Year : 2016

FUNCTIONAL IMPACT OF GENETIC VARIATION ON GENE EXPRESSION

Andreas Gschwind

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Centre Intégratif de Génomique (CIG)

FUNCTIONAL IMPACT OF GENETIC VARIATION ON GENE EXPRESSION

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Andreas GSCHWIND

Master of Science de l'Université de Berne

Jury

Prof. Jean-François Demonet, Président
Prof. Alexandre Reymond, Directeur de thèse
Prof. Ioannis Xenarios, expert
Prof. Niko Beerenwinkel, expert

Lausanne 2016

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président · e	Monsieur	Prof.	Jean-François	**Demonet**
Directeur · rice de thèse	Monsieur	Prof.	Alexandre	**Reymond**
Experts · es	Monsieur	Prof.	Ioannis	**Xenarios**
	Monsieur	Prof.	Niko	**Beerenwinkel**

le Conseil de Faculté autorise l'impression de la thèse de

Monsieur Andreas Gschwind

Master of science de l' Université Berne

intitulée

FUNCTIONAL IMPACT OF GENETIC VARIATION ON GENE EXPRESSION

Lausanne, le 8 janvier 2016

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Jean-François Demonet

Acknowledgements

I would like to thank all the collaborators with whom I had the chance to work and who contributed greatly to my thesis. Without them the work presented in this thesis would not have been possible. I would like to express my gratitude to Alexandre Reymond for giving me the opportunity to pursue this thesis and providing invaluable assistance. My thanks go to all present and former members of Alexandre Reymonds laboratory with whom I had the pleasure to work, with a special mention to Eugenia Migliavacca, who was an outstanding mentor. I also express my gratitude to the members of my PhD committee, namely Alexandre Reymond, Ioannis Xenarios, Niko Beerenwinkel and Jean-François Demonet, who gave invaluable advice for the work presented in this thesis. Many thanks also go to the Vital-IT team for their services and technical support. I would like to thank all my colleagues, friends and family members for all their support during my time as a PhD student, both on a professional and personal level.

Summary

Numerous links between genetic variants and phenotypes are known and genome-wide association studies dramatically increased the number of genetic variants associated with traits during the last decade. However, how changes in the DNA perturb the molecular mechanisms and impact on the phenotype of an organism remains elusive. Studies suggest that many trait-associated variants are in the non-coding region of the genome and probably act through regulation of gene expression. During my thesis I investigated how genetic variants affect gene expression through gene regulatory mechanisms. The first chapter was a collaborative project with a pharmaceutical company, where we investigated genome-wide copy number variation (CNVs) among Cynomolgus monkeys (*Macaca fascicularis*) used in pharmaceutical studies, and associated them to changes in gene expression. We found substantial copy number variation and identified CNVs linked to tissue-specific expression changes of proximal genes. The second and third chapters focus on genetic variation in humans and its effects on gene regulatory mechanisms and gene expression. The second chapter studies two human trios, where the allelic effects of genetic variation on genome-wide gene expression, protein-DNA binding and chromatin modifications were investigated. We found abundant allele specific activity across all measured molecular phenotypes and show extended coordinated behavior among them. In the third chapter, we investigated the impact of genetic variation on these phenotypes in 47 unrelated individuals. We found that chromatin phenotypes are organized into local variable modules, often linked to genetic variation and gene expression. Our results suggest that chromatin variation emerges as a result of perturbations of *cis*-regulatory elements by genetic variants, leading to gene expression changes. The work of this thesis provides novel insights into how genetic variation impacts gene expression by perturbing regulatory mechanisms.

Résumé

De nombreux liens entre variations génétiques et phénotypes sont connus. Les études d'association pangénomique ont considérablement permis d'augmenter le nombre de variations génétiques associées à des phénotypes au cours de la dernière décennie. Cependant, comprendre comment ces changements perturbent les mécanismes moléculaires et affectent le phénotype d'un organisme nous échappe encore. Des études suggèrent que de nombreuses variations, associées à des phénotypes, sont situées dans les régions non codantes du génome et sont susceptibles d'agir en modifiant la régulation d'expression des gènes. Au cours de ma thèse, j'ai étudié comment les variations génétiques affectent les niveaux d'expression des gènes en perturbant les mécanismes de régulation de leur expression. Le travail présenté dans le premier chapitre est un projet en collaboration avec une société pharmaceutique. Nous avons étudié les variations en nombre de copies (CNV) présentes chez le macaque crabier (*Macaca fascicularis*) qui est utilisé dans les études pharmaceutiques, et nous les avons associées avec des changements d'expression des gènes. Nous avons découvert qu'il existe une variabilité substantielle du nombre de copies et nous avons identifié des CNVs liées aux changements d'expression des gènes situés dans leur voisinage. Ces associations sont présentes ou absentes de manière spécifique dans certains tissus. Les deuxième et troisième chapitres se concentrent sur les variations génétiques dans les populations humaines et leurs effets sur les mécanismes de régulation des gènes et leur expression. Le premier se penche sur deux trios humains, père, mère, enfant, au sein duquel nous avons étudié les effets alléliques des variations génétiques sur l'expression des gènes, les liaisons protéine-ADN et les modifications de la chromatine. Nous avons découvert que l'activité spécifique des allèles est abondante abonde dans tous ces phénotypes moléculaires et nous avons démontré que ces derniers ont un comportement coordonné entre eux. Dans le second, nous avons examiné l'impact des variations génétiques de ces phénotypes moléculaires chez 47 individus, sans lien de parenté. Nous avons observé que les phénotypes de la chromatine sont organisés en modules locaux, qui sont liés aux variations génétiques et à l'expression des gènes. Nos résultats suggèrent que la variabilité de la chromatine est due à des variations génétiques qui perturbent des éléments *cis*-régulateurs, et peut conduire à des changements dans l'expression des gènes. Le travail présenté dans cette thèse fournit de nouvelles pistes pour comprendre l'impact des différentes variations génétiques sur l?expression des gènes à travers les mécanismes de régulation.

Contents

Introduction

Variation in biology

One of the definitions of life is its ability to adapt and evolve over time [113]. As a result of this, living systems show a tremendous amount variation, even though core genetic, cell biological, and developmental processes are largely conserved [114]. Variation in biology is found from the molecules of a cell, between individuals of a population to ecosystems. For centuries this tremendous display of variation fascinated naturalists and people such as Carl Linnaeus (1707 - 1778) spent their lives defining and cataloguing different species and taxa. However, the reason for this variability was unknown until scientists proposed the theory of evolution by natural selection during the 1850s. The most popular piece of work obviously being Charles Darwins ″On the Origin of Species″ published in 1859. This new theory unified the source and effects of variation, where differences among individuals would become the basis for adaptation through natural selection, leading to more diversity and eventually to the generation of new species [115]. One of the key criteria of this theory is that much of this variation must be heritable. No agreed on model of inheritance existed at that time [116], and even Darwin stated in the first chapter of ″On the Origin of Species″ that ″The laws governing inheritance are quite unknown.″ [115]. Only after experiments on the inheritance of traits, scientists such as Gregor Mendel (1822 - 1884) discovered the laws of inheritance. Even though the underlying biochemical mechanisms remained unknown, this knowledge laid the foundation for the science of genetics.

Genetic variation

Thanks to the discovery of the DNA double helix in 1953 by Watson and Crick [117] we now have very good knowledge about how genetic information is encoded and transmitted from one generation to another. Also we have a much more profound understanding of the nature of genetic differences between organisms. Different types of chemical mutations within DNA molecules can occur, creating de novo variation within the genetic code and evolutionary forces act on this variation, shaping the observe diversity between organisms [118]. Mutations can occur at any size, ranging from point mutations affecting single nucleotides, over insertions or deletions (indels) and microsatellites affecting several nucleotides to large structural variants (SV) such as duplications and deletions (Copy Number Variants, CNVs), inversions or translocations affecting large chunks of the genome [119, 120, 1]. Novel genetic variation is constantly generated by random mutations within the genome at specific mutation rates. Reported estimates of the mutation rate for point mutations in humans range from 1-3 x 10-8 per base pair per generation [121] and each person is estimated to carry on average $\sim$60 de novo point mutations that occured in the germline of their parents [122]. Estimated mutation rates for other types of variants are reported to be around 2.94 indels (120 bp) and 0.16 SVs (>20 bp) per generation [123].

Various evolutionary forces constantly act on a populations genetic variation. Positive selection promotes beneficial variants, balancing selection maintains variation of specific alleles and purifying selection removes deleterious variation from the gene pool. Additionally to these directed forces, random genetic drift adds stochastic fluctuation of allele frequencies [118]. Positive selection on a molecular level gained a lot of attention during recent decades, however its importance in shaping the observed genetic variability is questioned [124]. Given the random nature of mutations beneficial mutations are expected to be rare, while strongly deleterious mutations are likely to be removed from the population by purifying selection. Therefore most existing variation is expected to have very little to no effect on an individuals evolutionary fitness (neutral variation) [125] and being driven rather by random genetic drift than directed selection [126].

Despite this expectation, a small fraction of the genetic variation might still have functional implications for its carrier. Given the complexity of genomes, it is intuitive that random mutations can act through a variety of functional elements [122]. Probably the most classic mechanism is via a direct physical change of the protein sequence encoded by a gene. This can for instance be achieved if a point mutation occurs within the coding sequence of a gene and leads to a non-synonymous substitution, meaning its nucleotide change is crucial to the resulting amino acid sequence. Alternatively, mutations can lead to a stop codon and prematurely terminate the protein translation, or insertions and deletions can induce a frame shift during protein translation. Because of their potential easily change or destroy a proteins function, the effects of such mutations can be very strong. Many of such mutations were found to be associated with various diseases. A classical example is cystic fibrosis, where mutations in the CFTR gene alter the resulting proteins function and lead to severe symptoms in multiple tissues [127]. Thanks to extensive sequencing of the human exome, various diseases, such as developmental disorders were linked to specific mutations within the coding sequence of genes [128].

Because of the severity of such mutations, protein-coding sequences are usually depleted of mutations and more conserved across individuals and species than non-coding regions[129]. Much of the genetic variation lies therefore in the non-coding regions of the genome [130], where its potential implications are much harder to understand. Especially during recent years, genome-wide association studies associated phenotypic changes with genetic variants in non-coding regions, and it was concluded that their effects are most likely regulatory [71]. Rather than changing the sequence of a gene, these variants affect the expression level of a gene and therefore act through changing the stoichiometry within cells. One of the most popular examples is probably the widely spread lactose persistence in humans, where mutations upstream of the lactase-phlorizin hydrolase gene (LCT) were associated with continued expression of lactase during adulthood [131]. Further studies revealed signatures of positive selection for these variants in European-derived populations [132], making it a prime example of molecular evolution and how genetic variants can have an effect on gene expression regulation and affect an individuals phenotype. Genetic variation was associated with gene expression changes across many species and is likely to be a major driver of phenotypic variation [133, 134], disease [130] and evolution [114, 72].

From genomics to systems genetics

Systematic, genome-wide analysis of genetic and gene expression variation became feasible when the human genome was sequenced and the first array based high throughput methods appeared. This allowed assessing genetic and transcriptomic variation systematically throughout the genome and across many samples. Large efforts were made to catalogue genetic variation [48, 135] among humans, providing powerful resources to geneticists. Today, genome-wide association studies (GWAS) are able to associate phenotypes with genomic variation in thousands of individuals, leading to an impressive catalog of such associations[136]. Many of these associations are of medical relevance and genetic components for many common diseases were identified [137–140], explaining typically 10%30% of the traits heritability [137]. Also genetic bases for traits such as drug-efficacy were discovered, having potentially promising implications for personalized medicine [141, 142]. However, given their limitation of being statistical associations, they do not provide much functional explanation of the underlying molecular mechanisms. Many trait-associated variants detected in GWAS are found in non-coding regions of the genome, suggesting that their effects are most likely regulatory [71].

Novel genome-wide techniques such as chromatin immunoprecipitation (ChIP) and RNA sequencing allowed systematic investigation of regulatory mechanisms and their effects on gene expression. Projects like ENCODE, which aims at cataloguing all functional elements in the human genome [49], pioneered the way into the new era of functional genomics. Rather than investigating static properties of the genome such as the DNA sequence, this field focuses on the dynamic aspects genomics, mainly gene regulation and expression. Numerous examples of expression quantitative loci (eQTLs), where the expression level of a gene is associated with specific genetic variants, were reported [143, 144, 2]. Structural variants, such as CNVs, affect whole segments of the genome and therefore can have big effects on gene regulatory elements. Many cases of CNVs being associated with gene expression changes were reported for both diseases [145, 3] and regulatory variation in natural populations [4, 5]. Using gene expression levels as an intermediate phenotype between the genome and the organism, this approach allows a more profound understanding of the functional implications of genetic variation [146].
Another field in modern life science that focuses on dynamic aspects and interactions is systems biology, which aims at understanding biology by focusing on properties arising from the interaction of the systems component rather than the characteristics of isolated parts [147, 148]. Driven by the ever-increasing capabilities of quantitative molecular data and novel computational, statistical and mathematical analyses, systems biology became more relevant than ever. Quickly its potential for genomics was recognized and implementations of systems biology approaches were proposed to unravel the genomes function [149, 150]. This led to a field known as systems genetics, which takes advantage of integrating multiple quantitative molecular phenotypes, such as gene and protein expression, DNA-protein binding or metabolite levels, using a wide range of statistical methods. [151].
Following this philosophy, studies used quantitative, genome-wide molecular data to bridge the gap between genetic variation and gene expression regulation. Specific chromatin signatures, such as DNA-protein binding or histone modifications, have been linked to functional elements in the genome, illustrating their role in gene regulation [49]. Recent studies showed extensive genetic control of such signatures [50–52], highlighting their importance in understanding how genetic variants perturb molecular mechanisms and impact an individuals phenotype. Towards this goal, extensive integration of data on multiple molecular layers from different cell types and species in combination with sophisticated analytical methods will be crucial.

The work of this thesis

The goal of this thesis was to extend the knowledge of how genetic variation impacts on an organisms phenotype through perturbing gene expression regulation. This was done in three different projects, each of them investigating different aspects. The first project focuses on CNVs in Cynomolgus monkeys and their association with tissue specific gene expression levels. This project was carried out as collaboration with the pharmaceutical company Roche, where our role was to analyze the experimental data provided by them. This is to our knowledge the first study, which specifically investigates CNVs in this species and their potential implication for the animals.

The second and third projects were also collaborative efforts, but this time involving academic research groups. Both projects were focused the impact of genetic variation on gene expression changes through perturbation of gene regulatory elements. We generated genome-wide data on gene expression levels, protein-DNA binding events and histone modifications known to be linked to gene expression regulation. Using different computational analysis, we mapped extensive genetic effects on all measured molecular phenotypes. Our results suggest that genetic variants are likely to act through regulatory elements such as transcription factor binding sites, leading to a subsequent change in chromatin landscape and gene expression levels.

Chapter 1

Copy number variation in Cynomolgus monkeys linked to tissue specific gene expression

1.1 Preface

This first chapter focuses on CNVs and their tissue specific effects on gene expression levels in Cynomolgus monkeys (*Macaca fascicularis*) used in pharmaceutical studies. This study was conducted as a collaboration with the pharmaceutical company Roche, in which they provided the experimental data and we were in charge of the data analysis. The analytical results and text presented in this chapter are my contribution to this project. We assessed genome-wide copy number variation among 24 individuals using array comparative genome hybridization (aCGH) data combined with a customized analytical pipeline. Detected CNVs were then associated with gene expression levels of five different tissues (heart, kidney, liver, lung and spleen) obtained from gene expression array experiments in a *cis*-eQTL analysis framework. We find substantial copy number variation among the studied individuals, which is associated to tissue specific changes in gene expression levels. Of note is, that some of the CNVs are associated with expression changes of multiple genes within genomic regions.

1.2 Introduction

Copy number variations (CNVs) are genetic differences in the normal population displayed as microscopically invisible deletions or amplifications of stretches of genomic DNA ranging from 1 kilobase up to the megabase scale [1]. CNVs are commonly found in the genomes of humans

[6], primates [7], rodent [5], or even flies like Drosophila melanogaster [8]. In humans, more than 2.3 million different CNVs mapping to ∼200000 genomic regions have so far been identified [9]. They significantly contribute to genetic variation, covering more nucleotide content per genome than single nucleotide polymorphisms (e.g. approximately 0.8% of the length of the human genome differs between two human individuals) [10]. Furthermore CNVs exhibit a higher per-locus mutation rate than SNPs [11]. Since CNVs can reside in genomic regions harboring genes they can alter gene dosage, disrupt coding sequences or modify the level and timing of gene expression for genes within the CNV [12, 13] and on its flanks [3–5, 14–16]. These effects of CNVs are difficult to understand and not necessarily predictable, but relevant for many diseases [17–21] and pharmacological responses like in the case of CYP2D6 CNVs [22].

Cynomolgus monkeys (*Macaca fascicularis*) are well-established translational models for biomedical research and drug testing. These non-human primates are one of the closest animal model to humans with high genetic similarity (∼93% in nucleotide sequence identity), similar anatomies, and very similar physiologies [23–25]. These animals offer great promise as models for many aspects of human health and disease. Cynomolgus monkeys are outbred species, caught in the wild in many different places of peninsular Southeast Asia, the Philippines, and Mauritius, and used to found and continuously refresh breeding programs [25, 26]. They exhibit substantial levels of genetic variation which can affect the outcome and interpretation of biomedical studies [27–29]. Understanding of the contribution of this variation to phenotypes is lagging behind in Cynomolgus monkeys compared to the knowledge about human genetic and genomic variation [25]. Genome-wide catalogs of single nucleotide polymorphisms (SNPs) start to emerge for Cynomolgus monkeys with more and more genome sequencing projects published [23, 30–33]. However, information on structural variants, such as CNVs, is not available for Cynomolgus monkeys despite their prominent role in phenotypic variation. In this study, we assess for the first time genome-wide copy number variation among Cynomolgus monkeys from cohorts used in pharmaceutical studies using a custom 4.2 million probes CGH array. To investigate the potential functional implications of the detected copy number variation, we used a Cynomolgus monkey specific gene expression microarray to associate CNV genotypes with expression changes of proximal genes using a *cis* expression quantitative trait loci (*cis*-eQTL) mapping approach.

1.3 Material and Methods

1.3.1 Animal samples

All tissue samples used in this study were taken from untreated animals of GLP drug-safety studies in accordance with current animal welfare standards. Tissue samples (heart, kidney, liver, lung, spleen) for CGH and expression analysis were obtained from Cynomolgus monkey breeding centers located in the Philippines (4 females and 4 males), in Vietnam (2 males, 2 females), in China (4 females, originating from Southeast Asia), or in Mauritius (4 females and 4 males) (Figure 1.1A). Blood samples for CGH analysis were taken originated from individuals obtained from centers located in Mauritius (25 males). Details (gender, weight, age, origin) of all animals and their suppliers are on record and were part of the data submitted to public databases.

1.3.2 NimbleGen Gene Expression Analysis

Cynomolgus monkey tissues were homogenized in tubes prefilled with 1.4 mm ceramic beads and QiaGen's lysis reagent RLT using a FastPrep-24 instrument (MP Biomedicals, Solon, OH, USA). Total RNA from lysates was extracted using the RNeasy Mini kit combined with DNase treatment on a solid support (Qiagen Inc., Valencia, CA, USA). RNA quality assessment and quantification was performed using microfluidic chip analysis on an Agilent 2100 bioanalyzer (Agilent Technologies Inc., Santa Clara, CA, USA). On a Biomek FXp workstation (Beckman Coulter Inc., Brea, CA, USA), 10 ng of total RNA was used to prepare cDNA with the NuGen Ovation Pico WTA System V2 (NuGEN Technologies, Inc., SanCarlos, CA, USA), followed Cy3 labeling of cDNA with the Roche NimbleGen One Color DNA Labeling Kit. NimbleGen 12x135K gene expression microarrays were hybridized with 4 μg of Cy3-labeled cDNA for 16 h at 42°C and were washed and dried according to the manufacturer's instruction. Microarray data was collected by confocal scanning using the Roche NimbleGen MS200 Microarray scanner at 2 μm pixel resolution (Roche NimbleGen, Inc., Madison, WI, USA). NimbleGen probe intensities were subjected to Robust Multi-Array Analysis (RMA) with background correction and quantile normalization as implemented in the NimbleScan Software, version 2.6 (Roche NimbleGen, Inc., Madison, WI). Averaged gene-level signal intensities were summarized into gene calls and $\log_2$ transformed.

1.3.3 Comparative Genomic Hybridization Arrays

Cynomolgus monkey spleen tissues were homogenized in tubes prefilled with 1.4 mm ceramic beads and QiaGen's lysis reagent ALT using a FastPrep-24 instrument (MP Biomedicals, Solon, OH, USA) and then incubated with Proteinase K at 55°C for 1 h followed by RNAse A treatment at 25°C for 2 min (Qiagen Inc., Valencia, CA). Cynomolgus monkey blood specimens (200 μl) were incubated at 70°C for 10 min in QiaGen's lysis reagent ALT with Proteinase K and RNAse A. Genomic DNA from lysates was extracted using the QIAamp Mini kit (Qiagen Inc., Valencia, CA, USA). Assessment of unfragmented, high molecular weight DNA and quantification was performed using microfluidic chip analysis on an Agilent 2100 bioanalyzer (Agilent Technologies Inc., Santa Clara, CA, USA).0.5 μg of DNA from one animal tissue at a time and 0.5 μg of reference DNA - pooled DNA from blood specimens of 25 male Cynomolgus monkeys - were used for labeling by an isothermal Klenow fill-in reaction with either Cy3 or Cy5 random nonamer primer using the Roche NimbleGen Dual color labeling kit (Roche NimbleGen, Inc., Madison, WI). Labeling hybridization controls were spiked-in as quality controls for copy number variation detection (Roche NimbleGen, Inc., Madison, WI). NimbleGen 4.2M CGH microarrays were hybridized with 34 μg of Cy3- and 34 μg of Cy5-labeled DNA for 72 h at 42°C. After hybridization, microarrays were washed and dried according to the manufacturer's instruction, whereat 150 mM 1,3,5-Triaza-7 phospha-adamantane was included in the last washing step to avoid interference of ozone with the Cy5 dye during drying and scanning. Microarray data was collected by confocal scanning using the Roche NimbleGen MS200 Microarray Scanner at 2 μm pixel resolution (Roche NimbleGen, Inc., Madison, WI, USA).

1.3.4 aCGH normalization

aCGH probe intensities were subjected to LOESS spatial correction, background correction, and q-spline normalization as implemented in the NimbleGen DEVA software, version 1.2 (Roche NimbleGen, Inc., Madison, WI). The data was then additionally normalized for probe GC-content following [10]. To estimate the effect of probe GC-content on the measured log2 ratios, linear models were fitted for each array according to following formula:

$$log_2(R_i) = \alpha + \beta_1 GC_i + \beta_2 GC_i^2 + \epsilon_i \tag{1.1}$$

Where $\log_2(R_i)$ is the measured $\log_2$-ratio of an aCGH probe i, α the intercept, GC_i the probes GC content and ϵ_i a random error. The estimated effect of the probe GC-content was then subtracted from the measured $\log_2$-ratios (i.e. residualized). Furthermore the data was normalized for wave artifacts along chromosomes as described by [34]. This was done by fitting a local regression (LOESS) model for each chromosome and array separately to estimate the effect of chromosomal position on the measured $\log_2$-ratios:

$$log_2(R_i) = g(pos_i) + \epsilon_i \tag{1.2}$$

Where g is the local regression function, pos_i denotes the position of the probe i on the chromosome and ϵ_i a random error. Because the fraction of the data used in each local window (neighborhood) during model fitting is a crucial parameter, the normalization was performed across different fractions. The best was then selected based on signal-to-noise ratio (SNR) improvements before and after normalization using a CNV test set. The test set consisted of CNVs called based on the probe GC-content normalized data using all three callers with standard settings and the results were processed in the same way as the final CNV calls (see: 1.3.5). Only CNVs detected in at least 2 individuals were retained for more confident CNV calls. For each individual, the signal-to-noise ratio of each aCGH probe in each CNV of the test set was calculated in the following way:

$$SNR_i = \frac{|log_2(R_i)|}{\sigma_{Ci}} \tag{1.3}$$

SNR_i denotes the signal-to-noise ratio of a given probe i, and σ_{Ci} the standard deviation of all probes on the same chromosome as probe i. The average SNR of all CNV probes per CGH array was used as metric to evaluate the normalization performance. To visualize and to assess the quality of the normalized data, principal component analysis (PCA) and hierarchical clustering of the euclidean distance between $\log_2$-ratios of samples were used.

1.3.5 CNV calling

CNV calling was performed 3 inherently different approaches to mitigate method specific errors: R-GADA [35] was used with the following parameters: alpha=0.2, T=4.5, minseglen=5. DNAcopy [36] was used with minseglen=5, undosd=3, undoprune=0.05 and data smoothing was applied prior to CNV calling. CopyMap [37] was used with r=20, T=4, m=5, a=2.1, P=0.001. Further for R-GADA and DNAcopy z-scores were calculated for all CNV calls based on the mean $\log_2$-ratio of the CNV, and only CNVs with z-scores >1.5 or <-1.5 retained. For CNVs called by CopyMap a carrier probability of at least 0.8 was required. The three obtained CNV calling profiles per individual were then merged and only CNVs called by at least two methods were kept, and loci with conflicting copy number states were removed. These resulting profiles were then further merged between individuals to obtain CNV regions that could be genotyped across individuals. In cases where an individual carried more than one CNV in a CNV region, the locus was marked as a complex locus and removed from subsequent steps. Additionally

CNV loci located on the X chromosome or within array probe gaps larger than 500 kb +/- 250 kb (e.g. centromeres) were removed. To avoid potential calling mistakes the median $\log_2$-ratio of each CNV was used as genotype rather than the discrete copy number state provided by the CNV calling methods. PCA was again used to visualize and evaluate the inferred CNV genotypes.

1.3.6 eQTL analysis

To assess the potential functional impact of copy number variants, we associated the inferred CNV genotypes with the expression level of proximal genes in each of the five tissues by using a *cis* expression quantitative trait loci (*cis*-eQTL) approach. No complex CNV loci were used for that purpose and in order to avoid outlier driven results only CNVs called in at least two individuals were retained. The expression level of each gene was tested for associations with CNVs within 1Mb of its transcription start site (TSS) using following linear model:

$$E_{it} = \alpha + \beta_1 C_i + \beta_2 G_{it} + \beta_3 A_i + \epsilon_{it} \tag{1.4}$$

Where E_{it} is the measured expression level of gene E in tissue t for the i^{th} individual, C_i is the genotype of a proximal CNV C for the i^{th} individual and ϵ_{it} a random error. To account for non-genetic systematic variation between samples, the loadings of the first principal component for individual i for both the expression levels of all genes the same tissue (G_{it}) and the genotypes of all CNVs (A_i) were added as covariates. An adapted version of the fastQTL software [38] was used to test all possible associations using this model. Correction for multiple testing was carried out in two steps, where first local permutations were applied to correct for multiple variants per gene [38] and then the false discovery rate (FDR) (qvalue R-package, Storey J., 2015) was calculated per tissue to account for multiple tested genes. Only eQTLs below an FDR of 10% (qvalue <0.1) were considered as significant. To further investigate the impact of CNVs on the gene expression landscape, genes within the regions of detected eQTLs were investigated for further associations with the eQTL CNV. The expression levels of all genes within 1Mb from the TSS of an eQTL gene were tested for an association with the eQTL CNV with the same linear model as used for eQTL mapping. Bonferroni correction was calculated for all tested associations per region and association with a corrected p-value <0.05 were considered significant.

1.4 Results

1.4.1 NimbleGen Gene Expression Analysis

After RMA normalization gene expression data of each array were identically distributed, with identical median and standard deviation (Figure A.1). PCA generally did not reveal much separation of the gene expression data according to sample origin or array scan day (Figure A.2). Hierarchical clustering on the other hand revealed some limited clustering according to origin (Figure A.3).

1.4.2 aCGH normalization

The GC-content normalization estimated the proportion of the variance explained by the aCGH probe GC-content range from 0.0009 and 0.1501 depending on the array (R-squared, mean=0.065, SD=0.045, Table A.1). For the wave artifacts normalization, a fraction of 4000 probes per model fitting step resulted in the largest median SNR improvement (1.1%) and was therefore chosen (Figure A.4, Table A.2). When plotting the loadings of the first and second principal component and color-coding them according to sample origin and array scan date, the samples clustered mostly according to their origin (Figure A.5A). No strong clustering based on thee array scan date was observed expect for the two samples sI01776_F and s7828C_M, which both were separated from the other samples (Figure A.5B). The same separation was observed by hierarchal clustering of the aCGH data, where samples generally were grouped by geographical origin, except for three including sI01776_F and s7828C_M (Figure A.6). Therefore the samples sI01776_F and s7828C_M were defined as outliers and excluded from all following analyses.

1.4.3 CNV calling

The combination of the three CNV calling methods reported between 1,364 and 6,598 (mean= 3,116, SD=1,356) CNVs per individual with on average slightly more duplications (1,692) than deletions (1,424) (Figure 1.1B, Table A.3). These CNV calls led to a total of 17599 CNV regions after filtering, with only 292 regions excluded because they were classified as complex loci. Visualization of the first and second principle component of a PCA based on these CNV calls revealed a clear separation of sample sC30659_M (Figure A.7). This individual also showed a large number of deletion CNV calls (Figure 1.1B, Table A.3), resulting in an excess of deletions in the inferred CNV regions (Figure A.8). Based on these results, sC30659_M was also defined as outlier and removed from the data set, reducing the number of CNV regions to 15,183. The length of these loci ranged from 2.3 to 692.9 kb, with a median length of 8.35 kb (SD=15.1

kb) (Figure 1.1C) and in total, these CNV regions covered ∼4% (∼127 Mb) of the autosomal Cynomolgus monkey genome (Table A.4). 58% of CNV regions were detected in only 1 population, however 83% of them were restricted to one individual. 19% of CNV regions were found in two, 11% in three and 12% in all four populations (Figure 1.1D). PCA and hierarchical clustering based on the median $\log_2$-ratio of the resulting CNV regions showed clear separation of the samples by geographical origin (Figures 1.1E&F).

1.4.4 eQTL analysis

A total of 7,266 non-complex CNV loci with a minimal allele count of 2 were associated with the expression levels of 18,280 genes measured in all five tissues. The eQTL mapping reported 30 *cis*-eQTLs across all five tissues, ranging from one to eight *cis*-eQTLs per tissue (Figure 1.2A, Table A.5). Generally, eQTL genes showed lower average gene expression levels than the tissue average (Figure 1.2B), however this difference was only significant in heart eQTLs (Wilcoxen rank sum test, p = 0.015). The strongest associations were generally observed with CNVs in close distance of the TSS (Figure 1.2C), and the highest density of associations was observed around 200 kb upstream of the TSS (Figure 1.2D). Further investigation of associations within eQTL regions revealed a total of 10 additional associations in eight out of the 29 non-overlapping eQTL regions. Within these eight regions, on average 12.8% of genes were also associated with the eQTL CNV and all associations showed the same directionality as the eQTL. Among the most significant associations we found a group of olfactory receptor (OR) genes (OR4K17, OR5M9) on chromosome 7 and 14 as well as the ATP-binding cassette transporter 4 (ABCB4) on chromosome 3, also known as multidrug resistance protein 3 (MDR3) (Table A.5). In close proximity to the OR genes on chromosome 7 we detected a duplication event associated with expression changes of ORK17 in kidney and in lung and ORK14 in kidney (Figures 1.3A&B, A.9). Further investigation of this eQTL region revealed additional associations with ORK13 in lung and OR4L1 in both kidney and lung. For ABCB4 we detected a deletion ∼480 kb upstream associated with increased transporter expression in lung (Figures 1.3C, A.10).

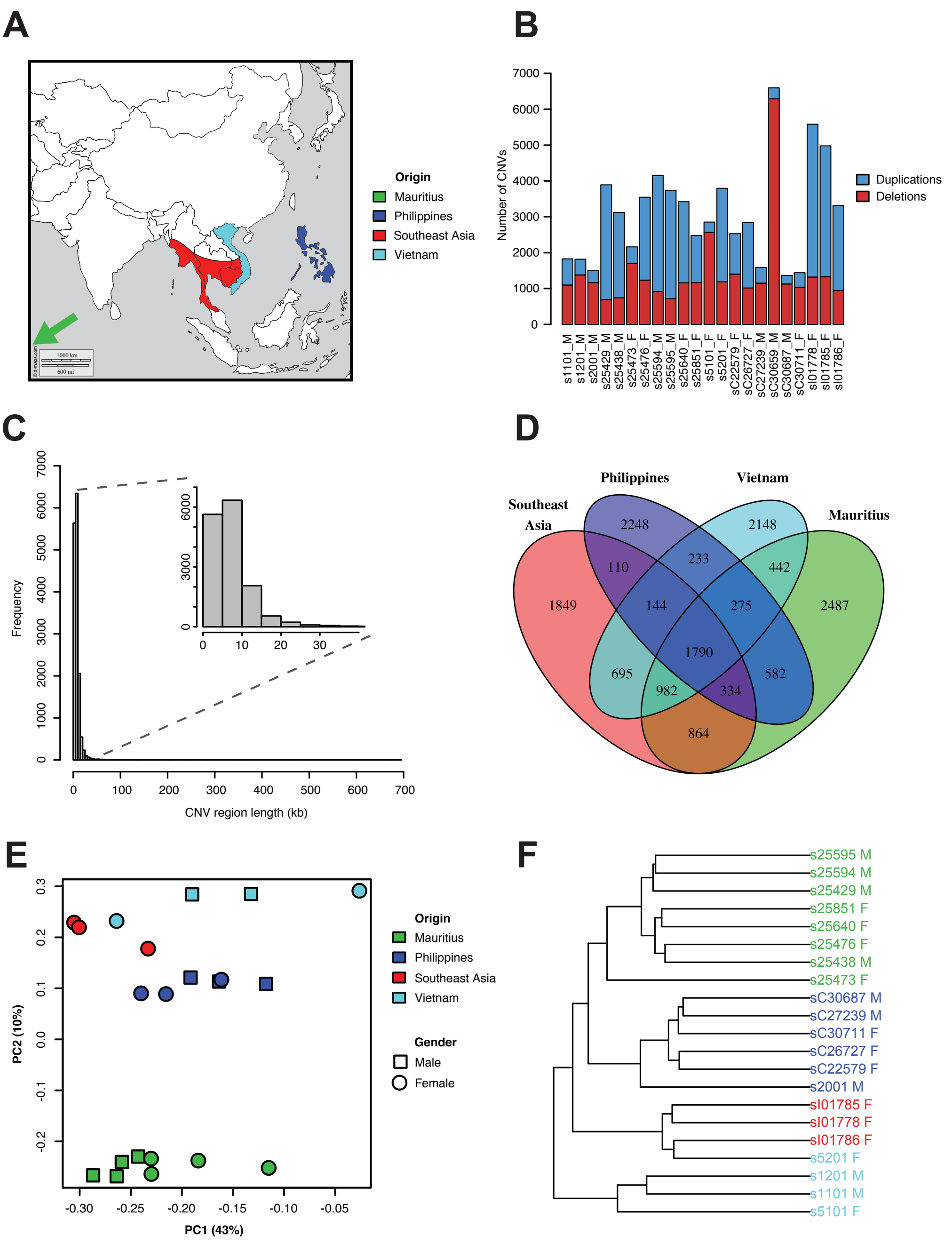


Figure 1.1: **CNV genotypes. A** Geographic origin of the four natural populations, from where the tested cynomolgus monkeys were caught. **B** Number of duplications and deletions detected per individuals by combining the three CNV calling approaches. **C** Size distribution of the inferred CNV regions across 21 individuals (n=15,183). **D** Number of CNV regions detected among and across the four different populations. **E** Loadings of the first and second principal component based on a PCA performed on the $\log_2$- ratio genotypes of all CNV regions (n=15,183) in the 21 individuals. **F** Hierachical clustering of the $\log_2$- ratio genotypes of all CNV regions (n=15,183) in the 21 individuals.

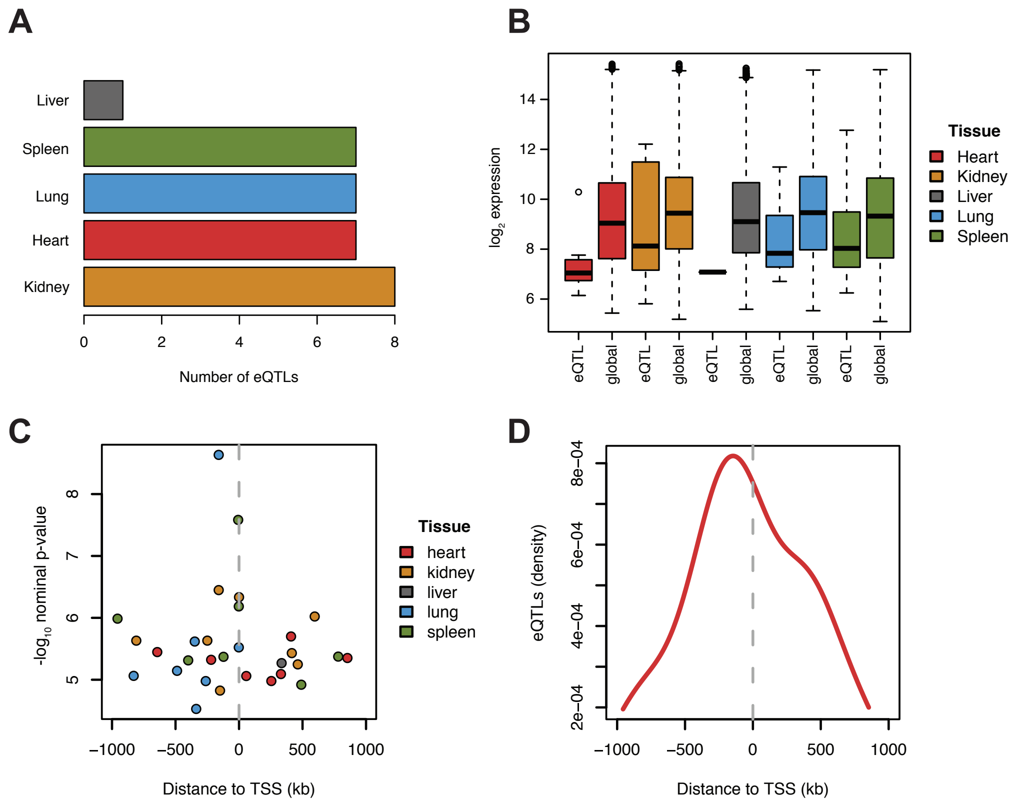


Figure 1.2: **eQTLs.** **A** Number of detected *cis*-eQTL per tissue under 10% false discovery rate (FDR). **B** Average expression levels of eQTL genes in each tissue versus the average expression level of all genes in the respective tissue. **C** Nominal p-value of all detected *cis*-eQTL as a function of the distance to the transcription start site (TSS) of the eQTL CNV to its associated gene. **D** Density of detected *cis*-eQTLs as a function of the distance to the transcription start site (TSS) of the eQTL CNV to its associated gene.

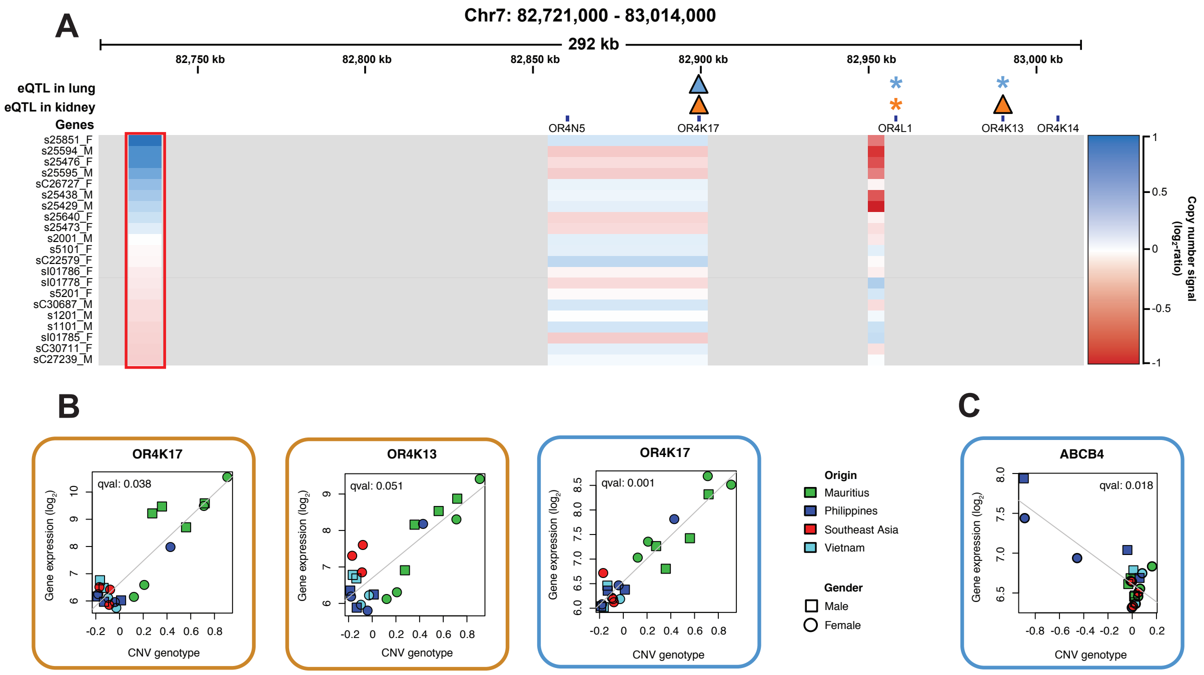


Figure 1.3: **eQTL loci.** **A** eQTL region on chromosome 7 containing olfactory receptor *cis*-eQTLs in both kidney (orange) and lung (blue). The red box highlights the CNV locus, which shows duplication events associated with gene expression changes of proximal olfactory receptor genes. Triangles indicate an association reported from the genome-wide *cis*-eQTL mapping, while stars indicate additional associations revealed by the eQTL region analysis. **B** eQTL associations for the three detected olfactory receptor *cis*-eQTLs on chromosome 7 in kidney and lung. CNV genotype represents the median $\log_2$-ratio of aCGH probes within the CNV. **C** eQTL association of ABCB4 with a close by deletion on chromosome 3 in lung.

1.5 Discussion

Using a microarray based CNV detection approach combining careful data normalization and three different CNV calling methods, we are able to assess copy number variation in Cynomolgus monkeys originating from different natural populations. According to our knowledge, this is the first study to investigate natural copy number variation in this species. We predominately find small CNVs among our individuals with a median length of 8.35 kb (SD=15.1 kb), which is by far smaller than the median gene locus size of 46.7 kb in Cynomolgus monkeys. This finding is in line with current research [39], which suggests that individuals from normal, healthy populations carry mostly short CNVs. In humans, short CNVs are more frequently generated de novo than large CNVs (>500 kb) [40], which indicates that they do not underlie strong purifying selection in contrast to potential deleterious large CNVs. This also highlights the importance of a meticulous CNV calling approach when using aCGH data, since we operate close to the resolution limit of the array with many CNVs only encompassing 5 array probes. Genotyping an individual for a defined CNV region is a difficult task, because it requires discretization of the continuous $\log_2$-ratio spectrum. We solve this problem by using the median $\log_2$-ratio of all probes within a CNV of an individual as its genotype, which avoids the problem of having to make a discrete statement about the copy number.

We are confident that our CNV genotypes represent true copy number variation, which is underlined by the clustering of the CNV genotypes by genetic background of our individuals (Figure 1.1F). As expected the island populations of Mauritius and the Philippines are clearly separated, while the separation between the Southeast Asian and Vietnamese main land populations is less pronounced. This is in line with the fact that these two populations share geographically adjacent biotops [25]. Additionally, individuals labeled as Southeast Asian might come from geographically less separated populations in Vietnam, Cambodia or Laos. Interestingly, the majority (83%) of population specific CNV regions were only detected in one individual, indicating that much of the population specific CNV loci are found at low frequency within the respective populations. Our findings indicate that one might face very different genetic backgrounds in animal experiments, depending on the population origin of the test animals.

However, even though we find extensive copy number variation among our samples, only a relatively small number of CNV loci are associated with gene expression level changes. This suggests that most copy number variation, similar to single nucleotide variants, have no effect on or link to gene expression regulation. However, given our small sample size of 21 individuals used for the *cis*-eQTL mapping, our statistical power is relatively low for genome-wide testing, which might also partially explain the low number of detected eQTLs. When looking at the association strength of our eQTLs, it becomes evident that the strongest associations are detected with variants close to their targets TSS (Figure 1.2C), which is in line with previous research [2, 4, 41]. This suggests, that for CNVs similar as for SNPs large-effect variants affect *cis*-regulatory elements in the immediate intergenic regions [2]. As expected we detect most significant associations close to the TSS (Figure 1.2D), maybe not because of the absence of more distal associations, but because these large-effect variants close to the TSS are easier to detect. Generally eQTL genes seem to show lower expression levels than the average gene expression level of the respective tissue, even though this difference is significant only in heart. This indicates that the expression of genes, which are required to be expressed at high level in a given tissue, is more tightly regulated. On one hand, this might be because purifying selection removes detrimental regulatory variants for these genes. On the other hand, buffering of genetic effects within the regulatory networks might be the case and the transcriptional machinery might compensate genetic effects if needed. When investigating the eQTL regions for additional associations, we discover additional genes for which the expression level is linked to CNV genotypes. These associations however do not pass a genome-wide correction for multiple testing, suggesting that indeed we would probably detect more *cis*-eQTLs with a larger sample size. Furthermore, these results also show that CNVs are regularly associated to the expression level of multiple genes within a genomic region. This might be because CNVs encompassing several kb potentially have a strong effect on regulatory elements and can easily affect multiple regulatory elements. Therefore, this finding could also be caused by other variants linked to the CNV and the resulting haplotype dependent associations.

One particularly interesting eQTL region is located on chromosome 7, where a duplication event located in proximity of a group of olfactory receptor (OR) genes affecting expression changes of OR4K13 and ORK17 in kidney and ORK17 in lung (Figure 1.3A&B). Although olfactory receptors are typically not expected to be expressed in internal organs, a recent study gave examples of such receptors to be expressed in organs involved in metabolic processes [42]. The validity of our findings is strengthened by the fact that we detect one of these associations in two independent tissues. Considering the expression level change of these OR genes from a very low level to a level close to the average global gene expression in duplication carriers led us to the hypothesis, that a copy number change in a *cis*-regulatory element might be responsible for activating gene transcription at this locus. When investigating this particular eQTL region for further associations, we detect additional associations of the CNV with OR4L1 in kidney and OR4L1, OR4K2 and OR4K13 in lung. This serves as a prime example of a CNV being linked to expression changes of multiple genes. Interestingly, additional CNVs were detected within this region (Figure 1.3A). A ∼50 kb CNV comprising both the OR4N5 and OR4K17 genes was detected, but was not associated to gene expression changes. Given the moderate change in the aCGH $\log_2$-ratio, it is possible that this CNV represents an amplification change in an array of copies, as such a change would lead to a weaker signal than expected from a classical duplication or deletion event. OR genes are well known to be highly variable in copy number among humans [43] and multiple genes have been reported to exist in high copy numbers [44]. Additionally, a deletion upstream of OR4L1 was detected in Mauritian individuals (Figure 1.3A), which at first seems to be in linkage with the duplication associated with gene expression changes. However, it is absent from one Filipino individual showing elevated expression levels of the OR genes, which excludes an association of this CNV with gene expression. Even though this is based on only one data point, it demonstrates the potential of breaking up the linkage between variants by including individuals from different populations. Interestingly, this genomic region was previously identified to be the origin of a chromosomal fission in the hominoid lineage, giving rise to the human chromosomes 14 and 15 [45, 46]. We show copy number and gene expression polymorphisms in the ancestral form, maybe linked to the genomic instability leading to the fission event.

Among our eQTLs, we also find ABCB4, which is known to act as tumor suppressor once overexpressed in lung cancer [47], and was shown to be regulated by epigenetic silencing. A deletion ∼480 kb upstream of this gene was linked to its expression, where it might disrupt such epigenetic silencing of ABCB4 and thus increase its expression (Figure 1.3C). Whether any of our detected associations are of physiological relevance in pharmaceutical studies remains unclear, however we find a link between CNVs and gene expression levels in organs such as the kidney, which plays a major role in drug excretion.

In summary, we detect substantial copy number variation in Cynomolgus monkey populations used in pharmaceutical studies, leading to a diverse and variable genetic background in such studies. We report several associations of CNVs with the expression levels of proximal genes. In some cases multiple genes within the same region are linked to the same CNV. Of note is a genomic region, which harbors several olfactory receptor genes showing an association with a close-by duplication event in both kidney and lung. Even though the physiological consequences remain unclear, our data suggests that CNVs shape the tissue transcriptomes of vitally important organs.

1.6 References

1. Scherer, S. W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nature genetics* **39,** S7–15 (July 2007).
2. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nature genetics* **39,** 1217–1224 (Oct. 2007).
3. Merla, G. *et al.* Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. *American Journal of Human Genetics* **79,** 332–341 (Aug. 2006).
4. Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315,** 848–853 (Feb. 2007).
5. Henrichsen, C. N. *et al.* Segmental copy number variation shapes tissue transcriptomes. **41,** 424–429 (Mar. 2009).
6. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–454 (Nov. 2006).
7. Gokcumen, O. & Lee, C. Copy number variants (CNVs) in primate species using array-based comparative genomic hybridization. *Methods* **49,** 18–25 (Sept. 2009).
8. Cardoso-Moreira, M., Arguello, J. R. & Clark, A. G. Mutation spectrum of Drosophila CNVs revealed by breakpoint sequencing. *Genome biology* **13,** R119 (2012).
9. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic acids research* **42,** D986–92 (Jan. 2014).
10. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (Apr. 2010).
11. Lupski, J. R. Genomic rearrangements and sporadic disease. *Nature genetics* **39,** S43–7 (July 2007).
12. Henrichsen, C. N., Chaignat, E. & Reymond, A. Copy number variants, diseases and gene expression. *Human molecular genetics* **18,** R1–8 (Apr. 2009).
13. Chaignat, E. *et al.* Copy number variation modifies expression time courses. *Genome research* **21,** 106–113 (Jan. 2011).
14. Cahan, P., Li, Y., Izumi, M. & Graubert, T. A. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nature genetics* **41,** 430–437 (Apr. 2009).

15. Orozco, L. D. *et al.* Copy number variation influences gene expression and metabolic traits in mice. *Human molecular genetics* **18,** 4118–4129 (2009).

16. Ricard, G. *et al.* Phenotypic consequences of copy number variation: insights from Smith-Magenis and Potocki-Lupski syndrome mouse models. *Plos Biology* **8,** e1000543 (2010).

17. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307,** 1434–1440 (Mar. 2005).

18. Aitman, T. J. *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439,** 851–855 (Feb. 2006).

19. Gilman, S. R. *et al.* Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70,** 898–907 (June 2011).

20. Levy, D. *et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70,** 886–897 (June 2011).

21. Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70,** 863–885 (June 2011).

22. Johansson, I. & Ingelman-Sundberg, M. Genetic polymorphism and toxicology–with emphasis on cytochrome p450. *Toxicological sciences : an official journal of the Society of Toxicology* **120,** 1–13 (Mar. 2011).

23. Ebeling, M. *et al.* Genome-based analysis of the nonhuman primate Macaca fascicularis as a model for drug safety assessment. **21,** 1746–1756 (Oct. 2011).

24. Shively, C. A. & Clarkson, T. B. The unique value of primate models in translational research. Nonhuman primate models of women's health: introduction and overview. *American journal of primatology* **71,** 715–721 (Sept. 2009).

25. Haus, T. *et al.* Genome typing of nonhuman primate models: implications for biomedical research. *Trends In Genetics* **30,** 482–487 (Nov. 2014).

26. Stevison, L. S. & Kohn, M. H. Determining genetic background in captive stocks of cynomolgus macaques (Macaca fascicularis). *Journal of Medical Primatology* **37,** 311–317 (Dec. 2008).

27. Liu, Y. Y. *et al.* Polymorphisms of CD3epsilon in cynomolgus and rhesus monkeys and their relevance to anti-CD3 antibodies and immunotoxins. *Immunology and cell biology* **85,** 357–362 (July 2007).

28. Menninger, K. *et al.* The origin of cynomolgus monkey affects the outcome of kidney allografts under Neoral immunosuppression. *Transplantation proceedings* **34,** 2887–2888 (Nov. 2002).

29. Drevon-Gaillot, E., Perron-Lepage, M.-F., Clément, C. & Burnett, R. A review of background findings in cynomolgus monkeys (Macaca fascicularis) from three different geographical origins. *Experimental and toxicologic pathology : official journal of the Gesellschaft für Toxikologische Pathologie* **58,** 77–88 (Nov. 2006).

30. Yan, G. *et al.* Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nature Biotechnology* **29,** 1019–U89 (Nov. 2011).

31. Ericsen, A. J. *et al.* Whole genome sequencing of SIV-infected macaques identifies candidate loci that may contribute to host control of virus replication. *Genome biology* **15,** 478 (2014).

32. Osada, N., Hettiarachchi, N., Adeyemi Babarinde, I., Saitou, N. & Blancher, A. Whole-genome sequencing of six Mauritian Cynomolgus macaques (Macaca fascicularis) reveals a genome-wide pattern of polymorphisms under extreme population bottleneck. *Genome biology and evolution* **7,** 821–830 (Mar. 2015).

33. Higashino, A. *et al.* Whole-genome sequencing and analysis of the Malaysian cynomolgus macaque (Macaca fascicularis) genome. *Genome biology* **13,** R58 (2012).

34. Marioni, J. C. *et al.* Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome biology* **8,** R228 (2007).

35. Pique-Regi, R., Cáceres, A. & González, J. R. R-Gada: a fast and flexible pipeline for copy number analysis in association studies. *BMC bioinformatics* **11,** 380 (2010).

36. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics (Oxford, England)* **23,** 657–663 (Mar. 2007).

37. Zöllner, S. CopyMap: localization and calling of copy number variation by joint analysis of hybridization data from multiple individuals. *Bioinformatics (Oxford, England)* **26,** 2776–2777 (Nov. 2010).

38. Ongen, H., Buil, A., Brown, A., Dermitzakis, E. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *bioRxiv,* 022301 (Aug. 2015).

39. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349,** aab3761–aab3761 (Sept. 2015).

40. Itsara, A. *et al.* De novo rates and selection of large copy number variation. *Genome research* **20,** 1469–1481 (Nov. 2010).

41. Stranger, B. E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS genetics* **8,** e1002639 (2012).

42. Wu, C. *et al.* Activation of OR1A1 suppresses PPAR-γ expression by inducing HES-1 in cultured hepatocytes. *The international journal of biochemistry & cell biology* **64,** 75–80 (July 2015).

43. Young, J. M. *et al.* Extensive copy-number variation of the human olfactory receptor gene family. *American Journal of Human Genetics* **83,** 228–242 (Aug. 2008).

44. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330,** 641–646 (Oct. 2010).

45. Rudd, M. K. *et al.* Comparative sequence analysis of primate subtelomeres originating from a chromosome fission event. *Genome research* **19,** 33–41 (Jan. 2009).

46. Ventura, M. *et al.* Neocentromeres in 15q24-26 map to duplicons which flanked an ancestral centromere in 15q25. *Genome research* **13,** 2059–2068 (Sept. 2003).

47. Kiehl, S. *et al.* ABCB4 is frequently epigenetically silenced in human cancers and inhibits tumor growth. *Scientific Reports* **4,** 6899 (Nov. 2014).

Chapter 2

Coordinated allelic variation across molecular phenotypes

2.1 Preface

Recent advances in genome-wide profiling of transcription factor (TF) binding and chromatin state have identified specific chromatin signatures related to various classes of functional elements in different cell types. However, their genetic basis and degree of variability across individuals remain largely unknown. We studied genome-wide enrichment profiles of transcription factor binding, chromatin marks, and different measures of transcription in lymphoblastoid cell lines from two human trios and eight unrelated samples. We quantified inter-individual variability in these phenotypes to understand both DNA sequence dependent and independent variation on DNA binding and transcription, chromatin state, and their interplay in an allele-specific framework. We find that different organizational layers of the genome show abundant allelic effects and strong allelic coordination between layers, with the genetic control of this coordination acting primarily through transcription factor binding.

This project in this chapter was carried out as a collaboration among four academic groups from Lausanne (Prof. Reymond, Prof. Deplancke, Prof. Hernandez) and Geneva (Prof. Dermitzakis). My contribution was in the analysis of the data, especially in ChIP-seq and RNA-seq data analysis, putative enhancer inference and how the different molecular phenotypes co-vary at functionally important elements of the genome. This study was published in the peer-reviewed journal *Science* in November 2013 [53].

Molecular phenotypes at functional elements

In order to assess the behavior of the measured molecular phenotypes at genomic regions linked to gene expression regulation, I defined promoter and putative enhancer loci. Promoter regions were defined as a 2.5 kb window centered on the transcription start sites of protein-coding and linc-RNA genes (n=13,720). Putative enhancers were created based on DNaseI hypersensitivity sites available for the trios [51]. DNAseI hypersensitivity sites were merged and sites outside exons and promoter regions of known transcripts were considered as putative enhancer sites. ChIP, mRNA and nascent transcription (GRO-seq) reads were then quantified within these promoter and putative enhancer sites and read counts were normalized across assays by calculating z-scores of the $\log_{10}$ transformed read counts. Based on these quantifications, spearman's rank correlation coefficients (ρ) were then calculated for each marker combination. We observe a clearly coordinated behavior among all our phenotypes at promoters of protein-coding and linc-RNA genes (Figure 2.1A). Most molecular phenotypes were positively correlated with each other except for H3K27me3, which was negatively correlated with the other phenotypes. This indicates the repressive function of H3K27me3, while the other molecular phenotypes are known to be associated with active transcription. Similar, yet weaker behavior was also found for putative enhancer sites. These results highlight the coordinated activity of chromatin phenotypes, transcription and gene expression at functionally relevant elements in the genome and suggest common functional implications.

Molecular phenotypes and gene expression regulation

Next, I investigated how the measured molecular phenotypes are linked to gene expression levels. The obtained ChIP-seq and Gro-seq read quantifications for promoter regions were normalized to 10,000,000 total mapped reads in each experiment and the promoters were grouped into percentiles according to the expression level of their genes. For each percentile the average RNA-seq quantification value and the number of ChIP-seq and GRO-seq reads were calculated for each marker. Obtained values for every percentile were plotted on a $\log_{10}$ scale. As expected we observed that all phenotypes except H3K27me3 were positively correlated with gene expression levels (Figure 2.1C). Our findings show that the measured chromatin phenotypes within promoter regions are linked to the expression levels of the associated gene. This suggests that these chromatin phenotypes are likely to be involved in the regulation of gene expression at these loci.

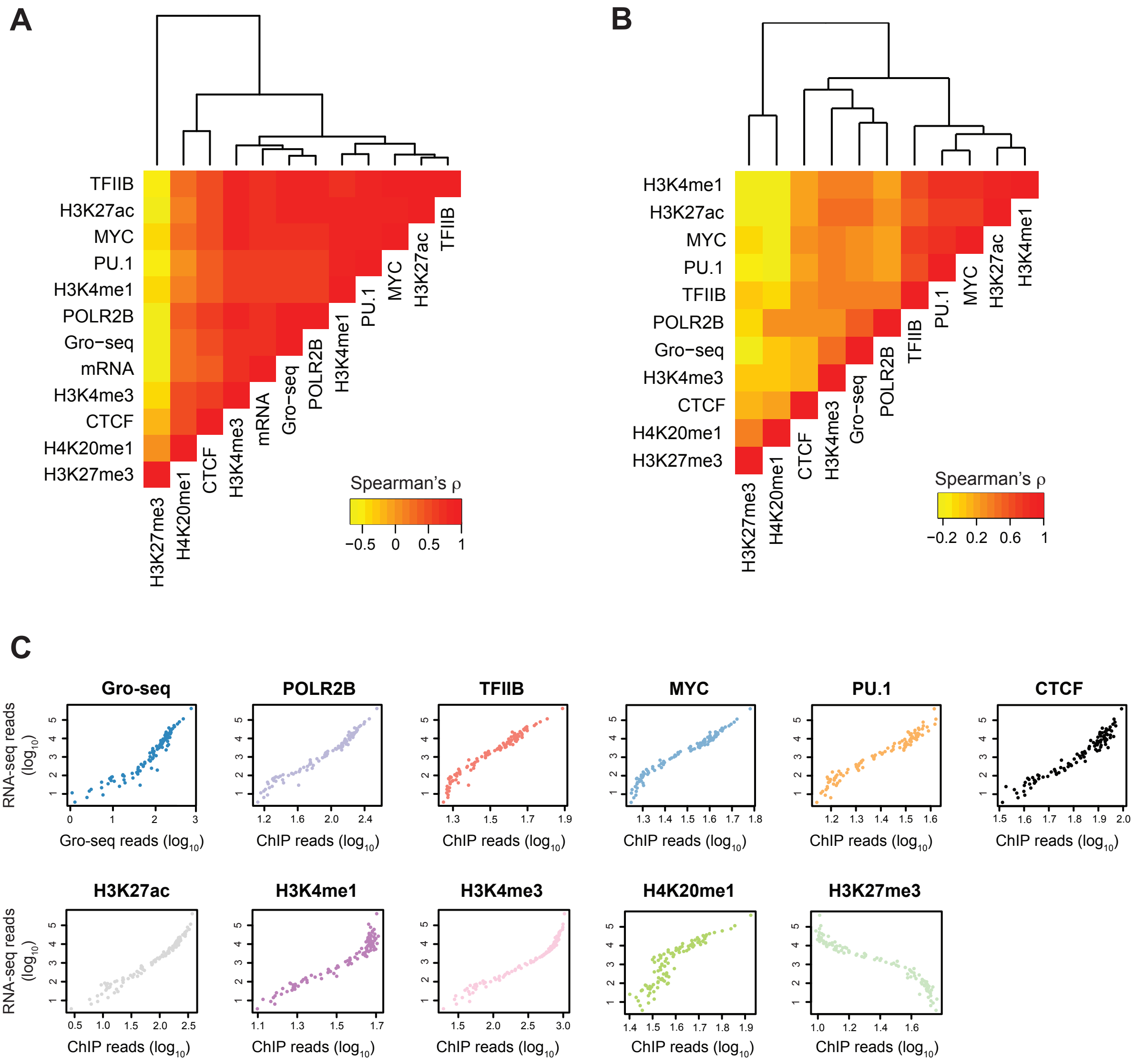


Figure 2.1: **Molecular phenotypes and gene expression. A,B** Genome-wide properties of the probed molecular phenotypes. Correlation of molecular marks at promoters (transcription start sites +/- 2.5 kb) for protein-coding and linc-RNA genes **A** and putative enhancers defined by DNaseI hypersensitivity sites **B**. Plotted values are Spearman correlation coefficients based on z-score transformed read densities for ChIP, mRNA and nascent transcription (GRO-seq) assays. **C** Relationship between gene expression (mRNA-seq) and genomic signals at promoters (transcription start site +/- 2.5 kb) of protein-coding and linc-RNA genes. Genes were grouped into percentiles according to their expression level and the average expression level and read density is shown for each percentile.

Article

Coordinated allelic variation across molecular phenotypes

Kilpinen H.[1,2*], Waszak S. M.[2,3*], Gschwind A. R.[2,4,*], Raghav S. K.[3], Witwicki R. M.[4], Orioli A.[4], Migliavacca E.[2,4], Wiederkehr M.[4], Gutierrez-Arcelus M.[1,2], Panousis N. I.[1,2], Yurovsky A.[1,2], Lappalainen T.[1,2], Romano-Palumbo L.[1], Planchon A.[1], Bielser D.[1], Bryois J.[1,2], Padioleau I.[1,2], Udin G.[3], Thurnheer S.[5], Hacker D.[5], Core L. J.[6], Lis J. T.[6], Hernandez N.[4], Reymond A.[4], Deplancke B.[2,3], and Dermitzakis E. T.[1,2]

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland

[2]Swiss Institute of Bioinformatics SIB, 1015 Lausanne, Switzerland

[3]Institute of Bioengineering, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland

[4]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland

[5]Protein Expression Core Facility, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland

[6]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14850, USA

[*]These authors contributed equally to this work

2.2 Abstract

DNA sequence variation has been associated with quantitative changes in molecular phenotypes such as gene expression, but its impact on chromatin states is poorly characterized. To understand the interplay between chromatin and genetic control of gene regulation we quantified allelic variability in transcription factor binding, histone modifications, and gene expression within humans. We found abundant allelic specificity in chromatin and extensive local, short-, and long-range allelic coordination among the studied molecular phenotypes. We observed genetic influence on most of these phenotypes, with histone modifications exhibiting strong context-dependent behavior. Our results implicate transcription factors as primary mediators of sequence-specific regulation of gene expression programs, with histone modifications frequently reflecting the primary regulatory event.

2.3 Main text

Functional genomic elements have been linked to specific chromatin signatures in different cell types [49], illustrating control of transcriptional processes through multiple layers of genome organization. While allele-specific gene expression is widespread [54], it has been difficult to pinpoint the upstream *cis*-regulatory variants and how they affect chromatin states. We performed chromatin immunoprecipitation (ChIP) of five histone post-translational modifications (hPTMs) (H3K4me1, H3K4me3, H3K27ac, H3K27me3, and H4K20me1), three transcription factors (TFs) (TFIIB, PU.1, and MYC), and the second largest RNA polymerase II subunit RPB2 [POLR2B] in lymphoblastoid cell lines (LCLs) (Figure B.1) in two parent-offspring trios [48]. A subset of the ChIP assays was also performed in eight additional unrelated individuals. We further profiled one of the trios with global run-on sequencing (GRO-seq), which measures nascent transcription at all transcribed regions (Figure B.2), and examined available DNaseI-seq and CTCF ChIP-seq data [51]. All 14 individuals were additionally profiled for messenger-RNA (mRNA) expression (supplementary information). Clustering of the molecular phenotypes along promoters and enhancers was consistent with published reports [49] (Figures B.3 - B.5).

We identified sites of allele-specific (AS) TF binding, hPTM, and transcription for all assays (supplementary information), ranging from 11-12% for TFs [51, 55] to 6-30% for hPTMs at heterozygous sites accessible for the analysis (median across all individuals) (Figure 2.2A, Figure

B.6). Notably, in the two trios, fewer AS effects were observed in mRNA (mRNA-seq, 5%) than in nascent transcripts (GRO-seq, 27-28%) (supplementary information), likely reflecting post-transcriptional modifications.

Multiple heterozygous SNPs overlapping regions of TF activity showed high consistency in allelic direction within individuals (Figure 2.2B, Figure B.7A, B.7B). AS consistency in nascent transcription and histone modifications was high up to several kb and decreased with distance (logistic regression, P <0.05, Figure B.7C). Strongest AS effects were enriched at promoters, while the allelic signals of marks of enhancer activity (PU.1, H3K4me1, H3K27ac) or heterochromatin (H3K27me3) showed a more dispersed distribution (Figure B.8). We also analyzed all accessible heterozygous SNPs overlapping known eQTLs from the 1000 genomes phase1 populations (supplementary information) [56] and observed an enrichment of allelic bias at eQTLs compared to non-eQTLs for TFs (P=0.016, Mann-Whitney U test) but not for hPTMs (Figure B.9), suggesting that a TF binding change is often causal to the gene expression change.

Linking hPTM signatures with specific DNA sequence features has proven difficult [57], but for sequence-specific TFs it is possible to assess whether the observed AS effects are due to motif-disrupting variants (Figure B.10). Categorization of significant AS binding sites, with respect to predicted TF motifs, revealed three classes of binding SNPs (B-SNPs): B-SNPs located either within (class I) or adjacent (class II) to predicted PU.1 and MYC consensus TF motifs, or B-SNPs in motif-devoid peaks (class III). Class I sites were enriched for B-SNPs compared to the other two classes (Figure B.11A, B.11B for PU.1, Figure B.12A, B.12B for MYC), suggesting that SNP-mediated disruption of the TF motif is likely causal to the observed AS binding activity. However, most TF AS binding events (70%, PU.1; 97%, MYC) appear triggered through TF consensus motif-independent mechanisms (Figure B.11A, B.12A) [52, 55]. For example, allelic binding cooperativity tests (supplementary information) revealed four additional motifs (NFKB1, POU2F2, PRDM1, STAT2), located proximal to the PU.1-bound site, which show covariance with AS PU.1 binding activity (FDR=5%; Figure 2.3A, Figure B.13) and collectively explain another 7.5% of AS PU.1 binding activity.

Despite a strong correlation between motif score differences and AS binding (Figure B.11C, B.12C; >90% expected direction), we observed that the majority of motif disrupting SNPs do not show significant allelic effects (Figure B.11A, B.12A). Therefore, we tested whether homotypic TF motifs (i.e., multiple motifs for the same TF) located within PU.1-bound regions might buffer the effects of motif-disrupting SNPs (supplementary information) [58, 59] and found that TF-bound regions with homotypic motifs exhibit fewer allelic effects (41% vs 25%; P = 0.0087, Mann-Whitney U test). In addition, the impact of SNPs on TF motifs scales with the likelihood to observe significant AS effects (Figure 2.3B, Figure B.12D), but this trend is not significant if a second, unaffected homotypic TF motif is located nearby (Figure 2.3B, Figure B.11D). These results suggest that homotypic motif clusters buffer the effect of genetic variation over several similar binding sites.

Next, we investigated the genetic component of allele-specific chromatin and binding signals and (i) compared direction of allelic bias at shared significant AS sites across ten unrelated individuals (Figure 2.4A) and (ii) tested for transmission of allelic effects from parents to children (Figure 2.4B, Figure B.17) [51]. Allelic directions at shared significant AS sites in the unrelated individuals were significantly correlated (P <0.05, Spearman correlation, Figure B.16A), with mRNA showing the highest degree of consistency in allelic directions between individuals followed by TF binding and histone modification, respectively (Figure 2.4A, B.14 - B.16). We observed evidence of significant parental transmission with all three regulatory TFs (ρ = 0.44-0.75, P <= 0.02, Spearman correlation; Figure 2.4C, Figure B.17), consistent with their strong sequence-dependence (4, 6). For hPTMs, evidence of transmission was detected for the active histone marks H3K4me1, H3K4me3, and H3K27ac (ρ = 0.12-0.21; P <= 0.02), but their level of transmission was lower than for TFs. Transmission signal for mRNA levels and nascent transcription was significant and comparable to TFs (ρ = 0.46 and 0.50; P = 0.0008 and P = 1.3e-07, respectively). We observed only weak transmission for POLR2B (Figure B.17), possibly due to the distinct activity states of the polymerase [60]. We determined the genetic control of the transmission signal of histone marks at known expression [56] and DNaseI sensitivity quantitative trait loci [50] (eQTLs and dsQTLs, respectively), since the former are enriched within TF binding sites [50]. Transmission of the active marks H3K4me1, H3K4me3, and H3K27ac was stronger near eQTLs and dsQTLs (ρ = 0.31-0.57) than genomewide (Figure 2.4D, Figure B.20), suggesting that the transmission behavior of the overall chromatin state depends on the properties of the underlying sequence. Collectively, these findings indicate coordinated and genetically driven changes between TF binding and histone modifications, and suggest that TFs are the primary determinants of regulatory interactions [61–63].

To further assess the extent of allelic coordination (AC) between distinct genomic regulatory layers, we calculated the correlation between AS effects across pairs of molecular phenotypes (Figure B.21). We observed that each testable phenotype exhibits significant correlation in allelic ratios with one or multiple phenotypes (Spearman's correlation; P <0.05). The majority of AC events reflect relationships between distinct regulatory layers that have also been observed quantitatively (e.g. POLR2B/H3K4me3 at promoters [64, 65]; GRO-seq/H3K4me1/H3K27ac at putative enhancers [66]) (Figure 2.5A, Figure B.21). These results support a strong allelic (i.e. local) interconnectivity between regulatory and general TFs, histone modifications, and transcription.

Expression QTLs (eQTLs) are often located distal to their target genes [67], indicating that allelic signals within regulatory layers might extend over short- and long distance. We examined Haplotypic Coordination (HC), defined as long-range coordination in allelic direction on the same chromosome, of AS effects at non-overlapping heterozygous sites (supplementary information) (Figure 2.5B, Figure B.21), and found that every TF and histone mark exhibits HC with one or more regulatory layer(s) around genes and their flanking regions (Figure B.21; Spearman's correlation $P < 0.05$). The degree of coordination varied between regulatory layers ranging from -0.24 (GRO-seq/CTCF; $P = 0.03$) to 0.64 (MYC/mRNA; $P = 2.9e\text{-}08$). The majority (>90%) of significant HC events were positive, i.e., the allelic bias co-occurred on the same haplotype (Figure 2.5B, Figure B.21). For 25% of assay pairs tested, the strength of HC was significantly correlated with the genomic distance between SNP pairs (logistic regression, $P < 0.05$; OR = 0.19-2.2) (Figure B.22). For example, the enhancer-associated histone marks H3K4me1 and H3K27ac showed allelic consistency up to 200 kb with the TF PU.1. Thus, a single or few variant(s) likely trigger long-distance allelic effects over many of the regulatory layers acting on a genomic region.

In summary, we observed abundant allele-specific activity across all regulatory layers. Parental transmission of the allelic effects suggests that DNA sequence variation affecting transcription, TF binding and histone modifications are largely transmitted from parents to children, with allelic histone effects showing more sensitivity to context-dependent effects compared to TFs. Coordinated allelic and haplotypic behavior at different functional elements of the genome suggest that TF binding, histone modifications, and transcription operate within the same allelic framework. This is consistent with the fact that a few TFs can induce cellular reprogramming and massive changes in the chromatin landscape [68], and that the maintenance of a transcription-permissive environment and transcriptional memory are independent of histone modifications [69]. Both histone modifications and TF binding are under genetic control, but histone modifications are more prone to stochastic, possibly transient effects and likely reflect [70], rather than define, coordinated regulatory interactions.

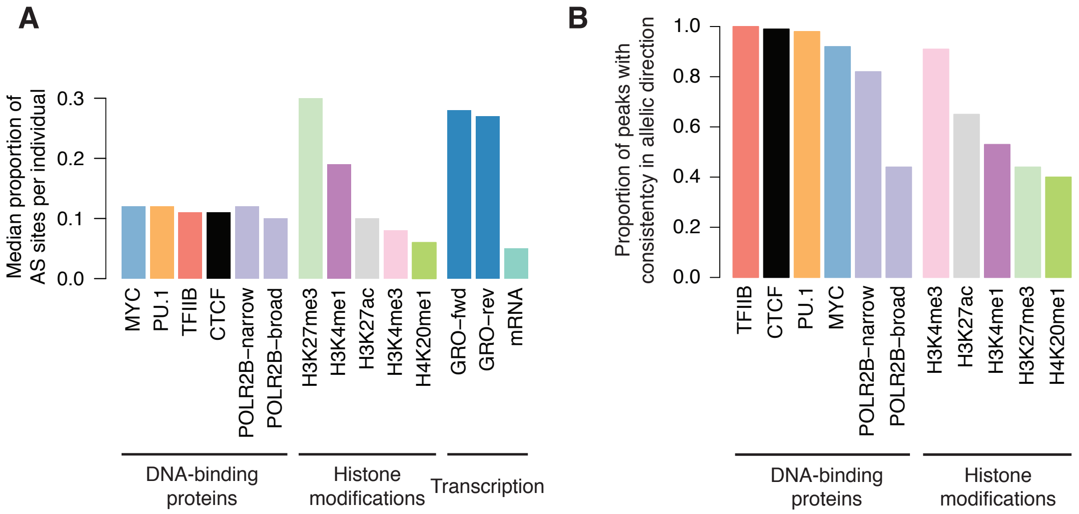


Figure 2.2: **Allele-specific (AS) activity within transcriptional and chromatin layers.** **A** Proportion of accessible heterozygous SNP sites showing significant AS activity (median across all individuals, n=3-14). **B** Consistency of allelic effects within genomic regions of TF binding and histone modification. Bars represent the proportion of peaks with a consistent allelic direction at two or more SNP sites.

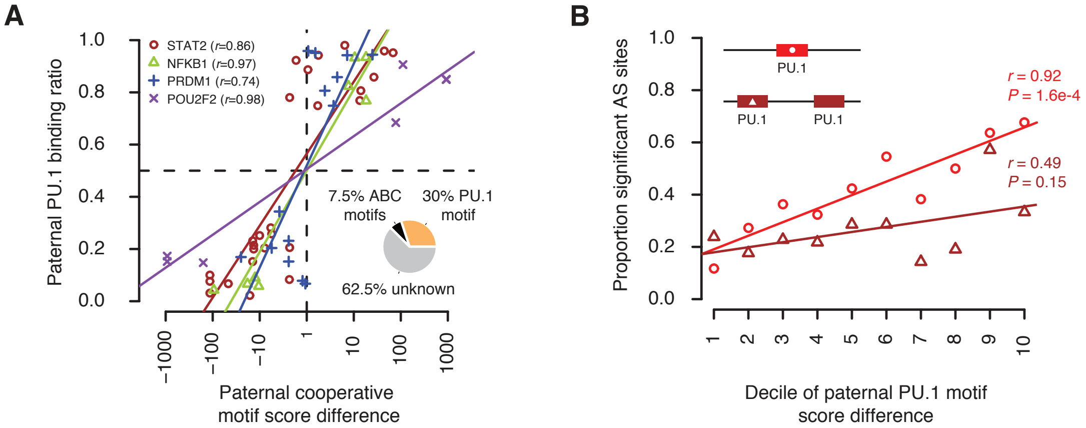


Figure 2.3: **DNA sequence properties at allele-specific (AS) PU.1 binding sites.** **A** SNPs in PU.1- and cooperative TF motifs are predictive of AS PU.1 binding (5% FDR) (5). **B** PU.1-bound regions (peaks) with homotypic PU.1 motifs show a weak response towards motif-disrupting SNPs. Motif-disrupting SNPs were split into two classes (one or two PU.1 motifs per peak) and grouped based on their motif impact (1, lowest; 10 highest).

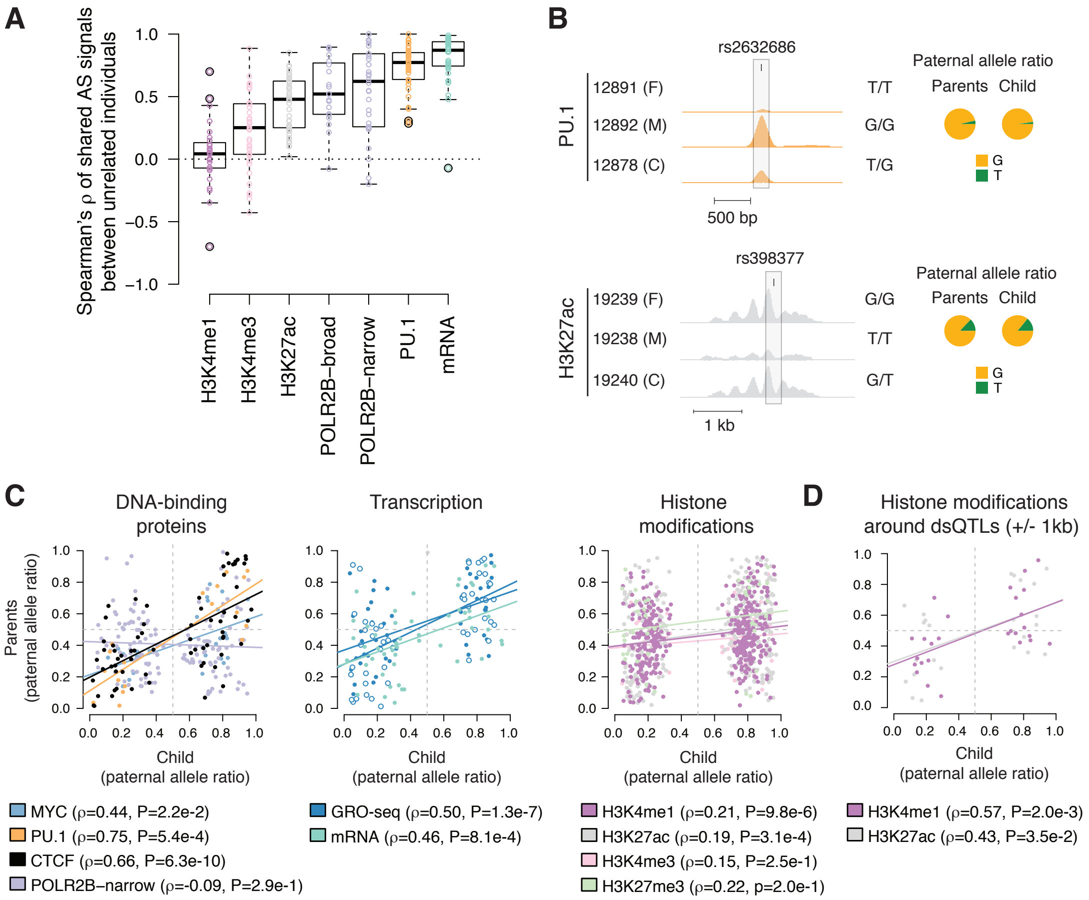


Figure 2.4: **Genetic component of allele-specific (AS) transcriptional and chromatin activity.** **A** Distribution of pairwise correlation coefficients of significant AS sites between all unrelated CEU individuals (n=10) for each molecular phenotype. Correlation of the reference allele ratio is calculated at shared significant AS SNP sites using Spearman rank correlation. **B-D** Correlation of the paternal allele ratio of the child and that inferred from the parents at SNP sites where parents are opposite homozygotes and the child has a significant allelic effect. **B** Examples of transmitted PU.1 and H3K27ac SNP sites. **C** Genome-wide transmission results. GRO-seq signal was analyzed separately for each strand (filled and empty points, forward and reverse strand, respectively; P-value represents combined data). **D** Transmission results of H3K4me1 and H3K27ac near DNase I sensitivity QTLs (+/- 1 kb window around the dsQTL).

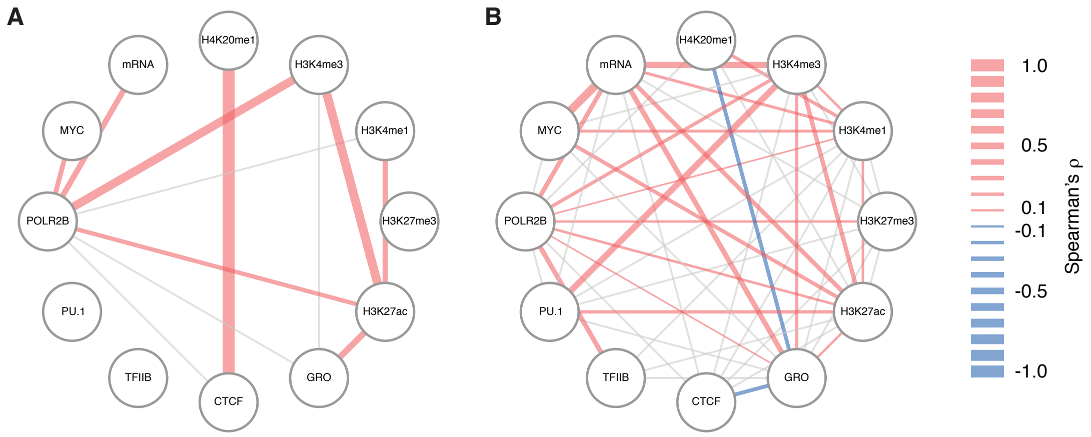


Figure 2.5: **Local, short- and long-range coordination between transcriptional and chromatin layers.** Results of allelic coordination **A** and haplotypic coordination **B** analysis at gene regions (genes +/- 50 kb) (5). Coordination of the allelic effect was considered between all pairs of assays. SNP sites within genomic regions were required to show a significant AS effect in both assays. Only assay pairs with >=20 SNPs were considered for the analysis. Significant Spearman rank correlation coefficients (P <0.05) between the paternal allele ratios of the SNP pairs are indicated with colored lines ranging in intensity from ρ = -1.0 (blue) to ρ = 1.0 (red). Non-significant correlations are indicated with grey lines and missing lines indicate lack of sufficient data points for analysis.

2.4 References

48. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (Oct. 2010).

49. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

50. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (Feb. 2012).

51. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328,** 235–239 (Apr. 2010).

52. Kasowski, M. *et al.* Variation in Transcription Factor Binding Among Humans. *Science* **328,** 232–235 (2010).

53. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342,** 744–747 (Nov. 2013).

54. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–U151 (2010).

55. Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome research* **22,** 860–869 (Mar. 2012).

56. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).

57. Henikoff, S. & Shilatifard, A. Histone modification: cause or cog? *Trends In Genetics* **27,** 389–396 (Oct. 2011).

58. Maurano, M. T., Wang, H., Kutyavin, T. & Stamatoyannopoulos, J. A. Widespread Site-Dependent Buffering of Human Regulatory Polymorphism. *PLoS genetics* **8** (Mar. 2012).

59. Spivakov, M. *et al.* Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome biology* **13** (2012).

60. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322,** 1845–1848 (2008).

61. Erwin, D. H. & Davidson, E. H. The evolution of hierarchical gene regulatory networks. *Nature reviews. Genetics* **10,** 141–148 (Feb. 2009).

62. Hobert, O. Maintaining a memory by transcriptional autoregulation. *Current Biology* **21,** R146–R147 (2011).

63. Ptashne, M. On the use of the word 'epigenetic'. *Current Biology* **17,** R233–R236 (2007).

64. Lin, C. Y. *et al.* Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151,** 56–67 (Sept. 2012).

65. Nie, Z. *et al.* c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells. *Cell* **151,** 68–79 (Sept. 2012).

66. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474,** 390–+ (2011).

67. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature genetics* **44,** 1084–+ (Oct. 2012).

68. Doege, C. A. *et al.* Early-stage epigenetic modification during somatic cell reprogramming by Parp1 and Tet2. *Nature* **488,** 652–655 (2012).

69. Margueron, R. & Reinberg, D. Chromatin structure and the inheritance of epigenetic information. *Nature reviews. Genetics* **11,** 285–296 (Apr. 2010).

70. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2** (2013).

Chapter 3

Population variation and genetic control of modular chromatin architecture in humans

3.1 Preface

To further decipher how genetic variation impacts gene expression and the chromatin landscape, we conducted a follow-up study where we investigated genetic variation, gene expression, DNA-binding and chromatin modifications among 54 unrelated individuals. We observed that the chromatin landscape is locally organized into chromatin modules (VCMs), which vary in their activity across individuals and are often correlated with gene expression levels of proximal genes. Both VCM activity and gene expression levels were found to be associated with genetic variation within their genomic region. The findings of this project were published in the peer-reviewed journal *Cell* in August 2015 [73]. Together with the two first authors, I was involved in the analysis of the data. My contribution to this article was to investigate the transmission of the genetic effects across molecular phenotypes using a Bayesian network approach.

Chromatin - gene expression models

I first sought to unravel in which order genetic variation perturbs the molecular layers. More precisely, I wanted to test three biologically relevant hypotheses: 1) That genetic variation impacts gene expression through changing chromatin activity, 2) that genetic variants change gene expression levels and chromatin activity follows this change and 3) that both chromatin activity and gene expression levels are affected by genetic variants independently (Figure 3.1A).

To this end we deviced a bayesian network approach to compute the probability of these three statistical models given our observed data. For each locus where a chromatin module was found to be associated with expression levels of a proximal gene and at least one genetic variant was reported to be associated with one of them, a ”triplet graph” was constructed. The three nodes of this graph consisted of the genotypes of the genetic variant (V), the chromatin activity summarized by the 1st principle component of the activities of all chromatin phenotypes within the module (C) and the expression levels of the gene (G). For each triplet graph we then scored each of the three following models using the log-likelihood.

$$P(V,C,G) = P(V)P(C|V)P(G|C) \tag{3.1}$$

$$P(V,C,G) = P(V)P(G|V)P(C|G) \tag{3.2}$$

$$P(V,C,G) = P(V)P(G|V)P(C|V) \tag{3.3}$$

Applying Bayes' theorem we then calculated the posterior probability of each model using an uninformative prior. A decision was made whenever one model exceeded a probability of 0.9 at a given locus (n=232, 72% of all loci). We observed that in 78% of our loci, the most probable model was that genetic variants impact gene expression through changing the chromatin state, while only in 18% chromatin activity was suggested to be the result of gene expression changes. The independent model was inferred in only around 4% of all cases (Figure 3.1B). These results suggest that the impact of genetic variation is very locus and context specific, since all inferred models are considered high confidence calls (posterior probability >0.9). However, in most of the cases genetic variants seem to affect regulatory elements such as enhancers, which lead to a change in chromatin activity upon activation and subsequently change gene expression. In 18% of cases, genetic variants might affect regulatory elements, such as promoter linked elements, which directly impact gene expression and chromatin activity changes as a result of this gene expression change. When comparing the composition of the chromatin modules at these loci, we do not observe considerable differences between loci under different causal models (Figure 3.1C). This suggests that the regulatory mechanisms do not differ between loci, but that their behavior rather depends on which elements are affected by genetic variants.

PU.1 binding motif disruption

To further investigate how genetic variants can affect regulatory elements and lead to changes in the chromatin landscape, we used data on PU.1 binding motif disruption and its downstream effects. We scanned all cases where PU.1 activity was correlated with the activity of other chromatin phenotypes. Whenever a genetic variant was found to be associated to the activity of both, we constructed again a triplet graph consisting of the genotypes (V), PU.1 activity (P) and the activity of the PU.1 associated chromatin phenotype (A). Using the same bayesian network approach we defined the most probable of the following models.

$$P(V,P,A) = P(V)P(P|V)P(A|P) \tag{3.4}$$

$$P(V,P,A) = P(V)P(A|V)P(P|A) \tag{3.5}$$

$$P(V,P,A) = P(V)P(P|V)P(A|V) \tag{3.6}$$

We then compared cases where the genetic variant disrupted the PU.1 binding motif versus cases where the genetic variant was located outside the PU.1 motif. For PU.1 associations with both H3K27ac and H3K4me1, both well known enhancer marks, we observed PU.1 binding motif disruption. For H3K27ac we observed a strong tendency that if the genetic variant disrupts the PU.1 binding site, the variant was likely to affect H3K27ac through PU.1 binding (Figure 3.1D). For H3K4me1 we observed the same trend but weaker, suggesting a tighter coupling between PU.1 and H3K27ac activity than PU.1 and H3K4me1 activity. These results serve as an example of how genetic variants can impact regulatory mechanisms which leads to changes in chromatin landscape activity and potentially to downstream changes in gene expression.

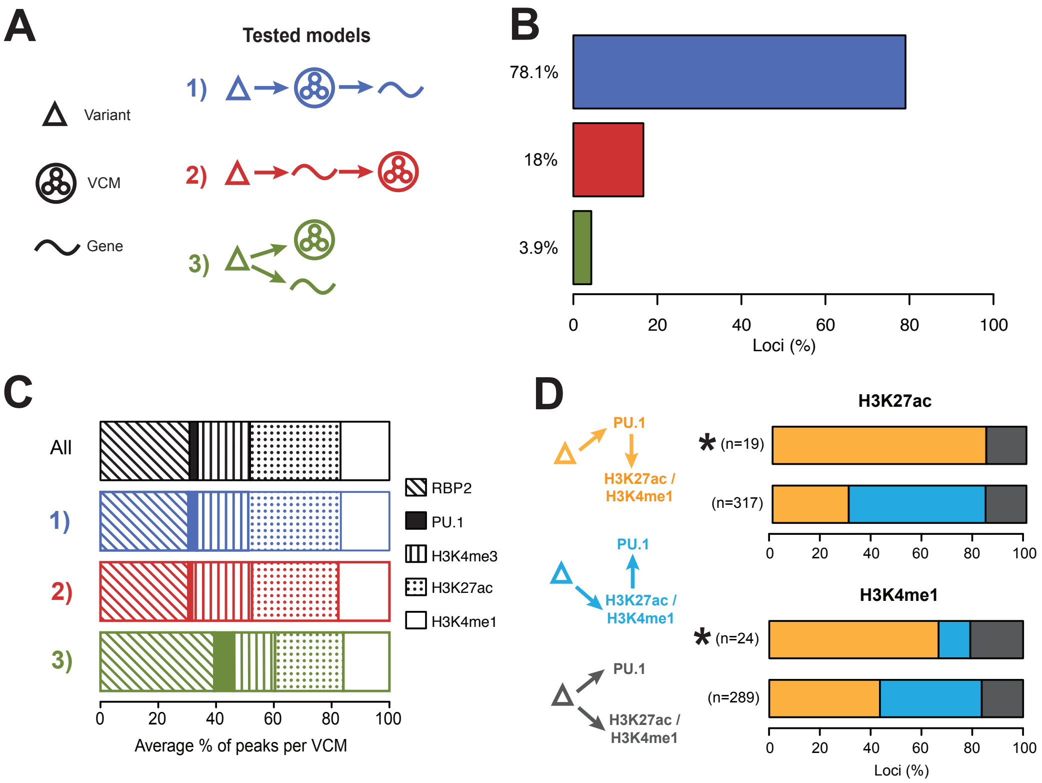


Figure 3.1: **Chromatin causality models. A** Graphical representation of the three tested models to decipher the causal order of the genetic effects at variable chromatin module (VCM) loci. **B** Proportion of most probable models inferred at each VCM locus. A minimal probability of 0.9 was required to make a decisive call, leading to a total of 232 loci, where a most likely model was inferred. **C** Composition of VCMs under the different inferred models. Shown is the average proportion of peaks per VCM for each individual assays. n All = 232, n 1) = 182, n 2) = 42, n 3) = 9. **D** Results of PU.1 motif disruption analysis for PU.1 peaks associated with H3K27ac or H3K4me1 activity. Top bar of each plot (*) shows the proportion of the most likely model in case of PU.1 motif disruption by the genetic variant, while the bottom bar shows the proportion of the most likely model for cases where the genetic variant did not disrupt the PU.1 binding motif.

Article

Population Variation and Genetic Control of Modular Chromatin Architecture in Humans

Waszak S. M.[1,2*], Delaneau O.[2,3*], Gschwind A. R.[2,4], Kilpinen H.[2,3], Raghav S. K.[1], Witwicki R. M.[4], Orioli A.[4], Wiederkehr M.[4], Panousis N. I.[2,3], Yurovsky A.[2,3], Romano-Palumbo L.[3] Planchon A.[3], Bielser D.[3], Padioleau I.[2,3], Udin G.[1], Thurnheer S.[5], Hacker D.[5], Hernandez N.[4], Reymond A.[4], Deplancke B.[1,2], and Dermitzakis E. T.[2,3]

[1]Institute of Bioengineering, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics SIB, 1015 Lausanne, Switzerland

[3]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland

[4]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland

[5]Protein Expression Core Facility, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland

[6]These authors contributed equally to this work

[*]These authors contributed equally to this work

3.2 Abstract

Chromatin state variation at gene regulatory elements is abundant across individuals, yet we understand little about the genetic basis of this variability. Here, we profiled several histone modifications, the transcription factor (TF) PU.1, RNA polymerase II, and gene expression in lymphoblastoid cell lines from 47 whole-genome sequenced individuals. We observed that distinct *cis*-regulatory elements exhibit coordinated chromatin variation across individuals in the form of variable chromatin modules (VCMs) at sub-Mb scale. VCMs were associated with thousands of genes and preferentially cluster within chromosomal contact domains. We mapped strong proximal and weak, yet more ubiquitous, distal-acting chromatin quantitative trait loci (cQTL) that frequently explain this variation. cQTLs were associated with molecular activity at clusters of *cis*-regulatory elements and mapped preferentially within TF-bound regions. We propose that local, sequence-independent chromatin variation emerges as a result of genetic perturbations in cooperative interactions between *cis*-regulatory elements that are located within the same genomic domain.

3.3 Introduction

Understanding the genetic contribution and molecular paths towards complex traits is one of the key outstanding challenges in biology. Genome-wide studies revealed that most common disease-associated genetic variants fall into gene regulatory sequences [71, 74–76] and affect transcriptional programs in disease-implicated cell types [67, 77]. Evolutionary studies have further uncovered several instances of gene regulatory changes that are causally implicated in complex phenotypes [72]. These changes are thought to originate mostly from variation in TF-DNA interactions, which are well known to mediate the spatiotemporal control of gene expression programs [78]. Understanding the extent of, and the mechanisms underlying, TF DNA binding variation is therefore key to elucidate the molecular determinants of complex phenotypes. Small-scale population- and family-based studies have shown that 5 to 25% of TF-DNA binding events exhibit intra- and inter-individual binding variation [52, 53, 55, 79, 80]. These studies, as well as those examining TF-DNA binding divergence among mammalian species (reviewed in Villar et al. [81]) showed that only a minority of this variation could be attributed to genetic differences within TF-bound sequences.

So far, few mechanisms have been proposed to clarify this phenomenon, and these are mostly centered on changes in either the local DNA structure or in collaborative interactions between co-bound TFs at *cis*-regulatory elements [53, 82–85]. Recently, others and we have observed that several chromatin state components exhibit a strong degree of coordinated allelic variation that extends over several thousands of base pairs [53, 80]. This observation suggests that variation in TF-DNA binding might be conditioned on the state of other *cis*-regulatory elements, but a general description of this effect has so far been hampered due to sparseness of allelic markers. Here, we measured ChIP-seq-based, population-level histone modification (HM) and TF enrichment patterns. Specifically, we mapped the regulatory TF PU.1, the second largest subunit of RNA polymerase II [RPB2], and three well-studied HMs often observed at enhancers and promoters (H3K4me1, H3K4me3, and H3K27ac) in lymphoblastoid cell lines (LCLs) derived from 47 unrelated European individuals whose genomes were sequenced in the frame of the 1000 Genomes Project [48]. In addition, we also profiled gene expression using mRNA sequencing in 46 LCLs. Our results provide unique insights into the mechanisms underlying variation in molecular activity at *cis*-regulatory elements, revealing that most of this variation results from alterations in the modular organization of the human genome.

3.4 Results

3.4.1 Population-level variation in molecular activity at *cis*-regulatory elements

To assess the extent of quantitative coordination in inter-individual chromatin variation at putative *cis*-regulatory elements, we performed an association analysis between molecular phenotypes, with ”molecular phenotype” being here defined as the normalized and covariate-corrected read depth of a histone-modified and TF-bound region, respectively. Specifically, we estimated the correlation levels between all TF-TF, HM-HM, and TF-HM combinations in 1 Mb *cis* windows (Figure 3.2A). We tested a total of 29 million associations between any two molecular phenotypes and estimated for each association pair the enrichment of low P values using $\pi 1$ statistics [86]. Estimates of $\pi 1$ ranged from 2.5% for PU.1-H3K4me3 to 11% for H3K4me1-H3K27ac (Figure C.1A), indicating extensive quantitative coordination in molecular activity levels between/at *cis*-regulatory elements. Moreover, molecular coordination decayed quickly with increasing genomic distance and was 20-fold more enriched between proximal *cis*-regulatory elements (<10 kb) than between any two *cis*-regulatory elements that were separated by 500 kb or more (Figure C.1B).

Overall, we detected 79,411 statistically significant, mostly positive (>99%) associations (at genome-wide correction) across all molecular association tests (Pearson r_{mean}=0.70, FDR 0.1%)

(Figures 3.2B&C and C.1C&D), involving on average 20% of all studied molecular regions (Figure C.1E). The histone mark H3K27ac exhibited the highest number and proportion of significant associations with other phenotypes (Figure C.1E-F), suggesting that this molecular phenotype is most sensitive to coordinated chromatin state perturbations. As expected, the TFs PU.1 and RPB2 were preferentially associated with enhancer- (H3K27ac / H3K4me1 for PU.1) and promoter-marking HMs (H3K27ac / H3K4me3 for RPB2), respectively (Figure C.1G). Except for RPB2-H3K4me3, the majority of all molecular associations were identified between non-overlapping *cis*-regulatory elements (Figure C.2A), which exhibit a log-normal distance distribution that preferentially centered around 45 kb (95%-CI: 7-308 kb) (Figures 3.2D and C.2B). The molecular association strength between covariable *cis*-regulatory elements decayed significantly with increasing distance (*r*=-0.19, P<2.2e-16, Figure C.2C). Overall, 25% of all molecular associations were found between promoters and enhancers (>5 kb from TSS), 25% within or between promoters, and 50% within or between putative enhancers (Figure 3.2E). These results suggest extensive molecular coupling between *cis*-regulatory elements and a strong degree of chromatin variation at enhancer-like regions.

The previous results suggest that chromatin state variation might reflect high-order genomic interactions. Using simple graph-based methods we could map individual molecular associations into 14,559 distinct ”Variable Chromatin Modules (VCMs)” that are composed of 25,417 distinct *cis*-regulatory elements (see Figures 3.2B&C and C.3A-C for examples). The median size of a single VCM was 4.2 kb and all VCMs together covered 5% (161 Mb) of the human genome. Although only 25% of VCMs were composed of multiple *cis*-regulatory elements (Figure C.3D), these ”multi-VCMs” captured the vast majority (78%) of molecular associations (Figure C.3E), were more likely to contain promoter- and enhancer-marking chromatin marks (Figure C.3F), and covered more DNA sequence (median size: 70 kb; Figure C.3G).

The majority of VCMs (56%) were exclusively composed of enhancer-marking signals (i.e. H3K4me1-PU.1, H3K4me1-H3K27ac, and H3K4me1-H3K27ac-PU.1) (Figure 3.3A), indicating that putative enhancers constitute the largest fraction of the variable epigenome in a single human population, which is consistent with comparative epigenomic studies across mammalian species [81].

To examine the extent of molecular coordination within VCMs, we tested whether the activity state of a VCM can be represented by a single quantitative phenotype, rather than by individual molecular phenotypes that define a VCM. We applied principal component (PC) analysis and extracted the first and second PC for each VCM (Figure 3.3B). We found that the first PC already explains on average 79% of the variability that is observed between molecular phenotypes of the same VCM (Figure 3.3C), suggesting that molecular activity is strongly coordinated within

VCMs.

This high degree of molecular coordination within VCMs implies a higher order chromatin organization, consistent with the now well-accepted notion that mammalian genomes are spatially arranged in distinct chromosomal contact domains [87, 88]. To test this hypothesis, we analyzed published, high-resolution, and genome-wide chromatin conformation data from a human lymphoblastoid cell line [88] and found that *cis*-regulatory elements with coordinated chromatin state variation were more preferentially embedded within the same chromosomal contact domain (odds ratio=14.9, P=2.2e-16, logistic regression) (Figure 3.3D; see Figure 3.2A and C.1H-I for examples). We also observed that *cis*-regulatory elements of the same VCM exhibited more frequently allelic chromatin biases along the same haplotype (OR=1.3, P=4.9e-5, logistic regression), further indicating that VCM define a regulatory unit. Moreover, analysis of genome-wide TF-DNA binding data of the architectural proteins CTCF and cohesin (RAD21/SMC3) [89] revealed a significant enrichment at *cis*-regulatory elements that participate in long-range (>300 - 500 kb) molecular associations (Figure C.2D-F). Together, these results support our hypothesis that VCMs represent a fine-grained, modular architecture of the variable human epigenome.

Next, we aimed to elucidate mechanisms that may be responsible for the emergence of VCMs. Here, we hypothesized that modular chromatin state dynamics may not only be driven by short-range cooperative TF-TF interactions, as shown earlier [53, 79, 84, 90], but also by interactions that act over long genomic distances and across *cis*-regulatory elements. To test this hypothesis, we investigated whether particular TF-TF pairs exhibited preferential enrichments at pairs of *cis*-regulatory elements that are part of the same VCM using experimentally defined TF-DNA binding data [49]. This analysis revealed 204 putative cooperative TF-TF pairs that are preferentially enriched at VCM-defined *cis*-regulatory elements (OR=1.1-3.2; P<0.05 after Bonferroni correction; Fisher's exact test) (Figure 3.3E). For example, NFKB emerged as the most cooperative TF among all tested factors and was preferentially associated with well known immunity-associated TFs (e.g., STAT3, BCL11A, BATF, and PU.1). Thus, our results suggest that modular chromatin dynamics occur within spatially organized domains of the genome and are likely in part mediated by long-range cooperative interactions between TFs that determine the molecular identity of a lymphoblastoid cell [91].

3.4.2 Chromatin variation reflects inter-individual variation in gene expression

To assess the functional impact of inter-individual chromatin state variation, we analyzed associations in *cis* between molecular phenotypes at *cis*-regulatory elements and gene expression

(TSS +/- 1 Mb). This analysis resulted in significant associations for 4,568 (22%) genes at a FDR of 0.1% (Figure C.3H and see Figures 3.4A and C.3I&J for examples). The vast majority (99%) of chromatin-gene associations were positive (i.e. higher gene expression levels correlated with stronger chromatin signals) (Figure C.3K), explained about half of the variation in gene expression (Figure C.3L), and correlated independently with multiple molecular events at *cis*-regulatory elements. Two thirds of all gene-associated *cis*-regulatory elements mapped outside of promoters (TSS +/- 2.5 kb) and thus likely pinpoint to putative enhancer-gene interactions (Figure 3.4B&C). We further measured allelic expression effects within individuals and observed that, consistent with coordinated allelic chromatin signals, that they are more concordant with allelic chromatin states at gene-associated regions than at random regions (OR=1.9, P=2e-10, logistic regression). Together, these results provide genome-wide evidence that population-level variation in chromatin states has functional consequences and that it is a potential approach to identify the gene targets of putative *cis*-regulatory elements.

We also observed that VCM states (as defined by the first PC) were associated with 3,580 genes in *cis* (TSS +/- 1 Mb; FDR 0.1%). This analysis has further allowed us to uncover that only 5% of ”enhancer VCMs” (H3K27ac-H3K4me1-PU.1) varied along with nearby genes, despite representing the most abundant class of VCMs. In strong contrast, variable promoter (H3K27ac-H3K4me3-RPB2) and promoter-enhancer (H3K27ac-H3K4me3-H3K4me1-RPB2-PU.1) VCMs correlated with gene expression in up to 80% of the cases (Figure 3.4D). Moreover, 23% of all gene-associated VCMs correlated with the expression levels of multiple genes (Figure C.3M), suggesting that these VCMs contain *cis*-regulatory elements that are potentially shared across genes. We also found that VCMs with several *cis*-regulatory elements were more likely to reflect variable gene expression (Spearman's ρ=0.91, P=1.8e-8) (Figure 3.4E), suggesting that both the type (promoter/enhancer), and the number of variable *cis*-regulatory elements are key determinants underlying the transcriptional state change of a gene.

We next assessed whether VCMs were located nearby specific sets of genes and found that VCMs embedding several *cis*-regulatory elements were highly enriched in immunity-related processes and pathways (Table C.2A&B), consistent with the biological nature of lymphoblastoid cells. Functional analysis of chromatin-associated genes further supported a strong enrichment of VCMs in immunity-related processes (Table C.2C).

3.4.3 Genetic control of chromatin state and gene expression variation

To identify potential mechanisms that explain variation in TF-DNA binding, HMs, VCM states, and gene expression, we mapped quantitative trait loci (QTLs) for all studied molecular phenotypes independently in a 500 kb *cis*-window around the center of a candidate *cis*-regulatory element (or TSS). We detected between 315 to 1,432 significant chromatin QTLs (cQTLs, i.e., tfQTLs and hmQTLs) and eQTLs at 10% FDR. This corresponds to 1.1% (H3K4me1) to 2.9% (mRNA) of the studied regions and explained around 40% of their variability (Figures 3.5 and C.4). Of note, the number of discovered QTLs significantly increased upon reduction of the *cis*-window size, yet at the expense of excluding distal effects (Figure C.4A-C). Indels and structural variants were significantly enriched among cQTLs (Figure C.5A), consistent with previous studies [52, 92]. We further used allele-specific analysis to validate cQTLs on a genome-wide scale [56]. We observed more significant allelic chromatin biases at cQTLs as compared to control sites (Figure 3.5C) and higher proportions of allelic chromatin biases at strong cQTLs (Figure 3.5D), thus supporting our cQTL inference. In addition, we mapped 1,173 vcmQTLs (8.1%) using the first PC as a quantitative trait (comprising 4,187 individual molecular phenotypes) and, surprisingly, none using the second PC despite observing a small enrichment of low P values (Figure C.4G). This suggests that the first VCM state captures the primary genetic contributions towards VCM activity. Overall, we found that all molecular phenotypes and in particular VCMs are affected by common genetic variants, supporting the hypothesis that a substantial proportion of coordinated chromatin state variation is driven by *cis*-acting genetic variation.

We further assessed the genomic location of cQTLs by measuring their distance relative to TF-targeted and histone modified regions. We found that the resulting distances exhibit bimodal log-normal distributions with the first mode centering very close to the mid-point of TF-bound sites (medians between 10-40 bp) and relatively close to the mid-point of HM regions (medians between 230-300 bp) and the second mode being located distally from its respective target region (medians between 20-30 kb) (Figure 3.5A). In contrast, when we tested the distance relative to the closest TSS (Figure C.5B), the log-normal bimodal signal completely disappeared, suggesting that the first mode derives from cQTLs falling within their respectively TF or HM-enriched target regions (Figure C.5C&D).

Although the proximally mapping cQTLs exhibited significantly stronger effect sizes than cQTLs located outside of their target elements (Figure 3.5B), they constituted only a minority (25%) of all cQTLs. For example, we found that only 33% of PU.1 QTLs mapped inside PU.1-bound regions, demonstrating that TF binding is strongly influenced by distal genetic effects. This complexity indicates that, like gene expression, sequence-specific TF-DNA binding can be considered as a complex trait, similar to other molecular and organismal traits. Moreover, we found that distally-acting cQTLs exhibited distances that matched the extent of coordination within VCMs, further supporting interactions across regions in the genome. These observations suggest a dual mode of action for cQTLs: strong cQTLs directly perturbing the proximal interactions that form the local chromatin signal and, more abundant, yet weaker *cis*-acting cQTLs exerting their effects over large distances (up to hundreds of kilobases). The latter process is likely involving several intermediate molecular processes that operate within the same VCM.

Given the high degree of quantitative coordination between chromatin state components of the same VCM, we assessed whether distinct molecular phenotypes were affected by the same cQTL. We estimated that half of all cQTLs are shared between two chromatin phenotypes, revealing a strong genetic basis for coordinated chromatin state variation across individuals (Figure 3.6A). In addition, we found that cQTL-eQTL sharing ranged from a relatively moderate (24% of PU.1 QTLs were also eQTLs) to a very high (73% of H3K4me3 QTLs were also eQTLs) degree (Figure 3.6A). These results demonstrate that only a small proportion of genetically variable TF-DNA binding events actually lead to measurable changes in gene expression, in line with recent TF knock-down studies carried out in LCLs [93]. They also suggest that promoter QTLs show very high specificity for genetic gene perturbations. The latter observation is consistent with the enrichment of complex trait-associated variants in cell type-specific H3K4me3 regions [94].

We further characterized the width and the depth of the QTL signal path by estimating the number of distinct molecular marks and phenotypes that were affected by the same cQTL and eQTL. We observed that the majority of QTLs affect several molecular marks (75%) (Figures 3.6B and C.6A) and molecular phenotypes of the same and/or different type (80%) (Figures 3.6C and C.6B). Instances of QTLs for which we did not identify cross talk between distinct molecular marks were of significantly lower effect sizes (Figure C.6C). In contrast, 99% of vcmQTLs were associated with multiple molecular marks and phenotypes, suggesting that they capture the deepest and widest range of genetic associations across all studied epigenomic components. Taken together, these results demonstrate that the majority of cQTLs perturb several chromatin state components at the same or across distinct *cis*-regulatory elements.

We next set out to identify which component is more likely to initiate the genetically induced molecular cascade. To do so, we estimated the enrichment of each QTL class being located within particular functional elements with the underlying reasoning that QTLs that overlap a functional element should initially affect that element first before their effect extends towards non-overlapping elements that belong to the same VCM. We found a clear enrichment signal in TF-bound regions for all types of QTLs. For instance, H3K27ac and H3K4me1 QTLs were seven times more likely to be located within PU.1-bound regions than expected by chance and vcmQTLs were nine times more enriched within PU.1-bound regions (Figure 3.6D). We independently validated this observation by testing for enrichment of QTLs in open chromatin regions and experimentally defined TF-bound regions (Figure C.6D&E). We found that vcmQTLs demonstrated the strongest enrichment at regions that were bound by PU.1, BATF, BCL11A, NFKB, MEF2A, and IRF4 (Figure C.6E), consistent with our observations that these TFs are specifically enriched at variable *cis*-regulatory elements (Figure 3.3E). Moreover, cQTLs that fell within TF-bound regions exhibited stronger effect sizes than those falling outside such regions (Figure C.6F), and we observed stronger enrichment of allelic biases at tfQTLs as compared to hmQTLs for each studied molecular mark (Figure 3.6E).

We next investigated the impact of TF motif disruption and its downstream effects onto other molecular phenotypes using Bayesian network modeling. We assessed all molecular associations that involve PU.1 and considered cases separately whereby PU.1 QTL variants disrupted a PU.1 binding site. We observed that PU.1-DNA binding variation was more likely to be causal to variation in H3K27ac and H3K4me1 signals in cases where the PU.1 motif was disrupted as compared to cases where the PU.1 QTL mapped elsewhere in the genome (Figure C.6G). Thus, these results indicate that sequence-specific TF-DNA interactions are an important driving force behind inter-individual chromatin state variation.

The previous sections demonstrated that genetic perturbation of few molecular phenotypes can be causal to changes in downstream molecular phenotypes, thus providing a potential explanation as to why most variation in chromatin state is likely independent of proximal sequence changes. VCMs provide the conceptual framework to test the hypothesis of few molecular phenotypes causing collateral changes to chromatin states across *cis*-regulatory elements. We therefore performed association analysis of vcmQTL variants with every molecular phenotype of the corresponding VCM and observed strong association signals with individual molecular phenotypes (Figure 3.7A&B). Moreover, we observed that the average QTL strength for individual molecular phenotypes scales significantly with the strength of vcmQTLs (ρ=0.91, P<2.2e-16), yet, one order of magnitude weaker (Figure C.6H). The latter observation suggests one or more of the following possibilities: (i) higher-order chromatin states are more reflective of genetic perturbations than individual molecular phenotypes; (ii) VCMs exhibit a genetically defined structure with few causal effects driving downstream molecular cascades; or (iii) VCMs constitute more accurate phenotype estimates, since the correlation structure represented as a PC is devoid of experimental and environmental noise independent of which molecular phenotype used.

To explore these possibilities, we contrasted the association strength of the same vcmQTL variant with VCM states and individual molecular events (Figure 3.7C). We further used the molecular association structure that defines VCMs to obtain a hierarchy of molecular interactions: (1)

entry phenotypes that exhibit the strongest association with vcmQTLs, (2) direct (1st degree) molecular phenotypes that are defined as being directly associated with the entry phenotype, and (3) indirect (2nd degree) molecular interactions that are associated with the entry phenotype via intermediate molecular associations. These analyses revealed that VCM entry phenotypes exhibit a similar association strength with vcmQTL variants as VCMs themselves, further supporting our observation that a single molecular phenotype can act as a seed for collateral changes within the respective VCM (Figure 3.7C, black boxplot). Interestingly, simulations demonstrated that PU.1 is most likely to act as an entry phenotype among our probed molecular marks (Figure C.6I). Consistent with a hierarchical view, we observed that the remaining molecular phenotypes are on average more weakly associated with vcmQTL variants than the overall VCM state and VCM entry phenotypes (Figure 3.7C, blue and orange boxplots). More specifically, 1st degree (direct) molecular phenotypes were more strongly associated with vcmQTL variants than 2nd degree (indirect) phenotypes.

We subsequently studied genetic variants that affect chromatin modules (vcmQTLs) and gene expression (eQTL) to obtain a global view of the *cis*-regulatory information flow. Bayesian modeling indicates that genetic variants affected gene expression levels through modulation of chromatin activity in 78% of the cases (Figure 3.7D), thus illustrating that genetic perturbation of chromatin states at *cis*-regulatory elements is in most cases causal to changes in gene expression. Finally, we found that all types of cQTLs are enriched in known complex disease susceptibility variants, especially in immune system disease variants (Figure 3.7E), providing direct functional genetic evidence that non-coding disease susceptibility variants exert their effects through perturbation of gene regulatory regions.

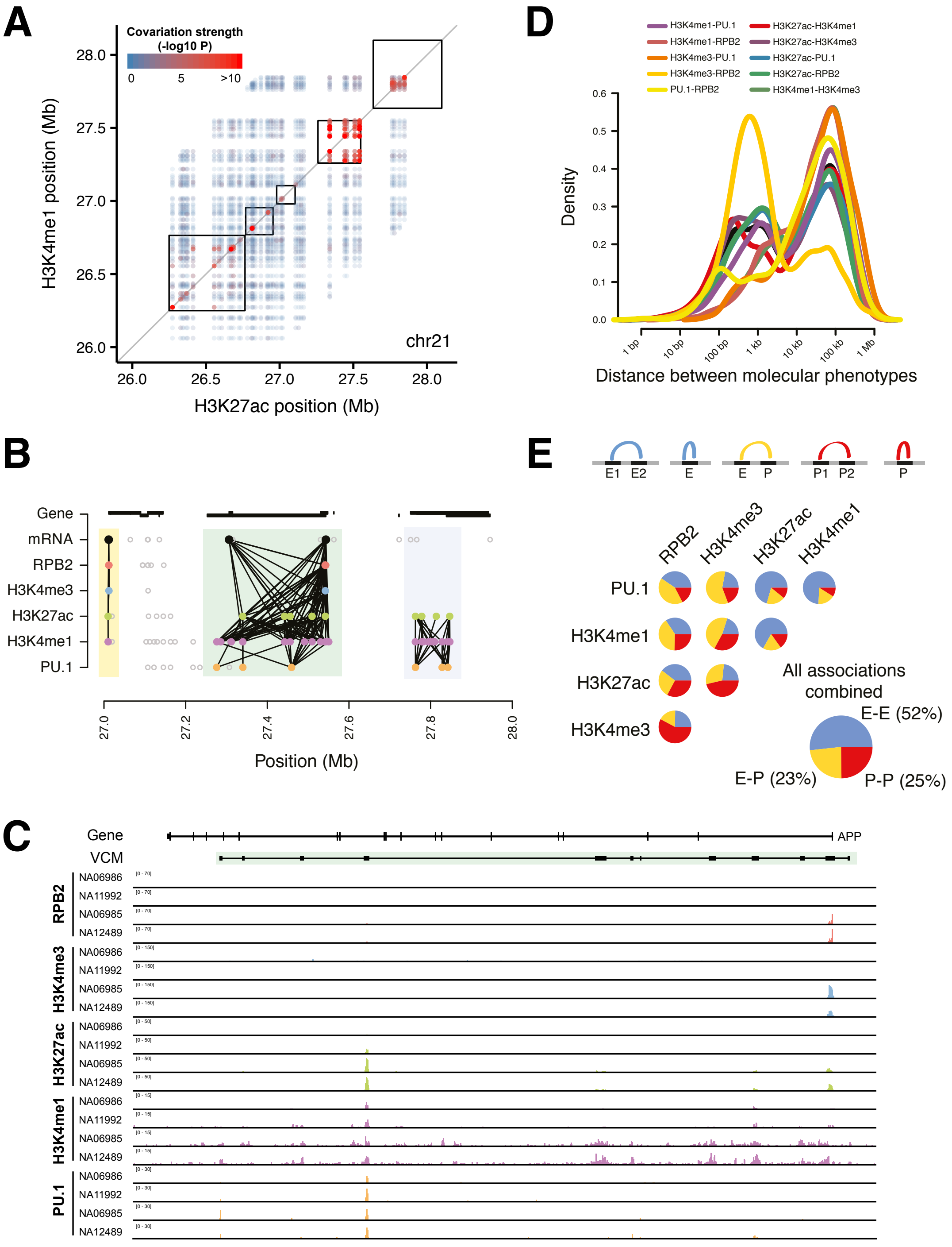


Figure 3.2: **Genome-wide associations among molecular phenotypes.** **A** Inter-individual co-association between the read depth at H3K27ac and H3K4me1 ChIP-seq peaks on chromosome 21 (26,000,000-28,000,000). The pairwise association strength (Pearson's P-value) is color-coded and ranges from blue (P=1) to red (P<1e-10). Chromosomal contact domains [88] are shown with black boxes. See Figure C.6H for molecular associations in this region based on other marks. **B** Significant associations between molecular phenotypes in a 1 Mb window on chr21

(27,000,000-28,000,000). Circles indicate variable (filled) or non-variable (open) enrichment of molecular marks (i.e. ChIP-seq peaks or gene expression). Lines connecting filled circles represent significant associations between molecular phenotypes (FDR 0.1%). **C** Selected individuals with either low (NA06986 and NA11992) or high (NA06985 and NA12489) enrichment of molecular marks around the APP gene locus. **D** Distance distribution between coordinated molecular phenotypes. **E** Annotation of *cis*-regulatory elements with coordinated enrichment of molecular marks into putative enhancers (E) and promoters (P). See also Figures C.1, C.2, and C.3.

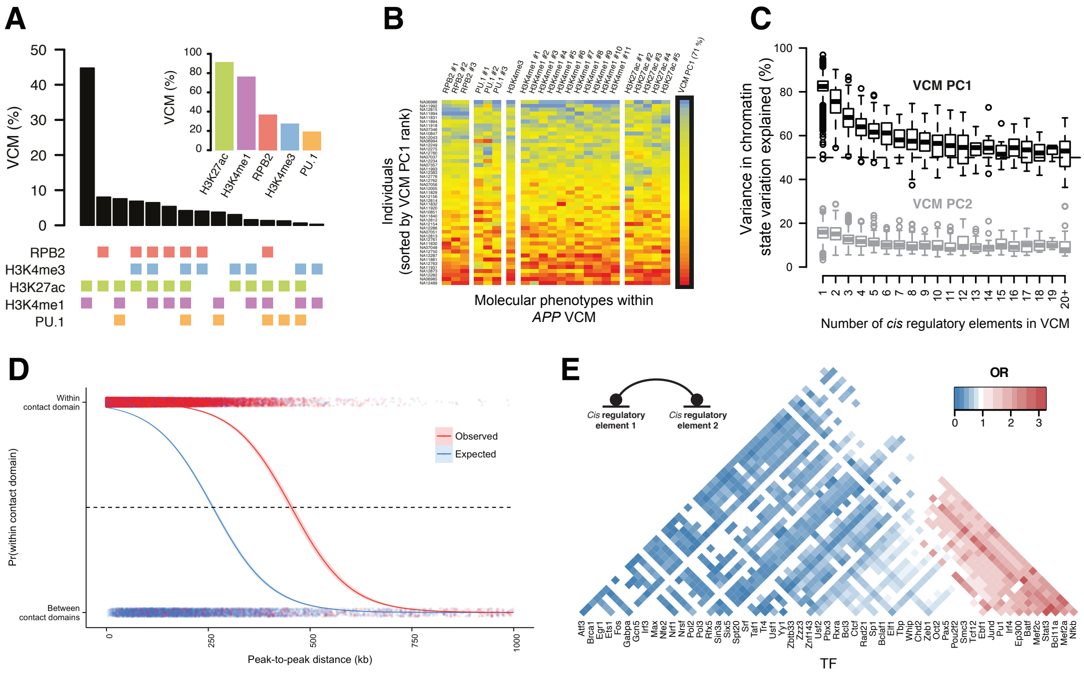


Figure 3.3: **Variable chromatin modules (VCMs).** **A** Molecular phenotype composition of VCMs. Bars (top) indicate the percentage of VCMs with specific combinations of molecular phenotypes (bottom). Inlet shows the percentage of VCMs with a specific molecular phenotype. **B** Coordination of molecular activity within VCMs. The heat map illustrates for 47 individuals (rows) the normalized signal of molecular marks (columns) that belong to the VCM spanning the APP gene locus (as shown in Figure 3.2B-C). Right column, the first principal component summarizes the majority (71%) of molecular variation within this VCM. **C** Percentage of molecular variation within VCMs that is explained by the first and second principal components. VCMs were divided according to the number of non-coding regions (domains). VCMs with >= 20 domains were grouped. **D** Enrichment of covariable *cis*-regulatory elements within chromosomal contact domains [88]. Red, covariable *cis*-regulatory elements; blue, random pairs of *cis*-regulatory elements. The probability indicates whether two covariable *cis*-regulatory elements are embedded within the same contact domain as opposed to two distinct contact domains. **E** Co-associations of TF pairs at non-overlapping, covariable *cis*-regulatory elements. Positive and negative odds ratios indicate significant enrichment/depletion of TF pairs (P<0.05 after Bonferroni correction). See also Figures C.2 and C.3.

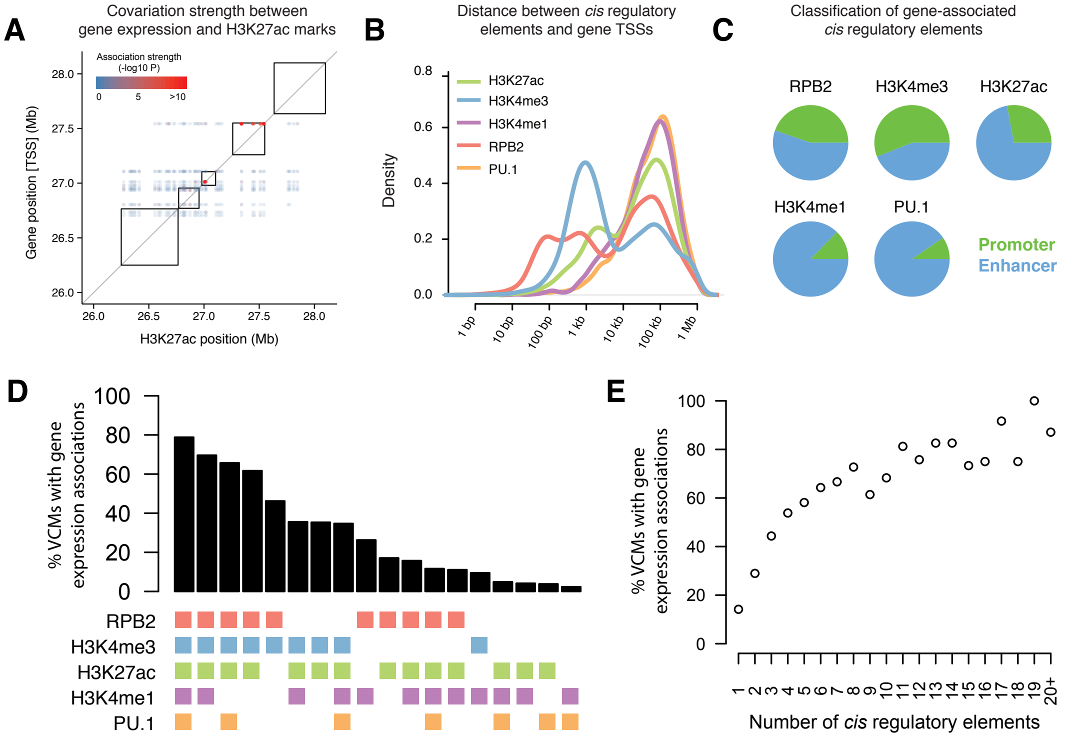


Figure 3.4: **Association between chromatin state and gene expression variation.** **A** Inter-individual co-variation between mRNA levels and H3K27ac enrichment signals at *cis*-regulatory elements on chr21 (26,000,000-28,000,000). The pairwise association strength (Pearson's P-value) is color-coded and ranges from blue (P=1) to red (P<1e-10) (legend, see Figure 3.2A). Chromosomal contact domains [88] are shown with black boxes. **B** Distance distribution in log-space between the transcription start site (TSS) and *cis*-regulatory elements with expression-linked molecular signals. **C** Classification of gene expression-linked *cis*-regulatory elements with molecular marks into putative enhancers and promoters (TSS +/- 5 kb). **D** Percentage of VCMs with gene expression associations (using the first principal component for VCM states) stratified by molecular compositions. **E**, Percentage of VCMs associated with gene expression stratified by VCM size (i.e., number of *cis*-regulatory elements that belong to a VCM). See also Figure C.3.

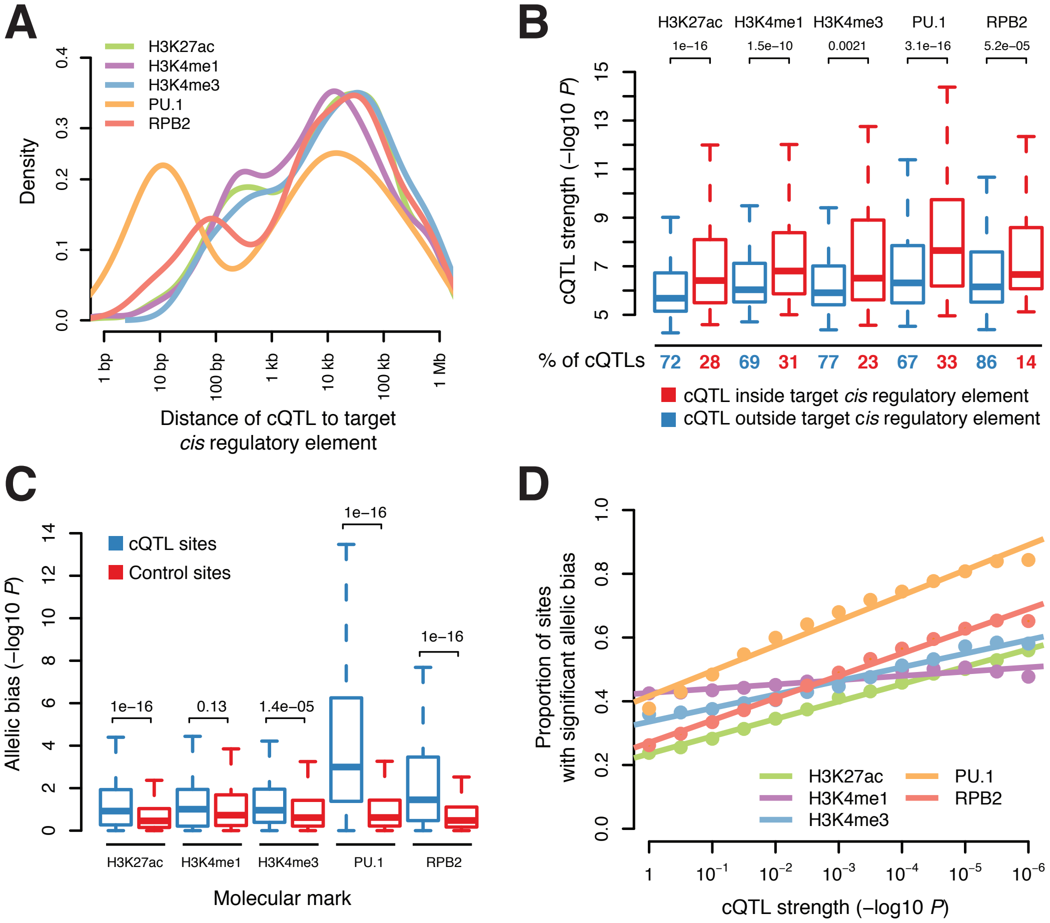
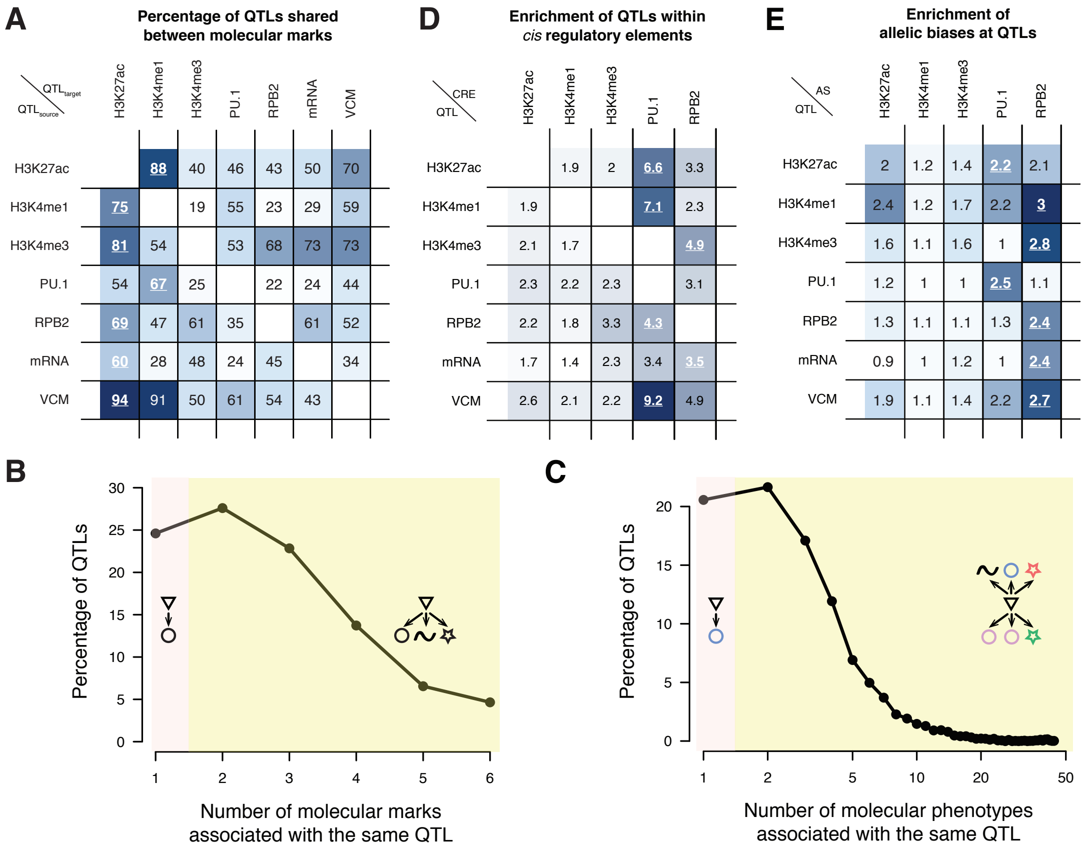


Figure 3.5: **Genetic control of chromatin state variation. A-B**, Quantitative trait locus (QTL) mapping for TF-DNA binding and HMs. **A** Bimodal distance distribution (in log10 space) between cQTLs and their associated *cis*-regulatory elements (FDR 10%). **B** Relationship between cQTL strength and genomic architecture. Boxplots demonstrate genetic association strength (-log10 P value) for QTL variants that map inside (red) and outside (blue) their target TF-bound/histone-modified regions. Percentages refer to the proportion of cQTLs that fall inside and outside their target regions. **C-D**, Allele-specific (AS) signals at cQTLs. **C** AS effect strength (-log10 binomial P-value) at heterozygous QTL (blue) and non-QTL variants (red). **D** Estimated frequency of AS effects (using π_1 statistics) at heterozygous variants as a function of cQTL strength (-log10 P value). For example, 83% of the heterozygous variants exhibit AS signals in PU.1-binding when considering genetic variants that are associated with PU.1-binding variation at P <10e-6. See also Figures C.4 and C.5.



QTL_{target} / QTL_{source}	H3K27ac	H3K4me1	H3K4me3	PU.1	RPB2	mRNA	VCM
H3K27ac		88	40	46	43	50	70
H3K4me1	75		19	55	23	29	59
H3K4me3	81	54		53	68	73	73
PU.1	54	67	25		22	24	44
RPB2	69	47	61	35		61	52
mRNA	60	28	48	24	45		34
VCM	94	91	50	61	54	43	

CRE / QTL	H3K27ac	H3K4me1	H3K4me3	PU.1	RPB2
H3K27ac		1.9	2	6.6	3.3
H3K4me1	1.9			7.1	2.3
H3K4me3	2.1	1.7			4.9
PU.1	2.3	2.2	2.3		3.1
RPB2	2.2	1.8	3.3	4.3	
mRNA	1.7	1.4	2.3	3.4	3.5
VCM	2.6	2.1	2.2	9.2	4.9

AS / QTL	H3K27ac	H3K4me1	H3K4me3	PU.1	RPB2
H3K27ac	2	1.2	1.4	2.2	2.1
H3K4me1	2.4	1.2	1.7	2.2	3
H3K4me3	1.6	1.1	1.6	1	2.8
PU.1	1.2	1	1	2.5	1.1
RPB2	1.3	1.1	1.1	1.3	2.4
mRNA	0.9	1	1.2	1	2.4
VCM	1.9	1.1	1.4	2.2	2.7

Figure 3.6: **Sharing of genetic associations between TF DNA binding, HMs, and gene expression. A** Estimation of QTLs that are shared between molecular marks. For example, 81% of H3K4me3 QTLs are also associated with H3K27ac marks. **B-C** Collateral impact of genetic variation on chromatin architecture and gene expression. **B** Percentage of tf-, hm-, and eQTLs being associated with multiple distinct molecular marks, i.e., DNA binding (PU.1, RPB2), HM levels (H3K4me1, H3K4me3, H3K27ac), and gene expression. For example, 75% of QTLs affect multiple marks. **C** Percentage of tf-, hm-, and eQTLs being associated with multiple molecular phenotypes (i.e. TF-binding, HM levels, and gene expression). For example, 7.5% of all QTLs affect >10 molecular phenotypes. **D** Enrichment of QTLs within active *cis*-regulatory elements. For example, vcmQTL variants map nine times more likely into PU.1-bound regions than expected by chance. **E** Estimation of allelic effect frequency (using π_1 statistics) at heterozygous QTL variants. For example, AS effects at H3K27ac sites are 2.2-fold more likely at PU.1 QTL variants as compared to all variants. See also Figures C.5 and C.6.

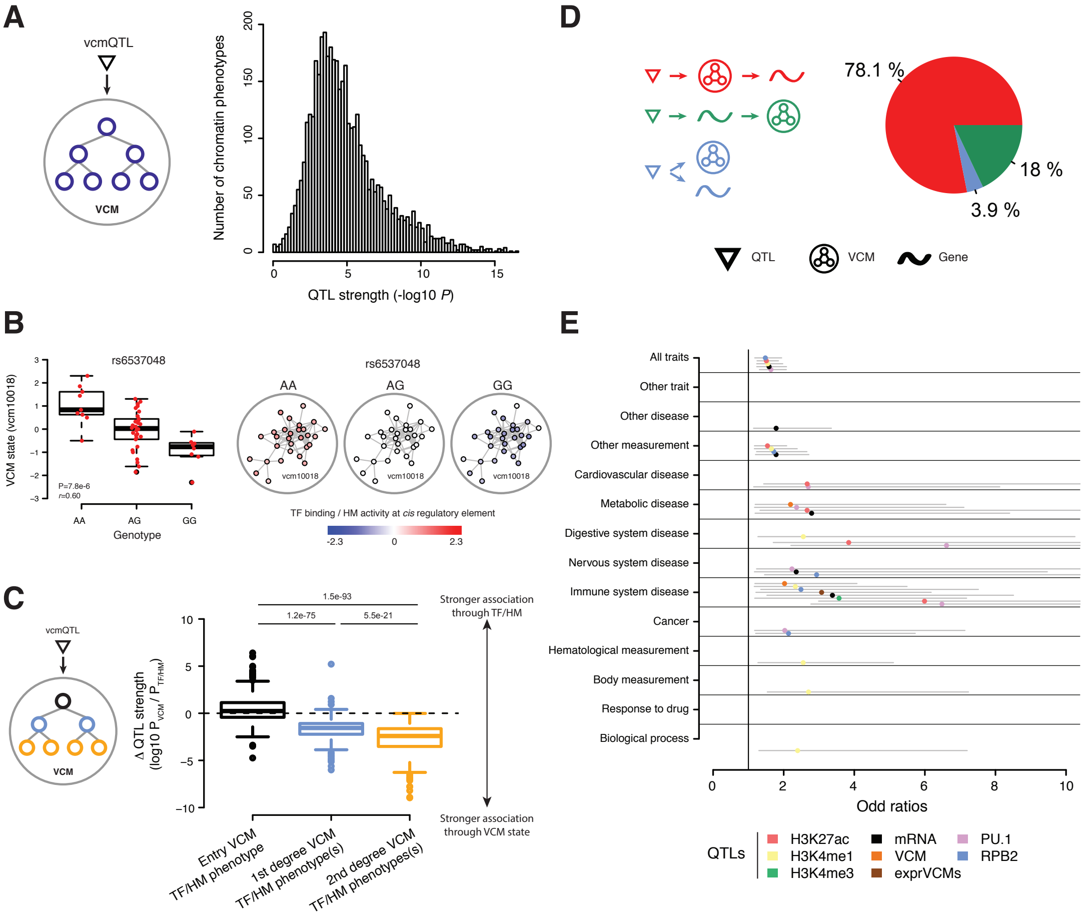


Figure 3.7: **Propagation of genetic signals through molecular phenotypes.** **A** Distribution of -log10-transformed association P-values for vcmQTL variants and VCM-defining molecular phenotypes. **B-C** Genetic variation exhibits direct and indirect effects on chromatin architecture. **B** Significant association between the SNP rs6537048 and the state of VCM vcm10018 (chr4:142,224,793-142,570,395) upstream of IL15. See Figure C.1I for molecular associations in this region based on all marks. Boxplot show the PCA-derived vcm10018 activity level split by genotype of the SNP rs6537048. Molecular marks within vcm10018 are themselves associated with rs6537048. Molecular association structure is shown together with rs6537048 genotype-averaged TF DNA binding and HM signals. Nodes define individual marks for specific molecular phenotypes (i.e., TF binding and HM) and grey lines significant associations between these molecular marks. **C** VCM associations are contrasted against the association strength of the same vcmQTL variant with individual molecular phenotypes (i.e. TF DNA binding and HM). The molecular association structure within VCMs is used to define three layers of molecular events (entry, 1st degree, and 2nd degree, see Main Text). Box plots show the ratio of genetic association strength between VCMs and the average of individual molecular phenotypes (i.e., $log_{10}P_{VCM}/P_{TF/HM}$). **D** Inference of causal relationships between VCM state and gene expres-

sion using Bayesian causality modeling. The frequency of the most likely model is shown. **E** Enrichment of cQTLs and eQTLs in complex disease susceptibility variants by trait class. See also Figure C.6.

3.5 Discussion

Our analyses uncovered extensive coordination in chromatin variation at and between *cis*-regulatory elements in a human population, revealing the existence of genomic compartments in the form of variable chromatin modules (VCMs). VCMs suggest a higher-order modular organization of gene regulation in the human genome, which is supported by the observation that VCMs are strongly enriched within chromosomal contact domains [88]. Interestingly, immunity-related genes are specifically associated with VCMs, consistent with the biological nature of LCLs. This finding implies that the resolution of topologically associated domains (TADs) that were so far detected [87, 88, 95] may extend to the level of individual genes (or sets of co-regulated genes), consistent with the observation of micro-topologies at the sub-Mb scale around key developmental genes in mouse embryonic stem cells [96]. Our data further suggest that population-level chromatin profiling might be an efficient strategy to assess putative chromatin interactions at high spatial resolution, complementing other molecular techniques aimed at mapping chromatin interactions [97], transcription-coupled chromatin remodeling events [98], TF-DNA binding-induced spreading of histone marks [99], and enhancer/promoter-gene interactions, respectively.

VCMs also provide a rational framework for explaining why regulatory events can vary independent of proximal sequence changes in molecular terms [52, 53, 55, 79, 80, 100]. Chromatin activity at *cis*-regulatory elements can be influenced by distally acting genetic variants of variable effect size, as we strongly suggest in this study for all analyzed molecular phenotypes. In addition, we found that the activity level of each VCM can be captured by a single quantitative phenotype, which suggests that molecular processes within each VCM (i.e. histone-mark deposition and TF binding) are subject to highly penetrant causal events. Our study provides strong support for the hypothesis that these events correspond in large part to genetic perturbations of TF-DNA interactions. This is based on the fact that vcmQTLs are (i) strongly enriched within TF-occupied regions, (ii) simultaneously perturb several layers of chromatin structure, and (iii) are in the majority of cases causal to the observed changes in gene expression. From this, a model emerges in which the perturbation of a single or a few TF-DNA interactions can act as a seed for coordinated, collateral regulatory changes within a respective VCM. We hypothesize that these changes are in large part mediated by long-range TF-TF cooperativity events given our observation that specific pairs of lineage-determining, signal-dependent, and architectural factors [89, 91, 101] are significantly enriched at VCM elements.

Interestingly, whereas ″promoter VCMs″ correlated frequently with gene expression, we found that only few ″enhancer VCMs″ were linked to nearby genes and only one quarter of PU.1 or H3K4me1 QTLs were shared with eQTLs. This finding may imply either (i) that abundant enhancer variation is of such small effect on target gene expression as to be undetectable given the sample size of this study, or (ii) that the affected enhancers are primed to conditionally regulate gene expression (for example in response to specific stimuli) [78, 102, 103]. Alternatively, these sequences may be subject to spurious regulatory activity, which would explain the findings that (i) only a minority of genetically variable TF-binding events result in differential gene expression (this work), (ii) a large portion of TF-DNA binding events have no functional consequence [93, 104], and (iii) TF binding sites tend to experience rapid turn-over [81, 105]. Another complementary interpretation involves the model of dose-dependent gene activation in which several TF binding sites in multiple elements cumulatively contribute to gene expression [106]. Under this model, loss of TF-DNA binding at one binding site would have little to no discernible functional consequence as long as the other implicated TF binding sites remain intact. This would in turn be consistent with our observation that VCMs involving multiple *cis*-regulatory elements were far more likely to correlate with gene expression variation than VCMs involving only one element.

Our present work on the discovery of molecular associations and cQTLs for key chromatin organization components in a human population sample provides unique insights and a novel framework for studying the molecular mechanisms underlying variable transcriptional programs between individuals.

3.6 Methods

3.6.1 Study samples

ChIP-seq and RNA-seq data were produced from lymphoblastoid cell lines (LCLs) of 54 samples from the [48]. All individuals were of European origin (Utah residents with ancestry from northern and western Europe and abbreviated as CEU). After excluding samples due to suspected swaps, contamination (see Supplementary Information C.3.4), or incomplete data availability (sample failed for a subset of assays) our final dataset consisted of 47 individuals for all ChIP assays and 46 individuals for gene expression measurements (Table C.1 for basic sample information).

3.6.2 ChIP-seq and mRNA-seq experiments

All sequencing assays (ChIP and mRNA) were produced from a single growth of LCLs and cell culture and cell fixation was performed as previously described [53]. The ChIPs for H3K27ac, H3K4me1, H3k4me3, PU.1 and RNA polymerase II (RPB2) were performed as described in the Supplementary Information C.1.1-C.1.3. RNA extraction was done following the procedure described in Supplementary Information C.2.1. Library preparation and sequencing done for ChIP and mRNA are described in detail in Supplementary Information C.1.4 and C.2.2, respectively. Short-read alignment for ChIP-seq and RNA-seq was performed using BWA 0.5.9 [107] against the hg19 build of the human reference genome supplemented with the Epstein-Barr virus (EBV) sequence. All sequencing data management was done using Samtools [108] (Supplementary Information C.1.5 and C.2.3). Summary of mapping statistics is provided in Table C.1B. All BAM files for this study have been submitted to the ArrayExpress Archive (`http://www.ebi.ac.uk/arrayexpress/`). The accession numbers are: E-MTAB-3656 (mRNA-seq data) and E-MTAB-3657 (ChIP-seq data).

3.6.3 From ChIP-seq experiments to molecular phenotypes

ChIP-seq peak calling was not directly performed in the current set of samples to avoid the issue of fuzzy peak boundaries. Instead, we used an independently derived peak set for each assay that is based on six 1000 Genomes Project Pilot individuals [53]. Quantifications for all peak-sample pairs were obtained by counting overlapping reads using HOMER [101], which resulted in a quantification matrix of size #samples x #peaks per assay (Supplementary Information C.1.6). Peak quantifications were scaled to adjust for differences in total library size and corrected for batch effects using PEER [109]. We empirically determined the optimal number of K PEER factors to be removed by finding the K leading to the highest number of discovered QTLs (Supplementary Information C.1.7).

3.6.4 From mRNA-seq experiments to molecular phenotypes

mRNA-seq data was quantified per sample based on GENCODE v15 (08/2012) gene annotations [110], resulting in a quantification matrix of size #samples x #genes. All genes with five samples (>10%) or more without any overlapping reads were removed and the remaining quantifications were scaled (10M reads) and corrected for batch effects (PEER K=15) (Supplementary Information C.2.4-C.2.5).

3.6.5 Genotype information

Genotypes for the 47 samples were obtained from two sources: (1) 34 with genome-wide sequence data from 1000 Genomes Phase1 v3 and (2) 13 other CEU samples with Illumina Omni2.5 genotype data. Both were merged by imputing untyped sequence variants into Illumina Omni2.5 data using IMPUTE2 [111]. Subsequently, all variants with minor allele frequency below 5% were removed. See Lappalainen et al. (2013) [56] for additional details on genotype processing.

3.6.6 Analytical methods for molecular phenotype-phenotype associations

Mapping molecular associations

To map associations between pairs of peaks, we proceeded as follows for each of the 15 possible unordered pairs of distinct molecular phenotypes. We measured inter-individual Pearson correlation and its significance (P-value) between quantifications of every possible pair of peaks within 1 Mb distance of each other. Then, we corrected for multiple-testing by controlling for a 0.1% false discovery rate using the R/qvalue package (Dabney A & Storey JD. qvalue: Q-value estimation for false discovery rate control. R package). Percentages (i.e. π_1 estimates) of truly associated pairs were also estimated as a by product (Supplementary Information C.3.1).

Building VCMs

We used graph theory to build VCMs and assumed that peaks are nodes and significant peak associations edges. Any two peaks belong to the same VCM as soon as there is a path (i.e. a sequence of edges) that links them together otherwise they belong to two distinct VCMs. Based on this, we implemented an iterative algorithm that assigns peaks to VCMs. Then, VCM state activity levels were obtained using principal component analysis (PCA) on quantifications of all peaks that belong to a VCM (Supplementary Information C.3.2).

Functional annotation of VCMs

We used the online service GREAT v2.0.2 to predict over-represented pathways and biological processes for VCM domains. Functional annotation of VCM-associated genes was performed using the online service ConsensusPathDB-human using the over-representation analysis module and gene ontology categories (BP level 2) (Supplementary Information C.3.9).

Enrichment in contact domains

We used high-resolution chromosomal contact domains for LCLs from Rao et al. (2014) [88] to estimate how more likely associated peak pairs occur within the same contact domain as compared to non-associated ones. To do so, we used logistic regression with within/between contact domains as the binary response, the association status (significant or not) as explanatory variable, and the peak-to-peak distance as a covariate (Supplementary Information C.3.11). TFs co-occurrence at VCM elements. We used the Fisher exact test to estimate enrichments of ENCODE TF-TF pairs at non-overlapping VCM elements (Supplementary Information C.3.13).

3.6.7 Analytical methods for quantitative trait loci (QTL)

Mapping QTLs

We mapped *cis*-acting QTLs by performing linear regressions between peak or gene quantifications and genotypes at all variant sites within 250 kb (*cis*-window around the gene TSS or the peak center). Then, we stored the best association for each peak/gene as a putative QTL and corrected (1) for multiple variants and (2) multiple peaks/genes being tested genome-wide. We used permutations and false discovery rate to correct for (1) and (2), respectively. In addition, we repeated this analysis multiple times with various *cis*-window sizes in order to determine the size providing the best trade-off between discovery power and distal effect capture (Figure C.4A-C; Supplementary Information C.3.3). This analysis has been performed using the software package FastQTL (`http://fastqtl.sourceforge.net/`).

Estimating proportion of shared QTLs

To see if a QTL for assay A1 is replicated in assay A2, we first found a A2 peak that matches the A1 peak by minimizing the distance between both and then we looked at the nominal P-value of association between the QTL and the matched A2 peak. By repeating this for all A1 QTLs, we can then estimate the proportion that is shared with A2 using the π_1 statistic (Supplementary Information C.3.5).

Detecting multiple effects of QTLs

To map out the peaks affected by a QTL, we measured association between the QTL and all features across all assays located within 250 kb, then divide the resulting P-values by the number of tested features (Bonferroni correction) and finally report as hits, associations with a P-value below the 0.05 threshold (Supplementary Information C.3.6).

Enrichment of QTLs within functional annotations

To measure how more likely than by chance a set of QTLs is located within a particular annotation, we developed an approach that corrects for the fact that QTLs and annotations are not uniformly distributed along the genome; the goal being to get more robust enrichment estimates. This method basically aims to find a null set of QTLs with some properties (e.g. distance to associated peak/gene) that match the original set (Supplementary Information C.3.7).

Enrichment of QTLs with GWAS hits

To measure how the QTL sets are enriched for GWAS hits, we used the NHGRI GWAS Catalog (Dec 8, 2014), generated 1,000 null sets of QTLs with matching properties (distance to associated peak/gene and MAF), and tested how often these null QTL sets overlap the GWAS hits as compared to the original QTL set. Note that two variants are assumed to overlap as soon as they are in high LD (Supplementary Information C.3.10).

QTL causality modeling

When a QTL is associated with two peaks (or genes), we inferred the most likely signal transmission path (i.e. the causal chain of events) through the two affected molecular phenotypes using Bayesian network modeling: we enumerate the three possible models (QTL =>A1 =>A2, QTL =>A2 =>A1 and QTL =>A1/QTL =>A2), estimate their respective likelihood, and assign the most likely model to each triplet (Supplementary Information C.3.8).

3.6.8 Analytical methods for allele-specific effects (ASE)

Mapping ASE

This was only performed on samples with sequence data (n = 34/47, Experimental Procedure 5) at heterozygous SNPs. Deviation from equilibrium (i.e. 50-50%) was characterized using binomial tests, accounting for multiple major sources of technical bias, such as reference allele mapping bias, clonal reads and non-unique mappability of reads as described previously [53, 56, 112] (Supplementary Information C.3.4). ASE analysis was also used as a QC step to identify putative sample swaps or contaminations.

Haplotypic ASE coordination

We looked at ASE measured at phased heterozygous SNPs falling within VCM peaks and assessed if the signal was consistent with the haplotype phase. In practice, we use logistic regression with concordance in allelic direction as response variable, association status (VCM/null) as explanatory variable and distance between peaks as covariates (Supplementary information C.3.12).

3.7 References

71. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (Sept. 2012).

72. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nature reviews. Genetics* **8,** 206–216 (Mar. 2007).

48. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (Oct. 2010).

49. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

52. Kasowski, M. *et al.* Variation in Transcription Factor Binding Among Humans. *Science* **328,** 232–235 (2010).

53. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342,** 744–747 (Nov. 2013).

55. Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome research* **22,** 860–869 (Mar. 2012).

56. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).

67. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature genetics* **44,** 1084–+ (Oct. 2012).

73. Waszak, S. M. *et al.* Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell* **162,** 1039–1050 (Aug. 2015).

74. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *The New England journal of medicine* **363,** 166–176 (July 2010).

75. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS genetics* **6,** e1000895 (Apr. 2010).

76. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics* **6,** e1000888 (Apr. 2010).

77. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nature genetics* **44,** 502–510 (May 2012).

78. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics* **13,** 613–626 (Sept. 2012).

79. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342,** 750–752 (Nov. 2013).

80. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342,** 747–749 (Nov. 2013).

81. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160,** 554–566 (Jan. 2015).

82. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nature reviews. Genetics* **16,** 197–212 (Apr. 2015).

83. Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* **503,** 487–492 (Nov. 2013).

84. Karczewski, K. J. *et al.* Cooperative transcription factor associations discovered using regulatory variation. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 13353–13358 (Aug. 2011).

85. Stefflova, K. *et al.* Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* **154,** 530–540 (Aug. 2013).

86. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 9440–9445 (Aug. 2003).

87. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (May 2012).

88. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (Dec. 2014).

89. Ong, C.-T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nature reviews. Genetics* **15,** 234–246 (Apr. 2014).

90. Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M. & Snyder, M. Genetic analysis of variation in transcription factor binding in yeast. *Nature* **464,** 1187–1191 (Apr. 2010).

91. Zhou, H. *et al.* Epstein-Barr virus oncoprotein super-enhancers control B cell growth. *Cell host & microbe* **17,** 205–216 (Feb. 2015).

92. Schlattl, A., Anders, S., Waszak, S. M., Huber, W. & Korbel, J. O. Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. **21,** 2004–2013 (Dec. 2011).

93. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The functional consequences of variation in transcription factor binding. *PLoS genetics* **10,** e1004226 (Mar. 2014).

94. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics* **45,** 124–130 (Feb. 2013).

95. De Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502,** 499–506 (Oct. 2013).

96. Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153,** 1281–1295 (June 2013).

97. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489,** 109–113 (Sept. 2012).

98. Smolle, M. & Workman, J. L. Transcription-associated histone modifications and cryptic transcription. *Biochimica et biophysica acta* **1829,** 84–97 (Jan. 2013).

99. Hathaway, N. A. *et al.* Dynamics and memory of heterochromatin in living cells. *Cell* **149,** 1447–1460 (June 2012).

100. Villar, D., Flicek, P. & Odom, D. T. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nature reviews. Genetics* **15,** 221–233 (Apr. 2014).

101. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38,** 576–589 (May 2010).

102. Calo, E. & Wysocka, J. Modification of Enhancer Chromatin: What, How, and Why? *Molecular cell* **49,** 825–837 (2013).

103. Shlyueva, D. *et al.* Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. *Molecular cell* **54,** 180–192 (Apr. 2014).

104. Farnham, P. J. Insights from genomic profiling of transcription factors. *Nature reviews. Genetics* **10,** 605–616 (Sept. 2009).

105. Dermitzakis, E. T. & Clark, A. G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Molecular biology and evolution* **19,** 1114–1121 (July 2002).

106. Spivakov, M. Spurious transcription factor binding: non-functional or genetically redundant? *BioEssays : news and reviews in molecular, cellular and developmental biology* **36,** 798–806 (Aug. 2014).

107. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25,** 1754–1760 (July 2009).

108. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25,** 2078–2079 (Aug. 2009).

109. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS computational biology* **6,** e1000770 (May 2010).

110. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22,** 1760–1774 (Sept. 2012).

111. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5,** e1000529 (June 2009).

112. Waszak, S. M. *et al.* Identification and removal of low-complexity sites in allele-specific analysis of ChIP-seq data. *Bioinformatics (Oxford, England)* **30,** 165–171 (Jan. 2014).

Conclusion

The three chapters presented in this thesis provide significant new insights into the extent of genetic variation and its link and potential implication for gene expression regulation. Each of the three chapters approaches the topic from a different angle, each providing unique views on the functional consequences of genetic variation in natural populations. In the first project we specifically looked at the extent of copy number variation in Cynomolgus monkey and used gene expression changes to assess their potential implications for the organism. Even though this species is very widely used in biomedical and pharmaceutical research, the extent of genetic variation and especially copy number variation within these animals has not been studied extensively so far. We find considerable copy number variation among the sampled individuals, which comes not unexpected, because unlike inbred laboratory mice strains, Cynomolgus monkeys used in pharmaceutical research are regularly captured in wild population across the world. In line with other studies [39], we discover predominately small variants of a few kilobases length as expected in healthy individuals from natural populations. The detected copy number variation clearly separates our individuals according to populations. This indicates a diverse genetic background in pharmaceutical studies when using Cynomolgus monkeys originating from different populations. Our results show that part of this variation is linked gene expression changes in vitally important tissues. Multiple copy number polymorphisms and associated gene expression changes within a cluster of olfactory receptor genes on chromosome 7 demonstrate intraspecific functional variation in a region well known for genomic rearrangements [45, 46, 152].

The physiological consequences of our findings remain unclear, but this study represents an important first step towards a better understanding of the biological differences among Cynomolgus monkeys used in pharmaceutical research. There is great interest in this from both an economical ethical perspective. Secondary effects because of unknown reasons usually lead to a stop of the development of the drug in question and therefore big financial loss. Better biological knowledge of the used test species can prevent this and has the potential to refine and reduce animal tests in pharmaceutical research. Moreover, this would also reduce unnecessary suffering of animals. The other two projects presented in this thesis approach the question of genetic effects on gene expression in humans, a very extensively studied species. This has the big advantage that we can incorporate much already present data and knowledge. Many studies concerning genetic variation and gene expression have been already performed with much greater sample size [56, 153]. The novelty of our projects comes from the integration of different additional molecular phenotypes to further investigate to molecular mechanisms linked to gene expression changes. Even though many studies focused on protein-DNA binding and chromatin modifications, the effect of genetic variation on these processes and the potential implications for gene regulation were not studied extensively before. In the project presented in chapter two, we investigated the behavior of chromatin phenotypes and gene expression and the implications of genetics by exploiting the family structure of the two trios. We observed that the molecular phenotypes are correlated in their activity in functionally annotated regions such as gene promoters and putative enhancers. Furthermore we showed extensive allele specific activity and allelic coordination among the studied molecular phenotypes. In the follow-up project presented in chapter three, we extended the previous study to specifically investigate the local coordination between chromatin components and their link to gene expression regulation in the context of genetic variation. We showed, that chromatin components are organized in local, variable modules, which are strongly enriched in chromosomal contact domains. The measured activity of these chromatin modules co-varies across individuals and numerous genetic associations were detected for single chromatin markers, chromatin modules and gene expression levels. By applying a Bayesian network approach, we could shed light onto the causal sequence of regulatory changes. We suggest that changes in the chromatin activity landscape via perturbation of regulatory elements by genetic variants are a likely mechanism underlying gene expression changes associated to genetic variation. The logical next step will now be to identify theses cis-regulatory elements and to assess their downstream causal effects.

Obviously our statistical approaches violate one of the key assumptions in causality inference, namely that all potential candidate variables must be sampled. Even though we surely cannot infer mechanistic causality between molecular components, we are confident that our results allow us to make statements about the sequence of observed events. A special case is PU.1 motif disruption, which allows us to make strong functional assumptions that the observed change in PU.1 activity is caused by disruption of the binding motif by genetic variants. This provides a prime example of how genetic variants can perturb regulatory elements leading to changes in chromatin activity and potential downstream changes in gene expression. The extent and importance of this mechanism on a genome-wide level, including binding sites of other transcription factors, remains uncertain, but our findings are in line with other studies which suggest perturbation of chromatin activity and gene expression regulation by genetic variants through transcription factor binding sites [52, 79, 154].

For all three presented projects, follow-up studies are a logical conclusion to further investigate the molecular mechanisms perturbed by genetic variants. However, this could be achieved in several ways. On one hand, one could add more breadth by sampling more individuals and increasing statistical power to discover additional associations and eventually explain more of the observed variability. On the other hand, one could add more depth by including additional functional assays and experiments to test specific hypotheses suggested by our results. To make the research truly conclusive, further integration of different molecular phenotypes representative of regulatory mechanisms in combination with computational methods will enable to pinpoint candidate molecular processes perturbed by genetic variants. This will provide new insight into causal links and generate hypotheses, which can be experimentally tested. In some cases such as transcription factor binding motif disruption this is rather straight-forward, however to establish causal links between downstream regulatory elements such as cofactor binding and histone modification will be much more challenging. These problems can only be solved by a strong interplay between computational and experimental research needed, where computational results generate new hypotheses that can be tested experimentally in a precise and conclusive manner. Such a hypothesis driven approach is central to the idea of systems biology, but is not easy to realize in modern genetics. Classical systems biology problems often deal with relatively small molecular networks, with much information on the individual components (e.g [155]). However, in genomic studies as presented here, we are often confronted with a multitude of candidate causal links, with little knowledge about their connection and behavior. Novel computational tools will be required to mine the massive genomic data and to identify relevant elements and interactions and to turn correlation based results into meaningful models of causality. Of course experimental technologies, which allow efficient testing of generated hypotheses in reasonable time will be crucial as well. Many classical experimental techniques such as for instance mouse or zebra fish models are not suited to test the huge number of genetic associations generated by todays technology. High-throughput functional assays such as for example high-throughput yeast one-hybrid (Y1H) and techniques from synthetic biology to assess specific regulatory hypotheses seem very promising. Once this tight interplay between computational and experimental biology is achieved, we will be able to realize systems genetics and to investigate the genomes function in a true systems biology approach.

In summary, the results presented in this thesis significantly extend the knowledge about the implications of genetic variation associated with gene expression. We provide novel insights into the extent of copy number variation and its link to gene expression levels in Cynomolgus monkeys, a key species in pharmaceutical research. Furthermore, we demonstrate how genetic variation, chromatin components and gene expression regulation are connected in humans and enlightened their interplay. Our results suggest that changes in chromatin activity are linked to changes in gene expression regulation, and that perturbations of regulatory elements by genetic variants are a likely cause. Overall, even though many new questions arose and many remain open, this work significantly contributed to the advancement in understanding the functional impact of genetic variation on gene expression.

Bibliography

Introduction

113. Chodasewicz, K. Evolution, reproduction and definition of life. *Theory in Biosciences* **133,** 39–45 (2014).

114. Kirschner, M. & Gerhart, J. Evolvability. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 8420–8427 (July 1998).

115. Darwin, C. *On the Origin of Species by Means of Natural Selection or the Preservation of Favored Races in the Struggle for Life* (Murray, London, 1859).

116. Larson, E. J. *Evolution* (Modern Library, Aug. 2006).

117. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171,** 737–738 (Apr. 1953).

118. Lewontin, R. C. *The Genetic Basis of Evolutionary Change* 1974.

119. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nature reviews. Genetics* **10,** 241–251 (Apr. 2009).

120. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature genetics* **36,** 949–951 (Sept. 2004).

1. Scherer, S. W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nature genetics* **39,** S7–15 (July 2007).

121. Campbell, C. D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nature genetics* **44,** 1277–1281 (Nov. 2012).

122. Shendure, J. & Akey, J. M. The origins, determinants, and consequences of human mutations. *Science* **349,** 1478–1483 (Sept. 2015).

123. Kloosterman, W. P. *et al.* Characteristics of de novo structural changes in the human genome. *Genome research* **25,** 792–801 (June 2015).

124. Hughes, A. L. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99,** 364–373 (Oct. 2007).

125. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nature reviews. Genetics* **8,** 610–618 (Aug. 2007).

126. Kimura, M. The neutral theory of molecular evolution: a review of recent evidence. *Idengaku zasshi* **66,** 367–386 (Aug. 1991).

127. O'Sullivan, B. P. & Freedman, S. D. Cystic fibrosis. *Lancet (London, England)* **373,** 1891–1904 (May 2009).

128. Filges, I. & Friedman, J. M. Exome sequencing for gene discovery in lethal fetal disorders - harnessing the value of extreme phenotypes. *Prenatal diagnosis* **35,** 1005–1009 (Oct. 2015).

129. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304,** 1321–1325 (May 2004).

130. Douglas, A. T. & Hill, R. D. Variation in vertebrate cis-regulatory elements in evolution and disease. *Transcription* **5,** e28848 (2014).

71. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (Sept. 2012).

131. Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nature genetics* **30,** 233–237 (Feb. 2002).

132. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics* **74,** 1111–1120 (June 2004).

133. Stranger, B. E. & Dermitzakis, E. T. The genetics of regulatory variation in the human genome. *Human genomics* **2,** 126–131 (June 2005).

134. Lappalainen, T. & Dermitzakis, E. T. Evolutionary history of regulatory variation in human populations. *Human molecular genetics* **19,** R197–203 (Oct. 2010).

72. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nature reviews. Genetics* **8,** 206–216 (Mar. 2007).

48. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (Oct. 2010).

135. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (Sept. 2010).

136. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* **42,** D1001–6 (Jan. 2014).

137. Hirschhorn, J. N. & Gajdos, Z. K. Z. Genome-wide association studies: results from the first few years and potential implications for clinical medicine. *Annual review of medicine* **62,** 11–24 (2011).

138. Moraes, C. F. *et al.* Lessons from genome-wide association studies findings in Alzheimer's disease. *Psychogeriatrics : the official journal of the Japanese Psychogeriatric Society* **12,** 62–73 (Mar. 2012).

139. Pan, S., Naruse, H. & Nakayama, T. Progress and issues of the genome-wide association study for hypertension. *Current medicinal chemistry* **22,** 1016–1029 (2015).

140. Buono, R. J. Genome wide association studies (GWAS) and common forms of human epilepsy. *Epilepsy & behavior : E&B* **28 Suppl 1,** S63–5 (July 2013).

141. Low, S.-K., Takahashi, A., Mushiroda, T. & Kubo, M. Genome-wide association study: a useful tool to identify common genetic variants associated with drug toxicity and efficacy in cancer pharmacogenomics. *Clinical cancer research : an official journal of the American Association for Cancer Research* **20,** 2541–2552 (May 2014).

142. Motsinger-Reif, A. A. *et al.* Genome-wide association studies in pharmacogenomics: successes and lessons. *Pharmacogenetics and genomics* **23,** 383–394 (Aug. 2013).

49. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

143. Stranger, B. E. & Raj, T. Genetics of human gene expression. *Current opinion in genetics & development* **23,** 627–634 (Dec. 2013).

144. Stranger, B. E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS genetics* **1,** e78 (Dec. 2005).

2. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nature genetics* **39,** 1217–1224 (Oct. 2007).

145. Migliavacca, E. *et al.* A Potential Contributory Role for Ciliary Dysfunction in the 16p11.2 600 kb BP4-BP5 Pathology. *American Journal of Human Genetics* **96,** 784–796 (May 2015).

3. Merla, G. *et al.* Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. *American Journal of Human Genetics* **79,** 332–341 (Aug. 2006).

4. Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315,** 848–853 (Feb. 2007).

5. Henrichsen, C. N. *et al.* Segmental copy number variation shapes tissue transcriptomes. **41,** 424–429 (Mar. 2009).

146. Westra, H.-J. & Franke, L. From genome to function by studying eQTLs. *Biochimica et biophysica acta* **1842,** 1896–1902 (Oct. 2014).

147. Kitano, H. Systems biology: A brief overview. *Science* **295,** 1662–1664 (Jan. 2002).

148. Chuang, H.-Y., Hofree, M. & Ideker, T. A decade of systems biology. *Annual review of cell and developmental biology* **26,** 721–744 (2010).

149. Sauer, U., Heinemann, M. & Zamboni, N. Genetics. Getting closer to the whole picture. *Science* **316,** 550–551 (Apr. 2007).

150. Schadt, E. E. Molecular networks as sensors and drivers of common human diseases. *Nature* **461,** 218–223 (Sept. 2009).

151. Civelek, M. & Lusis, A. J. Systems genetics approaches to understand complex traits. *Nature reviews. Genetics* **15,** 34–48 (Jan. 2014).

50. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (Feb. 2012).

51. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328,** 235–239 (Apr. 2010).

52. Kasowski, M. *et al.* Variation in Transcription Factor Binding Among Humans. *Science* **328,** 232–235 (2010).

Conclusion

52. Kasowski, M. *et al.* Variation in Transcription Factor Binding Among Humans. *Science* **328,** 232–235 (2010).

39. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349,** aab3761–aab3761 (Sept. 2015).

45. Rudd, M. K. *et al.* Comparative sequence analysis of primate subtelomeres originating from a chromosome fission event. *Genome research* **19,** 33–41 (Jan. 2009).

46. Ventura, M. *et al.* Neocentromeres in 15q24-26 map to duplicons which flanked an ancestral centromere in 15q25. *Genome research* **13,** 2059–2068 (Sept. 2003).

56. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).

79. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342,** 750–752 (Nov. 2013).

152. Giannuzzi, G. *et al.* Hominoid fission of chromosome 14/15 and the role of segmental duplications. *Genome research* **23,** 1763–1773 (Nov. 2013).

153. Bryois, J. *et al.* Cis and trans effects of human genomic variants on gene expression. *PLoS genetics* **10,** e1004461 (July 2014).

154. Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162,** 1051–1065 (Aug. 2015).

155. Raspopovic, J., Marcon, L., Russo, L. & Sharpe, J. Modeling digits. Digit patterning is controlled by a Bmp-Sox9-Wnt Turing network modulated by morphogen gradients. *Science* **345,** 566–570 (Aug. 2014).

Appendices

Appendix A

Copy number variation in Cynomolgus monkeys linked to tissue specific gene expression

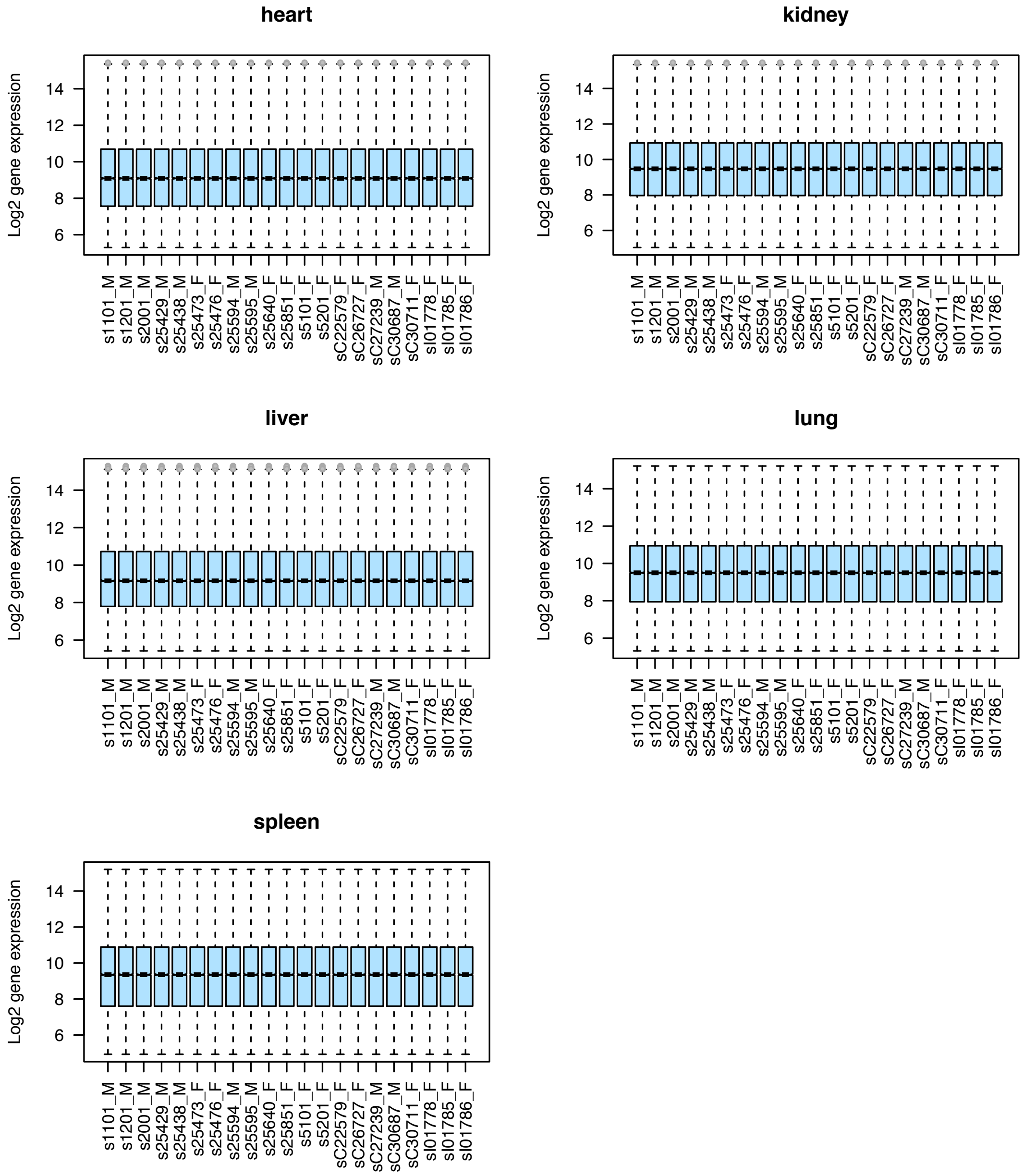


Figure A.1: **Gene expression data.** Distribution of gene expression data in all tissues for all samples used for eQTL mapping after RMA and quantile normalization.

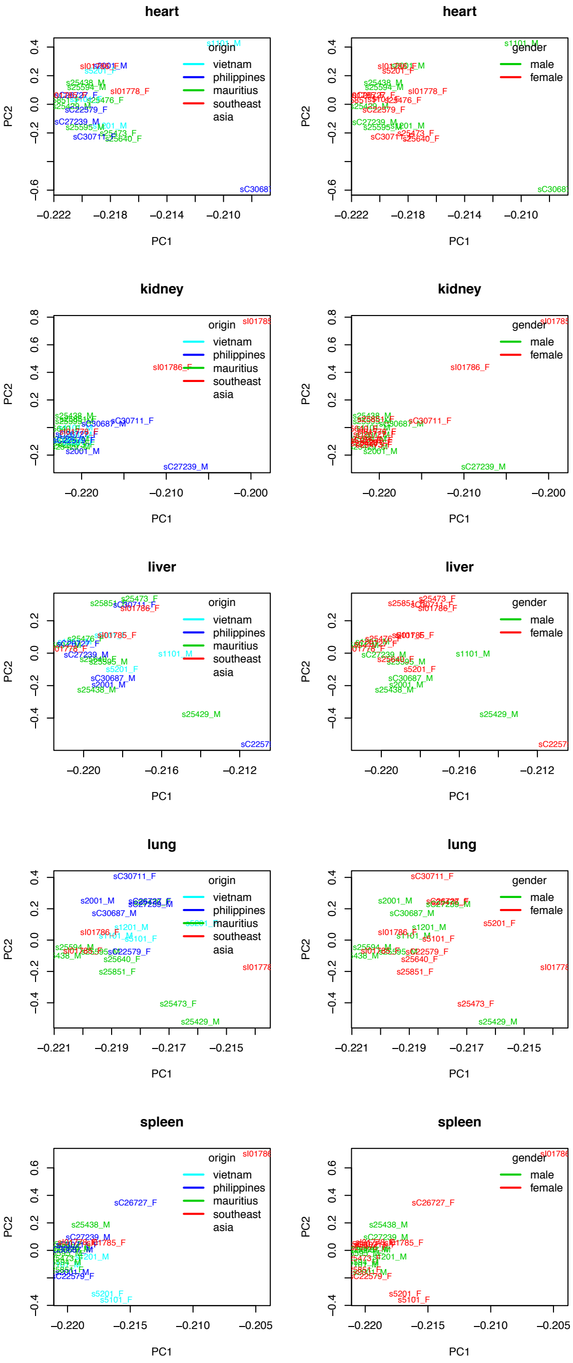


Figure A.2: **Gene expression PCA.** Loadings of the first and second principal component based on PCA preformed on gene expression data from all individuals used for eQTL mapping in all tissues separately. Color-coded for either sample origin or gender.

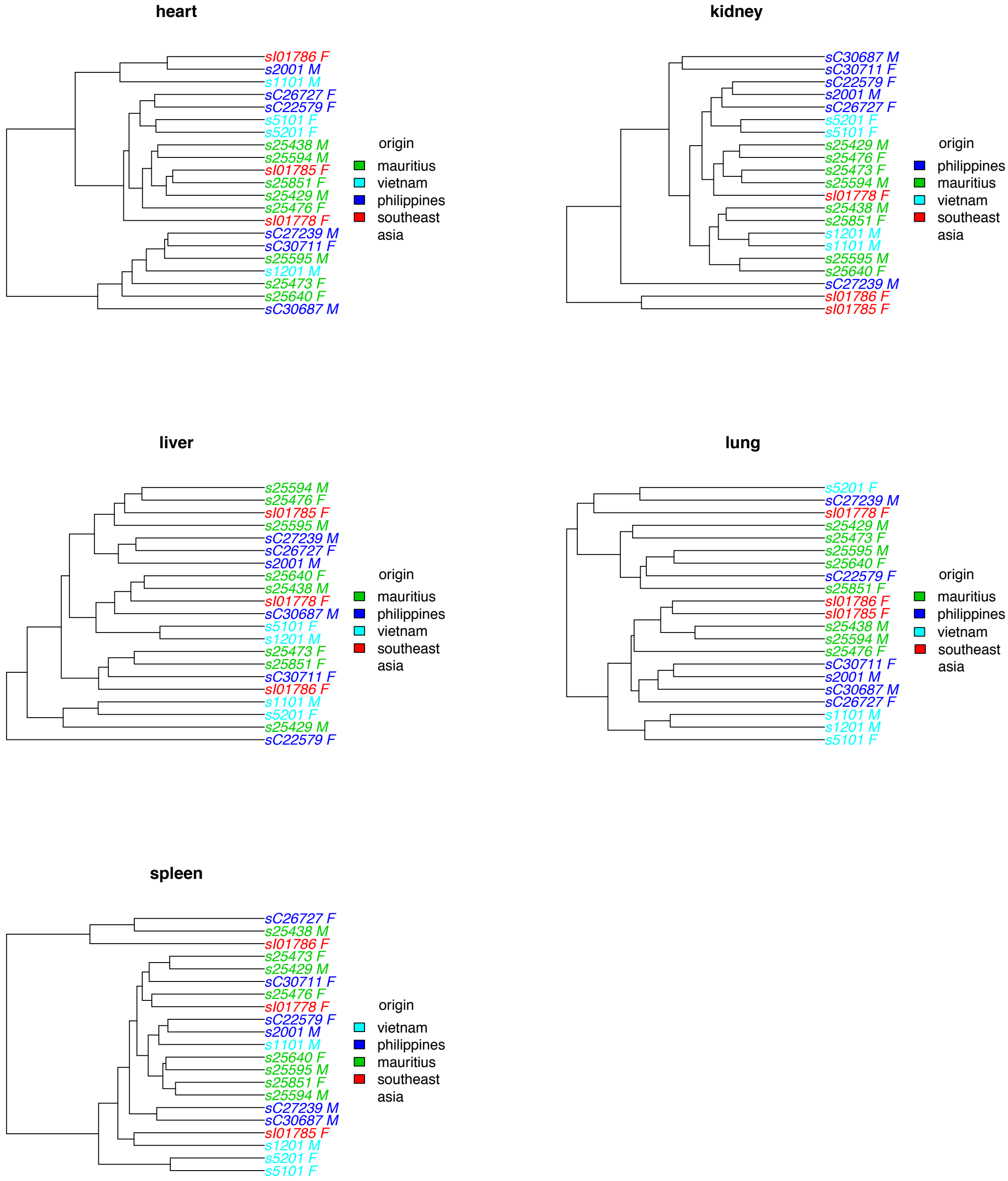


Figure A.3: **Gene expression clustering.** Hierarchical clustering of gene expression data from all individuals used for eQTL mapping in all tissues.

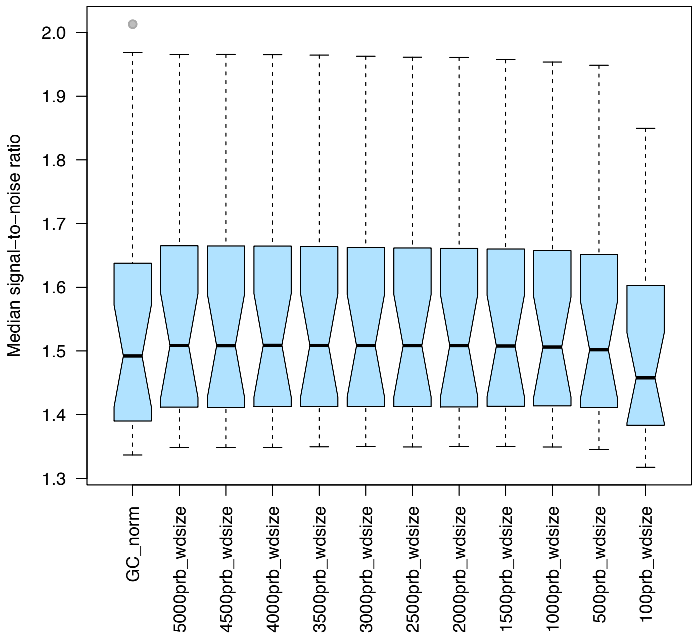


Figure A.4: **Wave artifact normalizations SNR.** Average signal-to-noise ratio (SNR) per sample for aCGH probes within CNVs called from probe GC- content normalized aCGH data across all tested LOESS fraction values.

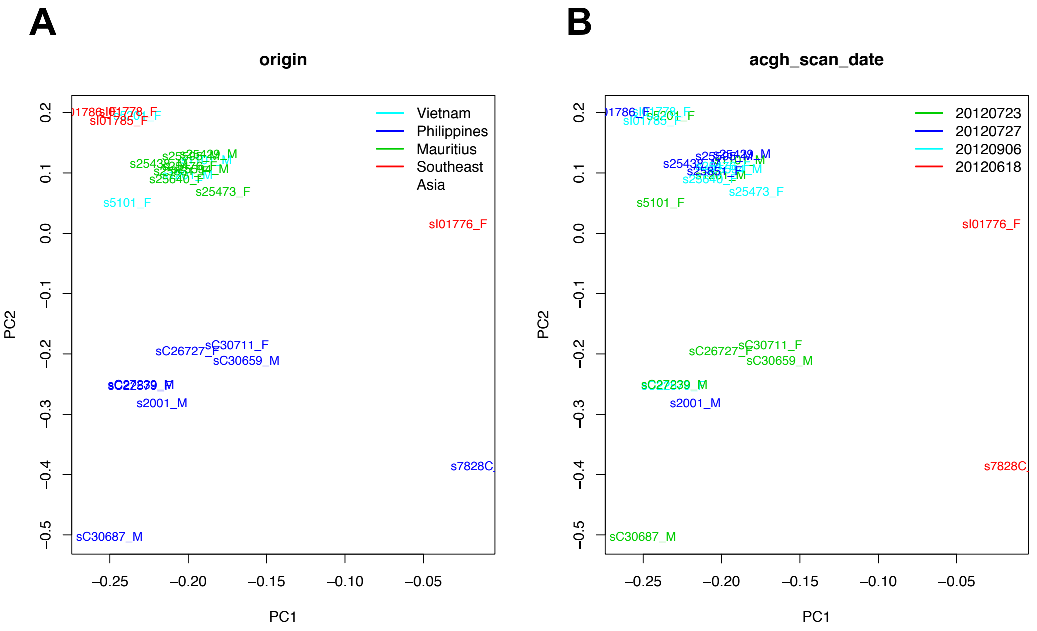


Figure A.5: **Normalized aCGH data PCA.** Loadings of the first and second principal component based on PCA performed with normalized aCGH data for all 24 samples. Color-coded for either sample origin **A** or aCGH scan data **B**.

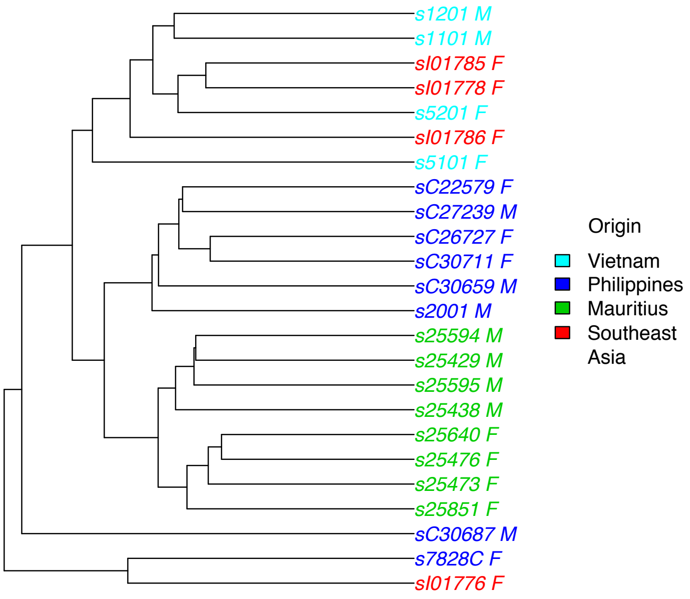


Figure A.6: **Normalized aCGH data clustering.** Hierarchical clustering based on Euclidean distance between normalized aCGH data from all 24 samples color-coded for sample origin.

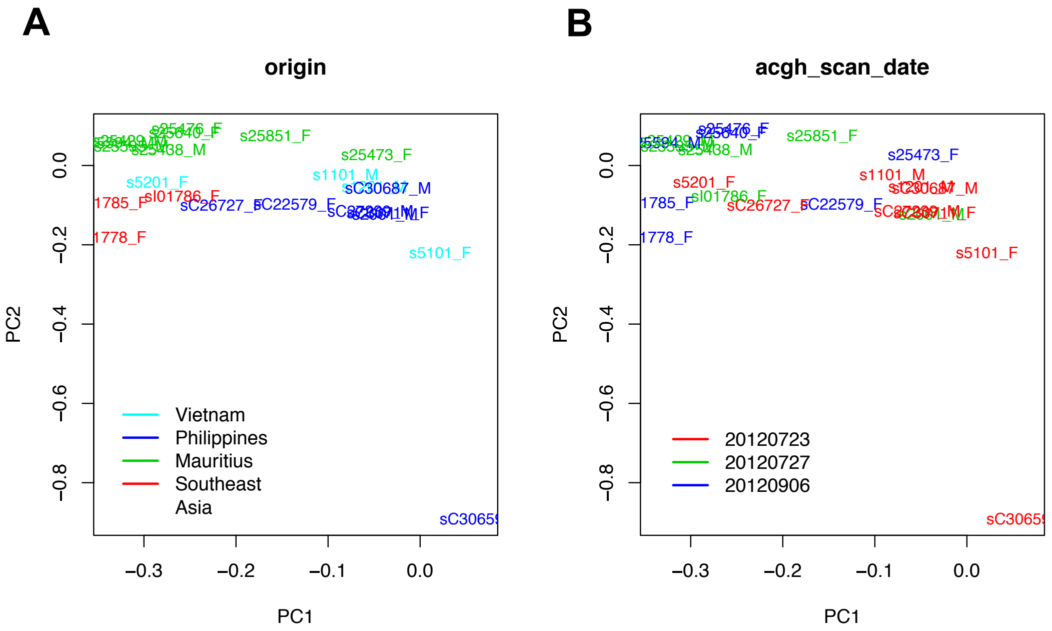


Figure A.7: **Initial CNV regions PCA.** Loadings of the first and second principal component based on PCA performed with CNV region (n=17,599) genotypes based on CNV calling with 22 samples. Color-coded for either sample origin **A** or aCGH scan data **B**.

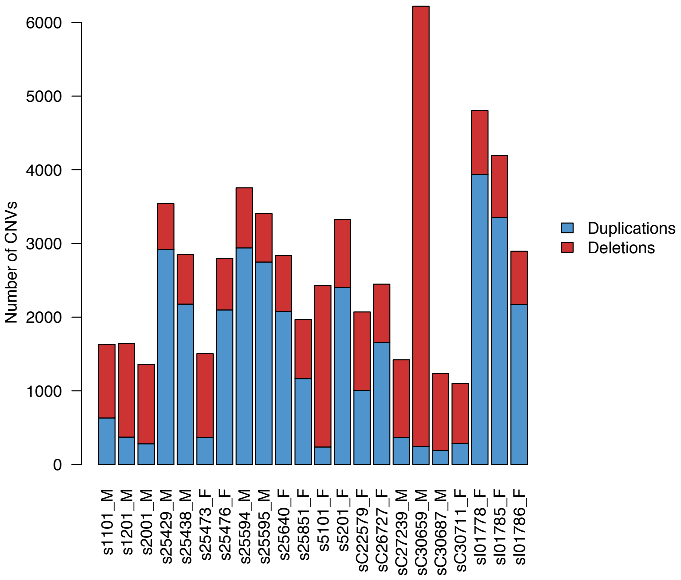
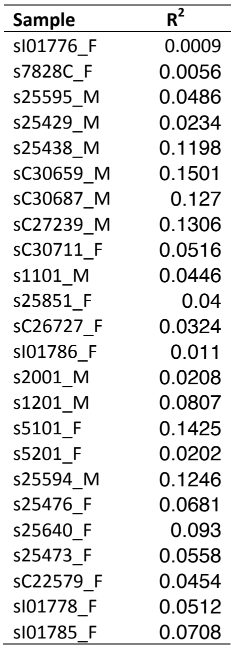


Figure A.8: **Initial CNV regions.** Number of deletion and duplications detected per individual for CNV regions (n=17,599) obtained from CNV calling with 22 samples.

Table A.1: **GC-content R-squared.** R-squared values obtained from linear models used to normalize for aCGH probe GC-content.



Sample	R^2
sl01776_F	0.0009
s7828C_F	0.0056
s25595_M	0.0486
s25429_M	0.0234
s25438_M	0.1198
sC30659_M	0.1501
sC30687_M	0.127
sC27239_M	0.1306
sC30711_F	0.0516
s1101_M	0.0446
s25851_F	0.04
sC26727_F	0.0324
sl01786_F	0.011
s2001_M	0.0208
s1201_M	0.0807
s5101_F	0.1425
s5201_F	0.0202
s25594_M	0.1246
s25476_F	0.0681
s25640_F	0.093
s25473_F	0.0558
sC22579_F	0.0454
sl01778_F	0.0512
sl01785_F	0.0708

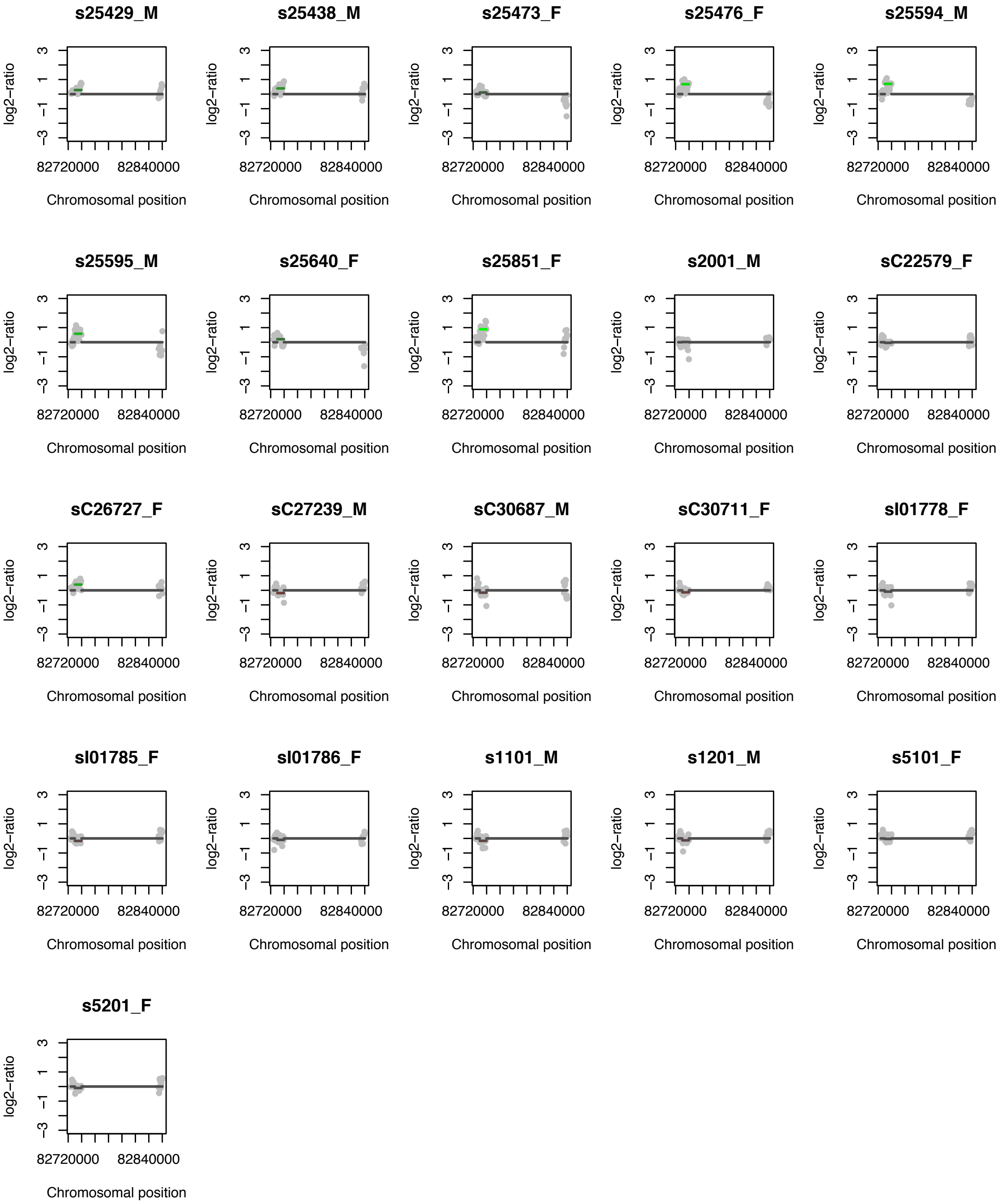


Figure A.9: **OR4K17 eQTL CNV.** Profiles of a CNV locus on chromosome 7 associated with expression changes of the OR4K17 and OR4K13 genes. CNV signals in comparison to a reference standard are displayed as $\log_2$-ratio along genomic positions of chromosome 7 (grey dots). Green or red bar intensity denotes higher, respectively lower median $\log_2$-ratio of probes within the CNV.

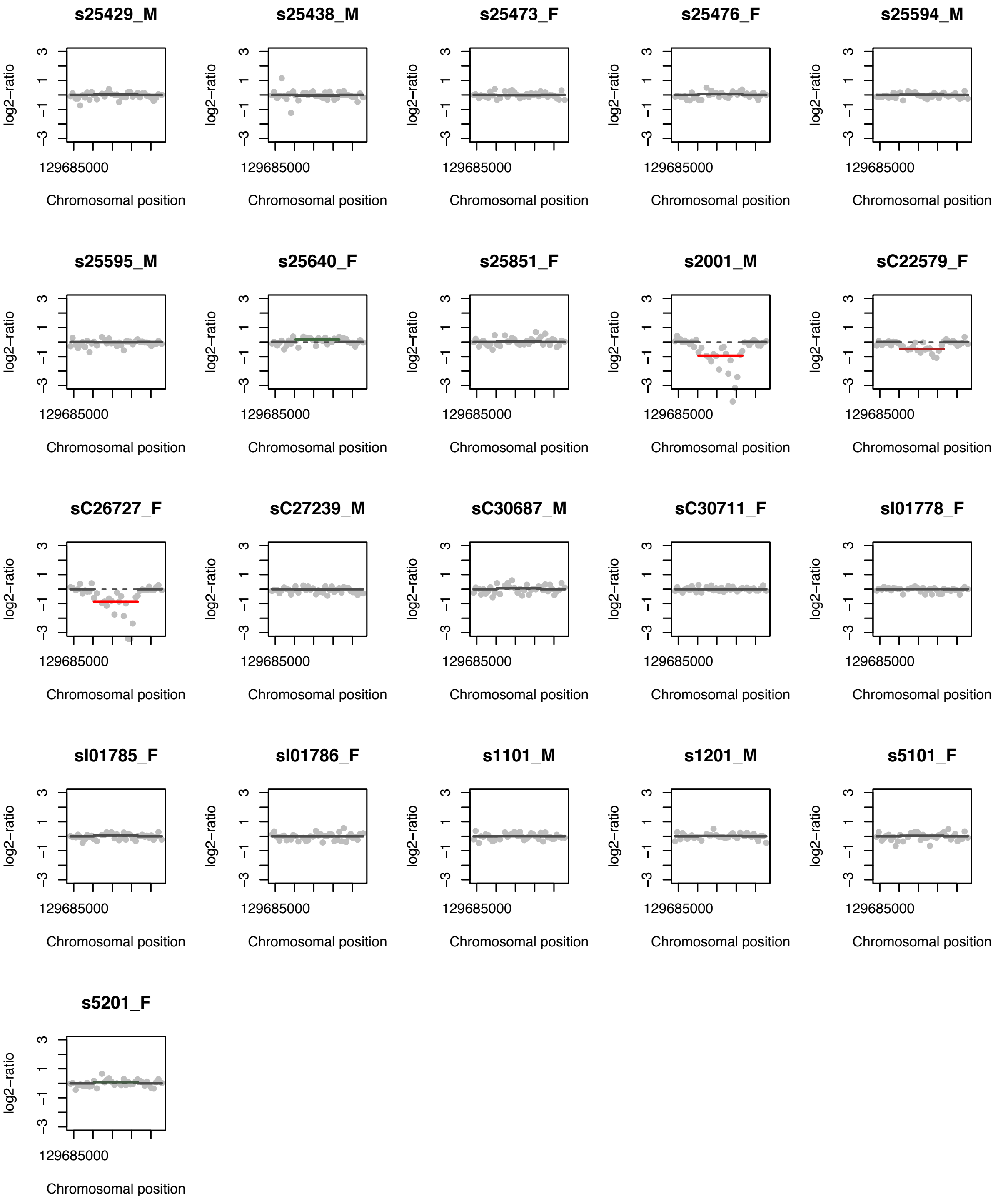


Figure A.10: **ABCB4 eQTL CNV.** Profiles of a CNV locus on chromosome 7 associated with expression changes of the ABCB4 gene. CNV signals in comparison to a reference standard are displayed as $\log_2$-ratio along genomic positions of chromosome 3 (grey dots). Green or red bar intensity denotes higher, respectively lower median $\log_2$-ratio of probes within the CNV.

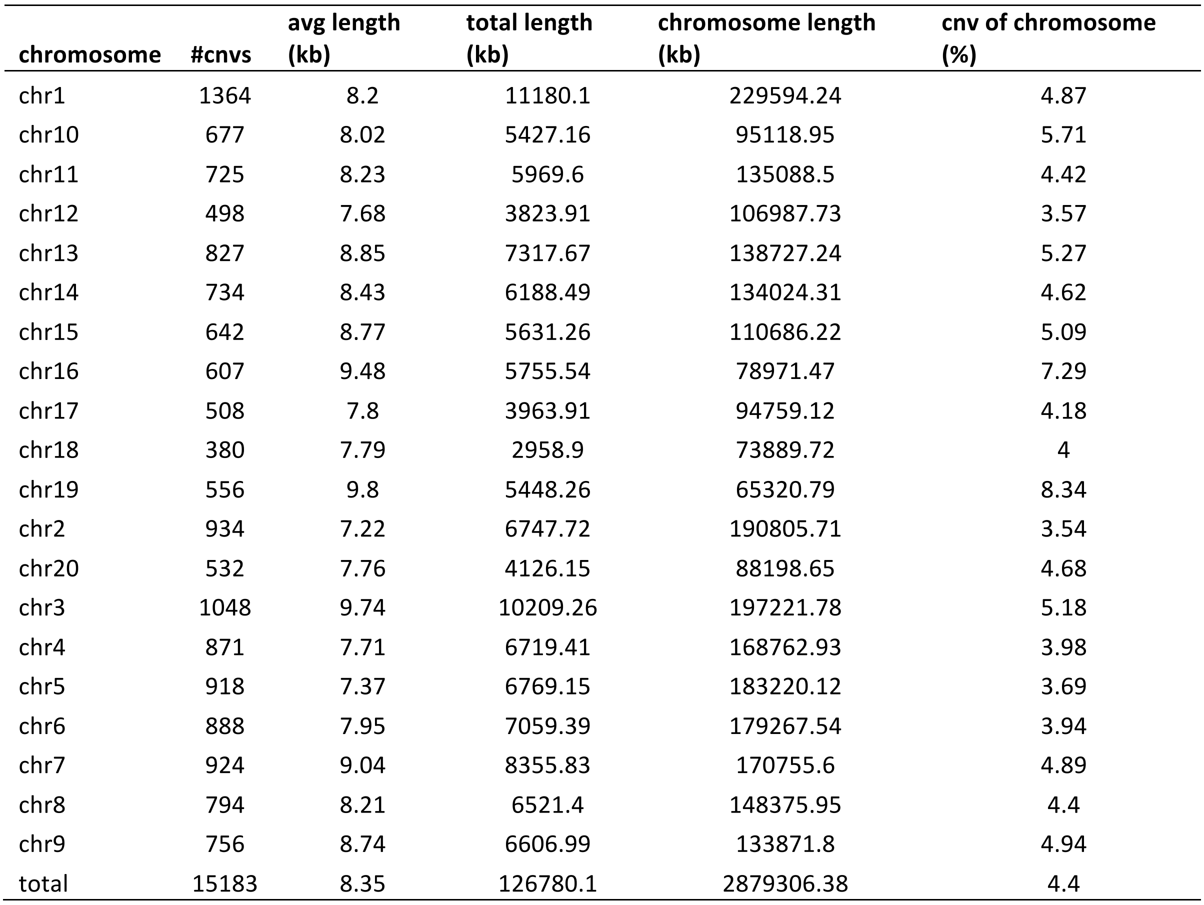
Table A.2: **Wave artifact normalizations SNR.** Median average signal-to- noise ratio (SNR) per sample for aCGH probes within CNVs called from probe GC-content normalized aCGH data across all tested LOESS fraction values.

Normalization	Median SNR
Probe GC-content only	1.4921
5000 probes	1.5084
4500 probes	1.5082
4000 probes	1.5089
3500 probes	1.5087
3000 probes	1.5083
2500 probes	1.5082
2000 probes	1.5083
1500 probes	1.5078
1000 probes	1.5061
500 probes	1.5018
100 probes	1.4577

Table A.3: **Number of detected CNVs.** Number of CNVs detected per individual by the three CNV calling methods prior to merging of individual CNV profiles calls into CNV regions across indivudals (n=22 samples).

Sample	Duplications	Deletions	Total CNVs
s1101_M	725	1099	1824
s1201_M	440	1378	1818
s2001_M	340	1170	1510
s25429_M	3205	687	3892
s25438_M	2385	741	3126
s25473_F	463	1700	2163
s25476_F	2316	1232	3548
s25594_M	3242	911	4153
s25595_M	3020	719	3739
s25640_F	2264	1158	3422
s25851_F	1311	1170	2481
s5101_F	291	2563	2854
s5201_F	2612	1185	3797
sC22579_F	1128	1402	2530
sC26727_F	1826	1015	2841
sC27239_M	438	1151	1589
sC30659_M	309	6289	6598
sC30687_M	237	1127	1364
sC30711_F	404	1038	1442
sI01778_F	4262	1321	5583
sI01785_F	3649	1327	4976
sI01786_F	2361	948	3309

Table A.4: **CNV regions per chromosome.** Number of CNV regions per chromosome, average and total CNV region length, and the percentage of the chromosome covered by these CNV regions.



chromosome	#cnvs	avg length (kb)	total length (kb)	chromosome length (kb)	cnv of chromosome (%)
chr1	1364	8.2	11180.1	229594.24	4.87
chr10	677	8.02	5427.16	95118.95	5.71
chr11	725	8.23	5969.6	135088.5	4.42
chr12	498	7.68	3823.91	106987.73	3.57
chr13	827	8.85	7317.67	138727.24	5.27
chr14	734	8.43	6188.49	134024.31	4.62
chr15	642	8.77	5631.26	110686.22	5.09
chr16	607	9.48	5755.54	78971.47	7.29
chr17	508	7.8	3963.91	94759.12	4.18
chr18	380	7.79	2958.9	73889.72	4
chr19	556	9.8	5448.26	65320.79	8.34
chr2	934	7.22	6747.72	190805.71	3.54
chr20	532	7.76	4126.15	88198.65	4.68
chr3	1048	9.74	10209.26	197221.78	5.18
chr4	871	7.71	6719.41	168762.93	3.98
chr5	918	7.37	6769.15	183220.12	3.69
chr6	888	7.95	7059.39	179267.54	3.94
chr7	924	9.04	8355.83	170755.6	4.89
chr8	794	8.21	6521.4	148375.95	4.4
chr9	756	8.74	6606.99	133871.8	4.94
total	15183	8.35	126780.1	2879306.38	4.4

Table A.5: **eQTL mapping results.** All detected *cis*-eQTL associations across all tissues are listed. Nominal pvalue denotes the uncorrected p-value obtained by the linear model, while beta pvalue is the adjusted p-value based on the permutation approach by fastQTL. Qvalue is the FDR corrected beta pvalue.

		variants		dist to	nominal		beta	
tissue	**gene_id**	**tested**	**cnv id**	**tss**	**pvalue**	**slope**	**pvalue**	**qvalue**
heart	LONP1_004793	9	chr19_4887102	-643606	3.57E-06	1.07803	3.29E-05	0.084325692
heart	TMPRSS11E_014058	3	chr5_61060708	-219549	4.76E-06	-1.00537	1.32E-05	0.084325692
heart	HOPX_139212	6	chr5_73528649	330121	8.09E-06	-1.09411	4.00E-05	0.084325692
heart	MAGEL2_019066	7	chr7_3747972	853506	4.45E-06	-2.34604	3.39E-05	0.084325692
heart	ODF1_024410	6	chr8_105576105	56991	8.70E-06	-3.64781	4.01E-05	0.084325692
heart	UGT1A6_205862	5	chr12_98452925	409599	2.00E-06	-0.844312	8.71E-06	0.084325692
heart	EHF_001206615	5	chr14_37953807	254455	1.05E-05	2.29301	2.01E-05	0.084325692
kidney	SGCB_000232	3	chr5_77742231	-149540	1.50E-05	0.691414	3.65E-05	0.071298371
kidney	OR4K17_001004715	10	chr7_82729888	-159990	3.56E-07	4.04066	3.25E-06	0.037883836
kidney	OR4K13_001004714	10	chr7_82729888	-249535	2.33E-06	2.80169	1.98E-05	0.05148385
kidney	CIDEB_014430	9	chr7_87700639	0	4.63E-07	2.78957	4.85E-06	0.037883836
kidney	LOXL2_002318	6	chr8_22711203	-808830	2.34E-06	4.37133	1.30E-05	0.046057843
kidney	C1orf190_001013615	7	chr1_49881978	595869	9.51E-07	1.04337	1.44E-05	0.046057843
kidney	FAIM3_001142473	7	chr1_164824826	415685	3.71E-06	5.8479	2.53E-05	0.056392257
kidney	GLIPR1_006851	3	chr11_73293567	462506	5.65E-06	-0.955227	1.47E-05	0.046057843
liver	GULP1_016315	2	chr12_52614781	335252	5.38E-06	-0.805831	4.25E-06	0.073700789
lung	ABCB4_018850	2	chr3_129690245	-487889	7.18E-06	-1.11717	2.69E-06	0.018352659
lung	CNTNAP2_014141	4	chr3_183849127	-337085	2.96E-05	-4.70647	3.50E-05	0.069867639
lung	KIF25_005355	4	chr4_165210711	-829498	8.65E-06	-5.01417	3.58E-05	0.069867639
lung	KLKB1_000892	6	chr5_179457643	-1783	3.00E-06	-2.82696	1.63E-05	0.050384526
lung	OR4K17_001004715	10	chr7_82729888	-159990	2.33E-09	2.29118	4.42E-08	0.000604078
lung	RABGAP1L_001243763	5	chr1_205287055	-348596	2.41E-06	0.248153	9.82E-06	0.044740414
lung	ASPHD2_020437	2	chr10_70197859	-261915	1.05E-05	-3.11456	1.84E-05	0.050384526
spleen	BMPER_133468	1	chr3_93158281	781027	4.23E-06	-3.55373	4.39E-06	0.02528205
spleen	KLKB1_000892	6	chr5_179457643	-1783	6.54E-07	-2.41492	3.41E-06	0.02528205
spleen	PLCB2_004573	8	chr7_18400455	-400166	4.86E-06	1.78734	3.18E-05	0.090040778
spleen	GSR_001195104	3	chr8_31487899	490152	1.20E-05	1.25321	3.44E-05	0.090040778
spleen	C11orf85_001037225	18	chr14_8575524	-956573	1.03E-06	-4.26057	2.25E-05	0.090040778
spleen	OR5M9_001004743	7	chr14_16700413	-6666	2.61E-08	-2.76906	3.98E-08	0.000688069
spleen	SPATA19_174927	9	chr14_132762123	-122283	4.26E-06	-0.708787	3.65E-05	0.090040778

Appendix B

Coordinated allelic variation across molecular phenotypes

B.1 Laboratory methods

B.1.1 Study sample

The lymphoblastoid cells lines (LCLs) used for the present study are a subset of those analyzed in the pilot phase of the 1000 Genomes project [48]. They encompass cells from two trios composed of father, mother and daughter, as well as eight unrelated individuals. The first trio and the unrelated samples are Utah residents from European ancestry (referred to as CEU), while the members of the second trio are Yoruban from Ibadan, Nigeria (YRI). The complete list of samples and related information is provided in Figure B.23.

B.1.2 Cell culture and fixation

For the majority of samples, all three sequencing assays (ChIP, RNA, small-RNA) were produced from a single growth of LCLs. GRO-seq was produced from a different batch of cells. Frozen cells were thawed and transferred to T25 flasks containing 15 ml of RPMI 1640 medium (Lonza, Vervier, Belgium) with 10% fetal calf serum (FCS). Cells were transferred to TubeSpin Bioreactor 50 tubes (TPP, Trasadingen, Switzerland) at a density of 0.3 x 106 cells/ml in 5 ml of same medium containing 10% FCS and 0.1% Pluronic F-68 (Sigma-Aldrich, St. Louis, MO, USA). The cultures were agitated at a shaking speed of 180 rpm in an ISF-4-W incubator shaker (Khner Shaker, Birsfelden, Switzerland) with 5% CO_2 and 85% humidity. When the cell density reached 2-3 x 106 cells/ml, the culture was diluted to 0.3 x 106 cells/ml and transferred to a 250-ml glass bottle (Schott Glass, Mainz, Germany) with the cap open by one quarter of a turn. The culture was agitated at 110 rpm in an ISF-4-W incubator shaker with 5% CO_2 but no humidity as described [156]. Eventually, the cells were scaled-up serially in 500-ml, 1-L, and 5-L glass bottles filled to a maximum of 40% of the nominal volume. For cell fixation, 2-L cultures at a density of 0.8-0.9 x 106 cells/ml in 5-L bottles were mounted on a shaker and agitated at 70 rpm at room temperature. Formaldehyde (Sigma-Aldrich) was slowly added to a final concentration of 0.8% and agitation was continued for 7 min. The fixation was quenched by addition of 2.5 M glycine (Rectolab, Servion, Switzerland) to a final concentration of 0.125 M, and the culture was agitated as before for 5 min. The cells were collected by centrifugation at 2000 rpm for 5 min at 4°C and then washed 4 times with cold PBS. The last centrifugation step was performed in 50-ml centrifuge tubes, each containing 50 x 106 cells. The final cell pellets were flash frozen in liquid nitrogen and stored at -80°C.

B.1.3 Chromatin immunoprecipitations (ChIP)

RNA polymerase II (POLR2B) and TFIIB

ChIPs were carried out as previously described [157] with a few modifications. Chromatin extracted from 5 x 107 cross-linked cells was sonicated to an average size of 200-700 bp. Sheared chromatin was then immunoprecipitated with 7 μg per 107 cells of an anti-Rpb2 antibody (sc-67318, Santa Cruz Biotechnology) or 7.5 μl per 107 cells of an anti-TFIIB antibody (rabbit CS396), described in [158]). Immunoprecipitated material was recovered with 2 mg per 107 cells of pre-blocked protein-A beads (17-0780-01, GE Healthcare) and washed twice with dialysis buffer, three times with IP wash buffer (see [157] for buffer compositions). After reversal of crosslinking and DNA purification, 10 ng of ChIP DNA was used for ChIP-seq libraries preparation.

PU.1, MYC, and H3K4me1

Cells were lysed in nuclei extraction buffer (50 mM HEPES-NaOH pH 7.5, 140 mM NaCl, 1 mM EDTA pH 8.0, 10% glycerol, 0.5% NP-40, 0.25% TritonX-100) supplemented with a protease inhibitor tablet (Roche) and phosphatase inhibitors (5 mM NaF, 1 mM β-glycerol phosphate and 1 mM sodium orthovanadate) for 10 min at 4°C on a shaker. The isolated nuclei were then washed using washing buffer (200mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 10 mM Tris-HCl pH 8.0) supplemented with protease and phosphatase inhibitors at RT for 10 min. Washed nuclei were resuspended in sonication buffer (1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 10 mM Tris-HCl pH 8.0 and 1% TritonX-100) containing protease and phosphatase inhibitors and the chromatin was fragmented using a Bioruptor sonicator (Diagenode) for 80 min using high amplitude and 30s ON & 30s OFF cycles to obtain 200-500 bp-sized fragments.

The fragmented chromatin was then centrifuged at 17,000xg for 5 min and clear supernatant was diluted with ChIP dilution buffer (1 mM EDTA pH 8.0, 10 mM Tris-HCl pH 8.0 and 1% TritonX-100 containing protease and phosphatase inhibitors) to get chromatin equivalent to 10 X 106 cells for each IP. All IPs were performed in duplicates. BSA and ssDNA (Salmon Sperm DNA)-preblocked protein-A sepharose (80 μl/IP) beads were added to the samples and incubated for 2h to remove non-specifically binding chromatin. To the supernatant, 5 μg/IP rabbit polyclonal anti-Myc antibody (Santa Cruz, Cat no: Sc- 764) was added to immunoprecipitate the chromatin complex at 4°C overnight. After incubation, 50 μl blocked protein-A sepharose beads were added to each sample and incubated for 90 min at 4°C to pull down the respective antibody-chromatin complexes. The beads were then washed four times with low salt wash buffer (20 mM Tris-Cl pH 8.0, 150 mM NaCl, 2 mM EDTA pH 8.0, 0.1% SDS, 1% TritonX-100) followed by two washes with high salt wash buffer (20 mM Tris-Cl pH 8.0, 500 mM NaCl, 2 mM EDTA pH 8.0, 0.1% SDS, 1% TritonX-100), lithium chloride wash buffer (10 mM Tris-Cl pH 8.0, 0.25 M LiCl, 1 mM EDTA pH 8.0, 1% NP-40, 1% sodium deoxycholate) and Tris- EDTA (TE) buffer (10 mM Tris-Cl pH 8.0, 1 mM EDTA pH 8.0). The c-Myc-bound chromatin complexes were eluted from beads for 30 min using 200 μl of elution buffer (100 mM sodium bicarbonate and 1% SDS in milliQ water). The chromatin was then reverse-crosslinked at 65°C overnight after adding 8μl of 5 M NaCl. The DNA was then purified from the reverse-crosslinked chromatin by proteinase-K and RNase digestion followed by purification using Qiagen DNA purification columns. The purified DNA was eluted in 30μl of Qiagen elution buffer. PU.1 and H3K4me1 ChIPs were performed with slight modifications in the protocol described above. We used 1% SDS instead of TritonX-100 in sonication buffer to increase the stringency of chromatin pull-down by the respective antibodies (PU.1 antibody from Santa Cruz, Cat no: 22805X and H3K4me1 from Abcam, Cat no: ab-8895) and the sonication was performed for 60 min instead of 80 min.

H3K4me3, H3K27me3, H3K27ac, and H4K20me1

ChIP was carried out largely as suggested in [159], with modifications made to automatize the procedure. Briefly, cells were lysed by addition of cell lysis buffer, then nuclei were washed and subsequently lysed using nuclei lysis buffer. Chromatin was sheared with Covaris S220 sonicator (Covaris Inc., MA, USA). Sonication efficiency was assessed by running a sample of de-crosslinked DNA on a 1.5% agarose gel. Fragmented chromatin was diluted 10 fold (5 fold in case of H3K27ac IP) in ChIP dilution buffer and immunoprecipitated using antibodies against H3K4me3 (Millipore 17-614; lot #JBC1793805), H3K27me3 (Millipore 17-622;

lot #DAM1731568), H3K27ac (Abcam ab4729; lot #GR71158) H4K20me1 (Abcam ab9051; lot #GR10999). The immunoprecipitation assays were performed on Diagenode SX-8G IP-Star Compact automated system using Auto Histone ChIP-seq kit (Diagenode s.a., Belgium). The minimum of 2 IPs of 106 cells (2x106 in case of H3K27ac) per cell line was used. Replicates were pooled following RNase A and proteinase K treatments. DNA was purified with Qiagen DNA purification kit (Qiagen N.V., Netherlands). DNA concentration was measured using Qubit apparatus (Life Technologies, CA, USA). Before proceeding with library preparation for sequencing, enrichment of the precipitated DNA was assessed by quantitative PCR. Of note automatization of the procedure to reach the necessary throughput required by this project did not significantly modify the results. Paralleled chromatin IP of 107 cells performed manually using Dynabeads magnetic beads (Life Technologies, CA, USA) to collect chromatin-antibody complexes showed concordant results. For example 88% of the nucleotides significantly enriched for H3K4me3 and H3K27me3 marks in the automated protocol were also identified as enriched by manual immunoprecipitation. Details of the comparisons between results obtained by the automated and the manual protocol are presented in Figure B.24.

B.1.4 ChIP-seq library preparation and sequencing

Trios

ChIP libraries were prepared for sequencing with the Illumina ChIP-seq sample preparation kit according to manufacturer's instructions. Sample concentration was re- measured prior to library preparation. The starting amount of ChIP DNA ranged from 6 ng to 10.5 ng per sample. The number of PCR cycles to amplify the libraries was either 18 (POLR2B) or 17 (all other assays). Library quality and average fragment size was confirmed with Bioanalyzer DNA analysis chips (25-1000 bp, Agilent). Sequencing was performed with one sample per lane on the Genome Analyzer IIx or on the HiSeq2000 (read length 36 bp, single-end), except H3K27ac, which was indexed (see below) and sequenced as pools of three per HiSeq lane.

Unrelated individuals

ChIP libraries were prepared with the TruSeq DNA sample prep kit (Illumina) and AD001-AD0012 indexing adapters set according to the manufacturer's recommendations. The starting amount of ChIP DNA used for library preparation ranged from 2.5 ng to 10.5 ng per sample. Library quality and average fragment size was confirmed with Bioanalyzer DNA analysis chips (25-1000 bp, Agilent). TruSeq libraries were subsequently multiplexed on Illumina HiSeq2000 lanes (three or four per lane for POLR2B and H3K27me3 assays and all other assays, respectively) (read length 36 bp, single-end). H3K4me1 libraries were sequenced twice in order to improve the coverage.

B.1.5 RNA extraction

Total RNA was extracted from cell pellets using the standard Trizol protocol (Invitrogen). RNA concentration was measured with the Qubit system (Invitrogen) and the quality of the samples confirmed with Agilent 2100 Bioanalyzer RNA 6000 Nano and small RNA (6-150 nt) analysis chips. All included RNA samples had a RNA integrity number (RIN) of 9.8 or more.

B.1.6 RNA-seq library preparation and sequencing

Libraries for RNA-seq were prepared with the Illumina TruSeq RNA sample preparation kit, according to manufacturer's instructions. 500 ng of total RNA was used for each library. Briefly, poly-A RNA is selected using poly-T oligo-attached magnetic beads, the RNA is cleaved, and converted to cDNA with first strand synthesis. After RNA digestion and second DNA strand synthesis, the fragments are end repaired and ligated to the adapters containing specific primer indexes. Finally, the cDNA libraries are amplified by PCR. Trio samples were sequenced as a single pool of six, whereas the eight additional unrelated samples were sequenced as part of pools with 12 libraries on the HiSeq (read length 49 bp, paired-end).

B.1.7 GRO-seq

We assayed the nascent transcriptome of LCLs from the three CEU trio individuals with Global Run-On Sequencing (GRO-seq) as previously described [60]. Nuclei were isolated and nucleotides washed off at 4°C, leaving RNA polymerases engaged in transcription bound to DNA. 5M nuclei per sample were then used for letting the polymerases run-on for $\sim$100 nts using Br-UTP, αP32-labeled CTP for tracking the nascent RNA through the experiment and sarkosyl for blocking new transcription initiation events. RNA was isolated and hydrolyzed. Subsequently, the nascent RNA was pulled down with agarose beads carrying antibodies for Br-UTP. 5′ and 3′ adapters were then ligated to nascent RNA, fulfilling another round of immuno-enrichment after each step. Finally, the nascent RNA library was converted to cDNA, PCR amplified and purified, yielding 2 libraries per sample. Given the way they were purified from the agarose gel, one library is of longer insert size (hereon termed ”long”) than the other (hereon termed ”short”). Libraries were sequenced with Illumina HiSeq2000 (read length 49 bp, paired-end) once (short libraries) or twice (long libraries), the three samples pooled in one lane.

B.2 Data Preprocessing and Quality Control

B.2.1 Genetic variation data

DNA variation data for the study sample was obtained from the 1000 Genomes project [48]. Genomic coordinates for the trios (data link 1, see below) were lifted over from b36 to hg19 genomic build using tools available in the GATK package [160, 161]. For the eight unrelated individuals, we used variants from the 1000 Genomes release 20100804 (data link 2), as some individuals used in the present study were not available in the phase 1 release. The number of variants available for each individual included in this study is provided in Figure B.23. Only SNP variants were used. We additionally used population- based variation data from the 1000 Genomes phase1 samples (data link 3) for quality control purposes in the allele-specific analyses (see Section B.3.2).

Variant data downloads:

1. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_07/trio/snps

2. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20100804

3. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521

B.2.2 Pre-processing and mapping of raw sequence data

RNA-seq and in house performed ChIP-seq

RNA-seq and ChIP-seq sequence reads (paired-end 49 bp and single-end 36 bp, respectively) were mapped against the standard hg19 build of the human reference genome [162] with BWA [163] using default parameters. We did not map to personalized reference genomes because reliable indel calls were not available for the reference construction for the given individuals at the time of study. We kept only uniquely mapping reads with a mapping quality (MAPQ) score of >= 10. For paired-end data we additionally required the reads to be properly paired in mapping. Samtools [108] was used for general data processing throughout the project. Summary of mapping statistics for each assay are provided in Tables B.1, B.2 and Figure B.25.

CTCF ChIP-seq and DNaseI hypersensitivity

Data for CTCF and DNaseI hypersensitivity for the two trios were obtained from a previous study [51]. For CTCF we obtained raw sequenced reads, which were mapped and processed the same way as the other ChIP-seq assays. Biological replicates were merged after mapping.

DNaseI data were obtained as original peak calls that were merged across biological replicates and samples into a metasample (see Section B.2.4).

B.2.3 Quantification of transcriptome data

RNA-seq

RNA-seq data was quantified based on Gencode v8 (03/2011) exon annotations [164]. In order to quantify exons in a non-redundant way, we created a set of merged exons from all protein coding and linc-RNA transcripts and merging any overlapping exons into new composite exons. We then counted the number of reads mapping to each exon, with each individual read from a pair contributing to the count.

GRO-seq

GRO-seq reads were trimmed to 39 bp due to abundant presence of adapter sequences and mapped to the genome with BWA [163]. Only read 1 was used for further analyses given that read 2 would be overlapping read 1 in many cases. Uniquely mapped reads with MAPQ >= 10 were then merged from the different sequenced libraries originating from the same sample, yielding a mean of 36.6 million reads per sample. Gene-based correlation of reads between any two samples was 0.96 (Spearman rank correlation). Final GRO-seq reads were overlapped with a selection of functional genomic features as defined in the Gencode v8 annotation [164] (Figure B.2). Namely, the following elements were queried: (a) genes (protein coding and linc-RNA genes only), (b) exons (merged exons from the genes used, correct strand required for overlap), (c) introns (anything not defined as exons within genes, correct strand required for overlap), and (d) putative enhancer elements (as defined in Section B.3.1). We additionally quantified antisense (within genes but on the opposite strand) and divergent (1kb upstream of TSS and opposite strand from the gene, excluding gene regions, correct strand required for overlap) transcription. Reads not falling into any of the mentioned categories were grouped as ”other”.

B.2.4 Quantification of ChIP-seq data

B.2.5 ChIP-seq peak calling

To call peaks from ChIP-seq data we merged final mapped reads (MAPQ>=10) from the six trio individuals into a metasample, excluding reads with identical start positions (i.e. duplicate reads). Two different peak calling algorithms were used depending on the specific properties of each assayed chromatin mark. Transcription factor-like peaks (MYC, PU.1, TFIIB, CTCF and POLR2B) as well as concise histone modification peaks (H3K4me1, H3K4me3, and H3K27ac) were called using HOMER [101] with the following parameters (MYC, PU.1, TFIIB, CTCF, POLR2B: -factor; H3K4me1, H3K4me3, H3K27ac: -region -size 1000 -minDist 2500). All TF-like peak calls were subsequently extended to the expected fragment length of 200 bp. Broad histone modifications domains (H3K27me3 and H4K20me1) were called using HMM-based RSEG [165]. Default settings were used with a maximum of 20 iterations for the training of the Hidden Markov Model. Deadzone correction was applied using the deadzone file for 36 bp reads and hg19 provided on the distributers webpage (`http://smithlab.usc.edu/histone/rseg/`). POLR2B ChIP-seq data were analyzed with both algorithms to capture the full scope of the RNA pol II-binding properties, i.e. promoter-associated narrow peaks versus broad domains covering the gene bodies (subsequently referred to as ″narrow″ and ″broad″, respectively). We included only chromosomal peaks, i.e. mapping to chromosomes 1-22, X or Y) and filtered away all peaks overlapping with know collapsed repeat regions [166] or genomic regions ″blacklisted″ by the ENCODE project (see Section B.3.2) [49], as both types of regions can cause bias in read mapping and subsequent peak calling. Peak calls are summarized in Figure B.26 and B.27 and pairwise overlap of peak calls among all assays are shown in Figure B.28.

B.3 Analytical Methods

B.3.1 Distribution of assays around the TSS of different classes of genes

TSS selection and transcriptional activity

We first sought to define a set of quantifiable transcription start sites which we could reliably associate with specific exons and, therefore, accurately determine levels of the associated transcript. TSSs were defined as the 5′ start of the first exon of each transcript annotated in Gencode version 8 [164]. TSSs were merged if belonging to the same gene and less than 100 bp apart from each other. For genes with only one TSS, or several but only one active TSS defined by overlap with RNA polymerase II peaks, this TSS was selected. In both cases the transcriptional activity for these TSSs was defined as RNA-seq reads mapping to all exons assigned to transcripts starting at this TSS. In cases of genes with several active TSSs, only the most 5′ TSS was selected to avoid confounding effects by other active transcripts. The transcriptional activity was defined by reads mapping to all exons that could be uniquely assigned to transcripts starting at this TSS. For loci with no active TSS, the one linked to the longest open reading frame was selected, and its expression was defined by RNA-seq reads mapping to exons uniquely assigned to that

TSS. All RNA-seq quantifications were normalized to reads per 100,000 bp.

Definition of putative enhancer elements

DNaseI hypersensitivity, a proxy for open chromatin structure, has been shown to define regulatory regions of the genome, including enhancer elements. Thus, we constructed two sets of putative regulatory elements (i.e. enhancers) by filtering the DNaseI hypersensitivity metapeaks [51] either for i) all annotated transcripts extending 2.5 kb upstream (intergenic enhancers only; type I), or ii) all exons, with 5′ exons of each transcripts extending 2.5 kb upstream (allowing putative intronic enhancers; type II). Type II enhancers were used throughout the paper, if not mentioned otherwise.

Marker quantification

ChIP-seq, RNA-seq and GRO-seq reads were counted within a 5 kb window centered on every quantifiable TSS (Gencode v8 annotation; nTSS=13,720), separately for protein- coding (nTSS=13,034) and linc-RNA (nTSS=686) genes. Putative enhancer loci were quantified in a similar manner within the defined enhancer site. Loci with no mapped reads were excluded from the following analyses.

Correlation heatmaps

For the correlation among markers at the TSS, read counts were normalized across assays by calculating z-scores of $\log_{10}$ transformed read counts within the 5 kb window centered on every TSS. Spearman's correlation coefficients were calculated for each marker combination. Putative enhancer loci were analyzed the same way. Clustering and heatmap was created using heatmap.2 for R (version 2.13) with hierarchical clustering based on Euclidean distance.

Gene expression versus marker binding

The obtained ChIP-seq marker quantifications for the TSSs were normalized to 10,000,000 total mapped reads in each experiment and the loci were grouped into percentiles according to their transcriptional activity. For each percentile the average RNA-seq quantification value and the number of ChIP-seq reads were calculated for each marker. Obtained values for every percentile were plotted on $\log_{10}$ scale.

B.3.2 Allele-specific analysis

General description

Allele-specific (AS) analysis was based on binomial testing of allelic ratios over heterozygous SNP sites of each individual. The analysis was limited to SNPs due to the lack of high quality indel calls for the individuals studied. We required both alleles to be observed in the data and included only SNP sites located within trio metasample peaks. Extensive filtering steps were taken to eliminate sources of bias in the analysis. We excluded sites overlapping with i) collapsed repeat regions [166], i.e. sequences that are present in a single copy in the reference genome, but which are present in multiple copies in reality (n=30,671), ii) ENCODE-defined blacklisted genomic regions (tracks 1-2, see below; merged total regions n=1,378; ChIP-seq data only), and iii) regions of general non-unique alignability given the read length of each assay (tracks 3-5, see below). Taking advantage of phased genotype data, we finally applied two additional simulation- based filtering steps to further exclude individual SNP sites susceptible to (i) mapping bias due to local haplotype effects (Section B.3.2) and (ii) low complexity library artifacts, which can lead to false positive allele-specific calls (Section B.3.2). A minimum coverage of 10-20 reads per site was required, depending on the sequencing depth of the assay in question. In all analyses, we used a minimum mapping and base quality threshold of 10. GRO-seq data was analyzed in a strand-specific manner, splitting the original reads per strand. An overview of the AS analysis pipeline is shown in Figure B.29.

ENCODE mapability track downloads:

1. `http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability`

Tracks used:

1. wgEncodeDacMapabilityConsensusExcludable.bed
2. wgEncodeDukeMapabilityRegionsExcludable.bed
3. wgEncodeCrgMapabilityAlign36mer.wig
4. wgEncodeCrgMapabilityAlign40mer.wig
5. wgEncodeCrgMapabilityAlign50mer.wig

Correction for reference allele mapping bias

To correct for bias caused by the preferential mapping of reads carrying the reference allele we calculated the estimated bias across all heterozygous sites with i) MAPQ >= 10, ii) sequencing base quality >= 10, iii) both alleles seen in sequence data, and iv) minimum coverage of eight reads per site. The estimates were calculated separately for each mapping quality bin and SNP allele combination (minimum 200 sites required for each category; if less, global estimate was used for that category) after down-sampling reads of sites in the top 25th coverage percentile in order to avoid the highest covered sites having a disproportionally large effect on the ratios. These matched estimates were then used as the expected ratios in the binomial test (instead of using 50-50) for each tested SNP.

Simulations to identify SNP sites susceptible to mapping artifacts

Two complementary strategies were adopted to identify and filter away SNP sites showing evidence of biased mapping of reads that might cause false allele-specific signals.

Personalized simulations We constructed all possible reads spanning both haplotypes in a sliding window of +/- respective read length (36 bp, 39 bp, 49 bp) around each phased heterozygous SNP site, separately for all six trio individuals. If other SNPs overlapped with this window, they were included in the simulated reads. We mapped the entire set of simulated reads back to the hg19 reference genome with the same parameters as the actual data and collected mapping statistics for each SNP site to identify those where a significant proportion (>5%) of the possible reads map incorrectly or not at all (Figure B.30). The obtained list of SNPs susceptible to mapping bias was used to filter the results of the AS analysis of the trio individuals.

Population-based simulations The same simulations were performed using all phased SNP and indel variants from 1000 Genomes phase1 (release 20110521, see Section B.2.3) data with a minor allele frequency >0.01 in either the European (EUR) or African (AFR) population. Instead of constructing all reads for the two haplotypes present in a single individual, all haplotypes present in the population were constructed. The obtained list of biased SNP sites was used to filter the results of the AS analysis of the eight unrelated CEU individuals. We additionally excluded all putative AS sites from the trio and unrelated individuals that were within a read length of an indel showing biased mapping in the population-based simulations (Figure B.30). This allowed us to account for indel effects indirectly, despite the lack of indel calls of the individuals in question.

Filtering of low complexity sites for allele-specific analysis

We devised a pipeline for heterozygous SNPs within each individual in order to identify and remove sites that show an enrichment of clonal reads around the SNP position in ChIP-seq and GRO-seq experiments. Clonal reads can lead to confounding effects during allele-specific analysis. Briefly, we employed two filtering steps in which we discarded biased SNPs. First, we removed SNPs independent of coverage that were covered by reads with less than five unique alignment start sites, and second SNPs, which showed an enrichment in clonal reads based on library-specific simulations (P <0.05). A stepwise histogram of AS ratios after each filtering step described in Section B.3.2 is presented in Figure B.31. See `http://updepla1srv1.epfl.ch/waszaks/absfilter` for more information.

eQTL overlap

All heterozygous SNP sites accessible for AS analysis were analyzed for overlap with known eQTL loci from the 1000 Genomes phase1 populations [56]. We compared the overlap with eQTL SNPs and non-eQTL SNPs matched for minor allele frequency and distance from the nearest TSS separately for TF assays (PU.1, TFIIB, MYC, CTCF) and hPTMs (H3K4me1, H3K4me3, H3K27ac, H4K20me1, and H3K27me3). eQTLs from the EUR and AFR populations were analyzed separately. Mann-Whitney U test between the allele ratios for eQTLs and null SNPs (EUR: n=6243 and n=7256; YRI: n=2045 and n=7248, respectively) was used to evaluate whether a significant bias in the allele ratios could be observed between the two groups of target sites (eQTL/null) and assays (TFs/hPTMs) (Figure B.9).

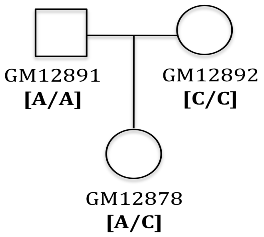
B.3.3 Parental transmission of allelic effects

Transmission per SNP site

Parental transmission of allelic effects was analyzed as described by McDaniell et al. [51]. Transmission was first analyzed at autosomal SNP sites where the child has a significant AS effect and the parents were homozygous for opposite alleles of this SNP. Standard AS QC criteria were applied to the child data (see Section B.3.2). For the parental reads, a MAPQ >= 10 was required but no minimum read coverage at the SNP site was applied. We calculated the ratio of reads covering each allele (maternal and paternal) in the parents and compared the paternal allele ratio of the parents to the paternal allele ratio in the child with Spearman rank correlation. A global scaling based on library size differences (i.e. number of usable reads) was applied to the read counts. The parental libraries were scaled to the child's library separately for each assay. This analysis (referred to as standard transmission) was then extended to SNP sites where the child has a significant AS signal but one parent is homozygous and the other heterozygous (referred to as extension 1) (Figure B.17). We included only parental heterozygous sites which were accessible for the AS analysis (i.e. fulfilled quality and coverage requirements for AS analysis) in order to have reliable read counts for each of the two alleles also in the parent(s). Analysis was performed genome-wide as well as at specific functional elements of the genome, namely promoters (Figure B.19) and putative enhancer elements (as defined in Section B.3.1) (Figure B.18), as well as known eQTLs [56] (see Section B.3.2) (Figure B.20) and dsQTLs [50] (n=8896) (Figure 2.4C, Figure B.20) (+/- 1000bp window around each QTL).

The allele ratios in each scenario were calculated as follows:

Standard transmission:



CHILD = A / (A + C)
PARENTS = AA / (AA + CC)

Extension 1: One parent homozygous, one heterozygous:

CHILD	= REF / (REF + NONREF)	[father = HOM REF]
	= NONREF / (REF + NONREF)	[father = HOM NONREF]
	= NONREF / (REF + NONREF)	[mother = HOM REF]
	= REF / (REF + NONREF)	[mother = HOM NONREF]
PARENTS	= 0.5 * HOM_REF / (0.5 * HOM_REF + HET_NONREF)	[father = HOM REF]
	= 0.5 * HOM_NONREF / (0.5 * HOM_NONREF + HET_ REF)	[father = HOM NONREF]
	= HET_NONREF / (0.5 * HOM_REF + HET_NONREF)	[father = HET; mother = HOM REF]
	= HET_REF / (0.5 * HOM_NONREF + HET_REF)	[father = HET; mother = HOM NONREF]

Transmission per haplotype

For chromatin marks we additionally tested parental transmission of allelic effects at sites where the child has a significant AS effect and the parents are homozygous for opposite alleles of the entire haplotype surrounding the target SNP in order to better capture long- range effects. No requirement for homozygosity was applied to the actual target SNP in the parents. To construct the parental haplotypes we used common SNPs with a minor allele frequency of 5% or greater in the 1000 Genomes phase1 EUR or AFR populations (for CEU and YRI trios, respectively). Windows of 5, 10, and 20 kb around the child SNP site were tested. Comparison of allele ratios was done as in the per site transmission test (as described in Section B.3.3). No significant improvement of the transmission signal compared to the standard per site analysis was discovered.

B.3.4 Transcription factor binding motif analysis

De novo motif identification

We performed *de novo* motif search on sequences around PU.1 and MYC peak maxima (+/- 100bp) using the software package MEME. We restricted the identification of de novo motifs to 1000 peaks with highest tag counts as identified by HOMER. We run MEME with the following settings: zero or one motif occurrence per peak (-zoops), maximum 10 de novo motifs (-nmotifs 10), minimum and maximum motif size 5 and 20 bp (-minw 5, -maxw 20), respectively, and we used the setting to perform the de novo search on the given and reverse complement strand (-revcomp). The highest scoring de novo motif PWMs were compared against the known PU.1 and MYC PWMs deposited in TRANSFAC or JASPAR using the online version of the motif comparison software TOMTOM (`http://meme.sdsc.edu/`). To check for motif overlap with common indels, we used a set of indel calls from the 1000 Genomes Project phase1 release [167] with a minor allele frequency $>$0.01 in either the European or African population. For this comparison, we used only peaks with a significant ASB signal not disrupting the motif (class II B- SNPs) and applied a window of +/- 50 bp around the motif. Of such PU.1 peaks, 1.5% overlap with common indels. Motifs in these peaks might be affected by indels, although the direct impact of rare indels on TF binding remains to be assessed.

Haplotype-specific motif analysis

We scanned within each individual the paternal and maternal haplotype for the occurrence of a PU.1 and MYC motif instance among all PU.1 and MYC peaks, which were tested for allele-specific effects, respectively. We used the ENCODE data-derived PU.1 and MYC PWMs for the motif scan [168]. Further, we used the phase information from the 1000 Genomes Project in case multiple SNPs were present within peaks, and discarded peaks if phase information was absent for one or multiple SNPs. Motif occurence was predicted using the software FIMO (part of the MEME package) and the default p-value threshold (P = 1e-4). In case a motif was predicted on only one haplotype, we performed another round of motif search with a soft p-value threshold of 0.1 in order to obtain the motif occurrence p-value for the alternative haplotype. In rare cases we observed that motif predictions overlapped and we decided to discard these sites to avoid ambiguity in later analysis.

Allele-binding cooperativity analysis

We tested for allele-binding cooperativity using the CEU and YRI trio PU.1 ChIP-seq data and the CEU trio MYC ChIP-seq data, respectively. We obtained 79 unique motifs, derived from 457 ChIP-seq data sets and 119 human TFs [168], and performed for each motif a haplotype-specific motif search within all ASB-significant PU.1 and MYC peaks. We restricted the motif search to 100 bp around the peak maxima. For each motif instance we calculated the difference between the paternal and maternal motif occurrence $-\log_{10}$ p-value. We considered only motif instances that caused a difference in motif score between the paternal and maternal haplotype (= polymorphic motifs). We combined data from all individuals and focused on 35 out of 79 motifs with at least five polymorphic motif instances. For each motif, we calculated Pearson correlations between differential motif scores and paternal TF binding ratios independently and applied Benjamini & Hochberg p-value adjustment to correct for multiple hypothesis testing (p.adjust function implemented in R). All TF motif analyses (Section B.3.4) were performed only in the trio samples.

B.3.5 Analyses of allelic consistency

Between unrelated individuals

Genome-wide consistency of the allelic ratios was analyzed between all pairs of unrelated CEU individuals (n=10) for heterozygous SNP sites with a significant AS effect (P <= 0.01) in i) both individuals (Figure 2.4A, Figure B.16), and ii) at least one of the two individuals (union of significant sites) (Figure B.14). Correlation of the reference allele ratios at these loci was calculated with Spearman correlation separately for each assay available for all 10 individuals. We also looked at the consistency separately in the trio parents (CEU and YRI) in order to include assays with data only from the trios. Both within and across trios comparisons were considered (Figure B.15). Examples are given in Figure B.16.

Across peaks within individuals and assays

In the evaluation of the consistency of AS effects within ChIP-seq peaks we took into account only peaks with at least two overlapping SNPs with a significant AS effect (P <= 0.01). A peak was considered consistent if all of the significant SNPs within that peak had a paternal ratio greater or smaller than 0.5. To summarize the peak consistency for each trio we considered the sum of the consistent peaks in each sample of the trio and divided by the sum of all the evaluated peaks for that trio (Figure 2.2B, Figure B.7). For assays with a variable peak length, we also tested how the consistency of the AS effects within the peak is affected by distance, i.e. peak length (Figure B.7).

Between different assays (AC and HC)

Allelic coordination (AC) and haplotypic coordination (HC) between all different pairs of assays were calculated using heterozygous SNP sites with a significant allele-specific signal in two assays. We define AC as a coordinated AS signal at the exact same SNP in two different assays, whereas in HC we compare the AS signal at two different SNP sites in two different assays within a given genomic window (Figure B.21). The following windows were analyzed for AC and HC: i) TSSs, as defined in Section B.3.1, ii) putative enhancer loci, allowing for intronic loci, as defined in Section B.3.1, and iii) the general vicinity of gene regions (protein coding and linc-RNA gene annotations from Gencode v8, +/- 50 kb upstream and downstream). For each window, we constructed all possible pairs of SNPs between all pairs of assays with available AS sites and correlated the paternal allele ratio for all comparisons using Spearman rank correlation. SNP pairs contributing to multiple windows per region (for e.g. to multiple overlapping TSSs) were included only once. Only significantly correlated comparisons with a minimum of 20 pairs genome-wide were considered ($P < 0.05$). We pooled the constructed SNP pairs across individuals, but analyzed the two trios and the eight unrelated individuals separately. To analyze if the degree of haplotypic coordination within and across assays correlates with genomic distance between the SNP pair, the correlation of genomic distance (log transformed absolute bp distance between the two SNPs) and the binary allelic consistency status (allelic ratio >0.5 or <0.5 in both SNPs = consistent, otherwise inconsistent) was analyzed using logistic regression. This analysis was performed only around gene regions (+/-50 kb).

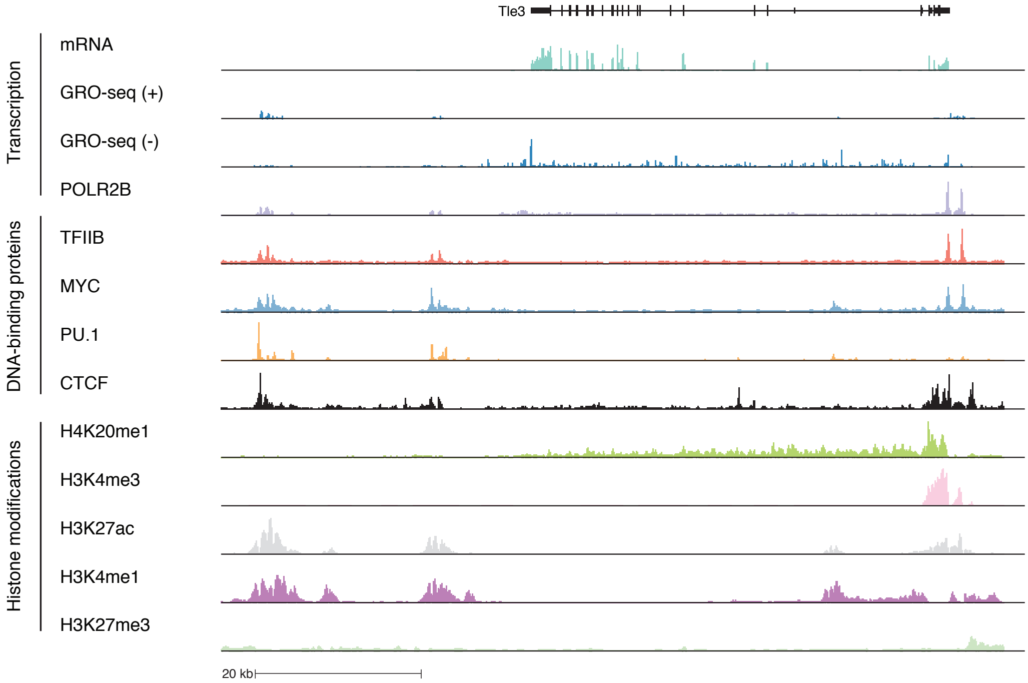
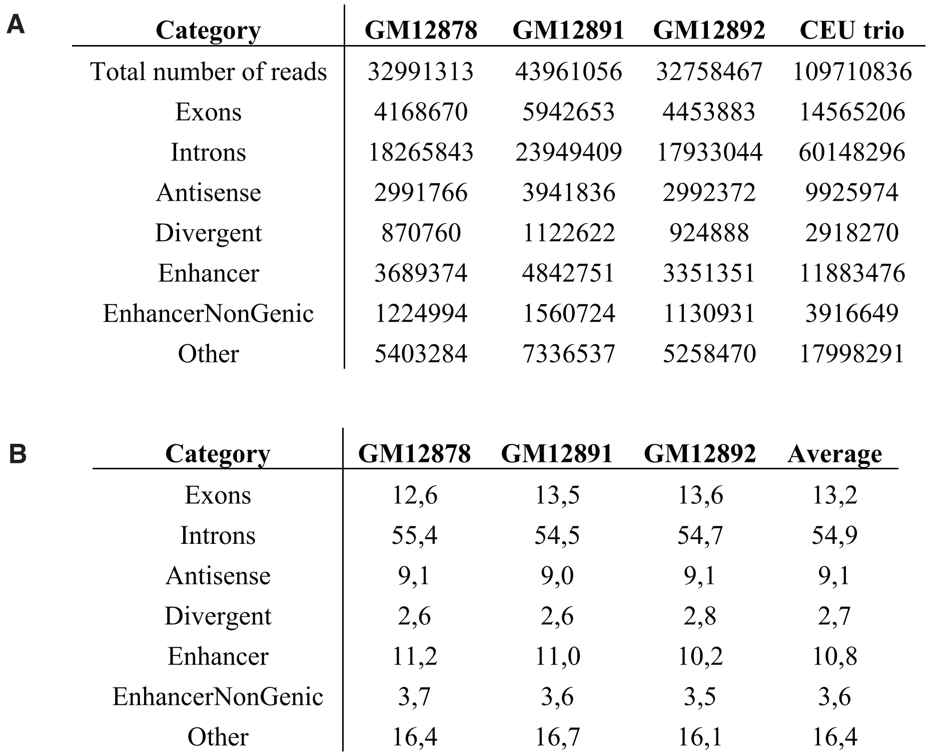


Figure B.1: Overview of the dataset generated and used in this study. Read density tracks from the trio metasamples are shown around the Tle3 gene locus. The y-axis scaling has been chosen for visualization purposes only. Track intensities are therefore not comparable among assays.

Figure B.2: Number **A** and **B** percentage of GRO-seq reads overlapping with genomic features based on Gencode v8 annotations. Categories: exons (requiring strandedness for overlap), introns (any region which is not an exon within genes, strandedness required), antisense (any region within genes in the opposite strand), divergent (1 kb upstream of TSSs and opposite strand of gene; minus any area that overlaps genes, requiring strandedness), enhancer (putative enhancers based on DNaseI hypersensitivity data, including intronic regions), enhancer-NonGenic (putative enhancers not overlapping genes), other (none of the above categories; may include pseudogenes or other transcripts not annotated as protein-coding or linc-RNA).



A

Category	GM12878	GM12891	GM12892	CEU trio
Total number of reads	32991313	43961056	32758467	109710836
Exons	4168670	5942653	4453883	14565206
Introns	18265843	23949409	17933044	60148296
Antisense	2991766	3941836	2992372	9925974
Divergent	870760	1122622	924888	2918270
Enhancer	3689374	4842751	3351351	11883476
EnhancerNonGenic	1224994	1560724	1130931	3916649
Other	5403284	7336537	5258470	17998291

B

Category	GM12878	GM12891	GM12892	Average
Exons	12,6	13,5	13,6	13,2
Introns	55,4	54,5	54,7	54,9
Antisense	9,1	9,0	9,1	9,1
Divergent	2,6	2,6	2,8	2,7
Enhancer	11,2	11,0	10,2	10,8
EnhancerNonGenic	3,7	3,6	3,5	3,6
Other	16,4	16,7	16,1	16,4

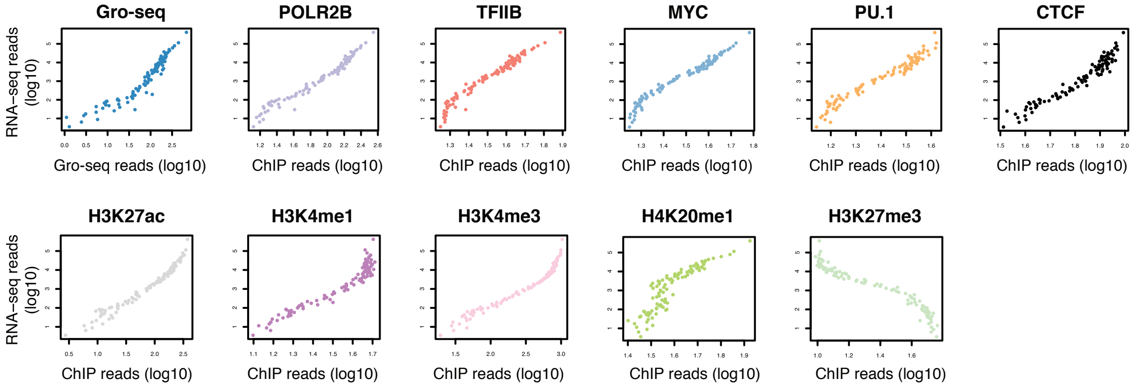


Figure B.3: Relationship between gene expression (mRNA-seq) and genomic signals at promoters (transcription start site +/- 2.5 kb) of protein-coding and linc-RNA genes. Genes were grouped into percentiles according to their expression level and the average expression level and read density is shown for each percentile.

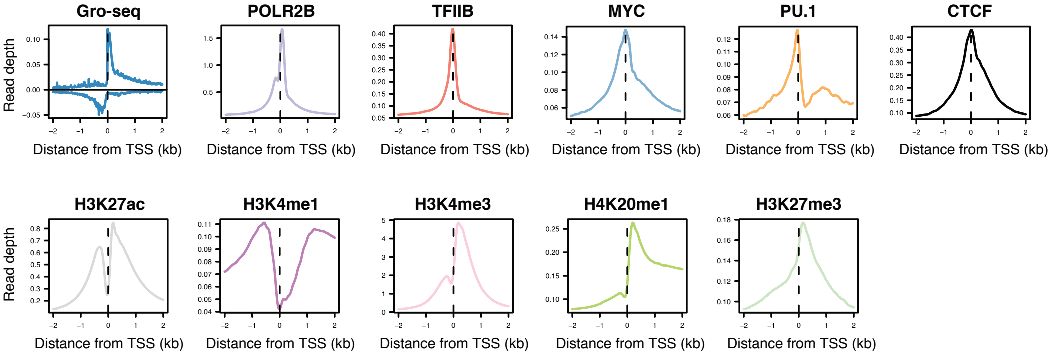


Figure B.4: ChIP fragment and GRO-seq read densities near transcription start sites of quantifiable protein-coding and linc-RNA genes.

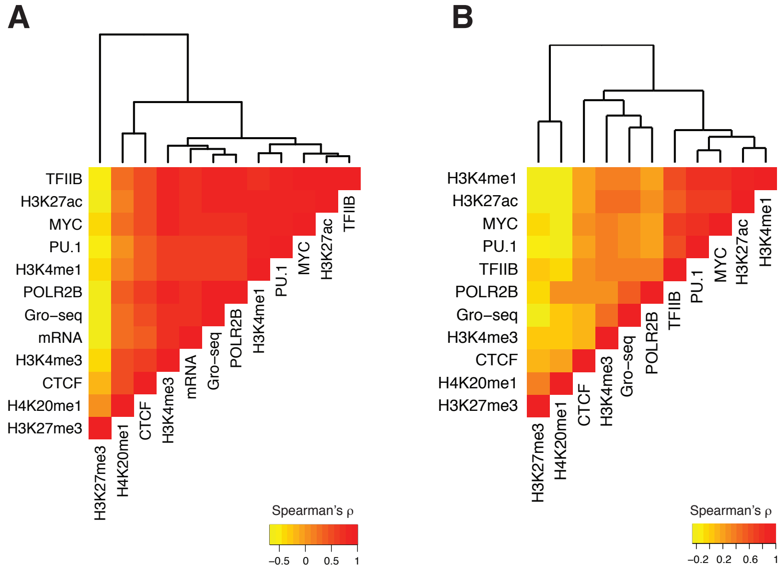


Figure B.5: Genome-wide properties of the probed molecular phenotypes. Correlation of molecular marks at promoters (transcription start sites +/- 2.5 kb) of protein-coding and lincRNA genes **A** and putative enhancers defined by DNaseI hypersensitivity sites **B** based on the trio metasample peaks. Plotted values are Spearman correlation coefficients based on z-score transformed read densities for ChIP, mRNA and nascent transcription (GRO-seq) assays.

Figure B.6: Summary of allele-specific (AS) effects discovered for each assay in the three sets of samples (CEU trio, YRI trio, unrelated eight CEU individuals). Accessible sites refer to heterozygous SNP sites that fulfill the general quality requirements of the AS analysis, whereas AS sites refer to significant AS effects detected. Numbers represent the mean of the samples in question. Ordering of assays by decreasing AS proportion in the CEU trio.

	CEU trio			**YRI trio**			**Unrelated eight individuals (CEU)**		
ASSAY	**Accessible**	**AS sites**	**Proportion AS**	**Accessible**	**AS sites**	**Proportion AS**	**Accessible**	**AS sites**	**Proportion AS**
H3K27me3	301	186	0.62	2911	957	0.33	22	3	0.16
POL2RB-narrow	1514	568	0.38	2630	935	0.36	2348	167	0.07
POL2RB-broad	7172	2371	0.33	12064	3497	0.29	9254	525	0.06
GRO-fwd	2689	754	0.28	NA	NA	NA	NA	NA	NA
GRO-rev	2492	705	0.28	NA	NA	NA	NA	NA	NA
MYC	1005	147	0.15	115	5	0.04	NA	NA	NA
PU.1	1510	160	0.11	917	83	0.09	930	154	0.17
CTCF	3315	335	0.1	4226	415	0.1	NA	NA	NA
H3K27ac	20381	1868	0.09	20600	2852	0.14	17545	1639	0.09
H4K20me1	426	36	0.08	359	34	0.1	NA	NA	NA
H3K4me1	15209	1103	0.07	30541	4486	0.15	179001	3417	0.19
RNA-seq	4963	258	0.05	7869	456	0.06	4061	190	0.05
TFIIB	249	12	0.05	75	7	0.09	136	19	0.14
H3K4me3	5194	197	0.04	8389	420	0.05	7594	1177	0.16

Figure B.7: Consistency of allele-specific (AS) effects within molecular phenotypes. **A** Consistency of AS effects within ChIP-seq peaks in the two trios and eight unrelated individuals. Data pooled across individuals. All peaks with at least two significant AS sites (P<=0.01) tested for the consistency of the allelic ratios. **B,C** Consistency of allele-specific (AS) effects over distance. **B** The consistency of the paternal allele ratio for significant (P<=0.01) AS sites within peaks is plotted against the quartile of peak length. Data pooled from the two trios per assay. Assays with fixed peak size are not included. The decrease in consistency as the domain length increases suggests that broad peaks may be more complex than their single peak definition implies. **C** Probability of allelic consistency between two AS sites within transcriptional and histone mark assays given the genomic distance between the SNPs. All pairs of SNPs within the regulatory landscape around gene regions (+/- 50 kb) were considered. Only assays showing significant correlation (logistic regression P <0.05) with distance are shown.

A

	Both trios (CEU and YRI)			Unrelated 8 CEU individuals		
Molecular phenotype	# tested peaks	# consistent peaks	% consistent peaks	# tested peaks	# consistent peaks	% consistent peaks
PU.1	47	46	98%	50	50	100%
TFIIB	1	1	100%	6	4	67%
MYC	13	12	92%	N/A	N/A	N/A
CTCF	102	101	99%	N/A	N/A	N/A
POLR2B-narrow	222	181	82%	79	73	92%
POLR2B-broad	2,548	1,120	44%	566	364	64%
H3K4me1	1902	1044	55%	452	256	57%
H3K4me3	120	109	91%	1266	769	61%
H3K27ac	1,079	704	65%	2041	1267	62%
H3K27me3	363	160	44%	NA	NA	NA
H4K20me1	10	4	40%	NA	NA	NA

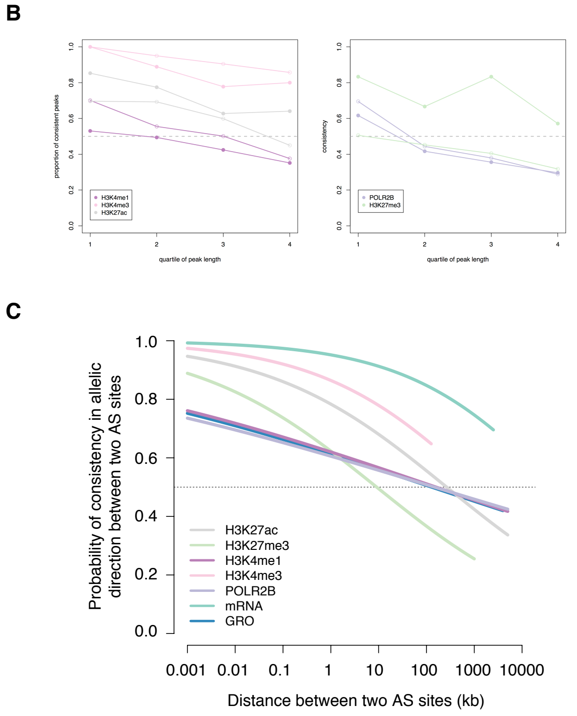


Table B.1: Sequencing statistics for all assays (ChIP-seq and RNA-seq) in the trio samples

Sample	Population	Sex	Assay	Antibody	Machine	Total reads	Usable reads (MAPQ>=10)	Proportion of usable reads
12878	CEU	F	H3K27ac	Abcam ab4729	HiSeq2000	70359451	62324767	0.89
12891	CEU	M	H3K27ac	Abcam ab4729	HiSeq2000	69108226	61426566	0.89
12892	CEU	F	H3K27ac	Abcam ab4729	HiSeq2000	43053435	35948291	0.83
19238	YRI	F	H3K27ac	Abcam ab4729	HiSeq2000	65470863	58583367	0.89
19239	YRI	M	H3K27ac	Abcam ab4729	HiSeq2000	44989382	38537159	0.86
19240	YRI	F	H3K27ac	Abcam ab4729	HiSeq2000	55112705	46932699	0.85
12878	CEU	F	H3K27me3	Millipore 17-622	HiSeq2000	172250298	106184985	0.62
12891	CEU	M	H3K27me3	Millipore 17-622	HiSeq2000	192263464	132431573	0.69
12892	CEU	F	H3K27me3	Millipore 17-622	HiSeq2000	217077781	113168651	0.52
19238	YRI	F	H3K27me3	Millipore 17-622	HiSeq2000	176973882	109249688	0.62
19239	YRI	M	H3K27me3	Millipore 17-622	HiSeq2000	159910142	103260822	0.65
19240	YRI	F	H3K27me3	Millipore 17-622	HiSeq2000	178851100	114117313	0.64
12878	CEU	F	H3K4me1	Abcam ab8895	HiSeq2000	238094924	199080946	0.84
12891	CEU	M	H3K4me1	Abcam ab8895	HiSeq2000	288209714	145366322	0.50
12892	CEU	F	H3K4me1	Abcam ab8895	HiSeq2000	292385011	118041373	0.40
19238	YRI	F	H3K4me1	Abcam ab8895	HiSeq2000	241109952	183615927	0.76
19239	YRI	M	H3K4me1	Abcam ab8895	HiSeq2000	235039807	185429357	0.79
19240	YRI	F	H3K4me1	Abcam ab8895	HiSeq2000	243792885	178829565	0.73
12878	CEU	F	H3K4me3	Millipore 17-614	GAII	38726870	24014808	0.62
12891	CEU	M	H3K4me3	Millipore 17-614	GAII	34841621	24698213	0.71
12892	CEU	F	H3K4me3	Millipore 17-614	GAII	40855708	28134136	0.69
19238	YRI	F	H3K4me3	Millipore 17-614	GAII	43293287	29959829	0.69
19239	YRI	M	H3K4me3	Millipore 17-614	GAII	42334884	31859790	0.75
19240	YRI	F	H3K4me3	Millipore 17-614	GAII	40548507	28168933	0.69
12878	CEU	F	H4K20me1	Abcam ab9051	GAII	39127290	26947519	0.69
12891	CEU	M	H4K20me1	Abcam ab9051	GAII	35353436	22113206	0.63
12892	CEU	F	H4K20me1	Abcam ab9051	GAII	29626106	19058044	0.64
19238	YRI	F	H4K20me1	Abcam ab9051	GAII	32263490	17900033	0.55
19239	YRI	M	H4K20me1	Abcam ab9051	GAII	33817709	20826482	0.62
19240	YRI	F	H4K20me1	Abcam ab9051	GAII	32541522	20029453	0.62
12878	CEU	F	MYC	Santa Cruz sc-764	HiSeq2000	184723869	135594208	0.73
12891	CEU	M	MYC	Santa Cruz sc-764	HiSeq2000	198511126	148086735	0.75
12892	CEU	F	MYC	Santa Cruz sc-764	HiSeq2000	193600103	148881581	0.77
19238	YRI	F	MYC	Santa Cruz sc-764	GAII	50756699	30109398	0.59
19239	YRI	M	MYC	Santa Cruz sc-764	GAII	43367433	29083069	0.67
19240	YRI	F	MYC	Santa Cruz sc-764	GAII	47055403	32494416	0.69
12878	CEU	F	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	HiSeq2000	178619999	127143929	0.71
12891	CEU	M	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	HiSeq2000	219047965	157243417	0.72
12892	CEU	F	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	HiSeq2000	240620717	179623084	0.75
19238	YRI	F	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	HiSeq2000	252527744	184041222	0.73
19239	YRI	M	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	HiSeq2000	235003947	174024139	0.74
19240	YRI	F	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	HiSeq2000	235656031	175312668	0.74
12878	CEU	F	PU.1	Santa Cruz sc-22805	GAII	45267160	32823189	0.73
12891	CEU	M	PU.1	Santa Cruz sc-22805	GAII	48394603	32906877	0.68
12892	CEU	F	PU.1	Santa Cruz sc-22805	GAII	48478749	34670209	0.72
19238	YRI	F	PU.1	Santa Cruz sc-22805	GAII	47955221	31216271	0.65
19239	YRI	M	PU.1	Santa Cruz sc-22805	GAII	45283181	30253430	0.67
19240	YRI	F	PU.1	Santa Cruz sc-22805	GAII	50289589	34103449	0.68
12878	CEU	F	TFIIB	rabbit CS396*	GAII	43250933	31111075	0.72
12891	CEU	M	TFIIB	rabbit CS396*	GAII	41341562	29937287	0.72
12892	CEU	F	TFIIB	rabbit CS396*	GAII	41859901	29976049	0.72
19238	YRI	F	TFIIB	rabbit CS396*	GAII	34146893	23302508	0.68
19239	YRI	M	TFIIB	rabbit CS396*	GAII	37190334	24966491	0.67
19240	YRI	F	TFIIB	rabbit CS396*	GAII	37680245	25954074	0.69
12878	CEU	F	CTCF	Millipore 07-729**	GAII	46021263	25977690	0.56
12891	CEU	M	CTCF	Millipore 07-729**	GAII	30244488	22854831	0.76
12892	CEU	F	CTCF	Millipore 07-729**	GAII	44885150	34535784	0.77
19238	YRI	F	CTCF	Millipore 07-729**	GAII	32377472	26150702	0.81
19239	YRI	M	CTCF	Millipore 07-729**	GAII	26628402	20306107	0.76
19240	YRI	F	CTCF	Millipore 07-729**	GAII	33399839	26516763	0.79
12878	CEU	F	RNA-seq	NA	HiSeq2000	37558398	24142888	0.643
12891	CEU	M	RNA-seq	NA	HiSeq2000	33455610	21651820	0.647
12892	CEU	F	RNA-seq	NA	HiSeq2000	40134722	25524620	0.636
19238	YRI	F	RNA-seq	NA	HiSeq2000	43166926	27595249	0.639
19239	YRI	M	RNA-seq	NA	HiSeq2000	37622042	23961104	0.637
19240	YRI	F	RNA-seq	NA	HiSeq2000	48889864	31420641	0.643

* See Schramm *et al.* 2000 for details.

** Data produced by McDaniell *et al.* 2010

Table B.2: Sequencing statistics for all assays (ChIP-seq and RNA-seq) in the eight unrelated individuals.

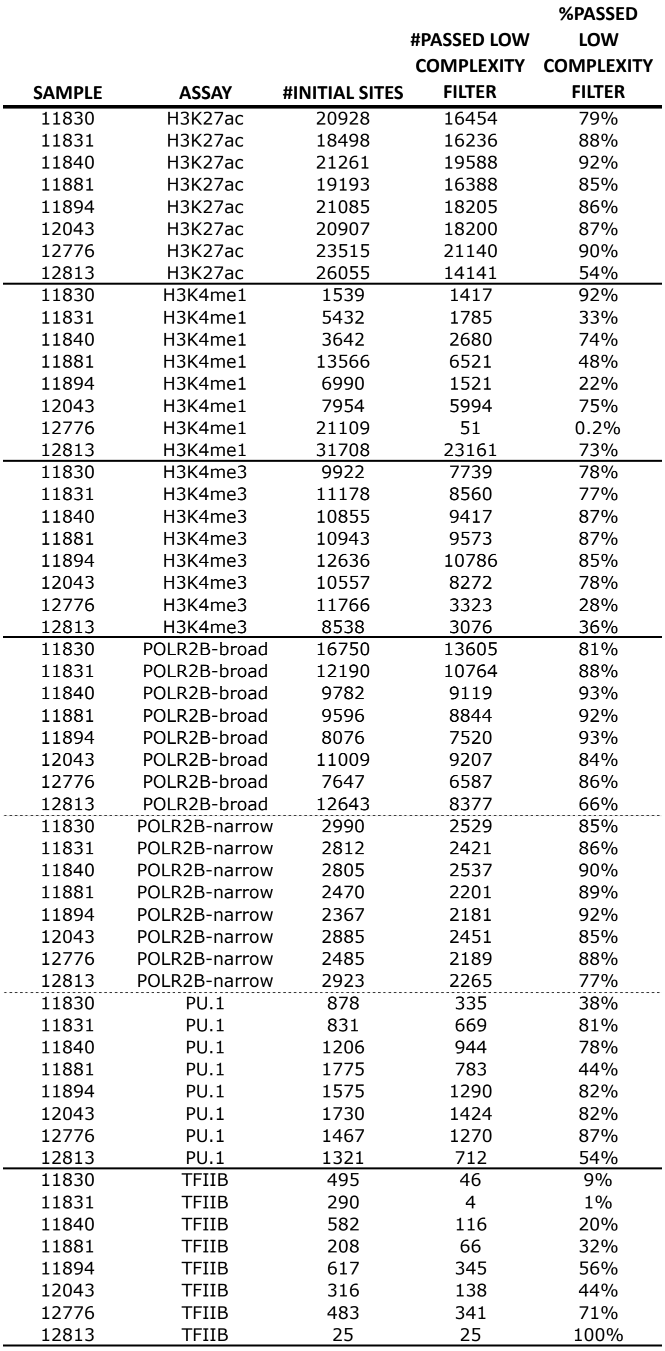
Sample	Population	Sex	Assay	Antibody	Total reads	Usable reads (MAPQ>=10)	Proportion of usable reads
11830	CEU	F	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	90274607	69375140	0.77
11831	CEU	M	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	74449496	56297515	0.76
11840	CEU	F	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	76235942	57469901	0.75
11881	CEU	M	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	62384167	48330942	0.78
11894	CEU	F	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	58461111	47037691	0.81
12043	CEU	M	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	64792841	44664006	0.69
12776	CEU	F	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	64591102	39909878	0.62
12813	CEU	F	RNA polymerase II (RPB2 subunit)	Santa Cruz sc-67318 (POLR2B)	91975825	59932875	0.65
11830	CEU	F	TFIIB	rabbit CS396*	47140228	35635483	0.76
11831	CEU	M	TFIIB	rabbit CS396*	52000647	24007661	0.46
11840	CEU	F	TFIIB	rabbit CS396*	51698503	34471644	0.67
11881	CEU	M	TFIIB	rabbit CS396*	43830419	33954992	0.78
11894	CEU	F	TFIIB	rabbit CS396*	69784956	48965638	0.70
12043	CEU	M	TFIIB	rabbit CS396*	42641208	31461390	0.74
12776	CEU	F	TFIIB	rabbit CS396*	48553552	36844244	0.76
12813	CEU	F	TFIIB	rabbit CS396*	50931075	38214596	0.75
11830	CEU	F	H3K27me3	Millipore 17-622	46438781	32852149	0.71
11831	CEU	M	H3K27me3	Millipore 17-622	57232570	32471588	0.57
11840	CEU	F	H3K27me3	Millipore 17-622	NA	NA	NA
11881	CEU	M	H3K27me3	Millipore 17-622	44670196	32443401	0.73
11894	CEU	F	H3K27me3	Millipore 17-622	56936419	30236924	0.53
12043	CEU	M	H3K27me3	Millipore 17-622	80945483	62434159	0.77
12776	CEU	F	H3K27me3	Millipore 17-622	75971667	59266493	0.78
12813	CEU	F	H3K27me3	Millipore 17-622	61877423	42302148	0.68
11830	CEU	F	H3K4me1	Abcam ab8895	161172573	128142508	0.80
11831	CEU	M	H3K4me1	Abcam ab8895	167212603	132761272	0.79
11840	CEU	F	H3K4me1	Abcam ab8895	177891799	139884277	0.79
11881	CEU	M	H3K4me1	Abcam ab8895	184522820	143159940	0.78
11894	CEU	F	H3K4me1	Abcam ab8895	145483060	117182105	0.81
12043	CEU	M	H3K4me1	Abcam ab8895	159590525	130881735	0.82
12776	CEU	F	H3K4me1	Abcam ab8895	149312151	121306615	0.81
12813	CEU	F	H3K4me1	Abcam ab8895	234604760	178669798	0.76
11830	CEU	F	H3K4me3	Millipore 17-614	58094045	46238912	0.80
11831	CEU	M	H3K4me3	Millipore 17-614	57455187	46252256	0.81
11840	CEU	F	H3K4me3	Millipore 17-614	74779896	59262845	0.79
11881	CEU	M	H3K4me3	Millipore 17-614	50453122	42122669	0.84
11894	CEU	F	H3K4me3	Millipore 17-614	75298221	62099764	0.83
12043	CEU	M	H3K4me3	Millipore 17-614	46299638	39486039	0.85
12776	CEU	F	H3K4me3	Millipore 17-614	73488265	60374223	0.82
12813	CEU	F	H3K4me3	Millipore 17-614	39792419	33345787	0.84
11830	CEU	F	PU.1	Santa Cruz sc-22805	53432856	16097627	0.30
11831	CEU	M	PU.1	Santa Cruz sc-22805	58703083	22347410	0.38
11840	CEU	F	PU.1	Santa Cruz sc-22805	43980556	32357862	0.74
11881	CEU	M	PU.1	Santa Cruz sc-22805	60927758	46474872	0.76
11894	CEU	F	PU.1	Santa Cruz sc-22805	66361339	48088769	0.73
12043	CEU	M	PU.1	Santa Cruz sc-22805	55867108	43040806	0.77
12776	CEU	F	PU.1	Santa Cruz sc-22805	67022786	46230170	0.69
12813	CEU	F	PU.1	Santa Cruz sc-22805	51062972	36225771	0.71
11830	CEU	F	H3K27ac	Abcam ab4729	59740687	52701922	0.88
11831	CEU	M	H3K27ac	Abcam ab4729	48785055	40724424	0.84
11840	CEU	F	H3K27ac	Abcam ab4729	63439157	52502660	0.83
11881	CEU	M	H3K27ac	Abcam ab4729	52619082	44519445	0.85
11894	CEU	F	H3K27ac	Abcam ab4729	54954122	46111359	0.84
12043	CEU	M	H3K27ac	Abcam ab4729	60157660	52354331	0.87
12776	CEU	F	H3K27ac	Abcam ab4729	60930972	51858110	0.85
12813	CEU	F	H3K27ac	Abcam ab4729	73452846	62992939	0.86
11830	CEU	F	RNA-seq	NA	35203512	21054217	0.60
11831	CEU	M	RNA-seq	NA	31872450	20253224	0.64
11840	CEU	F	RNA-seq	NA	38674322	19803844	0.51
11881	CEU	M	RNA-seq	NA	29541316	18842529	0.64
11894	CEU	F	RNA-seq	NA	28380900	17952424	0.63
12043	CEU	M	RNA-seq	NA	34284620	21200758	0.62
12776	CEU	F	RNA-seq	NA	40541082	25202817	0.62
12813	CEU	F	RNA-seq	NA	26419600	16889023	0.64

* See Schramm *et al.* 2000 for details.

Table B.3: Filtering of low complexity sites for allele-specific analysis in the CEU and YRI trios.

SAMPLE	ASSAY	#INITIAL SITES	#PASSED LOW COMPLEXITY FILTER	%PASSED LOW COMPLEXITY FILTER
12878	CTCF	3983	3683	92.5%
12891	CTCF	3247	2812	86.6%
12892	CTCF	3999	3448	86.2%
19238	CTCF	4854	4236	87.3%
19239	CTCF	4095	3737	91.3%
19240	CTCF	5334	4701	88.1%
12878	H3K27ac	27344	25165	92.0%
12891	H3K27ac	27251	24387	89.5%
12892	H3K27ac	15860	11587	73.1%
19238	H3K27ac	31058	22376	72.0%
19239	H3K27ac	22780	17058	74.9%
19240	H3K27ac	28944	22363	77.3%
12878	H3K27me3	4994	281	5.6%
12891	H3K27me3	24402	422	1.7%
12892	H3K27me3	21489	196	0.9%
19238	H3K27me3	4823	2625	54.4%
19239	H3K27me3	9816	3733	38.0%
19240	H3K27me3	3795	2372	62.5%
12878	H3K4me1	34727	31929	91.9%
12891	H3K4me1	7237	6693	92.5%
12892	H3K4me1	7239	7001	96.7%
19238	H3K4me1	31618	27137	85.8%
19239	H3K4me1	35898	30828	85.9%
19240	H3K4me1	35377	33654	95.1%
12878	H3K4me3	6606	6300	95.4%
12891	H3K4me3	4540	4322	95.2%
12892	H3K4me3	5350	4956	92.6%
19238	H3K4me3	8189	7753	94.7%
19239	H3K4me3	8211	7723	94.1%
19240	H3K4me3	10348	9688	93.6%
12878	H4K20me1	1798	1102	61.3%
12891	H4K20me1	538	133	24.7%
12892	H4K20me1	952	40	4.2%
19238	H4K20me1	606	311	51.3%
19239	H4K20me1	2427	704	29.0%
19240	H4K20me1	78	60	76.9%
12878	MYC	1912	1667	87.2%
12891	MYC	596	492	82.6%
12892	MYC	991	854	86.2%
19238	MYC	287	271	94.4%
19239	MYC	34	33	97.1%
19240	MYC	46	39	84.8%
12878	POLR2B-broad	13012	5086	39.1%
12891	POLR2B-broad	12286	4116	33.5%
12892	POLR2B-broad	16417	12310	75.0%
19238	POLR2B-broad	16930	12020	71.0%
19239	POLR2B-broad	22408	4174	18.6%
19240	POLR2B-broad	23579	19994	84.8%
12878	POLR2B-narrow	2091	1391	66.5%
12891	POLR2B-narrow	2082	1274	61.2%
12892	POLR2B-narrow	2648	1873	70.7%
19238	POLR2B-narrow	3489	2648	75.9%
19239	POLR2B-narrow	3659	1748	47.8%
19240	POLR2B-narrow	4446	3492	78.5%
12878	PU.1	1667	1508	90.5%
12891	PU.1	1453	1327	91.3%
12892	PU.1	1967	1691	86.0%
19238	PU.1	855	774	90.5%
19239	PU.1	729	630	86.4%
19240	PU.1	1536	1343	87.4%
12878	TFIIB	380	334	87.9%
12891	TFIIB	246	216	87.8%
12892	TFIIB	235	194	82.6%
19238	TFIIB	126	96	76.2%
19239	TFIIB	97	72	74.2%
19240	TFIIB	59	54	91.5%
12878	GRO-seq (fwd)	4014	2643	65.8%
12891	GRO-seq (fwd)	5142	3970	77.2%
12892	GRO-seq (fwd)	4261	1406	33.0%
12878	GRO-seq (rev)	3551	2373	66.8%
12891	GRO-seq (rev)	4802	3717	77.4%
12892	GRO-seq (rev)	3898	1265	32.5%

Table B.4: Filtering of low complexity sites for allele-specific analysis in the unrelated eight CEU individuals.



SAMPLE	ASSAY	#INITIAL SITES	#PASSED LOW COMPLEXITY FILTER	%PASSED LOW COMPLEXITY FILTER
11830	H3K27ac	20928	16454	79%
11831	H3K27ac	18498	16236	88%
11840	H3K27ac	21261	19588	92%
11881	H3K27ac	19193	16388	85%
11894	H3K27ac	21085	18205	86%
12043	H3K27ac	20907	18200	87%
12776	H3K27ac	23515	21140	90%
12813	H3K27ac	26055	14141	54%
11830	H3K4me1	1539	1417	92%
11831	H3K4me1	5432	1785	33%
11840	H3K4me1	3642	2680	74%
11881	H3K4me1	13566	6521	48%
11894	H3K4me1	6990	1521	22%
12043	H3K4me1	7954	5994	75%
12776	H3K4me1	21109	51	0.2%
12813	H3K4me1	31708	23161	73%
11830	H3K4me3	9922	7739	78%
11831	H3K4me3	11178	8560	77%
11840	H3K4me3	10855	9417	87%
11881	H3K4me3	10943	9573	87%
11894	H3K4me3	12636	10786	85%
12043	H3K4me3	10557	8272	78%
12776	H3K4me3	11766	3323	28%
12813	H3K4me3	8538	3076	36%
11830	POLR2B-broad	16750	13605	81%
11831	POLR2B-broad	12190	10764	88%
11840	POLR2B-broad	9782	9119	93%
11881	POLR2B-broad	9596	8844	92%
11894	POLR2B-broad	8076	7520	93%
12043	POLR2B-broad	11009	9207	84%
12776	POLR2B-broad	7647	6587	86%
12813	POLR2B-broad	12643	8377	66%
11830	POLR2B-narrow	2990	2529	85%
11831	POLR2B-narrow	2812	2421	86%
11840	POLR2B-narrow	2805	2537	90%
11881	POLR2B-narrow	2470	2201	89%
11894	POLR2B-narrow	2367	2181	92%
12043	POLR2B-narrow	2885	2451	85%
12776	POLR2B-narrow	2485	2189	88%
12813	POLR2B-narrow	2923	2265	77%
11830	PU.1	878	335	38%
11831	PU.1	831	669	81%
11840	PU.1	1206	944	78%
11881	PU.1	1775	783	44%
11894	PU.1	1575	1290	82%
12043	PU.1	1730	1424	82%
12776	PU.1	1467	1270	87%
12813	PU.1	1321	712	54%
11830	TFIIB	495	46	9%
11831	TFIIB	290	4	1%
11840	TFIIB	582	116	20%
11881	TFIIB	208	66	32%
11894	TFIIB	617	345	56%
12043	TFIIB	316	138	44%
12776	TFIIB	483	341	71%
12813	TFIIB	25	25	100%

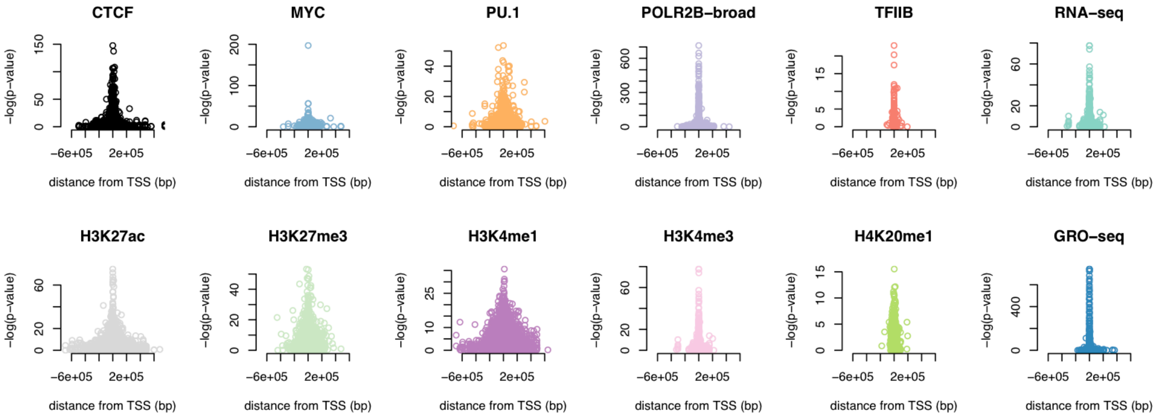


Figure B.8: Distance of allele-specific SNP sites from the closest annotated transcription start site. All accessible heterozygous sites per assay are plotted.

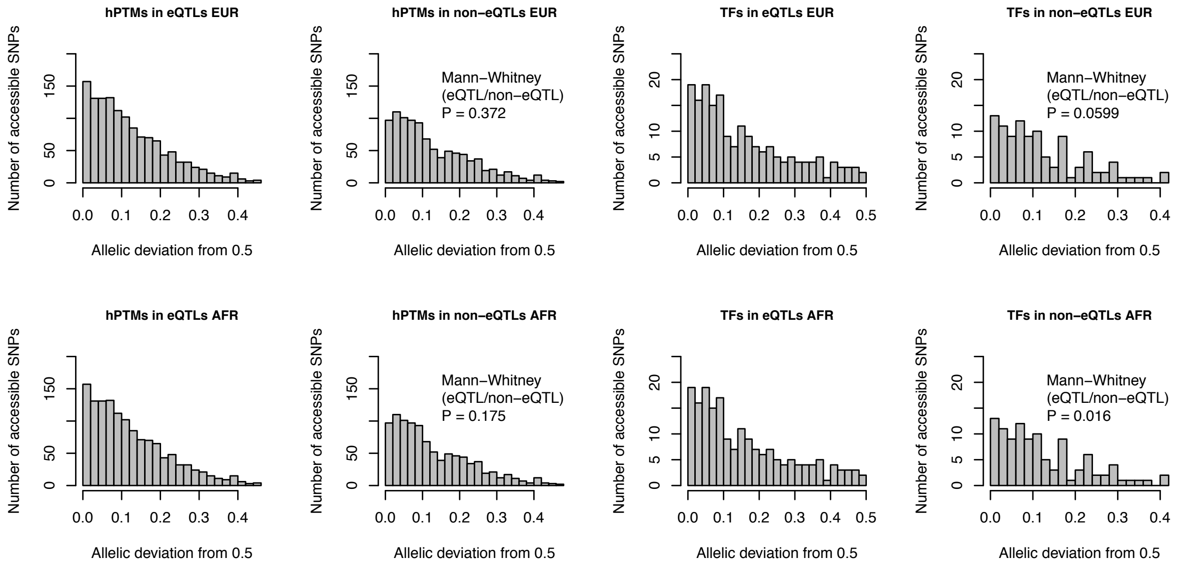


Figure B.9: All accessible heterozygous SNPs for transcription factors (PU.1, MYC, CTCF, TFIIB) and histone modifications (H3K4me1, H3K4me3, H3K27ac, H4K20me1, H3K27me3) overlapping known eQTL SNPs and matched non-eQTL SNPs in the 1000 Genomes Project phase1 European and African populations. Transcription factors show a slight enrichment of allelic bias at eQTLs compared to non-eQTLs from the African population (Mann-Whitney U test P = 0.016 between allele ratios for eQTLs and null) and a similar slight trend was observed in the European population. No enrichment was observed for hPTMs. Of the individual assays, CTCF contributes most of the overlapping TF sites and, if excluded, the enrichment at eQTLs does not remain significant (P = 0.85). Of the individual hPTMs, H3K4me1 shows a minor enrichment at African eQTLs(P = 0.01). Based on a large-scale sequencing-based eQTL study (7), the best eQTL variant per exon was also the causal variant for the observed expression change in 55% of EUR eQTLs and 74% of YRI eQTLs (same set of eQTLs was used for the current analysis). A more conservative estimate was 34% and 41%, respectively. In line with this, the observed enrichment of allelic bias at eQTLs for TFs in the current study is only significant at African eQTLs.

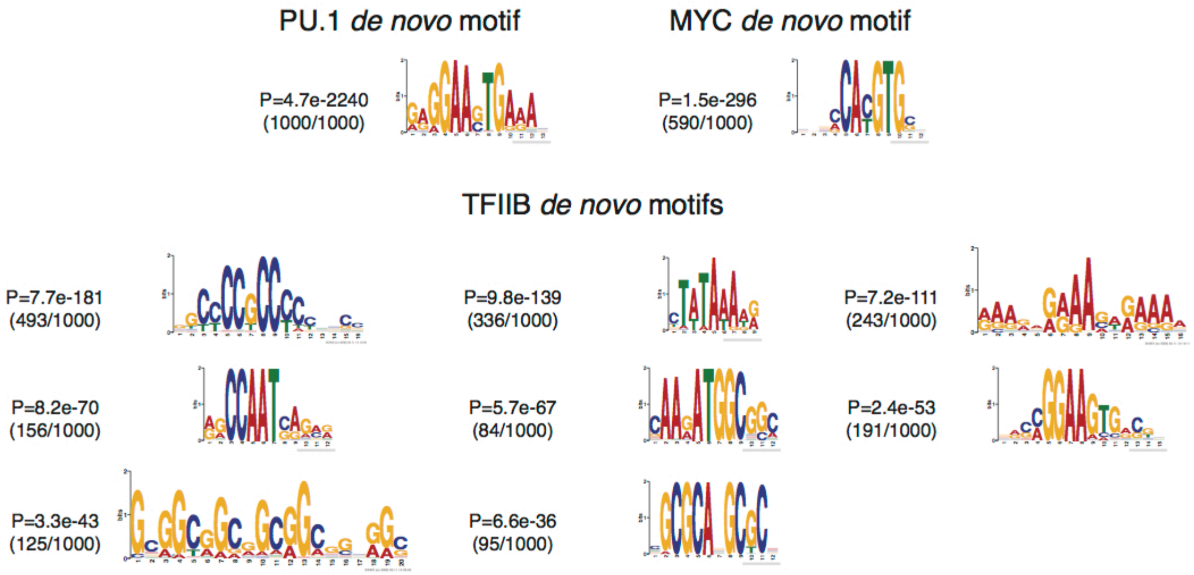


Figure B.10: *De novo* derived motifs from meta-sample PU.1, MYC, and TFIIB ChIP-seq peaks. To analyze the mechanism underlying allele-specific binding events we first derived de novo binding motifs for PU.1 and MYC and found that the inferred motifs were identical to the ones previously published [168]. *De novo* motif search was also conducted for TFIIB, which does not bind DNA directly on its own. Multiple known promoter motifs were discovered, consistent with the well-characterized association of TFIIB with POL2RB and the transcription initiation complex [169].

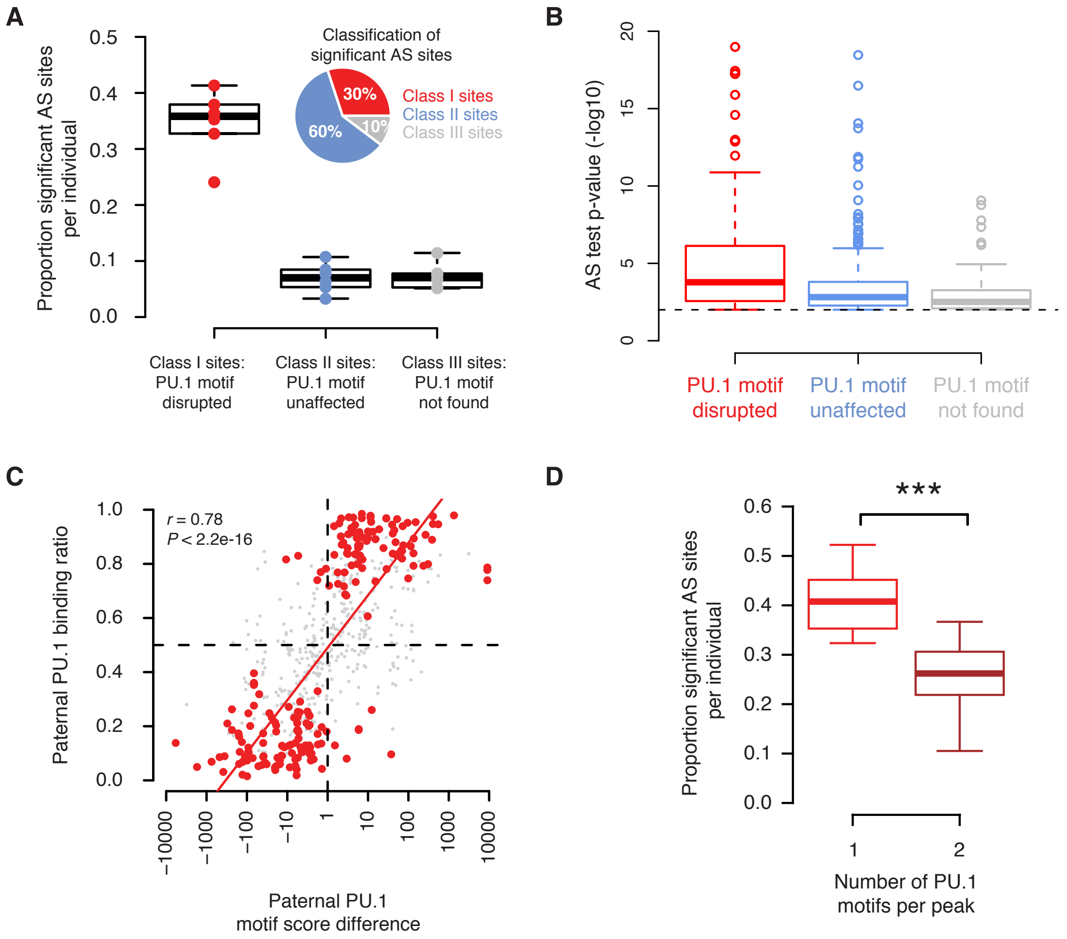


Figure B.11: Genome-wide analysis of allele-specific (AS) PU.1 binding. **A** Enrichment and classification (inlet) of significant AS PU.1 SNPs with reference to PU.1 binding site location within peaks. Data from trio individuals combined (n = 6). **B** ASB binomial test p-value distribution for significant B-SNPs. **C** PU.1 motif score changers are predictive of AS PU.1 binding. Ratio between paternal and maternal PU.1 PWM scores (x-axis) and proportion of reads mapping to the paternal allele (y-axis) (red, significant sites; grey, non-significant sites). **D** Peaks with multiple homotypic PU.1 motifs show reduced AS binding activity. Peaks were split into two groups, i.e., peaks with one and two PU.1 motifs, respectively, and the proportion of significant AS sites per group was calculated (y-axis). Significance was determined using the Mann-Whitney-U test.

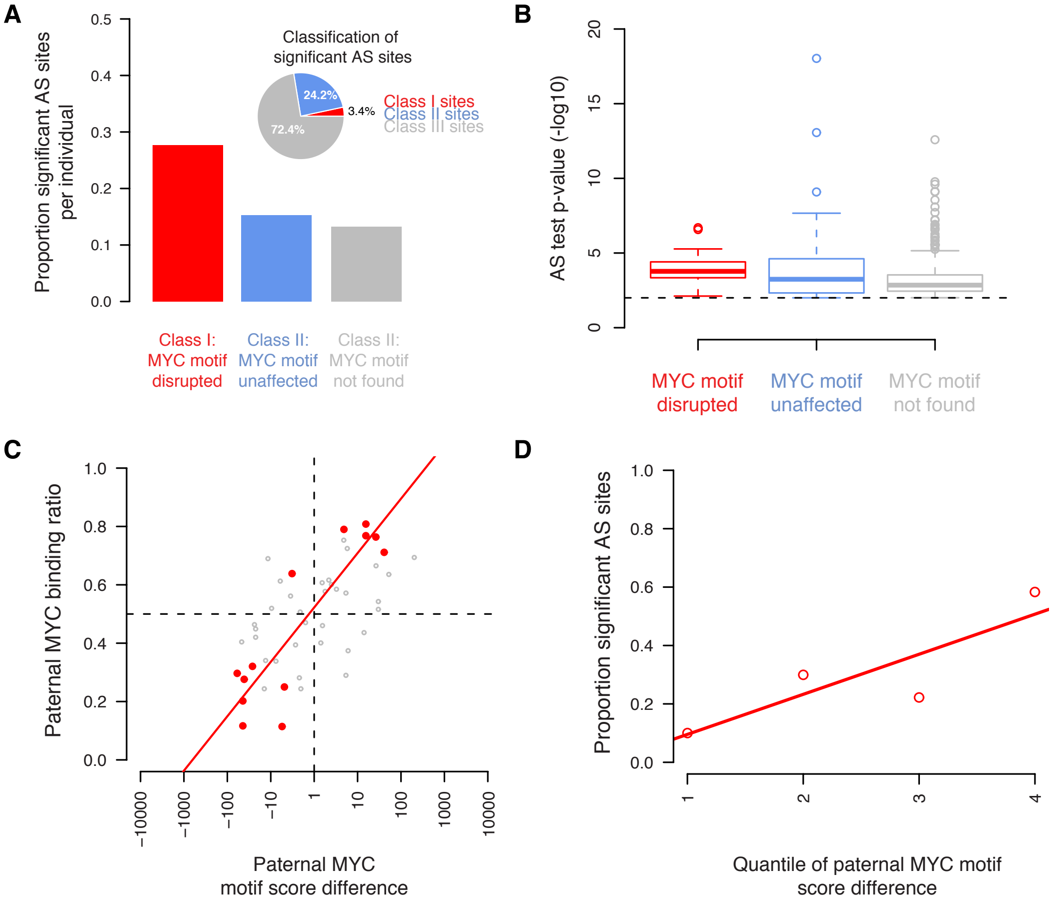


Figure B.12: Genome-wide analysis of allele-specific (AS) MYC binding. **A** Enrichment and classification (inlet) of significant AS MYC SNPs with reference to MYC binding site location within peaks. Data from CEU trio combined (n = 3). **B**. ASB binomial test p-value distribution for significant B-SNPs. **C** MYC motif score changers are predictive of AS MYC binding. Ratio between paternal and maternal MYC PWM scores (x-axis) and proportion of reads mapping to the paternal allele (y-axis) (red, significant sites; grey, non-significant sites). **D** B-SNPs with high impact on the MYC motif show more frequent signals of allele-specific binding. All SNPs located within MYC binding sites were grouped into quartiles (x-axis) and the fraction of significant B-SNPs per group was calculated (y-axis).

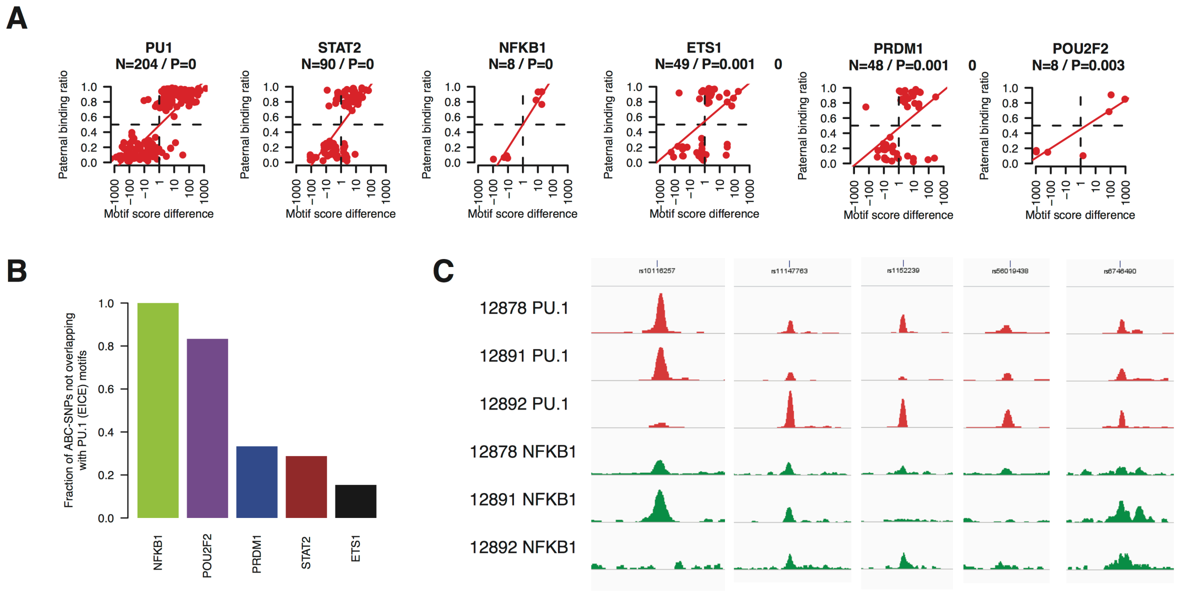


Figure B.13: Allele-specific binding cooperativity at PU.1-binding sites. **A** Covariable TF motifs within PU.1 peaks are predictive of allele-specific PU.1 binding. All accessible 35 ChIP-seq-derived TF motifs [168] were tested for allele-specific (AS) association between TF PWM score covariance and AS PU.1 binding activity. Data from CEU and YRI trios was combined and only significant PU.1 AS binding sites were considered. Tested motifs were sorted according to Pearson correlation P-value (left to right) and only significant motifs are shown (5% false discovery rate). The consensus PU.1 motif served as a positive control and ranked first among all tested motifs. The header of each panel indicates the motif, number of tested SNPs, and Pearson correlation test P-value. Of note, in some instances the impact of variants on co-operative motifs might be buffered due to heterotypic clusters of TFBS. **B** Overlap between predicted TF binding sites and PU.1 binding sites for significant covariant motifs. Binding sites discovered using the allele-specific binding cooperativity test pipeline were often shared between PU.1 and co- associated TFs indicating binding site ambiguity at significant PU.1 SNP sites. **C** Functional validation of predicted covariance between SNPs in NFKB1 motifs and PU.1 binding. All variable PU.1 peaks with unaffected (or not predicted) PU.1 motifs and an unambiguously disrupted NFKB1 motif were inspected for NFKB1 binding in the CEU trio. NFKB1 ChIP-seq datasets were obtained from Kasowski et al. 2010 [52].

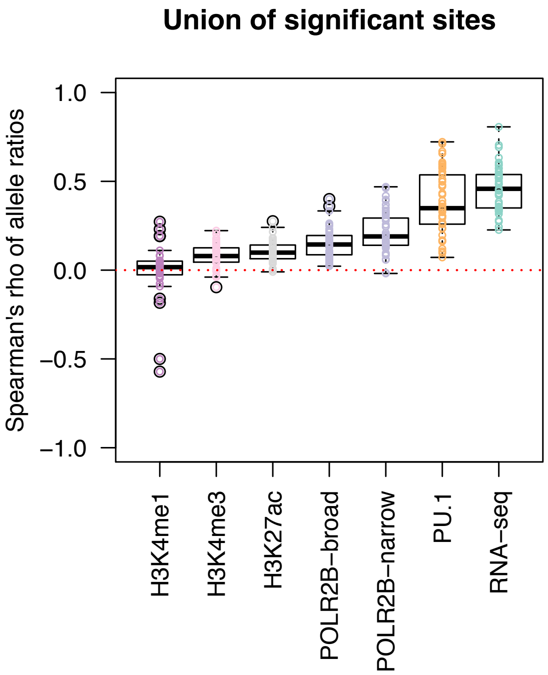


Figure B.14: Distribution of pairwise correlation coefficients of the union of significant allele-specific sites (i.e. significant in either one or both individuals) between all unrelated CEU individuals (n = 10) for each assay. Correlation of the reference allele is calculated for each comparison using Spearman's rank. Correlation is low for hPTMs but relatively high for PU.1 and mRNA, further supporting stronger genetic influence on TF binding and gene expression than chromatin marks.

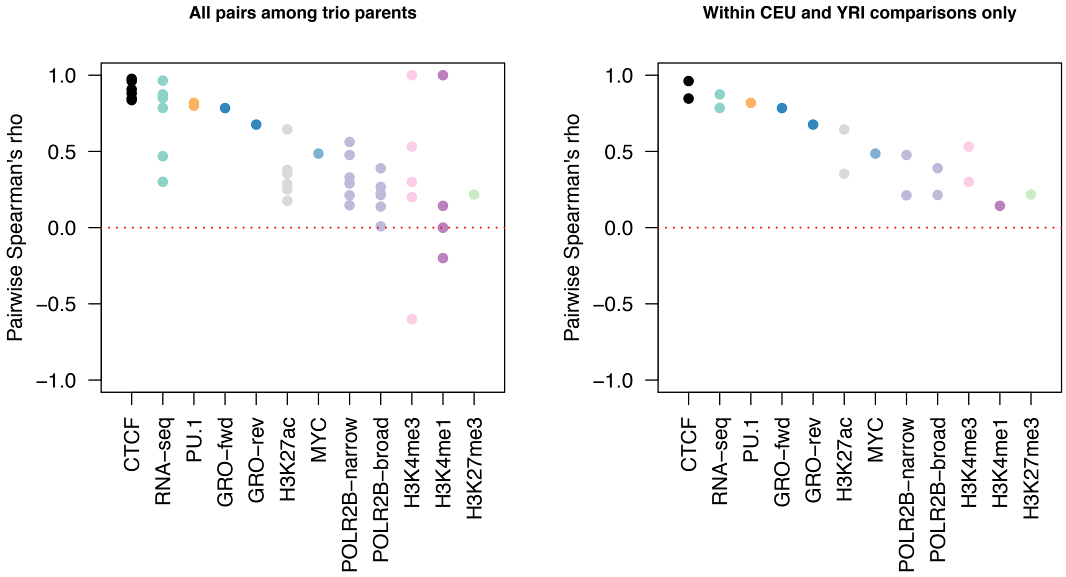


Figure B.15: Distribution of pairwise correlation coefficients of shared significant allele-specific sites between the trio parents (n = 4) for each assay. The correlation of the reference allele ratio at shared significant AS sites was calculated for each comparison using Spearman rank correlation. Left: All comparisons among the four parents. Right: Comparisons within each trio only.

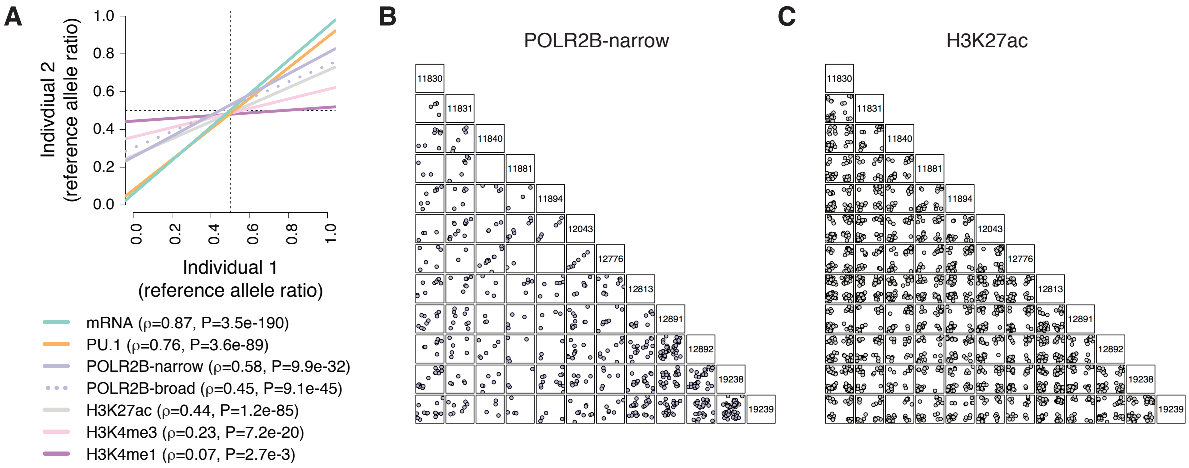


Figure B.16: Pairwise correlation of allele ratios in all unrelated individuals at heterozygous SNP sites with a shared significant allele-specific effect (P <= 0.01) in any two individuals. Correlation of the reference allele ratios at sites pooled across pairs of individuals (Spearman rank correlation) **A** and individual pairwise examples of POLR2B-narrow **B** and H3K27ac **C**.

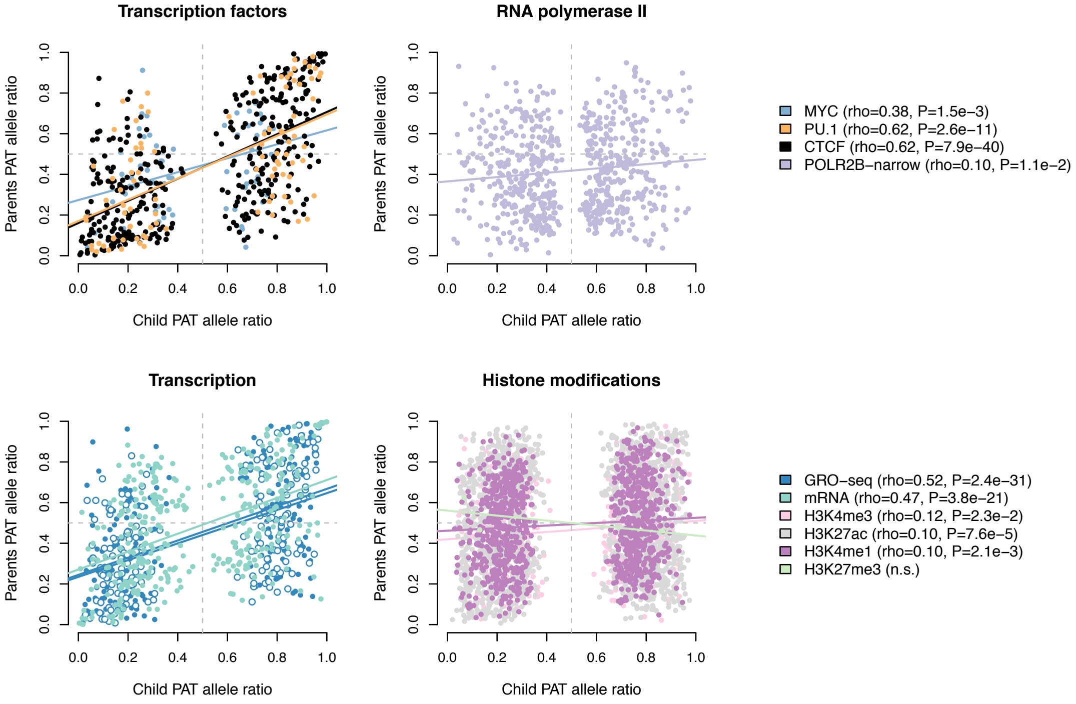


Figure B.17: Extension 1 of the parental transmission analysis: Transmission of allelic effects at SNP sites where child has a significant (P <= 0.01) AS effect, one parent is homozygous, and the other parent heterozygous. Results are shown for all accessible assays.

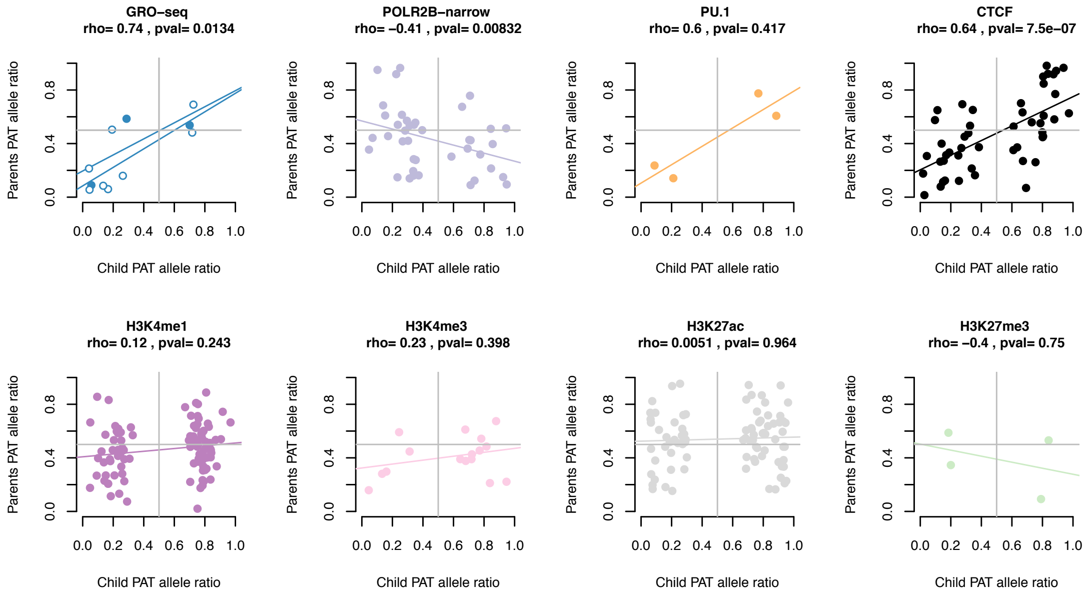


Figure B.18: Standard transmission analysis results for all accessible assays at putative enhancer loci allowing for intronic loci. GRO-seq signal is plotted separately for each strand (filled and empty points, forward and reverse strand, respectively), but rho and P-values represent both strands combined. mRNA-seq was not considered at enhancers.

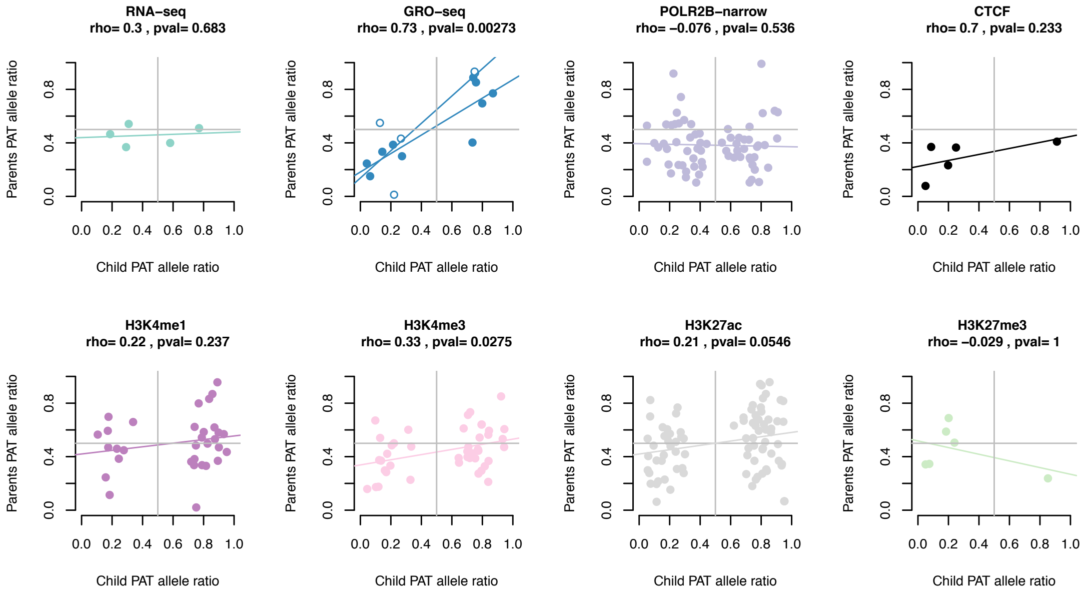


Figure B.19: Standard transmission analysis results for all accessible assays at promoters (transcription start site +/- 2.5 kb). GRO-seq signal is plotted separately for each strand (filled and empty points, forward and reverse strand, respectively), but rho and P-values represent both strands combined.

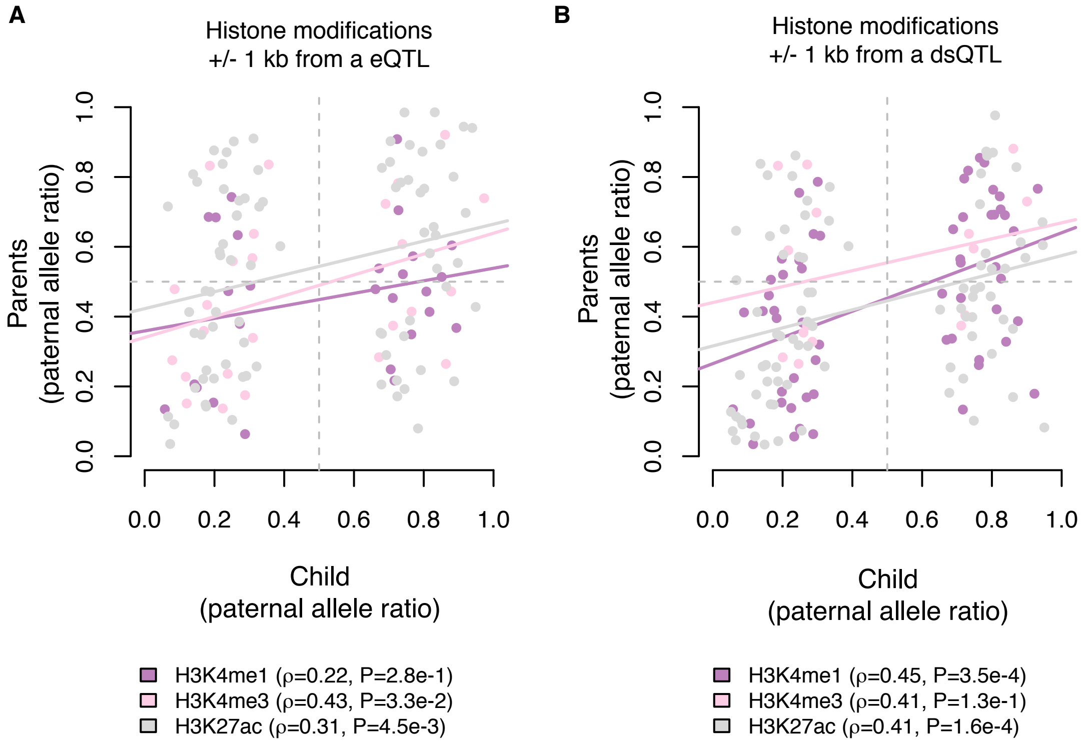


Figure B.20: Extension 1 transmission analysis results for all accessible histone modifications at known eQTLs and dsQTLs (+/- 1 kb). Analysis of standard transmission at eQTLs could not be performed due to low number of accessible sites.

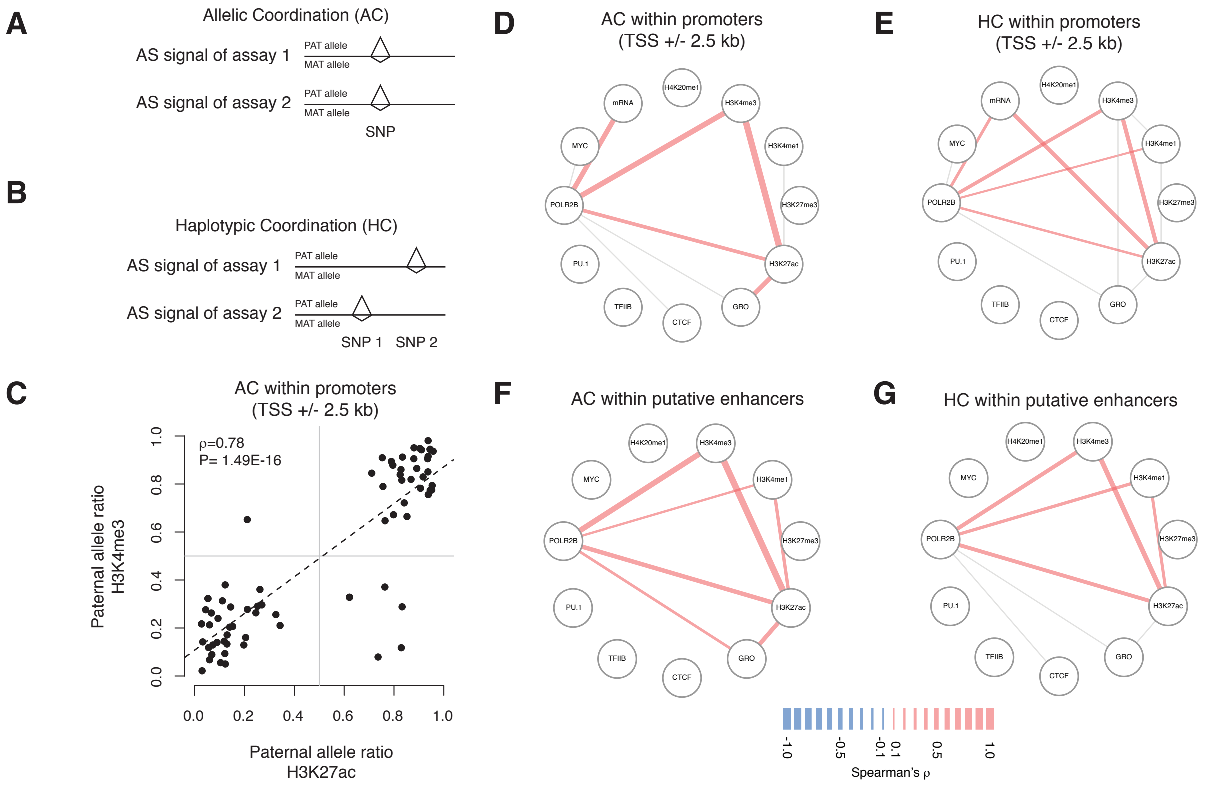


Figure B.21: Allelic coordination (AC) **A** and haplotypic coordination (HC) **B** at promoters (TSS +/- 2.5 kb) **D,E** and putative enhancers **F,G**. A minimum of 20 SNP pairs per comparison was required to perform an AC/HC test. Significant correlation coefficients (P <0.05, Spearman rank) are indicated with colored lines and non-significant correlations with gray. Missing lines indicate lack of sufficient data points for analysis. An example of significant AC between H3K27ac and H3K4me3 at promoters is provided in **C**. Sites pooled from the two trios.

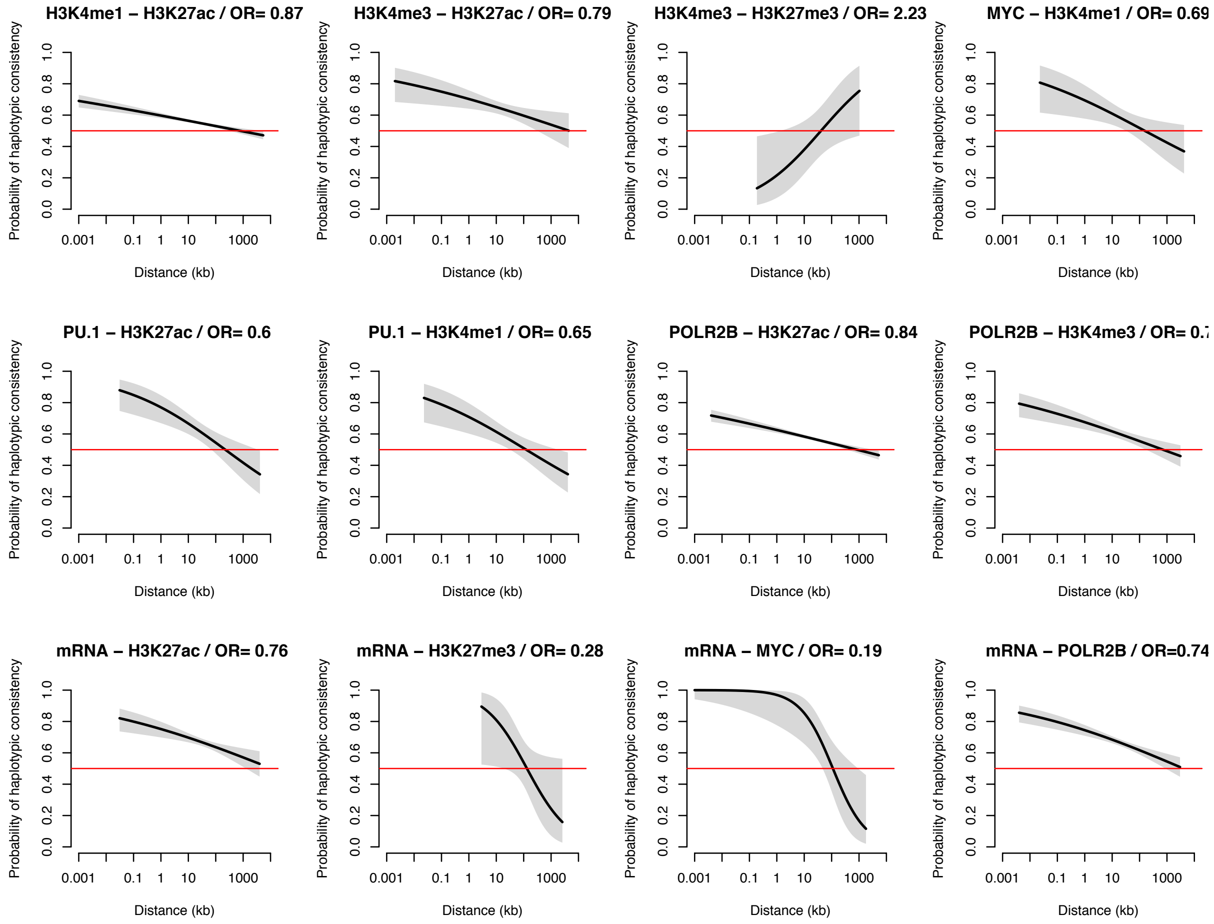


Figure B.22: Analysis of haplotypic consistency (HC) and genomic distance ($\log_{10}|bp|$) between all pairs of assays around and within gene regions (+/- 50 kb). Significantly correlated assay pairs are presented (logistic regression P <0.05). Grey areas show 95% confidence bands.

Figure B.23: SNP variants available for allele-specific analysis.

Sample ID	Population	Relation	Sex	1K dataset	HET SNPs	Phased HETs	% Phased
GM12878	CEU	child	F	Pilot2	1702593	1410467	82.8
GM12891	CEU	father	M	Pilot2	1630518	1338412	82.1
GM12892	CEU	mother	F	Pilot2	1667890	1375784	82.5
GM19238	YRI	mother	F	Pilot2	2065238	1720785	83.3
GM19239	YRI	father	M	Pilot2	2111292	1766838	83.7
GM19240	YRI	child	F	Pilot2	2226055	1881591	84.5
GM11830	CEU	unrelated	F	Pilot1	2067098	2035359	98.5
GM11831	CEU	unrelated	M	Pilot1	2009894	1991683	99.1
GM11840	CEU	unrelated	F	Pilot1	2025276	1979175	97.7
GM11881	CEU	unrelated	M	Pilot1	1952805	1943129	99.5
GM11894	CEU	unrelated	F	Pilot1	2078431	2041371	98.2
GM12043	CEU	unrelated	M	Pilot1	1966198	1951877	99.3
GM12776	CEU	unrelated	F	Pilot1	2062872	2012149	97.5
GM12813	CEU	unrelated	F	Pilot1	2019247	1972813	97.7

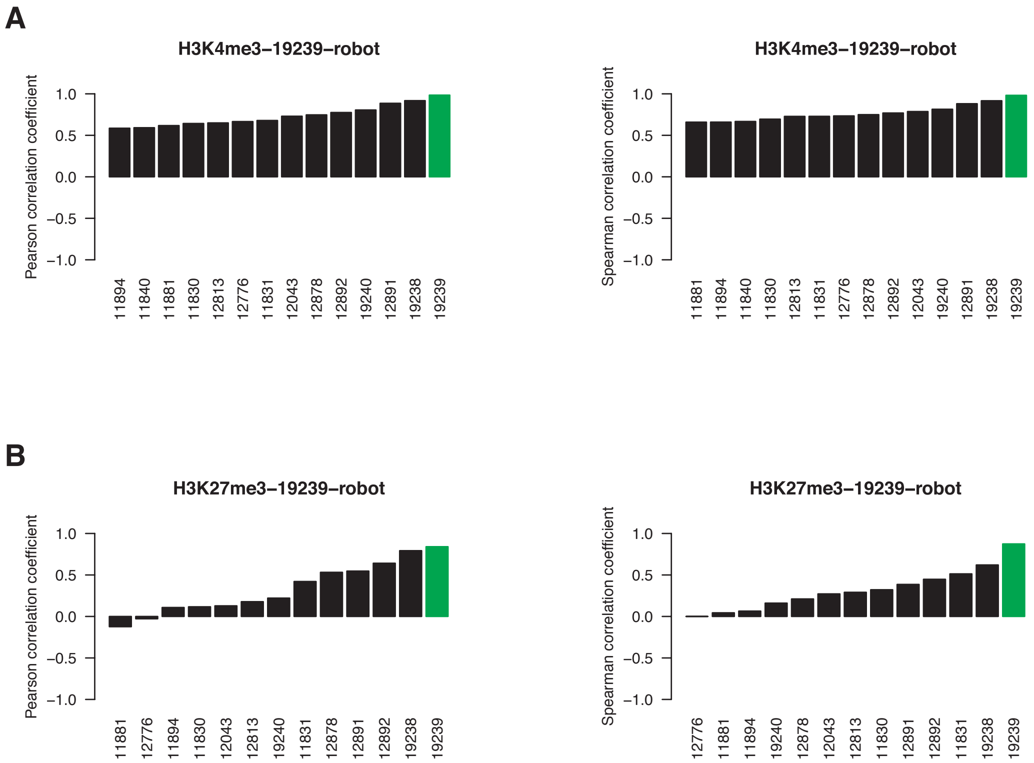


Figure B.24: Correlation of peak quantifications among different individuals and biological replicates of **A** H3K4me3 and **B** H3K27me3. Biological replicates (independent ChIP on the same cell preparation) are indicated in green. Less variation was observed between independent ChIP experiments than between any two unrelated individuals, suggesting low levels of technical variability.

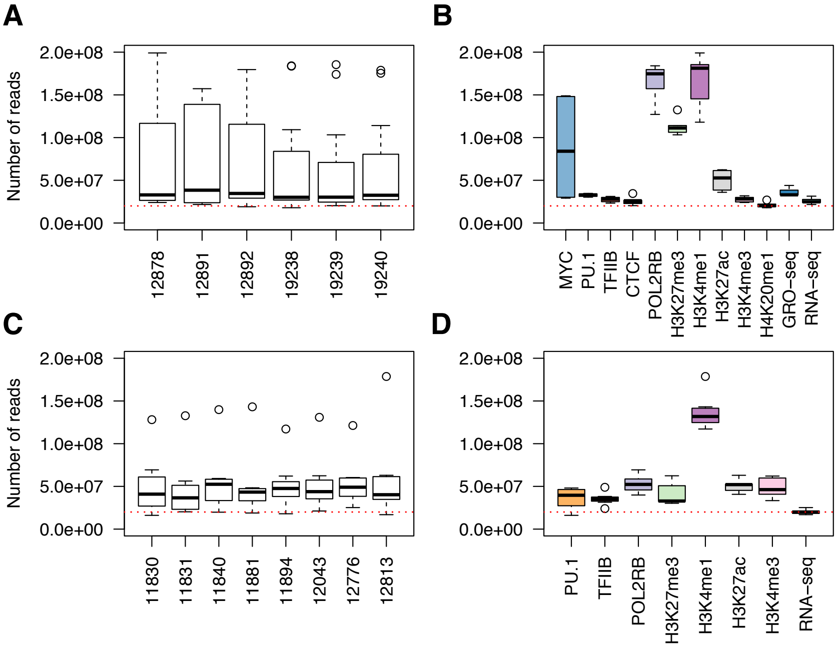


Figure B.25: Distribution of total number of uniquely mapping reads (MAPQ >= 10) per sample **A,C** and assay **B,D** for the two trios and the eight unrelated individuals. Dashed line is at 20 million reads.

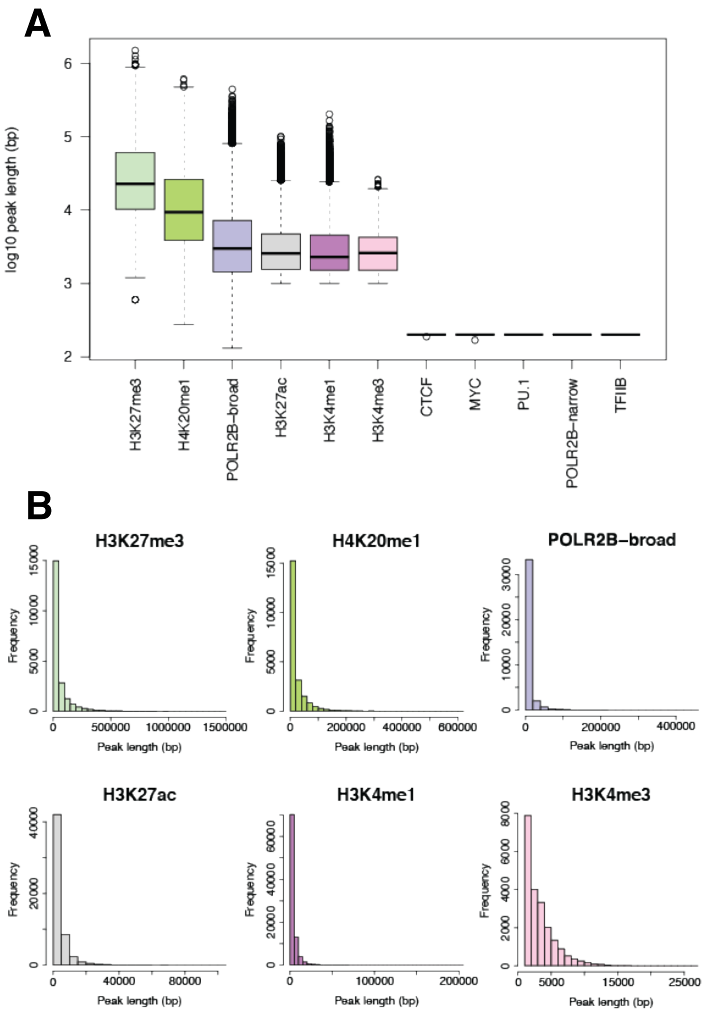


Figure B.26: Peak size distribution of **A** all ChIP-seq assays and **B** assays with variable peak lengths only. For DNA binding factors a fixed peak size of 200 bp was used. POLR2B peaks were called with both fixed and variable peak length (POLR2B-narrow and POLR2B-broad, respectively).

Figure B.27: Summary statistics of peaks called from the ChIP-seq trio metasamples.

	H3K27me3	H4K20me1	POLR2B broad	H3K27ac	H4K4me1	H3K4me3	CTCF	MYC	PU.1	POLR2B narrow	TFIIB
Mean length (bp)	58038.56	23230.78	8261.54	4219.56	4105.01	3283.32	200	200	200	200	200
Median length (bp)	22838	9350	3013	2575.5	2291	2592	200	200	200	200	200
Standard deviation (bp)	93974.3	37062.79	17082.74	5084.17	5543.65	2465.07	0.03	0.19	0	0	0
Number of peaks	21190	22128	36786	54840	90525	21167	145554	28014	57832	57241	18461
Minimal length (bp)	601	275	131	1000	1000	1000	188	168	200	200	200
Maximal length (bp)	1492283	610775	441208	100294	202718	26104	200	200	200	200	200
Genome covered (%)	39.73	16.61	9.82	7.47	12	2.25	0.94	0.18	0.37	0.37	0.12

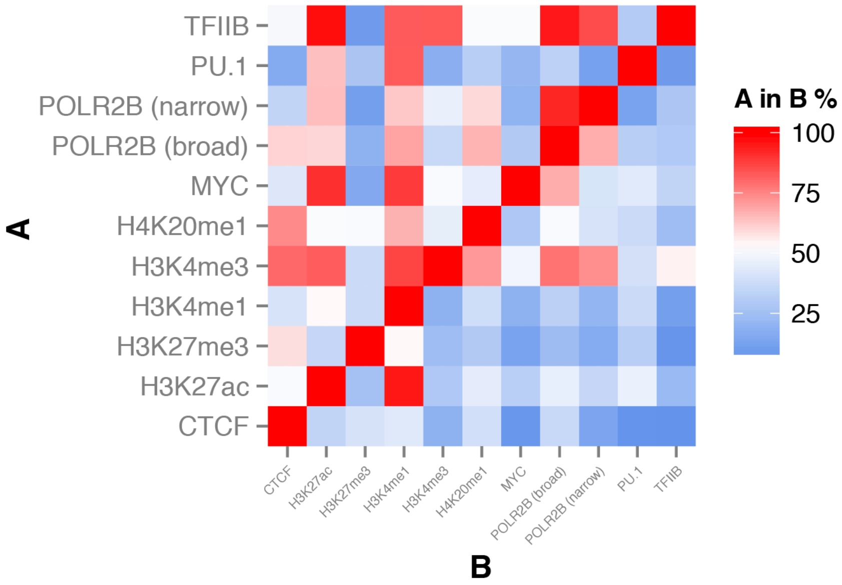


Figure B.28: Pairwise overlap of trio metasample peaks between all pairs of assays. Colors indicate the percentage of overlap between peak set A and peak set B.

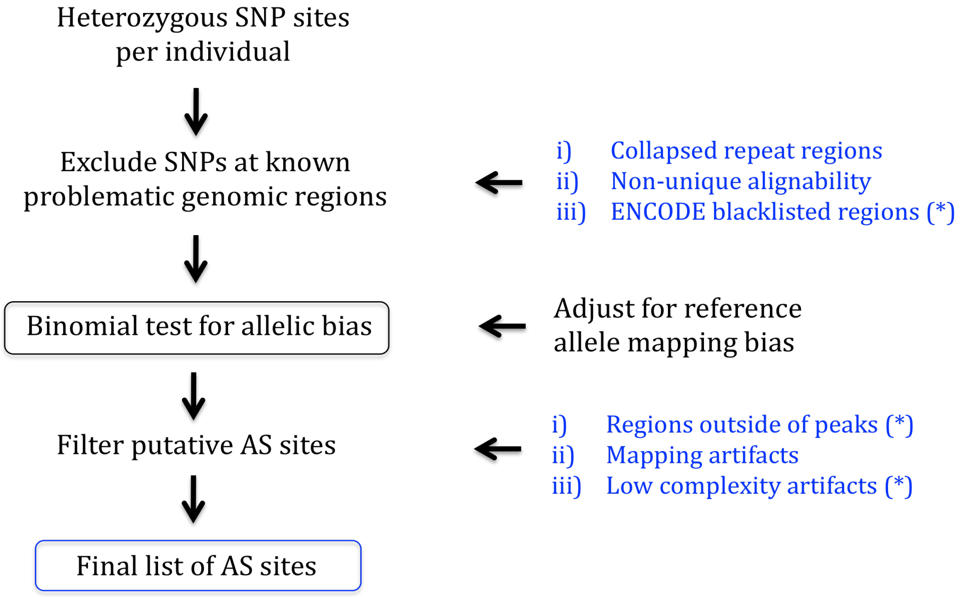

Heterozygous SNP sites
per individual
Exclude SNPs at known
problematic genomic regions
i) Collapsed repeat regions
ii) Non-unique alignability
iii) ENCODE blacklisted regions (*)
Binomial test for allelic bias
Adjust for reference
allele mapping bias
Filter putative AS sites
i) Regions outside of peaks (*)
ii) Mapping artifacts
iii) Low complexity artifacts (*)
Final list of AS sites

Figure B.29: Flow-chart of the allele-specific analysis pipeline and associated filtering steps. Steps marked with an asterisk (*) were not applied to RNA-seq.

Figure B.30: Summary of SNP sites susceptible to mapping bias in (A) population-based and (B) personalized simulations of local sequence effects. Results are presented separately for each read length (36 bp, 39 bp, and 49 bp for ChIP-seq, GRO-seq, and RNA-seq data, respectively).

A. Population-based

Sample	Biased SNPs			Biased INDELs			Total excluded		
	36 bp	*39 bp*	*49 bp*	*36 bp*	*39 bp*	*49 bp*	*36 bp*	*39 bp*	*49 bp*
1k CEU YRI MAF > 0.01	3044649	NA	1783818	665306	NA	583842	3199002	NA	1953882

B. Personalized

Sample	Biased SNPs			SNPs near biased INDELs			Total excluded		
	36 bp	*39 bp*	*49 bp*	*36 bp*	*39 bp*	*49 bp*	*36 bp*	*39 bp*	*49 bp*
NA12878	85310	60696	27155	75536	81985	81749	156288	139115	107015
NA12891	76184	53698	23696	75536	81985	81749	147602	132562	103762
NA12892	74696	51544	21048	75536	81985	81749	146707	130860	101601
NA19238	74503	50356	22604	84859	92527	93783	156174	140479	115050
NA19239	80824	55222	24890	84859	92527	93783	162115	145109	117219
NA19240	80932	55325	25001	84859	92527	93783	164219	146667	118228

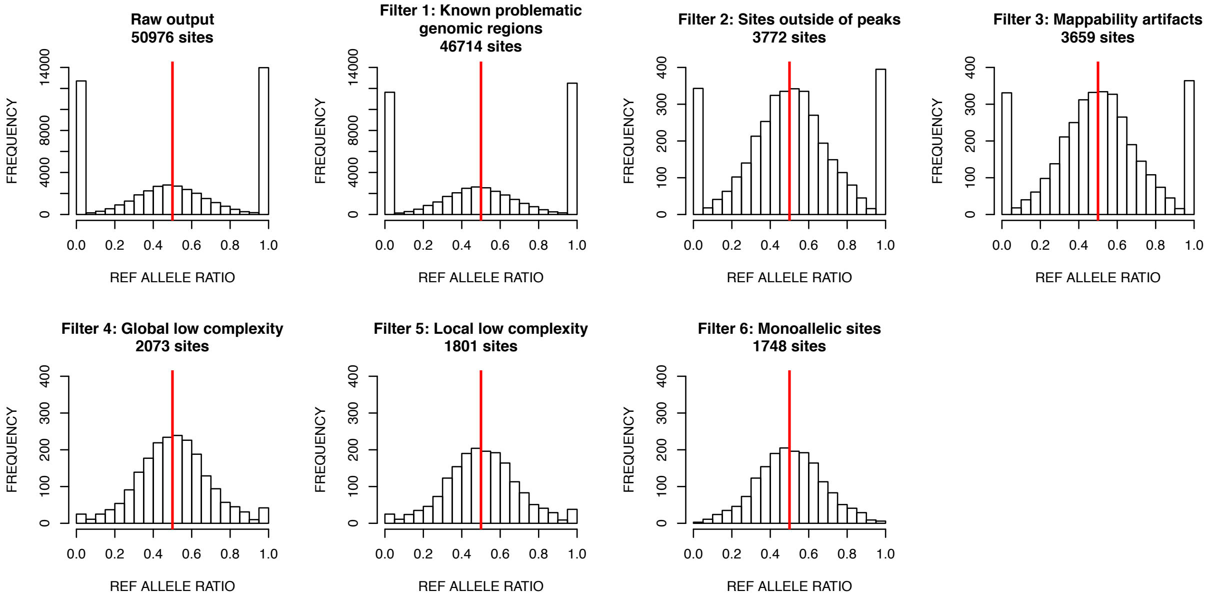


Figure B.31: Distribution of the reference allele ratio across all accessible heterozygous SNP sites after each step of filtering in the allele-specific analysis and the number of sites remaining after each step. A representative example from POL2RB-narrow GM19239 is shown.

B.4 References

48. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (Oct. 2010).

49. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

50. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (Feb. 2012).

51. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328,** 235–239 (Apr. 2010).

52. Kasowski, M. *et al.* Variation in Transcription Factor Binding Among Humans. *Science* **328,** 232–235 (2010).

56. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).

60. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322,** 1845–1848 (2008).

101. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38,** 576–589 (May 2010).

108. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25,** 2078–2079 (Aug. 2009).

156. Muller, N., Girard, P., Hacker, D. L., Jordan, M. & Wurm, F. M. Orbital shaker technology for the cultivation of mammalian cells in suspension. *Biotechnology and Bioengineering* **89,** 400–406 (2005).

157. Canella, D., Praz, V., Reina, J. H., Cousin, P. & Hernandez, N. Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. *Genome research* **20,** 710–721 (June 2010).

158. Schramm, L., Pendergrast, P. S., Sun, Y. L. & Hernandez, N. Different human TFIIIB activities direct RNA polymerase III transcription from TATA-containing and TATA-less promoters. *Genes & Development* **14,** 2650–2663 (2000).

159. O'Geen, H., Nicolet, C. M., Blahnik, K., Green, R. & Farnham, P. J. Comparison of sample preparation methods for ChIP-chip assays. *Biotechniques* **41,** 577–580 (Nov. 2006).

160. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43,** 491–+ (May 2011).

161. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297–1303 (Sept. 2010).

162. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (Feb. 2001).

163. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **26,** 589–595 (Mar. 2010).

164. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome biology* **7 Suppl 1,** S4.1–9 (2006).

165. Song, Q. & Smith, A. D. Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics (Oxford, England)* **27,** 870–871 (Mar. 2011).

166. Pickrell, J. K., Gaffney, D. J., Gilad, Y. & Pritchard, J. K. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics (Oxford, England)* **27,** 2144–2146 (Aug. 2011).

167. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (Nov. 2012).

168. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research* **22,** 1798–1812 (Sept. 2012).

169. Orphanides, G., Lagrange, T. & Reinberg, D. The general transcription factors of RNA polymerase II. *Genes & Development* **10,** 2657–2683 (Nov. 1996).

Appendix C

Population Variation and Genetic Control of Modular Chromatin Architecture in Humans

C.1 ChIP Sequencing Experiments

C.1.1 Chromatin Immunoprecipitation of RNA Polymerase II (RPB2)

ChIPs were carried out as previously described [53] with a few modifications. Chromatin extracted from 5 x 107 cross-linked cells was sonicated to an average size of 200-700 bp. Sheared chromatin was then immunoprecipitated with 7 μg per 107 cells of an anti-Rpb2 antibody (sc-67318, Santa Cruz Biotechnology). Immunoprecipitated material was recovered with 2 mg per 107 cells of pre-blocked protein-A beads (17-0780-01, GE Healthcare) and washed twice with dialysis buffer, three times with IP wash buffer [53] for buffer compositions). After reversal of crosslinking and DNA purification, 10 ng of ChIP DNA was used for ChIP-seq libraries preparation.

C.1.2 Chromatin Immunoprecipitations of PU.1 and H3K4me1

PU.1 and H3K4me1 ChIPs were carried out as previously [53]. Cells were lysed in nuclei extraction buffer (50 mM HEPES-NaOH pH 7.5, 140 mM NaCl, 1 mM EDTA pH 8.0, 10% glycerol, 0.5% NP-40, 0.25% TritonX-100) supplemented with a protease inhibitor tablet (Roche) and phosphatase inhibitors (5 mM NaF, 1 mM glycerol phosphate and 1 mM sodium orthovanadate) for 10 min at 4°C on a shaker. The isolated nuclei were then washed using washing buffer (200mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 10 mM Tris-HCl pH 8.0) supplemented with protease and phosphatase inhibitors at RT for 10 min. Washed nuclei were resuspended in sonication buffer (1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 10 mM Tris-HCl pH 8.0 and 1% SDS) containing protease and phosphatase inhibitors and the chromatin was fragmented using a Bioruptor sonicator (Diagenode) for 60 min using high amplitude and 30s ON & 30s OFF cycles to obtain 200-500 bp-sized fragments. The fragmented chromatin was then centrifuged at 17,000xg for 5 min and clear supernatant was diluted with ChIP dilution buffer (1 mM EDTA pH 8.0, 10 mM Tris-HCl pH 8.0 and 1% TritonX-100 containing protease and phosphatase inhibitors) to get chromatin equivalent to 10 X 106 cells for each IP. All IPs were performed in duplicates. BSA and ssDNA (Salmon Sperm DNA)-preblocked protein-A sepharose (80 μl/IP) beads were added to the samples and incubated for 2h to remove non-specifically binding chromatin. To the supernatant, 5 μg/IP of PU.1 antibody (Santa Cruz, Cat no: 22805X) or H3K4me1 antibody (Abcam, Cat no: ab-8895) was added to immunoprecipitate the chromatin complex at 4°C overnight. After incubation, 50 μl blocked protein-A sepharose beads were added to each sample and incubated for 90 min at 4°C to pull down the respective antibody-chromatin complexes. The beads were then washed four times with low salt wash buffer (20 mM Tris-Cl pH 8.0, 150 mM NaCl, 2 mM EDTA pH 8.0, 0.1% SDS, 1% TritonX-100) followed by two washes with high salt wash buffer (20 mM Tris-Cl pH 8.0, 500 mM NaCl, 2 mM EDTA pH 8.0, 0.1% SDS, 1% TritonX-100), lithium chloride wash buffer (10 mM Tris-Cl pH 8.0, 0.25 M LiCl, 1 mM EDTA pH 8.0, 1% NP-40, 1% sodium deoxycholate) and Tris-EDTA (TE) buffer (10 mM Tris-Cl pH 8.0, 1 mM EDTA pH 8.0). The antibody bound chromatin complexes were eluted from beads for 30 min using 200 μl of elution buffer (100 mM sodium bicarbonate and 1% SDS in milliQ water). The chromatin was then reverse-crosslinked at 65°C overnight after adding 8μl of 5 M NaCl. The DNA was then purified from the reverse-crosslinked chromatin by proteinase-K and RNase digestion followed by purification using Qiagen DNA purification columns. The purified DNA was eluted in 30 μl of Qiagen elution buffer.

C.1.3 Chromatin Immunoprecipitations of H3K4me3 and H3K27ac

ChIP was carried out largely as suggested in [53], with modifications made to automatize the procedure. Briefly, cells were lysed by addition of cell lysis buffer, then nuclei were washed and subsequently lysed using nuclei lysis buffer. Chromatin was sheared with Covaris S220 sonicator (Covaris Inc., MA, USA). Sonication efficiency was assessed by running a sample of de-crosslinked DNA on a 1.5% agarose gel. Fragmented chromatin was diluted 10-fold (5-fold in case of H3K27ac IP) in ChIP dilution buffer and immunoprecipitated using antibodies against H3K4me3 (Millipore 17-614; lot #JBC1793805) and H3K27ac (Abcam ab4729; lot #GR71158). The immunoprecipitation assays were performed on Diagenode SX-8G IP-Star Compact automated system using Auto Histone ChIP-seq kit (Diagenode S.A., Belgium). The minimum of 2 IPs of 106 cells (2x106 in case of H3K27ac) per cell line was used. Replicates were pooled following RNase A and proteinase K treatments. DNA was purified with Qiagen DNA purification kit (Qiagen N.V., Netherlands). DNA concentration was measured using Qubit apparatus (Life Technologies, CA, USA). Before proceeding with library preparation for sequencing, enrichment of the precipitated DNA was assessed by quantitative PCR.

C.1.4 Library Preparation and Sequencing

ChIP libraries were prepared with the TruSeq DNA sample prep kit (Illumina) and AR001-AR0027 indexing adapter set according to the manufacturer's recommendations. The starting amount of ChIP DNA used for library preparation ranged from 2.5 ng to 10.5 ng per sample. Library quality and average fragment size was confirmed with Bioanalyzer DNA analysis chips (25-1000 bp, Agilent). TruSeq libraries were subsequently multiplexed on Illumina HiSeq2000 lanes (three per lane, RPB2; four per lane all other assays) (read length 36 bp, single-end). A subset of all libraries was sequenced multiple times in order to improve coverage. The number of sequencing rounds for each sample as well as other experimental information is available in Table C.1A.

C.1.5 Short-Read Alignment

ChIP-seq reads (36 bp, single-end) were mapped against the hg19 build of the human reference genome supplemented with the Epstein-Barr virus (EBV) sequence with BWA 0.5.9 [107] using default parameters. If a sample had data from multiple lanes, reads were merged after mapping. We kept only uniquely mapping reads with a mapping quality (MAPQ) score of >= 10. Samtools [108] was used for general data processing throughout the project. The total number of usable reads for each assay and individual is summarized in Table C.1B. The average number of usable reads per individual is 54 +/- 10M for H3K27ac, 50 +/- 12M for H3K4me3, 134 +/- 19M for H3K4me1, 49 +/- 9M for PU.1, and 71 +/- 17M for RPB2 (mean +/- standard deviation).

C.1.6 Data Quantification

ChIP-seq peak calling was not performed in the current set of samples. Instead, we used an independently-derived peak set for each assay [53]. Briefly, mapped reads from six 1000 Genomes Project pilot individuals (two trios) were merged into a meta-sample for each assay, duplicate reads were removed, and peaks called using HOMER [101]. Obtained chromosomal peaks were filtered for collapsed repeat regions and genomic regions blacklisted by ENCODE (see [53] for details). To quantify these peaks in this study, read counts from ChIP-seq experiments were counted using HOMER within meta-peak regions across all unrelated individuals. The rational behind this approach was to (1) avoid the issue of fuzzy peak boundaries that would result from individual peak calling/merging and (2) to focus on common molecular events and QTLs given our sample size. Reads were shifted based on their estimated ChIP-fragment length and libraries were normalized by their size such that each library contains 10 million reads. Analysis in Section 6.9 was performed based on RPKM quantifications [170].

C.1.7 Data Normalization

Peak quantifications were first scaled to adjust for differences in total library size. We then applied PEER [109] to identify and regress out hidden confounding factors in each dataset. For QTL mapping we used PEER residuals that were first transformed to standard normal distribution. To estimate the best number of covariates (K) to correct for, we first ran PEER for chromosome 1 only across a range of values for K (0,1,3,5,7,10,13,15,20) and mapped cis-QTLs separately for each K [56]. We monitored the number of unique ChIP-seq QTL peaks obtained from each run (data not shown), and selected K=15 as the final number of covariates to correct for. PEER was then run with the genome-wide dataset of each assay, adding the mean to the model and using 100 iterations.

C.2 mRNA Sequencing Experiments

C.2.1 mRNA Extraction

Total mRNA was extracted from cell pellets using the standard Trizol protocol (Invitrogen). mRNA concentration was measured with the Qubit system (Invitrogen) and the quality of the samples confirmed with Agilent 2100 Bioanalyzer RNA 6000 Nano analysis chips.

C.2.2 Library Preparation and Sequencing

Libraries for mRNA-seq were prepared with the Illumina TruSeq mRNA sample preparation kit, according to manufacturer's instructions. 500 ng of total RNA was used for each library. Briefly, poly-A RNA is selected using poly-T oligo-attached magnetic beads, the mRNA is cleaved, and converted to cDNA with first strand synthesis. After RNA digestion and second DNA strand synthesis, the fragments are end repaired and ligated to the adapters containing specific primer indexes. Finally, the cDNA libraries are amplified by PCR. All samples were sequenced as part of pools with 12 libraries on the HiSeq (paired-end, read length 49 bp). Each library was sequenced twice to achieve sufficient coverage.

C.2.3 Short-Read Alignment

mRNA-seq sequence reads (paired-end 49 bp) were mapped against the hg19 build of the human reference genome supplemented with the Epstein-Barr virus (EBV) sequence with BWA 0.5.9 [107] using default parameters. We kept only uniquely mapping reads with a mapping quality (MAPQ) score of >= 10. Samtools [108] was used for general data processing throughout the project.

C.2.4 Data Quantification

mRNA-seq data was quantified based on GENCODE v15 (08/2012) gene annotations [110] and as previously described in Lappalainen et al. [56].

C.2.5 Data Normalization

All genes with more than 10% of the samples without a single overlapping read were removed from the analysis and all remaining quantified genes were normalized similarly to ChIP-seq data (section C.1.7), i.e. applying PEER with 15 covariates (K=15), adding the mean, and transforming PEER residuals to a standard normal distribution.

C.3 Analytical Methods

C.3.1 Molecular Phenotype-Phenotype Associations

To map associations between pairs of peaks, we proceeded as follows for each of the 15 possible unordered pairs of distinct molecular phenotypes (A1, A2). First, we measured inter-individual Pearson correlation between every possible pair of normalized quantification at peaks (p1, p2) such that (a) p1 and p2 belong to A1 and A2, respectively and (b) the genomic distance between p1 and p2 not exceeding 1 Mb. Note that the distances here were measured between the respective peak centers, excepted form mRNA for which we used the transcription start site (TSS). Then, we assessed how significant the correlations were different from 0 by calculating P-values using the R function cor.test and corrected them for multiple-testing by calculating the corresponding Q-values using the qvalue package R (Dabney A and Storey JD) in R. Finally, we could both estimate the percentage of truly associated pairs of peaks among the tests performed from the Q-value's π_1 estimate and extract all significant associations at a 0.001 false discovery rate [171].

C.3.2 Variable Chromatin Modules (VCMs)

To build VCMs from the discovered significant associations, we developed an approach based on graph theory. We first consider as nodes all peaks significantly associated and as edges all significant associations between peaks. Then, we defined as VCMs any two nodes for which there is a path (i.e. a sequence of edges) that links them together. Conversely, two nodes belong to two distinct VCMs as soon as there is not a single path linking them. In practice, we implemented an algorithm in R using the igraph package (`http://igraph.org/`) that performs this graph reconstruction by initially assigns a distinct VCM for every single node and then iteratively merges VCMs connected by edges. We derived VCM activity levels using principal component analysis (PCS) on normalized peak intensities using the prcomp function in R. The first and second VCM PC was further transformed to a standard normal distribution for QTL mapping (section C.3.3) and gene expression-VCM associations.

C.3.3 Mapping Chromatin and Expression Quantitative Trait Loci (QTL)

We mapped cis-acting QTLs within 250kb and 1 Mb of (a) the TSS for RNA or (b) the peak center for histone modifications and transcription factor binding sites. More specifically, we regressed genotypes linearly against peak quantifications for all variant site / phenotype pairs when the genomic distance between them was smaller than 1 Mb. Then, we stored for each peak the best association we found as a putative QTL. At this point, we had to correct for two distinct levels of multi-testing in order to determine whole genome significance of the putative QTLs: first, multiple variants sites are tested for association with a single peak and second, multiple peaks are tested genome wide. To correct for the first multiple-testing problem, we devised a permutation strategy in which we keep permuting quantifications for a peak until we either (a) find 100 more extreme association P-values than the observed (i.e. nominal) one or (b) reach a number of 100,000 permutations in total. Note that the genotype data stays unchanged throughout permutations in order to preserve the local LD structure in the tested cis-window. From this, we can then derive a corrected P-value for each peak that empirically quantifies how frequently a more extremely associated variant can be found via permutations; that is how likely the putative QTL we found can be obtained by chance. Then, we accounted for multiple testing at the level of peaks by estimating the minimum false discovery rate (FDR) at which each empirical P-value may be found significant; the Q-value. To compute them, we used the estimation method implemented in the qvalue package1. Once a Q-value was obtained for each peak, we can extract all QTLs at X% FDR by only keeping QTLs with a Q-value below X%. In practice, we implemented this QTL mapping strategy such that we can simultaneously test multiple cis-window sizes (10 kb, 20 kb, 50 kb, 100 kb, 200 kb, 500 kb, 1 Mb, 2 Mb) and FDRs in a single association run, thus choosing a good trade-off between both. We chose 10% FDR and 500 kb cis-windows which gave us a relatively high number of QTLs that can be located relatively far away from the phenotype location (Figure C.4A-C). This approach was implemented in the software package FastQTL (Ongen et al, 2015) available on `http://fastqtl.sourceforge.net/`.

C.3.4 Mapping Allele-Specific Chromatin and Gene Expression Effects

Allele-specific (AS) analysis was performed using genotypes from the 1000 Genomes Phase1 release v3, available for a subset of the samples (N = 34/47). The original VCF file was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ and the subset of individuals extracted with VCFtools, keeping only sites with a minor allele count >1 (-mac 1). This yielded a total of N = 9,609,399 SNP sites for analysis. Allele-specific analysis was performed at heterozygous SNP positions of each individual (average 2.1M sites) using a modified binomial test and accounting for multiple major sources of technical bias, such as reference allele mapping bias, clonal reads and non-unique mappability of reads as described previously [53, 56, 112]. To adjust the expected allele ratios in the binomial test for each site, the reference allele mapping bias was estimated separately for each pair of alleles and mapping quality bin, requiring a minimum of 250 sites per category (if less, a global average was used). A minimum of 10 reads per site were required and only bi-allelic sites overlapping peaks (ChIP-seq) and exons (mRNA-seq) were used for analysis. Allele-specific analysis was additionally used as a QC step to identify putative sample swaps or contaminations. We monitored the proportion of the heterozygous sites that appeared heterozygous also on the level of RNA/ChIP. Samples showing an unusually low proportion of heterozygosity in the RNA were flagged as possible swaps/contaminants. For the samples not included in the 1000 Genomes phase1 v3 release we used genotypes from the GEUVADIS project to perform AS analysis for QC purposes.

C.3.5 Estimating Pairwise Sharing of QTLs Between Molecular Phenotypes

We designed a method to estimate how QTLs are shared between pairs of assays based on the π_1 (1-π_0) statistic described in [86]. Specifically, we proceed in 4 steps: (1) we take for each feature (peak/gene) with a significant QTL from assay1, (2) we find the closest corresponding feature in assay2, (3) we compute the association P-value between the QTL and the assay2 feature we selected, and (4) we measure enrichment of low P-values in the resulting distribution via π_1 estimation.

C.3.6 Detecting Multiple Effects of QTLs

We counted (a) the number of distinct features (peaks/genes) and (b) the number of distinct molecular phenotypes affected by a given QTL by using the following method. We first measure association between the QTL and all features across all assays located within +/- 250 kb of the QTL. Then, we adjust the resulting association P-values for multiple tests using Bonferroni's method; we divide all P-values by the number of tested features. Finally, we consider features as associated when the their corrected P-value is below the 0.05 threshold. This gives us the number of features affected by the QTL and therefore by looking at which assay they belong to, the number of distinct molecular phenotypes affected.

C.3.7 Estimating Enrichment of QTLs Falling Within Features

To measure fold enrichment of assay1 QTLs located more likely than expected within features defined by assay2, we developed an approach that aims to correct for the fact that assay1 QTLs, assay1 features and assay2 features are not independently distributed along the genome which can therefore lead to false enrichment with nave methods. More specifically, we proceeded in five steps:

1. We ranked all association p-values within each cis-window for assay1 features that have a significant QTL.
2. We built a genomic 'segment' of associations for each cis-window w and each rank r with left and right boundaries matching the left- and right-most variant sites with a p-value of rank r. A segment basically contains only variants in full LD ($r^2 = 1$).
3. For each possible rank r, we counted the number of times the segments of rank r overlap assay2 features across all cis-windows. We find that after the 50th rank, the overlap counts converge around a stable value and we use the median from rank 50 to 200 as a null overlap count for the next step (Figure C.4F)
4. We estimate the odd ratios and significance of the enrichment by performing a fisher test between the rank 1 and null overlap counts. We consider as significant any enrichment with a Bonferroni corrected P value (corrected by the number of cells in the heatmap) below 0.05. All insignificant enrichments are not displayed in the heatmaps.

C.3.8 Modeling Causality with Bayesian Networks

To assess the most likely transmission model of genetic effects onto chromatin and gene expression, we constructed three-node graphs (triplets) for all VCMs that show significant correlations with gene expression. If at least one of the two was reported as a QTL, we combined the QTL variant, the VCM and the gene quantifications to create a triplet. If both the VCM and the gene were reported as QTLs with different variants, one variant was picked by chance. Posterior probabilities for all 25 possible directed graph models with 3 nodes were computed using the bnlearn package (v3.5) in R (v3.1.0). After a first assessment, posterior probabilities for three biologically relevant models were extracted, scaled to sum up to one and the most likely model per triplet was defined (posterior probability >= 0.9). For the PU.1 motif disruption analysis, triplets were constructed in a similar way, but based on significantly correlated single molecular phenotype pairs involving PU.1. Whenever at least one of the two phenotypes was reported as a QTL, a triplet was constructed consisting of the quantifications for the two phenotypes plus the QTL variant genotype. If both molecular phenotypes were reported as QTLs with different variants, one variant was picked by chance. For this analysis only triplets where the variant showed a significant association (P <0.05) with both molecular phenotypes were considered. Posterior probabilities for all 25 possible directed graph models with 3 nodes were computed using the bnlearn package (v3.5) for R (v3.1.0) and posterior probabilities for three biologically relevant models were extracted, scaled to sum up to one and the most likely model per triplet was defined. Finally the resulting triplet models were grouped according to whether a PU.1 binding site present in the reference genome was disrupted by the PU.1 QTL variant of the triplet or not. Reference PU.1 binding sites were predicted with the software fimo in the MEME suite (`http://meme.nbcr.net/meme/`) and using the PU.1 PWM described in [168]. Only H3K27ac and H3K4me1 showed enough sharing of QTLs with PU.1 for these analyses.

C.3.9 Functional Annotation of VCMs and VCM-Associated Genes

We used the online service GREAT v2.0.2 (`http://bejerano.stanford.edu/great/public/html/,defaultoptions`) to predict over-represented pathways and biological processes for VCM domains. Functional annotation of VCM-associated genes was performed using the online service ConsensusPathDB-human (`http://cpdb.molgen.mpg.de/`, default options) using the over-representation analysis module and gene ontology categories (BP level 2).

C.3.10 Estimating Enrichment of QTLs with GWAS Hits

All GWAS variants used in this study were obtained from the NHGRI GWAS Catalog (`http://www.genome.gov/gwastudies/`, Dec 8, 2014) and filtered out sites on the X and Y chromosome. Mappings between GWAS studies to Experimental Factor Ontology (EFO) terms were obtained from `http://www.ebi.ac.uk/fgpt/gwas/` using PUBMED IDs. For each of the eight QTL list we considered in our analysis:

1. We counted the number of QTLs overlapping GWAS hits, considering that two variants overlap as soon as they are in relatively strong LD (r^2 ¿= 0.5). This constitutes our observed overlap.

2. We generated 1,000 null sets of variants of the same size matched for frequency and distance to target molecular phenotype, making sure that the association between the variant and the target phenotype is not significant (P-value >0.1). Then, we used these sets to derive the null distribution of overlap with the GWAS hits using the same approach as described in step (i).

3. Given the observed and null overlaps obtained in step (i) and (ii), respectively, we calculated odds ratios and a two-tailed empirical P values of enrichment/depletion by looking at the position of the observed overlap within the null distribution. We declare a QTL list as significantly enriched/depleted with GWAS hits if the P value is below 0.05.

C.3.11 Enrichment of Molecular Associations Within Contact Domains

High-resolution chromosomal contact domains were obtained for LCLs from [88]. Overlapping contact domains were merged and the resulting contact domain sizes ranged from 65 kb to 2.7 Mb (median: 300kb). We then calculated the probability of VCM peak pairs occurring within the same as opposed to two different contact domains using logistic regression models (glm method as implemented in R). Distance-matched, non-significant molecular associations (nominal P-value >0.1) served as an expected null. Association status (vcm/null) and peak-to-peak distance were used as variables in the logistic regression model and location within/between contact domains as the binary response variable. Peak-to-peak distance was used as a variable to account for the fact that short-range molecular associations are more likely to be embedded within the same contact domain than long-range associations. Molecular associations whereby one or both peaks were not embedded within contact domains were excluded from the analyses in order to avoid boundary effects and unrecognized contact domains, respectively.

C.3.12 Haplotypic Coordination in Allelic Signals Between Cis-Regulatory Elements

AS effects were measured at phased heterozygous SNPs. We removed AS sites with monoallelic signals and only considered significant allelic biases at 1% FDR (calculated per assay). We collected significant AS sites at non-overlapping cis-regulatory elements of molecular associations defined based on VCMs and random controls (as defined in 6.11). We only considered associations whereby both cis-regulatory elements exhibited significant allelic effects. If several AS sites were located within one or both regions, then we selected a random pair of sites in order to avoid overrepresentation of cis-regulatory elements in our analysis. We used logistic regression models to test if regions defined by molecular associations exhibit higher levels of concordance in allelic directions than random control regions ($P<0.05$) using association status (VCM/null) and distance between cis-regulatory elements as predictor variables and coordination in allelic direction as the binary response variable (i.e. two cis-regulatory elements are defined as showing coordinated allelic effects if biases occur on the same haplotype).

C.3.13 Identification of Collaborative TFs at Variable Cis-Regulatory Elements

We intersected binding of 56 TFs (based on NA12878) (ENCODE Consortium, 2012) with non-overlapping covariable cis-regulatory elements (mid point +/- 1 kb) and tested for collaborative binding using Fisher exact tests. Non-significant (P-value >0.1), distance-matched pairs of cis-regulatory elements served as a null set. We only considered TF pairs that passed the Bonferroni corrected P-value threshold of 5% (nominal Fisher exact test P-value cutoff: 0.05/(0.5x56x56)=3.2e-5 based on 56 tested TFs). We also assessed if single TFs were specifically enriched around cis-regulatory elements (+/- 1 kb) that are involved in long-range associations using logistic regression models that take distance between cis-regulatory elements into account (P-value <0.05/56).

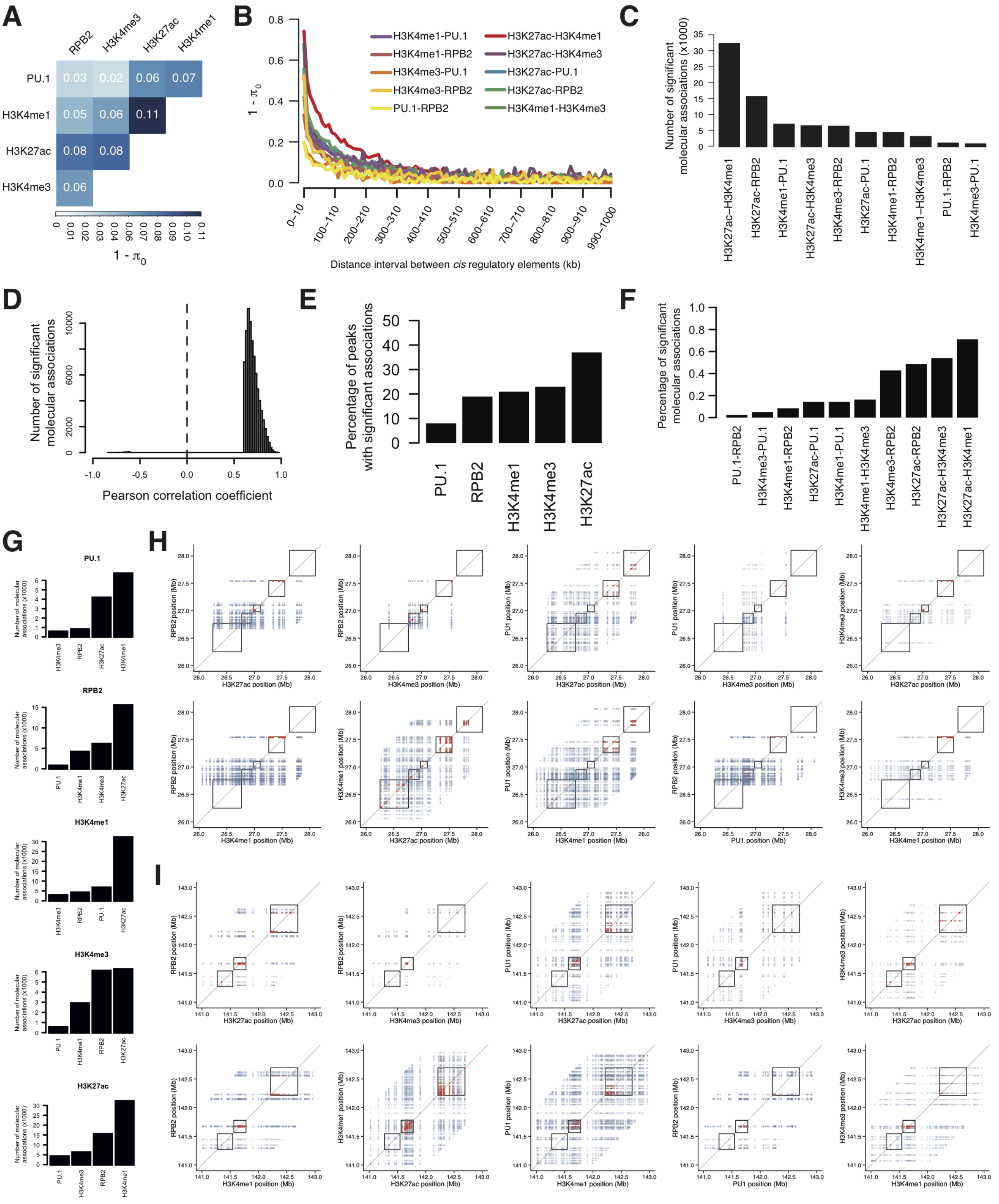


Figure C.1: **Characteristic of Molecular Phenotype Associations. A** Genome-wide enrichment of significant associations between molecular phenotypes (estimated using π_1). **B** Proportions of significant associations (π_1) between molecular phenotypes at spatially separated regulatory regions. Π_1 were estimated for all possible pairs of regulatory regions that are within 10 kb intervals at a specified distance (e.g. 100-110 kb). The center of a regulatory region was hereby considered as the reference position. **C** Total number of significant molecular phenotype associations. **D** Pearson correlation coefficient of all significant associations among molecular phenotypes. **E** Percentage of putative regulatory regions with significant molecular associations. **F** Percentage of significant associations among all tested phenotype-phenotype pairs. **G** Number of significant associations per molecular mark. **H-I** Pairwise molecular associations for all combinations of molecular mark (HMs, PU.1, RPB2) at two genomic regions: chr21:26,000,000-28,000,000 (H) and chr4:141,000,000-143,000,000 (I) (chromosomal contact domains from [88] are shown with black boxes). Related to Figure 3.2.

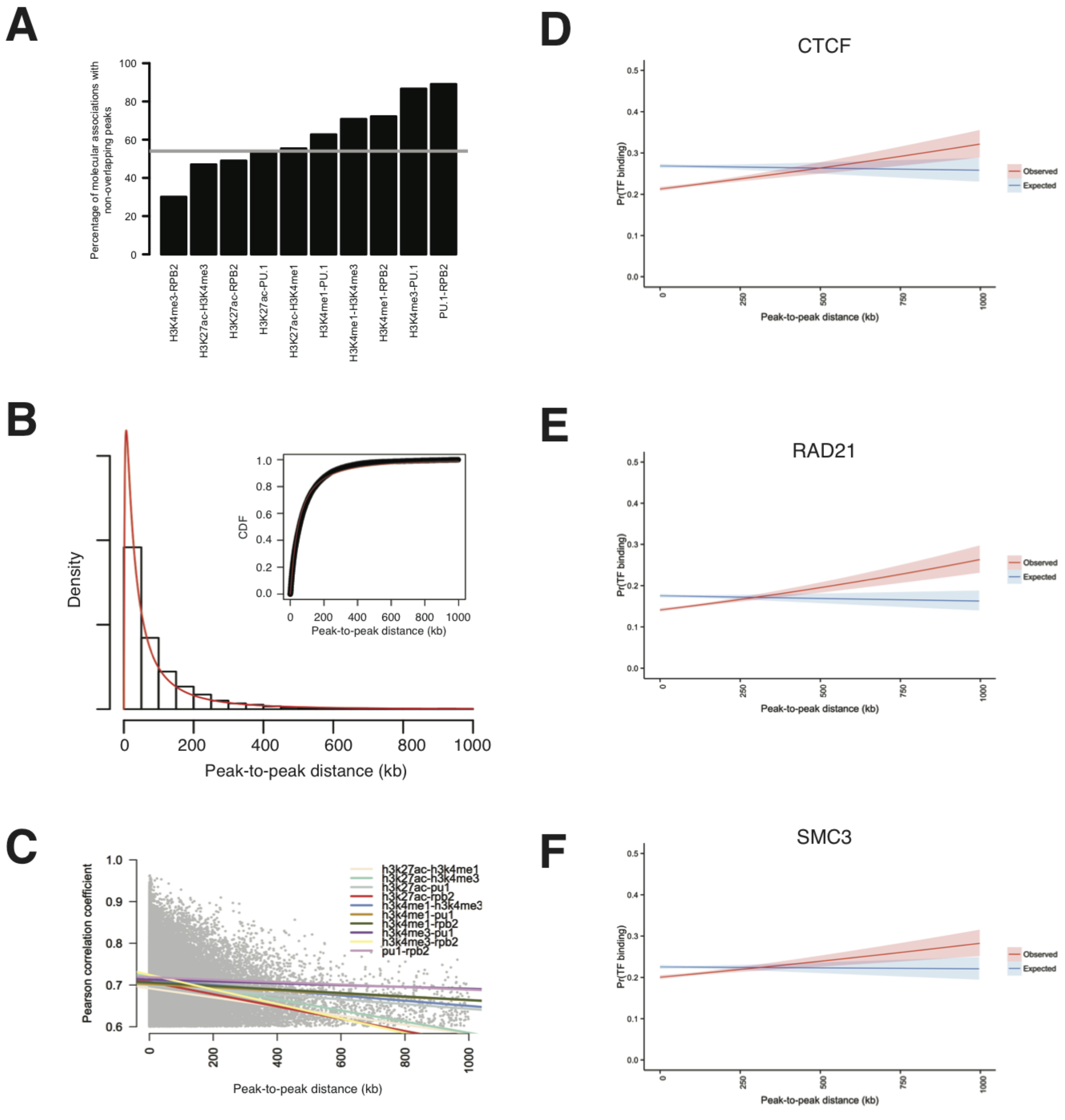


Figure C.2: **Long-Range Molecular Associations.** **A** Percentage of molecular associations between non-overlapping cis-regulatory elements. The grey line marks the overall percentage of non-overlapping associations. **B** Distance distribution between non-overlapping covariable cis-regulatory elements and best log-normal fit (logmean=10.72, logsd=1.38). **C** Correlation between association strength (Pearson's *r*) and distance between associated molecular phenotypes. Each dot represents a significant correlation pair across all tested molecular phenotype associations. **D-F** Enrichment of CTCF, SMC3, and RAD21 DNA binding events at sites of molecular association (peak center +/- 1 kb). Red line, significant molecular associations; blue line, distance-matched random control region pairs (see Methods). Logistic regression-based models indicate that all three TFs are significantly enriched at long-range molecular associations vs short-range associations. Related to Figures 3.2 and 3.3.

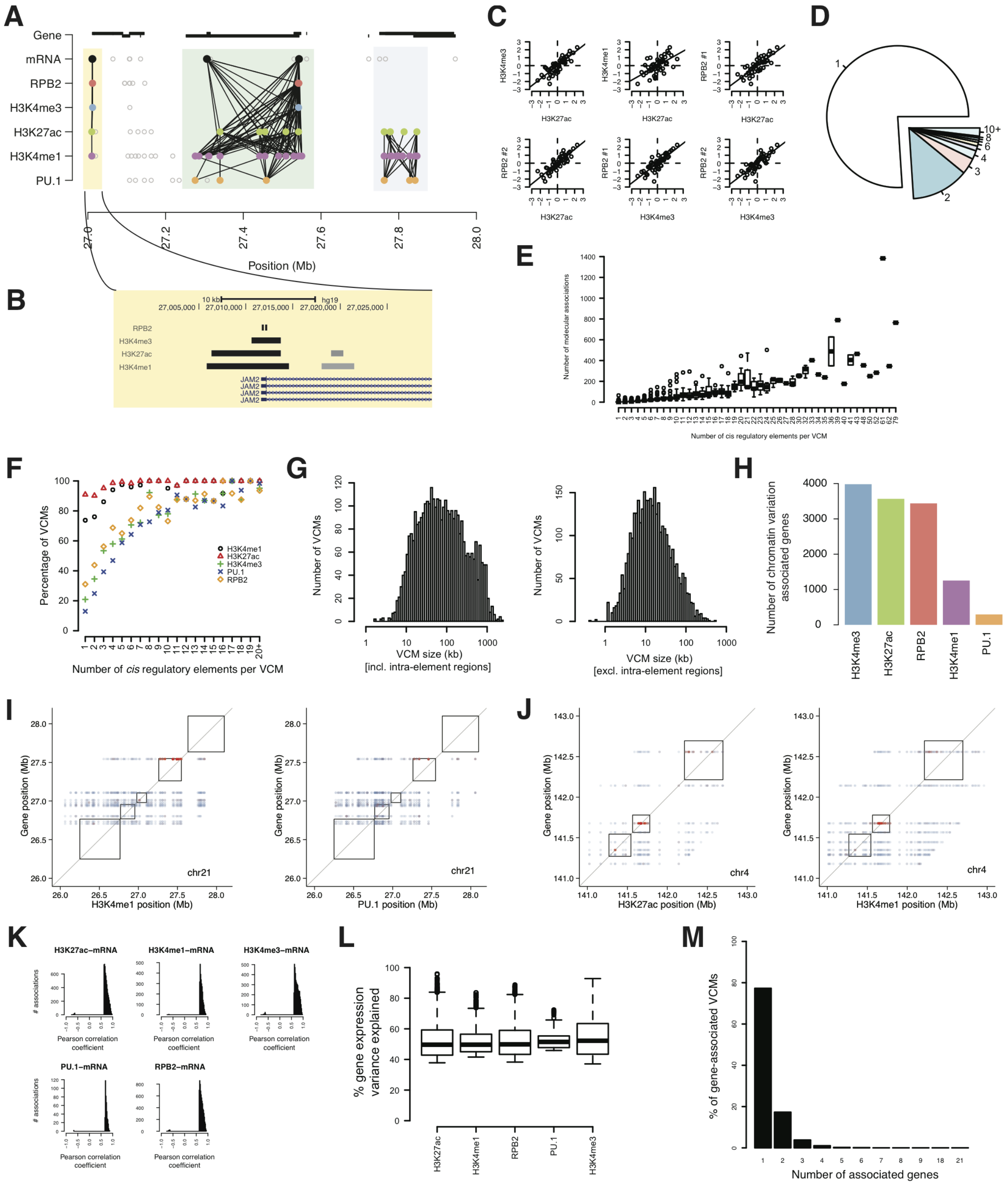


Figure C.3: **Characteristic of a Variable Chromatin Module (VCM) and its Co-association with Gene Expression. A-C** Examples of VCMs on chromosome 21 (marked by colored areas). Circles indicate the center positions of histone modified and TF-bound regions. Filled and open circles indicate molecular phenotypes with significant (FDR 0.1%) and non-significant associations, respectively. Lines connecting filled circles indicate significant associations. Detailed genomic view of VCM domains around the *JAM2* gene promoter (**B**) together with all significant pairwise molecular associations (**C**). **D** Number of cis-regulatory elements that define

VCMs. VCMs with 10 or more domains were grouped together. **E** Relationship between VCM size (number of cis-regulatory elements) and total number of molecular associations per VCM. **F** Relationship between number of cis-regulatory elements per VCM and phenotypic composition. For example, 20% of all VCMs defined by a single cis-regulatory elements contain the promoter mark H3K4me3. **G** Size distribution of cis-regulatory elements that are part of multi-element VCMs with and without intra-element regions **H** Number of significant associations between TFs/HMs and gene expression (FDR 0.1%). **I-J** Pairwise molecular associations are shown for selected gene-HM pairs at two genomic regions: **I** chr21:26-28Mb; **J** chr4:141-143Mb. **K** Pearson correlation coefficient of significant TF- and HM-gene expression associations. **L** Gene expression variance explained by TF DNA binding and HM variation. **M** Number of genes being associated with the same VCM. Related to Figure 3.4.

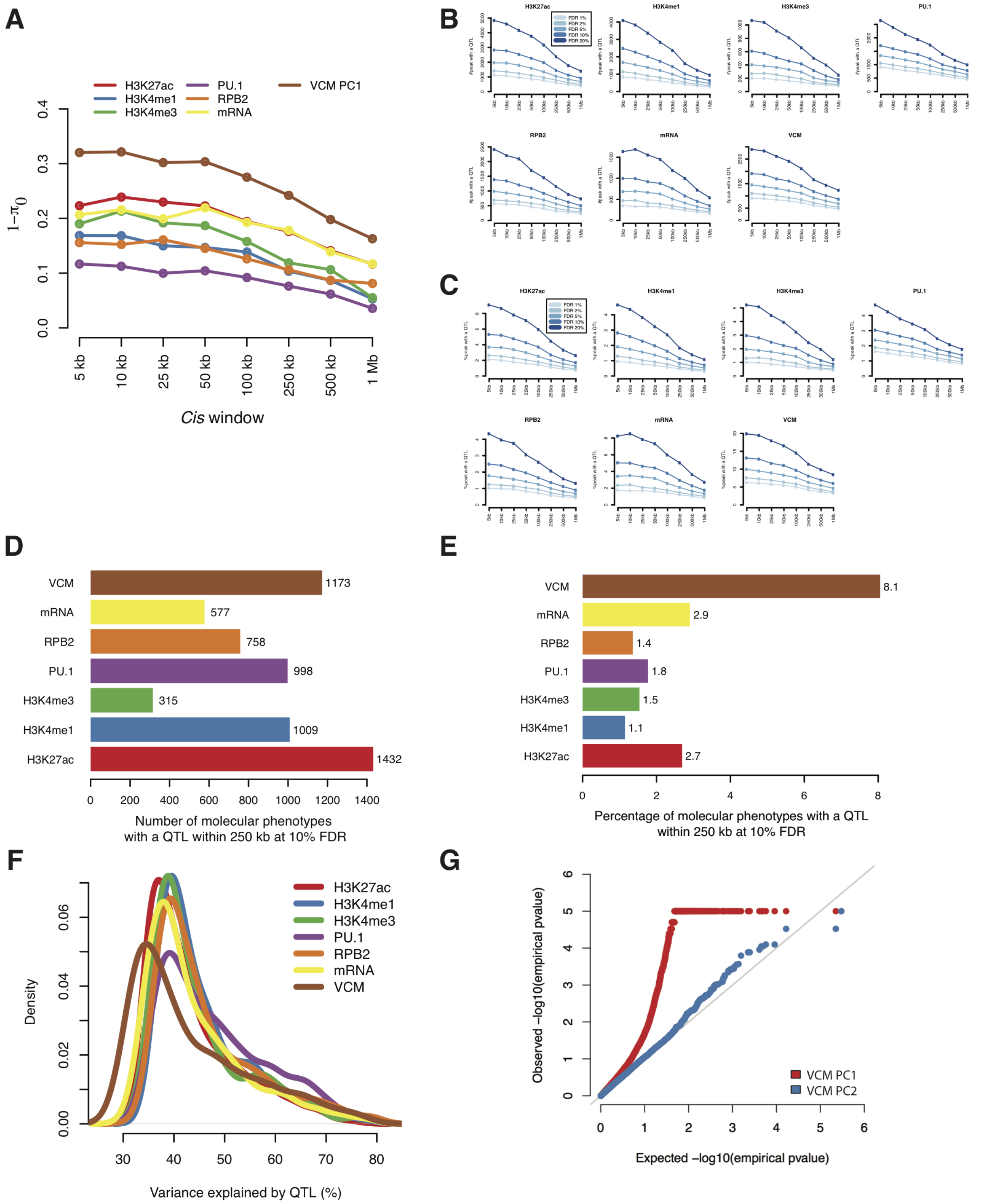


Figure C.4: **TF, HM, Expression, and VCM Quantitative Trait Loci (QTL). A** Enrichment for genetic associations per assay as a function of the cis-window size. We measured for each assay and cis-window combination the π_1 statistics [86] for the corresponding set of P-values. **B** Number of significant QTLs per assay as a function of FDR and cis-window size. **C** Percentage of molecular phenotypes per assay with a significant QTL as a function of the FDR and cis-window size. **D** Number of QTLs per assay at 10% FDR in 500 kb cis-windows (i.e., 2 x 250 kb). **E** Percentage of QTLs per assay at 10% FDR in 500 kb cis-windows (i.e., 2 x 250 kb). **F** Density distribution of variance explained by QTLs (FDR 10%, 500 kb cis-window). **G** QQ-plot of QTL association with the first and second VCM principal components. Related to Figure 3.5.

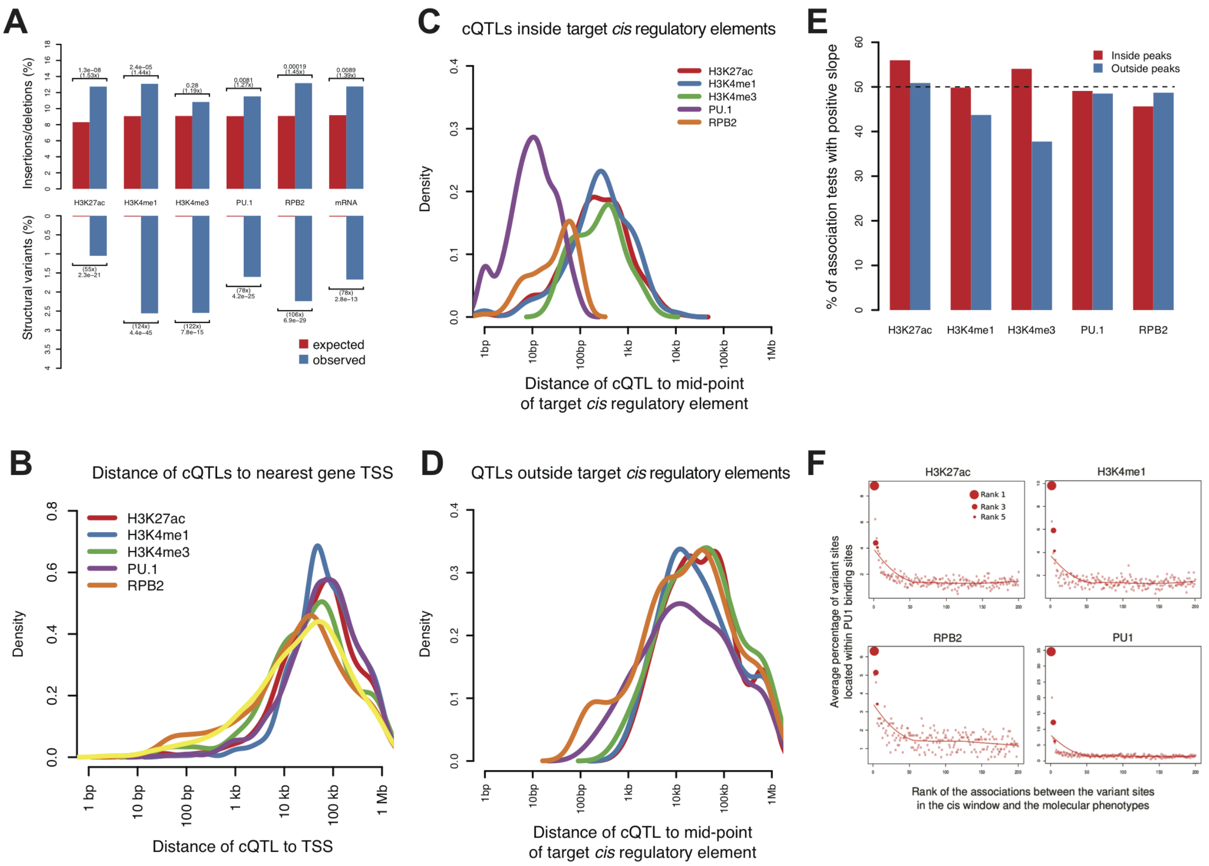


Figure C.5: **Characteristics of cQTLs A** Contribution of short insertions/deletions (top panel) and structural variants (bottom panel) to tf-, hm-, and eQTLs. Expected and observed proportions are shown in red and blue, respectively. For each assay / variant type combination, the fold enrichment and its significance is indicated above (top panel) or under (bottom panel) the corresponding bar. **B** Genomic distance density distribution between QTLs and the closest transcription start site (TSS). We plotted the measured distance on the $\log_{10}$ scale for each of the molecular assays. **C-D** Genomic distance density distribution between QTLs and their associated peaks. We plotted the measured distance on the log scale for each of the molecular assays and for all QTLs inside their associated non-coding regions (**C**) and all QTLs outside their associated non-coding regions (**D**). **E** Percentage of QTLs for which the cis-association test has a positive slope (i.e. positive regression β coefficient). A positive slope indicates higher alternative allele counts and thus higher quantification levels of the associated molecular phenotypes. The results are stratified for QTLs falling within (in red) or outside peaks (in blue). If there would have been a general mapping bias, we would have observed more negative cis-associations inside vs. outside target non-coding regions. **F** Frequencies at which variants overlap various types of non-coding regions as a function of their cis-association rank. Related to Figures 3.5 and 3.6.

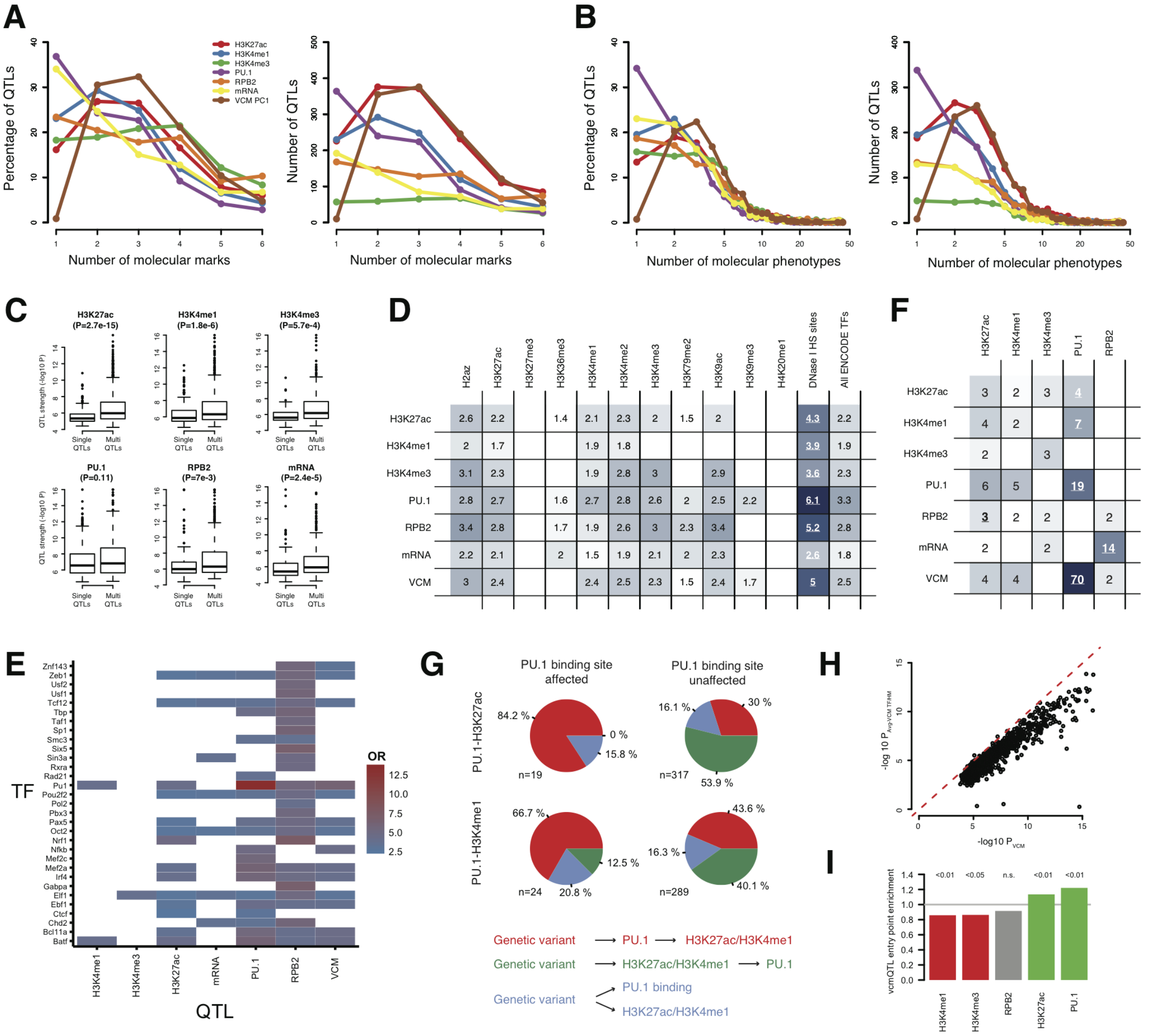


Figure C.6: **Characteristics of Shared QTLs for TF DNA Binding, HM, VCM States, and Gene expression.** **A-B** Percentage and number of distinct molecular marks (PU.1, RPB2, H3K4me1, H3K4me3, H3K27ac, mRNA) (**A**) and molecular phenotypes (TF-binding, HM, gene expression) (**B**) being affected by the same tf-, hm-, or eQTL. **C** Association strength distribution for isolated vs non-isolated QTLs. **D-E** Enrichment of cQTLs, eQTLs, and vcmQTLs within functional regions defined by ENCODE in LCLs (NA12878). We measured how often QTLs (rows) were located in functional regions (columns), estimated how often this occurred by chance, calculated the fold-change between both quantities, and corrected for multiple testing. High and low enrichment values are shown in dark and light blue, respectively. The ”ENCODE TFs” track regroups all ENCODE TFs into a single track. **F** Enrichment of cQTL, vcmQTL, and eQTL strength inside molecularly annotated non-coding regions. We calculated the fold-change in median P-values of cis-association for QTLs (rows) falling inside and outside non-coding regions (columns). For example, vcmQTLs falling inside PU.1-bound regions have a median

P-value that is 70 times smaller than those falling outside PU.1-bound regions. Dark and light blue show large and small fold changes in the median P-values, respectively. **G** Inference of causal relationships between PU.1 and H3K4me1/H3K27ac. The frequency of the most likely causal model is shown for instances where QTL variants affect reference PU.1 binding sites (left column) and instances where QTLs fall outside of reference PU.1 binding sites (right column), respectively. **H** Comparison between vcmQTL strength and average association strength between vcmQTL variants and VCM molecular events. The average cQTL strength scales linearly with the vcmQTL strength (r=0.93, P<2.2e-16), however, one order of magnitude weaker. **I** Enrichment of molecular phenotypes being entry events for vcmQTL variants. Association P-values between vcmQTL variants and each VCM member were calculated and molecular phenotypes with the smallest P-value were defined as entry events. A null distribution was estimated by randomly selecting a molecular phenotype within each VCM as the entry event (n=10,000). The fold enrichment defines how often a molecular mark was observed as the entry phenotype over random permutation. Significance is defined by empirical P-value. Related to Figures 3.6 and 3.7.

Table C.1: **Characteristics of ChIP-Seq Experiments.** **A** Number of IPs, library input DNA concentration, library size (in bp), and number of sequenced lanes per experiment. **B** Number of usable reads per ChIP assay and individual.

This supplementary excel table can be downloaded under:
`http://www.sciencedirect.com/science/article/pii/S0092867415009770`

Table C.2: **Functional Enrichment of VCMs and VCM-associated Genes.** **A-B** Functional enrichment (Pathway Commons, MSigDB Pathway, Gene Ontology Biological Processes) of single-domain VCMs **A** and multi-domain VCMs **B**. **C** Enrichment of Gene Ontology Biological Processes for VCM-associated genes. Related to Figures 3.4 and 3.5.

This supplementary excel table can be downloaded under:
`http://www.sciencedirect.com/science/article/pii/S0092867415009770`

C.4 References

53. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342,** 744–747 (Nov. 2013).

56. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).

86. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 9440–9445 (Aug. 2003).

88. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (Dec. 2014).

101. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38,** 576–589 (May 2010).

107. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25,** 1754–1760 (July 2009).

108. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25,** 2078–2079 (Aug. 2009).

109. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS computational biology* **6,** e1000770 (May 2010).

110. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22,** 1760–1774 (Sept. 2012).

112. Waszak, S. M. *et al.* Identification and removal of low-complexity sites in allele-specific analysis of ChIP-seq data. *Bioinformatics (Oxford, England)* **30,** 165–171 (Jan. 2014).

168. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research* **22,** 1798–1812 (Sept. 2012).

170. Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (July 2008).

171. Ayroles, J. F. *et al.* Systems genetics of complex traits in Drosophila melanogaster. *Nature genetics* **41,** 299–307 (Mar. 2009).