Serveur Académique Lausannois SERVAL serval.unil.ch

Author Manuscript

Faculty of Biology and Medicine Publication

This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Published in final edited form as:

Title: Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription.

Authors: Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI, Yurovsky A, Lappalainen T, Romano-Palumbo L, Planchon A, Bielser D, Bryois J, Padioleau I, Udin G, Thurnheer S, Hacker D, Core LJ, Lis JT, Hernandez N, Reymond A, Deplancke B, Dermitzakis ET

Journal: Science (New York, N.Y.)

Year: 2013 Nov 8

Issue: 342

Volume: 6159

Pages: 744-7

DOI: 10.1126/science.1242463

In the absence of a copyright statement, users should assume that standard copyright protection applies, unless the article contains an explicit statement to the contrary. In case of doubt, contact the journal publisher to verify the copyright status of an article.

serval
serveur académique lausannois

UNIL | Université de Lausanne
Faculty of Biology and Medicine

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2017 July 10.

Published in final edited form as:

Science. 2013 November 08; 342(6159): 744–747. doi:10.1126/science.1242463.

Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription

Helena Kilpinen[1,2,3,†], **Sebastian M Waszak**[2,4,†], **Andreas R Gschwind**[5,2,†], **Sunil K Raghav**[4], **Robert M Witwicki**[5], **Andrea Orioli**[5], **Eugenia Migliavacca**[5,2], **Michaël Wiederkehr**[5], **Maria Gutierrez-Arcelus**[1,2,3], **Nikolaos Panousis**[1,2,3], **Alisa Yurovsky**[1,2,3], **Tuuli Lappalainen**[1,2,3], **Luciana Romano-Palumbo**[1], **Alexandra Planchon**[1], **Deborah Bielser**[1], **Julien Bryois**[1,2,3], **Ismael Padioleau**[1,2,3], **Gilles Udin**[4], **Sarah Thurnheer**[6], **David Hacker**[6], **Leighton J Core**[7], **John T Lis**[7], **Nouria Hernandez**[5,*], **Alexandre Reymond**[5,*], **Bart Deplancke**[2,4,*], and **Emmanouil T Dermitzakis**[1,2,3,8,*]

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland [3]Institute of Genetics and Genomics in Geneva, University of Geneva, Geneva, Switzerland [4]Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland [5]Center for Integrative Genomics, Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland [6]Protein Expression Core Facility, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland [7]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA [8]Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Saudi Arabia

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.

[*]Correspondence to: emmanouil.dermitzakis@unige.ch (E.T.D); bart.deplancke@epfl.ch (B.D.), Alexandre.Reymond@unil.ch (A.R.); Nouria.Hernandez@unil.ch (N.H.).

[†]These authors contributed equally to this work.

Supplementary Materials:

Materials and Methods

Figures S1–S31

Tables S1–S2

References 24–39

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Functional genomic elements have been linked to specific chromatin signatures in different cell types (1), illustrating control of transcriptional processes through multiple layers of genome organization. While allele-specific gene expression is widespread (2), 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) (fig. S1) in two parent-offspring trios (3). 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 (fig. S2), and examined available DNaseI-seq and CTCF ChIP-seq data (4). All 14 individuals were additionally profiled for messenger-RNA (mRNA) expression (5). Clustering of the molecular phenotypes along promoters and enhancers was consistent with published reports (1) (figs. S3–S5).

We identified sites of allele-specific (AS) TF binding, hPTM, and transcription for all assays (5), ranging from 11–12% for TFs (4, 6) to 6–30% for hPTMs at heterozygous sites accessible for the analysis (median across all individuals) (Fig. 1A, fig. S6). Notably, in the two trios, fewer AS effects were observed in mRNA (mRNA-seq, 5%) than in nascent transcripts (GRO-seq, 27–28%) (5), likely reflecting post-transcriptional modifications.

Multiple heterozygous SNPs overlapping regions of TF activity showed high consistency in allelic direction within individuals (Fig. 1B, fig. S7A, S7B). AS consistency in nascent transcription and histone modifications was high up to several kb and decreased with distance (logistic regression, $P < 0.05$, fig. S7C). 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 (fig. S8). We also analyzed all accessible heterozygous SNPs overlapping known eQTLs from the 1000 genomes phase1 populations (5, 7) 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 (fig. S9), 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 (8), but for sequence-specific TFs it is possible to assess whether the observed AS effects are due to motif-disrupting variants (fig. S10). 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 (fig. S11A, S11B for PU.1, fig. S12A, S12B 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 (fig. S11A, S12A) (6, 9). For example, allelic binding cooperativity tests (5) revealed four additional motifs (NFKB1, POU2F2, PRDM1, STAT2), located proximal to the PU.1-bound site,

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

which show covariance with AS PU.1 binding activity (FDR=5%; Fig. 2A, fig. S13) and collectively explain another 7.5% of AS PU.1 binding activity.

Despite a strong correlation between motif score differences and AS binding (fig. S11C, S12C; >90% expected direction), we observed that the majority of motif disrupting SNPs do not show significant allelic effects (fig. S11A, S12A). 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 (5, 10, 11) 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 (fig. 2B, fig. S12D), but this trend is not significant if a second, unaffected homotypic TF motif is located nearby (Fig. 2B, fig. S11D). 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 (Fig. 3A) and (ii) tested for transmission of allelic effects from parents to children (Fig. 3B, fig. S17) (4). Allelic directions at shared significant AS sites in the unrelated individuals were significantly correlated ($P < 0.05$, Spearman's correlation, fig. S16A), with mRNA showing the highest degree of consistency in allelic directions between individuals followed by TF binding and histone modification, respectively (Fig. 3A, S14–S16). We observed evidence of significant parental transmission with all three regulatory TFs ($\rho = 0.44$–0.75, $P <= 0.02$, Spearman correlation; Fig. 3C, S17), consistent with their strong sequence-dependence (4, 6). For hPTMs, evidence of transmission was detected for the active histone marks H3K4me1, H3K4me3, and H3K27ac ($\rho = 0.12$–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 ($\rho = 0.46$ and 0.50; $P = 0.0008$ and $P = 1.3e{-}07$, respectively). We observed only weak transmission for POLR2B (fig. S17), possibly due to the distinct activity states of the polymerase (12). We determined the genetic control of the transmission signal of histone marks at known expression (7) and DNaseI sensitivity quantitative trait loci (13) (eQTLs and dsQTLs, respectively), since the former are enriched within TF binding sites (13). Transmission of the active marks H3K4me1, H3K4me3, and H3K27ac was stronger near eQTLs and dsQTLs ($\rho = 0.31$–0.57) than genomewide (Fig. 3D, fig. S20), 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 (14–16).

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 (Fig. S21). 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 (17, 18);

Author Manuscript

GRO-seq/H3K4me1/H3K27ac at putative enhancers (19)) (Fig. 4A, fig. S21). 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 (20), 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 (5) (Fig. 4B, fig. S21), and found that every TF and histone mark exhibits HC with one or more regulatory layer(s) around genes and their flanking regions (fig. S21; 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; $P = 2.9e{-}08$). The majority (> 90%) of significant HC events were positive, *i.e.*, the allelic bias co-occurred on the same haplotype (Fig. 4B, fig. S21). 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) (fig. S22). 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 (21), and that the maintenance of a transcription-permissive environment and transcriptional memory are independent of histone modifications (22). Both histone modifications and TF binding are under genetic control, but histone modifications are more prone to stochastic, possibly transient effects and likely reflect (23), rather than define, coordinated regulatory interactions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by the Swiss National Science Foundation (SNSF) (grant CRSI33_130326 to E.T.D, B.D., A.R, and N.H; grant 31003A_132958 to NH; grant 31003A 130342 to ETD; grant 31003A_129835 to AR), the European Research Council (ERC) (to E.T.D.), the Louis-Jeantet Foundation (to E.T.D), the European Molecular Biology Organization (fellowship to H.K.; EMBO ALTF 2010-337), the doctoral school of the Faculty of Biology and Medicine, University of Lausanne (fellowship to R.M.W), the NCCR Frontiers in Genetics Program (E.T.D, B.D.), the Japanese-Swiss Science and Technology Cooperation Program (ETHZ; to B.D.), Institutional support from the École Polytechnique Fédérale de Lausanne (EPFL; to B.D.). The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics. All data in this publication are available through ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) under accession

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

number YYY. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.T.D, B.D., A.R. and N.H.

References and Notes

1. Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57. [PubMed: 22955616]
2. Montgomery SB, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773. [PubMed: 20220756]
3. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061. [PubMed: 20981092]
4. McDaniell R, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. Science (New York, N.Y. 2010; 328:235.
5. Detailed information on methods and analyses can be found in the supplementary materials available on. Science Online.
6. Reddy TE, et al. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome research. 2012; 22:860. [PubMed: 22300769]
7. Lappalainen T, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature, in press.
8. Henikoff S, Shilatifard A. Histone modification: cause or cog? Trends Genet. 2011; 27:389. [PubMed: 21764166]
9. Kasowski M, et al. Variation in transcription factor binding among humans. Science (New York, N.Y. 2010; 328:232.
10. Maurano MT, Wang H, Kutyavin T, Stamatoyannopoulos JA. Widespread site-dependent buffering of human regulatory polymorphism. PLoS genetics. 2012; 8:e1002599. [PubMed: 22457641]
11. Spivakov M, et al. Analysis of variation at transcription factor binding sites in Drosophila and humans. Genome Biol. 2012; 13:R49. [PubMed: 22950968]
12. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science (New York, N.Y. 2008; 322:1845.
13. Degner JF, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012; 482:390. [PubMed: 22307276]
14. Erwin DH, Davidson EH. The evolution of hierarchical gene regulatory networks. Nat Rev Genet. 2009; 10:141. [PubMed: 19139764]
15. Hobert O. Maintaining a memory by transcriptional autoregulation. Curr Biol. 2011; 21:R146. [PubMed: 21334290]
16. Ptashne M. On the use of the word ‘epigenetic’. Curr Biol. 2007; 17:R233. [PubMed: 17407749]
17. Lin CY, et al. Transcriptional amplification in tumor cells with elevated c-Myc. Cell. 2012; 151:56. [PubMed: 23021215]
18. Nie Z, et al. c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells. Cell. 2012; 151:68. [PubMed: 23021216]
19. Wang D, et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature. 2011; 474:390. [PubMed: 21572438]
20. Grundberg E, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nature genetics. 2012; 44:1084. [PubMed: 22941192]
21. Doege CA, et al. Early-stage epigenetic modification during somatic cell reprogramming by Parp1 and Tet2. Nature. 2012; 488:652. [PubMed: 22902501]
22. Margueron R, Reinberg D. Chromatin structure and the inheritance of epigenetic information. Nat Rev Genet. 2010; 11:285. [PubMed: 20300089]
23. Gutierrez-Arcelus M, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. Elife. 2013; 2:e00523. [PubMed: 23755361]
24. Muller N, Girard P, Hacker DL, Jordan M, Wurm FM. Orbital shaker technology for the cultivation of mammalian cells in suspension. Biotechnology and bioengineering. 2005; 89:400. [PubMed: 15619325]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

25. Canella D, Praz V, Reina JH, 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. 2010; 20:710. [PubMed: 20413673]
26. Schramm L, Pendergrast PS, Sun Y, Hernandez N. Different human TFIIIB activities direct RNA polymerase III transcription from TATA-containing and TATA-less promoters. Genes Dev. 2000; 14:2650. [PubMed: 11040218]
27. O'Geen H, Nicolet CM, Blahnik K, Green R, Farnham PJ. Comparison of sample preparation methods for ChIP-chip assays. Biotechniques. 2006; 41:577. [PubMed: 17140114]
28. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics. 2011; 43:491. [PubMed: 21478889]
29. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010; 20:1297. [PubMed: 20644199]
30. Lander ES, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860. [PubMed: 11237011]
31. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010; 26:589. [PubMed: 20080505]
32. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078. [PubMed: 19505943]
33. Harrow J, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006; 7(Suppl 1):S4 1. [PubMed: 16925838]
34. Heinz S, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38:576. [PubMed: 20513432]
35. Song Q, Smith AD. Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics. 2011; 27:870. [PubMed: 21325299]
36. Pickrell JK, Gaffney DJ, Gilad Y, Pritchard JK. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics. 2011; 27:2144. [PubMed: 21690102]
37. Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56. [PubMed: 23128226]
38. Wang J, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome research. 2012; 22:1798. [PubMed: 22955990]
39. Orphanides G, Lagrange T, Reinberg D. The general transcription factors of RNA polymerase II. Genes Dev. 1996; 10:2657. [PubMed: 8946909]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

One sentence summary

Coordinated allelic variation across molecular phenotypes

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

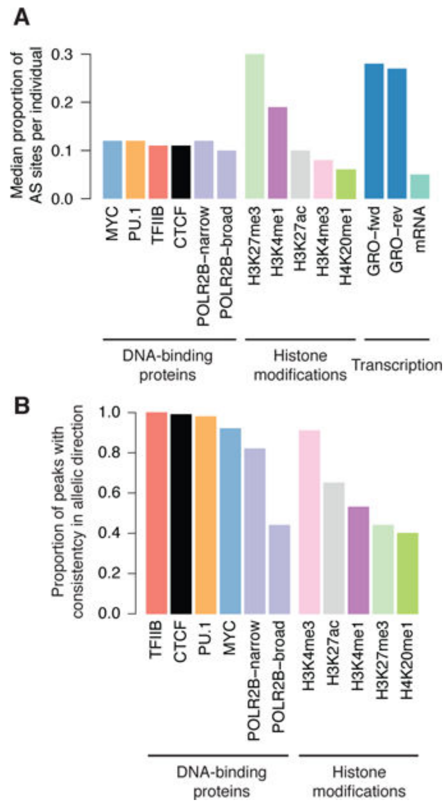

Figure 1. Allele-specific (AS) activity within transcriptional and chromatin layers

(**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.

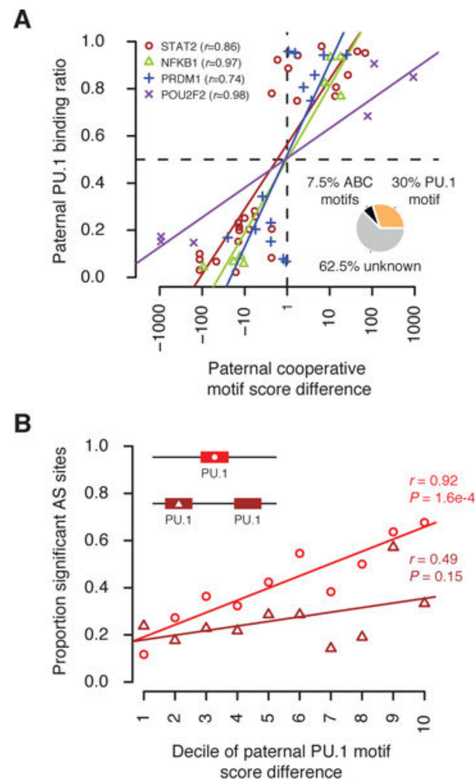

Figure 2. DNA sequence properties at allele-specific (AS) PU.1 binding sites

(**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).

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

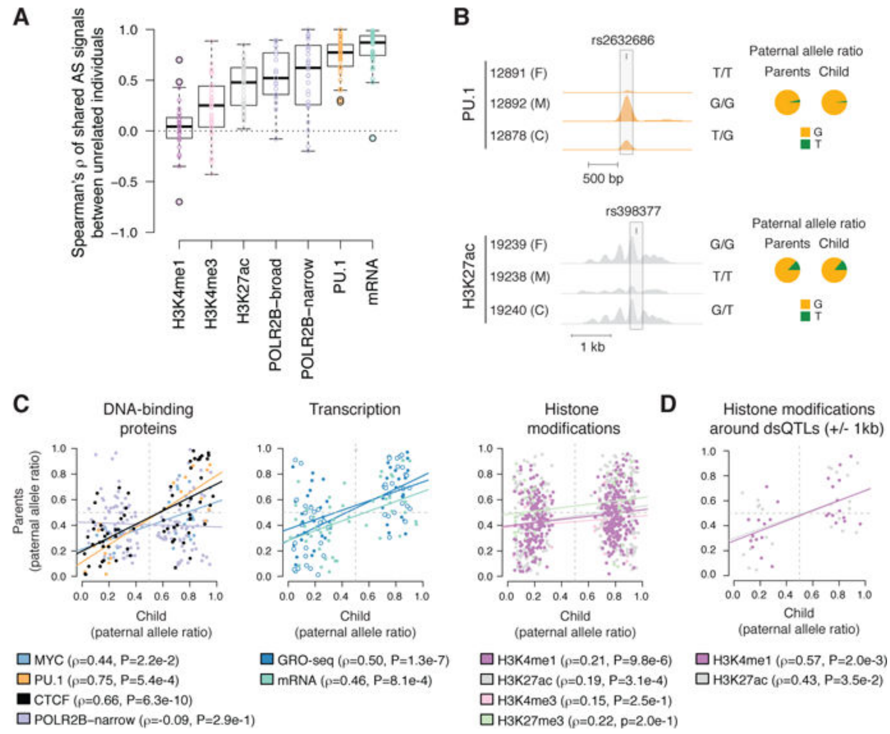

Figure 3. Genetic component of allele-specific (AS) transcriptional and chromatin activity

(**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).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

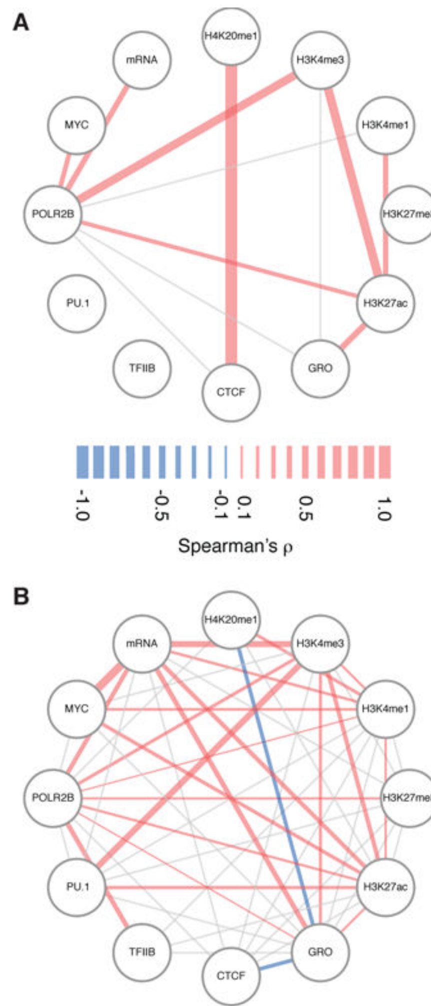

Figure 4. Local, short- and long-range coordination between transcriptional and chromatin layers

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 $\rho = -1.0$ (blue) to $\rho = 1.0$ (red). Non-significant correlations are indicated with grey lines and missing lines indicate lack of sufficient data points for analysis.