High-altitude Natural Selection Near Transcription Factor Footprints

Wanjun Gu

**Problem Statement and Hypothesis**

A comprehensive map of transcription factor (TF) footprints across the human genome has been created by Vierstra et al(1). Building upon the findings of Vierstra et al. that TFs occupy regions of DNA that are more mutable and likely to associate with traits, this research seeks to determine if such TF-occupied areas are also hotspots for genetic adaptations in high-altitude environments. Particularly, we examined a group of Andean highlanders whose ancestors have been living at high altitudes (>4200m) for hundreds of generations(2). In this study, we hope to show the presence of natural selection signals near these TF footprints but also to explore the specific biological processes and genes that are highlighted in these regions.

**Methods**

DNase I hypersensitivity site information and transcription factor footprint region information were downloaded from the supplementary data provided by Vierstra et al. and converted to human genome build hg19. The population genetics selection scan was done on the high-coverage (30x) whole genomes of 40 native Andean highlanders living in Cerro de Pasco, Peru (~ 4315 m). The selection scan combined the summary statistics of five different selection scan methods (FST, ΔDAF, iHS, ΔiHH, and XP-EHH) to obtain a Bayes factor () indicating the posterior probability of a certain genomic region under positive selection(3). An enrichment test was conducted on all the Andean genome markers with greater than zero. The test compares the abundance of genetic markers within the DNase I hypersensitivity sites (DHSs) and the TF footprint regions over the baseline abundance, which is the abundance of markers under null selection ( between 0 and 1) near DHS and TF footprint regions. Significant genetic markers from the selection scan were then annotated with non-pseudo genes that are within 100kb upstream and downstream. Biological pathway analysis was conducted using Cytoscape(4) and BiNGO(5) on regions demonstrating strong evidence of natural selection and overlapping with TF footprints.

**Results**

As shown in Figure XA, Markers with stronger evidence of natural selection tend to be enriched in DNase I hypersensitivity sites (DHS) and TF footprints. The enrichment of these markers becomes statistically significant in both DHS (p < 1.84E-5) and TF footprints (p < 8.39E-5) when these markers have greater than 12. All genetic markers from the selection scan with greater than 12 were then overlapped with TF footprint regions and annotated for non-pseudo genes. All genes in these regions under selection were used to conduct a biological pathway analysis. As shown in Figure XB, the pathway analysis highlights the role of DNA regulation, replication, and repair among genes in TF footprint regions that are also under natural selection in high-altitude conditions. In Figure XC, the pathways related to DNA regulation, replication, and repair are expanded to their third-degree neighbor and visualized in a network diagram. Among the top selected gene regions, it is worth pointing out that several genes related to the immune system and inflammation are underlined, including *IL17A* and *IL17F*. In addition, genes related to cell cycle and DNA replication (*MCM3*) as well as DNA Repair and Maintenance (*UNG* and *ALKBH2*) are also highlighted in this analysis.

**Discussion**

*Enrichment of Genetic Loci Near DHS and TF Footprints*

This study discovered a significant enrichment of genetically selected loci near DNAse I hypersensitivity sites (DHSs) and transcription factor (TF) footprints in high-altitude environments. This finding suggests that natural selection in these areas is not random but targeted at locations crucial for gene regulation. The proximity of these loci to TF footprints, which are direct sites for transcription factor binding, indicates an evolutionary pressure on gene regulation mechanisms, potentially improving the adaptability and survival of individuals living at high altitudes.

*Biological Pathways Related to DNA Regulation, Replication, and Repair*

Important biological pathways, including DNA regulation, replication, and repair, were identified as crucial in maintaining genomic integrity and responding to environmental stressors like low oxygen levels and increased UV radiation. This study posits that these pathways have evolved under selective pressure in high-altitude conditions, enabling Andean highlanders to better maintain cellular and genomic stability in such challenging environments.

*Highlighted Gene Adaptations and Their High-Altitude Implications*

The study particularly notes adaptations in genes like *IL17A*, *IL17F*, *MCM3*, *UNG*, and *ALKBH2*. *IL17A* and *IL17F* are linked to immune response and inflammation, suggesting an evolutionary advantage for enhanced immunity against respiratory infections common at high altitudes(6). *MCM3*, involved in DNA replication, indicates an adaptation for improved cellular proliferation and repair under hypoxic conditions(7). *UNG* and *ALKBH2*, essential for DNA repair, suggest an adaptation mechanism for increased UV-induced DNA damage at high altitudes, thereby reducing mutation risks and maintaining genomic integrity(8).

*Study Limitations and Future Directions*

A primary limitation of this research is the insufficient control for minor allele frequency (MAF) in the enrichment test, which could affect the interpretation of genetic markers' significance. Future studies should extend the enrichment analysis to different high-altitude cohorts and populations under various environmental pressures to validate and expand upon these findings. This approach would strengthen the evidence for natural selection's role in shaping genetic adaptations and enhance understanding of the environmental influences on human genetic evolution.

**Conclusion**

In conclusion, the study reveals a selective preference for regions near TF footprints in high-altitude Andean populations, primarily involving genes crucial for DNA replication, repair, and maintenance. This emphasizes the adaptive importance of gene regulation in response to unique environmental challenges faced at high altitudes.

**Figure Caption**

Figure X. Enrichment of high-altitude natural selection signals near transcription factor footprints. Fold enrichment compared to variants demonstrating null selection ( between 0 and 1) are plotted against in XA. The red dotted line indicates the baseline fold enrichment of 1. Variants within the bins that are both enriched in DNase I hypersensitivity sites and transcription factor footprints are denoted with \*. In XB, top 20 biological pathways overrepresented by genes within regions of strong evidence of natural selection are tabulated and ranked by their P-Value corrected for false discovery rate. In XC, the pathways associated with DNA regulation, replication, and repair have been extended to include their third-degree connections and are depicted in a network diagram. The sizes of the nodes indicate the number of study genes present in the pathway (bigger nodes indicate more genes) and the color of the nodes is associated with false-discovery-rate corrected P-Values (darker indicates more significant).

Reference:

1. Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Haugen E, et al. Global reference mapping of human transcription factor footprints. *Nature* (2020) 583:729–736.

2. Simonson TS. Altitude Adaptation: A Glimpse Through Various Lenses. *High Alt Med Biol* (2015) 16:125–137.

3. Bigham A, Bauchet M, Pinto D, Mao X, Akey JM, Mei R, Scherer SW, Julian CG, Wilson MJ, López Herráez D, et al. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet* (2010) 6:e1001116.

4. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* (2003) 13:2498–2504.

5. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* (2005) 21:3448–3449.

6. McGeachy MJ, Cua DJ, Gaffen SL. The IL-17 Family of Cytokines in Health and Disease. *Immunity* (2019) 50:892–906.

7. Semenza GL. The Genomics and Genetics of Oxygen Homeostasis. *Annu Rev Genomics Hum Genet* (2020) 21:183–204.

8. Marcinkowski M, Pilžys T, Garbicz D, Steciuk J, Zugaj D, Mielecki D, Sarnowski TJ, Grzesiuk E. Human and Arabidopsis alpha-ketoglutarate-dependent dioxygenase homolog proteins-New players in important regulatory processes. *IUBMB Life* (2020) 72:1126–1144.