High-throughput identification of human SNPs affecting regulatory element activity

van Arensbergen, Joris; Pagie, Ludo; FitzPatrick, Vincent D.; de Haas, Marcel; Baltissen, Marijke P.; Comoglio, Federico; van der Weide, Robin H.; Teunissen, Hans; Vosa, Urmo; Franke, Lude

Published in:
Nature Genetics

DOI:
10.1038/s41588-019-0455-2

Publication date:
2019

https://doi.org/10.1038/s41588-019-0455-2

nature genetics

High-throughput identification of human SNPs affecting regulatory element activity

Joris van Arensbergen[1]*, Ludo Pagie[1], Vincent D. FitzPatrick[2,3], Marcel de Haas[1], Marijke P. Baltissen[4], Federico Comoglio[1,5], Robin H. van der Weide[1], Hans Teunissen[1], Urmo Võsa[6,7], Lude Franke[6], Elzo de Wit[1], Michiel Vermeulen[4], Harmen J. Bussemaker[2,3] and Bas van Steensel[1]*

Most of the millions of SNPs in the human genome are non-coding, and many overlap with putative regulatory elements. Genome-wide association studies (GWAS) have linked many of these SNPs to human traits or to gene expression levels, but rarely with sufficient resolution to identify the causal SNPs. Functional screens based on reporter assays have previously been of insufficient throughput to test the vast space of SNPs for possible effects on regulatory element activity. Here we leveraged the throughput and resolution of the survey of regulatory elements (SuRE) reporter technology to survey the effect of 5.9 million SNPs, including 57% of the known common SNPs, on enhancer and promoter activity. We identified more than 30,000 SNPs that alter the activity of putative regulatory elements, partially in a cell-type-specific manner. Integration of this dataset with GWAS results may help to pinpoint SNPs that underlie human traits.

About 85 million SNPs have been identified in human genomes[1]. The vast majority of these are located in non-coding regions, and a typical human genome has about 500,000 variants with non-reference alleles overlapping regulatory elements such as enhancers and promoters[1]. It is becoming increasingly clear that such non-coding SNPs can have a substantial impact on gene regulation[2], thereby contributing to phenotypic diversity and a wide range of human disorders[3–5].

GWAS and expression quantitative trait locus (eQTL) mapping can identify candidate SNPs that may drive a particular trait or disorder[6,7] or the expression level of individual genes[3,8], respectively. Unfortunately, even the largest GWAS and eQTL studies rarely achieve single-SNP resolution, largely due to linkage disequilibrium (LD). In practice, tens to hundreds of linked SNPs are correlated with a trait. Although new fine-mapping techniques[9–11], integration with epigenomic data[12], deep learning computational techniques[13] and GWAS of extremely large populations can help to achieve higher resolution, pinpointing of the causal SNPs remains a major challenge.

Having a list of all SNPs in the human genome that have the potential to alter gene regulation would mitigate this problem. Ideally, the regulatory impact of SNPs would be measured directly. Two high-throughput methods have been employed for this purpose. First, changes in chromatin features, such as DNase sensitivity and various histone modifications, have been mapped in lymphoblasts or primary blood cells derived from sets of human individuals with fully sequenced genomes[14–20]. Here, the chromatin marks serve as proxies to infer effects on regulatory elements, with the caveat that a change in regulatory activity may not always be detected as a change in chromatin state, or vice versa. Furthermore, many traits do not manifest in blood cells, and other cell types are more difficult to obtain for epigenome mapping.

An alternative functional readout is to insert DNA sequence elements carrying each allele into a reporter plasmid. On transfection of these plasmids into cells, the promoter or enhancer activity of these elements can be measured quantitatively. Different cell types may be used as models for corresponding tissues in vivo. Large-scale versions of this approach are referred to as massively parallel reporter assays (MPRAs), which have been applied to screen tens of thousands of SNPs[21–25]. Each of these studies has yielded tens to, at most, several hundreds of SNPs that significantly alter promoter or enhancer activity. As these MPRA studies have covered only a tiny fraction of the genome, it is likely that many more SNPs with regulatory impact are to be discovered.

Here, we report application of an MPRA strategy with a >100-fold increased scale compared to previous efforts. This enabled us to survey the regulatory effects of 5.9 million SNPs in two different cell types, providing a resource that helps to identify causal SNPs among candidates generated by GWAS and eQTL studies. The data are available for download, and can be queried through a web application (https://sure.nki.nl).

Results

A survey of 5.9 million SNPs using SuRE. We applied our SuRE technology to systematically screen millions of human SNPs for potential effects on regulatory activity. SuRE is an MPRA with sufficient throughput to query entire human genomes at high resolution and high coverage[26]. Briefly, random genomic DNA fragments of a few hundred base pairs (bp) are cloned into a promoter-less reporter plasmid that, on transfection into cultured cells, only produces a transcript if the inserted genomic DNA fragment carries a functional transcription start site (TSS) (Fig. 1a). The transcript is identified and quantified by means of a random barcode sequence

[1]Division of Gene Regulation, Oncode Institute, Netherlands Cancer Institute, Amsterdam, the Netherlands. [2]Department of Biological Sciences, Columbia University, New York, NY, USA. [3]Department of Systems Biology, Columbia University Medical Center, New York, NY, USA. [4]Department of Molecular Biology, Oncode Institute, Radboud Institute for Molecular Life Sciences, Radboud University Nijmegen, Nijmegen, the Netherlands. [5]Department of Haematology, University of Cambridge, Cambridge, UK. [6]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. [7]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. *e-mail: joris.van.arensbergen@gmail.com; b.v.steensel@nki.nl

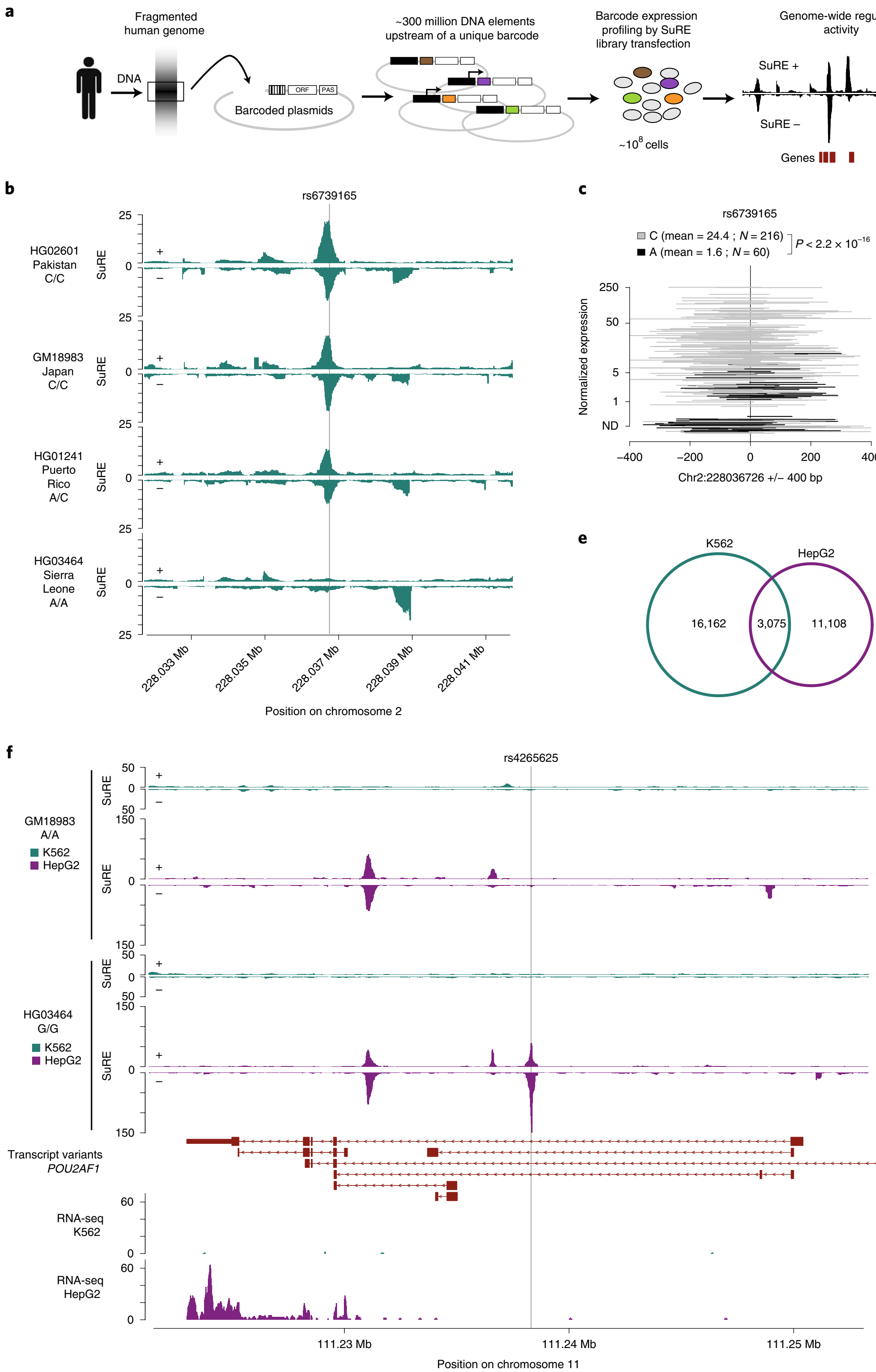


Fig. 1 | Identification of raQTLs by SuRE. a, Schematic representation of the SuRE experimental strategy. ORF, open reading frame; PAS, polyadenylation signal. Colors indicate different barcodes. SuRE yields orientation-specific activity information[26] (SuRE +/− tracks, right-hand panel). **b**, SuRE signals from the four genomes in an example locus, showing differential SuRE activity at raQTL rs6739165, depending on the allele (A or C) present. Mb, megabase. **c**, SuRE activity for all fragments containing rs6739165. ND, not detected. SuRE data of + and − orientations are combined. Values on the *y* axis were shifted by a random value between −0.5 and 0.5 to better visualize DNA fragments with the same value. $P < 2.2 \times 10^{-16}$, according to two-sided Wilcoxon rank-sum test. **d**, Same data as in **c**, but only the expression value for each fragment is shown (without the addition of the random value). Red lines indicate mean values. **e**, Numbers of raQTLs in K562, HepG2 or both. **f**, Example of a locus showing differential SuRE activity for two genomes in HepG2 only. Below the SuRE tracks known transcript variants of *POU2AF1* are indicated, and RNA-seq data from K562 and HepG2 (data from ref. [28]).

that is unique for every insert, allowing for a multiplexed readout of millions of random DNA fragments. Importantly, because active promoters as well as enhancers generate transcripts, activity of both types of elements can be assayed quantitatively by SuRE[26].

To survey a large cross-section of SNPs present in the human population, we chose four sequenced genomes of individuals from four different populations[1] (Fig. 1b). From each genome we generated two independent SuRE libraries that each contained ~300 million random genomic DNA fragments of 150–500 bp (Supplementary Fig. 1a and Supplementary Table 1). In these libraries a total of 2,390,729,347 unique genomic DNA fragments were sequenced from both ends, mapped to the reference genome and linked to their unique barcode. Among these fragments, 1,103,381,066 carried at least one SNP for which we identified both alleles in our libraries. These libraries enabled us to test promoter/enhancer activity of both alleles of 5,919,293 SNPs, which include 4,569,323 (57%) of the ~8 million known common SNPs (minor allele frequency (MAF) > 5%) worldwide[1]. Importantly, each SNP allele is covered by 122 different genomic DNA fragments on average (Supplementary Fig. 1b,c), which provides substantial statistical power.

We introduced these libraries by transient transfection into human K562 and HepG2 cell lines. K562 is an erythroleukemia cell line with strong similarities to erythroid progenitor cells[23]. HepG2 cells are derived from a hepatocellular carcinoma, and serve as an approximate representation of liver cells. After transfection of the SuRE libraries into each cell line we isolated messenger RNA and counted the transcribed barcodes by Illumina sequencing. Three independent biological replicates yielded a total of 2,377,150,709 expressed barcode reads from K562 cells, and two biological replicates yielded 1,174,138,611 expressed barcode reads from HepG2 cells.

Identification of thousands of SNPs with regulatory impact. From these data we first constructed strand-specific tracks of SuRE enrichment profiles for each of the four genomes (Fig. 1b). This revealed thousands of peaks that generally colocalize with known enhancers and promoters (Supplementary Fig. 1d), as reported previously[26]. For a subset of peaks, the magnitude varied between the four genomes and showed a correlation with a particular allele of a coinciding SNP. For example, in K562 cells we detected a strong SuRE signal overlapping with SNP rs6739165 in the genomes that are homozygous for the C allele, but no signal in the genome that is homozygous for the A allele and an intermediate signal in the genome that is heterozygous for this SNP (Fig. 1b).

To systematically annotate SNPs we combined the complete SuRE datasets from the four genomes for each transfected cell line. The sequencing data of the SuRE libraries then enabled us to group, for each SNP, the overlapping genomic DNA fragments by the two alleles (Fig. 1c,d). This allowed us to identify SNPs for which fragments carrying one allele produced significantly different SuRE signals compared to those carrying the other allele. Because all of these fragments differ in their start and end coordinate, the activity of each allele is tested in a multitude of local sequence contexts, providing not only statistical power but also biological robustness. For each SNP we calculated a *P* value and we used a random permutation strategy to estimate false discovery rates (FDR) (Supplementary Fig. 1e,f). We also required that the strongest allele showed an average SuRE signal of at least fourfold over background. We refer to the resulting SNPs at FDR < 5% as reporter assay QTLs (raQTLs).

This analysis yielded a total of 19,237 raQTLs in K562 cells and 14,183 in HepG2 cells (Fig. 1e). The average allelic fold change of these SNPs was 4.0-fold (K562) and 7.8-fold (HepG2) (Supplementary Fig. 1g,h). In 72% of cases the SuRE effect could be assigned to a single SNP; when SNPs were spaced less than ~200 bp apart, their effects could typically not be resolved (Supplementary Fig. 1i).

Most raQTLs were detected in either K562 or HepG2 cells, but not in both (Fig. 1e). The overlap may be underestimated due to the arbitrary FDR and expression cutoffs used to define these sets (Supplementary Fig. 1j). Nevertheless, many SNPs show clear cell-type-specific effects (Supplementary Fig. 1j–l). For example, rs4265625[G] creates regulatory activity in HepG2 only (Fig. 1f). This is interesting, because rs4265625 lies in *POU2AF1*, a gene that has been linked to primary biliary cirrhosis—a liver disease—in a GWAS[27]. In about 1,300 instances, a K562-specific raQTL and a HepG2-specific raQTL are in strong LD ($R^2 > 0.8$). An interesting possibility is that the two raQTLs in such linked pairs may contribute to the regulation of a common target gene, but each in a cell-type-specific manner.

raQTLs are enriched for known regulatory elements. We systematically analyzed the overlap of the raQTLs with known regulatory elements in K562 cells[28]. Compared to all SNPs analyzed, raQTLs showed 5–15-fold enrichment for promoter- and enhancer-related chromatin types, and depletion for repressed or transcribed chromatin types (Fig. 2a). We also observed strong enrichment of raQTLs in DNase hypersensitive sites (DHS) (Fig. 2b,c). This is consistent with SuRE signals overlapping with enhancers or promoters, as shown previously[26].

Some of the raQTLs are heterozygous in the genome of K562 cells. For these SNPs we investigated whether the allelic imbalance observed by SuRE was reflected in a corresponding imbalance in the DHS signal. For example, the SuRE signal at rs12985827, a non-coding SNP in an intron of the *APC2* gene, has a strong bias for the C allele (the alternative allele, ALT) over the T allele (the reference allele, REF) (Fig. 2d). Indeed, it shows a similar allelic imbalance for DHS (Fig. 2e). Among the 616 raQTLs that were heterozygous in K562 and showed sufficient DNase-seq coverage, we observed a strong skew for higher DNase sensitivity at the more active allele, compared to 616 heterozygous non-raQTL control SNPs that overlapped DHSs (Fig. 2f,g). We found a similar but less pronounced trend in H3K27ac[28] and ATAC-seq[29] data from K562 cells (Supplementary Fig. 2). We conclude that available epigenomic data is generally consistent with the SuRE results.

Altered transcription factor binding sites at raQTLs. The observed changes in enhancer or promoter activity are likely explained by SNPs changing transcription factor (TF) binding motifs. For example, the T allele of rs12985827 disrupts an EGR1 binding motif and leads to reduced SuRE activity (Figs. 3a and 2d). To investigate this systematically, we made use of the computationally predicted changes in TF motifs in the SNP2TFBS database[30]. Among the set of raQTLs in K562 and HepG2 cells, 31 and 38% are predicted to alter the motif of at least one TF, respectively. This is a 1.6-fold and 1.9-fold larger proportion than for all SNPs, respectively (Fig. 3b). Moreover, 67 and 69% of the raQTLs showed concordance between the predicted effect on motif affinity and SuRE expression, that is, the allele with the weakest motif resemblance had the lowest SuRE expression (Fig. 3c). We note that 100% concordance should not be expected in this analysis, for example, because not all TF binding motifs are known, some may be misannotated and some TFs can act as repressors.

We expected motifs in raQTLs to reflect the sets of TFs that are selectively active in the respective cell types. Indeed, raQTLs in K562 were enriched for motifs of TFs that are primarily active in the erythroid blood lineage, such as GATA and STAT factors, while raQTLs in HepG2 cells were enriched for motifs of TFs that are specific for liver cells, such as HNF factors (Fig. 3d). Disruptions of the TP53 motif also appeared to be relatively more consequential in HepG2, which is presumably related to the known inactivation of *TP53* in K562 cells[31] but not in HepG2 cells[32]. Together, these data point to a general concordance between the detected changes in SuRE activity and predictions based on sequence motif analysis.

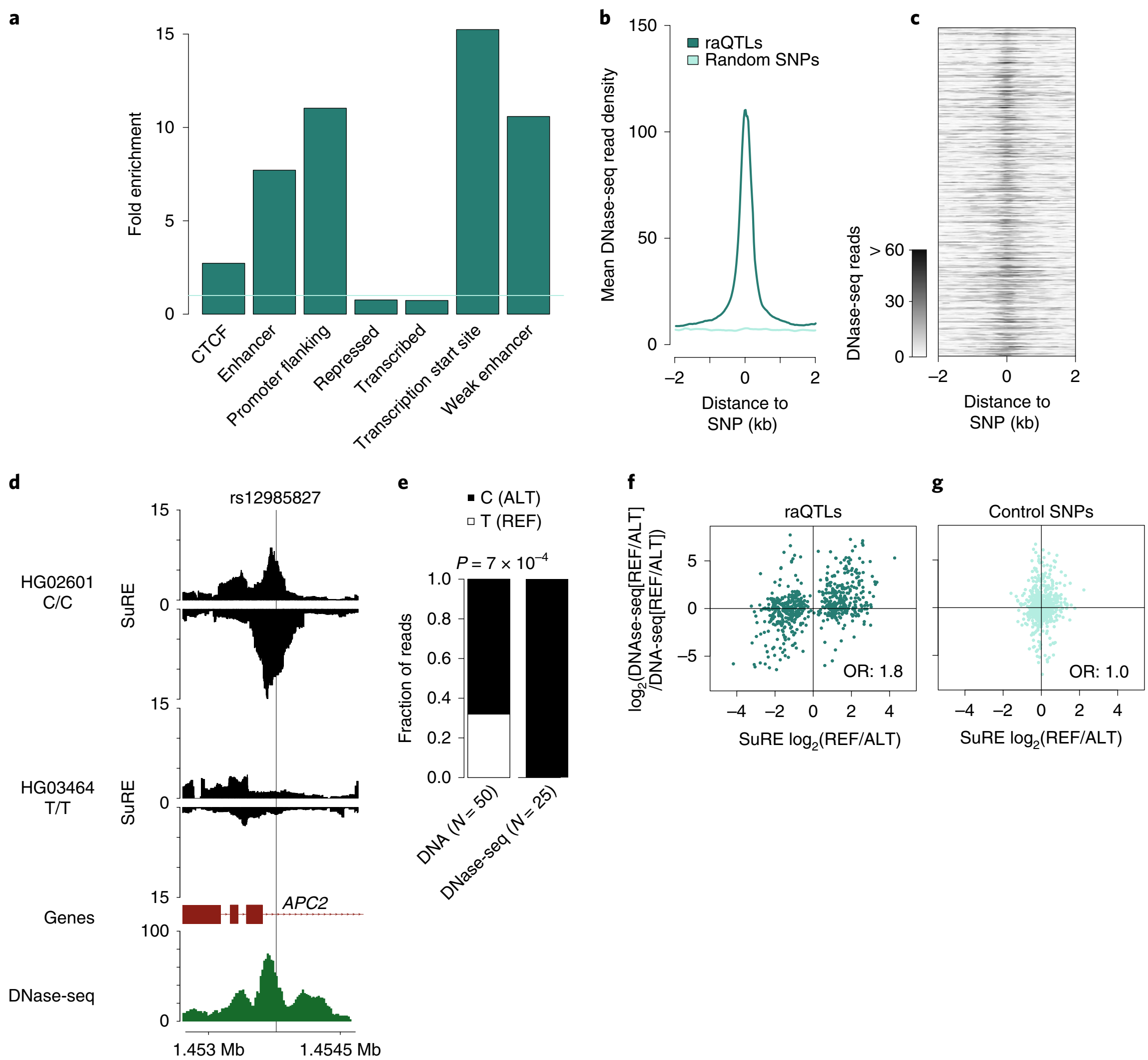


Fig. 2 | Correlation of SuRE signals with local chromatin states. **a**, Enrichment or depletion of 19,237 raQTLs among major types of chromatin in K562 (ref. [28]) relative to all SNPs analyzed. All values are significantly different from 1 ($P < 2.2 \times 10^{-16}$, two-sided Fisher's exact test). **b**, Average profile of DNase-seq enrichment for the 19,237 raQTLs compared to an equally sized random set of analyzed SNPs. **c**, DNase-seq signals aligned to the 19,237 raQTLs, sorted by their *P* value according to our SuRE analysis (lowest *P* value on top). **d**,**e**, Example of an SNP with differential SuRE activity for the two alleles, overlapping with a DNase-seq peak in K562 cells (**d**) and showing only DNase sensitivity for one allele, even though both alleles are present in the K562 genome. Note that K562 cells are aneuploid, hence the balance of the alleles in input DNA may not be 1:1 (**e**). *P* value according to two-sided Fisher's exact test. **f**, Comparison of allelic imbalance of SuRE signals and DNase-seq signals (normalized for genomic DNA allelic read counts) for 616 raQTLs for which K562 cells are heterozygous. OR, odds ratio. **g**, Same as in **f** but for a random set of 616 control SNPs overlapping with a DNase-seq peak. DNase-seq data in **b**–**g** are from ref. [28].

No evidence for strong negative selective pressure on raQTLs. It has been observed that genes that do not have *cis*-eQTLs are more likely to be loss-of-function (LOF)-intolerant genes, possibly reflecting selection against variants acting on such genes[8,33]. We found that the fraction of SNPs that are raQTLs was significantly, but only slightly, lower in the proximity of LOF-intolerant genes than in the proximity of LOF-tolerant genes (Supplementary Fig. 3a,b). However, for a set of control SNPs that were matched to the raQTLs for their SuRE activities and coverage in the SuRE libraries we observed a similar pattern (Supplementary Fig. 3a,b). This suggests that the overall density of active regulatory elements, rather than elements affected by SNPs, is lower near LOF-intolerant genes. Genome wide, we observed slightly lower MAFs for our raQTLs as compared to matched SNPs, but only for the K562 dataset and not for the HepG2 dataset (Supplementary Fig. 3c,d). This modest under-representation of raQTLs in the human population is consistent with recent computational predictions[13] and may point to a slight negative selection pressure. Taken together, we found no evidence for strong negative selective pressure on raQTLs.

Integration of SuRE with eQTL maps. Next, we integrated our SuRE data with eQTL mapping data from the GTEx Project[8]. We compared SuRE data from K562 and HepG2 cells with eQTL data from the most closely related tissues, that is, whole blood and liver, respectively. Strong similarity between the two data types should not be expected, because in the GTEx data each gene with significant associations (eGene) is linked to 101 eQTL SNPs on average, of which only one or a few may be causal. Rather, we regard the SuRE data as a filter to identify the most likely causal SNPs among the large number of eQTL candidates.

For each raQTL, the $\log_2$(ALT/REF) SuRE signal may be concordant with the eQTL signal (that is, having the same sign as the slope of the eQTL analysis) or discordant (having the opposite sign). The simplest interpretation of concordance is that the SNP alters a

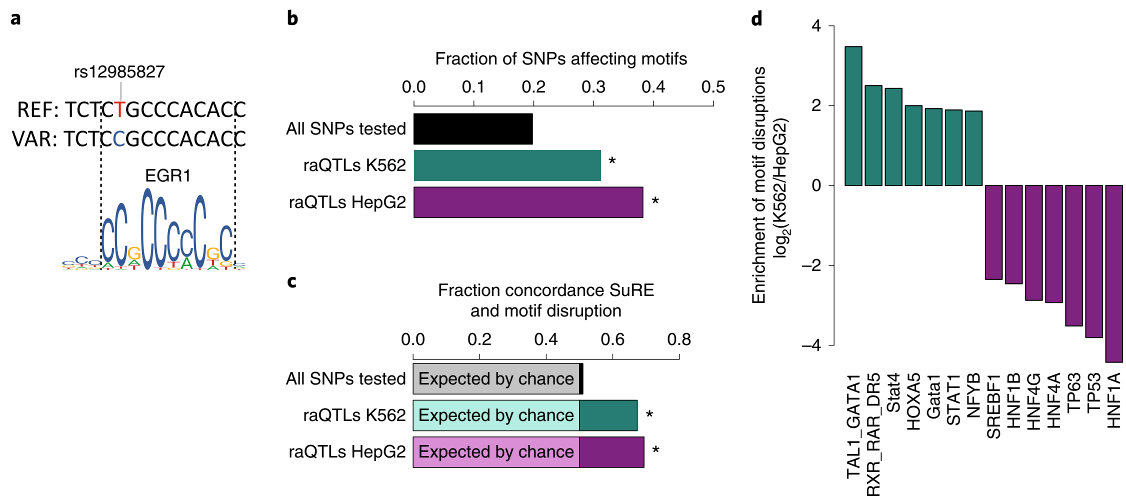


Fig. 3 | Concordance of SuRE data and predictions based on TF binding motifs. **a**, Comparison of the sequence flanking raQTL rs12985827 (same SNP as in Fig. 2d,e) and the sequence logo for EGR1. The T allele disrupts a conserved nucleotide in the EGR1 binding motif. **b**, Compared to all SNPs ($n=5{,}919{,}293$), raQTLs in K562 ($n=19{,}237$) and HepG2 ($n=14{,}183$) both overlap preferentially with computationally predicted alterations of TF binding motifs according to SNP2TFBS[30]. $^{*}P<2.2\times10^{-16}$, according to two-sided Fisher's exact test. **c**, Concordance between the predicted increase or decrease in TF binding according to SNP2TFBS and the observed effect in SuRE, assuming that decreased TF binding typically leads to decreased activity of a regulatory element. $^{*}P<2.2\times10^{-16}$, according to two-sided Fisher's exact test. **d**, TF motif alterations that are preferentially present among raQTLs in either K562 or HepG2 cells. Only the seven most enriched TF motifs for each cell type are shown.

regulatory element that positively regulates the eGene; if an allele reduces the activity of the element then it will also reduce the activity of the eGene. Discordance may point to mechanisms that are more indirect, for example, when an SNP alters the promoter of an antisense transcript that in turn interferes with the sense expression of the eGene. In line with this interpretation, concordant raQTLs are enriched near the TSSs of the eGenes (Supplementary Fig. 4a,b). The slightly higher odds ratios for HepG2 versus liver may be due to a stronger similarity of HepG2 to liver cells than of K562 to blood cells. Because discordant effects are more difficult to interpret, we further focused on SNPs with concordant effects.

Candidate causal SNPs in eQTL maps and their putative mechanism. For several physiologically relevant eGenes we highlight the most likely causal SNPs based on our SuRE data, and we provide insights into the potential underlying mechanisms.

A first example is *XPNPEP2*, a gene associated with the risk of angioedema in patients treated with an angiotensin-converting enzyme inhibitor[34]. The GTEx project has linked the expression of this gene in liver to 33 eQTL SNPs, of which 30 are covered by SuRE (Fig. 4a). Of these, a single SNP (rs3788853) located ~2 kilobases (kb) upstream of the TSS stands out by a strong (approximately fivefold difference between the two alleles) and concordant effect on SuRE activity in HepG2 cells (Fig. 4a). This SNP has previously been demonstrated to alter the activity of an enhancer that controls *XPNPEP2* (ref. [34]). Our independent identification of this SNP indicates that SuRE can successfully pinpoint a functionally relevant SNP among a set of eQTL SNPs.

A second example is the *ABCC11* gene, which encodes a transmembrane transporter of bile acids, conjugated steroids and cyclic nucleotides[35]. GTEx data from liver identified 281 eQTL SNPs surrounding the *ABCC11* TSS (Fig. 4b). SuRE data covered 219 (77.9%) of these, and identified as most likely candidates a cluster of three SNPs within a ~200 bp region located in an intron of *ABCC11* (Fig. 4c). Of these, rs11866312 is a likely candidate because the C allele is predicted to disrupt a binding motif for FOXA1 (Fig. 4d), a pioneer TF that is highly expressed in liver but not in blood[8]. Indeed, virtually no SuRE effect of these SNPs is observed in K562 cells (data not shown).

A third example is the neighboring genes *YEATS4* (encoding a transcription regulator) and *LYZ* (encoding lysozyme, an antibacterial protein). Overlapping sets of eQTL SNPs were identified for these genes in whole blood (Fig. 4e and Supplementary Fig. 4c). Among these, SuRE in K562 cells identified two neighboring raQTLs (rs623853 and rs554591) located ~400 bp downstream of the *YEATS4* TSS, which both show concordance with the eQTL data (Fig. 4e).

To identify TFs that might be responsible for the differential SuRE activity at these two raQTLs, we conducted an in vitro binding proteomic analysis[36,37]. Briefly, we immobilized double-stranded oligonucleotides carrying each of the two alleles on beads, and incubated them with nuclear extract from K562 cells. After washing to remove unbound proteins, we used on-bead trypsin digestion followed by dimethyl stable isotope labeling[38] and quantitative mass spectrometry to identify proteins that preferentially associated with one of the two SNP alleles. In this assay, at rs623853 the weaker A allele caused strong loss of binding of Ets-like factors (Fig. 4f), consistent with a disruption of the cognate motif (Fig. 4g). The A allele also showed moderately increased binding of several other proteins. At rs554591 the C allele caused strong loss of ZNF787 and gain of several other factors including KLF and SP proteins (Supplementary Fig. 4d). The variants of both SNPs may thus cause altered binding of TFs and their cofactors, leading to altered enhancer activity. K562 cells are heterozygous for both rs623853 and rs554591, but no significantly different DHS signal is detectable for either of the alleles (Supplementary Fig. 4e,f).

These examples illustrate that SuRE can prioritize SNPs as likely causal candidates from a set of tens to hundreds of eQTL SNP candidates. By integrating our data with the GTEx datasets, SuRE identified at least one raQTL among the eQTL SNPs for 20.0% of the 8,661 eGenes in whole blood, and for 11.1% of the 4,000 eGenes in liver.

Integrating SuRE data with GWAS data. SuRE may also help to identify candidate causal SNPs in GWAS data. We focused on a large GWAS that identified 6,736 lead SNPs and more than 1 million linked significant SNPs associated with at least one of 36 blood-related traits[39]. These SNPs occurred in LD clusters, each on average

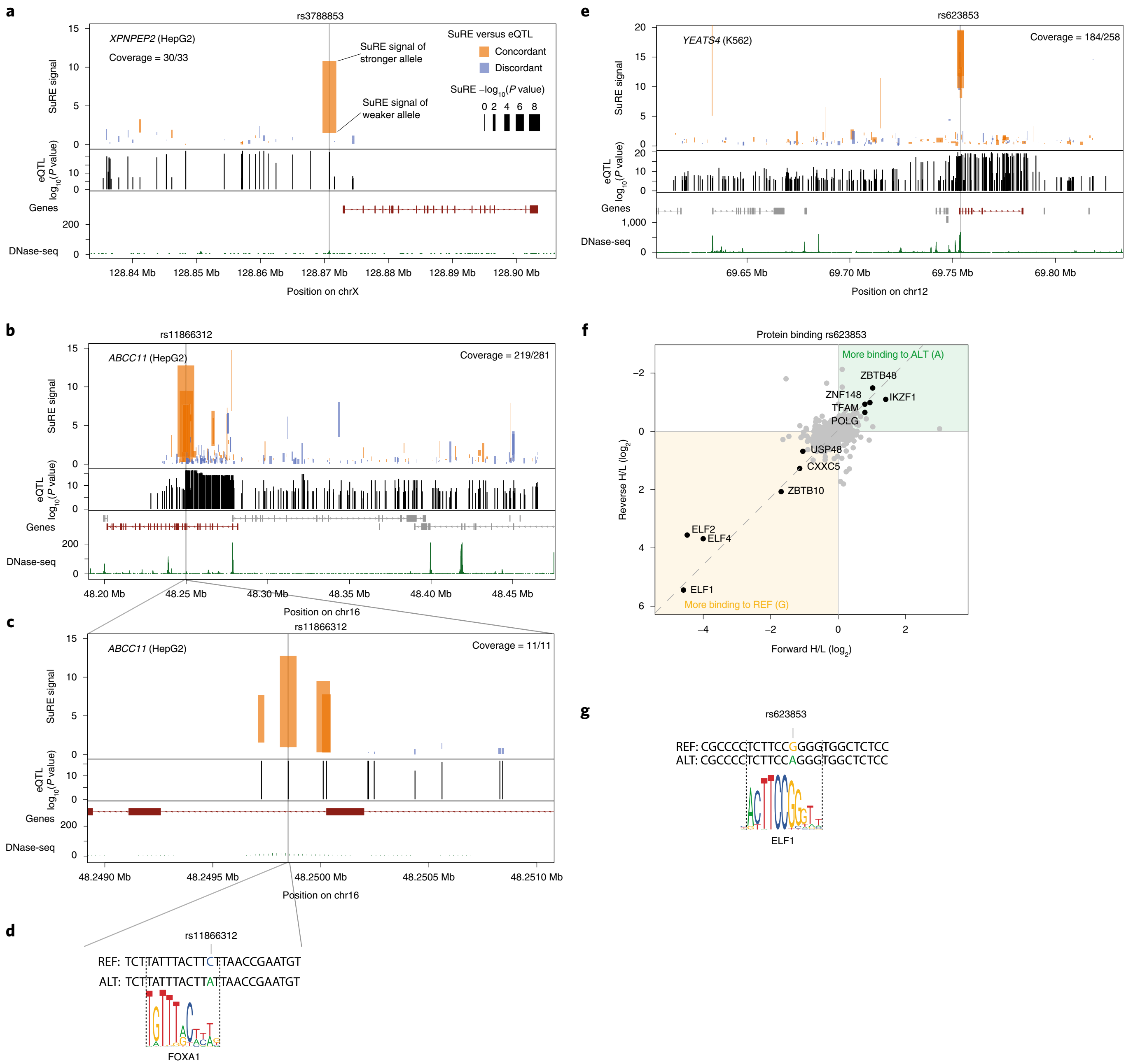


Fig. 4 | Candidate causal SNPs identified by SuRE among large sets of eQTL SNPs. **a**, SuRE signals in HepG2 cells for eQTL SNPs previously identified for the *XPNPEP2* gene in liver according to GTEx v.7 (ref. [8]). SuRE data in HepG2 cells. Top and bottom of each bar indicate the SuRE signal of the strongest and weakest allele, respectively. Width of the bars is proportional to the $-\log_{10}$(P value) obtained by a two-sided Wilcoxon rank-sum test; color indicates whether the eQTL effect orientation is concordant or discordant with the SuRE effect orientation (top). Positions of significant eQTL SNPs with the associated eQTL $-\log_{10}$(P values) according to GTEx v.7 (ref. [8]) (middle). Gene annotation of *XPNPEP2* and DNase-seq data from HepG2 cells[28] (bottom). **b**, Same as **a**, but for *ABCC11* (dark red). **c**, Zoomed in portion of **b**. **d**. Sequence of rs11866312 ± 12 bp aligned to the binding motif of FOXA1. **e**, Same as **a** but for *YEATS4* with SuRE data from K562 cells and eQTL data from whole blood (GTEx v.7 (ref. [8])). eGenes are shown in the bottom panels in dark red and all other genes in gray; coverage numbers in the top panels indicate the number of SNPs with SuRE data out of the total number eQTL SNPs (**a**–**c**,**e**). **f**, Mass spectrometry analysis of proteins from a K562 cell extract binding to 25-bp double-stranded DNA oligonucleotides containing either the A or G allele of rs623853. The experiment was performed once with heavy labeling of proteins bound to the A allele and light labeling of proteins bound to the G allele (*x* axis), and once with reverse labeling orientation (*y* axis). ELF, Ets-like factors. **g**, Sequence of the DNA probes used in **f** aligned to the binding motif of ELF1 (ref. [46]).

consisting of 158 SNPs and represented by a single statistically most significant (lead) SNP. The lead SNPs are not necessarily the causal SNPs, but are more likely to be the causal SNPs or in strong LD with the causal SNP(s). We therefore searched within a 100-kb window from each lead SNP for overlap between significant GWAS SNPs and raQTLs. For 1,238 out of 6,736 lead SNPs this yielded at least one linked raQTL. These raQTLs were preferentially located close to the lead SNPs, compared to a set of matching control SNPs (Fig. 5a). Overall, the enrichment of SuRE raQTLs among the total set of GWAS SNPs did not significantly exceed that of the matching control set of SNPs (1,188), but this was to be expected considering that only one or a few of the, on average, 158 significant SNPs may be true causal SNPs.

One example where SuRE provided a clear candidate among the GWAS SNPs is rs4572196, which is within 100 kb of 11 lead SNPs associated with various mature red blood cell traits, such as 'hemoglobin concentration' and 'hematocrit'[39]. In none of the 11 GWAS associations is rs4572196 the lead SNP, but in SuRE the G allele shows an approximately eightfold higher activity than the A allele ($P = 2.0 \times 10^{-8}$) and it is the only SNP in the region with a *P* value below our cutoff (Fig. 5b). By in vitro proteomics we identified several proteins with differential binding activity to the two rs4572196 alleles. JUN proteins showed stronger binding to the G allele (Fig. 5c), as one might predict based on the JUN binding motif (Fig. 5d). The GWAS demonstrated a positive association between the reference A allele and hemoglobin concentration and red blood cell counts[39]. Interestingly, rs4572196 lies ~11 kb upstream of *SH2B3/LNK*, a gene that inhibits erythropoiesis in mouse[40]. SuRE identifies the A allele as the weak allele, potentially reducing *SH2B3* expression. We cannot rule out that other SNPs in the region, for example, among those not included in our SuRE data, also play a causal role.

Another example is rs3748136, which, together with 66 other SNPs in this locus, was found to be linked to blood counts of reticulocytes. Among the 59 SNPs covered in our data, rs3748136 is the only significant SuRE hit, with the A allele showing an ~18-fold higher activity than the G allele ($P = 7.5 \times 10^{-20}$) (Fig. 5e). K562 cells are heterozygous for this SNP and, indeed, show a strong allelic imbalance in DHS-seq, with the A allele being the more active allele (Fig. 5f). In vitro binding proteomic analysis identified JUN and BACH1 proteins as more strongly bound to the A allele (Fig. 5g), consistent with the G allele disrupting predicted binding motifs for both proteins (Fig. 5h). Both BACH1 and JUN proteins are highly expressed in whole blood and in K562 cells[8,28]. Reanalysis of chromatin immunoprecipitation (ChIP) data for BACH1 (ref. [28]) showed a complete allelic imbalance for BACH1 binding and the same was found for JUND (Fig. 5i,j). eQTL analysis of whole blood[8] has linked the A allele of rs3748136 to elevated expression of the nearby non-coding RNA gene *NR_125431* (Fig. 5k). To further test this, we modified the G allele in K562 cells to an A allele by CRISPR-Cas9 editing[41]. We found that expression of *NR_125431* in K562 cells shows considerable clone-to-clone variation (Supplementary Fig. 5a–c). To account for this we performed G to A substitution in a stable K562 clone. This modification led to a fourfold upregulation of *NR_125431* (Fig. 5l).

Finally, we overlaid the HepG2 SuRE data with a recent GWAS that linked SNPs to the risk of hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC), a type of cancer prevalent in Asia[42]. This study identified 61 candidate SNPs in an ~200-kb region that includes the *HLA class 1* locus. Of the 50 SNPs covered by our SuRE data, two are raQTLs in HepG2 (Fig. 6). These raQTLs are intergenic; both ALT alleles show reduced SuRE signals and are associated with higher HCC risk. Besides the *HLA* genes that are important for antigen presentation, other genes in this region could be affected by these raQTLs and play a role in HCC. For example, *ZNRD1* and its antisense transcript have been implicated in expression of HBV mRNA and proliferation of HBV-infected HepG2 cells[43].

Discussion

By surveying 5.9 million SNPs from four entire human genomes we identified about 30,000 SNPs that alter regulatory activity of enhancers or promoters. The data are available for download, and can be queried through a web application (https://sure.nki.nl).

Because 90% of these raQTLs were identified in only one of the two tested cell lines, it is likely that extension of this survey to other cell types will increase the number of raQTLs substantially. It is thus conceivable that several percent of all human SNPs may have an impact on the activity of regulatory elements in at least one cell type.

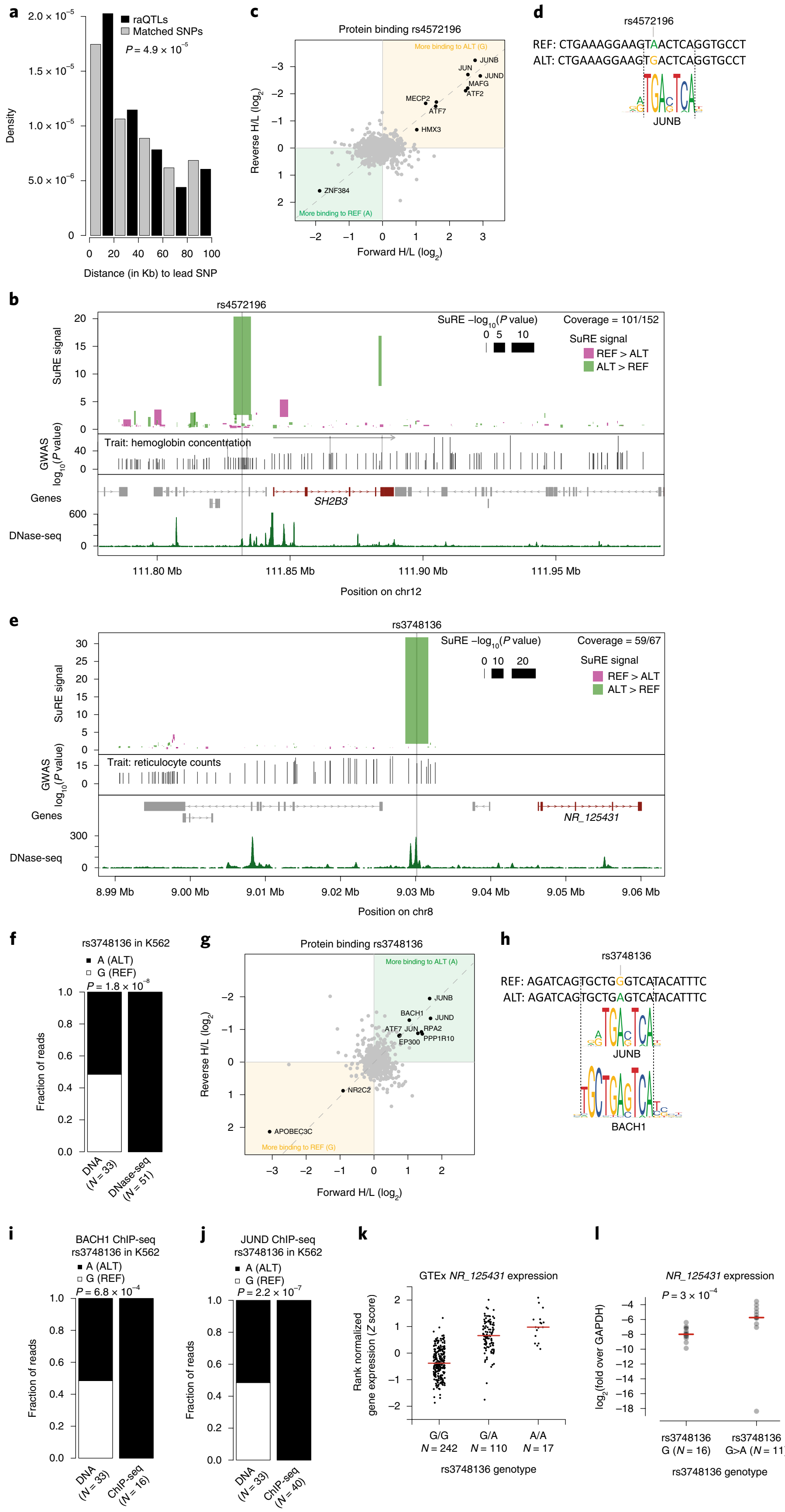
While the redundant design of SuRE increases the odds that a robust and biologically representative measure of SNP effects is obtained, we note that SuRE signals arising from enhancers are generally weaker than those from promoters, and it cannot be ruled out that certain enhancers cannot be detected by SuRE at all. Thus, our approach may be better powered to detect effects of SNPs on promoters compared to enhancers. Furthermore, SuRE infers effects of SNPs on enhancer activity indirectly, by virtue of the ability of most enhancers to act as autonomous TSSs. Although this feature generally correlates with the potency of enhancers to activate a *cis*-linked promoter, this correlation is not perfect[44] and, in some enhancers, both activities may be differentially affected by particular SNPs.

Like most other MPRAs, SuRE queries all DNA elements in a plasmid context and in cultured cell lines, which may yield different results compared to a proper genomic context and tissue context. Integration with multiple orthogonal datasets can help provide confidence in the relevance of candidate SNPs. Epigenomic data, sequence motif analysis and in vitro binding mass spectrometry[36,37] can serve this purpose, and, in addition, provide key insights into the mechanisms of action.

We foresee several additional applications of these SuRE data. First, there are many other eQTL studies and GWAS that may be overlaid with the SuRE maps. Second, in addition to SNPs, small insertions and deletions (indels) may be analyzed. While, in human genomes, such indels occur at an ~20-fold lower frequency than SNPs[1], their individual regulatory impact may be more potent, as they tend to disrupt TF binding motifs more dramatically. Third, our datasets may be useful for studying the regulatory grammar of

Fig. 5 | Candidate causal SNPs identified by SuRE among large sets of GWAS SNPs. a, Distribution of distances between lead SNPs for blood traits[39] and raQTLs (black) and a set of matched control SNPs (gray). *P* value was obtained with a two-sided Wilcoxon rank-sum test. raQTLs in K562 cells are modestly enriched near blood GWAS lead SNPs. *P* value was obtained using two-sided Wilcoxon rank-sum test. **b**, Overlay of SuRE and GWAS data for a cluster of GWAS SNPs linked to hemoglobin concentration[39]. SuRE data in K562 cells. Top and bottom end of each bar indicate the SuRE signal of the strongest and weakest allele, respectively. Color of the bars indicates which allele is stronger. Width of the bars is proportional to $-\log_{10}$(*P* value) (top). Positions of significant GWAS SNPs with the associated $-\log_{10}$(*P* values)[39] on the *y* axis (middle). Gene annotation (dark red: *SH2B3*) and DNase-seq data from K562 cells[28] (bottom). **c**, Protein binding analysis, as in Fig. 4f, for rs4572196. **d**, Sequence of the probes used in **c** aligned to sequence logo for JUNB. **e**, Same as **b** but for a cluster of SNPs associated with reticulocyte counts by GWAS[39]. **f**, Fraction of reads containing each of the two alleles of rs3748136 in K562 genomic DNA and K562 DNase-seq reads. *P* value was obtained with a two-sided Fisher's exact test. **g**, Same as **c** but for rs3748136. **h**, Sequence of the probes used in **g** aligned to binding motifs for JUNB and BACH1. **i**, Fraction of reads containing each of the two alleles of rs3748136 in K562 genomic DNA (left) and K562 ChIP-seq reads for BACH1 (right). **j**, Same as **i** but for ChIP-seq reads for JUND. ChIP data are from (ref. [28]). **k**, Association between alleles of rs3748136 and *NR_125431* expression in whole blood according to GTEx[8]. Red lines indicate median. **l**, Expression of *NR_125431* in subclones derived from K562 clone BL_2 subjected to CRISPR/Cas9 editing of rs1053036. Sixteen unaltered subclones and 11 G>A edited subclones were assayed by RT–qPCR of *NR_125431* (normalized to *GAPDH*). *P* value was obtained with a two-sided Wilcoxon rank-sum test. Red lines indicate medians. One G>A subclone appeared to have reverted to the completely inactive state seen in many K562 clones initially derived from the cell pool (Supplementary Fig. 5c).

TFs, as they cover natural genetic variation in thousands of regulatory elements. For example, the SuRE data may be used to refine computational predictions of SNP effects[13,30,45]. Finally, it will be interesting to expand this type of analysis to individuals with a genetic disorder to capture additional disease-relevant variants that might not be found in the general population.



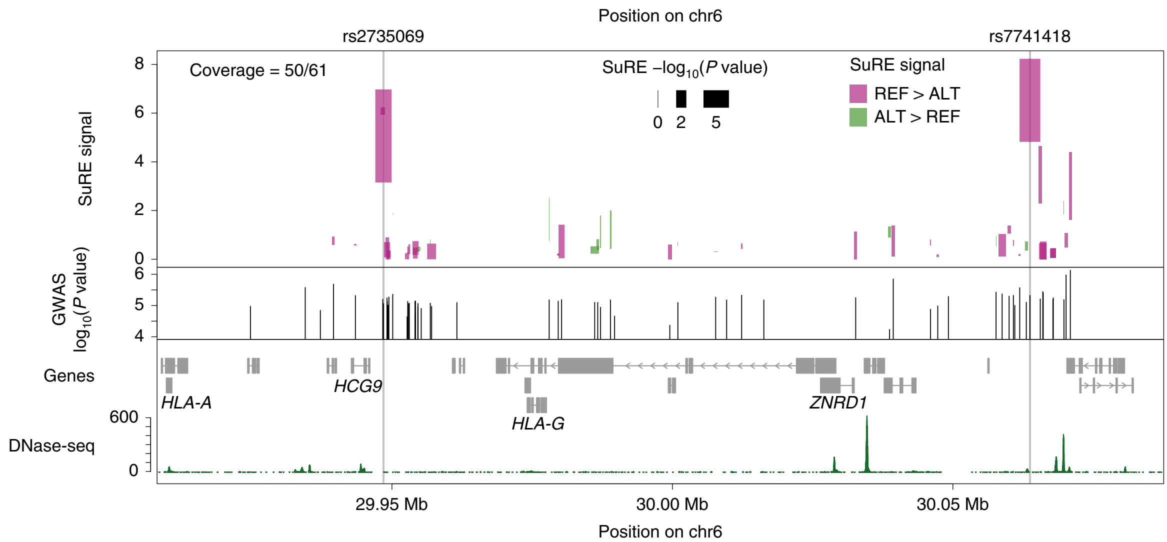


Fig. 6 | Candidate causal SNPs identified by SuRE among GWAS SNPs for HCC. Comparison of SuRE and GWAS data for a cluster of GWAS SNPs linked to HCC[42]. SuRE data in HepG2 cells. The top and bottom end of each bar indicate the SuRE signal of the strongest and weakest allele, respectively. Color of the bars indicates which allele is stronger. Width of the bars is proportional to $-\log_{10}$(P value) obtained by a two-sided Wilcoxon rank-sum test (top). Positions of significant GWAS SNPs with the associated $-\log_{10}$(P values)[42] on the y axis (middle). Gene annotation track and DNase-seq data from HepG2 cells[28] (bottom).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0455-2.

Received: 30 October 2018; Accepted: 24 May 2019;
Published online: 28 June 2019

References

1. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
3. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).
4. Miguel-Escalada, I., Pasquali, L. & Ferrer, J. Transcriptional enhancers: functional insights and role in human disease. *Curr. Opin. Genet. Dev.* **33**, 71–76 (2015).
5. Deplancke, B., Alpern, D. & Gardeux, V. The genetics of transcription factor DNA binding variation. *Cell* **166**, 538–554 (2016).
6. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
7. Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS era: from association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
8. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
9. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
10. Nishizaki, S. S. & Boyle, A. P. Mining the unknown: assigning function to noncoding single nucleotide polymorphisms. *Trends Genet.* **33**, 34–45 (2017).
11. Farh, K. K. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
12. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin* **8**, 57 (2015).
13. Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
14. Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
15. McVicker, G. et al. Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–749 (2013).
16. Kilpinen, H. et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744–747 (2013).
17. Kasowski, M. et al. Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).
18. Waszak, S. M. et al. Population variation and genetic control of modular chromatin architecture in humans. *Cell* **162**, 1039–1050 (2015).
19. Grubert, F. et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* **162**, 1051–1065 (2015).
20. Gate, R. E. et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).
21. Vockley, C. M. et al. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.* **25**, 1206–1214 (2015).
22. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).
23. Ulirsch, J. C. et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).
24. Liu, S. et al. Systematic identification of regulatory variants associated with cancer risk. *Genome Biol.* **18**, 194 (2017).
25. Zhang, P. et al. High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing. *Nat. Commun.* **9**, 2022 (2018).
26. van Arensbergen, J. et al. Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* **35**, 145–153 (2017).
27. Nakamura, M. et al. Genome-wide association study identifies TNFSF15 and POU2AF1 as susceptibility loci for primary biliary cirrhosis in the Japanese population. *Am. J. Hum. Genet.* **91**, 721–728 (2012).
28. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
29. Liu, X. et al. In situ capture of chromatin interactions by biotinylated dCas9. *Cell* **170**, 1028–1043.e19 (2017).
30. Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **45**, D139–D144 (2017).
31. Law, J. C., Ritke, M. K., Yalowich, J. C., Leder, G. H. & Ferrell, R. E. Mutational inactivation of the p53 gene in the human erythroid leukemic K562 cell line. *Leuk. Res.* **17**, 1045–1050 (1993).
32. Westerink, W. M., Stevenson, J. C., Horbach, G. J. & Schoonen, W. G. The development of RAD51C, Cystatin A, p53 and Nrf2 luciferase-reporter assays in metabolically competent HepG2 cells for the assessment of mechanism-based genotoxicity and of oxidative stress in the early research phase of drug development. *Mutat. Res.* **696**, 21–40 (2010).
33. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
34. Cilia La Corte, A. L. et al. A functional XPNPEP2 promoter haplotype leads to reduced plasma aminopeptidase P and increased risk of ACE inhibitor-induced angioedema. *Hum. Mutat.* **32**, 1326–1331 (2011).
35. Chen, Z. S., Guo, Y., Belinsky, M. G., Kotova, E. & Kruh, G. D. Transport of bile acids, sulfated steroids, estradiol 17-β-D-glucuronide, and leukotriene C4

by human multidrug resistance protein 8 (ABCC11). *Mol. Pharmacol.* **67**, 545–557 (2005).
36. Makowski, M. M. et al. An interaction proteomics survey of transcription factor binding at recurrent TERT promoter mutations. *Proteomics* **16**, 417–426 (2016).
37. Makowski, M. M. et al. Global profiling of protein–DNA and protein–nucleosome binding affinities using quantitative mass spectrometry. *Nat. Commun.* **9**, 1653 (2018).
38. Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protoc.* **4**, 484–494 (2009).
39. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429.e19 (2016).
40. Maslah, N., Cassinat, B., Verger, E., Kiladjian, J. J. & Velazquez, L. The role of LNK/SH2B3 genetic alterations in myeloproliferative neoplasms and other hematological disorders. *Leukemia* **31**, 1661–1670 (2017).
41. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
42. Sawai, H. et al. Genome-wide association study identified new susceptible genetic variants in HLA class I region for hepatitis B virus-related hepatocellular carcinoma. *Sci. Rep.* **8**, 7958 (2018).
43. Wen, J. et al. Expression quantitative trait loci in long non-coding RNA ZNRD1-AS1 influence both HBV infection and hepatocellular carcinoma development. *Mol. Carcinog.* **54**, 1275–1282 (2015).
44. Nguyen, T. A. et al. High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* **26**, 1023–1033 (2016).
45. Ward, L. D. & Kellis, M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* **44**, D877–D881 (2016).
46. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).

Acknowledgements
We thank the NKI Genomics Core Facility and Research IT team for technical support, the RHPC facility of the Netherlands Cancer Institute for providing computational resources and members of our laboratories for helpful discussions. Supported by ERC Advanced Grant no. 694466 (to B.v.S.); ERC Starting Grant no. 637587 (to E.d.W.); NIH grant no. R01HG003008 and Columbia University's Vagelos Precision Medicine Pilot Program (to H.J.B.). F.C. was supported by a Swiss National Science Foundation postdoctoral fellowship (no. P2EZP3_165206). J.v.A., L.P., M.d.H., M.P.B., F.C., R.H.v.d.W., H.T., E.d.W., M.V. and B.v.S. are part of the Oncode Institute, which is partly funded by the Dutch Cancer Society (KWF).

Author contributions
J.v.A. designed and performed experiments, analyzed data and wrote the manuscript. L.P., V.D.F. and H.J.B. developed algorithms and analyzed data. M.d.H. M.P.B., M.V., R.H.v.d.W., H.T., F.C., U.V., E.d.W. and L.F. generated and/or analyzed data. F.C. developed the web application. B.v.S. designed experiments, analyzed data and wrote the manuscript.

Competing interests
J.v.A. is founder of Gen-X B.V. (http://www.gen-x.bio/). E.d.W. is co-founder and shareholder of Cergentis B.V. F.C. is a co-founder of enGene Statistics GmbH.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0455-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.v.A. or B.v.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

SuRE library preparation and barcode-to-fragment association. SuRE libraries were generated as described previously[26]. DNA was isolated from lymphoblast cell lines HG02601, GM18983, HG01241 and HG03464 obtained from Coriell Institute, fragmented and gel-purified to obtain ~300-bp elements. For each genome, two SuRE libraries were generated, each of an approximate complexity of 300 million fragment–barcode pairs. This was done by transformation in CloneCatcher DH5G electrocompetent *Escherichia coli* cells (Genlantis, catalog no. C810111), or in E. cloni 10G cells (Lucigen, catalog no. 60107-1). Barcode-to-fragment association was done as described previously[26], except that because of the smaller genomic insert size no digest with a frequent cutter was required. Thus, after I-CeuI digest and self-ligation we immediately proceeded to the I-SceI digest.

Cell culture and transfection for SuRE. K562 cells were cultured and transfected as described[26]. HepG2 (ATCC, catalog no. HB 8065) were cultured according to the supplier's protocol and transiently transfected in the same manner as K562 cells, except that program T-028 was used for nucleofection, 7.5 µg of plasmid was used for each 5 million cells and cells were collected 48 h after transfection. One hundred million cells were transfected for each replicate. Every 3 months all cells in culture were screened for mycoplasma using PCR (Takara, catalog no. 6601).

Illumina sequencing. Paired-end (PE) sequencing (150 bp) of SuRE libraries was done by Novogene on the HiSeq-X platform, generating about 1 billion reads per library. Standard full genome sequencing and allele calling for the K562 cell line were done by Novogene on the HiSeq-X10 platform with PE 150-bp reads amounting to approximately 100 gigabases or an ~30-fold coverage of the genome. Single-end sequencing on reverse-transcribed, PCR-amplified barcodes was done by the NKI Genomics Core Facility on a HiSeq2500 machine.

Sequencing data processing. PE reads of the SuRE libraries (for associating genomic positions and barcodes for each SuRE fragment) and single-end reads (SE reads) of the PCR-amplified barcodes (representing raw SuRE expression data) were processed to remove adapter and vector backbone sequences, using cutadapt (v.1.9.1)[47]. PE and SE reads were discarded if the barcode sequence contained Ns or the sequence was not exactly 20 nucleotides. The remaining sequences in the PE reads are combinations of barcode sequences and genomic DNA sequences, whereas the SE reads only yield barcode sequences. The latter barcodes are simply recorded and counted. The genomic DNA sequences of the SuRE libraries were mapped to the reference genome sequence (hg19, including only chr1–22, chrX), using bowtie2 (v.2.3.2)[48], with a maximum insert length set to 1 kb. Read pairs with either the forward or the reverse genomic DNA sequence less than six nucleotides, and read pairs not aligned as 'proper pair', were discarded. To prevent allelic biases in alignment we used WASP[49] and SNP annotations from the 1000 Genome Project (Supplementary Table 3, external data source 17) to discard all reads potentially resulting in biased alignments.

The resulting associations of barcode sequence–genomic position pairs were further processed as follows:

(1) Identical barcodes associated with multiple alignment positions were discarded except for the most abundant barcode–position pair.
(2) Different barcodes associated with the exact same alignment position were merged; that is, the barcode sequence associated with this position was set to the most frequent barcode sequence in the set, and the total number of PE reads in the set was used as count for this barcode–position pair.

Next, the barcodes identified in the SE reads were matched to the barcodes in the remaining barcode–position pairs, and 'SuRE count' tables were generated associating barcode sequences, genomic positions and counts for associated PE reads and matched SE reads for each of the biological replicates.

SNP annotation. The fragments, specified in the SuRE count tables were further annotated with SNP positions and base identities. For this annotation only SNPs were considered that were single-nucleotide, bi-allelic and the alternative allele in at least one of all four considered genomes. SNPs homozygous for the reference allele in all four genomes were discarded. For each SNP in such a fragment we determined its base identity as observed in the actual sequence reads. Some fragments are too long to be entirely covered by the PE reads. In these cases the unidentified SNPs were assigned the IUPAC representation of both alleles; if the two alleles were known to be identical in the genome then that was used to construct the particular library (based on annotation by the 1000 Genome Project; Supplementary Table 3,external data source 17), this inferred allele was used for annotation. At the time of finalizing this article we noticed that, due to a minor coding error, ~0.4% of the SuRE fragments were incorrectly annotated to carry both the REF and ALT allele of an SNP. This may cause a very slight underestimate of the total number of raQTLs, but it does not alter any of the reported conclusions.

SuRE data analysis and visualization. Data analysis and figure production was mostly done using various R (https://www.R-project.org) and BioConductor[50] packages.

Generating BigWigs of SuRE enrichment profiles. For each strand separately, we determined the cDNA barcode count for all SuRE fragments overlapping a given position. This total was divided by the total counts of the SuRE fragments measured in the SuRE library (iPCR barcode counts) to give the SuRE enrichment score. Within each transfection replicate the genome-wide normalized SuRE enrichment score was scaled to a mean of 1. Transfection replicates were then averaged to yield a genome-wide normalized SuRE enrichment profile per SuRE library. Then, the library replicates were also averaged to yield a genome-wide normalized SuRE enrichment profile per genome. BigWig files were generated from these datasets. This analysis was done disregarding SNPs and is therefore independent of the identification of the raQTLs (see next section).

Identification of raQTLs. *Equalization of cDNA barcode sequencing depth.* To minimize biases that might be caused by excessive differences in sequencing depth, complementary DNA reads of some samples were subsampled. First, for each transfection replicate the relative sequencing depth of cDNA barcodes was determined as the total cDNA barcode counts divided by the corresponding library complexity (that is, the number of unique fragments identified in the library). Then, samples with a relative cDNA sequencing depth that exceeded the mean of all samples by more than one standard deviation (that is, all K562 and HepG2 transfection replicates for genome HG02601 library 1, and all HepG2 transfection replicates for genome HG02601 library 2) were down-sampled to the mean relative cDNA read depth.

SuRE signal normalization per fragment. For each fragment in each of the eight SuRE libraries we normalized the iPCR barcode count by dividing it by the total iPCR sequencing depth of that library, expressed as reads per billion. Similarly, we normalized cDNA barcode counts per transfection replicate by dividing it by the total cDNA sequencing depth, again expressed as reads per billion. Next, for each fragment we calculated the SuRE signal (S_{SuRE}), which is the mean of all transfection replicates of the ratio of the normalized cDNA barcode count over normalized iPCR barcode count. Since both values used to obtain this ratio are expressed as reads per billion, we interpret the resulting S_{SuRE} values as an enrichment score.

Aggregation of SuRE signals by SNP allele and P value calculation. Using this S_{SuRE} per fragment, a mean $\overline{S}_{SuRE}$ was calculated for each allele of each SNP as the mean S_{SuRE} of all fragments containing that SNP allele. Also, for each SNP a two-sided Wilcoxon rank-sum test was performed comparing the set of S_{SuRE} values of all fragments containing the REF allele with the set of S_{SuRE} values of all fragments containing the ALT allele. The resulting vector of P values (one for each SNP) is referred to as $\boldsymbol{P}_t$. In addition, to estimate the FDR, the same two-sided Wilcoxon rank-sum test was applied once after random shuffling of the S_{SuRE} of the fragments among the two alleles, yielding a vector of random P values referred to as $\boldsymbol{P}_r$.

Criteria for raQTL definition. We then focused on those SNPs for which both alleles were covered by at least ten fragments and no more than 999 fragments, and for which at least one of the alleles met the criterion $\overline{S}_{SuRE} > 4$ (170,118 SNPs in K562; 395,756 SNPs in HepG2). For FDR = 5%, we then chose the lowest P value cutoff P_{cut} for which the number of SNPs with $\boldsymbol{P}_t < P_{cut}$ was at least 20 times larger than the number of SNPs with $\boldsymbol{P}_r < P_{cut}$. We refer to this set of SNPs as raQTLs. The same procedure was applied separately for data from K562 and HepG2 cells. For K562, P_{cut} was 0.006192715 and for HepG2 P_{cut} was 0.00173121.

Enrichment of raQTLs in ENCODE classes. For Fig. 2a we used the GenomicRanges package of BioConductor[50] to determine the enrichment or depletion of raQTLs in ENCODE chromatin classes (Supplementary Table 3, external data source 7) compared with all 5.9 million SNPs that we assessed.

Comparison of SuRE to DNase-seq, H3K27ac and ATAC-seq allelic imbalance. In Fig. 2 we plotted the DNase-seq signal around the raQTLs using external data source 4 (see Supplementary Table 3) and BioConductor package CoverageView (v.1.4.0) and 25-bp windows. Figure 2c was generated from the same data. For the analysis of allelic imbalance in the DNase-seq signal we combined three available experiments (Supplementary Table 3, external data source 1–3) and extracted, from the bam files, the reads that overlapped an SNP that was found to be heterozygous in K562 in our own genome-sequencing analysis (see section 'Illumina sequencing'). We focused on those raQTLs for which we found at least 20 DNase-seq reads and quantified the ratio of reads containing the REF allele over the reads containing the ALT allele, after adding a pseudocount of 1 for each allele. Similarly, from our own genome sequencing of K562 we quantified the ratio of reads containing the REF allele over reads containing the ALT allele, after adding a pseudocount of 1 for each allele. Finally, the DNase allelic imbalance was calculated as the DNase-seq allele ratio over the genomic allele ratio.

For the SuRE data, the allelic imbalance was calculated as the ratio of S_{SuRE} for the REF allele over the S_{SuRE} of the ALT allele (since both these values are already normalized for coverage in the libraries).

To obtain a matching set of control SNPs we intersected a DNase peak annotation (Supplementary Table 3, external data source 7) with our SNPs, and we retrieved the DNase-seq allele counts for the 2,500 overlapping SNPs with the highest SuRE P_t values. We required at least 20 DNase-seq reads covering the SNP, and from the resulting set we randomly selected a subset of SNPs of the same size as the set of raQTLs (Fig. 2f). We applied the same analysis of allelic imbalance to this control set as we did to the raQTLs. The comparison in Supplementary Fig. 2 with H3K27ac (Supplementary Table 3, external data source 20 and 21) and ATAC-seq (Supplementary Table 3, external data source 22 and 23) was done in the same manner as that described for the DNase-seq data, except that we required only ten reads to cover the SNP, since these data are of approximately 5–10 times lower sequencing depth. We excluded a third ATAC-seq replicate (GSM2695562), since, for this replicate, we did not observe a similar pattern of enrichment as for the other ATAC-seq replicates. The mean enrichment profiles shown in Supplementary Fig. 2c–e were generated with BioConductor package CoverageView (v.1.4.0).

Re-mapping of BACH1 and JUND ChIP-seq data. Fastq files were downloaded from the SRA repository (Supplmentary Table 3, external data sources 13 and 14) using fastq-dump from the SRA-tools package (v.2.8.2, https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/). For BACH1 we downloaded data from datasets SRR502556 and SRR502557, and for JUND from datasets SRR502542 and SRR502543. Reads were aligned to the human reference sequence (hg19, including only chr1–22, chrX), using bowtie2 (v.2.3.2)[48] with default settings.

Allele frequencies. MAFs of SNPs were obtained from the 1000 Genomes Project (Supplementary Table 3, external data source 17). Common SNPs were defined as SNPs with MAF > 0.05. Of the 5,919,293 SNPs in our SuRE dataset, 4,569,323 classify as common. This is ~57% of the estimated ~8 million common SNPs according to the 1000 Genomes Project[1]. The proportion of raQTLs for which the SuRE effect could be resolved to a single SNP was calculated as the fraction of raQTLs for which neither neighbor SNP was also a raQTL.

TF motif analysis. We made use of SNP2TFBS (Supplementary Table 3, external data source 16)[30] to identify all SNPs for which there was a difference in predicted affinity for a TF between the REF and ALT allele. For each SNP we only considered the TF listed first in SNP2TFBS (that is, the TF with the biggest absolute difference in motif score). For the comparison of motif disruptions in K562 and HepG2 we identified all raQTLs of K562 and HepG2 that caused motif disruptions, we down-sampled the K562 disruptions to the same number as the HepG2 disruptions (since K562 has more raQTLs) and we plotted the ratio of the counts (plus one pseudocount) for the seven most extreme ratios for each cell type. Sequence logos were obtained from external data source 19 (see Supplementary Table 3).

Compiling a set of control SNPs that are matched to the significant SNPs. For the analyses in Supplementary Fig. 3 and Fig. 5a we compared the set of raQTLs to a control set of matching SNPs that was selected as follows. We ranked all SNPs, first by the S_{SuRE} (rounded to whole numbers) of the strongest allele, second by the number of fragments containing the least covered allele and third by the number of fragments containing the most covered allele. This is intended to rank them based on regulatory element activity (S_{SuRE}) and our sensitivity to detect a significant difference (coverage of the alleles). Then we identified the raQTLs along this ranking and selected both direct neighbors, removed the raQTLs and down-sampled the resulting set to yield the same number of matched SNPs as raQTLs.

raQTL density around LOF-tolerant and LOF-intolerant genes. For the definition of LOF-tolerant and LOF-intolerant genes, we used external data source 11 (Supplementary Table 3) and the same cutoffs as previously reported[33], classifying genes as having pLI scores of >0.9 as intolerant and <0.1 as tolerant. We then matched these genes to GENCODE v.19 annotation (Supplementary Table 3, external data source 12) to identify the corresponding TSSs. Using the GenomicRanges package of BioConductor[50] we defined regions around these TSSs of ±100 kb and subtracted all possible exons (as defined in GENCODE v.19). In the resulting regions we then determined the fraction of all SNPs that are a raQTL, and the fraction that belonged to the set of matched control SNPs. This analysis was done for both LOF-intolerant and LOF-tolerant genes.

Integration with eQTL data. GTEx eQTL data[8] (release v.7; external data source 18, Supplementary Table 3) were downloaded on 27 January 2018. For whole blood we used the extracted file Whole_Blood.v7.signif_variant_gene_pairs.txt.gz and for liver we used Liver.v7.signif_variant_gene_pairs.txt.gz.

Gene annotation tracks in Fig. 4 were generated by the Gviz package of BioConductor, function BiomartGeneRegionTrack(). ENCODE[28] DNase-seq data used in this figure are from external data sources 4 and 15 (Supplementary Table 3).

Integration with GWAS data. Overlap between SNPs identified in the GWAS study by Astle et al.[39], was obtained by searching for significant SuRE SNPs within 100 kb of each of the 6,736 lead SNPs identifying 1,238 lead SNPs within 100 kb of at least one significant SuRE SNP (Supplementary Table 3, external data source 6). The window of 100 kb was chosen to be substantially larger than the typical size of an LD block. For these lead SNPs we calculated the distance to SNP with the lowest P value in our SuRE data (Fig. 5a). As a control we did the same procedure for the set of matched SuRE SNPs. In Fig. 5 only those SNPs with significant P values in the GWAS (cutoff: $P < 8.31 \times 10^{-9}$; ref. 39) are shown. Note that what we refer to as 'lead SNPs' are called 'conditionally independent index-variant associations' in the original GWAS[39]. Gene annotation tracks in Fig. 5 were generated by the Gviz package of BioConductor, function BiomartGeneRegionTrack(). DNase-seq data used are from external data source 4 (Supplementary Table 3).

CRISPR/Cas9-mediated editing of rs3748136. We performed our CRISPR experiments on a K562 subclone in which *NR_125431* was active (subclone BL_2) because initial experiments revealed that in the K562 pool, *NR_125431* is expressed in only ~25% of the cells (Supplementary Fig. 5c). Five million of the BL_2 cells were nucleofected as described above with 2 µg of vector pX330-U6-Chimeric_BB-CBh-hSpCas9, a gift from F. Zhang (Addgene, plasmid no. 42230)[41], encoding Cas9 and the chimeric guide RNA; and 20 pmol repair template (see Supplementary Table 2 for nucleotide sequences of guide RNAs). Cells were then cultured for 3 d in the presence of 1 µM of DNA-PK inhibitor (NU7441; Cayman Chemical Company) and then expanded for another 5 d without this inhibitor. For genotyping we used a PCR amplicon that included SNP rs453301, ~250 bp downstream of rs3748136 that was also heterozygous in K562. After confirming editing efficiency for the population of cells using Sanger sequencing and TIDE analysis[51], single cells were cloned out. After expansion, clones were genotyped on the single PCR amplicon, and classified as successfully edited when they were heterozygous (that is, not edited) at rs453301 but homozygous for rs3748136, or they were classified as wild type when both loci were still heterozygous. Of note, we identified many clones in which our CRISPR editing caused deletions around the targeted SNP; these were discarded. Successfully edited clones and wild-type clones were then analyzed for RNA expression by quantitative PCR with reverse transcription (RT–qPCR). Since the chromosome that was edited at rs3748136 was the only chromosome showing expression of *NR_125431* to begin with, we are looking at the increased expression from that chromosome after editing (see main text), even though the RT–qPCR is not allele specific. See Supplementary Table 2 for oligonucleotide sequences used.

Targeted locus analysis. To determine the phasing of the K562 genome around rs1053036, targeted locus analysis was performed essentially as described[52]. Briefly, roughly 5 million K562 cells were cross-linked with 4% formaldehyde and cut with NlaIII. After ligation, the template was de-cross-linked and further digested with NspI. The second ligation yields circular DNA, which is used as input for the inverse PCR reaction. We performed two PCR reactions: the first with primers adjacent to rs1053036, located in the last exon of *NR_125431* and the second with primers adjacent to rs3748136, located in the intergenic region (Supplementary Table 2). The PCR amplicons were combined and we generated sequencing libraries using the KAPA High Throughput Library Preparation Kit (Roche, catalog no. 7961901001). We generated 2 × 150-bp PE sequences on an Illumina MiSeq. Sequence reads were mapped to hg19 using BWA-SW[53]. The resulting bam files and the K562 vcf file (obtained from whole-genome sequencing at Novogene) were used as input for HapCUT2 (ref. [54]) with the --hic option turned on to phase the alleles.

RT–qPCR. RNA was isolated from 1–5 million cells using Trisure (Bioline, catalog no. BIO-38033). DNase digestion was performed on ~1.5 µg of RNA with 10 units of DNase I for 30 min (Roche, catalog no. 04716728001) and DNase I was inactivated by addition of 1 µl of 25 mM EDTA and incubation at 70 °C for 10 min. cDNA was produced by adding 1 µl of 50 ng µl^{-1} random hexamers and 1 µl of dNTPs (10 mM each) and incubated for 5 min at 65 °C. Then, 4 µl of first strand buffer, 20 units of RNase inhibitor (ThermoFisher Scientific, catalog no. EO0381), 1 µl of Tetro reverse transcriptase (Bioline, catalog no. BIO-65050) and 2 µl of water were added and the reaction mix incubated for 10 min at 25 °C followed by 45 min at 45 °C and heat inactivation at 85° for 5 min. Quantitative PCR (qPCR) was performed on the Roche LightCycler480 II using the Sensifast SYBR No-ROX mix (Bioline, catalog no. BIO-98020). All expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method and normalized to the internal control *GAPDH*. See Supplementary Table 2 for oligonucleotide sequences used for qPCR.

DNA affinity purification and liquid chromatography–mass spectrometry (LC–MS) analysis. Nuclear extracts were generated from K562 cells essentially as described[55]. Briefly, cells were washed with PBS and then resuspended in 5× cell pellet volumes of hypotonic buffer A (10 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl). After incubation for 10 min at 4 °C, cells were collected by centrifugation and resuspended in two pellet volumes of buffer A supplemented with 0.15% NP40. Cells were then lysed by dounce homogenization using 35 strokes with a type B (tight) pestle on ice. Crude nuclei were collected by centrifugation and then lysed in two pellet volumes of buffer C (420 mM NaCl, 20 mM HEPES pH 7.9, 20% (v/v) glycerol, 2 mM $MgCl_2$, 0.2 mM EDTA, 0.1% NP40, EDTA-free complete protease inhibitors (Roche) and 0.5 mM DTT) by rotation for 1 h at 4 °C. After centrifugation for 20 min at 21,000*g*, nuclear extract was collected as the soluble fraction. This extract was then aliquoted, snap-frozen and stored at −80 °C until further use.

Oligonucleotides for the DNA affinity purifications were ordered from Integrated DNA Technologies with the forward strand containing a 5′-biotin moiety (see Supplementary Table 2). DNA affinity purifications and on-bead trypsin digestion were performed on 96-well filter plates essentially as described[37]. Tryptic peptides from SNP allele pull-downs were desalted using Stage (stop and go extraction) tips and then subjected to stable isotope dimethyl labeling on the Stage tips[38]. Matching light and heavy peptides were then combined and samples were finally subjected to LC–MS and subsequent data analyses using MaxQuant[56] and R, essentially as described[36].

Statistics. For the identification of raQTLs we tested the difference in SuRE expression of fragments containing the reference allele and the fragments containing the alternative allele using a two-sided Wilcoxon rank-sum test (see also section 'Identification of raQTLs'). In addition, the same two-sided Wilcoxon rank-sum test was applied once after random shuffling of fragments among the two alleles. Using all obtained *P* values from the real comparisons and the shuffled comparisons, we estimated the FDR. In Fig. 2a, a two-sided Fisher's exact test was performed, yielding $P < 2.2 \times 10^{-16}$ for each of the comparisons. The number of raQTLs overlapping with each of these states was CTCF: 667; enhancer: 1,052; promoter flanking: 74; repressed: 11,024; transcribed: 1,414; transcription start site: 2,189; weak enhancer: 576, while the number of all SNPs overlapping was CTCF: 79,437; enhancer: 44,296; promoter flanking: 2,178; repressed: 4,701,204; transcribed: 626,914; transcription start site: 46,631; weak enhancer: 17,665. For comparisons of alleles observed for DNA sequencing and DNase-seq, ChIP-seq or RNA-seq in Figs. 2e and 5f,i,j and Supplementary Figs 4e,f and 5b, *P* values were obtained using a one-side Fisher's exact test.

The *P* values in Fig. 3b,c were obtained using a two-sided Fisher's exact test. For the comparison in Fig. 5a, a two-sided Wilcoxon rank-sum test was done to compare the distances to the lead SNP for the raQTLs with the distances of the matched SNPs to the lead SNP.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw SuRE sequencing data are available at GEO (https://www.ncbi.nlm.nih.gov/geo/) under accession GSE128325. SuRE count tables, BigWig files for visualization of SuRE data tracks in genome browsers, lists of raQTLs and a table with SuRE data for all 5.9 million SNPs are available from the Open Science Framework (https://osf.io/w5bzq/wiki/home/?view). SuRE data can also be queried and visualized at https://sure.nki.nl. URLs to external data sources are listed in Supplementary Table 3.

Code availability

Scripts are available on https://github.com/vansteensellab/SuRE-SNV-code. Software used is described in the relevant methods section and in the Nature Research Reporting Summary.

References

47. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
48. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
49. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
50. Huber, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
51. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, e168 (2014).
52. de Vree, P. J. et al. Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nat. Biotechnol.* **32**, 1019–1025 (2014).
53. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
54. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
55. Dignam, J. D., Lebovitz, R. M. & Roeder, R. G. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res.* **11**, 1475–1489 (1983).
56. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

nature research

Corresponding author(s): Bas van Steensel, Joris van Arensbergen

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Paired end reads (PE reads) of the SuRE libraries (for associating genomic positions and barcodes for each SuRE-fragment), and single end reads (SE reads) of the PCR amplified barcodes (representing raw SuRE expression data), were processed to remove adapter and vector backbone sequences, using cutadapt (V1.9.1). The gDNA sequences of the SuRE libraries were mapped to the reference genome sequence (hg19, including only chr1-22, chrX), using bowtie2 (V2.3.2). To prevent allelic biases in alignment we used WASP. Full genome sequencing by Novogene of the K562 genome uses the GATK pipeline to call variants.
Data analysis	Data analysis and figure production was mostly done using various R (https://www.R-project.org) and BioConductor packages. Scripts are available on https://github.com/vansteensellab/SuRE-SNV-code

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequencing data are available at GEO accession GSE128325. SuRE count tables and a table with processed data for all 5.9 million SNPs is available from OSF https://osf.io/w5bzq/wiki/home/?view

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Complexities of the libraries were chosen to provide good coverage for each genome
Data exclusions	No data was excluded, except for some sequencing reads for samples that were too deeply sequenced. These were downsampled; see methods.
Replication	We looked at correlation of SuRE expression at promoters, assuming most will not be affected by variants, and found good correlation between individuals
Randomization	we do not control the classification of DNA fragments as REF or VAR
Blinding	no blinding was used since assignemtn as REF or VAR alleles is done computationally.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	lymphoblast cell lines HG02601, GM18983, HG01241 and HG03464 obtained from Coriell Institute, HepG2 (#HB 8065; ATCC), K562 (ATCC, #CCL-243)
Authentication	Only confirmed for lymphoblast cell lines. SuRE data provides genome-wide genome sequence.
Mycoplasma contamination	all were found negative during our regular (~ every 3 months) checks
Commonly misidentified lines (See ICLAC register)	none