Pre-Clinical Research Report

Journal of INTERNATIONAL MEDICAL RESEARCH

Knowledge database assisted gene marker selection for chronic lymphocytic leukemia

Journal of International Medical Research
2018, Vol. 46(8) 3358–3364
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0300060518783072
journals.sagepub.com/home/imr
SAGE

Xixi Xiang[1], Yu-Ping Wang[2], Hongbao Cao[3,4] and Xi Zhang[1]

Abstract

Objective: To investigate whether previously curated chronic lymphocytic leukemia (CLL) risk genes could be leveraged in gene marker selection for the diagnosis and prediction of CLL.

Methods: A CLL genetic database (CLL_042017) was developed through a comprehensive CLL-gene relation data analysis, in which 753 CLL target genes were curated. Expression values for these genes were used for case-control classification of four CLL datasets, with a sparse representation-based variable selection (SRVS) approach employed for feature (gene) selection. Results were compared with outcomes obtained by using analysis of variance (ANOVA)-based gene selection approaches.

Results: For each of the four datasets, SRVS selected a subset of genes from the 753 CLL target genes, resulting in significantly higher classification accuracy, compared with randomly selected genes (100%, 100%, 93.94%, 89.39%). The SRVS method outperformed ANOVA in terms of classification accuracy.

Conclusion: Gene markers selected from the 753 CLL genes could enable significantly greater accuracy in the prediction of CLL. SRVS provides an effective method for gene marker selection.

Keywords

Chronic lymphocytic leukemia (CLL), sparse representation, variable selection, disease prediction, case-control classification, genetic databases, gene markers

Date received: 29 November 2017; accepted: 24 May 2018

[1]Center of Hematology, The Second Affiliated Hospital of Army Military Medical University, No 83 Xinqiao Street, Shapingba District, Chongqing, 40037, China
[2]Department of Biomedical Engineering, Tulane University, New Orleans, LA, USA
[3]Department of Genomics Research, R&D Solutions, Elsevier Inc., Rockville, MD, USA
[4]Unit on Statistical Genomics, NIMH/NIH, Bethesda, MD, USA

Corresponding author:
Xi Zhang, Center of Hematology, The Second Affiliated Hospital of Army Military Medical University, No 83 Xinqiao Street, Shapingba District, Chongqing, 40037, China.
Email: x.zhang@gousinfo.com

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Introduction

Chronic lymphocytic leukemia (CLL) is the most frequent B-cell leukemia, which affects men more frequently than women.[1] The disease often occurs in elderly patients, and rarely affects children.[2] Despite the efforts of many genetic studies, the molecular abnormalities and genetic mechanics of CLL remain largely unknown.[3] Most CLL patients are diagnosed without symptoms, with the exception of a high white blood cell count in a routine blood test. Consequently, early CLL could easily remain untreated.[4] Therefore, there is an urgent need for biomarker identification to facilitate early prediction of CLL.[5]

In the past, hundreds of genes/proteins have been linked to CLL. Mutations of some risk genes, including IL4 and TP53, have been frequently reported as important markers for the pathogenic development of CLL.[6,7] These genes may serve as biomarkers for multiple other diseases,[7,8] thus decreasing their specificities as biomarkers for the prediction of CLL. Additionally, many CLL-gene relationships have been reported, but few can be replicated (e.g., PRKCD and TGFBR2[9,10]), reflecting the heterogeneity of CLL and the variance of CLL-related genetic changes among patients.[11] Moreover, a number of novel CLL risk genes are identified each year,[12] facilitating the development of an enriched genetic database for CLL.

The purpose of this study was to investigate whether previously reported CLL genes could be leveraged as a database for gene marker selection, specifically targeting early diagnosis of CLL. We hypothesized that if these CLL genes are effective for the prediction of CLL, gene markers selected from among them should enable significant accuracy in differentiating CLL cases from controls.

Methods

Development and analysis of CLL_042017

Figure 1 presents the database schema of the curated database CLL_042017. The database contains 753 genes (**CLL_042017→Related Genes**) that were collected as CLL target genes; each of these genes has at least one reference to support its relationship with CLL (3,078 references in total; see **CLL_042017→Ref for Disease-Gene Relation**). The CLL-gene relations were identified by using Pathway Studio (www.pathwaystudio.com).[13] The database also includes 235 drugs (**CLL_042017→Related Drugs**), 97 diseases (**CLL_042017→Related Diseases**), and 88 pathways (**CLL_042017→Related Pathways**). The information of 2,756 supporting references for CLL-Drug relations is provided in **CLL_042017→Ref for Related Drugs**. The reference information includes titles and related

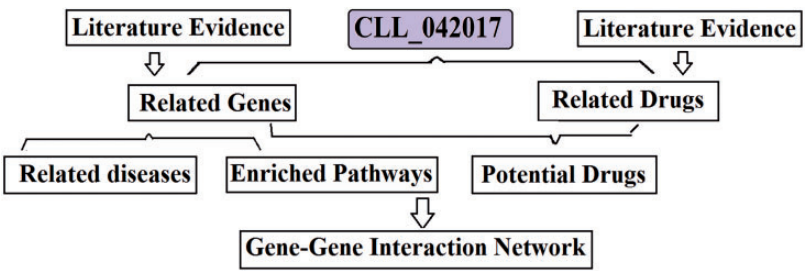

Figure 1. Chronic lymphocytic leukemia (CLL) genetic database schematic.

sentences where a relationship has been identified. The current CLL_042017 is online, available at http://gousinfo.com/database/Data_Genetic/CLL_042017.xlsx. For a more detailed description of the database, please refer to **CLL_042017→Database Note**.

SRVS for gene vector selection

A sparse representation-based variable selection (SRVS) algorithm (described in detail elsewhere)[14] was used to rank the 753 CLL target genes, on the basis of a given experimental dataset. For each gene, a sparse weight is assigned by SRVS. The gene vector, composed of the top n genes by SRVS, is the genetic marker for a CLL case/control group, where n is the number of genes corresponding to the maximum classification ratio (CR) as defined in Eq. (1).

$$classification\ ratio\ (CR) = \frac{\#correcly\ classified\ subjects}{\#total\ subjects} \quad (1)$$

Gene expression data

In this study, we used 4 RNA gene expression datasets to evaluate classification performance with CLL target genes; these datasets were GSE2466, GSE19147, GSE50006, and GSE8835. The datasets were selected by using the Illumina BaseSpace Correlation Engine (http://www.illumina.com) and are publicly available at the NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/). The data selection criteria were as follows: 1) Sample organism was Homo sapiens; 2) Data type was RNA expression; 3) Experiment design was CLL case *vs.* normal control. From each dataset, expression data of normal controls and CLL patients were extracted and used for case/control classification. Genes of each dataset were limited to CLL target genes curated within the database CLL_042017. The key statistics of the four datasets are summarized in Table 1.

The gene expression profiles of the four gene expression datasets are also included in CLL_042017: **CLL_042017→GSE2466, GSE19147, GSE50006, and GSE8835**. Within each dataset, the SRVS-generated weights (SRVSScore) and analysis of variance (ANOVA)-generated p-value score (PValueScore; logic transferred p-values: -10*log(p-value)) are also presented. The p-value for a gene is generated from the one-way ANOVA of the case/control comparison with the corresponding expression data. An SRVSScore and a PValueScore represent the significance of a gene in the dataset, according to SRVS and ANOVA methods, respectively.

CLL case/control classification

To identify the best gene vector and the corresponding classification accuracy (CR), the CLL target genes were first ranked by SRVSScore in descending order. Then, Euclidean distance-based

Table 1. Statistics of four gene expression datasets.

NCBI GEO ID	GSE2466	GSE19147	GSE50006	GSE8835
#CLL case/control	72/11	25/8	188/32	42/24
#genes from CLL_042017	564	624	685	624
Sample source	Peripheral blood lymphocytes	Peripheral blood CD3+T cells	leukemia cells	Peripheral blood CD4 T cells and CD8 T cells
Sample population	Austria	Germany	USA	USA

multivariate classification[18] was performed for each dataset, followed by leave-one-out (LOO) cross-validation. In each run of LOO, gene expression data of one subject were used for testing; the remaining data were used for training. The inputs of the classifier are the top n (n=1, 2 ...) genes, such that the CR of using the top n genes could be identified. A permutation of 5,000 runs was then conducted to test the hypothesis that a randomly selected gene set with a similar size can reach equal or higher CR. For each subset of genes, the permutation p-value was calculated as $n0/nT$, where $n0$ was the number of runs generating CR higher than that of the gene subset; nT was the total number of runs (5000 in this study). The gene vector that generated the highest CR was the best gene subset selected from the gene expression dataset, according to the SRVS method.

Following the same process, the best gene subset was identified for each dataset by the ANOVA approach. For comparison purposes, a CR baseline was also generated by using randomly selected gene sets of n (n=1, 2 ...) genes. For each point of the CR baseline, the value was the mean of 300 CRs by randomly selected genes from all genes within the dataset.

Results

CLL case/control classification

Figure 2 presents the classification results. Table 2 summarizes the results of LOO cross-validation of the two gene-ranking methods on four datasets, where the maximum CRs, corresponding numbers of top genes, and permutation p-values of the two methods are provided.

Figure 2 establishes that, compared with the CRs generated by randomly selected gene sets, the genes selected from CLL target genes by both SRVS and ANOVA can demonstrate significantly higher classification accuracies. Notably, by using only the top genes with highest SRVSScore/PValueScore, the highest CRs were acquired (See Figure 2 and Table 2); adding more genes with lower scores may not necessarily

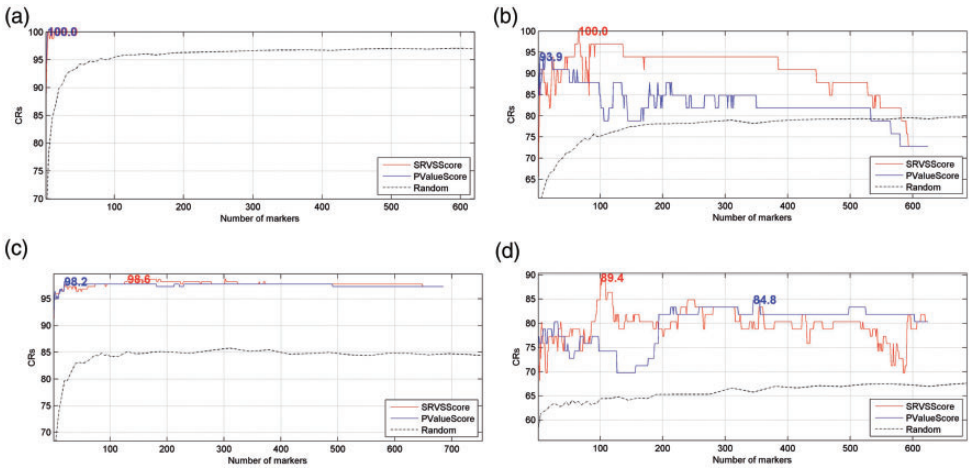

Figure 2. Comparison of different metrics through leave-one-out (LOO) cross-validation. Genes were ranked in ascending order according to SRVSScore or PValueScore, for sparse representation-based variable selection (SRVS) or analysis of variance (ANOVA), respectively. (a) GSE 2466, (b) GSE 19147, (c) GSE 50006 and (d) GSE 8835.

Table 2. LOO cross-validation and permutation results

	GSE2466 (case/control:72/11)		GSE19147 (case/control:25/8)		GSE50006 (case/control:188/32)		GSE8835 (case/control:42/24)	
	SRVS	ANOVA	SRVS	ANOVA	SRVS	ANOVA	SRVS	ANOVA
MaxCRs	100.00	100.00	100.00	93.94	98.64	98.18	89.39	84.85
# Selected Genes	4	3	65	3	131	20	101	345
p-value	0.001	0.0002	~0	0.0016	0.0014	0.0012	~0	~0
Unique genes from all datasets (%)	25% (1/4)	66.67% (2/3)	52.31% (34/65)	33.33% (1/3)	75.57% (99/131)	40% (8/20)	97.03% (98/101)	95.94% (331/345)
Overlap genes of two methods (%)	0% (0/4)	0% (0/3)	3.08% (2/65)	66.67% (2/3)	15.27% (20/131)	100% (20/20)	65.35% (66/101)	19.13% (66/345)

SRVS, sparse representation-based variable selection; ANOVA, analysis of variance.

improve classification accuracy. These results revealed the validity of both SRVS and ANOVA methods. Moreover, it was noted that SRVSScore outperformed PValueScore in terms of CR (Table 2).

Table 2 also shows that, for each dataset, the top genes selected by both methods could be significantly different (**CLL_042017→Venn Diagram**). For the SRVS method, the unique genes selected for the four datasets ranged from 25% to 97.03%; the range was 33.33% to 95.94% for the ANOVA method (Table 2→**Unique genes from all datasets (%)**). These results suggested that there were factors that could affect the gene marker selection, which is worthy of further study. It is also notable that, for a given dataset, gene markers selected by SRVS and ANOVA could differ (Table 2→**Overlap genes of two methods (%)**). This suggests that SRVS performs differently and more effectively than ANOVA.

Discussion

CLL affects approximately one million people globally, but remains poorly diagnosed at early stages. In the past, many studies have been performed with the aim of developing targeted molecular therapy for CLL[6,7]; hundreds of risk genes have been identified. Most of these genes are active within CLL-related genetic pathways, and many have been used as drug targets for the treatment of CLL. However, patients may demonstrate genetic variation, even in the same disease, implying the need for personalized treatment.[16] Therefore, for a given CLL patient/patient group, feature (gene) selection is important for diagnosis and treatment. Thus far, few studies have been conducted to test the validity of curated CLL risk genes for use as genetic markers in diagnosis and prediction of CLL.

In this study, we first conducted comprehensive literature data mining in 3078 scientific articles, which identified 753 CLL target genes. Gene set enrichment analysis showed that the majority of these genes (594/753) were significantly enriched within multiple genetic pathways that were associated with CLL (p-value<3e-13; q=0.001 for false discovery rate (FDR)). For instance, there are 230 genes significantly enriched within eight cell apoptosis pathways (p-value<5.2e-14; q=0.001 for FDR).[15] There were also 240 genes enriched within eight pathways/gene

sets related to cell growth and proliferation (p-value<6.8e-015)[16] and 218 genes enriched within immune response (p-value <8.7e-029).[17] More pathways and related information can be identified at **CLL_042017→Related Pathways.**

Sub-network enrichment analysis (SNEA; http://pathwaystudio.gousinfo.com/SNEA.pdf) showed that 717 of 753 genes significantly overlapped with risk genes linked to each of the 97 diseases (p-value<1.6e-100; q=0.001 for FDR; **CLL_042017→Related Diseases**). Many of these 97 diseases are cancers of different types, and many were related to CLL, including rheumatoid arthritis,[18] breast cancer[19] and multiple myeloma.[20]

Within CLL_042017, there were 235 known CLL drugs/small molecules (**CLL_042017→Related Drugs**) that have been evaluated within clinical trials and have demonstrated effectiveness in treating CLL. These 235 drugs demonstrated significant overlap (22 overlapped drugs; p-value=8.20e-23) with the top 100 potential drugs/small molecules (**CLL_042017→ Potential Drugs**), whose gene subnetworks were significantly enriched within the 753 CLL genes. Additionally, many of the 753 CLL genes were target genes of known CLL drugs. For instance, rituximab induces apoptosis of CLL cells by inhibiting the expression of BCL2.[21] These results supported a possible association between CLL and the 753 target genes.

CLL case/control classification was conducted on four independent gene expression datasets, with two algorithms for gene selection within the 753 CLL gene pool: SRVS method and ANOVA. The basic theory for feature (gene) selection is that not all 753 genes will exhibit mutations for a given CLL patient/patient group; therefore, it is not appropriate to use all as target genes in the diagnosis and treatment.

Compared with randomly selected genes, these selected by both SRVS and ANOVA led to significantly higher prediction power (permutation p-value<0.0014 for SRVS and permutation p-value<0.0016 for ANOVA; CRs of SRVS vs. ANOVA: 100% *vs.* 100%, 100% *vs.* 93.94%, 98.64% *vs.* 98.18% and 89.39% *vs.* 84.85%, for the four datasets, respectively), as shown in Table 2. These results indicated that genetic markers selected from the 753 CLL target genes possess significant power for the diagnosis and prediction of CLL. Moreover, SRVS outperforms ANOVA in terms of CR. This implies the effectiveness of the SRVS method for gene marker selection for CLL.

Gene markers selected by both SRVS and ANOVA methods demonstrated substantial uniqueness (>25%) across different datasets (Table 2). This indicates that, in addition to the genomic specificity of each patient group, there may be other factors that affect the gene marker selection, which merit further study. As shown in Table 1, the four datasets were acquired from different blood cells and different patient populations. This may contribute to variations in the gene marker selection results (Table 2).

In conclusion, our study suggested that gene markers selected from the 753 CLL genes could provide high accuracy in the prediction of CLL, and that SRVS is an effective method for gene marker selection in CLL diagnosis and prediction.

Declaration of conflicting interest

The authors declare that there is no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ORCID iD

Xi Zhang http://orcid.org/0000-0002-2533-8759

References

1. Kipps TJ, Stevenson FK and Wu CJ. Chronic lymphocytic leukaemia. *Nat Rev Dis Primers* 2017; 3: 16096.
2. Mauro FR, Foa R and Giannarelli D. Clinical characteristics and outcome of young chronic lymphocytic leukemia patients: a single institution study of 204 cases. *Blood* 1999; 94: 448–454.
3. Kashyap MK, Kumar D and Villa R. Targeting the spliceosome in chronic lymphocytic leukemia with the macrolides FD-895 and pladienolide-B. *Haematologica* 2015; 100: 945–954.
4. Lehmann S, Ogawa S, Raynaud SD, et al. Molecular allelokaryotyping of early-stage, untreated chronic lymphocytic leukemia. *Cancer* 2008; 112: 1296–1305.
5. Kambouris ME, Pavlidis C and Skoufas E. Culturomics: a new kid on the block of OMICS to enable personalized medicine. *OMICS* 2018; 22: 108–118. doi: 10.1089/omi.2017.0017.
6. Mainou-Fowler T, Proctor SJ, Miller S, et al. Expression and production of interleukin 4 in B-cell chronic lymphocytic leukaemia. *Leuk Lymphoma* 2001; 42: 689–698.
7. Rosenfeld MR, Malats N and Schramm L. Serum anti-p53 antibodies and prognosis of patients with small-cell lung cancer. *J Natl Cancer Inst* 1997; 89: 381–385.
8. Bahl R, Arora S and Nath N. Novel polymorphism in p21(waf1/cip1) cyclin dependent kinase inhibitor gene: association with human esophageal cancer. *Oncogene* 2000; 19: 323–328.
9. Ringshausen I, Schneller F and Bogner C. Constitutively activated phosphatidylinositol-3 kinase (PI-3K) is involved in the defect of apoptosis in B-CLL: association with protein kinase Cdelta. *Blood* 2002; 100: 3741–3748.
10. Spaner DE. Amplifying cancer vaccine responses by modifying pathogenic gene programs in tumor cells. *J Leukoc Biol* 2004; 76: 338–351.
11. Ustun C, Gotlib J and Popat U. Consensus opinion on allogeneic hematopoietic cell transplantation in advanced systemic mastocytosis. *Biol Blood Marrow Transplant* 2016; 22: 1348–1356.
12. Galletti G, Caligaris-Cappio F and Bertilaccio MT. B cells and macrophages pursue a common path toward the development and progression of chronic lymphocytic leukemia. *Leukemia* 2016; 30: 2293–2301.
13. Lorenzi PL, Claerhout S, Mills GB, et al. A curated census of autophagy-modulating proteins and small molecules: candidate targets for cancer therapy. *Autophagy* 2014; 10: 1316–1326.
14. Cao H, Duan J and Lin D. Sparse representation based biomarker selection for schizophrenia with integrated analysis of fMRI and SNPs. *Neuroimage* 2014; 102(Pt 1): 220–228.
15. Schimmer AD, Munk-Pedersen I, Minden MD, et al. Bcl-2 and apoptosis in chronic lymphocytic leukemia. *Curr Treat Options Oncol* 2003; 4: 211–218.
16. Klein U, Lia M and Crespo M. The DLEU2/miR-15a/16-1 cluster controls B cell proliferation and its deletion leads to chronic lymphocytic leukemia. *Cancer Cell* 2010; 17: 28–40.
17. Riches JC, Ramsay AG and Gribben JG. Immune reconstitution in chronic lymphocytic leukemia. *Curr Hematol Malig Rep* 2012; 7: 13–20.
18. Voulgari PV, Vartholomatos G and Kaiafas P. Rheumatoid arthritis and B-cell chronic lymphocytic leukemia. *Clin Exp Rheumatol* 2002; 20: 63–65.
19. Dialani V, Mani K and Johnson NB. Chronic lymphocytic leukemia involving the breast parenchyma, mimicker of invasive breast cancer: differentiation on breast MRI. *Case Rep Med* 2013; 2013: 603614.
20. Barlogie B and Gale RP. Multiple myeloma and chronic lymphocytic leukemia: parallels and contrasts. *Am J Med* 1992; 93: 443–450.
21. Pedersen IM, Buhl AM, Klausen P, et al. The chimeric anti-CD20 antibody rituximab induces apoptosis in B-cell chronic lymphocytic leukemia cells through a p38 mitogen activated protein-kinase-dependent mechanism. *Blood* 2002; 99: 1314–1319.