Genetic classification of acute myeloid leukemia by differential intergenic long noncoding RNA expression

Saioa Arza-Apalategi*, Daan Gilissen*, Anne van der Grinten*, Seline van den Oever, Branco Heuts, Erik van den Akker, Marieke Griffioen, Joop Jansen, Joost Martens, Bert van der Reijden

Long noncoding RNAs (lncRNAs) are >200 nucleotides in size, lack protein coding potential, and represent ~2% of the human genome. LncRNAs control important biological processes such as cell division, differentiation and apoptosis. They do this by binding chromatin, RNA and proteins to regulate all steps from gene expression to protein synthesis. Changes in lncRNA expression have been clearly implicated in malignant cell transformation including acute myeloid leukemia (AML).

For the most comprehensive AML lncRNA analysis to date, this study used RNA-seq data of 898 AML samples from four cohorts (AML-05, TARGET, BEAT and TCGA). First, we developed a bioinformatic pipeline for the discovery of unannotated lncRNAs. Only spliced, intergenic lncRNAs were included to minimize false positives. Canonical transcription marks of the novel lncRNA were assessed through CAGE-seq, and H3K4me1, -me3, and -K27ac ChIP on 6 KMT2A::MLLT3 AML samples. For further analysis, novel lncRNA and known lncRNA from GENCODE were combined. Next, we established whether genetically defined AML classes show specific lncRNA expression patterns using UMAPs and weighted gene co-expression network analysis (WGCNA). To determine how the identified lncRNA sets depend on the action of mutated transcription factors, we analyzed lncRNA expression following dTAG induced KMT2A::MLLT3 degradation and retinoic acid induced PML::RARA degradation in publicly available AML models.

Using our lncRNA discovery workflow, we identified 1560 novel lncRNAs and thereby expanded the known intergenic lncRNAs in the 898 AML samples by 27%. In the 6 KMT2A::MLLT3 validation samples, 220/1560 novel lncRNA were expressed, of which 60% showed overlap with at least one histone mark indicative of active promoters, in line with what has been described for lncRNA. Among the 1000 most variable lncRNAs, which were used for UMAP analysis, the novel lncRNA were overrepresented (352/1000), marking the importance of lncRNA discovery. The UMAP projections of the protein coding transcripts and lncRNAs both showed a very robust clustering pattern of samples according to genetic subclasses. In line with this, WGCNA identified sets of lncRNAs specific to these classes. Finally, lncRNA expression in KMT2A::MLLT3 and PML::RARA samples was found significantly altered upon degradation of the fusion protein ($p<0.01$). Thus, specific lncRNA expression patterns for AML depend on mutated transcription factors.

We conclude that lncRNAs show a similar degree of transcriptional changes compared to protein coding genes in AML and that mutated transcription factors are key to these changes. Given that lncRNA play pivotal roles in a plethora of biological processes it will be important to determine whether changes in their expression contribute to AML development.