International Journal of Endorsing
HEALTH SCIENCE RESEARCH

IJEHSR

Original Article

Systematic identification of regulatory elements in Leukemia.

Komal Tara[1], Nisar Ahmed Shar[2] & Amna Amin Sethi[1]

[1]Department of Biomedical Engineering, NED University of Engineering and Technology, Karachi-Pakistan.

[2]Department of Computer & Information System Engineering, NED University of Engineering and Technology, Karachi-Pakistan.

Doi: 10.29052/IJEHSR.v12.i1.2024.30-38

Corresponding Author Email: nisarshar@neduet.edu.pk

Received 02/12/2022
Accepted 23/05/2023
First Published 15/12/2023

OPEN ACCESS

Abstract

Background: Leukemia is a type of cancer that originates from the bone marrow's blood-forming stem cells. Cancer cells don't follow the usual cellular differentiation and function pathway, so they supersede the healthy cells. Depending on the population and maturity level of abnormal cells, leukemia can be classified as Acute or Chronic.

Methodology: We mapped defined leukemia mutations from the COSMIC- The Catalogue of Somatic Mutations to perspective regulatory elements. GeneCards - Human Genes Database and Factorbook were utilized in this work to extract data to analyze gene-centric data related to AML and CML. The chromosomal location, Ensembl version, GRCh37 coordinates, and detailed examination of exons, introns, promoter binding regions, and enhancers were used to investigate co-expressed and differentially expressed genes. Downloaded from Factorbook, the expression levels of healthy and sick genes were compared to known transcription factor binding patterns.

Results: When translational control genes become mutated, they begin performing their function excessively, leading to uncontrolled cell proliferation and an accumulation of immature cells in the blood that are unable to perform any function on the one hand and interfering with healthy cells' ability to function optimally due to overpopulation and growth on the other. There is a group of genes whose expression level declines as they are affected by the gene, suggesting that these genes should function as insulators or silencers under normal circumstances. The data for Myc and Max genes were extracted from the Human Genome database and sorted using different techniques to find the common regulatory regions (CRRs). These CRRs were then divided into distinct categories based on the degree to which they co-expressed or their level of expression.

Conclusion: Regulatory elements have been identified depending on the values of their expression level and how they are changing concerning the control group. This work will help in understanding the guidelines of blood malignancy at the cellular level by recognizing administrative destinations and are, in this manner, possible focuses for the treatment plans and precession accuracy medication.

Keywords

Cancer Mutations, Gene Expression, Gene Regulation, Leukaemia, Regulatory Regions.

Check for updates

Introduction

Genetic changes are often the cause of many malignancies, such as lymphoma, leukemia, and solid tumors[1]. Proto-oncogenes, tumor suppressor genes, DNA repair genes, and other genes can all be affected by DNA mutations, which can result in cancer. Every gene's expression level can change, disrupting normal cellular functions and leading to cancer development. The ENCODE-based project's objective is to recognize and thereby correlate the variation of these regulatory genes to the development of disease[2].

Leukemia is a blood cancer that occurs when myeloid or lymphoid blood lineages multiply and develop improperly. Leukemia is the most frequent childhood cancer and one of the most prevalent diseases in adults. AML kills 80% of elderly people and 60% who are under the age of 60[3]. Transcriptomes contain all RNA sequences derived from genetic code. They capture the complexities of gene expression beyond simple nucleic acid replication. At the molecular oncogenesis level, any change in the ruling of gene expression interrupts the assembly of essential proteins, affecting normal cellular activities and evolving toward cancer[4]. A segment of a nucleic acid molecule called the regulatory sequence controls the expression levels of a particular gene.

Many kinds of cancer and the genes that regulate them are the subject of our study. The number of cancer types exceeds 100. Researchers have learned that certain mutations are common in various cancers as they have obtained a deeper knowledge of the molecular abnormalities that lead to cancer[5]. Our primary goal is to identify these genes, categorize them based on their expression levels, and then compare those levels to the typical values obtained from the GTEx site to draw conclusions about the relevant set of genes and the degree of their expression in relation to disease[6].

Methodology

Datasets & Recognition of Regulatory Elements

Repositories of genetic information and many sequential genome components are used to extract data for the current study[7]. Gene-centric data has been extracted from the GeneCards-Human Genes Database to determine which ubiquitous genes are responsible for AML and CML.

KIT, FLT3, NPM1, CEBPA, RAS, WT1, BAALC, CBL, ERG, MN1, DNMT, TET2, IDH, ASXL1, and PTPN11 were among the responsible genes and transcription factors identified after searching and filtering for AML. Ubiquitous genes identified via bedtool for CML are BRC-ABL, ABL1, RUNX1, NRAS, SETPB1, MIR17, KRAS, MIR20A, MIR10A, SF3B1, IFNA1, CSF3R, and PDGFRB. Recurrent mutational genes identified after sorting via bedtool as FLT3, C-KIT, and RAS. After retrieving individual responsible genes, their ensemble IDs were noted in Ensembl GRCh38.p13 format. Ensembl is a bioinformatics project that organizes biological knowledge through massive genomic sequences[8]. Co-expressed genes were studied by observing their chromosomal location, Ensembl version, and GRCh37 coordinates and observed in detail for exons, introns, the promotor binding region, and enhancers[9].

Data Collection

The data for the expression levels of normal and diseased genes has been downloaded from Factorbook. It is a transcription factor (TF)--a centric web-based library of integrative analysis connected with ENCODE ChIP-seq data[10]. We have considered the NB4 cell line for research and transcription factors because it belongs to the tier 3 cell line, which is less explored yet novel. GM12878 and K562 are control groups, and NB4 is under study in this research. Two transcription factor data available for NB4 are MYC and MAX.

It comprises motifs, chromatin characteristics, histone modification patterns, DNase I cleavage imprints, and nucleosome orientation profiles. Searched genes are then compared against known

transcription factor binding sites (TFBS). Overlapped genes are then discarded and not considered for further study[11]. Considering the novel ones for further study.

In later data mining, the expression level data for different cancer types were mined from COSMIC-Cancer Gene Census. We have considered AML, CML, B-Cells Non-Hodgkin Lymphoma Blood, Prostate Adeno Carcinoma, Bladder Transitional Cells Carcinoma, and Breast Carcinoma for study and to compare how certain groups of genes expressed differently in different types of cancer[12].

Transcription Factors for NB4

MYC genes and proteins are extremely important in the treatment of cancers. Except for early response genes, Myc upregulates gene expression across the board[13]. In the context of Myc hyperactivation, inactivation of the SUMO-activating enzyme (SAE1 / SAE2) causes cell death[11]. BET inhibitors have been effectively utilized to limit Myc activity in pre-clinical cancer models, and they are now being tested in clinical studies. Myc is an oncoprotein that has a role in cell growth, differentiation, and death. The dimers fight for a shared DNA E box, and the dimer forms a transformation that creates a complicated transcriptional regulatory mechanism. Any changes in this gene cause pheochromocytoma in the past[14]. This gene's pseudogene may be found on chromosome 7's extended arm. Multiple transcript variations result from alternative splicing. Here, we are intending to convert coordinates into a peak center. Therefore, the size of the resulting coordinate should be one base pair for all binding sites[15].

Cis- Regulatory Regions CRRs

We have determined CRRs by downloading data for Myc and Max, sorting them through bed tools individually, and then intersecting them to find the overlapping locations. Then, find the protein size by converting coordinates into a peak center[16]. Later, the peak center file of Max intersected with the sorted Myc file to get CRR114. CRR2 was obtained by intersecting the peak center file of Myc intersected with sorted Max[14]. CRRs are then calculated by intersecting CRR1 & CRR2 with CRR Data, which is downloaded. Schematics for the determination of CRRs are depicted in Figure 1.

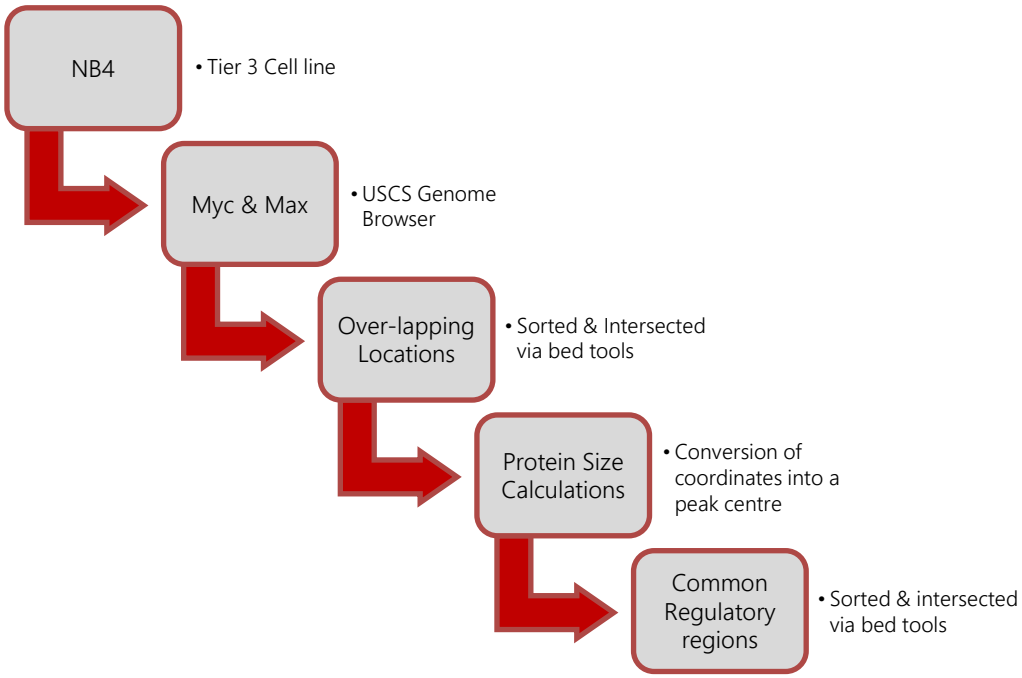

Figure 1: Determination of CRRs.

Single Nucleotide Polymorphisms (SNPs)

We first downloaded the data from the Genotype-Tissue Expression (GTEx) project to discover potential leukemia SNPs. Then, to measure the degree to which the data is distributed in proportion to the mean, the standard deviation is calculated[16]. The main data set has around 700 genes for different cancer types. Then, genes were categorized according to their expression level, which are the genes that co-expressed into different categories.

A t-test was used for statistical analysis due to the small number of explored genes[17]. Candidate gene expression level values related to different types of cancer are being compared to find out how distinguishable they are genes normal expression values are normally distributed as shown in Figure 2 and 3 respectively.

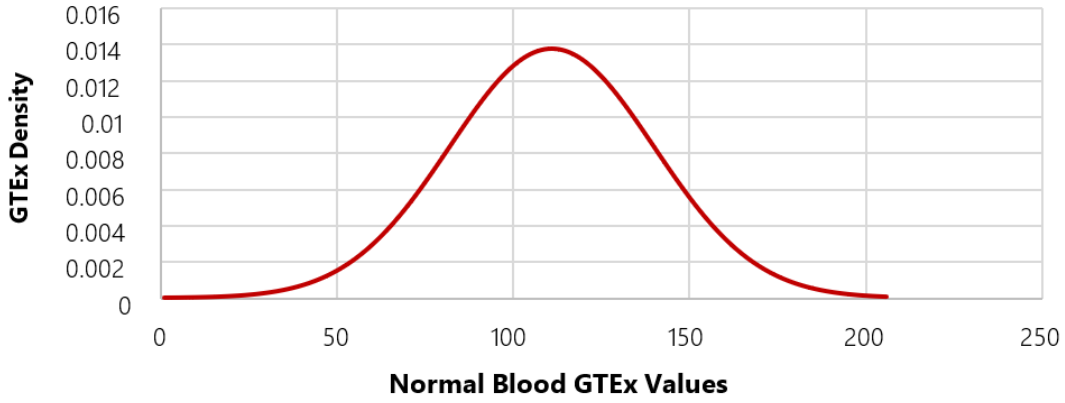

Figure 2: Assumption for T-test that the control data is regularly distributed.

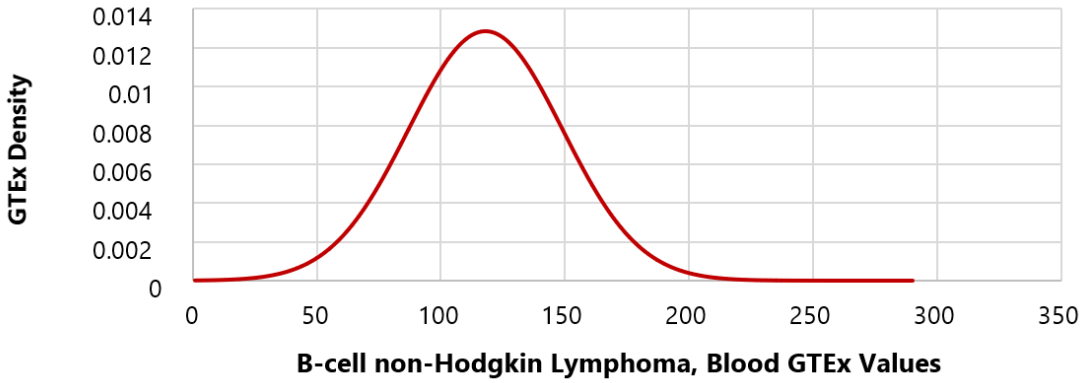

Figure 3: Assumption for T-test that the sample data is regularly distributed.

Results

Once the calculated results are verified, as shown in Table 1, regulatory elements can be chosen depending on the values of their expression level and how they are changing concerning the control group[18]. There is a group of regulatory elements whose expression level remains the same for cancer variants, so it can be suggested that they have some other role besides translation or expression. They are entitled as co-expressed genes by the fact that they are co-expressing with each other without altering or hindering one another's function[19].

Table 1: Statistical Analysis for P, T & SMD Values.

Parameters	B-Cell Non-Hodgkin Lymphoma, Blood	Chronic Lymphocytic Leukemia, Blood	Lymphoma, Blood	Normal-Blood (GTEx), Blood
Average	17.33	20.82	22.93	10.14
Std	31.03	25.55	36.77	28.85
Variance	963.02	652.74	1,351.77	832.23
Count	50.00	47.00	49.00	50.00
V/ N	19.26	13.89	27.59	16.64
Vn/Nn + V/N	35.90	30.53	44.23	
Sqrt	5.99	5.53	6.65	
Mean N - Mean	7.18	10.68	12.78	
Degree of freedom	98	95	97	
Critical value	1.98	1.98	1.98	1.98
M1-M2	7.18	10.68	12.78	
Pooled Std	29.96038871	27.30055859	33.004858	
Cohen's ds	0.239716516	0.391178965	0.387354232	
T-score	1.64	2.96	2.43	
Significance level	0.05	0.1, 0.05	0.05	
P- value	0.195	0.005*	0.020*	

*Significant at $p<0.05$

Under-expressed genes are the group of genes whose expression level decreases as they become mutated, so it can be interpreted that these genes should be working as silencers or insulators in normal conditions or they got mutated in the manner that the optimal level of their expression is required for normal functioning is not producing so the cancer is arising[20]. Then they come to the genes whose expression level elevated speedily after the onset of cancer; those must be the group of translational control genes, so after getting mutated, they start performing their function excessively, which results in uncontrolled cell proliferation and accumulation of immature cells in the bloodstream which are unable to perform any function on the one hand and also disturbing healthy cell to perform optimally due to overpopulation and unwanted growth[21].

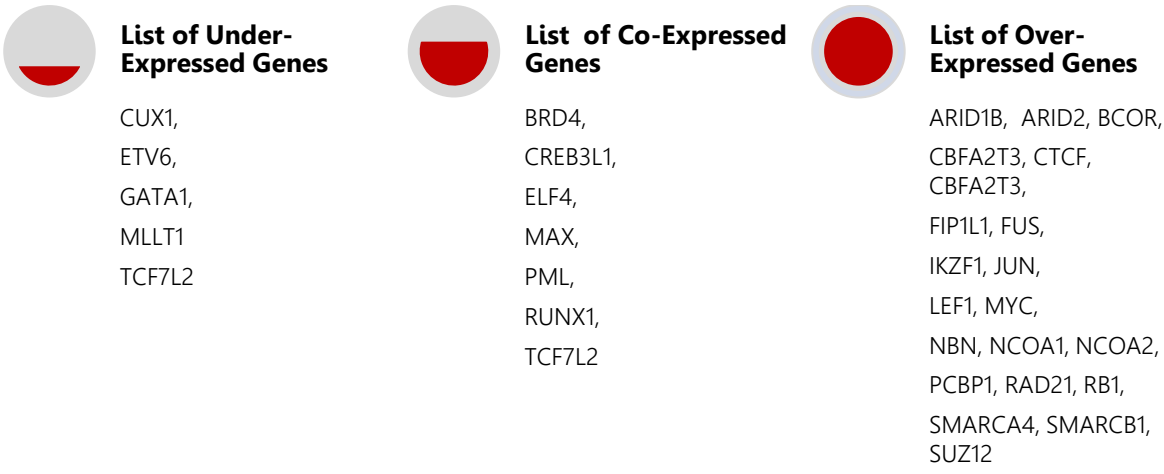

Figure 4: List of genes categorized as per their expression level.

DNA sequencing enables us to identify many elements important for genetic regulation at the genome scale and examine the genomes of different types of cancer[22]. The technique given here is relevant to two major challenges the data addresses. The first stage is to go from a qualitative understanding of potential regulatory domains to functional comprehension of specific genes[23].

Discussion

The discovery of expression quantitative trait loci (QTL) has revolutionized human genetics by having a broad, conveniently accessible, and explainable molecular relationship between genetic variants and organismal phenotypes[24]. Its application in illness investigations to provide a biological perspective has prompted subsequent research to broaden the variety of molecular phenotypes to be examined. Mutations impacting regulatory areas may be equally essential in tumorigenesis as those influencing protein-coding regions or active RNA molecules[25].

CUX1 may influence morphogenesis, gene expression, differentiation, and cell cycle progression. ETV6 is necessary for hematopoiesis and the development of the vascular network. This gene has been linked to a high range of chromosomal alterations linked to leukemia. GATA1 The protein plays an important role in erythroid development by regulating the switch of fetal hemoglobin to adult hemoglobin[26]. Mutations in this gene have been associated with X-linked dyserythropoietic anemia and thrombocytopenia. MLLT1 It is expected to play a role in transcriptional control. TCF7L2 The protein is thought to be involved in blood glucose control. This gene's genetic variations are associated with a higher probability of type 2 diabetes[27].

The family of genes identified in this study are co-expressed. BRD4 gene was determined and associated with respiratory system cancer in children. CREB3L1 inhibits the proliferation of virus-infected cells, which may contribute to the limitation of virus spread[28]. ELF4 The encoded protein is necessary for developing and operating natural killer cells and innate immunity. The transcription factor and tumor suppressor PML phosphoprotein are located in nuclear bodies. The RUNX1 transcription factor interacts with the core region of several enhancers and promoters. The expression of this gene's protein is thought to be necessary for the development of healthy hematopoiesis. Several types of leukemia have been associated with chromosomal translocations affecting this gene[29].

A group of over-expressed genes with a translational level that increases with the onset of cancer are considered translational control genes and have a greater impact on how the genes are expressed. ARID1B, this gene plays a role in cell cycle activation. ARID2 functions in transcriptional regulation, cell lineage gene regulation, cell cycle control, chromatin structure modification, and embryonic patterning[30]. BCOR, the protein produced by this gene, has been discovered as a transcription repressor, which is necessary for germinal center development and may impact apoptosis. CBFA2T3 This gene encodes a member of the family of myeloid translocation genes that interacts with DNA-bound transcription factors and recruits a wide range of corepressors to aid in transcriptional repression. This gene could reduce breast tumors. FIP1L1 These gene fusions and chromosomal losses are responsible for some forms of hyper-eosinophilic syndrome. Chromatin remodeling and IKZF1 are related. This protein controls the development of lymphocytes and is only expressed in the hemo-lymphopoietic system. JUN It has an incredibly similar protein to the viral protein that directly binds to specific target DNA sequences to regulate gene expression[29]. NCOA1 and NCOA2, the protein this gene produces, function as a transcriptional co-factor for hormone and nuclear hormone sites. SMARCA4 and SMARCB1 for the transcriptional activation of genes ordinarily inhibited by chromatin.
Suz12 This zinc finger gene was discovered at the breakpoints of a recurrent chromosomal translocation in endometrial stromal sarcoma patients. TBL1XR1 The protein encoded by this gene, which is thought to be a nuclear receptor corepressor, is required for the transcriptional activation of many transcription factors. Mutations

in these genes have been connected to several autism-related illnesses. TCF 12 The expression of genes particular to a given lineage may be regulated by this encoded protein, which is generated in several organs. ZMYM3 This gene resides on chromosome X. It has been preserved throughout vertebrate evolution and is most common in the brain[31]. The gene produces the zinc finger protein ZNF384, which is probably a transcription factor. Candidate genes are grouped as per their expression level, illustrated in Figure 4.

This study demonstrates that modeling using data compendia such as ENCODE can discover genomic areas possibly more clearly related to gene expression and suggest relationships to the responsive genes[32]. EQTL analysis is one of the simplest ways to identify possible point mutations at susceptibility loci, and it offers evidence of unique higher efficacy for risk SNPs. For eQTL evaluation, we used TCGA data[31].

The experts must critically analyze the information from the web-based compendium. Still, manual curation through a molecular pathology lab is the gold standard for validating any bio-informatics-based project. Specialists perform extensive literature searches to collect, rearrange, analyze, standardize, and categorize mutation data, phenotypic information, and clinical facts[33]. Digital systems cannot reconcile terminology discrepancies or determine if the findings are statistically accurate and meaningful to the predictive results based on the data accessible through online databases. Results will be considered more accurate if verified through molecular pathology labs[34]. As this is a systematic study and identification, it can be considered a baseline or guide map for future implementation or for designing and scheduling protocols for laboratory studies and experimentation. This is especially useful when molecular pathology laboratories encounter unusual variations or variants with uncertain significance[35,36]. In these circumstances, users can use their discretion to accept or disagree with online data for that specific variation.

Conclusion

Several genes are identified and grouped as per their expression level. This type of categorization can be used to design precise quality drugs for tumors arising from a mutation in the same group of responsible genes. Precision medicine in cancer uses relevant information about a patient's tumor to aid in diagnosis, therapy planning, and assessing how effectively the plan is working and creating a prognosis. For molecular pathology laboratories assessing somatic next-generation sequencing (NGS) testing for precision oncology applications, having easy and rapid access to data and evidence is critical. The examination can also be expanded to help in the early diagnosis and categorization of AML and CML, enhancing the quality of life and life expectancy.

Conflicts of Interest

We have no conflicts of interest to disclose.

Acknowledgment

We are obliged to the respondents of our study.

Funding

There was no funding, and the Department of Biomedical Engineering, NED University of Engineering and Technology, Karachi, supported the use of the existing facilities.

References

1. Ailles LE, Gerhard B, Hogge DE. Detection and characterization of primitive malignant and normal progenitors in patients with acute myelogenous leukemia using long-term coculture with supportive feeder layers and cytokines. Blood. Am J Hematol.1997;90:2555–2564.
2. Mattick JS. Non-coding RNAs: the architects of eukaryotic complexity. EMBO rep. 2001;2:986–991. doi: 10.1093/embo-reports/kve230.
3. Guo L, Du Y, Chang S, Zhang K, Wang J. rSNPBase: a database for curated regulatory SNPs. Nucleic Acids Res. 2014;42:D1033–D1039. doi: 10.1093/nar/gkt1167.
4. Simon JM, Gomez NC. Epigenetic analysis in Ewing sarcoma. In Ewing Sarcoma.: Springer;2021. p.285–302.

5. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M. The UCSC genome browser database: update 2011. Nucleic acids res. 2010;39(suppl_1):D876–D882. doi: 10.1093/nar/gkq963.
6. Stern C. Boveri and the early days of genetics. Nature. 1950;166:446–446. doi: 10.1038/166446a0.
7. Consortium ENCODEP. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS biol. 2011;9:e1001046. doi: 10.1371/journal.pbio.1001046.
8. Consortium ENCODEP, others. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. nature. 2007;447:799-816. doi: 10.1038/nature05874.
9. van Helden J, André B, Collado-Vides J. A web site for the computational analysis of yeast regulatory sequences. Yeast. 2000;16:177–187. doi: 10.1002/(SICI)1097-0061(20000130)16:2<177::AID-YEA516>3.0.CO;2-9.
10. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome res. 2002;12:996–1006. doi: 10.1101/gr.229102.
11. Cleary JD, Ranum LPW. Repeat associated non-ATG (RAN) translation: new starts in microsatellite expansion disorders. Curr Opin Genet Dev. 2014;26:6–15. doi: 10.1016/j.gde.2014.03.002.
12. Artz A, Ridgeway JA. Managing the Continuum of Myeloid Malignancies. JAdPrO. 2018;9:345.
13. Aplan PD. Causes of oncogenic chromosomal translocation. TiG. 2006;22:46–55. doi: 10.1016/j.tig.2005.10.002.
14. Xu C, Fu H, Gao L, Wang L, Wang W, Li J, Li Y, Dou L, Gao X, Luo X, Jing Y. BCR-ABL/GATA1/miR-138 mini circuitry contributes to the leukemogenesis of chronic myeloid leukemia. Oncogene. 2014;33:44–54. doi: 10.1038/onc.2012.557.
15. Keilwagen J, Posch S, Grau J. Accurate prediction of cell type-specific transcription factor binding. Genome biol. 2019;20:1–17. doi: 10.1186/s13059-018-1614-y.
16. Zelent A, Greaves M, Enver T. Role of the TEL-AML1 fusion gene in the molecular pathogenesis of childhood acute lymphoblastic leukaemia. Oncogene. 2004;23:4275–4283. doi: 10.1038/sj.onc.1207672.
17. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of reporting P values in the biomedical literature, 1990-2015. Jama. 2016;315:1141–1148. doi: 10.1001/jama.2016.1952.
18. Diaz de Arce AJ, Noderer WL, Wang CL. Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. Nucleic acids res. 2018;46:985–994. doi: 10.1093/nar/gkx1114.
19. Wu P. Inhibition of RNA-binding proteins with small molecules. Nature Rev Chem. 2020;4:441–458. doi: 10.1038/s41570-020-0201-4.
20. Maher B. ENCYCLOPAEDIA THE. Nature. 2012;489(7414):46-8. doi: 10.1038/489046a.
21. Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, Blue SM, Chen JY, Cody NA, Dominguez D, Olson S. A large-scale binding and functional map of human RNA-binding proteins. Nature. 2020;583:711–719. doi: 10.1038/s41586-020-2077-3.
22. Karolchik D, Hinrichs AS, Kent WJ. The UCSC genome browser. Curr Protoc Bioinformatics. 2012;40:1–4. doi: 10.1002/0471250953.bi0104s40.
23. Raghavan SC, Lieber MR. DNA structure and human diseases. Front Biosci. 2007;12:4402–4408. doi: 10.2741/2397.
24. Fan J, Li R, Zhang CH, Zou H. Statistical foundations of data science: Chapman and Hall/CRC;2020.
25. Ishida M, Miyamoto M, Naitoh S, Tatsuda D, Hasegawa T, Nemoto T, Yokozeki H, Nishioka K, Matsukage A, Ohki M, Ohta T. The SYT-SSX fusion protein down-regulates the cell proliferation regulator COM1 in t (x; 18) synovial sarcoma. Mol Cell Biol. 2007;27:1348–1355. doi: 10.1128/MCB.00658-06.
26. Onizuka T, Moriyama M, Yamochi T, Kuroda T, Kazama A, Kanazawa N, Sato K, Kato T, Ota H, Mori S. BCL-6 gene product, a 92-to 98-kD nuclear phosphoprotein, is highly expressed in germinal center B cells and their neoplastic counterparts. 1995;86(1):28-37.
27. Hiom K, Melek M, Gellert M. DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. Cell. 1998;94:463–470. doi: 10.1016/s0092-8674(00)81587-1.
28. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012; 489: 91–100.
29. Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature. 2016; 529: 110–114. doi: 10.1038/nature16490.

30. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. Stat Appl Genet Mol Biol. 2011;10.
31. Krzywinski M, Altman N. Comparing samples–part I: robustly comparing pairs of independent or related samples requires different approaches to the t-test. Nat methods. 2014;11:215–217. doi: 10.1038/nmeth.2858.
32. Korkmaz G, Lopes R, Ugalde AP, Nevedomskaya E, Han R, Myacheva K, Zwart W, Elkon R, Agami R. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat Biotechnol. 2016;34:192–198. doi: 10.1038/nbt.3450.
33. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M. Role of noncoding sequence variants in cancer. Nat Rev Genet. 2016;17:93–108. doi: 10.1038/nrg.2015.17.
34. Sano K. Structure of AF3p21, a new member of mixed lineage leukemia (MLL) fusion partner proteins-implication for MLL-induced leukemogenesis. Leukemia & Lymphoma. 2001;42:595–602. doi: 10.3109/10428190109099319.
35. Feingold EA, Pachter L. The ENCODE project Science.2004;306:636–640.
36. Management Group Liefer Laura A. 51 Wetterstrand Kris A. 51 Good Peter J. 51 Feingold Elise A. 51 Guyer Mark S. 51 Collins Francis S. 52, Baylor College of Medicine Human Genome Sequencing Center*, Washington University Genome Sequencing Center*, Broad Institute*, Children's Hospital Oakland Research Institute*, Gerstein M. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. nature. 2007;447(7146):799-816. doi: 10.1038/nature05874.