MICA polymorphism identified by whole genome array associated with NKG2D-mediated cytotoxicity in T-cell large granular lymphocyte leukemia

Aaron D. Viny,[1,2] Michael J. Clemente,[2] Monika Jasek,[2] Medhat Askar,[4] Hemant Ishwaran,[3] Amy Nowacki,[3] Aiwen Zhang,[4] and Jaroslaw P. Maciejewski[1,2,5]

[1]Cleveland Clinic Lerner College of Medicine of Case Western Reserve University, Departments of [2]Translational Hematologic and Oncologic Research, [3]Quantatative Health Science, [4]Allogen Laboratories, and [5]Hematologic Oncology and Blood Disorders, Taussig Cancer Institute, Cleveland Clinic, Cleveland, USA

Citation: Viny AD, Clemente MJ, Jasek M, Askar M, Ishwaran H, Nowacki A, Zhang A, and Maciejewski JP. MICA *polymorphism identified by whole genome array associated with NKG2D-mediated cytotoxicity in T-cell large granular lymphocyte leukemia. Haematologica 2010;95(10): 1713-1721. doi:10.3324/haematol.2010.021865*

SUPPLEMENTARY APPENDIX

Single nucleotide polymorphism array analysis

Illumina Human NS-12 Genotyping Beadchip arrays (Illumina Inc., San Diego, CA, USA) and Affymetrix 250K arrays (Affymetrix Inc., Santa Clara, CA, USA) were used for analysis. For SNP array hybridization, 50 ng of purified granulocyte DNA were used according to the manufacturer's specifications. Illumina array data from 33 patients were analyzed using a modified 'random forests' analysis with two independent control populations - healthy age-matched controls (n=56) and disease controls with aplastic anemia (n=48) - as described below. Affymetrix array data from 52 patients were analyzed using an automated analysis with registry controls from the Framingham Heart Database (n=238, NHLBI, Bethesda, MD, USA), as well as Affymetrix-generated control genotypes. For both SNP array platforms, data were excluded from samples with call rates less than 89% and individual SNP were excluded with levels less than 95% across all samples.

Random forests for single nucleotide polymorphism array analysis

Averaging over trees, in combination with the randomization used in growing the base tree learner, enables random forests to approximate large classes of decision functions. In order to preserve a low generalization error, and mitigate high false positive rates, regularization was imposed on the number of candidate SNP used to split a node within a tree. This analysis yields no P-values as this is a hypothesis-generating procedure. The outcome used for the analysis was disease status (LGL leukemia or disease-free) and x-variables in the regression comprised SNP data. A two-class random forests' analysis using SNP data to predict disease status was implemented as follows. A randomly selected 80% subset of the data was chosen. Using these data, 1000 classification trees were grown as outlined previously.1 Specifically, each tree was grown from a randomly selected bootstrap sample of the data. The tree was grown to full size using the Gini index for splitting nodes. Each node of a tree was split using a randomly selected number of SNP. We used 117 SNP, which roughly equaled the square-root of the total number of SNP in the full data set (13,705). Once the 1000 trees had been grown, majority voting was used to predict disease status. Namely, the predicted class label (diseased or not diseased) in a terminal node of a tree yielded the predicted disease status for an individual for that tree. The predicted value for disease status was that label having a majority vote across all 1000 trees. The entire random forests' procedure was repeated 1000 times independently and results averaged over the runs. Computations were implemented using the R-package randomForest.2,3 Random forests and Classification and Regression Trees (CART) have previously been validated as statistical algorithms but have not been applied in the context of SNP array analysis.4,5

We ranked SNP on the basis of their variable importance (VIMP). VIMP measures how predictive a variable is after adjusting for all other variables in the model and indicates how effective the variable is for predicting outcome on new data.1,2,4,6 The top 15 SNP in each of the independent control populations were identified and their gene of origin was examined for potential biological relevance.

A second, independent statistical strategy was applied to Illumina and Affymetrix data using Exemplar statistical software (Sapio Science, Baltimore, MD, USA). For this analysis, SNP which violated the Hardy-Weinberg equilibrium were excluded ($P<0.01$) as were SNP which deviated from population stratification. From this analysis, a lambda-curve was generated and the lambda correction factor was applied to all remaining SNP for further analysis. Remaining SNP were subjected to $\chi 2$ analysis algorithms and ranked according to the Exemplar-generated score.

Allele-specific polymerase chain reaction amplification

To ensure the validity of the genotype calls and to exclude any identified SNP that was due to processing or technical error, calls generated using the Illumina SNP array were confirmed using traditional PCR genotyping. Top-ranking SNP identified though analysis of SNP-A technology were confirmed using allele-specific PCR amplification. Primers were designed using the Tetra-Primer web-based primer design.7 Allele-specific PCR primers for SNP rs1063635 included a forward outer primer (FOP), 5'-TCCAATTCTGCTAGAGTCCCAGCCTG-3', a reverse outer primer (ROP), 5'-AAGCACCAGCACTTTCCCTGAAAAAAAG-3', a forward inner primer (FIP), 5'-CTGTTCCTCTCCCCTCCTTAGAGGTGG-

3', and a reverse inner primer (RIP), 5'-TAGCAGGTGAACCTCTGCTCCTCTCATT-3'. The internal control product from the two outer primers resulted in a 460 bp product, the G-allele from the FIP and ROP resulted in a 215 bp product, and the A-allele from RIP and FOP resulted in a 300 bp product. Homozygosity was determined by identification of only one allele product and heterozygosity was determined by the identification of both products of the G and A-allele. The presence of the internal control product was required for genotype calling. Primers were also designed for other SNP identified in our analysis.

References

1. Breiman L. Machine Learning. Random Forests, 2001:5-32.
2. Liaw AaW, M. Classification and regression by randomForest. Rnews. 2002;Sect. 18-22.
3. Liaw AaW, M. Random Forest. [R-Package] 2007 [cited; 4.5-18:[Available from: http://cran.r-project.org
4. Breiman L, Friedman J.H., Olshen, R.A., and Stone, C.J. Classification and Regression Trees. Belmont, California: Wadsworth, 1984.
5. Breiman L. Statistical Modeling: the two cultures (with discussion). Statistical Science. 2001;16(3):199-231.
6. Ishwaran H. Variable importance in binary regression trees and forests. Electronic Journal of Statistics. 2007(1):25-31.
7. Ye S, Dhillon S, Ke X, Collins AR, Day IN. An efficient procedure for genotyping single nucleotide polymorphisms. Nucleic acids research. 2001;29(17):E88-8.