our PGNnet algorithm [15].

This integrative “sequence-regulator-network” study revealed SEMA3A (semaphoring 3A) as a novel target of EZH2-silencing in t-AML. The fact that EZH2 and SEMA3A are inversely expressed in vivo is supported by previous data in mouse hematopoietic stem cells and human prostate cancer [16,17]. We predict that the loss of EZH2 silencing on SEMA3A augments sensitivity to both chemo- and radiotherapy, and thus may contribute to therapy-related AML pathogenesis. We also show that a group of SEMA3A-coexpressed genes, including
HOXA11, are up-regulated in t-AML and have been reported as EZH2 targets. We expect further validation both in vitro and in vivo.

Result

Identification of SEMA3A, a loss-of-EZH2-mediated silencing gene in leukemia

To systematically screen functional elements of histone-modified EZH2 targets in leukemia, we used the data in the Encyclopedia of DNA Elements (ENCODE, genome.ucsc.edu/ENCODE/) Project [18]. We identified 104,370 genomic regions that are enriched for both EZH2 and H3K27me3 in the leukemia cell line (K562) and 53,360 regions in the lymphoblastoid cell line (GM12878). The large number of enriched regions (peaks) suggests substantial downstream effects of EZH2 repression. EZH2 occupancy and presence of H3K27me3 at promoters directly silences the transcription of targeted genes, which has been observed in leukemia and other tumors [19,20]. Given this, we predict that genes adjacent to these non-coding regions are EZH2 repressive target candidates.

To discover target candidates, we associated coding genes residing in a given search radius to the identified regions, using a “seq2gene” mapping strategy (Figure 1A). Seq2gene considers the possibility that genes in both directions from each intergenic cis-regulatory element may fall under control, given the observation that enhancers reside on average 120-thousand base pairs (bps) away from their regulatory targets and act independently of their orientation in mammals [2,14,21]. This consideration resulted in mapping around 90% of EZH2 and H3K27me3 co-mediated loci to neighboring genes within an arbitrary distance of 150k-bp on both sides, of which only 31% are coding genes (Figure 1B, the ENSEMBL Hg19 assembly and definition).

To focus on EZH2-mediated coding-gene silencing that is specific to leukemia, we compared the candidates in leukemia with lymphoblastoids. Only around 10% of the identified ~17,400 EZH2 repressed coding-gene candidates are disease-specific, resulting in 1624 genes specific to lymphoblastoid but not leukemia (Figure 1C). Gene Ontology enrichment analysis suggests a loss of leukemia-specific repressive control on the molecular function termed “hematopoietin/interferon-class (D200 domain) cytokine receptor binding” (GO:0005126, FDR = 0.0017, count = 12), reflecting a cell quiescence-involved, generic cancer metastatic mechanism [22,23].

We subsequently identified SEMA3A as an EZH2 repressive target of interest. This identification is derived from sequence-based analysis and transcriptional evidence (Figure 1D). From the ChIP-seq peaks in the lymphoblastoid cell line, we observed EZH2 occupancy and presence of H3K27me3 adjacent to the transcription start site of SEMA3A and 20 other genes (Additional file: Table S1). These 21 genes significantly over-represent genes highly expressed in prostate cancer cells after knockdown of EZH2 [16] (p = 0.009, OR = 5.8) and genes down-regulated in fibroblasts expressing mutant forms of ERCC3 after UV irradiation [24] (P = 0.0046, OR = 7.1). As ERCC3 could help increase the sensitivity of cancer to radiation therapy, loss of EZH2-repression of these ERCC3 targets in t-AML indicates an increased radiosensitivity. Specifically, the transcriptional expression of SEMA3A is negatively EZH2-dependent in both human cancer and mouse model in vivo. For example, Merchan et al. generated mouse models that allow gain-of-function of Ezh2 in the haematopoietic system [17], and we identified a 2.2-fold decrease of Sema3a expression in Ezh2+ mice compared with wild-types (Q-value = 0.05, the limma test [25]). However, in the leukemia cell line, a peak with both EZH2 and H3K27me3 enrichment has not been observed in the human genomic region within 150kbp distance to SEMA3A, suggesting a leukemia-specific loss of EZH2-silencing on SEMA3A.

High-expression of SEMA3A may contribute to t-AML pathogenesis by augmenting chemosensitivity and radiosensitivity

We observed the loss of EZH2 and H3K27me3 enrichment on the SEMA3A promoter in leukemia (Figure 1C-D) and hypothesized that this loss rescues SEMA3A expression and facilitates leukemogenesis after chemoradiotherapy.

To validate this hypothesis, we investigated CD34+ cells from 28 t-AML patients including 8 with -7/del7q abnormality, and 24 normal controls (Table 1). Expression profiles of samples collected from different laboratories were adjusted for batch effects [26] (Additional file: Fig. S1) and then t-AML samples were compared to normal controls. There are 370 significantly up-regulated and 686 down-regulated genes (Q-value<0.05, FC≥2 or ≤0.5) when comparing -7/del7q t-AML samples with controls.

Significantly, SEMA3A shows up-regulation in patients with t-AML (Q-value = 1.1e-9, FC = 2.3, Figure 2A), even in patients with deletion of chromosome 7 or loss of chromosome 7q (FC = 2.2, Q-value = 4.6e-5, Figure 2B). In contrast, EZH2 was significantly down-regulated in t-AML (Q-value = 2.3e-07 and 0.00032, FC = 0.55 and 0.54, respectively). Besides SEMA3A, there are another 7 genes (AGR2, EVX, HOXA11, MET, PGAM2, BRAF, and UPP1) residing on chromosome 7 are significantly up-regulated in t-AML even with -7/del7q abnormality (Figure 2B-C). Three of them (HOXA11, MET, BRAF) are potential oncogenes currently being observed for common copy-number gains in a meta-analysis of copy number alterations across a panel of different cancer cell lines and tumor samples [27]. Their high expression suggests that a loss of EZH2...
Figure 1 Identification of SEMA3A, a loss-of-EZH2-mediated silencing gene in leukemia. A) The seq2gene strategy links genomic regions to coding genes with annotation. B) For both cell lines, more than 90% of the EZH2 and H3K27me3 co-binding regions are linked to intergenic coding and non-coding genes (red arc and box), 31% of which are coding-genes. C) Leukemic and lymphoblastic cell-specific EZH2 and H3K27me target genes. D) SEMA3A loses EZH2-silencing in leukemia but with EZH2-dependent expression changes in two independent experiments (the white circles). The gray circle represents 1624 predicted co-binding sites specific to lymphoblastoid cells identified in panel C.
regulation dominates the expression changes of SEMA3A and HOXA11.

More importantly, significantly higher expressed SEMA3A was previously reported in chemosensitive cancers than in chemoresistant tumors [28, 29]. Additionally, an identified EZH2-H3K27me3-enriched promoter region of SEMA3A (Chr7:83,814,596-83,835,002) covers a microsatellite marker that is significantly associated with acute adverse effects following radiotherapy in cancer patients [30] (Figure 3A). Note that the Phylop score, corresponding to cross-species genome conservation [31], is relatively high within this marker (Figure 3B), suggesting that it is functionally important. This observation agrees with the previous finding that biochemical, evolutionary, and genetic approaches provide complementary information for defining functional DNA segments [32]. This intergenic region (D7S0338i, chr7:83,825,594-83,825,895, Hg19 assembly) is 1.5k-bp upstream of the transcription start site of the SEMA3A gene. Evidence from normal skin cells has proven that SEMA3A knockdown enhances radiation resistance, suggesting an increased radiosensitivity with loss-of-silencing on SEMA3A in leukemia [30].

The stem cell self-renewal HOX gene family has been described as a major downstream target of EZH2. Unlike HOXA11, the other four HOX family genes (HOXA3, HOXA4, HOXA5, and HOXA9) are down-regulated when comparing -7/del7q t-AML cells against controls (Q-value<0.001, fold change (FC)<0.4, Figure 2C). Of note, sublethally irradiated Hoxa9-/- mice exhibited prolonged suppression of hematopoiesis and developed persistent pancytopenia [33], indicating an enhanced sensitivity to ionizing irradiation in t-AML cells with deficient HOXA9.

**Genes co-regulated with SEMA3A in t-AML are enriched in EZH2 repressed targets**

To further study SEMA3A function in t-AML, we investigated the functional enrichment among genes sharing an expression pattern with SEMA3A. Previously, we developed a phenotype-genotype network analysis (PGNet) algorithm to define a group of genes that share significant concurrence of expression pattern with respect to sample grouping (a phenotype of interest) and gene regulation (a genotype of interest) [15]. The PGNet algorithm was successfully applied to identify epigenetic regulators, despite the fact that transcriptional signatures of epigenetic regulation is subtle [34, 35], thus vetting the method for our similar such application. Using PGNet, we defined 66 genes that meet two criteria among 28 patients and 24 controls: 1) systematic co-expression with SEMA3A, and 2) higher expression in t-AML than in normal samples (among the top 150 ranks for both statistics). The similarity of the orders of these two gene expression statistics are significant at the top ranks but not the bottom ranks (empirical p = 0.037, permutation times = 1000). Specifically displayed in Figure 4A are the numbers of genes in which these two top ranks overlap. The overlap size is drawn as a step function over the respective ranks. Top ranks correspond to up-regulation in t-AML and positive correlation with SEMA3A, and bottom ranks correspond to down-regulation and negative correlation. The PGNet algorithm also identified 63 genes when comparing -7/del7q t-AML patients (n = 8)

---

**Table 1. Six studies pertaining to CD34+ cells in t-AML and normal controls**

| Journal  | Platform | PMID   | GSE24006 | GSE3077 | GSE17054 | E-TABM-978 | Qian | GSE23025 |
|----------|----------|--------|----------|---------|----------|------------|------|----------|
| leukemia | Hgu133+2 | 21177505 | 2011 | 2011 | 2011 | 2011 | 2011 | 2011 |
| Nat. Med | Hgu133a  | 21873988 | 2009 | 2009 | 2009 | 2009 | 2009 | 2009 |
| PNAS     | Hgu133+2 | 19218430 | 2011 | 2011 | 2011 | 2011 | 2011 | 2011 |
| Cancer Cell | HsHT-12 | 21251617 | 2011 | 2011 | 2011 | 2011 | 2011 | 2011 |
| Cancer Cell | Hgu99av2 | 12417757 | 2011 | 2011 | 2011 | 2011 | 2011 | 2011 |
| Cancer Cell | Hgu133+2 | 22094254 | 2011 | 2011 | 2011 | 2011 | 2011 | 2011 |

---

* Among 3 author collected samples, the one from a patient with breast cancer was excluded and the other two were included.

PBSC: peripheral blood stem cells; BM: bone marrow samples
to controls and modifying the second criteria appropriately (top 150 for both statistics, empirical \( p = 0.024 \)). Of note, 52 genes overlap between the two defined gene-sets (Figure 4B, Additional file: Table S2), suggesting a common transcriptional regulatory mechanism in t-AML with or without chromosome 7 loss.

Importantly, we found three functional biomodules among the 52 genes correlated with \( \text{SEMA3A} \) (Fisher’s exact test, \( p \leq 0.001 \), count>5, Figure 4C, Table 2). One module includes nine genes (\( \text{EIF2AK2, GULP1, CAP2, GJA1, SEMA3A, KRT34, SERPINB7, BBC3, KLK10} \), \( p = 0.001 \)) that were up-regulated in PC3 cells (prostate cancer) after knockdown of EZH2 by RNAi [16]. Shown in Figure 4B2, the expression patterns of these EZH2-repressed genes (blue lines) are correlated with \( \text{SEMA3A} \) (red line) and are higher in t-AML than in controls. These genes constitute a biomodule representing the expression pattern of a loss of EZH2-silencing. Another identified module is 12 TATA box binding protein genes with promoter regions (defined as ±2kb around the transcription start site) containing the motif TATAAA, including \( \text{SEMA3A} \) and \( \text{HOX11A} \) (\( p = 4.6e-4 \), Figure 4C). The HOX gene family is recognized as a major downstream target of EZH2 [12]. This supports the notion of \( \text{SEMA3A} \)-correlated \( \text{HOX11A} \) up-regulation in t-AML.

The third and most significant biomodule is 13 genes up-regulated in vascular smooth muscle cells by \( \text{MAPK8} \) (\( p = 1.6e-5 \)) [36]. \( \text{MAPK8} \), also known as \( \text{JNK1} \), encodes many...
transcripts and is activated by various cell stimuli. The bio-
module of MAPK8-induced and t-AML highly expressed
genes supports the observation that MAPK8 is involved in
carcinogenesis [37]. The regulatory mechanism linking
MAPK8 and SEMA3A remains unclear. Additionally, a
previous study shows that the EZH2 "loss-of-function"
mutation contributes to formation of the leukemic stem

cell by mediating self-renewal of myeloid progenitors [9],
indicating that the loss of EZH2 silencing on SEMA3A
contributes to leukemogenesis in t-AML.

Method
Data
The chromatin immunoprecipitation sequencing (ChIP-
seq) peaks for EZH2 occupancy and presence of H3K27me3 were downloaded from the Encyclopedia of
DNA Elements (ENCODE, Release 3, hg19 assembly)
[38]. We focused on human blood cell lines: K562 from
leukemia and GM12878 from lymphoblastoid cell of a
female donor.

We collected published expression levels of CD34+
progenitor cells from 28 t-AML patients and 24 healthy
control samples (Table 1) [39-43]. Samples from
patients with breast cancer or lymphoma before the
development of t-AML were excluded from the control
group [39,40].

Functional gene-sets were defined by the MsigDB
database [44].

Identifying target genes of transcription regulators or
histone marks
To define the EZH2 repressed targets in a cell line, we
performed a two-step analysis: 1) intersecting significant
ChIP-seq peaks of EZH2 and H3K27me3 to find the

Figure 3 Genomic view of the SEMA3A promoter. A) The SEMA3A promoter is enriched with EZH2 and H3K27me3 in the lymphoblastoid cell
line (GM12878) but not in the leukemic cell line (K562). B) This promoter region loses EZH2-silencing and covers a microsatellite marker, D7S0338i,
which is significantly associated with acute adverse effects during radiotherapy in cancer patients.
Figure 4 PGNet identifies three biomodules. A) The significance exists among top ranks but not the bottom ranks when comparing orders of two genome-wide gene expression statistics, one is the correlation with SEMA3A and the other is the differential expression between t-AML and control. The black line records the overlap among the top 150 ranks, while the orange line is the estimated overlap by chance. In addition, the expected overlap and 95% confidence intervals derived from a hypergeometric distribution are shown. B) 52 intersecting genes were commonly identified when the second statistic of the PGNet algorithm comparing t-AML to controls or -7/del7q t-AML to controls (subpanel 1). The expression values of these 52 genes (gray) are lines along with t-AML and control samples in subpanel 2, the red line sketches SEMA3A and the blue lines outline the other EZH2 targets in prostate cancer. C) Functional enrichment analysis identifies three sub-sets of genes from SEMA3A and the 52 genes, each sub-set significantly over-represented one MSigDB defined functional gene-sets (p < 0.001, count>5). Genes are marked as circles, and their functional terms are marked as squares. The expression pattern of EZH2 targets is highlighted as blue lines in Fig B2.
common enriched genomic regions; 2) finding candidate target genes and annotating them using the seq2gene strategy (Figure 5). Note that the bisection method is used to perform a binary search among exon and transcript annotations. To perform a search with respect to exon and transcript separately, we have prepared the “exon.table” and “transcript.table” files based on the ENSEMBL general feature format for end users (Figure 5). Both files use ENSEMBL IDs as the key index. This analysis was performed for each cell line separately.

Analyzing transcriptomic data

Data pre-processing. The normalized expression profiles were downloaded from the Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/) or ArrayExpress (www.ebi.ac.uk/arrayexpress/). The raw data from the Qian’s dataset was processed using the global rank-invariant normalization (GRSN) [45]. Gene expression profiles were log2-transformed if authors hadn’t already done so. When collapsing probes to genes for each dataset, probes with the same Entrez IDs were collapsed to the maximum mean expression per gene using an empirically recommended method [46] and the Bioconductor annotation package biomaRt (v2.18.0), resulting in a set of 8442 human Entrez genes measured across all cohorts. The batch effects due to multi-datasets were removed by an empirical Bayes method using the Bioconductor package sva [26,47] (Additional file: Fig. S1). Then a smaller data space of 4221 genes was considered for the following analysis by keeping the half of genes with the highest interquartile range [48].

Identifying biomodule

Step 1 to build biomodules: transcriptional association. To identify the potential upstream regulators or downstream targets of the genes that drive the tumorigenesis in t-AML, we applied the PGNet method [49]. PGNet evaluates the similarity of the gene orders between two independently ordered lists. Specifically, we used PGNet to compare genome-wide correlated expressions with a seed gene and the differential expressions between two sample groups and yielded a regulatory network of genes that are mutually associated. To identify the potential downstream regulatory targets of SEMA3A, we inputted it as a seed gene together with the phenotype information (t-AML: n = 28, normal control: n = 24) into the PGNet system to infer a regulatory network [49]. Differential expression (DE) was estimated using the Bioconductor package Limma [25] and co-expression (CE) was evaluated by the Pearson correlation test on the log-transformed data. The resulting p-values were adjusted by a Q-value for false discovery rate in the multiple testing problem [50,51]. The significance of the similarity between the two statistics (coefficient and fold change) was estimated using an empirical p-value based on permutation (n = 1000), using the Bioconductor package OrderedList [35]. We also tested the differential expression between -q/del(7q) t-AML (n = 8) and normal controls to determine whether a regulatory biomodule for t-AML is consistent despite loss of chromosome 7.

Step 2 to build biomodules: genomic or functional enrichment. The over-representation of functional gene-sets among the identified biomodule was evaluated using a conditional hypergeometric distribution test (Additional file: Fig. S2). Note that only the true background, the common genes within MSigDB and those covered by the experiment, were used for the test. We used a threshold of Q-value<0.001 and count>5 for significance over MSigDB defined functional gene-sets [44].

Discussion

The ENCODE project increasingly produces genomic data on transcription factor binding, chromatin structure and histone modification. Interpreting transcriptional regulation in relation to chromatin modifications has been recognized as a powerful strategy to discover and understand intergenic regulatory elements (reviewed by Kellis et al. [32]). However, ChIP-seq enrichment for chromatin modification and differential expression for transcriptional control may provide complementary information. Thus, the simple overlap strategy, that we here used, may discover only limited targets [52].
expect additional strategies to reveal more candidates than just a single locus to elucidate genomic function in human biology and diseases. To address this challenge, we piloted the application of the PGNet algorithm to build genomic and functional related as well transcriptionally associated gene-sets (biomodules). The identified biomodule of EZH2-suppressed targets that up-regulated in t-AML in conjunction with *SEMA3A*, in turn, evaluated our hypothesis that *SEMA3A* is a critical EZH2-silencing target.

---

**Algorithm**: seq2gene

**Input**: peaks, exon.table, transcript.table, search radius

**Output**: peak with annotated gene information

1. **for** i := 1 to length(peaks) **do begin**:
2.   m = peakleft
3.   n = peakright
4.   middle = (m+n)/2
5.   locate the nearest exon(J) for peak(i) by the basic bisect algorithm
6.   if peak(i) resides inside exon(J)
7.     report peak(i) with exon(J)
8.   **endif**
9.   **while** exon(x) intersecting with peak(i)
10.   report peak(i) with exon(x)
11.   exon(x) = the closest exons (left or right)
12. **endwhile**
13. locate the nearest transcript(H) by the basic bisect algorithm
14. if peak(i) resides outside transcript(H)
15.   report peak(i) with transcript(H), intergenic region *
16. else
17.   report peak(i) with transcript(H), intron region
18. **endif**
19. **for** transcripts(t) within the position of transcript(H)± search radius
20.   if peak(i) resides outside transcripts(t)
21.     report peak(i) with transcripts(x), intergenic region *
22.   else
23.     report peak(i) with transcripts(x), intron region
24. **endif**
25. **end**
26. **end**

*: more details about distance, promoter and bidirectional region judgment

---

*Figure 5 Pseudo code of the seq2gene algorithm*
The reduced EZH2 expression in t-AML parallels reduced expression observed in primary AML or pre-leukemia, which was previously found in 78% of patients carrying either EZH2 inactive mutation or -7/del7q involving the EZH2 locus [9]. Importantly, primary AML patients who have lower EZH2 expression (either spliceosomal mutants or -7/del7q) show decreased H3K27 trimethylation and increased chromatin relaxation at specific gene loci accompanied by higher transcriptional activity [9]. Using the proposed “sequence-regulator-network” strategy, we identified SEMA3A as a new such gene target that loses epigenetically modified EZH2 silencing in t-AML. The identified gene locus covers a radiation sensitivity mark revealed by genome-wide association study [53]. However, HOXA9 was found to be overexpressed in cases of either EZH2 mutations or -7/del7q when compared to EZH2 wild-type [9], which differs from the observation in our t-AML samples. This observation indicates a commonly reduced histone modification and alternative leukemogenic regulation on HOXA9 between t-AML and primary AML. Both involve EZH2 and potentially some of its DNA-binding cofactors. We expect further validation both in vitro and in vivo.

A literature review suggests two possible mechanisms to explain why this epigenetically modified up-regulation of SEMA3A contributes to pathogenesis in therapy-related AML. The first could be a reduction of DNA repair capacity, given that SemA3A suppresses aggressiveness and migration in mice models [54,55] and thus triggers the sensitivity of leukemic cells to apoptosis signal [56], possibly via a MAPK8 regulated pathway (Figure 4C). On the other hand, a previous study found that SemA3A counteracted chemotherapy-induced activation of epithelial-mesenchymal transition (EMT) by improving cancer tissue oxygenation and extending the vascular normalization [57]. Therefore, the second mechanism could be the production of cancer stem cells (CSCs), given that cells that undergo EMT gain stem cell-like properties [58].

In summary, we predicted EZH2-silencing targets and their functions in t-AML by performing a novel computational integrative analysis. The analysis incorporates chromatin-based epigenetic regulation patterns in different cell lines with transcriptional expression alteration pertaining to t-AML. This integrative analysis promises to reveal novel functional elements in a complex and versatile regulatory system behind target gene selection and their tissue-specific expression.

Additional material

Additional file 1: Table S1. The 21 candidate genes predicted in Figure 1D. They exhibit not only microsatellite markers associated with radiosensitivity but also genomic regions enriched with Ezh2 and H3k27me3 in the lymphoblastoid cell line (K562). Genomic loci are based on the hg19/GRCh37 assembly.

Table S2. SEMA3A dependently differentially expressed genes in t-AML. Figure S1. Correction of batch effects. A) There are batch effects when integrating samples from different datasets, showing by the first two principal components derived from all genes. B) The dataset-dependent batch effects are removed after the correction. In both panels, one dot is one sample colored by the datasets. Figure S2. Conditional hypergeometric distribution test. Note that the test uses the common genes (A+B+C+D) covered by both MSigDB and an experiment of interest. FGS: functional gene-set; DE: differentially expressed.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
We thank Joshua Stevens-Stein for his careful proof-reading. We acknowledge the assistance of Lorenzo Pesce for running BEAGLE supported by the NIH under grant [S10 RR029030-01].

Declarations
Publication of this article has been funded by the National Institutes of Health [R21 CA167305]. This article has been published as part of BMC Medical Genomics Volume 8 Supplement 2, 2015: Selected articles from the 4th Translational Bioinformatics Conference and the 8th International Conference on Systems Biology (TBC/ISB 2014). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcmedgenomics/supplements/v8/S2.

Published: 29 May 2015

References
1. Wittkopp PJ, Kalay G: Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergenc. Nat Rev Genet 2012, 13(1):59-69.
2. Visel A, Rubin EM, Pennacchio LA: Genomic views of distant-acting enhancers. Nature 2009, 461(7261):199-205.
3. Godley LA, Lanson RA: Therapy-related myeloid leukemia. Semin Oncol 2008, 35(4):418-429.
4. Ernst T, Chase AJ, Score J, Hidalgo-Curtis CE, Bryant C, Jones AV, Waghorn K, Zoi K, Ross FM, Reiter A, et al: Inactivating mutations of the histone methytransferase gene EZH2 in myeloid disorders. Nat Genet 2010, 42(8):722-726.
5. Jerez A, Sugimoto Y, Makishima H, Verma A, Jankowska AM, Przychodzen B, Visconte V, Tiu RV, O’Keefe CL, Mohamed A, et al: Loss of heterozygosity in 7q myeloid disorders: clinical associations and genomic pathogenesis. Blood 2012, 119(25):e1609-1617.
6. Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, Jones RS, Zhang Y: Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science 2002, 298(5595):1039-1043.
7. Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG, Ghish D, Pienta KJ, Sewalt RG, Otte AP, et al: The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature 2002, 419(6907):624-629.
8. Neer CG, Cao Q, Varambally S, Shen R, Ota I, Tomlins SA, Ghish D, Sewalt RG, Otte AP, Hayes DF, et al: EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. Proc Natl Acad Sci USA 2003, 100(20):11606-11611.
9. Khan SN, Jankowska AM, Mahfouz R, Dunbar AJ, Sugimoto Y, Hosono N, Hu Z, Cheriyath V, Vatolin S, Przychodzen B, et al: Multiple mechanisms deregulate EZH2 and histone H3 lysine 27 epigenetic changes in myeloid malignancies. Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK 2013, 27(6):1301-1309.
10. Ohm JE, McGarvey KM, Yu X, Cheng L, Schueler KE, Cope L, Mohammad HP, Chen W, Daniel WC, Yu W, et al: A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet 2007, 39(2):237-242.
11. Xu F, Li X, Wu L, Zhang Q, Yang R, Yang Y, Zhang Z, He Q, Chang C. Overexpression of the EZH2, RING1 and BMI1 genes is common in myelodysplastic syndromes: relation to adverse epigenetic alteration and poor prognostic scoring. Annals of Hematology 2011, 90(8):643-653.

12. Zardo G, Cimino G, Nervi C. Epigenetic plasticity of chromatin in embryonic and hematopoietic stem/progenitor cells: therapeutic potential of cell reprogramming. Leukemia - official journal of the Leukemia Society of America, Leukemia Research Fund, UK 2008, 22(8):1503-1518.

13. Arnold CD, Gerlich D, Stelzer C, Boryn UM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 2013, 339(6133):1074-1077.

14. de Laat W, Duboule D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature 2013, 502(7472):499-506.

15. Yang X, Huang Y, Chen JL, Xie J, Sun X, Lussier YA. Mechanism-anchored profiling derived from epigenetic networks predicts outcome in acute lymphoblastic leukemia. BMC Bioinformatics 2009, 10(Suppl 9):56.

16. Nuyten M, Beke L, Van Eynde A, Creulens M, Beulemans M, Van Hummelen P, Fiks F, Bollen M. The transcriptional repressor NPFIP1 is an essential element in EZH2-mediated gene silencing. Oncogene 2008, 27(10):1449-1460.

17. Herrera-Merchan A, Arzalliz L, Ligos JM, de Molina A, Dominguez O, Gonzalez S. EXPRESSION OF THE HISTONE METHYLTRANSFERASE EZH2 IN HEMATOPOIETIC STEM CELLS CAUSES MYELOPROLIFERATIVE DISEASE. Nature communications 2012, 3:623.

18. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004, 306(5696):636-640.

19. Gao SB, Xu B, Ding LH, Zheng Q, Zhang W, Cheng QJ, Li ZH, Su Q, Ji XW, Li JY, Yang YA, Wei Y, Liu SY, Chen ZJ, Yang Y. Overexpression of DNA polymerase iota regulated by JNK/c-Jun contributes to hypermutagenesis in bladder cancer. Plos One 2013, 8(7):e69317.

20. Yuan F, Xu Z, Yang M, Wei Q, Zhang Y, Yu J, Zhi Y, Liu Y, Chen Z, Yang J. Ectopic expression of the histone methyltransferase Ezh2 in hematopoietic stem cells causes myeloproliferative disease. Nature communications 2012, 3:623.

21. Vashisht S, Bagler G. Mechanism-anchored profiling derived from epigenetic networks predicts outcome in acute lymphoblastic leukemia. BMC Bioinformatics 2009, 10(Suppl 9):56.

22. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci USA 2014, 111(17):6131-6138.

23. Lawrence HJ, Christensen J, Fong S, Hu YL, Weissman I, Sauvageau G, Humphries RK, Largman C. Loss of expression of the Hoxa-9 homeobox gene impairs the proliferation and repopulating ability of hematopoietic stem cells. Blood 2005, 106(12):3988-3994.

24. Yang X, Sun X. Meta-analysis of several gene lists for distinct types of cancer: a simple way to reveal common prognostic markers. BMC Bioinformatics 2007, 8:118.

25. Lottaz C, Yang X, Schind S, Spang R. OrderListR-a bioconductor package for detecting similarity in ordered gene lists. Bioinformatics, 2006, 22(18):2315-2316.

26. Yoshimura K, Aoki H, Ikeda Y, Fuji K, Aoki Y, Furuhati A, Hoshi Y, Tanaka N, Ricci R, Ishihara T, et al. Regression of abdominal aortic aneurysm by inhibition of c-Jun-N-terminal kinase. Nat Med 2005, 11(12):1530-1538.

27. Greshock J, Nathanson K, Martin AM, Zhang L, Coukos G, Weber BL, et al. Global rank-invariant set profiling derived from Epigenetic Networks Predicts Outcome in Acute myeloid leukemia. Proc Natl Acad Sci USA 2002, 99(23):14925-14930.

28. Lui L, Li M, Sun C, Francisco L, Choiakraborty S, Sabado M, McDonald T, Gyoffy J, Chang K, Wang S, et al. Altered hematopoietic cell gene expression precedes development of therapy-related myelodysplasia/acute myeloid leukemia and identifies patients at risk. Cancer Cell 2011, 20(5):591-605.

29. Gentles AJ, Plewritsk SK, Majetti R, Alahdah AA. Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. JAMA 2010, 304(24):2706-2715.

30. Epper T, Takenaka K, Lechman ER, Waldron L, Nilsson B, van Galen P, Metzler KH, Poeppl A, Ling V, Beyene J, et al. Stem cell gene expression programs influence clinical outcome in human leukemia. Nat Med 2011, 17(9):1086-1093.

31. Majetti R, Becker MW, Tian Q, Lee TL, Yan X, Liu R, Chang JH, Hood L, Clarke MF, Weissman IL. Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. Proc Natl Acad Sci USA 2009, 106(9):3396-3401.

32. Liberzon A. A Description of the Molecular Signatures Database (MSigDB) Web Site. Methods Mol Biol 2014, 1150:153-160.

33. Pelz CR, Kulesz-Martin M, Bagby G, Sears RC. Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data. BMC Bioinformatics 2008, 9:320.

34. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, Horvath S. Strategies for aggregating gene expression data: the collapseFowers R function. BMC Bioinformatics 2011, 12:322.

35. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, German D, Baggeley K, Izarai Y. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 2010, 11(10):733-739.

36. Heydebreck Av, Huber W, Gentleman R. Differential Expression with the Bioconductor Project. In Encyclopedia of Genomics, Genomics, Proteomics and Bioinformatics. John Wiley & Sons/Jorde L, Little P, Dunn M, Subramaniam S 2009.

37. Yang X, Huang Y, Chen JL, Xie J, Sun X, Lussier YA. Mechanism-anchored Profiling derived from Epigenetic Networks Predicts Outcome in Acute Lymphoblastic Leukemia. BMC Bioinformatics 2009.

38. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 2003, 100(16):9440-9445.

39. Storey JD, Tibshirani R. Statistical methods for identifying differentially expressed genes in DNA microarrays. Methods Mol Biol 2003, 224:149-157.

40. Meyers CA, Liu XS. Identifying competing bias in next-generation sequencing methods for chromatin biology. Nat Rev Genet 2014, 15(11):709-721.

41. Aerts S, Haesueller M, van Voooren S, Griffith OL, Hulipua P, Jones S, Montgomery SB, Bergman CM, Open Regulatory Annotation C. Text-mining assisted regulatory annotation. Genome Biol 2008, 9(2):R31.
54. Mishra R, Thorat D, Soundararajan G, Pradhan SJ, Chakraborty G, Lohite K, Karnik S, Kundu GC: Semaphorin 3A upregulates FOXO 3a-dependent MelCAM expression leading to attenuation of breast tumor growth and angiogenesis. Oncogene 2014.

55. Nehil M, Paquette J, Tokuyasu T, McCormick F: High mobility group box 1 promotes tumor cell migration through epigenetic silencing of semaphorin 3A. Oncogene 2013.

56. Baratti MO, Moreira YB, Traina F, Costa FF, Verovski-Almeida S, Olalla-Saad ST: Identification of protein-coding and non-coding RNA expression profiles in CD34+ and in stromal cells in refractory anemia with ringed sideroblasts. BMC Med Genomics 2010, 3:30.

57. Maione F, Capano S, Regano D, Zentilin L, Giacca M, Casanovas O, Bussolino F, Serini G, Giraudo E: Semaphorin 3A overcomes cancer hypoxia and metastatic dissemination induced by antiangiogenic treatment in mice. J Clin Invest 2012, 122(5):1832-1848.

58. Mani SA, Guo W, Liao JS, Eaton EN, Ayyanan A, Zhou AY, Brooks M, Reinhard F, Zhang CC, Shipton M, et al: The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell 2008, 133(4):704-715.

doi:10.1186/1755-8794-8-S2-S6

Cite this article as: Yang et al.: Identification of epigenetic modifications that contribute to pathogenesis in therapy-related AML: Effective integration of genome-wide histone modification with transcriptional profiles. BMC Medical Genomics 2015 8(Suppl 2):S6.