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

Ivan Berest[1,4*], 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
[4] Collaboration for joint PhD degree between EMBL and Heidelberg University, Faculty of Biosciences

[*] equal contribution

Transcription factor (TF) activity is an important read-out of cellular signalling pathways and thus to assess 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 globally as transcriptional repressor or activator. To this end, we introduce a widely applicable genome-wide method *diffTF* to assess differential TF activity and to classify TFs as activator or repressor (available at https://git.embl.de/grp-zaugg/diffTF). This is done by integrating any type of genome-wide chromatin accessibility data with RNA-Seq data and in-silico predicted TF binding sites. We corroborated the classification of TFs into repressors and activators by three independent analyses based on enrichments of active/repressive chromatin states, correlation of TF activity with gene expression, and activator- and repressor-specific chromatin footprints. To show the power of *diffTF,* we present two case studies*:* First, we applied *diffTF* in to a large ATAC-Seq/RNA-Seq dataset comparing mutated and unmutated chronic lymphocytic leukemia samples, where we identified dozens of known (40%) and potentially novel (60%) TFs that are differentially active. We were also able to classify almost half of them as either repressor and activator. Second,

we applied *diffTF* to a small ATAC-Seq/RNA-Seq data set comparing two cell types along the hematopoietic differentiation trajectory (multipotent progenitors – MPP – versus granulocyte-macrophage progenitors – GMP). Here we identified the known drivers of differentiation and found that the majority of the differentially active TFs are transcriptional activators. Overall, *diffTF* was able to recover the known TFs in both case studies, additionally identified TFs that have been less well characterized in the given condition, and provides a classification of the TFs into transcriptional activators and repressors.

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 read-out of signaling pathways that regulate them (Kim et al., 2007). Transcriptional regulation is largely influenced by cell type specific features such as cofactors, cooperative binding partners and local chromatin environment (Whyte et al., 2013). Adding to this complexity, many TFs can act as transcriptional activators and repressors depending on the cell type and growth condition (Han et al., 2015, 2018). Thus, to correctly interpret the downstream effects of a change in abundance of a given TF, it is important to understand its global mode of action within the specific context of the study.

TFs are typically lowly abundant proteins, which makes it difficult to detect them in proteomics experiments (Kim et al., 2007; Teng et al., 2008), and even if they can be detected, their abundance and 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 genomic TF binding events, provides information only for one TF at a time and does not detect global changes in TF activity unless specific experimental normalisation methods are used (e.g. spike-ins (Bonhoure et al., 2014)). Neither proteomics nor ChIP-Seq experiments can give any insights into their mode of action. Finally, luciferase assays can measure the activity and mode of action for a specific TF at a specific location and are therefore fairly low throughput (Komatsu et al., 2010; Liu et al., 2011). Databases like *TRRUST* (Han et al., 2015, 2018) collect

annotations of regulation modes of TFs based on literature text mining and provide a comprehensive resource for well-studied TF-target interactions. However, for the vast majority of TFs, there is no consensus about their molecular functional mode of action. A general framework for determining differential activity of TFs between conditions and classifying TFs into transcriptional activators and repressors in a cell-type and condition-specific manner is currently lacking.

Towards closing these gaps, we have developed an approach called *diffTF* to estimate global changes in TF activities across conditions or cell types, and classify TFs into activators and repressors based on the integration of genome-wide chromatin accessibility or histone mark ChIP-Seq data with predicted TF binding sites and RNA-Seq data. We corroborated the classification of TFs into repressors and activators by three independent analyses. First, we showed that repressors and activators were enriched in repressive and active chromatin states, respectively. Second, we confirmed that expression levels of repressors were anti-correlated with their target genes while they were positively correlated with their activators. Third, we obtained activator- and repressor-specific chromatin footprints based on TFs with a known mode of action, and found that this agreed very well with the footprints obtained from the factors as classified by *diffTF*.

We applied this approach to two case studies, one comparing two patient cohorts each with a large number of heterogeneous samples, the second comparing two cell types along a differentiation trajectory each with a small number of homogeneous samples. For the first study, we obtained a large ATAC-Seq dataset of chronic lymphocytic leukemia (CLL) samples from Rendeiro et al. ((Rendeiro et al., 2016)) from > 50 patients and a total of over 1 billion reads and show that the quantification of differential TF activity by *diffTF* is highly robust with respect to a wide range of parameter settings. We recapitulate many known TFs associated with CLL and propose several novel TFs that are involved in processes related to CLL biology such as the circadian clock. Furthermore, were were able to classify ~40% of these TFs (186) into activators and repressors, thus reconciling some biological processes that seem driven by activators and repressors at the same time. For the second case study we performed ATAC- and RNA-Seq on murine multipotent progenitors (MPP) versus granulocyte-macrophage progenitors (GMP) in quadruplicate. Again, with *diffTF* we were able to identify the known driver TFs of the

differentiation process, and we found that the majority of the highly differentially active TFs are acting as transcriptional activators.

Finally, the approach has been successfully applied to identify TFs that are specifically associated with TET2, an enzyme involved in DNA demethylation (Rasmussen et al., 2018), and to identify novel driving factors in pulmonary artery hypertension (Reyes-Palomares et al., in preparation).

RESULTS

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

We define TF activity as the effect of a TF on the state of chromatin as measured by chromatin accessibility assays (e.g. ATAC-Seq, DNase-Seq) or ChIP-Seq for active chromatin marks (e.g. H3K27ac). This definition is based on our 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 TFBS (Grubert et al., 2015). Even though the exact mechanisms of how changes in TF affinity translate to the chromatin level are unknown, TF activity likely plays a causal role in mediating the effect of the DNA variant onto chromatin marks (Liu et al., 2015). 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 read-out for its change in activity (**Suppl. Fig. 1**). A similar concept has been proposed in other tools that estimate TF activity based on ATAC or DHS data (Baek et al., 2017); (Schep et al., 2017).

Based on this concept, we developed *diffTF,* which is a computational approach to globally assess differential TF activity between two conditions (basic mode; **Fig. 1a, Suppl. Fig. 2**) and classify TFs into activators and repressors (classification mode, see below; **Fig. 1b**). It is based on any data that measures active/open chromatin, putative binding sites for TFs of interest , and optionally, for the classification mode only, matched RNA-Seq data. Briefly, for the basic mode, it requires *in silico* TFBS that can be obtained using position weight matrices (PWMs) from a database such as *HOCOMOCO* (Kulakovskiy et al., 2013) and a PWM scanning algorithm such as *PWMScan* (Ambrosini et al., 2018) for all TFs, or from a database of ChIP-Seq data, such as *ReMap* (Griffon et al., 2015). 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 the GC content of the respective TFBS (**Suppl. Fig. 3**). The

significance is assessed using either an empirical or analytical procedure. The former assesses the significance of the differential TF activity by comparing the real data against the distribution of values obtained from permuting the sample labels. The analytical procedure, which is particularly useful if the number of samples is too low for performing a reasonable number of permutations, calculates the p-value explicitly based on a t-statistic and estimated variance (see Methods for details and guidelines and **Suppl. Fig. 4**). In the basic mode, *diffTF* outputs differential activity and p-value for each TF, which together allow the identification of a set of TFs that show a significantly higher activity in one of the conditions (**Fig. 1a**).

Conceptual derivation of using open chromatin and RNA to classify TFs as transcriptional activators and repressors

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 (Han et al., 2018) (**Suppl. Fig. 12b**) (Han et al., 2018) indicates that the cell type or other external factors are important in determining a TF's main mode of action. Therefore, we devised a cell-type specific and data-driven, multiomics approach to classify TFs into activators and repressors within the framework of *diffTF* that can be run on top of the basic mode (classification mode). Our classification framework is based on the fact that increasing abundance of an activator TF results in increased transcription of its target genes (and vice versa for repressors). Yet transcription is difficult to measure since a typical RNA-Seq experiment measures the steady-state RNA level regulated by transcription and degradation. We reasoned that measures of chromatin activity (such as accessibility) is a more direct read-out for the mode of action of TFs. Based on this, we implemented an activator/repressor classification scheme in *diffTF* using RNA-Seq data as an estimate for TF abundance. For each TF, we calculate the correlations across individuals between its expression level and the ATAC-Seq signal in its putative target peak (**Fig. 1b**). Each TF is then classified (i) as an activator when it shows an overall positive correlation with the ATAC-Seq signal at its putative target sites, or (ii) as a repressor for an overall negative correlation, or (iii) as undetermined if the distribution of correlations is not significantly different from the peaks that did not overlap its putative binding sites (see also Methods and **Suppl. Fig. 12a**). The assumptions underlying this classification are tested in the context of case-study I (see below).

Case-study I: Quantify differential TF activity in a large ATAC-Seq dataset in CLL

We sought to apply *diffTF* to a large ATAC-Seq data set in a biological setting that is well-studied so that we could assess the technical robustness and its power to recover relevant biological signal. To do so, we identified a large ATAC-Seq data-set comparing different subtypes of the extensively studied cancer chronic lymphocytic leukemia (CLL) (Rendeiro et al., 2016) as an ideal dataset.

Chronic lymphocytic leukemia (CLL) is one of the most frequent types of cancer in the Western world, particularly among the elderly. 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). In M-CLL, B-cells go through normal affinity maturation with the aid of T-helper cells and undergo multiple rounds of IgHV somatic hypermutation to produce high affinity B-cell receptors (BCR). This process is essential for their development, survival and growth (Neu and Wilson, 2016). In contrast, U-CLL B-cells reach their affinity maturation point in an unregulated manner, and without involvement of T-helpers (Chiorazzi and Ferrarini, 2011). Overall, this leads to worse clinical outcomes such as shorter survival time and higher frequency of relapse after treatment (Furman et al., 2014).

The CLL dataset is comprised of ATAC-Seq data for a cohort of 88 CLL patients stratified by the mutation status of the IgHV locus (34 U-CLL, 50 M-CLL, 4 unclassified). After data processing and quality control, 25 and 27 U-CLL and M-CLL samples remained, and we applied *diffTF* in the basic mode to identify the differences in TF activities between U-CLL and M-CLL (**Suppl. Fig. 5-7**, see Supplementary Methods for more details).

In total we identified 67 TFs that are differentially active (FDR < 10%) between the two subtypes (**Fig. 2a**; **Suppl. Table 1**). About ~41% of the differentially active TFs have previously been associated with CLL and mostly (90%) agree with the reported condition (i.e., mutated or unmutated; **Suppl. Table 2**), thus providing a strong biological validation of the approach. The remaining 59% may represent novel candidate TFs that can advance our understanding of the disease etiology in general and the biological differences of mutated and unmutated patients in particular.

Before focusing on the biological interpretation of the specific TFs, we used this dataset to assess the technical robustness of *diffTF* with respect to TF binding site predictions. First, we

compared the results of *diffTF* when using putative vs. ChIP-Seq validated TF binding sites since predicting TF binding sites is inherently noisy and may result in many false positive sites when compared to ChIP-seq experiments (Landt et al., 2012). We observed a very strong correlation of the resulting TF activity differences (r=0.81; **Fig. 2b**, **Suppl. Fig. 8**), which indicates that *diffTF* is robust with respect to false positive binding sites. Second, we assessed the parameters for TF binding site predictions and found that neither the nucleotide composition of the predicted binding sites for *PWMScan* (**Suppl. Fig. 9**) nor the motif database (*JASPAR* vs. *HOCOMOCO*) had a strong impact on the differential TF activity (r=0.87, 0.62 and 0.69, respectively, **Fig. 2c-e**). Third, we assessed whether the size of the region surrounding the TF binding site from where signal was extracted (ranging from just the 7-25 bp long binding site to additional 600 bp upstream and downstream) had an impact on the results. The resulting differential TF activities were strongly correlated (r>0.9 for 50-600 bp and r=0.76 for the binding sites only; see Supplement and **Suppl. Fig. 10**). Additional robustness tests are described in the supplement.

We also assessed the potential of *diffTF* to detect differential TF activities in experiments with little biological signal. For this, we removed high-signal regions (i.e. differentially accessible peaks at 5% FDR; see Methods) and compared the resulting differential TF activities to those of the full set. We found that they were very similar for both sets (r=0.89), thus demonstrating the power of *diffTF* to capture the differential TF activities by summarising the subtle changes in chromatin accessibility across many TFBS genome-wide (**Fig. 2f**).

Finally, we assessed the dependency of *diffTF* results on sample size and sequencing depth. Intriguingly, we found the results highly congruent across a wide range of sample sizes and sequencing depths, with the majority of the significant TFs from the full dataset changing in the same direction in the subsampled data (**Fig. 2g**, **Suppl. Fig. 11**). 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 the subsampling results are dataset-specific and difficult to generalize they provide guidelines for the applicability of *diffTF* and are in line with single-cell ATAC-Seq data analysis that also show robustness for low coverage at the level of genome-wide summary statistics (Mezger et al., 2018).

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

210 sensible approach to overcome limitations such as low coverage and little underlying biological
211 signal.
212

diffTF proposes many novel TF candidates

213 We next focused on the biological interpretation of the differentially active TFs (FDR 10%)
214 between M-CLL and U-CLL patients (**Fig. 2a**; **Suppl. Table 1**). Since TF binding motifs, which
215 are the basis of *diffTF* (and any other tool that is based on predicted TF binding sites), can be
216 very similar between TFs of the same family, we decided to group TFs into TF-motif families
217 using the PWM clustering tool *Rsat* (Medina-Rivera et al., 2015); clusters available at
218 https://bit.ly/2J9TaaK). The resulting clusters showed overall consistent activity changes within
219 a TF family (**Fig. 3a**), with the exception of cluster 17 that can be explained by a prominent split
220 into two branches early in the clustering into NFAT and NFκB factors, which show more activity
221 in U-CLL and M-CLL, respectively (**Fig. 3c**).

222 The most active TF cluster in U-CLL is the IRF family and STAT2 (cluster 40; **Fig. 3d**), both of
223 which have been associated with disease onset and progression, and harbour several CLL
224 susceptibility loci (Arvaniti et al., 2011; Havelange et al., 2011; Slager et al., 2013). It is followed
225 by the PAX TFs (cluster 54), which affect B-cell to plasma cell differentiation (PAX5), that is
226 linked with cell survival and poor prognosis in CLL (Ghamlouch et al., 2015). Another prominent
227 set of regulators are the members of the AP-1 complex (cluster 4), which increase proliferation
228 and play an important role in driving the invasive nature of U-CLL (Mittal et al., 2013). Finally, we
229 found c-MYC, which is involved in cell proliferation and differentiation and is highly abundant in
230 U-CLL (Landau et al., 2015; Yeomans et al., 2016).

231 For M-CLL, we identified TFs that regulate and possibly reduce apoptosis, regulate cell cycle
232 and suggest normal functionality of B-cells through the classical BCR, NF-kB and Wnt signaling
233 pathways. The most active TF family in M-CLL patients is that of the POU TFs, also known as
234 Oct factors (cluster 12; **Fig 3b**), which regulate B-cell development and immunoglobulin
235 production, therefore promoting survival of the lymphoma cells (Heckman et al., 2006). This is
236 followed by the ROR factors (cluster 36), which together with Wnt5a activate NF-kB-dependent
237 survival signaling in CLL (Minami et al., 2010), and the GATA family (cluster 16), which is known
238 to prime HSCs towards the lymphoid lineage and increase self-renewal of the stem cells in CLL

(Kikushige et al., 2011). Other examples include the EGR family, whose motifs are enriched in aberrantly hypomethylated CpG sites in CLL (Oakes et al., 2016), PPARD, which has recently been linked to M-CLL through its effect on metabolic pathways in cancer cells (Li et al., 2017), and members of the GLI family, which are part of the Hedgehog signaling pathway and regulate apoptosis, thereby supporting survival of M-CLL cells (Kern et al., 2015).

Among the novel factors associated with U-CLL, we found several TFs (i.e. BMAL1, CLOCK, and NR1D1) that are involved in the regulation of the circadian clock, which has recently been proposed as hallmark of cancer (El-Athman and Relógio, 2018). Moreover, we found members of the basic helix-loop-helix family, such as BHE40, a regulator of mitotic division (D'Annibale et al., 2014), which is essential for the development of B1-a cells (Kreslavsky et al., 2017) and TFAP4, TFE3 and TFEB, for which there are known cases of gene-fusions in renal cell carcinoma (Kauffman et al., 2014). Another set of novel TFs more active in M-CLL are associated with pathways relevant for cancer- and B-cells such as escape from apoptosis (ZN784) (Kasim et al., 2017), regulation of cell cycle progression (ZBTB6) (Chevrier et al., 2014), and selection of B-cells and promotion of fetal B lymphopoesis (ARID3A) (Zhou et al., 2015). 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 (Coscia et al., 2011).
In summary, these results show that *diffTF* is able to recapitulate much of the known biology of the two subtypes of CLL and, in addition, identifies several more 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 accompanied by decreased target gene transcription (**Fig. 1b**). To do so, we employed the classification mode of *diffTF,* which integrates the ATAC-Seq data with RNA-Seq to classify TFs as activators or

repressors. For this, we first needed to test the global assumption that repressors and activators have an opposing effect on chromatin accessibility that underlies our classification framework. For activators, the expectation is that an increased TF abundance will increase the accessibility at its targets sites. For a repressor, however, the relationship between abundance and accessibility at its binding site is less straightforward: On the one hand the binding of the factor itself will increase the accessibility locally, while on the other hand, repression is globally associated with closed chromatin. To understand the effect of repressors and activators on chromatin accessibility and derive general principles, we compared the accessibility footprint (Tn5 insertion sites) of a well-known repressor (REST) and a well-known activator (STAT2) that are active in our cell type. We observed that for the repressor REST, there is indeed an increase in accessibility at its motif, which likely reflects the binding of the TF itself. Importantly, however, the accessibility drops to below the genome-wide average within 10 bp from the center of the motif (**Fig. 4a, bottom**). In contrast, for the activator STAT2, we observed increased chromatin accessibility outside its core binding site, which only slowly drops to the genome-wide average over a distance of >100 bp from the center of the motif, likely representing the effect of the TF on opening the surrounding chromatin (**Fig. 4a, top**). This shows that, while there is an increase in accessibility for repressors at the immediate binding site, the surrounding chromatin is highly compact while it is open for the activator. This is in line with a previous observation on EGR and SP4 (Baek et al., 2017) and justifies our classification approach implemented in *diffTF*.

Applying this reasoning to the CLL dataset, where RNA-Seq data was available for eight individuals (after QC; see supplement), we were able to classify 44% of the expressed TFs as either activators or repressors (**Fig 4b-f**; n=186). Among the top activators are the IRFs, which are well known transcriptional activators (Yanai et al., 2012) and which showed the same footprint pattern as STAT2 (**Fig. 4e**). Among the top repressors, we found PAX5, which has been shown to repress the activity of BLIMP-1 (Yasuda et al., 2012) and also shows a footprint similar to the repressor REST (**Fig. 4b,e**). To assess the binding properties of activators and repressors globally we performed an unbiased footprinting analysis for all TFs deemed significant in CLL. Importantly, we found that the aggregate signal across all repressors produced a footprint similar to that of REST, while the footprint for activators looks similar to STAT2, again indicating that repressors and activators have very distinct open chromatin footprints (**Fig. 4f**). Clustering of the footprints of the individual TFs revealed four major classes. Class I is characterized by low levels of Tn5 insertions in the motif and high levels in the

adjacent regions and its members are mainly classified as activators. Class II comprises of TFs with very low accessibility overall (mainly repressors). Class III contains TFs with high accessibility at the binding site and low accessibility in the adjacent regions, mainly classified as repressors but including a few activators. Finally, Class IV comprises TFs with a footprint that neither resembles an activator nor a repressor (**Suppl. Fig. 13**). This clustering indicates that TFs with a Class I footprint are likely classified as activators. In contrast, Class III footprints (like REST) are more likely classified as repressor, even though there might be some activator TFs that with a similar footprint. Overall, it seems that TF footprints correlate well with the molecular mode of action of a TF as identified by *diffTF*.

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. 14**), supporting the hypothesis that the molecular function of a TF is highly variable. The exceptions are cluster 40, containing mainly members of the IRF family, and cluster 17 that contains both NFAT and NFKB TFs, which are mostly classified as activators. 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 4d**). However, since BMAL1 is an activator, while CLOCK and NR1D1 are repressors, all three circadian factors are consistently more active in M-CLL, albeit with a contrary effect on their target genes.

To assess the validity of our repressor/activator classification, we chose three independent approaches. First, we used *chromHMM* chromatin states for primary B-cells from the Epigenomic Roadmap (Roadmap Epigenomics Consortium et al., 2015) to assess whether activators and repressors are preferentially located in active and repressive states, respectively. Indeed, we observed that the fraction of TFBS in active chromatin states was significantly larger for activators than for repressors, and vice versa for heterochromatin and repressive states (**Fig. 5a**, see also **Suppl. Fig. 15**), thus corroborating our classification of their molecular mode of action. Second, we assessed whether the direction in gene expression changes of TFs between U-CLL and M-CLL was in agreement with their differential activity and molecular mode of action. Again, our observations were in line with our expectations: activators showed a positive correlation of activity and expression change ($r=0.19$, $P=0.05$) while repressors showed a negative relationship ($r=-0.32$, $P=0.0033$; **Fig. 5b**). Third, we checked whether the expression of

target genes of a given TF changes in the same direction as its activity calculated by *diffTF,* regardless of the TF's classification as activator or repressor, and again found that this was the case for both activators and repressors (**Fig. 5c,** see Methods). In summary, these observations provide three independent lines of evidence that our approach implemented in *diffTF* is able to classify TFs globally by their mode of action. The fact that the correlations are in the expected direction but not perfect are likely reflecting that TFs are also regulated on the post-transcriptional level and thus show the limitation of using gene expression as a proxy for the abundance of the active form of TFs.

Case study II: Applying diffTF to small scale multiomics dataset

To assess the applicability of diffTF to small datasets, we decided to apply it to the well-studied mouse hematopoietic system. We generated ATAC-Seq and RNA-Seq profiles of multipotent progenitor cells (MPP; $Lin^{-}cKit^{+}Sca1^{+}$; $CD150^{-}CD48^{+}$), an early hematopoietic progenitor population capable of supporting multilineage blood production (Sun et al., 2014), as well as the more differentiated and myeloid-restricted granulocyte-monocyte progenitors (GMP; $Lin^{-}cKit^{+}Sca1^{-}$; $CD16/32^{+}$). The profiles obtained were processed using an in-house ATAC-Seq pipeline and *diffTF* (using the analytical procedure due to the small number of samples) 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 RNA-Seq data to filter out non-expressed TFs. The differential signal is dominated by an increase activity of the members of the well-known class of master regulators of myelopoiesis, the CEBP family (C/ebpα,-β,-δ,-ε,-γ) in GMPs (**Fig. 6a**, **Suppl. Fig. 16**). In addition, we observed higher activity of the MYC/MYB factors, which are known to be exclusively active in the GMPs (Baker et al., 2014) and in NFIL3, which is involved in the generation of natural killer cells (Gascoyne et al., 2009). Conversely, MPPs show a higher activity for IRF/STAT, ZEB1 and ITF2 (part of the Wnt signaling) as well as TFs from the Homeodomain (HXB7,HXA10) and Forkhead (FOXO3) families, all of which are associated with self-renewal of hematopoietic stem cells (Sands et al., 2013).

The small number of samples made the correlation-based classification of the TFs into activators and repressors unreliable. Therefore, we devised a second - less quantitative - approach for activator/repressor classification that is based on the TF footprint and differential

361 RNA-Seq expression. In short, we determined whether TF activity and expression level co-vary
362 in the same direction and combined this with visual inspection of the pattern of its footprint
363 (**Suppl. Fig. 17**). This allowed us to identify the set of activators that showed a clear activating
364 footprint as observed for the Class I TFs in the CLL data (**Fig. 6b**). Similar to the Class II and III
365 for CLL, the pattern was less clear for the repressor footprint clusters, which contain both
366 potential repressors and activators. Interestingly, the most differentially active TFs between MPP
367 and GMP are mainly classified as activators (CEBPs, NFIL3, IRFs) or have mixed evidence (i.e.
368 activator footprint, but inconsistent directionality of expression and activity such as DBP and
369 HLF). The most differentially active repressor we identified is JUN, whose difference in activity is
370 far below the activators, indicating that the differentiation process from MPP to GMP is mainly
371 driven by transcriptional activators.

372 Overall, these results show that *diffTF* is able to identify the known TFs that drive the
373 differentiation from MPP to GMP, thus demonstrating its power to detect signals also for a small
374 number of samples. Furthermore, we show how a qualitative classification scheme of TFs into
375 activators and repressors that is primarily based on TF footprints can be useful in comparisons
376 where the small number of samples does not allow a correlation-based classification.
377

Comparison with similar tools

378 We compared the few tools with a similar focus (Baek et al., 2017; Heinz et al., 2010; Schep et
379 al., 2017) with *diffTF* (**Suppl. Table 5**) both qualitatively (*chromVAR*, *BagFooT*, *HOMER*) and
380 quantitatively (*HOMER* and *chromVAR*). Overall, *diffTF* provides a more flexible and tailored
381 analysis framework due to the extensive choice of parameters, diagnostic plots, TFBS-specific
382 results, visualizations, and pipeline adjustability. As mentioned above, it is unique in its ability to
383 directly integrate RNA-Seq with ATAC-Seq data to classify TFs into activators and repressors.
384 Due to its flexibility, *diffTF* is computationally expensive, and we provide detailed instructions on
385 memory and time requirements in the documentation.

386 We first compared *diffTF* with a more traditional TF motif analysis such as *HOMER* (Heinz et al.,
387 2010), which looks at motif enrichment in a set of differentially accessible peaks. Strikingly, no
388 enriched motifs were found in M-CLL with *HOMER*, while the few discovered in U-CLL
389 correlated significantly with differential TF activity as computed by *diffTF* (**Suppl. Fig. 18,** see

Methods). This analysis highlights the power of *diffTF* to capture more signal than standard motif enrichment approaches.

To compare *diffTF* with an approach that is also based on TF activities we chose *chromVAR* and *BaGFoot* (Baek et al., 2017). We were unfortunately unable to run and adopt the BagFoot workflow for our CLL data due to missing example files, an incomplete documentation and unresponsiveness from the authors. For *chromVAR*, the results correlate very well overall, with correlation coefficients between 0.75 and 0.93 (Pearson) and 0.69-0.88 (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. 19a-b**, see Methods). Differences likely arise due to distinct methodological divergences such as comparing fold-changes for peaks (*chromVAR*) vs. binding sites (*diffTF*) or whether to compare the TF-specific effect against the mean effect across all TFs (*diffTF*) or not (*chromVAR*; see also Methods and **Suppl. Fig. 19c-f**). 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.

DISCUSSION

We presented a genome-wide method for quantifying differences in differential TF activity for a large set of TFs simultaneously, 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 in two case studies that *diffTF* is able to recover a change in activity for the TFs expected to drive the biological processes, thus demonstrating the biological validity of the method. In addition, we have extensively tested and demonstrated the technical robustness of *diffTF*. 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.

417 Calculating differential TF activity based on aggregating signal across the genome has been
418 proposed before based on the expression of putative target genes of a certain TF (Boorsma et
419 al., 2008; Bussemaker et al., 2001). Using the effect on chromatin instead of expression has
420 several advantages: first chromatin is a much simpler trait since gene expression is the sum of
421 transcription and degradation, thus increasing the power to detect differences. Second, there
422 are much more peaks than genes, thus allowing for better statistics and signal to noise ratio.
423 Finally, the effect on chromatin is much more local than on gene expression – in particular in
424 mammalian genomes, where genes can be regulated by distal enhancers. We have compared
425 differential TF activity calculated based on the average expression change of the target genes to
426 the output of *diffTF* and found that while the direction of activity between both methods is highly
427 correlated, the signal is much lower when using the expression of target genes instead of
428 chromatin accessibility at putative binding sites.

429 To demonstrate the power of *diffTF* for large but heterogeneous datasets, we have applied it to
430 identify and characterise differences between M-CLL and U-CLL from a publicly available
431 dataset of ATAC-Seq (1bn reads, 52 patients) and RNA-Seq. It is noteworthy that a TF motif
432 enrichment analysis on the significantly differentially accessible peaks did not reveal any factor
433 to be significantly enriched in M-CLL, indicating that in this case (as probably in many
434 patient-control studies) the key TFs are not necessarily switching their target enhancers and
435 promoters on and off, but rather mis-regulating many regions to a lesser extent. The advantage
436 of *diffTF* is that it can detect a slight shift in activity of the TF even if the signal differences at
437 each binding site a very low and rarely significant. It does so by averaging across all of a TFs
438 putative binding sites and is therefore more powerful than conventional enrichment analyses.

439 We have devised an approach within the *diffTF* framework to classify TFs into activators and
440 repressors based on the correlation of their expression level (RNA-Seq) and the activity of their
441 putative binding sites (ATAC-Seq). This information is highly relevant when interpreting the
442 action of TFs since it is important to know whether an upregulation of a TF would have a
443 positive or negative effect on chromatin (and therefore transcriptional) activity. Notably, this
444 classification could work even for datasets for which insufficient RNA-Seq data are available –
445 as we have shown for the MPP-GMP case – by jointly investigating TF footprints, differential
446 expression of the TF and 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 cofactors, 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.

Since many ATAC-seq experiments have a rather low number of samples, we also assessed the power of *diffTF* to uncover biology in small (but more controlled) experiments. In particular, we have performed a *diffTF* analysis to compare murine MPP and GMP (4 replicates each). Again we identified the major TFs driving the differentiation, and were able to qualitatively classify TFs into activators and repressors - in a correlation-independent approach. This classification revealed that the bulk of the change in chromatin accessibility during the differentiation from MPP to GMP is driven by activators. This case-study demonstrates the applicability of *diffTF* to small-scale data.

While similar methods have been proposed for analysing ATAC-Seq data (Baek et al., 2017; Schep et al., 2017), 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, *diffTF* is insensitive to sequence and 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 are a significant help when interpreting the effects of up/down regulation of a particular factors. (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 might allow 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 calculates differential TF activity across conditions and classifies TFs based on their molecular mode of action into activators and repressors. With this, *diffTF* can aid in formulating testable hypotheses and ultimately improve the understanding of 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 supported by 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, K.H. supervised the experiments, J.Z. supervised the study and C.A., I.B., and J.Z. wrote the manuscript.

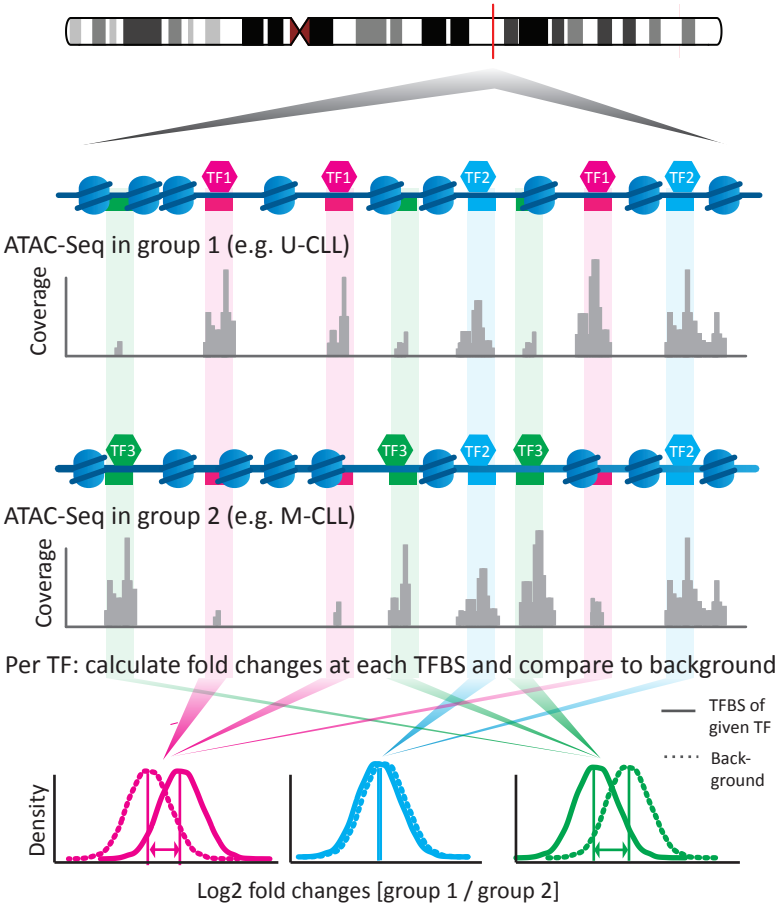
COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

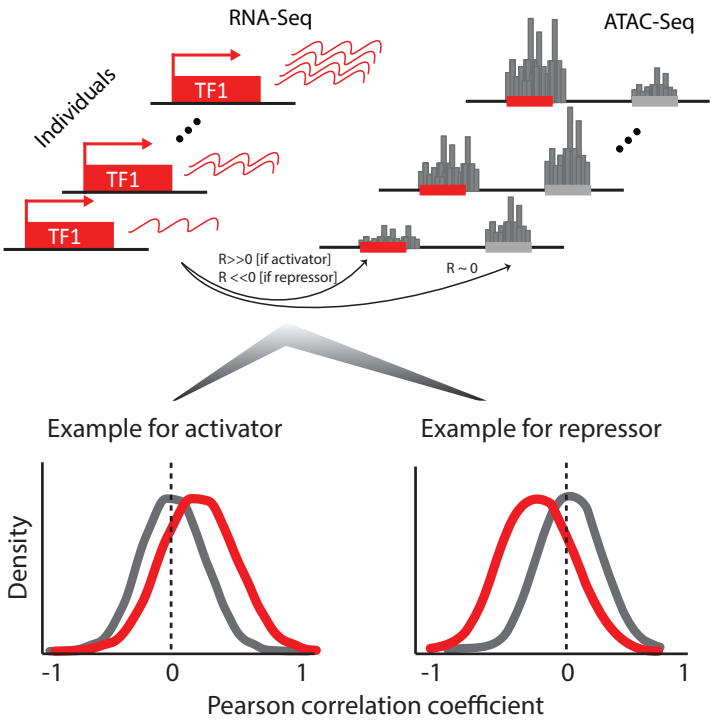
Figures and figure captions

Figure 1. *Schematic representation of the diffTF workflow*. **(a)** A simplified workflow illustrates the principle upon which *diffTF* is based: 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**). **(b)** *Schematic representation of the classification approach*: correlation of TF expression level with the accessibility of its target sites. If the distribution of correlations between a TF's RNA-level and the chromatin accessibility at its target sites is more positive than the background distribution (accessibility at non-target sites), the TF is classified as an activator in the particular biological environment; if negative, it is classified as a repressor. Correlations close to 0 are classified as undetermined.

A Schematic representation of *diffTF* - basic mode



B Schematic representation of *diffTF* - classification mode



516 **Figure 2.** *diffTF results for the CLL dataset, experimental validation and technical robustness*
517 *of the method.* **(a)** Volcano plot of differential TF activity between U-CLL (n=27) and M-CLL
518 (n=25) patients. The y-axis denotes statistical significance (-log10 transformed). TFs that pass
519 the significance threshold (5% FDR; dotted line) are labeled and colored according to their
520 novelty status (see text and **Suppl. Table. 2**). "#TFBS" denotes the number of predicted TF
521 binding sites in the peak regions for this analysis. (b)-(f): Technical robustness of *diffTF*.
522 Scatterplots of the differential TF activity from all TFs for two different *diffTF* analyses are
523 shown. Each point represents one TF. For (c-f), colors represent significance at 5% FDR (white
524 – not significant in either analysis; light green and light blue – significant for the analysis on the
525 x-axis or y-axis, respectively; purple – significant for both analyses). **(b)** Comparison of all
526 predicted TFBS and TFBS experimentally validated by ChIP-Seq data from ReMap. See also
527 **Suppl. Fig. 8**. **(c-d)** Comparison for different p-value thresholds in *PWMScan* to predict TFBS:
528 (c) standard vs. stringent (i.e., 1e-5 vs. 1e-6) and (d) standard vs. lenient (i.e., 1e-5 vs. 5e-5) for
529 a total of 628 TFs for which binding sites were retrieved for both scanning modes. **(e)**
530 Comparison of *diffTF* results based on *HOCOMOCO* v10 vs. *JASPAR* 2018 as input for the 412
531 TFs for which a motif was available in both databases. **(f)** Comparison of the full consensus
532 peak set (*allPeaks*) and only the non-differentially accessible peaks (*noDApeaks*; n=640 TFs
533 from *HOCOMOCO*). **(g)** Robustness analysis based on sequencing depth and sample size.
534 Each cell in the heatmap shows the fraction of TFs that showed the same direction of change as
535 in the full dataset for varying degrees of down-sampling sequencing depth and number of
536 samples, (5% FDR), averaged over 50 independent repetitions to minimize sampling noise.
537 Only TFs that were deemed significant in the full dataset are considered (see also **Suppl. Fig.**
538 **11**). Sequencing depth is shown as a fraction of the original data and median number of reads
539 across samples, while the number of samples is given as unmutated + mutated.

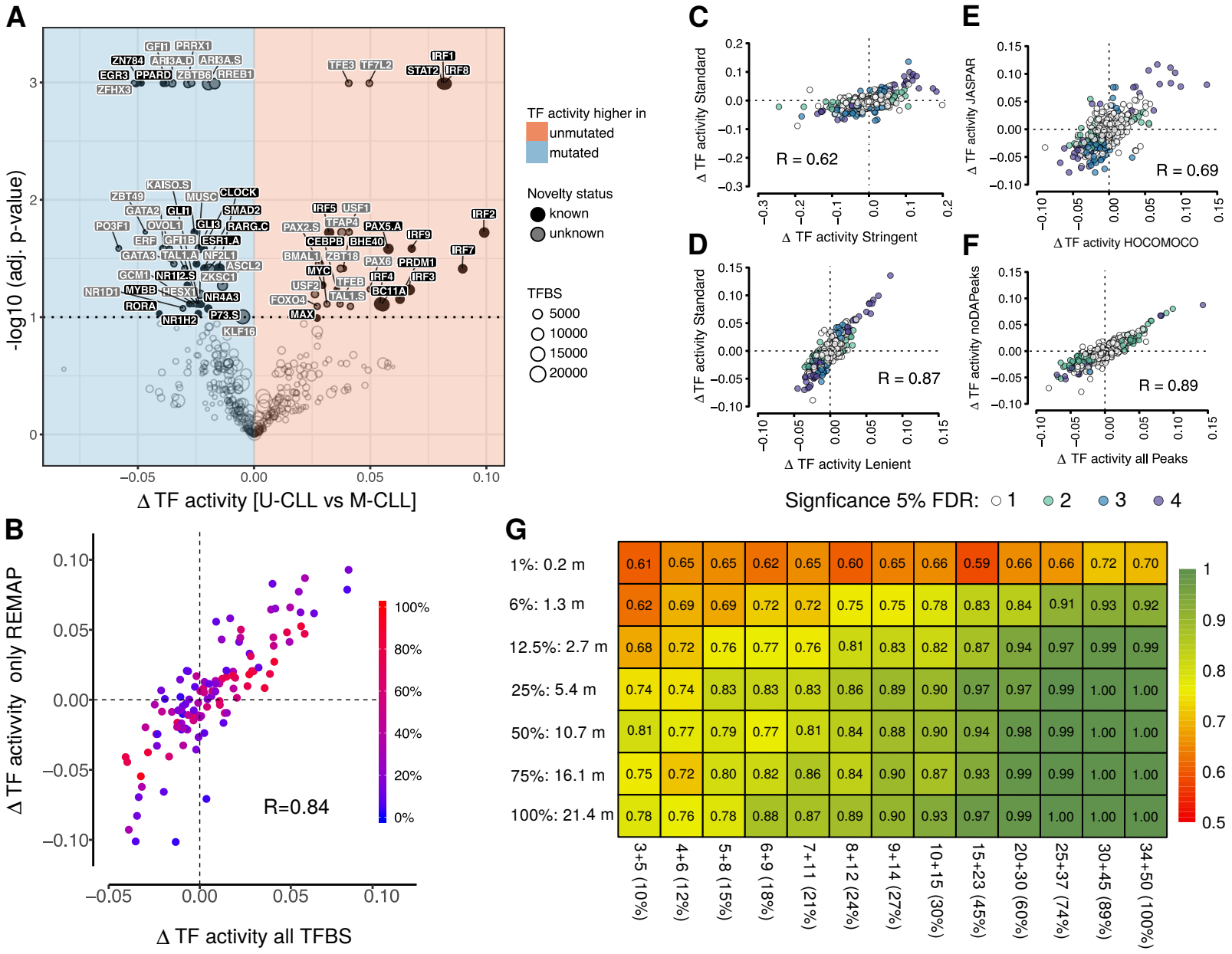
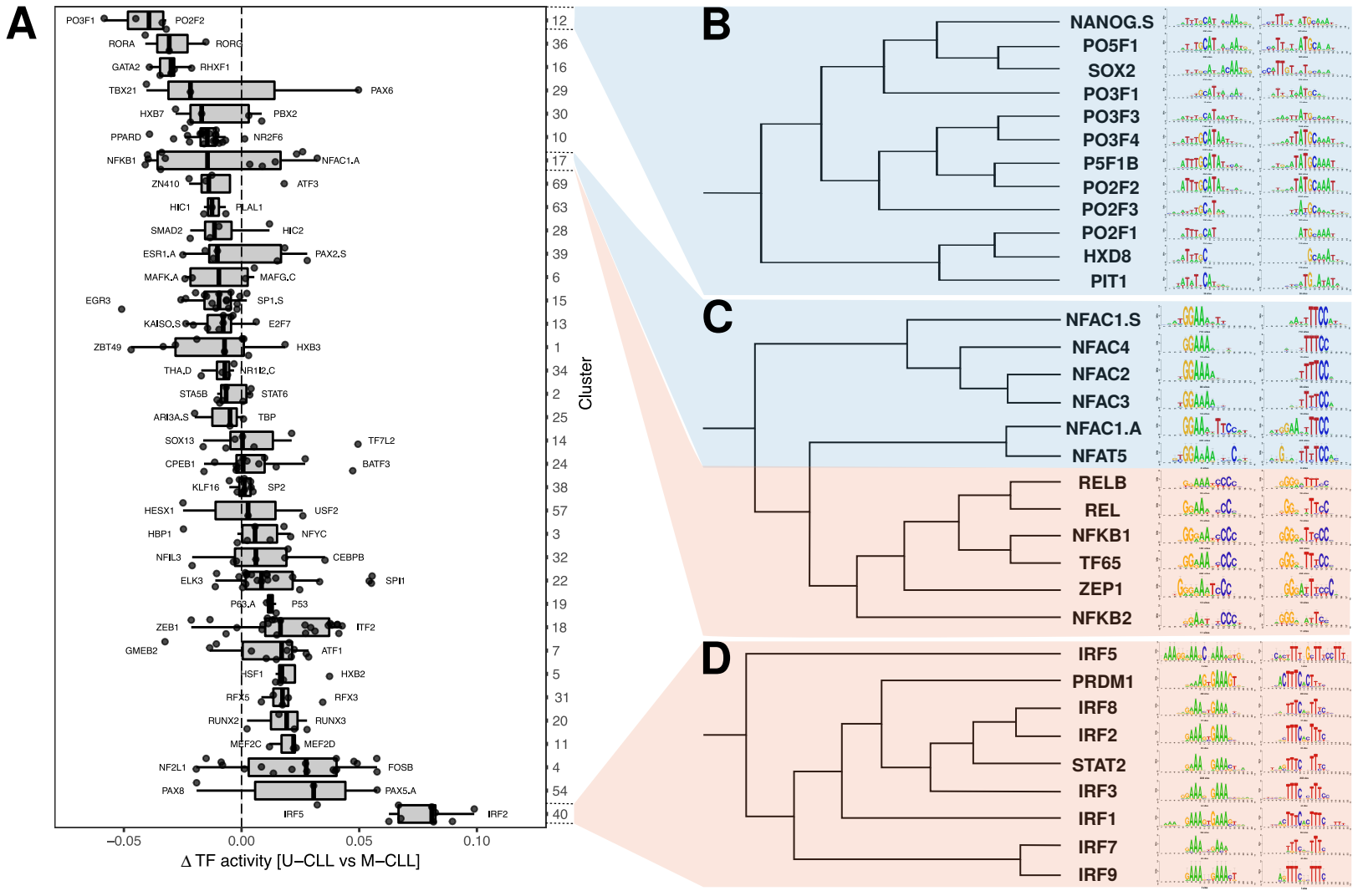

A
-log10 (adj. p-value)
Δ TF activity [U-CLL vs M-CLL]
GFI1
PRRX1
ZN784
ARI3A.D
ARI3A.S
EGR3
PPARD
ZBTB6
RREB1
ZFHX3
TFE3
TF7L2
IRF1
STAT2
IRF8
KAISO.S
ZBT49
MUSC
CLOCK
GATA2
GLI1
SMAD2
PO3F1
OVOL1
GLI3
RARG.C
ERF
GFI1B
ESR1.A
GATA3
TAL1.A
NF2L1
ASCL2
GCM1
NR1I2.S
ZKSC1
NR1D1
MYBB
HESX1
NR4A3
RORA
NR1H2
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%)

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


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

547 **Figure 4**. *Classification of TFs into activator or repressor based on RNA-Seq and ATAC-Seq*
548 *data.* **(a)** Exemplary footprints for a well-known activator (STAT2, top) and repressor (REST,
549 bottom). The x-axis depicts the distance in bp from the TFBS center, the y-axis denotes the
550 number of average Tn5 insertions, normalised to the library size and numbers of samples
551 across U-CLL and M-CLL. TFBS were predicted by *PWMscan* and only those overlapping with
552 open chromatin have been considered. The solid black line indicates the average insertion sites
553 within accessible chromatin. **(b)** Distributions of the Pearson correlations between TF
554 expression and ATAC-Seq signal at all putative TFBS (red line), and background distribution of
555 TFBS not containing the motif of interest (grey line) for two specific TFs, IRF2 (top) and PAX5
556 (bottom). **(c)** Summary of (b) across all TFs in the form of a heatmap showing the differences in
557 Pearson correlation of putative target peaks and background for each TF (i.e., subtracting the
558 black from the red line in (b)). Each TF is one row and is annotated as activator (green),
559 undetermined (grey) or repressor (red) as classified by *diffTF*. **(d)** Same as Fig. 2a, but with the
560 TFs labeled with their predicted role as activator (green), undetermined (black) or repressor
561 (red). See Fig. 2a for details. **(e)** Footprint analysis for an activator (IRF2, right) and repressor
562 (PAX5.A, left) as classified by *diffTF*. The top row shows the footprints separately for M-CLL and
563 U-CLL (blue and orange, respectively) based on the normalized number of Tn5 insertions, while
564 the bottom row highlights their differences (U-CLL - M-CLL) are shown. See (a) for axis
565 descriptions. **(f)** Summary footprint for all activators (top, green) and repressors (bottom, red).
566 See (a) for axis descriptions.

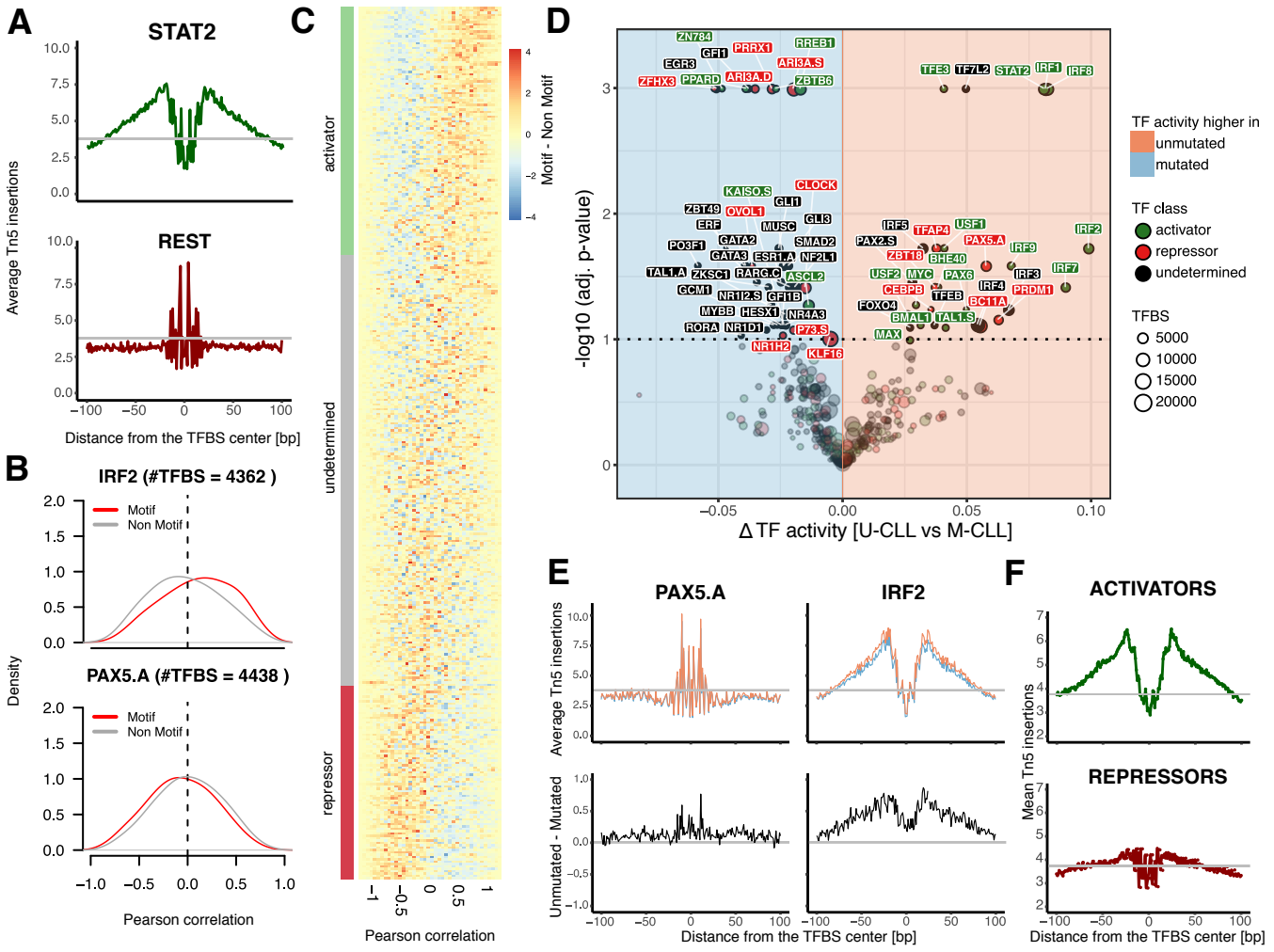

A
STAT2
REST
Average Tn5 insertions
Distance from the TFBS center [bp]
B
IRF2 (#TFBS = 4362)
PAX5.A (#TFBS = 4438)
Motif
Non Motif
Density
Pearson correlation
C
activator
undetermined
repressor
Motif - Non Motif
Pearson correlation
D
-log10 (adj. p-value)
Δ TF activity [U-CLL vs M-CLL]
ZN784
RREB1
GFI1
PRRX1
ARI3A.S
EGR3
ZFHX3
PPARD
ARI3A.D
ZBTB6
TFE3
TFL2
STAT2
IRF1
IRF8
KAISO.S
CLOCK
GLI1
ZBT49
OVOL1
ERF
GLI3
MUSC
IRF5
TFAP4
USF1
IRF2
PO3F1
GATA2
SMAD2
PAX2.S
PAX5.A
IRF9
GATA3
ESR1.A
NF2L1
ZBT18
BHE40
TAL1.A
ZKSC1
RARG.G
ASCL2
USF2
MYC
PAX6
IRF3
IRF7
GCM1
NR1I2.S
GFI1B
CEBPB
IRF4
PRDM1
MYB
HESX1
NR4A3
TFEB
BC11A
FOXO4
RORA
NR1D1
P73.S
BMAL1
TAL1.S
MAX
NR1H2
KLF16
TF activity higher in
unmutated
mutated
TF class
activator
repressor
undetermined
TFBS
5000
10000
15000
20000
E
PAX5.A
IRF2
Average Tn5 insertions
Unmutated - Mutated
Distance from the TFBS center [bp]
F
ACTIVATORS
REPRESSORS
Mean Tn5 insertions
Distance from the TFBS center [bp]

Figure 5. *Validations for the activator and repressor classification and downstream analyses.* **(a)** Boxplots showing the fraction TFBS overlapping with specific chromatin states as defined by *chromHMM* (Roadmap Epigenomics Consortium et al., 2015) for all activators (green) and repressors (red) are shown. Only *chromHMM* states with significant differences (Wilcoxon test; p-value < 0.05) between activators and repressors are displayed. **(b)** 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. **(c)** 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 correlation strength (see Fig. 4d). (OR=odds ratio, r=Pearson correlation coefficient).

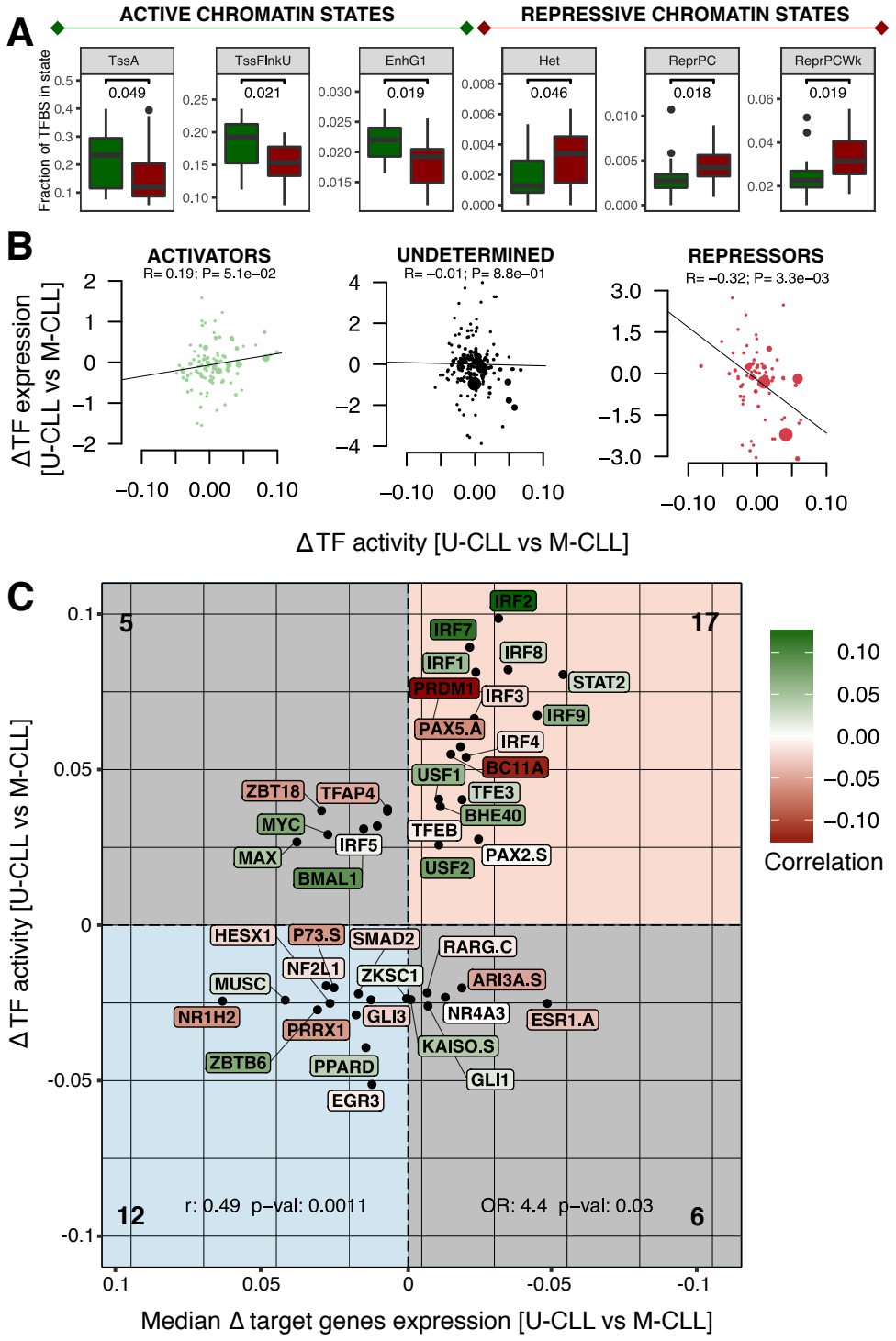
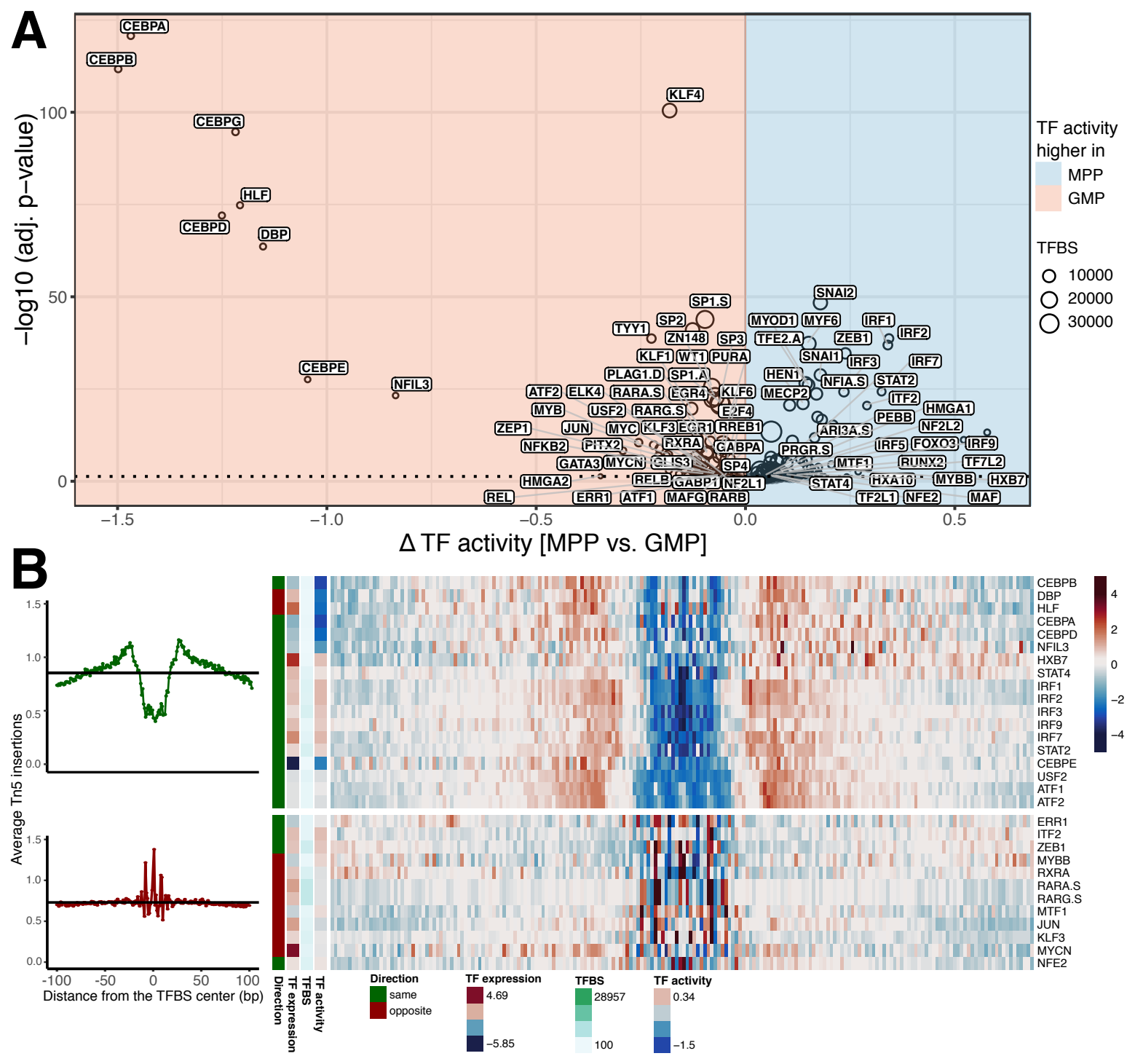

A
ACTIVE CHROMATIN STATES
REPRESSIVE CHROMATIN STATES
Fraction of TFBS in state
TssA
0.049
TssFlnkU
0.021
EnhG1
0.019
Het
0.046
ReprPC
0.018
ReprPCWk
0.019
B
ACTIVATORS
R= 0.19; P= 5.1e−02
UNDETERMINED
R= −0.01; P= 8.8e−01
REPRESSORS
R= −0.32; P= 3.3e−03
ΔTF expression [U-CLL vs M-CLL]
ΔTF activity [U-CLL vs M-CLL]
C
5
17
12
6
r: 0.49 p−val: 0.0011
OR: 4.4 p−val: 0.03
IRF2
IRF7
IRF8
IRF1
STAT2
PRDM1
IRF3
IRF9
PAX5.A
IRF4
BC11A
USF1
TFE3
ZBT18
TFAP4
BHE40
MYC
TFEB
IRF5
PAX2.S
MAX
USF2
BMAL1
HESX1
P73.S
SMAD2
RARG.C
NF2L1
ZKSC1
ARI3A.S
MUSC
NR1H2
GLI3
NR4A3
ESR1.A
PRRX1
KAISO.S
ZBTB6
PPARD
GLI1
EGR3
Correlation
Median Δ target genes expression [U-CLL vs M-CLL]
ΔTF activity [U-CLL vs M-CLL]

Figure 6. *diffTF recapitulates known TFs that drive the differentiation from MPP to GMP and shows a similar activator/repressor cluster as in the CLL data*. **(a)** Volcano plot of differential TF activity between MPP (n=4) and GMP (n=4) cells. Due to the high number of significant TFs only the most significant are labeled. The full list is available in Suppl. Table 4. **(b)** The footprints are for TFs in two selected clusters that represent the activators and repressors, respectively are shown as heatmap (right) and aggregate plots (left; see **Suppl. Fig. 17** for the full heatmap). Only TFs that were significantly differentially active and all significantly differentially expressed (adj. p-value < 0.05 for both) are displayed. Colors represent footprint strength, while white denotes the value of the genomic background in the consensus peakset. Clusters were defined using hierarchical clustering with the *ward.D2* method (clustering tree omitted for clarity)(see **Suppl. Fig. 17**). For the cluster summary footprints at the left, we divided each footprint value by the mean value of each cluster to highlight the differences in the surrounding chromatin structure. The direction of TF expression and TF activity is in analogy to what is described in the text. “Direction” denotes whether expression and TF activity have the same or opposite sign.


A
−log10 (adj. p−value)
Δ TF activity [MPP vs. GMP]
TF activity higher in
MPP
GMP
TFBS
10000
20000
30000
CEBPA
CEBPB
CEBPG
HLF
CEBPD
DBP
KLF4
CEBPE
NFIL3
SP1.S
SP2
TYY1
ZN148
SP3
KLF1
WT1
PURA
PLAG1.D
SP1.A
ATF2
ELK4
RARA.S
EGR4
KLF6
MYB
USF2
RARG.S
E2F4
ZEP1
JUN
MYC
KLF3
EGR1
RREB1
NFKB2
PITX2
RXRA
GABPA
GATA3
MYCN
GLIS3
SP4
HMGA2
RELB
GABP1
NF2L1
REL
ERR1
ATF1
MAFG
RARB
SNAI2
MYOD1
MYF6
IRF1
IRF2
TFE2.A
ZEB1
SNAI1
IRF3
IRF7
HEN1
NFIA.S
STAT2
MECP2
ITF2
HMGA1
PEBB
ARI3A.S
NF2L2
PRGR.S
IRF5
FOXO3
IRF9
MTF1
RUNX2
TF7L2
STAT4
HXA10
MYBB
HXB7
TF2L1
NFE2
MAF
B
Average Tn5 insertions
Distance from the TFBS center (bp)
Direction
TF expression
TFBS
TF activity
Direction
same
opposite
TF expression
4.69
−5.85
TFBS
28957
100
TF activity
0.34
−1.5
CEBPB
DBP
HLF
CEBPA
CEBPD
NFIL3
HXB7
STAT4
IRF1
IRF2
IRF3
IRF9
IRF7
STAT2
CEBPE
USF2
ATF1
ATF2
ERR1
ITF2
ZEB1
MYBB
RXRA
RARA.S
RARG.S
MTF1
JUN
KLF3
MYCN
NFE2

REFERENCES

Ambrosini, G., Groux, R., and Bucher, P. (2018). PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. Bioinformatics *34*, 2483–2484.

Arvaniti, E., Ntoufa, S., Papakonstantinou, N., Touloumenidou, T., Laoutaris, N., Anagnostopoulos, A., Lamnissou, K., Caligaris-Cappio, F., Stamatopoulos, K., Ghia, P., et al. (2011). Toll-like receptor signaling pathway in chronic lymphocytic leukemia: distinct gene expression profiles of potential pathogenic significance in specific subsets of patients. Haematologica *96*, 1644–1652.

Baek, S., Goldstein, I., and Hager, G.L. (2017). Bivariate genomic footprinting detects changes in transcription factor activity. Cell Rep. *19*, 1710–1722.

Baker, S.J., Ma'ayan, A., Lieu, Y.K., John, P., Reddy, M.V.R., Chen, E.Y., Duan, Q., Snoeck, H.-W., and Reddy, E.P. (2014). B-myb is an essential regulator of hematopoietic stem cell and myeloid progenitor cell development. Proc Natl Acad Sci USA *111*, 3122–3127.

Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I.M., Herr, W., Hernandez, N., Delorenzi, M., et al. (2014). Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res. *24*, 1157–1168.

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

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

Chevrier, S., Emslie, D., Shi, W., Kratina, T., Wellard, C., Karnowski, A., Erikci, E., Smyth, G.K., Chowdhury, K., Tarlinton, D., et al. (2014). The BTB-ZF transcription factor Zbtb20 is driven by Irf4 to promote plasma cell differentiation and longevity. J. Exp. Med. *211*, 827–840.

Chiorazzi, N., and Ferrarini, M. (2011). Cellular origin(s) of chronic lymphocytic leukemia: cautionary notes and additional considerations and possibilities. Blood *117*, 1781–1791.

Coscia, M., Pantaleoni, F., Riganti, C., Vitale, C., Rigoni, M., Peola, S., Castella, B., Foglietta, M., Griggio, V., Drandi, D., et al. (2011). IGHV unmutated CLL B cells are more prone to spontaneous apoptosis and subject to environmental prosurvival signals than mutated CLL B cells. Leukemia *25*, 828–837.

D'Annibale, S., Kim, J., Magliozzi, R., Low, T.Y., Mohammed, S., Heck, A.J.R., and Guardavaccaro, D. (2014). Proteasome-dependent degradation of transcription factor activating enhancer-binding protein 4 (TFAP4) controls mitotic division. J. Biol. Chem. *289*, 7730–7737.

El-Athman, R., and Relógio, A. (2018). Escaping circadian regulation: an emerging hallmark of cancer? Cell Syst. *6*, 266–267.

Furman, R.R., Sharman, J.P., Coutre, S.E., Cheson, B.D., Pagel, J.M., Hillmen, P., Barrientos, J.C., Zelenetz, A.D., Kipps, T.J., Flinn, I., et al. (2014). Idelalisib and rituximab in relapsed

631 chronic lymphocytic leukemia. N. Engl. J. Med. *370*, 997–1007.

632 Gascoyne, D.M., Long, E., Veiga-Fernandes, H., de Boer, J., Williams, O., Seddon, B., Coles,
633 M., Kioussis, D., and Brady, H.J.M. (2009). The basic leucine zipper transcription factor E4BP4
634 is essential for natural killer cell development. Nat. Immunol. *10*, 1118–1124.

635 Ghamlouch, H., Darwiche, W., Hodroge, A., Ouled-Haddou, H., Dupont, S., Singh, A.R.,
636 Guignant, C., Trudel, S., Royer, B., Gubler, B., et al. (2015). Factors involved in CLL
637 pathogenesis and cell survival are disrupted by differentiation of CLL B-cells into
638 antibody-secreting cells. Oncotarget *6*, 18484–18503.

639 Griffon, A., Barbier, Q., Dalino, J., van Helden, J., Spicuglia, S., and Ballester, B. (2015).
640 Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory
641 landscape. Nucleic Acids Res. *43*, e27.

642 Grubert, F., Zaugg, J.B., Kasowski, M., Ursu, O., Spacek, D.V., Martin, A.R., Greenside, P.,
643 Srivas, R., Phanstiel, D.H., Pekowska, A., et al. (2015). Genetic control of chromatin states in
644 humans involves local and distal chromosomal interactions. Cell *162*, 1051–1065.

645 Han, H., Shim, H., Shin, D., Shim, J.E., Ko, Y., Shin, J., Kim, H., Cho, A., Kim, E., Lee, T., et al.
646 (2015). TRRUST: a reference database of human transcriptional regulatory interactions. Sci.
647 Rep. *5*, 11432.

648 Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C.Y., Lee, M., Kim, E., et
649 al. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional
650 regulatory interactions. Nucleic Acids Res. *46*, D380–D386.

651 Havelange, V., Pekarsky, Y., Nakamura, T., Palamarchuk, A., Alder, H., Rassenti, L., Kipps, T.,
652 and Croce, C.M. (2011). IRF4 mutations in chronic lymphocytic leukemia. Blood *118*,
653 2827–2829.

654 Heckman, C.A., Duan, H., Garcia, P.B., and Boxer, L.M. (2006). Oct transcription factors
655 mediate t(14;18) lymphoma cell survival by directly regulating bcl-2 expression. Oncogene *25*,
656 888–898.

657 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C.,
658 Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription
659 factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell
660 *38*, 576–589.

661 Kasim, V., Xie, Y.-D., Wang, H.-M., Huang, C., Yan, X.-S., Nian, W.-Q., Zheng, X.-D., Miyagishi,
662 M., and Wu, S.-R. (2017). Transcription factor Yin Yang 2 is a novel regulator of the p53/p21
663 axis. Oncotarget *8*, 54694–54707.

664 Kauffman, E.C., Ricketts, C.J., Rais-Bahrami, S., Yang, Y., Merino, M.J., Bottaro, D.P.,
665 Srinivasan, R., and Linehan, W.M. (2014). Molecular genetics and cellular features of TFE3 and
666 TFEB fusion kidney cancers. Nat. Rev. Urol. *11*, 465–475.

667 Kern, D., Regl, G., Hofbauer, S.W., Altenhofer, P., Achatz, G., Dlugosz, A., Schnidar, H., Greil,
668 R., Hartmann, T.N., and Aberger, F. (2015). Hedgehog/GLI and PI3K signaling in the initiation

669 and maintenance of chronic lymphocytic leukemia. Oncogene *34*, 5341–5351.

670 Kikushige, Y., Ishikawa, F., Miyamoto, T., Shima, T., Urata, S., Yoshimoto, G., Mori, Y., Iino, T.,
671 Yamauchi, T., Eto, T., et al. (2011). Self-renewing hematopoietic stem cell is the primary target in
672 pathogenesis of human chronic lymphocytic leukemia. Cancer Cell *20*, 246–259.

673 Kim, R., Emi, M., and Tanabe, K. (2007). Cancer immunoediting from immune surveillance to
674 immune escape. Immunology *121*, 1–14.

675 Komatsu, M., Kurokawa, H., Waguri, S., Taguchi, K., Kobayashi, A., Ichimura, Y., Sou, Y.-S.,
676 Ueno, I., Sakamoto, A., Tong, K.I., et al. (2010). The selective autophagy substrate p62
677 activates the stress responsive transcription factor Nrf2 through inactivation of Keap1. Nat. Cell
678 Biol. *12*, 213–223.

679 Kreslavsky, T., Vilagos, B., Tagoh, H., Poliakova, D.K., Schwickert, T.A., Wöhner, M., Jaritz, M.,
680 Weiss, S., Taneja, R., Rossner, M.J., et al. (2017). Essential role for the transcription factor
681 Bhlhe41 in regulating the development, self-renewal and BCR repertoire of B-1a cells. Nat.
682 Immunol. *18*, 442–455.

683 Kulakovskiy, I.V., Medvedeva, Y.A., Schaefer, U., Kasianov, A.S., Vorontsov, I.E., Bajic, V.B.,
684 and Makeev, V.J. (2013). HOCOMOCO: a comprehensive collection of human transcription
685 factor binding sites models. Nucleic Acids Res. *41*, D195-202.

686 Landau, D.A., Tausch, E., Taylor-Weiner, A.N., Stewart, C., Reiter, J.G., Bahlo, J., Kluth, S.,
687 Bozic, I., Lawrence, M., Böttcher, S., et al. (2015). Mutations driving CLL and their evolution in
688 progression and relapse. Nature *526*, 525–530.

689 Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein,
690 B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the
691 ENCODE and modENCODE consortia. Genome Res. *22*, 1813–1831.

692 Li, Y.J., Sun, L., Shi, Y., Wang, G., Wang, X., Dunn, S.E., Iorio, C., Screaton, R.A., and Spaner,
693 D.E. (2017). PPAR-delta promotes survival of chronic lymphocytic leukemia cells in energetically
694 unfavorable conditions. Leukemia *31*, 1905–1914.

695 Liu, L., Jin, G., and Zhou, X. (2015). Modeling the relationship of epigenetic modifications to
696 transcription factor binding. Nucleic Acids Res. *43*, 3873–3885.

697 Liu, Z., Lee, J., Krummey, S., Lu, W., Cai, H., and Lenardo, M.J. (2011). The kinase LRRK2 is a
698 regulator of the transcription factor NFAT that modulates the severity of inflammatory bowel
699 disease. Nat. Immunol. *12*, 1063–1070.

700 Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J.A., Delerce, J.,
701 Jaeger, S., Blanchet, C., Vincens, P., Caron, C., et al. (2015). RSAT 2015: regulatory sequence
702 analysis tools. Nucleic Acids Res. *43*, W50-6.

703 Mezger, A., Klemm, S., Mann, I., Brower, K., Mir, A., Bostick, M., Farmer, A., Fordyce, P.,
704 Linnarsson, S., and Greenleaf, W. (2018). High-throughput chromatin accessibility profiling at
705 single-cell resolution. Nat. Commun. *9*, 3647.

706 Minami, Y., Oishi, I., Endo, M., and Nishita, M. (2010). Ror-family receptor tyrosine kinases in

707 noncanonical Wnt signaling: their implications in developmental morphogenesis and human
708 diseases. Dev. Dyn. *239*, 1–15.

709 Mittal, A.K., Chaturvedi, N.K., Rohlfsen, R.A., Gupta, P., Joshi, A.D., Hegde, G.V., Bociek, R.G.,
710 and Joshi, S.S. (2013). Role of CTLA4 in the proliferation and survival of chronic lymphocytic
711 leukemia. PLoS ONE *8*, e70352.

712 Neu, K.E., and Wilson, P.C. (2016). Taking the broad view on B cell affinity maturation. Immunity
713 *44*, 518–520.

714 Oakes, C.C., Seifert, M., Assenov, Y., Gu, L., Przekopowitz, M., Ruppert, A.S., Wang, Q.,
715 Imbusch, C.D., Serva, A., Koser, S.D., et al. (2016). DNA methylation dynamics during B cell
716 maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. Nat.
717 Genet. *48*, 253–264.

718 Rasmussen, K., Berest, I., Kessler, S., Nishimura, K., Simon-Carrasco, L., Vassilou, G.S.,
719 Pedersen, M.T., Christensen, J., Zaugg, J., and Helin, K. (2018). TET2 binding to enhancers
720 facilitates transcription factor recruitment in hematopoietic cells. BioRxiv.

721 Rendeiro, A.F., Schmidl, C., Strefford, J.C., Walewska, R., Davis, Z., Farlik, M., Oscier, D., and
722 Bock, C. (2016). Chromatin accessibility maps of chronic lymphocytic leukaemia identify
723 subtype-specific epigenome signatures and transcription regulatory networks. Nat. Commun. *7*,
724 11938.

725 Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A.,
726 Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of
727 111 reference human epigenomes. Nature *518*, 317–330.

728 Sands, W.A., Copland, M., and Wheadon, H. (2013). Targeting self-renewal pathways in myeloid
729 malignancies. Cell Commun. Signal. *11*, 33.

730 Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring
731 transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods *14*,
732 975–978.

733 Slager, S.L., Caporaso, N.E., de Sanjose, S., and Goldin, L.R. (2013). Genetic susceptibility to
734 chronic lymphocytic leukemia. Semin. Hematol. *50*, 296–302.

735 Sun, J., Ramos, A., Chapman, B., Johnnidis, J.B., Le, L., Ho, Y.-J., Klein, A., Hofmann, O., and
736 Camargo, F.D. (2014). Clonal dynamics of native haematopoiesis. Nature *514*, 322–327.

737 Teng, M.W.L., Swann, J.B., Koebel, C.M., Schreiber, R.D., and Smyth, M.J. (2008).
738 Immune-mediated dormancy: an equilibrium with cancer. J. Leukoc. Biol. *84*, 988–993.

739 Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee,
740 T.I., and Young, R.A. (2013). Master transcription factors and mediator establish
741 super-enhancers at key cell identity genes. Cell *153*, 307–319.

742 Yanai, H., Negishi, H., and Taniguchi, T. (2012). The IRF family of transcription factors:
743 Inception, impact and implications in oncogenesis. Oncoimmunology *1*, 1376–1386.

744 Yasuda, T., Hayakawa, F., Kurahashi, S., Sugimoto, K., Minami, Y., Tomita, A., and Naoe, T.
745 (2012). B cell receptor-ERK1/2 signal cancels PAX5-dependent repression of BLIMP1 through
746 PAX5 phosphorylation: a mechanism of antigen-triggering plasma cell differentiation. J.
747 Immunol. *188*, 6127–6134.

748 Yeomans, A., Thirdborough, S.M., Valle-Argos, B., Linley, A., Krysov, S., Hidalgo, M.S.,
749 Leonard, E., Ishfaq, M., Wagner, S.D., Willis, A.E., et al. (2016). Engagement of the B-cell
750 receptor of chronic lymphocytic leukemia cells drives global and MYC-specific mRNA
751 translation. Blood *127*, 449–457.

752 Zhou, Y., Li, Y.-S., Bandi, S.R., Tang, L., Shinton, S.A., Hayakawa, K., and Hardy, R.R. (2015).
753 Lin28b promotes fetal B lymphopoiesis through the transcription factor Arid3a. J. Exp. Med. *212*,
754 569–580.