**Machine learning reveals 3D regulatory mechanisms for height-associated haplotypes**

**Wanjun Gu, Erin N. Gilbertson, Rany S. Salem, John A. Capra**

**Background**

Variants associated with phenotypes in genome-wide association studies (GWAS) are predominantly non-protein-coding and regulatory in nature. There are many ways a variant could disrupt gene regulation, and one emerging mechanism is by perturbing the 3D genome, thereby affecting the gene expression of target genes. Some phenotypes, like body height, have enriched SNP heritability within topologically associated domains (TADs). Although it has been challenging to experimentally evaluate variants for effects on 3D contacts at genome-scale, the 3D genome structure can be predicted in silico using machine learning models based solely on DNA sequence information. This provides an opportunity to evaluate 3D genome disruption as a mechanism underlying height-associated loci genome-wide.

**Methods**

We analyzed genome-wide significant regions associated with body height (top p-value ≤ 1e-5) from the largest available GWAS. To enable haplotype-aware analyses, we used the NHLBI Trans-Omics for Precision Medicine (TopMed) sequencing data for ~50,000 participants to impute haplotypes for these significant loci across diverse populations, including Europeans, Africans, East Asians, South Asians, and Admixed/non-admixed Americans. We then predicted alterations in the 3D genome contacts caused by each common haplotype (count ≥ 30 in TopMed).

**Results**

We evaluated 9917 height-associated regions: 107 regions (top 1%) exhibited substantial divergence, and 17 (top 0.17%) demonstrated extreme disturbance of the 3D genome. The strongest divergence for a height-associated haplotype was near the *LCOR* gene on chromosome 10. A specific variant at this locus, rs7477274, likely disrupts 3D genome folding by altering the DNA-binding affinity of the CTCF transcription factor. In addition to LCOR, other significant disruptions were observed near the SLC41A2 and FGF2 regions, both of which are CTCF binding sites and expression quantitative trait loci.

**Conclusion**

We identify several haplotypes that likely influence variation in body height by modifying 3D genome folding. However, this functional mechanism is relatively rare among height GWAS hits. Our results demonstrate how in-silico mutagenesis based on powerful sequence-based machine learning models provides an efficient approach to fine-map GWAS signals and identify potentially functional variants and mechanisms.