Supplement to:

Genomic characterization of a case of acute myeloid leukemia with promyelocytic features characterized by expression of a novel *RARG-CPSF6* fusion.

Christopher A. Miller[1,2,3*], Christopher Tricarico[1*], Zachary L. Skidmore[2], Geoffrey L. Uy[1], Yi-Shan Lee[3], Anjum Hassan[3], Michelle O'Laughlin[2], Heather Schmidt[2], Ling Tian[1], Eric J. Duncavage[3], Malachi Griffith[2,4,5], Obi L. Griffith[1,2,4,5], John S. Welch[1,4], Lukas D. Wartman[1,2,5#]

[1]Department of Medicine, Washington University School of Medicine

[2]McDonnell Genome Institute at Washington University

[3]Department of Pathology and Immunology, Washington University School of Medicine

[4]Department of Genetics, Washington University School of Medicine

[5]Siteman Cancer Center, Washington University School of Medicine

* These authors contributed equally to this work

# Corresponding author. Email: Lwartman@wustl.edu

Supplemental Data

Dataset 1: Somatic SNVs and indels with readcounts. Tier 1 (coding) mutations are sorted to the top

Dataset 2: Copy number alterations derived from single-sample tumor analysis.

Dataset 3: Structural Variants in vcf format

Dataset 4: Gene fusions detected by INTEGRATE

Dataset 5: Gene expression values for all genes in this patient

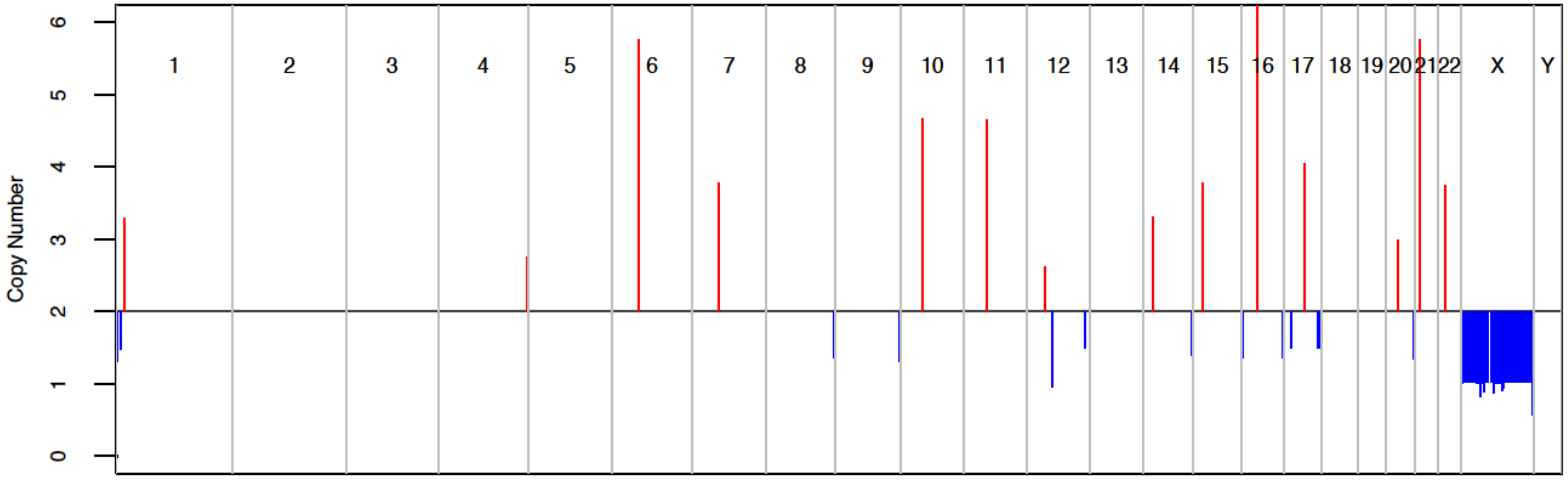
Dataset 6: Breakpoint PCR validation results

Dataset 7: Genes and expression values for the APL expression signature. Numbers at top are TCGA ids (i.e. 2840 is sample TCGA-AB-2840). “GTB14” is the *RARG*-truncated patient described in this study.

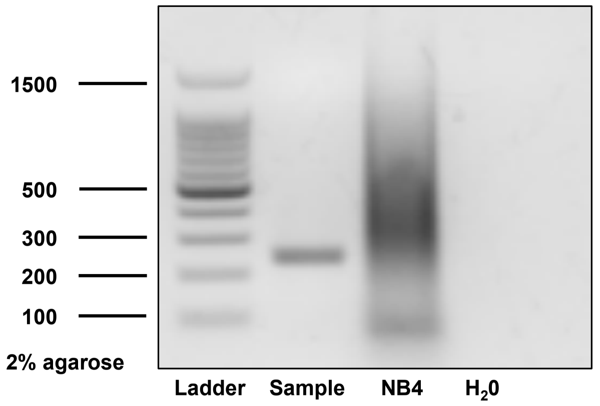
Supplemental Table 1.

Gene Name	AA Change	WGS Reference Count	WGS Variant Count	WGS Tumor VAF	RNA Reference Count	RNA Variant Count	RNA Tumor VAF
PCMTD1	p.L343F	94	26	21.67	315	0	0
BMPR1A	p.Y245*	35	27	42.86	3	0	0
WT1	p.S381*	34	28	45.16	18	20	52.63
WT1	p.E340*	31	27	45.76	20	27	57.45
NEAT1	-/T	32	13	28.89	12	0	0
GRIA4	p.R278C	41	23	35.38	0	0	0
KRTAP9-1	p.157in_frame_ins	61	12	16.44	0	0	0
ZNF433	p.T361I	25	27	51.92	4	2	33.33
SLC9B1P1	p.I179in_frame_del	3	3	50	0	0	0

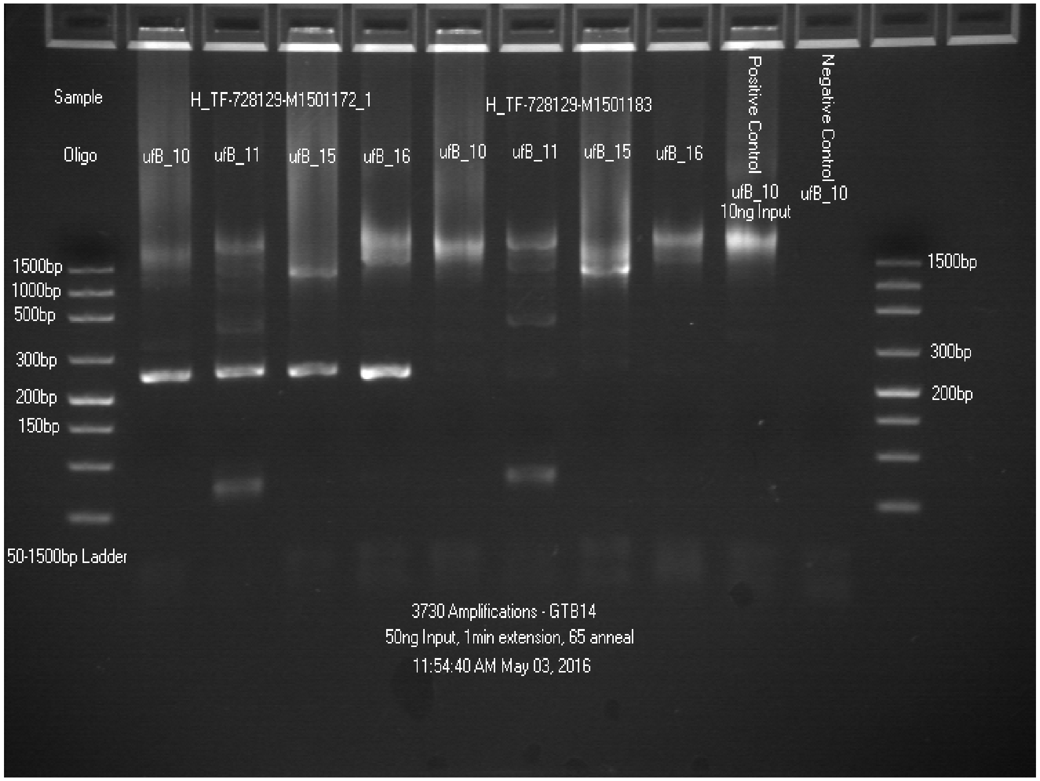
Supplemental Figure 1. Copy number analysis of WGS data showed scattered small amplifications and deletions throughout the genome. The subclonal loss of chromosome Y detected by conventional cytogenetics was not identified here.



Supplemental Figure 2. Reverse transcription PCR results showed the expected band at 237 bp, which confirmed the presence of the *RARG-CPSF6* fusion in the RNA of the sample but not control NB4 cells or water.



Supplemental Figure 3. *RARG-EIF4B* Genomic Fusion PCR Followed by Sanger Sequencing Results. PCR and primer sequences noted. Sample is depicted in the first four lanes of the gel with skin comparator in the next four lanes. The band at 250bp in the tumor sample is the expected size for the expressed fusion gene.



***RARG-EIF4B* Fusion Sequencing Results**

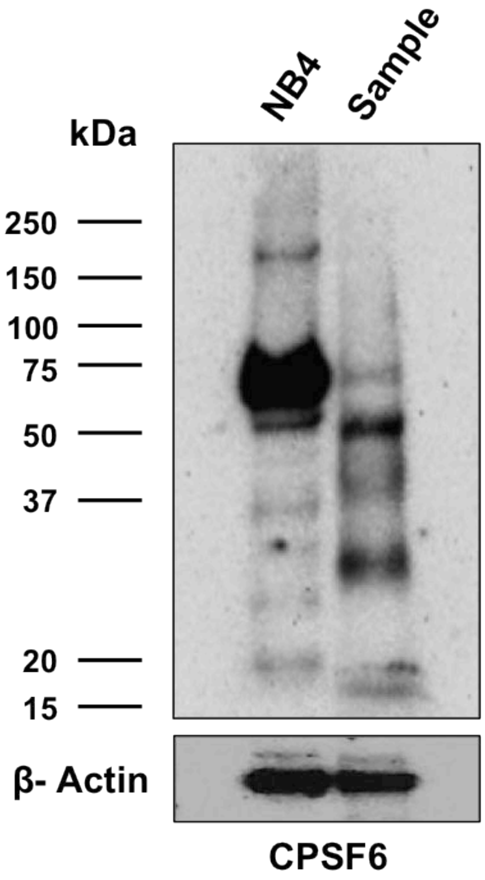
RARG
atatatttagagacaaagtcttgggcaaattgctttacttttttaatcttgtttgtaaactgggctattcttcacgtaccttg
taacgctgtcttgaggaCTAAATTAGTTAATATACAATGTAAAGCCAGGCACCATGGTG
CACATCTGTAGTTCCAGCTACTCAGGAGGCTGAGAGAGAAGAATGGTGGGA
ACCCGGGAGGCAGAGCTTGCAGGAGTGGAGATGGCGCCGCTGCACTCCA
GCCTGGGTgacagagtgagactccccatctcaaaaaaaaaaaaaaagcattggctaaggatgacatc
agtgga

EIF4B
atatatttagagacaaagtcttgggcaaattgctttacttttttaatcttgtttgtaaactgggctattcttcacgtaccttg
taacgctgtcttgaggaCTAAATTAGTTAATATACAATGTAAAGCCAGGCACCATGGTG
CACATCTGTAGTTCCAGCTACTCAGGAGGCTGAGAGAGAAGAATGGTGGGA
ACCCGGGAGGCAGAGCTTGCAGGAGTGGAGATGGCGCCGCTGCACTCCA
GCCTGGGTgacagagtgagactccccatctcaaaaaaaaaaaaaaagcattggctaaggatgacatc
agtgga

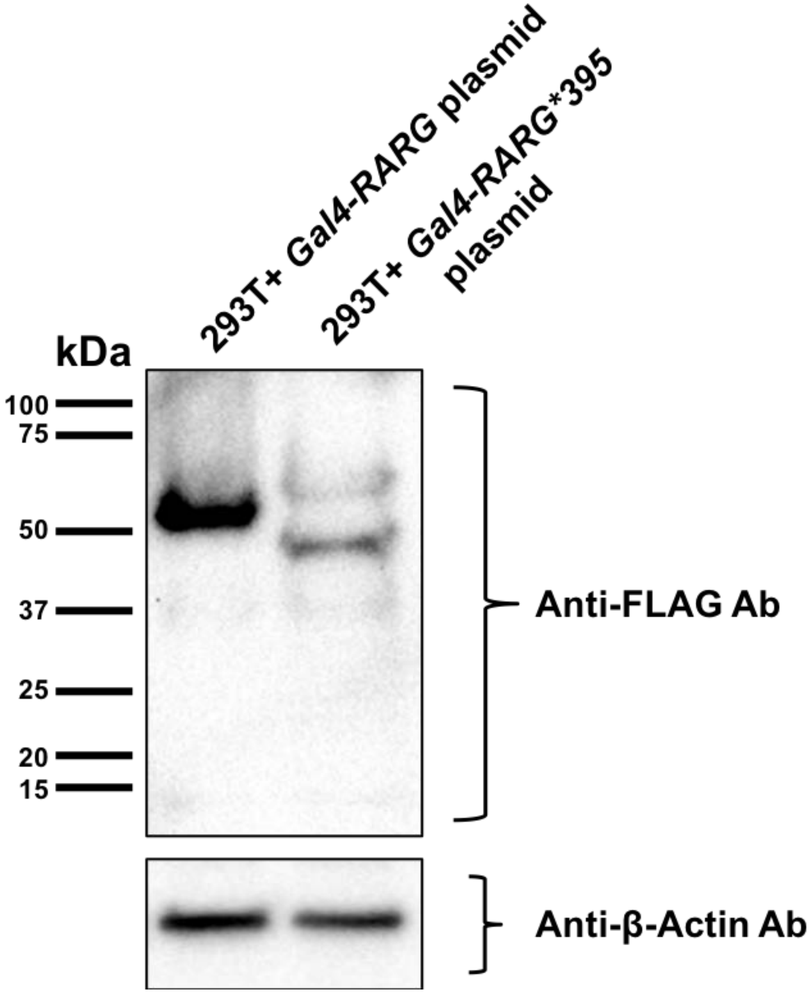
Overlap/Inverted Repeat
atatatttagagacaaagtcttgggcaaattgctttacttttttaatcttgtttgtaaactgggctattcttcacgtaccttg
taacgctgtcttgaggaCTAAATTAGTTAATATACAATGTAAAGCCAGGCACCATGGTG
C**ACATCTGTAGTTCCAGCTACTCAGGAGGCTGAGAGAGAA**GAATGGTGGG
AACCCGGGAGGCAGAGCTTGCAGGAGTGGAGATGGCGCCGCTGCACTCCA
GCCTGGGTgacagagtgagactccccatctcaaaaaaaaaaaaaaagcattggctaaggatgacatc
agtgga

Lane	Forward Primer Sequence	Reverse Primer Sequence
ufB_10	TGTAAAACGACGGCCAGTTGGGGAGTCTCACTCTGTCA	CAGGAAACAGCTATGACCACTGGGCTATTCTTCACGTACC
ufB_11	TGTAAAACGACGGCCAGTGCCATCTCCACTCACTGCAA	CAGGAAACAGCTATGACCACAAAGTCTTGGGCAAATTGCT
ufB_15	TGTAAAACGACGGCCAGTACTGGGCTATTCTTCACGTACC	CAGGAAACAGCTATGACCTGGGGAGTCTCACTCTGTCA
ufB_16	TGTAAAACGACGGCCAGTTCACGTACCTTGTAACGCTGT	CAGGAAACAGCTATGACCTGAGATGGGGAGTCTCACTCT

Supplemental Figure 4: Immunoblotting using an anti-CPSF6 antibody of the sample did not show evidence of an expressed RARG-CPSF6 fusion protein and only minimal evidence of WT CPSF6 protein. NB4 cells served as the control. The predicted size of the RARG-CPSF6 fusion is 78 kDa, and the predicted size of WT CPSF6 is 52 or 63 kDa, based on the size of isoforms 1 and 2 respectively. However, after post-translational modifications, CPSF6 migrates at 68 kDa. The experiment was repeated independently with similar results.



Supplemental Figure 5: Immunoblotting using an anti-FLAG antibody of 293T cells transduced with either the *Gal4-RARG* plasmid or the *Gal4-RARG*395* truncation plasmid confirms that both proteins were expressed. The experiment was repeated independently with similar results.



Supplemental Methods

Whole genome sequencing and somatic variant analysis

The patient was enrolled in a single-institution, tissue-banking protocol approved by the human studies committee at Washington University. He provided written informed consent for comprehensive sequencing studies, including whole genome sequencing (WGS) and RNA-Seq. WGS libraries were created using the Illumina TruSeq PCR-free kit and sequenced on a HiSeq X instrument. From WGS data, we obtained 72x haploid coverage of the tumor and 38x coverage of the skin. Sequence data was aligned to reference sequence build GRCh37-lite-build37 using BWA-MEM[1] version 0.7.10 (params: -t 8), then merged and deduplicated using Picard version 1.113 (https://broadinstitute.github.io/picard/). SNVs were detected using the union of four callers: 1) Samtools[2] version r982 (params: mpileup -BuDs) intersected with Somatic Sniper[3] version 1.0.4 (params: -F vcf –G -L -q 1 -Q 15) and processed through false-positive filter v1 (params: --bam-readcount- version 0.4 --bamreadcount-min-base-quality 15 --min-mapping-quality 40 --min-somatic-score 40), 2) VarScan[4] version 2.3.6 filtered by varscan-high-confidence filter version v1 and processed through falsepositive filter v1 (params: --bam-readcount-version 0.4 --bam-readcount-min-base-quality 15), 3) Strelka[5] version 1.0.11 (params: isSkipDepthFilters = 0), and 4) Mutect[6] v1.1.4. Indels were detected using the union of 4 callers: 1) GATK[7] somatic-indel version 5336 2) Pindel[8] version 0.5 filtered with Pindel somatic calls and VAF filters (params: --variant-freq-cutoff=0.08), and Pindel read support, 3) VarScan[4] version 2.3.6 filtered by varscan-high-confidence- indel version v1 and 4) Strelka[5] version 1.0.11 (params: isSkipDepthFilters = 0). SNVs and Indels were further filtered by removing artifacts found in a panel of 905 normal exomes[9], removing sites that exceeded 0.1% frequency in the 1000 genomes or NHLBI exome sequencing projects, and then using a bayesian classifier (https://github.com/genome/genome/blob/master/lib/perl/Genome/Model/Tools/Validation/IdentifyOutliers.pm) and retaining variants classified as somatic with a binomial log-likelihood of at least 10.

Copy number aberrations were detected using copyCat version 1.6.10 (https://github.com/chrisamiller/copyCat) in single sample mode (default params, except for --per-read-length --per-library, gapExpansion=1.3). Alterations smaller than 15 consecutive windows were filtered. Paired CN analysis was not possible because of library preparation artifacts in the normal sample, which resulted in uneven genomic coverage. Structural variants were detected using Lumpy[10] version 0.2.610, with MIN_MAPQ=20 and Discordant_z=5

RNA sequencing (RNA-Seq) analysis

RNA libraries were prepared using the TruSeq stranded kit and sequenced on one lane of an Illumina HiSeq 2500 v4. We obtained 265 million reads from RNA-Seq.

RNA-Seq data was aligned with Tophat[11] version 2.0.8 (denovo mode, params: --library-type frfirststrand--bowtie-version=2.1.0) and expression levels were calculated with Cufflinks[12] version 2.1.1 (params: --max-bundle-length 10000000 --max-bundle-frags 10000000). Gene fusions were detected using INTEGRATE[17] version 0.2.

We used Kalisto[18] version 0.43.0 to extract reads from the region of the predicted *RARG-CPSF6* breakpoints using a k-mer index built from the putative transcript fusion. We then re-aligned these reads with BWA-MEM[1] version 0.7.9a to a reference comprised of *RARG* wild-type transcript, *CPSF6* wild-type transcript, and predicted *RARG-CPSF6* fusion transcript (ENST00000425354; exons 1-9 and ENST00000435070; exons 6-10).

By manual review of the RNA-Seq data, the Integrated Genomics Viewer (IGV) view showed alignment against the predicted fusion transcript with a large number/variety of reads (>220) perfectly aligned to the predicted junction and supporting the fusion transcript.[19] We observed only three reads supporting wild-type *RARG* expression, which are likely indicative of a small number of contaminating benign cells (**Supplemental Data File 6**). The *EIF4B-RARG* fusion involves transcription of the first 6 exons of *EIF4B* and then picks up after the *RARG* locus with a cryptic exon. Only 8 amino acids are predicted to be added to *EIF4B* before a stop codon. Therefore, this cryptic exon (if even a real exon) is clearly out of frame. Loss-of-function of EIF4B is predicted. The cryptic exon does not BLAST to any expressed sequence tag (EST) and is only a partial 64-500 (94%) BLAST match against a refseq_rna for a macaque ncRNA uncharacterized transcript. There was no evidence of a reciprocal *CPSF6-RARG* fusion by WGS or RNA-Seq data. There was expression of the WT *CPSF6* transcript that was approximately equal to that of the *RARG-CPSF6* fusion.

The Cancer Genome Atlas (TCGA) RNA-Seq data was processed using TopHat and Cufflinks as previously described. R was used to compare values for each gene with at least three non-NA values in both APL (FAB type M3) samples and other subtypes using a Student's t-test. The 250 most significantly upregulated and downregulated genes were retained as a "signature" of APL (**Supplemental Data File 5**). The expression values from these 500 genes in this *RARG-CPSF6* fusion case were then added to the resulting matrix, and unsupervised hierarchical clustering was performed using the heatplot function from the "made4" R package. The co-clustering result was robust and also occurred when 100- or 250-gene signatures were defined in a like manner.

Data Deposition and Access

The sequence data for the leukemia and matched normal sample has been deposited in the database of Genotypes and Phenotypes (dbGaP) under accession number: phs000159.

Western blotting

NB4 cells were obtained from Dr. Timothy Ley. They were cultured in RPMI 1640 with 10% fetal calf serum. APL cells with t(15;17) were banked on the same protocol as the patient sample. NB4 or APL cells were rinsed with PBS, treated with 100 µM diisopropylfluorophosphate (Sigma-Aldrich, St. Louis, MO, USA), and resuspended in urea lysis buffer (7 M urea, 2 M thiourea, 30 mM Tris, pH 8.5) and then snap frozen in liquid nitrogen. Lysates were boiled in 6x loading buffer (0.01 M Tris-HCl, pH 6.8, 8% glycerol, 0.1 mg/mL bromophenol blue, 2% SDS, 1% β-mercaptoethanol) for 8 minutes at 100°C. 20 µg of protein lysate for each sample was loaded and then separated on a 4-15% precast gel (Bio-Rad Laboratories; Hercules, CA, USA) and then transferred onto PVDF blotting membrane (Amersham, GE Healthcare; Little Chalfont, United Kingdom). The membrane was incubated in 5% skim milk for 1 hr at room temperature and probed with an anti-RARG N-terminal antibody (1:1000; Aviva Systems Biology; San Diego, CA, USA); an anti-RARG C-terminal antibody (1:1000; Santa Cruz Biotechnology; Dallas, TX, USA); an anti-CPSF6 antibody (1:1000; Abcam; Cambridge, United Kingdom) or an anti-β-Actin antibody (1:10 000; Cell Signaling Technology; Danvers, MA, USA) overnight at 4°C. Immune complexes were revealed by peroxidase-linked anti-mouse IgG (1:10 000; GE Healthcare) or peroxidase-linked anti-rabbit IgG (1:10 000; GE Healthcare). Western blots were visualized by chemiluminescence using ECL Prime Western Blotting Detection Reagent (Amersham, GE Healthcare) and a ThermoFisher myECL Imager (Waltham, MA, USA).

RT-PCR validation of predicted *RARG-CPSF6* fusion followed by standard Sanger sequencing validation

RNA was prepped from cryopreserved bone marrow aspirate taken at the time of diagnosis using the Direct-zol™ RNA MicroPrep kit (Zymo Research; Irvine, CA, USA), and cDNA was made using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems; Foster City, CA, USA). Primers were purchased from Integrated DNA Technologies (Coralville, IA, USA). 40 ng of cDNA (20 ng/µL) was used for the PCR reaction. The predicted product is 237 bp.

RARG-CPSF6 F primer (in exon 9 of *RARG*)

5'-CCGAAAAAGTGGACAAGCTG-3'

RARG-CPSF6 R primer (in exon 6 of *CPSF6*)

5'-CAGCTAGAGGAGGAGGCAGA-3'

PCR conditions: 1. 95°C 3:00 2. 56°C 0:30 3. 72°C 1:00 4. 95°C 0:30 5. 72°C 10:00 6. GO TO 2 (35 cycles) 7. 4°C ∝.

***UAS-GFP* and *Gal4* constructs co-transduction experiments**

We generated a custom *Gal4-RARG* plasmid, in which we replaced the DNA binding domain of *RARG* with the *Gal4* domain, and a *Gal4*-truncated *RARG* plasmid (designated *RARG**395) as above except *RARG* was truncated after the 9th exon (Genewiz, LLC; South Plainfield, NJ, USA). The *UAS-GFP*, *Gal4-RARA*, and *Gal4-VP16* constructs have been previously described.[21] We used an MSCV-3xFlag-*Gal4*-plasmid-IRES-mCherry retrovirus and a *UAS-GFP* plasmid for these experiments. Transduction of the retrovirus was done as previously described[21], and the *UAS-GFP* plasmid was transfected into the 293T cells using Dharmafect™ 1 per manufacturer's instructions (GE Healthcare). We confirmed protein expression after transduction with the both the *Gal4-RARG* and *Gal4-RARG**395 plasmid by western blot using an anti-FLAG monoclonal antibody (Sigma-Aldrich; St. Louis, MO) (**Supplemental Figure 5**). Otherwise, experimental conditions and flow cytometry were done as described.[21] ATRA was purchased from Sigma-Aldrich. BMS961 was purchased from R&D Systems (Minneapolis, MN, USA). For statistical analysis, we used GraphPad Prism software (Version 7; La Jolla, CA, USA) using ANOVA with Bonferroni correction for multiple comparisons. Pairwise comparisons that are significant are indicated. *Gal4-VP16* was not included in the analysis due to absent values in the BMS961 and ATRA treatment conditions, as ANOVA requires all samples to have results for all conditions. All treatment conditions were significantly different from samples without *UAS-GFP* transfected, and these samples were removed from the analysis. For the PPRE-Luciferase and ApoA1-Luciferase co-transduction/transfection experiments, we performed them as above using lentiviral constructs expressing GFP, WT *RARG* or the truncated *RARG**395 (pLenti-C-mGFP vector, OriGene, Rockville, MD, USA).

Supplemental References

1. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 *Q-Bio.* 2013;http://arxiv.org/abs/1303.3997(accessed 13 Mar2015).

2. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinforma Oxf Engl.* 2009;**25**:2078–2079.

3. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinforma Oxf Engl.* 2012;**28**:311–317.

4. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012;**22**:568–576.

5. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinforma Oxf Engl.* 2012;**28**:1811–1817.

6. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol.* 2013;**31**:213–219.

7. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;**20**:1297–1303.

8. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinforma Oxf Engl.* 2009;**25**:2865–2871.

9. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012;**490**:61–70.

10. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* 2014;**15**:R84.

11. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009;**25**:1105–1111.

12. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc.* 2012;**7**:562–578.

17. Zhang J, White NM, Schmidt HK, et al. INTEGRATE: Gene fusion discovery using whole genome and transcriptome data. *Genome Res.* 2015;gr.186114.114.

18. Bray NL, Pimentel H, Melsted P and Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology.* 2016;**34**:525–527.

19. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative Genomics Viewer. *Nature Biotechnology.* 2011;**29**:24–26.

20. Ley TJ, Miller C, Ding L, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med.* 2013;**368**(22):2059-74.

21. Niu H, Hadwiger G, Fujiwara H, Welch JS. Pathways of retinoid synthesis in mouse macrophages and bone marrow cells. *J Leukoc Biol.* 2016;**99**(6):797-810.