Landscape of allele-specific transcription factor binding in the human genome

Sergey Abramov[1,2,3,*], Alexandr Boytsov[1,2,3,*], Dariia Bykova[4], Dmitry D. Penzar[1,2,3,4], Ivan Yevshin[5,6], Semyon K. Kolmykov[6,7], Marina V. Fridman[2], Alexander V. Favorov[2,8], Ilya E. Vorontsov[1,2], Eugene Baulin[3,9], Fedor Kolpakov[5,6], Vsevolod J. Makeev[2,3,10,11,+], Ivan V. Kulakovskiy[1,2,11,+]

1. Institute of Protein Research, Russian Academy of Sciences, Institutskaya 4, Pushchino, 142290, Russia
2. Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina 3, Moscow, GSP-1, 119991, Russia
3. Moscow Institute of Physics and Technology, Institutskiy per. 9, Dolgoprudny, 141700, Russia
4. Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Leninskiye gory 1-73, Moscow, 119234, Russia
5. BIOSOFT.RU LLC, Russkaya 41/1, Novosibirsk, 630090, Russia
6. Institute of Computational Technologies SB RAS, Lavrentieva 6, Novosibirsk, 630090, Russia
7. Institute of Cytology and Genetics SB RAS, Lavrentieva 10, Novosibirsk, 630090, Russia
8. Johns Hopkins University School of Medicine, 550 N Broadway, Baltimore, MD, 21205, USA
9. Institute of Mathematical Problems of Biology RAS - the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Vitkevicha 1, Pushchino,142290, Russia
10. State Research Institute of Genetics and Selection of Industrial Microorganisms of the National Research Center Kurchatov Institute, Pervy dorozgny proezd 1, Moscow, 117545, Russia
11. Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilova 32, Moscow, GSP-1, 119991, Russia

* equal contribution

\+ corresponding authors: vsevolod.makeev@vigg.ru, ivan.kulakovskiy@gmail.com

Keywords

Allele-specific binding, Human transcription factor, Single-nucleotide variant, SNV, Single-nucleotide polymorphism, SNP, Regulatory variant, ChIP-Seq, Allelic dosage

Abstract

Sequence variants in gene regulatory regions alter gene expression and contribute to phenotypes of individual cells and the whole organism, including disease susceptibility and progression. Single-nucleotide variants in enhancers or promoters may affect gene transcription by altering transcription factor binding sites. Differential transcription factor binding in heterozygous genomic loci provides a natural source of information on such regulatory variants. We present a novel approach to call the allele-specific transcription factor binding events at single-nucleotide variants in ChIP-Seq data, taking into account the joint contribution of aneuploidy and local copy number variation, that is estimated directly from variant calls. We have conducted a meta-analysis of more than 7 thousand ChIP-Seq experiments and assembled the database of allele-specific binding events listing more than half a million entries at nearly 270 thousand single-nucleotide polymorphisms for several hundred human transcription factors and cell types. These polymorphisms are enriched for associations with pathologies and often act as eQTLs, revealing molecular mechanisms and causality of the associations. Specifically, there is a special class of switching sites, where different transcription factors preferably bind alternative alleles, thus providing allele-specific molecular machinery for the target gene regulation.

Introduction

Sequence variants located in non-coding genome regions attract an increasing researchers' attention due to the frequent association with various traits, including predisposition to diseases[1,2]. Single-nucleotide variants (SNVs) in gene regulatory regions may affect gene expression[3] by altering binding sites of transcription factors (TFs) in gene promoters and enhancers and, consequently, efficiency of transcription[4].

On the one hand, parallel reporter assays allow massive assessment of variants in terms of gene expression alteration[5,6] but do not reveal particular TFs involved. On the other hand, there are multiple ways to assess if a single nucleotide substitution changes transcription factor binding affinity, from detailed measurements of the TF affinity landscape *in vitro*[7,8] to conventional experiments on individual sequence variants[9,10] and computational modeling[11–13]. However, it is not trivial to utilize these data for annotating SNV effects at the genome-wide scale in a cell-type-specific manner.

The functional effect of single nucleotide substitutions can be studied in heterozygous chromosome loci, where TFs differentially bind to sites in homologous chromosomes with alternative SNV alleles. Reliable evidence comes from modern *in vivo* methods based on chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq). ChIP-Seq provides a deep read coverage of TF binding regions, and non-perfect alignments of reads often carry single-nucleotide mismatches arising from heterozygous sites. Statistical biases between the numbers of mapped reads containing alternative SNV alleles reveal the so-called allele-specific binding events[1,14] (ASB, **Fig. 1A**).

Chromatin accessibility often serves as a proxy for the regulatory activity of a genomic region[15]. Massive assessment of allele-specific chromatin accessibility in more than 100 cell types[16] reported more than 60 thousands of significantly imbalanced sites. Yet, so far, only 10 to 20 thousand ASBs were reported per study (**Supplementary Table S1**), and the potentially vast landscape of allele-specific TF binding remains mostly unexplored.

Reliable identification of ASBs (the ASB calling) requires high read coverage at potential sites, which result either from deep sequencing of individual ChIP-Seq libraries or aggregating data across multiple experiments. Reprocessed ChIP-Seq data for hundreds of TFs and cell types are available in databases such as GTRD[17] and Remap[18],

opening a way to an integrative meta-analysis, which could yield raw statistical power to detect cell type- and TF-specific ASBs.

Straightforward meta-analysis of the ASBs has two major limitations. First, many ChIP-Seq data sets are obtained in aneuploid cell lines, and copy-number variants (CNVs) are common even for normal diploid cells. Both the chromosome multiplication and local CNVs affect the expected read coverage of the respective genomic regions[19] and bring about imbalanced read counts at SNVs, possibly generating false positive ASB calls (**Fig. 1A**). There exist strategies to reduce this bias (**Supplementary Table S2**), in particular, the known CNV regions can be filtered out[20] or predicted from a computational analysis of the corresponding genomic DNA[21,22] (which is often used as the ChIP-Seq control sample) and incorporated in statistical criteria when evaluating the potential ASB calls[19]. However, in many published experiments, the input DNA data control was omitted in favor of other controls, such as preimmune IgG, or had a limited sequencing depth making it useless for CNV predictions. Furthermore, currently, there are no systematic data on global (chromosome duplications) and local (CNVs) structural variations across all cell types with public ChIP-Seq data on TFs. Even when the external data on structural variation are available for particular cells, it is not guaranteed that the same estimates would be valid for ChIP-Seq data obtained elsewhere, since long-cultivated immortalized cell lines might keep accumulating unreported differences in genome dosage across chromosomes[23].

The other major problem in ASB calling is the so-called reference read mapping bias[21,24]. Standard read alignment tools generally map more reads to the alleles present in the reference genome assembly, as such mapping has lower or no mismatch penalties. To account for the reference read mapping bias, an ideal scenario involves mapping to individually reconstructed genomes[22,25] or computational simulations[20] that provide estimates of mapping probabilities to alternative alleles separately for each SNV (see **Supplementary Table S2** for an overview). Yet, these solutions are not applicable to pre-made read alignments (which are usually obtained with a simple reference genome) and hardly applicable to understudied cell types or particular samples that do not provide enough data to reconstruct an individual genome.

In this work, we present a novel framework for ASB calling from existing read alignments or pre-made variant calls, accounting for the allelic dosage of aneuploidy and CNVs, and read mapping bias. With this framework, we have performed a comprehensive meta-analysis to identify ASBs in the human ChIP-Seq data from the GTRD database[17]. The ADASTRA database (Allelic Dosage-corrected Allele-Specific human

Transcription factor binding sites, http://adastra.autosome.ru) provides ASB events across 674 human TFs (including epigenetic factors) and 337 cell types. We demonstrate that the single-nucleotide polymorphisms (SNPs) with ASBs often act as eQTLs and exhibit associations with various normal and pathogenic traits. A comparison of data for multiple TFs highlights the cases where different TFs preferentially bind to different alleles, i.e., when a single nucleotide substitution can change an entry point of the involved regulatory pathway. Finally, we discuss selected cases where the allele-specific binding at SNPs reveals molecular mechanisms of associations between SNPs and important medical phenotypes.

Results

We present a reproducible workflow for ASB calling and meta-analysis across human TFs and cell types (**Fig. 1B**). First, the variants are called from pre-made ChIP-Seq read alignments against the hg38 genome assembly. Next, the variant calls are filtered by excluding homozygous and low-covered variants, as well as variants absent from the dbSNP[26] common subset (as putative de novo point mutations). The filtered single-nucleotide variants from related ChIP-Seq data sets (sharing the cell type and particular wet lab) are used to identify the cell type features (aneuploidy and CNVs). A total set of variants is used to assess the global read mapping bias that is used as the basis for statistical model parameterization. Finally, ASB calling is performed separately for each ChIP-Seq experiment, and the resulting allele read bias P-values are aggregated using the Mudholkar-George's method[27] for each SNV, either at the TF-level (across ChIP-Seq data for a selected TF from all cell types) or the cell type-level (across ChIP-Seq data for a selected cell type for all TFs).

We used the workflow to process 7669 ChIP-Seq read alignments from GTRD covering 1025 human TFs and 566 cell types, and detected more than 2 hundred thousand ASBs at more than 2 hundred thousand of SNPs for various TFs and 3 hundred thousand ASBs for cell types passing an adjusted P-value of 0.05, see **Fig. 1C,D** for an overview. Reaching these numbers has become possible because of the large volume of the starting data (the filtered list of considered variant calls contained more than 54 million entries) and the advanced statistical framework that we describe below. An overview of the processed data sets and variant calls per transcription factor and cell type are shown in **Supplementary Fig. S1**.

Estimating aneuploidy and copy-number variation from single-nucleotide variant calls

Allele-specific binding is assessed against expected relative frequencies of reads supporting alternative alleles of a particular SNV in a particular genomic region. Assuming there were no read mapping bias, these expected frequencies would be mostly determined by the copy number of the respective genomic segments. In this study, we estimated the joint effect of local copy-number variation and global chromosome ploidy from the read counts at single-nucleotide variant calls, taking into account that the background for ASB calling is defined by the expected relative frequencies of the read counts supporting alternative alleles rather than by absolute allelic copy numbers.

Background Allelic Dosage

We introduce the *Background Allelic Dosage* (BAD) as the ratio of the major to minor allele dosage in the particular genomic segment, which depends on chromosome structural variants and aneuploidy. BAD can be estimated from the number of reads mapped at each allelic variant and does not require haplotype phasing. For example, if a particular genomic region has the same copy number of both alleles, e.g., 1:1 (diploid), 2:2, or 3:3, then it has BAD=1, i.e., the expected ratio of reads mapped to alternative alleles on a heterozygous SNV is 1. All triploid regions have BAD=2, and the expected allelic reads ratio is either 2 or ½. In general, if BAD of a particular region is known, then the expected frequencies of reads supporting alternative alleles are 1/(BAD+1) and BAD/(BAD+1).

Importantly, accounting for BAD provides an answer to the question of the necessity of overdispersion in the statistical evaluation of ASBs[19,22]. In fact, a large portion of overdispersion of read counts disappears once the variant calls are segregated according to BADs of the respective genomic segments (see Methods).

BAD calling with Bayesian changepoint identification

In this study, we present a novel method for reconstructing a genome-wide BAD map of a given cell type. The idea is to find genomic regions with approximately stable BAD using the read counts at single-nucleotide variant calls. Assuming that both differential chromatin accessibility and sequence-specific TF binding affect only a minor fraction of

variants, the read counts for most of the SNVs must be close to equilibrium and thus provide imprecise but multiple measurements of the background allelic dosage.

We have developed a Bayesian changepoint identification algorithm, which (1) segments the genomic sequence into regions of the constant BAD using dynamic programming to maximize the marginal likelihood and then (2) assigns the BAD with the maximal posterior to each segment (see Methods). An additional preprocessing employs distances between neighboring SNVs to exclude long deletions and centromeric regions from BAD estimation. The BAD caller in action is illustrated in **Fig. 2A** for two chromosomes using ENCODE K562 data (see the segmentation map of the complete genome with multiple deletions in **Supplementary Fig. S2**).

We performed the BAD calling for 2556 groups of variant calls, where each group consisted of calls obtained from ChIP-Seq alignments for a particular cell type and GEO series or ENCODE biosample ID (i.e., for K562 cells of different studies, the BAD calling was performed independently). In BAD calling, recurrent SNVs sharing dbSNP IDs and found in different datasets within the same group were considered as independent observations. To systematically assess the reliability of the resulting BAD maps, we compared the predicted BADs at all SNVs with the ground truth BADs estimated from COSMIC[28] CNV data for 76 matched cell types, with K562 and MCF7 being the most represented. For K562 and multiple other cell types, the Kendall τ_b rank correlation was consistently better for joint data sets with higher numbers of SNVs (**Fig. 2B**), which justifies the usage of read counts at SNVs as point measurements of BAD. However, BAD maps for MCF7 demonstrated poor agreement with COSMIC, independently of the number of SNVs in the dataset. We believe this is caused by the reported high genome instability of MCF7[29] that results in a strongly variable CNV pattern, varying chromosome counts, and, consequently, unstable BAD estimates in cells from different studies. Similar results were obtained in an additional comparison of BAD maps against independent microarray-based CNV estimates for major cell types[29], including 13 cell types matching across these data, COSMIC, and our study (**Supplementary Fig. S3A**).

As an alternative test, we used the predicted BAD maps as multiple binary classifiers for different BAD values. With the COSMIC data as the ground truth, we plotted a receiver operating characteristic (ROC) and a precision-recall curve (PRC) for each BAD (**Fig. 2C,D**). For the most widespread BADs (1, 2, and 3) covering more than 90% of candidate SNVs (**Supplementary Fig. S3B**), we reached >0.83 area-under-curve for ROC and 0.66-85 for PRC (**Supplementary Table S3**), proving the reliability of the predicted BAD maps.

ASB calling with the negative binomial mixture accounts for mapping bias

With BAD maps at hand, we segregated the variant calls from all datasets by BAD and by fixed read coverage either at reference or alternative alleles. Then, for each such set of SNVs, we fitted the background distribution as a mixture of two negative binomial distributions with BAD-determined *p* parameters (see Methods). ASBs were called independently for the reference (Ref-ASB) and the alternative (Alt-ASB) allele using separately fit background distributions for the fixed read counts at alternative and reference alleles, respectively, thus accounting for general read mapping bias.

Overview of the ADASTRA database

The results of the ASB calling are provided in the ADASTRA database (the database of Allelic Dosage-corrected Allele-Specific human Transcription factor binding sites). In ADASTRA, each dbSNP ID can have several ASB entries for different TFs or cell types. ADASTRA consists of two parts: the first part (TF-ASB, 233290 ASBs at 147909 SNPs) contains ASB obtained by aggregation of individual P-values for each TF over cell types. The listed ASBs passed multiple testing correction ($P < 0.05$ after Benjamini-Hochberg adjustment for the number of tested ASBs. P-value estimation (see below), aggregation, and multiple testing correction were performed separately for ASBs with preferred binding to the reference (Ref-ASB) and alternative (Alt-ASB) alleles, and for each TF. The other part of the database (CL-ASB, 351967 ASBs at 252469 SNPs) contains a similar aggregation of individual ASBs over TFs for each cell type.

TFs and cell types were unequally represented in the source data. Thus, the numbers of the resulting ASB calls were also biased towards most studied cell types and TFs (**Fig. 3A-B**), with the top contributions from CTCF for TFs and K562 for cell types. However, the top 10 TFs and top 5 cell types covered only half of ASB calls (for cell types) or less than a half of ASB calls (for TFs); thus, the produced data on ASB events is diverse across different samples.

Next, we assessed how ASBs and candidate SNVs are distributed in different genomic regions (**Fig. 3C**). Compared to all SNVs and tested candidate ASB sites, the significant ASBs were enriched in enhancers (~4x more than expected from the number of SNVs for which there were candidate ASBs, Fisher's exact test $P < 10^{-300}$) and promoters (~3x more than expected, $P < 10^{-300}$). We consider this observation consistent with both the actual location of functional transcription factor binding sites and deeper

coverage of the actual TF binding regions with ChIP-Seq reads. In fact, ASBs are likely to cluster at the scale of the typical ChIP-Seq peak width, as revealed by the distribution of pairwise distances between SNVs with and without ASBs, which has a bimodal shape (**Supplementary Fig. S4**).

We also compared the SNPs listed in ADASTRA with those of the previous ASB collections (**Supplementary Fig. S5A**). ADASTRA includes ASBs at 38%, 44%, and 63% of dbSNP SNPs reported as ASBs in AlleleDB[22], and collections published in[30] and[20], respectively. There is also a notable overlap between ADASTRA ASBs and sites of allele-specific DNA accessibility[16], as well as between ASBs and reporter assay quantitative trait loci (raQTLs)[6] (**Supplementary Fig. S5B-D**). Thus, in general, there is an overlap between ADASTRA ASBs and the existing data on regulatory SNPs, but the vast majority of ADASTRA data are novel.

Given the diversity of assessed TFs, it became possible to systematically compare SNVs carrying TF-level ASBs and identify the pairs of TFs sharing ASBs with the one-tailed Fisher's exact test (**Supplementary Fig. S6**). We did not observe any preference of shared ASBs for interacting TFs (checking the known protein-protein interactions from STRING-db[31]). Yet, some interacting proteins (such as CTCF-RAD21) often share ASBs, and the same holds for particular composite elements of binding sites such as AR-FOXA1[32]. Also, there is a major overlap between ASBs for chromatin-interacting epigenetic factors and related proteins, suggesting many of these events to be 'passengers' in regions of allele-specific chromatin accessibility with TFs bound only to the accessible chromosome.

Motif annotation is concordant with ASB calls

For transcription factors specifically interacting with DNA, it is possible to perform computational annotation of ASBs with TF-recognized sequence motifs[33]. When a strong binding site overlaps an ASB and the alternating alleles significantly change the motif prediction score, this ASB is likely to be a 'driver' event, that directly produces the asymmetry in the ChIP-Seq read counts from the different affinity of the TF to the alternative binding sites at homologous chromosomes, rather than from the chromosome-specific local chromatin accessibility, the case of a 'passenger' ASB. Furthermore, motif annotation allows to compare in a systematic way the actual observed ASB effect (the read counts) and the effect predicted by sequence analysis (the motif specificity), providing an independent evaluation of the reliability of ASB calls.

An ASB was considered as overlapping the TF motif occurrence if the TF position weight matrix (PWM) scored a hit with $P \leq 0.0005$ for any of the two alleles. The log ratio of P-values corresponding to PWM hits at alternative alleles was used as an approximation of the TF affinity fold change. **Fig. 4A** compares the ASB significance (X-axis, signed $\log_{10}$ FDR; the sign set positive for Alt-ASBs and negative for Ref-ASBs) with the log ratio of motif hits (Y-axis) for 218 TFs having at least 1 ASB within a motif hit. Predominantly, at heterozygous sites, alleles with more specific motif hits are covered with more ChIP-Seq reads, revealing the prevalence of motif-concordant ASB events (blue dots in **Fig. 4A**). Such concordance persists for more than 80% of SNVs with ASB allelic imbalance FDR < 5%, growing with decreasing ASB FDR and saturating at about 90% of SNVs (**Fig. 4B**). At 5% FDR, good motif concordance stands for many TFs, as illustrated by the top 10 TFs with the highest number of motif hits at ASBs (**Fig. 4C**). Importantly, even at larger FDR, there are more concordant than discordant ASBs.

Yet, for ~10-20% of SNVs, the motif hit odds-ratios (corrected for BAD) are discordant with the ASB disbalance, that is, more reads are attracted to the weaker motif hit (red dots in **Fig. 4A**, red bars in **Fig. 4B**). We believe that in such cases there are other contributors (allele-specific chromatin accessibility or indirect TF binding) to allele disbalance, thus overriding the contribution of the motif.

To quantify ASB allelic imbalance for BAD other than one, we defined the ASB effect size (ES) as follows (see Methods for details). For individual SNV (SNV in a single dataset):

$$ES_{\mathrm{Ref}} = \log_2(C_{\mathrm{Ref}} / E\,(C_{\mathrm{Ref}} \mid C_{\mathrm{Alt}})) \text{ and } ES_{\mathrm{Alt}} = \log_2(C_{\mathrm{Alt}} / E\,(C_{\mathrm{Alt}} \mid C_{\mathrm{Ref}})).$$

Here C_{Ref} and C_{Alt} are the read counts at the Ref and Alt alleles, and E is the expectation. For BAD = 1: $ES_{\mathrm{Ref}} \approx \log_2(C_{\mathrm{Ref}} / C_{\mathrm{Alt}})$.

The aggregated effect size of an ASB is calculated as a weighted mean of effect size values for the same allele for SNVs aggregated at the same genome position over TFs or cell types, with weights equal to negative logarithms of individual P-values, separately for each of the alleles.

This allows creating *staveplots* that compare the ASB effect size with minor-to-major allele changes in different sequence motif positions for all ASBs within significant motif hits. As an illustrative example (**Fig. 4D**), we show CEBPB ASBs on top of the CEBPB motif. Here the base is encoded with a color, which is the same in the motif logo diagram and in the stave graph, where Y-axis shows the ASB effect size. The most conserved motif positions 3-7-9-10 are almost unicolor, with the major allele being the same as the consensus letter in the motif. Less conserved positions allow for more

options, with position 6 being of special interest: it displays frequent T/C ASBs with C being the major allele. These cytosines belong to the core CG pair which is prone to spontaneous deamination. The produced mismatches are then protected from repair through enhanced CEBPB binding resulting in mutation fixation[34]. Such ASBs, on the one hand, confirm frequent mutagenesis of CEBPB binding sites, and, on the other hand, suggest the action of purifying selection that stabilizes such sites as heterozygous variants. The staveplots for other TFs are shown in **Supplementary Fig. S7**.

Machine learning predicts ASBs from sequence analysis and chromatin accessibility

With previously published ASB sets of smaller volumes, it was possible to predict allele-specific binding from chromatin properties and a sequence analysis[20]. To assess to what degree this holds for ADASTRA data, we applied machine learning with a random forest model[35] atop experimentally determined allele-specific chromatin DNase accessibility data[16], predicted allele-specific chromatin profile from DeepSEA[11], and sequence motif hits (**Supplementary Table S4**).

We were considering two binary classification problems: (1) general assessment, i.e., to predict if an SNV makes the ASB for any of the TFs or in any of the cell types, and (2) TF- and cell type-specific assessment, i.e., to predict if an SNV makes the ASB for the particular TF or in the particular cell type. Models for both problems were trained and validated using multiple single-chromosome hold-outs: iteratively for each of 22 autosomes, one autosome was selected for validation, and 21 other autosomes were used for training. At each iteration, the model performance was estimated at the held-out autosome, and the resulting receiver operating characteristics (ROC) and precision-recall curves (PRC) were averaged.

For the first problem, the performance at TF- and cell type-ASBs was 0.74 and 0.73 for the area under the receiver operating characteristic (auROC), and 0.44 and 0.56 for the area under the precision-recall curve (auPRC), respectively (see the plots in **Supplementary Fig. S8**). For the second problem, we used the top 10 TFs and top 10 cell types with the highest numbers of ASBs (**Supplementary Table S5**), and a dedicated model was trained for each TF and each cell type. The quality of the models (**Supplementary Table S5**) was different for different TFs and cell types, with the highest auROC of 0.72 and 0.81 for CTCF (of TFs) and HepG2 (of cell types), and the highest auPRC of 0.35 and 0.64 for CTCF and A549. Of note, RAD21 ASBs were also

predicted with very high reliability, as they are often located at the same variants as CTCF ASB.

We analyzed the feature importance and found that DeepSEA-derived features played an important role in the model scores, especially for the cell type-specific models, where features predicted for the matched cell types were automatically prioritized. In agreement with previous studies[16,20], all models received a significant contribution from allele-specific DNase chromatin accessibility data, and, for TFs, from the motif-based features, although there was no single dominant feature.

Disease-associated SNPs and eQTLs are enriched with ASBs

To assess if ASB facilitates the identification of functional regulatory sequence alterations, we annotated the ASB-carrying SNVs using data from several databases on phenotype-genotype associations: NHGRI-EBI GWAS catalog[36], ClinVar[37], PheWAS[38], and BROAD fine-mapping catalog of causal autoimmune disease variants[39]. We found a significant overlap between ASB-carrying SNPs and phenotype-associated SNPs, and the overlap significance increases with better ASB FDR or higher ASB effect size thresholds (**Supplementary Fig. S9**).

Next **(Fig. 5A)**, we merged the data of phenotype-genotype association databases and counted the number of known associations per SNP, considering SNVs of several classes: low-covered SNVs not tested for ASB (non-candidate sites); candidate sites that exhibit or not exhibit ASB from the datasets of a single TF; candidate sites from the datasets for two or more TFs that, again, exhibit ASB or do not; and finally, regulator switching ASBs, where different TFs prefer to bind alternative alleles, e.g., in different cell types. All variants were segregated into classes in regard to known associations: no known associations, with a single association, and with multiple associations.

We have found that the share of ASB variants with genetic associations was consistently higher than expected by chance (**Fig. 5A**), which apparently makes ASBs good candidates for prospection for causal SNVs. Specifically, the odds ratio between the observed and expected SNP numbers was specifically high for TF-switching ASBs, although only 1.5% of such ASBs were involved in two or more known GWAS associations. For many variants, there are no known associations with 'macro-phenotypes', as provided by GWAS studies, but there are data on molecular phenotypes like variations in mRNA levels. In fact, the effect of the so-called expression quantitative trait loci (eQTLs)[40] can be explained by alteration of TF binding affinity that is revealed by ASB. Using the same classification of SNVs as above, we tested ASB and

non-ASB SNVs for overlaps with GTEx[41] eQTLs and observed the same pattern as for phenotype associations, with the strongest enrichment of ASBs for which different TFs preferably bind alternative alleles **(Fig. 5B)**. The enrichment also grew stronger with the number of genes, mRNA levels of which were associated with the variant.

More than 80% of ASB SNVs with alternative alleles preferably bound by different TFs were found to be eQTLs in at least one cell type, whereas 10% of such ASB SNVs were eQTLs targeting 10 or more genes. A large fraction of genes associated with ASB eQTLs have been found among phenotype-associated genes from the ClinVar catalog[37] (3-fold enrichment as compared to random expectation, Fisher's exact test P ~0.0). It should be noted that as many as 2/3 of all phenotype-associated genes from ClinVar are eQTL target genes of ASB SNVs, and this constitutes 45% of all ASB-driven eQTL target genes.

We also studied the association of GWAS-tested phenotypes with all candidate SNVs, not necessarily significant ASBs, found in TF binding regions. To this end, we performed a general enrichment analysis for SNPs found in ChIP-Seq data of particular TFs within linkage disequilibrium blocks (LD-islands identified in [42]) using Fisher's exact test (see Methods). This way we identified TFs for which phenotype-associated SNVs were enriched within TF binding regions (**Fig. 5C,D**). For a number of TFs such association with phenotypes was reported in other studies. The examples include FOXA1 (involved in prostate development[43] and in our case, found associated with prostate cancer), IKZF1 (for which the protein damaging mutations are associated with leukemia), STAT1 (involved in the development of systemic lupus erythematosus[44]), and others. Practically in all cases one of the associated SNVs also acted as ASB of the respective TF, making ASBs strong candidates for the causal variants.

To illustrate how the functional role of regulatory SNPs can be highlighted with ASB data, we present several case studies. First, there is rs3761376 (G > A) that serves as a Ref-ASB for ESR1, which was already confirmed by electrophoretic mobility shift assay[45]. rs3761376 is located in the TFF1 gene promoter and was shown to reduce TFF1 expression through altered ESR1 binding, suggesting a molecular mechanism of the increased risk of gastric cancer [45].

Next, there is rs17293632 (C > T) that serves as a Ref-ASB for 25 different TFs and was previously reported to affect the chromatin accessibility in the adjacent region[46]. rs17293632 is associated with Crohn's disease. This SNP is located in SMAD3 intron and serves as an eQTL for SMAD3, AAGAB, and PIAS1 genes[41]. Interestingly, a variant of SMAD3 is also associated with Crohn's disease, particularly, with increased risk of

repeated surgery and shorter relapse[47]. Among the TFs displaying ASBs, there are JUN/FOS proteins with the ASB-concordant motif annotation. JUN/FOS form the AP1 pioneer complex [doi:10.1016/j.biochi.2003.09.006] that likely serves as a 'driver' for changes both in gene expression and chromatin accessibility, and is likely to cause ASB of all 25 TFs.

Apart from multi-TF ASBs which are linked to local chromatin changes, non-trivial cases can be found among TF-switching ASBs, e.g., rs28372852 located in the G elongation factor mitochondrial 1 (GFM1) gene promoter. According to ClinVar[37], this SNP is associated with combined oxidative phosphorylation deficiency. According to ADASTRA, rs28372852 serves as the Alt-ASB of CREB1 and Ref-ASB of MXI1, and in both cases, the allelic imbalance is concordant with the respective binding motifs. Also, GFM1 is the eQTL target for rs28372852. According to UniProt[48], CREB1 is a transcriptional activator, while MXI1 is a transcriptional repressor, suggesting that ASB can directly switch the gene expression activity. At the same time, UniProt reports four amino acid substitutions in GFM1 that are associated with combined oxidative phosphorylation deficiency. Thus, there is an example of loss-of-function by two different mechanisms, either directly through amino-acid substitutions, or due to altered gene expression caused by nucleotide substitution in the gene regulatory region.

Discussion

The functional annotation of non-coding variants remains a challenge in modern human genetics. Phenotype-associated SNPs found in GWAS are usually located in extensive linkage disequilibrium blocks, and reliable selection of causal variants cannot be done purely by statistical means. Additional data for the identification of causal variants come from functional genomics. In particular, an important class of causal variants consists of regulatory SNVs affecting gene transcription. For those variants, there are various approaches, e.g., parallel reporter assays, to obtain high-throughput data on molecular events caused by particular nucleotide substitution. Another common strategy is to check if a variant of interest falls into a known gene regulatory region detected by chromatin immunoprecipitation or chromatin accessibility assay followed by deep sequencing. By assessing the allele specificity, it is possible to further profit from these data through direct estimation of the effect that a particular allele has on the binding of relevant regulatory proteins or chromatin accessibility.

In this meta-study, for each SNV, we integrated the data by considering a TF bound to SNV in different cell types or a cell type and different TFs bound to the same SNV. Surprisingly, ASB identification through data aggregation had better sensitivity than standard ChIP-Seq peak calling at the level of individual data sets. Particularly, in GTRD, the ChIP-seq peak calls were gathered from 4 different tools (MACS, SISSRs, GEM, and PICS), but only 50-70% of significant ASBs were detected within peak calls (129378 of 233290 and 231445 of 351965 for TF-centric and cell type-centric aggregation), suggesting that up to half of ASBs could be lost if the ASB calling is restricted to the peak calls only.

Each particular ASB can either be a 'driver' directly altering TF binding affinity, or a 'passenger' with differential binding resulting from differential chromatin accessibility (in turn, caused by some neighboring SNVs), or a protein-protein interaction with the causal TF. In terms of machine learning, we expected the TF-ASBs to provide an easier prediction target since they could be mostly determined by the sequence motif of the respective TF. However, as found, the percentage of 'passenger' ASBs is rather large (e.g., 24662 out of 27233 CTCF ASBs lack significant CTCF motif hits), and the TF-specific models showed a limited ASB prediction quality. Further surprise came from cell type-specific models which displayed a notably higher performance. We interpret these data as follows: the cell-type ASBs are easier to predict by learning a small set of cell type-specific master regulators, while passenger TF-level ASBs are very diverse, as coming from data aggregation of many different cell types with varying cell type-specific features such as key TFs.

A general hindrance for ASB detection is the gene and chromosome duplications, which imitate ASB by varying the allelic dosage. The presented pipeline, to our knowledge, is the first control-free approach to reconstruct the genome map of background allelic dosage and to use it as a baseline for detecting allelic imbalance. Further on, such a pipeline might be applicable to other sequencing data that allow allele-specificity, e.g., analyses of allele-specific expression or chromatin accessibility. With matched cell types, BAD-corrected data on allele-specific chromatin accessibility will also allow for better classification of driver and passenger ASBs and better application of machine learning techniques.

Our collection of ASB events per se is also useful for other research areas involving TF-DNA interactions. First, ASBs provide unique in vivo data on differential TF binding and can be used for testing the predictive power of computational models for precise recognition of transcription factor binding sites[33]. Second, the transcription

factor binding not only affects transcript abundance, but also affects RNA splicing, localization, and stability[49,50]. Thus, ASBs may affect other levels of gene expression, particularly, the mRNA post-transcriptional modification: out of 65 RNAe-QTLs reported in[51], 4 are listed as ASBs in ADASTRA.

Last but not least, ADASTRA reports hundreds of TF-switching ASBs, where alternative alleles are preferably bound by different TFs. This possibility has been discussed previously[52] but, to our knowledge, we are first to report the genome-wide inventory of such events. Importantly, the respective SNVs exhibit the highest enrichment with phenotype associations. Probably these sites serve varying and allele-dependent molecular circuits. Further analysis of TF-switching ASBs in the scope of metabolic and regulatory pathway alterations will provide valuable insights into molecular mechanisms underlying particular normal and pathogenic traits.

Methods

Variant calling from GTRD alignments

We used 7669 pre-made short read alignments against hg38 genome assembly produced with bowtie2[53] and stored in the GTRD[17] database. PICARD was used for deduplication, followed by GATK base quality recalibration. Next, the variants were called with GATK Haplotype Caller, with dbSNP[26] (common variant set of the build 151) for annotation. The resulting variant calls were filtered to meet the following requirements: (1) an SNV must be biallelic and heterozygous (GATK annotation GT=0/1); (2) an SNV must have read coverage ≥ 5 at both the reference and alternative alleles; (3) an SNV must be listed as an SNP in the dbSNP 151 common set. Of note, we considered all eligible SNVs as candidate ASB, not necessarily located within ChIP-Seq peak calls.

We restricted ourselves with variants from the dbSNP common subset due to the following reasons: (1) allelic read counts at de novo mutations reflect the composition of the cell population (i.e. the fraction of cells carrying the mutation) rather than the local copy number ratio or allele-specific binding; (2) de novo point mutations within particular copies of duplicated segments (considering e.g. chromosome duplications) will exhibit allelic imbalance (e.g. in a tetraploid region with 2:2 ratio of allelic reads at SNPs, de novo mutations will likely exhibit the ratio of 1:3) and may lead to false-positive ASB calls.

Accounting for BAD

The observed distribution of ChIP-Seq allelic read counts on heterozygous SNVs significantly depends on aneuploidy and the CNV-profile of the cells (**Fig. 6A,B**). The modes of distribution correspond to the most represented copy number, e.g., the distribution is bimodal for mostly triploid K562 cells, **Fig. 6B**. However, the mixture of two Binomial distributions poorly approximates the data, showing a significant overdispersion. To systematically reduce the overdispersion from local CNVs and aneuploidy, we reconstructed the genome-wide background allelic dosage (BAD) maps from read counts at the heterozygous variants (see below). The distributions of the allelic read counts at SNVs segregated by BAD show a notably reduced overdispersion (**Fig. 6C, D**).

BAD calling with Bayesian changepoint identification

To construct genome-wide BAD maps from filtered heterozygous SNV calls, we developed a novel algorithm, the BAD caller by Bayesian changepoint identification (BABACHI).

At the first stage, BABACHI divides the chromosomes into smaller sub-chromosome regions by detecting centromeric regions, long deletions, loss of heterozygosity regions, and other regions depleted of SNVs. At this stage, only the distances between neighboring SNVs are taken into account and long gaps are marked. The sub-chromosome regions with less than 3 SNVs or chromosomes with less than 100 SNVs are removed. Next, BABACHI finds a set of changepoints in each sub-chromosome region that further divide it into smaller segments of stable BAD. The optimal changepoints are chosen to maximize the marginal likelihood to observe the experimental distribution of allelic read counts at the SNVs, given a region-specific (yet unknown) BAD persist in each region enclosed between neighboring changepoints. Finally, a particular BAD is assigned to each segment according to the maximum posterior.

The likelihood is calculated for the statistic $x = \min(C_{\mathrm{Ref}}, C_{\mathrm{Alt}})$, assuming C_{Ref} to be distributed according to the truncated Binomial distribution $\sim\mathrm{TruncatedBinom}(n, p)$ given that $C_{\mathrm{Ref}} + C_{\mathrm{Alt}} = n$, the number of reads overlapping the variant; the number of successes k is limited to $5 \le k \le n\text{-}5$ (the read coverage filter), and p is either $1/(\mathrm{BAD}+1)$ or $\mathrm{BAD}/(\mathrm{BAD}+1)$, matching one of the expected allelic read frequencies.

BAD of each segment is selected from the discrete set {1, 4/3, 3/2, 2, 5/2, 3, 4, 5, 6}, considering that the total copy number of a particular genomic region rarely exceeds 7. The prior distribution of BAD is assumed to be a discrete uniform, with the support being the same discrete set as above (non-informative prior). Details and mathematical substantiation of the algorithm are provided in the **Supplementary Methods**.

Practical BAD calling with the ADASTRA pipeline

To provide better genome coverage and robust BAD estimates, we merged the sets of variant calls from ChIP-Seq datasets produced in the same laboratory for the same cell type and in the same series (i.e., sharing either ENCODE biosample or GEO GSE ID). Different SNVs at the same genome position (either originating from different datasets or with different alternative alleles) were considered as independent observations. For each dataset, chromosomes with less than 100 SNVs were excluded from BAD calling and further analysis.

To assess the reliability of the BAD maps, for each BAD, we separately estimated Receiver Operating Characteristics and Precision-Recall curves. Here we considered the BAD maps as binary classifiers of SNVs according to BAD, with COSMIC CNV data as the ground truth. To plot a curve for $\mathrm{BAD}=x$, the following prediction score was used:

$S = L(\mathrm{BAD}=x) - \max_{y \neq x} L(\mathrm{BAD}=y)$, where L denotes the log-likelihood of the segment containing the SNV to have the specified BAD (**Fig. 2C, D**).

ASB calling with the Negative Binomial Mixture Model

To account for mapping bias, we fitted separate Negative Binomial Mixture Models for the scoring of Ref- and Alt-ASBs. For each BAD and each fixed read count at Ref- and Alt-alleles, we obtained separate fits using SNVs from all available datasets.

For every fixed read count value at a particular allele, we approximated the distribution of read counts mapped to the other allele as a mixture of two Negative Binomial distributions. The model estimates the number of successes x (the number of reads mapped to the selected allele) given the number of failures r (the number of reads mapped to the second allele) in the series of Bernoulli trials with probability of success p (for the first distribution in the mixture) or $1\text{-}p$ (for the second distribution in the mixture). The following holds for scoring Ref-ASBs at fixed Alt-allele read counts:

$$C_{Ref} \mid \text{fixed } C_{Alt} \sim w \cdot \mathrm{NegativeBinomial}(r, p) + (1 - w) \cdot \mathrm{NegativeBinomial}(r, 1\text{-}p)$$

$$P(C_{Ref} = x \mid \text{fixed } C_{Alt} = m, C_{Ref} \geq 5) = \binom{x + r - 1}{x} (w \cdot (1\text{-}p)^r \cdot p^x + (1\text{-}w) \cdot (1\text{-}p)^x \cdot p^r) / A$$

$$A = 1 - P(C_{Ref} < 5 \mid \text{fixed } C_{Alt} = m)$$

where p and $1\text{-}p$ were fixed to reflect the expected frequencies of allelic reads, namely, $1/(\text{BAD}+1)$ and $\text{BAD}/(\text{BAD}+1)$. The parameters r (number of failures) and w (weights of distributions in the mixture) were fitted with L-BFGS-B algorithm from *scipy.optimize*[54] package to maximize the model likelihood iteratively with boundaries $r > 0$ and $0 \le w \le 1$, assigning initial values of $r=m$ (number of reads on the fixed allele) and $w=0.5$, respectively. A is the normalization coefficient (necessary due to truncation) corresponding to allelic reads cutoff of 5. The goodness of fit was assessed by Root Mean Square Error of Approximation (RMSEA[55], **Supplementary Fig. S11**). Low-quality fits with RMSEA > 0.05 were discarded, fixing the parameters at $r=m$ and $w=1$, thereby penalizing the statistical significance of ASB at such SNVs, as fitted r is systematically lower than m (**Supplementary Fig. S12**). Of note, the values of r for distribution of reference allele read counts (with fixed alt-allele read counts) were systematically higher than those for alternative allele read counts (with fixed Ref-allele read counts), thus balancing the reference mapping bias. The obtained fitted models were used for statistical evaluation of ASB for alternative and reference alleles independently, with one-tailed tests. Examples of fits for BAD=1 and 2 are shown in **Fig. 6E, F**, with RMSEA < 0.02 for the fixed Ref/Alt read counts of 10.

Aggregation of ASB P-values from individual data sets

For each ChIP-Seq read alignment (except control data), we performed the ASB calling. Next, the SNVs were grouped by a particular TF (across cell types) or by a particular cell type (across TFs). A group of SNVs with the same position and alternative alleles was considered as an ASB candidate if at least one of the SNVs passed a total coverage threshold ≥20. Next, for each ASB candidate, we performed *logit* aggregation of individual ASB P-values[27], independently for Ref-ASB and Alt-ASB. Individual P-values of 1 were excluded from aggregation, and if none were left, the aggregated P-value for an SNV was set to 1.

Logit aggregation is the method of a choice, as it has two advantages. First, compared to Fisher's method, it cancels out symmetrical P-values like 0.01 and 0.99 to 0.5. Second, the pattern of evidence is not known in advance, significant ASB p-values can arise both from a small number of strongly imbalanced SNVs in deeply sequenced datasets and from a large number of weakly imbalanced SNVs in datasets with low or medium coverage. Compared to the similar Stauffer's method, the logit aggregation is less sensitive to the extreme P-values and can be considered a robust choice[56]. The

resulting aggregated P-values were FDR-corrected (Benhamini-Hochberg adjustment) for multiple tested SNVs separately for each TF and each cell type. SNVs passing 0.05 FDR for either Ref or Alt allele were considered ASB.

ASB effect size estimation

We define the ES separately for reference allele ASB ($\mathrm{ES}_{\mathrm{Ref}}$) and alternative allele ASB ($\mathrm{ES}_{\mathrm{Alt}}$) as the log-ratio of the observed number of reads to the expected number. To account for BAD and mapping bias, we use fitted Negative Binomial mixture at the fixed allele read counts:

$$\mathrm{ES}_{\mathrm{Ref}} = \log_2(C_{\mathrm{Ref}} / \mathrm{E}(C_{\mathrm{Ref}} \mid C_{\mathrm{Alt}})$$
$$\mathrm{ES}_{\mathrm{Alt}} = \log_2(C_{\mathrm{Alt}} / \mathrm{E}(C_{\mathrm{Alt}} \mid C_{\mathrm{Ref}})$$

In the basic case of BAD=1, the effect size can be approximated as the log-ratio of read counts, taking into account that the expectation bias due to the truncation is relatively small and r is close to the read count on the fixed allele:

$$\mathrm{ES}_{\mathrm{Ref}} \approx \log_2(\mathrm{C}_{\mathrm{Ref}} / \mathrm{C}_{\mathrm{Alt}})$$

In the case of BAD > 1, the same assumptions lead to the following estimation of the ES:

$$\log_2(C_{\mathrm{Ref}} \cdot \mathrm{BAD} / C_{\mathrm{Alt}}) \lesssim \mathrm{ES}_{\mathrm{Ref}} \lesssim \log_2(C_{\mathrm{Ref}} / (\mathrm{BAD} \cdot C_{\mathrm{Alt}}))$$

This holds due to the fact that for fixed BAD, C_{Ref} expectation is either $C_{\mathrm{Alt}} \cdot \mathrm{BAD}$ or $C_{\mathrm{Alt}} / \mathrm{BAD}$, depending on a haplotype. Therefore, the expectation of C_{Ref} according to the Negative Binomial Mixture Model is approximately $(1\text{-}w) \cdot C_{\mathrm{Alt}} \cdot \mathrm{BAD} + w \cdot C_{\mathrm{Alt}} / \mathrm{BAD}$.

The final ASB ES is estimated for SNVs with aggregated significance either across TFs or across cell types. The ES value is calculated as a weighted average of ES of individual SNVs in aggregation, with weights assigned as negative logarithms of individual P-values. ES is not assigned in the case if all individual P-values are equal to 1.

SNV and ASB annotation

Genomic annotation

To annotate SNVs according to their genomic location (**Fig. 3C**), we started with mapping SNVs to FANTOM5 enhancers and promoters[57]. The remaining set SNVs were annotated with ChIPseeker[58] with a hierarchical assignment of the following categories: Promoter (≤1kb), Promoter (1-2kb), Promoter (2-3kb), 5'UTR, 3'UTR, Exon, Intron, Downstream, Intergenic. For clarity, promoter (≤1kb) and 5'UTR categories were both tagged as 'Promoter'; Promoter (1-2kb) and Promoter (2-3kb) were both tagged as 'Upstream'.

Sequence motif analysis of ASBs

For TF-ASBs, we annotated the corresponding SNVs with sequence motif hits of the respective TFs. To this end, we used models from HOCOMOCO v11 core collection[59] and SPRY-SARUS[60] for motif finding. The top-scoring motif hit was taken considering both Ref and Alt alleles, and, at this fixed position, the *motif fold change* (FC) was calculated as the log-ratio of motif P-values at the reference and alternative variants so that the positive FC corresponded to the preference of the alternative allele.

To analyze the ASB motif concordance (**Fig. 4**), we considered the ASB SNVs ($\min(\mathrm{FDR}_{\mathrm{Ref}}, \mathrm{FDR}_{\mathrm{Alt}}) \le 0.05$) that overlapped the predicted TF binding site: ($\min(\text{motif P-value}_{\mathrm{Ref}}, \text{motif P-value}_{\mathrm{Alt}}) \le 0.0005$), and had $|\mathrm{FC}| \ge 2$. We defined the motif concordance/discordance as a match/mismatch of the signs of FC and $\Delta\mathrm{FDR} = \log_{10}(\mathrm{FDR}_{\mathrm{Alt}}) - \log_{10}(\mathrm{FDR}_{\mathrm{Ref}})$.

Annotation of ASBs with phenotype associations

To assess enrichment of ASBs within phenotype-associated SNPs, we used the data from four different SNP-phenotype associations databases, namely: (1) NHGRI-EBI GWAS catalog[36], release 8/27/2019 with EFO mappings[61] used to group phenotypes by their parent terms for **Fig. 5C, D**; (2) ClinVar catalog[37], release 9/05/2019; (3) PheWAS catalog[38]; (4) Finemapping catalog of causal autoimmune disease variants[62]. All entries were systematized in the form of triples <dbSNP ID, phenotype, database>. Next, the entries were annotated with the TF- or cell type-ASB data.

To evaluate TF-phenotype associations in detail, we used NHGRI-EBI GWAS Catalog and the following pipeline:

1) We filtered out TFs with less than two candidate ASBs, and phenotypes associated with less than two SNPs, resulting in 765 TFs and 2688 phenotypes suitable for the analysis. For each TF, we considered all SNPs with candidate ASBs passing the coverage thresholds.
2) For each pair of a TF and a phenotype, we calculated the odds-ratio and the P-value of the first one-tailed Fisher's exact test on SNPs with candidate ASBs considering two binary features: whether the SNP is associated with the phenotype, and whether the SNP is included in ASB candidates of the particular TF. The superset of SNPs was collected independently for each TF by gathering SNPs with candidate ASBs for all TFs but only from LD blocks[42] containing either TF-specific SNPs or phenotype-associated SNPs. The P-values were then FDR-corrected for multiple tested TFs separately for each phenotype.

Analysis of eQTL target genes

To evaluate ASB-driven eQTL target genes' associations with phenotypes, a one-tailed Fisher's exact test was performed on the enrichment of eQTL target genes of ASB SNPs (28588 genes according to GTEx[41]) with phenotype-associated genes from ClinVar catalog[37] (19657 genes, among all human genes from Ensembl[63] (60860 genes).

ASB prediction with machine learning

In our work, we used a standard software implementation of the random forest model from the scikit-learn package. The number of estimators was set to 500 and the other parameters were defaults. Three feature types were used (**Supplementary Table S4**): allele-specific chromatin DNase accessibility, synthetic data from neurons from the last layer of the DeepSEA[11], and HOCOMOCO motif predictions obtained with SPRY-SARUS[60]. As a global set of SNVs, we used 231355 dbSNP IDs overlapping between ADASTRA and Maurano et al. data[64], which provided allele-specific DNase accessibility. For the general model, we used SNVs with ASBs for any of TFs or in any of cell types as members of the positive class, and the remaining set of candidate SNVs as members of the negative class. For TF- and cell type-specific assessment, we defined ASB and non-ASB SNVs for a particular TF or in a particular cell type as the positive and negative class, respectively.

Data and software availability

The complete data on ASBs across TFs and cell types are aggregated and annotated in the ADASTRA database (http://adastra.autosome.ru/) and provided online: http://adastra.autosome.ru/soos/, the generated BAD maps are available at http://adastra.autosome.ru/soos/downloads.

The ADASTRA pipeline is available at GitHub: https://github.com/autosome-ru/ADASTRA-pipeline[65].

BABACHI segmentation software is available at GitHub: https://github.com/autosome-ru/BABACHI[66].

The code for machine learning analysis is available at GitHub: https://github.com/autosome-ru/ASB-ML[67].

The SPRY-SARUS motif scanner is available at GitHub: https://github.com/autosome-ru/sarus[60].

The reprocessed ChIP-Seq peaks and metadata are available in the GTRD database: http://gtrd.biouml.org.

Author contributions

SA and AB developed the computational framework and database; SA, AB, and IEV developed the website; DB, EB, IEV, AVF, and MVF performed the functional annotation and motif annotation of ASBs; DB and DP performed the machine learning analysis; IY and FK established the GTRD alignments processing; VJM and IVK designed and supervised the study. All authors participated in the manuscript preparation.

Conflict of interest

None declared.

Funding

This study was supported by RFBR grant 18-34-20024 to IVK (basic ADASTRA pipeline), RSF grant 20-74-10075 to IVK (machine learning application), RSF grant 19-14-00295 to FK (GTRD data extraction).

Acknowledgements

The authors thank Denis Litvinov for help in GTRD metadata processing and Evgenia Serebrova for help in manuscript preparation.

References

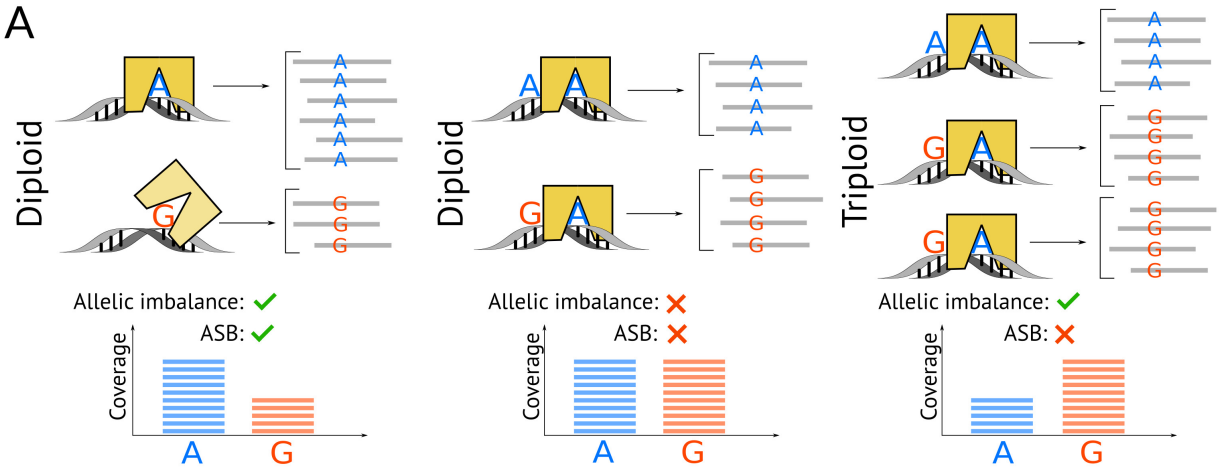
1. Ponomarenko, J. V. *et al.* rSNP_Guide: An integrated database-tools system for studying SNPs and site-directed mutations in transcription factor binding sites. *Hum. Mutat.* **20**, 239–248 (2002).
2. Cavalli, M. *et al.* Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. *Hum. Genet.* **135**, 485–497 (2016).
3. PCAWG Drivers and Functional Interpretation Working Group *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
4. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* **166**, 538–554 (2016).
5. Penzar, D. D. *et al.* What Do Neighbors Tell About You: The Local Context of Cis-Regulatory Modules Complicates Prediction of Regulatory Variants. *Front. Genet.* **10**, 1078 (2019).
6. van Arensbergen, J. *et al.* High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* **51**, 1160–1169 (2019).

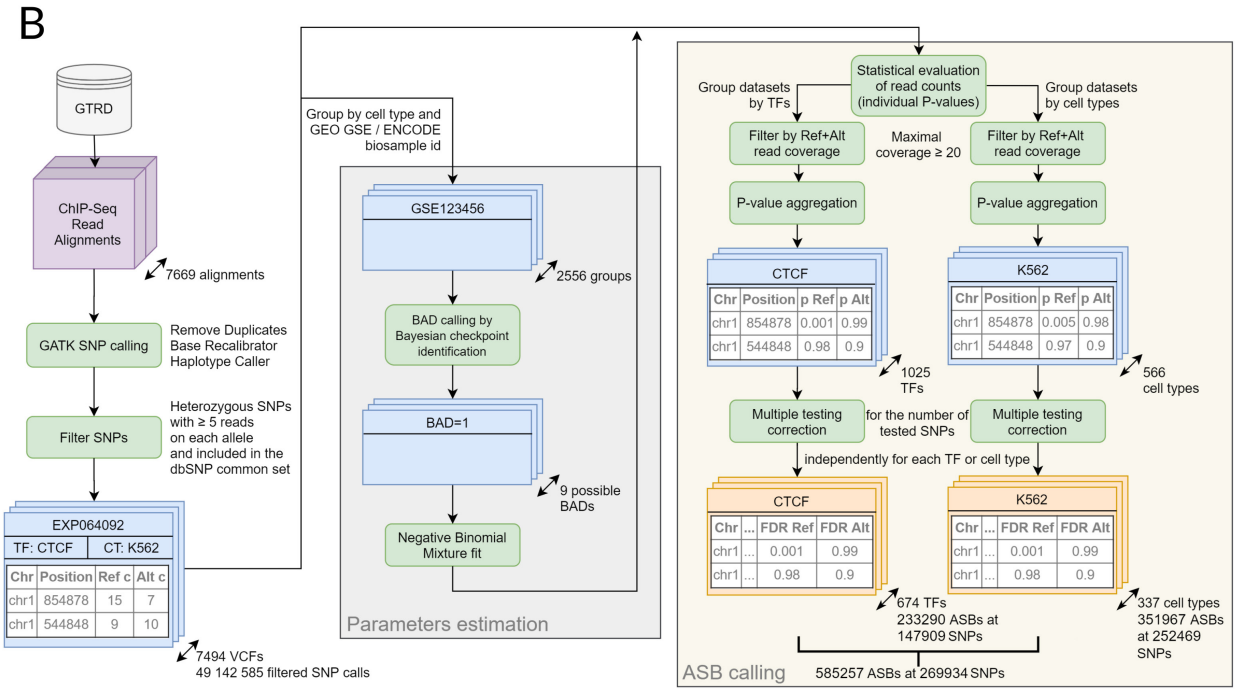
7. Bulyk, M. L. Protein Binding Microarrays for the Characterization of DNA–Protein Interactions. in *Analytics of Protein–DNA Interactions* (ed. Seitz, H.) vol. 104 65–85 (Springer Berlin Heidelberg, 2006).
8. Rockel, S., Geertz, M. & Maerkl, S. J. MITOMI: A Microfluidic Platform for In Vitro Characterization of Transcription Factor–DNA Interaction. in *Gene Regulatory Networks* (eds. Deplancke, B. & Gheldof, N.) vol. 786 97–114 (Humana Press, 2012).
9. Korneev, K. V. *et al.* Minor C allele of the SNP rs7873784 associated with rheumatoid arthritis and type-2 diabetes mellitus binds PU.1 and enhances TLR4 expression. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* **1866**, 165626 (2020).
10. Putlyaeva, L. V. *et al.* Potential Markers of Autoimmune Diseases, Alleles rs115662534(T) and rs548231435(C), Disrupt the Binding of Transcription Factors STAT1 and EBF1 to the Regulatory Elements of Human CD40 Gene. *Biochem. Mosc.* **83**, 1534–1542 (2018).
11. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **12**, 931–934 (2015).
12. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
13. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
14. McDaniell, R. *et al.* Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans. *Science* **328**, 235–239 (2010).
15. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
16. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* **47**, 1393–1401 (2015).
17. Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y. & Kolpakov, F. GTRD: a database on gene transcription regulation—2019 update. *Nucleic Acids Res.* **47**, D100–D105 (2019).
18. Chèneby, J. *et al.* ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.* gkz945 (2019) doi:10.1093/nar/gkz945.
19. de Santiago, I. *et al.* BaalChIP: Bayesian analysis of allele-specific transcription factor binding in cancer genomes. *Genome Biol.* **18**, 39 (2017).
20. Shi, W., Fornes, O., Mathelier, A. & Wasserman, W. W. Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.* gkw691 (2016) doi:10.1093/nar/gkw691.
21. Rozowsky, J. *et al.* AlleleSeq: analysis of allele - specific expression and binding in a network framework. *Mol. Syst. Biol.* **7**, 522 (2011).
22. Chen, J. *et al.* A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat. Commun.* **7**, 11101 (2016).
23. Liu, Y. *et al.* Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat. Biotechnol.* **37**, 314–322 (2019).
24. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).
25. Wei, Y., Li, X., Wang, Q. & Ji, H. iASeq: integrative analysis of allele-specificity of protein-DNA interactions in multiple ChIP-seq datasets. *BMC Genomics* **13**, 681 (2012).
26. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
27. George, E. O. & Mudholkar, G. S. On the convolution of logistic random variables. *Metrika* **30**, 1–13 (1983).
28. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
29. Varma, S., Pommier, Y., Sunshine, M., Weinstein, J. N. & Reinhold, W. C. High Resolution Copy Number Variation Data in the NCI-60 Cancer Cell Lines from Whole Genome Microarrays Accessible through CellMiner. *PLoS ONE* **9**, e92047 (2014).
30. Cavalli, M. *et al.* Allele specific chromatin signals, 3D interactions, and motif predictions for immune and B cell related diseases. *Sci. Rep.* **9**, 2695 (2019).
31. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
32. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390–394 (2011).
33. Wagih, O., Merico, D., Delong, A. & Frey, B. J. *Allele-specific transcription factor binding as a benchmark for assessing variant impact predictors*. http://biorxiv.org/lookup/doi/10.1101/253427 (2018) doi:10.1101/253427.
34. Ershova, A. S. *et al. Enhanced C/EBPs binding to C>T mismatches facilitates fixation of CpG mutations*. http://biorxiv.org/lookup/doi/10.1101/2020.06.11.146175 (2020) doi:10.1101/2020.06.11.146175.
35. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
36. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
37. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence.

Nucleic Acids Res. **46**, D1062–D1067 (2018).
38. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1111 (2013).
39. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
40. Brem, R. B. Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**, 752–755 (2002).
41. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
42. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* btv546 (2015) doi:10.1093/bioinformatics/btv546.
43. Pomerantz, M. M. *et al.* Prostate cancer reactivates developmental epigenomic programs during metastatic progression. *Nat. Genet.* **52**, 790–799 (2020).
44. Aue, A. *et al.* Elevated STAT1 expression but not phosphorylation in lupus B cells correlates with disease activity and increased plasmablast susceptibility. *Rheumatology* keaa187 (2020) doi:10.1093/rheumatology/keaa187.
45. Wang, W. *et al.* A functional polymorphism in *TFF1* promoter is associated with the risk and prognosis of gastric cancer: A functional polymorphism in *TFF1* promoter. *Int. J. Cancer* **142**, 1805–1816 (2018).
46. Gate, R. E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).
47. Fowler, S. A. *et al.* SMAD3 gene variant is a risk factor for recurrent surgery in patients with Crohn's disease. *J. Crohns Colitis* **8**, 845–851 (2014).
48. Consortium, T. U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
49. Bellofatto, V. & Wilusz, J. Transcription and mRNA Stability: Parental Guidance Suggested. *Cell* **147**, 1438–1439 (2011).
50. Zid, B. M. & O'Shea, E. K. Promoter sequences direct cytoplasmic localization and translation of mRNAs during starvation in yeast. *Nature* **514**, 117–121 (2014).
51. Belkadi, A. *et al.* Identification of genetic variants controlling RNA editing and their effect on RNA structure stabilization. *Eur. J. Hum. Genet.* (2020) doi:10.1038/s41431-020-0688-7.
52. Ameur, A., Rada-Iglesias, A., Komorowski, J. & Wadelius, C. Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP. *Nucleic Acids Res.* **37**, e85–e85 (2009).
53. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
54. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
55. Browne, M. W. & Cudeck, R. Alternative Ways of Assessing Model Fit. *Sociol. Methods Res.* **21**, 230–258 (1992).
56. Loughin, T. M. A systematic comparison of methods for combining p-values from independent tests. *Comput. Stat. Data Anal.* **47**, 467–485 (2004).
57. the FANTOM consortium *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 22 (2015).
58. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
59. Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).
60. Denisenko, N., Kulakovskiy, I. & Vorontsov, I. *autosome-ru/sarus: SPRY-SARUS v2.0.2*. (Zenodo, 2020). doi:10.5281/ZENODO.4015924.
61. Malone, J. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118 (2010).
62. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
63. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
64. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* **47**, 1393–1401 (2015).
65. Abramov, S. & Boytsov, A. *autosome-ru/ADASTRA-pipeline: release-Soos*. (Zenodo, 2020). doi:10.5281/zenodo.4008546.
66. Abramov, S. & Boytsov, A. *autosome-ru/BABACHI: release 1.3.7*. (Zenodo, 2020). doi:10.5281/ZENODO.4008544.
67. Penzar, D. *autosome-ru/ASB-ML: ASB-ML*. (Zenodo, 2020). doi:10.5281/ZENODO.4043865.

Figure 1. A scheme of allele-specific binding events, an overview of the ADASTRA pipeline, and its application to ChIP-Seq data.

(A) ChIP-Seq data allows detecting ASB events by estimating the imbalance of reads carrying alternative alleles. ASBs must be distinguished from sites where the allelic imbalance is caused by aneuploidy and copy-number variants.
(B) The scheme of the ADASTRA pipeline: variant calling in read alignments from GTRD, estimation of statistical model parameters and background allelic dosage, filtering, and statistical evaluation of candidate ASBs. ADASTRA generates two complementary datasets: transcription factor-ASBs (pairs of an SNP and a TF) and cell type-ASBs (pairs of an SNP and a cell type). SNPs are annotated according to dbSNP IDs.
(C, D) Number of SNPs (dbSNP IDs, Y-axis) with significant ASB events for various transcription factors **(C)** and for various cell types **(D)**. TFs or cell types (X-axis) are sorted by the number of SNPs.
GTRD: Gene Transcription Regulation Database, ADASTRA: Allelic Dosage-corrected Allele-Specific human Transcription factor binding sites.





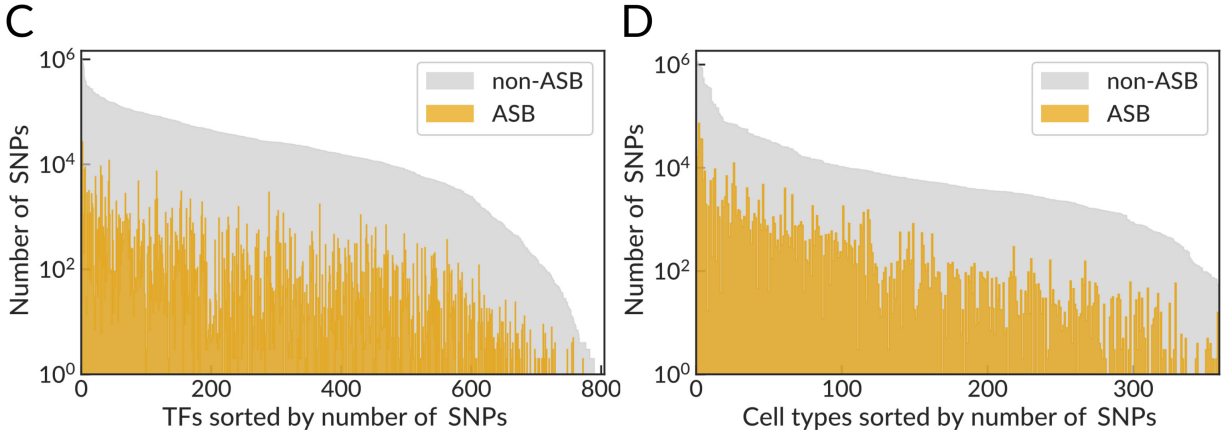


Figure 2. Bayesian changepoint identification allows reconstructing reliable genome-wide maps of background allelic dosage from single-nucleotide variant calls.

(A) BAD calling with bayesian changepoint identification applied to variant calls detected at chr2 and chr6 in K562 ENCODE data (ENCBS725WFV). X-axis: chromosome position, bp. Y-axis: the allelic disbalance of individual SNVs. Horizontal green lines (ground-level of the plots) indicate results of the initial stage of the algorithm: the detection of SNV-free regions including deletions, telomeric, and centromeric segments. Horizontal light-blue lines: predicted BAD. Orange dashes: 'ground truth' BAD according to the COSMIC data (when available).
(B) Y-axis: SNV-level Kendall τ_b rank correlation between the predicted BAD and the 'ground truth' BAD (COSMIC data). Each of 516 points denotes a particular group of related data sets of the same series (ENCODE biosample or GEO GSE ID) and the same cell type. X-axis: the number of SNV calls in a particular group of related data sets. Only SNVs falling into regions of known BAD (present in the COSMIC data) are considered, recurrent SNVs in several data sets are considered only once.
(C, D) Receiver Operating Characteristic and Precision-Recall curves for predicted BAD maps used as binary classifiers of individual SNVs according to BAD vs the 'ground truth' COSMIC data. To plot each curve, the score $S = L(\mathrm{BAD}{=}x) - \max_{y \neq x} L(\mathrm{BAD}{=}y)$, where L denotes log-likelihood, was used as the prediction score for thresholding. Colored circles denote the values obtained with the final BAD maps where particular BAD values were assigned to each segment according to the maximum posterior. Regions with BAD of 1, 3/2, 2, and 3 contain more than 97% of all candidate ASB variants.

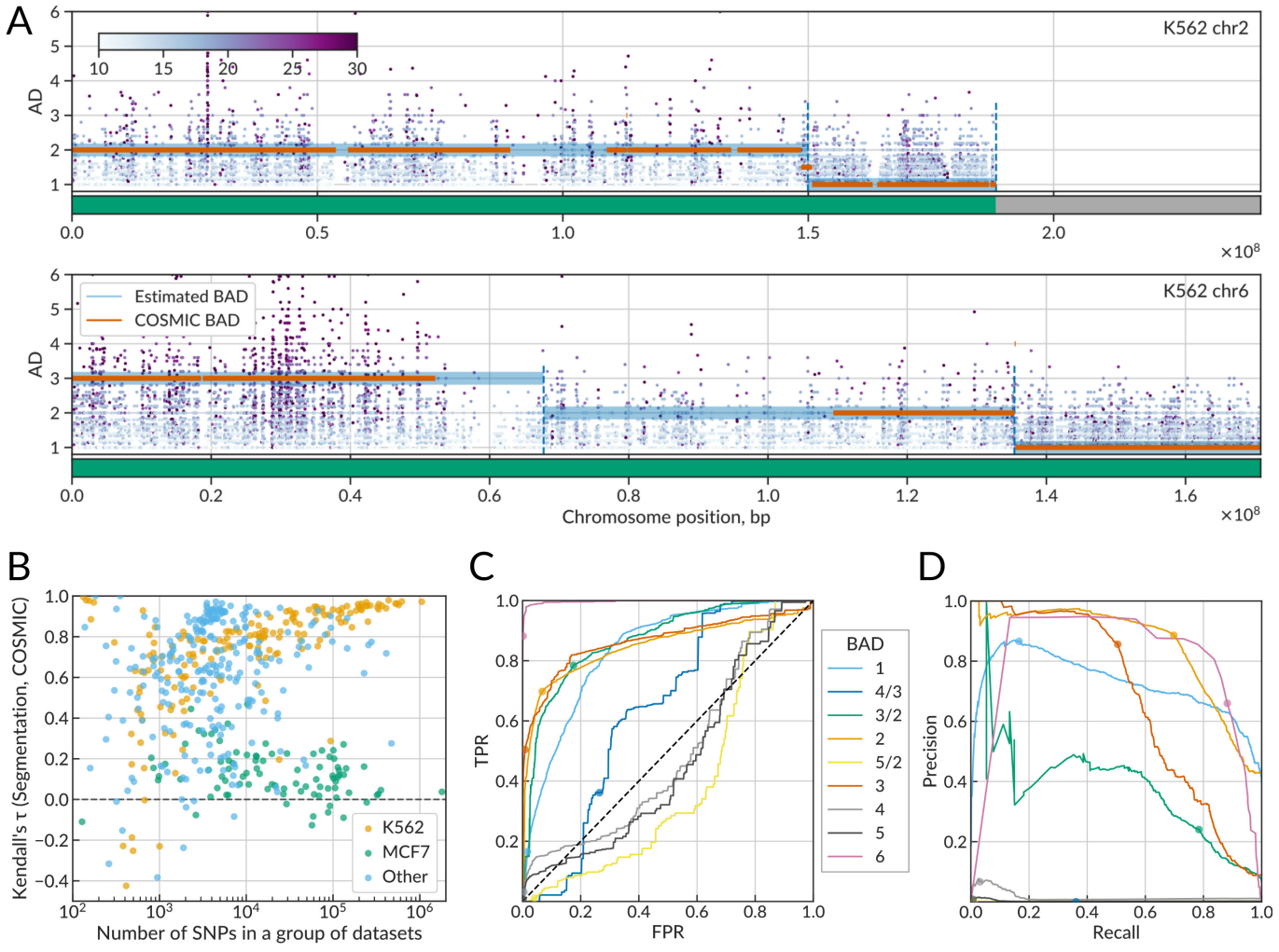


Figure 3. An overview of the ADASTRA ASBs and their genomic localization.

(A, B) The distribution of ASBs across TFs and cell types is not uniform. The top 8 TFs and top 5 cell types provide only nearly one third (TFs) or one half (cell types) of significant events. The bottom bars in each pair show the zoomed-in data for the top 8 TFs and top 5 cell types sorted by descending number of ASBs.
(C) The complete bars correspond to the full set of SNPs (unique dbSNP IDs) with significant ASBs. The ASBs are more often found in promoters and enhancers as compared to either SNVs with candidate ASBs or all detected SNVs. The percentage of ASB-carrying SNPs falling into particular types of genomic regions is shown on bar labels. Top bar: significant ASBs (passing 5% FDR, 269934 sites in total); middle bar: SNPs with candidate ASBs (passing the coverage thresholds and tested for significance, 2024836 sites in total); bottom bar: all SNPs detected in variant calling (4976303 sites in total).

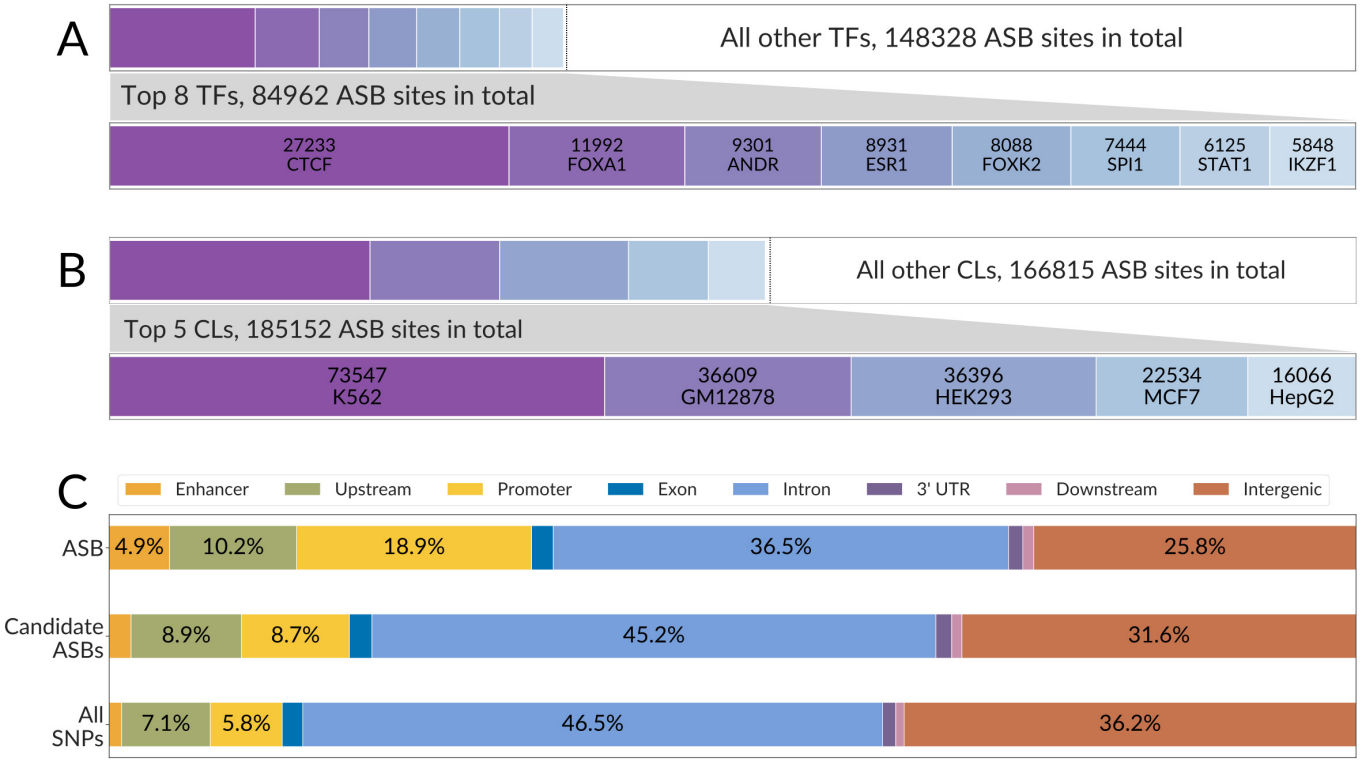


Figure 4. Motif annotation of SNVs agrees with TF-ASB calls.

(A) Scatterplot of the motif fold change (the predicted change in TF binding affinity) versus the ASB significance for TFs that have PWMs in HOCOMOCO v11 core collection. The plot shows only the ASBs that overlap the TF motif occurrence (TF motif PWM hit with P-value ≤ 0.0005). X-axis: signed ASB significance, the absolute value is max(-log10 FDR Ref ASB, -log10 FDR Alt ASB). The sign is set to negative if Ref ASB is more significant than Alt ASB (positive otherwise). Y-axis: motif fold change (FC) estimated as the $\log_2$-ratio of motif PWM hit P-values between the reference and the alternative alleles (the positive value corresponds to a higher affinity to alternative allele). The SNVs are marked as concordant (discordant) and colored in blue (red) if they exhibit significant ASBs (FDR ≤ 0.05), have motif |FC| ≥ 2, and the preferred allele of the ASB corresponds to (is opposite to) that of the TF motif.
(B) The fraction of discordant and concordant SNVs (Y-axis) and the total number of concordant SNVs among them depending on the ASB significance cutoff, $-\log_{10}$ FDR (X-axis).
(C) Barplot illustrating the proportion of SNVs with concordant and discordant ASBs for top 10 TFs with the largest total numbers of eligible SNVs.
(D) The *staveplot* for CEBPB illustrating motif analysis of significant ASBs. Each dot represents an SNV that is ASB for CEBPB that overlaps the predicted CEBPB binding site (P-value ≤ 0.0005) and has motif |FC| ≥ 2. The X-coordinate shows the SNV position in the motif (underlined by the motif logo), the individual dashed strings denote 4 possible minor alleles at each position, the color is defined by the major allele. The strand orientation of ASBs is aligned to the predicted motif hits. Y-axis shows the effect size of the ASB major allele.

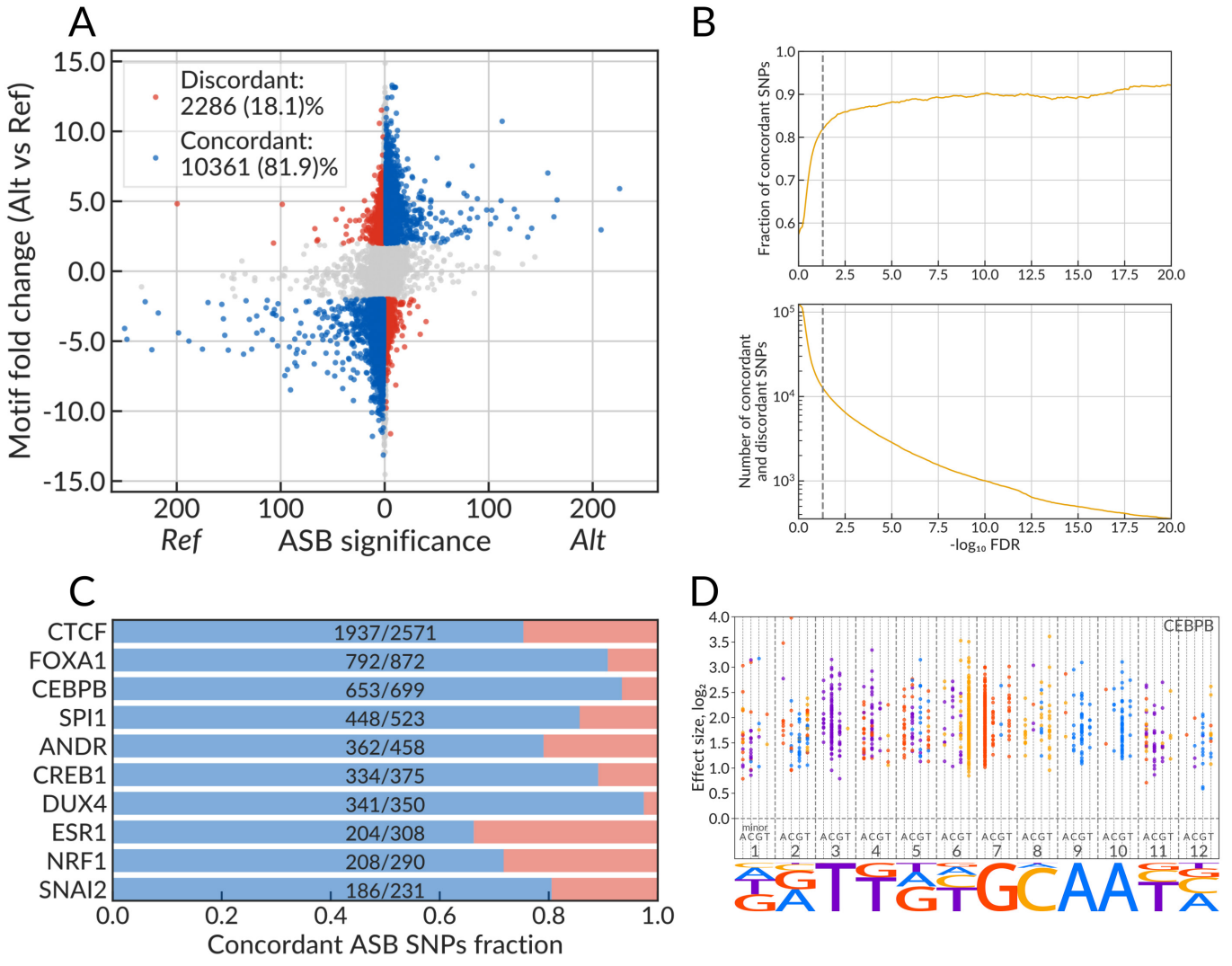
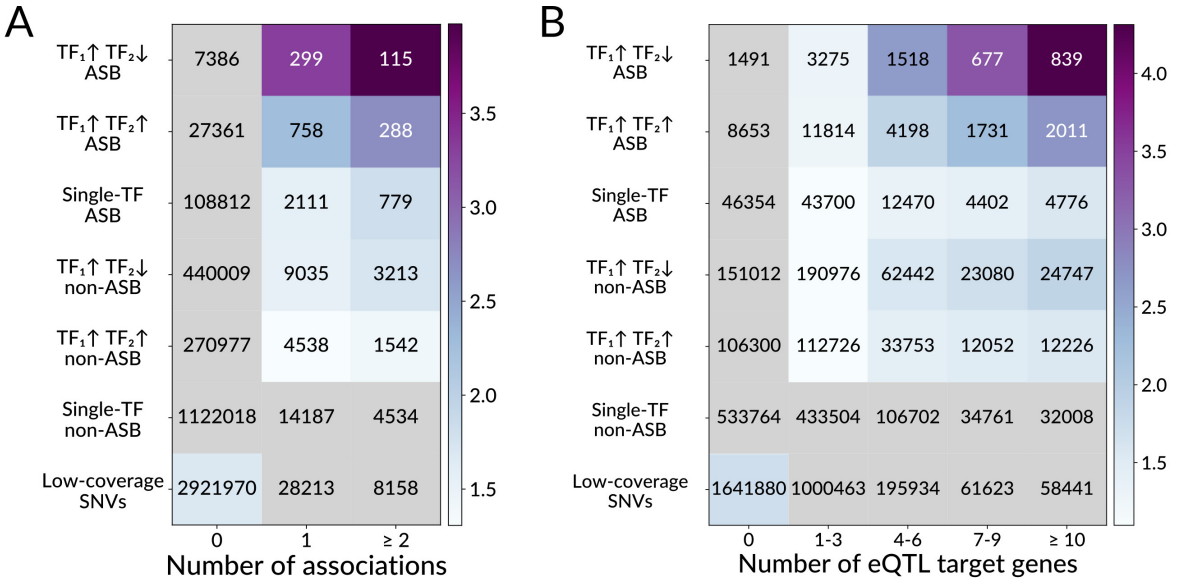
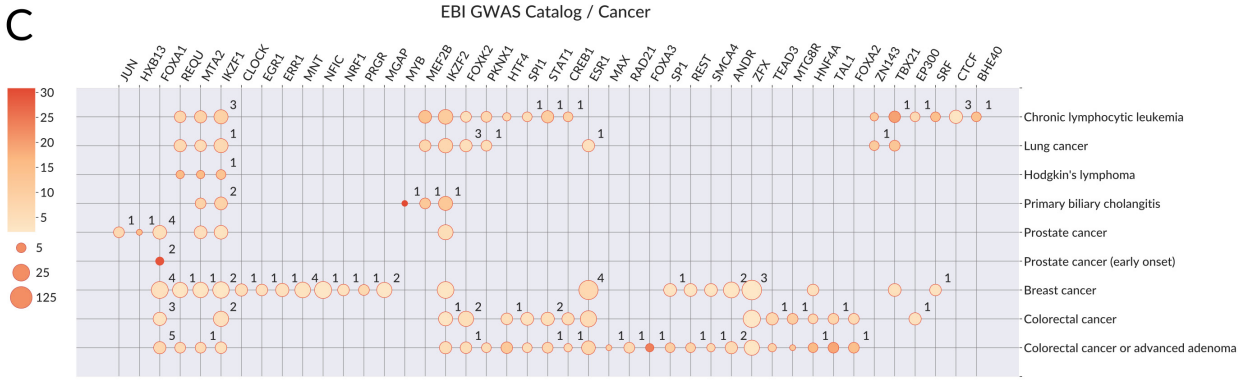


Figure 5. ASBs are enriched with pathogenic phenotype associations and eQTLs.

(A-B) Enrichment of ASBs among phenotype-associated and eQTL SNVs. Y-axis denotes several exclusive groups of SNPs: TF1 ↑ TF2 ↓ , SNVs carrying both Ref- and Alt-ASBs of different TFs, i.e. where at least two TFs prefer to bind alternating alleles; TF1 ↑ TF2 ↑ , SNVs carrying ASBs for at least two TFs preferring to bind the same allele; single-TF, SNVs with ASB of a single TF; Low-coverage SNVs that did not pass a total coverage threshold ≥ 20. Non-ASBs are SNVs with the TF-ASB FDR >0.05. X-axis: **(A)** the number of unique (dbSNP ID, trait, database) triples for a given SNV considering four databases of SNP-phenotype associations (EBI, ClinVar, PheWAS, and BROAD autoimmune diseases fine-mapping catalog); **(B)** the number of eQTL target genes according to GTEx eQTL data. The coloring denotes the odds ratios of the one-tailed Fisher's exact test for the enrichment of SNVs with associations for each group of ASBs (against all other SNVs in the table). The gray cells correspond to non-significant enrichments with P > 0.05 after Bonferroni correction for the total number of cells. The values in the cells denote the numbers of SNVs.

(C-D) Significant TF-phenotype associations estimated for ADASTRA SNVs and EBI-GWAS catalog data. Phenotypes categories: **(C)** cancer, **(D)** immune system disorder. X-axis: TFs, Y-axis: phenotypes. Each bubble represents a TF-phenotype pair with the SNVs found in TF ChIP-Seq data significantly enriched with the phenotype associations (the FDR-corrected P < 0.05 & odds ratio > 2) . The numbers in superscript show the number of TF-ASB sites associated with the phenotype. The area of the circles is proportional to the $\log_2$-number of the phenotype-associated SNVs found in TF ChIP-Seq data. The coloring scheme represents the odds ratios of the enrichment. TF-phenotype combinations w/o ASBs are not shown.





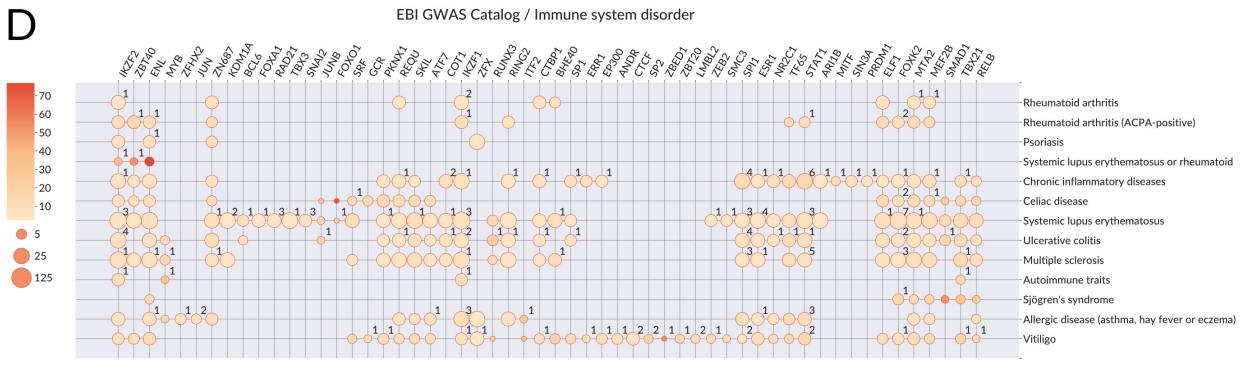


Figure 6. Distribution of read counts at SNVs significantly depends on background allelic dosage.

Each panel contains 3 plots: (1, left) A heatmap of allelic read counts colored by $\log_{10}$[number of SNVs that have the specified number of ChIP-seq reads] supporting the *reference* (X-axis) and *alternative* (Y-axis) alleles.
(2-middle, 3-right) Barplots of observed read counts at one of the alleles and the approximating distribution plot. Two barplots correspond to the two slices of the heatmap data, either by fixing the sum of reads at two alleles (A, B, C, D, diagonal slices along the dashed lines in the bottom left corner, approximated by the binomial mixture) or by fixing the read counts at one of the alleles (E, F, vertical and horizontal slices, approximated by the negative binomial mixture).
(A) Complete set of ADASTRA candidate ASB SNVs, no separation by BAD, the observed distribution can be interpreted as overdispersed binomial.
(B) K562 candidate SNVs, the distribution is similar to an overdispersed mixture of binomial distributions with $p=⅓$ and $p=⅔$ as K562 are mostly triploid.
(C) SNVs in diploid regions according to BAD=1, binomial distribution with $p=½$.
(D) SNVs in BAD-separated triploid regions (BAD=2), binomial mixture with $p=⅓$ and $p=⅔$.
(E) BAD-separated diploids (BAD=1), negative binomial distribution with $p=½$ (fit).
(F) BAD-separated triploids (BAD=2), negative binomial mixture with $p=⅓$ and $p=⅔$ (fit).
In all the cases the distributions are truncated, corresponding to the allelic read counts cutoff of 5.