Supplementary Methods

This study is part of the Leucegene project is an initiative approved by the research ethics boards of Université de Montréal and Maisonneuve-Rosemont Hospital. All sequencing data is available at GEO under accessing numbers GSE106272, GSE49642, GSE52656, GSE62190, GSE66917, and GSE67039.

RNA-sequencing

General workflow for sequencing, mutation analysis and transcripts quantification have been described previously [1]. Libraries were constructed with TruSeq RNA Preparation Kits (Illumina). Sequencing was performed on an Illumina HiSeq 2000 with 200 cycles paired end runs. Reads were mapped to the reference genome hg38 using STAR v2.7.1 [2]. Gene and transcript expression levels were quantified with RSEM version 1.3.2 [3]. Transcripts per million (TPM) were calculated for each transcript by normalizing for transcript effective length and library size. The TPM were used to compute gene aggregate value. Small polymorphisms were detected using FreeBayes version 1.3.1 [4]. km analyses were performed on reads trimmed of sequencing adapters and low quality 3' bases with Trimmomatic 0.38 in which the 31-mer were counted using Jellyfish 2.2.3 as previously reported [5–8].

Principal component analysis (PCA) was performed on log TPM of the differentially expressed genes (Supplementary Table 2) then Uniform Manifold Approximation and Projection (UMAP) was performed using a distance of 0.5 [9]. Samples were clustered on the PCA using the phenograph algorithm setting the k-nearest neighbors to 15 [10].

Exome

Tumor and normal gDNA were sequenced on NovaSeq6000 S4 with 100 cycles paired-end runs. Fastq reads were mapped to the reference genome GRCh38 using BWA-MEM [12]. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [11]. GATK version 4.1.3.0 was used for indel realignment and base quality score recalibration of the mapped reads. Mutect2 version 4.1.3.0 was used for variants calling in each sample in single-type mode [12].

Sanger sequencing

NPM1 mutations in patient samples were confirmed by sanger sequencing using forward primer ATGTTGCCCAGATTGGACTC and reverse primer GGCTCACAGACCCAATATCC for PCR amplification and sequencing.

Wild-type and mutant NPM1 overexpression

NPM1 expression vector was generated by DNA HiFi assembly (NEB E2621) of wild-type NPM1 PCR amplified into an expression vector downstream of a CMV promoter. Patient specific mutations were introduced by Q5 site directed mutagenesis (NEB E0554). Transfections in HEK-293t were performed in 6 well plates with 5ug of PEI using 3.75 ug of DNA on 100 000 cells grown on poly-lysine coated coverslips. Cells were fixed with cold methanol, stained with antibody against NPM1 (Santa Cruz sc-56622) and mounted using prolong gold with DAPI. Images were taken on a Leica inverted microscope at 100X magnification. For patient samples microscopy, 150 000 cells were cytospinned at 113g for 4 minutes with slow acceleration. After drying, slides were processed as above.

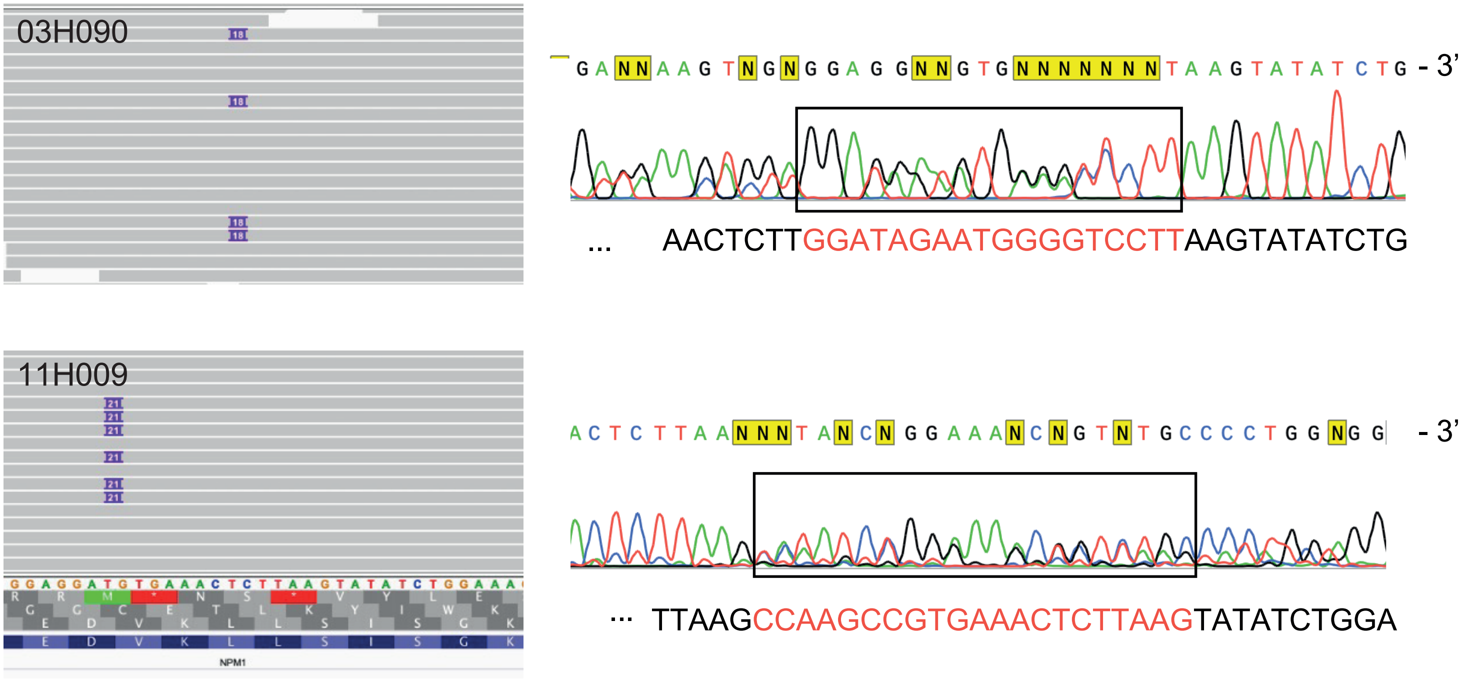
Supplementary Table 1 *NPM1* mutations in the Leucegene cohort
Supplementary Table 2 Differentially expressed genes in *NPM1*e12 AML
Supplementary Table 3 Co-occurring mutations in *NPM1*e5 mutations

Supplementary References

1. Lavallée V-P, Chagraoui J, MacRae T, et al. Transcriptomic landscape of acute promyelocytic leukemia reveals aberrant surface expression of the platelet aggregation agonist Podoplanin. *Leukemia*. 2018;32(6):1349–1357.
2. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
3. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
4. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907
5. Chen C, Khaleel SS, Huang H, Wu CH. Software for pre-processing Illumina next-generation sequencing short read sequences. *Source Code Biol. Med.* 2014;9:8.
6. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–2120.

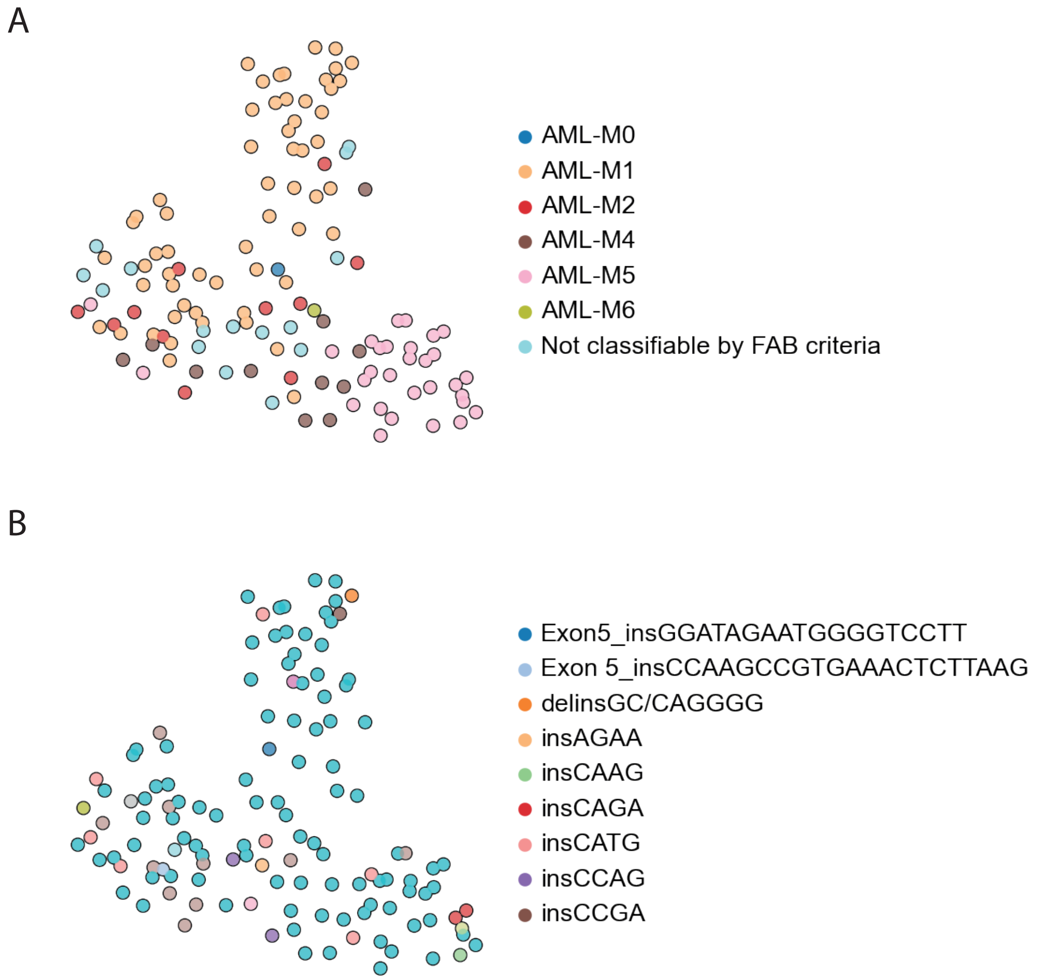
7. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27(6):764–770.
8. Audemard EO, Gendron P, Feghaly A, et al. Targeted variant detection using unaligned RNA-Seq reads. *Life Sci. Alliance*. 2019;2(4):.
9. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426v3
10. Levine JH, Simonds EF, Bendall SC, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*. 2015;162(1):184–197.
11. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997
12. Benjamin D, Sato T, Cibulskis K, et al. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* 861054; doi: https://doi.org/10.1101/861054

Supplementary Figure 1



Supplementary Figure 1. Confirmation of NPM1 e5 mutations. Integrative Genomic Viewer of exon sequencing data (left) showing the localization of the exon 5 insertion in the 2 primary AML patients. Sanger sequencing (right) of the same samples. 03H090 (top) is sequenced from the 3’ end and 11H009 (bottom) is sequenced from the 5’ end. The reference is shown above the chromatograph and the mutated sequence is written below. The insertion is in red and is highlighted by a box on the chromatograph.

Supplementary Figure 2



Supplementary Figure 2. Transcriptomic diversity in NPM1-mutated AML Uniform Manifold Approximation and Projection (UMAP) performed on the NPM1-mutated samples (n = 127) using the top differentially expressed genes from Supplementary Table 1 followed by principal component analysis and colored by French American British (FAB) classification (A) and mutation type (B).