Subtype-specific regulatory network rewiring in acute myeloid leukemia

Assi, Salam A.; Imperato, Maria Rosaria; Coleman, Daniel J. L.; Pickin, Anna; Potluri, Sandeep; Ptasinska, Anetta; Chin, Paulynn Suyin; Blair, Helen; Cauchy, Pierre; James, Sally R.; Zacarias-cabeza, Joaquin; Gilding, L. Niall; Beggs, Andrew; Clokie, Sam; Loke, Justin C.; Jenkin, Phil; Uddin, Ash; Delwel, Ruud; Richards, Stephen J.; Raghavan, Manoj

DOI:
10.1038/s41588-018-0270-1

Document Version
Peer reviewed version

Citation for published version (Harvard):
Assi, SA, Imperato, MR, Coleman, DJL, Pickin, A, Potluri, S, Ptasinska, A, Chin, PS, Blair, H, Cauchy, P, James, SR, Zacarias-cabeza, J, Gilding, LN, Beggs, A, Clokie, S, Loke, JC, Jenkin, P, Uddin, A, Delwel, R, Richards, SJ, Raghavan, M, Griffiths, MJ, Heidenreich, O, Cockerill, PN & Bonifer, C 2019, 'Subtype-specific regulatory network rewiring in acute myeloid leukemia', *Nature Genetics*, vol. 51, no. 1, pp. 151-162. https://doi.org/10.1038/s41588-018-0270-1

Subtype-specific regulatory network rewiring in acute myeloid leukemia

Salam A. Assi[1*], Maria Rosaria Imperato[1*], Daniel J. L. Coleman[1*], Anna Pickin[1], Sandeep Potluri[1], Anetta Ptasinska[1], Paulynn Suyin Chin[1], Helen Blair[2], Pierre Cauchy[1], Sally R. James[3], Joaquin Zacarias-Cabeza[1], Liam Niall Gilding[1], Andrew Beggs[1], Sam Clokie[4], Justin C. Loke[1], Phil Jenkin[5], Ash Uddin[5], H. Ruud Delwel[6], Stephen J. Richards[7], Manoj Raghavan[1,8], Michael J. Griffiths[4], Olaf Heidenreich[2], Peter N. Cockerill[1&] and Constanze Bonifer[1&]

[*] **Equal contribution**

[&] **Corresponding authors**

Acute myeloid leukemia (AML) is a heterogeneous disease caused by a variety of mutations in transcription factors, epigenetic regulators and signaling molecules. To determine how different mutant regulators establish AML subtype-specific transcriptional networks we performed a comprehensive global analysis of cis-regulatory element activity and interaction, transcription factor occupancy and gene expression patterns in purified leukemic blast cells. Here, we focussed on specific sub-groups of patients carrying mutations in genes encoding transcription factors (*RUNX1, CEBPA)* and signaling molecules (*FTL3-ITD, RAS, NPM1).* Integrated analysis of these data demonstrates that each mutant regulator establishes a specific transcriptional and signaling network unrelated to normal cells sustaining the expression of unique sets of genes required for AML growth and maintenance.

Introduction

Acute myeloid leukemia (AML) is characterized by blocked myeloid lineage differentiation and accumulation of leukemic blast cells. AML is a highly heterogeneous disease caused by different types of mutations affecting signaling pathways as well as transcriptional and epigenetic regulators[1-3]. Recurrent mutations include loss of function mutations in transcription factors (TFs) controlling hematopoietic development, such as RUNX1, GATA2 or C/EBPα[4], and gain of function mutations in signaling molecules such as FLT3, KIT, JAK2 and NRAS regulating inducible TFs such as NF-κB, STAT or AP-1 family members[5],[6,7]. The most common FLT3 mutations are internal tandem duplications (FLT3-ITD), which give rise to a constitutively active growth factor receptor[8,9] and often occur together with nucleophosmin1 mutations (NPM1). Another major group of mutations alters genes encoding epigenetic and chromatin regulators[10,11]. Genes belonging to this class play widespread roles in development and differentiation by controlling establishment, maintenance and extinction of lineage-specific gene expression programs. These include regulators of histone and DNA methylation such as MLL, EZH2, BCOR, TET2, DNMT3A and IDH1/2[11-17]. In normal cells, all common mutation targets cooperate to control the finely balanced gene expression changes essential for cell differentiation and lineage commitment.

TFs interact with defined target gene sequences and recruit epigenetic regulators to program specific chromatin states and mediate the coordinated activation and de-activation of cis-regulatory elements driving gene expression[18,19]. Distal cis-regulatory elements interact directly with proximal promoter elements, an arrangement that is both dynamic and robust[20,21]. From global studies examining a few selected types of AML we know that gene expression patterns and the epigenetic landscape differ from normal cells[22-27]. However, how the disruption of specific TF activity leads to a specific pattern of aberrant chromatin programming and changes in gene expression in AML is unclear. We do not know at the

global level which cis-regulatory elements are affected in their activity in different types of AML, how their activity is altered in patients carrying TF and signaling mutations, how such differential activity relates to the differentiation block in primary cells of actual patients and which factors maintain their transcriptional networks.

In this study we addressed these questions by collecting transcriptome, digital footprinting and chromatin conformation capture data from purified leukemic blasts from AML patients with defined transcription factor and signaling molecule mutations with the aim of defining the components of AML subtype-specific regulatory circuitries. Our study provides global insights into mutation-specific chromatin programming, and comprises a comprehensive resource of the transcriptional networks of different AML subtypes, highlighting pathways required for tumour maintenance

Results

AML with different mutant regulators adopt unique chromatin landscapes

In order to examine how specific TF and signaling mutations alter the epigenome of AML, we purified CD34+ or CD117+ leukemic blast cells from bone marrow or peripheral blood samples from a cohort of AML patients (Fig 1A). After determining the mutation status by targeted sequencing and cytogenetics (Table S1), we selected a cohort of patients with defined mutations in transcription factor, signaling and epigenetic regulator genes. Mutations included: *RUNX1* mutations affecting DNA-binding (D-type) or lacking the trans-activating domain (T-Type), t(8;21) translocations fusing the DNA binding domain of RUNX1 to the co-repressor ETO, inv(16) which fuses CBFβ to smooth muscle myosin heavy chain 11 (MYH11) protein, independent mutations of both alleles of the *CEBPA* gene whereby at one mutation leads to loss of DNA-binding activity[28] and FLT3-ITD with or without NPM1 mutations. One of the patients carrying a RUNX1 mutation (RUNX1-T-7) who had Non-

Hodgkin Lymphoma (NHL) was included as an alternative class of patient. To identify AML-specific gene regulatory networks we performed high read depth DNaseI-Seq (Fig.1B) and RNA-Seq (Fig S1A) on 29 samples comprising seven major groups, and at least one analysis on 12 additional samples, with mutations such as *NRAS, CBL, JAK2, SRFS2*, or inv(3), as defined in Table S1. Samples were compared to CD34+ mobilized peripheral blood stem cells (PBSCs) from the peripheral blood of two healthy individuals and to cord blood (CB) CD34+ cells. To provide the community with a data resource, we established an online database containing multiple data-sets including a genome browser (see Data Availability).

Unsupervised clustering revealed that distal DHSs clustered in different groups according to the class of mutations (Fig 1C). Samples with FLT3-ITD and/or NPM1 mutations represented one major group with sub-clusters for patients with NPM1 mutations or carrying two FLT3-ITD alleles, but excluding a FLT3-ITD patient carrying a RUNX1 mutation. DHSs from the t(8;21), inv(16) and *CEBPA* double mutant patients clustered as discrete groups within a larger group, indicating that these mutations affect similar pathways. Examples of these patterns can be seen for DHSs in the *POU4F1* locus in t(8;21) and *CEBPA*-mutated samples, and DHSs in the *FOXC1* locus in patients with FLT3-ID or NPM1 mutations (Fig. S1B). DHSs from patients with RUNX1 mutations were more heterogeneous and formed a larger cluster together with the PBSCs and the inv(3) patients. The NHL (RUNX1-T-7) and the NPM1/RAS sample were unrelated to any of the others. We further validated our findings by including an independently derived published ATAC-Seq data-set in our analysis[25], confirming that mutations in FLT3 underpin one major component of the clustering (Figure S2A). In contrast, the presence or absence of epigenetic mutations such as *DNMT3A* did not influence chromatin accessibility levels (Fig. S2B) or gene expression (data not shown).

Unsupervised clustering analysis of RNA-Seq data from the same patients (Fig S1C, S2C) revealed strong correlations between mutation-specific accessible chromatin landscapes and mutation-specific differential gene expression. This was again exemplified at the *POU4F1* and *FOXC1* loci (Fig. S1D) where the mRNA patterns correlated well with the chromatin profiles (Fig. S1B). We identified distinct patterns of expression for specific TF genes in different AML types (Fig S2D), with, for example a number of homeo-domain gene family members (*HOX, NKX, IRX* and *PBX* families) specifically up-regulated in the FLT3-ITD and NPM1-mutated patients. In summary, our comparative analyses show that aberrant TFs and chronic signaling impose distinct mutation group-specific programs of chromatin accessibility and gene expression, irrespective of the presence of other classes of mutations such as DNMT3A.

We next investigated, whether the mutation class and its associated DHS pattern correlated with a block at a specific stage of the differentiation. Here we again used published ATAC-Seq data[25] describing the open chromatin landscape of normal stem and progenitor cells (Fig 2A). Our DNaseI-Seq data correlated well with these data (Fig S3A and S3B), whereby CD34+ PBSC sequences clustered with hematopoietic stem cells (HSCs) and early progenitors but not monocytic cells. When compared to the various types of progenitor cells, t(8;21), inv(16), CEBPA(x2) and NPM1-mutated AML displayed distal element patterns most similar to those of normal GMPs, with some differentiation into monocytes (Fig 2B). In contrast, RUNX1 and FLT3-ITD/NPM1 mutated AML displayed a spread of lineage-specific patterns with little or no monocytic differentiation (Fig 2B). Gene-set enrichment analysis comparing the gene expression patterns of AML cells with the various progenitor stages confirmed that the mutation-group specific cistrome was mirrored by the gene expression pattern (Fig 2C). However, although AML subtypes showed some characteristics of normal progenitor cells, they still clustered away from normal cells (Fig S3B). Importantly, our mutation analyses showed no indications for the presence of

confounding major sub-clones in the purified undifferentiated AML cell population, as mutations were present at close to either a 50% or a 100% allele frequency (Dataset S1).

AML-specifically active cis-regulatory elements cluster into common and unique groups

We next examined which active cis-regulatory elements were specific for each AML subtype and which TF families were responsible for their activation. To this end, we defined the union of all AML-specific DHSs as compared to CD34+ PBSCs and performed k-mean clustering to identify unique and common DHSs shared between patients, which identified 20 distinct DHS clusters (Fig 3A). Less than half of these DHSs were found in any of the Corces et al.[25] progenitor data ATAC-Seq sets and the percentage overlap varied substantially between clusters (range 2 – 40%; Fig. 3C). These data indicated that AML cells did indeed reprogram their chromatin and adopted a separate identity compared to all stages of normal myeloid cells. We also verified mutation-specific clustering behaviour of our samples by comparing them with a recently published AML histone H3K27 acetylation data set containing samples with FLT3-ITD, RUNX1 and CEBPA double mutations. Mutation-specific cis-regulatory elements from this study[26] largely overlapped with the mutation-specific DHSs identified here (Fig. S3D). We then defined mutation specific groups of deregulated DHSs that were shared between the specific members of each of the seven major mutation groups defined in Table S1 (Figs. S4A and S4B) which were distributed between both the mutation-specific clusters and the shared clusters (Fig 2B) and were associated with differentially expressed genes (Dataset S2). As seen above in the clustering analysis, the t(8;21), inv(16), and *CEBPA* groups showed similar patterns whereby 914 upregulated DHSs were shared between the three groups (Fig. S4A). The FLT3-ITD, FLT3- ITD/NPM1 and NPM1 mutation groups also showed substantial overlap with 942 shared DHSs, and with only 19% of these DHSs also included in the 914 ITD/NPM1-specific DHSs. These AML-specific patterns showed little

similarity to normal myeloid differentiation as the majority of these specific sites were not up-regulated in GMPs relative to PBSCs (Supplementary note SN1B).

The presence of specific DHSs strongly correlated with the up-regulation of their nearest genes (Dataset S3), indicating that AML type-specific cis-regulatory elements drive the expression of AML type-specific genes (Fig S4C) as exemplified by a DHS at *POU4F1* (Fig S4D*)*. Fig S4E shows examples of AML-specific up-regulated genes of regulatory relevance, including those encoding growth factor receptors or TFs which were associated with the presence of AML type-specific DHSs, in this case *NFIX, POU4F1, MEIS1* and *FOXC1* (Dataset S4). The gene expression patterns of such genes were validated using publicly available data-sets (Fig S5).

Mutation-specific cis-regulatory elements display specific transcription factor occupancy patterns

To identify TF motifs responsible for the establishment of the different patterns, we performed digital footprinting analysis from high-read depth DNaseI-Seq data using our Wellington algorithm[29]. We began by creating a curated list of motifs based on several TF databases (Table S2) since closely related factors typically recognise identical sequences. We therefore selected single representative motifs for each of the different transcription factor families, so as to remove redundant motifs bound by multiple factors. Examples of footprints are depicted for NFI and ETS motifs at the *MDFI* locus in FLT3-ITD/NPM1-mutated AML (Fig. 4A) and for RUNX, NFAT and C/EBP motifs at the *C3AE1* locus in t(8;21) and CEBPA-mutated AML (Fig. S6A). The majority of AML type-specific DHSs within the 20 AML-specific DHS clusters contained footprints, as exemplified by a DHS at C3AE1 (Fig.S6B). For validation, we compared RUNX motif footprints with publicly available RUNX1 ChIP data from our studies (FLT3-ITD/NPM1[24], t(8;21)[30]) and others (inv(16)[31]) (Fig.S6C). Between 60% and 85% of footprinted RUNX motifs occurred in regions shown to bind

RUNX1. Additional motif enrichment analyses of up-and down regulated DHSs are shown in Supplemental Notes, Fig SN1 - 3.

We next evaluated occupied motifs for enrichment in any of the AML type-specific DHSs defined by the 20 DHS clusters (Fig. 4B). This analysis showed that motif occupancy patterns were highly AML type-specific. For example, the FLT3-ITD/NPM1-specific clusters 5 and 19 are enriched for occupied HOX, PBX, FOX/E-box and NFI motifs. Occupancy correlated with up-regulation of multiple homeo-domain genes together with *FOXC1* and *NFIX* (Fig S2C). We have previously shown that AP-1 is a crucial mediator of FLT3-ITD signaling[24]. Occupied AP-1 motifs are enriched in multiple clusters (01, 05, 07, 12, 13, 18, 19), many of which come from AML with signaling mutations. The same is true for NF-κB motifs which are enriched in clusters 01, 03, 06, and 08 which are shared amongst different AML groups. NF-κB and AP-1 factors are mediators of MAP kinase signaling which points to a wide-spread and specific activation of this signaling pathway not just in FLT3-ITD AML[24], but in other AML types. Composite ETS/E-box motifs are occupied in clusters 02, 17 and 20 (associated with the CBF/CEBPA groups), and also in clusters 03, 11 and 14. Finally, we observed significant occupancy of the motif for POU4F1 in clusters 02 and 20 containing samples from patients carrying the t(8;21) and CEBPA double mutations, but nowhere else (Fig4B). *POU4F1* has been shown to be aberrantly expressed in t(8;21) cells[32], but has so far not been linked to *CEBPA* double mutations. A similar differential occupancy picture was seen when footprints were clustered according to mutation-specific groups of DHSs (Fig. S6D). Inspection of motif occupancy of C/EBP motifs in AMLs where both alleles of *CEBPA* are mutated showed little, if any, reduction in overall motif occupancy, indicating a compensatory action of other C/EBP family members (Fig S6D) which are expressed in these cells (Dataset S4). This analysis connects the AML type-specific occupancy of the motifs to the AML subtype-specific expression of defined sets of TFs and demonstrates that

their expression is of functional relevance for the programming of chromatin at their target genes.

To examine the position of transcription factor occupancy patterns within the hematopoietic hierarchy, we correlated the presence of footprints specific for the different AML-subtypes with accessible chromatin regions present in precursor cells (Fig 4C) [25]. This analysis revealed unique factor occupancy patterns of AML cells compared to normal progenitor cell types. For example, HOX motifs within open chromatin regions observed in HSCs, MPPs and MEPs are occupied in the FLT3-ITD/NPM1 and RUNX1 groups, but not in the t(8;21) group, confirming the early block in differentiation (Fig 2B). Many of the samples, including NPM1, FLT3-ITD/NPM1 and t(8;21) cells, also displayed high AP-1 motif occupancy which is normally only seen in monocytes. POU4F1 is expressed in HSCs, MPPs, MEPs and in CLPs[25] and its binding motifs are occupied in t(8;21) and CEBPA double mutant cells, yet these AML cells also show strong occupancy of C/EBP motifs, which is normally a hallmark of GMPs and Monocytes.

In summary, our high-resolution digital footprinting analysis shows (i) that each AML subtype employs a different combination of factors binding to elements shared with different types of precursor cells and that (ii) lineage unrelated expressed TFs such as FOXC1, NFIX and POU4F1 participate in such cooperation.

AML type-specific cis-regulatory elements show differential intra-chromosomal interactions mediated by shared and specifically expressed transcription factors

The construction of gene regulatory networks requires the linking of cis-regulatory elements to their respective promoter[33]. We therefore examined whether the differential activity of cis-regulatory elements in different AML sub-types resulted in the formation of alternate cis-element interactions, and which TF families mediated such interactions. To this end we analysed cells from relapse patient sample t(8;21)-1R (Table S1), which maintained a gene

regulation network similar to the presentation sample t(8;21)-1 (Figs. 1C, S1C and S7A), and a patient carrying a FLT3-ITD/NPM1 mutation (ITD/NPM1-2, Table S1) using promoter-capture chromosomal structure analysis (CHi-C) and compared these data to a dataset derived from human CD34+ cells[34]. Most interactions occurred intra-chromosomally and did not differ at the global level (Fig. S7B). Fig 5A shows interactions across a segment of chromosome 2, projected on the DHS pattern, demonstrating the organization of this this region into topologically associated domains (TADs) that are separated by regions devoid of DHSs. This higher-level structure was unaffected by the type of AML.

Intra-chromosomal interactions driving gene expression are mediated by transcription factor complexes binding to cis-regulatory elements which exist as DHS. The proportion of DHSs involved in AML type-specific interactions varied between the DHS-clusters (Fig 3A, Fig S7C). Moreover, ~40% of all promoters showing differential interactions were associated with expressed genes (Fig S7D,E). Fig S7F shows a direct comparison between the data from the two patients and demonstrates that differential interactions (i) correlated with a differential DHS pattern and that (ii) this difference led to the expression of a differential set of genes with different GO terms (Dataset S5). Fig S7G shows an example of differential interactions within the *KLF2* gene which is differentially expressed between FLT3-ITD and t(8;21) and CD34+ cells.

On average 80% of all DHSs mapped in the t(8;21) AML, the FLT3-ITD AML and the CD34+ cells participated in interactions (Fig S7H). An an average of 17% of interactions were specific for each AML-type and not found in CD34+ PBSCs (Fig S7I), whereby half of these were unique to the type of AML (Fig S7J). To identify the TF families involved in regulating differential interactions we determined the proportions of enriched occupied motifs in the DHSs underlying interactions (Fig S7K). These analyses revealed (i) that hematopoietic TFs such as RUNX, ETS and C/EBP family members participated in differential interactions in both AML types, together with the ubiquitously expressed inducible

AP-1 factor family and the normally invariantly binding CTCF factor and (ii) that AML subtype-specifically expressed TFs participated in such interactions as well. In the FLT3-ITD AMLs this included HOX proteins and factors occupying FOX/E-box motifs. In the t(8;21) AMLs this included NF-κB and proteins binding to FOXO motifs as well as POU4F1, with some motifs being differentially occupied in the same DHS cluster.

Differential interactions drive AML subtype-specific expression of signaling genes but the majority of interactions are shared

In order to integrate differential interactions between promoters and DHSs, digital footprinting data and gene expression data, we assigned the respective DHSs to the promoter they interact with for the two patient classes as described in Fig 5B. We next used these interactions to link DHSs to their respective promoters for AML type-specifically expressed genes across all FLT3-ITD/NPM1 and t(8;21) patients as compared to CD34+ PBSCs. This analysis revealed that the vast majority of DHS underlying interactions between the three data-sets and those of individual patients were shared with an average level of more than 80% overlap (Fig S8A) confirming earlier observations that the global transcriptional network of related cells is also highly related[35,36]. Sub-type-specific DHSs participating in interactions clustered within their patient group, and related groups, but not with unrelated groups (Fig S8B), confirming that the two patients were representative for those groups. For both the FLT3-ITD/NPM1 and the t(8;21) sample the nearest promoter accounted for 65-74% of AML type-specific interactions driving the expression of genes that are up-regulated compared to CD34+ cells (Figure 5C). Similar results were seen for each of the 20 DHS clusters (Fig. S8C).

GO-term and KEGG-pathway analysis of expressed genes in the two types of AML (Fig 5D-G) revealed an AML type-specific core signature of genes being driven by specific cis-regulatory elements (for an extended gene list see Dataset S5). For both AML samples

285 these included genes involved in regulating pro-inflammatory pathways such as cytokine
286 receptor signaling and NF-κB signaling. FLT3-ITD cells also displayed an activated MAP
287 Kinase signaling signature whereas the t(8;21) signature also included RAP, RAS, PI3K and
288 FOXO signaling genes. Significantly, FOXO1 is already known to be part of the t(8;21) pre-
289 leukemic maintenance program[37]. Importantly, more that 50% of all genes within these
290 pathways were targets of RUNX1-ETO (Figure 5H)[30] linking them to the actual driver
291 mutation. A similar percentage of the genes within the FLT3-ITD/NPM1 core pathway are
292 bound by RUNX1 (Fig 5H) which is up-regulated in FLT3-ITD[24] (Dataset S5). This included a
293 number of growth factor receptor genes such as the normally T-cell specifically expressed
294 IL-2 receptor alpha chain which is specifically up-regulated in FLT3-ITD/NPM1 patients.
295 We noticed that ~83% of the DHSs which were involved in significant interactions in
296 each of the 3 samples (Fig S8D). We therefore merged all three CHi-C data-sets (Fig S8D)
297 to use this data to assign the DHS from the 20 clusters (Fig 3A) to their respective
298 promoters. The remaining 17% of DHS were assigned to the nearest promoter. Genes
299 associated with the DHS with confirmed interactions are listed in Dataset S6. GO-term and
300 KEGG-pathway analysis of such genes again showed activation of genes connected with
301 signaling processes such as an inflammatory response, regulation of MAPK activity and
302 cytokine regulation in all types of AML.
303
304 **Different types of AML are maintained by different transcription factor networks**
305 Constitutive and inducible transcription factors form regulatory circuitries and networks by
306 interacting with their own/or other regulatory genes[35]. Cancer cells are capable of
307 maintaining a stable regulatory network over extended periods of time, implying that the
308 expression of each member of such a network is tightly controlled and remains in balance.
309 Consequently, perturbation of the network components maintaining this balance may
310 destabilize leukemic cells thus offering novel therapeutic options. We therefore combined

311 footprinting, TF gene expression and where possible, CHi-C data to construct transcription
312 factor networks in normal CD34+ cells and the different AML subtypes by linking occupied
313 binding motifs on TF genes to specific TF families. The full network structure for each cell
314 type without filtering can be studied in detail via the weblink (http://bioinformatics-
315 bham.co.uk/tfinaml/). Comparison between the different AML subtypes and normal CD34+
316 cells identified interactions between TF sets that were either shared between AMLs and
317 CD34+ cells (Fig. S9) or were specific for each subtype (Fig. 6). These analyses suggested
318 that the AP-1 family network, which is known to integrate multiple MAPK signalling
319 pathways, is of central relevance for leukemic maintenance in all AML subtypes (Fig 6 B-G).
320 Interestingly, in each case these networks reveal tight links between AP-1 and KLF family
321 members that form another node of general relevance in each AML. POU4F1 and HLH
322 family factors that recognising MYC/MAX type E-boxes formed prominent nodes in t(8;21)
323 AML only, while HOX proteins, FOXC1, NFIX and the MAF family were exclusively
324 highlighted in FLT3-ITD and NPM1mut-associated AML. Specific nodes and edges were
325 also part of the normal precursor program (Fig S9). For example the link between the C/EBP
326 family and *NFIL3* was shared between the FLT3-ITD/NPM1 cells (Fig 6F) and CD34+
327 PBSCs (Fig S9F). A detailed discussion of the different network structures and the role of
328 different TF families with more examples can be found in Supplemental Note 5.
329
330 **Network analysis identifies transcription factors contributing to AML propagation**
331 We next used our network analyses to guide experiments validating the important role of
332 TFs forming network nodes that were either widely employed in AML, or which were AML
333 type-specific. To this end, we transduced three different AML cell lines and primary FLT3-
334 ITD AML cells with lentiviral vectors coding shRNAs targeting *POU4F1* (specific for
335 t(8;21)),or targeting *NFIX* or *FOXC1* (specific for FLT3-ITD), as well as control shRNAs.
336 NFIX is known to play a role in myeloid lineage specification[38] but has not been linked to

specific mutation types. *FOXC1* is an oncogene in its own right[39] and overexpression is observed in AMLs with FLT3-ITD mutations[24]. However, *NFIX* and *FOXC1* have not yet been directly linked to the maintenance of the FLT3-ITD AML-phenotype. We applied two distinct shRNA constructs per TF gene with all of them significantly reducing the corresponding TF transcript and protein levels in FLT3-ITD and t(8;21) cell lines (Fig S10A-F). Knockdown of *POU4F1* (Figs. S10 A and D) significantly inhibited the proliferation of t(8;21)-positive Kasumi-1 cells (Figs. 7A and S10G) in agreement with our previous findings[32]. Similarly, expression of NFIX shRNAs efficiently suppressed NFIX expression (Figs. S10B and E) and significantly impaired the proliferation of FLT3-ITD-positive MV4-11, but not FLT3-ITD-negative Kasumi-1 cells (Figs. 7B,C, S10H,I). We next tested the effect of transduction of shRNA constructs targeting these genes on the colony forming ability of patient $CD34^+$ cells carrying the FLT3-ITD/NPM1 mutations as well as on sorted $CD34^+$ PBSCs. Importantly, both NFIX and FOXC1 shRNA constructs reduced the colony forming ability of patient AML cells carrying the FLT3-ITD/NPM1 mutations, but not that of normal CD34+ HSP cells (Figs. 7D, E).

In addition to subtype-specific TFs such as POU4F1 or NFIX, our network analysis suggested that the AP-1 TF family is of general significance for all AML subtypes examined. AP-1 is a heterodimer formed by members of the FOS, ATF, JUN and JDP families of transcription factors and, consequently, challenging to target by defined RNAi approaches. In order to interfere with the binding of all AP-1 family members, we introduced an inducible version of a dominant negative FOS (dnFOS) protein[40,41]. Doxycyclin-mediated induction of dnFOS significantly inhibited proliferation of both t(8;21)-positive Kasumi-1 cells and FLT3-ITD expressing MV4-11 cell lines as compared to non-induced controls (Figs. 7F, G, S10J, K). Moreover, transduction of primary $CD34^+$ FLT3-ITD cells with a lentivirus encoding a constitutively expressed dnFOS reduced the colony forming ability of MV4-11 FLT3-ITD cells but not of $CD34^+$ HPSCs (Figs. 7H, I, S10L). Finally, we examined the significance of AP-1

for leukaemia propagation *in vivo*. To that end, we transplanted either Kasumi-1 or MV4-11 cells expressing a doxycycline-inducible dnFOS into immunodeficient RG mice followed by randomization into a doxycycline and untreated arm. In the case of Kasumi-1 transplantation, 6 out of 7 animals of the control group, but only 2 animals of doxycyclin-treated group developed granulosarcomas (Fig. 7J). Importantly, neither of the latter two tumours expressed dnFOS after DOX treatment (data not shown), further suggesting that induction of dnFOS was incompatible with tumour formation. Similarly, doxycycline treatment of mice transplanted with FLT3-ITD MV4-11 cells that harbored the *dnFOS* transgene inhibited the development of leukemia while all untreated mice rapidly developed tumours and had to be sacrificed (Fig. 7K). Taken together, these findings demonstrate the significance of AP-1 for several AML subtypes and emphasize the potential of transcriptional network analyses to predict TFs crucial for malignant propagation.

Discussion

In this study we define how aberrantly expressed TFs and signaling molecules shape the epigenetic landscape of different sub-types of primary AML. We show (i) that it is possible to use high-quality DNaseI footprinting analysis of purified AML blast cells to identify AML subtype specific TF networks, (ii) that such TF networks allow us to infer a dependency on specific factors for leukemic growth and (iii) that the global activation of signaling pathways in multiple types of AML parallels a growth dependency on AP-1 activity. This comprehensive integrative comparison of gene expression patterns, chromatin accessibility and TF occupancy of primary AML reveals a strong connection between leukemic classifier mutations and networks of TFs and signaling components. Moreover, mapping of cis-element promoter interactions by CHiC enabled assigning the majority of genes of all analysed subtypes to their correct promoter. It has long been known that different types of AML can be characterised by their gene expression and methylation patterns[42,43] suggesting

the existence of specific gene regulatory networks. However, our work now defines these networks in detail, and convincingly proves that leukemic drivers determine the regulatory phenotype by establishing and maintaining gene regulatory and signaling networks distinct from normal cells. Networks consist of shared and specific components and even involve regulators normally not expressed in myeloid cells, such as such as FOXC1 or POU4F1. Our validation experiments show that induced and aberrantly expressed TFs are not just bystanders, but are important for network maintenance and leukemic growth, thus harbouring novel therapeutic opportunities for targeted treatment.

A clinically relevant novel finding from our study is that *CEBPA* double mutant AML is epigenetically highly related to t(8;21) AML. The t(8;21) is driven by a single aberrant TF (RUNX1-ETO) which is sufficient to establish a pre-leukemic state and whose mechanism of action has been under scrutiny for many years. The most likely reason for this epigenetic congruence between t(8;21) and *CEBPA* double mutant AML is that both mutations target a common key control point of myeloid differentiation. RUNX1-ETO represses the *CEBPA* gene while C/EBPα is required for the differentiation response of t(8;21) cells to RUNX1-ETO knock-down[22]. Consequently, the two types of AML share a number of pathways, as exemplified by the expression of *POU4F1*, which could be translated into common therapeutic strategies.

The full set of target genes of RUNX1-ETO in t(8;21) is known and the t(8;21) specific epigenome and TF binding pattern has been extensively characterized[44]. A number of target genes relevant for the maintenance or establishment of the leukemogenic state have already been identified, including *FOXO1*, *UBASH3B, POU4F1,* and *LAT2* together with the members of the RUNX1-ETO complex [22,32,37,45-47]. Our current comparative study has validated these targets, highlighting the power of our methodology and has identified multiple new network components. However, for the other types of AML, in particular for the FLT3-ITD there had been insufficient knowledge of which genes and TFs are primarily

415 responsible for directing the AML type-specific gene regulatory networks. Here, we identified
416 a number of signaling and transcriptional components distinguishing FLT3-ITD from normal
417 blasts and from other types of AML comprising a rich resource for combination therapy
418 approaches. We examined the contribution to leukemic growth for two genes with AML type-
419 specific activity (*NFIX* and *FOXC1*) and showed that in every case their elimination resulted
420 in a growth reduction in AML but not normal cells, yet again confirming that each type of
421 AML stabilises a specific transcriptional network required for survival.

422 The AP-1 factor family has been known to play an important role in many types of
423 tumours [48] and our study shows that it is also of major importance for different types of AML.
424 FLT3-ITD MV4-11 cells have abundant levels of nuclear AP-1, and FLT3-ITD target genes
425 such as *CCNA1* are suppressed by MAP kinase inhibitors in these cells[24]. We have recently
426 shown that JUN scores highly in a siRNA dropout screen examining the requirements for
427 tumour development in t(8;21) AML (Martinez-Soria et al., in press). Moreover, FOS plays
428 an important role in the resistance against BCR-ABL inhibition in CML by activating
429 compensatory signaling pathways[49]. Since several growth factor and stress signal cascades
430 feed into AP-1, a targeted inhibition of all AP-1 binding may be less likely to lead to
431 resistance by rewiring of signalling pathways.

432 The classical picture of two-step leukemogenesis states that in AML a mutation
433 altering a differentiation trajectory cooperates with signaling mutations directing leukemic
434 growth[10,11]. Mutations in TFs which program chromatin directly, and epigenetic regulators
435 such as DNMT3A and TET2 which set up a specific global epigenetic landscape upon which
436 TFs act, fall into the first category while FLT3-ITD falls into the second. However, these
437 distinctions are now becoming blurred as from the viewpoint of the regulation of gene
438 expression, growth factor receptors elicit a strong influence on transcriptional activity via the
439 action of inducible TFs. Moreover, they play a dominant role in driving the differentiation
440 trajectory as their binding patterns, as exemplified by AP-1 family members, show an AML

441 sub-type specific occupancy signature that is uninfluenced by the presence or absence of
442 epigenetic regulator mutations (in this case DNMT3A)[24]. This is not to say that mutations in
443 such genes do not influence the developmental trajectory of AML and clinical outcomes, as
444 shown in CBF AML[50] since AML cells with such mutations acquire an altered DNA
445 methylation landscape that is likely to influence TF binding[51]. However, our data show that
446 the leukemic phenotype and self-renewal in different types of AML defined by differentially
447 activating a multitude of different and often lineage-unrelated signaling pathways and by
448 expressing lineage-unrelated TFs. From the viewpoint of finding therapeutic targets,
449 identifying such mutation-specific pathways will offer to eliminate their specific maintenance
450 program by targeting multiple pathways simultaneously. Our study provides a first step
451 towards this goal.

Methods

Patient samples and PBSC cell processing

Human tissue was obtained with the required ethical approval from the NHS National Research Ethics Committee. AML and PBSC samples used in this study were either surplus diagnostic samples, or were fresh samples obtained with specific consent from the patients. AML samples were obtained from either (i) the Haematological Malignancy Diagnostic Service (St James's Hospital, Leeds, UK, (ii) the Centre for Clinical Haematology, Queen Elizabeth Hospital Birmingham, Birmingham, UK, (iii) the West Midlands Regional Genetics Laboratory, Birmingham Women's NHS Foundation Trust, Birmingham, UK, or from iv) Erasmus University Medical Center, Rotterdam, The Netherlands. Mononuclear cells were purified on the same day that they were received, and in most cases also directly further purified using either CD34 or CD117 (KIT) magnetic antibodies, as previously described [24]. For some samples with greater than 92% blast cells the column purification was not performed. Mobilized PBSCs were provided by NHS BT, Leeds, and NHS BT, Birmingham.

Mutation detection

Mutated genes identified in each patient are summarized in Supplementary Table 1, together with the age, gender and white blood cell count for each patient. Mutations were identified by one of two different methods. The first batch of patients were assayed by targeted exon sequencing of 55 cancer-associated genes using 1212 pairs of previously defined PCR primers [24] for amplification using a RainDance Technologies platform. The mutation sequence data from this screen was analyzed using algorithms to detect either (i) nucleotide variants using the Genome Analysis Toolkit (GATK) [52] or insertions and deletions using Pindel [52]. Mutations were also screened against the COSMIC database of previously observed mutations (http://cancer.sanger.ac.uk/cosmic/). Subsequent samples were assayed using the Illumina Trusight myeloid panel of primers and processed by approaches

similar to those used for the first batch. All identified mutations are listed in Table S1. Some of these patients were also included in a previous publication from our laboratory, using different patients identification codes [24] to those used in the current study.

Cell lines

Cell lines were cultured in an incubator at 37^0C in GIBCO™ 1640 RPMI + Glutamax™ medium supplemented with 10% heat inactivated fetal calf serum (GIBCO), 100 U/ml Penicillin, 100 mg/ml Streptomycin.

Growth curve measurements

250000 MV4-11 or Kasumi-1 cells were cultured in RPMI supplemented with 10 % fetal calf serum, 2mM L-Glutamine, 100 U/ml penicillin and 100 mg/ml streptomycin. Cells were counted with Trypan Blue exclusion and split every 3 days to maintain them in the log phase of growth. For the inducible dnFOS, cells were counted and split every 2 days and 1.5 μg/ml of doxycycline was added.

Co-culture of Primary Cells with MS-5 feeders

Primary cells were maintained in co-culture with MS-5 cells [53] Briefly, cells were cultured in LTC medium (α-minimum essential medium (Lonza) supplemented with heat-inactivated 12.5% fetal calf serum (Gibco), heat-inactivated 12.5% horse serum (Gibco), penicillin and streptomycin, 200 mM glutamine, 57.2 μM β-mercaptoethanol (Sigma) and 1 μM hydrocortisone; (Sigma) supplemented with 20 ng/ml IL-3, granulocyte colony-stimulating factor (G-CSF) and thrombopoietin (TPO) in flasks pre-coated with MS-5 cells.

Lentiviral transduction and shRNA treatment

LEGO-iG-shRNA were generated by cloning shRNAs with the target sequences described below into the LEGO-iG vector[54]. LEGO-iG-dnFOS was generated by cloning the dnFOS insert, originally generated by Charles Vinson (National Cancer Institute, Bethesda, USA [40] into the LEGO-iG backbone. Inducible dnFOS was cloned into a pENTR backbone and then using Gateway Cloning to insert that into the Tet-on plasmid pCW57.1 (David Root,

Addgene plasmid #41393). Backbone vectors LEGO-iG and Inducible dnFOS then used to generate lentiviral particles using packaging and envelope genes on four separate plasmids: TAT, REV, GAG/POL and VSV-G[55].

shRNA Target sequences: shFOXC1_B GTCACAGAGGATCGGCTTGAA; shFOXC1_C GCCGCACCATAGCCAGGGCTT; shNFIX_B: GGAATCCGGACAATCAGAT; shNFIX_C GCAGTCTCAGTCCTGGTTCCT; shPOU4F1_C GCCGAGAAACTGGACCTCAAA; shPOU4F1: GCCGATTAACAAGACTGAAAT; shMM GCGCGATAGCGCTAATAATTT

For virus production, 293T Human Embryonic Kidney cells were cultured in Dulbecco's Modified Eagle Medium supplemented with 10 % fetal calf serum, 2mM L-Glutamine, 100 U/ml penicillin, 100 mg/ml streptomycin and 0.11 mg/ml Sodium pyruvate; and were seeded to achieve 70-80% confluency at time of transfection. HEK293T cells were transfected using the calcium phosphate co-precipitation of the five-plasmids (LEGO-iG with TAT, REV, GAG/POL and VSV-G) at a mass ratio of 24 µg : 1.2 µg : 1.2 µg : 1.2 µg : 2.4 µg per 150 mm diameter plate of cells. Viral supernatant was harvested after 24 h and subsequently every 12 h for 36 h prior to concentration with Centricon Plus 70 100 kDa filter (Millipore, USA), using the manufacturer's instructions. Concentrated viral particles were stored at 4 °C prior to lentiviral transduction. Cell lines were transduced with concentrated virus in the presence of 8 µg/ml polybrene by spinoculation at 1500 xG for 50 min. After 12 – 16 h incubation at 37 °C viral media was exchanged for fresh media. Cell sorting by FACS was performed to isolate GFP+ cells 3 days after transduction.

Primary cell samples were defrosted 24 h prior to transduction and co-cultured with MS-5 feeder cells in LTC medium. 6 well non-tissue culture treated plates were coated with 24 µg/ml retronectin (Takara Clontech) for 2 h prior to blocking with 2% BSA PBS for 30 min. The blocking buffer was washed off with HBSS (Gibco) containing 2.5% HEPES. 1 ml viral concentrate was applied to the retronectin coated plate by centrifugation at 2000 xG for 45

minutes, after which the concentrated viral supernatant was refreshed and the centrifugation repeated. Primary cells suspended to a concentration of 1 x 106 cells/ml in the remaining viral supernatant; supplemented with 20 ng/ul G-CSF, IL-3, TPO and 8 µg/ml polybrene, were then added to the plate and transduced by spinoculation at 1500 xG for 50 min. After 12 – 16 h incubation at 37 °C viral media was exchanged for fresh media. Cell sorting by FACS was performed to isolate GFP+ cells 3 days after transduction.

Colony Formation Assays of Primary Cells

Colony formation assays were performed on sorted cells by seeding at 2500 cells/ml in Methocult Express (Stem Cell Technologies). After 14 days colonies were counted.

Animal experiments

Immunodeficient $Rag2^{-/-}Il2r\gamma^{-/-}$129×Balb/c (RG) mice were housed in the Comparative Biology Centre (Newcastle University) under specific pathogen free conditions. All animal work was conducted in accordance with Home Office Project License PPL60/4552 by researchers who had completed approved Home Office training and held current Personal Licenses under the Animals (Scientific Procedures) Act 1986. Kasumi-1 pCW57.1-dnFOS cells were intrahepatically injected into 14 newborn (2 days old) RG mice at a cell dose of $2.5x10^5$ cells/mouse as described previously (Martinez Soria et al., 2009). Twelve days later, mice were randomized into two treatment groups, one given doxycycline 50 mg/kg three times per week intraperitoneally in an unblended fashion till the experimental endpoint. MV4-11 pCW57.1-dnFOS cells were intrafemorally injected into RG mice at a cell dose of $5x10^5$ cells/mouse followed by randomization into two groups. For the dox group doxycycline was added at a concentration of 2 mg/ml for the initial 3 days and at 0.2 mg/ml subsequently to drinking water containing 2% sucrose. Controls were given water containing 2% sucrose. Animals were humanely killed upon clinical signs of illness or at defined experimental endpoints.

RT- qPCR

RNA was extracted using the Machery-Nagel Nucleospin kit. 1µg RNA was used to make cDNA with 0.5µg OligoDT primer, Murine Moloney Reverse Transcriptase and RNase Inhibitor (Promega, USA) according to manufacturer's protocol. RT-PCR was performed using Sybr Green mix (Applied Biosystems, UK), at 2x dilution. Primers were used at 100nM concentration. A 7900HT system (Applied Biosystems, UK) was used to perform qPCR. Analyses were performed in technical duplicates using a standard curve derived from the untreated cell line.

Western Blotting

Protein lysates from cell lines were analysed by Western blot. Relevant primary antibodies against FOXC1 (Cell Signaling Technology - #8758) , NFIX (Invitrogen - #PA5-31234), POU4F1 (Santa Cruz Biotechnology – sc-8426) were used to detect target genes and GAPDH (mouse αGAPDH – Abcam – ab8245; rabbit αGAPDH – Cell Signaling Technology – 2118L) was used as a housekeeping gene. Secondary antibodies mouse anti-rabbit HRP (Rockland – 18-8816-31) and goat anti-mouse HRP (Jackson ImmunoResearch – 115-035-062) enabled detection and quantifications by densitometry using Imagelab software and a GelDoc imager.

RNA-Seq library preparation

RNA was extracted and analyzed from purified AML cells as previously described [23]

DNaseI-Seq

DNaseI digestions of permeabilized cells were performed as previously described[56]. Briefly, live cells were added directly to a solution of DNaseI (DPFF, Worthington) in dilute Nonidet P40, digested for 3 min at 22^{o}C, and the reactions then terminated by addition of SDS to 0.5%. DNaseI was typically used in the range of 2-6 µg/ml using a final 1.5 x 10^{7} cells/ml. DNaseI-Seq libraries were then prepared and validated essentially as previously described [30]. Libraries were run on Illumina sequencers.

Promoter capture HiC (CHi-C) from patient AML blasts

582 AML cells from patient peripheral blood were first purified by density gradient centrifugation
583 (Lymphoprep™) and then using CD34 antibody coupled beads. 5× 10^7 t(8;21) blasts (patient
584 t(8;21)-1R) and FLT3-ITD/NPM1 blasts (patient ITD/NPM1-2) were fixed in 37 ml of RPMI-
585 1640 supplemented with 15% FBS and 2% formaldehyde for 10 minutes at room
586 temperature. 6 ml of 1M glycine (0.125 M final concentration) was added to quench the
587 reaction and cells were incubated at room temperature for 5 min, followed by 15 minutes on
588 ice before pelleting the cells at 4 °C and washing them in ice cold PBS. Each sample was
589 flash frozen in liquid nitrogen, and stored at −80 °C. Cells were lysed in a tight dounce
590 homogeniser (ten cycles) with 3ml of cold lysis buffer (10 mM Tris-HCl pH 8, 10 mM NaCl,
591 0.2% Igepal CA-630, one tablet protease inhibitor cocktail (Roche complete, EDTA-free,
592 11873580001)). Cells were left on ice for five minutes then homogenised another ten times.
593 The lysed cells, in 3 ml lysis buffer, were added to 47ml of lysis buffer and incubated on ice
594 for 30 minutes with occasional mixing. Chromatin was pelleted and resuspended in 1ml of
595 1.25x NEBuffer 2 and split into four. Each sample was then pelleted at 1000 rpm and
596 resuspended in 358 µl of 1.25x NEBuffer 2. 11 µl 10% SDS was added and each tube was
597 incubated at 37°C for 60 minutes, rotating at 950 rpm. Samples were mixed by pipetting up
598 and down every 15 minutes. SDS was quenched with 75µl 10% Triton X-100 and incubated
599 at 37˚C for 60 minutes. HindIII digestion, biotinylation, ligation, crosslink reversal, promoter
600 capture and library preparation was performed exactly as described in[34].
601
602 **Bioinformatics analyses**
603 **DNaseI-Seq data analysis**
604 ***Alignment:*** DNaseI-seq sequences from all experiments were mapped onto the reference
605 human genome version hg38, with Bowtie version 2.3.1[57] using default parameters. Low
606 quality reads were trimmed prior to the alignment and the quality control (QC) statistics for

607 the samples were obtained using FastQC tools. Reads that were aligned to unique
608 chromosomal positions were retained.
609 ***Peak calling***. DNaseI Hypersensitive Sites (DHSs) were called with MACS2 using callpeak
610 function (nomodel, call-summits and q= 0.005 parameters)[58]. DHSs were allocated to genes
611 and to the gene promoter if it was within 2kb of the gene transcription start site (TSS), and
612 as distal otherwise. Overlaps between DHSs peaks were defined by requiring the summits of
613 two peaks to lie within +/-200 bp.
614 ***DNaseI-Seq peak set definition:*** To define a common set of coordinates covering all of the
615 significant distal DHSs investigated in this study, we merged all of the individual DNaseI-Seq
616 reads for all of the AML samples assayed by DNaseI-Seq. This data set was then used to
617 define the peak summits of 128,864 distal peaks, excluding promoters, which were detected
618 in the merged data. This approach was designed to maximize the precision and sensitivity of
619 the peak detection, allowing us to generate a single set of peak coordinates that (i) included
620 all the regions where peaks might be found, thereby reducing the level of false negatives,
621 and (ii) greatly diminished the number of false positives. The DNA read counts were then
622 determined for 400 bp windows centered on each peak for each AML sample and for the
623 PBSC samples. To account for the different number of reads in each of the samples; the
624 read counts were initially normalized for total read depth using DEseq2 [59]. Because most of
625 our individual DNaseI-Seq data sets encompassed in the range of 25,000 to 40,000
626 significant distal DHSs, we further normalized the values obtained on the basis of the
627 midpoint (12.5 percentile) of the top 25% of peaks (32,216 peaks).
628 ***Mutation-specific DNaseI-Seq peak set definition:*** We determined the average log2
629 values for 7 distinct subsets of AMLs that carried the same specific mutations in key
630 regulators, and which shared similar patterns of DHSs based on the DHS clustering analysis.
631 The samples included in each group are color-coded and listed in order in Table S1 for AML
632 samples with the following mutations: (i) 3 samples with FLT3-ITD but not NPM1 (#1 to 3), (ii)

6 samples with FLT3-ITD and NPM1 (# 1 to 6), (iii) 2 samples with NPM1 but not FLT3-ITD (# 1 and 2), (iv) 4 samples with t(8;21) (1 to 4), (v) 3 samples with inv(16) (# 1 to 3); (vi) 6 samples with RUNX1 or RUNX1 and CEBPA (1 to 6), and (vii) 3 samples with 2 CEBPA mutations (#1 to 3). To define mutation-specific subsets of specific DHSs, we identified peaks where the average log2 value both was at least 64 and at least 3-fold higher than in PBSCs. Downregulated DHSs are defined as being at least 3-fold less than in PBSCs. Samples were not included in these 7 specific groups in cases where, for example, 2 copies of the FLT3-ITD mutation were present, the NPM1 mutation was paired with a NRAS instead of the FLT3-ITD, RUNX1 mutations were paired with a JAK2 mutation, or where only a single CEBPA allele was mutated.

Clustering of DNaseI-Seq data: Clustering of DNaseI-seq samples was carried out using the merged distal DHSs. The number of reads that mapped to these DHSs was counted in a 400bp window centered on the DHS summit, and subsequently normalized to total sample size using DEseq2 [59]. Pearson correlation coefficients were then calculated for each pair of samples using the log2 of the normalized read counts, and then hierarchically clustered using Euclidean distance and complete linkage clustering of the correlation matrix in R.

K-mean clustering of AML specific DHSs: A combined set of up-regulated distal DHSs that defined as being at least 3-fold greater than in PBSCs was used to perform unsupervised k-mean clustering. The number of reads that mapped to these peaks was counted in a 400bp window centered on the DHS summit, and subsequently normalized to total sample size using DEseq2[59]. Clustering was done on rows (DHSs) while samples (columns) were ranked based on the hierarchical clustering in Figure 1C. Initially the read counts output from DEseq2 [59] was further quartile normalised using the “preprocessCore” package in R, The log2 of the normalised reads were clustered using *k*-means clustering with Euclidean distances (*stats* package in R) and the optimal number of clusters was determined to be 20 based on the lowest Bayesian Information Criterion (BIC) scores

659 (Schwarz, 1978). Each of the 20 clusters was then hierarchically clustered using the
660 "complete linkage" agglomeration method.

661 **ATAC sequencing data analysis**

662 ATAC-seq profiles of hematopoietic and leukemic cell types taken [25] were downloaded from
663 GEO with accession number GSE74912. ATAC-seq data of HSC, MPP, CMP, CLP, MEP,
664 GMP and Monocytes were downloaded and aligned to the human genome version hg38.
665 Aligned reads with the same cell line were merged and then ATAC peaks were obtained
666 using MACS2 with default parameter. Overlaps between DHS and ATAC peaks were
667 defined by requiring the summits of two peaks to lie within +/-200 bp. Pair-wise peak
668 overlaps between DHSs and ATAC peaks of hematopoietic *i* and *j* were performed in order
669 to calculate the fraction (M_{ij})

670 $M_{ij} = \frac{N_{ij}}{N_i}$ where N_{ij} is the total peaks that overlap, N_i is the total number peaks in set i
671 (DHSs) and N_j is the total peaks in j (ATAC). A matrix with the calculated fraction multiply by
672 100 was generated and a heatmap was plotted (Figure 2B) after hieratically clustered in R.
673 Clustering of DNaseI-seq and ATAC-seq samples (Figure S2A and Figure S3B) was carried
674 out using the merged distal DHSs as described earlier using the DNaseI-seq only.

675 **ChIP sequencing data analysis**

676 ChIP-Seq sequencing reads were downloaded from GEO with accession numbers
677 (GSM1581788, GSM1693378, GSM1466000)[24] (GSM722705, GSM722704)[30], the reads
678 were aligned to the human genome version hg38 with Bowtie version 2.3.1[57]. Reads that
679 mapped uniquely to the genome were retained and duplicated reads were removed using
680 the MarkDuplicates function in Picard tools (http://broadinstitute.github.io/picard/). Peaks
681 were identified with MACS version 1.4.2 [58] and DFilter software[60] with recommended
682 parameters (-bs=100 -ks=50 –refine). Peaks common to both peak calling methods were
683 considered for further analysis.

H3K27Ac ChIP data analysis

H3K27Ac ChIP data from[26] were downloaded from NCBI with accession number SRP103200. The raw reads were aligned to the human reference genome hg38 and density profiles were generated using *bedtools. The bedGraph* files were used to generate the H3K27Ac average coverage plotted a long side the DHSs of the 20 clusters (Figure S3D).

Digital genomic footprinting

Digital genomic footprinting was performed using the *Wellington_footprints* function of the Wellington algorithm[29] on High-depth AML and CD34+ PBSC DHSs. DHS footprints probability and DNase forward and reverse cut coverages, were generated using the *dnase_wig_tracks* function of Wellington. AML-specific footprints compared to PBSC CD34+ cells were identified using *wellington_bootstrap* function of Wellington. Mutation-specific footprints of the groups were identified by using the *Wellington_footprints* function using the merged reads of the Mutation-specific individual DNaseI-Seq of each group.

Motif identification

De novo motif analysis was performed on peaks using HOMER[61]. Motif lengths of 6, 8, 10, and 12 bp were identified in within ± 200 bp from the peak summit. The annotatePeaks function in HOMER was used to find occurrences of motifs in peaks. In this case we used known motif position weight matrices (PWM).

Motif co-localisation clustering: Motif co-localisation clustering was performed as previously described[22]. A motif position search was done within DHSs that are group mutation-specifically footprinted. The distance between the centres of each motif pairs was calculated and the motif frequency was counted if the first motif was within 50bps distance from the second motif. Z-scores were calculated from the mean and standard deviation of motif frequencies observed in random sets using bootstrap analysis, peak sets with a population equal to that of the footprinted peaks were randomly obtained from the merged footprints of all AML and CD34+ footprints sets. Motif search and motif frequencies

calculations were repeated 1000 times for each random set. A matrix was generated and Z-scores were displayed after hierarchical clustering as a heatmap with R.

Motif enrichment

To identify motifs that are relatively enriched in the distal footprinted DHSs of each of AML mutation groups (Figure S5) and the AML DHSs clusters (Figure 3A). For a given set *j* of footprints, we defined a motif enrichment score (ES_{ij}) for motif *i* in footprint set *j* as $ES_{ij} = \frac{n_{ij}/M_j}{\sum_j n_{ij}/\sum_j M_j}$ where n_{ij} is the number of footprints in each subset *j (j=1,2,…,12)* containing motif *i (i=1, 2,….,I), I* is the total number of motifs used in the test, and M_j the total number of peaks in each subset *j (j=1,2,…,30).* A matrix was generated and the motif enrichment scores were displayed as a heatmap after hierarchical clustering with Euclidean distance and complete linkage. The heatmap was generated using R. The statistical significance for a ES_{ij} score of a given motif *i* in peak set *j* is computed as Z-scores using bootstrapping (N=1000), where a random set of peaks is extracted from a global set of footprinted regions and *ES* is calculated. After N iterations the mean (μ_{ij}) and the standard deviation (σ_{ij}) are computed and the z-scores are computed as $Z_{ij} = \frac{ES_{ij} - \mu_{ij}}{\sigma_{ij}}$. The global set of regions is a merged set of all the AML footprints. These Z-scores are provided in Table S7.

RNA-seq data analysis

RNA-Seq reads were aligned to the human genome hg38 build with STAR[62] using ENCODE recommend parameters. Separate density profiles for the positive and negative strand were generated using bedtools. Cufflinks[63] was used to calculate the expression values as Fragments Per Kilobase per Million aligned reads (FPKM) from the aligned RNA-seq data. Mutation-specific group's gene-wise expression values were obtained using the *cuffdiff* function of cufflinks. The correlation between any two AML samples was obtained as the Pearson correlation coefficient of expression values over all genes. A correlation matrix was

735 thus generated for all the samples and hierarchically clustered to study the relationship
736 among samples as given in Figure S1C. Smooth scatter plots were generated in R.

737 ***Gene expression analysis***

738 Differentially expressed genes were extracted using the limma R package[64]. Genes were
739 said to be differentially expressed (DE) if there was a twofold change in expression between
740 any each of the AML patient sample or each of the mutation-specific group and the PBSC
741 CD34+ with a *p*-value less than or equal to 0.01 and with FPKM greater than 1 in at least
742 one AML sample. For each value of a DE gene a pseudo-count $\gamma = 0.1$ was added to the
743 FPKM values and the binary logarithm of this value was considered as the expression value
744 of the gene in each sample (*j*), $e_{ij} = log_2(FPKM_{ij} + \gamma)$. These DE values were then
745 clustered (Figure S1A) using hierarchical clustering with Euclidean distances (***stats*** package
746 in R). While Hierarchical clustering of transcription factors gene expression was carried out
747 on fold-changes for genes associated with at least a 2-fold change compared to the CD34+.

748 ***Gene set enrichment analysis***

749 A publically available RNA-seq data of hematopoietic cell types were downloaded from GEO
750 with accession number GSE74246. The downloaded RNA-seq data were processed in
751 similar way as described above. The GSEA software[65] was used to perform gene set
752 enrichment analysis on group of genes. Module map[66] implemented by Genomic software
753 was used to find which groups of genes are significantly up- or down-regulated using a
754 statistical test based on the hyper-geometric distribution the fraction of up or down regulated
755 is displayed as a heatmap (Fig 2C and Fig S4C).

756 ***Gene ontology (GO) analysis:*** Gene ontology (GO) analysis was performed using clueGO
757 tools[65] with Hypergeometric for overrepresentation and Benjamini and Hochberg (FDR)
758 correction for multiple testing corrections. KEGG Pathway network analysis was performed
759 using clueGO tools[65] with kappa score = 0.3. A right-sided enrichment (depletion) test based
760 on the hypergeometric distribution was used for terms and groups. The size of the nodes

761 reflects the number of genes within the term. The color of nodes reflects the enrichment
762 significance of the terms. The network is laid out using Cytoscape. The KEGG pathway
763 network figures for all DHS-cluster associated genes are shown in Table S6.

764 **Expression profiles from larger patient cohort datasets**

765 Microarray data from Verhaak et al.[42] were downloaded from GEO under the accession
766 number GSE6891. Patients were split according to their mutational status; Boxplots showing
767 the expression of the indicated genes in FLT3-ITD, NPM1, CEBPA, t(8;21), inv(16) and
768 NRAS mutation groups. The statistical significance of the difference in expression between
769 FLT3-ITD and other mutations was determined using an unpaired t-test.

770 **Promoter Capture HiC data analysis**

771 The CHi-C paired-end sequencing reads from ITD/NMP1-2 and t(8;21)-1R patients and a
772 publically available CD34+ dataset (accession numbers ERR436032 and ERR436025) were
773 put through *HiCUP* pipeline[67]. The raw sequencing reads were separated and mapped
774 against the human genome (hg38). The reads were then filtered for experimental artefacts
775 and duplicate reads, and then re-paired. Statistically significant interactions were called
776 using *GOTHiC* package[68] and HOMER software. This uses a cumulative binomial test to
777 detect interactions between distal genomic loci that have significantly more reads than
778 expected by chance, by using a background model of random interactions. This analysis
779 assigns each interaction with a p-value, which represents its significance. Differential
780 interactions were determined with HOMER[61] for t(8;21) using FLT3-ITD or CD34+ as
781 background and FLT3-ITD using t(8;21) or CD34+ as a background. A difference with a p-
782 value of less than 0.1 was deemed to be significant

783 **Transcription Factor Gene Regulatory Network Construction**

784 We identified a subset of 310 transcription factor (TF) genes that are expressed in one or
785 more of our AML samples. The gene names for transcription factors in human were obtained
786 from AnimalTFDB[69]. The 310 TFs were considered as nodes and the nodes coloured

according to their expression values at each AML subtype (Fig 6, Fig S9 and Fig SN5). Node border colour signifies whether the gene is up-regulated, down-regulated or invariant base on a 2-fold-change compared to CD34+ cells. Node border type indicates whether gene is differentially expressed in one AML subtype as compared to other subtypes. A directed edge from TF_a to TF_b indicates motif binding of a TF_a to the locus of the TF_b and the edge is prominently displayed if TF_a binds to the locus at that stage. The edge is classified and colour coded according to the significant of motif count enrichment.

Motif count enrichment for TFs network: Initially footprints for each AML subtype were identified by using the Wellington algorithm[29] and were annotated to their related promoter using CHi-C data where possible. Motif search within footprint coordinates where performed using HOMER[61]. The number of motifs per TF gene were counted and the significance of motif enrichment was identified using bootstrapping on random sampling, a random set of mapped motif were extracted from all union footprinted motif of all AML subtypes and the CD34+ cells. After 1000 iterations the mean, standard deviation and the z-scores are computed. Motif (TF_a) is linked to gene (TF_b) with only positive Z-score.

Motif count enrichment for up-regulated TFs: The correlations (r_e) between all TF genes based on FPKM values from the RNA-seq analysis were identified and the correlations (r_m) between all TF genes based on motif count binding were identified. The correlation coefficients were z-transformed using Fisher Z-transformation with "*FisherZ*" function in R. The average of the transformed z-scores of both gene expression and motif were transformed to correlation (r). All TF genes with a correlation coefficient equal or greater than a cut-off of 0.3 were considered. Then for each AML cell type, the differentially expressed TF genes among these correlated genes were identified. First with AML subtype as compared to CD34+ cells and second with AML subtype compared to the average expression of other subtypes includes the CD34+. Up-regulated genes with a 2-fold-change in expression were either compared to CD34+ or compared to other AMLs were considered

to construct the network. Motif enrichment for the correlated and up-regulated TF genes and edges were identified as described above.

List of used position weight matrices: A description of how the motifs were curated can be found in the legend of Table S2.

Data availability

Processed data will be available from our webserver

http://bioinformatics-bham.co.uk/tfinaml/

Password sal2018.

Raw data have been deposited at GEO under the accession number GSE108316.

References

1. Cancer Genome Atlas Research, N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**, 2059-74 (2013).
2. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *New England Journal of Medicine* **374**, 2209-2221 (2016).
3. Bonifer, C. & Cockerill, P.N. Chromatin Structure Profiling Identifies Crucial Regulators of Tumor Maintenance. *Trends Cancer* **1**, 157-160 (2015).
4. Rosenbauer, F. & Tenen, D.G. Transcription factors in myeloid development: balancing differentiation with transformation. *Nature reviews. Immunology* **7**, 105-17 (2007).
5. Cockerill, P.N. Receptor Signaling Directs Global Recruitment of Pre-existing Transcription Factors to Inducible Elements. *Yale J Biol Med* **89**, 591-596 (2016).
6. Ward, A.F., Braun, B.S. & Shannon, K.M. Targeting oncogenic Ras signaling in hematologic malignancies. *Blood* **120**, 3397-406 (2012).

7. Parikh, C., Subrahmanyam, R. & Ren, R. Oncogenic NRAS rapidly and efficiently induces CMML- and AML-like diseases in mice. *Blood* **108**, 2349-57 (2006).
8. Masson, K. & Ronnstrand, L. Oncogenic signaling from the hematopoietic growth factor receptors c-Kit and Flt3. *Cell Signal* **21**, 1717-26 (2009).
9. Gerloff, D. *et al.* NF-kappaB/STAT5/miR-155 network targets PU.1 in FLT3-ITD-driven acute myeloid leukemia. *Leukemia* **29**, 535-47 (2015).
10. Corces-Zimmerman, M.R., Hong, W.J., Weissman, I.L., Medeiros, B.C. & Majeti, R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc Natl Acad Sci U S A* **111**, 2548-53 (2014).
11. Shlush, L.I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328-33 (2014).
12. Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. *Nature structural & molecular biology* **20**, 1147-55 (2013).
13. Broske, A.M. *et al.* DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. *Nature genetics* **41**, 1207-15 (2009).
14. Bonifer, C. & Bowen, D.T. Epigenetic mechanisms regulating normal and malignant haematopoiesis: new therapeutic targets for clinical medicine. *Expert reviews in molecular medicine* **12**, e6 (2010).
15. Challen, G.A. *et al.* Dnmt3a is essential for hematopoietic stem cell differentiation. *Nature genetics* **44**, 23-31 (2012).
16. Ko, M. *et al.* Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839-43 (2010).
17. Shih, A.H., Abdel-Wahab, O., Patel, J.P. & Levine, R.L. The role of mutations in epigenetic regulators in myeloid malignancies. *Nature reviews. Cancer* **12**, 599-612 (2012).

18. Goode, D.K. *et al.* Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. *Dev Cell* **36**, 572-87 (2016).
19. Obier, N. & Bonifer, C. Chromatin programming by developmentally regulated transcription factors: lessons from the study of haematopoietic stem cell specification and differentiation. *FEBS Lett* **590**, 4105-4115 (2016).
20. Wilson, N.K. *et al.* Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532-44 (2010).
21. Rubin, A.J. *et al.* Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. *Nat Genet* **49**, 1522-1528 (2017).
22. Ptasinska, A. *et al.* Identification of a Dynamic Core Transcriptional Network in t(8;21) AML that Regulates Differentiation Block and Self-Renewal. *Cell Reports* **8**, 1974-1988 (2014).
23. Loke, J. *et al.* RUNX1-ETO and RUNX1-EVI1 Differentially Reprogram the Chromatin Landscape in t(8;21) and t(3;21) AML. *Cell Rep* **19**, 1654-1668 (2017).
24. Cauchy, P. *et al.* Chronic FLT3-ITD Signaling in Acute Myeloid Leukemia Is Connected to a Specific Chromatin Signature. *Cell Rep* **12**, 821-36 (2015).
25. Corces, M.R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**, 1193-203 (2016).
26. McKeown, M.R. *et al.* Superenhancer Analysis Defines Novel Epigenomic Subtypes of Non-APL AML, Including an RARalpha Dependency Targetable by SY-1425, a Potent and Selective RARalpha Agonist. *Cancer Discov* **7**, 1136-1153 (2017).
27. Martens, J.H. *et al.* ERG and FLI1 binding sites demarcate targets for aberrant epigenetic regulation by AML1-ETO in acute myeloid leukemia. *Blood* **120**, 4038-48 (2012).

888 28. Pulikkan, J.A., Tenen, D.G. & Behre, G. C/EBPalpha deregulation as a paradigm for
889 leukemogenesis. *Leukemia* **31**, 2279-2285 (2017).
890 29. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital
891 genomic footprints from DNase-seq data. *Nucleic Acids Res* **41**, e201 (2013).
892 30. Ptasinska, A. *et al.* Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-
893 wide changes in chromatin structure and transcription factor binding. *Leukemia* **26**,
894 1829-1841 (2012).
895 31. Mandoli, A. *et al.* CBFB-MYH11/RUNX1 together with a compendium of
896 hematopoietic regulators, chromatin modifiers and basal transcription factors
897 occupies self-renewal genes in inv(16) acute myeloid leukemia. *Leukemia* **28**, 770-8
898 (2014).
899 32. Dunne, J. *et al.* AML1/ETO proteins control POU4F1/BRN3A expression and function
900 in t(8;21) acute myeloid leukemia. *Cancer Res* **70**, 3985-95 (2010).
901 33. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome
902 conformation. *Science* **295**, 1306-11 (2002).
903 34. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-
904 resolution capture Hi-C. *Nat Genet* **47**, 598-606 (2015).
905 35. Chasman, D. & Roy, S. Inference of cell type specific regulatory networks on
906 mammalian lineages. *Curr Opin Syst Biol* **2**, 130-139 (2017).
907 36. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory
908 networks. *Cell* **150**, 1274-86 (2012).
909 37. Lin, S. *et al.* A FOXO1-induced oncogenic network defines the AML1-ETO
910 preleukemic program. *Blood* **130**, 1213-1222 (2017).
911 38. O'Connor, C. *et al.* Nfix expression critically modulates early B lymphopoiesis and
912 myelopoiesis. *PLoS One* **10**, e0120102 (2015).

913 39. Somerville, T.D. *et al.* Frequent Derepression of the Mesenchymal Transcription
914 Factor Gene FOXC1 in Acute Myeloid Leukemia. *Cancer Cell* **28**, 329-42 (2015).
915 40. Olive, M. *et al.* A dominant negative to activation protein-1 (AP1) that abolishes DNA
916 binding and inhibits oncogenesis. *J Biol Chem* **272**, 18586-94 (1997).
917 41. Obier, N. *et al.* Cooperative binding of AP-1 and TEAD4 modulates the balance
918 between vascular smooth muscle and hemogenic cell fate. *Development* **143**, 4324-
919 4340 (2016).
920 42. Verhaak, R.G.W. *et al.* Prediction of molecular subtypes in acute myeloid leukemia
921 based on gene expression profiling. *Haematologica* **94**, 131-134 (2009).
922 43. Figueroa, M.E. *et al.* DNA Methylation Signatures Identify Biologically Distinct
923 Subtypes in Acute Myeloid Leukemia. *Cancer cell* **17**, 13-27 (2010).
924 44. Lin, S., Mulloy, J.C. & Goyama, S. RUNX1-ETO Leukemia. *Adv Exp Med Biol* **962**,
925 151-173 (2017).
926 45. Goyama, S. *et al.* UBASH3B/Sts-1-CBL axis regulates myeloid proliferation in human
927 preleukemia induced by AML1-ETO. *Leukemia* **30**, 728-39 (2016).
928 46. Sun, X.J. *et al.* A stable transcription factor complex nucleated by oligomeric AML1-
929 ETO controls leukaemogenesis. *Nature* **500**, 93-7 (2013).
930 47. Essig, A., Duque-Afonso, J., Schwemmers, S., Pahl, H.L. & Lubbert, M. The
931 AML1/ETO target gene LAT2 interferes with differentiation of normal hematopoietic
932 precursor cells. *Leuk Res* **38**, 340-5 (2014).
933 48. Trop-Steinberg, S. & Azar, Y. AP-1 Expression and its Clinical Relevance in Immune
934 Disorders and Cancer. *Am J Med Sci* **353**, 474-483 (2017).
935 49. Kesarwani, M. *et al.* Targeting c-FOS and DUSP1 abrogates intrinsic resistance to
936 tyrosine-kinase inhibitor therapy in BCR-ABL-induced leukemia. *Nat Med* **23**, 472-
937 482 (2017).

938 50. Faber, Z.J. *et al.* The genomic landscape of core-binding factor acute myeloid
939 leukemias. *Nat Genet* **48**, 1551-1556 (2016).
940 51. Levis, M. *et al.* Results from a randomized trial of salvage chemotherapy followed by
941 lestaurtinib for patients with FLT3 mutant AML in first relapse. *Blood* **117**, 3294-301
942 (2011).
943 52. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-
944 generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
945 53. van Gosliga, D. *et al.* Establishing long-term cultures with self-renewing acute
946 myeloid leukemia stem/progenitor cells. *Exp Hematol* **35**, 1538-49 (2007).
947 54. Weber, K., Bartsch, U., Stocking, C. & Fehse, B. A multicolor panel of novel lentiviral
948 "gene ontology" (LeGO) vectors for functional gene analysis. *Mol Ther* **16**, 698-706
949 (2008).
950 55. Mostoslavsky, G. *et al.* Efficiency of transduction of highly purified murine
951 hematopoietic stem cells by lentiviral and oncoretroviral vectors under conditions of
952 minimal in vitro manipulation. *Mol Ther* **11**, 932-40 (2005).
953 56. Bert, A.G., Johnson, B.V., Baxter, E.W. & Cockerill, P.N. A modular enhancer is
954 differentially regulated by GATA and NFAT elements that direct different tissue-
955 specific patterns of nucleosome positioning and inducible chromatin remodeling. *Mol*
956 *Cell Biol* **27**, 2870-85 (2007).
957 57. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat*
958 *Methods* **9**, 357-9 (2012).
959 58. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137
960 (2008).
961 59. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and
962 dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

963 60. Kumar, V. *et al.* Uniform, optimal signal processing of mapped deep-sequencing
964 data. *Nat Biotechnol* **31**, 615-22 (2013).
965 61. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors
966 Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities.
967 *Molecular Cell* **38**, 576-589 (2010).
968 62. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21
969 (2013).
970 63. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with
971 RNA-seq. *Nat Biotechnol* **31**, 46-53 (2013).
972 64. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-
973 sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
974 65. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene
975 ontology and pathway annotation networks. *Bioinformatics* **25**, 1091-3 (2009).
976 66. Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional
977 activity of expression modules in cancer. *Nat Genet* **36**, 1090-8 (2004).
978 67. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res*
979 **4**, 1310 (2015).
980 68. Mifsud, B. *et al.* GOTHiC, a probabilistic model to resolve complex biases and to
981 identify real interactions in Hi-C data. *PLoS One* **12**, e0174744 (2017).
982 69. Zhang, H.M. *et al.* AnimalTFDB: a comprehensive animal transcription factor
983 database. *Nucleic Acids Res* **40**, D144-9 (2012).

Acknowledgements

This research was funded by a programme grant from Bloodwise (15001) to C.B. and P.N.C, as well as studentship awards from Cancer Research UK and Bloodwise to A.Pickin and N.G, respectively, a Kay Kendall Clinical Training Fellowship for J.L. and a MRC/Leuka Clinical Training Fellowship for S.P.

Author Affiliations

[1] Institute of Cancer and Genomic Sciences, University of Birmingham, B17 2TT, UK

[2] Northern Institute for Cancer Research, University of Newcastle, Newcastle, UK

[3]Section of Experimental Haematology, Leeds Institute for Molecular Medicine, University of Leeds, Leeds LS9 7TF, UK

[4] West Midlands Regional Genetics Laboratory, Birmingham Women's NHS Foundation Trust, Birmingham B15 2TG, UK

[5] CMT Laboratory NHS Blood & Transplant, Edgbaston, Birmingham, B15 2SG, UK

[6] Department of Hematology, Erasmus University Medical Center, Dr. Molewaterplein 50, 3015 GE Rotterdam, the Netherlands

[7] Haematological Malignancy Diagnostic Service, St. James's University Hospital, Leeds LS9 7TF, UK

[8] Centre for Clinical Haematology, Queen Elizabeth Hospital, Birmingham B15 2TG, UK, UK

Author contributions:

M.R.I., D.C., S.P., A.P, H.P., A. Pickin, N.G., J.L., P.S.C, R.R., S.R.J. performed experiments and generated data, H.R.D., M.R., S.R., M.G. P.J., A.U. provided patient samples, S.C., A,B, and P.N.C. conducted mutation analysis, S.A.A. and P.C. analysed data, O.H. supervised transplantation experiments and helped editing the manuscript, C.B. and P.N.C. conceived and directed the study and CB wrote the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Corresponding authors

Constanze Bonifer (c.bonifer@bham.ac.uk) and Peter N. Cockerill (p.n.cockerill@bham.ac.uk).

Figure Legends

Figure 1: Different types of AML adopt unique transcriptome and chromatin landscapes. (A) Experimental strategy. (B) UCSC Genome browser tracks of DNaseI-seq mapping in purified AML cells. (C) Hierarchical clustering of Pearson correlation coefficients of DNaseI accessible sequences from all patient samples with normalized read counts of DNaseI-Seq data for the different classes of mutations (left panel), right panel: list of mutations in cells from each patient

Figure 2: Different types of AML are blocked at different stages of differentiation and are regulated by different transcriptional network. (A) Hematopoietic hierarchy; shown are some of the precursor stages from which ATAC-seq and RNA-seq data were generated in Corces et al., 2016: Hematopoietic stem cells (HSC), common myeloid progenitors (CMP), common lymphoid progenitors (CLP), Megakaryocyte Erythrocyte Precursors (MEP) and Granulocyte Macrophage Precursors (GMP). (B) Clustering of the correlation of percentage of peak overlap between DNaseI-Seq and ATAC-seq data by first generating a matrix with all overlap percentages between all DHS peaks, and ATAC-seq peaks and then hierarchically clustering. (C) Gene set enrichment analysis for the differentially expressed genes that are at least 2-fold different compared to the normal CD34+ PBSCs. Up and down regulated gene expression patterns were tested for their similarity to specific pairs of progenitor RNA-seq data from Corces et al. 2016, representing different steps of differentiation. Up-regulated genes are shown in top panel and the bottom panel shows the down-regulated genes.

Figure 3: AML-specifically active cis-regulatory elements cluster into common and unique chromatin landscapes. (A) Heatmap depicting unsupervised K-mean clustering of the DNaseI-Seq log2 signals seen in each AML specific distal DHS peak in each AML sample compared to PBSCs. Clustering was done only on rows (DHS peaks) while samples were ranked based on the clustering in Figure 1C. A diagram on top of the heatmap shows the DHS peak population used for clustering. (B) A binary heatmap shows the overlap between the clusters from A and the DHSs of the 7 mutation classes which are deregulated compared to CD34+ve PBSCs as described in Fig S4A. (C) The percentage of DHS peaks that overlap with ATAC-Seq data from different progenitor types, DHS clusters from Figure 3A was overlapped with each of the progenitor ATAC peaks; these include CLP, CMP, GMP, MPP, LMPP, MEP and Monocyte populations.

Figure 4: AML-specifically active cis-regulatory elements display AML type-specific transcription factor occupancy patterns. (A) UCSC browser screen shot of the *MDFI* locus zooming in on an AML type-specific DHS (box). (B) Heatmap depicting the degree of motif enrichment after hierarchical clustering of motif occupancy in each of the 20 AML DHS clusters. Enrichment score was calculated by the level of motif enrichment in all the footprints of all high read-depth samples for each cluster, as compared to union of footprints in all experiments. (C) Enrichment analysis of motifs footprinted in AML subgroups which overlap with ATAC-Seq peaks present in precursor cells [25].

Figure 5: Capture HiC shows differences in locus-specific cis-regulatory interactions between different types of AML and normal cells. (A) Heatmaps showing the raw interactions of the promoter capture HiC data using purified patient blasts on chromosome 2 for the FLT3-ITD (FLT3-ITD/NPM1 patient) (left), t(8;21) (middle) and CD34+ (right), a UCSC tracks is shown below each heatmap. (B) Flow diagram shows the step for

identification of the differential interactions and the downstream analysis. (C) Percentage of up- and down-regulated genes with differential interactions from the FLT3-ITD and the t(8;21) compared to $CD34^+$. The bar figure shows also the percentage of the common genes for the FLT3-ITD and the t(8;21), the number of DEG is shown on top of each bar. (D) Top enriched GO terms for the up-regulated genes of the FLT3-ITD compared to the CD34+ as outlined in (A). (E) Network diagram of top KEGG pathways for the up-regulated genes of the FLT3-ITD compared to the CD34+ as outlined in (A). (F) Top enriched GO terms for the up-regulated genes of the t(8;21) compared to the CD34+ shown as outlined in (A) Network diagram of top KEGG pathways for the up-regulated genes of the t(8;21) compared to the CD34+ as outlined in (B). (H): percentage of RUNX1-ETO and RUNX1 targets amongst up-regulated genes with differential interactions.

Figure 6: Identification of transcription factor networks driving the expression of AML type-specific up-regulated TF genes

(A) Outline of analysis strategy. (B) t(8;21)-specific TF network, (C) CEBPA(x2)-specific TF, (D) INv(16) specific TF network, (E) Mutant RUNX1-specific TF network, (F) FLT3-ITD/NPM1 specific TF network, (G) NPM1-specific TF network

Factor families binding to the same motif as shown in Table S2 form a node contained within a circle. Arrows going outwards from the entire node highlight footprinted motifs in individual genes generated by any member of this factor family whereby the footprint was annotated to the gene using the CHiC data where possible, otherwise to the nearest gene. For selected nodes, the name of the underlying motif is highlighted in large grey letters. The expression level (FKPM) for the individual genes is depicted in white (low)/red (high) colour. An orange smooth ring around the circle indicates that this gene is specifically up-regulated in this type of AML compared to CD34+ PBSCs and/or other AML types, a dotted circle indicates a gene

that is up-regulated as compared to CD34+ cells. Genes with no outgoing arrows due to a lack of know binding motifs are highlighted by their octagon shapes.

7: Identification of AML type-specific TFs required for maintaining leukemic growth and colony forming ability.

(A - C) Histogram showing the growth curves of (A) Kasumi-1 cells after transduction with sh*POU4F1* and (B) of MV4-11 cells after transduction with *shNFIX*. (D, E) Histogram showing the number of colonies formed by a FLT3-ITD+ primary AML cell samples (D) or PBSCs (E) after transduction with shRNA targeting FOXC1, NFIX or a mismatch control. (F) Histogram showing the growth curve of Kasumi-1 cells transduced with either a doxycycline-inducible dominant negative FOS or an empty vector control (right panel) with and without 1.5 mcg/ml doxycycline. (G) Histogram showing the growth curve of MV4-11 cells transduced with either a doxycycline-inducible dominant negative FOS or an empty vector control (right panel) with and without 1.5 µg/ml doxycycline. (H,I) The expression of a dominant negative FOS causes a reduction in the colony forming ability of $CD34^+$ FLT3-ITD+ primary AML cells (H) but not $CD34^+$ PBSCs (I). All experiments were performed in triplicate (n=3) with * $p<0.05$, ** $p<0.01$. Error bars show 95% confidence intervals. (J) (J) Granulosarcoma formation in RG mice by Kasumi-1 expressing a doxycycline-inducible dnFOS. dnFOS was induced by intraperitoneal injection of doxycycline. (K) Survival curve for RG mice transplanted with MV4-11 cells expressing doxycycline-inducible dnFOS. dnFOS was induced by adding doxycycline to the drinking water. The control group did not develop any tumors during the observed time frame while all mice of the induced group had to be sacrificed.

Figure 1

A

Bone marrow aspirate or peripheral blood

Same day

Mono-nuclear cells

CD34⁺ or CD117⁺ AML blast cells

mRNA

Open chromatin
Digital footprints

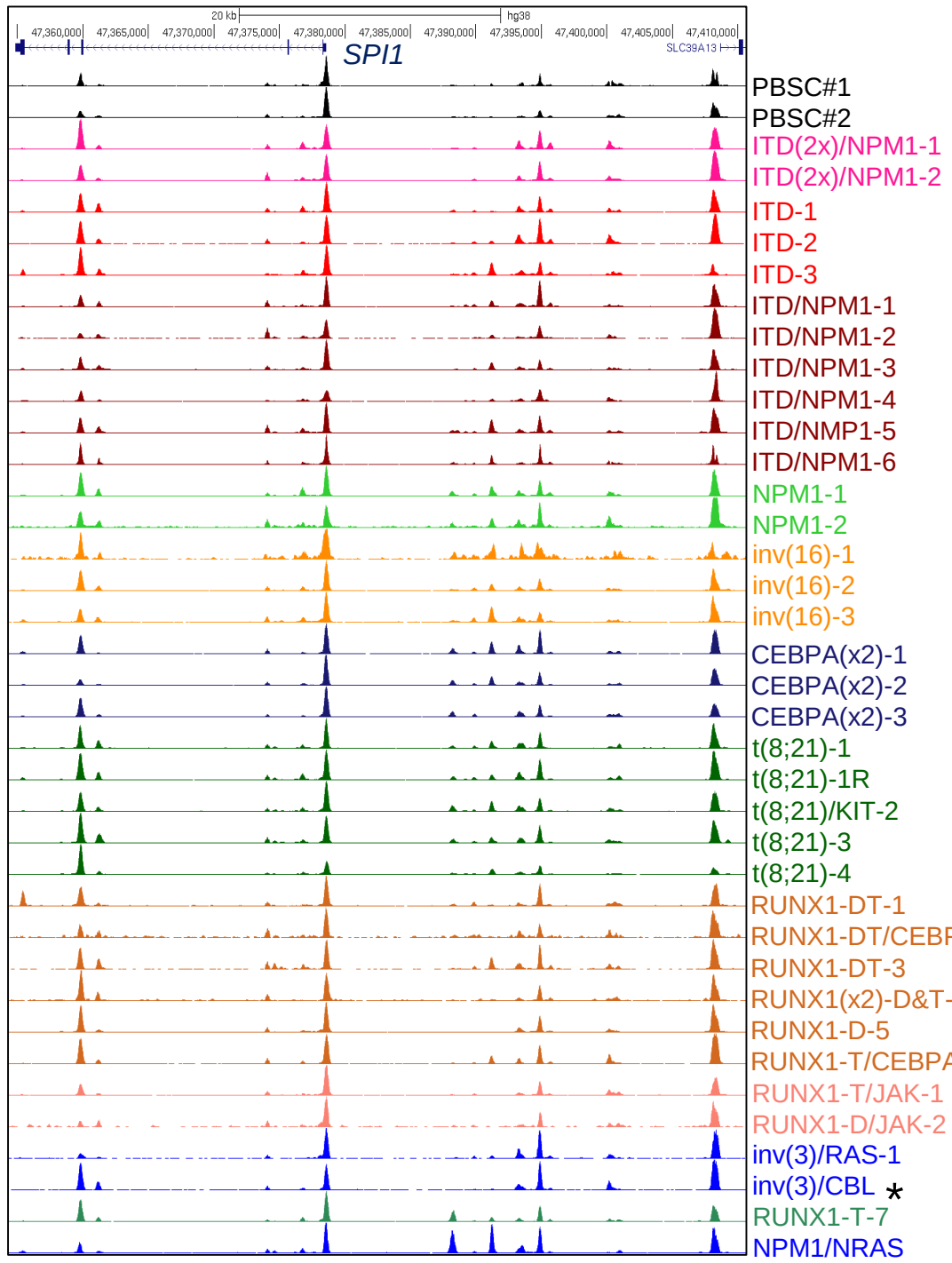
Mutations

Promoter-cis-element interactions

Integrative data analysis

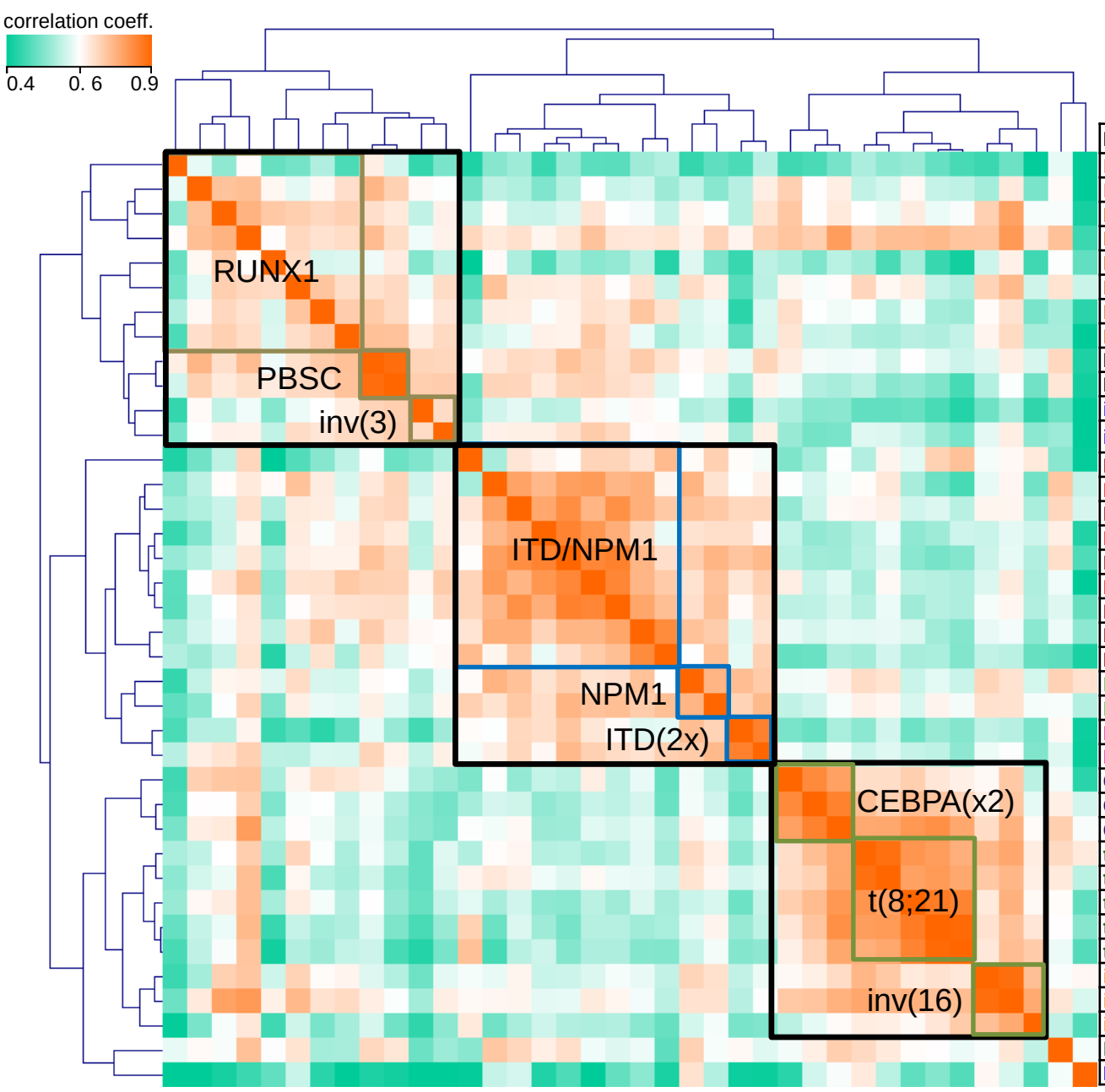
B

DHSs



C

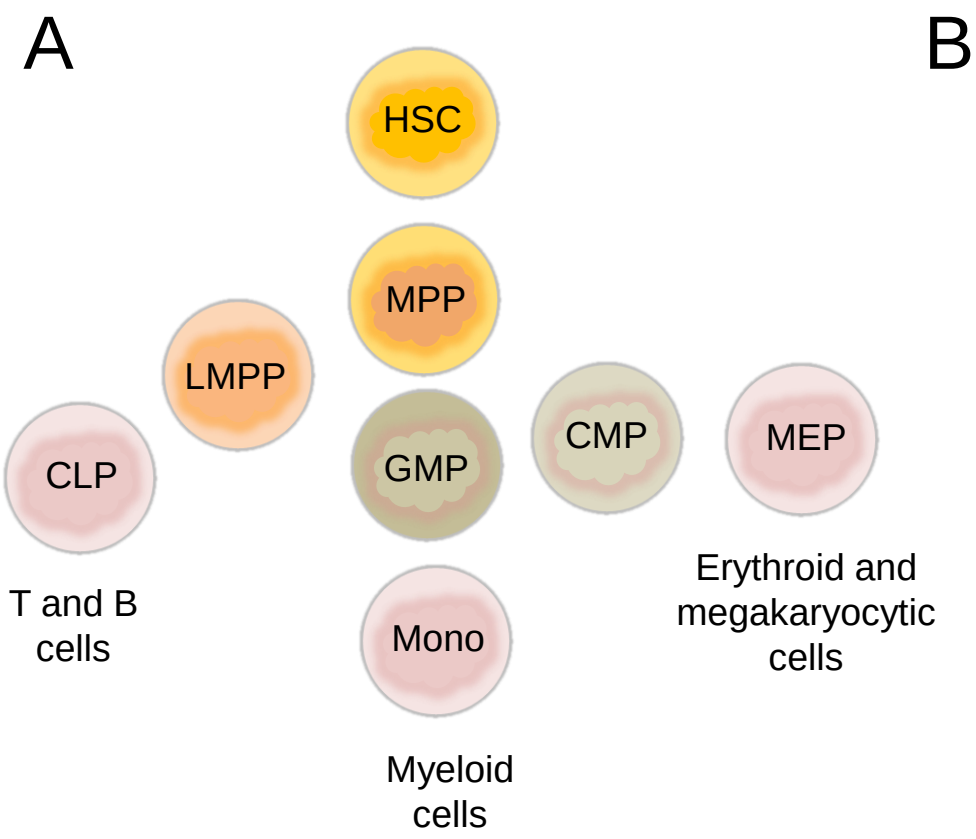
DNA-Seq correlation clustering



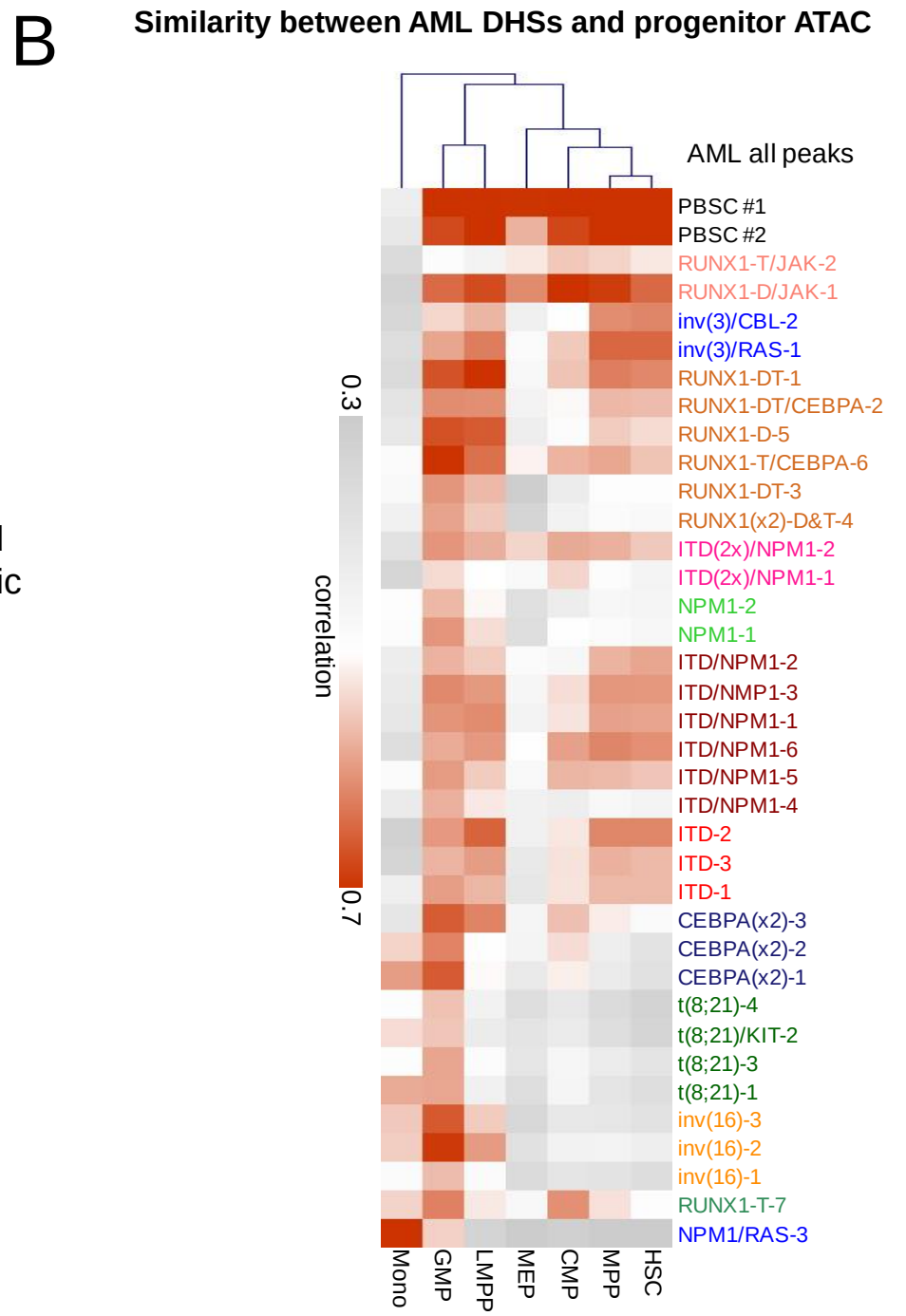
Mutations

Patient code	Signalling	NPM1	Chrom	RUNX	Other mutations
RUNX1-T/JAK-2	JAK2			RUNX1	TET2x2,TP53
RUNX1-D/JAK-1	JAK2		tri(21,9)	RUNX1	IDH2, SRSF2
RUNX1-D-5				RUNX1	IDH1, BCORL1x2, SRSF2x2
RUNX1-T/CEBPA-6	NRAS		tri(8)	RUNX1	CEBPA, EZH2
RUNX1(x2)-D&T-4				RUNX1x2	DNMT3A, IDH2, SRSF2
RUNX1-DT-3				RUNX1	
RUNX1-DT/CEBPA-2	FLT3-ITD			RUNX1	CEBPA, WT1x2, SF3B1, TP53
RUNX1-DT-1	FLT3		tri(13)	RUNX1	CREBBP, DNMT3A, SF3B1
PBSC-1					
PBSC-2					
inv(3)/RAS-1	NRAS		inv(3)		GATA2, SF3B1
inv(3)/CBL-2 *	CBL		inv(3)		SF3B1
ITD/NPM1-4	FLT3-ITD	NPM1			GATA2, DNMT3A
ITD-1	FLT3-ITD				DNMT3A, TET2x2, BCOR, TP53
ITD/NPM1-5	FLT3-ITD	NPM1			DNMT3A, BCOR
ITD-3	FLT3-ITD				DNMT3A
ITD/NPM1-6	FLT3-ITD	NPM1			WT1, DNMT3A, TET2, PHF6
ITD-2	FLT3-ITD		tri(13)		DNMT3A, TET2
ITD/NPM1-1	FLT3-ITD	NPM1			WT1, DNMT3A
ITD/NPM1-3	FLT3-ITD	NPM1			
ITD/NPM1-2	FLT3-ITD	NPM1			
NPM1-1		NPM1			IDH1
NPM1-2		NPM1			DNMT3A, TET2x2
ITD(2x)/NPM1-2	FLT3-ITDx2	NPM1			CEBPA, IDH2
ITD(2x)/NPM1-1	FLT3-ITDx2	NPM1			DNMT3A, IDH2
CEBPA(x2)-3					CEBPAx2, GATA2, TET2
CEBPA(x2)-2					CEBPAx2, GATA2
CEBPA(x2)-1					CEBPAx2
t(8;21)-1			t(8;21)		TET2
t(8;21)-1R	KIT		t(8;21)		TET2
t(8;21)-3	FLT3-TK		t(8;21)		
t(8;21)/KIT-2	KIT		t(8;21)		NOTCH1
t(8;21)-4			t(8;21)		
inv(16)-3			inv(16)		ASXL1
inv(16)-2			inv(16)		
inv(16)-1	KIT		inv(16)		
RUNX1-T-7 (NHL)			tri(21)	RUNX1	TET2x2, PHF6
NPM1/RAS-3	NRAS	NPM1			PTPN11, DNMT3A, IDH1

Figure 2



ATAC-Seq & RNA-Seq data from Corces et al., 2016



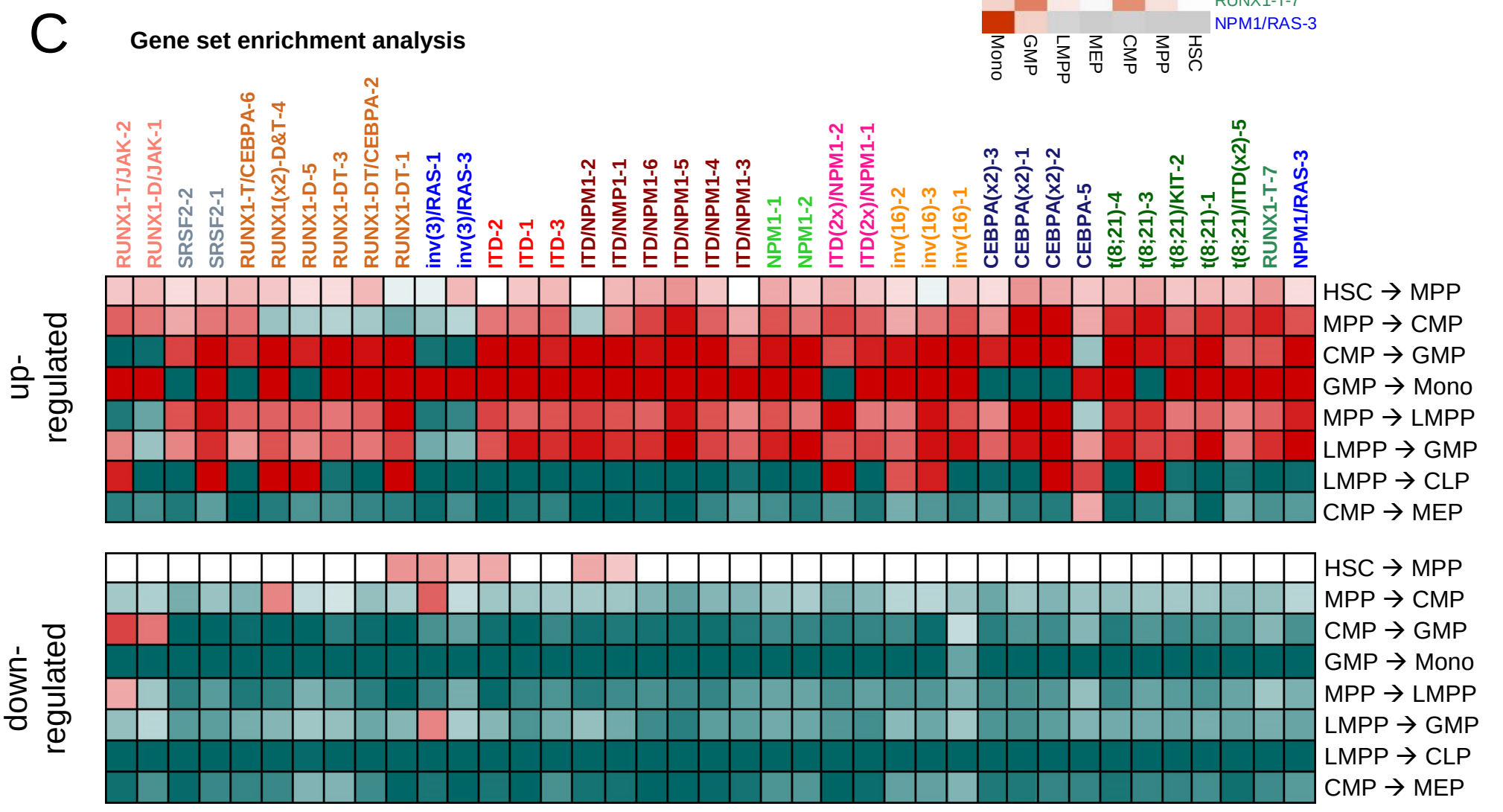
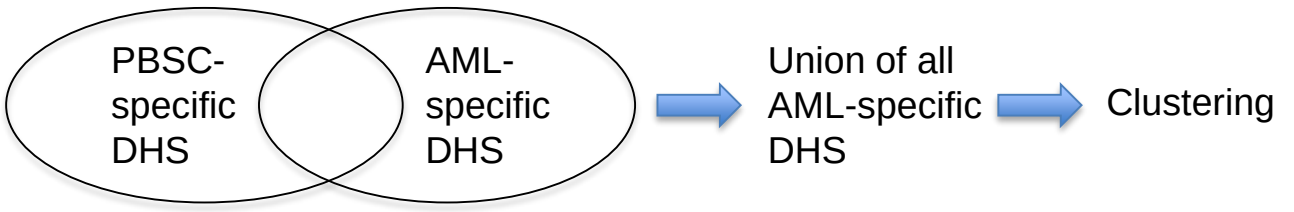
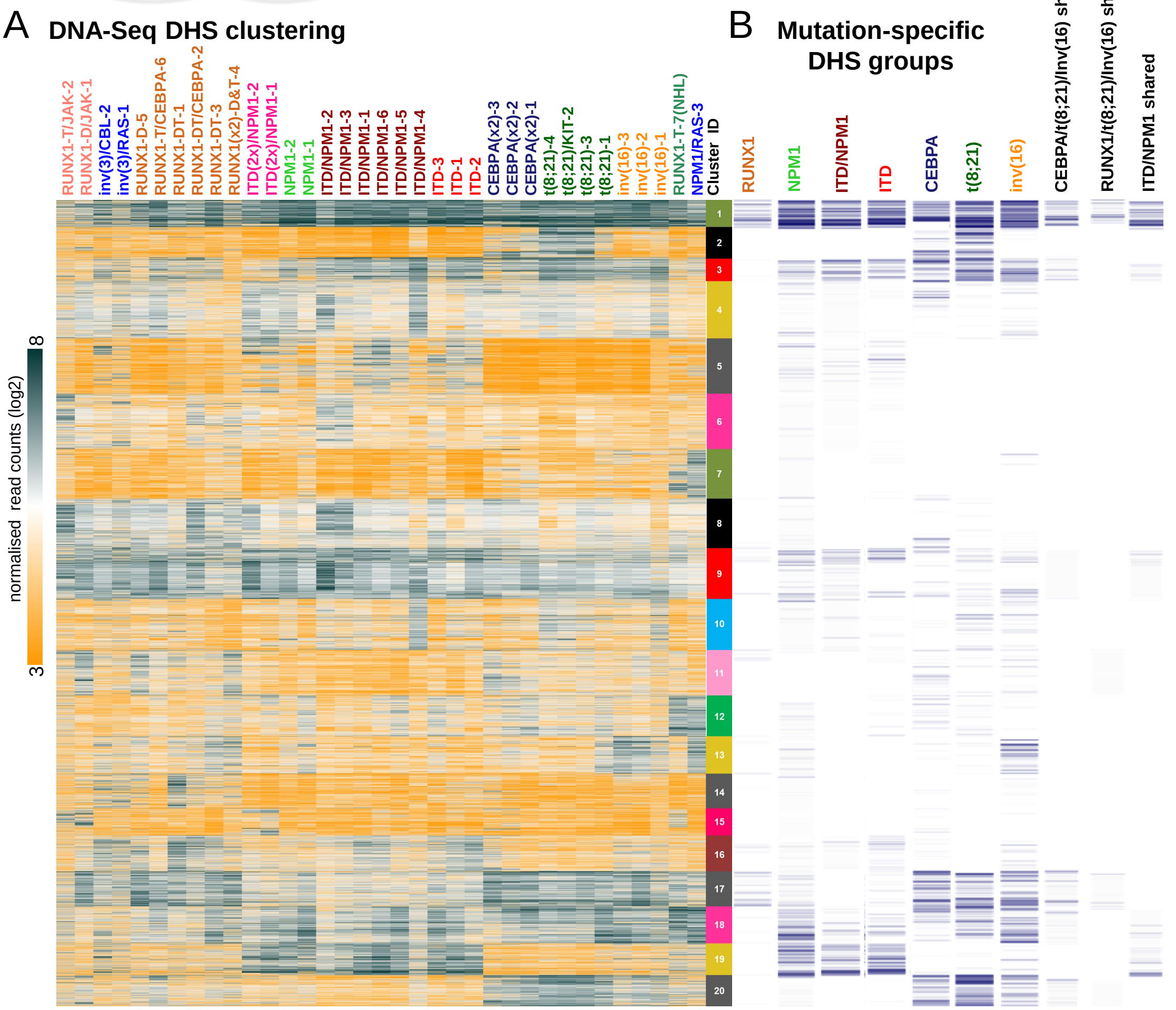


Figure 3


PBSC-specific DHS
AML-specific DHS
Union of all AML-specific DHS
Clustering


A
DNA-Seq DHS clustering
RUNX1-T/JAK-2
RUNX1-D/JAK-1
inv(3)/CBL-2
inv(3)/RAS-1
RUNX1-D-5
RUNX1-T/CEBPA-6
RUNX1-DT-1
RUNX1-DT/CEBPA-2
RUNX1-DT-3
RUNX1(x2)-D&T-4
ITD(2x)/NPM1-2
ITD(2x)/NPM1-1
NPM1-2
NPM1-1
ITD/NPM1-2
ITD/NPM1-3
ITD/NPM1-1
ITD/NPM1-6
ITD/NPM1-5
ITD/NPM1-4
ITD-3
ITD-1
ITD-2
CEBPA(x2)-3
CEBPA(x2)-2
CEBPA(x2)-1
t(8;21)-4
t(8;21)/KIT-2
t(8;21)-3
t(8;21)-1
inv(16)-3
inv(16)-2
inv(16)-1
RUNX1-T-7(NHL)
NPM1/RAS-3
Cluster ID
normalised read counts (log2)
8
3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
B
Mutation-specific DHS groups
RUNX1
NPM1
ITD/NPM1
ITD
CEBPA
t(8;21)
inv(16)
CEBPA/t(8;21)/Inv(16) shared
RUNX1/t(8;21)/Inv(16) shared
ITD/NPM1 shared

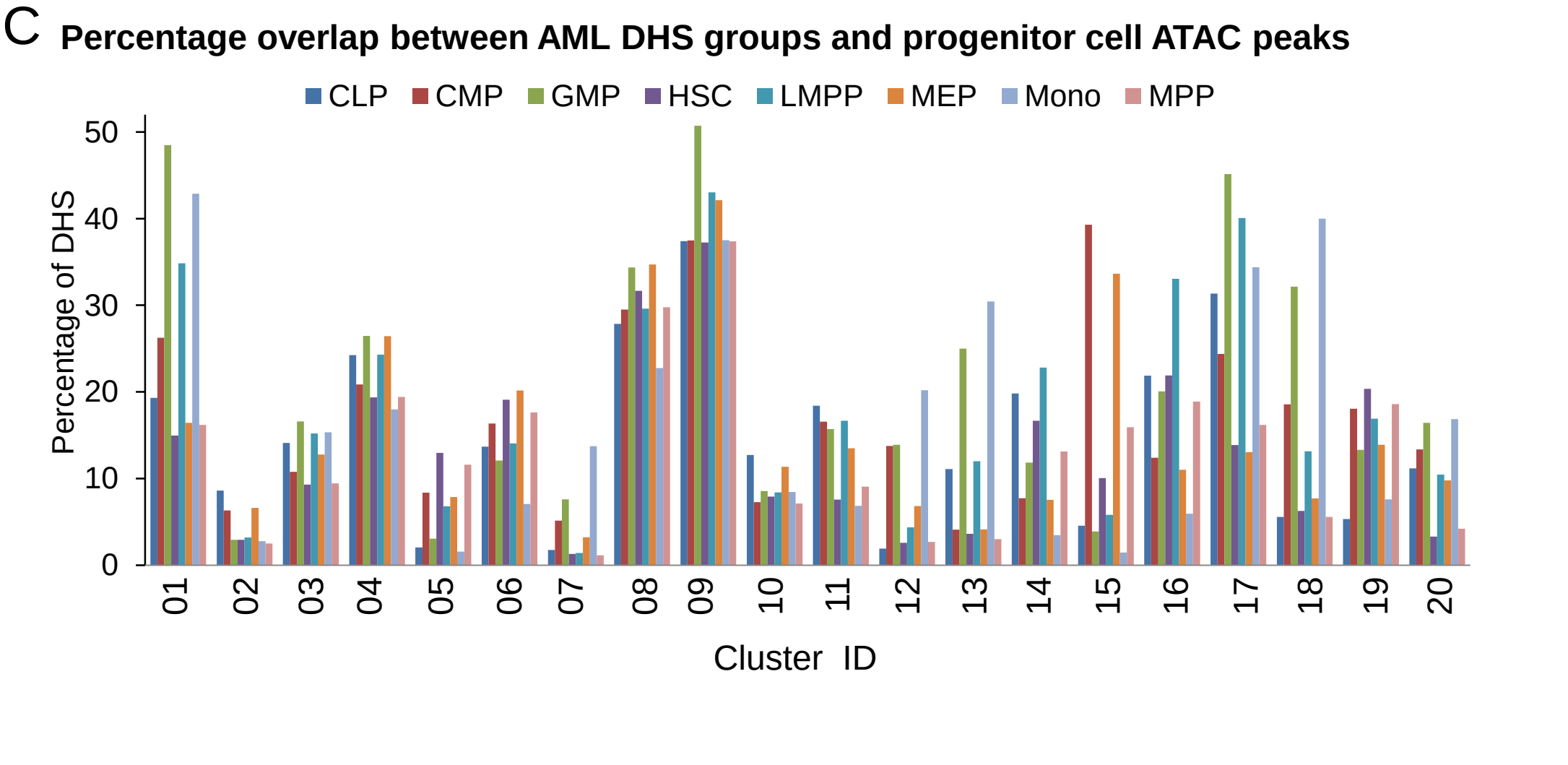
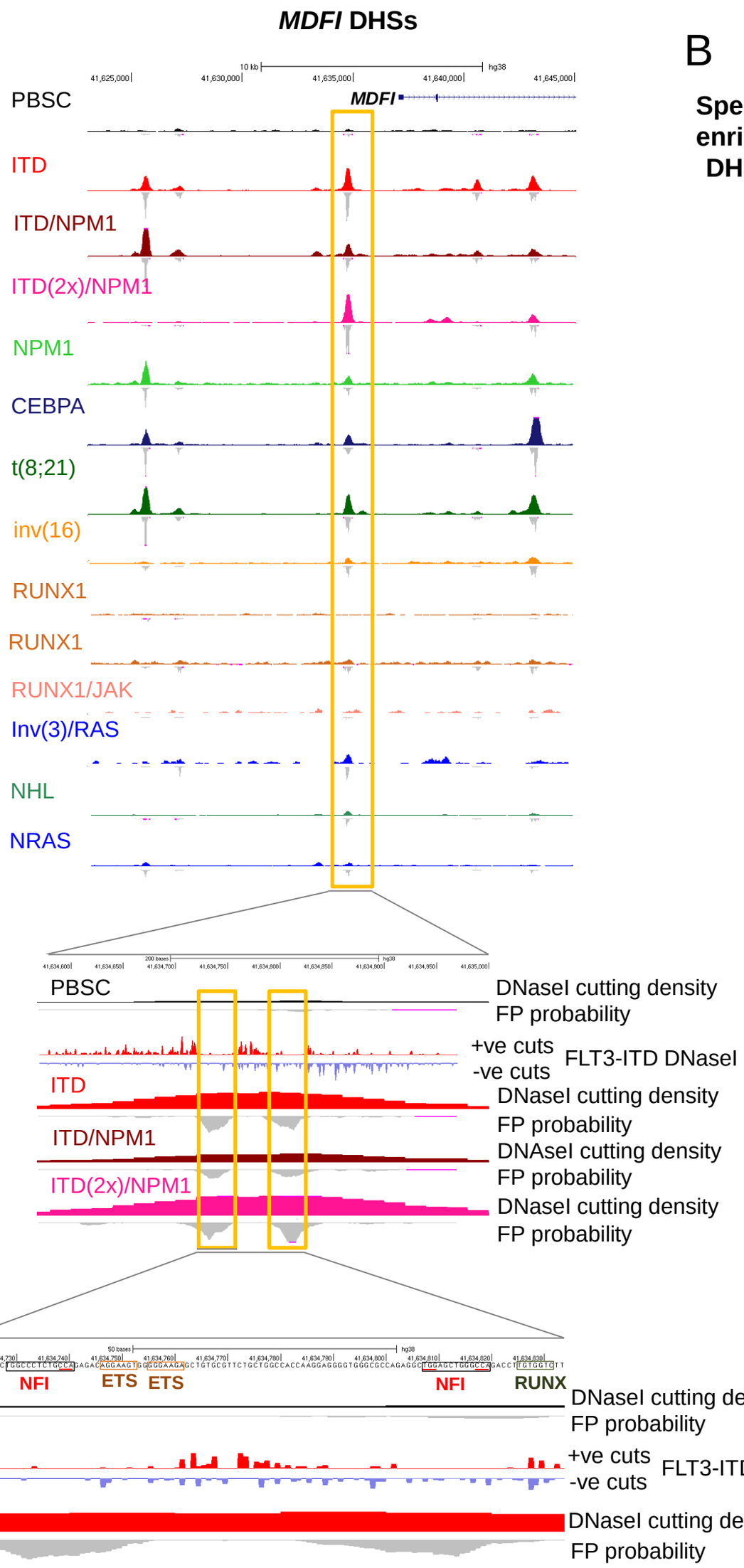

C
Percentage overlap between AML DHS groups and progenitor cell ATAC peaks
CLP
CMP
GMP
HSC
LMPP
MEP
Mono
MPP
Percentage of DHS
0
10
20
30
40
50
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
Cluster ID

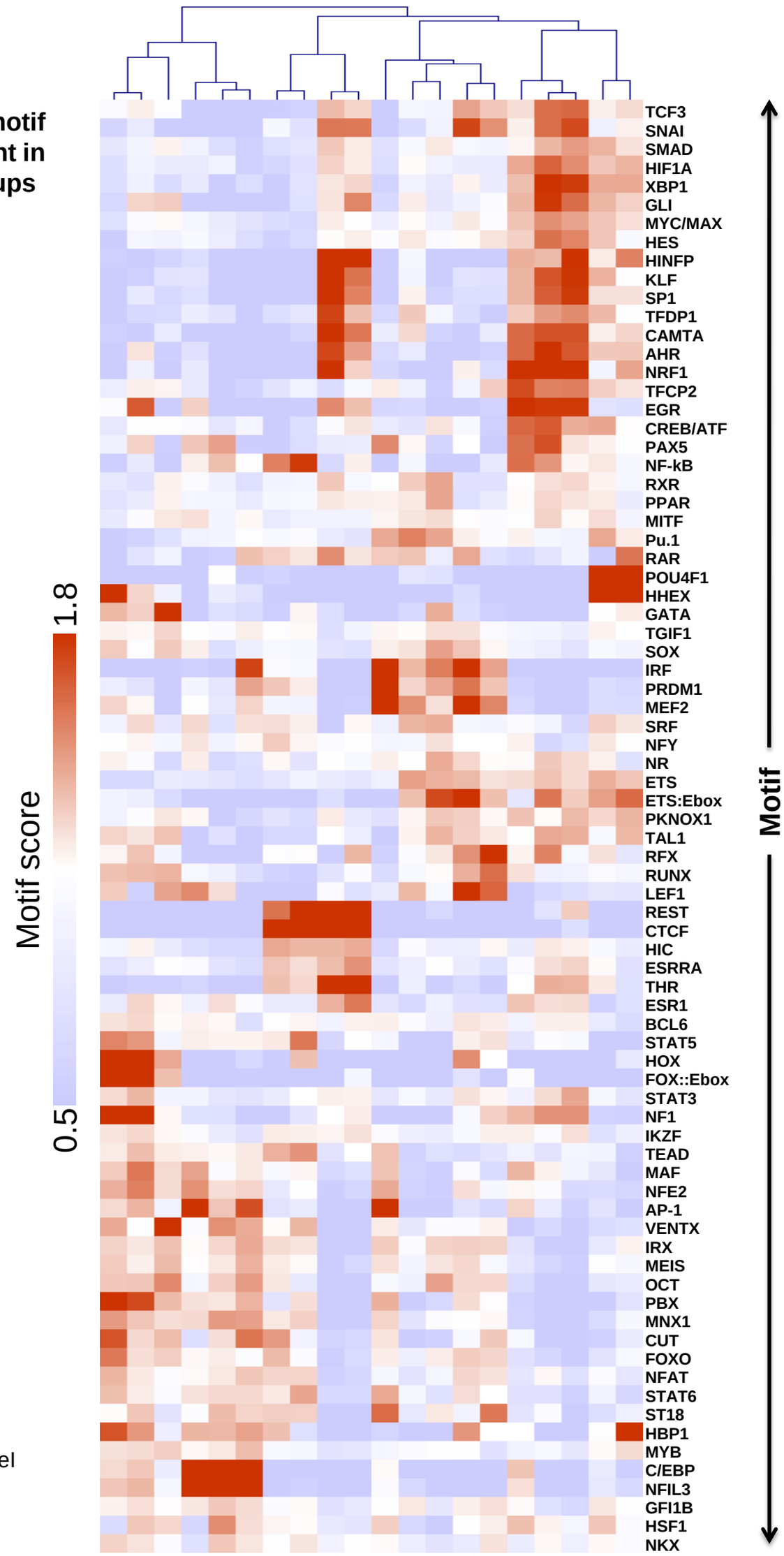
Figure 4

A



B

Specific motif enrichment in DHS groups

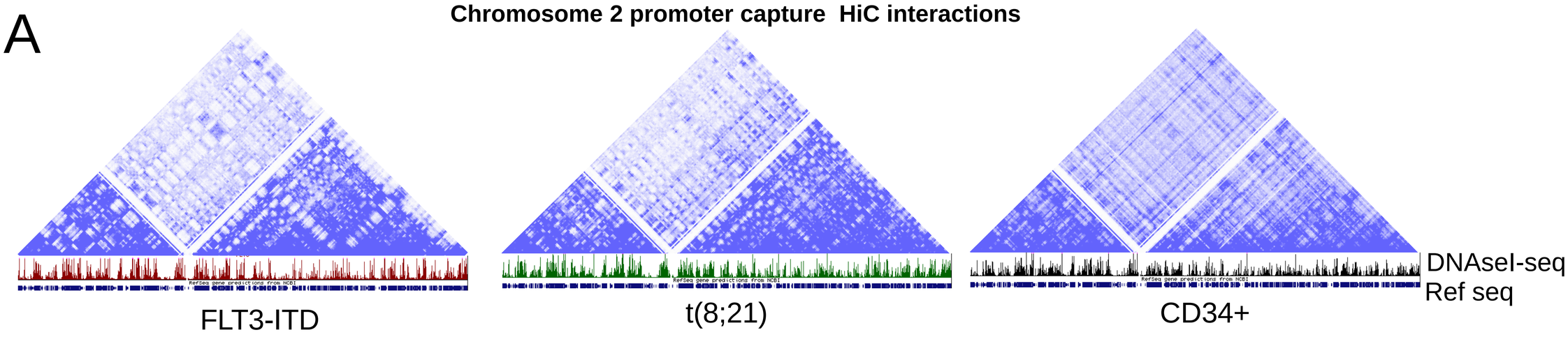


C

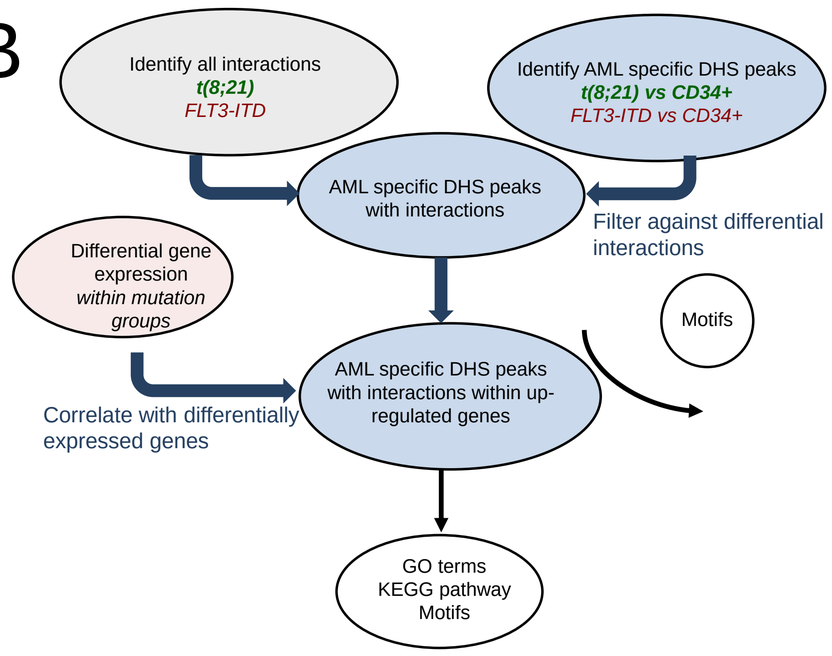
Enrichment of specific motifs footprinted in AML cells within ATAC peaks present in progenitor cells

Figure 5

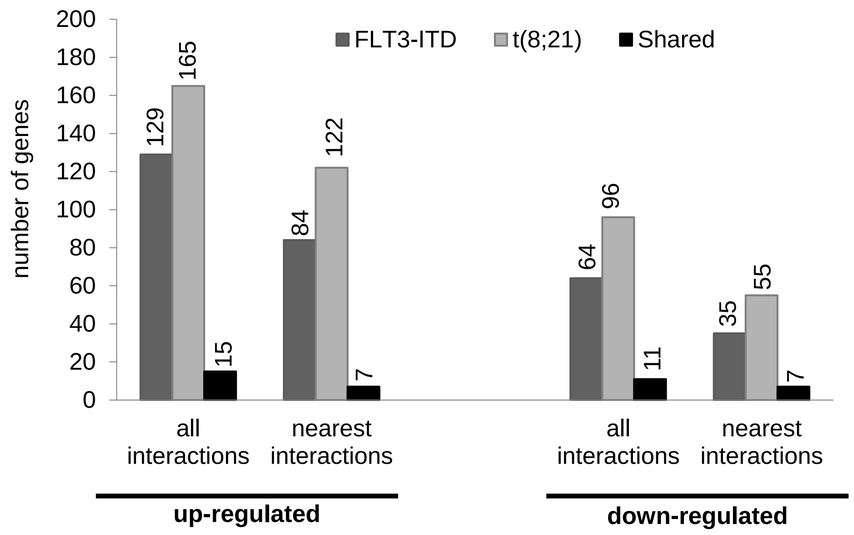
A



B



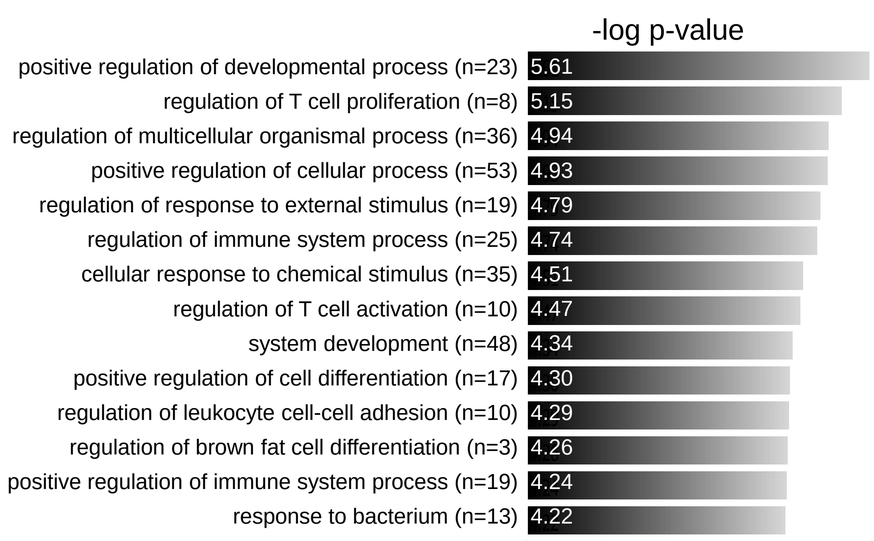
C



D

FLT3-ITD

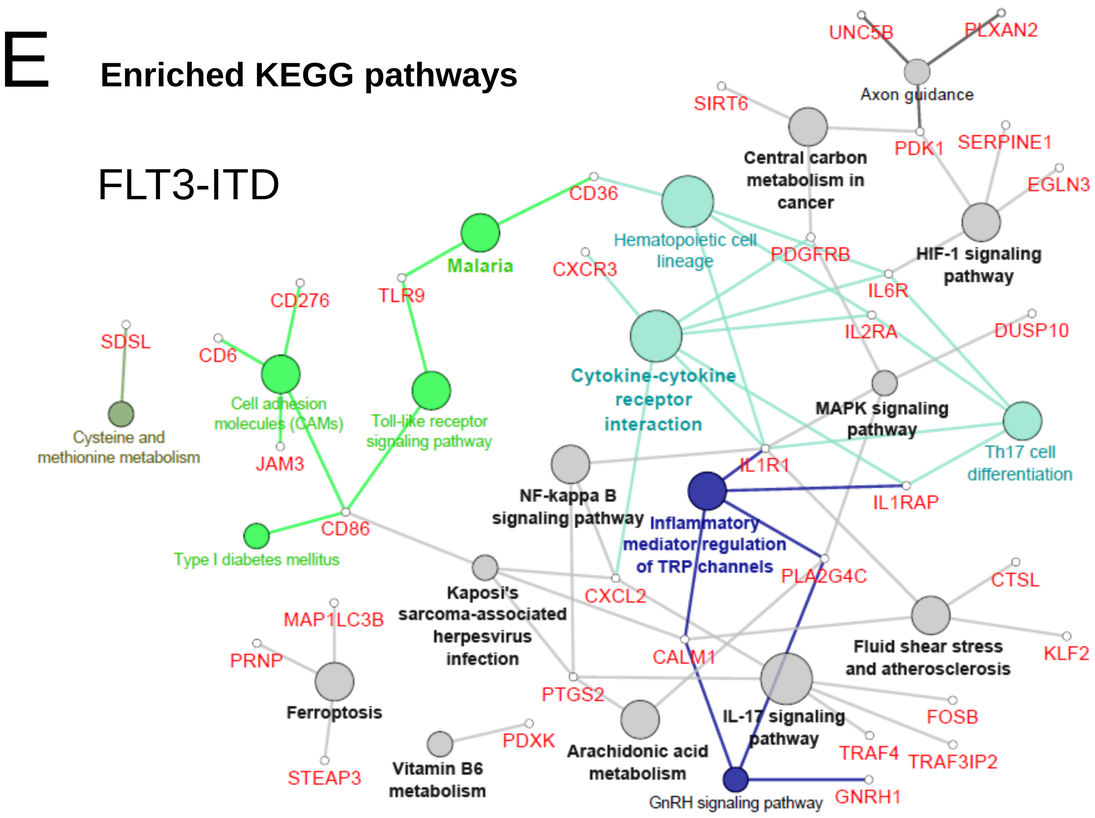
Enriched GO Terms



E

Enriched KEGG pathways

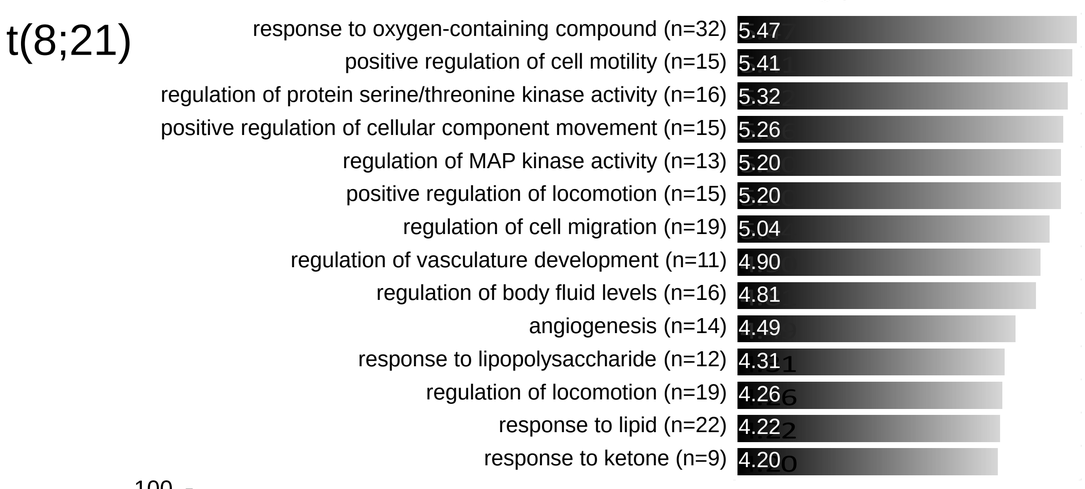
FLT3-ITD



F

Enriched GO Terms

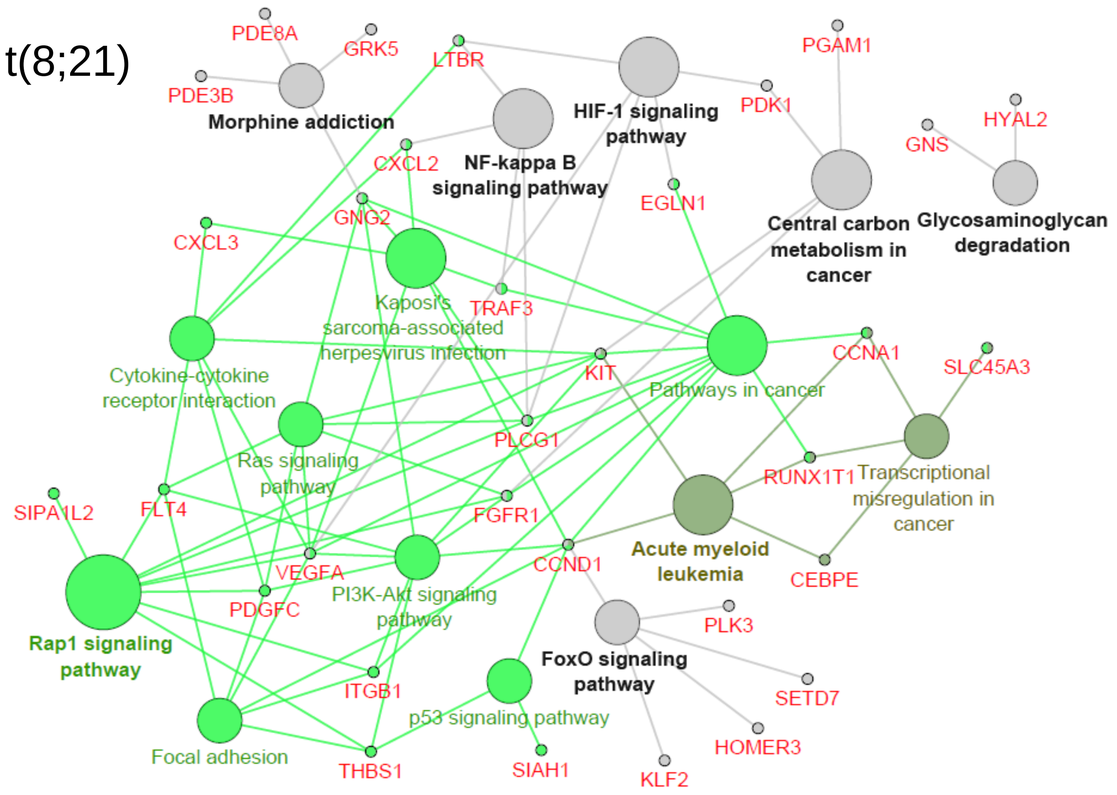
t(8;21)



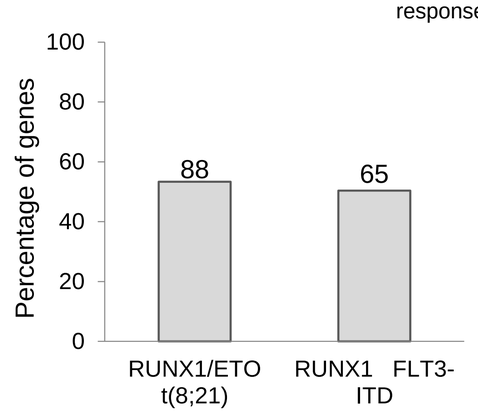
G

Enriched KEGG pathways

t(8;21)



H

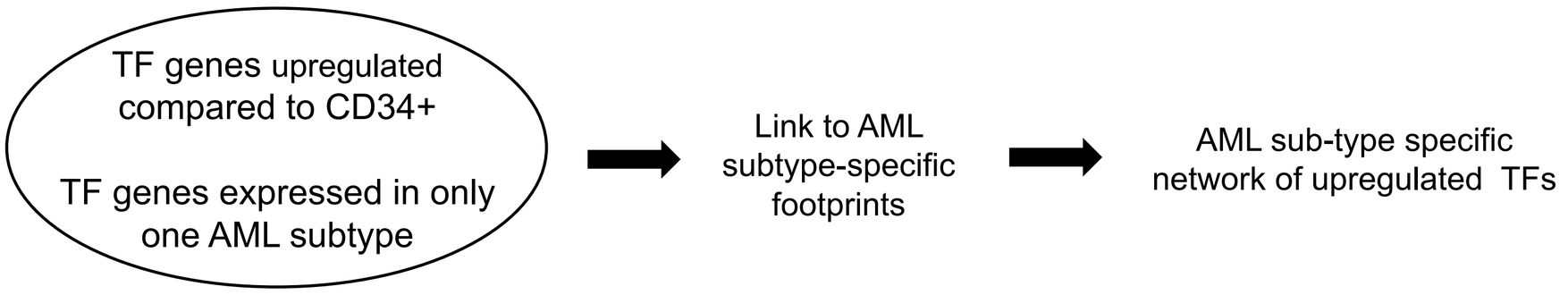


Node size significance

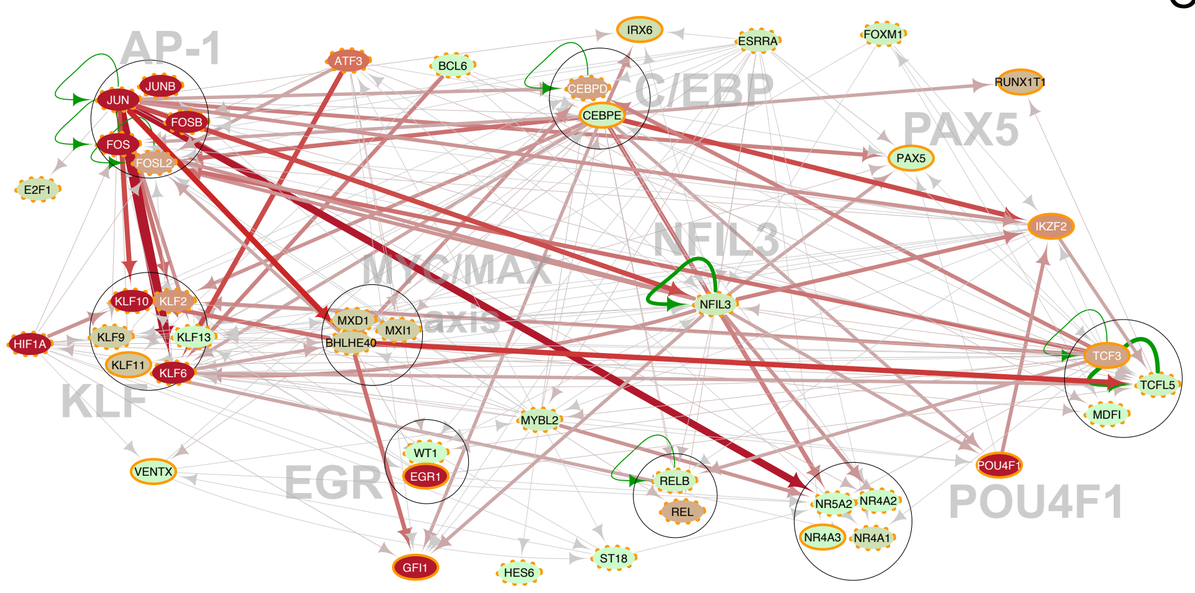
- p > 0.1
- p 0.05-0.1
- p 0.005-0.05
- p 0.0005-0.005
- p <0.005

Figure 6

A

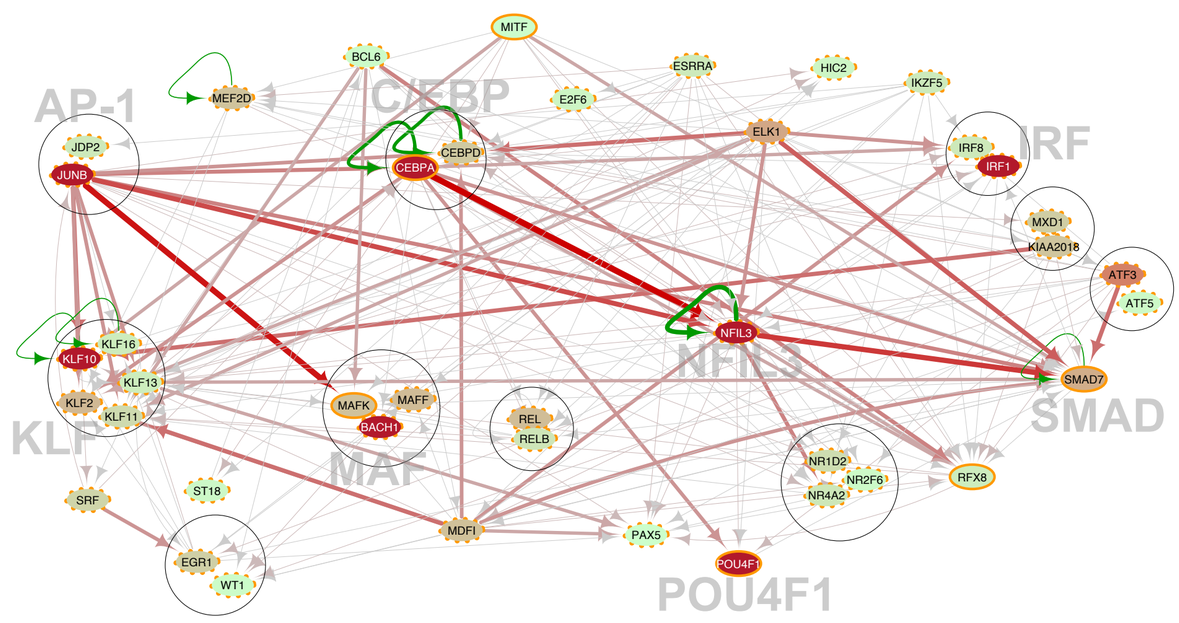


B



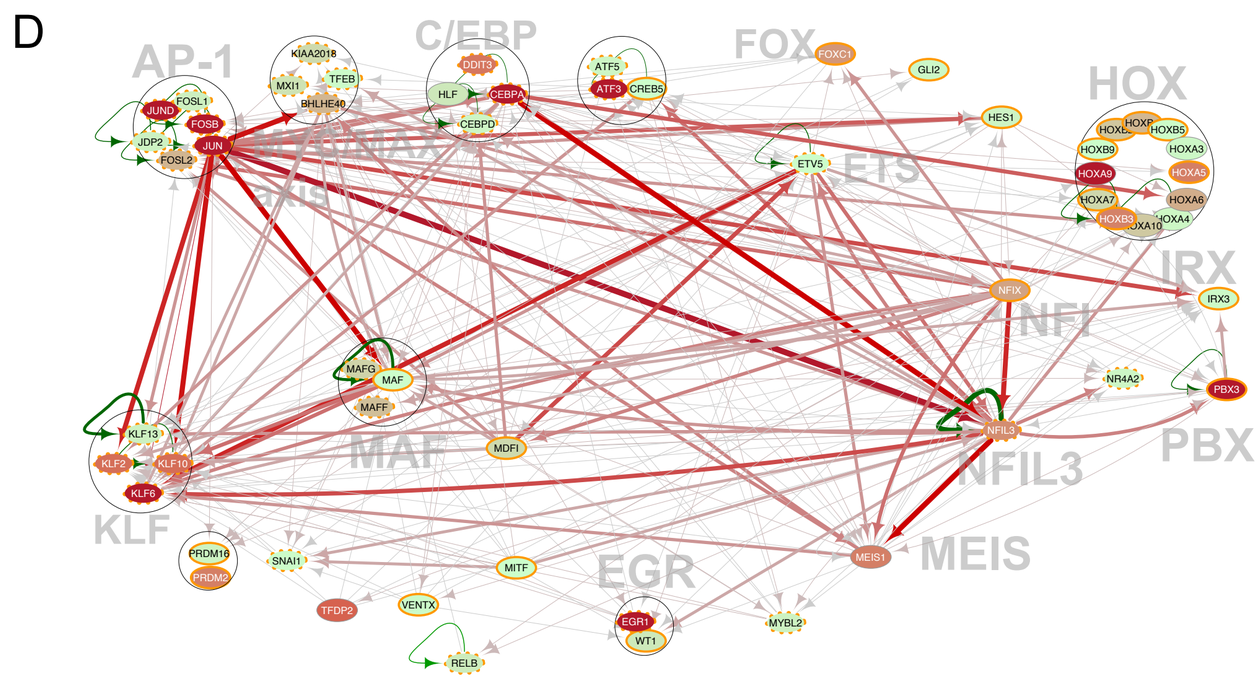
t(8;21) specific TF network

C



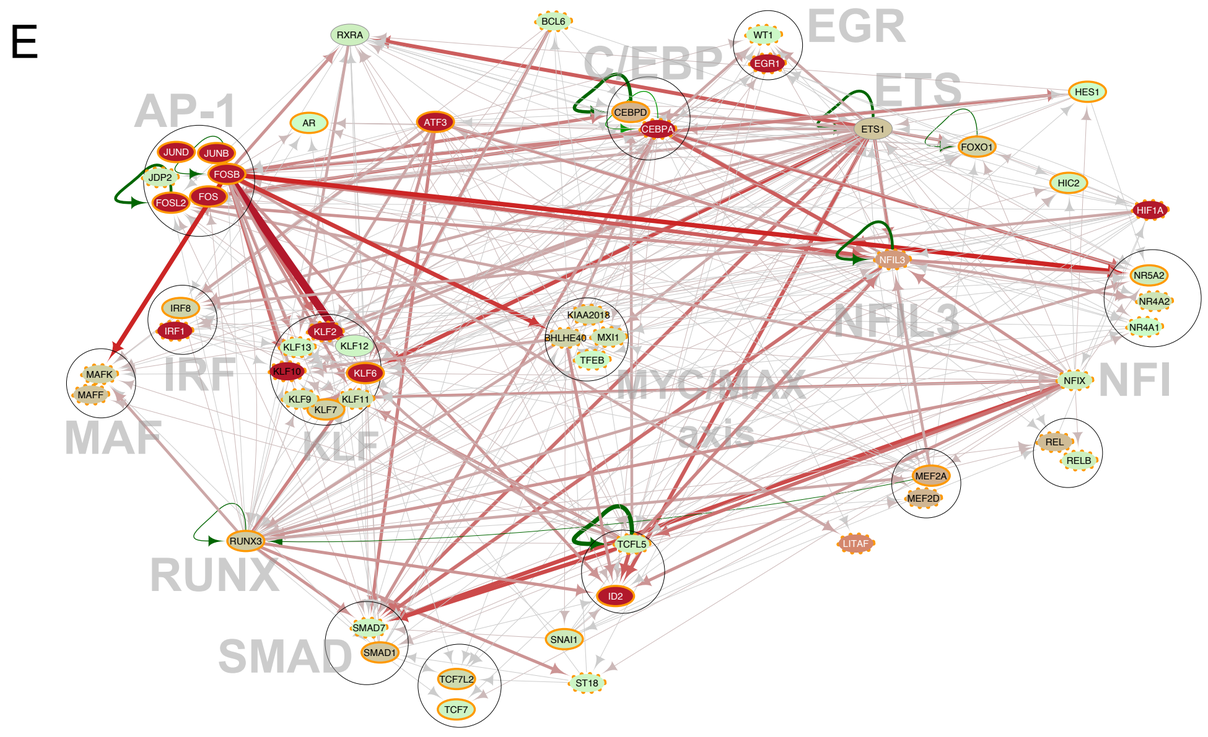
CEBPA(x2) specific TF network

D



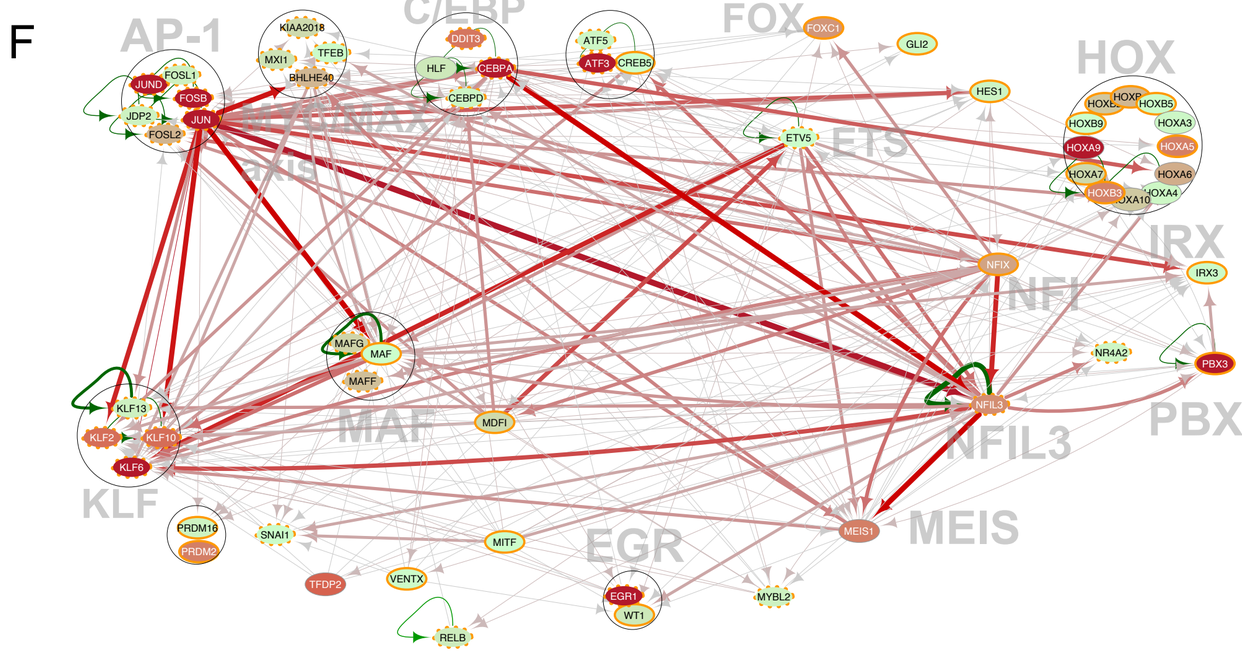
Inv(16) specific TF network

E



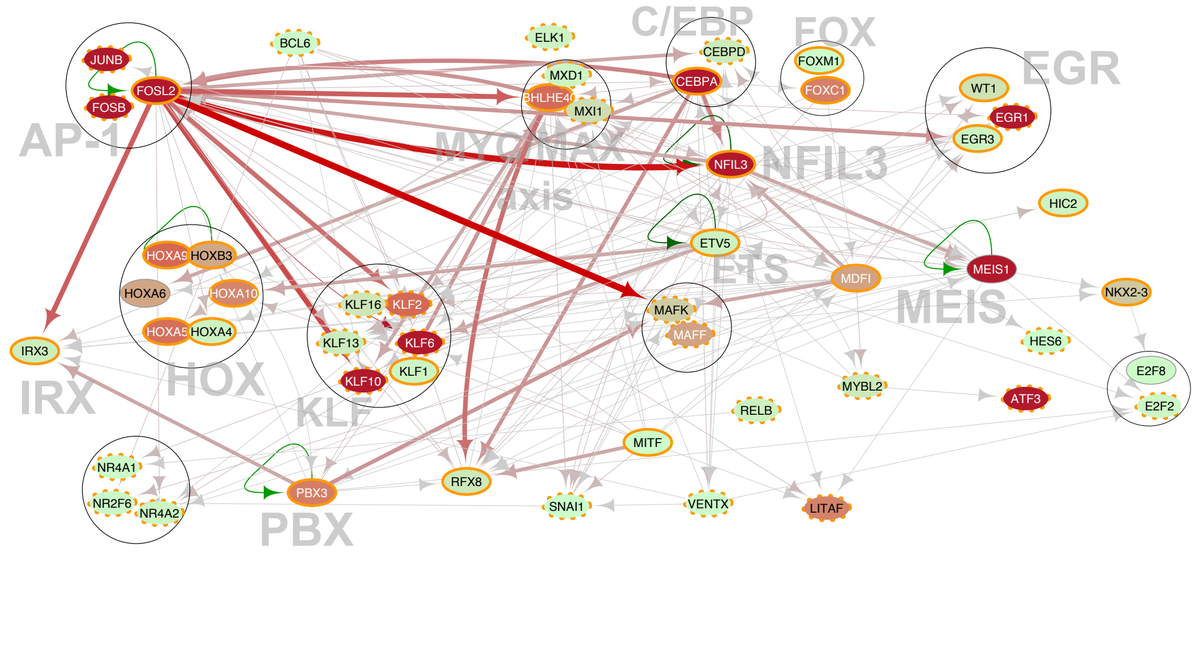
Mutant RUNX1 specific TF network

F



FLT3-ITD/NPM1 specific TF network

G



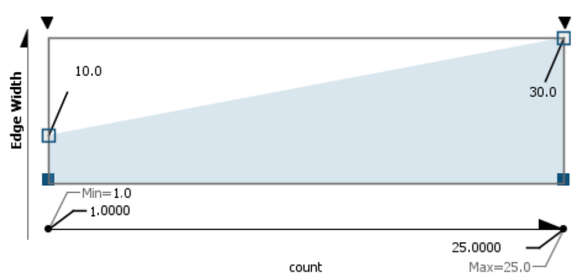
NPM1 specific TF network

node colour
fpkm
2
>30
white node colour:
gene not expressed (<2 fpkm)

border colour
down-regulated
up-regulated
invariant

Border type
up regulated AML vs CD34+ only
up regulated subtype specific group vs others

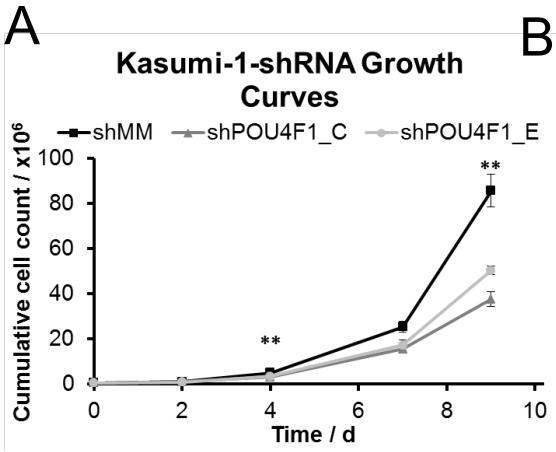
edge colour
1
count
>12



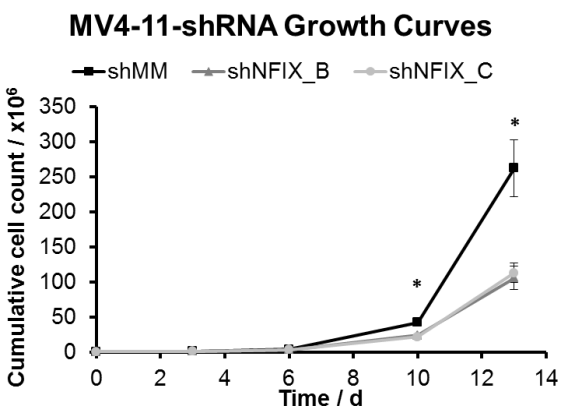
Node and edge attributes

Figure 7

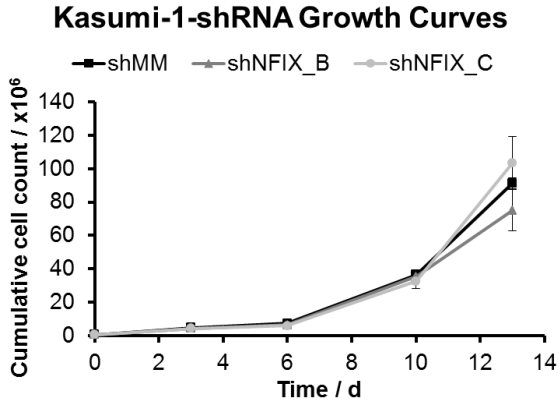
A



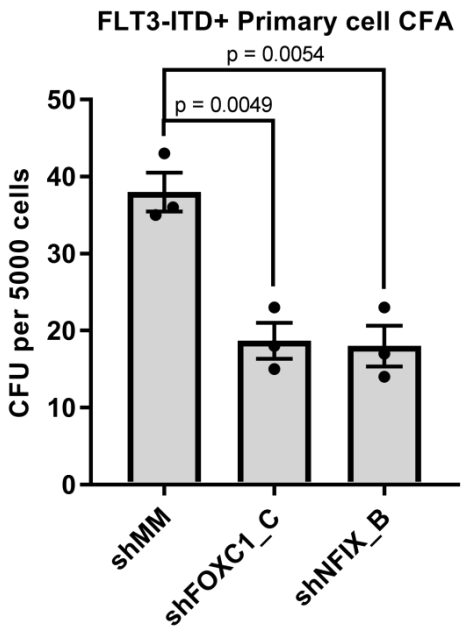
B



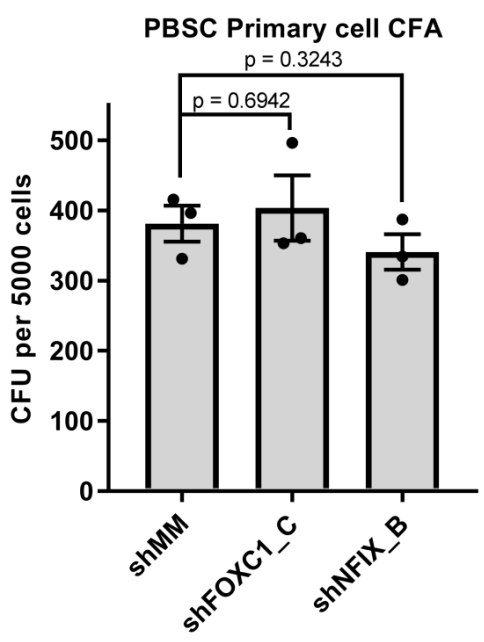
C



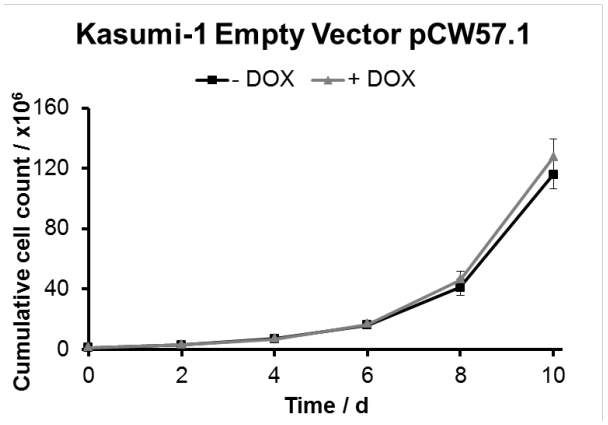
D

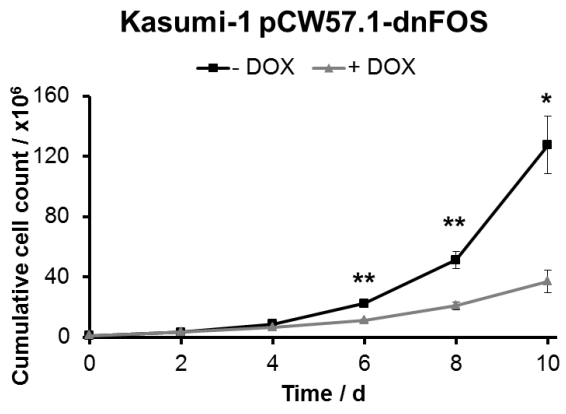


E

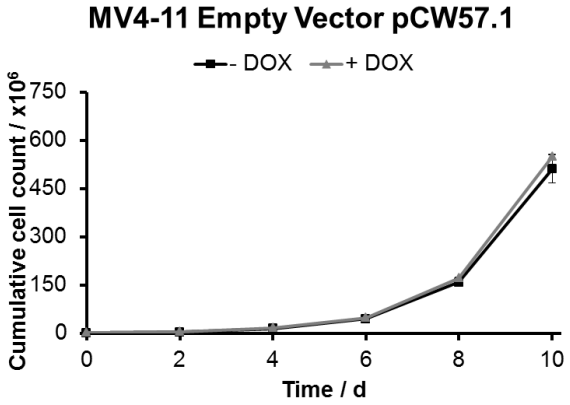


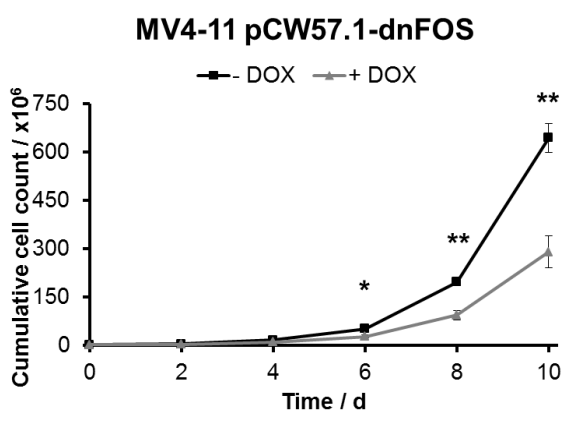
F



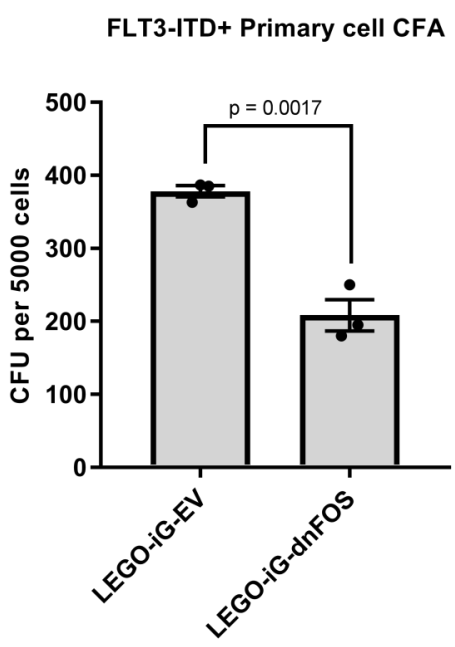


G

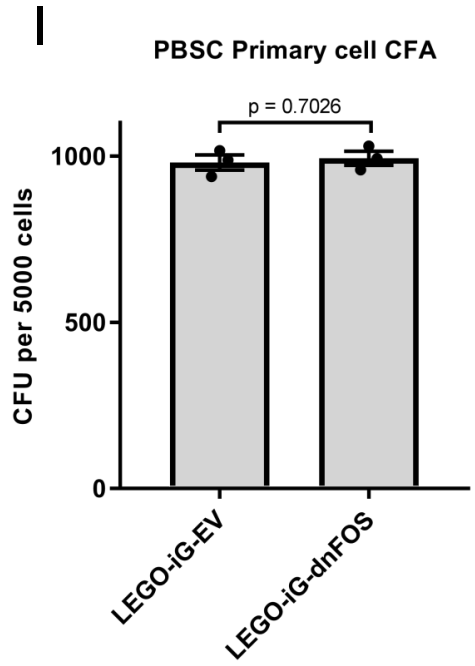




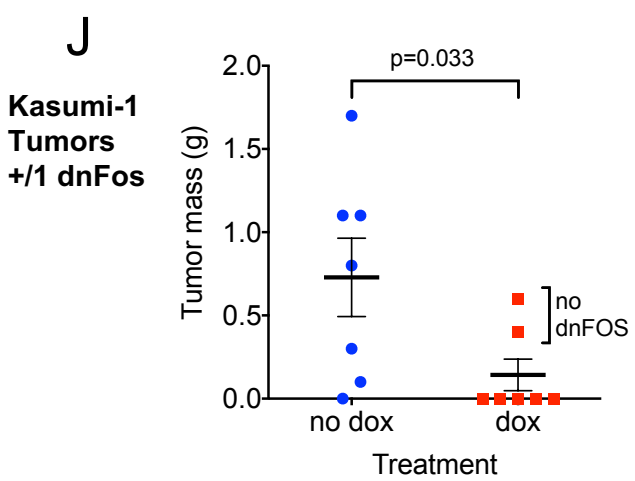
H



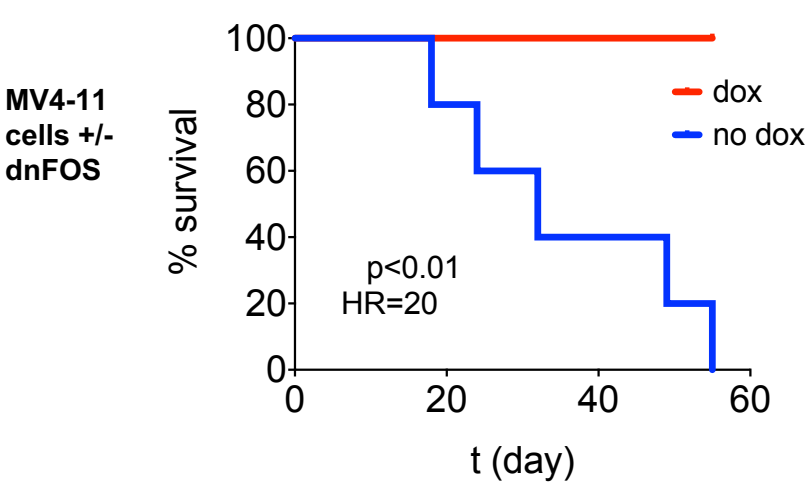
I



J



K



Long terminal repeat (LTR) elements are wide-spread in the human genome and have the potential to act as promoters and enhancers. Their expression is therefore under tight epigenetic control. We previously reported in classical Hodgkin Lymphoma (cHL) that a member of the THE1B class of LTR elements acted as a promoter for the proto-oncogene and growth factor receptor gene CSF1R and that expression of this gene is required for cHL tumour survival. However, to which extent and how such elements participate in globally shaping the unique cHL gene expression program is unknown. To address this question we mapped the genome-wide activation of THE1-LTRs in cHL cells using a targeted next generation sequencing approach (RACE-Seq). Integration of these data with global gene expression data from cHL and control B cell lines showed a unique pattern of LTR activation impacting on gene expression, including genes associated with the cHL phenotype. We also show that global LTR activation is induced by strong inflammatory stimuli. Together these results demonstrate that LTR activation provides an additional layer of gene deregulation in classical Hodgkin lymphoma and highlight the potential impact of genome-wide LTR activation in other inflammatory diseases.

Supplemental Materials

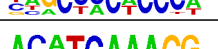
1. Supplemental Figures and Tables

Supplemental Table 1: Patient groups, mutation data, and clinical data. Patient codes depicted in color represent samples included in the seven major defined mutation groups, or which have either 2 FLT-ITD mutations or a mutation in either CBL or NRAS. This table also indicates samples where DHS-Seq and RNA-Seq data is either available (Y) or not available (N). Further details can be found in Supplemental data-set 1.

patient code	Signalling	NPM1	Chrom	RUNX	Other mutations	DHS Seq	RNA Seq	Age	Sex	wbc	case
ITD-1	FLT3-ITD				DNMT3A, TET2x2, BCOR, TP53	Y	Y	45	F	56	Rel
ITD-2	FLT3-ITD		tri(13)		DNMT3A, TET2	Y	Y	68	F	2	Pres
ITD-3	FLT3-ITD				DNMT3A	Y	Y	80	F	143	Pres
ITD/NMP1-1	FLT3-ITD	NPM1			DNMT3A, WT1	Y	Y	45	F	32	Pres
ITD/NPM1-2	FLT3-ITD	NPM1				Y	Y	61	F	7	Rel
ITD/NPM1-3	FLT3-ITD	NPM1				Y	Y	66	F	91	Pres
ITD/NPM1-4	FLT3-ITD	NPM1			GATA2, DNMT3A	Y	Y	65	F	21	Pres
ITD/NPM1-5	FLT3-ITD	NPM1			DNMT3A, BCOR	Y	Y	68	M	190	Pres
ITD/NPM1-6	FLT3-ITD	NPM1			WT1, DNMT3A, TET2, PHF6	Y	Y	58	F	195	Pres
NPM1-1		NPM1			IDH1	Y	Y	37	M	60	Pres
NPM1-2		NPM1			DNMT3A, TET2x2	Y	Y	75	M	94	Pres
t(8;21)-1			t(8;21)		TET2	Y	Y	72	M	29	Pres
t(8;21)/KIT-2	KIT		t(8;21)		NOTCH1	Y	Y	48	M	36	Pres
t(8;21)-3	FLT3-TK		t(8;21)			Y	Y	53	M	6	Pres
t(8;21)-4			t(8;21)			Y	Y	45	M	2	Pres
inv(16)-1	KIT		inv(16)			Y	Y	40	M	22	Pres
inv(16)-2			inv(16)			Y	Y	26	M	63	Pres
inv(16)-3			inv(16)		ASXL1	Y	Y	75	M	54	Pres
RUNX1-DT-1	FLT3		tri(13)	RUNX1	CREBBP, DNMT3A, SF3B1	Y	Y	68	M	112	Rel
RUNX1-DT/CEBPA-2	FLT3-ITD			RUNX1	CEBPA, WT1x2, SF3B1, TP53	Y	Y	83	M	68	Pres
RUNX1-DT-3				RUNX1		Y	Y	58	M	37	Pres
RUNX1(x2)-D&T-4				RUNX1x2	SRSF2, DNMT3A, IDH2	Y	Y	82	M	55	Pres
RUNX1-D-5				RUNX1	IDH1, BCORL1x2, SRSF2x2	Y	Y	65	M	8	Pres
RUNX1-T/CEBPA-6	NRAS		tri (8)	RUNX1	CEBPA, EZH2	Y	Y	75	M	107	Pres
CEBPA(x2)-1					CEBPAx2	Y	Y	76	F	238	Pres
CEBPA(x2)-2					CEBPAx2, GATA2	Y	Y	21	F	10	Pres
CEBPA(x2)-3					CEBPAx2, GATA2, TET2	Y	Y	75	M	106	Pres
ITD(2x)/NPM1-1	FLT3-ITDx2	NPM1			DNMT3A, IDH2	Y	Y	78	F	26	Pres
ITD(2x(/NPM1-2	FLT3-ITDx2	NPM1			CEBPA, IDH2	Y	Y	72	F	68	Pres
NPM1/RAS-3	NRAS	NPM1			PTPN11, DNMT3A, IDH1	Y	Y	30	F	4	Rel
inv(3)/RAS-3	NRAS		inv(3)		ETV6, SF3B1	N	Y	54	M	104	Pres
inv(3)/RAS-1	NRAS		inv(3)		GATA2, SF3B1	Y	Y	59	M	4	Rel
inv(3)/CBL-2	CBL		inv(3)		SF3B1	Y	N	34	F	21	Rel
t(8;21)/ITD(x2)-5	FLT3-ITD		t(8;21)		SMC1A	N	Y	43	M	86	Pres
RUNX1-D/JAK-1	JAK2		tri (21, 9)	RUNX1	IDH2, SRSF2	Y	Y	79	M	12	Pres
RUNX1-T/JAK-2	JAK2			RUNX1	TET2x2, TP53	Y	Y	77	F	79	Pres
RUNX1-T-7 (NHL)			tri (21)	RUNX1	TET2x2, PHF6	Y	Y	73	F	NA	Pres
CEBPA-5					CEBPA, DNMT3A	N	Y	79	F	40	Pres
SRSF2-1					IDH2, SRSF2	N	Y	67	M	2	Pres
SRSF2-2					SOCS1, DNMT3A, IDH2, SRSF2	N	Y	71	M	2	Pres
t(8;21)-1R	KIT		t(8;21)		TET2	Y	Y	72	M	29	Rel

Supplemental Table 2

List of used Position Weight Matrices

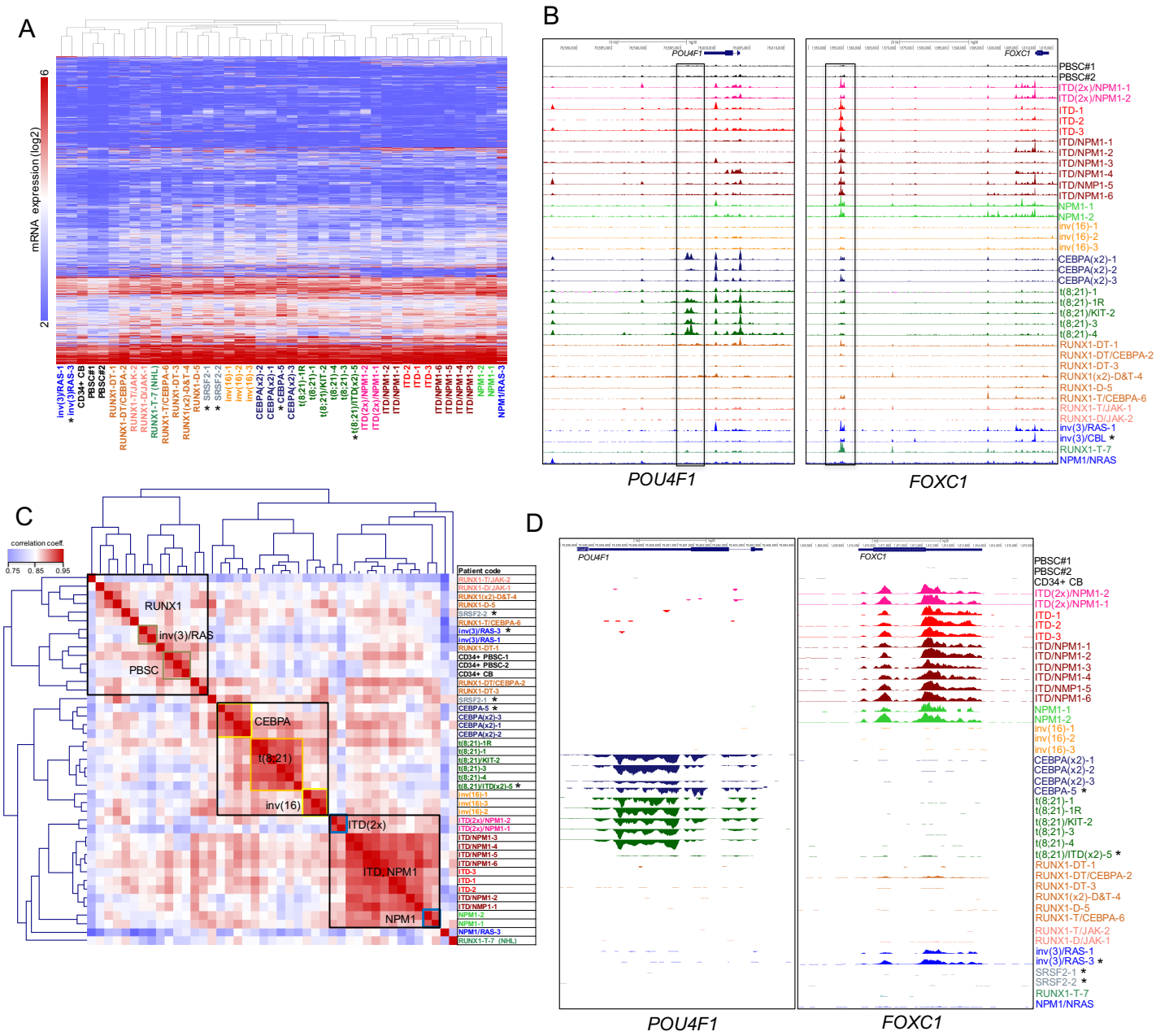
motif	logo	motif	logo	motif	logo
AHR		HSF1		PRDM1	
AP-1		IKZF		PU.1	
AR		IRF		RAR	
BCL6		IRX		REST	
CAMTA		KLF		RFX	
C/EBP		LEF1		RUNX	
CREB/ATF		MAF		RXR	
CTCF		MYC/MAX		SMAD	
CUT		MEF2		SNAI	
E2F		MEIS		SOX	
EGR		MITF		SP	
ESR1		MNX1		SRF	
ESRRA		MYB		ST18	
ETS		NF1		STAT3	
ETS:E-box		NFAT		STAT5	
EVI		NFE2		STAT6	
FOXO		NFIL3		TAL1	
FOX:E-box		NF-kB		TCF3	
GATA		NFY		TEAD	
GFI1B		NKX		TFCP2	
GLI		NR		TFDP1	
HBP1		NRF1		TGIF	
HES		OCT		THR	
HHEX		PAX5		VDR	
HIC1		PBX		VENTX	
HIF1A		PKNOX1		XBP1	
HINFP		POU4F1			
HOX		PPAR			

§

Supplemental Table 2: List of representative position weight matrices for TF families.

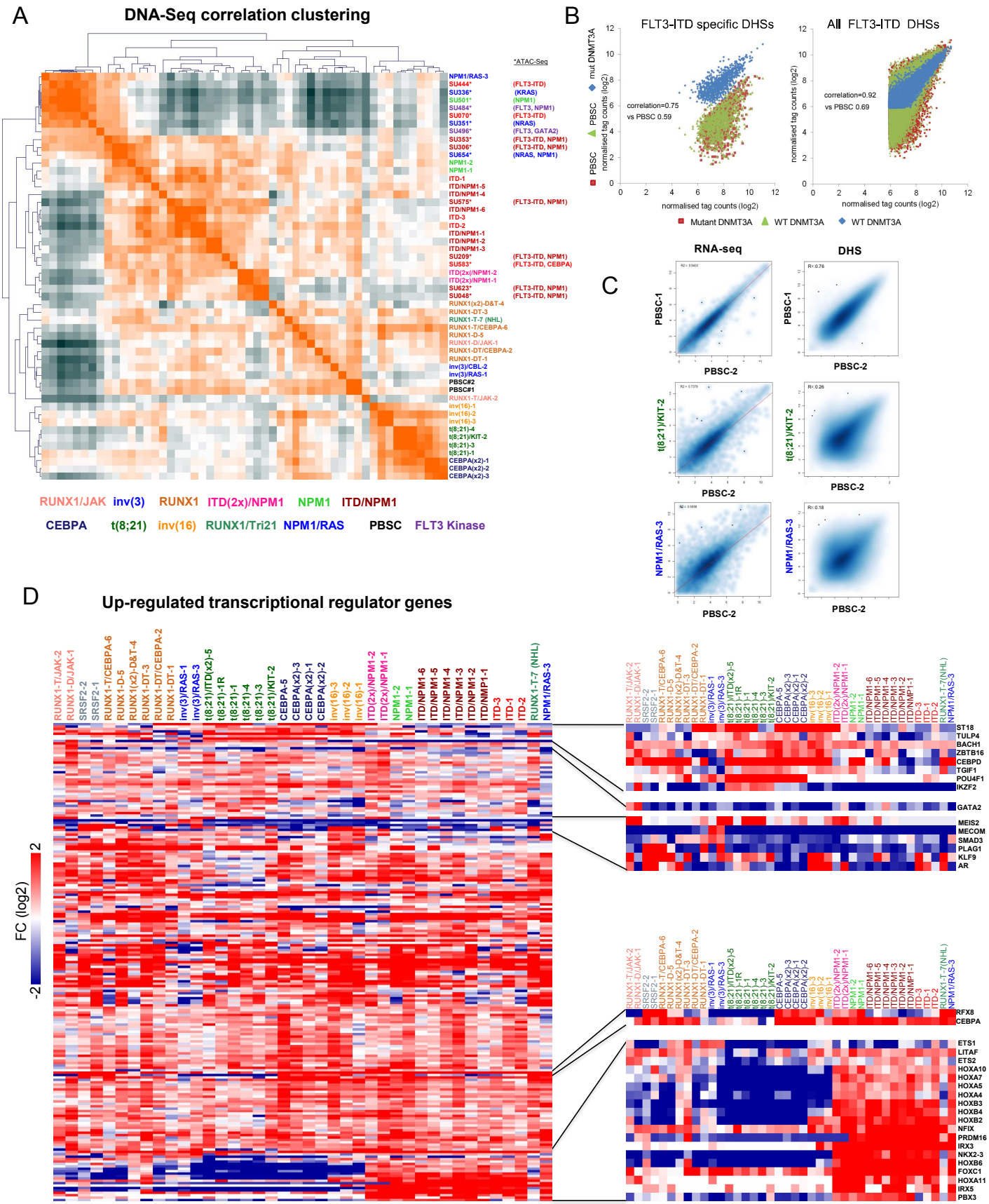
To improve the process of linking regulatory factors with their binding sites on DNA, we consolidated the different versions of transcription factor consensus binding sequences for closely related family members where the motif signatures are indistinguishable. For most transcription factor families there are typically various alternate subtly different versions of position weight matrices for not just different family members but also for the same factor from different data sets. The prevalence of so many different related consensus sequences is a major impediment to the construction of regulatory networks from genome-wide analyses of DNA elements. For the current study, we first identified a subset of almost 300 transcription factor genes that are expressed in one or more of our AML samples. We then inspected the motifs listed on either the HOMER or JASPER databases, motifs defined in a recent large-scale study of recombinant proteins (Jolma et al 2013, 10.1016/j.cell.2012.12.009), or motifs described in various other publications. We grouped together those factors where the motifs are essentially the same, and chose the best representative example for further analysis. These selections were often validated by referring to the large body of literature which is devoted to defining specific motifs, which also informed the choices of which orientation of motifs represented the conventional form used in publications. The JASPAR motifs were viewed via http://jaspar.genereg.net/. The HOMER motifs were viewed via http://homer.ucsd.edu/homer/motif/HomerMotifDB/homerResults.html.

Supplementary Figure 1



Supplemental Figure 1: Different types of AML adopt unique transcriptomes. (A) Hierarchical clustering of gene expression as determined by RNA-Seq of all patient samples. Clustering of log2 FPKM values for all differentially expressed genes changing expression at least 2 fold in at least one patient as compared to normal CD34+ PBSC. (B) UCSC genome browser screenshots of DNaseI-Seq in all AML patients with different classes of mutations and normal CD34+ PBSC at *POU4F1* (left panel) and *FOXC1* (right panel) locus. (C) Hierarchical clustering of Pearson correlation coefficient between all patient samples of RNA-Seq data: (left panel), right panel: list of mutations in cells from each patient. The correlation between any two patients was obtained with log2 FPKM expression values over all genes. (D) UCSC genome browser screenshots of RNA-Seq reads in AML patients at *POU4F1* (left panel) and *FOXC1* (right panel) locus. Asterisks denote samples for which the matching RNA-Seq or DNaseI-Seq data are unavailable.

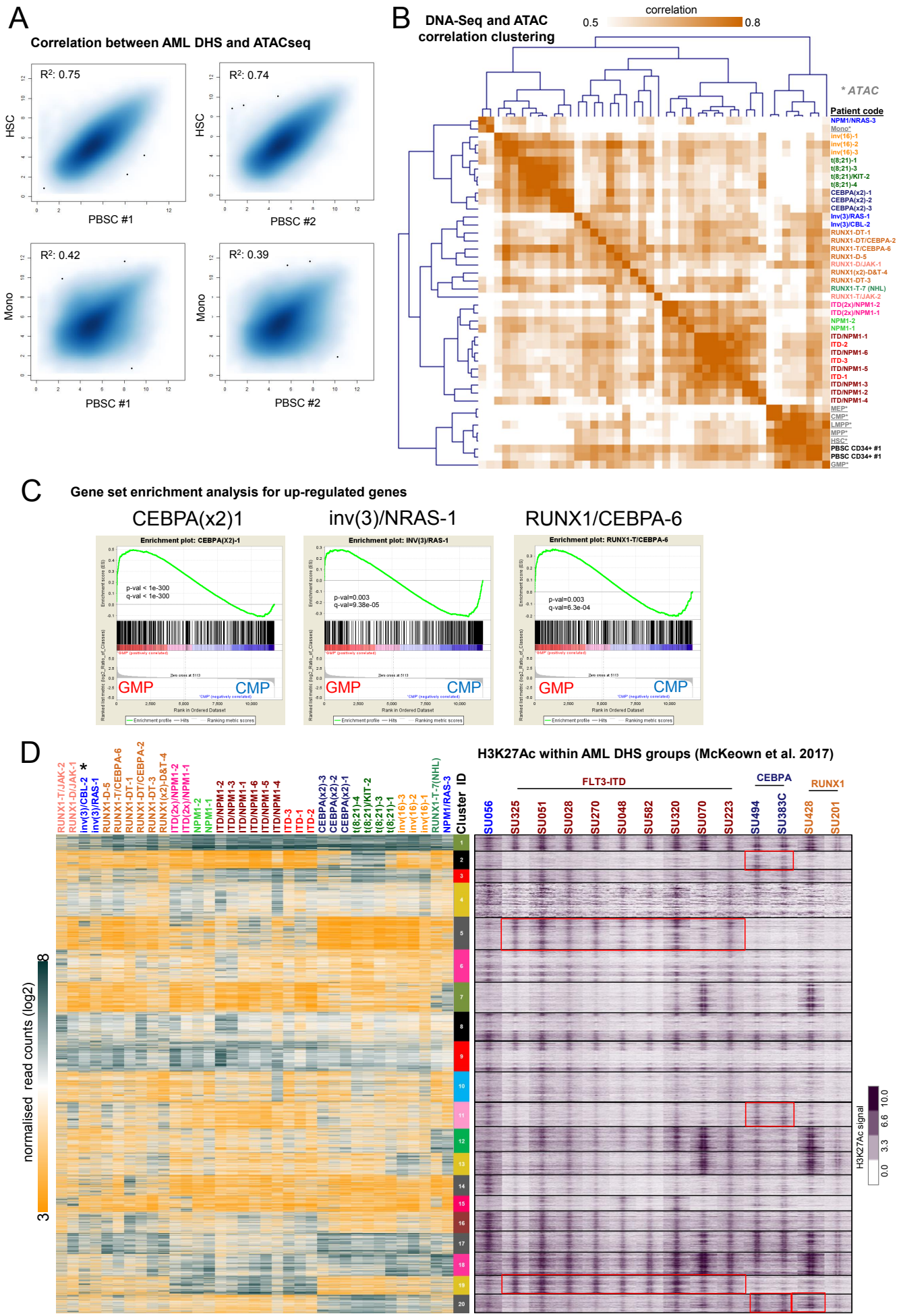
Supplementary Figure 2



Supplemental Figure 2: Different types of AML adopt unique transcriptome and chromatinlandscapes. (A) Hierarchical clustering of Pearson correlation coefficients of DNaseI

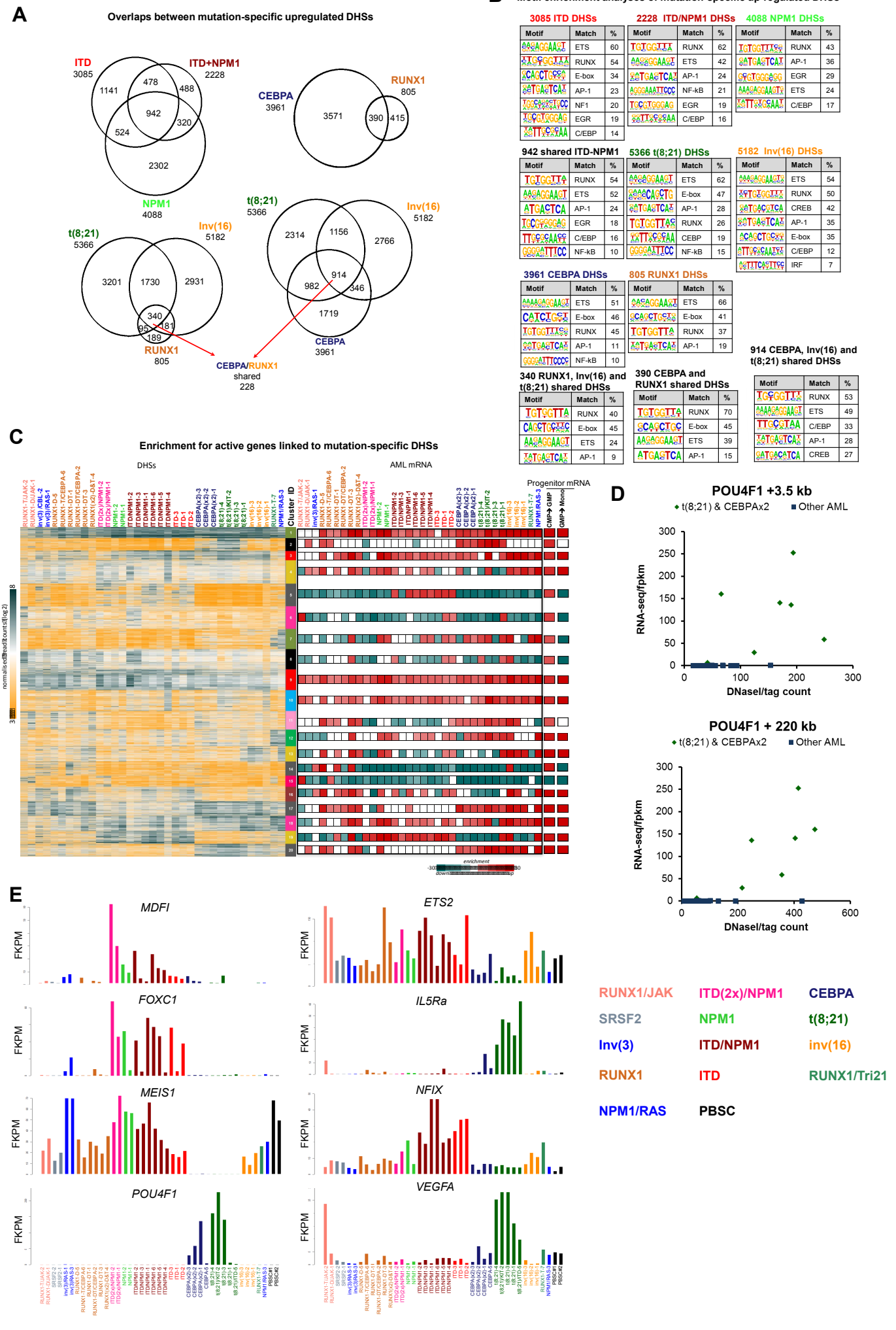
accessible sequences from all our patient samples with normalized read counts of DNase-Seq data for the different classes of mutations also including ATAC-Seq data from Corces et al., 2016 with similar mutations (SU(nnn), mostly FLT3-ITD). The mutation class is highlighted to the right of the panel and by a color code below the heatmap, again showing that specific elements from specific AML-types cluster together. Note the tight clustering of FLT3 and RAS mutant AML. (B) Scatter plots comparing the DNaseI tag count signals of patients with (11) and without (8) DNMT3 mutations against each other and against PBSCs as indicated by colored shapes. (C) Smooth scatter plots showing the correlation between DNase-Seq and RNA-Seq data from AML patients. Shown are CD34+ PBSC cells from individual #1 versus individual #2 (left plot), CD34+ PBSC from individual #2 versus a patient with NPM1 and NRAS mutation (right plot). RNA-Seq plots (top panel) and DNaseI-Seq plots (bottom panel). Other comparisons can be retrieved from the webserver. (D) Hierarchical clustering of log2 gene expression fold difference for all differentially transcription factor (TFs) and transcriptional regulator genes changing expression at least 2 fold in at least one patient as compared to normal CD34+ PBSC. Clustering was done only on rows (i.e., genes) while samples were ranked based on the clustering in Figure 1C. The heatmap colour is related to the degree of differential expression (fold-change (FC)). Red is up-regulated compared to normal CD34+ and blue is a down regulated TF.

Supplementary Figure 3



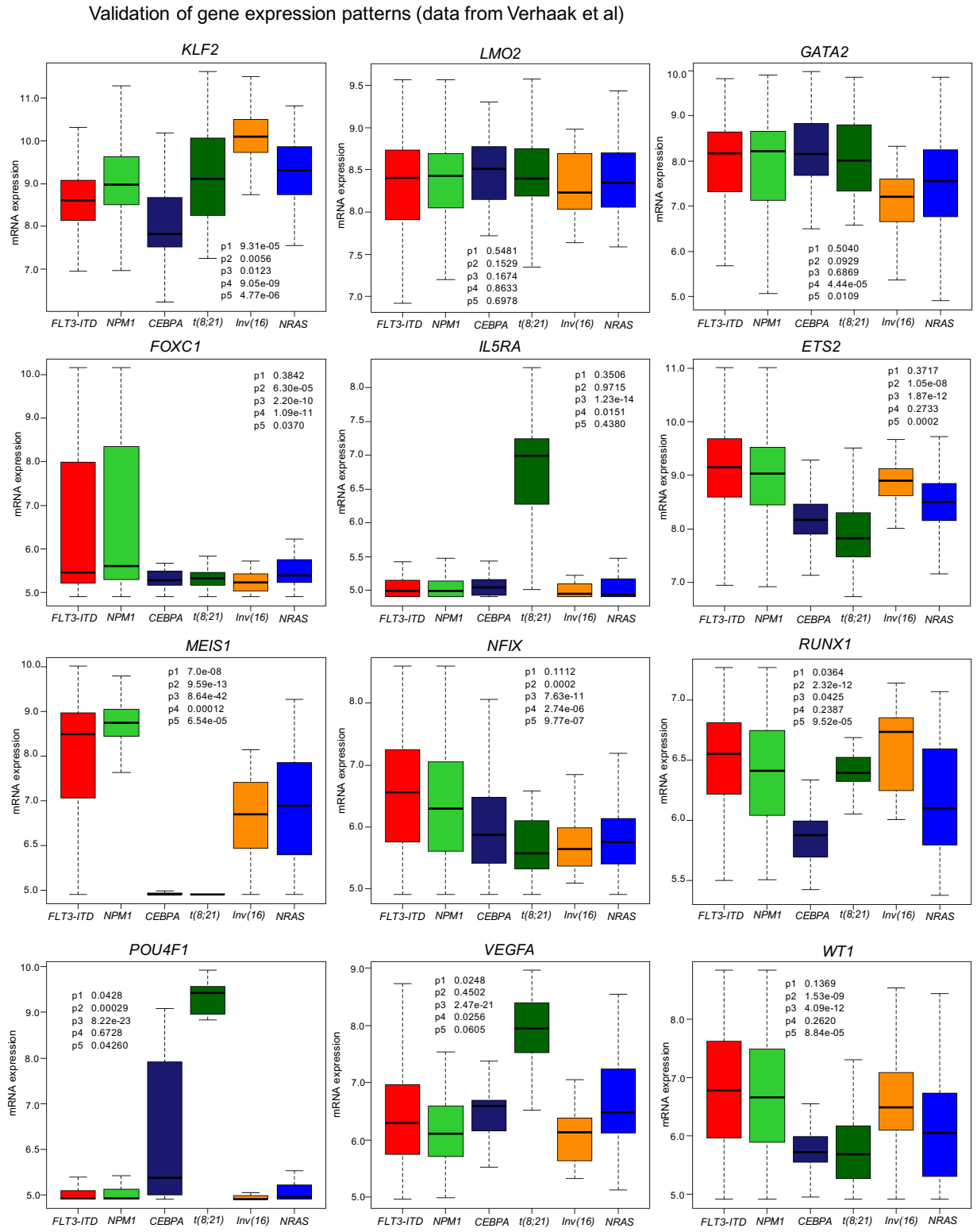
Supplemental Figure 3: Different types of AML are blocked at different stages of differentiation and correlation with publicly available data-sets. (A) Smooth scatter plots show the correlation between AML DNase-Seq and ATAC-Seq data. Top panel shows the DNaseI-Seq from normal CD34+ PBSC patient #1 & #2 versus the ATAC-Seq from Hematopoietic stem cells (HSC) and lower panel shows the DNaseI-Seq from normal CD34+ PBSC patient #1 & #2 versus the ATAC-Seq from Monocytes (Mono). (B) Hierarchical clustering of Pearson correlation coefficient between all patient samples of AML-Seq data plus the ATAC-seq data from Corces et al. The correlation between any two patients was obtained with normalized read counts calculated with +/- 200 bases from the peak center. (C) Gene set enrichment analysis for the up-regulated genes that are at least 2 fold difference compared to the normal CD34+. The AML up-regulated genes were tested for enrichment against the common myeloid progenitors (CMP) versus Granulocyte Macrophage Precursors (GMP) taken from Corces et al RNA-seq data. (D) Heatmap showing density enrichment of H3K27Ac peaks from McKeown et al., 2017 ranked according to the same coordinates of the DNase-Seq within the clusters (left heatmap), the H3K27Ac densities were plotted with a window size of +/- 2 kb around the DNaseI-Seq peaks summit. Selected AML-specific blocks of peaks are highlighted. The asterisk highlights samples inv(3)/CBL-2 for which RNA-Seq is available.

Supplementary Figure 4



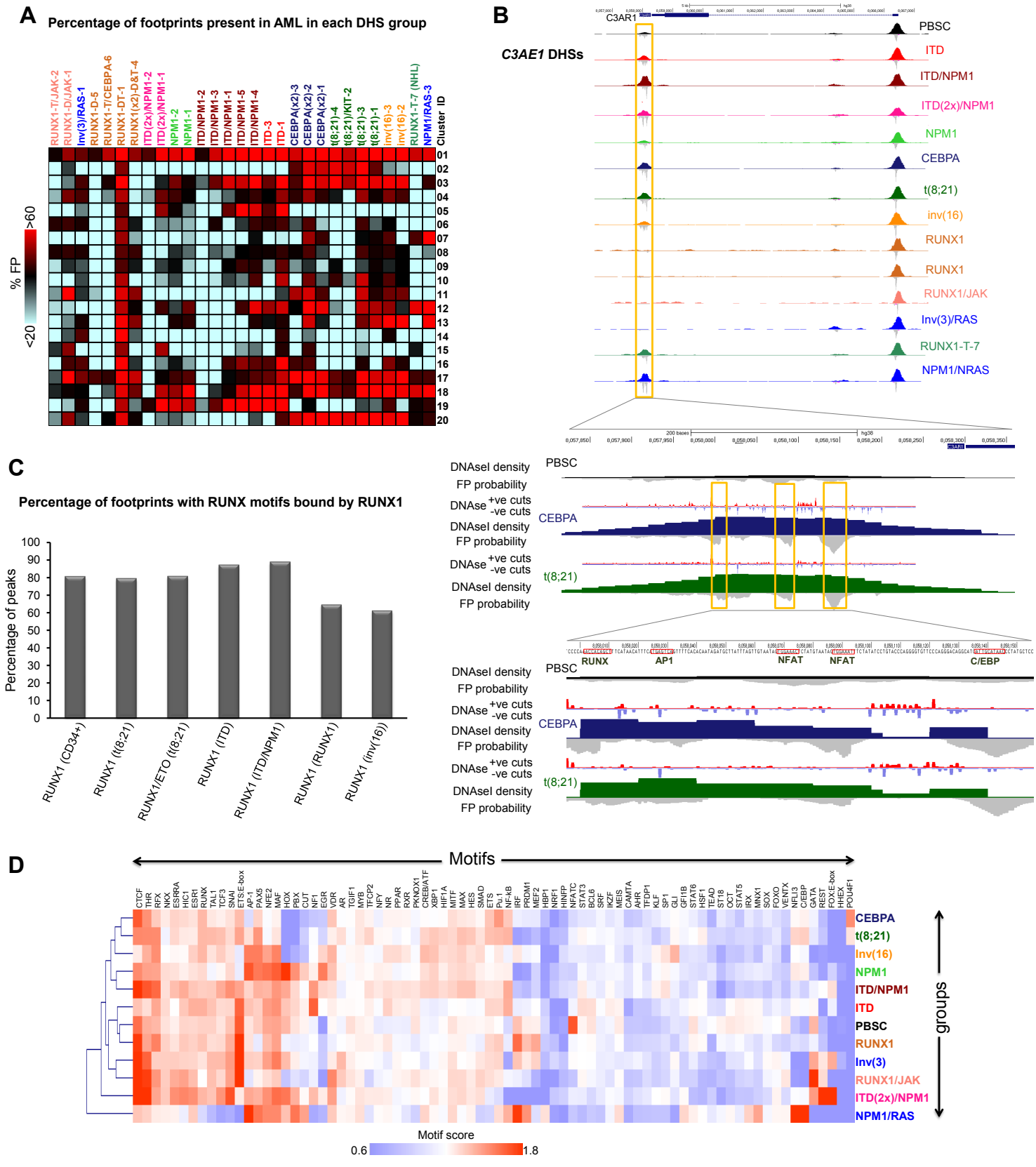
Supplemental Figure 4: AML-specifically active cis-regulatory elements cluster into common and unique chromatin landscapes and correlate with the upregulation of expression of the nearest genes. (A) Venn diagrams depicting the overlaps of subsets of DHSs which are up regulated compared to CD34+ve PBSCs within each of 7 mutation classes. These groups were generated as the average log2 values for 7 distinct subsets of AMLs that carried the same specific mutations in key regulators. These 7 mutation groups are defined on the basis of average values derived from 3 ITD patients, 6 ITD/NPM1 patients, 2 NPM1 patients, 4 t(8;21) patients, 3 inv(16) patients, 6 RUNX1 patients, and 3 patients with 2 CEBPA mutations. These groups are defined in Table S1 (note colour code). Up-regulated DHSs are defined as being at least 3-fold greater than in PBSCs, and have a DHS signal spanning a 400 bp window of at least 64. (B De novo motif search results using Homer for the up regulated DHSs classes and in overlapping deregulated DHSs for ITD and/or NPM1 and for CEBPA and RUNX1 that are shown in (A) the numbers indicate of percentage of each subset that contains the identified motif. (C) Gene set enrichment analysis for all expressed genes that are annotated to the DHSs identified in each of the AML specific 20 clusters, the enrichment scores (right panel) are aligned against each of the 20 clusters (left panel) that was initially described in Figure 3A. The target genes were tested for enrichment against all AML RNA-seq data; red color indicates that these genes are enriched with up-regulated genes compared to CD34+ PBSC. (D) Correlation of gene expression with DNaseI tag count as exemplified for the *POU4F1* gene (see also S1 B,D). (E) Bar figures depicting the expression level for some of the targets differentially expressed genes, the FPKM values were plotted on the y-axis for each AML samples used in this study, the color code identified each of the mutation classes.

Supplementary Figure 5



Supplemental Figure 5: Common and group-specific DHS associate with genes belonging to different functional groups. Boxplots validating gene expression patterns for some of the differentially expressed genes genes using gene expression data from Verhaak et al (2009). P-

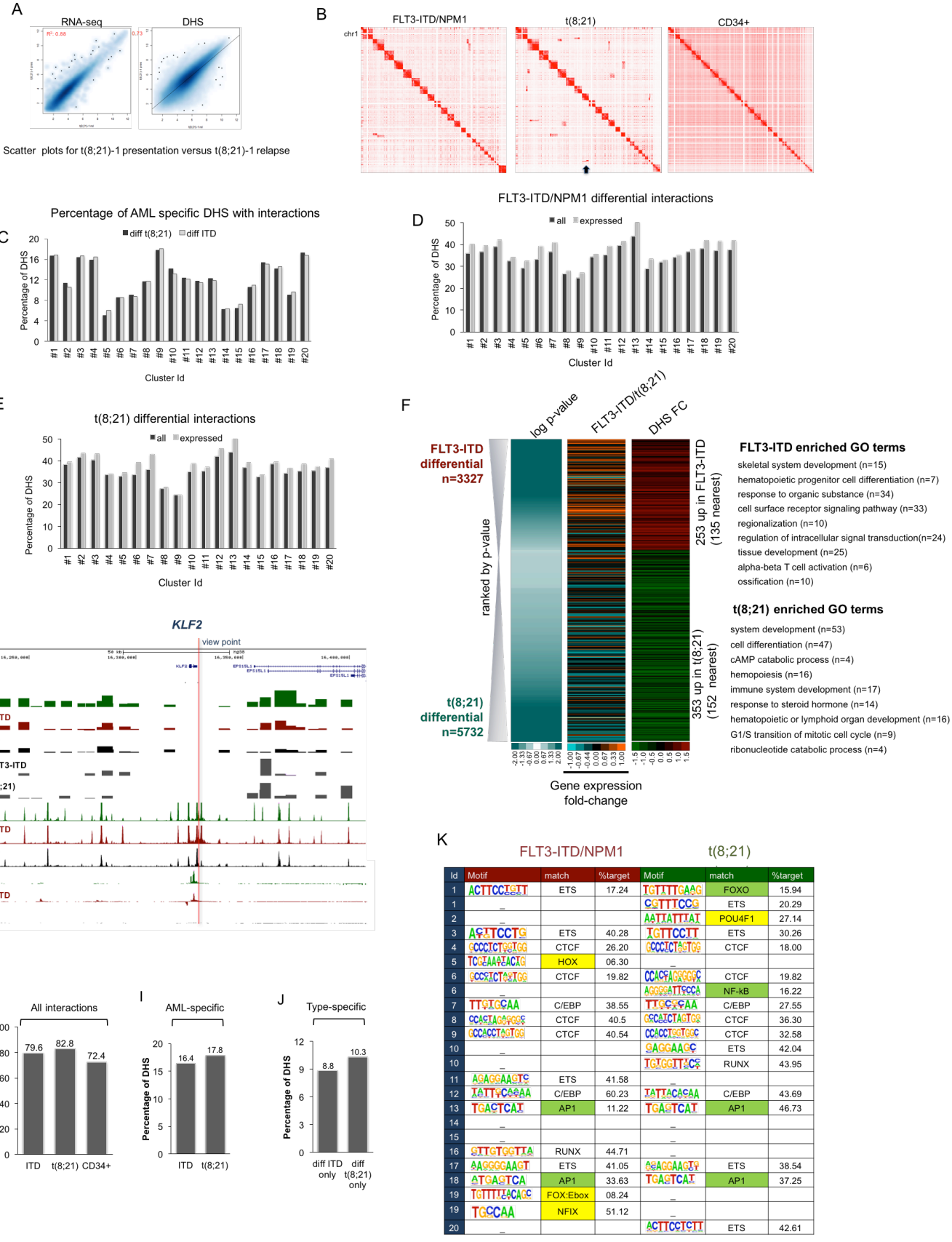
values highlighting the significance of differences are shown on each panel; the t-test was used to calculate the p-values.



Supplemental Figure 6: AML-specifically active cis-regulatory elements are characterized by specific transcription factor binding patterns. (A) Percentage of the footprints in the AML specific DHSs for the 20 clusters identified in Figure 3A. The footprints were identified using the Wellington algorithm. We first identified differential footprints for each AML sample compared to

the CD34+ PBSCs and then the percentage of these differential footprints in the DHS subsets in the 20 clusters was calculated. (B) UCSC genome browser screenshot of DNAseI-Seq data aligned with digital footprints at the *C3AE1* locus. The screenshot shows the DHSs for one patient from each group. Footprint probabilities as calculated by Wellington (Piper et al., 2015) are indicated as grey density below the lines. The bottom indicates the precise location of occupied RUNX, C/EBP and AP-1 footprints. (C) Percentage of footprints with RUNX motifs in the indicated AML-types peaks which are bound by RUNX1 or RUNX1-ETO in ChIP assays from [22,33,66]. (D) Heatmap depicting the degree of motif enrichment after hierarchical clustering of all (not just the specific) motif enrichments for each of the mutation-specific AML groups. Enrichment scores were calculated by the level of motif enrichment in all the footprints of all Hi-read depth samples for each group, as compared to the union of footprints in all experiments.

Supplementary Figure 7

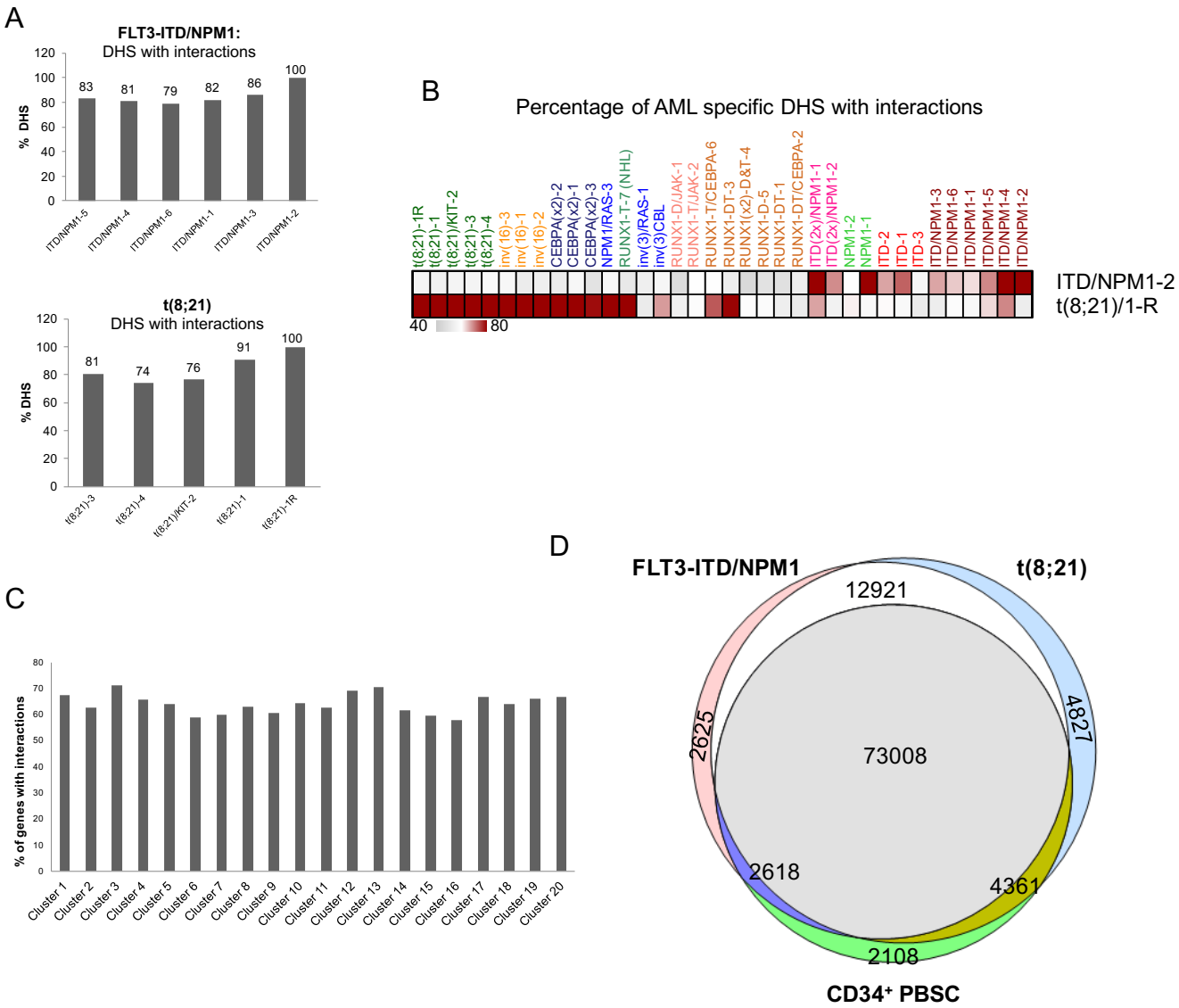


Id	Motif	match	%target	Motif	match	%target
1		ETS	17.24		FOXO	15.94
1					ETS	20.29
2					POU4F1	27.14
3		ETS	40.28		ETS	30.26
4		CTCF	26.20		CTCF	18.00
5		HOX	06.30			
6		CTCF	19.82		CTCF	19.82
6					NF-kB	16.22
7		C/EBP	38.55		C/EBP	27.55
8		CTCF	40.5		CTCF	36.30
9		CTCF	40.54		CTCF	32.58
10					ETS	42.04
10					RUNX	43.95
11		ETS	41.58			
12		C/EBP	60.23		C/EBP	43.69
13		AP1	11.22		AP1	46.73
14						
15						
16		RUNX	44.71			
17		ETS	41.05		ETS	38.54
18		AP1	33.63		AP1	37.25
19		FOX:Ebox	08.24			
19		NFIX	51.12			
20					ETS	42.61

Supplemental Figure 7: Capture HiC shows differences in cis-regulatory interactions between different types of AML and normal cells. (A) Smooth scatter plots show the correlation

between t(8;21)-1 presentation and t(8;21)-1 relapse AML DNAseI-Seq data. (B) Heatmaps show the raw overall inter and intra interactions of the promoter capture HiC data for all chromosomes for FLT3-ITD (ITD/NPM1-2, left), t(8;21) (middle) and CD34+ (right) across all chromosomes. (C) Bar figure showing the percentage of DHSs within each of the 20 clusters identified in Figure 3A that have differential interactions compared to CD34+, (D, E) percentage of DHSs within each of the 20 clusters interacting with the nearest gene within differential interactions for all genes expressed genes as identified by the RNA-Seq data. D: FLT3-ITD and E: t(8;21). (F) Heatmap of differential interactions ranked by the strength of interaction (-log p-value) from highly significant to less for the FLT3-ITD and from less significant to more significant for the t(8;21) (left panel) Plotted along-side are the gene expression fold-difference for the FLT3-ITD compared to the t(8;21) and for the FLT3-ITD (middle panel) and the DHS fold difference FLT3-ITD versus the t(8;21) (right panel). (F) The top enriched GO terms for the up regulated genes in the FLT3-ITD compared the t(8;21) where the DHSs differentially interact with the promoter of that gene. Similarly the bottom panel the top enriched GO terms for the up regulated genes in the t(8;21) compared to the FLT3-ITD such that the DHS is differentially interact with the promoter of that gene.. (G) UCSC genome browser showing a screenshot of *KLF2*. The top two tracks display the log p-value of the capture HiC interaction for *KLF2* promoter as viewpoint, the following two tracks display log p-value of the differential interaction of the t(8;21) and the FLT3-ITD compared to the CD34+. Shown are also the DNaseI-Seq and RNA-Seq data of t(8;21), FLT3-ITD and CD34+ PBSC. (H) Bar diagram showing the percentage of DHSs involved in significant interactions. (I) Bar diagram showing the percentage of DHSs involved in significant differential interactions compared to CD34+ cells. (J) Bar diagram showing the percentage of DHSs involved in significant differential interactions for DHSs unique to FLT3-ITD or t(8;21) DHSs compared to CD34+ cells, with DHS common to FLT3-ITD and t(8;21) being excluded. (K) Enriched footprinted motifs in DHS associated each of the 20 clusters involved in differential interactions for the two patients. Motifs for transcription factors normally not expressed in myeloid cells are highlighted in yellow, motifs for inducible factors are marked in green.

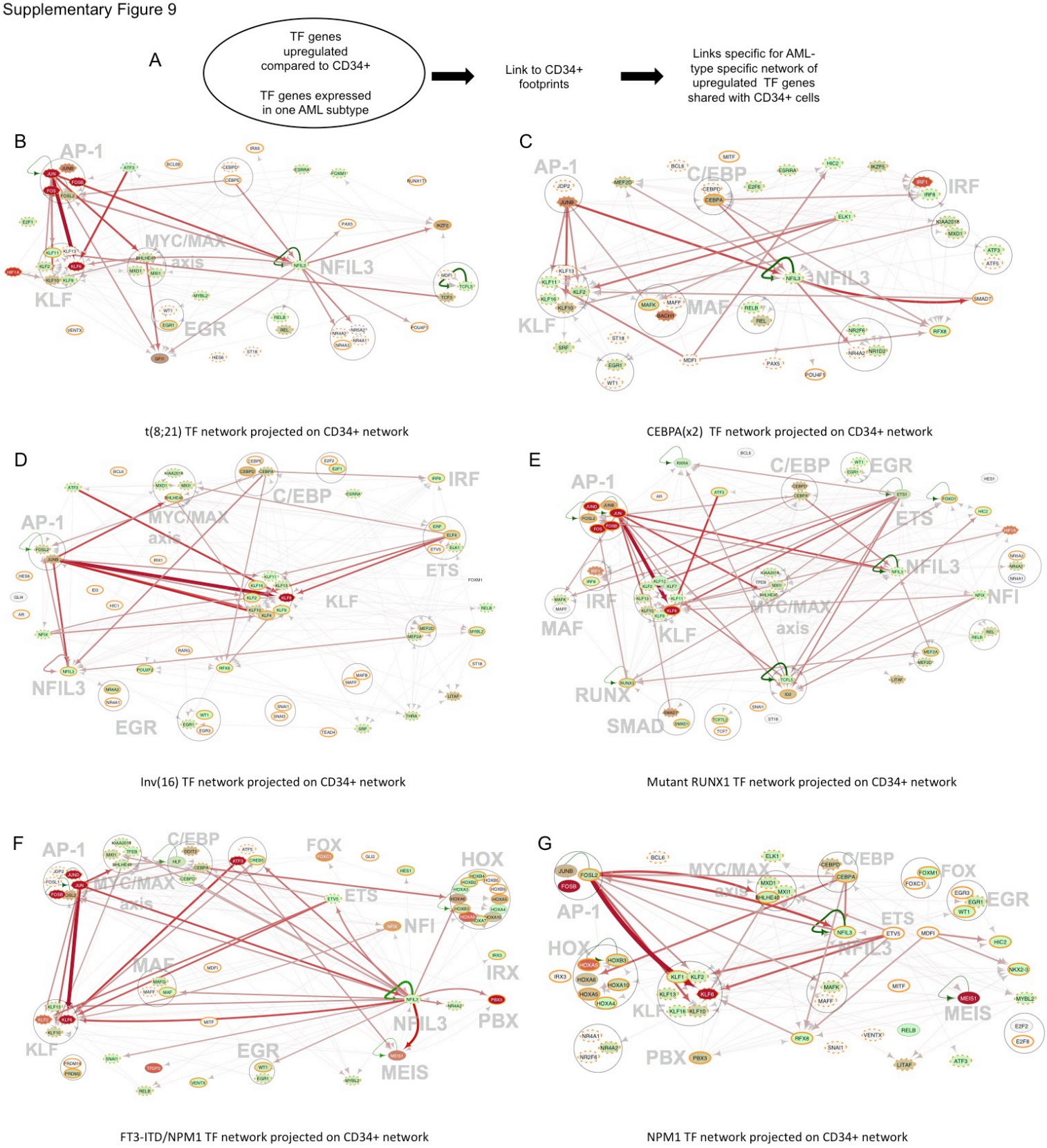
Supplemental Figure 8



Supplemental Figure 8: Interactions are representative for their patient groups and the majority of interactions are shared (A)

(A) Percentage of all DHSs with interactions present in each dataset of each individual patients, (B) Heat-map highlighting the percentage of AML-type specific DHS with interactions found in the different patient groups, indicating that the patient chosen for the Chi-C experiment are representative for each patient group. (C) Percentage of up-regulated gens associated with DHS clusters that have significant interactions in any of the three Chi-C experiments. (D) Overlap of all DHSs underlying interactions in all three samples as indicated demonstrating that the majority of interactions are the same in all three samples.

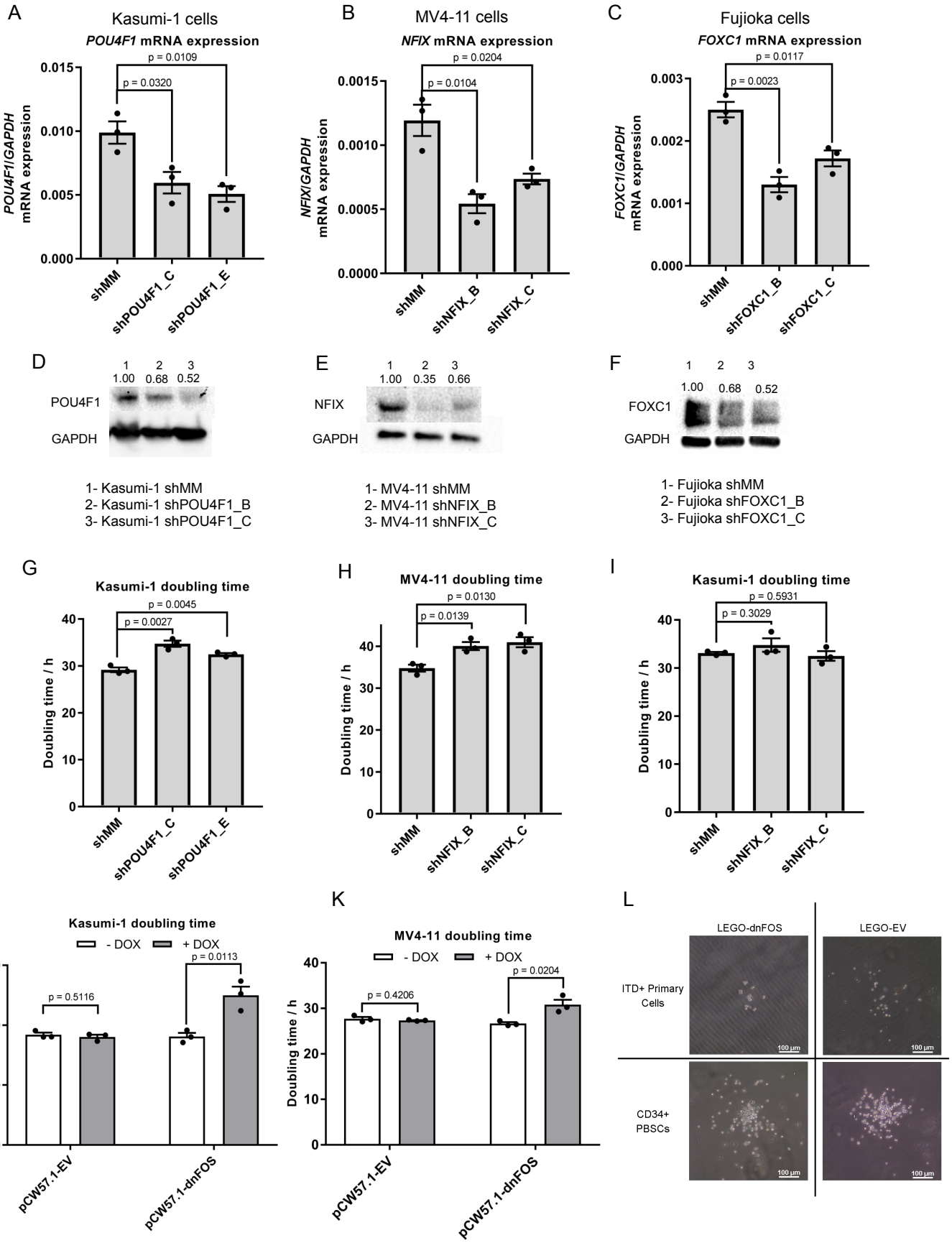
Supplementary Figure 9



Supplemental Figure 9: Identification of transcription factor network components driving the expression of TF genes in each AML subtype which are shared with CD34+ cells. Here we projected the links from the indicated AML subtypes onto the CD34+ footprints. Top panels: Analysis strategy. (A) Shared t(8;21) TF network, (B) Shared CEBPA(x2) TF, (C) Shared Inv(16) TF network, (D) Shared Mutant RUNX1 TF network, E) Shared FLT3-ITD/NPM1 TF network, (F) Shared NPM1- TF network. Factor families binding to the same motif as shown in Table S2 form a

node contained within a circle. Arrows going outwards from the entire node highlight footprinted motifs in individual genes generated by any member of this factor family whereby the footprint was annotated to the gene using the Chi-C data where possible, otherwise to the nearest gene. The expression level (FKPM) for the individual genes is depicted in white (low)/red (high) colour. An orange smooth ring around the circle indicates that this gene is specifically up-regulated in this type of AML compared to CD34+ PBSCs and/or other AML types, a dotted circle indicates a gene that is up-regulated as compared to CD34+ cells. Genes with no outgoing arrows due to a lack of know binding motifs are highlighted by an octagon shape. For a detailed guide to node and edge attributes: See legend of Figure 6.

Supplemental Figure 10



Supplemental Figure 10

AML type-specifically expressed transcription factors are required for leukemic growth. (A, B, C) Histograms showing *POU4F1* (A)*, FOXC1* (B) and *NFIX* (C) mRNA expression after transduction with the indicated shRNA and control lentiviruses in Kasumi-1, MV4-11 and Fujijoka cell lines, respectively. Note that Fujijoka cells express high levels of *FOXC1* and were only used to test the functionality of our lentiviral construct. *FOXC1* is not highly expressed in MV4-11 cells. (D-F) Western Blots showing the efficiencies of shRNA knock-down for FOXC1 (D), NFIX (E) and POU4F1 (F). (G - I) Histogram showing doubling time of t(8;21) Kasumi-1 cells after transduction with sh*POU4F1* (G)*,* MV4-11 cells after transduction with *shNFIX* (H) and of Kasumi-1 cells after transduction with *shNFIX* (I). (J, K): doubling times of Kasumi-1 (J) and MV4-11 cells (K) expressing a DOX inducible version of a dominant negative FOS peptide (dnFOS) (K,M) as well as empty control virus (L,O). All experiments were performed in triplicate. In all histograms n=3 * $p<0.05$, ** $p<0.01$, *** $p<0.001$. Error bars show 95% confidence intervals. (L) Pictures of representative colonies derived from FLT3-ITD patient cells and CD34+ PBSCs transduced with the indicated lentiviral vectors.

List of supplemental data files

Dataset S1: Summary of all AML mutation data

Dataset S2: Up and down-regulated genes associated with mutation groups (related to Figure 3 and Figure S4B)

Dataset S3: Number of differentially expressed genes for Figure 2C and Figure S4C

Dataset S4: List of transcriptional regulator genes showing AML type-specific expression (related to Figure S2C)

Dataset S5: Gene lists and GO terms for Figure 5 and Figure S7F.

Dataset S6: CHi-C-curated KEGG pathways and GO terms of DHS-cluster associated genes (related to Figure 3)

Supplemental Notes with five Figures