Identification and quantitative characterization of protein-DNA interactions with the native holdup assay: GATA1 regulation of the plasma membrane calcium pump

Boglarka Zambo[1*], Bastien Morlet[2], Luc Negroni[2], Orsolya Mózner[3], Balázs Sarkadi[3], Gilles Travé[1*], Gergo Gogl[2*]

[1] Équipe Labellisée Ligue contre le cancer, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), INSERM U1258/CNRS UMR 7104/Université de Strasbourg, Illkirch, France

[2] Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), INSERM U1258/CNRS UMR 7104/Université de Strasbourg, Illkirch, France

[3] HUN-REN Research Centre for Natural Sciences, Institute of Molecular Life Sciences, Budapest, Hungary

***Corresponding authors:** Boglarka Zambo: zambob@igbmc.fr, Gilles Travé: travegi@igbmc.fr, Gergo Gogl: gergo.gogl@inserm.fr

Polymorphisms in regulatory regions of the genome can alter RNA and protein expression and are implicated in a wide range of diseases. It has been recently discovered that a common haplotype in the erythroid-specific promoter of the *ATP2B4* gene alters the expression of Plasma Membrane Calcium ATPase 4 (PMCA4) in red blood cells and reduces malaria susceptibility. Identification of transcription factors that bind to such regions can be challenging and is of particular interest. Here, we introduce the DNA-native holdup assay, a versatile method that allows proteome-wide identification of proteins binding to short DNA fragments. It is a quantitative retention assay that can measure altered binding affinity to different polymorphic DNA sequences. Using this approach, we present direct evidence that GATA1 is the transcription factor whose binding is most affected by the haplotype in the erythroid-specific promoter of the *ATP2B4* gene. Furthermore, we demonstrate that this method is suitable for the convenient biophysical characterization of DNA binding of different mutant versions of transcription factors, allowing the rapid evaluation of mutations found in patients. Analyzing the DNA binding properties of six different mutants of the GATA1 protein leading to severe anemia in humans revealed altered DNA-binding properties for R307C and for the short version (GATA1s) lacking the N-terminal transactivation domain (N-TAD) of the protein. This method provides a simple quantitative approach to characterize DNA-protein interactions, and to assess the effect of polymorphic DNA variants or mutations in transcription factors on DNA binding.

Introduction

Genetic variations determine many physiological traits and can contribute to individual susceptibility to different diseases (1, 2). The vast majority of these variants are located outside the protein-coding regions, in specific regulatory parts of the genome, such as promoters and enhancers, and can regulate the transcription of one or more genes (3, 4). The first step to understand the mechanism of action of these regulatory variations is to identify the transcription factor whose binding is altered in the presence of the variant. In order to find the binding partners for a given site proteome-wide, one may choose an interactomic strategy (5, 6). However, commonly used interactomic methods are mainly restricted to study protein-protein interactions. Moreover, these approaches provide qualitative data, which means that they only discriminate between detected or undetected interactions, and in the case where genetic variations only cause subtle modulations in transcription factor binding, this resolution is insufficient (7). Concerning DNA-protein interactions, ChIP-seq is capable to identify specific binding sites for a given protein genome-wide, however it cannot determine affinity constants and measuring the effect of individual polymorphisms could be challenging (8, 9). By contrast, electrophoretic mobility shift assays can provide quantitative information on certain binding properties, yet they cannot be used to identify transcription factors because they only probe one protein at a time (10). Therefore, a versatile technique is needed to identify and simultaneously quantify the binding affinity of DNA-protein interactions to assess how polymorphic variants alter the DNA-binding properties of different transcription factors.

The *ATP2B4* gene encoding the calcium exporter protein PMCA4b contains an alternative erythroid-specific promoter, located just upstream of the second exon, that drives the expression of the calcium pump specifically in erythroid lineages (Figure 1A) (11, 12). In this erythroid promoter, a haplotype has been identified that causes lower expression of the protein in mature red blood cells, reduced mean corpuscular hemoglobin concentration in the blood and provides certain degree of protection against malaria (11, 13, 14). We have previously identified two SNP-containing sites in this region, referred to here as 4945 and 51 (abbreviations for SNP identification codes rs10751449/rs10736845 and rs10751451), which are predicted single and tandem GATA1 recognition sites altered by the haplotype (Figure 1A) (11). In reporter assays where this erythroid promoter was inserted to drive the expression of a luciferase, the mutation of any or both of these SNPs was enough to recapitulate the reduced expression observed with the full haplotype sequence (15). When we combined this luciferase assay with GATA1 overexpression, we found that GATA1 could efficiently enhance the expression from the WT promoter but to a much lesser extent when the haplotype-containing promoter was used, suggesting a GATA1-dependent expression control in this region (15). Although all these results suggest a critical role for GATA1 in the regulation of this erythroid promoter, there is still no direct evidence for GATA1 binding and how the haplotype interferes with this protein-DNA interaction.

GATA1 is an essential transcription factor involved in the differentiation of myeloid progenitor cells to mature blood cells in the bone marrow (16, 17). GATA1 is composed of two C4-type zinc fingers, referred to here as the N- and C-terminal fingers (NF and CF, respectively)

(18). The CF binds to DNA with high affinity and specificity through (A/T)GATA(A/G) sequences, while the NF displays weaker affinity to similar GATA sequences and appears to rather play a modulatory role by interacting with other proteins such as FOG1, or LMO2 (19–21). Complete loss of the GATA1 protein in mice is lethal at embryonic stage due to the absence of red blood cells, and mutations in the gene can alter proper erythropoiesis (22–25). In the past decades, many mutations have been identified in patients suffering from severe anemia, thrombocytopenia and/or β-thalassemia. These mutations are most commonly located in the NF (V205M, G208R/S, R216Q/W, D218G/N/Y) (26–33), but there are other mutations with similar phenotypes located right after the CF (insertion of PPFWQ at the position 290, or R307C/H/L) (34, 35). All of these mutants, which lie within either the NF or close to the CF, may directly influence DNA binding, but no systematic study has yet investigated this aspect on full-length protein under physiological conditions. In addition to these GATA1 mutations, a splicing defect (intronic mutations or the splice site substitution 332G-C often referred to as V74L), or mutation in the start codon (M1T) leads to the expression of a shorter GATA1 protein (GATA1s), which lacks the N-TAD part of the protein that may also affect the DNA binding properties of GATA1 indirectly (36–41). Expression of this short variant, rather than the full-length protein, induces Diamond-Blackfan anemia (DBA) and Down syndrome-associated myeloproliferative disorder (DS-TMD) (40–42).

Here we adapted our recently developed native holdup (nHU) assay to identify binding partners of a given short DNA oligonucleotide from a cellular extract (43). Just as the original assay dedicated to protein-protein interactions, this updated version is also suitable for the determination of binding constants, which enables the discrimination of stronger and weaker binders, allowing the study of variants that have only a mild effect on DNA binding (43). A major advantage of this method is that protein purification is not required to quantitatively assess a given DNA-protein interaction, which simplifies the investigation of DNA-binding properties of multiple mutant variants of a transcription factor. By analyzing the erythroid-specific promoter of the *ATP2B4* gene, we demonstrate that GATA1 is the transcription factor whose binding is most affected by the haplotype variants. In addition, we investigated six different disease-causing mutations in GATA1 and showed that the short version GATA1s and the R307C alter the DNA-binding properties of the protein.

Results

Proteome-wide affinity characterization of the *ATP2B4* promoter elements

We used the nHU assay coupled to mass spectrometry (nHU-MS) to measure the affinity interactome of the *ATP2B4* erythroid-specific alternative promoter elements (Figure 1B). The holdup is a comparative chromatographic approach where the analyte solution, that is the cell extract in the case of nHU, is mixed with a bait-saturated or control-saturated resin and the unbound prey fraction is measured from the liquid phase after incubation and rapid separation of the liquid and solid phases (Figure 1B) (43–46). From this fraction unbound value and from the estimated bait concentration during the experiment, the affinity of the bait-prey interaction can be estimated with good precision and accuracy. Previously, we performed the assay at relatively high bait concentration to capture weak and transient protein-protein interactions (43). However,

here we had to consider some aspects related to the nature of nucleic acid-protein interactions, which typically have a higher affinity than protein-protein interactions, and these interactions are highly dependent on ionic strength. For these reasons, we performed the nHU experiments at a lower estimated bait concentration (5 µM), using a more dilute cell extract (1 mg/ml) that the one we used for protein-protein assays. Besides, we also optimized cell lysis conditions to make it more optimal for capturing nucleic acid-protein interactions by introducing Mg^{2+} and K^{+} ions and by slightly lowering the ionic strength of the lysis buffer.

We used the wild-type and haplotype-containing double-stranded (ds) DNA constructs for the 4945 and the 51 fragments (26-mer and 31-mer, respectively), that we hybridized right before the nHU experiment (Figure 2A). Out of the synthetic single-stranded (ss) DNA constructs, only forward strands were biotinylated. To capture the interactions present in erythroid cells, K562 cell extracts were prepared. Although we considered using a nuclear extract, we hypothesized that the depth of standard MS measurements is going to be sufficient to capture most interactions of abundant nuclear proteins from total cell extracts and this way we do not introduce any sort of technical bias in our experiment. Apart from these minor modifications, we followed the original nHU protocol with biotinylated-dsDNA fragments as baits against biotin controls (Figure 1B). After a two-hour long incubation of bait saturated resins with cell extracts at 4 °C, we recovered the supernatant and subjected it to routine label-free quantitative proteomic measurements. In these total proteomic measurements, we managed to quantify the relative amounts of 4,560 proteins (Figure 2B, Supplementary Data 1). Binding threshold was set as discussed in details before, briefly, a hyperbolic threshold was used where significance has to reach a probability of at least 0.05, calculated by a two-tailed unpaired t-test, and prey concentration change needs to be greater than 95% of the measured changes (>2 σ) (43). Based on this threshold, 274 proteins showed significant binding to at least a single bait (Supplementary Data 1). Note that significant binding implies a statistically significant lower concentration after incubation with the bait than after incubation with the control, but the prey depletion is a constant value in a holdup reaction, determined by the bait concentration and the intrinsic affinity of the complex, consequently, there may be many more weak binding partners, marked as non-significant because of their small effect size.

Determination of sequence-specific partners in the *ATP2B4* promoter interactome

Substantially more partners were detected with the DNA baits than when performing native holdups with peptide baits, which is particularly surprising given the short size of the nucleic acids used in the assay (43, 47). Similar numbers of partners were only detected when promiscuous globular domains (e.g. SH3 or PDZ) or full-length proteins were used as bait (43, 48). A possible explanation for the large number of interaction partners is that nHU can capture indirect interactions. When large molecular machines are captured, all their subunits are identified as interaction partners (43). However, of the interactions identified, only a few large complexes have been identified, such as the RNA exosome complex, the RNA polymerase III complex, or components of the spliceosome, covering only a small fraction of all detected partners (Supplementary Data 1). Additionally, the different dsDNA-bait fragments showed remarkably similar interactome in qualitative terms. Among the 274 partners, 118 (43%) were identified as significant partners with all four dsDNA baits, while 152 (55%) and 190 (69%) proteins were identified as partners with three and two baits, respectively (Figure 2C).

In order to better understand the results of the nHU-MS experiments, we used the DAVID tool to perform Gene Ontology enrichment analysis for cellular compartments and molecular functions (Figure 2D, Supplementary Data 2) (49). Considering all 274 partners, we have found a great enrichment of proteins localized in the nucleus (N=190; P = 2.0 $\times 10^{-44}$) or more precisely in the nucleoplasm (N=183; P = 1.4 $\times 10^{-67}$). These partners were also greatly enriched with nucleic acid binding proteins, however surprisingly we identified more RNA binding proteins (N=177; P = 5.0 $\times 10^{-129}$) than DNA binding proteins (N=73; P = 7.4 $\times 10^{-24}$). Looking more closely at the 118 proteins that we identified as partners with all dsDNA baits, we also found that nuclear proteins (n=89; P = 5.4 $\times 10^{-25}$), RNA binding proteins (N=79; P = 1.1 $\times 10^{-58}$), and DNA binding proteins (N=37; P = 9.1 $\times 10^{-15}$) were highly enriched. Not surprisingly, we also identified many general DNA-binding proteins among the partners, such as multiple linker histone variants. Therefore, we concluded that the affinity interactomes of nucleic acid baits can be characterized quite exhaustively by simultaneously measuring sequence-specific and sequence-independent interactions using nHU.

Sequence-independent interactions are general properties of nucleic acids, related to their net charge and conserved topology. In sequence-independent interactions, molecular contacts are mostly made with the phosphate backbone and the contribution of the nucleic acid sequence is marginal. We assumed that the sequence-specific interaction partner critical for the erythroid-specific regulation of the *ATP2B4* gene is a transcription factor (TF). Based on a catalog of human TFs (50), we identified 32 TFs among the 274 proteins showing binding to at least one of the dsDNA baits (accounting for 11.7%, Supplementary Data 1). This is apparently a slightly higher ratio than the general occurrence of TFs in the entire genome (1,639 TFs among the 20,428 human proteins, accounting for 8.0%), and a much higher ratio than the measured occurrence of TFs in the explored portion of the erythroid proteome of K562 cells (141 detected TFs out of 4,560 detected proteins, accounting for 3.1%). However, even TFs can bind indirectly, or in a sequence-independent manner, similarly to other nucleic acid binding proteins discussed above. To identify the TFs that bind in a sequence-specific way, we exploited the quantitative aspect of the nHU assay by comparing the measured apparent affinities of the wild-type and haplotype dsDNA baits. By using a naive approach, we expected either a gain of affinity or a loss of affinity upon the SNPs. In addition, since we know that both sites contribute to the *ATP2B4* gene expression regulation in a similar manner, we assumed that the same interaction partner will be the critical regulator. Only a few TFs showed increased affinity for both haplotype baits, such as LIN28B or ATF2, but these showed minor changes in affinity rather than large perturbations (Table 1). Multiple TFs showed decreased affinities to the 51 site, such as ZNF24, or HOXB4, but these proteins did not show detectable binding to the 4945 site. Considering the 4945 segment, the interaction partner showing the highest down-regulation in affinity was the GATA1. This erythroid-specific TF also showed binding to the 51 site that displayed decreased affinity to the haplotype sequence, albeit only slight change was measured. Even if we ranked all transcription factors based on their cumulative ΔpK_{app} values of these sites (which is related to $\Delta\Delta G$), GATA1 showed the highest overall affinity change out of all partners (Table 1). From these observations, we concluded that the most likely sequence-specific interaction of these baits, which is expected to regulate the gene expression, is GATA1, in accordance with previous observations.

Ab ovo identification of sequence-specific interactions with nHU

As shown above, having clinically relevant variants allows us to distinguish sequence-specific relevant interactions from sequence-independent or irrelevant interactions. However, if we simply want to identify sequence-dependent interactions of a particular region without having such variants, it may be challenging due to the large number of DNA-binding partners caused by the chemical properties of the dsDNA molecule. To overcome this issue, we decided to use scrambled dsDNA baits, instead of free biotin, to saturate the control resin stock.

Thus, we repeated the nHU-MS experiments with the 51 WT fragment but using two additional scrambled DNA fragments as controls (Figure 3A-D, Supplementary Data 3). A major limitation that we faced during this assay, that due to detection limitations of mass spectrometry, we failed to detect GATA1 in the K562 proteome. For this reason, we performed Western-blot experiments using the same samples that we also measured with proteomics and with the same experimental strategy (same number of replicates, same statistical analysis) to quantify GATA1 protein specifically (Figure 3E, Supplementary Figure S1). Again, we detected a high number of interactors for all the three DNA probes compared to the biotin-saturated resin controls, but only in the case of 51 WT sequence we detected GATA1 as a binding partner (Figure 3B-D). In total, we had 79 partners that found to interact with all three dsDNA baits and 43 further proteins found to bind two of the dsDNA baits (Figure 3F). Therefore, just as in the first experiment, the measured interactomes of the different dsDNA baits are similar and we only found 16 specific binding partners for the 51 sequence. Of note, the design of scrambled control oligonucleotides is critical in this type of experiment. After scrambling, it is important to check that no new binding sites for other transcription factors have been created. As we have limited information about the consensus binding sites for all transcription factors present in the cell, it is possible that although our design was appropriate, the control probe may still be recognized by other transcription factors.

When we used the SCR1 or SCR2 separately as control samples, we only detected 10 or 4 statistically significant binding partners for the 51 WT sequence, respectively (Figure 3G-H, Supplementary Data 3). Among them, only 3 were common in both comparisons, namely TMEM258, CNBP and GATA1. TMEM258 is a transmembrane protein that is a common contaminant in mass spectrometry measurements, so it is likely that this protein was detected as a partner due to its aspecific binding to our probe. CNBP is a CCHC-type seven zinc finger nucleic acid binding protein that preferentially binds single-stranded oligonucleotides. The predicted binding motif consensus is somewhat consistent with our 51 WT sequence (51, 52). Due to this, we think it is possible that our annealing was not completely efficient, and considering that the collective binding of seven zinc fingers may increase the affinity due to avidity, the protein may appear as a sequence-specific binder in our study. Finally, GATA1 was the only protein that is a transcription factor and also binds to dsDNA. Thus, we were able to identify the relevant transcription factor that binds to our target sequence even without the use of sequence-disrupting mutations.

Biochemical characterization of the interactions between the *ATP2B4* promoter elements and GATA1

After identifying GATA1 as the most likely regulator of erythroid-specific expression of *ATP2B4*, we further characterized their interaction. Although the single point nHU experiment used in combination with mass spectrometry provides a proteome-wide estimation of the affinity interactome of a given oligonucleotide, it lacks certain biochemical details. Most importantly, to

convert the measured fractional prey depletion to apparent affinity, we had to assume a simple binding mechanism with no complications. In order to determine precisely the affinities, we performed titration nHU experiments and monitored GATA1 depletion from the liquid phase with Western blot (Figure 4A, Supplementary Figure S2). We found that both WT *ATP2B4* fragments displayed simple binding mechanism (matching a hyperbolic binding equation) but they also showed a partial GATA1 binding activity of cca. 75%, meaning that 25% of the endogenous GATA1 is unable to interact with the bait DNA fragments even at high bait concentrations (Figure 4B). We also found that the fragment 51, that contains two consecutive GATA motifs in tandem orientation, mediates higher affinity interactions with GATA1 with a dissociation constant of 17 nM, while the fragment 4945, that contains a single GATA motif, mediates a somewhat weaker affinity interactions with GATA1 with a dissociation constant of 100 nM (Figure 4B). By repeating the titration experiment with the haplotype sequences, we found that both promoter fragment displayed a weaker interaction. The haplotypic version of fragment 51, which still contains a single GATA motif, showed a dissociation constant of 130 nM with GATA1, which is about 10-fold weaker than the WT 51 sequence and similar to the affinity measured for WT 4945, which also contains a single motif. The haplotype version of fragment 4945, that do not contain any GATA motif, displayed only a very marginal interaction, approximately 100-times weaker than the WT sequence (Figure 4B).

Since DNA baits can interact with a wide range of prey molecules with high apparent affinity, including highly abundant molecules such as histones, there is a possibility of occupying a significant portion of the resin-immobilized bait molecule, causing competition between different preys affecting the outcomes of nHU experiments. To test whether this effect is present, we repeated the nHU titration experiment with the 4945 WT sequence but using 10-fold diluted (0.1 mg/ml total protein concentration) K562 cell extract (Figure 4C, Supplementary Figure S2). No evidence of this inhibitory effect was found, as similar affinities were obtained as before (K_{app}=200 nM). The charge of nucleic acids can also affect the saturation of the resin, and it may happen that our bait concentration estimation is lower than expected. To verify our measured affinities, we performed a nHU titration experiment using a palindromic DNA fragment that has been characterized before *in vitro*, using purified recombinant GATA1 protein with surface plasmon resonance (53). We obtained a dissociation constant (140 nM) very similar to that previously measured with SPR (150 nM), confirming that our bait concentration estimation is in the appropriate regime (Figure 4D, Supplementary Figure S2).

Characterization of the function effects of GATA1 mutations on the *ATP2B4* alternative promoter using luciferase assay

In order to shed light on the functional effects of different disease-associated GATA1 mutations on transcription, we performed a luciferase reporter assay in which K562 cells were co-transfected with a reporter plasmid containing the alternative promoter of the *ATP2B4* gene followed by the Firefly luciferase and with a HA-GATA1 construct containing the WT or a mutant GATA1 sequence (Figure 5A-B, Supplementary Data 4). A total of 10 different GATA1 constructs were generated, including the WT sequence, the short variant (GATA1s, or simply short) lacking the N-TAD, the V74L, which serves as a control since it does not exist in its natural form, as the short isoform is expressed as a result of this mutation, NF mutations (V205M, G208R, R216W, D218Y), the R307C mutation that is located right after the CF, and finally two artificial mutations disrupting the interaction with the DNA with either the NF (C228R) or the CF (C282R) by eliminating one of the cysteine coordinating the zinc ion.

Preliminary analysis showed that the largest window in the signal between the WT and HA-transfected cells could be obtained with the haplotype sequence, that still contains eight predicted single GATA1 binding sites, so we used this plasmid in our assay. 48 hours after transfection we lysed the cells and measured the amount of luciferase produced based on luminescence. The remaining lysates were subjected to Western blot analysis to determine the GATA1 expression levels for each condition. We did not see huge perturbations in the expression levels between the different mutant GATA1 proteins, except for the short variant, where we detected significantly lower expression in K562 cells, but still all proteins are produced stably under these conditions (Figure 5A, Supplementary Data 4, Supplementary Figure S3). The luciferase activity of GATA1 mutants was determined based on the luminescence signal compared to WT GATA1 (100%) and HA control (0%).

For the artificial C228R and C282R mutations, disrupting either the NF or CF DNA-binding, respectively, we detected a complete loss of function in the luciferase assay, highlighting the importance of DNA binding activity for both zinc fingers in function (Figure 5B). We also detected complete loss of function for the R216W and R307C mutants. We saw partial loss of function for the V205M and D218Y, located in the NF of the protein. No functional problems were observed with the artificial V74L and the G208R natural mutant, which had similar luciferase activity as the WT. Finally, an almost twofold increase in activity was observed for the short GATA1 variant in this study, despite Western blot results showing that cells express about half the amount of this variant, suggesting an approximately 4-fold increase in activity for this mutant (Figure 5A-B). Although this assay can help to identify mutations that cause functional aberrations, further binding studies are needed to distinguish which of these mutations cause defective binding to DNA, as it has been previously described that these variants can affect different regulatory mechanisms, such as interfering with the binding of other proteins essential for proper function (21, 30). Another possible explanation for these functional differences can be altered protein localization. It has been suggested before, that the R307C mutation can cause mis-localization, possibly by the disruption of a nuclear localization signal (35). To investigate this, we examined the co-localization of these HA-GATA1 variants with DAPI staining in transfected HEK293T cells, but in our assay all GATA1 variants localized to the nucleus and we could not observe any mis-localization of the different mutant proteins (Figure 5C, Supplementary Figure S4).

Determination of the effect of disease-associated GATA1 mutations on DNA binding properties by DNA-nHU

Determining the dissociation constant of GATA1 DNA binding with traditional assays for each mutant poses significant challenges. Performing biophysical measurements requires the recombinant production of all the different mutant versions of the protein. GATA1 is predominantly a disordered protein with two zinc-finger domains, which complicates its production in bacteria. A so-far unexploited aspect of the nHU approach is the possibility to rapidly and quantitatively examine the consequences of mutations on molecular interactions. Native holdup coupled to Western blot is a fast and relatively easy technique that requires only one interactant to be in a pure form, while the other molecule can be taken directly from cell extract. By using ectopic expression of our protein of interest with different tags, we can circumvent the need for high-quality antibodies specifically recognizing our proteins during the analysis.

In order to produce cell extracts for the nHU, we used the same HA-GATA1 constructs containing the different mutant versions of the protein than previously in the luciferase assay. We chose HEK293T cell line because it lacks detectable levels of endogenous GATA1 protein or mRNA, preventing any interference with our transiently expressed protein (15). The HA-tagged version of GATA1 was used because N-terminal tagging does not affect its DNA binding activity, and because the GATA1 antibody did not recognize the short isoform (53) (Supplementary Figure S5). HEK293T cells expressed all selected GATA1 mutants comparably to the wild type, and all mutants were detectable with the HA-tag specific antibody (Supplementary Figure S5). As baits in nHU experiments, we selected the WT and haplotype 51 dsDNA fragments as archetypical tandem, or single GATA motifs (Figure 2A). We performed 12-point nHU titration experiments similarly as before and analyzed the results by Western blot against the HA-tag (Figure 6A-B, Supplementary Figure S6-7). Importantly, the ectopically expressed HA-tagged WT GATA1 displayed nearly indistinguishable intrinsic binding affinities towards both tested baits compared to the naturally occurring (untagged) WT GATA1 found in K562 extracts (Figure 4A-B and Figure 6A-B).

Interestingly, we found that two disease-associated mutations impact DNA binding (Figure 6A-C). The R307C mutation displayed a two-fold decrease in affinity for the tandem GATA motif and approximately a 5.3-fold decrease in affinity for the single GATA motif. In contrast, the short isoform GATA1s exhibited a four-fold increase in affinity for the single GATA site, explaining the increased biological activity observed in the luciferase assay. The artificial mutation affecting the CF resulted in complete loss of activity on both DNA fragments, while the NF mutation reduced the affinity by two-fold for the tandem sequence and approximately 15-fold for the single GATA sequence. This observation aligns with previous findings that the NF alone has a very weak affinity for DNA and likely plays a stabilizing and modulatory role after the CF binds to the DNA (54). Notably, the single GATA site provided much larger differences between the mutant forms of GATA1 compared to the tandem, which could be due to synergistic effects between neighboring GATA sites.

Discussion

Opening new frontiers in nucleic acid research with native holdup

Our genome contains a myriad of regulatory regions composed of smaller elements that can interact with proteins to regulate transcription. Polymorphisms or mutations in these regions can contribute to individual susceptibility to different diseases. Yet, many of these regions are not yet characterized and we have no information about the identity of the binding partners of these elements. A possible way to get such an information is to perform ChIP-seq analyses with all putative transcription factors in our genome under multiple conditions, that is an enormous task and not available yet. Alternatively, *in silico* approaches can provide hints about cognate interaction partners of these regions. However, many DNA binding proteins have not yet been characterized extensively and these approaches do not provide any prediction for these proteins. Thus, these approaches are error prone and can lead to confusing conclusions. To experimentally find partners for a particular region, researchers had to rely on pulldown-based approaches. However, many nucleic acid-protein interactions display dynamic properties and can completely dissociate during the washing steps of a pulldown assay. Our approach, the nHU assay, provide a simple solution for these problems. It can be used similarly to a pulldown assay, but since it does not require washing steps, it can identify even highly transient interactions.

The nHU assay could not only identify the partner of the DNA bait, but it also provides a quantitative measurement of their intrinsic affinity. Such data is typically difficult to obtain and requires the production of purified proteins at high quantities. These are not requirements of a nHU assay that can be done with affordable synthetic DNA oligos, a small amount of commercially available resin and simple cell extracts. The results that these assays provide are comparable to conventional biophysical assays, such as ITC or SPR or other methods routinely used in the nucleic acid fields, such as mobility shift assays. In the past, multiple studies measured DNA binding to GATA1 using quantitative biophysical approaches, yet only with a few exceptions these studies were performed with isolated zinc fingers of GATA1 containing no or mild, non-disruptive mutations, often at non-physiological conditions (54, 55). In the case of nHU, there is no need to purify partner proteins, hence even highly unstable, yet biologically important full-length proteins can be studied at near native conditions, that would be not suitable for conventional methods. In conclusion, nHU offers new perspectives for studying nucleic acid-protein interaction and due to the simplicity of the assay, it can be effortlessly implemented in any cell biology laboratory.

Functional insight into GATA1 interactions, revealed by native holdup

In this study, we investigated the recently described erythroid-specific promoter of the *ATP2B4* gene encoding the calcium exporter protein PMCA4b (11, 12). In this erythroid promoter, a haplotype has been found that, by an unknown mechanism, reduces the expression of this protein in erythroid lineages and leads to reduced MCHC and susceptibility to malaria (13, 14). Previous ChIP-seq, bioinformatics and luciferase reporter analyses have suggested that loss of GATA1 binding may be responsible for this effect, but here we provide clear evidence that GATA1 binding is completely abolished at key sites in the presence of this haplotype. In the second part of our study, we used this interaction as a model system to study the mechanistic outcomes of clinically relevant GATA1 mutants.

Integrating the functional data from the luciferase assay with the binding information from nHU experiments provides a comprehensive understanding of the mechanism of action of these disease-associated mutations. Certain mutations localized to the NF can abolish the interaction between GATA1 and Friend of GATA-1 (FOG1) instead of the interaction between GATA1 and the DNA (25, 26, 56). Interestingly, out of the four NF mutations investigated here, three (V205M, R216W and R218Y) caused loss of function in the luciferase reporter assay, but we could not measure altered DNA binding for them (Figure 6C). These mutations structurally localize to the same surface of the protein, and probably they disrupt the binding of an important co-activator, such as FOG1, leading to reduced transcriptional activation by GATA1 indirectly. Of note, previous studies indicated that different substitutions at the same position can result in different activity due to structural compensation of the residue (55). In contrast to the NF mutations, two natural variants affected binding: the R307C variant and the short isoform. The R307C variant causes a decrease in affinity which manifests in a reduced cellular activity. This aligns well with the artificial Zn-finger mutations that abolish the DNA binding causing a complete loss of function. It has previously been suggested that mutation in this position may abolish a nuclear localization signal, but our results indicate that it is more likely that the DNA-binding capacity of this mutant is impaired, which may cause the slight mis-localization observed before (35). Indeed, structurally, it has been previously described that this arginine is located in the extension of the CF, protruding deep into the minor groove, and most probably contributes to the DNA binding (54) (Figure 6C). In contrast, increased DNA binding affinity

was observed for the short GATA1 variant, which also showed enhanced cellular activity in the luciferase assay. This allosteric effect on DNA binding can be explained by the removal of a potential inhibitory control in the N-terminal transactivation domain region of the protein. The shorter form of the protein was observed before in myeloproliferative disorders and leukemia, suggesting that this mutant may disrupt the delicate balance of regulation by GATA1 leading to cellular overproduction instead of anemia (40–42).

Limitations

In the nHU assay presented above, we used short synthetic dsDNA baits, but it the case of longer regulatory regions the use of nHU could be challenging. In practice, the length of synthetic oligonucleotides is difficult to extend beyond 200 nucleotides. Although it would be possible to use alternative ways to create larger biotinylated baits, the required reaction volumes would make such reaction impractical. Consequently, some interactions that require larger intact DNA fragments can be difficult to study with our approach, as well as the contribution of DNA flanking regions can be difficult to quantify. Similarly, the effects of distant cis regulatory elements can be lost due to short DNA fragments.

In addition, this assay suffers from the previously described limitations of nHU (43). First, bait concentrations are estimated based on empirical data. While this calculation may affect the absolute affinity values to some extent, but as done in the present work, comparison of affinities to varying sequences is quite accurate and precise, because the amounts on resins are well-controlled and reproducible. Second, only those proteins are screened that are sufficiently abundant to be able to quantified by bottom-up proteomic measurements, which may result in failing to identify critical partners. Then, due to the relatively small changes in concentrations during the nHU assay, some measured interactions are going to be appear as statistically not significant, because either, or both, the measured concentration change is small due to weak intrinsic affinities, or because the prey protein is difficult to quantify by proteomics. At last, the high sensitivity of the nHU approach can make it difficult to differentiate between direct and indirect partners, or between sequence specific and sequence aspecific partners. However, our study attempted to offer multiple solutions for these limitations.

Materials and Methods

Cell culturing

K562 cells (ATCC ref. CCL-243, RRID: CVCL_0004, cell line was a kind gift of Arnauld Poterszman) were grown in RPMI 1640 medium without HEPES (Gibco) completed with 10% fetal calf serum (FCS) and 40 µg/ml gentamicin and diluted 1:12 every 3rd/4th day. HEK293T cells (authenticated with 100% match as ATCC ref. CRL-3216, RRID: CVCL_0063) were grown in Dulbecco's modified Eagle's medium (DMEM, Gibco, glucose: 1 g/liter) completed with 10% FCS and 1% Penicillin-Streptomycin (Gibco), diluted 1:10 every 3rd/4th day. All cells were kept at 37°C and 5% CO_2.

Cell extract preparation for nHU assay

Cell extracts were prepared similarly as before (43) except the lysis buffer composition, which was optimized for DNA-protein interactions: 50 mM Hepes-KOH (pH 7.5), 100 mM NaCl, 50 mM KCl, 4 mM $MgCl_2$, 1% Triton X-100, 5× cOmplete EDTA-free protease inhibitor cocktail, 5 mM TCEP, and 10% glycerol. Briefly, K562 cells were collected by centrifugation (1000g × 5

min) and HEK293T cells were collected by scraping from T-175 flasks, quickly washed with PBS, and ice-cold lysis buffer was added. Lysates were sonicated (4 × 20 s with 1-sec-long pulses on ice) and incubated rotating (4°C for 30 min). The lysates were centrifuged (12,000 g 4°C for 20 min), and the supernatant was kept. Total protein concentration was measured by standard Bradford method (Bio-Rad protein assay dye reagent, ref. 5000006) using a bovine serum albumin (BSA) calibration curve (MP Biomedicals, ref. 160069) on a UV-1700 PharmaSpec spectrophotometer instrument. Extracts were diluted to 1 mg/ml concentration and were snap-frozen in liquid nitrogen and stored at −80°C.

Biotinylated DNA oligonucleotides and annealing for nHU

All DNA oligonucleotides were produced by Eurofins Genomics (Eurofins Genomics Germany GmbH, Ebersberg, Germany) and were high-performance liquid chromatography (HPLC) purified. The following oligonucleotides were used (5'-3' orientation, GATA1 recognition sequences are labelled green, Biot: biotin-group):

4945_WT:
For: Biot-TCAGCCCTCCG**TATC**GTCACCTACAC
Rev: GTGTAGGTGAC**GATA**CGGAGGGCTGA
4945_HAP:
For: Biot-TCAGCCCTCCG**TTTT**GTCACCTACAC
Rev: GTGTAGGTGAC**AAAA**CGGAGGGCTGA
51_WT:
For: Biot-GAATTGAGAGG**TATC**T**TATC**GCTCCCACTCC
Rev: GGAGTGGGAGC**GATA**A**GATA**CCTCTCAATTC
51_HAP:
For: Biot-GAATTGAGAGG**TATC**T**TACC**GCTCCCACTCC
Rev: GGAGTGGGAGC**GGTA**A**GATA**CCTCTCAATTC
SCR1:
For: Biot-CCACTCTGTTCTAGAGAAGCCTTCGTGAACT
Rev: AGTTCACGAAGGCTTCTCTAGAACAGAGTGG
SCR2:
For: Biot-ATGCCATTGCTTCATGCACGTATAGCCACTG
Rev: CAGTGGCTATACGTGCATGAAGCAATGGCAT
$CTRL_{palindrome}$:
For: Biot-GCGC**TATC**A**GATA**AGGCCTTG
Rev: CAAGGCCT**TATC**T**GATA**GCGC

For annealing, the two single-stranded oligonucleotides were mixed in 1:1.5 forward-reverse DNA ratio in a 12.5 µM final double-stranded oligonucleotide concentration in holdup (HU) buffer (50 mM Tris pH 7.5, 300 mM NaCl, 1 mM TCEP, 0.22-µm filtered). In a pre-heated thermoblock the samples were heated up to 95 °C for 5 minutes, then slowly cooled down within 100 minutes until they reached room temperature.

In general, 680 µl of 12.5 µM dsDNA were used for the saturation of 25 µl pure streptavidin resin (Streptavidin Sepharose High Performance, Cytiva). The estimated bait concentration of

the saturated resin is 10 µM once 25 µl pure bait-saturated streptavidin resin was mixed with 100 µl of cell extract.

Singe point and titration native holdup

For saturating resin with dsDNA or biotin (as control), 25 µl pure resin was mixed with either 680 µl of 12.5 µM dsDNA or 25 µl 1 mM biotin, and additional NaCl was added (at 0.5 M final concentration) to facilitate DNA fixation to the resin. Resin mixtures were incubated for 1 hour at room temperature, gently mixing every 15 minutes. Resins were briefly centrifuged (15 s, 2000 g), the supernatants were removed and the resins were depleted with biotin (10 resin volume, 3 min, holdup buffer supplemented with 100 µM biotin). Then, resins were washed four-times (10 resin volume, holdup buffer). To get a bait concentration of 5 µM, these pre-saturated resins were reconstituted in 6x resin volume of HU buffer and were mixed in a 1:1 ratio by adding 12.5 µl dsDNA to 12.5 µl control (biotin) pure resin. Finally, all liquid was removed and cell extracts were added in 1 mg/ml concentration in 1:4 pure resin:extract ratio (25 µl pure resin : 100 µl cell extract) and incubated for 2 hours at 4 °C with constant gentle agitation. After the incubation ended, the resins were separated from the supernatant by a brief centrifugation (15 s, 2000 g). 75% of the supernatants were quickly pipetted to a separate tube and centrifuged again briefly, and 50% of the original volume of the supernatant was kept in the end to reduce resin contamination in the final samples.

Titration experiments (coupled to Western-blot) were carried out similarly as before, by mixing control and bait-saturated resins and keeping the total resin volume constant (43). Twelve-point titrations were carried out. For the starting point, the DNA-saturated and biotin control resins were mixed in 1:1 (5 µM starting point) or 1:3 (2.5 µM staring point) ratio, typically 20+20 µl or 10+30 µl, respectively, calculated for the pure resin volume, in 4x resin volume in holdup buffer. Then, 1:1 serial dilution was carried out with the control resin (20+20 µl pure resin). The final tube contained only the biotin saturated resin as control. The resins were then briefly centrifuged (15 s, 2000 g), the supernatant was removed and cell lysates were added at a concentration of 1 mg/ml or 0.1 mg/ml, and further procedures were carried out as described for the single-point experiments (20 µl of resin was mixed with 80 µl of extract to obtain the correct concentration of bait during the incubation). All the experiments were carried out in Protein LoBind tubes (Eppendorf, refs. 022431064 and 022431081). All titrations were carried out with minimum two replicates.

Sample digestion and mass spectrometry (LC-MS/MS) analysis

MS analyses were performed as described before (43, 48). Briefly, nHU samples were precipitated with TCA, and the urea-solubilized, reduced and alkylated proteins were digested with trypsin at 2 M final urea concentration. Peptide mixtures were then desalted on C18 spin-column and dried on Speed-Vacuum. 100 ng peptide mixtures were analyzed using an Ultimate 3000 nano-RSLC coupled in line, via a nano-electrospray ionization source, with an Orbitrap Exploris 480 mass-spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with a FAIMS (high Field Asymmetric Ion Mobility Spectrometry) module. Data was collected in DDA (data dependent acquisition) mode, proteins were identified by database searching using SequestHT (Thermo Fisher Scientific) with Proteome Discoverer software (Thermo Fisher Scientific). Peptides and proteins were filtered with a false discovery rate (FDR) at 1%. Label-free quantification was based on the extracted ion chromatogram (XIC) intensity of the peptides.

All samples were measured in three technical triplicates. The measured XIC intensities were normalized based on median intensities of the entire dataset to correct minor loading differences. For statistical tests and enrichment calculations, not detectable intensity values were treated with an imputation method, where the missing values were replaced by random values similar to the 10% of the lowest intensity values present in the entire dataset. Unpaired two tailed t-tests, assuming equal variance, were performed on obtained $\log_2$ XIC intensities. All raw LC-MS/MS data have been deposited to the ProteomeXchange via the PRIDE database with identifier PXD055181. Obtained fold-change values were converted to apparent affinities using the hyperbolic binding equation and binding thresholds at 2σ were determined as described before in details (43).

Gene Ontology enrichment analysis

Gene Ontology enrichment analysis was performed using the DAVID tool with default parameters (49). Two independent analysis were performed using all the preys that were interacting with at least one bait, or using only those preys that were interacting with all baits, and both analyses were done against the same background (all proteins detected by the proteomic measurements). Only the Cell Component (CC_DIRECT) and Molecular Function (MF_DIRECT) terms were analyzed and all significant terms are included in supplementary data 2.

Identification of sequence-specific binders based on cumulative ΔpK_{app}

To calculate affinity differences between interactions measured with the WT and haplotype-carrying dsDNA baits, we used the apparent affinities, that were converted from the measured fold-change values. First, we identified all transcription factors among the interaction partners of the dsDNA baits using a recently reported comprehensive transcription factor catalog (50). Then, we collected the apparent affinities of all these proteins and calculated the overall effect of the haplotype by summing the affinity differences between WT and HAP sequences. For getting the affinity differences, we simply subtracted the measured pK_{app} value of the haplotype sequence from the measured pK_{app} value of the WT sequence, that is proportional to the $\Delta\Delta G_{app}$ values (48).

DNA plasmid constructs and mutagenesis

For the luciferase assay, the pGL3 vector containing the erythroid promoter region of the *ATP2B4* gene with the haplotype was created before (pGL3-ATP2B4-H1st plasmid, haplotype carrying version) (15). For the GATA1 mammalian plasmid, the sequence of GATA1 was cloned from the pIRES2-GATA1-GFP plasmid (15) into the pCI standard mammalian vector containing a HA-tag for N-terminal tagging with standard restriction cloning.

The different mutant forms of GATA1 were created by using standard site-directed mutagenesis on the pCI-HA-GATA1-WT plasmid with Platinum SuperFi II polymerase (Thermo Fisher Scientific).

Transfections

For the luciferase reporter assay, 1 x 10^5 K562 cells/well were seeded the day before the transfection on 24-well plates. The next day, cells were transfected with the pGL3-ATP2B4-H1st-haplo and pCI-HA-GATA1 WT and mutation carrying plasmid vectors using JetPRIME

reagent (Polyplus), according to the manufacturer's recommendations using 0.35 µg pGL3 and 0.15 µg pCI-GATA1 vector DNA and 1 µl JetPRIME reagent/well.

For the immunostaining of HA-GATA1 expressing cells, 24-well plates with coverslips were treated with 0.1 mg/ml poly-D-lysine (Sigma-Aldrich, ref. P6407) for 1 hour, washed twice with UltraPure water and allowed to dry completely. HEK293T cells were seeded on the plates the day before the transfection ($1x10^5$ cells/well). The transfection was carried out with JetPRIME reagent (Polyplus) according to the manufacturer's recommendations using 0.5 µg DNA and 1 µl of JetPRIME reagent/well. The medium was replaced with fresh medium after 5 hours of treatment with the transfection mixture to enhance cell survival.

For the cell extract preparation of HA-GATA1 WT and mutants expressing HEK293T cells, the cells were seeded on poly-D-lysine treated 100 mm cell culture dishes (2×10^6 cells/dish). The next day, the transfection was carried out with JetPRIME reagent (Polyplus) according to the manufacturer's recommendations using 5 µg DNA and 20 µl of JetPRIME reagent/well. The medium was replaced with fresh medium after 5 hours of treatment with the transfection mixture.

Luciferase reporter assay

24 hours after transfection, the K562 cells were collected by centrifugation (1 min x 1000g) and washed once with PBS. Then, 100 µl of 1x passive lysis buffer (5x PLB diluted with UltraPure water, Promega ref. E1941) were added to the cells and incubated for 15 minutes at room temperature under shaking. Lysates were pipetted up and down a few times and 20 µl lysate was transferred into 96-well white plates (Millipore ref. MSSWNFX40). Then, 100 µl luciferase assay reagent (Promega, ref. E1483) was added to each well and the firefly luminescence was measured immediately on a PHERAstar (BMG Labtech, Offenburg, Germany) microplate reader by using Lum Plus filter. The remaining cell lysates were kept for expression determination by Western-blot. In total, 3 biological (transfection) and 2 technical (from the same extract) replicates were performed. For statistics, the technical replicates were averaged and two-tailed t-tests were performed on these averaged values for each biological replicate (n=3). For statistical analyses, GraphPad Prism 10 software was used.

Western-blot

The samples were mixed with 4× Laemmli buffer (120 mM Tris-HCl pH 7, 8% SDS, 100 mM dithiothreitol, 32% glycerol, 0.004% bromphenol blue, and 1% β-mercaptoethanol) in a 3:1 ratio. Equal amounts of samples (generally 6 µg) were loaded on 12% acrylamide gels. Transfer was done using a Trans-Blot Turbo Transfer System and Trans-Blot Turbo RTA Midi 0.45 µm LF PVDF Transfer Kit (Bio-Rad, ref. 1704275). After 1 hour of blocking in 5% milk in TBS-Tween (1x Tris-buffered saline, 0.1% Tween 20), membranes were incubated overnight at 4°C in primary antibody in 5% milk in TBS-Tween. The following antibodies and dilutions were used: anti-GATA1 (1:10,000, abcam, ref. ab181544, RRID: AB_2920794), anti–glyceraldehyde-3-phosphate dehydrogenase (GAPDH, 1:5,000, Sigma-Aldrich, ref. MAB374, RRID: AB_2107445), and anti-HA.11 epitope tag (1:5,000, BioLegend, ref. 901501, RRID: AB_2565006). Membranes were washed three times with TBS–Tween and incubated at room temperature for 1 hour in secondary antibody solution [Jackson ImmunoResearch, peroxidase-conjugated Affinipure goat anti-mouse (H + L), ref. 115-035-146 RRID: AB_2307392 and goat anti-rabbit (H + L), ref. 111-035-003 RRID: AB_2313567] in 5% milk (concentration 1:10,000).

After washing three times with TBS-Tween, membranes were exposed to chemiluminescent horseradish peroxidase substrate (Immobilon, ref. WBKLS0100) and photographed in a docking system (Amersham ImageQuant 800). Densitometry was carried out on raw Tif images by using Fiji ImageJ 1.54f. Between different primary antibody labeling, the membranes were either exposed to 15% H_2O_2 to remove secondary signal (in the case of different species) or stripped with mild stripping buffer (15 g/liter glycine, 1 g/liter SDS, and 1% Tween 20, pH 2.2) to remove primary signal (in the case of same species). For native holdups, GAPDH was used as a non-binding control, but was not used to normalize GATA1 values as this induces technical errors coming from blotting issues rather than from differences between the samples, as discussed before (43). In case of inconsistencies in GAPDH values, the blot was repeated.

Immunostaining of HA-GATA1 transfected cells

24 hours after transfection, HEK293T cells were washed once with PBS and fixed with 4% PFA (Electron Microscopy Sciences, ref. 15710) in PBS for 10 minutes at room temperature. The cells were then washed three times with PBS and permeabilized with 500 µl/well of 0.2% Triton X-100 (Sigma-Aldrich, ref. T8787) in PBS for 10 minutes at room temperature. After permeabilization, the samples were blocked in blocking buffer [0.05 g/mL BSA (MP Biomedicals, ref. 160069), 0.1% Triton X-100 in PBS] for 1 hour at room temperature. Then, the cells were incubated overnight with the HA primary antibody (anti-HA.11 epitope tag, dilution: 1:1,000, BioLegend, ref. 901501, RRID: AB_2565006) at 4 °C in blocking buffer. The next day, the cells were washed three-times with PBS with 10 minutes incubations each time at room temperature. Then, the cells were incubated with the secondary antibody (Alexa Fluor 594–conjugated anti-mouse, dilution: 1:1000; Invitrogen, ref. A-11032, RRID: AB_2534091) in blocking buffer. After three-times 10 minutes PBS washes, the coverslips were mounted to microscopy slides by Vectashield mounting medium with 4′,6-diamidino-2-phenylindole (DAPI) (Vector laboratories, Burlingame, CA). Images were taken using a Leica SP5 confocal microscope (Leica Camera AG, Wetzlar, Germany) with an HCX PL APO 63×/1.40 to 0.60 oil objective using excitation at 405 nm (diode) and 594 nm (HeNe laser) and emission at 415 to 480 and 610 to 695 nm for DAPI and Alexa 594, respectively. Images were processed by the Fiji ImageJ software.

References

1. Timpson,N.J., Greenwood,C.M.T., Soranzo,N., Lawson,D.J. and Richards,J.B. (2018) Genetic architecture: The shape of the genetic contribution to human traits and disease. *Nat Rev Genet*, **19**, 110–124.

2. Bodmer,W. and Bonilla,C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*, **40**, 695–701.

3. Spielmann,M. and Mundlos,S. (2016) Looking beyond the genes: The role of non-coding variants in human disease. *Hum Mol Genet*, **25**, R157–R165.

4. Zhang,F. and Lupski,J.R. (2015) Non-coding genetic variants in human disease. *Hum Mol Genet*, **24**, R102–R110.

5. Stunnenberg,H.G. and Hubner,N.C. (2014) Genomics meets proteomics: Identifying the culprits in disease. *Hum Genet*, **133**, 689–700.

6. Gasperskaja,E. and Kučinskas,V. (2017) The most common technologies and tools for functional genome analysis. *Acta Med Litu*, **24**, 1–11.

7. Weimer,K., Zambo,B. and Gogl,G. (2023) Molecules interact. But how strong and how much? *BioEssays*, **45**, 2300007.

8. Furey,T.S. (2012) ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet*, **13**, 840–852.

9. Park,P.J. (2009) ChIP-seq: Advantages and challenges of a maturing technology. *Nat Rev Genet*, **10**, 669–680.

10. Alves,C. and Cunha,C. (2012) Electrophoretic mobility shift assay: analyzing protein-nucleic acid interactions. In Magdeldin,S. (ed), *Gel Electrophoresis - Advanced Techniques*. Rijeka, Croatia, pp. 205–229.

11. Zámbó,B., Várady,G., Padányi,R., Szabó,E., Németh,A., Langó,T., Enyedi,Á. and Sarkadi,B. (2017) Decreased calcium pump expression in human erythrocytes is connected to a minor haplotype in the ATP2B4 gene. *Cell Calcium*, **65**, 73–79.

12. Lessard,S., Stern,E.N., Beaudoin,M., Schupp,P.G., Sher,F., Ali,A., Prehar,S., Kurita,R., Nakamura,Y. and Baena,E. (2017) An erythroid - specific enhancer of ATP2B4 mediates red blood cell hydration and malaria susceptibility. *J Clin Invest*, **1**, 1–10.

13. Li,J., Glessner,J.T., Zhang,H., Hou,C., Wei,Z., Bradfield,J.P., Mentch,F.D., Guo,Y., Kim,C., Xia,Q., *et al.* (2012) GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African American children. *Hum Mol Genet*, **22**, 1457-1464.

14. Timmann,C., Thye,T., Vens,M., Evans,J., May,J., Ehmen,C., Sievertsen,J., Muntau,B., Ruge,G., Loag,W., *et al.* (2012) Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature*, **489**, 443–446.

15. Mózner,O., Zámbó,B. and Sarkadi,B. (2021) Modulation of the human erythroid plasma membrane calcium pump (Pmca4b) expression by polymorphic genetic variants. *Membranes*, **11**, 586.

16. Suzuki,M., Shimizu,R. and Yamamoto,M. (2011) Transcriptional regulation by GATA1 and GATA2 during erythropoiesis. *Int J Hematol*, **93**, 150–155.

17. Ferreira,R., Ohneda,K., Yamamoto,M. and Philipsen,S. (2005) GATA1 Function, a Paradigm for Transcription Factors in Hematopoiesis. *Mol Cell Biol*, **25**, 1215–1227.

18. Orkin,S.H. (1992) GATA-Binding Transcription Factors in Hematopoietic Cells. *Blood*, **80**, 575–581.

19. Liew,C.K., Simpson,R.J.Y., Kwan,A.H.Y., Crofts,L.A., Loughlin,F.E., Matthews,J.M., Crossley,M. and Mackay,J.P. (2005) Zinc fingers as protein recognition motifs: structural basis for the GATA-1/friend of GATA interaction.

20. Fox,A.H., Liew,C., Holmes,M., Kowalski,K., Mackay,J. and Crossley,M. (1999) Transcriptional cofactors of the FOG family interact with GATA proteins by means of multiple zinc fingers.

21. Wilkinson-White,L., Gamsjaeger,R., Dastmalchi,S., Wienert,B., Stokes,P.H., Crossley,M., Mackay,J.P. and Matthews,J.M. (2011) Structural basis of simultaneous recruitment of the transcriptional regulators LMO2 and FOG1/ZFPM1 by the transcription factor GATA1. *Proceedings of the National Academy of Sciences*, **108**, 14443–14448.

22. Crispino,J.D. and Weiss,M.J. (2014) Erythro-megakaryocytic transcription factors associated with hereditary anemia. *Blood*, **123**, 3080–3088.

23. Fujiwara,Y., Browne,C.P., Cunniff,K., Goff,S.C. and Orkin,S.H. (1996) Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *Proceedings of the National Academy of Sciences*, **93**, 12355–12358.

24. Weiss,M.J. and Orkin,S.H. (1995) Transcription factor GATA-1 permits survival and maturation of erythroid precursors by preventing apoptosis. *Proceedings of the National Academy of Sciences*, **92**, 9623–9627.

25. Ling,T. and Crispino,J.D. (2020) GATA1 mutations in red cell disorders. *IUBMB Life*, **72**, 106–118.

26. Nichols,K.E., Crispino,J.D., Poncz,M., White,J.G., Orkin,S.H., Maris,J.M. and Weiss,M.J. (2000) Familial dyserythropoietic anaemia and thrombocytopenia due to an inherited mutation in GATA1. *Nat Genet*, **24**, 266–270.

27. Del Vecchio,G.C., Giordani,L., De Santis,A. and De Mattia,D. (2005) Dyserythropoietic Anemia and Thrombocytopenia due to a Novel Mutation in GATA-1. *Acta Haematol*, **114**, 113–116.

28. Kratz,C.P., Niemeyer,C.M., Karow,A., Volz-Fleckenstein,M., Schmitt-Gräff,A. and Strahm,B. (2008) Congenital transfusion-dependent anemia and thrombocytopenia with myelodysplasia due to a recurrent GATA1G208R germline mutation. *Leukemia*, **22**, 432–434.

29. Mehaffey,M.G., Newton,A.L., Gandhi,M.J., Crossley,M. and Drachman,J.G. (2001) X-linked thrombocytopenia caused by a novel mutation of GATA-1. *Blood*, **98**, 2681–2688.

30. Yu,C., Niakan,K.K., Matsushita,M., Stamatoyannopoulos,G., Orkin,S.H. and Raskind,W.H. (2002) X-linked thrombocytopenia with thalassemia from a mutation in the amino finger of GATA-1 affecting DNA binding rather than FOG-1 interaction. *Blood*, **100**, 2040–2045.

31. Phillips,J.D., Steensma,D.P., Pulsipher,M.A., Spangrude,G.J. and Kushner,J.P. (2007) Congenital erythropoietic porphyria due to a mutation in GATA1: The first trans-acting mutation causative for a human porphyria. *Blood*, **109**, 2618–2621.

32. Freson,K., Devriendt,K., Matthijs,G., Van Hoof,A., De Vos,R., Thys,C., Minner,K., Hoylaerts,M.F., Vermylen,J. and Geet,C. Van (2001) Platelet characteristics in patients with X-linked macrothrombocytopenia because of a novel GATA1 mutation. *Blood*, **98**, 85–92.

33. Freson,K., Matthijs,G., Thys,C., Mariën,P., Hoylaerts,M.F., Vermylen,J. and Geet,C. Van (2002) Different substitutions at residue D218 of the X-linked transcription factor GATA1 lead to altered clinical severity of macrothrombocytopenia and anemia and are associated with variable skewed X inactivation. *Hum Mol Genet*, **11**, 147–152.

34. Abdulhay,N.J., Fiorini,C., Verboon,J.M., Ludwig,L.S., Ulirsch,J.C., Zieger,B., Lareau,C.A., Mi,X., Roy,A., Obeng,E.A., *et al.* (2019) Impaired human hematopoiesis due to a cryptic intronic GATA1 splicing mutation. *Journal of Experimental Medicine*, **216**, 1050–1060.

35. Ludwig,L.S., Lareau,C.A., Bao,E.L., Liu,N., Utsugisawa,T., Tseng,A.M., Myers,S.A., Verboon,J.M., Ulirsch,J.C., Luo,W., *et al.* (2022) Congenital anemia reveals distinct targeting mechanisms for master transcription factor GATA1. *Blood*, **139**, 2534–2546.

36. Chlon,T.M., McNulty,M., Goldenson,B., Rosinski,A. and Crispino,J.D. (2015) Global transcriptome and chromatin occupancy analysis reveal the short isoform of GATA1 is deficient for erythroid specification and gene expression. *Haematologica*, **100**, 575–584.

37. Wechsler,J., Greene,M., McDevitt,M.A., Anastasi,J., Karp,J.E., Le Beau,M.M. and Crispino,J.D. (2002) Acquired mutations in GATA1 in the megakaryoblastic leukemia of Down syndrome. *Nat Genet*, **32**, 148–152.

38. Matsuo,S., Nishinaka-Arai,Y., Kazuki,Y., Oshimura,M., Nakahata,T., Niwa,A. and Saito,M.K. (2021) Pluripotent stem cell model of early hematopoiesis in down syndrome reveals quantitative effects of short-form GATA1 protein on lineage specification. *PLoS One*, **16**, e0247595.

39. Hollanda,L.M., Lima,C.S.P., Cunha,A.F., Albuquerque,D.M., Vassallo,J., Ozelo,M.C., Joazeiro,P.P., Saad,S.T.O. and Costa,F.F. (2006) An inherited mutation leading to production of only the short isoform of GATA-1 is associated with impaired erythropoiesis. *Nat Genet*, **38**, 807–812.

40. Xu,G., Nagano,M., Kanezaki,R., Toki,T., Hayashi,Y., Taketani,T., Taki,T., Mitui,T., Koike,K., Kato,K., *et al.* (2003) Frequent mutations in the GATA-1 gene in the transient myeloproliferative disorder of Down syndrome. *Blood*, **102**, 2960–2968.

41. Shimizu,R., Engel,J.D. and Yamamoto,M. (2008) GATA1-related leukaemias. *Nat Rev Cancer*, **8**, 279–287.

42. Ludwig,L.S., Gazda,H.T., Eng,J.C., Eichhorn,S.W., Thiru,P., Ghazvinian,R., George,T.I., Gotlib,J.R., Beggs,A.H., Sieff,C.A., *et al.* (2014) Altered translation of GATA1 in Diamond-Blackfan anemia. *Nat Med*, **20**, 748–753.

43. Zambo,B., Morlet,B., Negroni,L., Trave,G. and Gogl,G. (2022) Native holdup (nHU) to measure binding affinities from cell extracts. *Sci Adv*, **8**, eade3828.

44. Vincentelli,R., Luck,K., Poirson,J., Polanowska,J., Abdat,J., Blémont,M., Turchetto,J., Iv,F., Ricquier,K., Straub,M.L., *et al.* (2015) Quantifying domain-ligand affinities and specificities by high-throughput holdup assay. *Nat Methods*, **12**, 787–793.

45. Charbonnier,S., Zanier,K., Masson,M. and Trave,G. (2006) Capturing protein– rotein complexes at equilibrium: The holdup comparative chromatographic retention assay. *Protein Expr Purif*, **50**, 89–101.

46. Gogl,G., Zambo,B., Kostmann,C., Cousido-Siah,A., Morlet,B., Durbesson,F., Negroni,L., Eberling,P., Jané,P., Nominé,Y., *et al.* (2022) Quantitative fragmentomics allow affinity mapping of interactomes. *Nat Commun*, **13**, 5472.

47. Zambo,B., Gogl,G., Morlet,B., Eberling,P., Negroni,L., Moine,H. and Travé,G. (2024) Comparative analysis of PDZ-binding motifs in the diacylglycerol kinase family. *FEBS J*, **291**, 690–704.

48. Zambo,B., Edelweiss,E., Morlet,B., Negroni,L., Pajkos,M., Dosztanyi,Z., Ostergaard,S., Trave,G., Laporte,J. and Gogl,G. (2024) Uncovering the BIN1-SH3 interactome underpinning centronuclear myopathy. *Elife*, **13**, RP95397.

49. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**, 44–57.

50. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The Human Transcription Factors. *Cell*, **172**, 650–665.

51. Armas,P., Coux,G., Weiner,A.M.J. and Calcaterra,N.B. (2021) What's new about CNBP? Divergent functions and activities for a conserved nucleic acid binding protein. *Biochimica et Biophysica Acta (BBA) - General Subjects*, **1865**, 129996.

52. Armas,P., Margarit,E., Mouguelar,V.S., Allende,M.L. and Calcaterra,N.B. (2013) Beyond the Binding Site: In Vivo Identification of tbx2, smarca5 and wnt5b as Molecular Targets of CNBP during Embryonic Development. *PLoS One*, **8**, e63234.

53. Hasegawa,A., Kaneko,H., Ishihara,D., Nakamura,M., Watanabe,A., Yamamoto,M., Trainor,C.D. and Shimizu,R. (2016) GATA1 Binding Kinetics on Conformation-Specific Binding Sites Elicit Differential Transcriptional Regulation. *Mol Cell Biol*, **36**, 2151–2167.

54. Wilkinson-White,L., Lester,K.L., Ripin,N., Jacques,D.A., Mitchell Guss,J. and Matthews,J.M. (2015) GATA1 directly mediates interactions with closely spaced pseudopalindromic but not distantly spaced double GATA sites on DNA. *Protein Science*, **24**, 1649–1659.

55. Campbell,A.E., Wilkinson-White,L., Mackay,J.P., Matthews,J.M. and Blobel,G.A. (2013) Analysis of disease-causing GATA1 mutations in murine gene complementation systems. *Blood*, **121**, 5218–5227.

56. Crispino,J.D., Lodish,M.B., MacKay,J.P. and Orkin,S.H. (1999) Use of Altered Specificity Mutants to Probe a Specific Protein–Protein Interaction in Differentiation: the GATA-1:FOG Complex. *Mol Cell*, **3**, 219–228.

Funding

This project was supported by the Ligue contre le cancer (équipe labellisée 2015 to G.T.) and the Agence Nationale de la Recherche (grants ANR-22-CE11-0026 HERC2-DOCKD-Modul and ANR-22-CE44-0018 Full-PDZ-PBM-HPVE6). B.Z. was supported by Fondation pour la Recherche Médicale (FRM, SPF202005011975). G.G. was supported by Inserm and Fondation ARC pour la recherche sur le cancer. The Light Microscopy Facility at the IGBMC imaging center, as a member of the national infrastructure France-BioImaging, was supported by the French National Research Agency (ANR-10-INBS-04). As a member of the IGBMC institute, the team also benefited from the French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INSB-05-01, from Instruct-ERIC, from IdEx Unistra (ANR-10-IDEX-0002), from SFRI-STRAT'US project (ANR 20-SFRI-0012), and from EUR IMCBio (ANR-17-EURE-0023) under the framework of the French Investments for the Future Program as a member of the Interdisciplinary Thematic Institute IMCBio, as part of the ITI 2021-2028 program of the University of Strasbourg, CNRS, and Inserm.

Acknowledgements

We thank the staff of the IGBMC Cell Culture and Light Microscopy Facility for their help in cell culturing and imaging.

Author contributions

B.Z. conceptualized the study, designed and performed most of the experiments. B.Z. and G.G. designed and performed all the nHU and did all the final data analyses. B.M. and L.N. performed the MS experiments and primary data analysis. O.M. and B.S. took part in conceptualization and helped in the luciferase assay. G.T. obtained funding for the experiments. G.T. and G.G. supervised the project. B.Z. and G.G. wrote the original draft, and all authors reviewed the manuscript.

Figure and table legends

Figure 1. Principles of *ATP2B4* gene regulation and native holdup using biotinylated oligonucleotide baits. **A.** Regulation of the *ATP2B4* gene. The *ATP2B4* gene contains two promoters, a general promoter that regulates expression in most cells and an alternative, erythroid-specific promoter located immediately upstream of the second exon of the gene. Previously, a haplotype located at this erythroid-specific promoter was identified that causes loss of expression only in erythrocytes. Two sites were found to be primarily responsible for this reduced expression (sites 4945 and 51). **B.** Basic principles of native holdup using biotinylated dsDNA bait. The first step is to produce total cell extracts. Then, annealing is performed using one biotinylated and one non-biotinylated ssDNA molecule. The annealed dsDNA is immobilized on streptavidin resin along with a control (biotin-treated) resin. These saturated resins are then incubated with the cell extract until the equilibrium is reached, after which the supernatant is rapidly separated and quantification is performed from the liquid phase. Finally, the unbound fractions are calculated by comparing the reaction containing dsDNA to the control (biotin-containing) reaction and converted to apparent affinity using the hyperbolic binding equation.

Figure 2. Affinity interactome determination of *ATP2B4* erythroid-specific alternative promoter elements measured by mass spectrometry-coupled nHU on K562 cell extracts. **A.** Oligonucleotides used in the native holdup assay. Both WT and haplotype (HAP) sequences were used for regions 4945 and 51. For 4945, the single GATA sequence is abolished by the haplotype, whereas for the tandem region 51, one GATA site is eliminated by the haplotype. **B.** Volcano plots of four experiments with the four different baits (n=3, injection replicates). Transcription factors are highlighted in green. For further data see Supplementary Data 1. **C.** Overlap of the interaction partners of the different baits, visualized with a Venn diagram (hyperbolic significance threshold was determined at 2σ). Surprisingly, most of the partners found were detectable with all four baits. **D.** Gene ontology enrichment was performed using the DAVID tool (49) on the proteins that were binding to at least one DNA bait. Most of the partners detected are general nucleic acid binding proteins that are located in the nucleus. For further data see Supplementary Data 2.

Table 1. Identification of sequence-specific binders by calculating cumulative ΔpK_{app} for transcription factors detected in the mass spectrometry coupled nHU experiments. Apparent affinities that belong to interactions below the significance threshold are colored in grey. To calculate the effect of the haplotype, the WT and haplotype (HAP) apparent affinity differences were summed and the measured pK_{app} value of the haplotype sequence was subtracted from the measured pK_{app} value of the WT sequence. GATA1 is highlighted in bold as it showed the largest negative effect due to the haplotype.

Figure 3. *Ab ovo* identification of sequence-specific interactions for a given DNA-oligonucleotide with nHU coupled to mass spectrometry. **A.** Oligonucleotides used in the experiment series. We used the 51_WT sequence and two scrambled versions (SCR1 and SCR2) to find the sequence specific recognition partners for the 51_WT bait in an unbiased manner. The GATA sequences are highlighted in green. **B.** Volcano plot visualization of the nHU result with 51_WT sequence compared to the biotin control. **C.** Volcano plot visualization of the nHU result with SCR1 sequence compared to the biotin control. **D.** Volcano plot visualization of the nHU

result with SCR2 sequence compared to the biotin control. All native holdup experiments were carried out with 2 biological replicates and their analytics were done with 3 technical replicates for each biological replicate (n = 2x3). GATA1 values were determined by Western-blots from the same samples and highlighted in green in each volcano plots. **E.** Western-blot analyses of GATA1 levels in the native holdup reactions subjected to mass spectrometry analyses in B-D (2 biological replicates and 3 technical replicates for each biological replicate, n = 2x3). GAPDH was used as non-binding loading control. **F.** The number of detected significant partners for each oligonucleotide baits visualized with a Venn diagram. Significance was determined based on hyperbolic significance threshold at 2σ. **G.** Volcano plot visualization of the native holdup results with the 51_WT bait compared to SCR1 bait. **H.** Volcano plot visualization of the native holdup results with the 51_WT bait compared to SCR2 bait. In both cases, GATA1 (green) has been identified as a significant sequence-specific binder (at 2σ). For further data see Supplementary data 3.

Figure 4. Biochemical characterization of DNA binding of GATA1 to the WT and haplotype-carrying *ATP2B4* promoter elements using nHU titration coupled to Western-blot on K562 cell extracts. **A.** Western-blot results of the titrations with the different baits showed in Figure 2A (n=2). **B.** Hyperbolic binding curves were fitted on fraction bound values (BI) to each data showed in A to calculate apparent affinities (K_{app}). 25-30% of the GATA1 protein found in the cell extract was inactive (*f*: active fraction). **C.** The results of the repeated titration experiments with diluted cell extracts (0.1 mg/ml, n=2). Left: 4945_WT bait, right: 4945 haplotype bait. **D.** Validation of measured affinity values and the estimated bait concentrations of the native holdup experiments, by a native holdup titration using a previously characterized palindromic GATA1 bait (n=2) (53).

Figure 5. Results of the luciferase reporter assays with WT and different mutant forms of GATA1 using K562 cells and a reporter plasmid containing the full haplotype promoter region of the *ATP2B4* gene and the localization of the R307C GATA1 protein. **A.** Expressions of mutation carrying HA-GATA1 proteins in the samples subjected to luciferase assay (3 biological and 2 technical replicates, two-tailed t-test on the average of the values measured for biological replicates, *: $p<0.05$) determined by Western-blot and expressed in percentage of WT (100%). For the Western-blot images see Supplementary Figure S3. **B.** The activity of the different mutant GATA1 proteins in the luciferase assays compared to HA transfected (0%) and HA-GATA1-WT (100%) transfected cells (3 biological and 2 technical replicates, two-tailed t-test on the average of the values measured for biological replicates, ns: non-significant, *: $p<0.05$, **: $p<0.01$). Grey: WT-like, green: GOF (gain of function), red: LOF (loss of function). For the raw data see Supplementary Data 4. **C.** Localization of the HA-tagged WT and R307C mutant GATA1 proteins in transfected HEK293T cells (scale bar: 20 μm). Magenta: HA-GATA1, green: DAPI. For further images see Supplementary Figure S4.

Figure 6. The results of nHU titrations coupled to WB with the different mutant forms of GATA1 using the 51_WT and haplotype sequences and the summary of the effects of different mutations. **A.** The hyperbolic binding equations were fitted to the native holdup titration data and plotted. The 51_WT oligonucleotide was used as archetypal tandem GATA motif. We used HEK293T cell extracts previously transfected with plasmids containing different mutant forms of GATA1 protein. For the Western-blot images and the results with other mutants see Supplementary Figure S6. **B.** Hyperbolic binding curves on titrations with the 51_HAP (haplotype) sequence as archetypical single GATA motif (performed as in A). For the Western-

blot images and the results with other mutants see Supplementary Figure S7. **C.** Summary of the results of the nHU titrations. Left: The different mutations in GATA1 is plotted on the structure of the protein (PDB: 3VD6) (54). Right: K_{app} values calculated from native holdup titrations with the 51_WT and 51_HAP sequences. Activity: transcription activation measured in the luciferase assay for the different mutant GATA1 proteins (ref: reference activity (WT), ns: non-significantly different (WT-like), GOF: gain of function, LOF: loss of function).

Supplementary Figure Legends

Supplementary Figure 1. Western-blot images of single point nHU experiments with 51_WT, SCR1 and SCR2 sequences.

Supplementary Figure 2. Western-blot images of titration nHU experiments with the 4945_WT, 4945_HAP, 51_WT, 51_HAP, or with the $CTRL_{PALINDROME}$ sequences on K562 cell extracts either at 1 mg/ml total protein concentration (upper panel) or 0.1 mg/ml total protein concentration (lower panel).

Supplementary Figure 3. Western-blot images on transfected K562 cell extracts that were used in luciferase assay, to determine the HA-GATA1 expressions.

Supplementary Figure 4. Additional images on the localization determination of different mutant versions of the GATA1. HEK293T cells were transfected with HA-GATA1 WT or mutant plasmids and immunostaining was carried out using an antibody recognizing the HA epitope (magenta), and the nucleus was visualized by using DAPI staining (green). Scale bar: 20 µm.

Supplementary Figure 5. Validation of transient expressions of the different mutant versions of GATA1 in HEK293T cells by Western-blot. The HA antibody recognizes all mutants, but the GATA1 antibody doesn't recognize the short version of GATA1 (GATA1s) as the epitope recognized by the antibody might located at the N-terminal part of the protein.

Supplementary Figure 6. A. Western-blot images of nHU titrations with the 51_WT sequence on cell extracts prepared of HA-GATA1 WT and mutant transfected HEK293T cells. **B.** Determined hyperbolic binding curves on titrations with the 51_WT bait and with the mutant GATA1 proteins that are not shown in Figure 6.

Supplementary Figure 7. A. Western-blot images of nHU titrations with the 51_HAP sequence on cell extracts prepared of HA-GATA1 WT and mutant transfected HEK293T cells. **B.** Determined hyperbolic binding curves on titrations with the 51_HAP bait and with the mutant GATA1 proteins that are not shown in Figure 6.

Supplementary Data Legends

Supplementary Data 1. Results of the nHU-MS experiments. The first four sheet contains the measured $\log_2$ fold change values, statistical significances ($-\log_{10}P$), calculated binding intensities, apparent affinities and significance thresholds. In cases where the significance threshold is smaller than the obtained statistical significance of the measurement, the apparent

affinity is considered significant and a numerical affinity value is displayed in the p$K_d$sign column. Proteins that are classified as transcription factors are highlighted in the last column. The last tab contains an assembled affinity matrix of all significant affinities.

Supplementary Data 2. Results of the GO enrichment. On the first sheet, an assembled affinity matrix of all significant affinities is shown with the last column showing the number of significant measurements for each prey. GO enrichment analysis was performed for all proteins with >0 values in this column or for the promiscuous proteins for proteins with a value of 4 in this column. The following sheets contain the significant GO terms for cellular compartments or for molecular functions.

Supplementary Data 3. Results of the nHU-MS experiments of the scrambled DNA fragments. Similarly to Data 1, each sheet contains the measured $\log_2$ fold change values, statistical significances ($-\log_{10}P$), calculated binding intensities, apparent affinities and significance thresholds and the significant affinities. The first three sheet contains the measurement of the SCR1, SCR2 and the 51WT baits against a biotin control. The last two sheets contain the measurement of the 51WT bait against the SCR1, or SCR2 controls.

Supplementary Data 4. Results of the luciferase assay measurements and the related Western-blots for expression determination. The first sheet contains the luminescence values measured with transfected K562 cells and an *ATP2B4* erythroid-promoter construct. The assay was performed with 3 biological and 2 technical replicates, data were normalized that the average of HA was 0 and the WT was 100 (%). The second sheet contains expression result by densitometry on Western blots, on the same samples as the luminescence was measured. For the Western-blot images see Supplementary Figure S3.

Figure 1.

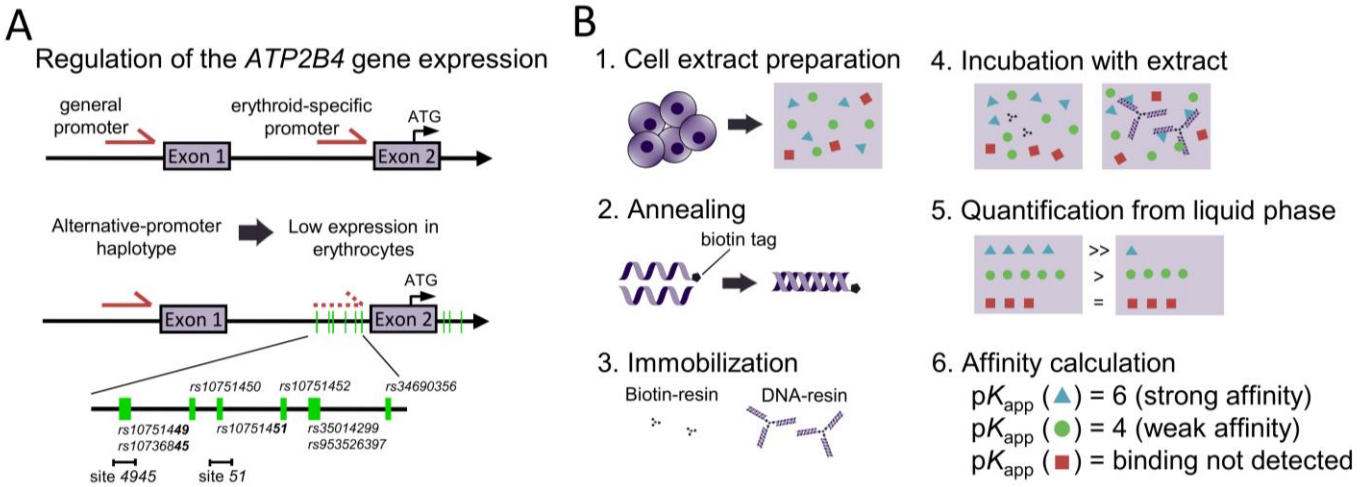

A
Regulation of the ATP2B4 gene expression
general promoter
erythroid-specific promoter
ATG
Exon 1
Exon 2
Alternative-promoter haplotype
Low expression in erythrocytes
ATG
Exon 1
Exon 2
rs10751450
rs10751452
rs34690356
rs10751449
rs10736845
rs10751451
rs35014299
rs953526397
site 4945
site 51
B
1. Cell extract preparation
2. Annealing
biotin tag
3. Immobilization
Biotin-resin
DNA-resin
4. Incubation with extract
5. Quantification from liquid phase
>>
>
=
6. Affinity calculation
pKapp (▲) = 6 (strong affinity)
pKapp (●) = 4 (weak affinity)
pKapp (■) = binding not detected

Figure 2.

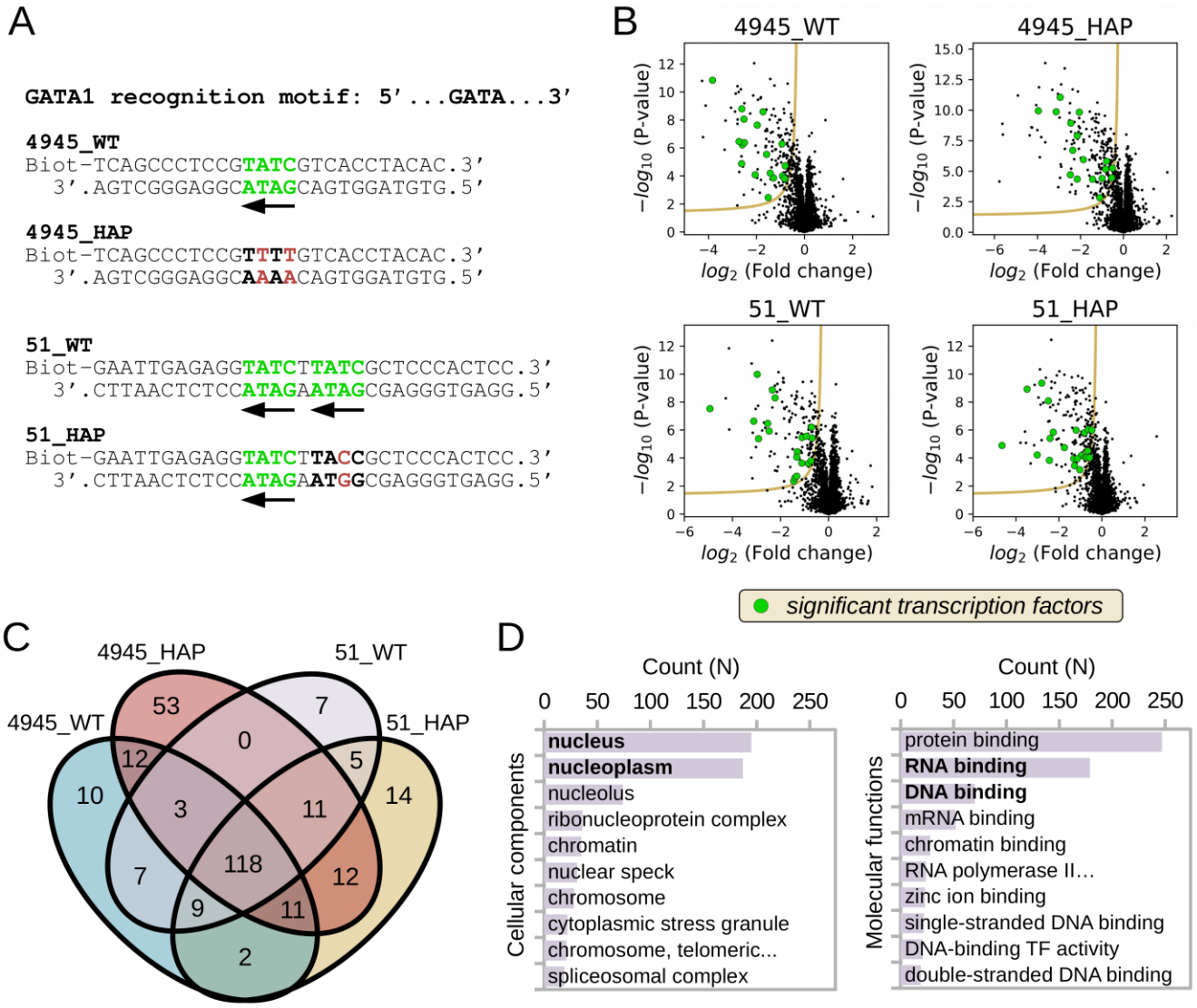

A
GATA1 recognition motif: 5'...GATA...3'
4945_WT
Biot-TCAGCCCTCCGTATCGTCACCTACAC.3'
3'.AGTCGGGAGGCATAGCAGTGGATGTG.5'
4945_HAP
Biot-TCAGCCCTCCGTTTTGTCACCTACAC.3'
3'.AGTCGGGAGGCAAAACAGTGGATGTG.5'
51_WT
Biot-GAATTGAGAGGTATCTTATCGCTCCCACTCC.3'
3'.CTTAACTCTCCATAGAATAGCGAGGGTGAGG.5'
51_HAP
Biot-GAATTGAGAGGTATCTTACCGCTCCCACTCC.3'
3'.CTTAACTCTCCATAGAATGGCGAGGGTGAGG.5'
B
4945_WT
4945_HAP
51_WT
51_HAP
−log10 (P-value)
log2 (Fold change)
significant transcription factors
C
4945_WT
4945_HAP
51_WT
51_HAP
53
7
0
12
5
10
3
11
14
118
7
12
9
11
2
D
Count (N)
Cellular components
nucleus
nucleoplasm
nucleolus
ribonucleoprotein complex
chromatin
nuclear speck
chromosome
cytoplasmic stress granule
chromosome, telomeric...
spliceosomal complex
Molecular functions
protein binding
RNA binding
DNA binding
mRNA binding
chromatin binding
RNA polymerase II...
zinc ion binding
single-stranded DNA binding
DNA-binding TF activity
double-stranded DNA binding

Table 1.

Identified transcription factor binding partners			Measured apparent affinities (pK_{app})				
Accession	**Protein**	**Gene**	**4945_WT**	**4945_HAP**	**51_WT**	**51_HAP**	**Effect of haplotype**
Q6ZN17	Protein lin-28 homolog B	LIN28B	6.01	6.46	6.14	6.31	$\Sigma\Delta pK_{app} < -0.4$
P17544	Cyclic AMP-dependent transcription factor ATF-7	ATF7	5.07	5.14	4.84	5.18	
P15336	Cyclic AMP-dependent transcription factor ATF-2	ATF2	5.60	5.80	5.12	5.32	$\lvert\Sigma\Delta pK_{app}\rvert < 0.4$
P16989	Y-box-binding protein 3	YBX3	5.66	5.84	5.87	6.08	
Q02447	Transcription factor Sp3	SP3	4.52	4.47	5.01	5.39	
Q96QR8	Transcriptional activator protein Pur-beta	PURB	4.64	4.93	4.85	4.88	
P67809	Y-box-binding protein 1	YBX1	5.46	5.53	5.48	5.67	
P17483	Homeobox protein Hox-B4	HOXB4	5.01	5.28	5.47	5.46	
P16220	Cyclic AMP-responsive element-binding protein 1	CREB1	5.76	5.96	5.36	5.41	
P17535	Transcription factor jun-D	JUND	5.80	5.92	5.37	5.47	
Q6DD87	Zinc finger protein 787	ZNF787	5.98	5.95	5.91	6.15	
Q00577	Transcriptional activator protein Pur-alpha	PURA	5.18	5.30	5.20	5.27	
Q8N5P1	Zinc finger CCCH domain-containing protein 8	ZC3H8	4.92	4.95	5.08	5.10	
P17482	Homeobox protein Hox-B9	HOXB9	6.01	6.13	6.12	5.95	
P17028	Zinc finger protein 24	ZNF24	5.18	5.39	5.54	5.24	
P85037	Forkhead box protein K1	FOXK1	5.26	5.16	5.10	5.04	
Q96JP5	E3 ubiquitin-protein ligase ZFP91	ZFP91	5.57	5.36	5.48	5.45	
O95785	Protein Wiz	WIZ	5.17	4.96	5.09	5.06	
Q8N1G0	Zinc finger protein 687	ZNF687	5.20	4.93	5.12	5.10	
Q9P016	Thymocyte nuclear protein 1	THYN1	6.42	6.19	6.77	6.68	
P17096	High mobility group protein HMG-I/HMG-Y	HMGA1	5.97	5.84	6.19	5.97	
P49711	Transcriptional repressor CTCF	CTCF	6.05	5.72	5.98	5.94	
Q9NP66	High mobility group protein 20A	HMG20A	4.98	4.87	5.26	4.98	
P49116	Nuclear receptor subfamily 2 group C member 2	NR2C2	5.22	4.89	4.64	4.51	$\Sigma\Delta pK_{app} > 0.4$
Q9H5H4	Zinc finger protein 768	ZNF768	5.53	5.19	5.25	5.10	
Q06330	Recombining binding protein suppressor of hairless	RBPJ	5.25	4.96	5.51	5.16	
P15976	**Erythroid transcription factor**	**GATA1**	**6.00**	**5.27**	**5.96**	**5.88**	

Figure 3.

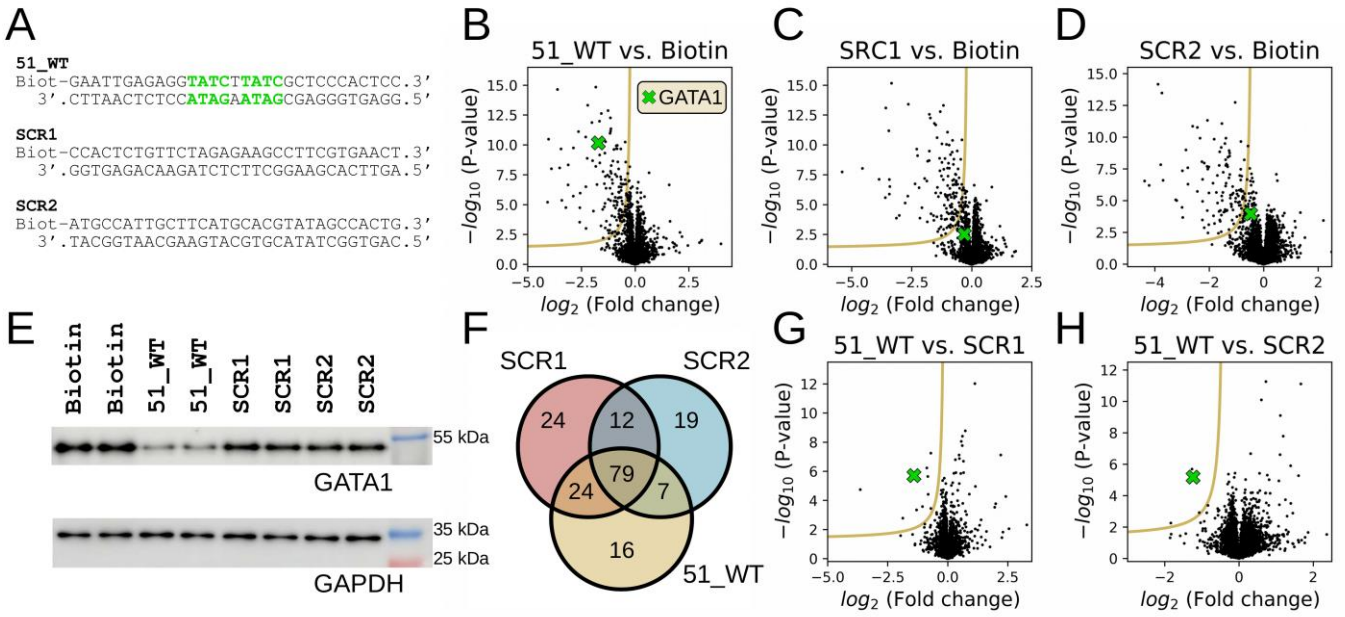

A
51_WT
Biot-GAATTGAGAGGTATCTTATCGCTCCCACTCC.3'
3'.CTTAACTCTCCATAGAATAGCGAGGGTGAGG.5'
SCR1
Biot-CCACTCTGTTCTAGAGAAGCCTTCGTGAACT.3'
3'.GGTGAGACAAGATCTCTTCGGAAGCACTTGA.5'
SCR2
Biot-ATGCCATTGCTTCATGCACGTATAGCCACTG.3'
3'.TACGGTAACGAAGTACGTGCATATCGGTGAC.5'
B
51_WT vs. Biotin
GATA1
C
SRC1 vs. Biotin
D
SCR2 vs. Biotin
−log10 (P-value)
log2 (Fold change)
E
Biotin
Biotin
51_WT
51_WT
SCR1
SCR1
SCR2
SCR2
55 kDa
GATA1
35 kDa
25 kDa
GAPDH
F
SCR1
SCR2
51_WT
24
12
19
79
24
7
16
G
51_WT vs. SCR1
H
51_WT vs. SCR2

Figure 4.

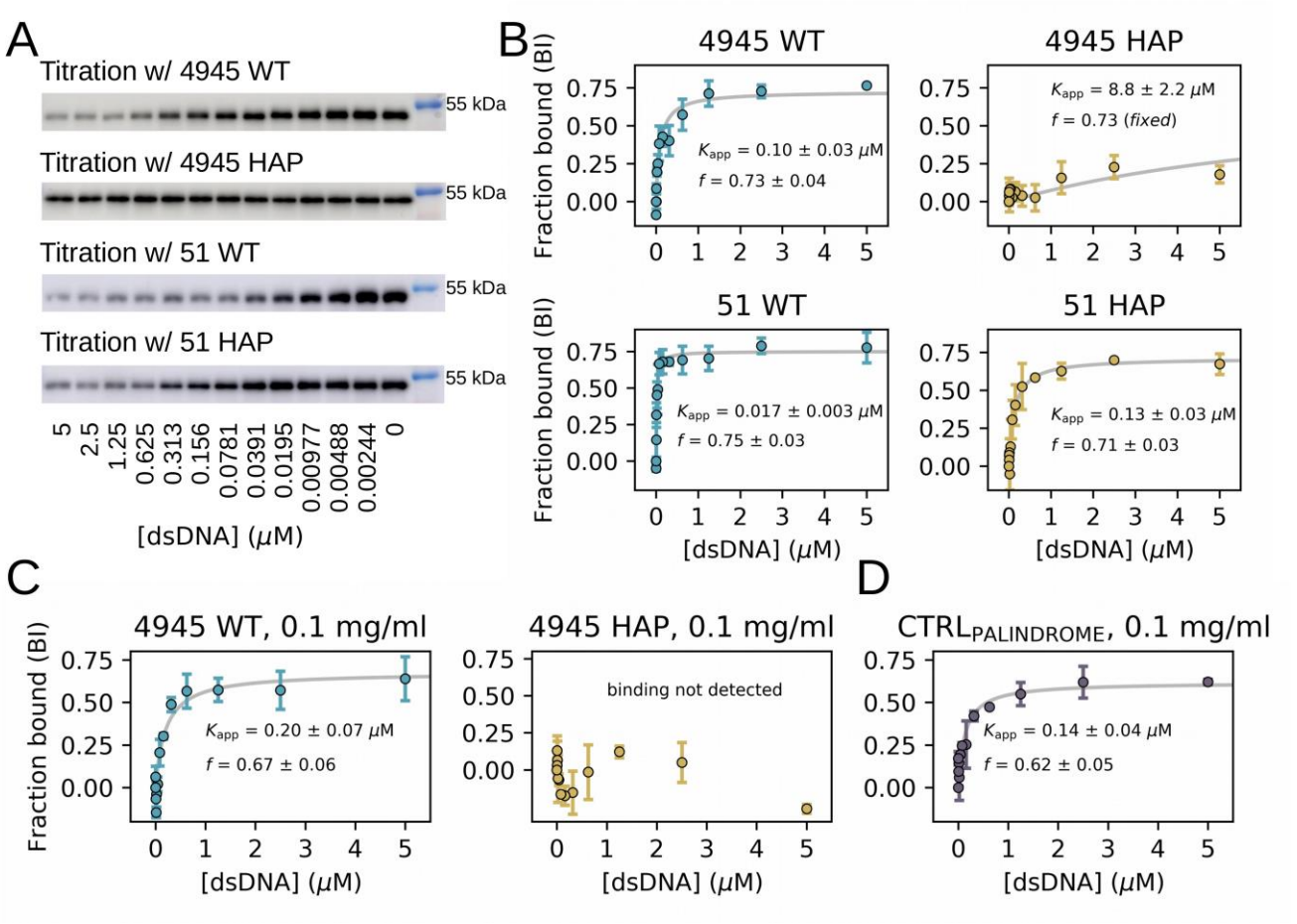

A
Titration w/ 4945 WT
55 kDa
Titration w/ 4945 HAP
55 kDa
Titration w/ 51 WT
55 kDa
Titration w/ 51 HAP
55 kDa
5
2.5
1.25
0.625
0.313
0.156
0.0781
0.0391
0.0195
0.00977
0.00488
0.00244
0
[dsDNA] (μM)
B
4945 WT
Fraction bound (BI)
K_{app} = 0.10 ± 0.03 μM
f = 0.73 ± 0.04
4945 HAP
K_{app} = 8.8 ± 2.2 μM
f = 0.73 (fixed)
51 WT
K_{app} = 0.017 ± 0.003 μM
f = 0.75 ± 0.03
51 HAP
K_{app} = 0.13 ± 0.03 μM
f = 0.71 ± 0.03
[dsDNA] (μM)
C
4945 WT, 0.1 mg/ml
Fraction bound (BI)
K_{app} = 0.20 ± 0.07 μM
f = 0.67 ± 0.06
4945 HAP, 0.1 mg/ml
binding not detected
[dsDNA] (μM)
D
$CTRL_{PALINDROME}$, 0.1 mg/ml
K_{app} = 0.14 ± 0.04 μM
f = 0.62 ± 0.05
[dsDNA] (μM)

Figure 5.

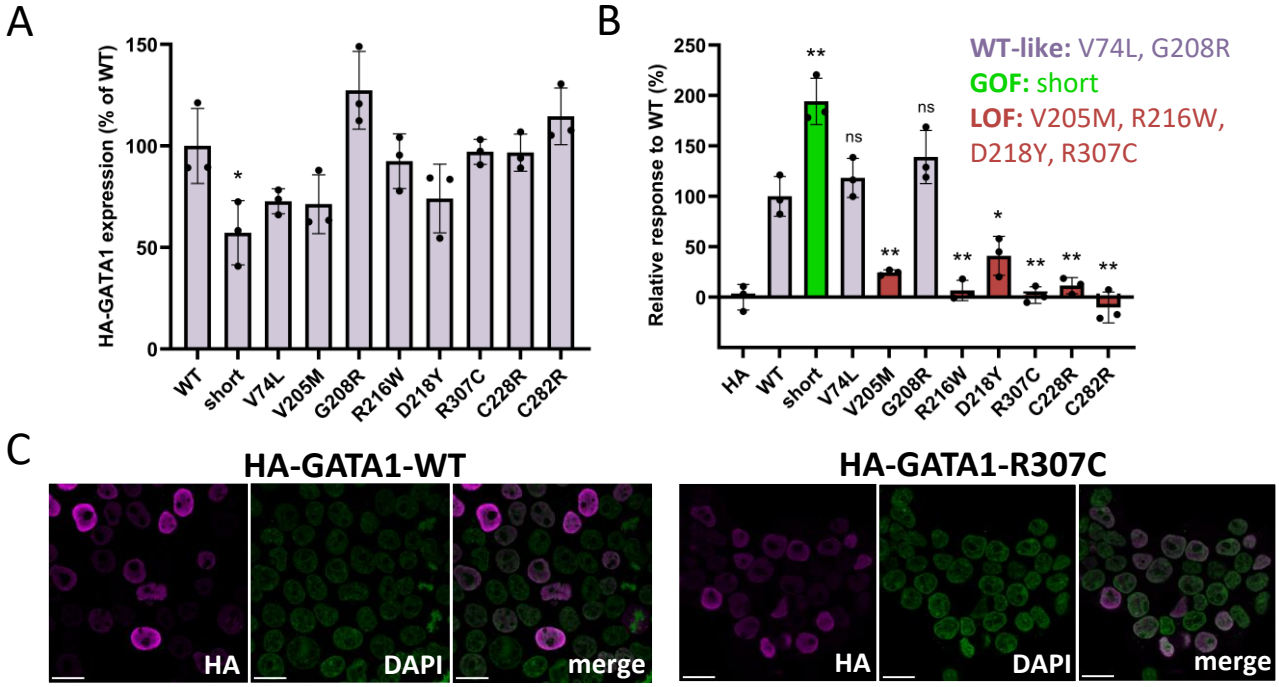
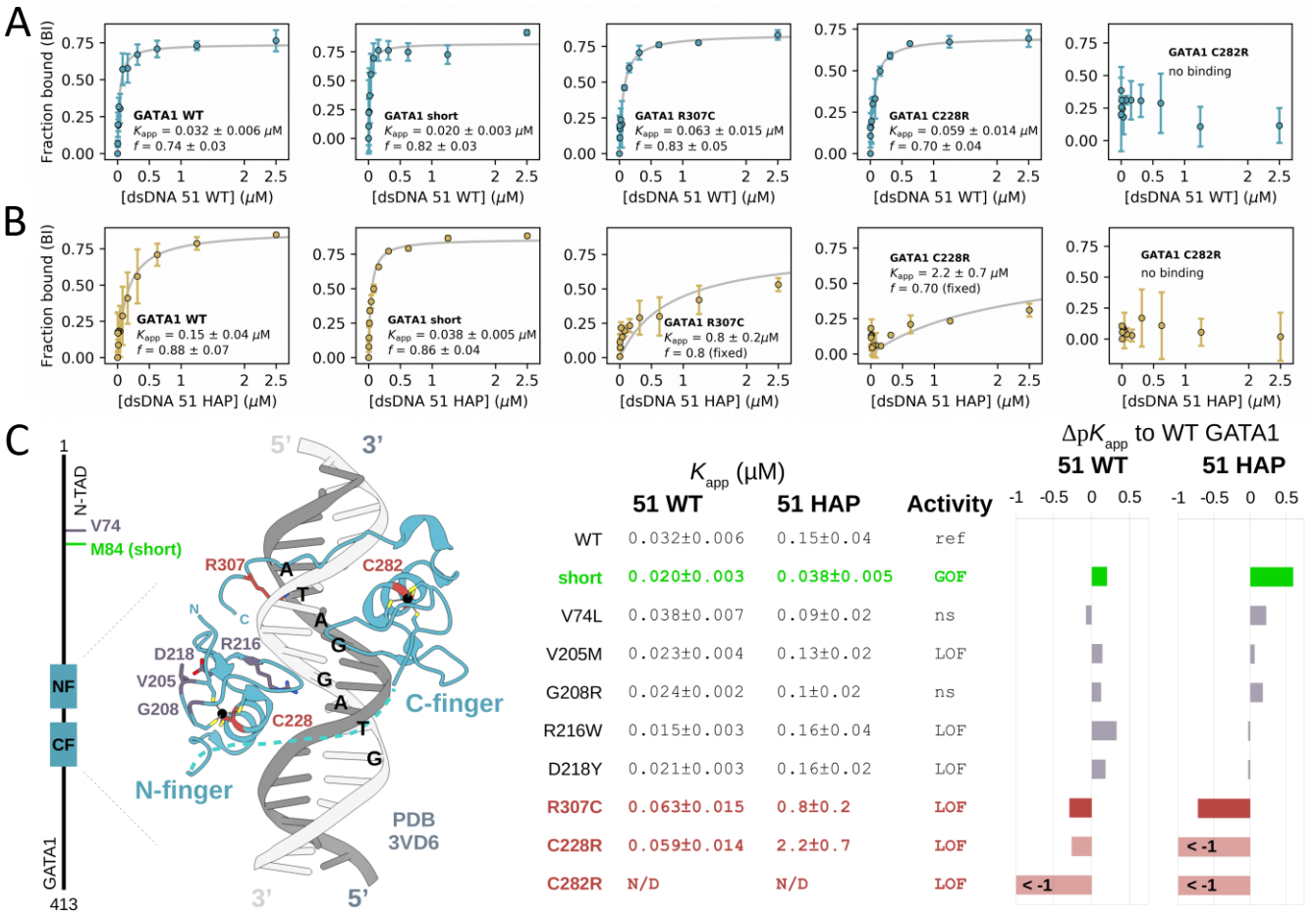

A
HA-GATA1 expression (% of WT)
150
100
50
0
*
WT
short
V74L
V205M
G208R
R216W
D218Y
R307C
C228R
C282R
B
Relative response to WT (%)
250
200
150
100
50
0
**
ns
ns
**
**
*
**
**
**
HA
WT
short
V74L
V205M
G208R
R216W
D218Y
R307C
C228R
C282R
WT-like: V74L, G208R
GOF: short
LOF: V205M, R216W, D218Y, R307C
C
HA-GATA1-WT
HA-GATA1-R307C
HA
DAPI
merge
HA
DAPI
merge

Figure 6.


A
Fraction bound (BI)
GATA1 WT
K_{app} = 0.032 ± 0.006 μM
f = 0.74 ± 0.03
GATA1 short
K_{app} = 0.020 ± 0.003 μM
f = 0.82 ± 0.03
GATA1 R307C
K_{app} = 0.063 ± 0.015 μM
f = 0.83 ± 0.05
GATA1 C228R
K_{app} = 0.059 ± 0.014 μM
f = 0.70 ± 0.04
GATA1 C282R
no binding
[dsDNA 51 WT] (μM)
B
GATA1 WT
K_{app} = 0.15 ± 0.04 μM
f = 0.88 ± 0.07
GATA1 short
K_{app} = 0.038 ± 0.005 μM
f = 0.86 ± 0.04
GATA1 R307C
K_{app} = 0.8 ± 0.2μM
f = 0.8 (fixed)
GATA1 C228R
K_{app} = 2.2 ± 0.7 μM
f = 0.70 (fixed)
GATA1 C282R
no binding
[dsDNA 51 HAP] (μM)
C
1
N-TAD
V74
M84 (short)
NF
CF
GATA1
413
5'
3'
R307
C282
N
C
D218
R216
V205
G208
C228
C-finger
N-finger
PDB
3VD6
K_{app} (μM)
51 WT
51 HAP
Activity
ΔpKapp to WT GATA1
51 WT
51 HAP
WT 0.032±0.006 0.15±0.04 ref
short 0.020±0.003 0.038±0.005 GOF
V74L 0.038±0.007 0.09±0.02 ns
V205M 0.023±0.004 0.13±0.02 LOF
G208R 0.024±0.002 0.1±0.02 ns
R216W 0.015±0.003 0.16±0.04 LOF
D218Y 0.021±0.003 0.16±0.02 LOF
R307C 0.063±0.015 0.8±0.2 LOF
C228R 0.059±0.014 2.2±0.7 LOF
C282R N/D N/D LOF
< -1
< -1
< -1