Variation in Transcription Factor Binding Among Humans

Maya Kasowski, *et al.*

Science **328**, 232 (2010);
DOI: 10.1126/science.1183621

Variation in Transcription Factor Binding Among Humans

Maya Kasowski,[1]* Fabian Grubert,[1,2]* Christopher Heffelfinger,[1] Manoj Hariharan,[1,2] Akwasi Asabere,[1] Sebastian M. Waszak,[3,4] Lukas Habegger,[5] Joel Rozowsky,[6] Minyi Shi,[1,2] Alexander E. Urban,[1,7] Mi-Young Hong,[1] Konrad J. Karczewski,[2] Wolfgang Huber,[3] Sherman M. Weissman,[7] Mark B. Gerstein,[5,6,8] Jan O. Korbel,[3,9]† Michael Snyder[1,2]†

Differences in gene expression may play a major role in speciation and phenotypic diversity. We examined genome-wide differences in transcription factor (TF) binding in several humans and a single chimpanzee by using chromatin immunoprecipitation followed by sequencing. The binding sites of RNA polymerase II (PolII) and a key regulator of immune responses, nuclear factor κB (p65), were mapped in 10 lymphoblastoid cell lines, and 25 and 7.5% of the respective binding regions were found to differ between individuals. Binding differences were frequently associated with single-nucleotide polymorphisms and genomic structural variants, and these differences were often correlated with differences in gene expression, suggesting functional consequences of binding variation. Furthermore, comparing PolII binding between humans and chimpanzee suggests extensive divergence in TF binding. Our results indicate that many differences in individuals and species occur at the level of TF binding, and they provide insight into the genetic events responsible for these differences.

Differences in gene expression have been observed in a variety of species (*1*–*3*). However, the extent to which transcription factor (TF) binding differences occur both among individuals and between closely related species, and the global relationship between TF binding and genetic variation, are largely unexplored (*4*). We used chromatin immunoprecipitation followed by sequencing (ChIP-Seq) to map nuclear factor κB (NFκB) and RNA polymerase II (PolII) binding sites in 10 humans: 5 are of European ancestry (including a parent-offspring trio), 2 of eastern Asian ancestry, and 3 of Nigerian ancestry (table S1); 9 of these have been analyzed by the HapMap (*5*) and the 1000 Genomes (*6*) projects, and one represents an individual for whom extensive structural variant (SV) maps are available (*7*, *8*). All individuals but one were females; in pairwise comparisons, modest differences in TF binding were observed between the male and 9 females; our analyses thus combined results from all 10 humans. For comparison we also analyzed PolII binding in one female chimpanzee.

We used stringent criteria to identify binding peaks (*9*), and clustered them into discrete binding regions (BRs) (*10*), yielding a total of 15,522 and 19,061 BRs for NFκB and PolII, respectively. Within BRs, most peaks were similar in position and magnitude among individuals (fig. S1A). However, significant differences in binding were observed (fig. S1A), and the Spearman correlation coefficients among replicates of different individuals (median values 0.79 and 0.90 for NFκB and PolII, respectively) were less than that of biological replicates of a given individual (median values 0.90 and 0.95, respectively) (fig. S2A and table S2). Seven and a half and 25% of the NFκB and PolII BRs, respectively, differed significantly between two individuals [analysis of variance test (*10*), Bonferroni-adjusted P value < 0.05; (*10*)] (fig. S3C), and many variable BRs exhibited more than twofold magnitude differences in binding (fig. S3D). Variable BRs for both NFκB and PolII, respectively, were often coassociated ($P < 1 \times 10^{-4}$; permutation test) (Fig. 1D and fig. S4), a correlation that is particularly strong for BRs that are less than 10 kb apart (fig. S4A). Variable NFκB and PolII regions were also often coassociated ($P = 2.80 \times 10^{-25}$, Kolmogorov-Smirnov test) (table S3 and fig. S4A), even though the NFκB and PolII data are from tumor necrosis factor–α (TNF-α)–treated and untreated cells, respectively. These results suggest that adjacent binding sites and BRs may influence one another, perhaps through cooperative binding or interactions with other proteins.

For both NFκB and PolII, BRs within 1 kb of transcription start sites (TSSs) of RefSeq genes showed less variability (6 and 25%, respectively) than intergenic peaks (8 and 28%) ($P < 1 \times 10^{-4}$; permutation test). TSS BRs also revealed stronger ChIP-Seq signals (1.2- and 2.3-fold, respectively), with many exceptions (fig. S5). The majority of binding regions (>70%) were occupied in two or more individuals, which argues against cell line artifacts (fig. S3B). The signal intensity for 40 and 53% of the BRs absent (that is, "lost") in one individual was similar to background for NFκB and PolII (*10*), respectively, suggesting complete absence of binding in these cases, rather than threshold effects.

BRs differing in TF occupancy among individuals often involve loci of potentially high interest. These include the *RPS26*, *BLK*, *SP140*, and *ZNF804A* genes for PolII, which have been associated with type 1 diabetes, systemic lupus erythematosus, chronic lymphatic leukemia, and schizophrenia, respectively, and *ORMDL3*, *PTGER4*, and *LOC253039* for NFκB, which are associated with asthma, Crohn's disease, and rheumatoid arthritis (*10*). Genes with variability in PolII binding showed a slight enrichment with immunity and defense functional gene categories ($P = 0.045$, Benjamini-Hochberg multiple testing correction) among target genes (*10*).

We examined the genetic contribution to binding variation using single-nucleotide polymorphisms (SNPs) from the 1000 Genomes project. Individual SNPs in NFκB and PolII BRs frequently affected binding (Fig. 1A and fig. S6A), and the number of SNPs in BRs correlated with the frequency of significant binding differences (Fig. 1B). SNPs altering the NFκB DNA binding motif had a strong effect, elevating the frequency of significant binding differences 2.4-fold. About 90% of the binding differences followed the expected trend in which better matches to the consensus motif yielded higher binding signals ($P < 1 \times 10^{-3}$) (Fig. 1C, table S4, and fig. S6B). SNPs that putatively affect binding are abbreviated as B-SNPs (binding SNPs).

We also searched for other associated DNA motifs, such as the Stat1 motif [previously associated with NFκB-binding (*11*)], TATA box, CAAT box, and GC box (*12*), and we performed de novo searches for enriched DNA motifs in BRs (*10*), which revealed BR enrichments for the NFκB motif and the GC box, along with additional motifs (fig. S7). We assessed the effect of genetic variation on each of the motifs. SNPs in the Stat1 motif markedly elevated the frequency of significant NFκB binding differences (1.3-fold enrichment; $P < 1 \times 10^{-3}$, permutation test) (Fig. 1B), and 71% of the alterations in the Stat1 motif changed NFκB binding in the expected direction; that is, improved Stat1 motif sequences increased NFκB binding ($P < 1 \times 10^{-3}$) (Fig. 1C, table S4, and fig. S6B). For PolII, SNPs in the CAAT box had a strong effect on binding (1.6-fold; $P < 1 \times 10^{-3}$), with 63% of cases displaying the correct trend, whereas SNPs in the TATA box and GC box had modest effects (1.5-fold and 1.3-fold, with 51 and 52%, respectively, exhibiting the correct trend). The significant covariance in the Stat1 motif with NFκB binding differences and the nuclear factor Y (NFY) CAAT box with PolII binding differences suggests a functional interaction of Stat1 with NFκB and NFY with PolII, respectively; the latter has been documented previously (*13*). We call this approach to examine covariation of motifs with variable binding regions the allele binding cooperativity test or ABC test.

We next analyzed the effect of SVs, >1-kb genomic segments displaying copy-number var-

[1]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520, USA. [2]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA. [3]Genome Biology Research Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [4]Department of Biotechnology and Bioinformatics, Weihenstephan-Triesdorf University of Applied Sciences, 85350 Freising, Germany. [5]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA. [6]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. [7]Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA. [8]Department of Computer Science, Yale University, New Haven, CT 06520, USA. [9]European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: jan.korbel@embl.de (J.O.K.); mpsnyder@stanford.edu (M.S.)

iants (CNVs) or balanced inversions (*7*, *8*, *14*, *15*). We probed high-density microarrays to identify CNVs in seven individuals (*10*) (table S5) and combined these with CNVs from another survey (*15*). CNVs significantly elevated the frequency of BR differences between individuals by 5.1- and 2.0-fold for NFκB and PolII, respectively ($P < 1 \times 10^{-4}$, permutation test) (Fig. 2, A and B, fig. S8, and table S6). Furthermore, the effect followed the correct trend in 90 and 80% of the respective NFκB and PolII cases (Fig. 2C); deletions reduced binding signals, whereas duplications elevated them. A combined set of high-resolution SVs identified by paired-end mapping (*7*, *14*) also exhibited enrichment in binding differences for deletions intersecting with NFκB and PolII BRs [3.2-fold and 1.7-fold, respectively ($P < 1 \times 10^{-4}$, permutation test)]. We observed a 2.8-fold significant enrichment in differential binding owing to inversions affecting NFκB BRs ($P < 1 \times 10^{-4}$, permutation test), and a slight, nonsignificant enrichment due to inversions affecting PolII BRs (Fig. 2B), suggesting that inversions may affect binding. SVs that are associated with binding are abbreviated B-SVs (binding SVs).

The total fraction of significant binding differences coinciding with genetic variations was 35% for NFκB and 26% for PolII (table S7 and fig. S6C). Thirty-four percent of the NFκB BRs intersect with SNP differences between corresponding regions in different individuals (1% intersect with a known TF motif, with SNPs falling both in the NFκB or the STAT1 motif) (table S8) and 3% with SVs (some SNPs coincide with SVs). Thus, genetic differences affecting the BR can be assigned to many, but not to the majority of, binding differences. Possible reasons for the remaining BR variation include *trans*-effects, epigenetic variation, as well as B-SNPs and B-SVs that were not ascertained. Some of the binding differences could be related to the different ages of the individuals.

We examined the effect of binding variation on gene expression by generating deep RNA-Seq data from each cell line (*10*) and comparing those data with binding data (Fig. 3A and fig. S9A). A significant correlation was observed (Spearman correlation coefficients of 0.475 and 0.461 for NFκB and PolII, respectively) (Fig. 3B, fig. S9B, and table S9), suggesting an influence of binding differences on mRNA abundance. Examples of correlated genes include *UGT2B17*, *GSTM1*, and *ZNF804A*, which encode glucuronic acid and glutathione transferases, and a gene linked to schizophrenia (*10*). However, a number of BR differences were not associated with differences in gene expression and presumably compensatory (for example, feedback) mechanisms influence the expression in these cases. We also examined the effect of B-SNPs with differences in both binding and gene expression and found that both NFκB and PolII binding and expression differences correlated with the presence of B-SNPs, including those in the NFκB and Stat1 motif (for NFκB) and the CAAT, GC, and TATA box (for PolII) (Spearman correlation coefficients: 0.48 to 0.82) (Fig. 3C and table S9). Copy number differences (that is, B-SVs) also correlated with gene expression, albeit the correlation was not as strong as that of copy number differences with binding (table S10), indicating a more-direct role for genetic variation on TF binding than on gene expression.

The observation that SNPs and SVs are frequently associated with binding differences suggests a crucial role of *cis* elements in the genetics of TF binding. We thus analyzed the segregation pattern of BR occupancy in the parent-offspring trio, and observed potential Mendelian segregation in >90% of BRs (fig. S10A), although this was difficult to determine with certainty, because not all alleles that are relevant to TF binding have been ascertained in the parents. In the child, 947 and 732 BRs were occupied by NFκB and PolII, respectively, but not in the parents. This is indicative of transgression in which a binding event was evident only in the offspring (Fig. 3, A and D, fig. S10B, and tables S11, S12, and S13).

We also examined whether some BRs are specific to certain populations. Although the number of individuals analyzed was small, the NFκB data revealed a total of 14 BRs that were specifically occupied or unoccupied in the African or Asian individuals (table S14). For PolII, the chimpanzee data were used to infer gains and losses relative to the likely ancestral state of binding, and a total of 68 population-specific occupancies (gains and losses) were identified in the three population groups (table S14). Overall, we found relatively few population-specific events, ~0.1 to ~0.4%, suggesting that most alleles affecting TF binding are shared among different populations.

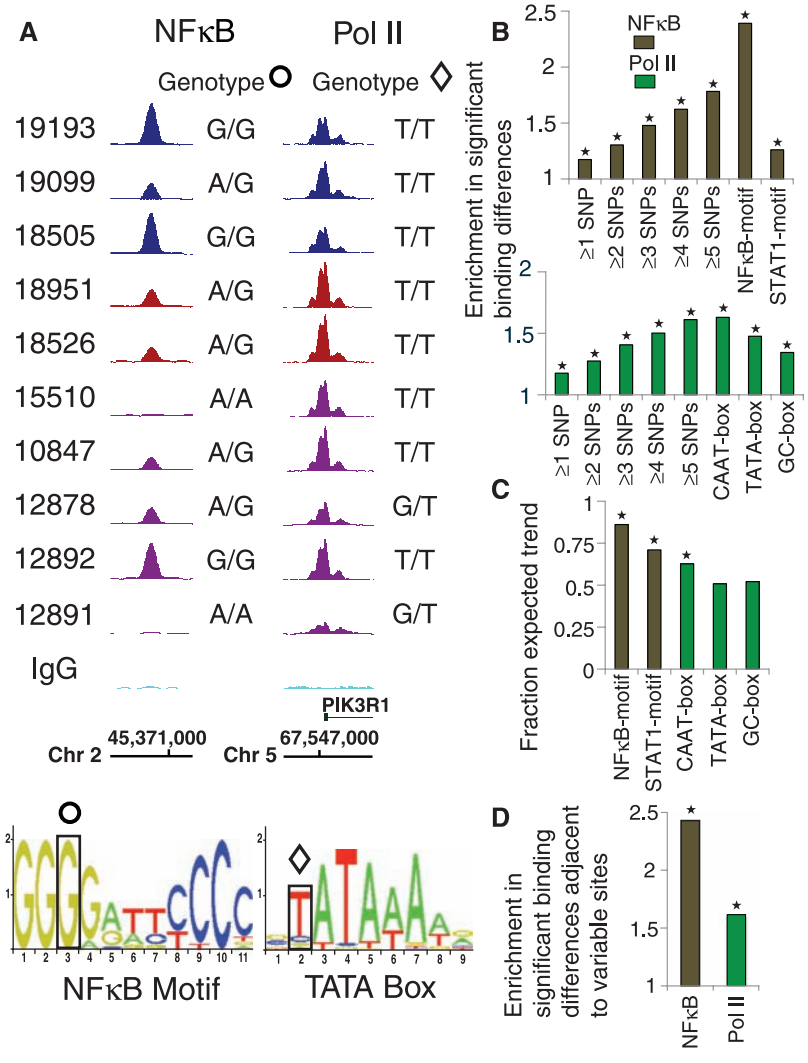

Fig. 1. Effect of SNPs on NFκB and PolII binding. (**A**) Signal tracks of a NFκB motif and a TATA box demonstrate effects of B-SNPs on TF binding, with correlations in the expected direction (that is, with correct trend). (**B**) Fold enrichments for cumulative SNP differences affecting BRs and for single SNPs affecting motifs, in pairwise comparisons between individuals relative to the overall frequency of binding differences for NFκB (7.5%) and PolII (25%). (**C**) B-SNPs affecting motifs frequently lead to binding differences with correct trend. *$P < 0.001$, based on randomization tests involving 10,000 permutations, that is, permutation tests. (**D**) BRs adjacent to differentially bound BRs are enriched for binding variation.

Because humans and chimpanzees exhibit 5 to 10% differences in gene expression (*16*), we also examined divergence of TF binding among primates by analyzing PolII binding in a single chimpanzee. We analyzed 15,418 (81%) of human BRs with corresponding syntenic regions in the chimpanzee genome. The majority of PolII BRs were occupied both in humans and chimp (fig. S11A). However, 32% of the BRs exhibited significant differences in binding (corrected P value < 0.05) (Figs. 2A and 4A), a figure higher than that for human PolII variation (25%). Genes near regions uniquely occupied in the chimp were enriched in the following functional categories: (i) nucleoside, nucleotide, and nucleic acid metabolism; and (ii) steroid metabolism (P values = 3.60×10^{-5} and 4.16×10^{-4}, respectively). Furthermore, BRs that were uniquely occupied in humans were significantly enriched in protein modification and mRNA transcription [Fischer Exact test (*10*), Benjamini-Hochberg P values = 2.22×10^{-89} and 9.08×10^{-139}, respectively] (table S15).

As in humans, relative differences identified in the chimpanzee were higher in intergenic BRs

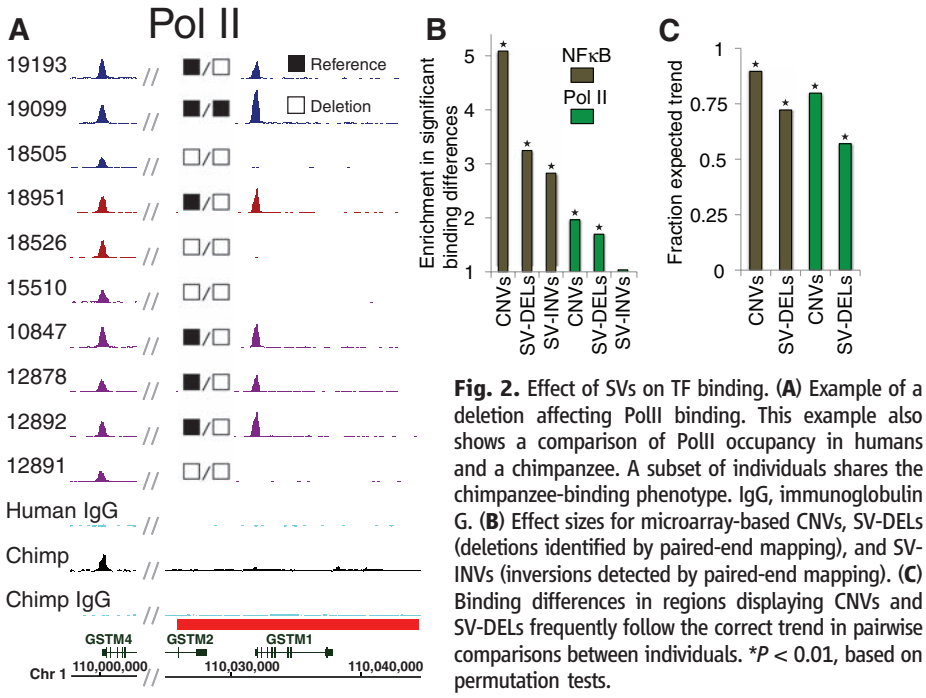

Fig. 2. Effect of SVs on TF binding. (**A**) Example of a deletion affecting PolII binding. This example also shows a comparison of PolII occupancy in humans and a chimpanzee. A subset of individuals shares the chimpanzee-binding phenotype. IgG, immunoglobulin G. (**B**) Effect sizes for microarray-based CNVs, SV-DELs (deletions identified by paired-end mapping), and SV-INVs (inversions detected by paired-end mapping). (**C**) Binding differences in regions displaying CNVs and SV-DELs frequently follow the correct trend in pairwise comparisons between individuals. *P < 0.01, based on permutation tests.

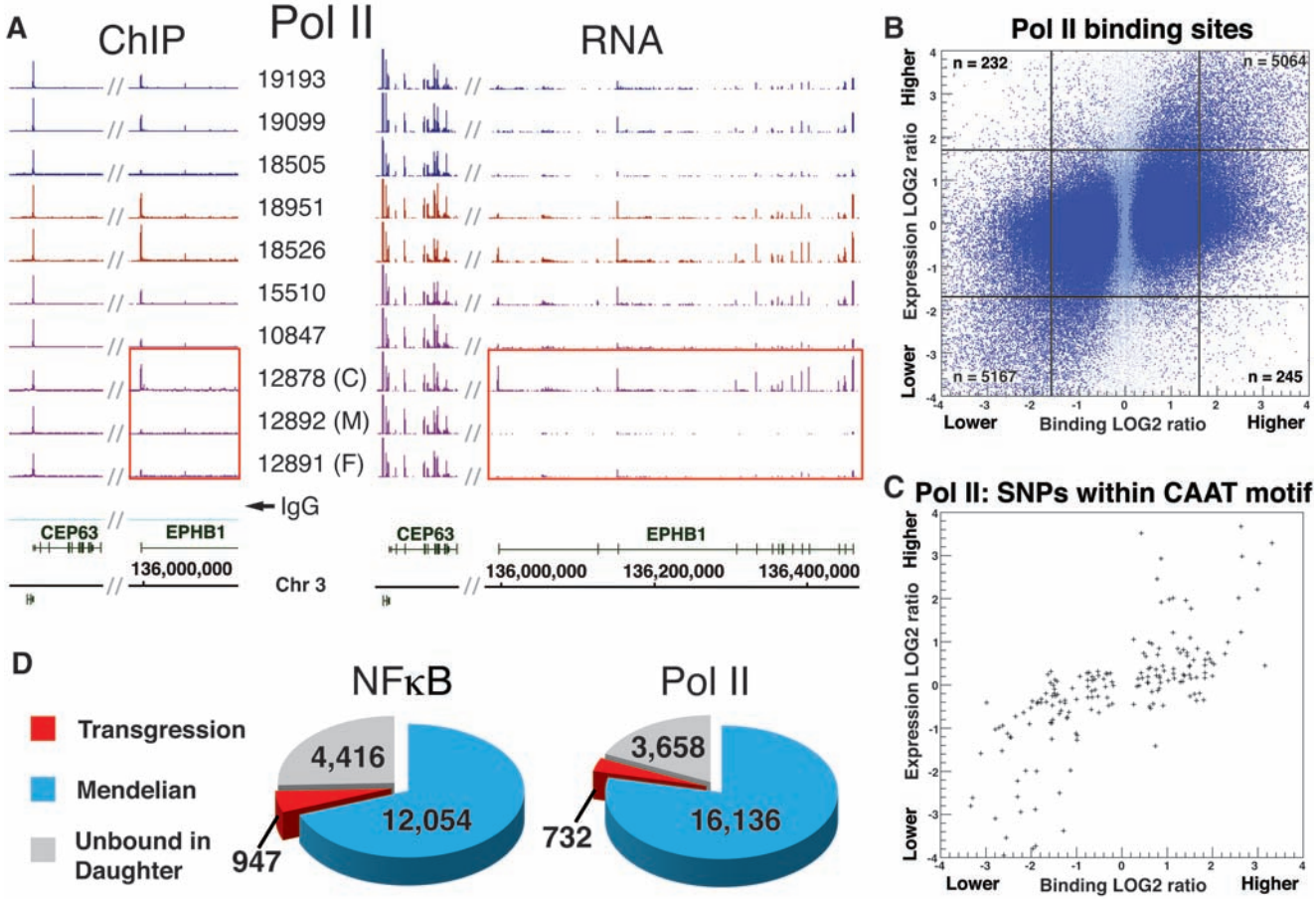

Fig. 3. Correlation and effect sizes of TF binding and gene expression. (**A**) Example showing a correlation of binding and expression. This figure also shows a transgression event, in which the daughter displays a strong increase in binding relative to the parents. Continuous signal tracks are shown in fig. S10C. (**B**) Regions with binding variation correlate with differences in expression. Dark blue dots, PolII BRs displaying significant differences in binding in pairwise comparisons between individuals; light blue dots, other BRs. The black lines demarcate data points that fall 2 SDs outside the binding ratio or gene expression distributions. Indicated counts (n) represent data points falling into the four corners for each data set. (**C**) Strong correlation between binding and gene expression at BRs in which a B-SNP intersects with the PolII-specific CAAT box. (**D**) Breakdown of segregation events in the trio showing the extent of BRs with candidate transgression events.

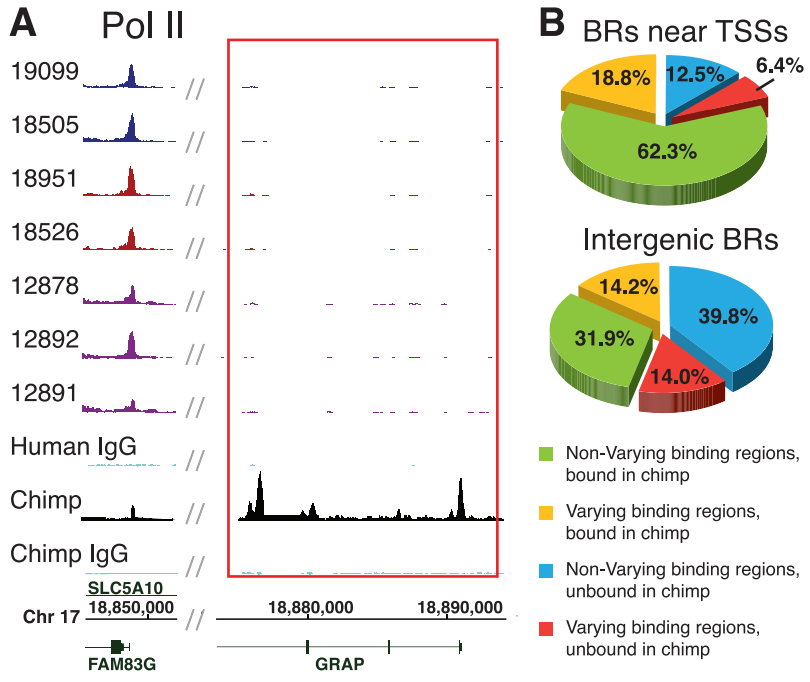

Fig. 4. Comparison of PolII binding in humans and a chimpanzee. (**A**) Signal tracks for a peak found only in the chimpanzee. All 10 individuals are shown in fig. S11B. (**B**) Pie charts displaying occupancy by PolII of genomic regions where the chimp and human genomes are in synteny.

relative to BRs within 1 kb of a TSS: 33% of the syntenic intergenic PolII BRs differed significantly from the human samples, compared with 31% near TSSs ($P < 1 \times 10^{-4}$; permutation test). Consequently, human BRs near TSSs were generally more likely to be scored as occupied in chimpanzee (81%) than intergenic BRs were (46%) (Fig. 4B). Furthermore, human BRs with strong binding signals (that is, many mapped reads) are more frequently occupied in the chimpanzee than those with weaker signals (fig. S11C), indicating either divergence of the weaker sites or signals that fell below the threshold at the low signal sites. Finally, we observed a general correlation between polymorphism and divergence in binding; that is, variable BRs in humans displayed, on average, more divergence from chimpanzee BRs (in terms of fold change in normalized read counts) than did non-variable BRs (Spearman test, 0.68; $P = 3.9 \times 10^{-7}$) (fig. S11D).

Our data demonstrate extensive contributions of genetic variations on TF binding, many of which are expected to be functional through their effect on gene expression. Overall, the differences observed here (7.5 and 25% for NFκB and PolII, respectively, for humans; 32% for human/chimpanzee) greatly exceed estimates for sequence variation in coding sequences [estimated as 0.025% for humans (*17*) and 0.71% for human/chimpanzee (*18*)], suggesting a strong role for binding variation in human diversity. Extending mapping of B-SNPs and B-SVs for these and additional transcription factors should further inform on the genetic underpinnings of phenotypic diversity in humans and provide insights into genetic causes of human disease.

References and Notes

1. B. E. Stranger *et al.*, *Science* **315**, 848 (2007).
2. M. V. Rockman, L. Kruglyak, *Nat. Rev. Genet.* **7**, 862 (2006).
3. D. A. Skelly, J. Ronald, J. M. Akey, *Annu. Rev. Genomics Hum. Genet.* **10**, 313 (2009).
4. A. R. Borneman *et al.*, *Science* **317**, 815 (2007).
5. International HapMap Consortium, *Nature* **449**, 851 (2007).
6. More information on the 1000 Genomes project can be found at http://1000genomes.org.
7. J. O. Korbel *et al.*, *Science* **318**, 420 (2007).
8. E. Tuzun *et al.*, *Nat. Genet.* **37**, 727 (2005).
9. J. Rozowsky *et al.*, *Nat. Biotechnol.* **27**, 66 (2009).
10. Materials and methods and supporting data are available on *Science* Online.
11. O. H. Krämer *et al.*, *Genes Dev.* **20**, 473 (2006).
12. A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, B. Lenhard, *Nucleic Acids Res.* **32** (Database issue), D91 (2004).
13. M. C. Faniello *et al.*, *J. Biol. Chem.* **274**, 7623 (1999).
14. J. M. Kidd *et al.*, *Nature* **453**, 56 (2008).
15. S. A. McCarroll *et al.*, *Nat. Genet.* **40**, 1166 (2008).
16. H. Creely, P. Khaitovich, *Prog. Brain Res.* **158**, 295 (2006).
17. S. Levy *et al.*, *PLoS Biol.* **5**, e254 (2007).
18. H. Watanabe *et al.*, *Nature* **429**, 382 (2004).

19. We thank the 1000 Genomes project for early data access. This research was funded by grants from NIH (M.S., S.W., and M.G.), and by funding from the European Molecular Biology Laboratory (J.K.), a March of Dimes Foundation Grant (A.U.), and the NIH Medical Scientist Training Program grant TG T32GM07205 (M.K.). M.K. was a Howard Hughes Medical Institute Medical Research Training Fellow. Data sets are available at the Gene Expression Omnibus (GEO) database with accession number GSE19486. M.S. is on the Scientific Advisory Board and a founder for both Affymetrix and Metagenomix.

Supporting Online Material

www.sciencemag.org/cgi/content/full/science.1183621/DC1
Materials and Methods
Figs. S1 to S16
Tables S1 to S20
References

20 October 2009; accepted 12 February 2010
Published online 18 March 2010;
10.1126/science.1183621
Include this information when citing this paper.

Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans

Ryan McDaniell,[1] Bum-Kyu Lee,[1] Lingyun Song,[2,3] Zheng Liu,[1]* Alan P. Boyle,[2] Michael R. Erdos,[4] Laura J. Scott,[4,5] Mario A. Morken,[4] Katerina S. Kucera,[2] Anna Battenhouse,[1] Damian Keefe,[6] Francis S. Collins,[4] Huntington F. Willard,[2] Jason D. Lieb,[7] Terrence S. Furey,[2] Gregory E. Crawford,[2,3]† Vishwanath R. Iyer,[1]† Ewan Birney[6]†

The extent to which variation in chromatin structure and transcription factor binding may influence gene expression, and thus underlie or contribute to variation in phenotype, is unknown. To address this question, we cataloged both individual-to-individual variation and differences between homologous chromosomes within the same individual (allele-specific variation) in chromatin structure and transcription factor binding in lymphoblastoid cells derived from individuals of geographically diverse ancestry. Ten percent of active chromatin sites were individual-specific; a similar proportion were allele-specific. Both individual-specific and allele-specific sites were commonly transmitted from parent to child, which suggests that they are heritable features of the human genome. Our study shows that heritable chromatin status and transcription factor binding differ as a result of genetic variation and may underlie phenotypic variation in humans.

Control of gene transcription is believed to be important in determining organismal phenotype and fitness. Variations in genomic DNA, such as single-nucleotide polymorphisms (SNPs), insertions, or deletions (indels), may act singly or in combination to influence gene regulation (*1*, *2*). These heritable variations have been thought to affect the binding of sequence-