

Analysis of Sequence Variation Underlying Tissue-specific Transcription Factor Binding and Gene Expression

Karen M. Lower¹, Marco De Gobbi¹, Jim R. Hughes, Christopher J. Derry, Helena Ayyub, Jacqueline A Sloane-Stanley, Douglas Vernimmen, David Garrick, Richard J. Gibbons & Douglas R. Higgs²

MRC Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Oxford, UK.

¹ These authors contributed equally to this work

² To whom correspondence should be addressed. Professor Douglas R. Higgs; MRC Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Headington OX3 9DS, UK.

E-mail: doug.higgs@imm.ox.ac.uk; Phone: +44 1865 222393; Fax +44 1865 222 424

Abstract

Whereas mutations causing monogenic disorders most frequently lie within the affected gene, sequence variation in complex disorders is more commonly found in non-coding regions. Furthermore, recent genome-wide studies have shown that common DNA sequence variants in intergenic regions are associated with “normal” variation in gene expression resulting in cell-specific and/or allele-specific differences. The mechanism by which such sequence variation causes changes in gene expression is largely unknown. We have addressed this by studying natural variation in the binding of key transcription factors (TFs) in the well-defined, purified cell system of erythropoiesis. We have shown that common polymorphisms frequently directly perturb the binding sites of key TFs, and detailed analysis shows how this causes considerable (~10 fold) changes in expression from a single allele in a tissue-specific manner. We also show how a single nucleotide polymorphism, located at some distance from the recognised TF binding site, may affect the recruitment of a large multiprotein complex and alter the associated chromatin modification of the variant regulatory element. This study illustrates the principles by which common sequence variation may cause changes in tissue-specific gene expression, and suggests that such variation may underlie an individual’s propensity to develop complex human genetic diseases.

Key Words

Transcriptional Regulation, Transcription Factor Binding, Polymorphism, Allele-Specific

Introduction

With few exceptions (such as X-inactivation and imprinting), all non-random differences in gene expression should ultimately be explained by a change in the DNA code. The numerous mutations that are now known to cause recognisable single gene disorders usually lie within the affected gene. Whilst some may exert their effect through creation (e.g. De Gobbi et al., 2006) or destruction (e.g. Benko et al., 2009) of a regulatory element, examples of single point mutations in regulatory elements causing human genetic disease are relatively rare. Recent genome-wide studies have shown that some common DNA sequence variants (single nucleotide polymorphisms [SNPs] and copy number variants [CNVs]) are associated with “normal” variation in gene expression (Kasowski et al., 2010; McDaniell et al., 2010). Changes in *trans*-acting factors may alter expression of both alleles of a target gene, whereas *cis*-acting variants most frequently cause changes in allele-specific expression.

In contrast to the single gene disorders, a large number of SNPs and CNVs associated with normal variation in gene expression, or associated with common complex diseases (such as diabetes, hypertension, and autoimmune disorders), often lie in non-genic regions. Whilst the mechanisms by which these sequence variants exert their effect are not well understood, they are expected to affect tissue-specific, *cis*-acting, regulatory elements. This is consistent with the observation that most normal variation in gene expression occurs in an allele-specific and tissue-specific manner (Dimas et al., 2009; Ge et al., 2009; Zhang et al., 2009). A key goal in current biology is to explain the mechanisms by which such common DNA sequence variants cause significant changes in gene expression and how this eventually relates to altered phenotypes associated with human disease. A major hurdle in analysing

such mechanisms is to identify and purify the appropriate cell type to study. Clearly a regulatory polymorphism will only affect expression when the regulatory element is active in a relevant cell type and at the appropriate stage of differentiation and development. In most cases, this information is not available and may be extremely difficult to establish.

To determine the principles by which common variants may cause allele-specific, cell-specific changes in gene expression we have studied the effects of natural DNA polymorphisms during formation of human red blood cells. Erythropoiesis provides an excellent model system in which to address such mechanisms. Purified progenitors can be obtained in large numbers and the tissue-specific regulatory elements (genome-wide) which control gene expression during erythropoiesis have been well characterised (Garriick et al., 2008; Hattangadi et al., 2011). These elements are identified by their association with DNase1 hypersensitive sites (DHS) and active chromatin marks. Furthermore, the underlying sequence contains common transcription factor (TF) binding sites including those for GATA1, Scl/TAL1 and Sp/XKLF proteins. These proteins, in turn, are thought to recruit multiprotein complexes which bind the regulatory elements and impose tissue-specific expression on their cognate genes, which are the building blocks of red blood cells.

Here we have compared the genome-wide occupancy of Scl/TAL1 in erythroid cells between two different individuals. We find that apparently non-pathological sequence variation in and around Scl/TAL1 binding sites determines whether multiprotein complexes do or do not bind particular regulatory elements. To establish the mechanism by which such polymorphisms may cause allele-specific, tissue-specific changes in gene expression, we have investigated one variant site in detail, showing how a single common polymorphism

mediates changes in the binding of many different components of a large multiprotein complex giving rise to allele- and tissue-specific expression of the gene that is regulated by this element.

Materials and Methods

Cell types. Erythroid cells were obtained using a two-phase culture system as previously described (Pope et al., 2000). Non-erythroid cells were EBV-transformed B lymphocyte cell lines.

ChIP and ChIP-Seq. Chromatin immunoprecipitation was performed as previously described (De Gobbi et al., 2007). Briefly, for one immunoprecipitation, 1×10^7 cultured primary human erythroblasts or EBV-transformed B lymphocytes were cross-linked with 1% formaldehyde for 10 minutes. DNA was sheared by sonication to fragments under 500 base pairs. Antibodies used were H3ac (06-599, Millipore), H3K4me3 (ab8580, Abcam), H3K4me1 (07-436, Millipore), RNA polymerase II (sc-9001, Santa Cruz), GATA-1 (sc-1234, Santa Cruz), and Scf/TAL1 (gifted by C. Porcher). ChIP DNA was analysed by quantitative real time PCR, calculated relative to input and normalized to a positive control. For details of primers and probe sequences see Table S1. ChIP DNA was processed for Illumina High-throughput sequencing according to Illumina protocol.

ChIP-Seq data analysis. Reads were aligned to the repeat masked hg18 reference genome (UCSC Genome Browser repeatmasker track) using bowtie (version 0.12.3) (Langmead et al.,

2009) with the `-m` reporting option set to 2. Replicates C1a and C1b yielded 4.7 million and 22 million, and C2a and C2b 4.1 million and 8.5 million uniquely aligned reads, respectively (Supp. Fig. S1A, GEO accession number GSE42390). Since it has been shown that in ChIP-Seq studies the sensitivity of peak detection increases with the number of mapped reads (Fujiwara et al., 2009, Bernstein et al., 2012), reads from all replicates were merged. Peak detection was performed as follows using SeqMonk Software

(<http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/>). SeqMonk, using the “contig probe generator” option (230 bp contig of overlapping reads with a depth cutoff of 3), was used to call peaks. Peaks in C1 and C2 were then quantified by normalising the number of reads in each peak for the total number of reads detected within peaks. Of the 4538 total peaks initially identified, in order to minimize background noise only peaks containing a minimum of 32 reads in either C1b or C2b were retained. These peaks were then overlapped and annotated with the Encode generated file

(<ftp://hgdownload.cse.ucsc.edu/apache/htdocs/goldenPath/hg18/encodeDCC/wgEncodeMapability/wgEncodeDukeRegionsExcluded.bed6.gz>), which contains regions in the genome which strongly overreact in high-throughput sequencing experiments due to large copy number differences between the real genome and the genome build, and therefore normalize poorly. After removal of these regions, 2936 peaks remained.

In order to identify differentially bound regions between C1 and C2, a two-classes paired-test Rank Product analysis (500 permutations, $FDR < 0.2$) was performed (MeV4.6 TM4 Software). Whilst an FDR of 0.2 is relatively permissive, we hypothesised this was reasonable given the relatively small number of peaks analysed, and that a subset of identified differences were subject to subsequent confirmation by alternative methods (e.g. RT-PCR).

Expression analysis. Expression analysis was carried out as previously described (Lower et al., 2009). The mean of expression of *NME4* in Group A samples (G/G) is set to 100%, and expression in Group B samples (G/C) is expressed relative to this mean, for each tissue analysed.

Electrophoretic mobility shift assay (EMSA). Nuclear extracts from K562 erythroleukaemia cells were prepared, and EMSAs were carried out as described previously (Crossley et al., 1996). Double strand oligonucleotides were prepared by annealing 28 bp oligonucleotides containing either the G or the C variant of the *NME4* intronic GC-box (chr16:389546-389573, hg18 build sequence; Table S2), radiolabeled and incubated with 3-5 µg nuclear extract protein. For antibody supershift assay, 0.4 µg of Sp1 (sc-59, Santa Cruz), Sp3 (sc-13018, Santa Cruz), EKLF and BKLF (both gifted by M. Crossley) antibody was added before addition of the probe, and samples were incubated on ice for 2 hours. The protein/DNA interaction were resolved on a 5% non denaturing polyacrylamide gel (37.5 :1 acrylamide:bis-acrylamide, 0.25X TBE), followed by exposure to x-ray film.

Statistical analysis. Significance of differences in expression between Groups A and B was calculated with a 2-tailed Student's t-test assuming non-equal variance.

Pyrosequencing. The ratio of expression of allele-specific transcripts of *NME4* was ascertained by pyrosequencing. Primer and dispensation information has been published previously (Lower et al. 2009). Peak height is directly proportional to the amount of nucleotide incorporated. Analysis was performed in duplicate and an average obtained.

Sequence information. All human sequence positions correspond to the International Human Genome Sequencing Consortium Human Mar. 2006 (hg18) Assembly sequence. The *NME4* gene corresponds to sequence position chr16:387193-390755; *eNME4* corresponds to sequence position chr16:389609-390755. SNP reference numbers refer to dbSNP build 132.

Results

The Genome-wide Occupancy of Scl/TAL1 In Human Primary Erythroid Cells

ChIP-Seq experiments using an Scl/TAL1 antibody were carried out in primary erythroid cells cultured from two individuals of Caucasian origin (C1 and C2). Cell surface expression of the transferrin receptor (CD71) and glycophorin A (GPA) was analysed in order to ensure that all cell populations were harvested at equivalent stages of erythroid differentiation (Supp. Fig. S2A). ChIP-Seq libraries were prepared from two biological replicates for each individual after validation by qPCR analysis of Scl/TAL1 binding at known targets within the α and β -globin loci (Supp. Fig. S2B).

Global analysis of these data sets (see Materials and Methods and Fig. S1A) identified a total of 2936 Scl/TAL1 bound regions from both individuals that were used for further analysis. This approach identified known functional Scl/TAL1 target sequences, previously reported in human and in mouse erythroid cells (including the α globin MCS-Rs, β globin LCR, ALAD, EPB42, GATA1, GYPA, KIT, KLF1, LMO2, PRDX2, RUNX1, Scl/TAL1, TRIM10) (Fig. 1 and Supp. Table S3). Compared to a recently published global map of

ScI/TAL1 binding in human pro-erythroblasts (Palii et al., 2011) we identified fewer binding sites, which can be mostly accounted for by the increased stringency of our analysis. Nevertheless, these peaks showed a very high overlap (80%) with the previously published data (Palii et al., 2011) (Supp. Fig. S1B), confirming that we had successfully identified *bona fide* ScI/TAL1 targets in this analysis.

To identify the specific sequences responsible for the chromatin occupancy by ScI/TAL1, we undertook MEME and DREME *de novo* motif finding analysis (Machanick and Bailey, 2011) (Supp. Fig. S3). We identified the previously known predominant ScI/TAL1 binding sequence (WGATAR; GATA motif), confirming that most often ScI/TAL1 binding is directed by interactions with GATA factors. Our dataset revealed the combinatorial CTG(n9)GATA motif as the most significant composite motif (MEME output in Supp. Fig. S3). In addition, other motifs were also significantly enriched at the ScI/TAL1 targets, as predicted by JASPAR CORE 2009 (Supp. Fig. S3). Amongst these were binding sites for known TFs (Sp/XKLF, RUNX1, NFE2), implying that the combined action of ubiquitous and tissue-specific TFs can determine the specificity of ScI/TAL1 binding *in vivo*.

ScI/TAL1 Occupancy Varies Between Individuals

Having produced a high-resolution map of ScI/TAL1 genomic targets in erythroid cells, we next investigated how many of these targets varied between these two individuals. We first ensured that the number of sequence reads within ScI/TAL1 peaks were closely correlated between both individuals (Fig. 2A). We found a very tight correlation between

both individuals ($R^2=0.89$) with few outliers. In order to detect differential Scl/TAL1 bound regions, we used a two-class paired-test Rank Product analysis (MeV4.6 TM4 Software). This method detects regions that are consistently different in the two sets of replicated experiments, independent of their numerical intensities. Fourteen regions were found to bind more in C1 compared to C2 (Fig. 2A, red dots) and 11 regions more in C2 compared to C1 (Fig. 2A, purple dots) (500 permutations, FDR=0.2). Examples of the differential binding signal in the two individuals are shown in Fig. 2B-C, Supp. Fig. S4A-D and Fig. S5-S6. Having found that 25/2936 (0.8%) of Scl/TAL1 binding regions differed significantly between these two individuals, we validated a sample of 14 differential peaks (7 from each set of differential bound regions) using standard ChIP-qPCR. All interrogated regions were validated as true differential binding sites as compared to a set of control peaks, which were found to be of similar intensity in the two individuals (Supp. Fig. S4E).

We next examined the DNA sequences underlying the observed differential binding. Given that we had identified both the combinatorial CTG(n9)GATA binding site, and an enrichment of other TF binding sites within our identified Scl/TAL1 peaks (Fig. S3), we chose to analyse the underlying sequence in both individuals, extending 500bp either side of the differential Scl/TAL1 peaks (Supp. Table S4). This revealed that 100% of the 14 differential binding regions validated by qPCR contained at least one putative GATA binding site (Table 1). We found that WGATAR motifs were directly disrupted by SNPs in five regions (Peaks 1-5, Table 1), most likely resulting in the observed inability to bind Scl/TAL1. In another region, the identified polymorphism created a new E-box site adjacent to a WGATAR site (Peak 6, Table 1), presumably resulting in this sequence gaining an ability to bind Scl/TAL1. SNPs were also found in relative close proximity (7-85 bp) to a GATA site in four regions (Peaks 7-10, Table 1), whereas strikingly, in two regions, SNPs were >200 bp

away from the GATA site (Peaks 11-12, Table 1). Finally, one region was deleted, as part of a >1kb deletion (Peak 13, Table 1), and in one region no sequence differences were identified within 500bp of the differential binding peak (Peak 14, Table 1). The latter case indicates that other factors (e.g. trans-acting, epigenetic or long-range interaction) may play a role in TF recruitment. In this respect, it is notable that 2 regions (Peaks 7 and 11) which were bound by Scl/TAL1 in C1 and were not bound in C2 were only 25kb apart from each other (Supp. Fig. S4C).

Of the 14 differential peaks analysed in detail, only six could be attributed to a SNP (Peaks 1-5, Table 1), or deletion (Peak 13, Table 1), directly affecting the canonical Scl/TAL1 WGATAR binding site. Given the relatively small size of both our sample numbers and the number of differential peaks identified, this indicates that the importance of the sequence surrounding TF binding sites, and the role this surrounding sequence may play in functional variability, may currently be underestimated. In order to unequivocally link a sequence polymorphism to a functional effect, it is essential to carry out molecular studies at a locus specific level. Therefore, to analyse this phenomenon in more detail, we turned to a gene that we have previously studied in detail, *NME4*, whose erythroid specific expression is regulated by an erythroid element that binds a multiprotein complex which includes Scl/TAL1.

***NME4* is Bound by Scl/TAL1 and Exhibits Allele-Specific Expression**

NME4 is a widely expressed gene (encoding a dinucleotide kinase (Milon et al., 1997) that lies 200kb downstream of the α globin genes (which are exclusively expressed in erythroid cells). During evolution, *NME4* has acquired a non-conserved GATA-1 site, which

forms a new element that recruits both Scl/TAL1 and GATA1 *in vivo*, resulting in the production of an alternative erythroid-specific transcript (Fig. 3A) (Lower et al., 2009). We have previously shown that in erythroid cells the distal enhancer of the α globin genes physically interacts (over a distance of 300kb) with the *NME4* allele *in-cis*, and significantly up-regulates (by a factor of ~10-fold) expression of the full length *NME4* transcript compared to non-erythroid cells (Lower et al., 2009).

We analysed expression of *NME4* in a number of normal individuals. In non-erythroid cells, all individuals (Groups A and B) expressed similar levels of *NME4* (Fig. 3B) consistent with the presence of two functional *NME4* genes. However, in erythroid cells, one group of individuals (Fig. 3B, Group B) displayed a level of expression approximately 50% of others (Fig. 3B, Group A). Pyrosequencing studies performed on a known synonymous SNP (rs14293:G>A, referred to here on as the tracking SNP), lying within both the widely expressed and erythroid-specific transcripts of *NME4*, showed that the observed reduction of the erythroid-specific expression in group B was due to specific reduction in the expression of one allele of *NME4* (the allele with A at the tracking SNP) (Fig. 3C).

In view of our global survey of sequence variants affecting TF binding at GATA sites, we hypothesized that the allele-specific reduction of *NME4* may be due to the local effect of a sequence polymorphism affecting either the known GATA site or its surrounding region. We therefore sequenced 350 bp either side of the element associated with the transcription start site of *eNME4* in all individuals. Within this region we identified 7 SNPs, all of which are documented in dbSNP (Supp. Table S5). Of the 7 SNPs (6 being polymorphic in at least 3/7 individuals screened), only one (rs2071914:G>C) was found to segregate with the

observed *NME4* expression pattern i.e. was homozygous (G/G) in 4 individuals with erythroid bi-allelic expression of *NME4* (Group A in Fig. 3) and heterozygous (G/C) in individuals with highly-skewed erythroid expression of *NME4* (Group B in Fig. 3). Interestingly, this SNP was contained within a predicted GC-box lying 55 bp upstream of the *eNME4* transcription start site (TSS) (Fig. 3A).

SNP rs2071914 Abrogates the Binding of TFs *in vitro*

The finding of a polymorphic GC-box in the vicinity of a canonical GATA binding site was analogous to that seen in a region identified as having differential Scl/TAL1 binding in the genome-wide analysis (Peak 7, Table 1). Therefore, we performed functional studies to explore how the GC box could be involved in recruiting a TF complex (Sp/XKLF, Scl/TAL1, GATA) at the *NME4* locus.

Given that many Sp/XKLF factors (e.g. Sp1, Sp3, BKLF and EKLF) are known to recognize and bind to GC-box elements, we hypothesized that the C variant of the SNP within the GC-box may be affecting the recognition of this site by such TFs. Therefore we carried out electrophoretic mobility shift assays (EMSAs) using double-strand oligos containing either SNP variant (Supp. Table S2). The G variant oligonucleotide was found to show a complex protein-binding pattern (Fig. 4A, lane 2), which with the addition of specific antibodies was shown to include Sp1, Sp3 and BKLF (Fig. 4A, lanes 3,4 and 6). Anti-EKLF antibody did not cause a shift of this protein-DNA complex as EKLF protein is not expressed in K562 cells; this observation supports antibody specificity within this *in vitro* system (Fig.4A, lane 5). In contrast to the G oligonucleotide, the C oligonucleotide displayed no

protein binding at all (Fig. 4A lane 8), confirming, at least *in vitro*, that this polymorphism abrogates recognition of this sequence as a functional GC box.

SNP rs2071914 Perturbs Scl/TAL1 Recruitment at *eNME4*

We then went on to analyse the effect of this polymorphism on TF binding *in vivo*. We were unable to identify a C/C homozygote at this SNP, but we were able to carry out ChIP in a number of G/C heterozygotes at SNP rs2071914 (referred to here-on as the functional SNP). Due to limited sample material, and as we had already established abrogation of Sp1, Sp3 and BCLF binding at the GC-box *in vitro*, we went onto analyse the binding of tissue-specific TFs (Scl/TAL1 and GATA-1), as well as activating histone modifications (H3ac and H3K4me3), and RNA polymerase II (PolII) recruitment, by ChIP analysis in primary erythroblasts. This confirmed that individuals heterozygous for the functional SNP (C/G) displayed approximately 50% of the level of enrichment of these TFs or histone modifications (Fig. 4B). This suggests that the presence of an intact GC-box can be a determinant for recruitment of functionally active tissue-specific TFs at neighbouring elements, which by themselves are not capable of any *in vivo* regulatory activity. These results give functional support that GATA and GC-box sites co-operate, and contribute towards recruiting and/or stabilizing TF binding in a tissue-specific manner.

In addition to being either homozygous or heterozygous at the functional SNP, and regardless of their level of expression of *NME4*, these individuals were all heterozygous at the tracking SNP (rs14293) (Fig. 3A). PCR products containing both functional and tracking SNPs were cloned and sequenced in order to establish the phase for each SNP. In all

individuals, the A allele of the tracking SNP was in phase with the C allele of the functional SNP. Due to their close proximity, the functional and tracking SNPs are likely to both be present within the same sonicated fragments of DNA (sonication was carried out to generate fragments of ~500bp). Therefore, the tracking SNP was used in a pyrosequencing quantitative assay to determine which allele of *NME4* was enriched in chromatin fractions immunoprecipitated by antibodies recognising TFs and active histone modifications. Genomic samples of all individuals (functional SNP heterozygotes and homozygotes) were analysed to confirm heterozygosity at the tracking SNP and to determine specificity of the assay (genomic, Fig. 4C). When assaying the ChIP material, individuals homozygous for the functional SNP had equal enrichment from both alleles (Fig. 4C, black dots). However, individuals heterozygous for the functional SNP clearly showed skewing towards enrichment for only one allele; in all cases, enrichment came from the allele of the tracking SNP which was linked to the G variant of the functional SNP (Fig. 4C, grey dots). This provides clear evidence that the C variant of the functional SNP (rs2071914) does indeed abrogate binding of these TFs *in vivo*. This results, in erythroid cells, in both the loss of *eNME4* transcription, in addition to the loss of the upregulated expression of full length *NME4*, highlighting the profound effect that SNPs which perturb Scl/TAL1 recruitment to chromatin can have on tissue-specific gene expression (Fig. 5).

Discussion

Using erythropoiesis as a cellular model, we have shown in detail how common sequence variation (SNPs and CNVs) can affect the binding of TF complexes at regulatory elements to

produce tissue-specific and allele-specific changes in gene expression with no significant change in phenotype. Detailed analysis of erythroid elements that are commonly bound by the TFs GATA, Scl/TAL1 and SpXKLF revealed the principles by which formation of multiprotein complexes at such elements may be perturbed by common SNPs. These common SNPs act in *cis* and directly affect, or are in close proximity to, motifs required for the recruitment of Scl/TAL1 (GATA site and E-box) or other associated TFs (GC-box). Nevertheless, in three regions identified in our genome-wide study, no DNA polymorphisms were found in close proximity (<100 bp) to GATA sites, suggesting that other factors might account for such binding differences. One possibility is that activation of such a regulatory element may depend on activation of another distal linked *cis*-acting element. This may occur when multiple, widely separated elements work together, providing different components of a multiprotein complex within the context of a transcription factory. Clearly, a regulatory SNP in one element could affect the recruitment of factors to another element, even though the sequence of the second element is unchanged. In this respect, it was of interest that 2 regions bound by Scl/TAL1 in C1 and not in C2 were only 25kb apart from each other, however the significance of this is not clear. Another possibility is that one allele has acquired an epigenetic change even though the sequence remains unaltered, as occurs at imprinted or X-inactivated loci (Jeon et al., 2012; Koerner and Barlow, 2010).

Searches for SNPs that affect binding of specific TFs naturally focus on the known consensus binding site for that factor. However, detailed analysis of the Scl/TAL1 polymorphic site at *NME4* demonstrates why this approach may be misleading. To examine the potential functional effect of SNPs lying beyond the predicted canonical binding site, we undertook detailed analysis of a locus-specific (*NME4*) regulatory SNP, which revealed how

such polymorphisms can have an effect on gene expression. The erythroid specific pentameric complex (containing Scl/TAL1) would be predicted to bind to the canonical GATA sequence (WGATAR) situated at the TSS of *eNME4*. However, an upstream polymorphism within a GC-box is clearly abrogating the binding of not only the GC-box binding TFs (SpX and xKLF) but also the erythroid specific TFs (Scl/TAL1 and GATA1) which should recognize and bind to the downstream GATA site. This upstream polymorphism clearly results in the loss of active chromatin histone modifications, PolII recruitment and transcription. This shows that the recognition of a TF binding site is clearly context dependent, and that co-operativity with proteins binding at elements in the surrounding sequence may play a crucial role for the recruitment of a specific TF. In light of the recent ENCODE data, which found that co-binding of one group of transcription factors is often affected by the binding of another transcription factor (Gerstein et al., 2012) it is likely that the situation we describe at *NME4* is a common occurrence in the genome, and functional studies such as these will be required to determine the true functional effect of this variation.

While sequence variation associated with monogenic diseases most frequently lie within the gene itself, sequence variation associated with common, complex disorders is more likely to occur in the non-coding regulatory regions of the genome. A key question to ask in this field is; what are the principles by which such variation leads to observed changes in tissue-specific, and often allele-specific, expression? The functional SNP within the GC-box at *NME4*, despite affecting the binding site of an ubiquitous TF, has no effect on gene expression in non-erythroid cells, implying that SNPs within *cis*-regulatory elements, in some cases, might be functional only in the presence of tissue-specific TFs. This finding has

significance beyond *NME4*, as analysis of the ENCODE data found that one third of all DNaseI hypersensitive sites (taken as an indication of TF binding) were found in only one cell type, revealing the extent to which regulation of the genome, and hence the effect of these SNPs, varies between tissues (Neph et al., 2012).

Interestingly, despite a very large change in gene expression, we find that no apparent physiological effect results from the presence of this SNP in erythroid cells. In addition, whilst we were not able to obtain samples from individuals homozygous for the non-functional version of the SNP, such individuals do exist (as annotated in dbSNP), and to the best of our knowledge have no associated pathology. Nevertheless, this is exactly the type of variation which may be found in association with predisposition to particular common diseases. Given the exponential growth in genome-wide association studies in which numerous SNPs are being implicated in the aetiology of common complex disorders, this study elucidates molecular mechanisms that might account for phenotypic diversity and highlights the importance of carrying out functional characterisation of non-coding polymorphisms associated with disease risk in clinically relevant tissues.

Acknowledgements

We thank Dr. C. Porcher for the kind gift of the Scl/TAL1 antibody; Prof. M. Crossley for the kind gift of the EKLF and BKLF antibodies; and the Computational Biology Research Group (CBRG), Oxford University, and Dr. E. Marchi for bioinformatic support. This work was supported by the Medical Research Council and the NIHR Biomedical Research Centre

Programme. K.M.L. was supported by an Oxford Nuffield Medical Fellowship, Oxford University.

References

Benko S, Fantes JA, Amiel J, Kleinjan DJ, Thomas S, Ramsay J, Jamshidi N, Essafi A, Heaney S, Gordon CT, McBride D, Golzio C et al. 2009. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat Genet* 41:359-364

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57-74.

Crossley M, Whitelaw E, Perkins A, Williams G, Fujiwara Y, Orkin SH. 1996. Isolation and characterization of the cDNA encoding BKLf/TEF-2, a major CACCC-box-binding protein in erythroid cells and selected other cells. *Mol Cell Biol* 16:1695-1705.

De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, de Jong P, Cheng JF, Rubin EM et al. 2006. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 312:1215-1217.

De Gobbi M, Anguita E, Hughes J, Sloane-Stanley JA, Sharpe JA, Koch CM, Dunham I, Gibbons RJ, Wood WG, Higgs DR. 2007. Tissue-specific histone modification and transcription factor binding in alpha globin gene expression. *Blood* 110:4503-4510.

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, Gagnebin M, Nisbett J et al. 2009. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325:1246-1250.

Fujiwara T, O'Geen H, Keles S, Blahnik K, Linnemann AK, Kang YA, Choi K, Farnham PJ, Bresnick EH. 2009. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell* 36:667-681.

Garrick D, De Gobbi M, Lynch M, Higgs DR. 2008. Switching genes on and off in haemopoiesis. *Biochem Soc Trans* 36:613-618.

Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, Le J, Koka V, Lam KC, Gagne V, Dias J, Hoberman R et al. 2009. Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat Genet* 41:1216-1222.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P et al. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* 489:91-100.

Hattangadi SM, Wong P, Zhang L, Flygare J, Lodish HF. 2011. From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood* 118:6258-6268.

Jeon Y, Sarma K, Lee JT. 2012. New and Xisting regulatory mechanisms of X chromosome inactivation. *Curr Opin Genet Dev* 22:62-71.

Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, Hong MY, Karczewski KJ et al. 2010. Variation in transcription factor binding among humans. *Science* 328:232-235.

Koerner MV, Barlow DP. 2010. Genomic imprinting-an epigenetic generegulatory model. *Curr Opin Genet Dev* 20:164-170.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memoryefficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25.

Lower, K, Hughes J, De Gobbi M, Henderson S, Viprakasit S, Fisher C, Goriely A, Ayyubb H, Sloane-Stanley J, Vernimmen D, Langford C, Garrick D et al. 2009. Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proc Natl Acad Sci USA* 106:21771-21776.

Machanick P, Bailey TL. 2011. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 27:1696-1697.

McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, Keefe D, Collins FS et al. 2010. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* 328:235-239.

Milon L, Rousseau-Merck MF, Munier A, Erent M, Lascu I, Capeau J, Lacombe ML. 1997. nm23-H4, a new member of the family of human nm23/nucleoside diphosphate kinase genes localised on chromosome 16p13. *Hum Genet* 99:550-557.

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman, RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489:83-90.

Palii CG, Perez-Iratxeta C, Yao Z, Cao Y, Dai F, Davison J, Atkins H, Allan D, Dilworth FJ, Gentleman R, Tapscott SJ, Brand M et al. 2011. Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. *Embo J* 30:494-509.

Pope SH, Fibach E, Sun J, Chin K, Rodgers GP. 2000. Two-phase liquid culture system models normal human adult erythropoiesis at the molecular level. *Eur J Haematol* 64:292-303.

Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, Li Z, Lee JH, Aach J, Leproust EM, Eggan K, Church GM et al. 2009. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods* 6:613-618.

Figure Legends

Figure 1. Detection of Scl/TAL1 binding sites in primary human erythroblasts.

Four examples of Scl/TAL1 ChIP-Seq peaks are shown. Shown for each locus are (from top to bottom): hg 18 genomic coordinate; RefSeq gene annotation; ChIP-Seq profile in C1 and C2. The sequencing reads are visualized as peaks and mapped along the specific sequence in GBrowse. Within each track, the peak height reflects the number of mapped reads.

Figure 2. Differential Scl/TAL1 binding sites in C1 and C2.

A. Correlation plot showing the number of reads per peak of all Scl/TAL1 bound regions identified in C1 and C2. Linear regression and R^2 correlation index are shown. In the right panel, which focuses on the data within the dashed lines, the differential bound regions, identified by non-parametric rank product statistical analysis, are highlighted in red ($C1 > C2$) and in purple ($C2 > C1$)

B-C. Examples of Scl/TAL1 differential peaks. The loci are annotated as in Fig. 1A. Differential peaks are indicated with an arrow.

Figure 3. Variable expression of *NME4* in normal individuals.

A. Schematic representation of the genomic structure of *NME4* and *eNME4*. Black boxes, exons; gray box, alternative erythroid-specific exon; full lines, introns; dashed lines, splicing of mature transcript; light gray lines, expanded region. The DNA sequence surrounding the transcription start of *eNME4* is shown. The predicted GC-box is in blue, and contains the functional G/C SNP (rs2071914:G>C, indicated by a vertical arrow). The GATA site underlying the *eNME4* transcription start site (TSS) is shown in red. The tracking A/G SNP

(rs14293:G>A, indicated by a vertical arrow) is also shown. The horizontal arrow represents the TSS of the erythroid-specific transcript. The red box represents the position of amplicon 389776 (referred to in Fig. 4).

B. Expression of *NME4* in erythroid and non-erythroid cells. Expression is normalised to *CD71* (erythroid) and to 18S (non-erythroid), and expressed relative to the mean expression of group A (see text). Error bars are +/- 1 SD. Group A, n=4, black bars. Group B, n=3, green bars

C. Allele-specific expression of *NME4*, as determined by pyrosequencing (see Materials and Methods for details) in non-erythroid (EBV) and erythroid cDNA. All samples are heterozygous (A/G) for SNP rs14293, except for A/A and G/G genomic DNA controls (indicated by *). Individuals in Group A (n=4; black dots) were found to be homozygous (G/G) and individuals in Group B (n=3; green dots) were found to be heterozygous (G/C) for SNP rs2071914.

Figure 4. The C variant of SNP2071914 abolishes binding of TFs *in vitro* and *in vivo*.

A. Electrophoretic mobility shift / super-shift assay. The 28bp double stranded oligonucleotides containing the native surrounding *NME4* sequence, differ only at the SNP rs2071914 (Table S2). Protein-DNA complexes were observed with the G allele upon addition of nuclear extract (lane 2), and specific complexes could be super-shifted with the addition of anti-Sp1 (lane 3), anti-Sp3 (lane 4) and anti-BKLF (lane 6) antibodies as indicated. There is no super-shift upon addition of anti-EKLF antibody as EKLF is not expressed in K562 cells (lane 5). Protein binding was completely abolished on the C allele oligonucleotide (lane 8). Addition of K562 nuclear extract is indicated by +.

B. Transcription factor binding, chromatin modifications and PolII at *eNME4*. Real-time analysis of ChIP enrichment at HS-40, α -globin promoters, *eNME4* and a negative control *DECR2*. Enrichment is relative to input and normalized to amplicons 162909 (H3ac, H3K4me3 and PolII) or 103432 (Scl/TAL1 and GATA1). Values represent the mean of 3 (G/G) or 4 (G/C) independent experiments +/- 1SD.

C. Proportion of the G allele (rs2071914) of *NME4* contributing to total ChIP enrichment, determined by pyrosequencing (see Materials and Methods for details). All samples are heterozygous (A/G) for the tracking SNP rs14293, except for an A/A and a G/G genomic control (indicated by *). Samples are grouped according to homozygosity (G/G, black dots; n=3) or heterozygosity (G/C, grey dots; n=4) for the functional SNP rs2071914. Note that the A allele of the tracking SNP was always found in phase with the C allele of the functional SNP (see text).

Figure 5. Schematic representation of the tissue-specific regulation of expression of *eNME4*.

A. When a G is present at SNP2071914, the functional GC box is able bind the xKLF/SP1/3 factors, the downstream GATA1 site is able to bind the erythroid specific transcription factors Scl/TAL1, the associated chromatin acquires activating modifications (such as H3K4me3 and H3ac), and *eNME4* is transcribed.

B. When a C is present at SNP2071914, the GC box is no longer functional, which in turn abrogates the functionality of the downstream GATA1 site. The erythroid-specific transcription factors do not bind, the activating chromatin modifications are lost, and *eNME4* transcription does not occur.

SNP2071914 is underlined; the functional GC box is shown in italics; the GATA1 site is shown in bold; the box-like shapes represent histone H3.

Supp. Figure S1. Experimental validation of Scl/TAL1 ChIP.

A. Outline of experimental design. See methods for details.

B. Venn diagram showing the overlap between the Scl/TAL1 binding sites detected by ChIP-Seq in this study and those identified in another Scl/TAL1 genome-wide binding analysis (Palii et al., 2011).

Supp. Figure S2. Characterization of erythroid cultures from C1 and C2.

A. Flow cytometry profile (CD71, x-axis, and GPA, y-axis) of C1 and C2 erythroblast cultures shows homogeneity of the maturation stage across the two cultures. The whole-cell culture was used for ChIP assays.

B. ChIP-qPCR validation at selected regions at the α and β globin loci. The y-axis represents the enrichment over input DNA, normalized to the HS-40 positive control sequence.

Supp. Figure S3. DNA motifs underlying Scl/TAL1 occupancy. For each Scl/TAL1 bound region, we extracted the peak sequence and used it for searching *de novo* motifs with the MEME/DREME software (<http://meme.ebi.edu.au/meme>). MEME is able to find relatively long composite motifs, whereas DREME is a motif discovery algorithm specifically designed to find short DNA-binding motifs. Logos identified by MEME/DREME motif search are shown together with the recognizing TF, as predicted by JASPAR CORE 2009, the corresponding p-values and the number of the Scl/TAL1 peaks harbouring each motif.

Supp. Figure S4. Examples of differential Scl/TAL1 binding.

A-D. The loci are annotated as in Fig. 1A. Differential peaks are indicated by a vertical arrow.

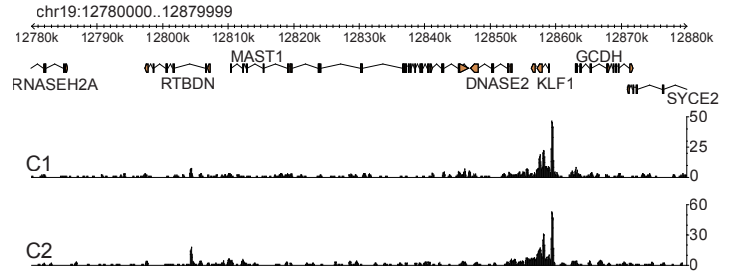
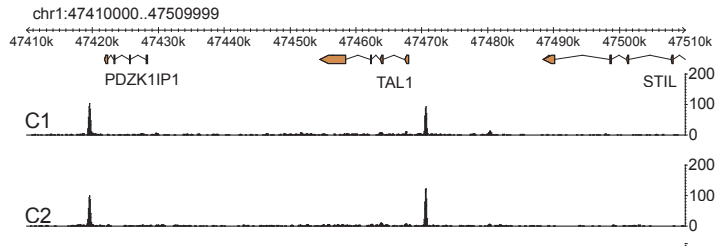
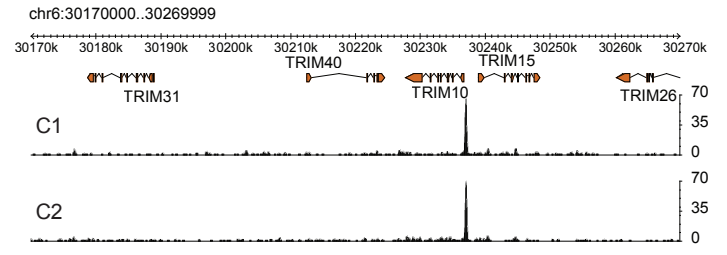
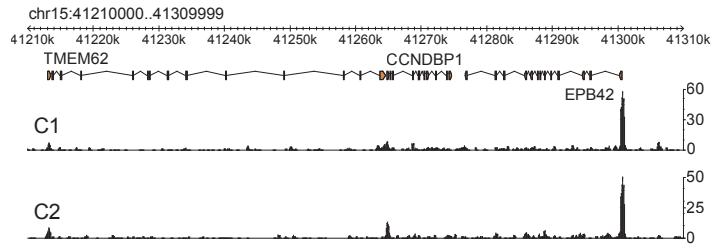
E. qPCR validation of a subset of Scl/TAL1 differentially bound regions. For each region, the ratio between the fold enrichment, as measured by comparison with HS-40 at α -globin locus, observed in C1 and in C2 is plotted on a log₂ scale. Del = region deleted in C1, C1/C2 ratio not calculated. Primer sequences can be found in Table S1.

Supp. Figure S5. Examples of differential Scl/TAL1 binding patterns in the two sets of replicates (C1>C2).

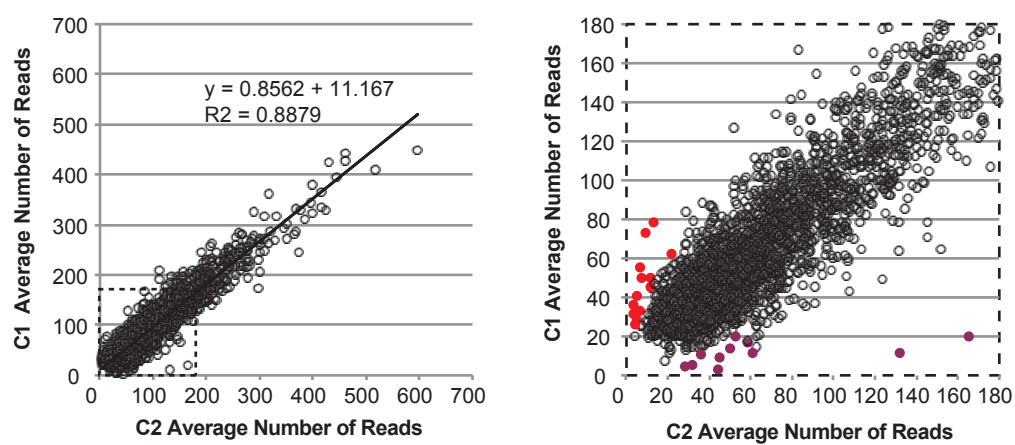
The locus is annotated as in Fig. 1A. Differential peaks are indicated by a vertical arrow.

Supp. Figure S6. Examples of differential Scl/TAL1 binding patterns in the two sets of replicates (C2>C1).

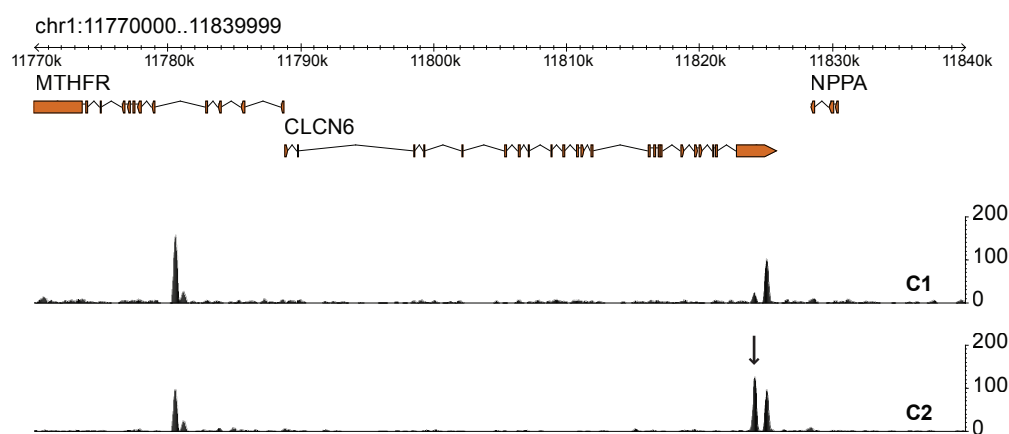
The locus is annotated as in Fig. 1A. Differential peaks are indicated by a vertical arrow.



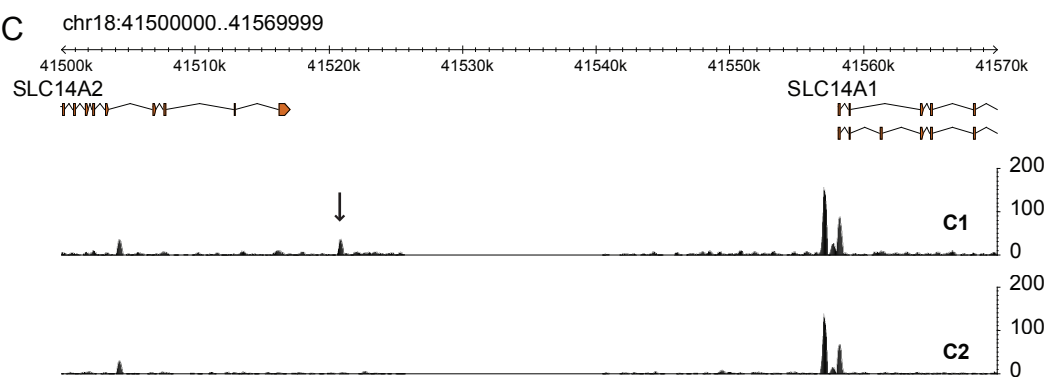
A



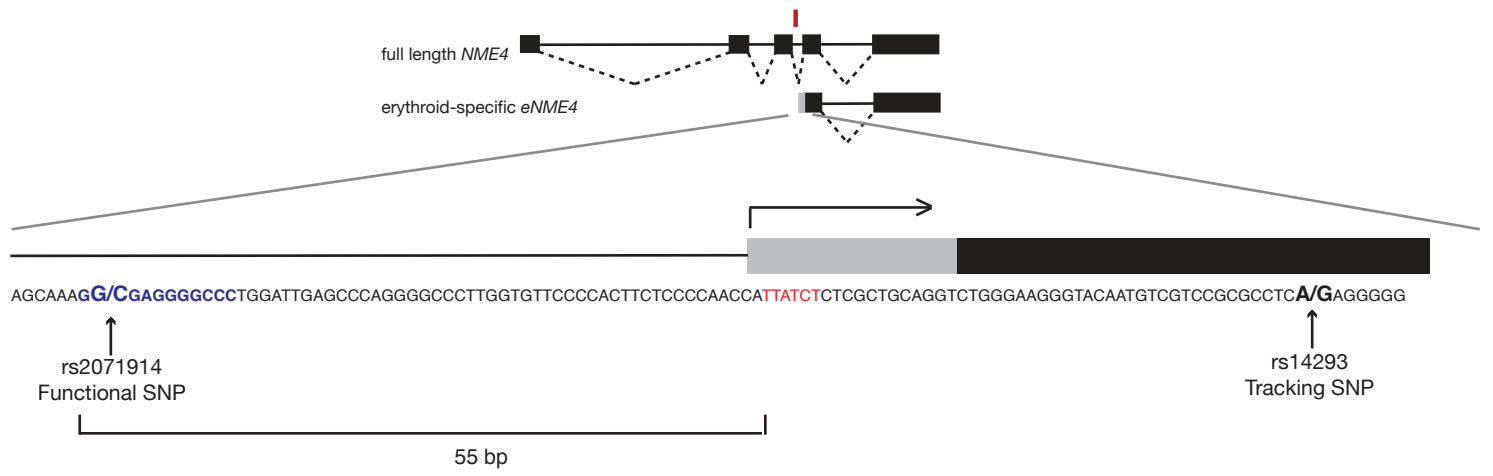
B



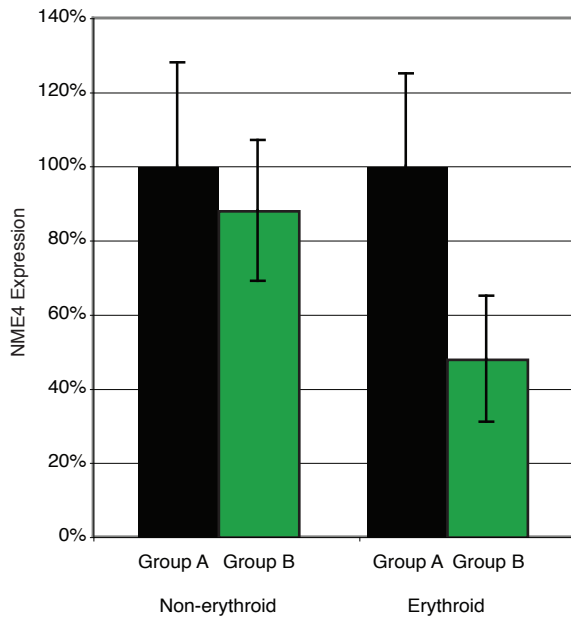
C



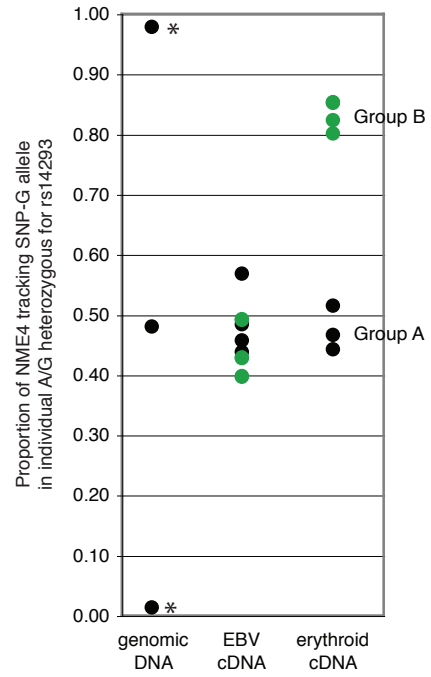
A



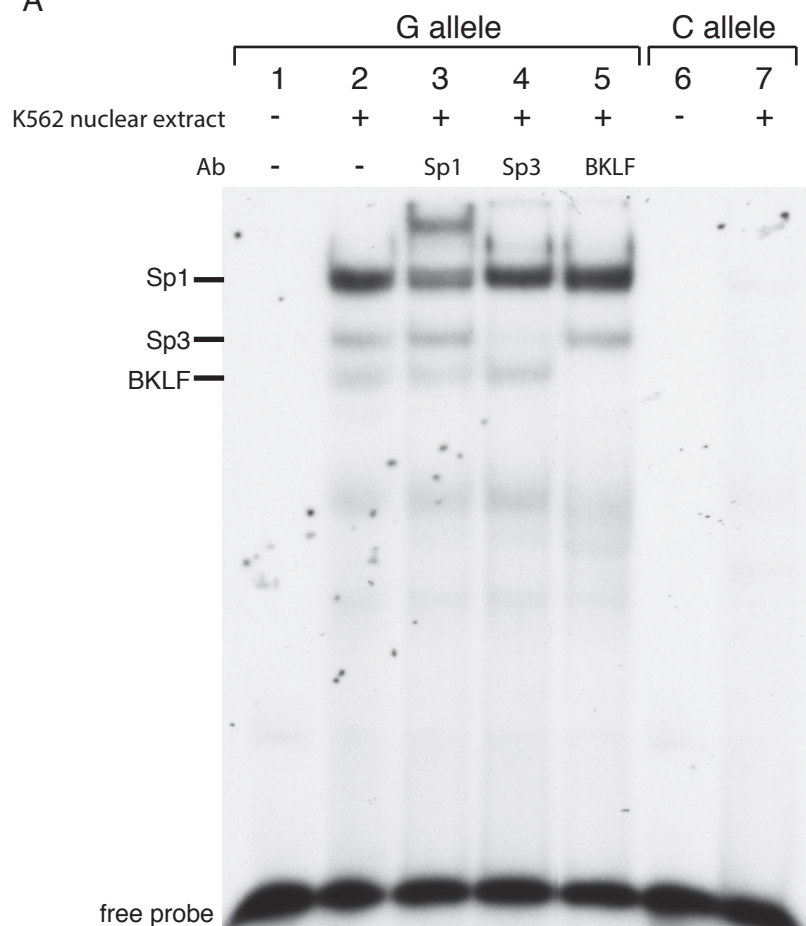
B



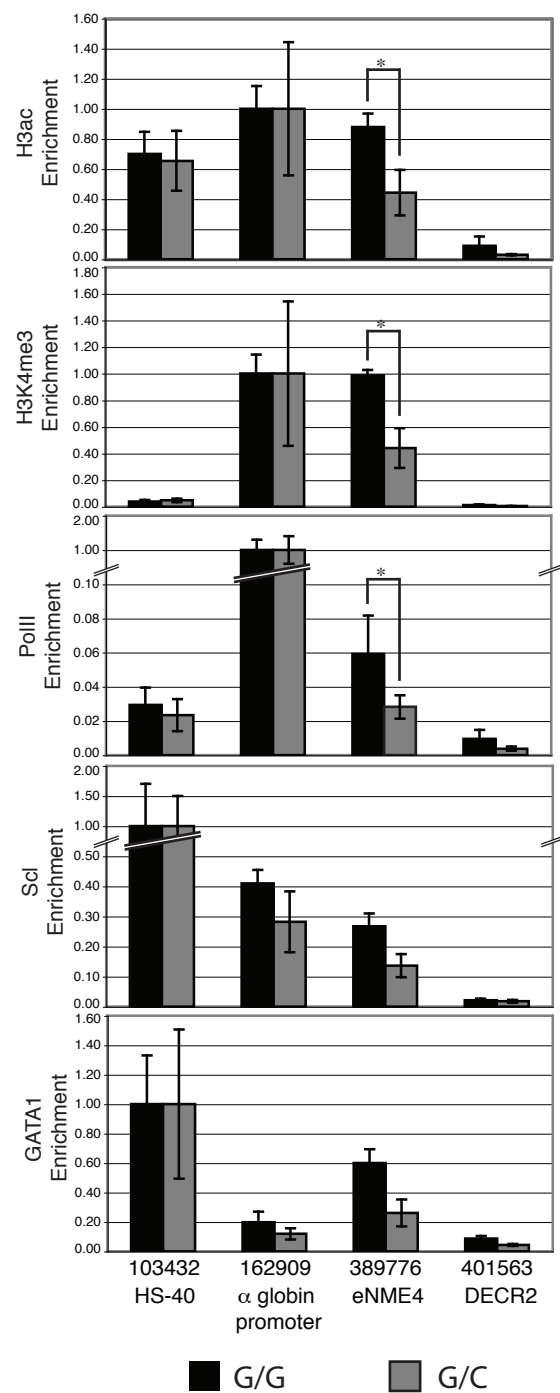
C



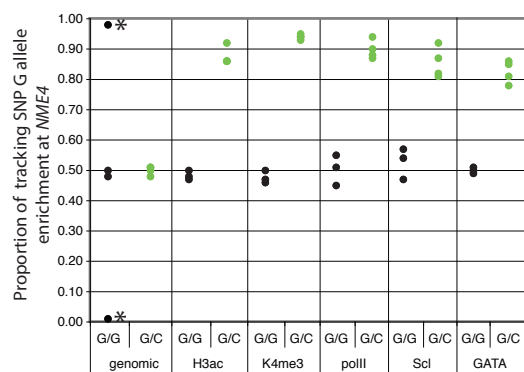
A



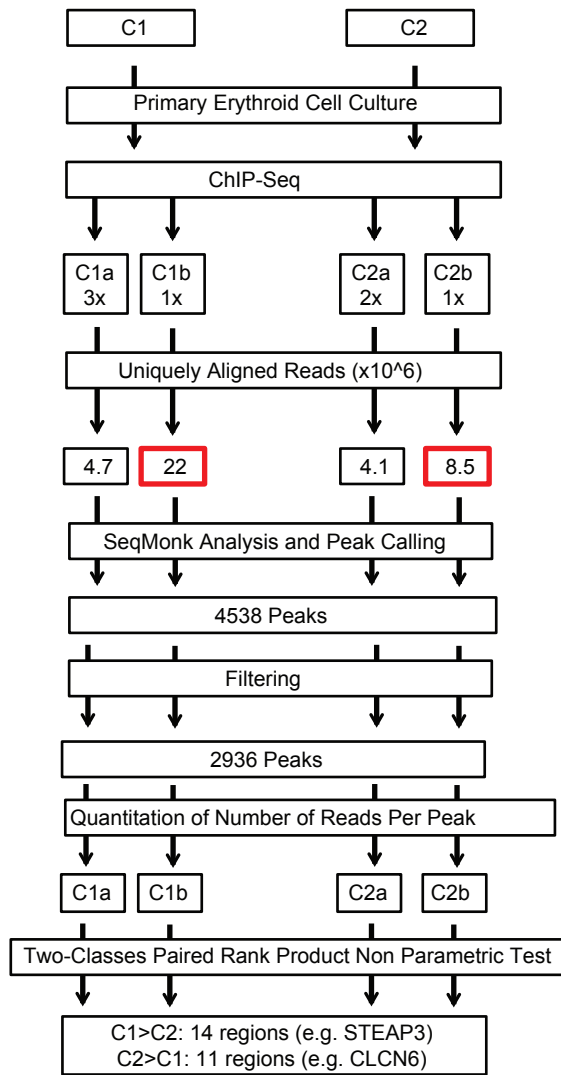
B



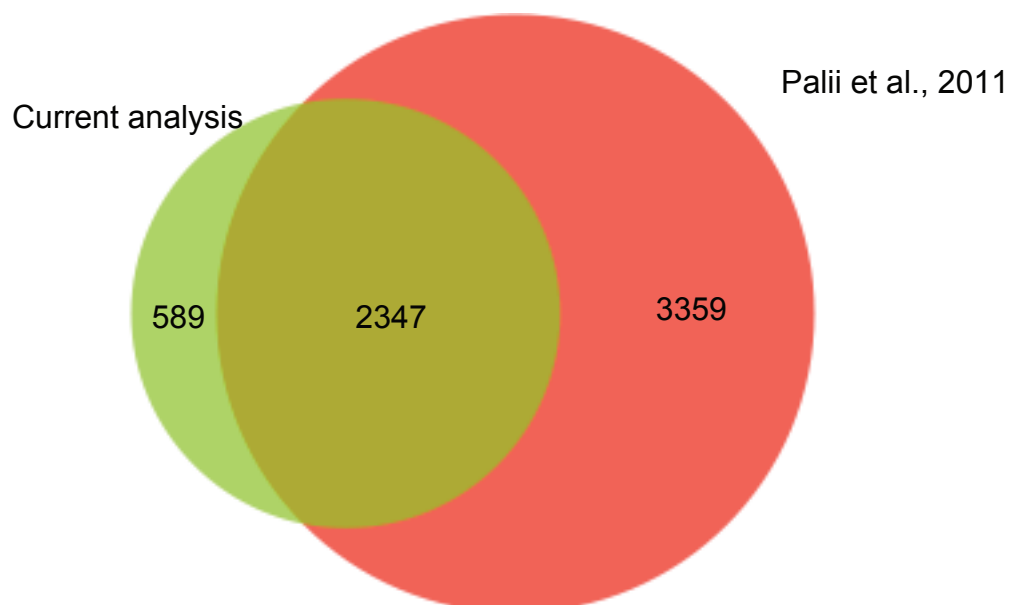
C

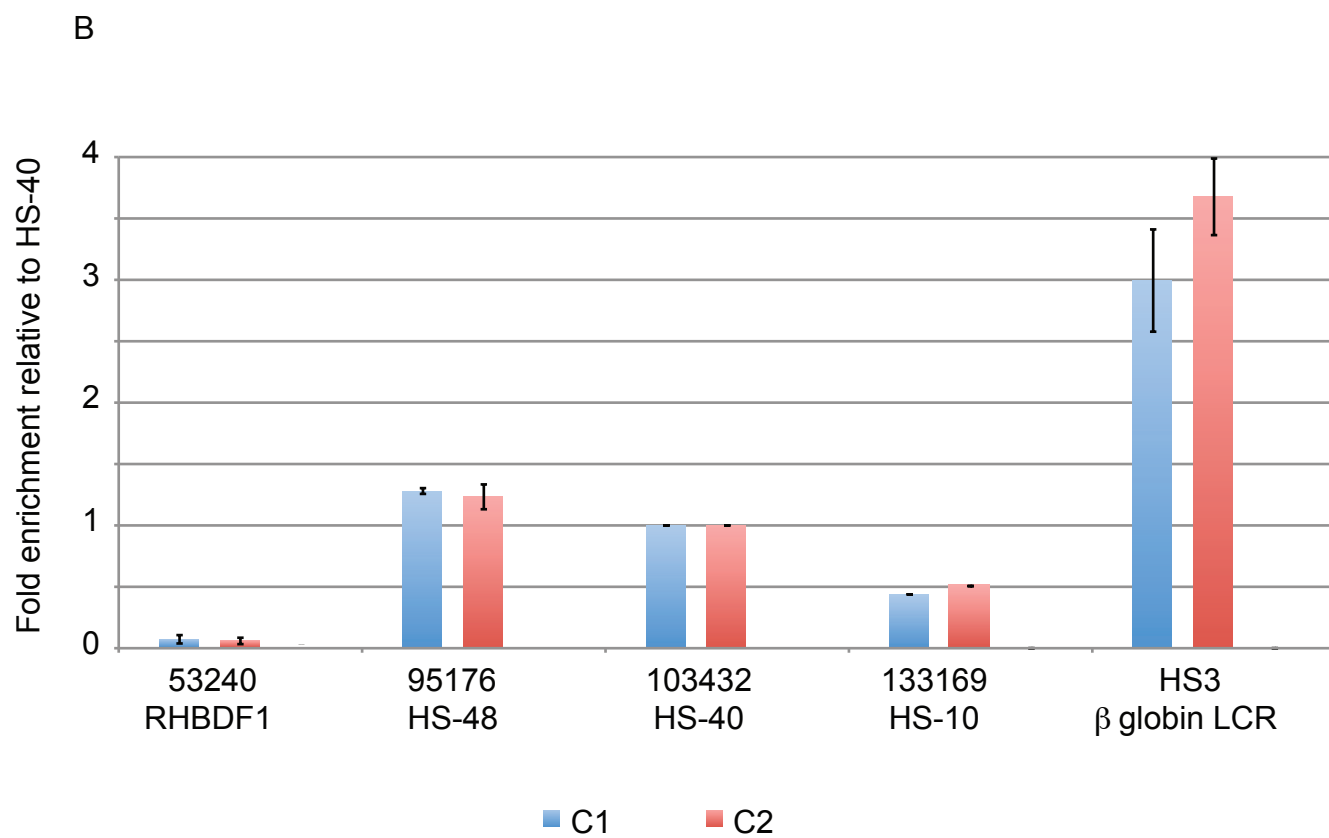
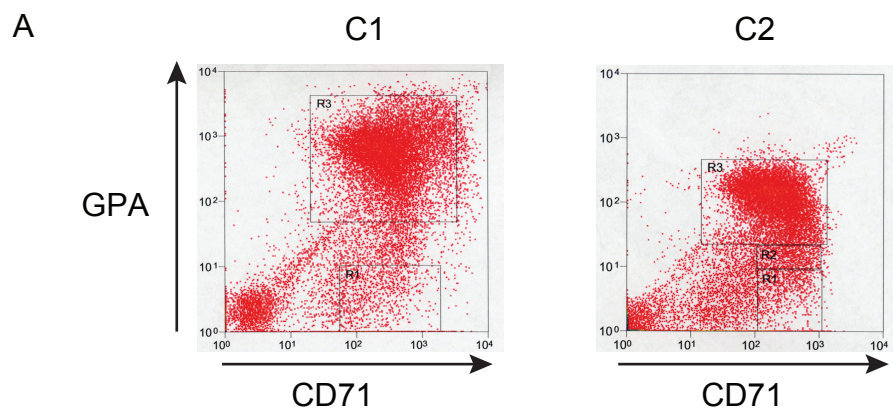


A

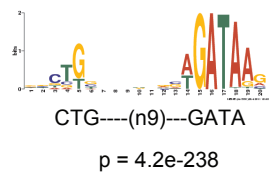


B

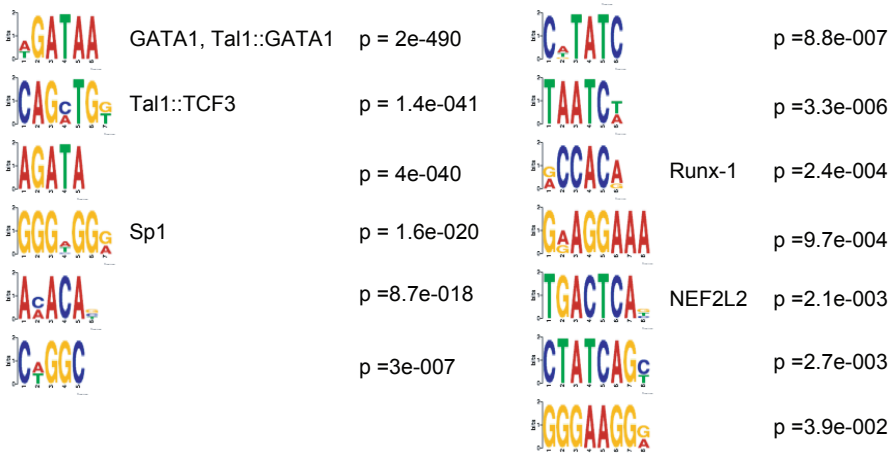


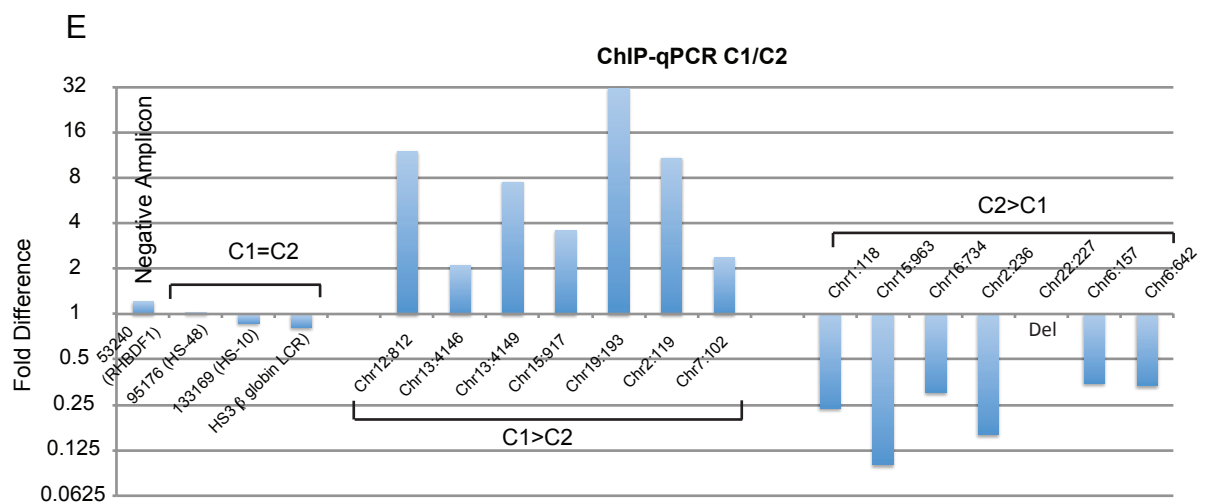
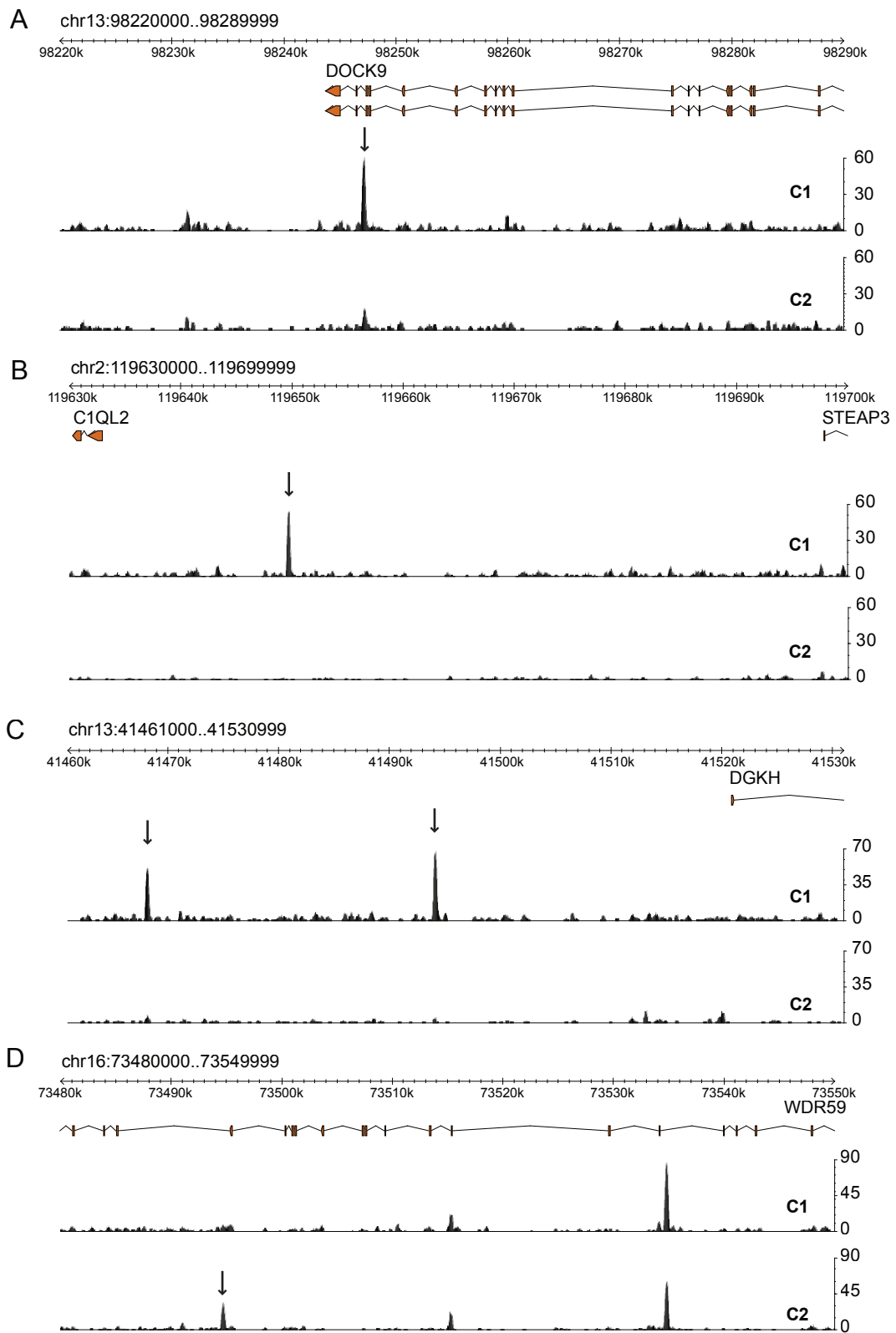


MEME Output



DREME Output





chr7:8110000..8249999

