Quantification of differential transcription factor activity and multiomic-based classification into activators and repressors: *diffTF*

Ivan Berest[1,*], Christian Arnold[1,*], Armando Reyes-Palomares[1], Giovanni Palla[1], Kasper Dindler Rasmussen[2,3], Kristian Helin[2,3] & Judith B. Zaugg[1]

[1] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg

[2] Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen

[3] Novo Nordisk Foundation Center for Stem Cell Biology, Copenhagen

[*] equal contribution

Transcription factor (TF) activity constitutes an important readout of cellular signalling pathways and thus for assessing regulatory differences across conditions. However, current technologies lack the ability to simultaneously assess activity changes for multiple TFs and in particular to determine whether a specific TF acts as repressor or activator. To this end, we introduce a widely applicable genome-wide method *diffTF* to assess differential TF binding activity and classifying TFs as activator or repressor by integrating any type of genome-wide chromatin with RNA-Seq data and in-silico predicted TF binding sites (available at https://git.embl.de/grp-zaugg/diffTF). We apply *diffTF* to a large ATAC-Seq dataset of mutated and unmutated chronic lymphocytic leukemia and identify dozens of TFs that are differentially active. Around 40% of them have a previously described association with CLL while ~60% constitute potentially novel TFs driving the different CLL subtypes. Finally, we validated the method experimentally using the well studied system of hematopoietic differentiation in mouse.

INTRODUCTION

Transcription factors (TFs) are important for orchestrating coordinated and dynamic responses to intra- and extracellular stimuli and regulating a multitude of biological processes. Indeed, since many signaling cascades end in the activation of a particular set of TFs, observing a change in overall TF activity can serve as a good readout of signaling pathways that regulate them[1]. However, even though many methods exist to investigate particular aspects of TF biology, none of them is able to globally measure the "activity" of all TFs. In addition, surprisingly little is known about whether particular TFs act as a transcriptional activator or repressor, despite some efforts that have been made by collecting published studies using text mining techniques[2]. TFs are typically lowly abundant proteins, which makes it difficult to detect them in proteomics experiments[1,3], and even if they can be detected, their abundance and TF activity do not necessarily correspond since TFs are highly regulated at the post-translational level. On the other hand, chromatin immunoprecipitation followed by sequencing (ChIP-Seq), which is the gold-standard technique for measuring the genomic TF binding events, provides information only for one TF at the time and it does not detect global changes in TF activity across conditions or cell types unless specific experimental normalisation methods are used (e.g. spike-ins[4]). Finally, luciferase assays only measure the activity of a specific TF at a specific location and are therefore fairly low throughput.

To correctly interpret the downstream effects of changes in TFs binding, it is crucial to know whether they act as transcriptional activator or repressor for the specific biological environment. Databases like TRRUST[2,5] collect annotations of TF modes' of regulation based on literature mining and provide a great resource for well-studied TF-target interactions. However, for the vast majority of TFs, there is no consensus about their molecular function, and most TFs are annotated as both transcriptional activators and repressors, depending on the target gene, cell type and growth condition. A general framework for classifying TFs based on their molecular mode of action in a cell-type and condition-specific manner is currently lacking.

Towards closing these gaps we have developed an approach called *diffTF* to integrate RNA-Seq and genome-wide chromatin accessibility or histone mark ChIP-Seq data to (i) estimate global changes in TF activities across conditions or cell types, and (ii) to classify TFs based on their molecular role into transcriptional activators and repressors. Applying this approach to a large-scale ATAC-Seq dataset of chronic lymphocytic leukemia (CLL) we first showed that *diffTF* is highly robust with respect to wide range of parameter settings, and that it is capable of

identifying many known TFs involved in CLL. In addition, we report many novel candidate TFs that have never been shown before to be associated with CLL. Furthermore, our classification into activators and repressors allowed us to specifically investigate the effects of transcriptional activators and repressors on open chromatin and derive general principles of regulatory modes. We have experimentally validated *diffTF* using the well characterized system of hematopoietic stem cell differentiation tree. Finally, the approach has already been successfully applied to identify TFs that are specifically associated with TET2, an enzyme involved in DNA demethylation[6].

RESULTS

Conceptual derivation of using open chromatin as read-out of differential TF activity

We define change in TF "activity" as changes in open or active chromatin as measured by ATAC-Seq, DNase-Seq or ChIP-Seq for H3K27ac or any other chromatin mark that is known to be affected by TFs. This definition is based on earlier work where we showed that genetic variants affecting H3K27ac signal across individuals (hQTLs) can be explained by disruptions of TF motifs whenever the hQTL-SNP overlaps with a TF binding site (TFBS)[7]. Even though the exact mechanisms of how changes in TF affinity translate to the chromatin level are unknown, TF binding likely plays a causal role in mediating the effect of the DNA variant onto chromatin marks[8]. By reversing this argument, we propose to use the aggregate changes in chromatin accessibility in the vicinity of putative binding sites of a TF as a readout for its change in "activity" (**Fig. 1a**).

Based on this concept, we developed *diffTF,* which is a computational approach to globally assess differential TF activity between two conditions (**Suppl. Fig. 1**). It is based on (i) any data that measures active/open chromatin, (ii) putative binding sites for TFs of interest, and optionally (iii) matched RNA-Seq data, which are used to classify TFs into transcriptional activators or repressors. Briefly, it requires *in silico* TFBS that can be obtained using position weight matrices (PWMs) from a database (e.g. *HOCOMOCO*[9]) and a PWM scanning algorithm (e.g. *PWMScan*[10]) for all TFs, or from a database of ChIP-Seq data, such as *ReMap*[11]. For each TFBS it then calculates the difference between two conditions and summarizes their change in accessibility across all binding sites of a given TF. In this step it also normalises for GC content of the respective TFBS (**Suppl. Fig. 2**). Finally, the significance is assessed using an empirical approach by permuting the sample labels or in an analytical way if the number of samples is too

low (see Methods and **Suppl. Fig. 3**). As a final result, *diffTF* outputs a differential activity value along with an empirical FDR for each TF, which allows to identify the set of TFs that show a significantly higher "activity" in either of the conditions (**Fig. 1b**).

Applying *diffTF* to a large ATAC-Seq and RNA-Seq dataset in chronic lymphocytic leukemia (CLL)

Chronic lymphocytic leukemia (CLL) is one of the most frequent common type of cancer of the Western world, particularly among old people. There are two major subtypes of CLL, which are defined based on the mutation status of the IgHV locus (mutated: M-CLL and unmutated: U-CLL). Overall, U-CLL associates with worse clinical outcomes such as shorter survival time and higher frequency of relapse after treatment[12]. In M-CLL, B-cells go through normal affinity maturation. With the aid of T-helper lymphocites, they undergo multiple rounds of IgHV somatic hypermutation to produce high affinity B-cell receptors (BCR), which are however clonally restricted. This process is essential for their development, survival and growth[13]. In contrast, U-CLL B-cells reach their affinity maturation point in an unregulated manner, and without involvement of T-helpers[14]. Recent studies have shown that the IgHV mutation status can be well predicted based on chromatin profiles[15].

Given the vast knowledge about CLL, we thought this is an ideal case to apply *diffTF* to (i) assess its power to recover relevant biological signal, (ii) gain more specific regulatory insights into the different subtypes of CLL and (iii) assess its technical robustness. We therefore obtained publicly available ATAC-Seq data for a cohort of CLL patients characterised by the mutation status of the IgHV locus[15]. ATAC-Seq was performed in lymphocyte samples derived from 52 CLL patients (27 mutated vs. 25 unmutated)[15]. For a subset of the patients we also have RNA-Seq data available[15].

TF activity differences between U-CLL and B-CLL are highly robust within a wide range of parameter settings

After extensive data processing and quality control (**Suppl. Fig. 4-6**), we applied *diffTF* to identify the differences in TF activities between U-CLL and M-CLL. In total we identified 67 TFs that are differentially active (FDR < 10%) between the two subtypes (**Fig. 2a**; **Suppl. Table 1**). Before going into the biological interpretation of the results we used this dataset to assess the technical robustness of *diffTF* with respect to a reasonable range of parameter settings.

Predicting putative TF binding sites is inherently noisy and it may result in many false positive predictions when compared to a gold standard such as ChIP-seq experiment for a given TF[16]. To address this point, we assessed the dependency of the *diffTF* results on the binding site predictions. We run *diffTF* only on binding sites that were bound by the TF based on ChIP-Seq data using all TFs comprised in *ReMap* and observed a very strong correlation of the resulting TF activity differences (R=0.81; **Fig. 2b**, **Suppl. Fig. 7**). This indicates that *diffTF* is robust with respect to false positive binding sites. Similarly, we found that neither the p-value cut-offs for motif scanning (**Suppl. Fig. 8**) nor the motif database (*JASPAR* vs. *HOCOMOCO*) had a strong impact on the differential TF activity (R=0.87, R=0.62 and R=0.69, respectively, **Fig. 2c-e**). Varying *diffTF* parameters such as the region size used for extracting the signal did not affect the resulting differential TF activities (see Supplement and **Suppl. Fig. 9**).

Next we sought to assess the power of *diffTF* to detect differential TF activities with respect to the level of accessibility differences across conditions. For this, we run it on a reduced set of peaks from which we removed all regions that were differentially accessible at a 5% FDR (see Methods) and found that the results were very similar for both sets (R=0.89). This demonstrates that *diffTF* can detect differences even in low signal data that have little or no differentially accessible peaks between conditions (**Fig. 2f**).

Finally, we assessed the power of *diffTF* with respect to sample size and sequencing depth. Intriguingly, we found the results highly congruent across a wide range of samples sizes and sequencing depths, with the majority of significant TFs in the full data showing the same direction of change in the subsampled data (**Fig. 2g**, **Suppl. Fig. 10**). Generally, the number of samples appears more important than read depth, and results using the full set were consistent for a coverage as low as 1-5 million processed reads per sample (see Methods). Although these results are dataset-specific and difficult to generalize they provide guidelines for the applicability of *diffTF*.

In summary, these results establish the robustness of *diffTF* in quantifying differences of TF activities, and demonstrate that aggregating signal across all binding sites is a powerful and sensible approach to overcome limitations such as low coverage and little underlying biological signal.

146

diffTF recovers known CLL TFs and proposes many novel candidates

147 After establishing the technical robustness of *diffTF* for identifying differentially active TFs, we
148 next focused on the biological interpretation of the 67 TFs that showed significant differential
149 activity at FDR 10% between M-CLL and U-CLL patients (**Fig. 2a**; **Suppl. Table 1**). Intriguingly,
150 about ~41% of the differentially active TFs have previously been associated with CLL and
151 mostly (90%) agree with the reported condition (i.e., mutated or unmutated; **Suppl. Table 2**),
152 thus providing a strong biological validation of the approach. The remaining 50% may represent
153 novel candidate TFs that can advance our understanding of the disease etiology in general and
154 the biological differences of mutated and unmutated patients in particular.
155 Since *diffTF* is based on aggregating signal across TF binding sites, which in turn are identified
156 by scanning for TF motifs, it is not able to distinguish between TFs with highly similar motifs.
157 Therefore, we used the PWM clustering tool Rsat[17] to group TFs into clusters to allow proper
158 interpretation of the results (clusters are available here https://bit.ly/2J9TaaK). Reassuringly, TF
159 clusters are overall consistent in terms of TF activity changes (**Fig. 3a**). Exceptions include
160 cluster 17 which can be explained by a prominent split into two branches early in the clustering
161 into NFAT and NFκB factors, which show more activity in U-CLL and M-CLL, respectively (**Fig.**
162 **3c**).
163 The most active TF cluster in U-CLL is the IRF family and STAT2 (cluster 40; **Fig. 3d**), which is
164 known to regulate cell-differentiation and induce apoptosis[18,19]. Several CLL susceptibility loci
165 have been described proximal to IRF genes, and IRF expression has been associated with
166 disease onset and progression[20–22]. The next most active cluster is the PAX TFs (cluster 54).
167 One of its members, PAX5, together with PRDM1 (IRF cluster), affects B-cell to plasma cell
168 differentiation, for which a disrupted regulation can lead to decreased cell survival and poor
169 patient prognosis[23]. Another prominent set of regulators are the members of the AP-1 complex
170 (cluster 4), which increase proliferation and have recently been shown to play an important role
171 in driving the invasive nature of U-CLL[24]. Other prominent CLL-associated TFs that were
172 identified with *diffTF* are c-MYC and CEBPB, which are involved in cell-proliferation and
173 differentiation. Specifically, high c-MYC mRNA levels have been detected in U-CLL[25,26]. Several
174 TFs with increased activity in U-CLL have been associated with massive U-CLL linked genomic
175 aberrations such as BCL11A (having a (2;14)(p16;q32) translocation[27]). Overall, for U-CLL, we

observe an increased activity for TFs that induce and increase apoptosis and have been reported to be dysregulated in CLL.

In contrast, for M-CLL we identified TFs that regulate and possibly reduce apoptosis, regulate cell cycle and suggest normal functionality of B-cells through the classical BCR, NF-kB and Wnt signaling pathways. The most active TF family in this set of patients are the POU TFs, also known as Oct factors (cluster 12; **Fig 3b**), which regulate normal B-cell development and immunoglobulin production, therefore promoting survival of the lymphoma cells[28]. The POU cluster is followed by the ROR factors (cluster 36), which together with Wnt5a activates NF-kB-dependent survival signaling in CLL[29], and the GATA family (cluster 16), which is known to prime HSCs towards the lymphoid lineage and increase self-renewal of the stem cells in CLL[30]. Other examples include the EGR family, for which motifs have been found to be enriched in aberrant hypomethylated CpG sites in CLL[31]. Interestingly, we also found PPARD, which has recently been linked to M-CLL through its effect on metabolic pathways in cancer cells[32]. We also found members of the GLI family significantly more active in M-CLL. These are part of the Hedgehog signaling pathway and regulate apoptosis, thereby supporting survival of M-CLL cells[33].

Among the novel factors associated with U-CLL, we found several TFs involved in the circadian clock, such as BMAL1 and CLOCK, which form a complex, as well as their repressor NR1D1, the regulation of which has recently been proposed as another hallmark of cancer[34]. Moreover, we found other members of the basic helix-Loop-Helix family, such as, BHE40, a regulator of mitotic division[35], which has recently been found to be essential for development of B1-a cells[36] and TFAP4, TFE3 and TFEB, for which there are known cases of gene-fusions in renal cell carcinoma[37]. Another set of novel TFs that we observed to be more active in M-CLL were associated with cancer- and B-cell relevant pathways such as escape from apoptosis (ZN784)[38], regulation of cell cycle progression (ZBTB6)[39], selection of B-cells and promotion of fetal B lymphopoesis (ARID3A)[40], and control of the B cell-mediated autoimmunity[41,42]. The GFI1 family (cluster 35) is less active in U-CLL and their expression and activation might influence and decrease rates of apoptosis in B-cells[43].

In summary, these results show that *diffTF* is able to recapitulate much of the known biology of the two subtypes of CLL and additionally identifies numerous additional factors that are likely to be implicated in the disease.

Determination of the molecular function of TFs: transcriptional repressors and activators

The paragraphs above have shown that *diffTF* can identify TFs that alter their activity across different types of CLL patients. However, to gain mechanistic insights into some of the regulatory differences between U-CLL and M-CLL, it is crucial to know whether a TF acts as activator, in which case a higher abundance would generally result in increased transcription of its target genes, or repressor, in which case an increase in abundance would be paralleled by decreased target gene transcription. Surprisingly little is known about whether a certain TF acts mostly as transcriptional repressor or activator, and based on literature text mining most TFs have been annotated multiple times as both activator and repressor[5], indicating that the cell type or other external factors are important in determining a TF's main mode of action (**Suppl. Fig. 11B**). Therefore, we have devised a cell-type specific and data-driven, multiomics approach within the framework of *diffTF* to classify TFs into activators and repressors based on RNA-Seq, ATAC-Seq and putative TFBS.

The rationale behind this classification is that increasing levels of an activator TF should result in more open, active chromatin at its targets sites and increased transcription of its target genes and vice versa for repressors. With an increase in abundance of an activator TF we therefore expect an increase in chromatin accessibility surrounding its binding sites while the chromatin surrounding binding sites of a repressing TF we would expect to close (**Fig. 4a**). To test this rationale, we investigated the ATAC-Seq profile surrounding binding sites of a well-known repressor (REST) and activator (STAT2) that are both expressed in our cell type (**Fig. 4b**). We observed that for REST, the accessibility drops to genome-wide average (within peak areas) within 10 bp from the center of the core motif, while only a small region at the binding site is slightly accessible, likely reflecting the binding of the factor itself. In contrast, for STAT2, we observed increased chromatin accessibility also outside the core binding site, which drops to genome-wide average over a distance of >100 bp from the center of the motif, likely representing the effect of the TF on opening chromatin. This indicates that, while there is a similar pattern of increase in accessibility at the exact binding site of a TF, the surrounding chromatin is highly compact for the repressor while it is open for the activator, thus supporting our rationale that repressors overall decrease chromatin accessibility around their binding sites.

To classify TFs into activators and repressors *diffTF* uses RNA-Seq data as an estimate for TF abundance. For each TF it then calculates the correlations across individuals between a TF's expression level and the ATAC-Seq signal in its putative target peak. A TF gets classified as activator when it shows an overall positive correlation with the ATAC-signal at its putative target sites, as repressor if it is overall negative, and as undetermined if the distribution of correlations is not significantly different from the peaks that did not overlap its putative binding sites (examples in **Fig 4c**, see also Methods and **Suppl. Fig. 11A**).

Applying this to the CLL data, where RNA-Seq data was available for eight individuals, we were able to classify 50% of the expressed TFs as either activators or repressors (**Fig 4c-f**; n=186). Among the top activators are the IRFs, which are well known transcriptional activators[44] (see **Fig. 4f** for a footprint of IRF2 and **Fig. 4c** for the correlation distribution that led to its classification). Among the top repressors we found PAX5, which has been shown to repress the activity of BLIMP-1 and thereby regulates plasma cell differentiation[45] (**Fig. 4f,c**). The circadian regulators provide an example of why it is important to know the mode of action of a particular TF: When analysing the differential TF activities, it appears as if BMAL1 is more active in M-CLL while the other two TFs (CLOCK and NR1D1) are more active in U-CLL (**Fig 4e**). However, since BMAL1 is an activator, CLOCK a repressor and NR1D1 borderline classified as repressor, we get a consistent result from this analysis whereby all three circadian factors are more "active" in M-CLL, albeit with a contrary effect on their target genes.

In addition to the TF-specific molecular insights and better interpretability of the data, this global, cell-type specific classification of TFs allowed us to explore more details about activators and repressors in general. Importantly, we found that the aggregate signal across all repressors produced a footprint similar to that of REST, while the one for activators looks similar to STAT2, thus again indicating that repressors and activators have very distinct open chromatin footprints (**Fig. 4g**). When investigating TF families, as defined above with the RSat clusters (**Fig. 3**), we found that TFs from the same PWM cluster are often classified both as activators and repressors (**Suppl. Fig. 12**), supporting the hypothesis that the molecular function of a TF is highly variable. The exceptions are clusters 40, containing mainly members of the IRF family and cluster 17 that contains both NFAT and NFKB TFs, which are mostly classified as activators.

As a validation for the repressor/activator classification, we tested two specific hypotheses: first we expected that activating TFs would show a positive correlation between their differential expression and differential TF activity of M-CLL vs U-CLL patients, while it would be

anticorrelated for repressors. Indeed, we found both a significant positive correlation for activators (r=0.19, P = 0.05) and a significant negative one for repressors (r=-0.32, P = 0.0033; **Fig. 5a**). Second, we expect that the expression of target genes of a given TF should go in the same direction as the differential TF activity as assessed by *diffTF*. Again, we found that this was the case for both activators and repressors (**Fig. 5b,** see Methods). These observations provide an additional, independent validation that *diffTF* is able to pick up differential TF activity.

It is important to note that TFs can often act as activators and repressors at different genomic loci e.g. depending on their co-factors, whereas here we predict their main mode of action based on the mean effect across all their predicted binding sites, and thereby lose any information about bifunctionality. Furthermore, since the classification is based on correlations, it is heavily dependent on variation in the RNA-Seq signal across individuals, thus biasing the TFs that can be classified towards those that are variable across individuals. As a consequence, TFs whose post-transcriptional regulation is not reflected in their transcript abundance will not get classified correctly. Another potential misclassification may happen because of the similarity of PWMs within a cluster, which makes it difficult to distinguish the exact effect of one TF while its expression level is uniquely defined. As an example we cite PRDM1, which is part of IRF family (cluster 40) and classified as very strong repressor, its footprint however looks more similar to a typical activator (data not shown), suggesting that it is not PRDM1 driving the ATAC-Seq signal in this case, but the IRFs. Thus, for distinguishing the functional roles of TFs from the same cluster/family further biochemical experiments will be needed. Despite these potential pitfalls, *diffTF* provides unique insights into the molecular mechanism of TFs on a global level.

Comparison with similar tools

There are a few tools with a similar or comparable focus[46–48] and we thought to compare them with *diffTF* (**Suppl. Table 5**). Overall, *diffTF* enables a more flexible and tailored analysis of the data and the question at hand due to the extensive choice of parameters, diagnostic plots, TFBS-specific results, visualizations, and pipeline adjustability. It is unique in its ability to directly integrate RNA-Seq with ATAC-Seq data to classify TFs into activators and repressors. Due to this flexibility, *diffTF* is computationally expensive, and we provide detailed instructions on memory and time requirements in the documentation.

To compare *diffTF* with a more traditional TF motif analysis we chose *HOMER*[48], which looks at motif enrichment in a set of differentially accessible peaks. Strikingly, no enriched motifs were found in M-CLL with *HOMER*, while the few discovered in U-CLL significantly correlated with differential TF activity as computed by *diffTF* (**Supp. Fig. 13,** see Methods). This analysis highlights the power of *diffTF* to capture more significant signal than standard motif enrichment-approaches.

For comparing *diffTF* with an approach that is also based on TF activities we chose *chromVAR*. Overall, the results correlate very well, with correlation coefficients between 0.65-0.9 for Pearson and 0.56-0.81 for Spearman, depending on the set of TFs (i.e., all TFs or only those deemed significant by *diffTF*, therefore predominantly removing TFs with low signal) and whether *chromVAR* deviations or deviation scores are compared against (**Suppl. Fig. 14**, see Methods). Differences likely arose from the way of what the scores represent or what they are compared against (i.e., compared against the mean effect in *diffTF* vs modelled background in *chromVAR*).

Overall, the comparable part of the results of *diffTF* are highly correlated with the output of other tools. However, *diffTF* goes one step beyond the currently available methods by classifying TFs based on their mode of regulation - activator or repressor, thus providing important additional insights into their molecular function.

Experimental validation in a well-studied system

Finally, we experimentally validated our approach using the well-studied mouse hematopoietic system. We generated ATAC-Seq profiles of multipotent progenitor cells (MPP; $Lin^{-}cKit^{+}Sca1^{+}CD150^{-}CD48^{+}$), an early hematopoietic progenitor population capable of supporting multilineage blood production[49], as well as the more differentiated and myeloid-restricted granulocyte-monocyte progenitors (GMP; $Lin^{-}cKit^{+}Sca1^{-}$; $CD16/32^{+}$). The obtained profiles were processed using an in-house ATAC-Seq pipeline and *diffTF* was subsequently used to identify TFs that are differentially active between MPP and GMP cells (see Online Methods). Due to the large number of significant TFs reflecting the high diversity between the two cell types, we used available RNA-Seq data to filter non-expressed TFs. The small number of samples (two) did not allow a faithful classification the TFs into activators and repressors since the classification is based on correlations across samples. In GMPs, members of the well-known class of master regulators of myelopoiesis, the CEBP family (C/ebpα,-β,-δ,-ε,-γ), showed a strong increase in

activity. In addition, we could detect increased activity of the NFKB (RELB, NFKB2)[50] as well as the AP-1 (JUN/FOS)[51] complexes, which act as another cue for further granulopoiesis (**Fig. 6**, **Suppl. Fig 15**). In addition, we also observed higher activity of the MYC/MYB factors, which are known to be exclusively active in the GMPs[52] as well as NFIL3, which is involved in the generation of natural killer cells[53]. Conversely, MPPs show a higher activity for IRF/STAT, ZEB1 and ITF2 (part of the Wnt signaling) as well as TFs from the Homeodomain and Forkhead families[54], all of which are associated with self-renewal of hematopoietic stem cells.

Overall, these results show that *diffTF* was able to identify the known TFs that drive the differentiation from MPP to GMP, thus demonstrating that it confidently identifies the biologically significant signals.

DISCUSSION

We here presented a genome-wide method for quantifying differences in differential TF activity simultaneously for a large set of TFs, and for classifying them into their molecular mode of action as transcriptional activators or repressors. The method is available for download at *https://git.embl.de/grp-zaugg/diffTF* along with a comprehensive documentation and example data.

We have shown the robustness of *diffTF* with respect to a large range of parameters, sample sizes and read depth. In particular, we have shown that *diffTF* is able to overcome the inherent noisiness of TF binding site predictions by aggregating data across all putative binding sites, as the results when using putative binding sites and ChIP-Seq validated sites were strongly correlated. Furthermore, we validated the estimation of differential TF activity by (i) comparing our results with known biological relationships between TFs and certain biological processes, (ii) by showing an overall correspondence between changes in transcript abundance and changes in the activity of a given TF, and finally experimentally (iii) by profiling different cell types along the well-known hematopoietic differentiation system and identifying the key driving factors.

To demonstrate the power of *diffTF* we have applied it to identify and characterise differences between M-CLL and U-CLL from a publicly available dataset of ATAC-Seq and RNA-Seq. It is noteworthy that a TF motif enrichment analysis on the significantly differentially accessible peaks did not reveal any factor to be significantly enriched in M-CLL, indicating that in this case (as probably in many patient - control studies) the key driver TFs are not necessarily switching their target peaks (enhancers and promoters) on and off, but rather mis-regulate many common

regions to a lesser extent. Although signal differences at each individual putative binding site can be very low and rarely significant, if there is a slight bias in activity of the TF, *diffTF* can still detect an overall difference in TF binding activity by averaging across all of its putative binding sites and is therefore more powerful than conventional enrichment analyses.

Unlike other methods that have been published on differential TF activity based on chromatin[46,47], we have devised an approach within the *diffTF* framework to classify TFs into activators and repressors based on the correlation of their expression level (RNA-Seq) and the activity of their putative binding sites (ATAC-Seq). This information is highly relevant when interpreting the action of TFs since it makes a difference of whether the signal is caused by a repressor that is less active in condition A or an activator that is more active in condition A. Notably, this classification could work even for data sets for which no RNA-Seq data are available, as one can use ATAC-Seq signal summarized across the gene body as proxy for its expression level. We have observed this correlation to be around 0.63 (Pearson).

Calculating differential TF activity based on aggregating signal across the genome is not a new idea and has been proposed several times based on aggregating the expression of putative target genes of a certain TF[55,56]. Using the effect on chromatin instead of expression has several advantages: first chromatin is a much simpler trait since gene expression is the sum of transcription and degradation. Second, there are much more peaks than genes, thus allowing for better statistics and signal to noise ratio. Finally, the effect on chromatin is much more local than on gene expression - in particular in mammalian genomes, where genes can be regulated by very distal enhancers. We have compared differential TF activity calculated based on the average expression change of the target genes to the output of *diffTF* and found that while the direction of activity between both methods is highly correlated, the signal is much lower when using the expression of target genes instead of chromatin accessibility at putative binding sites.

While similar methods have been proposed for analysing ATAC-Seq data[46,47], our method has several advantages when dealing with bulk ATAC-Seq data and can also be used for histone mark ChIP-Seq data: (i) Unlike other methods that calculate the background theoretically based on the genome-wide read depth, it is insensitive to sequence or locus dependent biases since we calculate a fold-change between conditions for each region, thus normalizing for local read depth biases. This is particularly advantageous for detecting small changes such as between two heterogeneous cohorts in patient-control studies. (ii) *diffTF* allows integration with matching RNA-Seq data to classify TFs into activators and repressors in a fully data-driven approach

within the same analysis framework. Such classifications greatly help in interpreting the effects of up/down regulation of a particular factors. And, even if classifying them is not possible, at least non-expressed TFs can be identified. (iii) *diffTF* provides the fold-change value for each TFBS which allows for easy retrospective follow-up analysis, e.g. identifying the most differential regions regulated by a specific set of TFs. (iv) Finally, our method allows to analyse time-course data in an additive manner by calculating the overall change of slope for each TF (see Methods).

Overall, with *diffTF* we present a multiomics data integration strategy of ATAC-Seq and RNA-Seq data that allows to calculate differential TF activity across conditions and classifying TFs based on their molecular mode of action into activators and repressors. With this, *diffTF* is laying the foundation for understanding the regulatory mechanisms that are driving the differences in cell state on a systems-wide scale.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank Bernd Klaus for help with the statistical part and EMBL for funding. A.R.P is recipient of a postdoctoral fellowship granted by Fundación Ramón Areces. G.P. was awarded the Otto Bayer Scholarship from Bayer Foundations.

AUTHOR CONTRIBUTIONS

J.Z. and I.B. conceived the study, C.A. and I.B. developed the computational framework and performed the analyses, A.R.P. and G.P. contributed to the development and analysis, K.D.R performed the experiments, J.Z. supervised the study and C.A., I.B, and J.Z wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests

Figures and figure captions

Figure 1. Concept and simplified workflow of *diffTF*. (a) Concept of the differential TF activity approach: Based on findings from chromatin QTL studies that have reported changes in chromatin often coincide with SNPs disrupting TF binding sites (TFBS), we propose to detect global differences in TF activity by investigating chromatin accessibility (or modifications) in close proximity to their TFBS. (b) A simplified workflow illustrates the principle *diffTF* is based on: it calculates a fold-change between two conditions for each binding site of a given TF and compares this distribution to a background set of fold-changes obtained from GC-content matched loci that do not contain the putative TFBS. The difference in distribution is assessed in significance and effect size and visualized in a volcano plot where the y-axis indicates statistical significance and the x-axis shows the effect size. (For a detailed workflow see Suppl. Fig. 1 and 2).

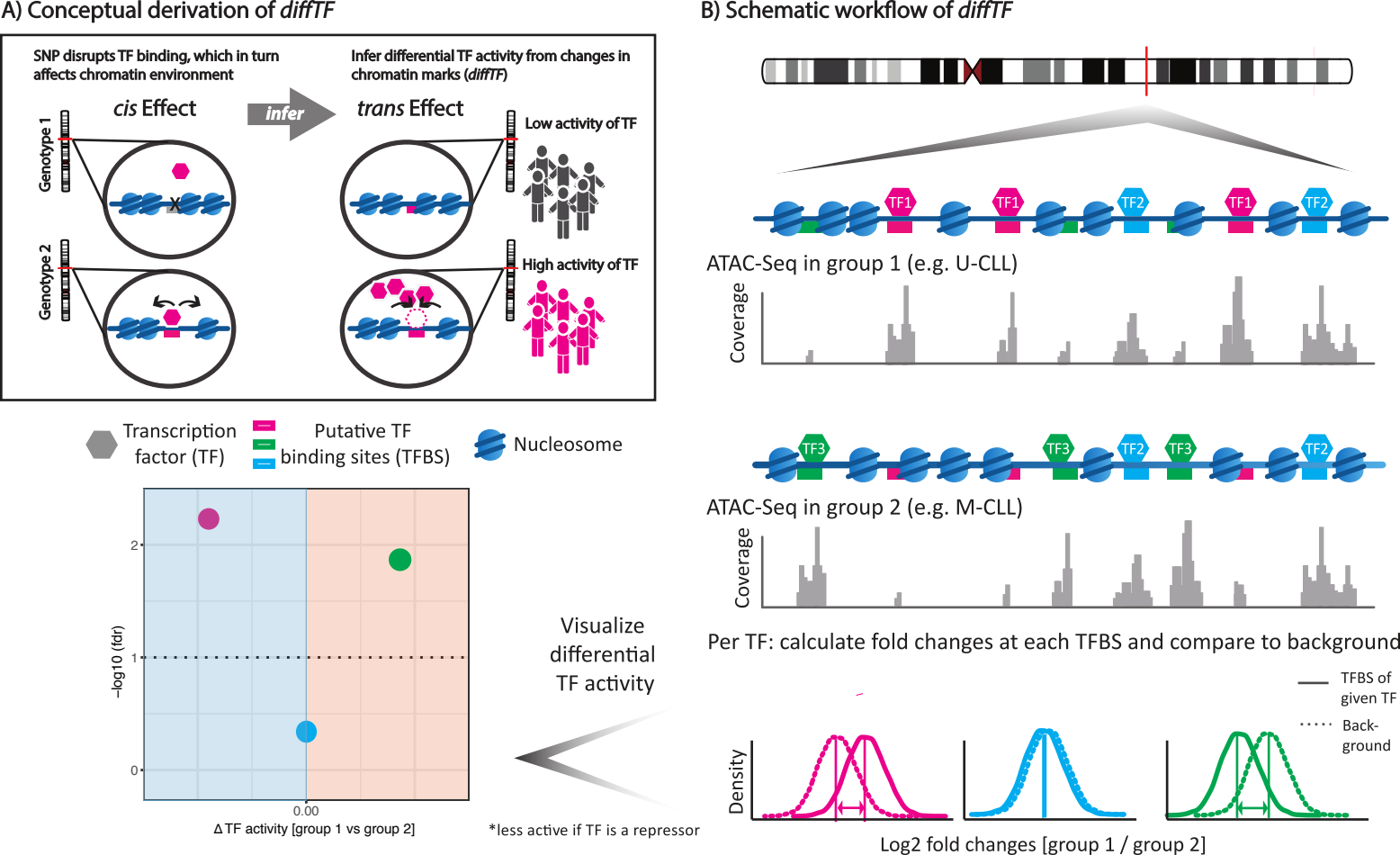
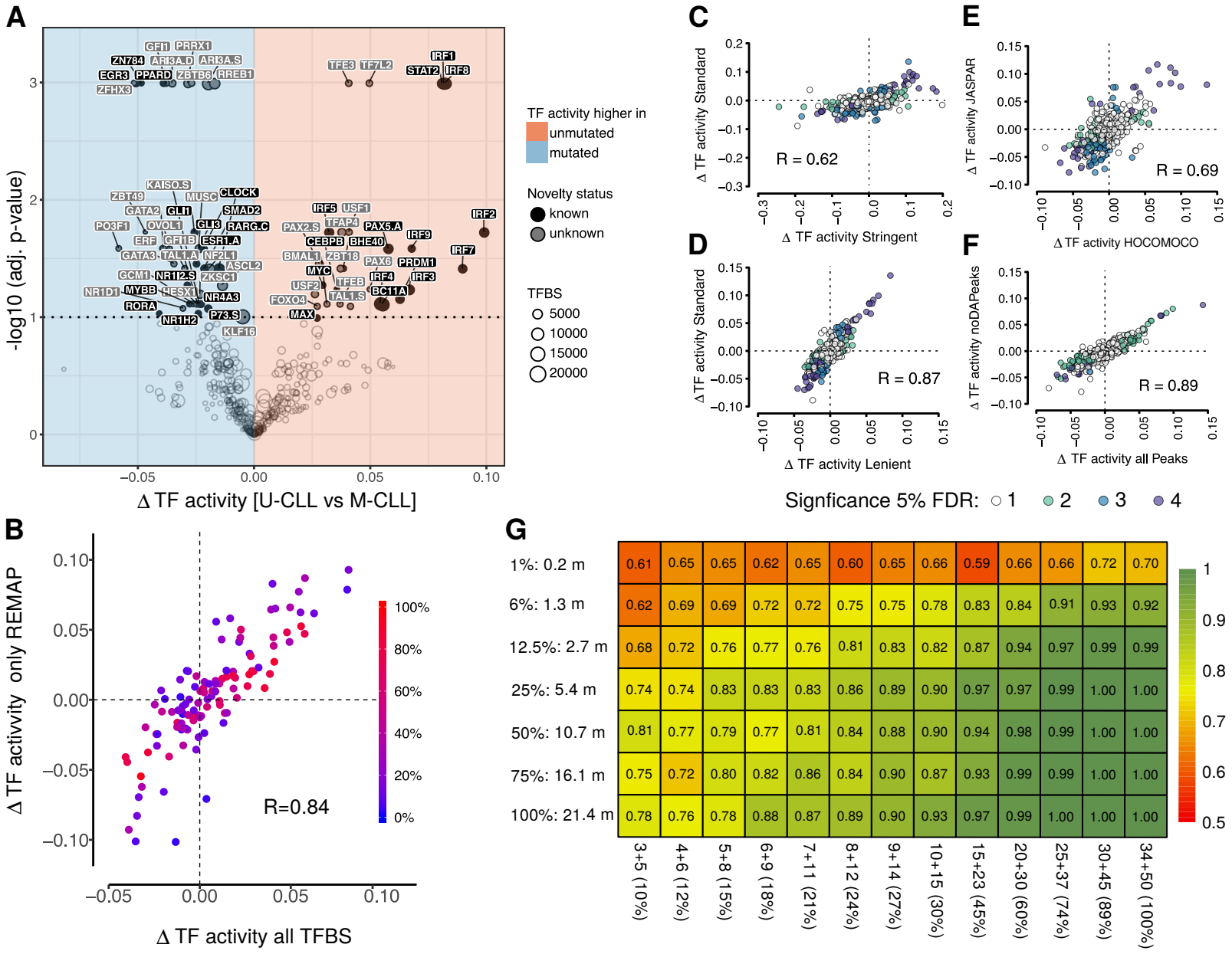
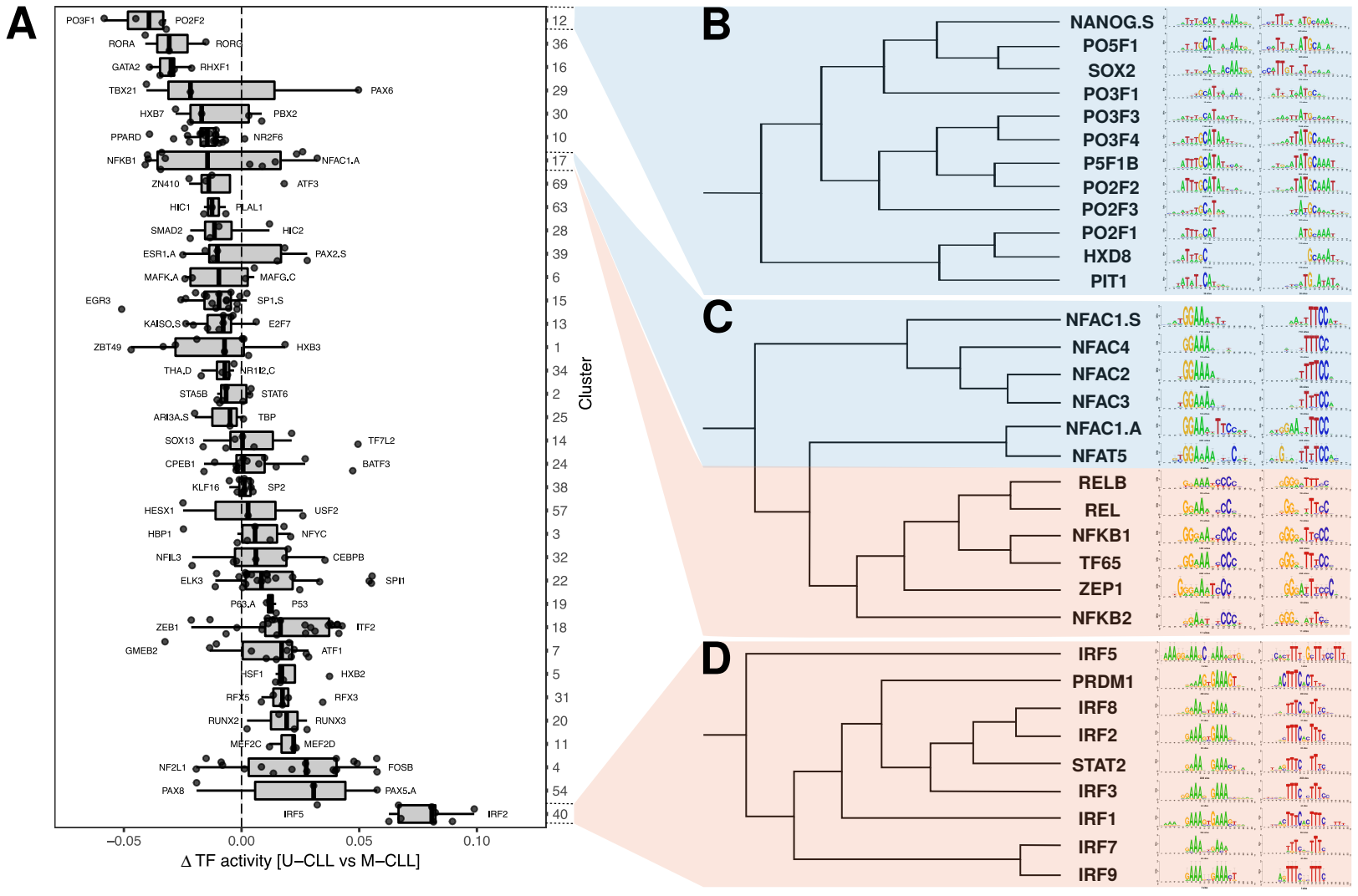

A) Conceptual derivation of diffTF
SNP disrupts TF binding, which in turn affects chromatin environment
Infer differential TF activity from changes in chromatin marks (diffTF)
cis Effect
infer
trans Effect
Genotype 1
Genotype 2
Low activity of TF
High activity of TF
Transcription factor (TF)
Putative TF binding sites (TFBS)
Nucleosome
−log10 (fdr)
Δ TF activity [group 1 vs group 2]
Visualize differential TF activity
*less active if TF is a repressor
B) Schematic workflow of diffTF
TF1
TF2
TF3
ATAC-Seq in group 1 (e.g. U-CLL)
ATAC-Seq in group 2 (e.g. M-CLL)
Coverage
Per TF: calculate fold changes at each TFBS and compare to background
TFBS of given TF
Back-ground
Density
Log2 fold changes [group 1 / group 2]

Figure 2. *diffTF* results for the CLL dataset, experimental validation and technical robustness of the method. (a) Volcano plot of differential TF activity between U-CLL (n=27) and M-CLL (n=25) patients.. The y-axis denotes statistical significance (-log10 transformed). TFs that pass the significance threshold (5% FDR; dotted line) are labeled and colored according to their novelty status (see Suppl. Table. 2). #TFBS denotes the number of predicted TF binding sites in the peak regions this analysis is based on. (b)-(f): Technical robustness of *diffTF*. Scatterplots of the differential TF activity from all TFs for two different *diffTF* analyses are shown. Each point represents one TF. For (c-f), colors represent significance at 5% FDR (white - not significant in either analysis; light green and light blue - significant for the analysis on the x-axis or y-axis, respectively; purple - significant for both analyses). (b) Comparison of all predicted TFBS and TFBS experimentally validated by ChIP-Seq data from ReMap. See also Suppl. Fig. 7. (c-d) Comparison for different p-value thresholds in PWMScan to predict TFBS: (c) standard (1e-5) vs. stringent (1e-6), (d) standard vs. lenient (5e-5) (n=628 TFs for which binding sites were retrieved for both scanning modes.) (e) Comparison of *diffTF* results based on *HOCOMOCO* v10 vs *JASPAR* 2018 as input (n=412 TFs, for which a motif was available in both databases.) (f) Comparison of the full consensus peak set (*allPeaks*) and only the non-differentially accessible peaks (*noDApeaks*; n=640 TFs from *HOCOMOCO*) (g) Robustness analysis based on sequencing depth and sample size. Each cell in the heatmap shows the fraction of TFs that showed the same direction of change as in the full dataset for varying degrees of down-sampling sequencing depth and number of samples, (5% FDR), averaged over 50 independent repetitions to minimize sampling noise. Only TFs that were deemed significant in the full dataset are considered (see also Suppl. Fig. 8). Sequencing depth is shown as fraction of the original data and median number of reads across samples, while the number of samples is shown in the format unmutated + mutated.


A
-log10 (adj. p-value)
Δ TF activity [U-CLL vs M-CLL]
GFI1
PRRX1
ZN784
ARI3A.D
ARI3A.S
EGR3
PPARD
ZBTB6
RREB1
ZFHX3
TFE3
TF7L2
IRF1
STAT2
IRF8
KAISO.S
ZBT49
MUSC
CLOCK
GATA2
GLI1
SMAD2
PO3F1
OVOL1
GLI3
RARG.C
ERF
GFI1B
ESR1.A
GATA3
TAL1.A
NF2L1
ASCL2
GCM1
NR1I2.S
ZKSC1
NR1D1
MYBB
HESX1
NR4A3
RORA
NR1H2
P73.S
KLF16
IRF5
USF1
TFAP4
PAX2.S
PAX5.A
IRF9
CEBPB
BHE40
IRF2
IRF7
BMAL1
ZBT18
PAX6
PRDM1
MYC
IRF4
IRF3
USF2
TFEB
BC11A
FOXO4
TAL1.S
MAX
TF activity higher in
unmutated
mutated
Novelty status
known
unknown
TFBS
5000
10000
15000
20000
B
Δ TF activity only REMAP
Δ TF activity all TFBS
R=0.84
100%
80%
60%
40%
20%
0%
C
Δ TF activity Standard
Δ TF activity Stringent
R = 0.62
D
Δ TF activity Standard
Δ TF activity Lenient
R = 0.87
E
Δ TF activity JASPAR
Δ TF activity HOCOMOCO
R = 0.69
F
Δ TF activity noDAPeaks
Δ TF activity all Peaks
R = 0.89
Signficance 5% FDR: 1 2 3 4
G
1%: 0.2 m
6%: 1.3 m
12.5%: 2.7 m
25%: 5.4 m
50%: 10.7 m
75%: 16.1 m
100%: 21.4 m
3+5 (10%)
4+6 (12%)
5+8 (15%)
6+9 (18%)
7+11 (21%)
8+12 (24%)
9+14 (27%)
10+15 (30%)
15+23 (45%)
20+30 (60%)
25+37 (74%)
30+45 (89%)
34+50 (100%)

455 **Figure 3**. Functional clustering of TFs based on the similarity of their PWMs. (a) Boxplot of
456 PWM clusters with at least 3 members as defined by RSAT for the differential TF activity
457 between U-CLL and M-CLL. In each cluster, the most negative and most positive TF is labeled.
458 (b) RSAT clustering output and tree for cluster 17. The cluster has two distinct subclusters,
459 representing the NFAT and NFKB families. (c) and (d) RSAT clustering output and tree for
460 cluster 12 (POU family) and 40 (IRF family), the most distinct clusters for M-CLL and U-CLL
461 patients, respectively).


A
PO3F1
PO2F2
RORA
RORC
GATA2
RHXF1
TBX21
PAX6
HXB7
PBX2
PPARD
NR2F6
NFKB1
NFAC1.A
ZN410
ATF3
HIC1
FLAL1
SMAD2
HIC2
ESR1.A
PAX2.S
MAFK.A
MAFG.C
EGR3
SP1.S
KAISO.S
E2F7
ZBT49
HXB3
THA.D
NR1I2.C
STA5B
STAT6
ARI3A.S
TBP
SOX13
TF7L2
CPEB1
BATF3
KLF16
SP2
HESX1
USF2
HBP1
NFYC
NFIL3
CEBPB
ELK3
SPI1
P63.A
P53
ZEB1
ITF2
GMEB2
ATF1
HSF1
HXB2
RFX5
RFX3
RUNX2
RUNX3
MEF2C
MEF2D
NF2L1
FOSB
PAX8
PAX5.A
IRF5
IRF2
12
36
16
29
30
10
17
69
63
28
39
6
15
13
1
34
2
25
14
24
38
57
3
32
22
19
18
7
5
31
20
11
4
54
40
Cluster
−0.05
0.00
0.05
0.10
Δ TF activity [U−CLL vs M−CLL]
B
NANOG.S
PO5F1
SOX2
PO3F1
PO3F3
PO3F4
P5F1B
PO2F2
PO2F3
PO2F1
HXD8
PIT1
C
NFAC1.S
NFAC4
NFAC2
NFAC3
NFAC1.A
NFAT5
RELB
REL
NFKB1
TF65
ZEP1
NFKB2
D
IRF5
PRDM1
IRF8
IRF2
STAT2
IRF3
IRF1
IRF7
IRF9

462 **Figure 4**. Classification of TFs into activator or repressor based on RNA-Seq and ATAC-Seq
463 data. (a) Schematic description of the classification approach: correlation of TF expression level
464 with the accessibility of its target sites. If the correlation of the RNA-level and the chromatin
465 peak openness is positive, the TF acts more as an activator in the particular biological
466 environment, if negative, it acts more as a repressor. Correlations around 0 are classified as
467 undetermined. (b) Exemplary footprints for a well-known repressor (REST, left) and activator
468 (STAT2, right), respectively. The x-axis depicts the distance in bp from the TFBS center, the
469 y-axis denotes the number of average Tn5 insertions, normalised to the library size and
470 numbers of samples across U-CLL and M-CLL. TFBS predicted by PWMscan and overlapping
471 with open chromatin have been used. Solid black line indicates the average insertion sites within
472 accessible chromatin. (c) Distributions of the Pearson correlations between TF expression and
473 ATAC-Seq signal at all putative TFBS (red line), and background distribution of TFBS
474 not-containing the motif of interest (grey line) for two specific TFs, IRF2 (left) and PAX5 (right).
475 (d) Heatmap of the differences in Pearson correlation of putative targets peaks and background
476 for each TF (i.e. subtracting the black from the red line in (c)). Each TF is one row and is
477 annotated as activator (green), undetermined (grey) or repressor (red) as classified by *diffTF*.
478 (e) Same as Figure 2a, but with their predicted role as activator (green), undetermined (black) or
479 repressor (red). See Figure 2a for details. (f) Footprint analysis for an activator (IRF2; right) and
480 repressor (PAX5; left) as classified by *diffTF*. The footprint based on the normalized number of
481 Tn5 insertions (top; blue: U-CLL, orange: M-CLL) and differences (U-CLL - M-CLL; bottom) are
482 shown. See b) for axis descriptions. (g) Summary footprint for all activators (top, green) and
483 repressors (bottom, red). See b) for axis descriptions.

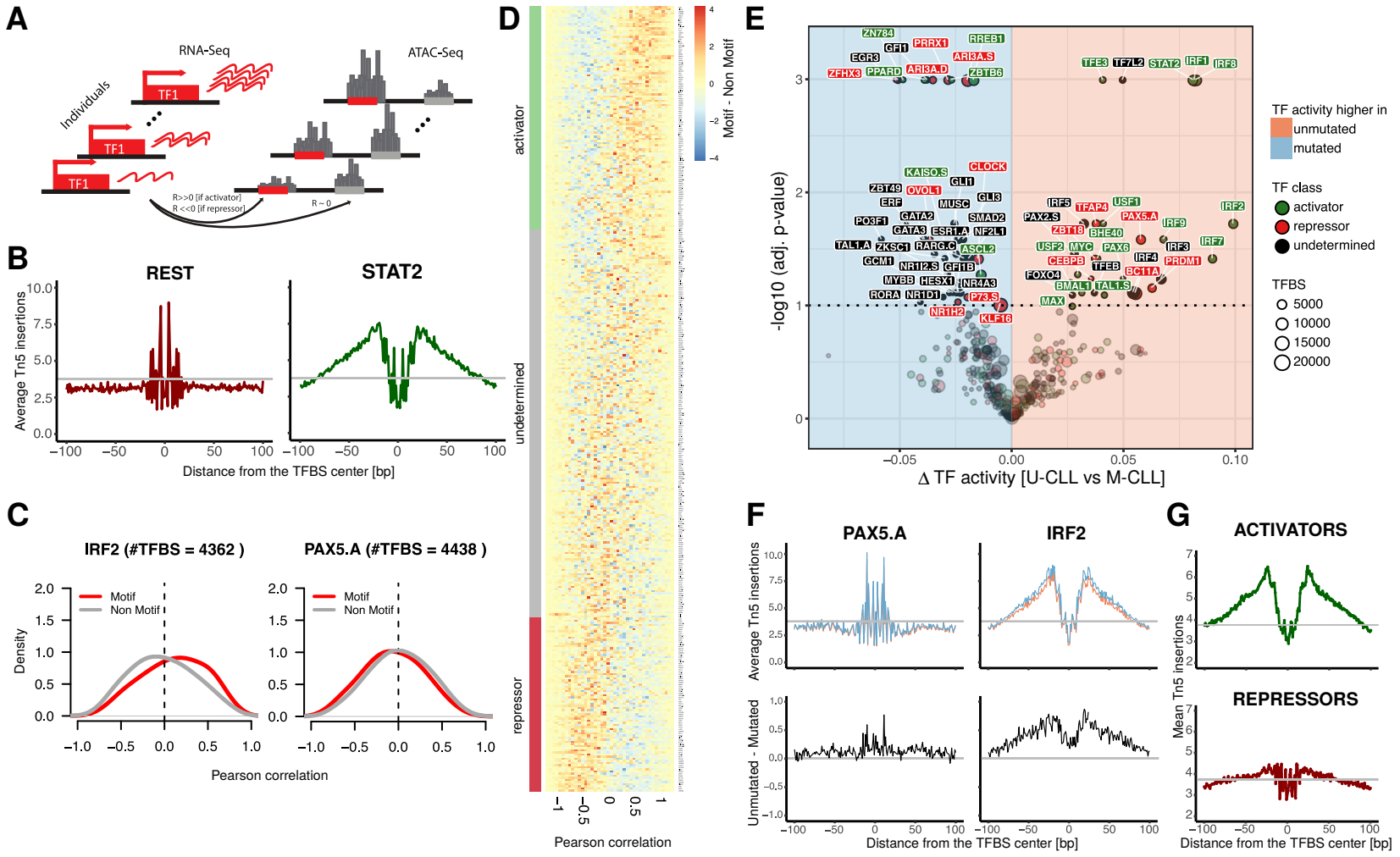
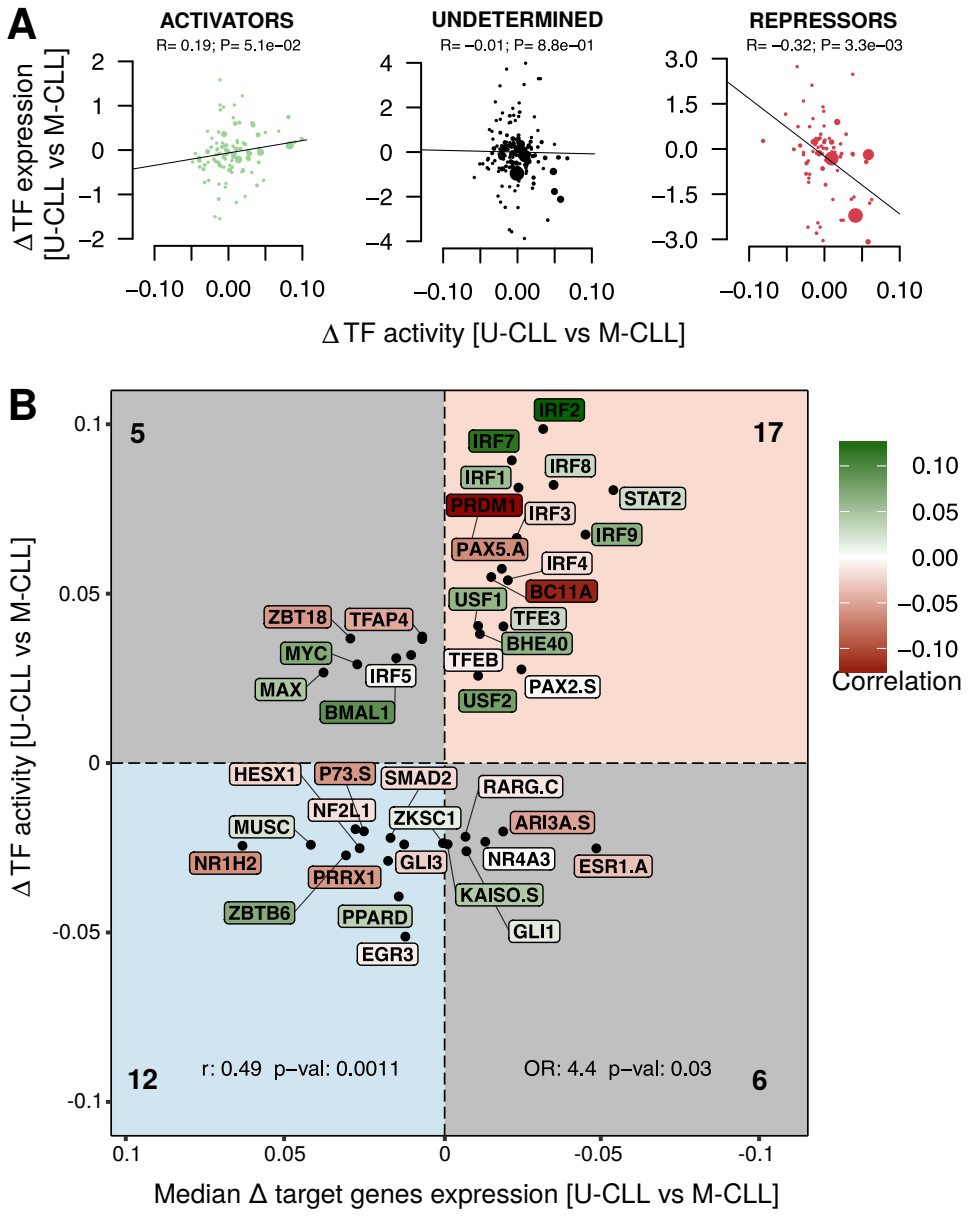
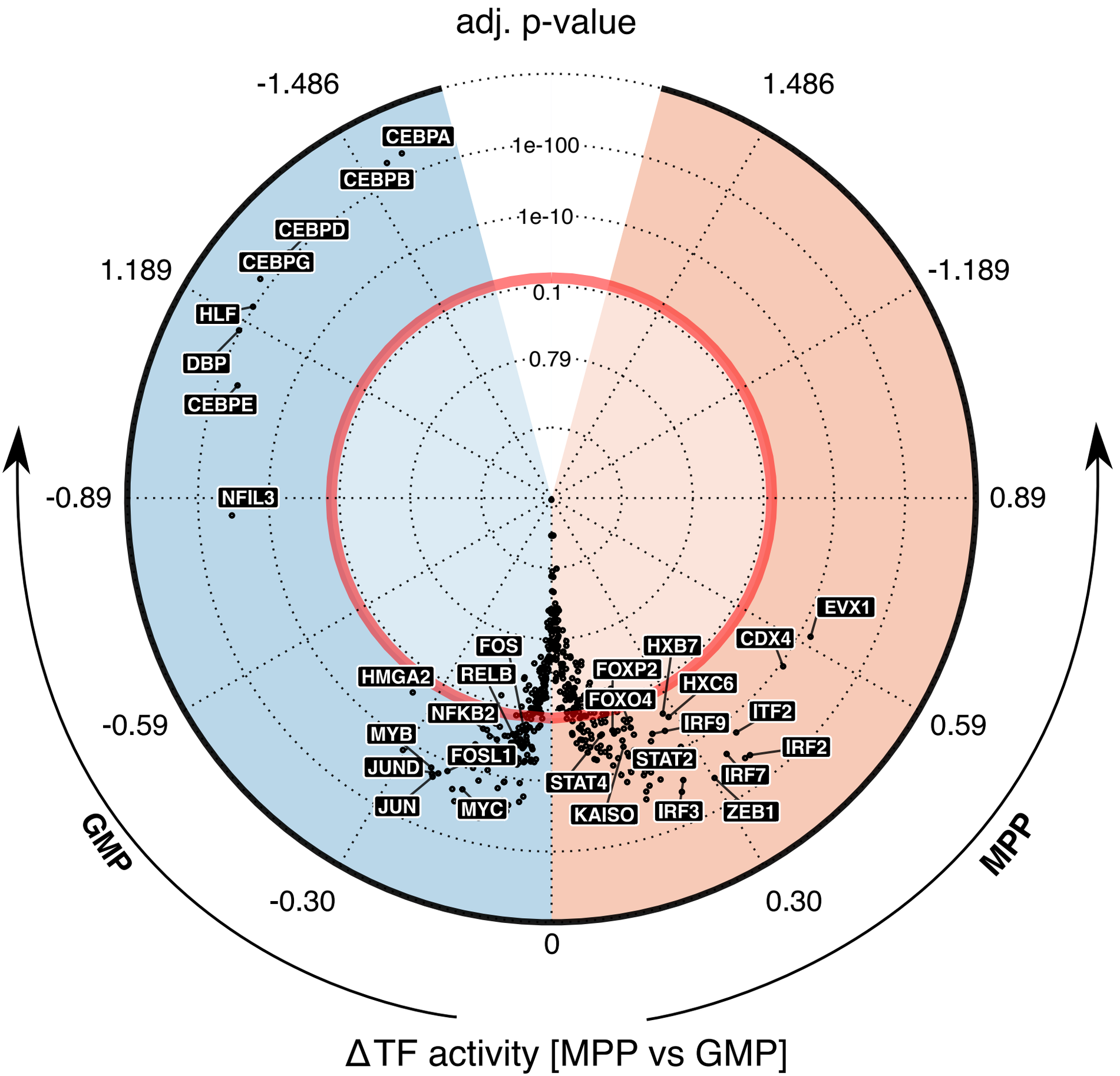

A
RNA-Seq
ATAC-Seq
Individuals
TF1
R>>0 [if activator]
R <<0 [if repressor]
R ~ 0
B
REST
STAT2
Average Tn5 insertions
Distance from the TFBS center [bp]
C
IRF2 (#TFBS = 4362)
PAX5.A (#TFBS = 4438)
Motif
Non Motif
Density
Pearson correlation
D
activator
undetermined
repressor
Motif - Non Motif
Pearson correlation
E
-log10 (adj. p-value)
Δ TF activity [U-CLL vs M-CLL]
TF activity higher in
unmutated
mutated
TF class
activator
repressor
undetermined
TFBS
5000
10000
15000
20000
F
PAX5.A
IRF2
Average Tn5 insertions
Unmutated - Mutated
Distance from the TFBS center [bp]
G
ACTIVATORS
REPRESSORS
Mean Tn5 insertions
Distance from the TFBS center [bp]

Figure 5. (a) Pearson correlation of the log2 fold changes from RNA-Seq and differential TF activity for activators (green, left), undetermined TFs (black, middle), and repressors (red, right). Only expressed TFs are shown. (b) Correlation of differential TF activity and median of the differential target gene expression of U-CLL against M-CLL. The x-axis shows the median target gene (TG) log2 fold-change, the y-axis denotes the differential TF activity. Each TF label is colored based on its activator/repressor status (green/red) on a continuous scale (dark to light) based on the the correlation strength (see Figure 4d). (OR=odds ratio, r=Pearson correlation coefficient).


A
ACTIVATORS
R= 0.19; P= 5.1e−02
UNDETERMINED
R= −0.01; P= 8.8e−01
REPRESSORS
R= −0.32; P= 3.3e−03
ΔTF expression [U-CLL vs M-CLL]
ΔTF activity [U-CLL vs M-CLL]
B
5
17
12
6
IRF2
IRF7
IRF8
IRF1
STAT2
PRDM1
IRF3
IRF9
PAX5.A
IRF4
BC11A
USF1
TFE3
ZBT18
TFAP4
BHE40
MYC
TFEB
IRF5
MAX
PAX2.S
BMAL1
USF2
HESX1
P73.S
SMAD2
RARG.C
NF2L1
ZKSC1
ARI3A.S
MUSC
NR4A3
ESR1.A
NR1H2
GLI3
PRRX1
KAISO.S
ZBTB6
PPARD
GLI1
EGR3
r: 0.49 p−val: 0.0011
OR: 4.4 p−val: 0.03
Correlation
ΔTF activity [U-CLL vs M-CLL]
Median Δ target genes expression [U-CLL vs M-CLL]

492 **Figure 6**. *diffTF* recapitulates known TFs that drive the differentiation from MPP to GMP.
493 Circular plot of differential TF activity between 4 wildtype GMP and 4 wildtype MPP samples are
494 shown. This is an alternative representation equivalent to the Volcano plot shown in Figure 2:
495 values on the x-axis denote the differential TF activity, with higher values indicating a higher
496 differential TF activity between the two conditions. TFs that are similarly active between GMP
497 and MPP samples are close to 0, while TFs more active in either condition are located in the left
498 and right part of the plot. The y-axis (radial position) denotes the statistical significance
499 (adjusted p-values). The significance threshold (5% FDR) is indicated as a red circular line. Due
500 to the high number of significant TFs we labeled only those discussed in the main text. The full
501 list is available in Suppl. Table 4.


adj. p-value
-1.486
1.486
1e-100
1e-10
0.1
0.79
1.189
-1.189
-0.89
0.89
-0.59
0.59
-0.30
0.30
0
CEBPA
CEBPB
CEBPD
CEBPG
HLF
DBP
CEBPE
NFIL3
EVX1
CDX4
HXB7
FOS
FOXP2
HMGA2
RELB
HXC6
FOXO4
NFKB2
ITF2
IRF9
MYB
IRF2
FOSL1
STAT2
JUND
IRF7
STAT4
JUN
MYC
KAISO
IRF3
ZEB1
GMP
MPP
ΔTF activity [MPP vs GMP]

502 **REFERENCES**

503 1. Kim, R., Emi, M. & Tanabe, K. Cancer immunoediting from immune surveillance to immune
504 escape. *Immunology* **121,** 1–14 (2007).
505 2. Han, H. *et al.* TRRUST: a reference database of human transcriptional regulatory
506 interactions. *Sci. Rep.* **5,** 11432 (2015).
507 3. Teng, M. W. L., Swann, J. B., Koebel, C. M., Schreiber, R. D. & Smyth, M. J.
508 Immune-mediated dormancy: an equilibrium with cancer. *J. Leukoc. Biol.* **84,** 988–993
509 (2008).
510 4. Bonhoure, N. *et al.* Quantifying ChIP-seq data: a spiking method providing an internal
511 reference for sample-to-sample normalization. *Genome Res.* **24,** 1157–1168 (2014).
512 5. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse
513 transcriptional regulatory interactions. *Nucleic Acids Res.* **46,** D380–D386 (2018).
514 6. Rasmussen, K. *et al.* TET2 binding to enhancers facilitates transcription factor recruitment
515 in hematopoietic cells. *BioRxiv* (2018). doi:10.1101/336008
516 7. Grubert, F. *et al.* Genetic control of chromatin states in humans involves local and distal
517 chromosomal interactions. *Cell* **162,** 1051–1065 (2015).
518 8. Liu, L., Jin, G. & Zhou, X. Modeling the relationship of epigenetic modifications to
519 transcription factor binding. *Nucleic Acids Res.* **43,** 3873–3885 (2015).
520 9. Kulakovskiy, I. V. *et al.* HOCOMOCO: a comprehensive collection of human transcription
521 factor binding sites models. *Nucleic Acids Res.* **41,** D195-202 (2013).
522 10. Ambrosini, G., Groux, R. & Bucher, P. PWMScan: a fast tool for scanning entire genomes
523 with a position-specific weight matrix. *Bioinformatics* (2018).
524 doi:10.1093/bioinformatics/bty127
525 11. Griffon, A. *et al.* Integrative analysis of public ChIP-seq experiments reveals a complex

526 multi-cell regulatory landscape. *Nucleic Acids Res.* **43,** e27 (2015).
527 12. Furman, R. R. *et al.* Idelalisib and rituximab in relapsed chronic lymphocytic leukemia. *N.*
528 *Engl. J. Med.* **370,** 997–1007 (2014).
529 13. Neu, K. E. & Wilson, P. C. Taking the broad view on B cell affinity maturation. *Immunity* **44,**
530 518–520 (2016).
531 14. Chiorazzi, N. & Ferrarini, M. Cellular origin(s) of chronic lymphocytic leukemia: cautionary
532 notes and additional considerations and possibilities. *Blood* **117,** 1781–1791 (2011).
533 15. Rendeiro, A. F. *et al.* Chromatin accessibility maps of chronic lymphocytic leukaemia
534 identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat.*
535 *Commun.* **7,** 11938 (2016).
536 16. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE
537 consortia. *Genome Res.* **22,** 1813–1831 (2012).
538 17. Medina-Rivera, A. *et al.* RSAT 2015: regulatory sequence analysis tools. *Nucleic Acids Res.*
539 **43,** W50-6 (2015).
540 18. Lin, T. S., Mahajan, S. & Frank, D. A. STAT signaling in the pathogenesis and treatment of
541 leukemias. *Oncogene* **19,** 2496–2504 (2000).
542 19. Benekli, M., Baumann, H. & Wetzler, M. Targeting signal transducer and activator of
543 transcription signaling pathway in leukemias. *J. Clin. Oncol.* **27,** 4422–4432 (2009).
544 20. Arvaniti, E. *et al.* Toll-like receptor signaling pathway in chronic lymphocytic leukemia:
545 distinct gene expression profiles of potential pathogenic significance in specific subsets of
546 patients. *Haematologica* **96,** 1644–1652 (2011).
547 21. Havelange, V. *et al.* IRF4 mutations in chronic lymphocytic leukemia. *Blood* **118,** 2827–2829
548 (2011).
549 22. Slager, S. L., Caporaso, N. E., de Sanjose, S. & Goldin, L. R. Genetic susceptibility to

550 chronic lymphocytic leukemia. *Semin. Hematol.* **50,** 296–302 (2013).
551 23. Ghamlouch, H. *et al.* Factors involved in CLL pathogenesis and cell survival are disrupted
552 by differentiation of CLL B-cells into antibody-secreting cells. *Oncotarget* **6,** 18484–18503
553 (2015).
554 24. Mittal, A. K. *et al.* Role of CTLA4 in the proliferation and survival of chronic lymphocytic
555 leukemia. *PLoS ONE* **8,** e70352 (2013).
556 25. Landau, D. A. *et al.* Mutations driving CLL and their evolution in progression and relapse.
557 *Nature* **526,** 525–530 (2015).
558 26. Yeomans, A. *et al.* Engagement of the B-cell receptor of chronic lymphocytic leukemia cells
559 drives global and MYC-specific mRNA translation. *Blood* **127,** 449–457 (2016).
560 27. Socola, F., Insuasti-Beltran, G., Henrich Lobo, R., Atrash, S. & Sasapu, A. Chronic
561 Lymphocytic Leukemia with Translocation (2;14)(p16;q32): A Case Report and Review of
562 the Literature. *Case Rep. Oncol. Med.* **2016,** 9037436 (2016).
563 28. Heckman, C. A., Duan, H., Garcia, P. B. & Boxer, L. M. Oct transcription factors mediate
564 t(14;18) lymphoma cell survival by directly regulating bcl-2 expression. *Oncogene* **25,**
565 888–898 (2006).
566 29. Minami, Y., Oishi, I., Endo, M. & Nishita, M. Ror-family receptor tyrosine kinases in
567 noncanonical Wnt signaling: their implications in developmental morphogenesis and human
568 diseases. *Dev. Dyn.* **239,** 1–15 (2010).
569 30. Kikushige, Y. *et al.* Self-renewing hematopoietic stem cell is the primary target in
570 pathogenesis of human chronic lymphocytic leukemia. *Cancer Cell* **20,** 246–259 (2011).
571 31. Oakes, C. C. *et al.* DNA methylation dynamics during B cell maturation underlie a
572 continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat. Genet.* **48,**
573 253–264 (2016).

574 32. Li, Y. J. *et al.* PPAR-delta promotes survival of chronic lymphocytic leukemia cells in
575 energetically unfavorable conditions. *Leukemia* **31,** 1905–1914 (2017).
576 33. Kern, D. *et al.* Hedgehog/GLI and PI3K signaling in the initiation and maintenance of
577 chronic lymphocytic leukemia. *Oncogene* **34,** 5341–5351 (2015).
578 34. El-Athman, R. & Relógio, A. Escaping circadian regulation: an emerging hallmark of
579 cancer? *Cell Syst.* **6,** 266–267 (2018).
580 35. D'Annibale, S. *et al.* Proteasome-dependent degradation of transcription factor activating
581 enhancer-binding protein 4 (TFAP4) controls mitotic division. *J. Biol. Chem.* **289,**
582 7730–7737 (2014).
583 36. Kreslavsky, T. *et al.* Essential role for the transcription factor Bhlhe41 in regulating the
584 development, self-renewal and BCR repertoire of B-1a cells. *Nat. Immunol.* **18,** 442–455
585 (2017).
586 37. Kauffman, E. C. *et al.* Molecular genetics and cellular features of TFE3 and TFEB fusion
587 kidney cancers. *Nat. Rev. Urol.* **11,** 465–475 (2014).
588 38. Kasim, V. *et al.* Transcription factor Yin Yang 2 is a novel regulator of the p53/p21 axis.
589 *Oncotarget* **8,** 54694–54707 (2017).
590 39. Chevrier, S. *et al.* The BTB-ZF transcription factor Zbtb20 is driven by Irf4 to promote
591 plasma cell differentiation and longevity. *J. Exp. Med.* **211,** 827–840 (2014).
592 40. Zhou, Y. *et al.* Lin28b promotes fetal B lymphopoiesis through the transcription factor
593 Arid3a. *J. Exp. Med.* **212,** 569–580 (2015).
594 41. Rathinam, C., Lassmann, H., Mengel, M. & Klein, C. Transcription factor Gfi1 restricts B
595 cell-mediated autoimmunity. *J. Immunol.* **181,** 6222–6229 (2008).
596 42. Gómez-Martín, D., Díaz-Zamudio, M., Galindo-Campos, M. & Alcocer-Varela, J. Early
597 growth response transcription factors and the modulation of immune response: implications

598 towards autoimmunity. *Autoimmun. Rev.* **9,** 454–458 (2010).
599 43. Coscia, M. *et al.* IGHV unmutated CLL B cells are more prone to spontaneous apoptosis
600 and subject to environmental prosurvival signals than mutated CLL B cells. *Leukemia* **25,**
601 828–837 (2011).
602 44. Yanai, H., Negishi, H. & Taniguchi, T. The IRF family of transcription factors: Inception,
603 impact and implications in oncogenesis. *Oncoimmunology* **1,** 1376–1386 (2012).
604 45. Yasuda, T. *et al.* B cell receptor-ERK1/2 signal cancels PAX5-dependent repression of
605 BLIMP1 through PAX5 phosphorylation: a mechanism of antigen-triggering plasma cell
606 differentiation. *J. Immunol.* **188,** 6127–6134 (2012).
607 46. Baek, S., Goldstein, I. & Hager, G. L. Bivariate genomic footprinting detects changes in
608 transcription factor activity. *Cell Rep.* **19,** 1710–1722 (2017).
609 47. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring
610 transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*
611 **14,** 975–978 (2017).
612 48. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime
613 cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,**
614 576–589 (2010).
615 49. Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514,** 322–327 (2014).
616 50. Paz-Priel, I., Ghosal, A. K., Kowalski, J. & Friedman, A. D. C/EBPalpha or C/EBPalpha
617 oncoproteins regulate the intrinsic and extrinsic apoptotic pathways by direct interaction
618 with NF-kappaB p50 bound to the bcl-2 and FLIP gene promoters. *Leukemia* **23,** 365–374
619 (2009).
620 51. Cai, D. H. *et al.* C/EBP alpha:AP-1 leucine zipper heterodimers bind novel DNA elements,
621 activate the PU.1 promoter and direct monocyte lineage commitment more potently than

622 C/EBP alpha homodimers or AP-1. *Oncogene* **27,** 2772–2779 (2008).

623 52. Baker, S. J. *et al.* B-myb is an essential regulator of hematopoietic stem cell and myeloid
624 progenitor cell development. *Proc Natl Acad Sci USA* **111,** 3122–3127 (2014).

625 53. Gascoyne, D. M. *et al.* The basic leucine zipper transcription factor E4BP4 is essential for
626 natural killer cell development. *Nat. Immunol.* **10,** 1118–1124 (2009).

627 54. Sands, W. A., Copland, M. & Wheadon, H. Targeting self-renewal pathways in myeloid
628 malignancies. *Cell Commun. Signal.* **11,** 33 (2013).

629 55. Bussemaker, H. J., Li, H. & Siggia, E. D. Regulatory element detection using correlation
630 with expression. *Nat. Genet.* **27,** 167–171 (2001).

631 56. Boorsma, A., Lu, X.-J., Zakrzewska, A., Klis, F. M. & Bussemaker, H. J. Inferring
632 condition-specific modulation of transcription factor activity in yeast through regulon-based
633 analysis of genomewide expression. *PLoS ONE* **3,** e3112 (2008).