Local ancestry transitions modify snp-trait associations

Alexandra E. Fish

Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN 37235, USA; Departments of Biological Sciences, Biomedical Informatics, and Computer Science, Vanderbilt University, Nashville, TN 37235, USA.
Email: alex.fish@vanderbilt.edu

Dana C.Crawford

Institute for Computational Biology, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Wolstein Research Building, 2103 Cornell Road, Suite 2530, Cleveland, OH 44106, USA
Email:dana.crawford@case.edu

John A. Capra*

Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN 37235, USA; Departments of Biological Sciences, Biomedical Informatics, and Computer Science, Vanderbilt University, Nashville, TN 37235, USA.
Email: tony.capra@vanderbilt.edu

William S. Bush*

Institute for Computational Biology, Department of Population and Quantitative Health Sciences, Case Western Reserve University, Wolstein Research Building, 2103 Cornell Road, Suite 2530, Cleveland, OH 44106, USA
Email:wsb36@case.edu

Genomic maps of local ancestry identify ancestry transitions – points on a chromosome where recent recombination events in admixed individuals have joined two different ancestral haplotypes. These events bring together alleles that evolved within separate continential populations, providing a unique opportunity to evaluate the joint effect of these alleles on health outcomes. In this work, we evaluate the impact of genetic variants in the context of nearby local ancestry transitions within a sample of nearly 10,000 adults of African ancestry with traits derived from electronic health records. Genetic data was located using the Metabochip, and used to derive local ancestry. We develop a model that captures the effect of both single variants and local ancestry, and use it to identify examples where local ancestry transitions significantly interact with nearby variants to influence metabolic traits. In our most compelling example, we find that the minor allele of rs16890640 occuring on a European background with a downstream local ancestry transition to African ancestry results in significantly lower mean corpuscular hemoglobin and volume. This finding represents a new way of discovering genetic interactions, and is supported by molecular data that suggest changes to local ancestry may impact local chromatin looping.

1. Introduction

Admixture occurs due to recent mixing of ancestral human populations, and admixed populations represent a unique opportunity to investigate epistasis, or genetic interactions, between alleles with different histories. Prior studies have shown that variants common to any one ancestral population (minor allele frequency/MAF > 5%) are typically shared between all populations,[1] though the frequency at which they occur can vary substantially among different ancestral groups.[2] Lower frequency variants (MAF < 5%) are much more likely to be population specific, and are more likely

[*] Authors contributed equally to this work

to occur in more ancestral populations. Both recombination rates and the sites of recombination also vary considerably by population; for example, there are more than 2,000 recombination hotspots observed in populations of West African descent, but not European descent populations.[3] Differences in both allele frequency and recombination hotspots between the ancestral populations of an admixed group result in combinations of variants that may have not been observed (or occurred very rarely) in either continental population separately. Given the extensive admixture between human populations over the past several hundred years, many allelic combinations on admixed chromosomes have had limited time to undergo purifying selection and thus may be more likely to influence human traits.

It remains unclear what role epistasis, or genetic interactions, plays in the architecture of human traits. However studies of model organisms suggest that genetic variants likely do not act in isolation, and that the genetic background of a variant may influence its phenotypic effect. For example, it is well-established that when human-derived disease-associated variants are introduced into mice, the phenotypic consequences vary between strains despite consistent environmental factors.[4] This could occur through a variety of mechanisms: strains may carry compensatory mutations that mitigate the effect of the variant, or a given threshold of genetic predisposition (i.e., burden) may be required for phenotypic effects to manifest. Within a natural population, genetic variants that mask the effect of a genomic region may permit potentially deleterious variants to arise. For example, genetic variants that are associated with decreased expression of a transcript can accumulate recently derived rare variants on the same haplotype with limited phenotypic impact.[5,6] Functional haplotypes of regulatory variants also form in which variants either cancel out one another's effects, or in which they both amplify their influence on a phenotype in the same direction.[7]

In this study, we explore admixed chromosomes for new combinations of variants with observable epistasis. As different ancestral chromosomes recombine, disease-associated alleles are placed onto new genetic backgrounds – here we specifically focus on recombination events in close physical distance to variants with established trait associations. Within a dataset of nearly 10,000 adults of African ancestry (an admixed group of European and African ancestral populations), we investigated whether transitions in local ancestry modify the effect of variants on electronic health record (EHR) derived phenotypes. We used genetic data from the Illumina Metabochip, a custom genotyping platform with dense genotyping of previously disease-associated genomic regions, to identify chromosomes with a transition in local ancestry. We then used linear regression models to evaluate the impact of local ancestry transitions on SNP-trait associations reported for African-descent populations in the NHGRI-EBI GWAS Catalog.[8]

2. Methods

2.1. *Subjects and Genotyping*

All samples used in this analysis were part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study, which used Vanderbilt University Medical Center's de-identified biorepository (BioVU) to link patient EHR data with blood-based DNA samples.[9] Details of the consent model and other human subjects issues are described elsewhere.[10] EAGLE selected 9,559 individuals for inclusion in based upon adminstratively-reported African American race,[11,12] rather than for specific health phenotypes, which minimizes ascertainment bias.[13] Samples were genotyped using the Illumina Metabochip, a custom array of almost 200,000 SNPs that targets genomic regions previously associated with type 2 diabetes, obesity, coronary artery disease, and other cardio-metabolic traits for fine-mapping purposes.[14] As part of quality control, variants were removed that did not have at least a 95% genotyping efficiency rate, or that did not vary in this dataset, leaving a total of 192,093 variants for analysis.

2.2. *Local Ancestry Determination*

Local ancestry was assigned using a two-step process: first, we phased the genotype data using SHAPEITv2[15] and the 1000 Genomes Phase 3 reference panel (available for download at https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference). There were 171,439 variants that were successfully phased; when variants failed it was typically due to inconsistencies with the reference panel. We then used RFMix[16] (v1.5.4) to assign local ancestry of the phased genetic data, using a window size of 0.1 cM, and a minimum node size of 5. Phased chromosomal haplotypes were matched to ancestral population reference panels. We used all Yoruba (YRI) and CEPH/European (CEU) individuals from 1000G, phase 3v5a, representing African (AFR) and European (EUR) ancestry, respectively.

2.3. *Electronic Phenotyping and Quality Control*

Using the GWAS Catalog,[18] we identified phenotypes and their corresponding EHR trait that had previous associations to regions fine-mapped by Metabochip (Table 1). Across the EHR, individuals have multiple measures for many quantitative traits, such as height/weight/BMI[19] and low-density lipoprotein (LDL) levels. For the majority of traits, we computed the median measurement for each year with data available in the EHR, and then computed the median of these scores. For more rarely collected quantitative traits (i.e. uric acid, serum albumin, etc), we took the median value over all entries. For all phenotypes, we removed clearly non-valid scores (i.e., scores of zero or one), and then removed outliers (those scores more than three standard deviations away from the mean) as quality control.

2.4 *Statistical analyses*

Given that local ancestry is specific to a given chromosome, we performed all analyses on the level of the chromosome, rather than the individual. We used linear regression to determine whether local ancestry transitions interacted with the allele to influence the phenotypes of interest, using the following model:

$$y = A + LA + TRANS + A * TRANS + PC_{1-3} + AGE + GENDER + BMI$$

where y is the phenotype of interest; A corresponds to the allele status (0, absence of the allele; 1, presence of the allele); LA corresponds to the local ancestry at the variant (0, African ancestry; 1, European ancestry); TRANS indicates the presence of a local ancestry transition within the Metabochip region (0, no transition; 1, no transition); A*TRANS represents the interaction term between the allele and local ancestry transition (1 indicates presence of both the allele and a local ancestry transition; 0 encompasses all other possibilities). AGE, GENDER, and the first three principal components (PC1-3) were included as covariates for all analyses. In the case of binary phenotypes, logistic regression was used. Because this analysis is designed as an investigation of known SNP-trait associations, we used a region-phenotype level Bonferroni multiple testing correction, correcting for the number of tests performed within each Metabochip region.

Table 1. GWAS Catalog traits with genetic associations in Metabochip regions. *median value was taken. In ** the median of yearly medians was taken.

GWAS Catalog Trait	Corresponding EHR Trait	Metabochip Region	Top Reported Gene
Urate levels	Uric Acid*	chr6:25235303-26141375	*SLC17A1*
Type 2 diabetes	PAGE T2D Algorithm[17]	chr11:2444094-2943115	*KCNQ1*
Red blood cell traits	Red Blood Cell Count*; Red Cell Distribution Width*	chr6:25235303-26141375	*HFE*
Iron status biomarkers	Total Iron Binding Capacity*	chr6:25235303-26141375	*HFE*
Weight	Weight**	chr16:53539509-54185787 chr18:57727147-58094636	*FTO* *MC4R*
Hematology traits	Albumin*; Alkaline Phosphatase*; Anion-gap*; Blood Urea Nitrogen*; Calcium*; Cloride*; CO2*; Creatinine*; GluBed*; Glucose*; Hgb*; Potasium*; Mean Corpuscular Hemoglobin*; Mean Corpuscular Volume*; Sodium*; RBC Count*; Red Cell Distribution Width*; Aspartate Aminotransferase*; Alanine Transaminase*; Total Bilirubin*; White Blood Count; MPV*; Platelet Count*; Total Iron Binding Capacity*	chr6:25235303-26141375	*HFE*
Mean platelet volume	MPV*	chr12:111290599-113206306 chr12:111505708-113105952 chr12:111681897-112225304 chr6:25235303-26141375	*ACAD10* *ACAD10* *ACAD10* *LRRC16A*
Height	Height**	chr7:27784039-28282062	*JAZF1*
Obesity-related traits	BMI**	chr16:53539509-54185787	*FTO*
Platelet count	Plt-Ct*	chr12:111290599-113206306	*SH2B3*
Coronary artery disease	Cases at least one ICD-9-CM Codes (410 – 414); all others were controls	chr12:111290599-113206306 chr12:111505708-113105952 chr12:111681897-112225304 chr13:110795080-111049623 chr18:57727147-58094636	*ALDH2* *ALDH2* *ALDH2* *RP11* *PMAIP1*
LDL cholesterol	First LDL-C measurement (with no mention of medication use)	chr1:109655637-110043693 chr1:109789347-109826136	*SORT1* *SORT1*
Body mass index	BMI**	chr12:111290599-113206306 chr12:111505708-113105952 chr12:111681897-112225304 chr16:53539509-54185787 chr18:57727147-58094636 chr3:122976919-123206919 chr3:123039584-123139034	*ALDH2* *ALDH2* *ALDH2* *FTO* *MC4R* *ADCY5* *ADCY5*

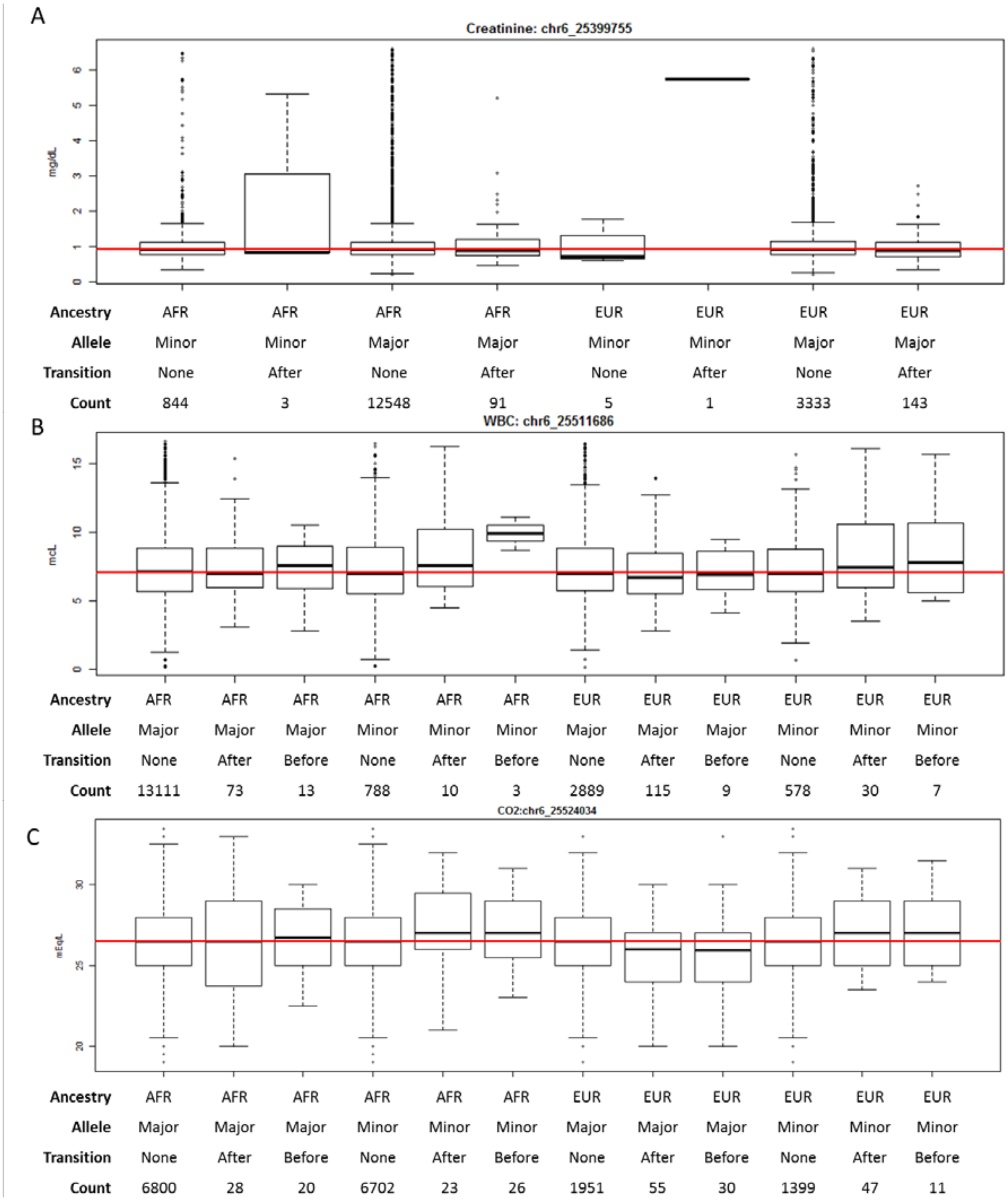

Figure 1. Stratified phenotype distributions reveal interactions between local ancestry transitions and variants regulating creatinine (A), white blood cell (WBC) counts (B), and CO_2 levels (C). Interactions are characterized by stratifying chromosomes based on: the local ancestry at the variant (EUR or AFR); the major or minor allele; and the presence and relative location (upstream/downstream) of a local ancestry transition on that chromosome within the broader Metabochip region. The number of chromosomes observed for each category is provided and reveals that the interactions for creatine and WBC are driven by a small number of chromosomes. The overall median value for each phenotype is represented with a red line. P-values for these interaction tests are given in Table 2.

3. Results

3.1. *Local ancestry transitions interact with variants to influence GWAS Catalog traits*

Using the GWAS Catalog,[18] we identified phenotypes that had previous associations to regions fine-mapped by the Metabochip. We analyzed Metabochip regions with at least 100 local ancestry transitions to provide power to detect the interaction of these transitions with risk alleles. We only considered associations that made reference to African ancestry in the study sample, were for phenotypes that could be readily derived from the EHR, and had at least 200 cases/values in the EHR. This resulted in 28 regions (Table 1), and a total of 57 trait-region pairs.

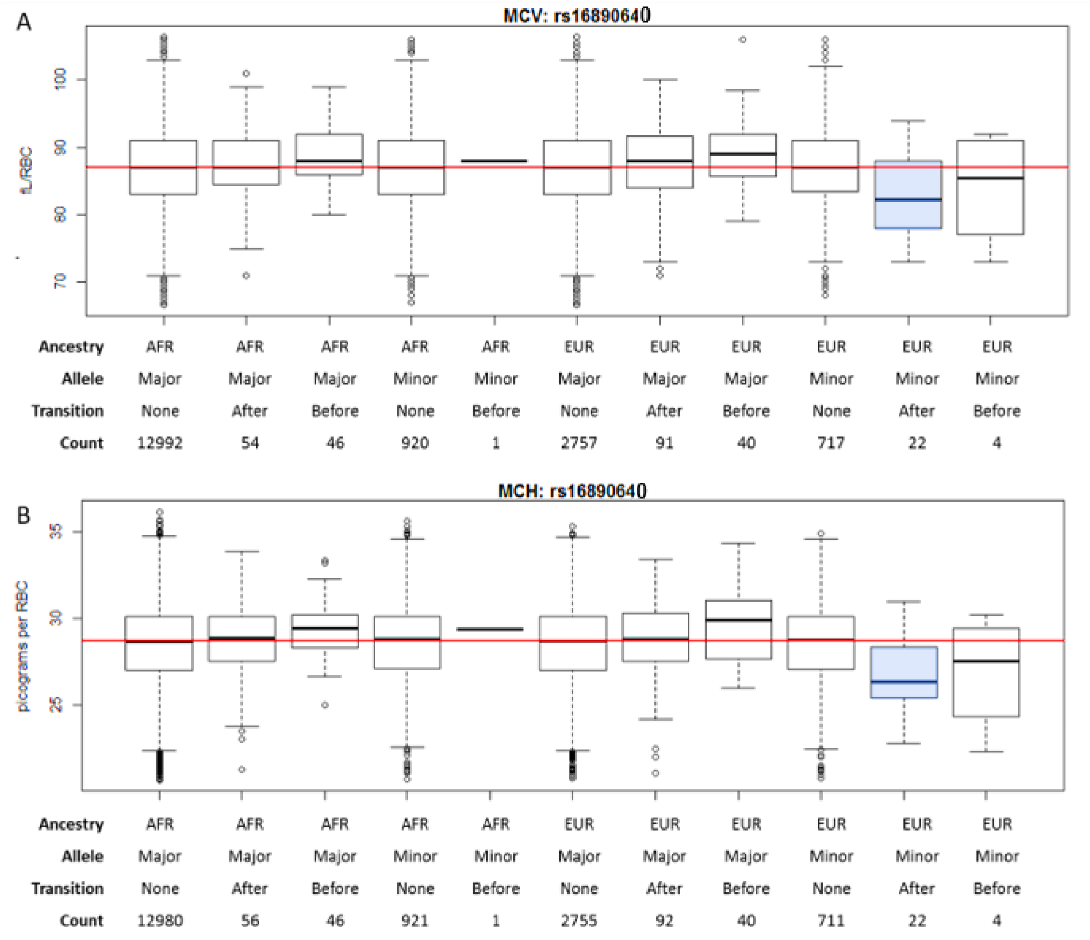

Figure 2. Chromosomes with the minor allele for rs16890640 on a European background and a downstream local ancestry transition are associated with lower MCV (A) and MCH (B). P-values for these interaction tests are given in Table 2.Only one chromosome category was significant with multiple testing corrections for each pairwise test: the minor allele of rs16890649, on a European ancestry, with a downstream local ancestry transition for MCH (p = 0.0024, shown in blue).

Due to both differential linkage disequilibrium (LD) structure between populations and differences in SNP coverage in the original GWAS reporting these associations, we tested all variants within the Metabochip region for association with the phenotype. For each trait-Metabochip region pair, there was at least one nominal genetic association suggesting genetic association within the region. We next evaluated the impact of local ancestry transitions on regional SNP-trait associations.

We identified five significant interactions between local ancestry transitions and the allele across all traits, using a Bonferroni multiple testing correction for all tests within each Metabochip region (adj *P* < 0.05) (Table 2). We then characterized these interactions, grouping chromosomes together based on their local ancestry, allele, and local ancestry transition status. For two interactions, significance is driven by chromosomes that are rarely observed. The interaction for rs9467458 with creatinine levels ($P = 6.90\text{E}^{-11}$, Figure 1A) is driven by a single chromosome of European ancestry containing the minor allele, and having a downstream local ancestry transition. The interaction for variant rs4712930 regulating white blood cell (WBC) counts ($P = 8.99\text{E}^{-05}$, Figure 1B) is attributable to three chromosomes with African ancestry containing the minor allele, and having an upstream ancestry transition. While these may represent real genetic effects, given the small number of observations within our data, we did not investigate these associations further. The interaction between the variant rs1410438 and CO_2 levels ($P = 7.83\text{E}^{-05}$, Figure 1C) shows an interesting pattern in which the minor allele on either ancestral background in the context of an ancestry switchpoint is associated with higher CO_2 levels. This result points to an effect of this region, but because there is no clear biological influence of the ancestry swichpoint, this region was not further investigated.

Table 2. Significant trait associations (all within chr6:25235303-26141375 region) showing SNP by ancestry transition interactions.

Trait	Index Variant	Allele p	Switchpoint p	Local Ancestry p	Switch x Allele p
Creatinine	chr6:25399755	0.759601	2.92E-10	0.639942	6.90E-11
WBC	chr6:25511686	0.170542	0.11056	0.892657	8.99E-05
CO2	chr6:25524034	0.169619	0.048373	0.234598	7.83E-05
MCV	chr6:25577310	0.218143	0.002092	0.153851	7.12E-05
MCH	chr6:25577310	0.071935	0.002658	0.209261	1.24E-05

The two remaining significant interactions for rs16890640 and mean corpuscular hemoglobin (MCH) and mean corpuscular volume (MCV) are illustrated in Figure 2. Both have low numbers of chromosomes in a category; however, these low frequency categories closely resemble the sample median and do not drive the significance of the effect. To identify the categories driving the interactions, we compared each category against the rest of the population with a Mann-Whitney U test. Only one chromosome category was significant after multiple testing correction for each pairwise test; for both traits the interaction is predominantly driven by the minor allele of rs16890640 on a background of European ancestry, with a downstream transition to African ancestry ($p = 0.0024$). A composite of Manhattan plots in the context of local ancestry transitions for MCH associations on this chromosome 6 region is shown in Figure 3.

Mean corpuscular hemoglobin (MCH) and mean corpuscular volume (MCV) are highly correlated with one another, and consequently, the interactions strongly resemble one another. We investigated

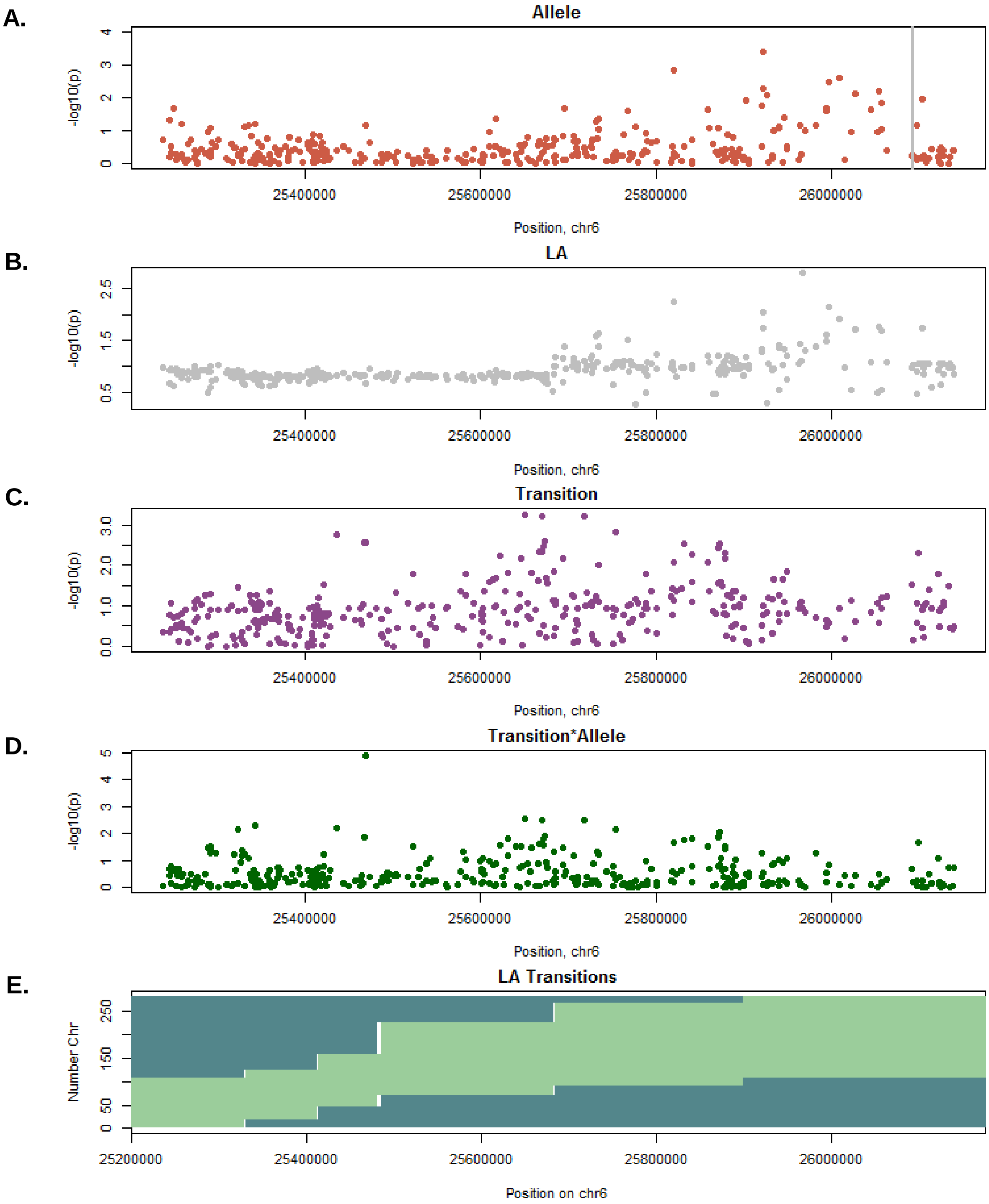

Figure 3. The association of genetic variants within chr6:25235303-26141375 with MCH interacts with local ancestry. Manhattan plots for the effect of the allele (A), local ancestry (B), presence of a local ancestry transition in the region (C), and an interaction between the allele and local ancestry transition (D) – please note differences in scale. The specific local ancestry transitions observed in this region are shown in panel E. Dark green indicates European ancestry along the chromosome; light green indicates African ancestry. rs1800562 (location marked by gray line in panel (A) has been associated with a variety of iron-related phenotypes.

the impact of the downstream transition to African ancestry by stratifying the 22 individuals from the significant chromosome category based on the location of their local ancestry transition. Local ancestry transitions occurred at three downstream points (Figure 3E). We observed a position-dependent effect wherein individuals with a transition to African ancestry at the point nearest to the variant (chr6:25481231) had lower MCH levels (Figure 4). Individuals with transitions to African ancestry at the two subsequent points began to approach the median MCH level. This suggests that the putative functional element interacting with rs16890640 is located between rs16890640 and the second transition point.

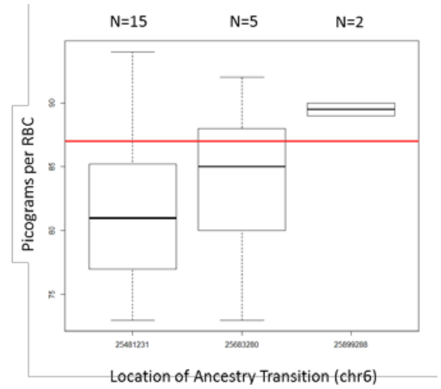

Figure 4. The effect of downstream local ancestry transitions on MCH is position-dependent. Median MCH level is shown in red. The number of chromosomes with European ancestry is provided above each boxplot.

3.2. *Local Ancestry Transitions May Affect Chromatin Looping*

To further characterize rs16890640 and explore potential biological mechanisms mediating its detected interaction with downstream local ancestry transitions, we analyzed its frequency in different populations and genomic context. rs16890640 is roughly three times more common in European-descent populations (EUR = 21%) than it is in African-descent populations (AFR = 8%) based on 1000 Genomes Project Phase 3 frequencies.[20] It occurs within an intron of *CARMIL1 (LRRC16A),* a cytoskeleton-associated protein involved in regulation of actin polymerization and in megakaryocyte development and platelet production (Reactome Pathway R-HAS-983231). rs16890640 falls within an observed binding site for MAFK, a transcription factor relevant to hemoglobin phenotypes (Figure 5); knock out of MAFK in mice results in reduced MCV and MCH levels.[21] Additionally, it is less than 500 base pairs upstream of a predicted insulator element; however, chromatin looping patterns indicate that contacts occur on either side of this putative insulator (Figure 6). Thus, rs16890640 occurs within a plausibly relevant genomic-region, and is more frequent in Europeans.

We also identified a relatively close (within 20 kb) GWAS-catalog variant associated to a related phenotype, serum transferrin levels (i.e., the amount of glycoproteins that bind free iron).[22] Notably, this variant (rs2274089) is flanked by the two recombination peaks that could result in a local ancestry transition in the area of interest (Figure 5). The first of these recombination peaks is observed in both European (CEU) and African (YRI) descent populations;

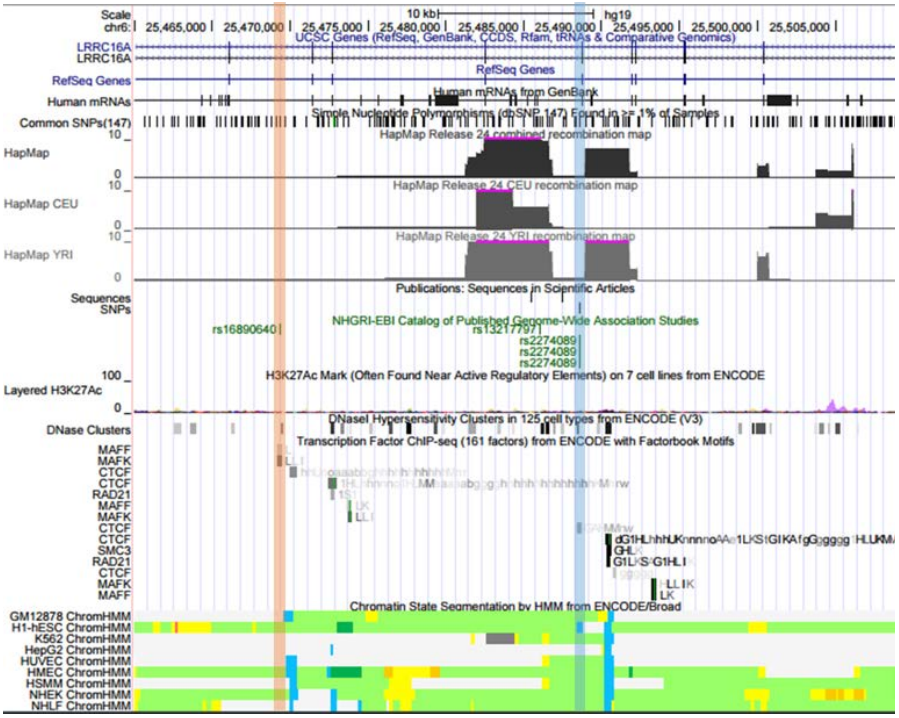
Scale
chr6:
10 kb
hg19
25,465,000
25,470,000
25,475,000
25,480,000
25,485,000
25,490,000
25,495,000
25,500,000
25,505,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
LRRC16A
LRRC16A
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Simple Nucleotide Polymorphisms (dbSNP 147) Found in >= 1% of Samples
Common SNPs(147)
HapMap Release 24 combined recombination map
HapMap
HapMap Release 24 CEU recombination map
HapMap CEU
HapMap Release 24 YRI recombination map
HapMap YRI
10
0
Publications: Sequences in Scientific Articles
Sequences
SNPs
NHGRI-EBI Catalog of Published Genome-Wide Association Studies
rs16890640
rs13217797
rs2274089
rs2274089
rs2274089
100
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
MAFF
MAFK
CTCF
CTCF
RAD21
MAFF
MAFK
CTCF
CTCF
SMC3
RAD21
CTCF
MAFK
MAFF
Chromatin State Segmentation by HMM from ENCODE/Broad
GM12878 ChromHMM
H1-hESC ChromHMM
K562 ChromHMM
HepG2 ChromHMM
HUVEC ChromHMM
HMEC ChromHMM
HSMM ChromHMM
NHEK ChromHMM
NHLF ChromHMM

Figure 5. Ancestry-specific recombination hotspots may disrupt functional elements pertinent to MCH and MCV. rs16890640 (orange line) is located within MAFF and MAFK binding sites, and is approximately 500 bp upstream of a predicted insulator element. rs16890640 interacts with a downstream local ancestry transition occuring at one of the two African-specific -recombination hotspots shown here, and is proximal to rs2274089 (blue line), a GWAS catalog variant for related traits which overlaps an insulator element.

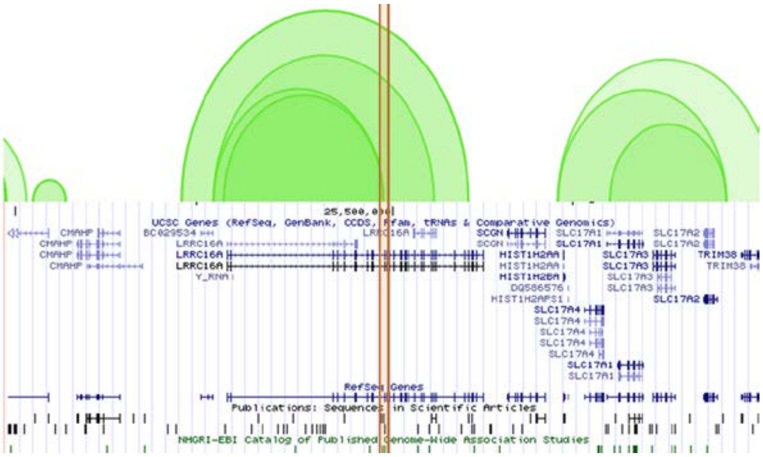

Figure 6. Local ancestry transitions may perturb regional chromatin looping patterns. The African-specific recombination peak region physically interacts with the *CARMIL1 (LRRC16A)* promoter based on ChIA-PET data for RAD21 in the GM12878 cell line. The GWAS variant rs2274089, associated with a relevant phenotype, is highlighted in orange.

however, the second recombination peak is African-specific. This African-specific recombination peak overlaps a ChromHMM predicted insulator element (Figure 5), based largely on the presence of CTCF binding. Chromatin looping data suggest this region may function as an enhancer: the region contacts the *CARMIL1 (LRRC16A)* promoter in GM12878 (Figure 6). Regardless of whether the region is an enhancer or insulator, it is engaged in regulatory chromatin looping, and we propose that the downstream transition to African ancestry introduces a haplotype that alters the regional chromatin conformation to modify the effect of rs16890640 on MCH and MCV levels.

4. Conclusion

In this study, we hypothesized that the recombination of historically isolated ancestral haplotypes in admixed populations would result in unique combinations of genetic variants that, since they have not been subject to evolutionary pressures, are more likely to influence phenotypes relevant to human health. We investigated this hypothesis in almost ten thousand adults of African ancestry, with both EHR-derived phenotypes and genetic data from targeted regions on the Metabochip. We identified a compelling example that suggests that combinations of haplotypes from different continental ancestries may interact with one another to influence hematological traits in humans.

Chromatin looping is one biological mechanism that may explain our observed statistical interaction between a SNP and a downstream transition to African ancestry. Chromatin looping establishes "domains" in which gene regulatory activity can be confined, keeping the promoters and enhancers for one gene from influencing another. It is possible that the African-specific recombination hotspot (which overlaps putative insulator elements) has introduced low-frequency genetic variants on African-descent haplotypes that alter the insulator's function or epigenetic state. With loss of the insulator, the regulatory variant rs16890640 is then able to engage in 'off-target' effects, which ultimately reduce MCH and MCV levels. Alternatively, an African haplotype may be simply carrying another functional variant that interacts with rs16890640 to influence MCH and MCV, regardless. Ultimately, molecular validation will be required to discern between possibilities.

This approach of examining the modifying role of local ancestry in single variant association studies has several limitations. First, the resolution of local ancestry transitions is limited by the density and proximity of variants along the chromosome that are captured by the genotyping array or imputation. Secondly, we collapsed all local ancestry transitions into a single variable, regardless of where the transition occurred within the region. This introduces additional noise within the data, as not all transitions may have the same effect. Furthermore, by examining each phased chromosome separately, we do not capture *trans* effects between them. In the future, additional models of local ancestry and variants may address these limitations.

The interaction we identified provides potential evidence for epistasis influencing health-related phenotypes in humans. The variant rs16890640 is not significantly associated with the phenotype on its own – it is only in combination with the downstream transition to African ancestry that an association to the phenotype is observed. While it is possible that this haplotype conformation tags a causal variant within this region, we consider this unlikely as nearby variants did not demonstrate a strong association. Instead, it highlights that admixed populations provide a unique opportunity to investigate epistasis, as novel combinations of variants are generated and population-specific recombination hotspots may disrupt functional haplotypes. This further highlights the need to perform genetic studies within admixed populations (as the variants may not have an effect in either continental population) to address health disparities and the role of epistasis in human health.

5. Acknowledgements

Admixture methods development work was supported by the NIH grants RF1 AG054074 and U01 AG052410. Metabochip data generation was supported by NIH U01 HG004798 and its ARRA supplements. The dataset(s) were obtained from Vanderbilt University Medical Center's BioVU, supported by institutional funding Vanderbilt CTSA grant UL1 TR000445 from NCATS/NIH.

6. References

1. Consortium, T. 1000 G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **135,** 0–9 (2012).
2. Hinds, D. A. *et al.* Whole-Genome Patterns of Common DNA Variation in Three Human Populations. *Science (80-.).* **307,** 1072 LP-1079 (2005).
3. Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476,** 170–175 (2011).
4. Doetschman, T. Influence of Genetic Background on Genetically Engineered Mouse Phenotypes. *Methods Mol. Biol.* **530,** 423–433 (2009).
5. Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E. T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet.* **7,** e1002144 (2011).
6. Lappalainen, T., Montgomery, S. B., Nica, A. C. & Dermitzakis, E. T. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am. J. Hum. Genet.* **89,** 459–63 (2011).
7. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* **24,** 1–13 (2014).
8. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).
9. Roden, D. M. *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* **84,** 362–9 (2008).
10. Pulley, J., Clayton, E., Bernard, G. R., Roden, D. M. & Masys, D. R. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin. Transl. Sci.* **3,** 42–8 (2010).
11. Dumitrescu, L. *et al.* Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. *Genet. Med.* **12,** 648–50 (2010).
12. Hall, J. B., Dumitrescu, L., Dilks, H. H., Crawford, D. C. & Bush, W. S. Accuracy of administratively-assigned ancestry for diverse populations in an electronic medical record-linked biobank. *PLoS One* **9,** e99161 (2014).
13. Crawford, D. C. *et al.* Leveraging Epidemiologic and Clinical Collections for Genomic Studies of Complex Traits. *Hum. Hered.* **79,** 137–46 (2015).
14. Voight, B. F. *et al.* The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genet.* **8,** 1–12 (2012).
15. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Meth* **10,** 5–6 (2013).
16. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am. J. Hum. Genet.* **93,** 278–288 (2013).
17. Kho, A. N. *et al.* Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J. Am. Med. Inform. Assoc.* **19,** 212–8 (2012).
18. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001-6 (2014).
19. Goodloe, R. J., Farber-Eger, E., Boston, J., Crawford, D. C. & Bush, W. S. Reducing clinical noise for body mass index measures due to unit and transcription errors in the electronic health record. *AMIA Jt Summits Transl Sci Proc* (2017).
20. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
21. Onodera, K., Shavit, J. A., Motohashi, H., Yamamoto, M. & Engel, J. D. Perinatal synthetic lethality and hematopoietic defects in compound mafG::mafK mutant mice. *EMBO J.* **19,** 1335–1345 (2000).
22. Benyamin, B. *et al.* Identification of novel loci affecting circulating chromogranins and related peptides. *Hum. Mol. Genet.* **26,** 233–242 (2017).