Analysis of mutational landscape of patients with chronic lymphocytic leukemia

A.V. Terskikh[1], A.A. Samsonova[2], A.A. Kanapin[2]
[1]-*Peter the Great St.Petersburg Polytechnic University, St. Petersburg, Russia*, anastasiya-terskih@mail.ru; [2]-*Department of Oncology, University of Oxford, Oxford, UK*, a.kanapin@gmail.com, a.a.samsonova@gmail.com

A precise understanding of the genomic features of chronic lymphocytic leukemia (CLL) may benefit the study of the disease's staging and treatment. Genomic landscape of CLL probably reflects either an unknown underlying biochemical mechanism playing a key role in CLL or multiple biochemical pathways independently driving the development of this tumor. The elucidation of either scenario may have important consequences on the clinical management of CLL. Our aim is to analyze mutational landscape of the disease to identify potential pathways driving the pathology.

Chronic lymphocytic leukemia is a clonal neoplasia of B-lymphocytes which accumulate mainly in the blood, bone marrow, lymph nodes and spleen [1]. Notably, these B-lymphocytes are differentiated, and can remain in an arrested state for several years after diagnosis. Two major molecular subtypes can be distinguished, characterized respectively by a high or low number of somatic hypermutations in the variable region of immunoglobulin genes [2]. This classification system was improved with the characterization of additional genomic and transcriptomic factors [3]. However, the genomic events that dictate the initiation and heterogeneous evolution of CLL remained partially unknown.

Next-Generation Sequencing (NGS) allows the comparison of the genome of tumor cells with the constitutive genome in normal tissues of the same patients. The variants present in the tumor genome and absent from the germinal genome are called somatic mutations, and constitute a requisite of cancer development. To detect somatic mutations we compared each tumor's VCF file with a corresponding normal VCF file and chose only those the variants which were presented only in a tumor. We detected 1682 nonsynonymous mutations and an average of 30 nonsynonymous mutations per patient. There a clear bias toward C > T / G > A substitutions as seen previously in other cancers [5].

To identify genes whose mutations were associated with leukemic tumorigenesis ("driver" mutations), we utilized penalized logistic regression [6]. We used treatment response as a binary output (1 – partial response, stable disease or progressive disease; 0 – complete response). The variables were frequency of protein domains (PFAM domains [7]) for each patient (observation). As there are far more variables (238) than observations (26), classical logistic regression does not work. The key idea in penalization methods is that overfitting is avoided by imposing a penalty on large fluctuations on the estimated parameters and thus on the fitted curve. We analyzed and compared two penalized methods: quadratic regularization or ridge regression [8] and LASSO method [9]. Three resampling methods: cross-validation (CV) [10], repeated CV [11], leave-one-out CV (LOOCV) [12] were subject to comparison as well. Thus, we identified 10 common protein domains with a corresponding genes list for both methods. The results are presented in Table 1.

Table 1. Common protein domains with corresponding genes list for ridge regression and LASSO method.

PFAM domain ID	Gene ID
PF00038	KRT14, KRT86, KRT36, KRT32
PF00047	KIR2DL1, DSCAM, KIR2DL4, VCAM1
PF00091	TUBA3D, TUBB8
PF00530	DMBT1
PF01055	MGAM
PF07686	IGSF3, IGHV5-51, TREML2
PF00054	CNTNAP3B, CELSR2, CNTNAP5
PF00566	USP6
PF01007	KCNJ12
PF06758	NBPF14, NBPF9, NBPF20, NBPF10, NBPF1, NBPF12, NBPF3

Then we used MetaCore software [13] to process these genes for identification of pathways driving the CLL development. MetaCore enrichment analysis consists of matching gene IDs with gene IDs in functional ontologies in MetaCore. Canonical pathway maps represent a set of signaling and metabolic maps covering human in a comprehensive way. All maps are created by Thomson Reuters scientists through a high-quality manual curation process based on published peer-reviewed literature. Experimental data is visualized on the maps as blue (for downregulation) and red (upregulation) histograms. The height of the histogram corresponds to the logarithm of p-value. The score represents an overrepresentation of genes in our list, included in MetaCore pathway. Thus, we identified the most significant pathways associated with CLL (Figure 1). For example, cytoskeletal signaling regulates several important cellular processes such as cell division, adhesion, polarity, migration, and movement. Immune response also plays a key in role cellular processes and attacks organisms and substances that invade body systems and cause disease

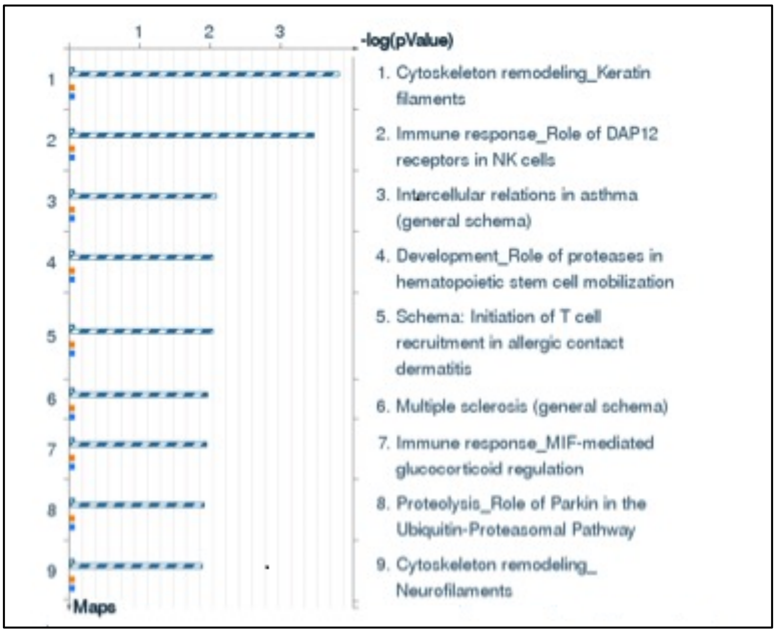

Figure 1. Pathway Enrichment scores.

Understanding of mechanisms driving the development of CLL will provide new biochemical targets for therapeutic intervention. It could change the strategy for treatment selection and monitoring and provide more successful healthcare with better outcomes for individual patients.

This research was supported by a grant of SpBSTU and Department of Oncology, University of Oxford.

References

1. G.Gaidano et. all (2012) Molecular pathologenesis of chronic lymphocytic leukemia, *J Clin Invest,* **122:**3432–3438.
2. U.Klein, R.Dalla-Favera (2008), Germinal centres: role in B-cell physiology and malignancy, *Nat Rev Immunol,* **8:**22–33.

3. R.N.Damle et. all (2007), CD38 expression labels an activated subset within chronic lymphocytic leukemia clones enriched in proliferating B-cells, *Blood,* **110:**3352–3359.
4. P.Danecek et. all (2011), The variant call format and VCFtools, *Bioinformatics,* **27:**2156–2158.
5. C. Greenman et. all (2007), Patterns of somatic mutation in human cancer genome, *Nature,* **446:**153–158.
6. A. Antoniadis (2003), Penalized Logistic Regression and Classification of Microarray Data, *University Joseph Fourier*.
7. M. Punta et. all (2012), The Pfam protein families database, *Nucleic Acids Res.,* **40:**290–301.
8. R. Tibshirani (2013), Modern regression 1: Ridge regression, *Data Mining,* **36:**462–662.
9. R. Tibshirani (1997), The LASSO method for variable selection in the cox model, *Statistics in medicine,* **16:**385–395.
10. A. Moore (2005), Cross-validation for detecting and preventing overfitting, *Carnegie Mellon University*.
11. J. Rodrigues (2010), Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation, *IEEE Trans. On Pattern Analysis and Machine Intelligence*.
12. M. Barron (2004), LOOCV: Stata module to perform Leave-One-Out Cross-Validation, *University of California, Santa Cruz*.
13. MetaCore Training Manual, Version 5.0, *GeneGo*.